Discussion for Question 1

Link: https://www.examtopics.com/discussions/amazon/view/43814-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 78 votes
- C: 29 votes

Discussion

Comment: Should it be C? Cost incurred by the company as a result of false positives (predicted churn, actual not churn) is less than the false negatives (predicted not churn, actual churn). Incentive cost < churn cost.

Replies:

Comment: The question says "the cost of churn is far greater than the cost of the incentive", so we want to identify all the true churns, in order to do something about it. We don't want there to be any true churns we didn't see. This means we want false negatives as low as possible. So we want false negatives < false positives. So A.

Replies:

Comment: Exactly, Count of False Positive should be greater than count of False Negative. In other words, cost / penalty for company is more when False Negative are predicted. So, Answer C - Cost incurred by the company as a result of False Positives is less than the False Negatives.

Replies:

Comment: prefect explaination

Comment: No, the text clearly says the cost of churn is "far" greater - not equal to. One incident of churn could be higher than 10 incidents of incentive. Ans = A

Replies:

Comment: According to your logic the answer is C. A false positive is no churn, a false negative is churn. So false negatives are the thing to avoid and are most expensive, hence, a false positive costs less than a false negative.

Comment: Fully Agree: FN = Predict Not churn, actual churn, high cost FP = Predict churn, actual not churn, pay incentive, low cost. so FP < FN, The answer is C.

Comment: The Answer is A. Reasons: 1. accurate is 86% 2. FN=4, FP= 10. The question is asking why this is a feasible model which means why this is working. So it is not asking the explaination of the unit cost of churn(FN) is greater than cost of incentive(FP). It is asking from the matrixs result, the number it self, FN(4) is less than FP(10). The model successfully keep a smaller number of FN regarding of FP.

Comment: A is correct answer

Comment: FN has a higher cost than FP, so A is a better choice than C.

Comment: Should be A. Since the cost of churn is much higher, the priority should be focused on minimizing FN and a viable model should be one with FN < FP, isn't it?

Comment: A) Because FN = 4 < FP = 10. FN are missed churns, and FP is misidentified churns.

Comment: A is the correct answer

Comment: cost of churn (churn cost) is greater than the cost of incentive (customers who do not churn)... the model predicts more false positives (customers who do not churn) than false negatives (customers who churn), Therefore, the costs of false negatives are greater than the costs of false positives, as churn is more expensive.

Comment: FN < FP

Comment: The question says "the cost of churn is far greater than the cost of the incentive", so we want to identify all the true churns, in order to do something about it. We don't want there to be any true churns we didn't see. This means we want false negatives as low as possible. So we want false negatives < false positives and we get exactly that in the model. Now this fact coupled with the fact that incentives are welcome rather than churn, in other words, cost / penalty for company is more when False Negative are predicted. So, Answer C - Cost incurred by the company as a result of False Positives is less than the False Negatives.

Comment: The closest answer to this rationale is: A. The model is 86% accurate and the cost incurred by the company as a result of false negatives is less than the false positives. Despite the answer options not matching the typical calculations of accuracy and precision, option A seems to be the most aligned with the company's goals if we consider the cost implications as more significant than the accuracy metrics alone. The company prefers a model has higher Recall score (10/14 this case 71.4%) than Precision score (10/20 this case 50%).

Comment: Cost incurred by the company is directly proportional to cost of churn which is directly proportional to number of false negatives. False positives are more acceptable than false negatives in this case.

Comment: accuracy is 86% so A or C. The cost of losing a customer is very high. Thus we do not want False Negatives (we do not want to predict no churn when there is churn). Thus the cost of a false positive is less than a false negative. Answer C

Comment: Will go with C. My opinion is the same as brunokiyoshi

Comment: Cost FN > cost FP so want to minimize FN

Comment: Precision is 50%, so B&D are wrong. Accuracy is 86% which left A&C FP is 10 & FN is 4 which mean A will be the right answer. https://dataaspirant.com/wp-content/uploads/2020/08/3_confusion_matrix.png

Comment: I passed today Sept 21, 2023.. score of 900.. With Very little experience in ML as I am in mgmt but prior dev experience as Architect and sys Admin.. Thanks to all contributors.. Make sure to focus on every Q here as there's more than 70% from here.. Also look at Udemy dumps and course on Udemy to study.

Discussion for Question 2

Link: https://www.examtopics.com/discussions/amazon/view/11248-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 55 votes

Discussion

Comment: B see https://en.wikipedia.org/wiki/Collaborative_filtering#Model-based

Comment: Content-based filtering relies on similarities between features of items, whereas colloborative-based filtering relies on preferences from other users and how they respond to similar items.

Comment: B is correct answer

Comment: B is correct answer

Comment: top it exam com

Comment: 'Collaborative filtering is a technique that can filter out items that a user might like on the basis of reactions by similar users.' Source: https://realpython.com/build-recommendation-engine-collaborative-filtering/#what-is-collaborative-filtering

Comment: A. NO - content-based filtering looks at similarities with items the user already looked at, not activities of other users B. YES - state of the art C. NO - too generic terms, everything is a model D. NO - combinative filtering does not exist

Comment: Collaborative filtering is a technique used by recommendation engines to make predictions about the interests of a user by collecting preferences or taste information from many users. The underlying assumption of the collaborative filtering approach is that if a person A has the same opinion as a person B on an issue, A is more likely to have B's opinion on a different issue than that of a randomly chosen person.

Comment: B. Build a collaborative filtering recommendation engine with Apache Spark ML on Amazon EMR.

Comment: I think it should be b

Comment: Content-based recommendations rely on product similarity. If a user likes a product, products that are similar to that one will be recommended. Collaborative recommendations are based on user similarity. If you and other users have given similar reviews to a range of products, the model assumes it is likely that other products those other people have liked but that you haven't purchased should be a good recommendation for you.

Comment: B. Build a collaborative filtering recommendation engine with Apache Spark ML on Amazon EMR. Collaborative filtering is a technique used to recommend products to users based on their similarity to other users. It is a widely used method for building recommendation engines. Apache Spark ML is a distributed machine learning library that provides scalable implementations of collaborative filtering algorithms. Amazon EMR is a managed cluster platform that provides easy access to Apache Spark and other distributed computing frameworks.

Comment: Build a collaborative filtering recommendation engine with Apache Spark ML on Amazon EMR. (TRUE) Collaborative filtering is a commonly used method for recommendation systems that aims to predict the preferences of a user based on the behavior of similar users. In the case described, the objective is to use users' behavior and product preferences to predict which products they want, making collaborative filtering a good fit. Apache Spark ML is a machine learning library that provides scalable, efficient algorithms for building recommendation systems, while Amazon EMR provides a cloud-based platform for running Spark applications. You can find more detail in https://www.udemy.com/course/aws-certified-machine-learning-specialty-2023

Comment: feature engineering is required, use model based

Comment: Answer is "B"

Comment: go for B

Comment: B is correct https://aws.amazon.com/blogs/big-data/building-a-recommendation-engine-with-spark-ml-on-amazon-emr-using-zeppelin/

Discussion for Question 3

Link: https://www.examtopics.com/discussions/amazon/view/8303-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 51 votes
- D: 20 votes

Discussion

Comment: Answer is B https://github.com/ecloudvalley/Building-a-Data-Lake-with-AWS-Glue-and-Amazon-S3

Replies:

Comment: you cannot use AWS glue for streaming data. Clearly B is incorrect.

Replies:

Comment: Even if the exam's answer is based on solution before AWS implemented the capability of AWS glue to process streaming data, this answer is still correct as Kinesis would output the data to S3 and Glue will pick it up from there and covert to parquet. Question does not say data must be converted to parquet in real time, it only says the csv data is received as a stream in real time.

Replies:

Comment: Actually question says "The source systems send data in CSV format in real time The Data Engineering team wants to transform the data to the Apache Parquet format before storing it on Amazon S3" same as saying data must be converted real time

Comment: AWS Glue can do it now (2020 May) https://aws.amazon.com/jp/blogs/news/new-serverless-streaming-etl-with-aws-glue/

Replies:

Comment: This link is in Japanese

Comment: the Approve Of B https://aws.amazon.com/blogs/aws/new-serverless-streaming-etl-with-aws-glue/

Comment: D is wrong as kinesis firehose can convert from JSON to parquet but here we have CSV. B is correct and here is another proof link: https://medium.com/searce/convert-csv-json-files-to-apache-parquet-using-aws-glue-a760d177b45f

Replies:

Comment: https://docs.aws.amazon.com/firehose/latest/dev/record-format-conversion.html You are right. https://docs.aws.amazon.com/firehose/latest/dev/record-format-conversion.html If you want to convert an input format other than JSON, such as comma-separated values (CSV) or structured text, you can use AWS Lambda to transform it to JSON first

Replies:

Comment: But there is no Lambda in D

Replies:

Comment: But there's a D in Lambda

Comment: Between B and D chose D. Because Firehose can't handle csv directly.

Comment: Between B and D chose D. Because Firehose can't handle csv directly.

Replies:

Comment: Between B and D chose B. Because Firehose can't handle csv directly.

Comment: Answer is B. https://docs.aws.amazon.com/firehose/latest/dev/record-format-conversion.html "If you want to convert an input format other than JSON, such as comma-separated values (CSV) or structured text, you can use AWS Lambda to transform it to JSON first."

Comment: u need glue to convert to parquet

Comment: D for sure, Firehose can convert csv to parquet

Comment: Answer is unfortunately B. firehose cannot convert coma separated CSV to parquet directly.

Comment: b is not goog but - >given the context of "finding the solution that requires the least effort to implement," option D is the most suitable choice. Ingesting data from Amazon Kinesis Data Streams and using Amazon Kinesis Data Firehose to convert the data to Parquet format is a serverless approach. It allows for automatic data transformation and storage in Amazon S3 without the need for additional development or management of data conversion logic. Therefore, under the given conditions, option D is considered the solution that requires the "least effort" to implement

Replies:

Comment: Kinesis Data Firehose doesn't convert anything, it rather calls a lambda function to do so which is the overhead we want to avoid. B is the correct answer.

Comment: Amazon Kinesis Data Streams is a service that can capture, store, and process streaming data in real time. Amazon Kinesis Data Firehose is a service that can deliver streaming data to various destinations, such as Amazon S3, Amazon Redshift, or Amazon Elasticsearch Service. Amazon Kinesis Data Firehose can also transform the data before delivering it, such as converting the data format, compressing the data, or encrypting the data. One of the supported data formats that Amazon Kinesis Data Firehose can convert to is Apache Parquet, which is a columnar storage format that can improve the performance and cost-efficiency of analytics queries. By using Amazon Kinesis Data Streams and Amazon Kinesis Data Firehose, the Mobile Network Operator can ingest the .CSV data from the source systems and use Amazon Kinesis Data Firehose to convert the data into Parquet before storing it on Amazon S3

Replies:

Comment: Firehose cannot natively do the conversion. It requires a Lambda function for that purpose.

Comment: D. Ingest .CSV data from Amazon Kinesis Data Streams and use Amazon Kinesis Data Firehose to convert data into Parquet. This is because Amazon Kinesis Data Firehose has built-in support for converting incoming streaming data into different formats like Parquet before storing it in S3. This requires less setup and management compared to the other options, making it a low-effort solution.

Comment: chat gpt4 is 'D'

Comment: Answer is B. Since Glue can work with real time data and work with csv files directly and store them in S3 For option D it needs a lambda function in between to convert csv to json then store to S3

Comment: D is the most straightforward solution. Kinesis Data Firehose directly supports converting streaming data into Parquet format and requires the least amount of setup and operational overhead compared to the other options. It eliminates the need for server management and manual job scheduling, providing a more seamless and low-effort solution for real-time data ingestion and transformation.

Comment: B is the right option. Using firehose , we will need to write a lambda to do the conversion as firehose only converts from JSON to parquet. Where as Glue has inbuilt integration for doing this conversion. This answer is verified using the recently launched Amazon Q

Comment: kinesis data firehose supports data Conversions from CSV / JSON to Parquet / ORC (only for S3)

Comment: The least effort solution to implement this use case would be option D - Ingest .CSV data from Amazon Kinesis Data Streams and use Amazon Kinesis Data Firehose to convert data into Parquet. Amazon Kinesis Data Firehose is a fully managed service that automatically delivers real-time streaming data to destinations such as Amazon S3, Redshift and Elasticsearch. It can ingest streaming data from Kinesis Data Streams and convert the data format to Parquet before delivering to S3. This avoids the need to setup and manage infrastructure for data ingestion and transformation. Compared to the other options: A) Using Kafka Streams and Kafka Connect would require more effort to setup and manage the Kafka cluster on EC2. B) Using Glue to convert data would require developing a Glue ETL job which is more complex than just using Firehose transformation feature. C) Setting up an EMR cluster and developing Spark Structured Streaming job is more complex than the serverless Firehose service.

Discussion for Question 4

Link: https://www.examtopics.com/discussions/amazon/view/12382-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 13 votes

Discussion

Comment: answer should be C

Comment: go for C

Comment: Ans should be c

Comment: A. Managing Kafka on EC2 is not compatible with least effort requirement B. Doable (in 2024) as Glue supports streaming ETL to consumes streams and supports CSV records -> https://docs.aws.amazon.com/glue/latest/dg/add-job-streaming.html C. Managing an EMR cluster imo is no compatible with least effort requirement D. Firehose supports kinesis data stream as source and it can use lambda to convert CSV records into parquet -> https://docs.aws.amazon.com/firehose/latest/dev/record-format-conversion.html I guess this is a bit old question, pre Glue streaming ETL support (2023) -> https://aws.amazon.com/about-aws/whats-new/2023/03/aws-glue-4-0-streaming-etl/ Thus I'll go for D

Comment: This blog wrote Japanese. but its said using LinearLearner for air pollution prediction. https://aws.amazon.com/jp/blogs/news/build-a-model-to-predict-the-impact-of-weather-on-urban-air-quality-using-amazon-sagemaker/

Comment: A. NO - kNN is not forecasting, it is similarities B. NO - RCF is for anomality detection C. YES - Linear Regression good for forecasting D. NO - we don't want to classify

Comment: The reason for this choice is that the Linear Learner algorithm is a versatile algorithm that can be used for both regression and classification tasks1. Regression is a type of supervised learning that predicts a continuous numeric value, such as the air quality in parts per million2. The predictor_type parameter specifies whether the algorithm should perform regression or classification3. Since the goal is to forecast a numeric value, the predictor_type should be set to regressor.

Comment: The HyperParameter is . Either “binary_classifier” or “multiclass_classifier” or “regressor”., there is no classifier so the answer is C

Comment: Ans should be c

Comment: a kNN will require a large value of k to avoid overfitting and we only have 1 year's worth of data - kNNs also face a difficult time extrapolating if the air quality series contains a trend If we had assurances there is no trend in the air quality series (no extrapolation), and we had enough data, then kNN should beat a linear model ... I am inclined to go for C just going off of the cue that "only daily data from last year is available"

Replies:

Comment: Agree with you analysis, to further expand it: we don't have info about dataset features based on "only daily data from last year is available" this let me think we could be in a situation where our dataset is made up by timestamp and pollution_value so KNN would be pretty useless in this situation.

Comment: Random cut forests in timeseries are used for anomaly detection, and not for forecasting. KNN's are classification algorithms. You would use the Linear Learner as a regressor, since forecasting falls into the domain of regression.

Replies:

Comment: I mean, you could use KNN's for regression, but for forecasting I don't think so

Comment: KNN isn't for time series predicting, go for A!

Replies:

Comment: Im sorry, I wanted to say go for C!

Comment: Creating a machine learning model to predict air quality To start small, we will follow the second approach, where we will build a model that will predict the NO2 concentration of any given day based on wind speed, wind direction, maximum temperature, pressure values of that day, and the NO2 concentration of the previous day. For this we will use the Linear Learner algorithm provided in Amazon SageMaker, enabling us to quickly build a model with minimal work. Our model will consist of taking all of the variables in our dataset and using them as features of the Linear Learner algorithm available in Amazon SageMaker

Comment: Answer should be A. k-Nearest-Neighbors (kNN) algorithm will provide the best results for this use case as it is a good fit for time series data, especially for predicting continuous values. The predictor_type of regressor is also appropriate for this task, as the goal is to forecast a continuous value (air quality in parts per million of contaminants). The other options are also viable, but may not provide as good of results as the kNN algorithm, especially with limited data. using the Amazon SageMaker Linear Learner algorithm with a predictor_type of regressor, may still provide reasonable results, but it assumes a linear relationship between the input features and the target variable (air quality), which may not always hold in practice, especially with complex time series data. In such cases, non-linear models like kNN may perform better. Furthermore, the kNN algorithm can handle irregular patterns in the data, which may be present in the air quality data, and provide more accurate predictions.

Comment: Seen on Dec. 1 exam

Comment: Answer is "C" !!!

Comment: answer C

Discussion for Question 5

Link: https://www.examtopics.com/discussions/amazon/view/9818-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- D: 49 votes

Discussion

Comment: Why not D? When the data encrypted on S3 and SageMaker uses the same AWS KMS key it can use encrypted data there.

Replies:

> **Comment:** should be D
>
> **Comment:** Should be D. Use Glue to do ETL to Hash the card number
>
> **Comment:** Answer would be D

Comment: D is correct

Comment: A. NO - no need for custom encryption B. NO - IAM Policies are not to encrypt C. NO - launch configuration is not to encrypt D. YES

Comment: The reason for this choice is that AWS KMS is a service that allows you to easily create and manage encryption keys and control the use of encryption across a wide range of AWS services and in your applications1. By using AWS KMS, you can encrypt the data on Amazon S3, which is a durable, scalable, and secure object storage service2, and on Amazon SageMaker, which is a fully managed service that provides every developer and data scientist with the ability to build, train, and deploy machine learning models quickly3. This way, you can protect the data at rest and in transit.

Comment: I think d is correct

Comment: It's D, KMS key can be used for encrypting the data at rest!

Comment: agreed with D

Comment: IMHO, the problem with the question is that it is not clear whether the credit card number is used in the model. In that case discarding is never a good option. Hashing should be a safe option to keep it in the learning path

Comment: It's gotta be D but C is a clever fake answer. Use PCA to reduce the length of the credit card number? That's a clever joke, as if reducing the length of a character string is the same as reducing dimensionality in a feature set.

Comment: Can Glue do redaction?

Replies:

> **Comment:** Just have the Glue job remove the credit card column.

Comment: Encryption on AWS can be done using KMS so D is the answer

Comment: D is correct

Comment: D..KMS fully managed and other options are too whacky..

Comment: D is correct

Comment: Ans D is correct

Comment: is this really a viable reference? so misleading.

Comment: I think it is D as using PCA to reduce the length of credit card numbers does not seem viable

---

Discussion for Question 6

Link: https://www.examtopics.com/discussions/amazon/view/11559-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 18 votes

Discussion

Comment: I think the answer should be C

Comment: The correct answer HAS TO be A The instances are running in customer accounts but it's in an AWS managed VPC while exposing ENI to customer VPC if it was chosen. See explanation at https://aws.amazon.com/blogs/machine-learning/understanding-amazon-sagemaker-notebook-instance-networking-configurations-and-advanced-routing-options/

Replies:

> **Comment:** Can't be A because A says "but they run outside of VPCs", which is not correct. They are attached to VPC, but it can either be AWS Service VPC or Customer VPC, or Both, as per the explanation url you provided.
>
> ### Replies:
>
> > **Comment:** This is exactly right. According to that document, if the notebook instance is not in a customer VPC, then it has to be in the Sagemaker managed VPC. See Option 1 in that document.
>
> **Comment:** Actually your link says: The notebook instance is running in an Amazon SageMaker managed VPC as shown in the above diagram. That means the correct answer is C. An Amazon SageMaker managed VPC can only be created in an Amazon managed Account.

Comment: A. NO. If the EC2 instance of the notebook was in the customer account, customer would be able to see it. Also, "they run outside VPCs" isn't true as they run in service managed VPC or can be also attached to customer provided VPC -> https://aws.amazon.com/blogs/machine-learning/understanding-amazon-sagemaker-notebook-instance-networking-configurations-and-advanced-routing-options/ B. NO, Notebooks are based on EC2 + EBS C. YES -> https://aws.amazon.com/blogs/machine-learning/understanding-amazon-sagemaker-notebook-instance-networking-configurations-and-advanced-routing-options/ D. NO, Notebooks are based on EC2 + EBS I also actually tested it in my account: I created a Notebook and attached it to my VPC, I was not able to see the EC2 instance behind the Notebook but I was able to see the its ENI with the following description "[Do not delete] Network Interface created to access resources in your VPC for SageMaker Notebook Instance ..."

Comment: A. NO - AEC2 instances within the customer account are necessarily in a VPCb B. NO - Amazon ECS service is not within customer accounts C. YES - EC2 instances running within AWS service accounts are not visible to customer account D. NO - SageMaker manages EC2 instance, not ECS

Comment: already given below

Comment: I am pretty sure the answer is A : Amazon SageMaker notebook instances are indeed based on EC2 instances, and these instances are within your AWS customer account. However, by default, SageMaker notebook instances run outside of your VPC (Virtual Private Cloud), which is why they may not be visible within your VPC. SageMaker instances are designed to be easily accessible for data science and machine learning tasks, which is why they typically do not reside within a VPC. If you need them to operate within a VPC, you can configure them accordingly, but this is not the default behavior.

Comment: The explanation for this choice is that Amazon SageMaker notebook instances are fully managed by AWS and run on EC2 instances that are not visible to customers. These EC2 instances are launched in AWS-owned accounts and are isolated from customer accounts by using AWS PrivateLink1. This means that customers cannot access or manage these EC2 instances directly, nor can they see the EBS volumes attached to them.

Comment: I think it should be c

Comment: Per https://docs.aws.amazon.com/sagemaker/latest/dg/studio-notebooks-and-internet-access.html it's C

Comment: Notebooks can run inside AWS managed VPC or customer managed VPC

Comment: C, check the digram in https://docs.aws.amazon.com/sagemaker/latest/dg/studio-notebooks-and-internet-access.html

Comment: When a SageMaker notebook instance is launched in a VPC, it creates an Elastic Network Interface (ENI) in the subnet specified, but the underlying EC2 instance is not visible in the VPC. This is because the EC2 instance is managed by AWS, and it is outside of the VPC. The ENI acts as a bridge between the VPC and the notebook instance, allowing network connectivity between the notebook instance and other resources in the VPC. Therefore, the EBS volume of the notebook instance is also not visible in the VPC, and you cannot take a snapshot of the volume using VPC-based tools. Instead, you can create a snapshot of the EBS volume directly from the SageMaker console, AWS CLI, or SDKs.

Replies:

Comment: what you described is C "This is because the EC2 instance is managed by AWS, and it is outside of the VPC."

Comment: Notebooks run inside a VPC not outside!

Comment: Definitely C

Comment: Sagemaker notebook instances run in AWS sagemaker service accounts: https://docs.aws.amazon.com/sagemaker/latest/dg/studio-notebooks-and-internet-access.html Additionally, A is wrong because sagemaker instances run in a VPC (not outside)... It is just in aws service account

Comment: AWS Managed services do not happen outside the customer account. So the Sagemaker Managed Instance is primarily attached to a Sagemaker managed VPC inside a customer account. There has never been a hint of cross-account operations taking place in basic sagemaker operations. And this will not be the only AWS managed service situated inside a customer AWS account. A is the logical answer here. He cannot see the instance, because it will not show up in the EC2 console, especially if it is not attached to any customer VPC

Comment: C is correct

Discussion for Question 7

Link: https://www.examtopics.com/discussions/amazon/view/11560-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 11 votes

Discussion

Comment: Agreed. Ans is B

Comment: The reason for this choice is that Amazon CloudWatch is a service that monitors and manages your cloud resources and applications. It collects and tracks metrics, which are variables you can measure for your resources and applications1. Amazon SageMaker automatically reports metrics such as latency, memory utilization, and CPU utilization to CloudWatch2. You can use these metrics to monitor the performance and health of your SageMaker endpoint during the load test.

Comment: the question is clear that the specialist is seeking for latency, memory utilization, and CPU utilization during the load test and the ideal answer for all of these is amazon cloud watch which give you all these metrics https://docs.aws.amazon.com/sagemaker/latest/dg/monitoring-cloudwatch.html

Comment: the question is clear that the specialist is seeking for latency, memory utilization, and CPU utilization during the load test and the ideal answer for all of these is amazon cloud watch which give you all these metrics https://docs.aws.amazon.com/sagemaker/latest/dg/monitoring-cloudwatch.html

Comment: I think it should be b

Comment: It's B, even the resources that aren't visible in a first try are visible if you use cloudwatch agent.

Comment: Should be B

Comment: agreed with B

Comment: B is the ans

Comment: Should be C right, as Cloudwatch does not have metrics for memory utilization.

Replies:

Comment: After further research, I think answer is B. While indeed true that Cloudwatch does not have metrics for memory utilization by default, you can achieve by installing CloudWatch agent on the EC2. The EC2 used by Sagemaker is pre-installed with Cloudwatch Agent.

Comment: I do not think that CloudWatch, by default, logs memory utilization. It does log CPU utilization. If memory utilization is required, then a separate agent needs to be installed to watch for memory. Hence, in this case, we have to write an agent if the answer has to be B. Else, C looks to be a better solution.

Comment: answer is B

Comment: Answer is B 100%; very straightforward method

Comment: B is correct. Don't need to use Kibana or QuickSight.

Comment: ans is B

Comment: B is correct

Discussion for Question 8

Link: https://www.examtopics.com/discussions/amazon/view/11771-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 12 votes

Discussion

Comment: B is correct

Comment: The correct answer HAS TO be B Using Glue Use AWS Glue to catalogue the data and Amazon Athena to run queries against data on S3 are very typical use cases for those services. D is not ideal, Lambda can surely do many things but it requires development/testing effort, and Amazon Kinesis Data Analytics is not ideal for ad-hoc queries.

Comment: The reason for this choice is that AWS Glue is a fully managed service that provides a data catalogue to make your data in S3 searchable and queryable1. AWS Glue crawls your data sources, identifies data formats, and suggests schemas and transformations1. You can use AWS Glue to catalogue both structured and unstructured data, such as relational data, JSON, XML, CSV files, images, or media files2.

Comment: I think it should be b

Comment: AWS Glue is a fully managed ETL service that makes it easy to move data between data stores. It can automatically crawl, catalogue, and classify data stored in Amazon S3, and make it available for querying and analysis. With AWS Glue, you don't have to worry about the underlying infrastructure and can focus on your data. Amazon Athena is an interactive query service that makes it easy to analyze data in Amazon S3 using standard SQL. It integrates with AWS Glue, so you can use the catalogued data directly in Athena without any additional data movement or transformation.

Comment: B is the easiest. We can use Glue crawler.

Comment: Answer B

Comment: Querying data in S3 with SQL is almost always Athena.

Comment: If AWS asks the question of querying unstructured data in an efficient manner, it is almost always Athena

Comment: B. I don't think that you even need Glue to transform anything. Just use Glue to define the schemas and then use Athena to query based on those schemas.

Comment: answer is B

Comment: SQL on S3 is Athena so answer is B for sure

Comment: B is right

Comment: Answer is B. Queries Against an Amazon S3 Data Lake Data lakes are an increasingly popular way to store and analyze both structured and unstructured data. If you want to build your own custom Amazon S3 data lake, AWS Glue can make all your data immediately available for analytics without moving the data. https://aws.amazon.com/glue/

Comment: Correct Ans is D...Kinesis Data Analytics can use Lambda to transform and then run the SQL queries..

Replies:

> **Comment:** May I know why you are taking complex route?

Comment: Can Glue Crawler process unstructured data?

Replies:

> **Comment:** https://aws.amazon.com/glue/ - See Use cases; Queries against an Amazon S3 data lake Data lakes are an increasingly popular way to store and analyze both structured and unstructured data. If you want to build your own custom Amazon S3 data lake, AWS Glue can make all your data immediately available for analytics without moving the data.
>
> **Comment:** You could use a glue job to transform your unstructured data into more appropriate formats. Also, depending on your data, you might be able to create a custom classifier in glue, which will be able to crawl your data - this works particularly well in semi-structured cases, say for log files.

Discussion for Question 9

Link: https://www.examtopics.com/discussions/amazon/view/9656-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 8 votes

Discussion

Comment: Answer is A. The answer to this question is about Pipe mode from S3. The only options are A and C. As AWS Glue cannot be use to create models which is option C. The correct answer is A

Comment: Answer is A.

Comment: B. Generate daily precision-recall data in Amazon QuickSight, and publish the results in a dashboard shared with the Business team. This solution leverages QuickSight's managed service capabilities for both data processing and visualization, which should minimize the coding effort required to provide the Business team with the necessary insights. However, it's important to note that QuickSight's ability to calculate the precision-recall data depends on its support for the necessary statistical functions or the availability of such calculations in the dataset. If QuickSight cannot perform these calculations directly, option C might be necessary, despite the increased effort.

Comment: A. YES - pipe mode is best to start inference before the entire data is transferred; the only drawback is if multiple training jobs are done in sequence (eg. different hyperparamater), the data will be downloaded again B. NO - we want to use SageMaker first for initial training C. NO - We first want to test things in SageMaker D. NO - the SageMaker notebook will not use the AMI so the testing done is useless

Comment: The reason for this choice is that Pipe input mode is a feature of Amazon SageMaker that allows you to stream data directly from an Amazon S3 bucket to your training instances without downloading it first1. This way, you can avoid the time and space limitations of loading a large dataset onto your notebook instance. Pipe input mode also offers faster start times and better throughput than File input mode, which downloads the entire dataset before training1.

Comment: I think it should be a

Comment: It's A, pipe mode is for dealing with very big data.

Comment: A, PIPE is to do that sort of modeling

Comment: When data is already in S3 and next it should move to Sagemaker.. so option A is suitable

Comment: Answer is A. B, C & D can be dropped because there is no integration from/to Sage Maker train job (model).

Comment: Gotta be A. You need to use Pipe mode but Glue cannot train a model.

Comment: AAAAAAAAAa

Comment: ans is A

Comment: Will you run AWS Deep Learning AMI for all cases where the data is very large in S3? Also what role is Glue playing here? Is there a transformation? These are the two issues for options B C and D. I believe they do not represent what is required to satisfy the requirements in the question. The answer definitely requires the pipe mode, but not with Glue. I go with A https://aws.amazon.com/blogs/machine-learning/using-pipe-input-mode-for-amazon-sagemaker-algorithms/

Comment: go for A

Comment: Agree with A.

Comment: A is correct

Discussion for Question 10

Link: https://www.examtopics.com/discussions/amazon/view/11376-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 10 votes

Discussion

Comment: Answer is B as the data for a SageMaker notebook needs to be from S3 and option B is the only option that says it. The only thing with option B is that it is talking of moving data from MS SQL Server not RDS

Replies:

> **Comment:** https://www.slideshare.net/AmazonWebServices/train-models-on-amazon-sagemaker-using-data-not-from-amazon-s3-aim419-aws-reinvent-2018
>
> **Replies:**
>
> > **Comment:** Please look at the slide 14 of that link, although the data source from DynamoDB or RDS, it is still need to use AWS Glue to move the data to S3 for SageMaker to use. So, the right anwser should be B.
>
> **Comment:** I agree. As from the ML developer guide I just read, it is the MYSQL RDS that can be used as SQL datasource.

Comment: Option B (exporting to S3) is typically more flexible and cost-effective for large-scale or complex data needs (Which is our case - production), while Option A (direct connection) can be simpler and more immediate for real-time or smaller-scale scenarios like testing.

Comment: A. NO. It is doable, but this is not the best approach. B. YES C. NO. Pushing data to DynamoDB would not make it easier to access data D. NO. Pushing data to ElastiCache would not make it easier to access data

Comment: For Amazon S3, you can import data from an Amazon S3 bucket as long as you have permissions to access the bucket. For Amazon Athena, you can access databases in your AWS Glue Data Catalog as long as you have permissions through your Amazon Athena workgroup. For Amazon RDS, if you have the AmazonSageMakerCanvasFullAccess policy attached to your user's role, then you'll be able to import data from your Amazon RDS databases into Canvas. https://docs.aws.amazon.com/sagemaker/latest/dg/canvas-connecting-external.html

Replies:

Comment: https://aws.amazon.com/about-aws/whats-new/2024/04/amazon-sagemaker-studio-notebooks-data-sql-query/

Comment: A. NO - SageMaker can only read from S3 B. YES - AWS Data Pipeline can moved from SQL Server to S3 C. NO - SageMaker can only read from S3 and not DynamoDB D. NO - SageMaker can only read from S3 and not ElastiCache

Comment: This approach is the most scalable and reliable way to train a model using data stored in Amazon RDS. Amazon S3 is a highly scalable and durable object storage service, and Amazon Data Pipeline is a managed service that makes it easy to move data between different AWS services. By pushing the data to Amazon S3, the Specialist can ensure that the data is available for training the model even if the Amazon RDS instance is unavailable.

Comment: I think it should be b

Comment: It's B, even if Microsoft SQL Server is a strange name for RDS, it's a possible database to use there and the data for sagemaker needs to be in S3!

Comment: In Option B approach, the Specialist can use AWS Data Pipeline to automate the movement of data from Amazon RDS to Amazon S3. This allows for the creation of a reliable and scalable data pipeline that can handle large amounts of data and ensure the data is available for training. In the Amazon SageMaker notebook, the Specialist can then access the data stored in Amazon S3 and use it for training the model. Using Amazon S3 as the source of training data is a common and scalable approach, and it also provides durability and high availability of the data.

Comment: B is the correct answer. Official AWS Documentation: "Amazon ML allows you to create a datasource object from data stored in a MySQL database in Amazon Relational Database Service (Amazon RDS). When you perform this action, Amazon ML creates an AWS Data Pipeline object that executes the SQL query that you specify, and places the output into an S3 bucket of your choice. Amazon ML uses that data to create the datasource."

Comment: While B is a valid answer, It is also possible to make a SQL connection in a notebook and create a data object so A could be a valid answer too https://stackoverflow.com/questions/36021385/connecting-from-python-to-sql-server https://www.mssqltips.com/sqlservertip/6120/data-exploration-with-python-and-sql-server-using-jupyter-notebooks/

Replies:

Comment: you need to choose the best answer, not any valid answer. Often, many of the answers are valid solutions, but are not best practice.

Comment: B is correct. MS SQL Server is also under RDS.

Comment: B is right

Comment: B it is

Comment: I'll go with B

Discussion for Question 11

Link: https://www.examtopics.com/discussions/amazon/view/8304-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 5 votes

Discussion

Comment: answer should be C Collaborative filtering is for recommendation, LDA is for topic modeling

Comment: In natural language processing, the latent Dirichlet allocation (LDA) is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. Amazon SageMaker Random Cut Forest (RCF) is an unsupervised algorithm for detecting anomalous data points within a data set Neural network is used for image detection Answer is C

Comment: From the doc: "You can use LDA for a variety of tasks, from clustering customers based on product purchases to automatic harmonic analysis in music." https://docs.aws.amazon.com/sagemaker/latest/dg/lda-how-it-works.html

Comment: A. NO - LDA is for topic modeling B. NO - NN is a too generic term, you want Neural Collaborative C. YES - Collaborative filtering best fit D. NO - Random Cut Forest (RCF) for anomalities

Comment: Collaborative filtering is a machine learning technique that recommends products or services to users based on the ratings or preferences of other users. This technique is well-suited for identifying customer shopping patterns and preferences because it takes into account the interactions between users and products.

Comment: I think it should be c

Comment: C, always when talk about recommendation you can think about collaborative patterns!

Comment: A LDA used before collaborative filtering is largely adopted. 1) the input data that we have doesn't lend itself to collaborative filtering - it requires a set of items and a set of users who have reacted to some of the items, which is NOT what we have 2) recommendation is just one thing that we want to do. What about trends? 3) collaborative filtering isn't one of the pre-built algorithms (weak argument, admittedly)

Comment: collaborative

Comment: C. Easy question.

Comment: its a appropriate use case of Collaborative filtering

Comment: this is C

Comment: I'm thinking that it is A because: 1) the input data that we have doesn't lend itself to collaborative filtering - it requires a set of items and a set of users who have reacted to some of the items, which is NOT what we have 2) recommendation is just one thing that we want to do. What about trends? 3) collaborative filtering isn't one of the pre-built algorithms (weak argument, admittedly)

Comment: Answer is C, demographics, past visits, and locality information data, LDA is appropriate

Replies:

Comment: Collaborative filtering is appropriate

Comment: Answer A might be more suitable than other https://docs.aws.amazon.com/zh_tw/sagemaker/latest/dg/lda-how-it-works.html

Replies:

Comment: Not convinced with A. Answer C seems to be a better fit than A for recommendation model (LDA appears to be a topic-based model on unavailable data with similar patterns) https://aws.amazon.com/blogs/machine-learning/extending-amazon-sagemaker-factorization-machines-algorithm-to-predict-top-x-recommendations/

Comment: Also found this article which implemented LDA for analysing shopping trends, patterns.https://arxiv.org/pdf/1810.08577.pdf

Replies:

Comment: That said ..I'm splijt between LDA and Collaborative filtering(CF). One point is that CF is a method which will feed into Recommendeer system.

Discussion for Question 12

Link: https://www.examtopics.com/discussions/amazon/view/10005-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 9 votes

Discussion

Comment: B seems to be okay

Comment: A. NO - Linear regression is not best for classification B. YES - Classification C. NO - we want supervised classification D. NO - there is nothing to Reinforce from

Comment: Option B. This is a scenario for supervised learning model as data is labelled and only A, B are supervised learning algorithms from the options. Linear learning is to predict time series data and distribution is selecting which class the input belongs to. Hence most suitable is to use Binomial distribution model in this case.

Comment: The reason for this choice is that classification is a type of supervised learning that predicts a discrete categorical value, such as yes or no, spam or not spam, or churn or not churn1. Classification models are trained using labeled data, which means that the input data has a known target attribute that indicates the correct class for each instance2. For example, a classification model that predicts customer churn would use data that has a label indicating whether the customer churned or not in the past. Classification models can be used for various applications, such as sentiment analysis, image recognition, fraud detection, and customer segmentation2. Classification models can also handle both binary and multiclass problems, depending on the number of possible classes in the target attribute3.

Comment: The question is not clear. Actually we have 2 tasks here - group into categories (clustering) and predict if customers will churn/not churn (classification). If we had to simply do classification, why there was mentioned to group into categories?

Comment: This is definitely a classification problem

Comment: B is correct

Comment: B - it's a Binary Classification problem. Will the customer churn: Yes or No

Comment: 100% is B since it is about labelled data

Comment: i think the key is "the company has labeled the data" so this is classification, so it's B

Comment: B is okey

Comment: B is correct

Discussion for Question 13

Link: https://www.examtopics.com/discussions/amazon/view/45385-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: A is correct answer. Please Refer: https://machinelearningmastery.com/decompose-time-series-data-trend-seasonality/

Comment: A; the problem is bias, not trends

Comment: agreed, this seems to be A. there is similarity between the blue and green lines as far as capturing trend and seasonality is concerned. It just seems that if assumption is that the model is a linear regression model then just the intercept is off by a few units.

Comment: A. The model predicts both the trend and the seasonality well

Comment: The problem is Bias not trends or sesonality!

Comment: A is right, both trend (rising) and seasonality is there

Comment: C is correct answer

Replies:

> **Comment:** A is correct answer. Not C

Comment: The trend is up, so isn't it correctly predicted? And the seasonality is also in sync, the amplitude is wrong.

Comment: A is right. trend and seasonality are fine, level is the one the model gets wrong

Comment: Should be C

Comment: Should be A

Discussion for Question 14

Link: https://www.examtopics.com/discussions/amazon/view/43907-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 7 votes

Discussion

Comment: Answer is C. SVM sample use case is to put the dimensions into a higher hyperplane that can separates it. Seeing how separable it is, SVM can be used for it.

Comment: You can use a support vector machine (SVM) when your data has exactly two classes. An SVM classifies data by finding the best hyperplane that separates all data points of one class from those of the other class. The best hyperplane for an SVM means the one with the largest margin between the two classes. Margin means the maximal width of the slab parallel to the hyperplane that has no interior data points.

Comment: Well, C is the correct answer. This example is a classical one to use SVM.

Comment: SVM for RBF mode is the answer!

Comment: Answer is C

Comment: Textbook C

Comment: C. more reading for using non-linear kernel and separate samples with a hyperplane in a higher dimension space: https://medium.com/pursuitnotes/day-12-kernel-svm-non-linear-svm-5fdefe77836c

Comment: C seems right

Comment: answer is C

Comment: Agree. The answer is A. https://www.surveypractice.org/article/2715-using-support-vector-machines-for-survey-research

Comment: This is a good explanation of SVM https://uk.mathworks.com/help/stats/support-vector-machines-for-binary-classification.html

---

Discussion for Question 15

Link: https://www.examtopics.com/discussions/amazon/view/11279-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 10 votes

Discussion

Comment: Important things to note here is that 1. "The Data in S3 Needs to be Accessible from VPC" 2. "Traffic should not Traverse internet" To fulfill Requirement #2 we need a VPC endpoint To RESTRICT the access to S3/Bucket - Access allowed only from VPC via VPC Endpoint Even though Sagemaker uses EC2 - we are NOT asked to secure the EC2 :) So the answer is A

Comment: Between A & B, the answer should be A. From here: https://docs.aws.amazon.com/vpc/latest/userguide/vpc-endpoints-s3.html#vpc-endpoints-s3-bucket-policies We can see that we restrict access using DENY if sourceVpce (vpc endpoint), or sourceVpc (vpc) is not equal to our VPCe/VPC. So we are using a DENY (choice A) and not an ALLOW policy (choice B). Choices C, D we eliminate because they don't address S3 access at all.

Comment: A. YES - We first create a S3 endpoint in the VPC subnet so traffic does not flow through the Internet, then on the S3 bucket create an access policy that restricts access to the given VPC based on its ID B. NO - we don't want to be specific to an instance C. NO - the S3 bucket is on AWS network, you cannot change the NACL for it D. NO - not all instances in a VPC will necessarily have the same principal that can be specified in the policy

Comment: Definetly A

Comment: Well, but removing methodology, only A remains: The question never cited EC2

Comment: Per https://docs.aws.amazon.com/AmazonS3/latest/userguide/example-bucket-policies-vpc-endpoint.html it's A

Comment: In Option A, the Machine Learning Specialist would create a VPC endpoint for Amazon S3, which would allow traffic to flow directly between the VPC and Amazon S3 without traversing the public internet. Access to the S3 bucket containing PII can then be restricted to the VPC endpoint and the VPC using a bucket access policy. This would ensure that only instances within the VPC can access the data, and that the data does not traverse the public internet. Option B and D, allowing access from an Amazon EC2 instance, would not meet the requirement of not traversing the public internet, as the EC2 instance would be accessible from the internet. Option C, using Network Access Control Lists (NACLs) to allow traffic between only the VPC endpoint and an EC2 instance, would also not meet the requirement of not traversing the public internet, as the EC2 instance would still be accessible from the internet.

Comment: The question do not mention EC2 at all, so should be A

Comment: I think it should be B. Traning instance is a EC2 instance and need to be set an endpoint to load the data from S3.

Comment: AWS security is a conservative security model, which implies that access are denied by default rather than granted by default. We have to explicitly allow access to a AWS resource. Additionally, B talks about allowing access FROM the VPC to S3 while A talks about allowing access from S3 to VPC (which is not what we need). So, B.

Replies:

> **Comment:** Um, no. A VPC endpoint is outbound from the VPC to a supported AWS service.

Comment: Will go with B

Comment: Betting on B here, we should control access from VPC, not to VPC.

Comment: A! Restricting access to a specific VPC endpoint The following is an example of an Amazon S3 bucket policy that restricts access to a specific bucket, awsexamplebucket1, only from the VPC endpoint with the ID vpce-1a2b3c4d. The policy denies all access to the bucket if the specified endpoint is not being used. The aws:SourceVpce condition is used to specify the endpoint. https://docs.aws.amazon.com/AmazonS3/latest/userguide/example-bucket-policies-vpc-endpoint.html

Comment: Can't be B. You simple cannot enable access to an endpoint to some selected instance. So A.

Replies:

> **Comment:** We shouldn't use private IP in bucket policy.
>
> **Comment:** B does not say enable access TO the VPC endpoint. It says to allow access FROM the endpoint. So B is the correct answer. A talks about restricting access TO the VPC endpoint, so that option is irrelevant. We're worried about access TO the S3 bucket, not access to the VPC. The question is not poorly-worded, but it is tricky and you need to read it carefully.

Comment: I also vote A.

Comment: A found here "You can control which VPCs or VPC endpoints have access to your buckets by using Amazon S3 bucket policies. For examples of this type of bucket policy access control, see the following topics on restricting access." https://docs.aws.amazon.com/AmazonS3/latest/dev/example-bucket-policies-vpc-endpoint.html

Comment: agree with A

Discussion for Question 16

Link: https://www.examtopics.com/discussions/amazon/view/12378-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- D: 9 votes

Discussion

Comment: Answer is D. Should the weight be increased or reduced so that the error is smaller than the current value? You need to examine the amount of change to know that. Therefore, we differentiate and check whether the slope of the tangent is positive or negative, and update the weight value in the direction to reduce the error. The operation is repeated over and over so as to approach the optimal solution that is the goal. The width of the update amount is important at this time, and is determined by the learning rate.

Comment: maybe D ?

Comment: When the learning rate is set too high, it can lead to oscillations or divergence during training. Here's why: High Learning Rate: A high learning rate means that the model's parameters are updated by a large amount in each training step. This can cause the model to overshoot the optimal parameter values, leading to instability in training. Oscillations: If the learning rate is excessively high, the model's updates can become unstable, causing it to oscillate back and forth between parameter values. This oscillation can prevent the model from converging to an optimal solution. To address this issue, you can try reducing the learning rate. It's often necessary to experiment with different learning rates to find the one that works best for your specific problem and dataset. Learning rate scheduling techniques, such as reducing the learning rate over time, can also help stabilize training.

Comment: If the learning rate is too high, the model weights may overshoot the optimal values and bounce back and forth around the minimum of the loss function. This can cause the training accuracy to oscillate and prevent the model from converging to a stable solution. The training accuracy is the proportion of correct predictions made by the model on the training data.

Comment: Answer is A. A high learning rate means that the model parameters are being updated by large magnitudes in each iteration. As a result, the optimization process may struggle to converge to the optimal solution, leading to erratic behavior and fluctuations in training accuracy.

Comment: If learning rate is high, the accuracy is fluctuated because the value of loss function moves back and forth over the global minimum.

Comment: The big learning rating overshoot in true minima.

Comment: A high learning rate can cause oscillations in the training accuracy because the optimizer makes large updates to the model parameters in each iteration, which can cause overshooting the optimal values. This can result in the model oscillating back and forth across the optimal solution.

Comment: D Learning rate is too high. Textbook example of learning rate being too high. Lower Learning_rate will take more iterations, or longer to train, but will settle in place.

Comment: 12-sep exam

Comment: D: per supuesto

Comment: A company sells thousands of products on a public website and wants to automatically identify products with potential durability problems. The company has 1.000 reviews with date, star rating, review text, review summary, and customer email fields, but many reviews are incomplete and have empty fields. Each review has already been labeled with the correct durability result. A machine learning specialist must train a model to identify reviews expressing concerns over product durability. The first model needs to be trained and ready to review in 2 days. What is the MOST direct approach to solve this problem within 2 days? A. Train a custom classifier by using Amazon Comprehend. B. Build a recurrent neural network (RNN) in Amazon SageMaker by using Gluon and Apache MXNet. C. Train a built-in BlazingText model using Word2Vec mode in Amazon SageMaker. D. Use a built-in seq2seq model in Amazon SageMaker.

Replies:

> **Comment:** Is A valid option?

Comment: D is correct. big batch size make local minia.

Comment: it is a multiple answer question and answer should be both A and D

Comment: Answer is D 100%; learning rate too high will cause such an event

Comment: The answer is D, from the Coursera deep learning specialization (course 2 - improving Deep NN)

Comment: I think D is the answer: https://www.jeremyjordan.me/nn-learning-rate/

Discussion for Question 17

Link: https://www.examtopics.com/discussions/amazon/view/8306-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 7 votes

Discussion

Comment: the MOST efficient means to you don't need to coding, building infra All of sevices are manage by AWS is good, Transcribe, Amazon Translate, and Amazon Comprehend Answer is A

Replies:

Comment: Agree, Answer is A

Comment: A is not 100% correct. You don't need to translate Spanish. Amazon Comprehend supports Spanish.

Replies:

Comment: Arguably, you still need a translation since the person doesn't speak Spanish.

Replies:

Comment: I think there is no need to use Amazon translate because sometimes the translation is not accurate. It means some information gets lost.

Replies:

Comment: Given the question, I believe that is necessary: look at the enphase of not understanding spanish. besides that, even with some information lost, you will at least understand something.

Comment: A. YES - Comprehend is supervised so user must understand through Translate B. NO - seq2seq is for generation and not classification C. NO - Amazon SageMaker Neural Topic Model is unsupervised topic extraction, will not give sentiment against user-defined classes D. NO - BlazingText is word2vec, does not give sentiment classes

Comment: It's A: 1.Amazon Transcribe - to convert Spanish speech to Spanish text. 2.Amazon Translate - to translate Spanish text to English text 3.Amazon Comprehend - to analyze text for sentiments

Comment: It's A 100%

Comment: It's A: 1.Amazon Transcribe - to convert Spanish speech to Spanish text. 2.Amazon Translate - to translate Spanish text to English text 3.Amazon Comprehend - to analyze text for sentiments

Comment: Transcribe: Speech to text Translate: Any language to any language Comprehend: offers a range of capabilities for extracting insights and meaning from unstructured text data. Ex: Sentiment analysis, entity recognition, KeyPhrase Extraction, Language Detection, Document Classification

Comment: absolutely need STT(transcribe), translation(translate), and sentimental analysis(comprehend)

Comment: A - confirmed by ACG

Comment: I agree that the answer is A

Comment: answer is a

Comment: A; D is wrong because The Amazon SageMaker BlazingText algorithm provides highly optimized implementations of the Word2vec and text classification algorithms.

Comment: The Question/Anwser is not poorly as someone mentioned. --Even though Comprehend can do the analysis directly on Spanish (no need of translate) but if comprehend does analysis and the resulting words are still in spanish , it will no help the employee as he doesnt know Spanish.So the transalate after transcribe will help Employee understand what is being analyzed by Comprehend in next step. So read the question carefully before jumping to conclusions. it will save you an Exam :)

Comment: I don't get this question. Comprehend supports Spanish natively. There is no need for Translate, and translate would actually reduce effectiveness of sentimental analysis. However, BCD are all invalid choices.

Comment: A because Comprehend can provide sentiment analysis

Comment: A, https://aws.amazon.com/getting-started/hands-on/analyze-sentiment-comprehend/

Comment: Amazon Comprehend is needed for sure; answer is A 100%

Discussion for Question 18

Link: https://www.examtopics.com/discussions/amazon/view/9805-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 10 votes

Discussion

Comment: https://docs.aws.amazon.com/sagemaker/latest/dg/sagemaker-dg.pdf page 55: If you plan to use GPU devices, make sure that your containers are nvidia-docker compatible. Only the CUDA toolkit should be included on containers. Don't bundle NVIDIA drivers with the image. For more information about nvidia-docker, see NVIDIA/nvidia-docker. So the answer is B

Replies:

Comment: Yeah, it's B. But the page in the developer guide is page number 201 (209 in pdf). Second bullet point at the top.

Comment: Answer is B. below is from AWS documentation, If you plan to use GPU devices for model training, make sure that your containers are nvidia-docker compatible. Only the CUDA toolkit should be included on containers; don't bundle NVIDIA drivers with the image. https://docs.aws.amazon.com/sagemaker/latest/dg/your-algorithms-training-algo-dockerfile.html

Comment: https://docs.aws.amazon.com/sagemaker/latest/dg/sagemaker-dg.pdf page 55

Replies:

Comment: page 570 On a GPU instance, the image is run with the --gpus option. Only the CUDA toolkit should be included in the image not the NVIDIA drivers. For more information, see NVIDIA User Guide.

Comment: Answer B Load the CUDA toolkit only, not the drivers. Ref GPU section : https://docs.aws.amazon.com/sagemaker/latest/dg/studio-byoi-specs.html

Comment: A. NO - the drivers are not necessary (https://docs.aws.amazon.com/sagemaker/latest/dg/your-algorithms-training-algo-dockerfile.html) B. YES - it is about using the CUDA library, need to use proper base image (https://medium.com/@jgleeee/building-docker-images-that-require-nvidia-runtime-environment-1a23035a3a58) C. NO - file structure irrelavant to GPU D. NO - SageMaker config, irrelevant to Docker

Comment: The reason for this choice is that NVIDIA-Docker is a tool that enables GPU-accelerated containers by automatically configuring the container runtime to use NVIDIA GPUs1. NVIDIA-Docker allows you to build and run Docker containers that can fully access the GPUs on your host system. This way, you can run GPU-intensive applications, such as deep learning frameworks, inside containers without any performance loss or compatibility issues.

Comment: I think it should be b

Comment: B is correct!

Comment: As per aws documentation, answer is B, and A is even explicitly not recommended

Comment: NVIDIA-Docker is a Docker container runtime plugin that allows the Docker container to access the GPU resources on the host machine. By building the Docker container to be NVIDIA-Docker compatible, the Docker container will have access to the NVIDIA GPU resources on the Amazon EC2 P3 instances, allowing for accelerated training of the ResNet model.

Comment: To leverage the NVIDIA GPUs on Amazon EC2 P3 instances for training with Amazon SageMaker, the Docker container must be built to be compatible with NVIDIA-Docker. NVIDIA-Docker is a wrapper around Docker that makes it easier to use GPUs in containers by providing GPU-aware functionality. To build a Docker container that is compatible with NVIDIA-Docker, the Specialist should install the NVIDIA GPU drivers in the Docker container and install the NVIDIA-Docker runtime on the EC2 instances.

Comment: As referred in other comments ans is B

Comment: ANS B As mentioned byi other users

Comment: As per me answer is B

Comment: The answer is for sure B - as mentioned by others. And this is clearly stated in the docs

Comment: Ans. is B.

Comment: https://docs.aws.amazon.com/sagemaker/latest/dg/sagemaker-dg.pdf is clear "DO NOT BUNDLE NVIDIA DRIVERS WITH THE IMAGE" Details are found in https://github.com/NVIDIA/nvidia-docker A is wrong, C and D are out. Looks more like a B than anything

---

Discussion for Question 19

Link: https://www.examtopics.com/discussions/amazon/view/10011-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: Ans. A is correct

Comment: Answer is A. An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds

Comment: The reason for this choice is that a ROC curve is a graphical plot that illustrates the performance of a binary classifier across different values of the classification threshold1. A ROC curve plots the true positive rate (TPR) or sensitivity against the false positive rate (FPR) or 1-specificity for various threshold values2. The TPR is the proportion of positive instances that are correctly classified, while the FPR is the proportion of negative instances that are incorrectly classified.

Comment: ROC curve is for defining the threshold.

Comment: A surely

Comment: Question is about classification so confusion matrix would come into mind; A is the answer

Comment: A is indeed correct see https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters: • True Positive Rate • False Positive Rate True Positive Rate (TPR) is a synonym for recall and is therefore defined as follows: TPR = TP/TP+FN False Positive Rate (FPR) is defined as follows: FPR = FP/FP+TN

Comment: It is A.

Comment: obviously A

Comment: Root Mean Square Error (RMSE) Ans. c

Replies:

Comment: I think RMSE is for regression model

Discussion for Question 20

Link: https://www.examtopics.com/discussions/amazon/view/9825-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- D: 20 votes

Discussion

Comment: the solution is word embedding. As it is a interactive online dictionary, we need pre-trained word embedding thus the answer is D. In addition, there is no mention that the online dictonary is unique and does not have a pre-trained word embedding. Thus I strongly feel the answer is D

Comment: D is correct. It is not a specialized dictionary so use the existing word corpus to train the model

Comment: A. One-hot word encoding vectors: These vectors represent words by marking them as present or absent in a fixed-length binary vector. However, they don't capture relationships between words or their meanings. B. Producing synonyms: This would involve generating similar words for each word manually, which could be time-consuming and might not cover all possible contexts. C. Word embedding vectors based on edit distance: This approach focuses on how similar words are in terms of their spelling or characters, not necessarily their meaning or context in sentences. D. Downloading pre-trained word embeddings: These are vectors that represent words based on their contextual usage in a large dataset, capturing relationships between words and their meanings.

Comment: Pre-trained word embeddings, such as Word2Vec, GloVe, or FastText, capture the semantic and contextual meaning of words based on a large corpus of text data. By downloading pre-trained word embeddings, the Specialist can leverage the semantic relationships between words to provide meaningful word features for the nearest neighbor model powering the widget. Utilizing pre-trained word embeddings allows the model to understand and display words used in similar contexts effectively.

Comment: correct D ay tupoy

Comment: words that are used in similar contexts will have vectors that are close in the embedding space

Comment: A. NO - one-hot encoding is a very early featurization stage B. NO - we don't want human labelling C. NO - too costly to do from scratch D. YES - leverage exiting training; the word embeddings will provide vectors than be used to measure distance in the downstream nearest neighbor model

Comment: D is correct

Comment: I also believe that D is the correct answer. No reason to create word embeddings from scratch

Comment: 1. One-hot encoding will blow up the feature space - it is not recommended for a high cardinality problem domain. 2. One still needs to train the word features on large bodies of text to map context to each word

Comment: D. Download word embeddings pre-trained on a large corpus. Word embeddings are a type of dense representation of words, which encode semantic meaning in a vector form. These embeddings are typically pre-trained on a large corpus of text data, such as a large set of books, news articles, or web pages, and capture the context in which words are used. Word embeddings can be used as features for a nearest neighbor model, which can be used to find words used in similar contexts. Downloading pre-trained word embeddings is a good way to get started quickly and leverage the strengths of these representations, which have been optimized on a large amount of data. This is likely to result in more accurate and reliable features than other options like one-hot encoding, edit distance, or using Amazon Mechanical Turk to produce synonyms.

Comment: 12-sep exam

Comment: DDDDDDDDDDDDD

Comment: D for sure

Comment: Definitely D.

Comment: A)It requires that document text be cleaned and prepared such that each word is one-hot encoded. Refhttps://machinelearningmastery.com/what-are-word-embeddings/

Comment: I don't see how one-hot encoding works; I would say D 100% B & C are definitely wrong

---

Discussion for Question 21

Link: https://www.examtopics.com/discussions/amazon/view/10012-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- AD: 5 votes

Discussion

Comment: AD is correct

Comment: CloudTrail is use to track scientist how ofthe they deploy a model CloudWatch for monitoring GPU and CPU so answer is A & D

Comment: I think AWS Config is still not the service designed to track how often Data Scientists are deploying models, nor does it track operational performance metrics like GPU and CPU utilization or the invocation errors of SageMaker endpoints. and AWS CloudTrail continues to be the service that will track and record user activity and API usage, which includes deploying models in Amazon SageMaker. So the answers are still A and D - CloudTrail and CloudWatch.

Comment: AD is correct

Comment: A. YES - to track deployements B. NO - AWS Health is to track AWS Cloud itself (eg. is a zone down ?) C. NO - AWS Trusted Advisor to give recommendations on infra D. YES - for errors E. AWS Config

Comment: I also believe that A and D are correct. Can someone please explain to me the main differences between CloudWatch and CloudTrail? I find the documentation a bit confusing about it

Comment: Option E AWS Config to record all resource types, then the new resources will be automatically recorded in your account. Option A CloudTrail is use to track scientist how of the they deploy a model Option D CloudWatch for monitoring GPU and CPU

Comment: Log Amazon Sagemaker API Calls with AWS CloudTrail - https://docs.aws.amazon.com/sagemaker/latest/dg/logging-using-cloudtrail.html

Comment: I wouldn't be so sure about CloudTrail, AWS Configs also tracks Sagemaker and the resource "AWS::Sagemaker::Model"

Replies:

Comment: just seen, this was release 4 days ago... https://aws.amazon.com/about-aws/whats-new/2022/06/aws-config-15-new-resource-types/

Comment: A&D CloudWatch and ClouTrail

Comment: AD Are Correct.

Comment: absolutely

Comment: cloudtrail and cloudwatch, no thinking

Discussion for Question 22

Link: https://www.examtopics.com/discussions/amazon/view/9826-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: D is correct. Question has "'simple transformations, and some attributes will be combined" and Least development effort. Kinesis analytics can get data from Firehose, transform and write to S3 https://docs.aws.amazon.com/kinesisanalytics/latest/java/examples-s3.html

Replies:

Comment: I can't find any information that indicate Kinesis data analytics taking data from firehose

Comment: Best explanation here, kudos.

Comment: The best way to transform data is before it arrives to S3 so D should be best answer. But D is not completed. It should have another Firehose to deliver results to S3.

Comment: Ans is D Amazon Kinesis Data Analytics provides a serverless option for real-time data processing using SQL queries. In this case, by inserting a Kinesis Data Analytics stream downstream of the Kinesis Data Firehose stream, the retail chain can easily perform the required simple transformations on the ingested purchasing records.

Comment: The best answer is to use a lambda, but the letter D can do it very good too in the absence of the lambda option.

Comment: I go with D. A tough question, though. And C are definitely out. They key to the question is that it does not say that the transformed data needs to be stored again in S3. It just needs to be sent to the model for training after being transformed. So a Kinesis Data Analytics stream is appropriate to do the transformation.

Comment: Legacy data -- Firehose -- Kinesis Analytics -- S3.This happens in near real time before the data ends up in S3. --Legacy data -- Firehose -- S3 is already happening (mentioned in first line in question), adding Kinesis Data Analytics to do simple transformation joins using SQL on the incoming data is the LEAST amount of work needed. Kinesis Data analytics can write o S3. here is the AWS link with working example.Even Though Udemy tutorial said it cannot write directly to S3 :) . https://docs.aws.amazon.com/kinesisanalytics/latest/java/examples-s3.html

Comment: It seems that LEAST developmnet effort: https://aws.amazon.com/fr/blogs/big-data/preprocessing-data-in-amazon-kinesis-analytics-with-aws-lambda/ and GRETAST development effort: https://aws.amazon.com/fr/blogs/big-data/optimizing-downstream-data-processing-with-amazon-kinesis-data-firehose-and-amazon-emr-running-apache-spark/

Comment: It's D https://aws.amazon.com/blogs/big-data/preprocessing-data-in-amazon-kinesis-analytics-with-aws-lambda/ "In some scenarios, you may need to enhance your streaming data with additional information, before you perform your SQL analysis. Kinesis Analytics gives you the ability to use data from Amazon S3 in your Kinesis Analytics application, using the Reference Data feature. However, you cannot use other data sources from within your SQL query."

Replies:

Comment: I believe, kinesis should be used only in case of live data stream and this is not the case here. So as per me D shouldn't be the answer. I think A should be the answer as AWS storage gateway is something which is used alongwith on premise applications to move data to s3. Then glue can be used to transform the data.

Replies:

Comment: With option A, you would be changing the legacy data ingestion, a huge development effort. Remember, you're talking about 20,000 stores.

Comment: It is D.

Comment: I think the answer is D, because require the LEAST amount of development effort.

Comment: it's D, kinesis analytic can easily connect with firehose

Comment: why not A. it seems good to me

Replies:

Comment: "require stores to capture data locally using S3 gateway" - for 20k stores this creates a HUUUGE operational overhead and development effort, definitely wrong

Comment: D is correct...rest all need some kind of manual intervention as well as they are not simple..Firehose allows transformation as well as moving into S3

Comment: I think the answer is B. D would be correct if they didn't want to transform the legacy data from before the switch, but it seems like they do. Choosing D would mean that you'd have to use an EC2 instance or something else to transform the legacy data along with adding the Kinesis data analytics functionality. Also, there is no real-time requirement so daily transformation is fine.

Replies:

Comment: Its D, because with KDA you can transform the data with SQL while with EMR you need to write code, considering the requirement of "least development effort", so D

Comment: I think the answer is B. D would be correct if they didn't want to transform the legacy data from before the switch, but it seems like they do. Choosing D would mean that you'd have to use an EC2 instance or something else to transform the legacy data along with adding the Kinesis data analytics functionality. Also, there is no real-time requirement so daily transformation is fine.

Replies:

Comment: You can use Lambda instead of EC2. So D should be OK. https://aws.amazon.com/blogs/big-data/preprocessing-data-in-amazon-kinesis-analytics-with-aws-lambda/

Comment: If the question is "least cost" then B, but the question is "least develope effort, then you want to keep original architeture. I agree that for daily ETL instead of real-time, and large dataset, B is better option.

Comment: "LEAST amount of development effort" , EMR is no complicated to LEAST

Comment: can be B

Comment: Amazon Kinesis Data Analytics can not send data to S3 directly - it needs something like Kinesis Data Firehose after it.

Discussion for Question 23

Link: https://www.examtopics.com/discussions/amazon/view/8307-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: C might be much suitable softmax is to turn numbers into probabilities. https://medium.com/data-science-bootcamp/understand-the-softmax-function-in-minutes-f3a59641e86d

Comment: C is right. Softmax function is used for multi-class predictoins

Comment: A. NO - Dropout is to prevent overfitting B. NO - L1 regularization is to prevent overfitting C. YES - Softmax will give probabilities for each class D. NO - Rectified linear units (ReLU) is an activation function

Comment: Softmax is the correct answer.

Comment: Multiclassification with probabilities is about softmax!

Comment: Softmax is for probability distribution

Comment: it should be C. Softmax Softmax converts outputs to Probabilites of each classification

Comment: absolutely C

Comment: Absolute C.

Comment: This is as easy a question as you will likely see on the exam, Everyone has the right answer here.

Comment: C --> Softmax. Let's go over the alternatives: A. Dropout --> Not really a function, but rather a method to avoid overfitting. It consists of dropping some neurons during the training process, so that the performance of our algorithm does not become very dependent on any single neuron. B. Smooth L1 loss --> It's a loss function, thus a function to be minimized by the entire neural network. It's not an activation function. C. Softmax --> This is the traditional function used for multi-class classification problems (such as classifying an animal into one of 10 categories) D. Rectified linear units (ReLU) --> This activation function is often used on the first and intermediate (hidden) layers, not on the final layer. In any case, it wouldn't make sense to use it for classification because its values can exceed 1 (and probabilities can't)

Comment: C, Softmax is the best suitable answer Ref: The softmax function, also known as softargmax[1]:184 or normalized exponential function,[2]:198 is a generalization of the logistic function to multiple dimensions. It is used in multinomial logistic regression and is often used as the last activation function of a neural network to normalize the output of a network to a probability distribution over predicted output classes, based on Luce's choice axiom.

Comment: You guys are right, the answer is C since it automatically provides the output with a confidence interval... Relu could be used as well but it needs to be coded in to provide the probabilities https://medium.com/@himanshuxd/activation-functions-sigmoid-relu-leaky-relu-and-softmax-basics-for-neural-networks-and-deep-8d9c70eed91e

Comment: Definitely C

Comment: Definitely softmax.

Comment: Are you sure it is C?` The output should be "[the probability that] the input image belongs to each of the 10 classes." And not the most likely class with the highest probability, which would be the result of softmax layer.

Replies:

Comment: Yes, softmax returns indeed a vector of probabilities.

Comment: C, everyone with basic knowledge in neural network can easily see that

---

Discussion for Question 24

Link: https://www.examtopics.com/discussions/amazon/view/8308-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: RMSE says about the error value but not the sign of error. The question is to find whether the model overestimates or underestimates - I guess residual plots clearly show that answer B

Comment: Answer is B. Residual plot distribution indicates over or under-estimations

Comment: Residual plots shows mistake by mistake!

Comment: B - Residual plots it is - https://docs.aws.amazon.com/machine-learning/latest/dg/regression-model-insights.html

Comment: Residual Plots (B). AUC and Confusion Matrices are used for classification problems, not regression. And RMSE does not tell us if the target is being over or underestimated, because residuals are squared! So we actually have to look at the residuals themselves. And that's B.

Replies:

Comment: Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out

these residuals are. In other words, it tells you how concentrated the data is around the line of best fit. Root mean square error is commonly used in climatology, forecasting, and regression analysis to verify experimental results. 1) Squaring the residuals. 2) Finding the average of the residuals. 3) Taking the square root of the result.

Replies:

Comment: Residual Plots (B). would have to be my answer

Comment: residual plot https://docs.aws.amazon.com/machine-learning/latest/dg/regression-model-insights.html

Comment: https://stattrek.com/statistics/dictionary.aspx?definition=residual%20plot#:~:text=A%20residual%20plot%20is%20a,nonlinear%20model%20is%20more%20appropriate. Answer is B

Comment: without a second thought residual plot

Comment: The answer is B. Refer to Exercise 7.2.1.A https://stats.libretexts.org/Bookshelves/Introductory_Statistics/Book%3A_OpenIntro_Statistics_(Diez_et_al)./07%3A_Introduction_to_Linear_Regression/7.02%3A_Line_Fitting%2C_Residuals%2C_and_Correlation

Comment: Residual plot it is Option B

Comment: Residual plot

Comment: B is the correct answer!!!! RMSE has the S in it that is square... that vanishes the above below factor of the prediction. Answers C and D are for other type of problems

Comment: It should be B. The residual plot will be give whether the target value is overestimated or underestimated.

Comment: Answer is C. https://www.youtube.com/watch?v=MrjWcywVEiU

Replies:

Comment: Answer is B. Your vid shows a technique that is useful for defining integrals and has NOTHING to do linear regression. Also, it over-/underestimates the area under the curve, NOT the target value.

Comment: Good grief, AUC is used for classification not regression.

Comment: B..Residual helps to find out whether the model is underestimating or overestimating

Comment: answer is B

Comment: Go for B. Residual spots can handle this problem

Discussion for Question 25

Link: https://www.examtopics.com/discussions/amazon/view/43910-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 12 votes

Discussion

Comment: C is the correct answer because gaussian naive Bayes can do this nicely.

Replies:

Comment: of course it doesn't mention the gaussian here and refers to naive bayes in general, but I'm still positive with C.

Comment: Answer should be A:. B: LINEAR SVM is a linear classifier -> All of these have a linear decision boundary (so it's just a line y = mx+b). This leads to a bad recall and so A must be the right choice.

Comment: Answer in my opinion is A A Decision Tree Classifier can handle complex decision boundaries and does not assume any particular distribution of data. It is well-suited for cases like this where the decision boundary is non-linear, as seen with the clear separation between the normal and fraudulent transactions. A Naive Bayesian classifier, on the other hand, assumes independence among features and typically performs better when data is normally distributed, which might not be the case here given the data's clustering pattern.

Comment: From Claude 3 Haiku: A. NO, decision trees may struggle to capture the linear separability of the classes. B. NO, Linear SVM may not be able to fully exploit the class separation due to its linear decision boundary. C. YES, The Naive Bayesian classifier tends to perform well in situations where the classes are linearly separable. This model requires the features are independent and this is the case D. The single Perceptron with a sigmoidal activation function may not be able to capture the complex class distributions as effectively as the Naive Bayesian classifier.

Comment: Answer by Claude3: In contrast, the Decision Tree (A) and Linear SVM (B) models are generally more robust to overfitting and can achieve a better balance between recall and precision, but they may not necessarily have the highest recall for the minority class. Considering the importance of maximizing recall for the fraudulent class in this use case, the Naive Bayesian Classifier (C) could be a valid choice, although it may come with the trade-off of lower precision and potentially higher false positive rates.

Comment: highest recall. So A

Comment: Only A (DT) is non-linear among the mentioned algorithms.

Comment: Given the visualized data, the Decision tree (Option A) is likely the best model to achieve the highest recall for the fraudulent class. It can handle complex patterns and create rules that are more suited for clustered and potentially non-linearly separable classes. Recall is a measure of a model's ability to capture all actual positives, and a decision tree can be tuned to prioritize capturing more of the fraudulent cases at the expense of making more false-positive errors on the normal cases.

Comment: if it was highest precision: Given these considerations, the best model for precision would likely be a Support Vector Machine with a non-linear kernel, such as the RBF (Radial Basis Function) kernel. This model can tightly fit the boundary around the fraudulent class, minimizing the inclusion of normal transactions in the fraudulent prediction space, and thus potentially achieving high precision. Precision is sensitive to the false positives, and the flexibility of SVMs with non-linear kernels to create a tight and precise boundary can help to minimize these.

Comment: GPT 4 Answer is Decision Tree. Considering the goal is to achieve the highest recall for the fraudulent class, which means we aim to capture as many fraudulent cases as possible even if it means getting more false positives, a Decision Tree would likely be the best option. This is because it can adapt to the complex shape of the class distribution and encapsulate the majority of the fraudulent class within its decision boundaries. Recall is a measure of a model's ability to capture all actual positives, and the decision tree's complex boundary setting capabilities make it well-suited for maximizing recall in this case.

Comment: I'm going with A. As pointed out in this article, Naive Bayes performs poorly with non-linear classification problems. The picture shows a case where the classes are not linearly separable. Decision Tree will probably give better results. https://sebastianraschka.com/Articles/2014_naive_bayes_1.html

Comment: Highest recall for fraudulent class means that Precision for Fraudulent predictions can be low. So basically just two conditions Transaction Month nearly greater than 8 and age of accounts greater than 8 can help identify the fraudulent class but it will classify most of non-fraudulent cases as fraudulent.

Comment: hERE iTS tHE rECALL. iF iTS fOR pRECISION, i wOULD hAVE gONE wITH sVM. sO tHE cORRECT aNSWER IS c, nAVIE bAYES.

Comment: A. NO - Decision tree would create boundaries perpendicular to the axes; not great for an oval B. NO - No linear separation here, unless we increase the space dimensation C. YES - Naive Bayesian classifier create clean boundaries (https://martin-thoma.com/comparing-classifiers/) D. NO - It would need many hidden layers (https://medium.com/@amanatulla1606/unraveling-the-magic-how-multilayer-perceptron-captures-non-linearity-in-data-6a4d385f7592)

Comment: Option c

Comment: it's A. Decision Trees can do this easily

Comment: This is the answer given by Bing Chat:Based on the figure provided in the link and the information given in your message, it appears that a Naive Bayesian classifier would have the highest recall with respect to the fraudulent class. This is because a Naive Bayesian classifier can handle overlapping class distributions and can work well when there is a clear separation between classes as shown in the figure. However, it's important to note that model selection should also take into account other factors such as precision and overall accuracy.

Discussion for Question 26

Link: https://www.examtopics.com/discussions/amazon/view/10264-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- D: 12 votes

Discussion

Comment: This is a very tricky question. The idea is to reconfigure the ranges of the hyperparameters. A refers to a feature, not a hyperparameter. A is out. C refers to training the model, not optimizing the range of hyperparameters. C is out. Now it gets tricky. D will let you find determine what the approximately best tree depth is. That's good. That's what you're trying to do but it's only one of many hyperparameters. It's the best choice so far. B is tricky. t-SNE does help you visualize multidimensional data but option B refers to input variables, not hyperparameters. For this very tricky question, I would do with D. It's the only one that accomplishes the task of limiting the range of a hyperparameter, even if it is only one of them.

Replies:

Comment: It's good to see someone keeping a thoughtful and curious mind to this question. I too have the same conclusion, not an easy question.

Comment: But, how do you optimize hyperparameters without training experiments? That is why C is the best option. You get a value for each unique combination of hyperparameters.

Comment: B is also wrong as t-SNE picture is not actionable - good visual but ... that's it. try pictures here https://lvdmaaten.github.io/tsne/

Comment: When you are tuning hyperparameters you are literally training multiple models and searching for the best ones.

Comment: B doesn't make sense I think it's D

Comment: A. No, doesn't help to set/reduce hyperparameter value/range B. No, honestly this is gibberish to me C. No, doesn't help to reduce hyperparameter value range D. YES, this help me understand how to set max tree depth hyperparameter

Comment: Option C. See it is doing a scatter plot on the metric for each iteration. Each iteration is running with a certain set of hyper parameters. So if I plot this. and I find which iteration has the best metric, I could simply pick up those set of hyperparameters. D will only led to the tuning of maximum tree depth. I am not sure which option would satisfy the goal to decrease cost but just looking at maximum tree depth doesnt seem right to me. It might be a way to just look at the tree depth and tune just that parameter and since you are only tuning 1 paramter, it may be cheaper, but would that lead to a usable model? I think it should be option C.

Comment: On what basis the correct answers are provided in this platform? Are they assuming this is the correct answer or it is taken from somewhere ?

Comment: D IS THE CORRECT

Comment: Option D, can also be useful in hyperparameter tuning for tree-based ensemble models, especially if the maximum tree depth is one of the hyperparameters you want to optimize. However, when the goal is to decrease training time and costs by reconfiguring input hyperparameter ranges, a scatter plot showing the performance of the objective metric over each training iteration (Option C) is generally more directly related to the hyperparameter tuning process. It helps you track how the model's performance changes during hyperparameter tuning, which is critical for making decisions about which hyperparameter ranges to explore further. Option D is valuable for understanding the relationship between maximum tree depth and the objective metric, but it might not provide as comprehensive insights into the overall hyperparameter tuning process compared to Option C.

Comment: A. NO - it is about data discovery B. NO - it is about data discovery C. MIGHT - (NO) is a training iteration the overnight training the question is referring to ? (YES) Or each HPO training within each night ? D. YES - the less ambiguous answer

Comment: I think that C should be the right answer. The specialist can monitor how the model works by changing hyperparameters' values in each training iteration.

Comment: Option D

Comment: A and B are wrong, because is totally out of question context. C is for monitoring a model, it doesn't help to change your HP range. D is the only answer that applies to the question.

Comment: I think it should be c

Comment: By plotting the performance of the objective metric (AUC) over each training iteration, the Specialist can analyze how different hyperparameter configurations affect the model's performance. This visualization helps in understanding which hyperparameter combinations lead to better results and allows the Specialist to identify areas of improvement.

Comment: D: By analyzing this relationship, the Specialist can adjust the range of maximum tree depth values used during hyperparameter tuning to decrease training time and costs.

Comment: D Seems like the best answer. When answer is considered correct who is making that call an is there any justification provided for us to learn from?

Comment: It's about parameters, not about dimensionality.

Comment: BING chat chooses this answer, and provides an explanation:The correct answer is D. A scatter plot showing the correlation between maximum tree depth and the objective metric. This visualization will help the Machine Learning Specialist to understand how changes in maximum tree depth affect the performance of the model with respect to the objective metric (AUC). By analyzing this relationship, the Specialist can adjust the range of maximum tree depth values used during hyperparameter tuning to decrease training time and costs.

Discussion for Question 27

Link: https://www.examtopics.com/discussions/amazon/view/11820-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- BCF: 5 votes

Discussion

Comment: B C F should be correct.

Comment: I will select B, C, F 1- Apply words stemming and lemmatization 2- Remove Stop words 3- Tokensize the sentences https://towardsdatascience.com/nlp-extracting-the-main-topics-from-your-dataset-using-lda-in-minutes-21486f5aa925

Comment: A. NO - word2vec works on raw data B. YES - case here is not significant C. YES - will help reduce dimensionality D. NO - word2vec will do it by itself E. NO - One-hot encoding is for classification F. YES - word2vec takes tokens as input

Comment: Data need to be tokenized and cleaned!

Comment: B, C F is the correct

Comment: BCF correct. D is not correct (Pay attention to “in a repeatable manner” in the question.)

Comment: B/C/F. D should not be performed because spell check is a subjective thing. You don't know for sure what the word was supposed to be if you have a typo.

Comment: I saw this exact question on "whizlabs" practice exam and correct options were B/C/F

Comment: https://towardsdatascience.com/an-implementation-guide-to-word2vec-using-numpy-and-google-sheets-13445eebd281 Data Preparation — Define corpus, clean, normalise and tokenise words To begin, we start with the following corpus: “natural language processing and machine learning is fun and exciting” For simplicity, we have chosen a sentence without punctuation and capitalization. Also, we did not remove stop words “and” and “is”. In reality, text data are unstructured and can be “dirty”. Cleaning them will involve steps such as o removing stop words, o removing punctuations, o convert text to lowercase (actually depends on your use-case), o replacing digits, etc. o After preprocessing, we then move on to tokenising the corpus Answer: B, C, F

Replies:

Comment: BCF is 100% correct

Comment: Correct answers are B, C and F

Comment: The correct answer is B, C and F A: POS tagging has nothing to do with word2vec D: fixing "quck" to "quick" only works for that specific word F: word2vec can use CBOW or skipgram, so no need to have one-hot decoding here

Replies:

Comment: sorry E: word2vec can use CBOW or skipgram, so no need to have one-hot decoding here

Comment: BCF is correct

Comment: B, C F correct

Comment: B, C, and F are correct answers. I have done this question many times in many practice tests.

Comment: B, C, F are my choice. D is also possible but not as widely used as others.

Comment: Why C is not included in the answer? ABCD, all are correct answers

Replies:

Comment: "choose three"

Discussion for Question 28

Link: https://www.examtopics.com/discussions/amazon/view/10014-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 23 votes
- A: 5 votes

Discussion

Comment: SSML is specific to that particular doument, like W3C an be pronounced as "World Wide Web Consortium" using $_{W3C}$ in that specific document and when you create a new document, you need to format again. But with LEXICONS, you can upload a lexicon file once and ALL the FUTURE documents can just have W3C and that will be pronounced as "World Wide Web Consortium".. so answer is B, because the question asks for "future" documents.

Replies:

Comment: A.The document section for "Pronouncing Acronyms and Abbreviations". Source: https://docs.aws.amazon.com/polly/latest/dg/supportedtags.html

Comment: absolutely, B is the correct choice.

Comment: For the exact reason you state, the correct answer is A. For every different document, a particular acronym may mean something different so you must have a solution that is document-specific.

Replies:

Comment: It is the same business, so the acronyms are not expected to change from document to document

Comment: The correct answer is B, as explained by VB.

Comment: I think the answer is B. https://docs.aws.amazon.com/polly/latest/dg/managing-lexicons.html https://www.smashingmagazine.com/2019/08/text-to-speech-aws/

Replies:

Comment: Lifted from the above link - "Your text might include an acronym, such as W3C. You can use a lexicon to define an alias for the word W3C so that it is read in the full, expanded form (World Wide Web Consortium)." Clearly this is the same use case.

Comment: Answer : B https://aws.amazon.com/blogs/machine-learning/customize-pronunciation-using-lexicons-in-amazon-polly/ Use SSML tag which is great for inserting one-off customizations or testing purposes. We recommend using Lexicon to create a consistent set of pronunciations for frequently used words across your organization. This enables your content writers to spend time on writing instead of the tedious task of adding phonetic pronunciations in the script repetitively.

Comment: SSML supports phonetic pronunciation. Seems to me A is correct too. https://docs.aws.amazon.com/polly/latest/dg/supportedtags.html#phoneme-tag

Comment: B IS ANSWER

Comment: B is the correct, A hardan cixdi debil?

Comment: This issue can be faced with both methods described in A and B. Though the answer A refers to the "current" document while the question regards "future" documents, so I think the right answer is B.

Comment: Letter B is correct to ensure that acronyms or terms are pronounced correctly. Letter A works, but look at the catch: It's asked for future documents, but it mentions converting only current ones to SSML format, while future ones would be in plaintext.

Comment: Company using plaintext and Future document means plaintext! So only Custom Lexicon will help.

Comment: https://docs.aws.amazon.com/polly/latest/dg/managing-lexicons.html this explains acronym.

Comment: I believe it should be lexicon. Can you share how you tag the correct answer?

Comment: Key here being "for future documents", answer should be B as SSML is for a specific document only

Comment: this should be multiple choice question which answer is a AND b

Comment: response B A pronunciation lexicon is a list of words and their correct phonetic pronunciation that can be used to improve the accuracy of text-to-speech conversion. In this case, the Machine Learning Specialist can create a custom lexicon for the company's acronyms and upload it to Amazon Polly. This will ensure that the acronyms are pronounced correctly in the future announcements.

Comment: Should be B

Comment: With Amazon Polly's custom lexicons or vocabularies, you can modify the pronunciation of particular words, such as company names, acronyms, foreign words, etc. To customize these pronunciations, you upload an XML file with lexical entries. {rpmimcoatopm ;exocpms enable you to customize the pronunciation of words. Amazon Polly provides API operations that you can use to store lexicons in an AWS region. Those lexicons are then specific to that particular region. References: https://docs.aws.amazon.com/polly/latest/dg/managing-lexicons.html https://aws.amazon.com/blogs/machine-learning/create-accessible-training-with-initiafy-and-amazon-polly/

Comment: Ref: https://docs.aws.amazon.com/polly/latest/dg/managing-lexicons.html Your text might include an acronym, such as W3C. You can use a lexicon to define an alias for the word W3C so that it is read in the full, expanded form (World Wide Web Consortium).

Discussion for Question 29

Link: https://www.examtopics.com/discussions/amazon/view/10037-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- BE: 5 votes

Discussion

Comment: The model must have been overfitted. Regularization helps to solve the overfitting problem in machine learning (as well as data augmentation). Correct answers should be BE.

Replies:

Comment: agree on BE

Comment: Agreed 100%

Comment: Answer: BE https://www.kdnuggets.com/2019/12/5-techniques-prevent-overfitting-neural-networks.html 5 techniques to prevent overfitting: 1. Simplifying the model 2. Early stopping 3. Use data argumentation 4. Use regularization 5. Use dropouts

Comment: agreed with vetal

Comment: A. NO - vanishing gradient is somebody bad they might happen and prevent convergence, we don't want that or something we can add explicitly. it is a result of the learning B. YES - we have a overfitting problem so more training examples will help C. NO - we already have good accuracy on the training set D. NO - gradient checking is to find bugs in model implementation E. YES - we have a overfitting problem

Comment: B. Perform data augmentation on the training data. (it should add validation data as well) data should be distributed among train validation and test.

Comment: Answer B&E looks good

Comment: B & E is the correct ans

Comment: BE is exact

Replies:

Comment: BE are the correct answers

Comment: Looks like B and D are correct.. For D -> https://www.youtube.com/watch?v=P6EtCVrvYPU

Replies:

Comment: gradient checking doesn't resolve the issue, but adding it will confirm / deny the issue. So, it helps to validate the issue but not resolve. I would say B, E are correct

Comment: L2 regularization tries to reduce the possibility of overfitting by keeping the values of the weights and biases small.

Comment: why not becuase of vanishing gradient?

Replies:

Comment: Vanishing gradients are a problem when training a NN. Answer A mentions that the solution should be to add that, which is not possible. Correct solution is BE. https://www.kdnuggets.com/2022/02/vanishing-gradient-problem.html

Comment: This is L2 Regularization....Do you think this is the right answer?

Comment: agree BE

Discussion for Question 30

Link: https://www.examtopics.com/discussions/amazon/view/8316-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- CEF: 14 votes
- ACF: 12 votes

Discussion

Comment: THE ANSWER SHOUD BE CEF IAM ROLE, INSTANCE TYPE, OUTPUT PATH

Replies:

Comment: Why not A? You don't need to tell Sagemaker where the training data is located?

Replies:

Comment: You need to specify the InputDataConfig, but it does not need to be "S3" I think the reason why A and B are wrong, not because data location is not required, but because it doesn't need to be S3, it can be Amazon S3, EFS, or FSx location

Comment: Should be C, E, F From the SageMaker notebook example: https://github.com/aws/amazon-sagemaker-examples/blob/master/introduction_to_amazon_algorithms/semantic_segmentation_pascalvoc/semantic_segmentation_pascalvoc.ipynb # Create the sagemaker estimator object. ss_model = sagemaker.estimator.Estimator(training_image, role, train_instance_count = 1, train_instance_type = 'ml.p3.2xlarge', train_volume_size = 50, train_max_run = 360000, output_path = s3_output_location, base_job_name = 'ss-notebook-demo', sagemaker_session = sess)

Replies:

Comment: It says InstanceClass - CPU/GPU in the question, not InstanceType

Comment: instance type has default value.

Comment: From here https://docs.aws.amazon.com/zh_tw/sagemaker/latest/dg/API_CreateTrainingJob.html .. the only "Required: Yes" attributes are: 1. AlgorithmSpecification (in this TrainingInputMode is Required - i.e. File or Pipe) 2. OutputDataConfig (in this S3OutputPath is Required - where the model artifacts are stored) 3. ResourceConfig (in this EC2 InstanceType and VolumeSizeInGB are required) 4. RoleArn (..The Amazon Resource Name (ARN) of an IAM role that Amazon SageMaker can assume to perform tasks on your behalf...the caller of this API must have the iam:PassRole permission.) 5. StoppingCondition 6. TrainingJobName (The name of the training job. The name must be unique within an AWS Region in an AWS account.) From the given options in the questions.. we have 2, 3, and 4 above. so, the answer is CEF.

Replies:

Comment: This is the best explanation that CEF is the right answer, IMO. The document at that url is very informative. It also specifically states that InputDataConfig is NOT required. Having said that, I have no idea how the model will train if it doesn't know where to find the training data, but that is what the document says. If someone can explain that, I'd like to hear the explanation.

Replies:

Comment: If I see this question on the actual exam, I'm going with AEF. The model absolutely must know where the training data is. I have seen other documentation that does confirm that you need the location of the input data, the compute instance and location to output the model artifacts.

Replies:

Comment: but you also need to specify the service role sagemaker should use otherwise it will not be able to perform actions on your behalf like provisioning the training instances.

Comment: Perfect explanation. It is CEF

Comment: The question is asking about built in algorithms. It should be ADE. See https://docs.aws.amazon.com/zh_tw/sagemaker/latest/dg/API_CreateTrainingJob.html

Comment: for "3. ResourceConfig", only VolumeSizeInGB is required. So, it's not about the instance type. Check: https://docs.aws.amazon.com/zh_tw/sagemaker/latest/APIReference/API_ResourceConfig.html

Comment: Going with cef

Comment: ANSWER IS CEF Here from Amazon docs InputDataConfig An array of Channel objects. Each channel is a named input source. InputDataConfig describes the input data and its location. Required: No OutputDataConfig Specifies the path to the S3 location where you want to store model artifacts. SageMaker creates subfolders for the artifacts. Required: Yes ResourceConfig - Identifies the resources, ML compute instances, and ML storage volumes to deploy for model training. In distributed training, you specify more than one instance. Required: Yes

Comment: CEF https://docs.aws.amazon.com/sagemaker/latest/APIReference/API_CreateTrainingJob.html#API_CreateTrainingJob_RequestParameters

Comment: Based on https://docs.aws.amazon.com/sagemaker/latest/APIReference/API_CreateTrainingJob.html Required parameters are: - AlgorithmSpecification (registry path of the Docker image with the training algorithm) - OutputDataConfig (path to the S3 location where you want to store model artifacts) - ResourceConfig (resources, including the ML compute instances and ML storage volumes, to use for model training) - RoleArn - StoppingCondition (time limit for training job) - TrainingJobName Thus, the answer is: C E F wording for option E is inaccurate "EC2 instance class specifying whether training will be run using CPU or GPU" but they do it on purpose

Comment: The input channel and output channel are mandatory, as the training job needs to know where to get the input data from and where to publish the model artifact. IAM role is also needed, for AWS services. others are not mandatory, validation channel is not mandatory for instance in case of unsupervised learning, likewise hyper params can be be auto tuned for as well as the ec2 instance types can be default ones that will be picked

Comment: As they narrowed it to S3, A is incorrect BUT when submitting Amazon SageMaker training jobs using one of the built-in algorithms, it is a MUST to identify the location of training data. While Amazon S3 is commonly used for storing training data, other sources like Docker containers, DynamoDB, or local disks of training instances can also be used. Therefore, specifying the location of training data is essential for SageMaker to know where to access the data during training. So the right answer is CEF for me for this case... However if A was saying identify the location of training data, I think option A would be included in the MUST parameter.

Comment: InputDataConffig is optional in create_training_job.Please check thte parameters that are required. So answer is CEF: https://docs.aws.amazon.com/sagemaker/latest/APIReference/API_CreateTrainingJob.html

Comment: InputDataConffig is optional in create_training_job.Please check thte parameters that are required. So answer is SEF: https://docs.aws.amazon.com/sagemaker/latest/APIReference/API_CreateTrainingJob.html

Comment: Input is required only when calling Fit method. When initializing the Estimator, we do not need input

Comment: I open the sagemaker and tested. A C F B is not needed for non-supervised algorithm.

Comment: C, E, F The trick was the training channel, but all the data channel are passed during when actually training the model using fit method

Comment: E is not important, some models could simply work on the default of CPU. A is a must and E is a must too. C is important for permission handling on S3 etc. It has to be A, C, F

Replies:

> **Comment:** Correction, having gone thru the doc more closely, there is no default for instance type. So the choices should be A, C, E.

Comment: A. The training channel identifying the location of training data on an Amazon S3 bucket: This is where SageMaker will get the input data for training the model. C. The IAM role that Amazon SageMaker can assume to perform tasks on behalf of the users: This role provides SageMaker the necessary permissions to access AWS resources. F. The output path specifying where on an Amazon S3 bucket the trained model will persist: After training, the model artifacts need to be saved in a specified S3 bucket location.

Comment: Please go through the lab https://catalog.us-east-1.prod.workshops.aws/workshops/63069e26-921c-4ce1-9cc7-dd882ff62575/en-US/lab2

Comment: ACF is answer

Discussion for Question 31

Link: https://www.examtopics.com/discussions/amazon/view/14807-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 6 votes

Discussion

Comment: Answer is B. Athena is best in Parquet format.

Comment: You can improve the performance of your query by compressing, partitioning, or converting your data into columnar formats. Amazon Athena supports open source columnar data formats such as Apache Parquet and Apache ORC. Converting your data into a compressed, columnar format lowers your cost and improves query performance by enabling Athena to scan less data from S3 when executing your query

Comment: A. NO - slower B. YES - Parquet native in Aethena/Presto C. NO - Compressed JSON D. NO - no built-in support

Comment: according to: https://dzone.com/articles/how-to-be-a-hero-with-powerful-parquet-google-and the query run time over parquet file was 6.78 seconds while it was 236 seconds on the same data but stored on csv file which mean that parquet file is 34x faster than csv file

Comment: B it is

Comment: Answer is B. https://aws.amazon.com/tw/blogs/big-data/top-10-performance-tuning-tips-for-amazon-athena/ But why does this question relate to Machine Learning?

Replies:

> **Comment:** Because you must explore data very quickly using SQL in order to run EDA / analyze data for ML purposes. Those explorations can inform on selecting features that can be used for modeling purposes.

Discussion for Question 32

Link: https://www.examtopics.com/discussions/amazon/view/46347-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: I choose b

Comment: Correct answer is B. Example: Mon | Tue | Wed 1 0 0 0 1 0

Comment: Easy Peasy

Comment: Any categorical feature needs to be converted using One Hot Encoding and NOT label encoding.

Comment: Originally I put A, (believing to be able to format it as (0,1,2,3,4,5,6), or something as it mentioned it to convert the column, but later I realized Binarization is only designed for continuous or numerical data. Even though one-hot encoding will create 6 more columns it is correct. B is correct.

Comment: B 1000000 = Mon 0100000 = Tue 0010000 = Wed 0001000 = Thur 0000100 = Fri 0000010 = Sat 0000001=Sun

Comment: why not A? 001 010, 011

Replies:

> **Comment:** i thought of this at first, but chatgpt's explanation changed my mind In summary, if the names of days represent nominal categorical variables, one-hot encoding is generally the preferred choice. It maintains distinctiveness, is interpretable, and ensures that each day is clearly represented as a separate binary feature. Binary encoding may be considered for memory efficiency, especially when dealing with a large number of ordinal categories, but it should be used with caution as it introduces an ordinal relationship between categories, which may or may not align with the nature of the data. Ultimately, the choice between the two methods should align with the specific needs of your analysis and the data's characteristics.

Comment: B is the obvious answer

Comment: Binary encoding would've been a correct answer but it is not here & Binarization is used for continuous variables. leaving w/ option B

Comment: B is wrong. You do not need to one hot encode the variable in random trees. If you do so, you tree must be very deep, which is not efficient. The correct answer is C!

Replies:

> **Comment:** Stop misleading people, the question already asked to convert the data into binary. C is not even remotely close to be correct

Comment: "The Specialist want to convert the Day Of Week column in the dataset to binary values." You are misreading the question. The answer is B.

Discussion for Question 33

Link: https://www.examtopics.com/discussions/amazon/view/10409-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- CD: 5 votes

Discussion

Comment: I think it should be CD C: because we need a balance dataset D: The number of positive samples is large so model tends to predict 0 (negative) for all cases leading to False Negative problem. We should minimize that. My opinion

Comment: I think it should be CD C: because we need a balance dataset D: The number of negative samples is large so model tends to predict 0 (negative) for all cases leading to False Negative problem. We should minimize that. My opinion

Comment: C,D is correct (percentage of the positive class is key to decide which case we are interested in) This question, positive class (Pay) is 0.01% as compared to 99.99(not pay) , as a result, we have to pay attention to Pay because if we miss 0.01% out, we didn't get revenue. it is a false negative. In contrast to these questions, it positive class (Pay) is 40% as compared to negative class (60% not pay), it is avoidable to emphasize on 40% (if model predict as payment but in reality customer neglect), we won't get revenue the amount from false positive)

Comment: I think is CD

Comment: C and D. Hopefully, no one honestly thinks that B is a good answer. Never expose test data to the training set or vice versa. C is right because of the highly imbalanced training set. D is right because you want to minimize false negatives, maximize true positives, maximize recall of the positive class. I'm not sure why anyone's worried about precision in this case.

Comment: CD The model has 99% accuracy because it's simply predicting that everyone's a negative. Since almost everyone's a negative, it will get almost everyone right. So we need to penalize the model for predicting that someone is a negative when it is not (i.e. penalize false negatives). So that's D. Also, it would be really nice to have more positives -- one way to do that is to follow option C.

Comment: CD 100%

Comment: CD C:imbalance of test (1000 positive, 999000 negative = 0.1% positive) thus C to increase that D :also to reduce generalizing, since everyone says no, the model would generalize to no, but increasing the penalty of a false negative would reduce generalizing..

Comment: It is needed to diminish the FP, because they are player predicted to pay and in reality will not pay. So FP should impact the cost metric more. CE should be the answer.

Comment: CD are correct for sure.

Comment: It is C,E... we want to find all paying customers, which are positives, so we have to punish incorrectly finding negatives, which is E

Comment: CD although i am worried about the noise being introduced as it could skew the data nevertheless no better answer is given

Comment: CD We need high recall so that we do not miss many Positive cases. In that case we need to have less False Negative(FN) therefore it should have high impact on cost function.

Comment: in my view, CD are answers C: of course, handle the imbalanced dataset D: right now, model accuracy is 99%, it means model predict everything is negative leading to FN problem, so we need to minimize it more in cost function

Comment: CD, FN are valuable players, we should care more on FN

Comment: Is my assumption right here? ACTUAL ---------- P PAY NPAY R ---------- E PAY TP FP D I NPAY FN TN C ----------

Comment: C,D correct

Discussion for Question 34

Link: https://www.examtopics.com/discussions/amazon/view/10054-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 18 votes
- D: 13 votes

Discussion

Comment: Dropping the Age feature is a NOT ATOLL a good idea - as age plays a critical role in this disease as per the question Dropping 10% of data is NOT a good idea considering the fact that the number of observations is already low. The Mean or Median are a potential solutions But the question says that "Disease worsens after age 65 so there is a correlation between age and other symptoms related feature" So that means that using Unsupervised Learning we can make pretty good prediction of "Age" So the answer is D Use K-Means clustering

Replies:

Comment: https://www.displayr.com/5-ways-deal-missing-data-cluster-analysis/ B is correct

Comment: Replacing the age with mean or median might bring a bias to the dataset. Use k-means clustering to estimate the missing age based on other features might get better results. Removing 10% available data looks odd. Why not D?

Comment: B. Replace the age field value for records with a value of 0 with the mean or median value from the dataset: This method allows for retaining all patient records while addressing the anomaly. It is a standard approach for dealing with missing or incorrect values in a way that preserves the integrity of the dataset. B. GPT answer

Comment: B-chatgpt

Comment: The question tries to mislead by adding information around the feature correlation. K-means clustering is not meant for imputing data. Hence answer should be B, that would be the right way of handling the missing value.

Comment: Using k-means clustering to handle missing features is not directly applicable to this scenario. K-means clustering is a method for grouping data points into clusters based on similarity, and it's not typically used for imputing missing values.

Comment: add/ comment why? b ? - >replacing the age field value for records with a value of 0 with the mean or median value from the dataset, is generally the best approach among the given options. It allows the preservation of the dataset size and leverages the remaining correct data points, assuming age is a crucial predictor in this context. However, it's vital to perform this imputation carefully to avoid introducing bias. Median is often preferred in this scenario to mitigate the impact of outliers.

Comment: The best way to handle the missing values in the patient age feature is to replace them with the mean or median value from the dataset. This is a common technique for imputing missing values that preserves the overall distribution of the data and avoids introducing bias or reducing the sample size. Dropping the records or the feature would result in losing valuable information and reducing the accuracy of the model. Using k-means clustering would not be appropriate for handling missing values in a single feature, as it is a method for grouping similar data points based on multiple

Comment: mean or median is for outliers so D

Comment: Obviously B, why would you use a clustering algorithm to predict a value? D just doesn't make sense

Comment: B is correct.K-means is unsupervised and used mainly for clustering. KNN would have been more accurate. It can be used to predict a value. since knn is not present i think it is mean median value

Comment: B is correct or KNN, but dont K means

Comment: A. NO - unless we want to loose 10% of the data B. NO - age is predictive, so using the means we would introduce a bias C. NO - age is predictive D. YES - better quality than B, it is likely that other physiological values can help predict the age

Comment: k-means should give the best estimation of the age. Using mean would reduce the correlation between outcome and age for the model.

Comment: How can it be when there is a labelled outcome, which means this is Supervised and K-Means is for UnSupervised. So only possible answer should be B

Comment: Both A and B are correct. But, I noted that at the end of the question is mentioned that all other features are OKAY, so is reasonable to do this simple imputation.

Comment: B is correct, K-NN could have helped instead of k-means

---

Discussion for Question 35

Link: https://www.examtopics.com/discussions/amazon/view/10056-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 5 votes

Discussion

Comment: Ans: A (S3) is most cost effective

Comment: A : S3 cost effective + athena (not c redshift dont support unstructured data)

Comment: 'cost effective' --> AWS S3

Comment: A. YES - S3 + Athena/Presto B. NO - no SQL support C. NO - expensive to scale D. NO - DynamoDB is NoSQL

Comment: AWS S3 + Athena will do it

Comment: The most appropriate storage scheme for this scenario is option A: Store datasets as files in Amazon S3. Amazon S3 is a highly scalable and cost-effective object storage service that can store a large amount of data. S3 can scale automatically to accommodate a large number of datasets, making it a good option for storing the training data used in machine learning models. Additionally, S3 supports SQL querying through Amazon Athena or Amazon Redshift Spectrum, allowing data scientists to easily explore the data.

Comment: "store a large amount of training data commonly used in its machine learning models".. well it cannot be anything other than S3. Athena can query S3 cataloged data with SQL commands. Anwser is A

Comment: Amazon Redshift is not cost-effective.

Comment: I would say C https://docs.aws.amazon.com/redshift/latest/mgmt/working-with-clusters.html "For workloads that require ever-growing storage, managed storage lets you automatically scale your data warehouse storage capacity without adding and paying for additional nodes."

Replies:

> **Comment:** Amazon Redshift is a fast, fully managed data warehouse that makes it simple and cost-effective to analyze all your data using standard SQL and your existing Business Intelligence (BI) tools. It allows you to run complex analytic queries against petabytes of structured data using sophisticated query optimization, columnar storage on high-performance storage, and massively parallel query execution. Most results come back in seconds.
>
> **Comment:** Data warehouse is not needed. For exploring data using SQL, you can use Athena

Comment: s3 is right

Comment: A, S3 is most cost effective

Discussion for Question 36

Link: https://www.examtopics.com/discussions/amazon/view/10057-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- D: 5 votes

Discussion

Comment: Ans: D

Comment: 'retrained using the original training data plus new data'

Comment: I believe it should be B 1. The model performance has diminished gradually over the past few months, indicating the data distribution may have changed since initial deployment over a year ago. This is a classic sign of concept drift. 2. The model architecture and training procedure have remained unchanged since initial deployment. Updating the hyperparameters is a lighter approach than retraining the model from scratch, and can help prevent further performance deterioration if done periodically to adapt to changes in user preferences and product inventory.

Comment: Answer is D

Comment: D is the answer!

Comment: D is the answer. There has been a data drift resulting from new customer segment visiting the site. So, the model needs to be updated periodically with new data from the website.

Comment: DDDDD. D :D

Comment: Incremental training. D.

Comment: Periodically Re-Fit D

Comment: agree with D

Comment: D is correct

Discussion for Question 37

Link: https://www.examtopics.com/discussions/amazon/view/10058-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 9 votes

Discussion

Comment: Ans: A seems to be reasonable

Comment: A looks correct but it is missing for "Interactive analytics of historical data"

Replies:

> **Comment:** but C is missing for "real-time analythics"

Replies:

Comment: and also C is saying historical data analytics for Kinesis Data analytics which is real-time analytics not historical, so the answer might not C but the answer is A

Comment: Once you insert real-time data to ES, you can see historical data from Kibana dashboard.

Comment: AWS Glue as data catalog, then you can analyze historical data, such as running sql with Athena.

Comment: A. YES - Amazon Kinesis Data Analytics is for real-time data insights B. NO - Amazon Athena has no data catalog C. NO - Amazon Kinesis Data Streams and Amazon Kinesis Data Analytics is not for historical data insights D. NO - Amazon Athena has no data catalog

Comment: Athena can not be used for data catalog, so B and D are wrong. A and C are equals, but it's well known that Kinesis DS and Analytics are used together for real time solutions, which is mentioned in the question / answer, but lack on C.

Comment: All are bad options, but A can do it.

Comment: AWS Glue is a fully managed extract, transform, and load (ETL) service that makes it easy to move data between data stores. It can be used as a data catalog to store metadata information about the data in the data lake. Amazon Kinesis Data Streams and Amazon Kinesis Data Analytics can be used together to collect, process, and analyze real-time streaming data. Amazon Kinesis Data Firehose can be used to deliver streaming data to destinations such as Amazon ES for clickstream analytics. Finally, Amazon EMR can be used to run big data frameworks such as Apache Spark and Apache Hadoop to generate personalized product recommendations.

Comment: A or C https://aws.amazon.com/blogs/big-data/retaining-data-streams-up-to-one-year-with-amazon-kinesis-data-streams/

Comment: Athena can do Interactive analytics on Historical data, but here its only use is "Athena as the data catalog" and this is the work of Glue data catalog using its crawlers, so it cannot be B or D. --So its either A or C -- Now Kinesis data Streams/Analytics is know for real time data analytics but if it is reading from data already stored in S3 using DMS then we can say it is getting historical data. -- Here I am not very clear if Kinesis part will happen on incoming data before S3 or After data persists to S3 and Kinesis reads it through S3-->DMS--Kinesis data stream -- Kinesis analytics-->Firehose. But still insights are always on real-time/current data based on historical data trends , so the statement in C "Analytics for historical data insights" is in-correct in general . Hence ANSWER is :A

Comment: A is correct, for those asking the difference between A and D, D talks about using kinesis stream and data analytics to create historical analysis.... waste of money no?

Comment: Answer = A

Comment: A it is

Comment: it's A, ES can perform clickstream analytics and EMR can handle spark job recomendation at scale

Comment: Only C and D mention interactive analytics of historical data. Glue won't provide personalised recommendation so it is C

Comment: What is the difference between the solution in A or C ????

Replies:

Comment: A is real time data analytics with Kinesis Data analytics and C is saying historical data which is wrong

Comment: Looks like C Amazon ES has KIbana which supports click stream

Replies:

Comment: A is Correct

Discussion for Question 38

Link: https://www.examtopics.com/discussions/amazon/view/8317-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- CD: 5 votes

Discussion

Comment: You might be spent a lot of money for ask AWS A.CHANGE built-in image OR B.Create a support case. The effectial way BOTH RELATIVE TO SageMaker Estimator C.DOCKER OR BRING YOUR CODE BY D.SageMaker with TensorFlow Estimator THE BEAUTYFUL ANSWER ARE C AND D

Comment: I will go for C & D

Comment: Option A is not possible because the built-in image classification algorithm cannot be customized. Option B is not feasible because it is not possible to change the default image classification algorithm through a support case. Option E is also not a recommended approach because it involves manually installing software on an EC2 instance rather than using the managed services provided by SageMaker.

Comment: The effectial way BOTH RELATIVE TO SageMaker Estimator C.DOCKER OR BRING YOUR CODE BY D.SageMaker with TensorFlow Estimator

Comment: This question ask for 2 ways not a set of actions. So may be confused.

Comment: Anwers AD go to: https://docs.aws.amazon.com/sagemaker/latest/dg/docker-containers.html

Comment: CD and also A says it but in a more general term....

Comment: https://aws.amazon.com/blogs/machine-learning/transfer-learning-for-custom-labels-using-a-tensorflow-container-and-bring-your-own-algorithm-in-amazon-sagemaker/

Comment: C and D are correct

Comment: https://docs.aws.amazon.com/zh_tw/sagemaker/latest/dg/your-algorithms.html https://aws.amazon.com/tw/blogs/machine-learning/transfer-learning-for-custom-labels-using-a-tensorflow-container-and-bring-your-own-algorithm-in-amazon-sagemaker/ https://docs.aws.amazon.com/zh_tw/sagemaker/latest/dg/tf.html

---

Discussion for Question 39

Link: https://www.examtopics.com/discussions/amazon/view/8318-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 19 votes

Discussion

Comment: DROPOUT HELPS PREVENT OVERFITTING https://keras.io/layers/core/#dropout THE BEAUTIFUL ANSER SHOULD BE B.

Replies:

Comment: agree. it should be B

Comment: https://kharshit.github.io/blog/2018/05/04/dropout-prevent-overfitting Answer is B 100%

Comment: Increasing dropout rate will reduce complexity of the model which inturn reduces overfitting

Comment: This is clearly B, dont get why the answer is marked as D.

Comment: Regularization will seek to obtain similar accuracies in train and test sets. Anything else will make the overfitting worse

Comment: B is correct, D so stup*d answer

Comment: A. NO - accuracy on training set is high B. YES - increased dropout rate => reduce model complexity => less overfitting C. NO - we want to reduce model complexity D. NO - the model converged

Comment: I don't understand why the highlighted "right" answer is D. To increase the number of epochs will make the situation even worse than it is; dropout is the right action to take in this case

Comment: B is correct

Comment: agree, B makes more sense here

Comment: Definitely B because overfitting comes from complex model that captures patterns of training data well. But D is getting this model more complex, worsening overfitting.

Replies:

> **Comment:** Correct my reasoning! D is worsening overfitting because it feeds more data after overfitting arises. D is used for underfitted models.

Comment: Increasing Epoch only makes things worse on a overfitting model. You should perform regularization by introducing drop outs to generalize the model.

Comment: Option B is the correct answer because increasing the dropout rate at the flatten layer helps prevent overfitting by randomly dropping out units during training, effectively creating a more robust model that can generalize better to new data. Dropout is a regularization technique that helps prevent overfitting by forcing the model to learn redundant representations of the data. By increasing the dropout rate at the flatten layer, the model becomes more generalized, which should help to improve the testing accuracy.

Comment: Overfitting occurs when a model is too complex and memorizes the training data instead of learning the underlying pattern. As a result, the model performs well on the training data but poorly on new, unseen data. Increasing the dropout rate, a regularization technique, can help combat overfitting by randomly dropping out some neurons during training, which prevents the model from relying too heavily on any single feature.

Comment: Model is overfitting, I will go with option B, increasing epoch will cause more overfitting

Comment: it should answer B.

Comment: 12-sep exam

Discussion for Question 40

Link: https://www.examtopics.com/discussions/amazon/view/8319-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: THE ANSWER SHOULD BE B. YOU DON'T NEED TO THROUGH LAMBDA TO INTERGE CLOUDTRAIL Log Amazon SageMaker API Calls with AWS CloudTrail https://docs.aws.amazon.com/sagemaker/latest/dg/logging-using-cloudtrail.html

Comment: Agreed B for the following reasons # CloudTrail logs captured in S3 without any code/lambda # The custom metrics can be published to Cloudwatch...in this case it would be a test for overfit on MXNET .... which will set off an alarm .... which can then be subscribed on SNS

Comment: https://docs.aws.amazon.com/fr_fr/sagemaker/latest/dg/training-metrics.html#define-train-metrics It detects hardware resource usage issues (such as CPU, GPU, and I/O bottlenecks) and non-convergent model issues (such as overfitting, disappearing gradients, and tensor explosion). why couldn't the answer be D, as this covers all of the requirements, and B seems to add an extra step with adding push code, when it already has a builtin metric for overfitting.

Replies:

> **Comment:** Custom metric Need to built and pushed.

Comment: A. NO - CloudTrail has built-in SageMaker API calls tracking, no lambda needed B. YES - the chain works C. NO - CloudTrail has built-in SageMaker API calls tracking, no lambda needed D. NO - CloudTrail has not specific Amazon SageMaker integration to detect overfitting

Comment: Option B

Comment: "least amount of code and fewest steps?" I think it's D.

Replies:

> **Comment:** Agreed, with less code effort.

Comment: I would consider D as well. You can just setup a SNS that is triggered by a built-in action like here: https://docs.aws.amazon.com/sagemaker/latest/dg/debugger-built-in-actions.html You can see that overfitting is a built-in rule for MXNet from here: https://docs.aws.amazon.com/sagemaker/latest/dg/debugger-built-in-rules.html Not that B is not working. Maybe the question was prior to this new solution.

Replies:

> **Comment:** The loss_not_decreasing, overfit, overtraining, and stalled_training_rule monitors if your model is optimizing the loss function without those training issues. If the rules detect training anomalies, the rule evaluation status changes to IssueFound. You can set up automated actions, such as notifying training issues and stopping training jobs using Amazon CloudWatch Events and AWS Lambda. For more information, see Action on Amazon SageMaker Debugger Rules. https://docs.aws.amazon.com/sagemaker/latest/dg/use-debugger-built-in-rules.html

Comment: It's B.

Comment: AWS CloudTrail provides a history of AWS API calls made on the account. The Machine Learning team can use AWS CloudTrail to log Amazon SageMaker API calls to Amazon S3. They can then use CloudWatch to create alarms and receive notifications when the model is overfitting. To ensure auditors can view the Amazon SageMaker log activity report, the team can add code to push a custom metric to Amazon CloudWatch. This provides a single place to view and analyze logs across all the services and resources in the environment.

Comment: B. cloudwatch + metrics from sagemaker + sns https://docs.aws.amazon.com/fr_fr/sagemaker/latest/dg/training-metrics.html#define-train-metrics

Comment: B requires the least amount of code and satisfies all conditions

Comment: What does this line do? "Add code to push a custom metric to Amazon CloudWatch"

Replies:

> **Comment:** It creates a metric for overfitting (accuracy of training data and accuracy of test data).

Comment: Its not B. Why would you use CloudTrail? Having used Lambda for API calls I'm inclined to agree with the original answer, C.

Replies:

> **Comment:** https://docs.aws.amazon.com/sagemaker/latest/dg/logging-using-cloudtrail.html

> **Comment:** Because that is the only job of CloudTrail - to log actions taken on your AWS account. So why need a Lambda function to trigger it?

Comment: B it is

Comment: B it is

Comment: Agree on B

Comment: ALL AWS Service's API calls are logged to CloudTrail automatically.

Discussion for Question 41

Link: https://www.examtopics.com/discussions/amazon/view/11851-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 6 votes

Discussion

Comment: C is correct

Comment: You want to reduce features/dimension so PCA is the answer

Comment: C is the way

Comment: C is correct. D could be correct if the correlation is used to omit features.

Comment: PCA and T-SNE are for solving the curse of dimensionality mencioned here!

Comment: I assume PCA is for unsupervised learning!...and the scenario in the question looks like supervised learning

Replies:

Comment: data (x, y) --> (PCA) --> preprocessed data(x', y) --> learning why not for supervised learning?

Comment: Tricky. The sentence 'many features are highly correlated with each other' is no use.

Replies:

Comment: It's. PCA removes such correlation.

Comment: Answer C: Read through this carefully "What should be done to reduce the impact of having such a large number of features?" only answer comes in mind PCA

Comment: Of course, it's PCA.

Comment: PCA is the solution. So, answer is C

Discussion for Question 42

Link: https://www.examtopics.com/discussions/amazon/view/8339-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 17 votes

Discussion

Comment: A If you have information about the average (mean) number of things that happen in some given time period / interval,Poisson distribution can give you a way to predict the odds of getting some other value on a given future day

Comment: Ans: A https://brilliant.org/wiki/poisson-distribution/

Comment: A is correct

Comment: Poisson distribution is discrete, and gives the number of events that occur in a given time interval

Comment: A. YES - Poisson distribution is discrete, and gives the number of events that occur in a given time interval B. NO - Uniform distribution is continuous, we want discrete C. NO - Normal distribution is continuous we want discrete D. NO - Binomial distribution give the probability that a random variable is A or B (possibly in with different weight)

Comment: Option A indeed

Comment: ANSWER IS A https://www.investopedia.com/terms/d/discrete-distribution.asp

Comment: The Poisson distribution is commonly used for count data, which is the case here as we are interested in the number of minutes New Yorkers wait for a bus. The Poisson distribution is characterized by a single parameter, lambda, which represents the mean and variance of the distribution. In this case, the mean is 3 minutes, so we would set lambda to 3. The Poisson distribution assumes that events occur independently of each other, which is a reasonable assumption in this case since the waiting time for each individual is likely to be independent of the waiting time for others.

Comment: The Poisson distribution is a discrete probability distribution that is commonly used to model the number of events that occur in a fixed interval of time, given an average rate of occurrence. Since the buses cycle every 10 minutes and the mean wait time is 3 minutes, it is reasonable to assume that the number of minutes New Yorkers wait for a bus can be modeled by a Poisson distribution.

Comment: 100% A, as discrete, while binomial has to be binary data (success or failure)

Comment: A is a discrete distribution

Comment: I do choose Poisson. A.

Comment: 12-sep exam

Comment: Answer is A .. these types on footfalls ,etc ..answer always Poisson-distribution

Comment: I agree that the answer is A: because the Poisson distribution is a probability distribution that is used to show how many times that buses are likely to occur over a specified period.

Comment: Definitely A, Poisson.

Comment: A!!!!!!!!

Discussion for Question 43

Link: https://www.examtopics.com/discussions/amazon/view/8343-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 5 votes

Discussion

Comment: NAT gateway COULD GO OUT TO THE INTERNET AND DOWNLOAD BACK MALICIOUS D. IS NOT A GOOD ANSWER. THE SAFE ONE IS ANSWER C. ASSOCIATE WITH VPC_ENDPOINT AND S3_ENDPOINT

Comment: C is correct We must use the VPC endpoint (either Gateway Endpoint or Interface Endpoint)to comply with this requirement "Data communication traffic must stay within the AWS network". https://docs.aws.amazon.com/sagemaker/latest/dg/notebook-interface-endpoint.html

Comment: A. NO - We don't place a S3 bucket in a VPC, it is always in AWS Service Account B. NO - without an S3 VPC endpoint, traffic will go through the Internet C. YES - we need endpoints for both SageMaker and S3 to avoid Internet traffic D. NO - we need endpoints for both SageMaker and S3 to avoid Internet traffic

Comment: Option C

Comment: C is the correct. A is not so correct, because it's possible to communicate two different VPCs inside AWS network (which is not optimized).

Comment: This configuration would meet the company's requirements for security, as the notebook instance would be placed within a private subnet in a VPC, and data communication traffic would stay within the AWS network through the use of VPC endpoints for S3 and Amazon SageMaker. Additionally, the VPC would not have internet access, further reducing the security risk.

Comment: C - "and data communication traffic must stay within the AWS network." that discards D

Comment: Answer should be C. Because, Security team don't want Internet Access, Option-D has NAT and will get to Internet somehow. Also connecting S3 and SageMaker EC2 instance via VPC endpoints is best way to secure the resources.

Comment: Using a NAT gateway is the old way to do it. Option C is the way to do it now. https://cloudacademy.com/blog/vpc-endpoint-for-amazon-s3/#:~:text=Accessing%20S3%20the%20old%20way%20%28without%20VPC%20Endpoint%29,has%20no%20access%20to%20any%20outside%20public%20resources

Comment: "and data communication traffic must stay within the AWS network" , NAT gateway will always go over the Internet to access S3.with NAT you can put your instances in private subnet and NAT itself in public subnet , but still in order to access S3 it will go over the internet. SO answer cannot be D. -- C is the only correct option here , as S3 VPC endpoints is a real thing "google it" and it sole purpose is to create route from VPC endpoint to S3 , without going over the Internet.

Comment: C is correct answer. D is only applicable -"If your model needs access to an AWS service that doesn't support interface VPC endpoints or to a resource outside of AWS, create a NAT gateway and configure your security groups to allow outbound connections. " https://docs.aws.amazon.com/sagemaker/latest/dg/host-vpc.html

Comment: D is correct

Comment: Answer is D, read third paragraph https://docs.aws.amazon.com/sagemaker/latest/dg/appendix-notebook-and-internet-access.html

Comment: NAT is the way that a VPC connect to internet and other ASW service when there is NO INTERNET ACCESS FOR VPC. Thus the answer is D.

Comment: "concerned that internet-enabled notebook instances create a security vulnerability where malicious code running on the instances could compromise data privacy." NAT Gateway does not mitigate this risk!

Replies:

> **Comment:** This is the correct answer. If this answer is confusing, study some of the associate exams before going for this one. VPC endpoint and NAT gateway are similar, but NAT gateway is for giving resources in the VPC the chance to initiate connections with the internet, whereas a VPC endpoint only allows it to go to other AWS services, which is the best solution for this question.

Comment: C: If you configure your VPC so that it doesn't have internet access, models that use that VPC do not have access to resources outside your VPC. If your model needs access to resources outside your VPC, provide access with one of the following options: If your model needs access to an AWS service that supports interface VPC endpoints, create an endpoint to connect to that service. For a list of services that support interface endpoints, see VPC Endpoints in the Amazon VPC User Guide. For information about creating an interface VPC endpoint, see Interface VPC Endpoints (AWS PrivateLink) in the Amazon VPC User Guide. If your model needs access to an AWS service that doesn't support interface VPC endpoints or to a resource outside of AWS, create a NAT gateway and configure your security groups to allow outbound connections. For information about setting up a NAT gateway for your VPC, see Scenario 2: VPC with Public and Private Subnets (NAT) in the Amazon Virtual Private Cloud User Guide.

Comment: what is the difference between A & C? are both answers OK?

Replies:

> **Comment:** It is not enough for sagemaker to communicate to S3 if both of them are inside the same VPC. Sagamaker inside a VPC needs to create a endpoint to connect to other AWS services which has endpoint too.

Discussion for Question 44

Link: https://www.examtopics.com/discussions/amazon/view/8348-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- BCF: 12 votes

Discussion

Comment: Yes, answer is BCF

Comment: Go for BCF

Comment: I think here the point is around the definition of "feature combinations". If you refer to it as "combine the features to generate a smaller but more effective feature set" this would end up to a smaller feature set thus a good thing for overfitting. However, if you refer to it as "combine the features to generate additional features" this would end up to a larger feature set thus a bad thing for overfitting. Also, in some cases you implement feature combinations in your model (see hidden layers in feed-forward network) thus increasing model complexity which is bad for overfitting. To me this question is poorly worded. I would pick F as my best guess is that you need to implement feature combination in your model, thus decreasing feature combination decrease complexity hence improving with overfitting issue

Replies:

> **Comment:** Great callout - what exactly the Feature combination is performing has not been elaborated It can be: Using PCA or t-SNE, it is essentially optimizing features - good to address overfitting, and should be done Or, it can be: Using Cartesian Product, features are being combined to create additional features - this will aid overfitting and should NOT be done. Wish questions and answer options are written clearly so that there is no room for ambiguity. Especially, taking into account that in real life, these kind of communication/write-up will trigger follow-up questions until addressed satisfactorily.

Comment: About option E: When increasing feature combinations, the goal is not to simply add more features indiscriminately, which could indeed lead to overfitting. Instead, it involves selecting and combining features in a way that captures important patterns and relationships in the data. When done effectively, increasing feature combinations can help the model generalize better to unseen data by providing more informative and discriminative features, thus reducing the risk of overfitting.

Comment: If your model is overfitting the training data, it makes sense to take actions that reduce model flexibility. To reduce model flexibility, try the following: Feature selection: consider using fewer feature combinations, decrease n-grams size, and decrease the number of numeric attribute bins. Increase the amount of regularization used. https://docs.aws.amazon.com/machine-learning/latest/dg/model-fit-underfitting-vs-overfitting.html

Comment: Best choices are B (Increase regularization), C (Increase dropout), and F (Decrease feature combinations), as these techniques are effective in reducing overfitting and improving the model's ability to generalize to new data.

Comment: BCE The model has learnt training data. One approach is to increase complexity by increasing the features or remove some features to increase bias. In deep learning, i thinking increasing feature set is more workable.

Comment: B-C-F. All of those options can be used to reduce model complexity and thus: overfit

Comment: its BCF

Comment: BCF is correct.

Comment: Increasing regularization helps to prevent overfitting by adding a penalty term to the loss function to discourage the model from learning the noise in the data. Increasing dropout helps to prevent overfitting by randomly dropping out some neurons during training, which forces the model to learn more robust representations that do not depend on the presence of any single neuron. Decreasing the number of feature combinations helps to simplify the model, making it less likely to overfit.

Comment: I see all the comments for BCF, although when you look at F it just says decrease 'feature combinations', not features themselves. In one way to decrease feature combinations results in having more features (less feature engineering), which in turn will cause more overfitting. Unless the question in badly worded, saying less feature combinations just mean those combinations, which components will not be used, then it has to be BCE.

Replies:

> **Comment:** Decrease feature combinations - too many irrelevant features can influence the model by drowning out the signal with noise

> **Comment:** Increasing the number of feature combinations can sometimes improve the performance of a model if the model is underfitting the data. However, in this context, it is not likely to be a solution to overfitting.

Comment: BCF - Always remember in case of overfitting - reduce features, Add regularisation and increase dropouts.

Comment: BCE: The main objective of PCA (technic to feature combination) is to simplify your model features into fewer components to help visualize patterns in your data and to help your model run faster. Using PCA also reduces the chance of overfitting your model by eliminating features with high correlation. https://towardsdatascience.com/dealing-with-highly-dimensional-data-using-principal-component-analysis-pca-fea1ca817fe6

Replies:

Comment: AWS Documentation explicitly mentions reducing feature combinations to prevent overfitting - https://docs.aws.amazon.com/machine-learning/latest/dg/model-fit-underfitting-vs-overfitting.html It's B C F

Comment: B/C/F Easy peasy.

Comment: BCF 100%

Comment: BCF F explained in AWS document: Feature selection: consider using fewer feature combinations, decrease n-grams size, and decrease the number of numeric attribute bins. Increase the amount of regularization used https://docs.aws.amazon.com/machine-learning/latest/dg/model-fit-underfitting-vs-overfitting.html

---

Discussion for Question 45

Link: https://www.examtopics.com/discussions/amazon/view/8351-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 12 votes

Discussion

Comment: Kinesis Data Analytics NO PARQET FORMAT, BESIDES THAT JSON NO NEED TO STORE IN S3. RDS ISN'T serverless ingestion and analytics solution ANSWER IS A.

Comment: I thinks it should be A please check https://aws.amazon.com/blogs/big-data/analyzing-apache-parquet-optimized-data-using-amazon-kinesis-data-firehose-amazon-athena-and-amazon-redshift/

Comment: A. Create a schema in the AWS Glue Data Catalog of the incoming data format. Use Amazon Kinesis Data Firehose to buffer and transform the streaming JSON data to a columnar format like Apache Parquet or ORC using the AWS Glue Data Catalog before delivering to Amazon S3. Analysts can then query the data using Amazon Athena and connect to BI dashboards using the Athena JDBC connector. This solution is serverless, manages high-velocity data streams, supports SQL queries, and connects to BI tools—all while being highly available.

Comment: A. YES - we need a catalog to create parquet (https://docs.aws.amazon.com/firehose/latest/APIReference/API_SchemaConfiguration.html) B. NO - no need for extra staging C. NO - no need for extra staging D. NO - we need a catalog

Comment: Option A

Comment: A is correct. For those selecting B, answer me: how exactly the json will be stored in the S3? It's not mentioned in the answer. For me it's an incomplete solution.

Comment: This solution leverages AWS Glue to create a schema of the incoming data format, which helps to buffer and convert the records to a query-optimized, columnar format without data loss. The Amazon Kinesis Data Firehose delivery stream is used to stream the data and transform it to Apache Parquet or ORC format using the AWS Glue Data Catalog, and the data is stored in Amazon S3, which is highly available. The Analysts can then query the data directly from Amazon S3 using Amazon Athena, and connect to BI tools using the Athena JDBC connector. This solution provides a serverless, scalable, and cost-effective solution for real-time streaming data ingestion and analytics.

Comment: Since you want to buffer and convert data so A is correct answer. No other option is fulfilling this requirement

Comment: I go for A. However, I am not sure why AWS Glue is very important here given that Firehose can convert JSON to parquet.

Replies:

Comment: If I haven't remembered correctly. Athena requires a schema of the S3 object to perform SQL query. That's probably why we need Glue for the schema

Replies:

Comment: once you ingest the data using Kinesis Firehose, you can set "generate table" and automatically create Glue schema. I think both Glue and Firehose can do data conversion from JSON to parquet.

Comment: Why AWS Glue is needed? Firehose could convert to parquet directly...

Comment: Kinesis Data Analytics is near real-time, not real time

Comment: Answer is ”A”

Comment: The difference between "real-time" and "near-real-time" is pretty semantic(60s). The fact that the data comes through kinesis data streams (real time) is implied as the only valid input to firehose.

Replies:

Comment: Mind you, "the ingestion process must buffer and transform incoming records from JSON to a query-optimized, columnar format" That is exactly what kinesis firehose does. "Kinesis Data Firehose buffers incoming data before delivering it to Amazon S3. You can configure the values for S3 buffer size (1 MB to 128 MB) or buffer interval (60 to 900 seconds), and the condition satisfied first triggers data delivery to Amazon S3." See link: https://aws.amazon.com/kinesis/data-firehose/faqs/#:~:text=Kinesis%20Data%20Firehose%20buffers%20incoming,data%20delivery%20to%20Amazon%20S3.

Comment: Data Firehose is always Near Real Time not Real Time. The prompt clearly states that process must be done in Real Time.

Comment: Why A? Firehose is near real-time, and not real-time which is a requirement

Replies:

Comment: There is no requirement for real time processing. It says the data is in real time but the processing of that data should buffer

Comment: ANSWER is A -- and every statement in it is accurate. Firehose does integrate with GLue data catalog and it also "Buffers" the data . "When Kinesis Data Firehose processes incoming events and converts the data to Parquet, it needs to know which schema to apply." This is achived by glue data catalog and athena and it works on real-time data ingest.See link below. https://aws.amazon.com/blogs/big-data/analyzing-apache-parquet-optimized-data-using-amazon-kinesis-data-firehose-amazon-athena-and-amazon-redshift/

Comment: https://aws.amazon.com/blogs/aws/new-serverless-streaming-etl-with-aws-glue/ A is the answer imo

---

Discussion for Question 46

Link: https://www.examtopics.com/discussions/amazon/view/10080-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 8 votes

Discussion

Comment: C looks correct since multiple imputation can be performed based on the related variable as given in the question

Comment: Multiple Imputation by Chained Equations or MICE, as per udemy this is always the best answer of all

Comment: Why not D: Doesn't Account for Relationships: Mean substitution doesn't take into account the potential relationships between variables. In the scenario you provided, it's believed that other columns could help in reconstructing the missing data. Using only the mean of the missing column doesn't leverage this potential inter-column relationship. Assumption of Missing Completely at Random (MCAR): Mean substitution often operates under the assumption that the data is Missing Completely at Random (MCAR). In reality, data might be missing for a reason, and that reason might relate to other observed variables. Using mean substitution in such cases can introduce biases.

Comment: A. NO - Listwise deletion is just dropping rows B. NO - does not reconstruct the data based on other fields C. YES - by definition D. NO - does not reconstruct the data based on other fields

Comment: MICE is the algorithm to choose here

Comment: Option C

Comment: Multiple imputation is a statistical technique for handling missing data that involves generating multiple versions of the dataset with missing values filled in, and then combining the results to produce a single, complete dataset. This approach takes into account the relationship between variables in the dataset, and uses statistical models to predict missing values based on the information in other columns. This helps to preserve the integrity of the dataset by

avoiding the introduction of bias or systematic error into the results.

Comment: I am trying to understand why Mean Substitution is not the solution. Imputation typically uses the mean if the missing data is random, implying the substitution is not biased.

Replies:

> **Comment:** Reason is if you replace 30% of the missing values , likely you will bias the variable.
>
> **Comment:** Mean substitution is limited to the current column. In this case, the requirement is to impute missing data from other columns

Comment: If it's handling missing data then imputation comes into play Answer is C 100%

Comment: https://www.countants.com/blogs/heres-how-you-can-configure-automatic-imputation-of-missing-data/ C

Comment: it's C

Comment: A common strategy used to impute missing values is to replace missing values with the mean or median value. It is important to understand your data before choosing a strategy for replacing missing values. https://docs.aws.amazon.com/machine-learning/latest/dg/feature-processing.html

---

Discussion for Question 47

Link: https://www.examtopics.com/discussions/amazon/view/8370-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 15 votes

Discussion

Comment: NAT CLOUD GO OUT TO THE INTERNET, IT STILL CANNOT PREVENT DOWNLOAD MALICIOUS BY YOURSELF. THE RIGHT ANSWER IS C. C.INTERFACE VPC ENDPOINT https://docs.aws.amazon.com/sagemaker/latest/dg/sagemaker-dg.pdf (516) https://docs.aws.amazon.com/zh_tw/vpc/latest/userguide/vpc-endpoints.html

Replies:

> **Comment:** Not sure if C is correct in this particular scenario. From https://docs.aws.amazon.com/sagemaker/latest/dg/sagemaker-dg.pdf Page 202 of the SageMaker Guide has: If you allowed access to resources from your VPC, enable direct internet access. For Direct internet access, choose Enable. Without internet access, you can't train or host models from notebooks on this notebook instance unless your VPC has a NAT gateway and your security group allows outbound connect
>
> ### Replies:
>
> > **Comment:** A may the right answer
> >
> > **Comment:** There are two possible solutions, but the safer solution and easier is trough VPC endpoints. You can connect to your notebook instance from your VPC through an interface endpoint in your Virtual Private Cloud (VPC) instead of connecting over the internet. When you use a VPC interface endpoint, communication between your VPC and the notebook instance is conducted entirely and securely within the AWS network. And there is not problem that the notebooks does not have public internet. Because Amazon SageMaker notebook instances support Amazon Virtual Private Cloud (Amazon VPC) interface endpoints that are powered by AWS PrivateLink. Each VPC endpoint is represented by one or more Elastic Network Interfaces (ENIs) with private IP addresses in your VPC subnets. Each VPC endpoint is represented by one or more Elastic Network Interfaces (ENIs) with private IP addresses in your VPC subnets... so the Answer is C.

Comment: C is correct. "The VPC interface endpoint connects your VPC directly to the Amazon SageMaker API or Runtime without an internet gateway, **NAT** device, VPN connection, or AWS Direct Connect connection." https://docs.aws.amazon.com/sagemaker/latest/dg/interface-vpc-endpoint.html

Comment: The answer is C. - If you want to allow internet access, you must use a NAT gateway with access to the internet, for example through an internet gateway. - If you don't want to allow internet access, create interface VPC endpoints (AWS PrivateLink) to allow Studio Classic to access the following services with the corresponding service names. You must also associate the security groups for your VPC with these endpoints. This is exactly what's written in the ref. doc given in the answer section of the question. (Check page Security and Permissions 1120- 1121) https://docs.aws.amazon.com/pdfs/sagemaker/latest/dg/sagemaker-dg.pdf

Comment: C. To enable Amazon SageMaker service without enabling direct internet access to Amazon SageMaker notebook instances, while adhering to a corporate data security policy that restricts internet communication, the company can: C. Create Amazon SageMaker VPC interface endpoints within the corporate VPC. This option involves setting up VPC (Virtual Private Cloud) interface endpoints for Amazon SageMaker within the corporate VPC (Virtual Private Cloud). This is done using AWS PrivateLink, which allows private connectivity between AWS services using private IP addresses. By creating VPC interface endpoints, the traffic between the corporate VPC and Amazon SageMaker does not traverse the public internet, thereby meeting the corporate data security requirements.

Comment: A would allow instances in a private subnet to initiate outbound internet traffic. This is against the requirement of no direct internet access.

Comment: NAT means data will go to internet. C is the right choice.

Comment: Option c

Comment: Only C, endpoints.

Comment: C is correct, NAT allow outband traffic pass through internet.

Comment: To prevent SageMaker from providing internet access to your Studio notebooks, you can disable internet access by specifying the VPC only network access type when you the onboard to Studio or call CreateDomain API. As a result, you won't be able to run a Studio notebook unless your VPC has an interface endpoint to the SageMaker API and runtime, or a NAT gateway with internet access, and your security groups allow outbound connections.

Replies:

> **Comment:** If you want to allow internet access, you must use a example through an internet gateway. If you don't want to allow internet access, NAT gateway with access to the internet, for create interface VPC endpoints (AWS PrivateLink) to allow Studio to access the following services with the corresponding service names. You must also associate the security groups for your VPC with these endpoints.
>
> **Comment:** To disable direct internet access, under Direct Internet access, simply choose Disable – use VPC only , and select the Create notebook instance button at the bottom. You are ready to go. from: https://aws.amazon.com/blogs/machine-learning/customize-your-amazon-sagemaker-notebook-instances-with-lifecycle-configurations-and-the-option-to-disable-internet-access/#:~:text=To%20disable%20direct%20internet%20access%2C%20under%20Direct%20Internet%20access%2C%20simply,running%2C%20without%20direct%20internet%20access.

Comment: A VPC interface endpoint is a private connection between a VPC and Amazon SageMaker that is powered by AWS PrivateLink. With a VPC interface endpoint, traffic between the VPC and Amazon SageMaker never leaves the Amazon network.

Comment: Page 3438 of https://docs.aws.amazon.com/sagemaker/latest/dg/sagemaker-dg.pdf

Comment: VPC Interface endpoints

Comment: If the question just had the last sentence, the answer would be A or C, per this page:https://docs.aws.amazon.com/sagemaker/latest/dg/appendix-notebook-and-internet-access.html. "To disable direct internet access, you can specify a VPC for your notebook instance. By doing so, you prevent SageMaker from providing internet access to your notebook instance. As a result, the notebook instance won't be able to train or host models unless your VPC has an interface endpoint (PrivateLink) or a NAT gateway, and your security groups allow outbound connections." HOWEVER, the question has more context that internet access is not allowed by the corporate policy. ("When you use a VPC interface endpoint, communication between your VPC and the notebook instance is conducted entirely and securely within the AWS network.") Therefore, the answer must be ONLY C.

Comment: Answer is C. From https://docs.aws.amazon.com/sagemaker/latest/dg/interface-vpc-endpoint.html -> "The VPC interface endpoint connects your VPC directly to the SageMaker API or Runtime without an internet gateway, NAT device, VPN connection, or AWS Direct Connect connection. The instances in your VPC don't need public IP addresses to communicate with the SageMaker API or Runtime."

Comment: I see a lot of people employing pretzel logic to try to explain why they should be using NAT. The question states no internet communication. Period. No internet means no NAT. Answer is C.

Comment: The point here is that SM notebooks need internet access to download updates and some open data. Although C is ok, it won't allow SM notebook to download such data. NAT GW will allow this action. I go with A.

Discussion for Question 48

Link: https://www.examtopics.com/discussions/amazon/view/10081-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 18 votes

Discussion

Comment: Ans B sounds correct

Comment: In transfer learning, a pre-trained model is used as a starting point to train a new model on a different task, typically using a smaller dataset. The pre-trained model contains weights that have been learned from a large amount of data on a related task, and these weights can be leveraged to train the new model more efficiently. To re-train the model with the custom data, the Specialist should initialize the model with pre-trained weights in all layers, as these weights can provide a good starting point for the new task. The Specialist should then replace the last fully connected layer, which is responsible for making the final predictions, as this layer will likely need to be modified to reflect the new task. By keeping the pre-trained weights in the other layers, the Specialist can take advantage of the knowledge learned from the previous task, and potentially speed up the training process.

Comment: Transfer learning helps accelerate the training and at this point, model has yet to learn from the new data. So, all layers including the fully-connected by replaced. Eventually, the training will update the fully-connected layer. The question is about initialization, so we should initialize the fully-connected layers too.

Comment: A. NO - random weights does not allow transfer learning B. YES - the last layer gives the final classes, we want to have new classes C. NO - random weights does not allow transfer learning D. NO - the last layer gives the final classes, we want to have new classes

Comment: Option B

Comment: For Transfer Learning, A and C are incorrect because we restart the model. The correct is letter B

Comment: B. The reason is, fine-tuning a model means to use the weights/biases trained before. also no matter which strategy you go for in transfer learning (fine-tuning or feature extraction) you always replace the last or last few layers.

Comment: The task is to " to re-train it with the custom data". That means, it is not transfer learning anymore. The "transfer learning" is just a title to make a question tricky. So, in this case we should randomize the weights and retrain whole model from scratch on custom user's images only. The correct answer is C.

Replies:

> **Comment:** I think retraining revers in this context to the training on the custom data that the expert as already conducted before thinking about transfer learning.

Comment: The task is to " to re-train it with the custom data". That means, it is not transfer learning anymore. The "transfer learning" is just a title to make a question tricky. So, in this case we should randomize the weights and retrain whole model from scratch on custom user's images only. The correct answer is C.

Comment: The fully connected layer will need to be trained from scratch to incorporate the features of his domain problem (Car models)

Comment: 12-sep exam

Comment: D is the best - here is why Question is not to design a final production with deep lense - it is to use it as a dev platform to comeup with a edge ML vs. dump load all to S3 -which is very wasteful! AWS did not mae deeplense as a toy for devs! it is meant to help companies experiment with edge ML And then copy and reuse the open hardware platfom

Comment: one of the method to implement transfer learning

Comment: I will go with B, we are mainly concerned with the output layer for us to get the desired results, hence we need to replace it.

Comment: B is correct

Comment: Actually, it should be NONE of IT!.... it should be like B with exception that 20-40% top layers should be retrained :) -- this is classic transfer learning setup, so B is the answer here.

Comment: Since it is transfer learning where you retain knowledge from a solved problem, weights are to be pre-trained. So A and C are wrong. Between B and D, D keeps the last layer but that is not what you want since the question mentions a change of general objects to more specific types. So answer is B

Discussion for Question 49

Link: https://www.examtopics.com/discussions/amazon/view/8374-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 9 votes

Discussion

Comment: Answer is "A". C and D are out as DeepLens is not offered as a commercial product. It is purely for developers to experiment with. From https://aws.amazon.com/deeplens/device-terms-of-use/ " (i) you may use the AWS DeepLens Device for personal, educational, evaluation, development, and testing purposes, and not to process your production workloads;" A is correct as it's will analyse live video streams instead of images. From https://aws.amazon.com/rekognition/video-features/ "Amazon Rekognition Video can identify known people in a video by searching against a private repository of face images. "

Replies:

> **Comment:** Agreed

> **Comment:** Agree as well, besides that: (D) uses Rekognition with Image mode, which is wrong for this case.

Comment: Why not A? DeepLens is for development purpose and much more expensive than just a camera. They are referring to 1000 camera in production scale?

Replies:

> **Comment:** A bit off topic but yeah, how could you justify using deep lens for production. Cameras have viewing angles, weather proofing, network connectivity issues (Wifi only), infra red for low lighting conditions, no power over ethernet? Using Deeplens would be laughable for a full production system.

> **Comment:** C is the correct answer. We could use A, since it is for security service, DeepLens allows to notify the security (through aws lamda) immediately when it sees non employee at the office location. So C is more appropriate for the problem than A.

> ### Replies:
>
> > **Comment:** DeepLens is for developers only, it is not available as a commercial product.

Comment: The correct answer is D. Very tricky one but re-read the 2nd sentence in the question; “Images from the cameras were uploaded to Amazon S3 and tagged using Amazon Rekognition, and the results were stored in Amazon ES.” So, we have ‘images' as training data, not videos. This is why it can not be option C - where it says to use Amazon Recognition Video. The only option mentioning Amazon Recognition Image is the option D. Also check: https://docs.aws.amazon.com/rekognition/latest/dg/what-is.html “...For example, each time a person arrives at your residence, your door camera can upload a photo of the visitor to Amazon S3. This triggers a Lambda function that uses Amazon Rekognition API operations to identify your guest. You can run analysis directly on images that are stored in Amazon S3 without having to load or move the data.”

Comment: A is answer

Comment: A not B: Use Amazon Rekognition Video instead of Amazon Rekognition Image in this case.

Comment: A is correct!

Comment: DeepLens is overkill for mass systems

Comment: A. NO - thousands of cameras would choke network bandwidth B. NO - thousands of cameras would choke network bandwidth C. YES - DeepLens is made for edge computing; it might be EOL / Not commercially available, but if they did not want you to use DeepLens the question would not have come in the first place D. NO - use Amazon Rekognition Video directly instead of Amazon Rekognition Image

Comment: Why A and not B? Can someone please explain it?

Comment: Option A

Comment: From Chat GPT The solution that the agency should consider is option A: Use a proxy server at each local office and for each camera, and stream the RTSP feed to a unique Amazon Kinesis Video Streams video stream. On each stream, use Amazon Rekognition Video and create a stream processor to detect faces from a collection of known employees and alert when non-employees are detected. By using a proxy server at each local office and streaming the RTSP feed to individual Amazon Kinesis Video Streams video streams, the agency can efficiently handle the large number of video cameras in different office locations. Using Amazon Rekognition Video, the agency can create a stream processor to detect faces from a collection of known employees. This allows for real-time identification of non-employees based on facial recognition. Alerts can then be generated when non-employees are detected, ensuring that the agency is able to identify and respond to potential security threats in real-time.

Comment: I initially thought it is C but looks like A makes more sense here.

Comment: The DeepLens Service will reach EOL at the end of Jan 2024, so more than likely that this question will not be asked in the exam

Comment: D is the answer now, DeepLens is used for situations like this!

Replies:

Comment: Maybe, its EOL Jan 2024

Comment: Think big picture - you tested something (let say code python) and ready to implement into prod will you move python code or java code! Here in this particular case, they tested with actual video camera and they did not say deeplense so answer is A! For knowledge sake if they say in real exam it is tested with deeplense ---then ideal solution should be model inference happening at deeplense itself with search against existing employees and send back model inference when it detect new faces who are not employees back to cloud may be S3

Comment: Answer is "A". As mentioned in below user comment, DeepLens is not offered as a commercial product. https://aws.amazon.com/deeplens/device-terms-of-use/

Comment: Answer is C based on this exact same article answer: https://docs.aws.amazon.com/rekognition/latest/dg/streaming-video.html Rekognition stream processor doesn't need Lambda.

Discussion for Question 50

Link: https://www.examtopics.com/discussions/amazon/view/8376-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 16 votes

Discussion

Comment: All of the questions in the preceding examples rely on having example data that includes answers. There are times that you don't need, or can't get, example data with answers. This is true for problems whose answers identify groups. For example: "I want to group current and prospective customers into 10 groups based on their attributes. How should I group them? " You might choose to send the mailing to customers in the group that has the highest percentage of current customers. That is, prospective customers that most resemble current customers based on the same set of attributes. For this type of question, Amazon SageMaker provides the K-Means Algorithm. https://docs.aws.amazon.com/sagemaker/latest/dg/algos.html Clustering algorithms are unsupervised. In unsupervised learning, labels that might be associated with the objects in the training dataset aren't used. https://docs.aws.amazon.com/sagemaker/latest/dg/algo-kmeans-tech-notes.html THE ANSWER COULD BE B.clustering on customer profile data to understand key characteristic

Replies:

Comment: Yes, Clustering seems to be more appropriate in this scenario than recommender system

Replies:

Comment: Collaborative filtering recommendation system is also unsupervised

Comment: https://towardsdatascience.com/customer-segmentation-with-machine-learning-a0ac8c3d4d84 B

Comment: Option C. This is not purely unsupervised, as clustering would be, because we have current and past customer profiles to go on. We want to find new customers by finding similar profiles on social media. So it is supervised to some extent. It's not a cluster problem; it is user-user collaborative filtering. The key is to recognize that this is not clustering. You're not blindly trying to group people. You have existing profiles that you are comparing them to.

Comment: It is B. Recommendation Engines: Traditionally focus on suggesting products/services to existing customers based on past behavior.

Comment: Clustering is right

Comment: C would be an answer if wanted to send the promo to the existing customers. But we want to find potential customers. And we can do it only by comparing existing customers with potential customers. It can be done by creating clusters of existing customers and measuring the distance to those clusters for the new potential users. So my answer is B

Comment: A. NO - Linear Regression not best to understand relationships between data B. NO - it is supervised (we know premiums received vs. claims paid, so can assign users to GOOD or BAD), so no clustering C. YES - A recommendation engine in AWS lingua is Amazing Recommender (https://docs.aws.amazon.com/personalize/latest/dg/what-is-personalize.html - "Creating a targeted marketing campaign") and can create user segments D. NO - not as good as C

Comment: B for me

Comment: Recommendation engines is perfect for customers we have, but for implementing a machine learning model to identify potential (new customers on social media) this requires clustering and segmentation. https://neptune.ai/blog/customer-segmentation-using-machine-learning

Comment: Based on the link below, it must be C https://medium.com/voice-tech-podcast/a-simple-way-to-explain-the-recommendation-engine-in-ai-d1a609f59d97

Comment: We are divided, but I stick with B.

Comment: I think it should be c

Comment: recommender system would help here, as we already have details of all customers

Comment: it should be C - recommender system would be better fit here.

Comment: We should use recommendation system to find key characteristics only among company users (past and present). At this step we don't take any users from the web. After we finish processing this CF model we identify key characteristics (important features?) and only after that, we will start looking for similar users on the web.

Comment: I would use clustering technique to identify which customers in my database are the target audience and get similar customer profiles from the social media dataset. Its a lot simpler

Comment: recommendation engines can use either supervised or unsupervised learning. I can't find any reason to NOT use recommendation engine???

Comment: I am still not sure between B & C, reading this post https://medium.com/@daniilkorbut/recommendation-system-algorithms-ba67f39ac9a3 seems clustering is possible but there are a few things to consider: clustering is done on large dataset while here is not mentioned. also it's an unsupervised algorithm, here we have existing customer's data.

Replies:

Comment: we have customer data but not target. these are all positive customers (the one who bought sth.) not negative ones (the one who didnt) if we had the one who didnt we could use C.

Discussion for Question 51

Link: https://www.examtopics.com/discussions/amazon/view/8379-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- D: 18 votes

Discussion

Comment: HOW MANY/MUCH, THOSE ARE REGRESSION TOPIC, LOGISTIC FOR 0/1,YES/NO https://docs.aws.amazon.com/zh_tw/machine-learning/latest/dg/regression-model-insights.html THE ANSWER SHOULD BE D.

Replies:

> **Comment:** agree. RCF is mostly used for anomaly detection or separate outliers

Comment: Amazon SageMaker Random Cut Forest (RCF) is an unsupervised algorithm for detecting anomalous data points within a data set Answer is D 100%

Comment: The Answer is D. Random Cut Forest is for Anomaly Detection

Comment: D should be the answer

Comment: How many units should give this away as Linear regression

Comment: I do not see any hint of anomalies here, we are looking for a number to be predicted, this seems to be the reason of the correct answer https://docs.aws.amazon.com/quicksight/latest/user/how-does-rcf-generate-forecasts.html

Comment: How can the right answer be B? That Random Cut Forest is an algorithm written for anomaly detection.

Comment: option D

Comment: D is the correct. B is for outlier detection only.

Comment: It sounds like Linear regression problem and Random Cut is more known for anomaly detection while it can do other types of ML. The answer seems to be strange with no explanation.

Comment: D is correct!

Comment: D. Linear regression would be the appropriate machine learning approach to solve this problem of predicting the number of units of a particular part to be produced each quarter. Linear regression is a supervised learning algorithm used for predicting continuous variables based on input features. In this case, the historical sales data can be used as input features, and the number of units produced each quarter can be used as the continuous target variable.

Comment: definitely D.

Comment: This is a regression problem where the goal is to predict a continuous outcome, which in this case is the number of units of a particular part that should be produced each quarter. Linear regression is a simple and commonly used approach to solve such problems, where a linear relationship is established between the independent variables (e.g., historical sales data) and the dependent variable (e.g., number of units of a part to be produced).

Comment: D, RCF answers here just link one article where RCF is implemented to find outliers in time series, or are able to deduce trends, but here they mention already labelled data, RCF is unsupervised, so that data would go to waste.

Comment: Honestly, i think these are all bad answers. It should be time series modeling methods.

Comment: The answer is D. B is out of it as RCF is used for anomaly detection. Logistic Regression is for Classification mainly. Only linear regression can be used if Time series algorithms are not part of the options.

Discussion for Question 52

Link: https://www.examtopics.com/discussions/amazon/view/8382-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 7 votes

Discussion

Comment: BOTH A AND B ARE ANSWERS. BUT external Apache Hive MIGHT BE NOT SERVERLESS SOLUTION. The AWS Glue Data Catalog is your persistent metadata store. It is a managed service that lets you store, annotate, and share metadata in the AWS Cloud in the same way you would in an Apache Hive metastore. The Data Catalog is a drop-in replacement for the Apache Hive Metastore https://docs.aws.amazon.com/zh_tw/glue/latest/dg/components-overview.html BEAUTIFUL ANSWER IS A.

Replies:

> **Comment:** I am thinking about Answer C, because events can be triggered by cloudwatch w/Glue metastore
>
> ### Replies:
>
> > **Comment:** if we use Flexible as key word ..Using Lambda might be a constraint
> >
> > **Comment:** you can't schedule AWS Batch with CloudWatch
> >
> > ### Replies:
> >
> > > **Comment:** srr, looks like you can apart from Cron, the argument should be AWS Batch aren't SERVERLESS
> > >
> > > **Comment:** We can schedule batch with cloud watch events.

Comment: Answer is A. Lamda is the preferred way of implementing event-driven ETL job with S3, when new data arrives in S3, it notifies lamda which can start the ETL job.

Replies:

> **Comment:** agree, event-driven means Lambda, CloudWatch alarms are just to trigger alarms based on log analysis.

Comment: A. YES - all integrated components B. NO - missing a component to invoke the Lambda C. NO - CloudWatch will not trigger when there is a new file to process D. NO - CloudWatch will not trigger when there is a new file to process

Comment: A for me

Comment: Note that the question asks for a serverless system. In this case, the letters B, C and D are wrong, as they bring options that are managed: AWS Batch (managed) and external Apache Hive (even more managed). For event-driven AWS ETL solutions that are serverless, activation through the Lambda function is recommended, so the correct alternative is Letter A. Note that CloudWatch Alarms only activates from log evaluation, which is not mentioned in the question.

Comment: I will chose A, I think C & D is wrong, you can use Amazon CloudWatch Event to trigger lambda but not CloudWatch alarm.

Comment: Batch is more for configurations and other kinds of things by scheduling than event driven and batch data processing with ETL, the answer is A.

Comment: Found this supporting A - Lambda used to trigger ETL job after crawler completes. The crawler starts on schedules or events (files arriving).

Comment: Based on Majority discussion

Comment: Quite confused between A&C since they all workable solution. In below AWS Blog, even mix the CloudWatch + Lambda to use the Glue. For key word event trigger, prefer CloudWatch https://aws.amazon.com/blogs/big-data/build-and-automate-a-serverless-data-lake-using-an-aws-glue-trigger-for-the-data-catalog-and-etl-jobs/ https://docs.aws.amazon.com/glue/latest/dg/automating-awsglue-with-cloudwatch-events.html

Replies:

> **Comment:** Agreed. CloudWatch could trigger event to launch Lambda. Refer to: https://docs.aws.amazon.com/lambda/latest/dg/services-cloudwatchevents.html
>
> **Comment:** cloudwatch and lambda function can work together to trigger event. But AWS batch cannot independently conduct ETL and require other service. when it comes to ETL, glue is much easier choice than Batch

Comment: Answer is A 100%

Comment: A is preferred. Lambda can trigger ETL pipelines: https://aws.amazon.com/glue/

Comment: A is correct...Lambda is event driven and Glue is serverless as opposed to Hive

---

Discussion for Question 53

Link: https://www.examtopics.com/discussions/amazon/view/10082-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 15 votes

Discussion

Comment: the answer is B. using Hovord distribution results in less coding effort

Comment: Answer is B. "minimize coding effort and infrastructure changes" If we use DeepAR then the code and infra has to be changed to work with DeepAR.

Comment: A. NO, this will not address training dataset continuous increase B. NO, this will require code effort and infrastructure change C. YES, a built-in model ensure low code effort, so only infrastructure change needed* D. This will not work * they say current model accuracy is acceptable, we doo expect good results with DeepAR as it allows to automatically pick among 5 different models what works best for the customer

Replies:

Comment: DeepAR doesn't pick among 5 models. However, I still think that switching to DeepAR can assure accuracy and minimize coding effort as the model is built-in

Comment: A comes with minimum changes, but it wont scale. B code changes are minimum but infrastructure still needs to be changed to achieve a distributed solution. C. Is even more significant infra and code change. D. wont work. It is really subjective and tricky. Could be A or B, depending on what change is considered "SMALL". For scalability, B seems better. for quick win A could work. I keep going back and forth.

Comment: A. NO - one time shot and not scalable B. YES - best practice C. NO - DeepAR is for forecasting D. NO - code will not benefit from parallelization without change

Comment: option B

Comment: Note that we want to increase training speed, minimize code and infrastructure modification effort on AWS. Letter A would only delay the problem and increase costs too much. The solution that best translates the problem would be Letter B: we would keep the code in tensorflow and use Horovod to make our training faster through parallelization. Letter D is too complex and would change the execution infrastructure a lot and Letter C would be too abrupt a turn as we would throw our model away.

Comment: A is better option even though B helps. Firstly, you only have One GPU, in this case distributed training Horovod doesn't help much; Secondly, the question is about minimize "coding effort" not minimize budget. adding distributed framework require much more coding, but increase gpu instance only require single click.

Comment: Horovod distribution is accepted by sagemaker, making easy to implement!

Comment: Hovord distribution will allow the Machine Learning Specialist to take advantage of Amazon SageMaker's built-in support for Horovod, which is a popular, open-source distributed deep learning framework. Implementing Horovod in TensorFlow will allow the Specialist to parallelize the training across multiple GPUs or instances, which can significantly reduce the time it takes to train the model. This will allow the company to meet its requirement to update the model on an hourly basis, and minimize coding effort and infrastructure changes as it leverages the existing TensorFlow code and infrastructure, along with the scalability and ease of use of Amazon SageMaker.

Comment: Are is there a 23X differential between the weakest and strongest GPU in AWS? (and allow for future growth). I don' tthink so.

Comment: Answer:C- built-in sagemaker DeepAR model. minimize coding & infra changes.

Replies:

Comment: But they are happy with it - just want it to go faster. Not throw the whole thing out.

Comment: the answer is B. using Hovord distribution results in less coding effort

Comment: Most likely , it is A because it is based on AWS teachnoloy, why we have to use open source we exam AWS ML , the answer should be relevant to AWS technology inevitably https://aws.amazon.com/sagemaker/distributed-training/

Comment: This one reminds me of an old saying by Yogi Berra: "When you come to a fork in the road, take it." If you see Horovod as an option in a question about scaling TF, take it. Answer is B.

Comment: I Think it's B https://aws.amazon.com/blogs/machine-learning/launching-tensorflow-distributed-training-easily-with-horovod-or-parameter-servers-in-amazon-sagemaker/ & https://aws.amazon.com/blogs/machine-learning/multi-gpu-and-distributed-training-using-horovod-in-amazon-sagemaker-pipe-mode/

Comment: Seen similar question on udemy/whizlab , its always Horvord when Tensorflow needs scaling. ANWSER is B

Discussion for Question 54

Link: https://www.examtopics.com/discussions/amazon/view/8384-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- D: 6 votes

Discussion

Comment: RECALL IS ONE OF FACTOR IN CLASSIFY, AUC IS MORE FACTORS TO COMPREHENSIVE JUDGEMENT https://docs.aws.amazon.com/zh_tw/machine-learning/latest/dg/cross-validation.html ANSWER MIGHT BE D.

Replies:

Comment: Not might be, but should be D

Comment: AUC is to determine hyperparams in a single model, not compare different models.

Comment: AUC/ROC work well with special case of Binary Classification not in general

Replies:

Comment: AUC is to compare different models in terms of their separation power. 0.5 is useless as it's the diagonal line. 1 is perfect. I would go with F1 Score if it was an option. However, taking Recall only as a metric for comparing between models, would be misleading.

Comment: why not C?

Replies:

Comment: it's a classification problem, mape is for regression

Comment: option D

Comment: AUC is the best metric.

Comment: Area Under the ROC Curve (AUC) is a commonly used metric to compare and evaluate machine learning classification models against each other. The AUC measures the model's ability to distinguish between positive and negative classes, and its performance across different classification thresholds. The AUC ranges from 0 to 1, with a score of 1 representing a perfect classifier and a score of 0.5 representing a classifier that is no better than random. While recall is an important evaluation metric for classification models, it alone is not sufficient to compare and evaluate different models against each other. Recall measures the proportion of actual positive cases that are

correctly identified as positive, but does not take into account the false positive rate.

Replies:

> **Comment:** chatgpt answers, all your answers are from chatgpt

Comment: D. AUC is always used to compare ML classification models. The others can all be misleading. Consider the cases where classes are highly imbalanced. In those cases accuracy, misclassification rate and the like are useless. Recall is only useful if used in combination with precision or specificity, which what AUC does.

Comment: Its Accuracy,Precision,Recall and F1 score , there is no metion of AUC/ROC for comparing models in many articles , so ANSWER is A

Replies:

> **Comment:** When you draw the ROC graph, you're considering True and False Positive Rate. The first one is also called Recall ;)

Comment: D. AUC is scale- and threshold-invariant, enabling it compare models. https://towardsdatascience.com/how-to-evaluate-a-classification-machine-learning-model-d81901d491b1

Comment: Actually A, B and D seem to be correct

Comment: Probably D https://towardsdatascience.com/metrics-for-evaluating-machine-learning-classification-models-python-example-59b905e079a5

Comment: why not B?

Comment: Answer should be D..ROC is used to determine the diagnostic capability of classification model varying on threshold

Comment: Should be A. A is the only one that generally works for classifcation. AUC only works with binary classification.

Replies:

> **Comment:** Could be, you mean in a multiclass clasification problem. But in that con context recall directly can't be compare because first you have to decide recall of what of the classes, in a 3 classes problem we have 3 recalls or you suppose a weighted recall or average recall ?. Do you think in that ?
>
> ### Replies:
>
> > **Comment:** Also in multi-class classification, if you follow an One-vs_Rest strategy you can still use AUC. https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html#sphx-glr-auto-examples-model-selection-plot-roc-py
>
> **Comment:** Actually AUC could be generalized for multi-class problem. https://www.datascienceblog.net/post/machine-learning/performance-measures-multi-class-problems/

Comment: Correct Answer is D. Another benefit of using AUC is that it is classification-threshold-invariant like log loss. https://towardsdatascience.com/the-5-classification-evaluation-metrics-you-must-know-aa97784ff226

Discussion for Question 55

Link: https://www.examtopics.com/discussions/amazon/view/10083-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 14 votes
- B: 9 votes

Discussion

Comment: Ans C is reasonable

Comment: Agree with C. Quicksight cannot handle 100TB each day.

Comment: Ans C Because Quicksight Can't handle 100 TB even in Entiripise Quotas for SPICE are as follows: 2,047 Unicode characters for each field 127 Unicode characters for each column name 2,000 columns for each file 1,000 files for each manifest For Standard edition, 25 million (25,000,000) rows or 25 GB for each dataset For Enterprise edition, 1 billion (1,000,000,000) rows or 1 TB for each dataset https://docs.aws.amazon.com/quicksight/latest/user/data-source-limits.html

Comment: QuickSight can handle large volumes of data for analytics and visualizations. Some key points: QuickSight scales seamlessly from hundreds of megabytes to many terabytes of data without needing to manage infrastructure. It uses an in-memory engine called SPICE to enable high performance analytics on large datasets. so the choice is B

Comment: B. Generate daily precision-recall data in Amazon QuickSight, and publish the results in a dashboard shared with the Business team. This solution leverages QuickSight's managed service capabilities for both data processing and visualization, which should minimize the coding effort required to provide the Business team with the necessary insights. However, it's important to note that QuickSight's ability to calculate the precision-recall data depends on its support for the necessary statistical functions or the availability of such calculations in the dataset. If QuickSight cannot perform these calculations directly, option C might be necessary, despite the increased effort.

Comment: The question does not ask for processing of 1Tb data. it asks for visuals/predications of that data. So B

Comment: C. Considering the large volume of data (100 TB daily), Option C seems to be the most appropriate solution

Comment: B it's not correct because of 100tb data size. C is the answer: https://docs.aws.amazon.com/quicksight/latest/user/data-source-limits.html

Comment: ANs c is correct

Comment: A. NO - we want a dashboard for business B. NO - 100TB is very large, it will not fit in memory (1TB max for SPICE dataset) or return within the 2min limit if delegated to a DB (https://docs.aws.amazon.com/quicksight/latest/user/data-source-limits.html) C. YES - best combination; EMR can distribute the computation of precision-recall for each slice of data D. NO - ES cannot help to generate precision-recall

Comment: although C is tempting but goes with B due to less effort

Replies:

> **Comment:** it is not about the least effort only, since the least effort solution here will not get your job done, look at the quick sight max data it can deal with when it compared to EMR which is built to deal with Big data.

Comment: using quick sight for creation of the precision recall with 100 TB every day cann't be done since the max size for quick sight to deal with is : For Standard edition, 25 million (25,000,000) rows or 25 GB for each dataset For Enterprise edition, 1 billion (1,000,000,000) rows or 1 TB for each dataset acc to AWS documentation : https://docs.aws.amazon.com/quicksight/latest/user/data-source-limits.html but we can do it with EMR and latterly use quick sight to visualize the results

Comment: Looking at the QuickSight documentation: it has a limit of 1 TB per dataset. So it's necessary a previous layer. Letter C is the correct one.

Comment: It's 100TB daily, need EMR to reduce, option C is correct.

Comment: Quicksight can handle maximum 1TB data set only. We have 100TB data set so we need EMR. https://docs.aws.amazon.com/quicksight/latest/user/data-source-limits.html

Comment: c is correct answer

Comment: B is the correct answer according to this resource and even according to ChatGPT, GPT-4 and Claude+ https://aws.amazon.com/blogs/big-data/diligent-enhances-customer-governance-with-automated-data-driven-insights-using-amazon-quicksight/

Replies:

> **Comment:** so? chatgpt is a parrot that says whatever you want to listen

---

Discussion for Question 56

Link: https://www.examtopics.com/discussions/amazon/view/10084-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: C is okay

Comment: Anwer is C. Most Amazon SageMaker algorithms work best when you use the optimized protobuf recordIO format for the training data. https://docs.aws.amazon.com/sagemaker/latest/dg/cdf-training.html

Comment: option C

Comment: The Specialist should transform the dataset into the RecordIO protobuf format. This format is optimized for use with SageMaker and has been shown to improve the speed and efficiency of training algorithms. Using the RecordIO protobuf format is a best practice for preparing data for use with Amazon SageMaker, and it is specifically recommended for use with the built-in algorithms.

Comment: I would assume the issue is the transformation. It can be nasty slow between pandas / csv / numpy. Go to protobuf.

Comment: C is the best

Comment: Agree with C

Discussion for Question 57

Link: https://www.examtopics.com/discussions/amazon/view/8386-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 19 votes

Discussion

Comment: NO CORRECT TRAINING DATA, MORE WORKS JUST WASTE TIME. ONE OF THE REASONS FOR POOR ACCURACY COULD BE INSUFFICIENT DATA. THIS CAN BE OVERCOME BY IMAGE AUGMENTATION. IMAGE AUGMENTATION IS A TECHNIQUE OF INCREASING THE DATASET SIZE BY PROCESSING (MIRRORING, FLIPPING, ROTATING, INCREASING/DECREASING BRIGHTNESS, CONTRAST, COLOR) THE IMAGES. HTTPS://MEDIUM.COM/DATADRIVENINVESTOR/AUTO-MODEL-TUNING-FOR-KERAS-ON-AMAZON-SAGEMAKER-PLANT-SEEDLING-DATASET-7B591334501E ANSWER A. ADD MORE TRAINING DATA FOR ROTATION IMAGES COULD BE A WAY TO DEAL WITH ISSUE

Replies:

Comment: agree with A

Comment: The key phrase might be "constant test set", so you can't increase training set by shrinking the size of test set. Thus the only feasible choice is to increase training time by increasing the number of epochs => answer B.

Replies:

Comment: A . Increase the training data by adding variation in rotation for training images. It never says to move the images from Test data set (because it is constant)... only variations are added to the images..so, A is correct.

Comment: The problem is images are upside down and misclassified. If right side up then the model would classify correctly. This can only be fixed ba rotating not by trying to recognise upside down cat more times.

Replies:

Comment: What's your answer B?

Comment: is it possible no using MAYUS? it is annoying

Comment: Donald, your caps lock is on.

Replies:

Comment: LOL :D

Comment: Okay, was funny

Comment: A is answer

Comment: Data Augmentation would fix the missing conditional data

Comment: ChatGPT says the answer is A. Trust a model to answer an ML question correctly! ;)

Comment: how come more epochs it better than augmentation?

Comment: option A

Comment: The question is clear and the answer is clear as well

Comment: should be A

Comment: More epochs is not a good approach to fundamental data issues

Comment: the Specialist can apply data augmentation techniques to increase the training data by adding variation in rotation for training images. This technique will allow the model to learn to recognize cats in various orientations, including upside down.

Comment: Adding more variation in rotation to the training data can help the model to learn how to classify cats in different orientations, including when they are held upside down. This can improve the model's ability to identify cats in this position and reduce the misclassification rate for images in which the cats are upside down. By adding more rotation to the training data, the model can be trained to generalize better to new images, including those with cats in different orientations. This can help to reduce overfitting and improve the model's overall performance.

Comment: Only logical answer 100% A.

Comment: More data is a good answer. A

Comment: Answer is "A""

Comment: Answer is A

Comment: This is a clear case of Data Augumentation solution.

Comment: Common step in CNN, Image augmentation. A.

Discussion for Question 58

Link: https://www.examtopics.com/discussions/amazon/view/10085-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: the answer is C. as the main point of the question is data transformation to Parquet format which is done by Kinesis Data Firehose not Data Stream. Coming to the data store the data store in Kinesis Data Stream is only for couple of days so it does not serve the purpose here

Replies:

Comment: The storage part will be taken care of by S3 anyway. Firehose would just transform to Parquet on the fly.

Comment: Firehose

Comment: Not sure Firehose can store the data .... Data Stream can store the data. Someone please explain the answer

Replies:

Comment: Firehose is to Store the data. Stream requires other service to do that.

Comment: Kinesis Data Streams can Store for up to 365 days, While Firehouse sends it to S3. Which is correct?

Comment: Firehose can do it if the data is in JSON or ORC format initially!

Comment: It should be KDS

Comment: Amazon Kinesis Data Firehose is a fully managed service that can automatically load streaming data into data stores and analytics tools. It can ingest real-time streaming data such as application logs, website clickstreams, and IoT telemetry data, and then store it in the correct format, such as Apache Parquet files, for exploration and analysis. This makes it a suitable option for the requirement described in the question.

Comment: B https://github.com/ravsau/aws-exam-prep/issues/10

Comment: B) Only Amazon Kinesis Data Streams can store and Ingest data. We don't need to apply any transformation; the question asks to ingest and store data in Apache Parquet format, There is no assumption that the data coming in a different format than parquet.

Replies:

Comment: KDS cant store to s3 https://stackoverflow.com/questions/66097886/writing-to-s3-via-kinesis-stream-or-firehose

Comment: It is C with no doubt https://aws.amazon.com/about-aws/whats-new/2018/05/stream_real_time_data_in_apache_parquet_or_orc_format_using_firehose/

Comment: It appears all agree that the answer is between Firehose and Analytics. Kinesis Firehose is used for ingestion. Both firehose and analytics can store, only firehose can ingest. https://docs.aws.amazon.com/firehose/latest/dev/record-format-conversion.html shows firehose can store parquet to S3

Replies:

Comment: It appears all agree that the answer is between Firehose and Analytics. Data Streams handle stuff like event data, clickstream etc. Its not interested in special format, the focus is speed. The question did not talk of transformation, only ingestion. Kinesis Firehose is used for ingestion. Both firehose and analytics can store, only firehose can ingest. https://docs.aws.amazon.com/firehose/latest/dev/record-format-conversion.html shows firehose can store parquet to S3

Comment: Think just like this -- batch process Glue ETL and Streaming process Firehose ETL .....covert to parquet or any other format.

Comment: C for Firehose

Comment: Just in case https://acloud.guru/forums/aws-certified-big-data-specialty/discussion/-KhI3MgPEo-FY5rfgl3J/what_is_difference_between_kin

Comment: Amazon Kinesis Data Firehose can convert the format of your input data from JSON to Apache Parquet or Apache ORC before storing the data in Amazon S3. https://github.com/awsdocs/amazon-kinesis-data-firehose-developer-guide/blob/master/doc_source/record-format-conversion.md

Comment: I would go with B. Kinesis data streams stores data, while Firehose not.

Replies:

Comment: It's the other way around. Firehouses stores data; data streams does not.

Discussion for Question 59

Link: https://www.examtopics.com/discussions/amazon/view/74274-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 67 votes
- C: 21 votes

Discussion

Comment: C would be my answer here. Rescaling each set independently could lead to strange skews. Training set, Test set and Evaluation set should be on the same scale

Replies:

Comment: You're right. test set and val set should be rescaled on the same scale. But the scale value should be extracted by only statistical value from training data. I think C means that the rescaling stage is affected by the values from the whole data (with val, test set) So, I think B is correct

Comment: https://stackoverflow.com/questions/49444262/normalize-data-before-or-after-split-of-training-and-testing-data C also leads to data leakage. You are using the test data to scale everything. So part of the data in the test set is used to scale for when you build the model on the training and check against the validation set.

Comment: If you Rescale all the data first you are going to do data leakage by showing all the variance of data with in training. The rescaling needs to be after splitting the data and not before it

Comment: The best practice is --> to split the dataset into training, validation, and test sets first, and then rescale the training set and apply the SAME scaling to the validation and test sets. This ensures that the scaling parameters (e.g., mean and standard deviation for standardization or min and max values for min-max scaling) are calculated only based on the training set to prevent data leakage and maintain the integrity of the evaluation process. By following this approach, you prevent information from the validation and test sets from influencing the scaling parameters, which could lead to data leakage and overestimation of model performance. Keeping the scaling consistent across all subsets ensures a fair evaluation of the model's generalization performance on new, unseen data.

Comment: Answer is B. The other options have shortcomings: A: Random sampling is a good practice, but it doesn't address the issue of feature scaling. Also, rescaling should occur after splitting the data. C: Rescaling the entire dataset before splitting could lead to data leakage, where information from the validation/test sets inadvertently influences the training process. D: Rescaling the sets independently would lead to inconsistencies in scale across the training, validation, and test sets, which could negatively impact model performance and evaluation.

Comment: OPTION C. Rescale the dataset. Then split the dataset into training, validation, and test sets. Explanation: Rescaling the dataset: This is the first step to address the varying statistical dispersion among features. By rescaling, you ensure that all features are on a similar scale, which is important for many machine learning algorithms. Splitting into training, validation, and test sets: After rescaling, the dataset is split into training, validation, and test sets. This ensures that the model is trained on one set, validated on another set, and tested on a third set. This separation helps evaluate the model's performance on unseen data. Option C ensures that the rescaling is applied before splitting the data, ensuring consistency in the scaling across different sets. This approach prevents data leakage and provides a more accurate representation of how the model will perform on new, unseen data.

Comment: Validation and test set should be scaled as per parameters used for scaling of training set. Independent scaling of test set would mean that drift of model in production will be way quicker and is not recommended in data science

Comment: B is correct, scale on train and apply the others. prevent to data leakage

Comment: Answer B, C is not a good data science practise.

Comment: We need firstly split the data to avoid data leakage from test/eval sets, then rescale data in all sets using statistics from training set

Comment: I think the right answer here is B. We need to split the dataset into Training, Validation and Test set. Then we can only scale (by using some technique) data contained in the Training set. Data that belong to Validation and Test set must be scaled by using the parameters used on the training. For example, if we want to apply a standardization, we can do that only on the Training set as we should not be allowed to use mean and standard deviation computed on Validation/Test set. We must act as we don't own those data!

Comment: option B

Comment: Data Science 101: (A) Given the question, doesn't solve the magnitude problem. (B) Correct (C) Data Leakage (D) It's not correct, still data leakage.

Comment: Tricky question, but, D, definitely! B: You can't apply the same scaling to the validation and test sets 'cause you may suffer data leakage! C: You shouldn't rescale the whole dataset then split into training, validation and test, it's not a good practice and may suffer data leakage as well. D: You're first splitting the whole dataset and applying rescaling individually, preventing any data leakage and each set is rescaled based in your own statistics.

Replies:

Comment: Theoretically, you should not have Test set data at Training time (when you're doing the scaling), so how do you think to do that? What if you will not have an entire Test set, but you will receive each new row at a time?

Comment: but you are leaking information from validation samples between themselves.

Comment: From Bing chat (and it makes complete sense) "Based on the search results, I think the best sequence of steps for the data scientist to take is B. Split the dataset into training, validation, and test sets. Then rescale the training set and apply the same scaling to the validation and test sets. This sequence of steps ensures that the data scientist can evaluate the model performance on different subsets of data that have not been used for training or tuning. It also ensures that the data scientist can rescale the features to have a common scale without introducing any data leakage from the validation or test sets. Rescaling the features can help improve the accuracy of some machine learning algorithms that are sensitive to the magnitude or distribution of the data, such as distance-based methods or gradient-based methods 1.

Comment: You want to measure how the model performs on new data. Scaling with the test set is a no-no.

Comment: B or D, I dont understand the semantics of "independently" and the effect it would have. It's most def not done before because of data leakage. https://www.linkedin.com/pulse/feature-scaling-dataset-spliting-arnab-mukherjee/

Discussion for Question 60

Link: https://www.examtopics.com/discussions/amazon/view/8392-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 5 votes

Discussion

Comment: ANSWER B. YOU COULD INSTALL DOCKER-COMPOSE (AND NVIDIA-DOCKER IF TRAINING WITH A GPU) FOR LOCAL TRAINING HTTPS://SAGEMAKER.READTHEDOCS.IO/EN/STABLE/OVERVIEW.HTML#LOCAL-MODE HTTPS://GITHUB.COM/AWSLABS/AMAZON-SAGEMAKER-EXAMPLES/BLOB/MASTER/SAGEMAKER-PYTHON-SDK/TENSORFLOW_DISTRIBUTED_MNIST/TENSORFLOW_LOCAL_MODE_MNIST.IPYNB

Replies:

Comment: None of these links are working

Comment: https://aws.amazon.com/blogs/machine-learning/use-the-amazon-sagemaker-local-mode-to-train-on-your-notebook-instance/ B

Comment: Correction it will be B, while D is possible, it cannot exactly mimic the sagemaker env, with docker all the configuration and libs will be available to the user which would be an ideal working setup for the DS to work with.

Comment: You can easily download the notebook instance, and work locally using jupyter notebook configured on your laptop which is one the advantages of using sagemaker, and that is what Amazon also promotes imo.

Comment: Both Amazon Q (AWS Expert) and ChatGPT insist on D. Plus all the links that I see here about Docker/Git and stuff, they either not working or deprecated so far. Not to mention their complexity to my eyes. Thus, I will go for D.

Comment: the local mode of sagemaker SDK: https://sagemaker.readthedocs.io/en/stable/overview.html#local-mode B

Comment: Option B

Comment: B, https://github.com/aws/sagemaker-tensorflow-serving-container

Comment: It's B

Comment: Answer : D

Comment: why not D?

Replies:

Comment: My assumption is that D there is no way to test the code. You need the Sagemaker SDK in order to utilize dockerized container of Tensorflow from Sagemaker is my best guess.

Replies:

Comment: Cannot be D. If you used Jupyter notebook, you are unable to use it without internet access.

Replies:

Comment: That is incorrect, once jupyter notebook is configured you can use it offline.

Comment: Agreed for B

Discussion for Question 61

Link: https://www.examtopics.com/discussions/amazon/view/8394-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 21 votes

Discussion

Comment: I WOULD LIKE TO CHOOSE ANSWER A. https://aws.amazon.com/tw/blogs/machine-learning/use-the-built-in-amazon-sagemaker-random-cut-forest-algorithm-for-anomaly-detection/

Replies:

Comment: Donald, do you know your CAPS LOCK has been on the whole time?

Replies:

Comment: I know why his caps lock has been on :D to enter the "I am not robot" code easier :D

Replies:

Comment: yes, but it works with minus also...

Comment: Answer is A. As the word anamoly talks about Random Cut Forest in the exam and that can be done in a cost effective manner using Kinesis Data Analytics

Comment: The question says REAL TIME events doesn't that eliminate Data Firehose as it is technically NEAR real time but not real time like Data Stream? Though Random Cut Forest seems like the best option for anomaly detection. I'm torn between A and B

Comment: Kinesis Firehose and Data Analytics with random cut forest should do it.

Comment: A. Based on these considerations, Option A is the most efficient way to accomplish the tasks. It provides a seamless, real-time data ingestion and processing pipeline, leverages machine learning for anomaly detection, and efficiently stores data in a data lake, meeting all the key requirements of the cybersecurity company.

Comment: ONLY A

Comment: B not as efficient for real-time processing and storing results as using Kinesis services.

Comment: At least B is a possible solution, but A will not work as KDF doesn't support KDA as a destination service https://docs.aws.amazon.com/firehose/latest/dev/create-name.html . In my opinion, KDF should always be the latest Kinesis Service in a streaming pipeline

Replies:

> **Comment:** KDF does support KDA as destination

Comment: A has all the required steps

Comment: A. YES - Firehose can pipe into KDA, and KDA supports RCF B. NO - RCF best for anomality detection C. NO - no need for intermediary S3 storage D. NO - no need for intermediary S3 storage

Comment: option A

Comment: A is the correct. One tip for the exam: When you see Data Streaming, possibly the solution should contains a Kinesis Service. B is too much complex!

Comment: Makes sense to select A here.

Comment: I strongly believe A is the right answer. At a minimum there should be some justification provided for your answer.

Comment: Amazon Kinesis Data Firehose is a fully managed service for streaming real-time data to Amazon S3 and can handle the ingestion of large amounts of data in real time. Kinesis Data Analytics Random Cut Forest (RCF) is a fully managed service that can be used to perform anomaly detection on streaming data, making it well suited for this use case. The results of the anomaly detection can then be streamed to Amazon S3 using Kinesis Data Firehose, providing a scalable and cost-effective data lake for later processing and analysis.

Replies:

> **Comment:** The problem with A, is that there is that KDF doesn't support KDA as a destination service https://docs.aws.amazon.com/firehose/latest/dev/create-name.html . In my opinion, KDF should always be the latest Kinesis Service in a streaming pipeline

Comment: I would select A

Comment: B is too resource intensive for that use case. I choose A, but I think the data should be better ingested using Kinesis streams

---

Discussion for Question 62

Link: https://www.examtopics.com/discussions/amazon/view/11384-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 10 votes

Discussion

Comment: A is correct. Kinesis Data Analytics can use lamda to convert GZIP and can run SQL on the converted data. https://aws.amazon.com/about-aws/whats-new/2017/10/amazon-kinesis-analytics-can-now-pre-process-data-prior-to-running-sql-queries/

Comment: A is correct: https://aws.amazon.com/about-aws/whats-new/2017/10/amazon-kinesis-analytics-can-now-pre-process-data-prior-to-running-sql-queries/ "To get started, simply select an AWS Lambda function from the Kinesis Analytics application source page in the AWS Management console. Your Kinesis Analytics application will automatically process your raw data records using the Lambda function, and send transformed data to your SQL code for further processing. Kinesis Analytics provides Lambda blueprints for common use cases like converting GZIP ..."

Comment: If gaining real-time insights involves complex analytics or custom processing, Amazon Kinesis Data Analytics with AWS Lambda is likely a more suitable choice. If the requirements can be met with simpler data transformations, Amazon Kinesis Data Firehose might provide a more straightforward and potentially lower-latency solution. In other words, if this data is in GZIP files and the processing requirements are relatively simple, Amazon Kinesis Data Firehose might be a more straightforward and efficient choice. GZIP files typically contain compressed data, and if our primary objective is to ingest, transform, and load this data into other AWS services for real-time insights, Kinesis Data Firehose provides a managed and streamlined solution that can handle GZIP compression.

Replies:

> **Comment:** The answer can be A , please comment if you have more clarity. After searching more, I also found out the following: (I have missed the SQL requirement in the question) Use Amazon Kinesis Data Analytics if you need SQL-based processing and advanced analytics capabilities for streaming data. Use Amazon Kinesis Data Firehose if your primary requirement is to deliver, transform, and load streaming data into various AWS destinations with simplified configurations, but not for SQL-based processing.

Comment: A is correct, why D xiyarsan sen?

Comment: A is correct

Comment: "allow the use ohttps://www.examtopics.com/exams/amazon/aws-certified-machine-learning-specialty/view/13/#f SQL to query the stream with the LEAST latency?" Well, the only solution that presents SQL query is (A). It's a description of KDA.

Comment: the term "lease latency" is the the hidden point. with Glue we can have near real-time but Kinesis data analytics will give you real-time transformation with internal lambda

Comment: A is correct, with KDA you can run sql queries in the data during the streaming (real-time SQL queries).

Comment: D. Amazon Kinesis Data Firehose to transform the data and put it into an Amazon S3 bucket would be the best solution for allowing the use of SQL to query the stream with the least latency. Amazon Kinesis Data Firehose can be configured to transform the data before writing it to Amazon S3 in real-time. Once the data is in S3, it can be queried using SQL with Amazon Athena, which is a serverless query service that allows running standard SQL queries against data stored in Amazon S3. This approach provides the lowest latency compared to other options and requires minimal setup and maintenance.

Replies:

> **Comment:** Query has to be run on stream so firehose not possible.

Comment: A is correct.

Comment: And somehow "transformation" is added to the answer as a requirement when it clearly was not part of the requirement from the question.

Comment: AAAAAAA

Comment: what about "LEAST latency"?

Comment: A is correct. you can pre-process data prior to running SQL queries with Kinesis Data Analytics and Lambda (more or less) is always a best practice :)

Comment: Answer is B. Kinesis Data Analytics does not do any transformation, it is only for querying. Glue ETL can have scripts that can transform the data

Replies:

> **Comment:** But we need to run SQL on real time stream data.

Comment: so you need lambda

Discussion for Question 63

Link: https://www.examtopics.com/discussions/amazon/view/10089-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 21 votes

Discussion

Comment: Ans: A XGBoost multi class classification. https://medium.com/@gabrielziegler3/multiclass-multilabel-classification-with-xgboost-66195e4d9f2d CNN is used for image classificaiton problems

Comment: Answer is A. This a classification problem thus XGBoost and the fact that there are six categories SOFTMAX is the right activation function

Comment: Deep convolutional neural networks (CNNs) are primarily used for image processing tasks. Given that the dataset provided is structured/tabular in nature (with features like dimensions, weight, and price) and does not mention image data, a CNN is not the most appropriate choice.

Comment: A. YES - perfect fit, multisoftmax the highest probability class is assigned B. NO - CNN is for imaging C. NO - regression forest is for continuuous variables, we can discrete classification D. NO - it is classification, not forecasting

Comment: Option A XGBoost multi class classification

Comment: A is the answer.

Comment: The XGBoost algorithm is a popular and effective technique for multi-class classification. The objective parameter can be set to multi:softmax, which uses a softmax objective function for multi-class classification. This will train the model to predict the probability of each product belonging to each category, and the most probable category will be chosen as the final prediction. A deep convolutional neural network (CNN) (B) is a powerful technique commonly used for image recognition tasks. However, it is less appropriate for tabular data like the dataset provided.

Comment: A, CNN is used for image classification. It would be suitable if we were classifying products using pictures of them.

Comment: https://xgboost.readthedocs.io/en/stable/parameter.html

Comment: B - CNN is used for dataset that have "local intermediate features" ex) images, or textCNN, etc C - We need classfication model, not regression model D - RNN is used for dataset that have sequential features A is correct

Comment: A is the best option here. Only 1200 items and 6 classes are not enough data to involve a deep neural architecture for classification.

Comment: Ans- A … For multiclassification - multi: SoftMax

Comment: That is a classification problem so A is the answer

Comment: Easy one. A is correct

Comment: A is correct

Comment: Definitely A.

Comment: 100% is A; the the others are clearly wrong Convolutional Neural Network (ConvNet or CNN) is a special type of Neural Network used effectively for image recognition and classification Recurrent neural networks (RNN) are a class of neural networks that is powerful for modeling sequence data such as time series or natural language

---

Discussion for Question 64

Link: https://www.examtopics.com/discussions/amazon/view/8395-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- D: 22 votes
- C: 10 votes

Discussion

Comment: D is correct. Amazon Comprehend syntax analysis =/= Amazon Comprehend sentiment analysis. You need to read choices very carefully.

Replies:

Comment: We're looking only to improve the validation accuracy and Comprehend syntax analysis would help that because the word set is rich and the sentiment carying words infrequent. We're not looking to replace the sentiment analysis tool with Comprehend.

Comment: AWS COMPREHEND IS A NATURAL LANGUAGE PROCESSING (NLP) SERVICE THAT USES MACHINE LEARNING TO DISCOVER INSIGHTS FROM TEXT. AMAZON COMPREHEND PROVIDES KEYPHRASE EXTRACTION, SENTIMENT ANALYSIS, ENTITY RECOGNITION, TOPIC MODELING, AND LANGUAGE DETECTION APIS SO YOU CAN EASILY INTEGRATE NATURAL LANGUAGE PROCESSING INTO YOUR APPLICATIONS. HTTPS://AWS.AMAZON.COM/COMPREHEND/FEATURES/?NC1=H_LS JUST THROUGH AMAZON COMPREHEND IS MUCH EASY THAN OTHER THE MUCH MORE CONVENIENT ANSWER IS A.

Replies:

Comment: Agree Also Keyword is TOOL rest are frameworks

Comment: Both Amazon Comprehend and the TF-IDF with a classifier solution are valid. If ease of use and pre-trained capabilities are high priorities, Comprehend is a solid option. If customization and dataset-specific nuances are crucial, building a custom model with TF-IDF may be needed. Since Comprehend is a tool, I am going with A.

Comment: D. Scikit-learn term frequency-inverse document frequency (TF-IDF) vectorizer Here's why: TF-IDF Vectorizer: This tool from Scikit-learn is effective in handling issues of rich vocabularies and low frequency words. TF-IDF down-weights words that appear frequently across documents (thus might be less informative) and gives more weight to words that appear less frequently but might be more indicative of the sentiment. This approach can enhance the model's ability to focus on more relevant features, potentially improving validation accuracy.

Comment: C I think c is correct. stemming involves reducing words to their root or base form, and stop word removal involves removing common words (e.g., "the," "and," "is") that may not contribute much to sentiment analysis. By using NLTK for stemming and stop word removal, you can simplify the vocabulary and potentially improve the model's ability to capture sentiment from the remaining meaningful words. A - syntax and entity recognition wont solve the scenario B - blaze text for words. D - capturing the importance of words in a document collection. frequency of a word in a document.

Comment: D is the correct guys

Comment: Amazon Comprehend's syntax analysis and entity detection are more about understanding the structure of sentences and identifying entities within the text rather than tackling the problem of a rich vocabulary with low average frequency of words. TF-IDF vectorization is a technique that can help reduce the impact of common, low-information words in the dataset while emphasizing the importance of more informative, less frequent words. This could potentially improve the validation accuracy by addressing the identified problem.

Comment: A. YES - he works on an application and not a model, Amazon Comprehend is the ready-to-use tool he wants; TF-IDF is built-in B. NO - word2vec will be challenged with low frequency terms; GloVe and FastText are better for that C. NO - the vocabulary is righ, so stemming and stop word removal will not address the core issue D. NO - right approach, but that is not "a tool"

Comment: Option D. This approach can help in reducing the impact of words that occur frequently in the dataset and increasing the impact of words that occur less frequently. This can help in improving the accuracy of the model.

Comment: The anwer is B. Blazing text can hadle OOV words as explained below. https://docs.aws.amazon.com/sagemaker/latest/dg/blazingtext.html

Comment: This is an AWS exam, so why would you choose anything other than A or B, and based on the link, it looks like B most likely

Comment: The passage “low average frequency of words” points directly to the use of TF-IDF. Letter A deviates from what the question proposes and is discarded. Letter B proposes a radical change in my POV. Letter C does not solve the passage mentioned at the beginning. Letter D is correct.

Comment: The Amazon SageMaker BlazingText algorithm provides highly optimized implementations of the Word2vec and text classification algorithms. The Word2vec algorithm is useful for many downstream natural language processing (NLP) tasks, such as *****sentiment analysis, named entity recognition, machine translation, etc. Text classification is an important task for applications that perform web searches, information retrieval, ranking, and document classification.

Comment: I would say since the buzzword "low average frequency" comes up, the safe choise would be the tfid vectorizer. I go for D.

Comment: The Scikit-learn term frequency-inverse document frequency (TF-IDF) vectorizer is a widely used tool to mitigate the high dimensionality of text data. Option A, Amazon Comprehend syntax analysis, and entity detection, can help in extracting useful features from the text, but it does not address the issue of high dimensionality. Option B, Amazon SageMaker BlazingText cbow mode, is a tool for training word embeddings, which can help to represent words in a lower dimensional space. However, it does not directly address the issue of high dimensionality and low frequency of words. Option C, Natural Language Toolkit (NLTK) stemming and stop word removal, can reduce the dimensionality of the feature space, but it does not address the issue of low-frequency words that are important for sentiment analysis.

Comment: Emphasis is on the rich words - so stemming can help reduce these to more common words. Blazing Text in cbow mode doesnt seem relevant is about providing words given a context. And TF-IDF I'm not sure would do anything except highlight the problem you are already having?

Comment: D. Scikit-learn term frequency-inverse document frequency (TF-IDF) vectorizer would be the best tool to use in this scenario. The TF-IDF vectorizer will give less weight to the less frequent words in the dataset, and allow the more informative and frequent words to have a greater impact on the sentiment analysis. This can help to improve the validation accuracy of the model.

Discussion for Question 65

Link: https://www.examtopics.com/discussions/amazon/view/10090-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: Ans: C; Normalization is correct

Replies:

Comment: Ans is not C. What is listed there is the definition of STANDARDIZATION. Normalization just scales and is not useful for reducing the effect of outliers

Replies:

Comment: nevermind ignore this

Comment: Guys, I passed the exam today. It is a tough one but there are many questions here. Good luck everyone! Thank examtopics

Replies:

Comment: Hi Phong! Please add my skype: haison8x

Comment: Ans: C; Normalization is correct

Comment: C (Yep, STANDARDIZATION is the correct name) That's an odd question for me

Comment: ans C is correct.

Comment: ANS should be C as Normalization work best in case of amplitude diff

Comment: Hi, guys, First thanks this website for the information it provided. However, the ML exam has updated most of the questions. only 20+ questions here are included in today's test. Anyway, it is still helpful. GOOD LUCK EVERYONE!

Replies:

Comment: So there are 40+ other questions on the exam that aren't included in Examtopics?

Comment: QUESTION 69 A large consumer goods manufacturer has the following products on sale: • 34 different toothpaste variants • 48 different toothbrush variants • 43 different mouthwash variants The entire sales history of all these products is available in Amazon S3. Currently, the company is using custom-built autoregressive integrated moving average (ARIMA) models to forecast demand for these products. The company wants to predict the demand for a new product that will soon be launched. Which solution should a Machine Learning Specialist apply? A. Train a custom ARIMA model to forecast demand for the new product. B. Train an Amazon SageMaker DeepAR algorithm to forecast demand for the new product. C. Train an Amazon SageMaker k-means clustering algorithm to forecast demand for the new product. D. Train a custom XGBoost model to forecast demand for the new product. Correct Answer: B

Replies:

Comment: https://aws.amazon.com/blogs/machine-learning/forecasting-time-series-with-dynamic-deep-learning-on-aws/ Answer: B

Comment: QUESTION 68 An agency collects census information within a country to determine healthcare and social program needs by province and city. The census form collects responses for approximately 500 questions from each citizen. Which combination of algorithms would provide the appropriate insights? (Select TWO.) A. The factorization machines (FM) algorithm B. The Latent Dirichlet Allocation (LDA) algorithm C. The principal component analysis (PCA) algorithm D. The k-means algorithm E. The Random Cut Forest (RCF) algorithm Correct Answer: CD

Replies:

Comment: I think the answer is A and B. The census question and answer will be in text. Use LDA (unsupervised algorithm) which takes the census question/answer and groups them into categories. Use the categorization to group the people and identify similar people. Use the Factorization Machine to group the people. For each person identify if they answer a question or not. Find the total questions they answered and that will be the Target variable. Now the problem is similar to movie recommendation (consider each question a movie and the total number of questions answered will be the Rating). Based on the questions a Person answered, Factorization Machine groups the people. Findings from both the algorithms can be used to compare and identify the people for the social programs.

Replies:

Comment: FM is mainly used in recommendation system to find hidden variables between two known variables to find correlation between two variables.

Comment: it's CD

Comment: https://aws.amazon.com/blogs/machine-learning/analyze-us-census-data-for-population-segmentation-using-amazon-sagemaker/ Answer: C and D

Comment: QUESTION 67 A. Use AWS Lambda to trigger an AWS Step Functions workflow to wait for dataset uploads to complete in Amazon S3. Use AWS Glue to join the datasets. Use an Amazon CloudWatch alarm to send an SNS notification to the Administrator in the case of a failure. B. Develop the ETL workflow using AWS Lambda to start an Amazon SageMaker notebook instance. Use a lifecycle configuration script to join the datasets and persist the results in Amazon S3. Use an Amazon CloudWatch alarm to send an SNS notification to the Administrator in the case of a failure. C. Develop the ETL workflow using AWS Batch to trigger the start of ETL jobs when data is uploaded to Amazon S3. Use AWS Glue to join the datasets in Amazon S3. Use an Amazon CloudWatch alarm to send an SNS notification to the Administrator in the case of a failure. D. Use AWS Lambda to chain other Lambda functions to read and join the datasets in Amazon S3 as soon as the data is uploaded to Amazon S3. Use an Amazon CloudWatch alarm to send an SNS notification to the Administrator in the case of a failure. Correct Answer: A

Comment: QUESTION 67 A Machine Learning Specialist is developing a daily ETL workflow containing multiple ETL jobs. The workflow consists of the following processes: • Start the workflow as soon as data is uploaded to Amazon S3. • When all the datasets are available in Amazon S3, start an ETL job to join the uploaded datasets with multiple terabyte-sized datasets already stored in Amazon S3. • Store the results of joining datasets in Amazon S3. • If one of the jobs fails, send a notification to the Administrator. Which configuration will meet these requirements?

Comment: QUESTION 66 A Machine Learning Specialist must build out a process to query a dataset on Amazon S3 using Amazon Athena. The dataset contains more than 800,000 records stored as plaintext CSV files. Each record contains 200 columns and is approximately 1.5 MB in size. Most queries will span 5 to 10 columns only. How should the Machine Learning Specialist transform the dataset to minimize query runtime? A. Convert the records to Apache Parquet format. B. Convert the records to JSON format. C. Convert the records to GZIP CSV format. D. Convert the records to XML format. Correct Answer: A

Discussion for Question 66

Link: https://www.examtopics.com/discussions/amazon/view/19369-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 5 votes

Discussion

Comment: Answer A seems correct...

Replies:

Comment: sorry, the link https://aws.amazon.com/blogs/big-data/prepare-data-for-model-training-and-invoke-machine-learning-models-with-amazon-athena/

Comment: A (Most queries will span 5 to 10 columns only)

Comment: Option A

Comment: clue is: most queries will span 5 to 10 column while there are 200 columns. Indicating Data Warehouse means columner storage. Option A is correct.

Comment: A. See https://aws.amazon.com/blogs/big-data/analyzing-data-in-s3-using-amazon-athena/

Discussion for Question 67

Link: https://www.examtopics.com/discussions/amazon/view/26038-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: A: Correct. S3 events can trigger AWS Lambda function. B: Wrong. There's nothing to do with SageMaker in the provided context. C: Wrong. AWS Batch cannot receive events from S3 directly. D: Wrong. Will not meed the requirement: "When all the datasets are available in Amazon S3..." https://docs.aws.amazon.com/step-functions/latest/dg/tutorial-cloudwatch-events-s3.html

Replies:

Comment: Actually, I think that D does meet the requirement of waiting until all datasets are in S3, BUT you do need Glue to join the datasets. Answer is still A.

Comment: I agree. Step Functions can be used to implement a workflow. In this case, wait for all the datasets to be loaded before triggering the glue job.

Comment: Option A

Comment: Batch isn't event driven, answer is A.

Comment: If EMR were present I would have chose that because of the size of dataset, else is Glue

Replies:

Comment: exactly, this is also where I got confused. Since Glue is not good at handling such large dataset, multiple terabyte-sized datasets + multiple ETL jobs + daily

Comment: A. The answer omits stuffs like Lambda functions and Event Bridge. https://aws.amazon.com/blogs/big-data/orchestrate-multiple-etl-jobs-using-aws-step-functions-and-aws-lambda/

Comment: https://d1.awsstatic.com/r2018/a/product-page-diagram-aws-step-functions-use-case-aws-glue.bc69d97a332c2dd29abb724dd747fd82ae110352.png

---

Discussion for Question 68

Link: https://www.examtopics.com/discussions/amazon/view/17281-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- CD: 9 votes

Discussion

Comment: C: (OK) Use PCA for reducing number of variables. Each citizen's response should have answer for 500 questions, so it should have 500 variables D: (OK) Use K-means clustering A: (Not OK) Factorization Machines Algorithm is usually used for tasks dealing with high dimensional sparse datasets B: (Not OK) The Latent Dirichlet Allocation (LDA) algorithm should be used for task dealing topic modeling in NLP E: (Not OK) Random Cut Forest should be used for detecting anormal in data

Comment: https://aws.amazon.com/blogs/machine-learning/analyze-us-census-data-for-population-segmentation-using-amazon-sagemaker/ Answer: C and D

Comment: Option C and D

Replies:

Comment: The answer depends on the type of question is it is open ended then would need LDA hence B and D but if the question is a feature then PCA should work

Comment: C and D are the way

Comment: CD, C - for reduce number of columns. D - for data clustering

Comment: C. The principal component analysis (PCA) algorithm D. The k-means algorithm PCA is a dimensionality reduction technique that can be used to identify the underlying structure of the census data. This algorithm can help to identify the most important questions and provide an overview of the relationship between the questions and the responses. K-means is an unsupervised learning algorithm that can be used to segment the population into different groups based on their responses to the census questions. This algorithm can help to determine the healthcare and social program needs by province and city based on the responses collected from each citizen. These algorithms can help to provide insights into the patterns and relationships within the census data, which can inform decision making for healthcare and social program planning.

Comment: Reduce dimensionality and cluster subjects.

Comment: This is the same question as Topic 2 Q3

Replies:

Comment: how to reach Topic 2 every questions here seem to belong to topic 1

Discussion for Question 69

Link: https://www.examtopics.com/discussions/amazon/view/17280-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 10 votes

Discussion

Comment: B https://docs.aws.amazon.com/sagemaker/latest/dg/deepar.html "...When your dataset contains hundreds of related time series, DeepAR outperforms the standard ARIMA and ETS methods. You can also use the trained model to generate forecasts for new time series that are similar to the ones it has been trained on."

Comment: "You can also use the trained model to generate forecasts for new time series that are similar to the ones it has been trained on" https://docs.aws.amazon.com/sagemaker/latest/dg/deepar.html

Comment: 'autoregressive integrated moving average (ARIMA)' <--> DeepAR. https://docs.aws.amazon.com/sagemaker/latest/dg/deepar.html

Comment: B - DeepAR is based on GluonTS, and can use multiple time series for learning

Comment: Option B

Comment: DeepAr for new products forever!

Comment: The DeepAR algorithm is a powerful time series forecasting algorithm that is designed to handle multiple time series data and can handle irregularly spaced time series data and missing values, making it a good fit for this task. Additionally, the large amount of sales history data available in Amazon S3 makes the use of a deep learning algorithm like DeepAR more appropriate.

Comment: B https://docs.aws.amazon.com/sagemaker/latest/dg/deepar.html

Comment: It is B

Comment: This is the same question as Topic 2 Q4

Discussion for Question 70

Link: https://www.examtopics.com/discussions/amazon/view/43708-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 13 votes

Discussion

Comment: Should be C. "You don't need to specify the AWS KMS key ID when you download an SSE-KMS-encrypted object from an S3 bucket. Instead, you need the permission to decrypt the AWS KMS key. When a user sends a GET request, Amazon S3 checks if the AWS Identity and Access Management (IAM) user or role that sent the request is authorized to decrypt the key associated with the object. If the IAM user or role belongs to the same AWS account as the key, then the permission to decrypt must be granted on the AWS KMS key's policy." https://aws.amazon.com/premiumsupport/knowledge-center/decrypt-kms-encrypted-objects-s3/?nc1=h_ls

Comment: Should be C. I think it is not possible to assign a key directly to a Sagemaker notebook instance like D suggests. Normally in AWS in general, IAM roles are used to do so. So C.

Comment: 'IAM role' principle of least privilege (PoLP)

Comment: IAM roles securely provide temporary AWS credentials that services (like SageMaker notebooks) can assume to access other resources. This avoids using long-lived access keys or directly embedding API keys into code. KMS Key Policy: This policy controls access to your KMS key. Granting the notebook's role permission within this policy lets SageMaker decrypt the data when reading from S3.

Comment: Seems to follow the best cloud authorization practice

Comment: IAM role associated with the SageMaker notebook instance must be given permissions in the KMS key policy to decrypt the data using the KMS key that was used for encryption.

Comment: answer is C

Comment: Assign an IAM role to the Amazon SageMaker notebook with S3 read access to the dataset. Grant permission in the KMS key policy to that role. To read data from Amazon S3 that is encrypted with AWS KMS, the Amazon SageMaker notebook instance needs to have both S3 read access and KMS decrypt permissions. This can be achieved by assigning an IAM role to the notebook instance that has the necessary policies attached, and by granting permission in the KMS key policy to that role.

Comment: C only.

Comment: Should be C. The reference doc provided did not have any information about assigning keys to the notebook. Doing so become very cumbersome as you can have 100's of notebooks and its not scalable. Someone needs to moderate these answers.

Comment: To allow an Amazon SageMaker notebook instance to read a dataset stored in an Amazon S3 bucket that is protected with server-side encryption using AWS KMS, the ML Specialist should assign an IAM role to the Amazon SageMaker notebook with S3 read access to the dataset. The IAM role should have permissions to access the S3 bucket and the KMS key that was used to encrypt the data. This role should be granted permission in the KMS key policy to allow it to decrypt the data.

Comment: To encrypt the machine learning (ML) storage volume that is attached to notebooks, processing jobs, training jobs, hyperparameter tuning jobs, batch transform jobs, and endpoints, you can pass a AWS KMS key to SageMaker. If you don't specify a KMS key, SageMaker encrypts storage volumes with a transient key and discards it immediately after encrypting the storage volume. For notebook instances, if you don't specify a KMS key, SageMaker encrypts both OS volumes and ML data volumes with a system-managed KMS key.

Replies:

> **Comment:** I correct myself- Option C is correct: Background AWS Key Management Service (AWS KMS) enables Server-side encryption to protect your data at rest. Amazon SageMaker training works with KMS encrypted data if the IAM role used for S3 access has permissions to encrypt and decrypt data with the KMS key. Further, a KMS key can also be used to encrypt the model artifacts at rest using Amazon S3 server-side encryption. Additionally, a KMS key can also be used to encrypt the storage volume attached to training, endpoint, and transform instances. In this notebook, we demonstrate SageMaker encryption capabilities using KMS-managed keys. resource: https://github.com/aws/amazon-sagemaker-examples/blob/main/advanced_functionality/handling_kms_encrypted_data/handling_kms_encrypted_data.ipynb Option D is correct if sagemaker does the encryption, if you are dealing with encrypted data then C is 100% correct.

Comment: C. Assign an IAM role to the Amazon SageMaker notebook with S3 read access to the dataset. Grant permission in the KMS key policy to that role. To access the encrypted dataset in Amazon S3, the Amazon SageMaker notebook instance must have the appropriate permissions. This can be achieved by assigning an IAM role to the notebook with read access to the dataset in Amazon S3 and granting permission in the KMS key policy to that role. This ensures that the notebook has the necessary permissions to access the encrypted data in Amazon S3, while adhering to best practices for securing sensitive data.

Comment: agreed with C

Comment: Answer is C : Open the IAM console. Add a policy to the IAM user that grants the permissions to upload and download from the bucket. You can use a policy that's similar to the following: https://aws.amazon.com/premiumsupport/knowledge-center/s3-bucket-access-default-encryption/ (number 2)

Comment: Seems to be D https://docs.aws.amazon.com/sagemaker/latest/dg/encryption-at-rest-nbi.html

Comment: Not D as if you assign the key in the notebook, that's not secure, it will make the encryption ineffective. Instead, you assign the access permission by using IAM.

Discussion for Question 71

Link: https://www.examtopics.com/discussions/amazon/view/43716-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 8 votes

Discussion

Comment: B it is .

Replies:

> **Comment:** I agree, B is serverless and reuses Pyspark. Similar example shown here: https://docs.aws.amazon.com/glue/latest/dg/aws-glue-programming-python-samples-medicaid.html

Comment: A is not correct because Minimize the number of servers that will need to be managed. EMR is not server-less. B is correct. AWS Glue supports an extension of the PySpark Python dialect for scripting extract, transform, and load... C is not correct because using Lambda for ETL you will not be able to Reuse existing PySpark logic D is not correct because Kinesis is not server-less. And you can not Reuse existing PySpark logic

Comment: Answer is A, as B clearly mentions that Pyspark code is written with leverage from already existing code. Also, the server architecture used currently is on-premises which will have more servers that solution A.

Comment: Amazon Kinesis Data Analytics is more suited for real-time processing and streaming data. The given use case does not indicate a need for real-time processing, so this might not be the best fit. Furthermore, it doesn't support PySpark natively.

Comment: Voted B based on the serverless (minimum servers) and https://docs.aws.amazon.com/glue/latest/dg/aws-glue-programming.html

Comment: Indeed B using Glue

Comment: B is the correct. A you have to manage EMR, so it's wrong. D you don't use Spark, so it's wrong. C you will not be using Spark, so it's wrong.

Comment: B ticks all boxes. Minimize servers -> AWS managed services -> Glue.

Comment: Option A would be the best response for this scenario. This solution allows the Data Scientist to reuse the existing PySpark logic while migrating the ETL process to the cloud. The raw data is written to Amazon S3, and a Lambda function is scheduled to trigger a Spark step on a persistent EMR cluster based on the existing schedule. The PySpark logic is used to run the ETL job on the EMR cluster, and the results are output to a processed location in Amazon S3 that is accessible for downstream use. This solution minimizes the number of servers that need to be managed, and it allows for a seamless migration of the existing ETL process to the cloud.

Comment: Option D is wrong it should be B

Comment: D cannot be answer as there is no streaming data or Realtime processing.

Comment: the answer is b

Comment: Answer should be B. Serverless, on a regular schedule (no real time requirement), reuses PySpark code in Glue ETL script.

Comment: Answer is B as they specifically ask about reusing existing PySpark, which can be done with Glue

Comment: https://docs.aws.amazon.com/glue/latest/dg/creating_running_workflows.html

Comment: It is B. ! "Minimize number of servers to be managed". B is a Serverless solution which fulfils other requirements!

Comment: I like both A & B however with B you would need to rewrite the Pyspark code to account for the ETL process you are now introducing, so it would not be using the original code. Both A & B are using managed services. Answer B would also require a notebook instance to get set up as there is no direct integration of Pysprak in Glue so there are some assumptions being made. On Balance, I am leaning towards answer A

Discussion for Question 72

Link: https://www.examtopics.com/discussions/amazon/view/74983-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- AC: 19 votes

Discussion

Comment: AC - correct answer

Comment: I think ACE are all correct

Comment: A. YES - standard for overfitting B. NO - we have already too much overfitting C. YES - feature elimination can reduce model complexity and thus overfitting D. NO - that does dimensionnality reduction to 2D or 3D, for visualization; we want more than a few features E. NO - LDA is an alternative to logistic regression; it may not address overfitting

Comment: A due to fitting C Recursive feature elimination (RFE) is a wrapper method that iteratively removes features based on their importance scores from a classifier. RFE starts with all features and then eliminates the least important ones until a desired number of features is reached. This can help to reduce the dimensionality of the dataset and improve the model performance by removing irrelevant or redundant features. The Marketing team can then interpret the model by looking at the remaining features and their importance scores.

Comment: AC are the correct

Comment: How can we add features to the dataset provided.... we can't make them up from thin air. Hopefully the moderators can provide some insight on this. I was thinking of paying for this site but the answers are all over the place.

Comment: A. Add L1 regularization to the classifier and C. Perform recursive feature elimination are the methods that can be used to improve the model performance and satisfy the Marketing team's needs. Explanation: A. Adding L1 regularization to the logistic regression classifier can help to improve the model performance and reduce overfitting. This can also help to highlight the relevant features for churn prediction as L1 regularization can shrink the coefficients of irrelevant features to zero. C. Recursive feature elimination can be used to select the most relevant features for the model. This can help to improve the model performance and highlight the relevant features for churn prediction.

Comment: A. Adding L1 regularization can help to reduce overfitting by shrinking the coefficients of less important features towards zero, which can improve the model's generalization performance on the validation set. C. Recursive feature elimination is a feature selection technique that removes the least important feature at each iteration and trains the model on the remaining features until a desired number of features is reached. This method can be used to identify the most relevant features for the prediction task and reduce the dimensionality of the dataset, leading to improved model performance and interpretability for the Marketing team.

Comment: AC - Key: logistic regression model = non linear in terms of Odds and Probability, however it is linear in terms of Log Odds. Key: Large gap between training & validation = overfitting => 5 techniques to prevent overfitting: 1. Simplifying the model | 2. Early stopping 3. Use data argumentation | 4. Use regularization | 5. Use dropouts A - yes to avoid overfitting (although i am thinking it is talking about regressor) Not B - add feature will lead to overfitting C - feature elimination - prevent overfitting Not D - t-SNE is a nonlinear dimensionality reduction technique Not E - find feature correlation only - Linear discriminant analysis (LDA), normal discriminant analysis (NDA), or discriminant function analysis is a generalization of Fisher's linear discriminant, a method used in statistics and other fields, to find a linear combination of features that characterizes or separates two or more classes of objects or events.

Comment: L1 won't do naturally the feature elimination? I guess AB

Comment: why not A & D? or C & D? does not t-SNE grant the marketing team's wish for visualization of relationships? or are we to presume that A&C are best as C (recursive feature elimination) grants us some visualization of feature importance.

Comment: AC is correct

Comment: overfitting: add regularization, remove features

Discussion for Question 73

Link: https://www.examtopics.com/discussions/amazon/view/43717-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- D: 23 votes

Discussion

Comment: D near-real time

Replies:

> **Comment:** The main problem with D is that Amazon Kinesis Data Firehose can not be a source service for Amazon Kinesis Data Analytics. The answer would be correct if it said "Using Amazon Kinesis Data Stream to ingest data, using Amazon Kinesis Data Analytics for defect detection and using Amazon Kinesis Data Firehose for storing data for further Analysis" https://docs.aws.amazon.com/firehose/latest/dev/create-name.html
>
> **Replies:**

Comment: Actually Kinesis Data Firehose can be used for Data Ingestion. So the correct option is still D

Comment: Glad we are all in agreement D is the correct answer

Comment: D - firehose for near realtime

Comment: Kinesis Data Firehose is a fully managed service that can ingest streaming data and load it into destinations like S3, Redshift, Elasticsearch. and with Kinesis Data Analytics and RCF and then Data Firehose again to store on S3. D is the best choice.

Comment: https://docs.aws.amazon.com/managed-flink/latest/java/get-started-exercise-fh.html

Comment: Kinesis seems like the only viable option

Comment: The answer is D. Since, data is continuously coming in Kinesis datafirehose is our streaming application (also we need near Real time defect detection and storage in S3) and anomaly detection can be done by kinesis data application (RCF algorithm).

Comment: D, near real-time ingestion is the key

Comment: A. NO - AWS IoT will first store the data, then make it available for Analytics/Jupyter (https://docs.aws.amazon.com/iotanalytics/latest/userguide/welcome.html); so not real-time B. NO - not realtime to store the data before analytics C. NO - not realtime to store the data before analytics D. YES - real-time pipe, RCF best for anomalities

Comment: How can someone use S3 for ingestion? Firehose is the right answer

Comment: This option meets the requirements of performing near-real time defect detection, storing all the data for offline analysis, and handling 200 performance metrics in a time-series. Amazon Kinesis Data Firehose is a fully managed service that can ingest streaming data from various sources and deliver it to destinations such as Amazon S3, Amazon OpenSearch Service, and Amazon Redshift. Amazon Kinesis Data Analytics is a service that can process streaming data using SQL or Apache Flink applications. Amazon Kinesis Data Analytics provides a built-in RANDOM_CUT_FOREST function, a machine learning algorithm that can detect anomalies in streaming data1. This function can handle high-dimensional data and assign an anomaly score to each record based on how distant it is from other records1. The anomaly scores can then be delivered to another destination using Kinesis Data Firehose or consumed by other applications using Kinesis Data Streams.

Comment: D is the correct If the question says "data streaming", "real time data" or "near real time" you should look for kinesis services. B and C are totally wrong: It's not possible to use S3 to ingestion, only storage.

Comment: D, https://docs.aws.amazon.com/kinesisanalytics/latest/sqlref/sqlrf-random-cut-forest.html

Comment: At a minimum the moderators should put some explanation when the community vote overwhelmingly for a different option.

Comment: Option D is not necessarily incorrect, but it may not be the most effective approach to perform near-real time defect detection in this scenario. Here are some potential drawbacks of this approach: Amazon Kinesis Data Firehose is primarily used for data ingestion and delivery to other services, and may not be the best choice for real-time analysis. Using Amazon Kinesis Data Analytics for anomaly detection may be less flexible than using Amazon SageMaker, which provides a wide range of algorithms and models for anomaly detection. Random Cut Forest (RCF) is a popular anomaly detection algorithm used for time-series data, and Amazon SageMaker provides an RCF implementation that can be used for anomaly detection in real-time or offline. While Amazon Kinesis Data Analytics also provides RCF, using Amazon SageMaker may be a better choice for scalability and flexibility.

Comment: Yes, option C can provide near real-time defect detection. Amazon SageMaker's Random Cut Forest (RCF) algorithm is designed to work with streaming data and can detect anomalies in near real-time. It can process data in batches as small as a single data point, making it well-suited for real-time anomaly detection. In this scenario, if the manufacturing process is generating data in real-time, it can be ingested into Amazon S3 and processed by Amazon SageMaker's RCF algorithm, allowing for near real-time detection of critical manufacturing defects during testing.

Replies:

Comment: this is ridiculous. How can you store in s3 and then conduct real-time analysis?

Comment: Firehose to ingest in near real time and RCF to do Anomaly is the best approach

---

Discussion for Question 74

Link: https://www.examtopics.com/discussions/amazon/view/43916-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: CE is the right answer. ECR uses ECS internally while using SGM.

Replies:

Comment: CE based on criteria and this documentation: https://docs.aws.amazon.com/sagemaker/latest/dg/sagemaker-mkt-create-model-package.html "For Location of inference image, type the path to the image that contains your inference code. The image must be stored as a Docker container in Amazon ECR. For Location of model data artifacts, type the location in S3 where your model artifacts are stored."

Replies:

Comment: the answer is correct but the explanation is completely wrong. The question is about how to create your own algorithm using container, not "put the inference in market" (which is your resource link). the right citation should be "Adapting your own training container": create s3 to store model articraft, and push code to ECR.

Comment: CE IS THE CORRECT ANSWER 100%

Comment: Amazon ECR is a fully managed container registry service that allows users to store, manage, and deploy Docker container images. Amazon SageMaker supports using custom Docker images for training and inference, which can contain the user's own training algorithm and any external assets or dependenciesAd1. The user can push their Docker image to Amazon ECR and then reference it in their Amazon SageMaker training job configurationAd1.

Comment: CE is correct!

Comment: ECR for the code, S3 for the parameters!

Comment: C contain the algorithm's image and E contain algorithm's parameters.

Comment: The location of the model artifacts. Model artifacts can either be packaged in the same Docker container as the inference code or stored in Amazon S3. Not so sure.

Comment: https://aws.amazon.com/blogs/machine-learning/bringing-your-own-custom-container-image-to-amazon-sagemaker-studio-notebooks/ If you wish to use your private VPC to securely bring your custom container, you also need the following: A VPC with a private subnet VPC endpoints for the following services: Amazon Simple Storage Service (Amazon S3) Amazon SageMaker Amazon ECR AWS Security Token Service (AWS STS) CodeBuild for building Docker containers Answer C+E

Comment: For me CD. needs storage and create a custom docker using ECR to store it.

Replies:

Comment: Sorry, CE is correct.

Replies:

Comment: Sagemaker will spin up the instances needed with the right image. No need to use ECS. CE is right

---

Discussion for Question 75

Link: https://www.examtopics.com/discussions/amazon/view/43915-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: C is correct . SageMakerVariantInvocationsPerInstance = (MAX_RPS * SAFETY_FACTOR) * 60 AWS recommended Saf_fac =0 .5

Comment: Answer C: SageMakerVariantInvocationsPerInstance = (MAX_RPS * SAFETY_FACTOR) * 60 https://docs.aws.amazon.com/sagemaker/latest/dg/endpoint-scaling-loadtest.html

Comment: To calculate the SageMakerVariantInvocationsPerInstance setting, we can use the following equation from the web search results1: SageMakerVariantInvocationsPerInstance = (MAX_RPS * SAFETY_FACTOR) * 60 Where MAX_RPS is the maximum RPS that the variant can handle, SAFETY_FACTOR is the safety factor that we choose to ensure that we don't exceed the maximum RPS, and 60 is to convert from RPS to invocations-per-minute. Plugging in the given values, we get: SageMakerVariantInvocationsPerInstance = (20 * 0.5) * 60 SageMakerVariantInvocationsPerInstance = 10 * 60 SageMakerVariantInvocationsPerInstance = 600 Therefore, the Specialist should set the SageMakerVariantInvocationsPerInstance setting to 600.

Comment: SageMakerVariantInvocationsPerInstance = (MAX_RPS * SAFETY_FACTOR) * 60

Comment: SageMakerVariantInvocationsPerInstance = (MAX_RPS * SAFETY_FACTOR) * 60 (20RPS * 0.5Safety Factor) * 60 (10)*60 = 600 Answer C

Comment: Maximum request at peak time = 20 RPS = 20x60 = 1200RPM Safety factor of 0.5 = 1200*0.5 = 600 Basic setting of parameter = 600 (requests per minutes)

Discussion for Question 76

Link: https://www.examtopics.com/discussions/amazon/view/43874-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- D: 13 votes

Discussion

Comment: I think the right answer is D

Comment: D is correct. C is not the best the answer because the question states that tuning parameters doesn't help a lot. Transfer learning would be better solution!

Comment: A. NO, transfer learning helps, word2vec > TD-ITF as the first keeps into account part of the word context (there is a hyperparameter for this) B. LTSM delivers better results wrt GRU which is in turn a compromise architecture to balance accuracy with training time/cost C. Heperparameters tuning has been already applied, this will not help D. YEs, transfer learning will help and word3vec is better option in this scenario

Comment: How are the 'correct' answers being provided? I'm seeing so many answers that seem to be wrong and usually, the community vote seems to be correct. This is kind of frustrating.

Comment: Word2vec is a technique that can learn distributed representations of words, also known as word embeddings, from large amounts of text data. Word embeddings can capture the semantic and syntactic similarities and relationships between words, and can be used as input features for neural network models. Word2vec can be trained on domain-specific corpora to obtain more relevant and accurate word embeddings for a particular task.

Comment: From my perspective, B and C are wrong because the DS already tried something close to this. D is correct.

Comment: I don't think High Dimensionality is take care by C2V; TF-IDF is required. A.

Comment: Transfer learning, in my experience, has been a good way to boost performance when hyperparameter tuning did not work.

Comment: The case ask for predicting labels for sentences, the appropriate algo should be "Text Classification" Which, just as "wrod2vec,i part of Blazing Text.

Comment: The answer should be D. My reasoning is that by using a word embedding which is trained on domain specific material, the embeddings between two words are more domain specific. This means that relations (good or bad) are represented in a better way, which also means that the model should be able to predict the results in a more accurate way.

Comment: both A & D "seem" correct, but word2vec takes ORDER of words into acc (to some extent)--while TF-IDF does not. Thus max boost is from D. B,C are wrong because the DS has tried several network architectures (aka LSTM) and hyperparameter tuning (aka option C)

Comment: i think answer is A as The model reviews multi-page text documents

Replies:

Comment: I think that the general tf-idf vectors cannot be directly adapted to the deep learning model, because of the large dimension in vector values

Comment: I think it should be B A/D are false flags because the question doesn't specify what kind of data engineering is currently done on the inputs, as a baseline Per wikipedia, for GRUs, "GRUs have been shown to exhibit better performance on certain smaller and less frequent datasets", which fits the context of a particular energy sector

Comment: why not B??

Replies:

Comment: Generally, LSTM has the better performance then GRU in large datasets such as multi-page documents. GRU has advantages of memory allocation and training time.

Comment: Early stopping can give the model better performance, but I think that the model needs more condition like patience value for early stopping. This is because the model doesn't always show the performance at its maximum when the validation loss stops decreasing.

Comment: It cannot be C, because hyper parameter tuning didnt work as given in question. Also, A and D are same, however, word2vec model internally implements tf-idf much more efficiently. So answer got to be D

Replies:

Comment: but they need to classify the whole sentence i think for such a case we use object2vec not word2vec, but since it's not available in the answers, B is the only answer left.

Comment: I go for C

Comment: Agree. D seems more reasonable, as word2vec provides content of the sentences which is very important for evaluation of the risk.

Discussion for Question 77

Link: https://www.examtopics.com/discussions/amazon/view/43721-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- BC: 12 votes

Discussion

Comment: should be BC

Replies:

Comment: Agreed, AWS Example: https://docs.aws.amazon.com/datapipeline/latest/DeveloperGuide/what-is-datapipeline.html

Comment: It is obviously B and C, I am frustrated with the number of wrong answers. Why the moderator's answers keep being super weird?

Comment: B for near real-time C for hourly

Comment: The right answer is BC

Comment: AWS Data Pipeline (Option C) can be used to move the hourly data, as it provides a way to move data from various sources to Amazon EMR for processing. Amazon Kinesis (Option B) can be used to process data in near-real time, as it is a real-time data streaming service that can handle large amounts of incoming data from multiple sources. The data can be fed to Amazon EMR MapReduce jobs for processing.

Comment: should be BC

Comment: Kinesis for near realtime data and pipeline for the other data moved hourly.

Comment: AWS ES is an elastic search , it is nothing to do with this question.

Comment: Kinesis data into EMR: https://docs.aws.amazon.com/emr/latest/ReleaseGuide/emr-kinesis.html

Comment: BC. easy

Comment: I believe the answer is BC

Comment: BD. Data Pipeline is to orchestrate the workflow, how can that feed data to the MR jobs?

Comment: Answer is B and C

Comment: Answer is for sure BC

Comment: Ans is BC. (https://aws.amazon.com/jp/emr/?whats-new-cards.sort-by=item.additionalFields.postDateTime&whats-new-cards.sort-order=desc)

Discussion for Question 78

Link: https://www.examtopics.com/discussions/amazon/view/43938-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 14 votes

Discussion

Comment: Ans : A Refer the below : https://sagemaker-workshop.com/custom/containers.html

Comment: A https://sagemaker-workshop.com/custom/containers.html

Comment: You need the container to be hosted on ECR.

Comment: A. YES - the inference code is built after inspecting the coefficient of the Linear Model (or, alternatively, the model can be serialized via pickle and the inference code is simply to unserialized the mode); ECR is only registry supported by SageMaer; tagging the Docker image with the registry hostname (eg. docker tag image1 public.ecr.aws/g6h7x5m5/image1) is required so that the docker push command knows where to push the image B. NO - no need to compress; image must be on ECR C. NO - no need to compress; image must be on ECR D. NO - image must be on ECR

Comment: A is the right answer

Comment: For SageMaker to run a container for training or hosting, it needs to be able to find the image hosted in the image repository, Amazon Elastic Container Registry (Amazon ECR). The three main steps to this process are building locally, tagging with the repository location, and pushing the image to the repository.

Comment: A for sure.

Comment: Answer is A.

Comment: Docker Hub is a repository so ANS D makes no sense. Option A is the way to go.

Discussion for Question 79

Link: https://www.examtopics.com/discussions/amazon/view/43940-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 17 votes

Discussion

Comment: B is the right answer

Comment: B to use as storage with policies

Comment: EMR/HDFS is not more 'flexible' than S3

Comment: A. NO - volume too big for a DB B. YES C. NO - instance access will not control HDFS access D. NO - EFS does not use IAM policies (it is unix)

Comment: S3 indeed

Comment: S3 always

Comment: I would say the answer is B not because of the cost on EMR,. that is also a current answer. however: "most processing flexibility" indicates that S3 is a better option. because all ML solutions and work flows integrate with S3. it hasn't spoken what the ML solution and which services so I take the safe side and go with S3

Comment: C is not affordable because it is ephemeral storage. https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-plan-file-systems.html "HDFS is used by the master and core nodes. One advantage is that it's fast; a disadvantage is that it's ephemeral storage which is reclaimed when the cluster ends. It's best used for caching the results produced by intermediate job-flow steps."

Replies:

Comment: the question does not require long-term storage.

Comment: C is correct. it says real time data and to be used for ml process so EMR more suitable. also S3 bucket policies not same as IAM users so B is not correct.

Replies:

Comment: Why will you need to spin up servers (EMR) just to store visual data for ML?

Comment: I think Amazon EMR is more appropriate, as the data scheme stated is a big data scheme. https://aws.amazon.com/emr/?whats-new-cards.sort-by=item.additionalFields.postDateTime&whats-new-cards.sort-order=desc

Replies:

Comment: IAM support is required for storage feature , that is not possible as per options described as IAM is supported for HDFS for the instance running on top of it, hence B should be correct

Comment: B is the right answer

Comment: S3 is the easy, scalable and secure option to store the image data.

Comment: B is the right answer

Comment: B is an appropriate choice

Discussion for Question 80

Link: https://www.examtopics.com/discussions/amazon/view/43919-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 10 votes

Discussion

Comment: Answer C. Need reduce the features preserving the information on it this is achieve using PCA.

Replies:

Comment: without losing a lot of information from the original dataset since when PCA retains information?

Comment: PCA helps to speed up the training

Comment: Answer is A, because one must avoid information loss that PCA or autoencoders introduce through new features (https://www.i2tutorials.com/what-are-the-pros-and-cons-of-the-pca/). Otherwise, I would perform C.

Replies:

Comment: If you REMOVE highly correlated features(that means in pairs), the model lost a lot of information.

Comment: A doesn't have sense. Self-correlation is for times series data, not for pair correlation

Comment: Answer is C

Comment: Answer C PCA (Principal Component Analysis) takes advantage of multicollinearity and combines the highly correlated variables into a set of uncorrelated variables. Therefore, PCA can effectively eliminate multicollinearity between features. https://towardsdatascience.com/how-do-you-apply-pca-to-logistic-regression-to-remove-multicollinearity-10b7f8e89f9b#:~:text=PCA%20(Principal%20Component%20Analysis)%20takes,effectively%20eliminate%20multicollinearity%20between%20features.

Comment: Option C

Comment: An autoencoder is a type of neural network that can learn a compressed representation of the input data, called the latent space, by encoding and decoding the data through multiple hidden layers1. PCA is a statistical technique that can reduce the dimensionality of the data by finding a set of orthogonal axes, called the principal components, that capture the most variance in the data2. Both methods can transform the original features into new features that are lower-dimensional, uncorrelated, and informative.

Comment: C is the correct. Self-correlation is for time series, which is not mention here. Besides that, even if was correlation only, try to do this in thousand features...

Comment: A . run correlation matrix and remove highly correlated features.

Comment: PCA for feature reduction

Comment: is it just me or is every 15th answer here PCA?

Comment: Using an autoencoder or PCA can help reduce the dimensionality of the dataset by creating new features that capture the most important information in the original dataset while discarding some of the noise and highly correlated features. This can help speed up the training time and reduce overfitting issues without losing a lot of information from the original dataset. Option A may remove too many features and may not capture all the important information in the dataset, while option B only rescales the data and does not address the issue of highly correlated features. Option D is not a feature engineering technique and may not be an effective way to reduce the dimensionality of the dataset.

Comment: PCA builds new features starting from high correlated ones. So it matches the question

Comment: It's C. The Data Scientist should use principal component analysis (PCA) to replace the original features with new features. PCA is a technique that reduces the dimensionality of a dataset by projecting it onto a lower-dimensional space, while preserving as much of the original variation as possible. This can help to speed up the training time of the model and reduce overfitting issues, without losing a significant amount of information from the original dataset.

Comment: C: PCA is the solution

Comment: Correction to C. Removing correlated features from hundreds of columns will be tedious and time consuming. PCA is the way to go here. Apologies for the flip

Comment: Answer is A. Eliminate features that are highly correlated. This will not compromise the quality of the feature space as much as PCA would.

Comment: Answer is C

Discussion for Question 81

Link: https://www.examtopics.com/discussions/amazon/view/43921-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- D: 14 votes

Discussion

Comment: For me answer is D, adjust to higher weight for class of interest: https://androidkt.com/set-class-weight-for-imbalance-dataset-in-keras/. More data may/may not be available and a data labeling job will take time.

Comment: I believe is C, because we already made all changes possible in MLP hidden layers and the results have not improved then we must change model so XGBoot seems the best option

Comment: In this case, the data scientist is training a multilayer perceptron (MLP), which is a type of neural network, on a dataset with multiple classes. The target class of interest is unique compared to the other classes within the dataset, but it does not achieve an acceptable recall metric. Recall is a measure of how well the model can identify the relevant examples from the minority class. The data scientist has already tried varying the number and size of the MLP's hidden layers, which has not significantly improved the results. A solution to improve recall must be implemented as quickly as possible.

Comment: The fastest one is D

Comment: "quickly as possible" mean do not change to new stuff, so it's D.

Comment: Not C, as the question ask for a quick solution. I accept D.

Comment: Answer C : https://towardsdatascience.com/boosting-techniques-in-python-predicting-hotel-cancellations-62b7a76ffa6c

Comment: Adding class weights to the MLP's loss function balances the class frequencies in the cost function during training, so the optimization process focuses more on the underrepresented class, improving recall.

Comment: I have done this before, class weights help with unbalanced data. Only logical solution that would help if not done, XGBoost could be different, but who knows, both NNs and XGBoost have comparable performance. Answer D!

Comment: In this example, it is necessary to improve recall as soon as possible, so instead of creating additional datasets, it is effective to change the weight of each class during learning.

Comment: C: 'distinct' indicates we can simplify this as a binary classification problem; then, NN is just overkill. plus, retraining a NN is much slower than training an XGboost model

Comment: I feel answer is B. Question says Target is different than the input data which is hint for anomaly detection.

Replies:

Comment: stop overthink

Comment: I believe the answer is C because we need to use hyperparameters to improve model performance.

Comment: In case of the quickest possible way, D seems fine. For XGBoost, it will take a bit of time to code again

Comment: For me Answer A. Why no other model instead xgBoost, the model need more labeled data to be trained and learn more positive examples.

Replies:

Comment: A is incorrect. Even if you hire Amazon Mechanical Turk, you won't have more data. This question is NOT asking about "labeling".

Discussion for Question 82

Link: https://www.examtopics.com/discussions/amazon/view/43922-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 11 votes

Discussion

Comment: Answer B.

Comment: B IS NOT CORRECT! Return the probability. Not the 1 or 0. D IS THE CORRECT ANSWER.

Replies:

Comment: Regression Classification is a made-up term, any binary classifier makes decisions based on probability score.

Comment: there is nothing like regression classification. (instead it should have said logistic regression). It should be Binary. i.e., either fraud or non fraud. Even with probabilities, we have a threshold to decide the class.

Comment: Logistic regression will give the probability, and logistic regression is a binary classification algorithm. https://machinelearningmastery.com/logistic-regression-for-machine-learning/

Comment: I always see that the community voting is more appropriate and the moderator answer looks out to be on wrong side. I see this for almost in 1 out of 5 questions. Which answer should we consider here as right one ??

Comment: Its definitely a classification problem, and between Binary and Streaming classification. Binary classification makes more sence

Comment: Binary classification

Comment: B, easy.

Comment: B, obviously! from sklearn.linear_model import LogisticRegression log_reg = LogisticRegression() log_reg.fit(X_train, y_train) log_reg.predict_proba(X_test) =)

Comment: The correct solution obviously is binary classification. For the comment above that says that binary classfication doesn't returns a probablity (for example SVM(classification) only returns a class and logistic, RFClassifier, XGBoostClassifier gives a probability and also a class given a threshold), you should ask yourself if that a regressor model returns always a probability, that is, if there is a restriction in a regressor model to predict values only in [0,1].

Comment: The Specialist is trying to determine whether a given transaction is fraudulent or not, which is a binary outcome (yes or no). Therefore, the problem should be framed as binary classification. The goal is to predict the probability of a transaction being fraudulent or not, and based on that, the Specialist can make a binary decision (fraudulent or not).

Comment: This is just binary classification, I don't understand how it could be anything else

Comment: It's B. This business problem can be framed as a binary classification problem, where the goal is to predict whether a given transaction is fraudulent (positive class) or not fraudulent (negative class). The model should output a probability for each transaction, indicating the likelihood that it is fraudulent.

Comment: should be D

Comment: Logistic regression models the probability of the default class (e.g. the first class). For example, if we are modeling people's sex as male or female from their height, then the first class could be male and the logistic regression model could be written as the probability of male given a person's height.

Comment: I think the answer is B, fraud has various cases which hard to define. So, Classification result will be fraud or not fraud. If Multi-category classfication, must define case of fraud in detaily

Replies:

Comment: More specifically, anomaly detection model will be needed

Comment: I believe the answer is B: it a binary classification problem because we are classifying an observation into one of two categories and the target variable in this problem is limited to two options: fraudulent or not fraudulent

Comment: well, regression classification is bullshit, I hope they formulate their questions better on the real exam. binary classification gives probability between 0 and 1

Comment: D for me. I think they want to talk about Logistic Regression, which is used as classification.

Discussion for Question 83

Link: https://www.examtopics.com/discussions/amazon/view/43923-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: B is the correct answer.

Comment: Linear regression

Comment: B, the only model for regression in the options.

Comment: Answer B

Comment: Answer B.

Discussion for Question 84

Link: https://www.examtopics.com/discussions/amazon/view/43942-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 12 votes

Discussion

Comment: B is correct answer .

Replies:

Comment: why B is the correct answer and not C?

Replies:

Comment: A square matrix is singular, that is, its determinant is zero, if it contains rows or columns which are proportionally interrelated; in other words, one or more of its rows (columns) is exactly expressible as a linear combination of all or some other its rows (columns), the combination being without a constant term.

Comment: For example. If you have two variables, X and Y, and you have two data points. You want to solve the problem: aX1+bY1 = Z1, aX2 + bY2 = Z2. However, if Y=2X -> Y1 = 2X1, Y2 = 2X2, then problem becomes: aX1+bY1 = Z1, a*2X1 + b*2Y1 = Z2 = 2*Z1. So you end up with only one function: aX1+bY1=Z1, meaning there will be more than one answer for (a, b). If you are familiar with linear algebra, it's easier to express the concept.

Comment: Agree, B.

Comment: B: If two features in the dataset are perfectly linearly dependent, it means that one feature can be expressed as a linear combination of the other. This can create a singular matrix during optimization, as the linear model would be trying to fit a linear equation to a dataset where one variable is fully determined by the other. This would lead to an ill-defined optimization problem, as there would be no unique solution that minimizes the sum of the squares of the residuals. This could lead to problems during training, as the model would not be able to find appropriate parameter values to fit the data.

Comment: Option B

Comment: The presence of linearly dependent features means that they are redundant, and provide no additional information to the model. This can result in a matrix that is not invertible, which is a requirement for solving a linear least squares regression problem. The presence of a singular matrix can also cause numerical instability and make it impossible to find an optimal solution to the optimization problem.

Comment: linera dependence creates singular matrix that causes problems at the moment we fit the modle

Comment: https://towardsdatascience.com/multi-collinearity-in-regression-fe7a2c1467ea B - two features are perfectly linearly dependent = singular matrix during optimization Not D - Not100% correct (as Multicollinearity happens when independent variables in the regression model are highly correlated to each other) they can still be independent variables

Comment: Consider one of the 5 assumptions of linear regression. This situation violates the assumption of "No multicollinearity between feature variables" Hence, D

Comment: B. See the multicollinearity problem in wikipedia https://en.wikipedia.org/wiki/Multicollinearity (second paragraph)

Comment: This issue is overfitting.

Discussion for Question 85

Link: https://www.examtopics.com/discussions/amazon/view/44056-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: B is the correct answer. Straightforward!

Comment: https://docs.aws.amazon.com/machine-learning/latest/dg/multiclass-model-insights.html

Comment: to be able to understand this Multiclass Model Insights and to be able to answer this question : True class-frequencies in the evaluation data: The second to last column shows that in the evaluation dataset, 57.92% of the observations in the evaluation data is Romance, 21.23% is Thriller, and 20.85% is Adventure. Predicted class-frequencies for the evaluation data: The last row shows the frequency of each class in the predictions. 77.56% of the observations is predicted as Romance, 9.33% is predicted as Thriller, and 13.12% is predicted as Adventure. REF: https://docs.aws.amazon.com/machine-learning/latest/dg/multiclass-model-insights.html

Comment: 12-sep exam

Comment: The image can be found here: https://vceguide.com/what-is-the-true-class-frequency-for-romance-and-the-predicted-class-frequency-for-adventure/

Comment: No image is there!

Comment: WHy there is no image? Admin. Please fix it.

Comment: B is correct

Comment: A seems to be correct

Discussion for Question 86

Link: https://www.examtopics.com/discussions/amazon/view/43943-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 9 votes

Discussion

Comment: C seems correct as per documentations.

Comment: I would answer C: https://docs.aws.amazon.com/sagemaker/latest/dg/your-algorithms-training-algo-dockerfile.html "To configure a Docker container to run as an executable, use an ENTRYPOINT instruction in a Dockerfile. SageMaker overrides any default CMD statement in a container by specifying the train argument after the image name"

Comment: I thought it was D, but it is C. It's not D because we copy the TRAINING code into /opt/ml/code/train.py

Comment: In Docker, the ENTRYPOINT instruction is used to specify the executable that should be run when the container starts. However, Amazon SageMaker expects the training script to be launched by specific commands provided by SageMaker itself, rather than relying solely on the Docker container's ENTRYPOINT. The convention for using Docker containers with Amazon SageMaker is to copy the training script and associated resources to specific directories within the container, such as /opt/ml/code, and let SageMaker manage the execution of the training process. I would go with D

Comment: C is correct

Comment: C is correct

Comment: Amazon SageMaker requires that a custom algorithm container has an executable named train that runs your training program. This executable can be configured as an ENTRYPOINT in the Dockerfile, which specifies the default command to run when the container is launched.

Comment: Amazon SageMaker requires that a custom algorithm container has an executable named train that runs your training program. This executable can be configured as an ENTRYPOINT in the Dockerfile, which specifies the default command to run when the container is launched.

Comment: you are all wrong, it is D based on,https://docs.aws.amazon.com/sagemaker/latest/dg/adapt-training-container.html

Comment: To package a Docker container for use with Amazon SageMaker, the training program should be configured as an ENTRYPOINT named train in the Dockerfile. This means that the training program will be automatically executed when the container is launched by Amazon SageMaker, and it can be passed command-line arguments to specify hyperparameters or other training settings.

Comment: The recommended option to package the Docker container for Amazon SageMaker is to configure the training program as an ENTRYPOINT named train. This is because ENTRYPOINT allows you to specify a command that will always be executed when the Docker container is run, ensuring that the training program will always run when the container is launched by Amazon SageMaker. Additionally, naming the ENTRYPOINT "train" is a convention used by Amazon SageMaker to identify the main training script.

Comment: It's C

Comment: C for sure as per AWS docs: > In your Dockerfile, use the exec form of the ENTRYPOINT instruction: > ENTRYPOINT ["python", "k-means-algorithm.py"]

Comment: C is correct

Comment: option C https://github.com/awsdocs/amazon-sagemaker-developer-guide/blob/master/doc_source/your-algorithms-training-algo-dockerfile.md

---

Discussion for Question 87

Link: https://www.examtopics.com/discussions/amazon/view/43728-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- BD: 6 votes

Discussion

Comment: I would go with B,D. Refer to quantile binning and log transform below. https://towardsdatascience.com/understanding-feature-engineering-part-1-continuous-numeric-data-da4e47099a7b

Replies:

Comment: agree B&D. both are strategies to eliminate the effect of skewing

Comment: Agree with B &D B binning for age D for make income in normal dist

Comment: SHOULD BE C,D

Comment: D. Logarithmic Transformation: Addresses the right-skewed income and age distributions. The log function compresses large values, reducing the impact of outliers and making the distributions closer to normal. B. Numerical Value Binning: Useful for the age distribution. By grouping ages into bins (e.g., 20-29, 30-39, etc.), you reduce the impact of the right skew caused by fewer older individuals. While it doesn't achieve a perfectly normal distribution, it often makes the feature more interpretable and manageable for modeling.

Comment: B and D

Comment: Agree with B &D B binning for age D for make income in normal dist

Comment: BD is correct

Comment: A and E, it asks incorrectly

Comment: B & D. Reasonable explanation in below discussion.

Comment: BD With age, always do quantile binning With skewed data, always use log.

Comment: B because we have skewed data with few exeptions D log transform can change distribution of data not C - because there is no indicaiton in the text, that data is following any of the HIGH DEGREE polynomial distribution like x^ 10

Comment: should be c and d

Comment: polynomial transformations can also be used for skewed data. https://machinelearningmastery.com/polynomial-features-transforms-for-machine-learning/

Comment: It seems the ans are C,D https://anshikaaxena.medium.com/how-skewed-data-can-skrew-your-linear-regression-model-accuracy-and-transfromation-can-help-62c6d3fe4c53

Discussion for Question 88

Link: https://www.examtopics.com/discussions/amazon/view/73881-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 21 votes

Discussion

Comment: I think C will be answer, because we even don't know how many layers now, so apply L1,L2 and dropouts layer will be first resort to solve overfitting. If it still does not work, then to reduce layers

Comment: D: D is the correct answer. C could be the answer only if it is a regression problem. You cannot apply L1 (Lasso regression) and L2 (Ridge regression) to classification problems. However, you can use dropout here.

Replies:

Comment: Why do you think it works only for regression problems? L1/L2 regularizations are just adding penalties to loss functions. I don't see any problems with applying it to DL model

Comment: C Regulization

Comment: if you see overfit think regularization.

Comment: C is the correct answer: The overfitting problem can be addressed by applying regularization techniques such as L1 or L2 regularization and dropouts. Regularization techniques add a penalty term to the cost function of the model, which helps to reduce the complexity of the model and prevent it from overfitting to the training data. Dropouts randomly turn off some of the neurons during training, which also helps to prevent overfitting.

Comment: D can work, but C is a better answer!

Comment: C and D both seems to be correct but, seems like removing layer is first step in to optimization https://www.kaggle.com/general/175912 d

Comment: C. Apply L1 or L2 regularization and dropouts to the training" because regularization can help reduce overfitting by adding a penalty to the loss function for large weights, preventing the model from memorizing the training data. Dropout is a regularization technique that randomly drops out neurons during the training process, further reducing the risk of overfitting.

Comment: "The first step when dealing with overfitting is to decrease the complexity of the model. To decrease the complexity, we can simply remove layers or reduce the number of neurons to make the network smaller." https://www.kdnuggets.com/2019/12/5-techniques-prevent-overfitting-neural-networks.html

Comment: Deep learning tuning order: 1. Number of layers 2. Number of neurons (indirectly implements dropout) 3. L1/L2 regularization 4. Dropout

Replies:

Comment: the problem is overfitting, not HP Tuning.

Comment: Here we are looking to reduce the Overfitting to improve the generalization. In order to do so, L1(or Lasso) regression has always been a good aide.

Replies:

Comment: This is not a regression problem at all.

Comment: C, Regularization and dropouts should be the first attempt

Comment: Yes, C is right here. Regularization and Dropouts

Comment: C is the answer

Discussion for Question 89

Link: https://www.examtopics.com/discussions/amazon/view/43864-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 12 votes

Discussion

Comment: A is correct. tSNE can do segmentation or grouping as well. Refer: https://towardsdatascience.com/an-introduction-to-t-sne-with-python-example-5a3a293108d1

Comment: A is definitely the correct answer. Pay attention to what the question is asking: "whether there are natural groupings for these columns across all customers and visualize the results as quickly as possible" The key point is to visualize the "groupings"(exactly what t-SNE scatter plot does, it visualize high-dimensional data points on 2D space). The question does not ask to visualize how many groups you would classify (K-Means Elbow Plot does not visualize the groupings, it is used to determine the optimal # of groups=K).

Comment: option A

Comment: B doesn't even answer the question: how are you going to see your customer groups in an elbow plot

Replies:

> **Comment:** Elbow plot helps you identify the correct number of clusters during K-Means clustering. The clustering happens basis of all the features and thus group employees. This is to help your understanding. And the correct answer however is still tSNE becuase the question focuses on identifying relationships/similarities between the features / columns in the dataset. The correct answer is A

Comment: Euclidean Distance suffers for high dimensional data. tSNE can suffers as well, but from my perspective is the correct one.

Comment: Elbow plot will not help visualize groups, only try to predict an optimal number of clusters. I think A is a better choice here

Comment: A. The t-SNE algorithm is a popular tool for visualizing high-dimensional datasets, as it can transform high-dimensional data into a 2D scatter plot, which makes it easier to visualize and understand the relationships between data points. The scatter plot produced by t-SNE can be interpreted as a map that reveals the structure of the data, showing whether there are natural groupings or clusters within the data. Option A is the quickest and simplest way to visualize the data in a meaningful way, allowing the Specialist to gain insights into the data more efficiently.

Comment: A is correct

Comment: 12-sep exam

Comment: A as k-means elbow is erroneous. It does not helping here. Scatter plot and t-sne is the right answer

Comment: An elbow plot (B) will not give you what the question is asking for. A scatter plot will, and t-SNE is first for visualizing before dimensionality reduction.

Comment: A is correct as k means suffer from curse of dimensionality and t-she will be a better option.

Comment: The B,C,D plots are meaningless wrt the problem —> A

Comment: t-SNE suffers curse of dimensionality and is indicated for small datasets

Comment: Additionally the numeric features don't require "embedding". I think they meant to write "standardize"

Comment: Rooting for A

Comment: B & D are wrong--because data contains "thousands of columns" and using k-means with euclidean suffers from "curse of dimensionality" Thus leaving A & C, you CANNOT viz clusters/groups/segments in a line graph so correct answer is A

Discussion for Question 90

Link: https://www.examtopics.com/discussions/amazon/view/43866-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 10 votes

Discussion

Comment: Answer is C. https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-plan-instances-guidelines.html

Replies:

> **Comment:** It's definitely C. The fact that this site indicates A is a clear sign that answers are just randomly selected, it would make zero sense to spot-instance the master node for an EMR cluster. Make sure you look at discussions for all of these questions.

Comment: C is the correct answer. "Long-Running Clusters and Data Warehouses If you are running a persistent Amazon EMR cluster that has a predictable variation in computational capacity, such as a data warehouse, you can handle peak demand at lower cost with Spot Instances. You can launch your master and core instance groups as On-Demand Instances to handle the normal capacity and launch task instance groups as Spot Instances to handle your peak load requirements."

Comment: According to :https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-plan-instances-guidelines.html The task nodes process data but do not hold persistent data in HDFS. If they terminate because the Spot price has risen above your maximum Spot price, no data is lost and the effect on your cluster is minimal. When you launch one or more task instance groups as Spot Instances, Amazon EMR provisions as many task nodes as it can, using your maximum Spot price. This means that if you request a task instance group with six nodes, and only five Spot Instances are available at or below your maximum Spot price, Amazon EMR launches the instance group with five nodes, adding the sixth later if possible.

Comment: The correct answer is C

Comment: I don't get why the wrong answer are still not updated after more than 1 year of everyone showing docs proving answer C..

Replies:

> **Comment:** 1 and a half year and still wrong.. Incredible!

Comment: Long-running clusters and data warehouses If you are running a persistent Amazon EMR cluster that has a predictable variation in computational capacity, such as a data warehouse, you can handle peak demand at lower cost with Spot Instances. You can launch your primary and core instance groups as On-Demand Instances to handle the normal capacity and launch the task instance group as Spot Instances to handle your peak load requirements.

Comment: Only task nodes can be deleted without losing data.

Comment: C, If you want to cut cost on an EMR cluster in the most efficient way, use spot instances on the task nodes because it, task nodes do not store data so no risk of data loss

Comment: For Long running jobs, you do not want to compromise the Master node(sudden termination) or the core nodes (HDFS data loss). Spot Instances on 20 task nodes are enough cost savings without compromising the job. Hence, C

Comment: If your primary concern is the cost, then you can run the master node on spot instances.

Replies:

> **Comment:** Adding the related reference from the AWS documentation: Master node on a Spot Instance The master node controls and directs the cluster. When it terminates, the cluster ends, so you should only launch the master node as a Spot Instance if you are running a cluster where sudden termination is acceptable. This might be the case if you are testing a new application, have a cluster that periodically persists data to an external store such as Amazon S3, or are running a cluster where cost is more important than ensuring the cluster's completion.
>
> #### Replies:
>
> > **Comment:** In the question , there are no specific conditions mentioned except the concern with the COST, thus I think the answer should be A.

Comment: Answer: C. https://aws.amazon.com/getting-started/hands-on/optimize-amazon-emr-clusters-with-ec2-spot/ Amazon recommends using On-Demand instances for Master and Core nodes unless you are launching highly ephemeral workloads.

Comment: Answer should be C.

Comment: Only master node is incorrect. Either use all on spot or only task or core on spot. As per: https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-plan-instances-guidelines.html Better to use only task node on spot for long running tasks/jobs

Replies:

Comment: Answer is C you should only run core nodes on Spot Instances /*when partial HDFS data loss is tolerable*/ Question is what "Should" be launched as spot instance

---

Discussion for Question 91

Link: https://www.examtopics.com/discussions/amazon/view/43867-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 11 votes

Discussion

Comment: This is a supervised problem and needs labels. Can't use clustering to find when faults can happen. CNN is for images not for timeseries data here. Hence, A seems appropriate.

Replies:

Comment: Agree, the answer is A

Comment: AGREE WITH YOU

Comment: A. YES - RNN good for time series as we want to use previous input B. NO - we know the class (fault) ahead of time, it is supervised C. NO - CNN is for images D. NO - seq2seq is for word generation

Comment: Answer is A

Comment: A recurrent neural network (RNN) is a more suitable choice than a convolutional neural network (CNN) because the data collected from the engines is a sequence of values over time, and the goal is to predict a future event (an engine fault). RNNs are designed to handle sequential data and can learn patterns and dependencies over time, making them well-suited for time-series data like this. On the other hand, CNNs are designed for image processing and are not ideal for sequential data.

Comment: Answer should be A

Comment: It can only be A. Agree with the comments before

Comment: Obviously A

Comment: A - obviously.

Comment: Seq2Seq also uses RNN under the hood, BUT option D. did not mention anything about "adding labels"--which is required here--hence --> A

Comment: A is correct. CNN is for images and RNN is for timeseries.

Comment: AAAAAAAAAAA https://towardsdatascience.com/how-to-implement-machine-learning-for-predictive-maintenance-4633cdbe4860

Comment: I think A is correct

Comment: It is A

Discussion for Question 92

Link: https://www.examtopics.com/discussions/amazon/view/43930-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- D: 14 votes

Discussion

Comment: D should be the more comprehensive answer. If it's not correlated, you can't make use of it in a linear regression A lot of others say B, but low variance can also be due to the nature/typical magnitudes of the variable itself

Replies:

Comment: Correlation indicates only linear relation, but, there might be non linear as well. To exploit it in the Linear Regression, you can take the variables to some power or run some non linear preprocessing on it, and you don't have to change the algorithm for it. So, answer B seem much more solid for me.

Comment: I think the problem with B is that what is considered "low variance"? The features are on different scales.

Comment: Answer B. Is not the best solucion prior can use other analysis. https://community.dataquest.io/t/feature-selection-features-with-low-variance/2418 If the variance is low or close to zero, then a feature is approximately constant and will not improve the performance of the model. In that case, it should be removed. Or if only a handful of observations differ from a constant value, the variance will also be very low.

Replies:

Comment: Low variance does not mean the feature is not important, right? If variance of target true value is also small and the correlation between above feature and target, the feature can be important feature.

Replies:

Comment: it does. If feature and target are correlated and you expect the target to change, the feature must have some sort of variance. Otherwise it means feature is almost constant so does target.

Comment: D is the best answer as it is mentioned multivariable linear regression applied where correlation is strong between dependent and independent variables.

Comment: D: We should remove features that are strongly correlated with each other and weakly correlated with the target: https://androidkt.com/find-correlation-between-features-and-target-using-the-correlation-matrix/ You can evaluate the relationship between each feature and target using a correlation and selecting those features that have the strongest relationship with the target variable.

Comment: I think D is the correct answer. If I remember correctly, Benjamini-Hochberg Method is essentially answer D if you consider the Hypothesis to be: the feature is powerfully influential to the target. My problem with B is that the variance can be easily affected by the scale. In the question, the number of bedroom's variance is very low, while the sqrt of the house has a high variance, both of these could be very useful. Furthermore, zip codes are included, and it is safe to assume the variance of zip codes can be high, but the information is very limited, especially if you use them as numerical instead of categorical features.

Comment: B is correct but the answer in D is better.

Comment: D is preferred over C because the goal is to predict the sale price of houses, which is the target variable. By checking the correlation of each feature against the target variable, the machine learning specialist can identify which features are most relevant to the prediction of the sale price and which are less relevant. Removing features with low correlation to the target variable helps reduce the complexity of the model and potentially improve its accuracy. On the other hand, a heatmap showing the correlation of the dataset against itself (C) doesn't directly address the relevance of the features to the target variable, and so it's not as effective in reducing the complexity of the model.

Comment: Answer should be D, THIS is feature elimination /selection during feature Enginering. Choice c is so close just to confuse test takers to pick the wrong choice! See below C and D answers -- C should have been correct if the question asked about how to visualize correlation among independent variables! PROVIDED second sentence in C needs to be removed or to say which feature you will eliminate in such case then the one with low correlation against target out of those two. C. Build a heatmap showing the correlation of the dataset against itself. Remove features with low mutual correlation scores. D. Run a correlation check of all features against the target variable. Remove features with low target variable correlation scores.

Comment: The multiple regression model is based on the following assumptions: There is a linear relationship between the dependent variables and the independent variables The independent variables are not too highly correlated with each other yi observations are selected independently and randomly from the population Residuals should be normally distributed with a mean of 0 and variance σ

Comment: I think the answer is D. If the model is a decision tree or something like that, I don't think it is possible to make a decision based only on the direct correlation with the target variable. But in multiple linear regression, the only thing that matters is the relationship between the target variable and the feature variable. B, if the standard deviation is small but not zero, then we have information.

Comment: B is correct.

Comment: To eliminate extraneous information. So, the answer is D.

Comment: Correct answer is D. The reason B is wrong because it is difficult to reason out why would you plot a histogram? Absolutely unnecessary step and distraction choice.

Comment: D is not the proper answer. Here is why: It says that it is comparing with the target variable (dependent variable), which implies it is comparing the correlation between the dependent and independent variables. This type of comparison is usually done after a model is constructed in order to prevent assessing the predictive strength of the model. To compare the target label, the label you wish to predict, with the other variables before - is premature and will likely result in weakening your model. Variables with low variance has very less information and the inclusion of which will likely weaken the model performance. Hence, B.

Comment: Answer is D. https://deep-r.medium.com/difference-between-variance-co-variance-and-correlation-ea0b7ddbaa1

Comment: Answer C. Heatmaps is used to visualize for correlation matrix https://towardsdatascience.com/better-heatmaps-and-correlation-matrix-plots-in-python-41445d0f2bec

Replies:

Comment: but is mentioned, "Remove features with low mutual correlation scores." which is wrong you should drop features with high correlation scores. so Answer is D

Comment: The problem with correlation tasks is it capture linear relations only. So, I would go with B

Discussion for Question 93

Link: https://www.examtopics.com/discussions/amazon/view/43870-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 12 votes

Discussion

Comment: Due to straight angles, I would choose Decision tree. See https://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html#sphx-glr-auto-examples-classification-plot-classifier-comparison-py

Replies:

Comment: From your link it is obvious that the best answer is still SVM with RBF kernel. In your link the SVM-RBF got 88% accuracy on the 'square-like' dataset whereas the Decision tree achieved only 80%. Answer is SVM with RBF kernel

Replies:

Comment: note the data from sklearn link is shaped as a ball of mass not a square. the RBF kernel would be better but the question shows a square. Decision tree should be better fit for this problem.

Comment: B - Decision tree - is not the best answer. If you use decision tree to do clustering, every time you need to partition the space into 2 parts. Hence you will split the space into 3*3. The red points in the center box and the black points will fall into the 8 boxes around it. The black points will be identified as 8 different classes. C is the correct answer. SVM with non-linear kernel is appropriate for non-linear clustering. Even if the shape is close to rectangular. SVM with non-linear kernel will be ale to approximate the rectangular boundary shape.

Replies:

Comment: Your statement "The black points will be identified as 8 different classes" does not make a lot of sense because the leaf node in a tree with be 1 of 2 classes, not 8 different classes just because they are visually in one place or the other

Comment: The tree works like this with this branch with 4 nodes: Age > 49? Y Age > 51? N Transaction > 28? Y Transaction > 31? N Positive Correct answer is B.

Comment: As the positive cases can be interpreted and separated from non positive ones by decision tree easily. SVM would have made sense if the two classes were inseparable or had complex relationship in data.

Comment: It is C SVM with RBF Kernel can classify this image. For decision tree, it will be more difficult

Comment: From the visual information provided, an SVM with an RBF kernel (Option C) would likely be the best choice because it can handle the circular class distribution. The RBF kernel is especially good at dealing with such scenarios where the boundary between classes is not linear.

Comment: Answer C B. Decision Tree: Decision trees can capture non-linear patterns and are capable of splitting the feature space in complex ways. They can be very effective if the decision boundary is not linear, but they might also overfit if the decision boundary is too complex. C. SVM with RBF Kernel: An SVM with a radial basis function (RBF) kernel is designed to handle non-linear boundaries by mapping input features into higher-dimensional spaces where the classes are more likely to be separated by a hyperplane. Given the clustered nature of the classes in the image, an SVM with an RBF kernel would likely be able to separate the classes with a higher degree of accuracy.

Comment: SVM-RBF is the correct solution

Comment: Support vector machine (SVM) with a radial basis function kernel would likely have the highest accuracy for this task because it can handle the non-linear separation required by the data.

Comment: I will lean with C

Comment: Answer is B as Decision tree can attain 100% accuracy in this case.

Comment: SVM with RBF and proper C and Gamma value can accomodate this square shape (https://vitalflux.com/svm-rbf-kernel-parameters-code-sample/)

Comment: confusing between SVN or devision tree. learning towards C

Comment: Answer C In general, SVMs are a good choice for tasks where accuracy is critical, such as fraud detection and medical diagnosis. Decision trees are a good choice for tasks where interpretability is important, such as customer segmentation and product recommendation.

Comment: A Decision tree produces a stepwise boundary that consists of rectilinear splits in the feature space that are perpendicular to the axes. The boundary is determined by the conditions specified in the tree. Support Vector Machines (SVM) produce a non-linear boundary in the feature space that separates the classes. The boundary is defined as the maximum margin hyperplane, which is the line that maximally separates the classes while having the greatest margin between the classes. The boundary can be linear, polynomial, radial basis function (RBF) or other types of non-linear functions, depending on the choice of kernel and the configuration of the SVM.

Comment: Correct Answer is C

Comment: SVM with RBF kernel is the best answer here. KNN is also a solution, but it is not one of the options

Comment: This answer is decision tree due to the Square decision boundaries.

Discussion for Question 94

Link: https://www.examtopics.com/discussions/amazon/view/43931-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- BDF: 8 votes

Discussion

Comment: when looking at an overfitting issue : https://www.kdnuggets.com/2019/12/5-techniques-prevent-overfitting-neural-networks.html 1. Simplifying The Model (reduce number of layers) 2. Early Stopping 3. Use Data Augmentation 4. Use Regularization (L1 + L2) 5. Use Dropouts So looking at the options: B, D, F

Comment: BDF !!!

Comment: looking at last 100 questions many answers were wrong , thanks to the discussion forum to provide correct answer

Comment: I would say BCD or BDF

Comment: Agree with BDF

Comment: Over fitting problem. All the options B, D, F reduce over fitting.

Comment: In what world is ACE the answer ?

Comment: BDF is the answer

Comment: BDF is the correct

Comment: ADE is absolutely wrong. 50 layers is already overfitting the model. We cannot increase the number of layers again.

Comment: BDF is the correct answer

Comment: should be BDF

Comment: One of the correct answer is showing as A. I wanted to understand how A(Choose Higher Number of Layers) is the correct Answer ?

Comment: I believe the answer is BCE because the model is overfitting.

Replies:

Comment: C might not be, because the model yielded 99% accuracy on the training set

Comment: choose smaller learning rate c, d, f,

Replies:

Comment: ignore answer is correct BDF

Comment: BDF!!!

Comment: It is supposed to be BDF

Discussion for Question 95

Link: https://www.examtopics.com/discussions/amazon/view/43932-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 5 votes

Discussion

Comment: Answer A.

Comment: The answer is Early Stopping. Stopp the training before accuracy start do decrease.

Replies:

Comment: Appreciates your explanation. Cheers

Comment: I will go with A Early stopping is a powerful technique to prevent overfitting. It involves monitoring the model's performance on a validation dataset during training. If the validation loss starts increasing or plateaus, early stopping stops further training. This ensures that the model doesn't overfit to the training data. Based on the graph, if the validation loss begins to stagnate or increase after a certain number of epochs, enabling early stopping could lead to better generalization.

Comment: early stopping before error increase

Comment: Early stopping

Comment: Early stopping

Comment: A: stop the training process of a neural network before it reaches the maximum number of epochs or iterations; in this case stop close to 64 Epochs.

Comment: Early stopping and not increasing epochs.

Comment: Answer is "A"

Comment: Early Stopping can improve the model?

Comment: A is the answer

Comment: I would go for A

Comment: I would choose A.

Discussion for Question 96

Link: https://www.examtopics.com/discussions/amazon/view/43948-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 6 votes

Discussion

Comment: I would choose A. See: https://www.itl.nist.gov/div898/handbook/pmd/section4/pmd442.htm and https://blog.minitab.com/blog/the-statistics-game/checking-the-assumption-of-constant-variance-in-regression-analyses

Comment: Ans. is A. High-degree polynomial transformation.

Replies:

Comment: I think so too.

Comment: Answer A. One of the key assumptions of linear regression is that the residuals have constant variance at every level of the predictor variable(s). If this assumption is not met, the residuals are said to suffer from heteroscedasticity. When this occurs, the estimates for the model coefficients become unreliable https://www.statology.org/constant-variance-assumption/

Comment: Agree with A

Comment: https://blog.minitab.com/en/the-statistics-game/checking-the-assumption-of-constant-variance-in-regression-analyses

Comment: Kind of like heteroskedasticity, anyways it is A.

Comment: Answer is A. It does not has constant variance !

Comment: A is correct answer

Comment: These images are broken. I cannot review the question properly!

Comment: D is best answer all x values are scattering as a whole , no matter what x is https://www.statisticshowto.com/residual-plot/ if you take all x values to plot histogram , it will be bel-curv.

Comment: And it does NOT mean linear regression is not appropriate. It means your linear regression model is biased due to several reasons.

Replies:

Comment: yes, it does. One of the main assumptions is homoscedasticity.

Comment: 100% A

Comment: I will choose A , because the data is heteroscedastic. It violates a key assumption of linear regeression

Comment: A. https://www.originlab.com/doc/origin-help/residual-plot-analysis

Comment: Do not have content variance https://stats.stackexchange.com/questions/52089/what-does-having-constant-variance-in-a-linear-regression-model-mean

Comment: Answer is A. As x raises, the residuals become higher and higher...

Comment: Some Good Reading https://www.andrew.cmu.edu/user/achoulde/94842/homework/regression_diagnostics.html Ans is A

Replies:

Comment: Thank you for sharing.

Discussion for Question 97

Link: https://www.examtopics.com/discussions/amazon/view/43962-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- CDF: 27 votes
- CEF: 20 votes

Discussion

Comment: C - voice and text interface E - understanding F - Speech to text

Replies:

Comment: Why would I need to transcribe while I have Lex that do the NLU part? It would be more reasonable to select Either Connect (B) or Polly (D) if the specs to generate output speech.

Comment: E - is more to express the "feeling" or "mood". We would rather need something, that can speak to the customer. So my suggestion is c,d,f

Replies:

Comment: The question states that the company wants to "provide executives with an enhanced experience so they can use natural language to get data from the reports." The key phrase here is "use natural language," which implies that the executives will be interacting with the system using human-like language, either written or spoken. To understand and interpret natural language inputs from users, whether written or spoken, the system needs to have natural language understanding (NLU) or natural language processing (NLP) capabilities. Without NLU/NLP capabilities, the system would not be able to make sense of the executives' natural language queries and extract the relevant information to retrieve data from the reports and dashboards. Services like Amazon Lex and Amazon Comprehend are specifically designed to provide NLU and NLP functionalities, respectively. Amazon Lex uses NLU models to understand the intent and extract relevant information from user inputs, while Amazon Comprehend provides NLP capabilities to analyze and extract insights from text data.

Comment: If we need to build written and spoken interfaces we need : F - Transcribe (speech to text) D- Polly (text ot speech) And for chatbot: E - Lex

Replies:

Comment: *C - Lex So C,D,F

Replies:

Comment: I second that, the keyword here is "conversational interface". so, no conversation without Amazon Lex

Comment: Alexa for Business: Handles the voice interaction, converting spoken queries into text and providing the voice interface that executives use to interact with the BI application. Amazon Lex: Processes the text input (converted by Alexa) and understands the intent behind the queries, enabling the conversational interface. Amazon Polly: Optional but useful if you want to convert the textual responses from the BI application back into spoken responses, providing a complete voice-based interaction.

Comment: Lex for bot service, Polly for text-to-speech (answer) and Transcribe for speech-to-text (question).

Comment: I believe Answer should be CDF C: Lex D: Polly F: Transcribe

Comment: For a BI application where executives can ask questions using written and spoken interfaces, the following combination of services would be suitable: Amazon Lex (Option C): To build the core conversational interface that understands and processes natural language queries. Amazon Polly (Option D): To provide spoken responses to written queries, giving a more interactive experience for users who are not using the voice interface. Amazon Transcribe (Option F): To convert spoken queries into text that can be understood by Amazon Lex. These three services would work together to provide a comprehensive conversational interface that allows for both text and voice interactions, meeting the requirements of the scenario provided.

Comment: C. Amazon Lex: It provides advanced deep learning functionalities of automatic speech recognition (ASR) for converting speech to text, and natural language understanding (NLU) to recognize the intent of the text, enabling you to build applications with highly engaging user experiences and lifelike conversational interactions. D. Amazon Polly: This service turns text into lifelike speech using deep learning. It would enable the BI application to deliver the answers to the executives' questions in a spoken format. F. Amazon Transcribe: This is an automatic speech recognition (ASR) service that makes it easy for developers to add speech-to-text capability to their applications. This would be necessary for the BI application to interpret spoken questions from the executives.

Comment: CDF -> CEF. you dont need comprehend in this scenario.

Comment: Amazon Lex (C): This service is crucial for building conversational interfaces. It provides the capabilities to understand and interpret user input in natural language, which is essential for understanding the questions asked by executives. Amazon Transcribe (F): For a spoken interface, you need a service that can convert speech into text. Amazon Transcribe does exactly this, allowing the system to process spoken questions by converting them into text that can then be interpreted by Amazon Lex. Amazon Polly (D): To enhance the user experience by responding to inquiries not only in text but also in spoken form, Amazon Polly is ideal. It converts text responses into lifelike speech, allowing the system to verbally communicate with the executives. Together, these three services (Amazon Lex, Amazon Transcribe, and Amazon Polly) will enable a comprehensive conversational interface for the BI application, catering to both written and spoken queries and responses

Comment: why does aws use muliplt service for tts and stt?

Comment: No, don't need E Comprehend because the report has already been generated.

Comment: Answer is CEF --> Input can be speech but the output to the user will be text (as nothing specific is mentioned) using Lex for conversational interface, Transcribe to convert speech to text (if input is speech) and Comprehend for insights from text

Comment: CEF is correct

Comment: I will go with: lex for the chat interface comprehend for getting insights from reports Polly for text-to-speech transformation https://aws.amazon.com/blogs/machine-learning/deriving-conversational-insights-from-invoices-with-amazon-textract-amazon-comprehend-and-amazon-lex/

Comment: Amazon Polly is essential for providing spoken responses in a conversational interface, it doesn't directly handle the natural language understanding and processing aspect, which is why it wasn't included as one of the top

three services for building the conversational interface in this scenario. Correct is C, E, F

Comment: A. NO - Alexa for Business B. NO - Amazon Connect for call centers C. YES - Amazon Lex for chatbots D. YES - Lex Text-to-Speech E. NO - Amazon Comprehend is for topic extraction and sentiment analysis, Transcribe already does it F. YES - Transcribe Speech-to-Text

Replies:

Comment: Transcribe does not do sentiment analysis and topic extraction it just generates transcript from speech so we need Amazon Comprehend

Comment: Agree with CDF

Discussion for Question 98

Link: https://www.examtopics.com/discussions/amazon/view/73977-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 16 votes

Discussion

Comment: the answer is B - the model is underfitting = high bias, so we want to reduce it C is wrong because the intention is not to increase variance which equals overfitting (using a more complex model would be good, but to reduce bias not increase variance)

Comment: I would think ImageNet network is good enough already, so more data

Comment: A. NO - HPO has already been done though grid search B. YES - 150 images is very small; need x10 that C. NO - need bigger training set D. NO - what would the new model be ?

Comment: More data to training set

Comment: Letter B is the correct one. We can add more data with data augmentation. Letter A would be a repetition of what has already been done. Letter C is impractical. Letter D is starting from scratch without need.

Comment: I think it should be D: "Train a new model using the current neural network architecture". Because apples data is very specific and ImageNet weights will be to generic there. We still can leave ImageNet weights for an initial configuration but the model should be retrained from scratch.

Comment: 450 images should be fine. HPO for me.

Comment: bOTH VALIdation set and train set performing equally but performance not good. So the basic problem here is high bias (train error) and high variance (test error). Ideally we want both low, but there is trade-off need to be cautious to avoid overfitting. So this problem needs solution for Low bias first (so training performance improves with decent) for later to figure out whether that leads to overfit or not when you test it,! Answer choice B

Comment: why not A? https://aws.amazon.com/about-aws/whats-new/2022/07/amazon-sagemaker-automatic-model-tuning-supports-increased-limits-improve-accuracy-models/

Comment: not B, c is corect

Comment: Given that the model can't even fit the training set properly, it would be convenient to amplify the layers that are trained. If I understood the phrasing correctly, I would go with C.

Comment: C, accuracy on training set is low, model not complex enough

Replies:

Comment: It only has 150 photos for training, more complex neural network won't help

Comment: B is more accurate, while adding more complexity for model is viable but you don't want to increase variance

---

Discussion for Question 99

Link: https://www.examtopics.com/discussions/amazon/view/73978-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- D: 21 votes

Discussion

Comment: Data Augumentation is the way to go here. How does converting to grayscale help? What if the colors of the items are relevant in object identification???

Comment: data augemntation

Comment: D is correct How can I make the decision to use gray images if the question doesn't even indicate whether the images are colored or not? and even so, colored images are important to ensure more accuracy in training than compared to gray imagens. Due that the model is underfitting, more data like indicated the option D is the correct action.

Comment: C: "Attach different colored labels to each item, take the images again, and build the model" It is also kind of augmentation. It is even better than just inverting and translating existing samples.

Replies:

Comment: But it's done in real life and your manual work would be lost.

Comment: shouldnt it be reduced to 2 variables , taking image of empty shelf and non empty and that should do it ?

Comment: D is of course the right answer, grayscale only won't help anything

Comment: D is the CORRECT ANSWER https://research.aimultiple.com/data-augmentation/

Comment: Data augmentation is correct. we need more samples

Comment: D is correct

Comment: D is my answer for this. A can help but it'll need more than that.

Comment: D, i guess

Discussion for Question 100

Link: https://www.examtopics.com/discussions/amazon/view/45178-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: B - stratified k-fold cross-validation will enforce the class distribution in each split of the data to match the distribution in the complete training dataset.

Comment: B is the correct answer. Use Stratified k-Fold Cross-Validation for Imbalanced Classification. Stratified train/test splits is an option too. But the question is specifically asking "cross-validation" strategy.

Comment: for imbalanced data. Stratified k-fold cross-validation ensures that the distribution of the target variable is the same in each fold. This is important for binary classification problems, where the target variable is imbalanced. In this case, the disease is seen in only 3% of the population. This means that if we do not use stratified k-fold cross-validation, then there is a risk that the training and validation sets will not be representative of the actual population.

Comment: B https://towardsdatascience.com/understanding-8-types-of-cross-validation-80c935a4976d

Comment: Stratified cross validation is for unbalanced data like this!

Comment: Why K=5?

Replies:

Comment: K=5 is just standard

Comment: Yes, B...

Discussion for Question 101

Link: https://www.examtopics.com/discussions/amazon/view/44886-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 12 votes

Discussion

Comment: Answer is A https://aws.amazon.com/blogs/compute/gpu-workloads-on-aws-batch/

Replies:

Comment: Makes most sense

Comment: I would go for D. As far as I know Fargate does not support GPU computing.

Replies:

Comment: It does support GPU https://docs.aws.amazon.com/batch/latest/userguide/fargate.html

Replies:

Comment: this is wrong information. It does not support GPU

Comment: the problem that fargate is serverless which mean you can't control its compute capabilities

Comment: fargate doesnt support GPU. https://github.com/aws/containers-roadmap/issues/88

Comment: AWS batch will easily satisfy the requiremtns

Comment: Fargate doesn't support GPU. So go with AWS Batch and DLC (Deep Learning Container)

Comment: A. NO - Fargate provides batch functionnalities already fully integrated with ECS B. NO - too low level C. YES - AWS Deep Learning Containers are optimized; AWS Fargate is serverless (so less ops complexity); Spot best for cost D. NO - ECS service scheduler is not serverless

Replies:

Comment: Answer is A. C is not correct. GPU resources aren't supported for jobs that run on Fargate resources.

Comment: A and C are both great answers but when it comes to cost I believe A is the more cost effective solution. So A is my answer

Comment: Automate the workload by scheduling the job to run once a week using AWS Batch's built-in scheduler or a cron expression. Optimize the performance by using AWS Deep Learning Containers that are tailored for GPU acceleration and deep learning frameworks. Reduce the cost by using Spot Instances that offer significant savings compared to On-Demand Instances. Handle failures by using AWS Batch's retry strategies that can automatically restart the job on a different instance if the Spot Instance is interrupted.

Comment: Answer is A. (But the question is tricky) A and D are both correct solutions but pay attention to the words - "Senior managers are concerned about the complexity of the solution's resource management and the costs". With Cost is everything simple - use Spot instances, with resource management - use higher abstraction servicr AWS Batch is a management/abstraction layer on top of ECS and EC2 (and some other AWS resources). It does some things for you, like cost optimization, that can be difficult to do yourself. Think of it like Elastic Beanstalk for batch operations. It provides a management layer on top of lower-level AWS resources, but if you are comfortable managing those lower level resources yourself and want more control over them it is certainly an option to use those lower-level resources directly.

Comment: Why not C? AWS Fargate is oriented to manage resources as you need.

Comment: A looks good to me https://aws.amazon.com/blogs/compute/deep-learning-on-aws-batch/

Comment: b: for those who think its B because of spot instance interruption, ready question phrase "The job can be paused, restarted, and continued at any time in the event of a failure, and is run from a central queue." between a and c: at the time of this question i doubt if fargate supported GPU, even if it did I choose aws batch for job and fargate for services/apps that need to run all the time. a is answer

Comment: Option A is the most cost-effective architecture as it uses GPU-compatible Spot Instance which is the lowest cost compute option for GPU instances in the AWS cloud. AWS Batch is a fully managed service that schedules, runs, and manages the processing and analysis of batch workloads. The use of AWS Deep Learning Containers enables the technology startup to use pre-built, optimized Docker containers for deep learning, which reduces the complexity of the solution's resource management and eliminates the need for repeated processing.

Comment: Answer is A. Option B is similar to A, but it uses a low-cost GPU-compatible EC2 instance rather than a container, which may not be as flexible or scalable as using containers.

Comment: Answer is A https://aws.amazon.com/blogs/compute/gpu-workloads-on-aws-batch/

Comment: https://aws.amazon.com/blogs/compute/gpu-workloads-on-aws-batch/ There you have it

Comment: spot instance can be out of service any time, which makes it unsuitable for workflow on a regular basis. Anwer is B

Discussion for Question 102

Link: https://www.examtopics.com/discussions/amazon/view/44105-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 11 votes

Discussion

Comment: B seems correct based on the elbow method

Replies:

Comment: I agree, most likely B.

Comment: Hi all, I am not able to see the image. It is broken for me. Is it possible for someone to share the image.

Comment: The closest no.to the elbow is clearly 4

Comment: B is correct

Comment: Elbow method

Comment: B: https://en.wikipedia.org/wiki/Elbow_method_(clustering)

Comment: Elbow is more visible at 4

Comment: B seems correct

Comment: https://www.analyticsvidhya.com/blog/2021/01/in-depth-intuition-of-k-means-clustering-algorithm-in-machine-learning/

Comment: Elbow method

Comment: . B seems correct.

Comment: B seems better

Comment: number 4 is the elbow of the hand, B is correct

Comment: because the elbow method is a heuristic method its open to debate as to where the correct bend in the cluster is. It's a good tool to use with lower computation cost than computing the silhouette score. When looking at https://www.youtube.com/watch?v=qs8nfzUsW5U instead of eyeballing where the bend is, he calculates where the difference between scores is smaller than the 90th persentile

---

Discussion for Question 103

Link: https://www.examtopics.com/discussions/amazon/view/44106-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 5 votes

Discussion

Comment: A. Fastest route must Amazon Services.

Replies:

Comment: Amazon Mechanical Turk is an Amazon service

Comment: A. YES - Rekognition with built-in labels for images & video, Transcribe to convert sound to text and Comprehend for Topic Modeling B. NO - complicated C. NO - complicated D. NO - complicated

Comment: AWS services for fastest route

Comment: why not C? C. Use Amazon Transcribe to convert speech to text. Use the Amazon SageMaker Neural Topic Model (NTM) and Object Detection algorithms to tag data into distinct categories/classes.

Replies:

Comment: this takes time and u need to have atleast some technical ML expertise

Comment: I will choose B, https://aws.amazon.com/cn/getting-started/hands-on/machine-learning-tutorial-label-training-data/

Comment: Mechanical Turk is the most accurate, but the three services in letter A is the fastest!

Comment: A. Use Amazon Rekognition, Amazon Comprehend, and Amazon Transcribe to tag data into distinct categories/classes is the fastest route to index the assets. These AWS services provide pre-built machine learning models that can be used to tag the content in the archive without the need for building custom models from scratch. This option would be faster than using custom models with the AWS Deep Learning AMI and Amazon EC2 GPU instances, or using Amazon Mechanical Turk for human labeling. Additionally, the use of pre-built models reduces the need for machine learning expertise, aligning with the company's goal of accelerating efforts by its in-house researchers.

Comment: A. Option B is for those without ML experience. But "researchers who have limited machine learning expertise", so A is better.

Comment: A. The most straight forward use of services.

Comment: I would have said B, but in B it says "label footage" which means it ignored the rest of the data, so i'd go with A

Replies:

Comment: The question said "a very large archive" meaning a lot of money to pay for labour. B won't be as fast as machine, plus you only label the footage, ignored other stuff.

Comment: Would go for A

Comment: I will go with B

Comment: Correct answer is A

Comment: B. as no one in-house is an expert and It probably is the fastest way to get there

Replies:

Comment: Take into consideration that it is "a very large archive"

Discussion for Question 104

Link: https://www.examtopics.com/discussions/amazon/view/43964-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: Agreed, B it is. See https://medium.com/slalom-data-analytics/amazon-kinesis-data-streams-auto-scaling-the-number-of-shards-105dc967bed5 One shard can Ingest 1 MB/second or 1,000 records/second. So 100 KB * 100 = 10 MB (10 shards required)

Comment: 100 KB * 100 = 10 MB 1 MB/second 10 / 1 = 10 shards.

Comment: Each shard in Amazon Kinesis Data Streams can support up to 1,000 transactions per second. The data needs to be ingested at up to 100 transactions per second, so we need at least 1 shard. However, we also need to consider the size of the JSON data blob. Each JSON data blob is 100 KB in size, and each shard can only store up to 1 MB of data. This means that we need to have at least 10 shards, so that each shard can store 100 KB of data.

Comment: B - Max. ingestion per shard = 1000 KB/s --> 100 Records * 100 KB = 10.000 KB --> 10.000 KB / 1000 KB/per Shard = 10 Shards

Comment: 10 should be correct.

Comment: B is correct

Comment: 100 kb * 100 t/second = 10000 kb = 10 mb 10mb / max_threshold_per_shard (1 mb) = 10 shards

Comment: Reference: https://docs.aws.amazon.com/streams/latest/dev/service-sizes-and-limits.html

---

Discussion for Question 105

Link: https://www.examtopics.com/discussions/amazon/view/43965-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- D: 9 votes

Discussion

Comment: I would say D, because of correlations and dependencies between features. See https://towardsdatascience.com/basics-of-bayesian-network-79435e11ae7b and https://www.quora.com/Whats-the-difference-between-a-naive-Bayes-classifier-and-a-Bayesian-network?share=1

Replies:

Comment: I agree, makes moste sense

Comment: It should be D. Naive Bayes is called naive because it assumes that each input variable is independent. This is a strong assumption and unrealistic for real data; however, the technique is very effective on a large range of complex problems.

Comment: In this case, the absolute values of the Pearson correlation coefficients range between 0.1 to 0.95. This means that some of the features are statistically dependent. Therefore, a full Bayesian network is a better model for the underlying data than a naive Bayesian model.

Comment: In a full Bayesian network, features are connected to each other by edges that represent their conditional dependence relationships. A full Bayesian network is useful when the relationships between the features are complex, non-linear or when they are not conditionally independent. In this situation, where the Pearson correlation coefficients range between 0.1 and 0.95, it suggests that there are dependencies between the features, indicating that a full Bayesian network would be appropriate to capture the relationships between the features and model the data.

Comment: distinction between Bayes theorem and Naive Bayes is that Naive Bayes assumes conditional independence where Bayes theorem does not. This means the relationship between all input features are independent . The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

Comment: A naive Bayesian model, since some of the features, are statistically dependent.

Comment: D. Naive bayes - features are independent given the class.

Comment: I would say, B. Naive Bayes assumes conditional independence and not statistical

Replies:

Comment: you mean (a) naive bayes not (b)

Comment: This is also a good source of information to help build your understanding https://www.simplypsychology.org/correlation.html

---

Discussion for Question 106

Link: https://www.examtopics.com/discussions/amazon/view/44249-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 7 votes

Discussion

Comment: I would say B. Logarithmic transformation converts skewed distributions towards normal

Comment: I would go with B. For right skewed distributions -> Logrithmic transformation For left skewed distributions -> exponential transformations=

Comment: The linear regression model assumes that the errors are normally distributed. The plot of the feature shows that the errors are not normally distributed. The logarithmic transformation can be used to transform the errors to be normally distributed. The exponential transformation, polynomial transformation, and sinusoidal transformation cannot be used to transform the errors to be normally distributed.

Comment: B when the feature data is not normally distributed, applying a logarithmic transformation can help to normalize the data and satisfy the assumptions of the linear regression model.

Comment: 'A' would make it considerably worse.

Replies:

Comment: Exponential transformation would make it exponentially worse. :D

Comment: Log Normal Distribution => Log() => Normal Distribution

Comment: B is correct answer

Comment: This is B, as this feature seems skewed while others have a regular distribution according to the question. The log transformation will reduce this features skewness.

Comment: I think it's B. reference: https://corporatefinanceinstitute.com/resources/knowledge/other/positively-skewed-distribution/#:~:text=For%20positively%20skewed%20distributions%2C%20the,each%20value%20in%20the%20dataset. "For positively skewed distributions, the most popular transformation is the log transformation. The log transformation implies the calculations of the natural logarithm for each value in the dataset. The method reduces the skew of a distribution. Statistical tests are usually run only when the transformation of the data is complete."

Comment: I would also go for B, as Log transformation is often mentioned, when we are talking about right (positive) skewness.

Discussion for Question 107

Link: https://www.examtopics.com/discussions/amazon/view/44289-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: B lower max_depth is the correct answer. D min_child_weight means something like "stop trying to split once your sample size in a node goes below a given threshold" Lower min_child_weight, the tree becomes more deep and complex. Increase min_child_weight, the tree will have less branches and less complexity.

Comment: The max_depth parameter controls the maximum depth of the decision trees in the XGBoost model. A higher max_depth value will result in more complex decision trees, which can lead to overfitting.

Comment: Overfitting occurs when a model performs well on the training data but poorly on unseen or test data. In the context of XGBoost, reducing the max_depth parameter helps prevent overfitting. The max_depth parameter controls the maximum depth of the trees in the ensemble. A smaller max_depth value limits the complexity of the trees, making them less likely to memorize the noise in the training data and improve generalization to unseen data.

Comment: It is B

Comment: B: overfitting problem.

Comment: 12-Sep Exam.

Comment: When a model overfits, the solutions are: 1. Reduce model flexibility and complexity 2. Reduce the number of feature combinations 3. Decrease n-grams size 4. Decrease the number of numeric attribute bins 5. Increase the amount of regularization 6. Add dropout

Comment: B. 30-deep tree is crazy; normally it's 6-7 no more

Comment: A. Increase the max_depth parameter value. (This would increase the complexity resulting in overfitting) B. Lower the max_depth parameter value. (This would reduce the complexity and minimize overfitting) C. Update the objective to binary:logistic. it depends on what the target(s) generally you would have a binary classification for fraud detection but there is nothing to say you can't have a multi class so there is not enough information given. D. Lower the min_child_weight parameter value. (This would reduce the complexity and minimize overfitting) I find that there are 2 correct answers to this question which does not help B & D

Replies:

Comment: Ans : B , Lower values avoid over-fitting. No for D - Larger values avoid over-fitting.

Comment: Thus, those parameters can be used to control the complexity of the trees. It is important to tune them together in order to find a good trade-off between model bias and variance

Comment: min_child_weight is the minimum weight (or number of samples if all samples have a weight of 1) required in order to create a new node in the tree. A smaller min_child_weight allows the algorithm to create children that correspond to fewer samples, thus allowing for more complex trees, but again, more likely to overfit.

Comment: max_depth is the maximum number of nodes allowed from the root to the farthest leaf of a tree. Deeper trees can model more complex relationships by adding more nodes, but as we go deeper, splits become less relevant and are sometimes only due to noise, causing the model to overfit.

Discussion for Question 108

Link: https://www.examtopics.com/discussions/amazon/view/44530-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- D: 13 votes

Discussion

Comment: I think the answer is D - RCF works together with Data Analytics, and sliding window helped on new information

Replies:

Comment: better to say "RCF is a built-in algorithm/function in Kinesis Data Analytics"

Comment: D uses the built-in RCF algorithm, which is designed for anomaly detection on streaming data and can adapt to changing patterns over time.It does not require any training data or preprocessing steps, as the RCF algorithm can learn from the streaming data directly. It uses a sliding window, which allows for continuous updating of the anomaly scores based on the most recent data points. It leverages the Amazon Kinesis Data Analytics service, which provides a scalable and managed platform for running SQL queries on streaming data. Option A requires training an RCF model on historic data, which may not reflect the current web traffic patterns. It also adds complexity and latency by invoking a Lambda function for each record.

Comment: Answer D

Comment: Letra B está descartada, pois trás um modelo supervisionado de classificação para um problema não supervisionado. Letra C trás outro modelo que não é recomendado também, em comparação ao RCF. A solução mais fácil de implementar e que atinge os critérios pedidos é a Letra D. Letra A está errada, pois usamos KDS para ingestão apenas.

Comment: the data scientist needs to identify unusual web traffic patterns in real-time and adapt to changing web patterns over time. Amazon Kinesis Data Analytics provides real-time analytics capabilities on streaming data. The Amazon Random Cut Forest (RCF) SQL extension is designed for anomaly detection in streaming data, which fits the requirement to calculate an anomaly score for each web traffic entry.

Comment: Answer is D "The algorithm starts developing the machine learning model using current records in the stream when you start the application. The algorithm does not use older records in the stream for machine learning, nor does it use statistics from previous executions of the application." https://docs.aws.amazon.com/kinesisanalytics/latest/sqlref/sqlrf-random-cut-forest.html

Comment: D it is. easy one

Comment: RCF is dynamic and adapts with time. D seems more appropriate.

Comment: It is A. The only way to handle the historic data is using sagemaker and you can preprocess a data stream using a lambda.

Replies:

Comment: But, the question does not require using historical data. BTW, it only has unlabled historic data, and unlabled data is not really useful training a detection model.

Comment: Definitly D, Data Anaytics is using RCF, Using window for selecting data with SQL

Comment: One more reason to select D, not A, is there is no Lambda function to preprocess record in Kinesis Data Stream.

Replies:

Comment: That's not true: https://docs.aws.amazon.com/kinesisanalytics/latest/dev/lambda-preprocessing.html

Comment: " Adapt unusual event identification to changing web patterns over time." -> option A does not satisfy this, only mentions build the model once

Comment: The data scientist has access to unlabeled historic data to use, if needed. D has no mention of this. Also, A says the lambda function provides data enrichment. For me it's A.

Comment: A and D both seems to works. But A does not satisfy requirement 2, adapt to patterns over time. Since the model is only trained on old data. So D may be better.

Comment: It is definitely D

Discussion for Question 109

Link: https://www.examtopics.com/discussions/amazon/view/44290-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 9 votes

Discussion

Comment: I think it should be C as the final outcome among 200 categories is already know. No need to build a classification model. It's pure forecasting problem.

Replies:

Comment: he said for a few what about the unclassified many i think we need to make classification for the rest first as it will help us with forecasting later with month to month forecasting

Comment: C is my answer. No need to do classification. Because you know whether the insurance has a claim or not in the dataset. The claim contents do not provide additional information.

Comment: forcasting

Comment: It's pure forescasting problem.

Comment: I would say no machine learning model needed at all. Just using count group by categories SQL is enough

Comment: FinalOutcome 1 2 . 200 RecordID, FinalOutcome, Date, ClaimContents 1 2 . 100000 Note: claim content has partial information, only for few of 200 categories predict how many claims to expect in each category from month to month, a few months in advance We dont need the claim contents, we have all we need from first 3 columns to train a forecast model c

Comment: Forecasting using claim IDs and timestamps to identify how many claims in each category to expect from month to month. The problem requires the prediction of the number of claims in each category for each month, which is a time series forecasting problem. The timestamps and record IDs can be used to model the underlying patterns in the data, and the model can be trained to predict the number of claims in each category for future months based on these patterns. While the claim contents might provide additional information, the fact that partial information is only available for a few categories suggests that this information might not be enough to build a robust model, and that it might not be possible to apply supervised learning to all 200 categories. Instead, the model should be trained on the time series data (claim IDs and timestamps) for all categories, and the claim contents can be used to improve the accuracy of the model only for the categories for which such information is available.

Comment: how can a forecasting/classification model can be based on the claim ID? (that should be unique)

Comment: it's a forecasting problem, not a classification one

Comment: predict how many claims to expect in each category from month to month, a few months in advance C is the only one mentioning forecasting

Comment: D is correct. Multi-label classification to impute the missing claim contents, then forecasting what we want. C is missing the imputation part.

Replies:

> **Comment:** The question is, can we get something useful out of the handful of 200s and will this impact the forecast as we could forecast the numbers without...

Comment: It is true that the final outcome is known. But C does not use the partial information from the 200 categories. Reinforcement learning currently is state of the art in stock prediction and other time series. Why waste valuable information? For me it's B.

Comment: This is a supervised learning approach: Supervised learning problems can be further grouped into regression and classification problems. Classification: A classification problem is when the output variable is a category, such as “red” and “blue” or “disease” and “no disease.” Regression: A regression problem is when the output variable is a real value, such as “dollars” or “weight.”

Discussion for Question 110

Link: https://www.examtopics.com/discussions/amazon/view/44917-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 8 votes

Discussion

Comment: B is the correct answer. A/B testing with Amazon SageMaker is required in the Exam. In A/B testing, you test different variants of your models and compare how each variant performs. Amazon SageMaker enables you to test multiple models or model versions behind the `same endpoint` using `production variants`. Each production variant identifies a machine learning (ML) model and the resources deployed for hosting the model. To test multiple models by `distributing traffic` between them, specify the `percentage of the traffic` that gets routed to each model by specifying the `weight` for each `production variant` in the endpoint configuration.

Comment: I would answer B, it seems similar to this AWS example: https://docs.aws.amazon.com/sagemaker/latest/dg/model-ab-testing.html#model-testing-target-variant

Comment: Option B

Comment: This solution allows the Data Science team to build and host multiple models in Amazon SageMaker, which is a fully managed service for training, deploying, and managing machine learning models. The team can then create an endpoint configuration with multiple production variants, which are different versions of the models. By programmatically updating the endpoint configuration, the team can control the portion of inferences served by the different models. This allows them to evaluate the models against their business goals and measure their long-term effectiveness without having to make changes at the application layer.

Comment: Answer D as it is said “the team intends to run numerous versions in parallel for extended periods of time,” so batch transform

Replies:

> **Comment:** How can you create a single endpoint for batch transforms? this answer is nonsensical.
>
> ### Replies:
>
> > **Comment:** It is possible to create a single endpoint for AWS Batch transforms. Here are the key steps: Create an interface endpoint for AWS Batch in your VPC using the AWS CLI or console. The endpoint service name will be in the format of com.amazonaws. .batch . When creating the endpoint, assign an IAM role with necessary permissions to make calls to the Batch API. You can then submit batch transform jobs to AWS Batch referencing resources in both public and private subnets of the VPC. The endpoint ensures private connectivity to Batch. The single endpoint allows chaining multiple transforms together in a pipeline efficiently without needing internet access. New transforms can be added without redeploying the endpoint. AWS Batch will automatically provision the required compute environments like EC2 instances or containers to run the transforms and scale as needed based on job requirements.

Comment: it says,"host a sleep monitoring application", it is the host which means online, not batch, b is correct

Comment: The possibility to alter the percentage of inferences supplied by the models. Which method achieves these criteria with the LEAST amount of effort?

Comment: B. Easy

Comment: Think anser is D, below is from the Sagemaker doc. "https://docs.aws.amazon.com/sagemaker/latest/dg/batch-transform.html" Use Batch Transform to Test Production Variants To test different models or various hyperparameter settings, create a separate transform job for each new model variant and use a validation dataset. For each transform job, specify a unique model name and location in Amazon S3 for the output file. To analyze the results, use Inference Pipeline Logs and Metrics.

Replies:

> **Comment:** The question talks about the LEAST amount of effort. In this case, there will be as many transform jobs required to be built as there are variants. That may not be the least amount of effort.

Discussion for Question 111

Link: https://www.examtopics.com/discussions/amazon/view/44063-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: C is my answer. Pay attention that the question is asking for 2 things: 1. detect specific types of weeds 2. detect the location of each type within the field. Image Classification can only classify images. Object detection algorithm: 1.identifies all instances of objects within the image scene. 2.its location and scale in the image are indicated by a rectangular bounding box. Data format for Computer Vision algorithms in SageMaker: Recommend to use RecordIO.

Comment: Record IO preffered and also object detection due to several types of weeds

Comment: The goal is to detect specific types of weeds and their locations within a field, which is a task that requires object detection, rather than image classification. Object detection algorithms are designed to identify objects and their locations within an image, whereas image classification algorithms only categorize an entire image into various classes. Single-shot multibox detectors (SSD) are a type of object detection algorithm that are well-suited for real-time inferencing and have been shown to be effective for a variety of object detection tasks. By preparing the images in RecordIO format and using Amazon SageMaker, the company can easily train, test, and validate the model, making it easier to deploy the model in a scalable and secure environment.

Comment: C is the right answer. you need to detect location

Comment: C You can detect the type of weeds and the location within the field.

Comment: If they had an answer with "Faster R-CNN" then it would be different. This is a good article talking about SSD, Faster R-CNN, R-FCN and others which is a good read. https://jonathan-hui.medium.com/object-detection-speed-and-accuracy-comparison-faster-r-cnn-r-fcn-ssd-and-yolo-5425656ae359

Comment: I would go with answer C .. SSD are new architectures faster than the old CNN https://towardsdatascience.com/understanding-ssd-multibox-real-time-object-detection-in-deep-learning-495ef744fab

Comment: I would select answer A, situation is very similar to this one: https://aws.amazon.com/blogs/machine-learning/building-a-lawn-monitor-and-weed-detection-solution-with-aws-machine-learning-and-iot-services/

Replies:

Comment: I think it's better go with C, since the question also ask for the location of the weed on the field, while the example you posted is just a classifier.

Replies:

Comment: since field is divided in to 10x10 grid I felt A is more suitable.

Replies:

Comment: So you expect precisely one weed per grid (total 100) of an entire field? If a field is a hectare, then each grid would be 100m2

Comment: The llink says clearly only classification and not talking about object detection. Answer should be C

Discussion for Question 112

Link: https://www.examtopics.com/discussions/amazon/view/44064-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 12 votes

Discussion

Comment: I would select B, based on the following AWS examples: https://aws.amazon.com/blogs/iot/industrial-iot-from-condition-based-monitoring-to-predictive-quality-to-digitize-your-factory-with-aws-iot-services/ https://aws.amazon.com/blogs/iot/using-aws-iot-for-predictive-maintenance/

Comment: B is my answer. For latency-sensitive use cases and for use-cases that require analyzing large amounts of streaming data, it may not be possible to run ML inference in the cloud. Besides, cloud-connectivity may not be available all the time. For these use cases, you need to deploy the ML model close to the data source. SageMaker Neo + IoT GreenGrass To design and push something to edge: 1. design something to do the job, say TF model 2. compile it for the edge device using SageMaker Neo, say Nvidia Jetson 3. run it on the edge using IoT GreenGrass

Comment: without relying on internet connectivity.

Comment: The described solution will be solved by an edge solution as internet relability is low. IoT Greengrass is the best solution for the edge inference.

Comment: This is an edge solution, having as little traffic with AWS resources in regions. For this, start thinking IoT Greengrass and Sagemaker Neo, and you'll be halfway there. Answer is B, no doubt

Comment: B is the answer, obviously

Comment: This solution requires edge capabilities and to be able to run the inference models in near real-time. SageMaker Neo is a deployable unit on the edge architecture (IoT Greengrass) which can host the runtime inference model.

Comment: A: not a complete solution a lot of details is missed C: daily batch training is huge defect in this solution D: writing to dynamoDB and invoking endpoint make this solution slower than using an IoT Green Grass Answer: B

Comment: I would choose B because IoT reduce latency because they work on local machine

Comment: I would choose B

Discussion for Question 113

Link: https://www.examtopics.com/discussions/amazon/view/44065-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 8 votes

Discussion

Comment: I would select B. Based on the following AWS documentation it appears this is the right approach: https://sagemaker.readthedocs.io/en/stable/frameworks/tensorflow/using_tf.html https://github.com/aws-samples/amazon-sagemaker-script-mode/blob/master/tf-horovod-inference-pipeline/train.py

Comment: B is my answer. Reading Data filenames = ["s3://bucketname/path/to/file1.tfrecord", "s3://bucketname/path/to/file2.tfrecord"] dataset = tf.data.TFRecordDataset(filenames)

Comment: option B

Comment: Letters C and D need code development and are therefore discarded. As we want a scalable data storage, it is recommended to use the Letter B, since S3 is scalable. Letter A is wrong as your personal computer is not scalable.

Comment: Where had the Capslock Donald gone? I kinda miss his answers

Comment: Internet connectivity issue: then how IOT can be a solution? (Correct answer should be A)

Comment: Amazon SageMaker script mode enables training a machine learning model using a script that you provide. By using the unchanged train.py script and putting the TFRecord data into an Amazon S3 bucket, you can easily point the Amazon SageMaker training invocation to the S3 bucket without reformatting the training data. This option avoids the need to rewrite the train.py script or to prepare the data in a different format. It also leverages the scalability and cost-effectiveness of Amazon S3 for storing large amounts of data, which is important for training machine learning models.

Replies:

Comment: thank you chatgpt

Comment: B, obviously

Comment: I like answer B

Comment: Why not A? Why can't we train it from local?

Replies:

Comment: Sagemaker to my understanding requires the data to be in S3.

Comment: B. https://aws.amazon.com/about-aws/whats-new/2019/01/amazon-sagemaker-batch-transform-now-supports-tfrecord-format/

Comment: Unfortunilty you can't use the script unchanged, there are some things that need to be added: 1. Make sure your script can handle --model_dir as an additional command line argument. If you did not specify a location when you created the TensorFlow estimator, an S3 location under the default training job bucket is used. Distributed training with parameter servers requires you to use the tf.estimator.train_and_evaluate API and to provide an S3 location as the model directory during training. 2. Load input data from the input channels. The input channels are defined when fit is called. ## https://sagemaker.readthedocs.io/en/stable/frameworks/tensorflow/using_tf.html Beause of the pre-rec Ans A and B are an easy disqualifcation. There is no need to change the training format so option C is a red herring Ans is D Not the most obvious answer

Replies:

Comment: according your explaination, the correct answer should be B

Comment: It mentions using sagemaker in "script mode" Which is different from working on Sagemaker using python SDK.

Discussion for Question 114

Link: https://www.examtopics.com/discussions/amazon/view/45463-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- D: 7 votes

Discussion

Comment: D. CNN - image

Comment: yeah, nothing creepy about a company wanting to do this :)

Comment: D - image

Comment: D is the correct Well, I did those exams topics questions for 2 weeks and still easier compared to Maarek exams that simulate AWS MLS. Is that correct? Can I expect the exam to be easier as it's in exams topics?

Comment: Convolutional neural networks (CNNs) are specifically designed for image recognition tasks and have been highly successful in detecting patterns and features within images. CNNs are particularly effective in capturing spatial patterns and visual features from images

Comment: Answer is "D"

Discussion for Question 115

Link: https://www.examtopics.com/discussions/amazon/view/45464-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 11 votes

Discussion

Comment: A is the correct answer. Because in this case, it is not the problem with the existing historical data (event value, event type(click or not)), the sales do not keep growing and now you need to obtain more recent interactive data. An event tracker specifies a destination dataset group for new event data.

Replies:

> **Comment:** I agree.. A is the right choice.. The model need the real time data to adjust to create recommendations..

Comment: here is the receipt: https://docs.aws.amazon.com/personalize/latest/dg/maintaining-relevance.html

Comment: A real time data

Comment: A is the right choice.

Comment: A. Use the event tracker in Amazon Personalize to include real-time user interactions. By using the event tracker in Amazon Personalize, the data scientist can collect real-time user interactions, including clicks, views, and purchases, and use these interactions to update the model and generate more accurate recommendations. This could help address the decrease in sales after deploying a new solution version, as the model can be updated to reflect the latest customer behavior. Additionally, including real-time user interactions can help the model better respond to changes in customer behavior and provide more relevant and personalized recommendations, which can help increase sales.

Comment: easy one. A

Comment: A. https://docs.aws.amazon.com/personalize/latest/dg/recording-events.html

---

Discussion for Question 116

Link: https://www.examtopics.com/discussions/amazon/view/44301-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- AC: 11 votes

Discussion

Comment: A&C...>https://aws.amazon.com/blogs/machine-learning/securing-all-amazon-sagemaker-api-calls-with-aws-privatelink/

Comment: A - VPC endpoint policy can limit the access to specific group of user/roles Not B - setting iam user policy can limit user access other aws service but not secure the traffic C - “specific” sets of instances - means security rules in instance level Not D - ACL (access control list) allows or denies specific inbound or outbound traffic at the subnet level. Not E - VPC is configured with public subnet, adding interface without limit the traffic means not secure

Comment: A. YES - for users B. NO - the users should access more than just SageMaker C. YES - for instances D. NO - ACL are not supported for SageMaker endpoint (only S3, RDS, EKS, etc.) E. NO - endpoint is already there

Comment: A. Add a VPC endpoint policy to allow access to the IAM users: This will specify the permissions for the IAM users to access the Amazon SageMaker Service API through the VPC endpoint. C. Modify the security group on the endpoint network interface to restrict access to the instances: By configuring the security group, the specialist can control which instances are allowed to communicate with the SageMaker Service API through the VPC endpoint.

Comment: Should be A & D n0? We want to configure the endpoint - first to allow IAM users, second to control access to instances. Since Security Groups are attached to instances (not VPCs) and only allow allow rules - it should be D.

Comment: Yes, A & D are correct. A> This will limit access to only names IAM users. It is like defining all for given principals as below: { "Statement": [{ "Effect": "Allow", "Principal": "*", "Action": "*", "Resource": "*" }] } D-> To restrict access to certain instances or IP address you define deny rule at NACL level. Here VPC Interface endpoint is in subnet (the only subnet in VPC). So modify NACL configurations at this subnet level. Security group are only for allowing the traffic not for deny so so C is incorrect.

Comment: security group cannot restrict access explicitly, C?

Replies:

> **Comment:** i mean A, D

Comment: Security Group controls instance level access. The question requires instance level access. The VPC endpoint is already set up. It needs a policy attachment for particular IAM Users. I would have preferred this to be IAM Roles instead of Users, as a more appropriate question. Nevertheless, answer is A & C.

Comment: A say allow access TO the IAM users? That's wired, why to the IAM users? How do you access them?

Comment: The VPC endpoint is already available waiting to be configured. No need to add one. A and E are out. Furthermore if an IAM endpoint is not set, a default one will be provided and you can't have more than 1 IAM policy but can modify the one that's available. -Restric access to only calls coming from the VPC, then modify the security group to give access to user group or roles that need access to that notebook. I think the answer is B and C

Replies:

> **Comment:** A says add a VPC endpoint policy, not add an endpoint.

Comment: https://docs.aws.amazon.com/vpc/latest/userguide/vpc-endpoints-access.html https://docs.aws.amazon.com/sagemaker/latest/dg/notebook-interface-endpoint.html#nbi-private-link-policy

https://docs.aws.amazon.com/vpc/latest/userguide/integrated-services-vpce-list.html

---

Discussion for Question 117

Link: https://www.examtopics.com/discussions/amazon/view/44073-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 5 votes

Discussion

Comment: ASK : Extract Data over IPsec So we need an ETL + Site to site VPN GLUE is an ETL service but can it connect to PostgreSQL? yes https://docs.aws.amazon.com/glue/latest/dg/aws-glue-programming-etl-connect.html#aws-glue-programming-etl-connect-jdbc How to connect Glue to an on-site DB https://aws.amazon.com/blogs/big-data/how-to-access-and-analyze-on-premises-data-stores-using-aws-glue/ My Answer would be A Anser C only makes a 443 (SSL) connection so does not meet the IPsec requirement

Comment: A? IPSec needs to be covered as well

Replies:

Comment: Yes. https://aws.amazon.com/blogs/big-data/how-to-access-and-analyze-on-premises-data-stores-using-aws-glue/. 'A' is the correct answer.

Comment: A: https://medium.com/awsblackbelt/loading-on-prem-postgres-data-into-amazon-s3-with-server-side-filtering-c13bcee8b769

Comment: B - Doesnt take care of only nonsensitive data being allowed to leave the on-premise. C - Uses SSL and not IPSec. D - like B transfers all data. Hence the correct answer is A

Comment: I will go with B Site to Site VPN -> IPsec requirement AWS Glue -> connect and catalog PostgressSQL Pyspark -> remove sensitive information. AWS glue supports pyspark

Comment: The best option is to use AWS Database Migration Service (AWS DMS) with table mapping to select PostgreSQL tables with no sensitive data through an SSL connection. Replicate data directly intoAmazon S3. This option meets the following requirements:It ensures that only nonsensitive data is transferred to the cloud by using table mapping to filter outthe tables that contain sensitive data1.It uses IPsec to secure the data transfer by enabling SSL encryption for the AWS DMS endpoint2.It uploads the data to Amazon S3 each day for model retraining by using the ongoing replicationfeature of AWS DMS3

Comment: but glue can not filter out the data during the ingestion and hence option A wouldn't be the right one! I would go for B

Comment: B Both A and C are not correct... due that the question is not talking about tables with no sensitive data... and that DMS tipically act on the data on AWS side, the right answer is B AWS Glue connects to the PostgreSQL database, allowing the removal of sensitive data using a PySpark job BEFORE securely ingesting the data into Amazon S3, thus aligning with the requirements.

Replies:

Comment: I think the issue with this answer would be that the data actually leaves the DC and enters the glue service before sensitive data is redacted. - Which makes me lean A

Comment: Option c

Comment: A: https://aws.amazon.com/blogs/big-data/doing-data-preparation-using-on-premises-postgresql-databases-with-aws-glue-databrew/

Comment: A. Create an AWS Glue job to connect to the PostgreSQL DB instance. Ingest tables without sensitive data through an AWS Site-to-Site VPN connection directly into Amazon S3. This solution meets the requirements of data localization regulations and secure data transfer. By creating an AWS Glue job to connect to the PostgreSQL DB instance, the machine learning specialist can extract tables without sensitive data. By using a Site-to-Site VPN connection, the data can be securely transferred from the on-premises data center to Amazon S3, where it can be used for model retraining. This solution ensures that any sensitive data remains in the on-premises data center, and that only non-sensitive data is uploaded to the cloud.

Comment: IPSec means VPN

Comment: Answer is A. IPsec is not the same as SSL. Site to site VPN is for IPsec: https://aws.amazon.com/vpn/site-to-site-vpn/ Also Glue can directly connect to Postgres and upload to S3: https://aws.amazon.com/blogs/big-data/how-to-access-and-analyze-on-premises-data-stores-using-aws-glue/

Comment: A is the answer

Comment: Between A and C, I pick A because IPSec requires VPN; otherwise DMS is a better option

Comment: Between A and C, I pick A because IPSec requires VPN; otherwise DMS is a better option

Comment: 100% A, DMS is not supporting S3, but glue is. Also, Site to Site is supporting IPSec.

Discussion for Question 118

Link: https://www.examtopics.com/discussions/amazon/view/45611-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- AD: 7 votes

Discussion

Comment: I would choose A and D, however both of them is not possible at the same time. The question is ambiguous, it could mean which two options, but no necessarily both. A - If you want Amazon Forecast to evaluate each algorithm and choose the one that minimizes the objective function, set PerformAutoML to true. D - The following algorithms support HPO: - > DeepAR+.

Replies:

Comment: If custom forecast types are specified, Forecast evaluates metrics at those specified forecast types, and takes the averages of those metrics to determine the optimal outcomes during HPO and AutoML. For both AutoML and HPO, Forecast chooses the option that minimizes the average losses over the forecast types. During HPO, Forecast uses the first backtest window to find the optimal hyperparameter values. During AutoML, Forecast uses the averages across all backtest windows and the optimal hyperparameters values from HPO to find the optimal algorithm. https://docs.aws.amazon.com/forecast/latest/dg/metrics.html

Comment: It is A and D, there are no weekly data, they have only monthly data and can not switch horizon to 4

Comment: A. YES - DeepAR+ most likely to be chosen, but worth a try B. NO - increasing the forecast horizon is not likely to improve the 3 months we want C. NO - we want monthly, nto weekly D. YES E. NO - there are no missing values

Comment: The changes to the CreatePredictor API call that could improve the MAPE are option A and option D. By setting PerformAutoML to true, you can enable Amazon Forecast to automatically explore different algorithms and choose the best one for your data and business problem. By setting PerformHPO to true, you can enable Amazon Forecast to perform hyperparameter optimization (HPO) and tune the algorithm parameters to improve the accuracy of the predictor. These options can help you find the optimal configuration for your forecast model without manually specifying the algorithm or the hyperparameters.

Comment: A. Set PerformAutoML to true. D. Set PerformHPO to true. Setting PerformAutoML to true will enable Amazon Forecast to automatically select the best algorithm and hyperparameters for your data and problem. This can help improve the MAPE by finding the optimal combination of algorithm and hyperparameters that minimize prediction error. Setting PerformHPO to true will enable Amazon Forecast to perform a hyperparameter optimization search to find the best combination of hyperparameters that result in the best prediction performance. This can help improve the MAPE by finding the optimal combination of hyperparameters that minimize prediction error.

Comment: A. Looking for better algorithms performance D. Hyperparameters optimization

Comment: 12-sep exam

Comment: Why are not B and C? The question asks about modifications that increase MAPE (thats bad): B - If FH is larger, error will increase C - Data is based on months, change that will make erros on forecasting values E - There is no data gap so is useless A - Selec best between all should DECREASE MAPE D - Tunning hyperparms will DECREASE MAPE

Comment: A&D...>By default, Amazon Forecast uses the 0.1 (P10), 0.5 (P50), and 0.9 (P90) quantiles for hyperparameter tuning during hyperparameter optimization (HPO) and for model selection during AutoML. If you specify custom forecast types when creating a predictor, Forecast uses those forecast types during HPO and AutoML. If custom forecast types are specified, Forecast evaluates metrics at those specified forecast types, and takes the

averages of those metrics to determine the optimal outcomes during HPO and AutoML. For both AutoML and HPO, Forecast chooses the option that minimizes the average losses over the forecast types. During HPO, Forecast uses the first backtest window to find the optimal hyperparameter values. During AutoML, Forecast uses the averages across all backtest windows and the optimal hyperparameters values from HPO to find the optimal algorithm.

Comment: C. ForecastFrequency M- MONTHLY W- WEEKLY D. PerformHPO Whether to perform hyperparameter optimization (HPO). HPO finds optimal hyperparameter values for your training data. The process of performing HPO is known as running a hyperparameter tuning job. The default value is false. In this case, Amazon Forecast uses default hyperparameter values from the chosen algorithm. E. FeaturizationMethodName The name of the method. The "filling" method is the only supported method.

Replies:

> **Comment:** But for option C, according to the Developer Guide, The forecast frequency must be greater than or equal to the TARGET_TIME_SERIES dataset frequency. and the training data is monthly data, so ForecastFrequency can not be less than Monthly.

Comment: ABE can be excluded. CD is my answer. A. PerformAutoML If you want Amazon Forecast to evaluate each algorithm and choose the one that minimizes the objective function, set PerformAutoML to true. The objective function is defined as the mean of the weighted losses over the forecast types. By default, these are the p10, p50, and p90 quantile losses. When AutoML is enabled, the following properties are disallowed: AlgorithmArn HPOConfig PerformHPO TrainingParameters B. ForecastHorizon Specifies the number of time-steps that the model is trained to predict. The forecast horizon is also called the prediction length. For example, if you configure a dataset for daily data collection (using the DataFrequency parameter of the CreateDataset operation) and set the forecast horizon to 10, the model returns predictions for 10 days. The maximum forecast horizon is the lesser of 500 time-steps or 1/3 of the TARGET_TIME_SERIES dataset length.

Discussion for Question 119

Link: https://www.examtopics.com/discussions/amazon/view/44077-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 13 votes

Discussion

Comment: I would answer A. Target and metadata must be in two files and loaded from S3, based on documentation: https://docs.aws.amazon.com/forecast/latest/dg/dataset-import-guidelines-troubleshooting.html

Replies:

> **Comment:** 1. I cannot find any evidence support the seperate file defination. 2. A,B,C all seperate datasets, this explanation is weak.

Comment: Amazon Forecast requires the input data to be separated into a target time series dataset and an item metadata dataset. The target time series dataset should include the time series data that you want to use for forecasting, such as inventory demand in this case. The item metadata dataset should include the metadata that describes the items in the time series, such as product IDs, categories, and attributes. Therefore, the data scientist should use ETL jobs in AWS Glue to separate the dataset into a target time series dataset and an item metadata dataset. Both datasets should be uploaded as .csv files to Amazon S3, which is a suitable storage option for input data to Amazon Forecast.

Replies:

> **Comment:** thank you chatgpt

Comment: I would go with A Input formats for forecast -> Json, CSV and paraquet (Selects A & eliminates B, C, D) Data needs to be split in target time series dataset and an item metadata dataset

Comment: Target and metadata must be in two files

Comment: Letter A is correct, as it uses a specific transformation service (AWS Glue) and saves it in a cloud database for AWS Forecast to access. By default in ML, our storage option will be AWS S3 (unless caveats or issue specifications). That said, we discard B and C. Letter D is discarded due to the format requested by AWS Forecast being csv.

Comment: I would vote for A

Comment: The answer is A. According to the https://docs.aws.amazon.com/forecast/latest/dg/forecast.dg.pdf , page 51. Target Time Series Dataset: Required: timestamp, item_id, demand Additional: lead_time Item Metadata Dataset: item_id, category

Replies:

> **Comment:** You can find the same question with the picture at https://ccnav7.net/a-data-scientist-wants-to-use-amazon-forecast-to-build-a-forecasting-model-for-inventory-demand-for-a-retail-company/

Comment: The correct answer is A "Forecast supports only the comma-separated values (CSV) file format. You can't separate values using tabs, spaces, colons, or any other characters. Guideline: Convert your dataset to CSV format (using only commas as your delimiter) and try importing the file again."

Comment: lead time belongs to related time series, as its not a target variable

---

Discussion for Question 120

Link: https://www.examtopics.com/discussions/amazon/view/44078-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 7 votes

Discussion

Comment: B is correct. Redeploy with CPU and add elastic inference to reduce costs. See: https://aws.amazon.com/machine-learning/elastic-inference/

Comment: redeploying the model on a P3dn instance is the best approach to ensure the provisioned GPU resources are being utilized effectively.

Comment: My vote is B Elastic inference - provides cheper accelration that full GPU - works with M class machines - Works with Tensorflow, MXNet, pytorch, image classification and object detection algorithms

Comment: Elastic Inference has been depreciated since Apr 2023.

Comment: can reduce the cost and improve the resource utilization of your model, as Amazon Elastic Inference allows you to attach low-cost GPU-powered acceleration to Amazon EC2 and Amazon SageMaker instances to run inference workloads with a fraction of the compute resources. You can also choose the right amount of inference acceleration that suits your needs, and scale it up or down as needed.

Comment: Amazon Elastic Inference allows you to attach low-cost GPU-powered acceleration to EC2 and Sagemaker instances, to reduce the cost of running deep learning inference. You can choose any CPU instance that is best suited to the overall compute and memory needs of your application, and then separately configure the right amount of GPU-powered inference acceleration. This would allow you to efficiently utilize resources and reduce costs.

Comment: Elastic inference enables GPU only when load increases. With 50% utilisation there is no need to deploy P3 as the base inference machine.

Comment: Agreed with B

Comment: 12-sep exam

Comment: Answer: B Explanation: https://aws.amazon.com/machine-learning/elastic-inference/

Comment: B: production mostly needs CPU with EI rather than GPU machines

Comment: B..>Amazon Elastic Inference (EI) is a resource you can attach to your Amazon EC2 CPU instances to accelerate your deep learning (DL) inference workloads. Amazon EI accelerators come in multiple sizes and are a cost-effective method to build intelligent capabilities into applications running on Amazon EC2 instances.

Comment: B is correct

Discussion for Question 121

Link: https://www.examtopics.com/discussions/amazon/view/44080-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- D: 14 votes

Discussion

Comment: I would select D. See AWS documentation: https://docs.aws.amazon.com/sagemaker/latest/dg/nbi-add-external.html

Comment: by excluding wrong options: A you might not have access to the EC2 instance => out B no automation => out C only the default kernel, which limits the DS => out => D

Comment: option D

Comment: D. https://docs.aws.amazon.com/sagemaker/latest/dg/nbi-add-external.html

Comment: Install custom environments and kernels on the notebook instance's Amazon EBS volume. This ensures that they persist when you stop and restart the notebook instance, and that any external libraries you install are not updated by SageMaker. To do that, use a lifecycle configuration that includes both a script that runs when you create the notebook instance (on-create) and a script that runs each time you restart the notebook instance (on-start).

Comment: Even the link given suggest Option D

Comment: Please ignore my previous comment, the answer is D

Comment: Key word here is how can the developer "guarantee"?? He guarantees that by including the install commands as part of the notebook. So, against the grain, I stand with B

Replies:

> **Comment:** Scratch that. The Answer is D

Comment: D should be the answer

Comment: https://docs.aws.amazon.com/sagemaker/latest/dg/nbi-add-external.html upvoted 1 times

Comment: D "automatically" is the key here and using lifecycle configuration https://docs.aws.amazon.com/sagemaker/latest/dg/nbi-add-external.html

Comment: It is D, although is not even the best answer in my opinion. Although by default conda packages are installed in ephemeral storage, you can change that default behaviour. I did that in my last project and we created our own conda environment that persisted between shutdowns.

Comment: based on the refernce given under the answer its D not B

Comment: You can install packages using the following methods: 1-Lifecycle configuration scripts 2-Notebooks – The following commands are supported. %conda install %pip install 3-The Jupyter terminal – You can install packages using pip and conda directly.

Comment: NOT B ...>/etc/init contains configuration files used by Upstart. ANS...>D

Comment: Its for sure D

Comment: D https://docs.aws.amazon.com/sagemaker/latest/dg/nbi-add-external.html

---

Discussion for Question 122

Link: https://www.examtopics.com/discussions/amazon/view/44079-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: B ,You can use the FindMatches transform to find duplicate records in the source data. A labeling file is generated or provided to help teach the transform.

Comment: B it is. Reasonable explanation.

Comment: option B Find matches

Comment: It's B, how it's using Glue to clean the data the easiest way will be use Glue's ML FindMatches extension to do this too.

Comment: It is B.

Comment: Agree. Please refer to: https://aws.amazon.com/blogs/big-data/integrate-and-deduplicate-datasets-using-aws-lake-formation-findmatches/

Discussion for Question 123

Link: https://www.examtopics.com/discussions/amazon/view/75336-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- BD: 16 votes

Discussion

Comment: B and D

Comment: Compensate for imbalance and optimize on AUC. This is a class imbalance problem, not an overfitting problem.

Replies:

> **Comment:** totally right, overfitting has nothing to do so there is no need to reduce tree depth

Comment: A. NO - that will not address FN specifically but also FP B. YES - changing weight is best practice for class imbalance C. NO - there is no underfitting at 99.1% accuracy D. YES - AUC will address recall, which takes into account FN rate E. NO - there is no overfitting at 99.1% accuracy

Comment: Step B: Increasing the XGBoost scale_pos_weight parameter to adjust the balance of positive and negative weights can help the model deal with the imbalanced dataset. According to the XGBoost documentation, this parameter controls the balance of positive and negative weights, and is useful for unbalanced classes. A typical value to consider is sum(negative instances) / sum(positive instances). In this case, since there are 100 times more non-fraudulent transactions than fraudulent ones, setting scale_pos_weight to 100 can make the model more sensitive to the minority class and reduce false negatives. Step D: Changing the XGBoost eval_metric parameter to optimize based on Area Under the ROC Curve (AUC) can help the model focus on improving the true positive rate and the true negative rate, which are both important for fraud detection. According to the XGBoost

Comment: Step B: Increasing the XGBoost scale_pos_weight parameter to adjust the balance of positive and negative weights can help the model deal with the imbalanced dataset. According to the XGBoost documentation, this parameter controls the balance of positive and negative weights, and is useful for unbalanced classes. A typical value to consider is sum(negative instances) / sum(positive instances). In this case, since there are 100 times more non-fraudulent transactions than fraudulent ones, setting scale_pos_weight to 100 can make the model more sensitive to the minority class and reduce false negatives. Step D: Changing the XGBoost eval_metric parameter to optimize based on Area Under the ROC Curve (AUC) can help the model focus on improving the true positive rate and the true negative rate, which are both important for fraud detection.

Comment: I have some doubts about D and E. Precision-Recall AUC is better than AUC curve in imbalanced classes. Then, I choose E

Comment: Option A and Option E are unlikely to help reduce false negatives. Option C, increasing max_depth, may lead to overfitting, which could make the model worse. Option D, changing the eval_metric to optimize based on AUC, could help improve the model's ability to discriminate between the two classes. Option B, increasing the scale_pos_weight parameter to adjust the balance of positive and negative weights, can help the model better handle imbalanced datasets, which is the case here. By increasing the weight of positive examples, the model will learn to prioritize correctly classifying them, which should reduce the number of false negatives.

Comment: BD, I have done this before, but it would be better to use Average Precision(AP) instead of AUC, but it is better than other answers.

Comment: 12-sep exam

Comment: Compensate for imbalance and overwriting.

Comment: B and E

Comment: B. Increase the XGBoost scale_pos_weight parameter to adjust the balance of positive and negative weights is the correct answer. According to https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost_hyperparameters.html, scale_pos_weight controls the balance of positive and negative weights. It's useful for unbalanced classes.

Discussion for Question 124

Link: https://www.examtopics.com/discussions/amazon/view/44186-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 11 votes

Discussion

Comment: I agree with an answer of C Attention mechanism. The disadvantage of an encoder-decoder framework is that model performance decreases as and when the length of the source sequence increases because of the limit of how much information the fixed-length encoded feature vector can contain. To tackle this problem, in 2015, Bahdanau et al. proposed the attention mechanism. In an attention mechanism, the decoder tries to find the location in the encoder sequence where the most important information could be located and uses that information and previously decoded words to predict the next token in the sequence.

Comment: C. By tuning attention-related hyperparameters (such as attention type, attention layer size, and dropout), the model can focus on relevant parts of the input sequence during translation.

Comment: A. NO - n-grams are more the opposite, it is to capture local information B. NO - it could help, but not best C. YES - best practice (https://docs.aws.amazon.com/sagemaker/latest/dg/seq-2-seq-hyperparameters.html) D. NO - weight are for vectorized words, they do not relate to sequences

Comment: Action C: Adjusting hyperparameters related to the attention mechanism can help improve the translation quality for long sentences, because the attention mechanism allows the decoder to focus on the most relevant parts of the source sentence at each time step. According to the Amazon SageMaker documentation, the seq2seq algorithm supports several types of attention mechanisms, such as dot, general, concat, and location. The data scientist can experiment with different values of the hyperparameters attention_type, attention_coverage_type, and attention_num_hidden to find the optimal configuration for the translation task.

Comment: C. Adjust hyperparameters related to the attention mechanism. The seq2seq algorithm uses an attention mechanism to dynamically focus on relevant parts of the input sequence for each output sequence element. Increasing the attention mechanism's ability to learn dependencies between long input and output sequences might help improve the translation quality for long sentences. The data scientist could try adjusting relevant hyperparameters such as attention depth or attention scale, or try a different attention mechanism such as scaled dot-product attention, to see if that improves the translation quality for long sentences.

Comment: i go with C

Comment: Ans: C Explanation: https://docs.aws.amazon.com/sagemaker/latest/dg/seq-2-seq- howitworks.html

Comment: This is such a niche question for a niche market. Geared towards someone who specializes in NLP.

Comment: c is correct https://docs.aws.amazon.com/sagemaker/latest/dg/seq-2-seq-howitworks.html

Comment: I believe the answer is C https://docs.aws.amazon.com/sagemaker/latest/dg/seq-2-seq-howitworks.html

---

Discussion for Question 125

Link: https://www.examtopics.com/discussions/amazon/view/44081-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- DE: 9 votes

Discussion

Comment: D, E is the answer. we need to make the recall rate(not precision) high.

Comment: To maximize detection of fraud in real-world, inbalanced datasets, D and E should always be applied. https://en.wikipedia.org/wiki/Sensitivity_and_specificity https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-imbalanced-classification/

Replies:

Comment: Note, True positive rate = Sensitivity = Recall

Replies:

Comment: that is not correct unfortunately Recall is = Sensitivity = False Negative which is a Type II error Precision = specificity = False Positive which is a Type I error I do agree that in the real world you would focus on Recall/sensitivity ie. reducing type II errors. However, in the question, they want to reduce the False Positives so you would need to focus on precision and specificity minimizing type I errors

Replies:

Comment: My ANS would be AB

Replies:

Comment: I came a cross this article and I would say it might support choosing A&B as answers: https://medium.com/datadriveninvestor/rethinking-the-right-metrics-for-fraud-detection-4edfb629c423

Comment: specifity is different than precision

Comment: "accurately capture positives" means maximize TPR. A. NO - Specificity = TN / (TN + FP) is a measure of negative cases B. NO - FPR = FP / Total C. NO - given the class imbalance, overall accuracy would not help D. YES - not sure if we need that on top of E, but other options are eliminated anyway E. YES - TPR = TP / Total, what we want

Comment: The data scientist should use True positive rate and Area under the precision-recall curve to optimize the model. The true positive rate (TPR) is the proportion of actual positives that are correctly identified as such. It is also known as sensitivity or recall. In this case, it is important to capture as many fraudulent transactions as possible, so the TPR should be maximized. The area under the precision-recall curve (AUPRC) is a measure of how well the model is able to distinguish between positive and negative classes. It is a good metric to use when the classes are imbalanced, as in this case where only 2% of transactions are fraudulent. The AUPRC summarizes the trade-off between precision and recall across all possible thresholds. Accuracy and specificity are not good metrics to use when the classes are imbalanced because they can be misleading. The false positive rate (FPR) is also not a good metric to use because it does not take into account the number of true negatives.

Comment: Metric D: Area under the precision-recall curve (AUPRC) is a good metric to use for imbalanced classification problems, where the positive class is much less frequent than the negative class. Precision is the proportion of positive predictions that are correct, and recall (or true positive rate) is the proportion of positive cases that are detected. AUPRC summarizes the trade-off between precision and recall for different decision thresholds, and a higher AUPRC means that the model can achieve both high precision and high recall. Since the company's goal is to accurately capture as many positives as possible, AUPRC can help them evaluate how well the model performs on the minority class. Metric E: True positive rate (TPR) is another good metric to use for imbalanced classification problems, as it measures the sensitivity of the model to the positive class. TPR is the same as recall, and it is the proportion of positive cases that are detected by the model. A higher TPR means that the model can identify more fraudulent transactions, which is the company's goal.

Comment: The goal is to accurately capture as many fraudulent transactions (positives) as possible. To optimize the model towards this goal, the data scientist should focus on metrics that emphasize the true positive rate and the area under the precision-recall curve. True positive rate (TPR or sensitivity) is the proportion of actual positive cases that are correctly identified as positive by the model. A higher TPR means that more fraudulent transactions are being

captured. The precision-recall curve is a graph that shows the trade-off between precision and recall for different thresholds.

Replies:

> **Comment:** Precision is the fraction of correctly identified positive instances among all instances the model has classified as positive. Recall, also known as the true positive rate, is the fraction of positive instances that are correctly identified as positive by the model. A higher area under the precision-recall curve indicates that the model is making fewer false positive predictions and more true positive predictions, which aligns with the goal of the financial company to accurately capture as many fraudulent transactions as possible.

Comment: agreed with DE

Comment: Why not A and D? - Specificity shows us how the FNR is - AUC PR includes Precision and Recall which shows us the ratio of TP to TP/FP and TP to TP / FN

Comment: Answer: D&E

Comment: I meant say D&E not BD

Comment: I believe the answer is B&D, which equals F1. F1 combines precision and Sensitivity.

Comment: D&E is the only choices that takes False Negatives into considration

Replies:

> **Comment:** TPR is already included in the AUC PR TNR is not included in all others besides A

Comment: Recall and TPR D and E

Comment: AB is the answer

Discussion for Question 126

Link: https://www.examtopics.com/discussions/amazon/view/46341-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- D: 8 votes

Discussion

Comment: I will go with D, "cannot be accessed and transferred to a remote host by malicious code accidentally installed on the training container" Based on the following link: https://aws.amazon.com/blogs/security/secure-deployment-of-amazon-sagemaker-resources/ "EnableNetworkIsolation – Set this to true when creating training, hyperparameter tuning, and inference jobs to prevent situations like malicious code being accidentally installed and transferring data to a remote host."

Comment: If you enable network isolation, the containers can't make any outbound network calls, even to other AWS services such as Amazon S3. Additionally, no AWS credentials are made available to the container runtime environment. In the case of a training job with multiple instances, network inbound and outbound traffic is limited to the peers of each training container. SageMaker still performs download and upload operations against Amazon S3 using your SageMaker execution role in isolation from the training or inference container.

Replies:

> **Comment:** ahaha this link literally contains the answer For example, a malicious user or code that you accidentally install on the container (in the form of a publicly available source code library) could access your data and transfer it to a remote host.

Comment: 'network isolation' make sense

Comment: A. NO - Remove Amazon S3 access permissions from the SageMaker execution role. B. NO - Encrypting the weights has nothing to do with protecting the training data C. NO - If the dataset is encrypted, one may still hack SageMaker instance and get access to uncrypted data D. YES - Enable network isolation for training jobs, data is protected end-to-end

Comment: Network isolation

Comment: It's D, not C because encrypted can be stole.

Comment: I choose D. More document about it: https://docs.aws.amazon.com/sagemaker/latest/dg/mkt-algo-model-internet-free.html

Comment: Answer is D. https://aws.amazon.com/blogs/security/secure-deployment-of-amazon-sagemaker-resources/ search for 'isolation' and there is a security parameter : EnableNetworkIsolation talking about this.

Comment: I would choose C

Comment: most likely it is C. https://docs.aws.amazon.com/sagemaker/latest/dg/data-protection.html

Replies:

> **Comment:** incorrect; you CAN transfer encrypted files even w/o a key D is a better option

---

Discussion for Question 127

Link: https://www.examtopics.com/discussions/amazon/view/44091-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 5 votes

Discussion

Comment: I would answer C, because of the requirement that authorized users should only have access. These users will comprise the private workforce of AWS Ground Truth. See documentation: https://docs.aws.amazon.com/sagemaker/latest/dg/sms-workforce-private.html

Replies:

> **Comment:** agree C
>
> **Comment:** Yes it is C
>
> **Comment:** Agree C

Comment: Answer is C. The question mentions that "to detect *areas* of concern on patients' CT scans", that can be achieved by bounding box instead of image classification. bounding box: https://docs.aws.amazon.com/sagemaker/latest/dg/sms-bounding-box.html image classification: https://docs.aws.amazon.com/sagemaker/latest/dg/sms-image-classification.html

Replies:

> **Comment:** The concern here is not about "object detection" or "image classification". it is about using "ground truth" and "private workforce"

Comment: The principal key is that Mechanical Turk workforce does not ensure privacy of the CT Scans, and Ground Truth does.

Comment: This option would allow the medical imaging company to create a private workforce, which can ensure that only authorized users have access to the scans, and to use Amazon SageMaker Ground Truth to create a labeling job, which would simplify the labeling pipeline process.

Comment: 12-sep exam

Comment: C - GroundTruth and privacy concerns

Discussion for Question 128

Link: https://www.examtopics.com/discussions/amazon/view/44092-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: I agree with C, given we are evaluating model inferences (predictions). See https://aws.amazon.com/augmented-ai/ and https://aws.amazon.com/blogs/machine-learning/automated-monitoring-of-your-machine-learning-models-with-amazon-sagemaker-model-monitor-and-sending-predictions-to-human-review-workflows-using-amazon-a2i/

Replies:

Comment: yeap, it literally says it there Loan or mortgage applications, tax forms, and many other financial documents contain millions of data points which need to be processed and extracted quickly and effectively. Using Amazon Textract and Amazon A2I you can extract critical data from these forms

Comment: why not a?

Replies:

Comment: i think ground truth can do same task instead of Augmented AI

Comment: the differences rely on the function of these two service. Ground Truth is used for "labeling" typically, text or image label: if the service cannot automatically label the data, it send to ground truth and wait for human to label it. but A2I is for validate prediction. The model already predict the results and human then add views to it.

Replies:

Comment: https://docs.aws.amazon.com/textract/latest/dg/a2i-textract.html https://aws.amazon.com/blogs/machine-learning/using-amazon-textract-with-amazon-augmented-ai-for-processing-critical-documents/

Comment: The answer is C, Augmented AI is made for review ML predictions!

Comment: By routing the low-confidence predictions to Amazon Augmented AI, the company can reduce the time to process the loan applications by leveraging human intelligence to review and validate the predictions. This way, the company can quickly address any errors or mistakes that Amazon Textract might make, reducing the time to process loan applications.

Comment: correct is C

Discussion for Question 129

Link: https://www.examtopics.com/discussions/amazon/view/45615-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 20 votes
- A: 12 votes

Discussion

Comment: C is the correct answer. # of shard is determined by: 1. # of transactions per second times 2. data blob eg. 100 KB in size 3. One shard can Ingest 1 MB/second

Comment: the answer should be A - the reason why shards are not the right answer is the lack of ProvisionedThroughputExceeded exceptions that occur when a KDS has too few shards. The scenario talks about a consistent pace of delivery into S3 and a rising backlog of data (which indicates KDS stream is still able to ingest data) in the stream, hence the S3 write limit per prefix is at fault: https://www.amazonaws.cn/en/kinesis/data-streams/faqs/#:~:text=Q%3A%20What%20happens%20if%20the%20capacity%20limits%20of%20a%20Kinesis%20stream%20are%20exceeded%20while%20the%20data%20producer%20adds%20data%20to%20the%20stream https://docs.aws.amazon.com/AmazonS3/latest/userguide/optimizing-performance.html

Replies:

Comment: from https://aws.amazon.com/kinesis/data-firehose/faqs/?nc1=h_ls Q: How often does Kinesis Data Firehose read data from my Kinesis stream? A: Kinesis Data Firehose calls Kinesis Data Streams GetRecords() once every second for each Kinesis shard. // and the number of records per GetRecords() is at most 10.000 => having n shards you will get at most 10.000n records to firehose per sec. => hence firehose instead of s3 could be the limiting factor. => i'd also go with inc shards as the first choice (to not having to change the S3 consumers)

Comment: shards is a concept in kinesis data stream. But here the topic mention "There also is an increasing backlog of data for Kinesis Data Streams and Kinesis Data Firehose to ingest" So even firehose has large backlogs, which means the limit comes from the S3. So A.

Comment: The bottle neck is not at data ingestion (i.e. Kinesis shards), but in write to S3, which throughput is bound by prefixes used.

Comment: A is not solving the issue, the bottleneck locate not in S3 but in the KDS, so we should solve the problem at the KDS, the Shards

Comment: I think the question is very ambiguous. "There also is an increasing backlog of data for Kinesis Data Streams and Kinesis Data Firehose to ingest.", that suggest the backlog is on the client-side (even before reaching KDS). Any component down the chain can be a bottlneck (KDS shrad, Firehose, S3). There is just no way to know in my opinion, but increasing shard is certainly the easiest to try without impact the storage structure in S3 and possibly breaking the app.

Comment: this is my key word to solve this problem : There also is an increasing backlog of data for Kinesis Data Streams and Kinesis Data Firehose to ingest. so increasing the shards to ingest is the solution

Comment: no of shards

Comment: A is not correct, because "There also is an increasing backlog of data for Kinesis Data Streams and Kinesis Data Firehose to ingest", the backlog is totally not caused by S3 performance, but the shard issue.

Comment: To increase ingest

Comment: The increasing backlog of data for Kinesis Data Streams and Kinesis Data Firehose indicates that the ingestion rate is slower than the data production rate. Therefore, the next step to improve the data ingestion rate into Amazon S3 is to increase the capacity of Kinesis Data Streams by increasing the number of shards. This will increase the parallelism of data processing, allowing for a higher throughput rate. Option C is the correct answer. Option A is incorrect because increasing the number of S3 prefixes for the delivery stream will not directly affect the ingestion rate into S3.

Comment: To improve the data ingestion rate into Amazon S3, the ML specialist should consider increasing the number of shards for the Kinesis data stream. A Kinesis data stream is made up of one or more shards, and each shard provides a fixed amount of capacity for ingesting and storing data. By increasing the number of shards, the specialist can increase the overall capacity of the data stream and improve the rate at which data is ingested.

Comment: C is the correct answer

Comment: Clearly S3 is a bottleneck. S3 has parallel perfromance acrtoss prefixes, thus increasing throughput

Comment: It seems S3 is the bottlneck. Adding more prefixes will help: https://docs.aws.amazon.com/AmazonS3/latest/userguide/optimizing-performance.html

Comment: 12-sep exam

Comment: The question seems to indicate the problem in the ability of S3 to load the data. Therefore, I think the answer is A. https://docs.aws.amazon.com/firehose/latest/dev/dynamic-partitioning.html

---

Discussion for Question 130

Link: https://www.examtopics.com/discussions/amazon/view/44093-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- D: 14 votes
- B: 12 votes

Discussion

Comment: I would select B, straight from this AWS example: https://aws.amazon.com/blogs/machine-learning/building-a-customized-recommender-system-in-amazon-sagemaker/

Replies:

Comment: the blog didn't mentioned anything about sample selection. how is B arrived?

Comment: I think the answer is D because customers by only 4-5 products every 5-10 years so it doesn't make sense to get 10% interactions for each user as a test set.

Replies:

Comment: B. Recommendation should use the historcial to predict the furture action. B is using the older records to prediect the newer records. D is using 90% user to predict other 10%, 90% is irrelevant to other 10%.

Comment: Yes, agree. Answer should be D

Comment: There is no difference between A and D, so I prefer B as the answer

Comment: The best way to split the dataset into a training and test set for this use case is to randomly select 10% of the users and split off all interaction data from these users for the test set. This is because the company relies on a steady stream of new customers, so the test set should reflect the behavior of new customers who have not been seen by the model before. The other options are not suitable because they either mix old and new customers in the test set (A and B), or they bias the test set towards users with less interaction data . References: Amazon SageMaker Developer Guide: Train and Test Datasets Amazon Personalize Developer Guide: Preparing and Importing Data

Comment: Primary concern is to evaluate the model's performance on completely new users then option D would be more appropriate.

Comment: I'd also take time into consideration, since even for such long-lived products there might be trends or regulations or whatever that make customers prefer one over the other. => A,D are out C will not give you a test set of desired size => out => B

Comment: If the primary concern is to evaluate the model's performance on completely new users (which seems to be the case for the company in question), then option D would be more appropriate.

Comment: I would choose D. According to the question, because of the product nature, the company doesn't rely on customer-product historical interactions for recommendations. It relies on customer explicit preferences, which are gathered on the first sign-up. The company wants to make recommendations for these new users. It is the main source of revenue for the company. To conduct thorough testing company needs to simulate the new users, not existing ones. To do it we need to randomly choose some percentage of users and remove all of their transactions from the train set. And use their transactions only in test.

Comment: By selecting the most recent interactions for each user, you are simulating the scenario of having new customers in your test set. This method allows you to assess how well the model generalizes to both existing and new users.

Comment: A. NO - the data is denormalized and users' preferences are present in multiple rows in the interactions; if we split off interactions, we introduce leakage as the same user will be present in train & test A. NO - the data is denormalized and users' preferences are present in multiple rows in the interactions; if we split off based on the interaction, we introduce leakage as the same user will be present in train & test C. NO - bias D. YES - no bias and user based

Comment: A NO introduces a bias in the training set (old interactions) vs. test set (new interactions) C NO will have a very sparse test set B NO the same user will be present in the training and test set; we want a user-based model, not an interaction-based one, so a user should belong to only one set D YES - last remaining option.

Comment: Changing to B

Comment: Between B and D but the issue is 4-5 transaction every 5-10 years. Hence last 10% transaction is difficult. So going for D

Comment: I would select B as it is time series data. Order might be important. So for each user, last 10% of transactions ordered by date could be a good answer.

Comment: You want different users in training and in testing datasets, which is C or D. In addition, B is wrong since you cannot take 10% of 4-5 transactions per customer. Actually, between B, C and D, only in D you can get exactly 10%.

Comment: This method is appropriate because it takes into account the unique buying behavior of each customer and is likely to reflect the latest preferences of the customer. It ensures that the test set contains a representative sample of the most recent customer preferences, which is important in this use case where customer preferences change infrequently over time.

Comment: B makes the most business sense. Since customers buy products every 4-5 years, it makes sense to be able to predict future sales from really old data. splitting the test set to be only recent interactions is the best way to test model performance from historically 'recent' data

---

Discussion for Question 131

Link: https://www.examtopics.com/discussions/amazon/view/44155-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- ADE: 20 votes

Discussion

Comment: ADF - the concepts in ADF are explained in detail on the official Amazon Exam Readiness Exam Readiness: AWS Certified Machine Learning - Specialty. Amazon official materials do not mention other concepts in BCE.

Replies:

Comment: ADE for sure. F is for encryption and not data egress.

Comment: I agree with ADF. SCP is to control access to a service, it's not related to securing data.

Comment: As per official document only 4 ways to do data egress Enforcing deployment in VPC,Enforcing network isolation,Restricting notebook pre-signed URLs to IPs,Disabling internet access Correct Ans - ADE Read Controlling data egress section Link - https://aws.amazon.com/blogs/machine-learning/millennium-management-secure-machine-learning-using-amazon-sagemaker/

Comment: F - it takes care of data sitting in sagemaker env which is encrypted but E ensures that the srvices or its resources cannot be accessed outside of the allowed IP's

Comment: My vote for ADF

Replies:

Comment: I changed my selection It is truly ADE. I read the link provided by rahulw230

Comment: A = VPC endpoints are well know safety mechanism in SM so traffic doesn't leave AWS B = service control policy can restrict access at org level D = Network isolation limits training model access only to S3

Comment: To control data egress from SageMaker, the ML engineer can use the following mechanisms: Connect to SageMaker by using a VPC interface endpoint powered by AWS PrivateLink. This allows the ML engineer to access SageMaker services and resources without exposing the traffic to the public internet. This reduces the risk of data leakage and unauthorized access1 Enable network isolation for training jobs and models.

Comment: Question is wrong A, B, E and D are all valid to a point.

Comment: The more I see it, the more likely I will go with ABD, the only answers than address the data egress issue

Comment: For those who are sure that is E, please explain how you can use pre-signed urls to restrict IP's, from my understanding it is a time based access to your S3 objects, you can policies to control access, like SCP (Service Control Policy), Isolation is definitely one option so that leaves F (Encrypting in transit and Encrypting objects) as the only possible solution as BDF

Comment: A and D are for sure. The challenge between E and F. E restrict access to the notebook hence indirectly control who access it and can access data but encrypting the data is more direct way to protect the egress of the

data. hence leaning more towards F

Comment: Not F because the que4stion is "to control data egress". F (encryption) is not egress control.

Comment: A, D, F are the mechanisms that the ML engineer can use to control data egress from SageMaker. B, C, and E do not directly control data egress from SageMaker. SCPs restrict access to AWS services, disabling root access on the SageMaker notebook instances improves security, and restricting notebook presigned URLs to specific IPs used by the company adds another layer of security, but none of these mechanisms control data egress from SageMaker.

Comment: https://aws.amazon.com/blogs/machine-learning/millennium-management-secure-machine-learning-using-amazon-sagemaker/

Comment: Are the correct

Comment: According to the subheadings in this case study: https://aws.amazon.com/blogs/machine-learning/millennium-management-secure-machine-learning-using-amazon-sagemaker/ The relevant options are: Controlling data egress: Enforcing deployment in VPC: This does not require a VPN to be enabled Enforcing network isolation Restricting notebook pre-signed URLs to IPs Disabling internet access Enforcing encryption: Enforcing job encryption: sagemaker:VolumeKmsKey

Comment: I think option E "Restrict notebook presigned URLs to specific IPs used by the company. " is ambiguous. The official blog post at https://aws.amazon.com/blogs/machine-learning/millennium-management-secure-machine-learning-using-amazon-sagemaker/ does indicate that the objective can be achieved by limiting access to the notebook by blocking some IPs. The point is not presigned URLs, but blocking IPs. The option E however puts more emphasis on presigned URLs.

Comment: Per thr following link, it is ADE - see section on controlling data egress - https://aws.amazon.com/blogs/machine-learning/millennium-management-secure-machine-learning-using-amazon-sagemaker/

Discussion for Question 133

Link: https://www.examtopics.com/discussions/amazon/view/76354-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 7 votes

Discussion

Comment: C is the correct answer. You definitely need Amazon Textract service which eliminate options B & D. Between A & C - Comprehend will quicker.

Comment: Textract and Comprehend will do the job

Comment: Keywords 'Amazon Textract' and 'Amazon Comprehend'

Comment: C indeed due to least effort

Comment: C is correct

Comment: I think C

Comment: I go for C

Comment: C is the best answer, textract is to extract data from documents and comprehend to understand the filling, objective or origin of a file.

Comment: C is correct, you can extract Entity information easily with Comprehend. https://aws.amazon.com/comprehend/features/

Discussion for Question 134

Link: https://www.examtopics.com/discussions/amazon/view/74974-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 24 votes
- A: 22 votes

Discussion

Comment: A; IAM assigned to SageMaker Notebook instance can be passed to other SageMaker jobs like training; processing, automl, etc.,

Replies:

> **Comment:** Why should the IAM permission be assigned to create S3, when the data is already stored in S3? It only require permission to read and write data in S3. I believe A is incorrect.

Comment: based on answers from here

Comment: Option A: The IAM role is created with the necessary permissions to create Amazon SageMaker Processing jobs, read and write data to the relevant S3 bucket, and access the KMS CMKs and ECR container image. The IAM role is attached to the SageMaker notebook instance, which allows the notebook to assume the role and create the Amazon SageMaker Processing job with the necessary permissions. The Amazon SageMaker Processing job is created from the notebook, which ensures that the job has the necessary permissions to read data from S3, process it, and upload it back to the same S3 bucket. Option B is close, but it's not entirely correct. It mentions creating an IAM role with permissions to create Amazon SageMaker Processing jobs, but it doesn't mention attaching the role to the SageMaker notebook instance. This is a crucial step, as it allows the notebook to assume the role and create the Amazon SageMaker Processing job with the necessary permissions.

Comment: The correct solution for granting permissions for data preprocessing is to use the following steps:Create an IAM role that has permissions to create Amazon SageMaker Processing jobs. Attach therole to the SageMaker notebook instance. This role allows the ML specialist to run Processing jobsfrom the notebook code1 Create an Amazon SageMaker Processing job with an IAM role that hasread and write permissions to the relevant S3 bucket, and appropriate KMS and ECR permissions.This role allows the Processing job to access the data in the encrypted S3 bucket, decrypt it with theKMS CMK, and pull the container image from ECR23 The other options are incorrect because theyeither miss some permissions or use unnecessary steps. For example

Comment: Least priv.

Comment: A. Create an IAM role with S3, KMS, ECR permissions and SageMaker Processing job creation permissions. Attach it to the SageMaker notebook instance: This option seems comprehensive as it includes all necessary permissions. However, attaching this role directly to the SageMaker notebook instance would not be sufficient for the Processing job itself. The Processing job needs its own role with appropriate permissions. B. Create two IAM roles: one for the SageMaker notebook with permissions to create Processing jobs, and another for the Processing job itself with S3, KMS, and ECR permissions: This option is more aligned with best practices. The notebook instance and the Processing job have different roles tailored to their specific needs. This separation ensures that each service has only the permissions necessary for its operation, following the principle of least privilege.

Comment: The processing job may not run on the notebook instance. AWS will provide resources to execute the job. So A is wrong. B.

Comment: If we follow the principle of Least Privillege, B is correct. The notebook instance does not need access to S3 and KMS given that it is only needed to trigger the processsing Job.

Comment: Not A b/c it does not indicate perms given to the Job via IAM role. => I went with B.

Comment: My answer is B. The notebook instance doesn't need access to S3 and ECR. This access is needed for Processing Job only. And as a best practice of least privilege I'll choose B

Comment: where permissions are granted to the SageMaker Processing job itself and not to the notebook instance. This approach offers better security and control over permissions, making it the preferred choice for running SageMaker Processing jobs with the required access to S3, KMS, and ECR. (Follows the principle of least privilege and have more control over permissions.

Comment: It says "Amazon SageMaker Processing job that is triggered from code in an Amazon SageMaker notebook." - so A or C. There is no need to create an S3 endpoint (C), that is only to allow traffic over the internet. So A.

Comment: Confusing between A and B. Leaning to B The main difference between A and B is the IAM role that is attached to the SageMaker notebook instance. In A, the role has permissions to access the data, the container image, and the KMS CMK. In B, the role only has permissions to create SageMaker Processing jobs. This means that in A, the notebook instance can potentially access or modify the data or the image without using a Processing job, which is not desirable. In B, the notebook instance can only create Processing jobs, and the Processing jobs themselves have a separate IAM role that grants them access to the data, the image, and the KMS CMK. This way, the data and the image are only accessed by the Processing jobs, which are more secure and controlled than the notebook instance.

Comment: Letters C and D are wrong, as they bring VPC, something that is not mentioned in the problem. Letter A is correct, since Letter B asks for the creation of two different IAM roles.

Replies:

Comment: What is the problem with creating two different IAM roles?

Comment: Option A ensures that the role has the necessary permissions to access the required resources (S3, KMS, ECR) and that the notebook has the ability to create a processing job in SageMaker seamlessly. It also follows the principle of "least privilege" by granting only the necessary permissions to perform the task without exposing more access than required.

Comment: Probably A is simpler than B. Per https://docs.aws.amazon.com/sagemaker/latest/dg/security-iam-awsmanpol.html#security-iam-awsmanpol-AmazonSageMakerFullAccess One IAM Role can do everything.

Replies:

Comment: It's rarely a best prectice

Comment: B is least privilege, since the notebook only needs access to sagemaker processing and the processing instance needs access to S3, KMS and ECR.

Discussion for Question 135

Link: https://www.examtopics.com/discussions/amazon/view/74279-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 19 votes

Discussion

Comment: This is correct according to official documentation. https://docs.aws.amazon.com/sagemaker/latest/dg/nbi-software-updates.html

Comment: Amazon SageMaker periodically tests and releases software that is installed on notebook instances, such as Jupyter Notebook, security patches, AWS SDK updates, and so on. To ensure that you have the most recent software updates, you need to stop and restart your notebook instance, either in the SageMaker console or by calling StopNotebookInstance.

Comment: By stopping and restarting the SageMaker notebook instance, it will automatically apply the latest security and software updates provided by SageMaker. This process refreshes the underlying infrastructure, ensuring that the notebook instance is running with the most up-to-date software and security patches. It is a simple and effective way to comply with the security team's mandate for using the latest updates.

Comment: C per Developer Documentation https://gmoein.github.io/files/Amazon%20SageMaker.pdf Page44

Discussion for Question 136

Link: https://www.examtopics.com/discussions/amazon/view/74070-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 8 votes

Discussion

Comment: A Images passed to Amazon Rekognition API operations may be stored and used to improve the service unless you unless you have opted out by visiting the AI services opt-out policy page and following the process explained there https://docs.aws.amazon.com/rekognition/latest/dg/security-data-encryption.html

Replies:

Comment: So the answer is A

Comment: https://docs.aws.amazon.com/organizations/latest/userguide/orgs_manage_policies_ai-opt-out.html

Comment: Yes, but server-side encryption doesn't protect at transit. Only client-side encryption can do it.

Replies:

Comment: Ok, I see "encryption in transit" mean HTTPS: Amazon Rekognition API endpoints only support secure connections over HTTPS. All communication is encrypted with Transport Layer Security (TLS).

Comment: Absolutely A. Rekognition API endpoints only support secure connections over HTTPS and all communication is encrypted in transit with TLS

Comment: B is correct one

Comment: client-side encryption requires you to manage the encryption and decryption of your data yourself and is an overkill. Will go with Server side encryption. Recognition already encrypts data in transit

Comment: B https://docs.aws.amazon.com/rekognition/latest/dg/collections.html You can opt-out of AI data usage of aws through organizations settings.

Comment: Option A is correct

Comment: Also, when the images are used with Amazon Rekognition. they need to be encrypted in transit A server-site encryption doesn't encrypt images in transit. onyly when they are already uploaded to the S3. Only client-side encryption can encrypt the images before they are moving to AWS cloud.

Replies:

Comment: You forgot about removing the possibility of Rekognition training.

Replies:

Comment: client side encryption means the key is stored on the client side. AWS has no key, how can they train?

Comment: According to Rekognition FAQs, You may opt out of having your image and video inputs used to improve or develop the quality of Amazon Rekognition and other Amazon machine-learning/artificial-intelligence technologies by using an AWS Organizations opt-out policy. https://aws.amazon.com/rekognition/faqs/

Comment: how is it A???

Discussion for Question 137

Link: https://www.examtopics.com/discussions/amazon/view/74926-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- D: 27 votes

Discussion

Comment: Answer is D: A is not correct since restaurant has limited bandwidth B is not correct since cannot enable Rekognition service on DeepLens C is not correct the same reason as A

Replies:

Comment: B is correct with Rekognition integrated with Deeplens and no extra configuration needed. (https://aws.amazon.com/blogs/machine-learning/building-a-smart-garage-door-opener-with-aws-deeplens-and-amazon-rekognition/)

Replies:

Comment: In this blog, rekognition service is not running on Deeplens. It said "After you deploy the sample object detection project into AWS DeepLens, you need to change the inference (edge) Lambda function to upload image frames to Amazon S3. ... and Rekognition would do its work from the Cloud to image frames on S3."... It would still consume lots of bandwidth. So B is NOT correct.

Comment: I also agree with D. B is incorrect due to that it's no need to do "person is recognized". It just needs to count the number of people.

Comment: AWS will not recommend to use Deeplense in production. From https://aws.amazon.com/deeplens/device-terms-of-use/

Replies:

Comment: aws doesn't allow use in production but in evaluation. can we accept counting number of people as an evaluation?

Comment: https://aws.amazon.com/deeplens/device-terms-of-use/

Comment: The best solution for building a line-counting application for use in a quick-service restaurant is to usethe following steps:Build a custom model in Amazon SageMaker to recognize the number of people in an image. AmazonSageMaker is a fully managed service that provides tools and workflows for building, training, anddeploying machine learning models. A custom model can be tailored to the specific use case of linecounting and achieve higher accuracy than a generic model1 Deploy AWS DeepLens cameras in therestaurant to capture video

Comment: B. AWS DeepLens with Local Amazon Rekognition and AWS Lambda: AWS DeepLens is designed for local processing and can run models at the edge (i.e., on the device itself). This setup would enable local analysis of the video feed without the need to stream the video to the cloud, thus conserving bandwidth. Amazon Rekognition and Lambda can then be used to analyze the footage and send notifications. This option aligns well with the bandwidth limitations. D. Custom Model on AWS DeepLens with AWS Lambda: Deploying a custom model built in SageMaker to AWS DeepLens allows for local processing of video data. This option also avoids the bandwidth issue by processing data on the device. However, developing a custom model might be more complex than using pre-built solutions like Amazon Rekognition.

Comment: Rekognition is a managed service. It uses API's and can't be deployed locally on devices. What we need here is local inference on the camera. AWS DeepLens comes pre-installed with a high performance, efficient, optimized inference engine for deep learning using Apache MXNet.

Comment: I would go with A, As DeepLenght is not for production workloads, we are left with A or C. A requires less effort.

Comment: B : https://aws.amazon.com/ko/blogs/machine-learning/building-a-smart-garage-door-opener-with-aws-deeplens-and-amazon-rekognition/

Comment: Based on the requirements, the best solution is option D. This option uses AWS DeepLens cameras to capture video and process it locally on the device, without sending any video streams to external services. This reduces the bandwidth consumption and avoids impacting other operations in the restaurant. The option also uses a custom model built in Amazon SageMaker to recognize the number of people in an image, which can be more accurate and tailored to the specific use case than a generic face detection model. The option also deploys an AWS Lambda function to the cameras to use the model to count people and send an Amazon Simple Notification Service (Amazon SNS) notification if the line is too long.

Comment: it is D. "The restaurant locations have limited bandwidth for connections to external services and cannot accommodate multiple video streams without impacting other operations." So, using Amazon Kinesis Video Streams is not a solution here. Ok, DeepLens dissapears in 2024...but this questions is for 2022... In the real world, the restaurant would buy good signal internet and use answer C, which is better solution.

Comment: C is Answer

Comment: AWS DeepLens will reach end-of-ilfe in 31/01/2024 so, I don't think this question will even appear in the exam.

Comment: Deeplens + lambda + model inference

Comment: After giving this some thought, I am thinking D. Tricky, my initial answer was C. But D is a better solution - given DeepLens and counting the number of people.

Comment: https://aws.amazon.com/ko/blogs/machine-learning/building-a-smart-garage-door-opener-with-aws-deeplens-and-amazon-rekognition/

Replies:

Comment: According to this link, Answer should be D, because we can directly deploy model in Deep lense to count the number of people instead a use of rekognition.

Comment: https://aws.amazon.com/blogs/machine-learning/optimize-workforce-in-your-store-using-amazon-rekognition/ B

Comment: B is the most suitable answer. A and C can be ignored because of requirement to use Amazon Video Streams which will not go well with low internet bandwidth. DeepLens is already compatible with Rekognition so better to use it rather than creating a custom model on SageMaker.

Replies:

Comment: https://aws.amazon.com/deeplens/community-projects/Customer_Counter/

Comment: B is possible. not stream only send detected. https://aws.amazon.com/ko/blogs/machine-learning/building-a-smart-garage-door-opener-with-aws-deeplens-and-amazon-rekognition/

Discussion for Question 138

Link: https://www.examtopics.com/discussions/amazon/view/74280-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- D: 22 votes

Discussion

Comment: I believe this is a problem to do with scaling out (increasing the number of instances), cooldown period should be increased. https://docs.aws.amazon.com/autoscaling/ec2/userguide/Cooldown.html

Comment: https://aws.amazon.com/blogs/machine-learning/configuring-autoscaling-inference-endpoints-in-amazon-sagemaker/

Comment: Option D

Comment: The issue is related to scaling out, specifically the fact that new instances are being launched before the existing ones are ready. To address this issue, the ML team could consider increasing the minimum number of instances, reducing the target value for CPU utilization, or increasing the warm-up time for the instances. These actions can help to ensure that new instances are not launched until the existing ones have reached a stable state, which can prevent performance issues and ensure the reliability of the service.

Replies:

Comment: Option D, which suggests increasing the cooldown period for the scale-out activity, could potentially help to address this issue by ensuring that the new instances are not launched too quickly. Option A, which suggests decreasing the cooldown period for the scale-in activity and increasing the maximum capacity of instances, is not an appropriate solution to the problem described. Decreasing the cooldown period for scale-in activity would result in instances being terminated too quickly, and increasing the maximum capacity of instances would not necessarily prevent new instances from being launched too quickly.

Comment: Agreed with D. should be increased not decreased

Comment: Answer is "D"

Comment: Definitely D.

Discussion for Question 139

Link: https://www.examtopics.com/discussions/amazon/view/74921-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 32 votes
- A: 22 votes

Discussion

Comment: https://docs.aws.amazon.com/sagemaker/latest/dg/model-ab-testing.html

Replies:

Comment: after reviewing it maybe C not A

Comment: https://sagemaker-examples.readthedocs.io/en/latest/sagemaker_endpoints/a_b_testing/a_b_testing.html Should be A

Comment: Answer is C, hosting two models under single endpoint has less operational overheads than two hosting endpoints

Comment: The Answer is A. The question says "Developers want to introduce a new version of the model for a limited number of users who subscribed to a..." In order to introduce a new production version with least overhead you have to create a production variant by using CreateEndpointConfig operation and set the InitialVariantWeight to 0. You then specify the TargetVariant parameter for InvokeEndpoint calls for users who subscribed to the preview feature and gradually update the weight. https://docs.aws.amazon.com/sagemaker/latest/dg/model-ab-testing.html preview feature of the app

Comment: -CreateEndPointConfig with initial weight to 0 prohibits any traffic to new variant - TargetVariant Parameter in the endpoint calls made by selected users ensures new variant be used - Change of InitialWeight causes gradual release of new variant

Comment: Obviously C

Comment: https://docs.aws.amazon.com/sagemaker/latest/dg/deployment-best-practices.html You can modify an endpoint without taking models that are already deployed into production out of service. For example, you can add new model variants, update the ML Compute instance configurations of existing model variants, or change the distribution of traffic among model variants. To modify an endpoint, you provide a new endpoint configuration. SageMaker implements the changes without any downtime. For more information see, UpdateEndpoint and UpdateEndpointWeightsAndCapacities. According to this doc, new variants can be deployed with UpdateEndpoint, and weights can be updated with UpdateEndpointWeightsAndCapacities. Though for using UpdateEndpoint we need to create an endpoint config. I will go with C

Comment: The company can implement the testing model with the least amount of operational overhead by using Option A. The developers can update the ProductionVariant data type with the new version of the model by using the CreateEndpointConfig operation with the InitialVariantWeight parameter set to 0. They can specify the TargetVariant parameter for InvokeEndpoint calls for users who subscribed to the preview feature. When the new version of the model is ready for release, they can gradually increase InitialVariantWeight until all users have the updated version

Comment: The best option for the company to implement the testing model with the least amount of operational overhead is option C. Option C uses the SageMaker feature of production variants, which allows the company to test multiple models on a single endpoint and control the traffic distribution between them. By setting the DesiredWeight parameter to 0 for the new version of the model, the company can ensure that only users who subscribed to the preview feature will invoke the new version by specifying the TargetVariant parameter. When the new version of the model is ready for release, the company can gradually increase the DesiredWeight parameter until all users have the updated version. This option minimizes the operational overhead by avoiding the need to create and manage additional endpoints, load balancers, or DNS records.

Comment: C is correct. The existing model will be updated using parameter DesiredWeightAndCapacity for new production variant and lead to less operational effort.

Comment: This one is tricky, but I think it is testing the difference between UpdateEndpointWeightsAndCapacities and ProductionVariant UpdateEndpointWeightsAndCapacities: Updates variant weight of one or more variants associated with an existing endpoint, or capacity of one variant associated with an existing endpoint https://docs.aws.amazon.com/sagemaker/latest/APIReference/API_UpdateEndpointWeightsAndCapacities.html ProductionVariant: Identifies a model that you want to host and the resources chosen to deploy for hosting it. If you are deploying multiple models, tell SageMaker how to distribute traffic among the models by specifying variant weights. https://docs.aws.amazon.com/sagemaker/latest/APIReference/API_ProductionVariant.html So it must be A, because the variant must exist before it is updated This link gave me confidence to choose A https://docs.aws.amazon.com/sagemaker/latest/dg/model-ab-testing.html

Comment: I agree with C.

Comment: Please see step 4: https://docs.aws.amazon.com/sagemaker/latest/dg/model-ab-testing.html & in option A it's mentioned that we set initial_weight to 0 which isn't true as the value should be 1.

Comment: I did not found the InitialVariantWeight, only DesiredWeight, therefore is C: https://docs.aws.amazon.com/sagemaker/latest/APIReference/API_DesiredWeightAndCapacity.html

Comment: https://docs.aws.amazon.com/sagemaker/latest/dg/model-ab-testing.html Step 4: Increase traffic to the best model Now that we have determined that Variant2 performs better than Variant1, we shift more traffic to it. We can continue to use TargetVariant to invoke a specific model variant, but a simpler approach is to update the weights assigned to each variant by calling UpdateEndpointWeightsAndCapacities.

Replies:

Comment: Update should be the correct action to this change.

Discussion for Question 140

Link: https://www.examtopics.com/discussions/amazon/view/74071-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 13 votes

Discussion

Comment: B? because ip insights algorithm is unsupervised learning that don't need label

Comment: B; IP Insights for IP address anomaly detection

Comment: Amazon SageMaker IP Insights is an unsupervised learning algorithm that learns the usage patterns for IPv4 addresses. It is designed to capture associations between IPv4 addresses and various entities, such as user IDs or account numbers. You can use it to identify a user attempting to log into a web service from an anomalous IP address, for example. Or you can use it to identify an account that is attempting to create computing resources from an unusual IP address. Trained IP Insight models can be hosted at an endpoint for making real-time predictions or used for processing batch transforms.

Comment: Answer should be B

Comment: Agree with the comments beliw

Comment: B. https://docs.aws.amazon.com/sagemaker/latest/dg/ip-insights.html

---

Discussion for Question 141

Link: https://www.examtopics.com/discussions/amazon/view/74993-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 27 votes

Discussion

Comment: C; Glue can use FindMatches transformation to find duplicates

Replies:

Comment: It says "Each dataset contains records with a unique structure and format.", so C would not be correct.

Replies:

Comment: but thats exactly the use of FindMatches: The FindMatches transform enables you to identify duplicate or matching records in your dataset, even when the records do not have a common unique identifier and no fields match exactly

Comment: It is C as described in the tutorial - https://docs.aws.amazon.com/glue/latest/dg/machine-learning-transform-tutorial.html LakeFormation can also invoke a FindMatches algorithm (because it manages Data Ingestion through Glue), but we don't have a data lake in this example. No one would build a whole Data Lake - a process that takes days - only to find some matching records.

Comment: Option C

Comment: Lake Formation helps clean and prepare your data for analysis by providing a Machine Learning (ML) Transform called FindMatches for deduplication and finding matching records. For example, use FindMatches to find duplicate records in your database of restaurants, such as when one record lists "Joe's Pizza" at "121 Main St." and another shows "Joseph's Pizzeria" at "121 Main." You don't need to know anything about ML to do this. FindMatches will simply ask you to label sets of records as either "matching" or "not matching." The system will then learn your criteria for calling a pair of records a match and will build an ML Transform that you can use to find duplicate records within a database or matching records across two databases. https://aws.amazon.com/lake-formation/features/

Comment: AWS Lake Formation FindMatches is a new machine learning (ML) transform that enables you to match records across different datasets as well as identify and remove duplicate records, with little to no human intervention Ans is D

Replies:

> **Comment:** Thing is, FindMatches is not a custom transformation in LakeFormation. And LakeFormation transforms are actually Glue jobs

Comment: D is correct

Discussion for Question 142

Link: https://www.examtopics.com/discussions/amazon/view/74388-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 22 votes

Discussion

Comment: B appears to be correct according to the official source. https://docs.aws.amazon.com/sagemaker/latest/dg/notebook-interface-endpoint.html#notebook-private-link-restrict

Comment: Its A. This solutions works for all users, no more configurations needed.

Comment: Going with B because - underlying notebook instances are managed by aws and can't apply security groups - updating IAM policy only restricts connection only from VPC endpoints

Comment: The issue is that the notebook instances' security group allows inbound traffic from any source IP address, which means that anyone with the authorized URL can access the notebook instances over the internet. To fix this issue, the data science team should modify the security group to allow traffic only from the CIDR ranges of the VPC, which are the IP addresses assigned to the resources within the VPC. This way, only the VPC interface endpoints and the resources within the VPC can communicate with the notebook instances. The data science team should apply this security group to all of the notebook instances' VPC interfaces, which are the network interfaces that connect the notebook instances to the VPC.

Comment: B. notebook instances are controlled by AWS service accounts and hence no access to those instances

Comment: A. Modify the notebook instances' security group: This approach involves adjusting the security group settings to only allow traffic from the VPC's CIDR ranges. By applying this security group to all of the notebook instances' VPC interfaces, it ensures that only traffic originating from within the VPC can access the notebook instances. This is a viable solution because it directly restricts access based on the source of the traffic. B. Create an IAM policy for VPC endpoint access: This solution involves crafting an IAM policy that restricts certain SageMaker actions to only the VPC endpoints. However, this approach might not fully address the issue of external access to the notebook instances themselves. It's more about controlling who can create or describe notebook instances, rather than restricting network access.

Replies:

> **Comment:** BUT according to here, it should be A: https://docs.aws.amazon.com/sagemaker/latest/dg/notebook-interface-endpoint.html
>
> ### Replies:
>
> > **Comment:** should be B*

Comment: B is talking about a policy to allow. It doesn't ban anything, it's only about allow.... So the answer can't be B. A

Comment: A. NO - it is not possible the security group of the instances, they are managed by SageMaker and will not appear in the console B. YES - https://docs.aws.amazon.com/sagemaker/latest/dg/notebook-interface-endpoint.html#notebook-private-link-restrict C. NO - subnets cannot be converted from public to private D. NO - ACL are for the notebooks, not the network

Comment: Based on my search, the answer is A. Modifying the notebook instances' security group to allow traffic only from the CIDR ranges of the VPC is a way to restrict access to anyone outside the VPC1. Amazon VPC interface endpoints enable you to privately connect your VPC to supported AWS services and VPC endpoint services powered by AWS PrivateLink without requiring an internet gateway, NAT device, VPN connection, or AWS Direct Connect connection2. However, they do not prevent users from accessing the notebook instances using presigned URLs3. Therefore, options B, C and D are not correct.

Replies:

> **Comment:** guys the right answer is B according to this reference: https://docs.aws.amazon.com/sagemaker/latest/dg/notebook-interface-endpoint.html#notebook-private-link-restrict To restrict access to only connections made from within your VPC, create an AWS Identity and Access Management policy that restricts access to only calls that come from within your VPC. Then add that policy to every AWS Identity and Access Management user, group, or role used to access the notebook instance.

Comment: This question may be old based on this https://aws.amazon.com/blogs/machine-learning/customize-your-amazon-sagemaker-notebook-instances-with-lifecycle-configurations-and-the-option-to-disable-internet-access/ but you can still remove all other allowed access and just add the VPC ciders to the SGs as there is an explicit Deny for anything not explicitly allowed.

Comment: Option B creates an IAM policy that allows the sagemaker:CreatePresignedNotebookInstanceUrl and sagemaker:DescribeNotebookInstance actions from only the VPC endpoints. These actions are required to access the notebook instances through the Amazon SageMaker console or the AWS CLI1. By applying this policy to all IAM users, groups, and roles used to access the notebook instances, the data science team can ensure that only authorized users within the VPC can connect to the notebook instances across the internet.

Comment: Modifying the notebook instances' security group to allow traffic only from the CIDR ranges of the VPC ensures that only connections from within the VPC are permitted. This restricts access to the notebook instances from individuals outside the VPC, effectively securing the communication and preventing unauthorized access. On the other hand, Option B, creating an IAM policy for sagemaker:CreatePresignedNotebookInstanceUrl and sagemaker:DescribeNotebookInstance actions from VPC endpoints, does not address the issue of restricting direct internet access to the notebook instances. IAM policies manage permissions for AWS service actions and resources, but they do not control network-level access.

Comment: "...the data science team realizes that individuals outside the VPC can still connect to the notebook instances across the internet.." B states - "Create an IAM policy that allows the sagemaker:CreatePresignedNotebookInstanceUrl and sagemaker:DescribeNotebookInstance actions from only the VPC endpoints" Ok, so now the individuals outside the VPC can't create a CreatePresignedNotebookInstanceUrl or DescribeNotebookInstance, but does that stop them from StopNotebookInstance or DeleteNotebookInstance operations? For option A, we only allow traffic from the VPC

Comment: The problem about a is that "You can specify allow rules, but not deny rules." https://docs.aws.amazon.com/vpc/latest/userguide/security-group-rules.html#security-group-rule-characteristics Therefore, you cannot restrict the unauthorized access

Comment: Should be security group thing

Comment: The answer should be A according to this source "Select the Subnet and Security group(s) as part of the VPC setting. To disable direct internet access, under Direct Internet access, simply choose Disable – use VPC only" https://aws.amazon.com/blogs/machine-learning/customize-your-amazon-sagemaker-notebook-instances-with-lifecycle-configurations-and-the-option-to-disable-internet-access/

Comment: Answer is A, as per the following link https://aws.amazon.com/blogs/machine-learning/customize-your-amazon-sagemaker-notebook-instances-with-lifecycle-configurations-and-the-option-to-disable-internet-access/

Discussion for Question 143

Link: https://www.examtopics.com/discussions/amazon/view/76355-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 6 votes

Discussion

Comment: C is correct answer. Straight forward to use KMS.

Comment: "The company wants AWS to maintain the root of trust for the master keys" The reason A is wrong. So C

Comment: option C

Comment: Using customer managed keys in AWS KMS will allow the company to maintain the root of trust for the master keys, and AWS KMS will log key usage. This ensures that the encryption keys used to encrypt the ML data volumes and model artifacts are properly managed and secured. Additionally, using customer managed keys allows the company to have greater control over the encryption process.

Replies:

Comment: "AWS Security Token Service (AWS STS) to create temporary tokens" - AWS STS also using KMS keys.

Comment: https://docs.aws.amazon.com/kms/latest/developerguide/security-logging-monitoring.html

---

Discussion for Question 144

Link: https://www.examtopics.com/discussions/amazon/view/74392-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- D: 22 votes

Discussion

Comment: D. AWS Glue can connect with DynamoDB and join both data sets together via Glue Studio. Requiring minimal overheads

Comment: D. AWS Glue can connect with DynamoDB and join both data sets together via Glue Studio. Requiring minimal overheads

Comment: Option D with AWS Glue crawlers and ETL job provides a straightforward and efficient way to merge the data from DynamoDB and Amazon S3 into a format suitable for training the Amazon SageMaker model with minimal administrative overhead.

Comment: D. https://aws.amazon.com/blogs/big-data/accelerate-amazon-dynamodb-data-access-in-aws-glue-jobs-using-the-new-aws-glue-dynamodb-elt-connector/

Comment: 12-sep exam

Discussion for Question 145

Link: https://www.examtopics.com/discussions/amazon/view/74394-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 29 votes

Discussion

Comment: A should be the answer

Replies:

Comment: it's a sentiment analysis problem => comprehend

Comment: BlazingText can also do supervised text classification

Comment: Built-in BlazingText model using Word2Vec mode in Amazon SageMaker would likely be quicker to set up compared to using Amazon Comprehend for this specific use case. Since the problem statement mentions that the review data is already labeled with the correct durability result, preparing the training data should be relatively straightforward. Additionally, as a built-in algorithm, BlazingText is optimized and pre-configured for text classification tasks, reducing the need for extensive customization and configuration compared to using Amazon Comprehend for this specific use case. It's important to note that while BlazingText may be quicker to set up for this particular task, Amazon Comprehend offers a broader range of NLP capabilities and may be more suitable for other NLP tasks or scenarios where more customization and flexibility are required. However, given the time constraint of 2 days and the specific requirement of identifying product durability concerns from reviews, training a built-in BlazingText model using Word2Vec mode in Amazon SageMaker is likely to be the more direct and quicker approach to get a working solution set up and running.

Comment: Given the time constraint of 2 days and the need for a quick solution, the most direct approach would be to choose an option that provides a ready-to-use solution without the need for extensive customization or training. Among the given options, the most direct approach would be: C. Train a built-in BlazingText model using Word2Vec mode in Amazon SageMaker. This option allows you to leverage a pre-built model (BlazingText) that is optimized for text classification tasks. Word2Vec mode is suitable for analyzing text data and can quickly provide insights into sentiment or, in this case, concerns over product durability. This approach minimizes the need for extensive data preprocessing and model tuning, allowing you to focus on training and deploying the model within the given timeframe.

Comment: Using a existing model to do the task in 2 days. A

Comment: I would say A

Comment: A. YES - Amazon Comprehend with multi-class mode and Augmented manifest file B. NO - Gluon is for timeseries C. NO - still a lot of work after generating embedding D. NO - seq2seq is to generate text, we want to classify

Comment: To solve the problem in 2 days, and dealing with sentiment analysis so A will be the right answer using the comprehend AWS Comprehend is a natural language processing (NLP) service that uses machine learning to discover insights from text. It provides a range of functionalities, including detecting language and sentiment, extracting named entities and key phrases, and tagging parts of speech5. AWS Comprehend can automatically break down concepts like entities, phrases, and syntax in a document, which is particularly helpful for identifying events, organizations, persons, or products referenced in a document

Comment: The most direct approach to solve this problem within 2 days is option A, train a custom classifier by using Amazon Comprehend. By doing so, you can use Amazon Comprehend, a natural language processing (NLP) service that uses machine learning to find insights and relationships in text, to create a custom classifier that can identify reviews expressing concerns over product durability. You can use the labeled reviews as your training data and specify the durability result as the class label. Amazon Comprehend will automatically preprocess the text, extract features, and train the classifier for you. You can also use Amazon Comprehend to evaluate the performance of your classifier and deploy it as an endpoint. This way, you can train a model to solve this problem within 2 days without requiring much coding or infrastructure management.

Comment: A: You can customize Amazon Comprehend for your specific requirements without the skillset required to build machine learning-based NLP solutions. Using automatic machine learning, or AutoML, Comprehend Custom builds customized NLP models on your behalf, using training data that you provide.

Comment: Comprehend can do Custom Classification

Comment: Comprehend can do Sentiment Analysis

Comment: The answer is C, because of the amount of data, and the time constraint. C is the most efficient solution. Conventionally A would be the right answer, but given the time constraint the answer is C.

Comment: I would say blaze text. Cuz comprehend needs custom code, so we have only 2 days.

Comment: https://docs.aws.amazon.com/comprehend/latest/dg/how-document-classification.html

Comment: If the problem needs to be solved in 2 days I would avoid going with any customised solution which would eliminate A and B. As the data is labelled already we don't need an unsupervised algorithm therefore eliminating C. Which leaves us with D

Replies:

Comment: its exactly the opposite, because its needs to be ready in 2 day I would use Comprehend ;) You don't need to write code, you have the data already available, so its faster then D

Discussion for Question 146

Link: https://www.examtopics.com/discussions/amazon/view/74078-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- D: 17 votes

Discussion

Comment: D is the answer. The unrecognized words are synonyms for "comedy", so they should be added as synonyms under the comedy slot type see the excerpt: "For each intent, you can specify parameters that indicate the information that the intent needs to fulfill the user's request. These parameters, or slots, have a type. A slot type is a list of values that Amazon Lex uses to train the machine learning model to recognize values for a slot. For example, you can define a slot type called "Genres." Each value in the slot type is the name of a genre, "comedy," "adventure," "documentary," etc. You can define a synonym for a slot type value. For example, you can define the synonyms "funny" and "humorous" for the value "comedy."" https://docs.aws.amazon.com/lex/latest/dg/howitworks-custom-slots.html

Comment: D? can not be C.Amazon Lex doesn't support the AMAZON.LITERAL or the AMAZON.SearchQuery built-in slot types. https://docs.aws.amazon.com/lex/latest/dg/howitworks-builtins-slots.html

Replies:

Comment: https://docs.aws.amazon.com/lex/latest/dg/howitworks-custom-slots.html

Comment: D is the answer.

Comment: The best way to fix the problem is option D, add the unrecognized words as synonyms in the custom slot type. By doing so, you can map different words that have the same meaning to the same slot value, without changing the Lambda code or data in DynamoDB. For example, you can add “funny”, “fun”, and “humor” as synonyms for the slot value “comedy”. This way, Amazon Lex can understand the category spoken by users and pass it to the Lambda function that queries the DynamoDB table for a list of book titles. Option A, adding the unrecognized words in the enumeration values list as new values in the slot type, is not a good choice because it would create new slot values that do not match the existing categories in the DynamoDB table. For example, if you add “funny” as a new value in the slot type, Amazon Lex would pass it to the Lambda function, which would not find any book titles for that category in the DynamoDB table.

Comment: C is the answer AMAZON.SearchQuery As you think about what users are likely to ask, consider using a built-in or custom slot type to capture user input that is more predictable, and the AMAZON.SearchQuery slot type to capture less-predictable input that makes up the search query. The following example shows an intent schema for SearchIntent, which uses the AMAZON.SearchQuery slot type and also includes a CityList slot that uses the AMAZON.City slot type. Make sure that your skill uses no more than one AMAZON.SearchQuery slot per intent. The Amazon.SearchQuery slot type cannot be combined with another intent slot in sample utterances. Each sample utterance must include a carrier phrase. The exception is that you can omit the carrier phrase in slot samples. A carrier phrase is the word or words that are part of the utterance, but not the slot, such as "search for" or "find out".

Comment: The ML specialist should add the unrecognized words as synonyms in the custom slot type. This will allow Amazon Lex to understand the user's intent even if they use synonyms for the predefined slot values. By adding the synonyms, Amazon Lex will recognize them as variations of the predefined slot values and map them to the appropriate slot value. This approach can be a quick and effective way to improve the accuracy of the chatbot's understanding of user requests without having to change the Lambda code or the data in DynamoDB.

Comment: B is correct

Discussion for Question 147

Link: https://www.examtopics.com/discussions/amazon/view/74395-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- D: 21 votes

Discussion

Comment: D is correct according to official documentation. https://docs.aws.amazon.com/greengrass/v1/developerguide/ml-inference.html

Replies:

Comment: A-C: excluded out, Direct Connect is expensive

Comment: Option D eliminates the need for internet connection since the inference is done on the edge component, and the results are directly forwarded to the web service. This approach also reduces the need for larger instances and direct connect connections, thus being the most cost-effective solution.

Discussion for Question 148

Link: https://www.examtopics.com/discussions/amazon/view/75091-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 13 votes

Discussion

Comment: Agree with the Answer C. Attach the policy to the IAM roal associated with the notebook.

Comment: c is the right answer

Comment: Amazon SageMaker notebook ARN , I don't think there is such a thing. So A is not right . So C

Comment: C. Attach policy to IAM role associated with the notebook: This is a standard and recommended approach in AWS. By attaching a policy to the IAM role that the SageMaker notebook instance assumes, you can precisely control the notebook's access to the specific S3 bucket. This method follows the AWS best practice of using IAM roles for managing permissions and also allows for easier management and scalability. A. Add an S3 bucket policy: This approach involves modifying the S3 bucket policy to grant permissions directly to the SageMaker notebook instance's ARN. While this method can effectively grant access, it is less flexible and scalable compared to using IAM roles. It directly ties the bucket's access policy to a specific resource (the notebook instance), which might not be ideal for managing access in a larger environment.

Comment: The best way for the data scientist to securely access data stored in a specific Amazon S3 bucket from an Amazon SageMaker notebook instance is option C, attach the policy to the IAM role associated with the notebook that allows GetObject, PutObject, and ListBucket operations to the specific S3 bucket. By doing so, the data scientist can use IAM role-based access control to grant permissions to the notebook instance to access the S3 bucket without exposing any credentials or keys. The data scientist can also limit the scope of the permissions to only the necessary operations and resources, following the principle of least privilege.

Comment: Option A suggests adding an S3 bucket policy, but it is not the recommended way to grant permissions to specific IAM roles associated with SageMaker notebook instances. Bucket policies are generally used for granting cross-account access or public access, not for specifying access for specific IAM roles.

Comment: An IAM policy cannot attach to an ARN. An IAM policy can only attach to an IAM role or an IAM user. So the answer is C

Comment: A - we allow access to specific notebook. AIM role policy can be global and related to all user notebooks.

Replies:

Comment: On the other hand, in C they state "specific S3 bucket" and in the A - only "an S3 bucket". Maybe in A they add global policy to allow access to all S3 buckets?

Comment: AC are both correct answer, but A is better than C, mostly due to the limitation of IAM policy. IAM policies: The maximum size of an IAM policy document is 6,144 characters. You can attach up to 10 policies to an IAM user, role, or group.

Comment: Option C ensures that the notebook instance is granted permission to access the S3 bucket without the need to provide credentials. Option A is incorrect because it suggests adding a bucket policy that grants permission to a specific IAM principal, which is less secure than granting permission to an IAM role.

Replies:

> **Comment:** I dont agree with this. Restrict bucket access only to limited principal is much secure than grant specific IAM prinicap. Restrict specific principal eliminate other visits, but grant specific IAM user permission does not exclude other visit.

Comment: 12-sep exam

Comment: C is correct

Comment: Quoting the book "Data Science on AWS": "Generally, we would use IAM identity-based policies if we need to define permissions for more than just S3, or if we have a number of S3 buckets, each with different permissions requirements. We might want to keep access control policies in the IAM environment. We would use S3 bucket policies if we need a simple way to grant cross-account access to our S3 environment without using IAM roles, or if we reach the size limit for our IAM policy. We might want to keep access control policies in the S3 environment." A would be the choice then.

Replies:

> **Comment:** Based on this logic indeed A would be better.
>
> **Comment:** I am not sure but in question we don't have cross-account situation?
>
> **Comment:** For A, only some operations are allowed, no specified users or roles have been granted this permission for these operations.

Comment: B is the answer

Replies:

> **Comment:** Only "securely access" is required, not encryption.

Comment: A - for me

Discussion for Question 149

Link: https://www.examtopics.com/discussions/amazon/view/74079-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- D: 18 votes

Discussion

Comment: D is the correct answer. Following from the previous comment. The company wants to minimize the infrastructure and data science resources needed to evaluate the messages. Therefore any custom services would be eliminated (A and B). Similarly DynamoDB would add complexity to the infrastructure there C is eliminated. leaving D

Comment: Option D is the right answer. Following are the key terms in question to notice, sentiment expressed in social media posts --> Comprehend configure alarms based on various thresholds --> CloudWatch (can send alerts without SNS) wants to minimize the infrastructure and data science resources --> AWS S3

Comment: The best services for the data science team to use to deliver this solution are option D, trigger an AWS Lambda function when social media posts are added to the S3 bucket, call Amazon Comprehend for each post to capture the sentiment in the message and record the sentiment in a custom Amazon CloudWatch metric and in S3, and use CloudWatch alarms to notify analysts of trends. By doing so, the data science team can use Amazon Comprehend, a natural language processing (NLP) service that uses machine learning to find insights and relationships in text, to evaluate the sentiment expressed in social media posts. Amazon Comprehend can detect positive, negative, neutral, or mixed sentiment from text input. The data science team can also use AWS Lambda, a service that lets you run code without provisioning or managing servers, to trigger a function when posts are added to the S3 bucket and call Amazon Comprehend for each post.

Comment: Amazingly D is possible - https://catalog.us-east-1.prod.workshops.aws/workshops/4faab440-8c3a-4527-bd11-0c88a6e6213c/en-US/30-build-the-application/400-send-sentiment-to-cloudwatch I was so sure of option C, because sending a sentiment to a custom CloudWatch metric just didn't make any sense. But you learn something new everyday.

Comment: This is a puzzling question, as both answers C and D miss essential steps: C is missing DynamoDB Streams to capture new records D is missing a notification mechanism like SNS, as CloudWatch Alarms alone can only be used as a trigger, but are not sufficient for notification I agree that A and B should be eliminated for requiring data science development

Comment: I also do agree that D is correct answer. In A, why we are adding extra dependency of Dynamo DB.

Comment: D, blazing text is not for sentiment analysis. The Amazon SageMaker BlazingText algorithm provides highly optimized implementations of the Word2vec and text classification algorithms. The Word2vec algorithm is useful for many downstream natural language processing (NLP) tasks, such as sentiment analysis, named entity recognition, machine translation, etc. Text classification is an important task for applications that perform web searches, information retrieval, ranking, and document classification.

Replies:

> **Comment:** BlazingText can do sentiment analysis: https://docs.aws.amazon.com/sagemaker/latest/dg/blazingtext.html

---

Discussion for Question 150

Link: https://www.examtopics.com/discussions/amazon/view/74996-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 21 votes

Discussion

Comment: B, SageMaker Model Debugger is used to generate SHAP values

Replies:

> **Comment:** https://aws.amazon.com/blogs/machine-learning/ml-explainability-with-amazon-sagemaker-debugger/

Comment: I believe C is the right answer, it is simpler and more accurate than B.

Replies:

> **Comment:** It will show only importance of features not their contribution to the final score

Comment: The best option is to use Amazon SageMaker Studio to rebuild the model and deploy it at an endpoint. Then, use Amazon SageMaker Model Monitor to store inferences and use the inferences to create Shapley values that help explain model behavior. Shapley values are a way of attributing the contribution of each feature to the model output. They can help the credit team understand why the model makes certain decisions and how the features affect the model outcomes. A chart that shows features and SHapley Additive exPlanations (SHAP) values can be created using the SHAP library in Python. This option is the most operationally efficient because it leverages the existing XGBoost training container and the built-in capabilities of Amazon SageMaker Model Monitor and SHAP library.

Comment: A. NO - too complicated to compute SHAP B. YES - Debugger supports built-in SHAP C. NO - too complicated to compute SHAP D. NO - too complicated to compute SHAP

Comment: Option B utilizes Amazon SageMaker Studio to build and train the model, and it also activates Amazon SageMaker Debugger, which allows calculating and collecting Shapley values. These Shapley values will help explain accurately why the model denies credit to certain customers. Generating a chart that displays the features and their SHAP values will provide a visual and clear explanation of the impact of each feature on the model's decisions, making it easier for the credit team with limited data science skills to understand.

Comment: Either A or B Sage Maker Monitor require no experience so A is preferred while B can provide more details but depend if require knowledge to use it.

Replies:

Comment: More towards B

Comment: SageMaker Model Monitor is a tool that helps monitor the quality of model predictions over time by analyzing data inputs and outputs during inference. It can detect and alert when data drift or concept drift occurs, and can identify features that are most responsible for the changes in model behavior. Model Monitor can be used to continuously monitor and improve model performance, and can be integrated with SageMaker endpoints or SageMaker Pipelines. SageMaker Debugger is a tool that helps debug machine learning models during training by analyzing the internal states of the model, such as weights and gradients, as well as the data inputs and outputs during training. It can detect and alert when common training issues occur, such as overfitting or underfitting, and can identify the root causes of these issues. Debugger can be used to improve model accuracy and convergence, and can be integrated with SageMaker training jobs.

Replies:

Comment: After reconsideration, it is actually B. https://aws.amazon.com/blogs/machine-learning/ml-explainability-with-amazon-sagemaker-debugger/

Comment: Debugger because we are in the context of "training data"

Replies:

Comment: There are so many explanations, but most of them are just superfacial, focusing on what service is related to SHAP. This is the only one really answer the difference between A and C. 1. Both SagaMaker Model Monitor and Debugger can explain model, can generate SHAP. so it should be either A or C 2. Monitor is about inference. After deploy the model, we may find some attributes start to contribute more to the model, contradict to the training dataset. This case we use SageMaker Model Monitor. But our problem is not about deploying, is still in training stage. We only want to figure out why some customer with specfic characteristics are more likely to get loan, in other words, certain feature contribute more to the prediction. It is C !!!! If you don't fully understant the question, stop explaining !!!

Replies:

Comment: not comparing between A and C, should be A and B

Comment: C is the straight forward and simpler.

Comment: Why not C? 'C' is the most easiest way to find out.!

Comment: This is debugger's work

Comment: Option A suggests using Amazon SageMaker Model Monitor to store inferences and create Shapley values that can help explain the model's behavior. This option can be more operationally efficient because it doesn't require the credit team to understand the complexities of Shapley values, and it doesn't necessarily slow down the model's inference time.

Replies:

Comment: After a review, I go with option B

Discussion for Question 151

Link: https://www.examtopics.com/discussions/amazon/view/75436-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- D: 28 votes

Discussion

Comment: I will go with A. Refer to link : https://aws.amazon.com/comprehend/features/

Replies:

Comment: whoever select A misunderstant "Custom classification", it is model for custom classificaiton, not submitting your own script!!!! and for the above reply with document, read document first.

Comment: Agree. A is my answer. 1. part of speech tagging : https://docs.aws.amazon.com/comprehend/latest/dg/API_PartOfSpeechTag.html 2. Key phas extraction https://docs.aws.amazon.com/comprehend/latest/dg/how-key-phrases.html 3. custum classification algorithm https://docs.aws.amazon.com/comprehend/latest/dg/how-document-classification.html

Comment: D is the answer. Using Apache MXNet rules out Comprehend from making the classification task

Replies:

Comment: any reference?

Replies:

Comment: "Automatically improve performance with optimized model training for popular frameworks like TensorFlow, PyTorch, and Apache MXNet." https://aws.amazon.com/cn/machine-learning/containers/

Comment: Amazon Comprehend is a natural language processing (NLP) service that can perform part-of-speech tagging and key phrase extraction tasks. AWS Deep Learning Containers are Docker images that are pre-installed with popular deep learning frameworks such as Apache MXNet. Amazon SageMaker is a fully managed service that can help build, train, and deploy machine learning models. Using Amazon Comprehend for the text preprocessing tasks and AWS Deep Learning Containers with Amazon SageMaker to build the custom classifier is the solution that can be built most quickly to meet the requirements. References: Amazon Comprehend AWS Deep Learning Container

Comment: The Custom classification in AWS Comprehend cannot choose algorithm, you cannot use your own algorithm in it. You only feed dataset to it. So A is wrong. The data science team want to use their own MXNET model, so D.

Comment: Will go with A

Comment: A is the most quickly solution.

Comment: We have to solve two NLP problems: part-of-speech tagging and key phase extraction. Note that the custom classifier already exists and has been trained! The question asks that it be done as quickly as possible, so the idea is to use a ready-made service. Letter A is wrong, as it uses another service compared to the already created model to classify. Letter B requires development and therefore would not be the fastest solution. Letter C is wrong for the same reason as Letter A, in addition it proposes an unsupervised service (LDA) for a supervised problem. Letter D is correct.

Comment: Therefore, option D is the most efficient solution for building a NLP application that meets the requirements of the data science team.

Comment: Quickest A

Replies:

Comment: Latest is A.

Comment: The other mxnet model is the key

Comment: option D is the most appropriate answer, given that the team has already written and trained a custom classification algorithm using Apache MXNet. Option D allows the team to use Amazon Comprehend for part-of-speech tagging and key phrase extraction, while also using AWS Deep Learning Containers with Amazon SageMaker to build and deploy the custom classifier.

Comment: D for me. Question says "The preprocessed text WILL be input to a custom classification algorithm that the data science team has already written and trained using Apache MXNet". So for some reason they want to use MXNet to do the classification, not Amazon Comprehend. So using MXNet for classification is a part of their requirement. How do we meet these requirements quickly? Well, use Amazon Comprehend for part-of-speech and key phrase tasks; and use container for the MXNet stuff.

Replies:

Comment: I had selected "A" in my first go, thanks for understanding the question. Although, comprehend does all three, since they have already built custom classification, we only need to provide solution for first two. D for me too.

Comment: The question did not make it clear whether the new solution has to use the custom model that the team built or not.

Comment: A for me

Comment: Agreed with A, Comprehend 3 functions

Comment: A for me. https://docs.aws.amazon.com/comprehend/latest/dg/how-document-classification.html

Comment: D for me

Discussion for Question 152

Link: https://www.examtopics.com/discussions/amazon/view/75321-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 18 votes

Discussion

Comment: Dimensions are too high. Use PCA

Comment: A should be the answer to avoid the curse of dimensionality

Comment: Easy choice. Always choose PCA for dim reduction

Comment: the best feature engineering strategy for the ML specialist to use with Amazon SageMaker is to apply dimensionality reduction by using the PCA algorithm.

Comment: Selected Answer: A Given that the dataset has 1,020 features and 200 of them are highly correlated, it is likely that the dataset suffers from multicollinearity. In such cases, dimensionality reduction techniques like principal component analysis (PCA) can be used to transform the data into a lower dimensional space without losing much information. Therefore, option A, "Apply dimensionality reduction by using the principal component analysis (PCA) algorithm" is the most appropriate feature engineering strategy for the ML specialist to use with Amazon SageMaker. This would help reduce the computational complexity of the model, improve model performance, and help to avoid overfitting.

Comment: A. Apply dimensionality reduction by using the principal component analysis (PCA) algorithm. Since the dataset has many features, and a significant number of them have high correlation scores, the model may suffer from the curse of dimensionality. To reduce the dimensionality of the dataset, the specialist can use a technique like PCA, which reduces the number of features while still retaining the maximum amount of information. PCA can help remove redundant features and improve the model's performance by reducing the chances of overfitting. Additionally, since there are no missing values and a small percentage of duplicate rows, no data cleaning techniques like anomaly detection or dropping the features are required. Concatenating features with high correlation scores is not an appropriate strategy since it may lead to collinearity issues.

Comment: A PCA: PCA is a linear dimensionality reduction technique (algorithm) that transforms a set of correlated variables (p) into a smaller k (k

Comment: Choosing C is answer by ExamTopics is completely laughable.

Comment: I think it's A.

Discussion for Question 153

Link: https://www.examtopics.com/discussions/amazon/view/75020-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- D: 9 votes

Discussion

Comment: D - over fitting problem.

Comment: The specialist should consider using some form of regularization to fix this issue. Regularization techniques such as dropout or L2 regularization can help prevent overfitting, which can occur when the model performs well on the training data but poorly on the validation data. Option A, a longer training time, might not necessarily fix the issue and could lead to overfitting if the model is already performing well on the training data. Option B, making the network larger, could also lead to overfitting and may not be necessary if the current network architecture is sufficient to perform the classification task. Option C, using a different optimizer, might not necessarily fix the issue and could lead to slower convergence or worse performance. Therefore, option D, using some form of regularization, is the most appropriate solution to consider in this situation.

Comment: some form of regularization

Comment: I wouldn't go with D since it doesn't seem an overfitting problem considering training accuracy is not so high. So the main problem here is to get an higher accuracy even on training set. I would go with A or B

Comment: A - IMO it's an underfitting problem, as training accuracy is not better than baseline error (human accuracy). Would consider B as well, but it may actually decrease accuracy.

Comment: typical overfitting problem

Comment: typical overfitting problem

Comment: C - It is not a overfitting problem as the training accuracy stands at 90%, which is at same level of human performance. That means the algorithm used is not optimized for this problem. So, some other algorithm should applied for this problem.

Comment: I'd go A. Regularization could not guarantee higher validation accuracy.

Comment: I believe answer is B , because clearly it is a overfiting problem , if we reduce complexity the accurate will reduce close to 80% ... But human works can reach up to 90% .

Replies:

Comment: I mean looks like a overfitting problem....

Discussion for Question 154

Link: https://www.examtopics.com/discussions/amazon/view/74991-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 14 votes
- B: 10 votes

Discussion

Comment: C and E B needs to build custom model

Replies:

Comment: The SageMaker seq2seq algorithm is a supervised learning algorithm. And it needs to train then translate. translate can directly use to translate from Spanish to English

Comment: The question did not say you cannot build a custom model. They have a ML specialist, so building a custom model shouldn't be a problem.

Comment: It asked 2 answers, but I can see only one answer. Please advise. Thanks!

Comment: C & E based on comments, but you are not allowed to select multiple choices.

Comment: C and E - Use translate so that text is in common language - In options with translate only Comprehend and NTM allow for topic modeling (C & E) Other options Blazingtext is for text classification, not topic modelling, LDA is requireres user specified topics and Lex is for conversational interfaces

Comment: C & E Option C (Amazon Translate and Amazon Comprehend): This is a strong combination. Amazon Translate can be used to translate Spanish comments into English, and then Amazon Comprehend, which supports topic modeling, can be used to identify the most discussed topics. Option E (Amazon Translate and Amazon SageMaker Neural Topic Model): This is also a viable combination. Amazon Translate would handle the translation of Spanish comments, and the Neural Topic Model (NTM) in Amazon SageMaker can then be used for topic modeling. NTM uses neural networks for topic discovery and is well-suited for analyzing large sets of text data.

Comment: B and E I dont think amazon comprehend can do topic modelling. LDA is used for topic modelling

Comment: BCE are all right. https://docs.amazonaws.cn/en_us/sagemaker/latest/dg/algos.html LDA and NTM are all topic modeling tools.

Comment: A. NO - BlazingText is word2vec, will not do topic modeling alone B. NO - Translate better than custom seq2seq C. NO - NTM better than LDA used by Comprehend D. NO - Lex is for chatbots E. YES

Comment: The right answers are C & E The other steps are not suitable because: A. The BlazingText algorithm is for word embeddings and text classification, not topic modeling. B. The LDA algorithm is an unsupervised learning algorithm that requires a user-specified number of topics. D. Amazon Lex is for building conversational interfaces, not extracting topics from content

Comment: Correct answer is BE

Comment: It has to be B + C .. for spanish to English use Translate. For Topics it has to be LDA.

Replies:

Comment: Sorry and NTM.. in that case, C is a winner for translation.. then pick E to be consistent.. final answer B + E.

Comment: For me: B - C - E are correct: it's solved translation + topic modelling. The question is not well construct from my POV.

Comment: C and E

Comment: B + E, I think

Replies:

Comment: Per https://docs.aws.amazon.com/comprehend/latest/dg/topic-modeling.html it's C +E.

Comment: C and E

Comment: C and E, The most common topic modeling algorithm is called Latent Dirichlet Allocation (LDA), Amazon Comprehend uses a Latent dirichlet allocation-based learning model to determine the topics in a set of documents. https://docs.aws.amazon.com/comprehend/latest/dg/topic-modeling.html. Despite its popularity and success, one limitation of this algorithm is that it makes an assumption that the distribution of words in a document follows a Dirichlet distribu- tion.The NTM algorithm relaxes this assumption and aims to learn a latent representa- tion without prior assumptions.

Comment: Option A suggests using an unsupervised learning algorithm, BlazingText, to find the topics independently of language. This can be a good approach if the content is in multiple languages and we don't want to translate them first. Option C suggests using Amazon Comprehend, which can detect the language and then perform topic modeling on the content in that language. If any comments are in Spanish, they will be translated to English using Amazon Translate before being fed to Amazon Comprehend. Together, these options cover both cases - comments in English and comments in Spanish, without having to translate everything.

Replies:

Comment: In Option A, BlazingText has two modes: word2vec, text classification. https://docs.aws.amazon.com/sagemaker/latest/dg/blazingtext.html Word2Vec is an unsupervised learning algorithm, but it is for word embeddings, so it is not appropriate to find the topics from corpus. Text classfication isn't also appropriate to find the topics, because there is no given specified topics. So Option A is not an answer

---

Discussion for Question 155

Link: https://www.examtopics.com/discussions/amazon/view/75415-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 8 votes

Discussion

Comment: I would go with B: https://docs.aws.amazon.com/sagemaker/latest/dg/model-monitor-create-baseline.html

Replies:

Comment: Agree, the answer is B. From the document, the violation file contains several checks and "The violations file is generated as the output of a MonitoringExecution" . https://docs.aws.amazon.com/sagemaker/latest/dg/model-monitor-interpreting-violations.html.

Comment: The baseline job computes baseline statistics and constraints for the new training set. By using this updated baseline, Model Monitor can better detect any drift or violations in the production traffic.

Comment: B. Run the Model Monitor baseline job again on the new training set: This is a key step after retraining the model. Since the model has been retrained with a new dataset, the baseline against which its predictions are compared should also be updated. Running the baseline job again on the new training set and configuring Model Monitor to use this new baseline will ensure that the monitoring is relevant to the current state of the model and the data it's processing. D. Retrain the model again with a combination of the original and new training sets: While retraining the model can be a good approach in some scenarios, there's no indication in this case that the issue lies with the model's performance itself. The issue seems to be with the Model Monitor's baseline not aligning with the current model.

Comment: https://docs.aws.amazon.com/sagemaker/latest/dg/model-monitor-interpreting-violations.html

Comment: running the Model Monitor baseline job again on the new training set and configuring Model Monitor to use the new baseline, is the most appropriate step to resolve the violations and ensure the SageMaker endpoint's performance is in line with expectations.

Comment: Running the Model Monitor baseline job again with the new training set and configuring Model Monitor to use the new baseline is a valid option to resolve the violations. By running the baseline job with the new training set, a new baseline is created, which can be used to compare with the new data to detect any drifts in the data distribution. Then, the updated baseline can be set as the new baseline for monitoring the endpoint. So, option B is also a valid solution to resolve the violations.

Discussion for Question 156

Link: https://www.examtopics.com/discussions/amazon/view/74397-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- AC: 35 votes

Discussion

Comment: AC would be my answer. As half the stores have only been open for 6 months, no seasonality would be captured. The aggregation of the daily also removes trends we see during the week which is also not great when we are looking for the daily predicated sales figure

Replies:

Comment: B - no reason to assume there is not enough variance D - missing data can be assumed to be 0, no need to ask for empty data E - no reason to ask for two years of data having one already

Comment: I would go for AD A : Many stores have been in business for < 6 months --> unable to capture seasonality D : Zero sales are also sales records and will result in bias if omitted.

Comment: Since half of the stores are 6 months old seasonality would be a problem for them. instead of omitting weeks with no sales could lead to bias, requesting zero entries will help in predicting better

Replies:

Comment: I changed my mind. It should be C and D. Since both of them foundation aspect of training.

Comment: A as missing seasonality is an issue for the majority of the stores. D as we need to impute zeros as we would otherwise miss data. C won't do anything on performance.

Comment: The factors that will adversely impact the performance of the forecast model are: Sales data is aggregated by week. This will reduce the granularity and resolution of the data, and make it harder to capture the daily patterns and variations in sales volume. The data scientist should request daily sales data from the source database to enable building a daily model, which will be more accurate and useful for the prediction task. Sales data is missing zero entries for item sales. This will introduce bias and incompleteness in the data, and make it difficult to account for the items that have no demand or are out of stock. The data scientist should request that item sales data from the source database include zero entries to enable building the model, which will be more robust and realistic

Comment: C. Aggregated Weekly Data: Since the objective is to predict daily sales volume, weekly aggregated data might mask important daily trends and variations. Requesting daily sales data will provide a finer granularity of information that is crucial for building an accurate daily sales prediction model. D. Missing Zero Entries for Item Sales: The omission of weeks with no sales can lead to biased predictions, as the model might not correctly account for periods of no sales. Including zero entries for item sales would provide a more accurate representation of sales patterns, including the absence of sales, which is valuable information for the model. Based on this analysis, the factors that would most adversely impact the model's performance are the aggregated weekly data (Option C) and the omission of weeks with no sales (Option D).

Comment: A - six months is likely not enough to detect clear seasonality C - Can do weekly from daily but cant reliably do daily from weekly

Comment: Letters A and C are correct: we want to do a daily model (our base is on weeks) and we need to deal with new stores VS old stores. It is important to emphasize that the letter D also makes sense: we need to know the days when there were no sales, however the way it is written means saving lines (days of sales) with zero in the database, which is not practical.

Comment: the two factors that will adversely impact the forecast model's performance are seasonality detection for new stores and the aggregation of sales data on a weekly basis. The data scientist should request categorical data to relate new stores with historical data and request daily sales data from the source database to build a daily model, respectively, to mitigate these issues effectively.

Comment: AD. rest makes no sense.

Comment: A. Since more than half of the stores have been in business for less than 6 months, it will be challenging to detect seasonality patterns for these new stores. Therefore, one solution is to request categorical data to relate new stores with similar stores that have more historical data. This will help the model to identify common patterns and accurately forecast sales for new stores. C. Since the sales data is aggregated by week, it may not be possible to identify daily patterns or trends. Hence, one solution is to request daily sales data from the source database to enable building a daily model. This will help the model to identify daily patterns and improve its forecasting accuracy.

Comment: I go with CD. How could we ignore the days with 0 sales? The model should be trained so that it can predict 0 sales days as well.

Comment: B, C, D are possible. A couldn't be an answer because the model must predict daily sales volumes while A says 'Request categorical data'.

Discussion for Question 157

Link: https://www.examtopics.com/discussions/amazon/view/74934-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- CDF: 14 votes
- CEF: 11 votes

Discussion

Comment: B CE is correct.

Comment: The questions says "The images for each product are classified according to specific product lines." why do we need Amazon Rekognition Custom Labels then?

Comment: Option C is correct because augmenting the images in the dataset can help the model learn more features and generalize better to new products. Image augmentation is a common technique to increase the diversity and size of the training data. Option E is correct because Amazon Rekognition Custom Labels can train a custom model to detect specific objects and scenes that are relevant to the business use case. It can also leverage the existing models from Amazon Rekognition that are trained on tens of millions of images across many categories. Option F is correct because class imbalance can affect the performance and accuracy of the model, as it can cause the model to be biased towards the majority class and ignore the minority class. Applying oversampling or undersampling can help balance the classes and improve the model's ability to learn from the data

Comment: assuming improve accuracy of the (existing) solution

Comment: Hopefully final answer this time CEF. was initially looking for D but changed to E now

Comment: C & F for sure the confusion between D and E but lets go for D as E will need more steps

Comment: The question asks for quick solutions and to improve the classifier's accuracy. Since we want a quick fix, I'm going to avoid solutions that requires a new model implementation. Therefore, the alternatives that can improve the performance of the current classification are: Letter F, C and D. Letters B and E would bring a new development cost from zero and Letter A does not solve the classification problem.

Replies:

Comment: NVM, D is wrong!

Comment: the three steps that would improve the accuracy of the solution are C (data augmentation), D (image normalization and scaling), and F (addressing class imbalances)

Replies:

Comment: See community answer is CEF due to images all same dimension so D removed.

Comment: CEF : C&F for Overfitting; E : "Rekognition DetectLabel" is the general image labeling capability of Amazon Rekognition, which provides predefined labels for common objects and concepts out-of-the-box. On the other hand, "Rekognition Custom Labels" allows you to create custom models to detect specific labels or objects that are not covered by the default labels,

Comment: CEF better choose

Comment: This is CDF. No idea why this is unclear here.

Comment: The problem is about "Overfitting", because the new products doesn't work well. It is not about simply improve model accuracy. C is great answer, augmentation is for overfitting. D is wrong, because normalization of pixel is not for overfitting, and "all images have the same dimensions. no need for scaling, they are already scaled. F is for imbalance data. if the data is imbalanced, they should perform poor on both training and testing data(new product). And the new product should perform bad only on those cold category, not overall poor performance. B and E are all about Rokognition, one is Rekognition Detect label, a built-in image classification model; one is Rekognition Custom Labels, a pre-trained with fine-tuning model.

Comment: you fix images (C), train Rekognition with these images (E) and finally infer to get the classes (B)

Comment: Amazon Rekognition custom label model requires time, expertise, and resources, often taking months to complete. Additionally, it often requires thousands or tens-of-thousands of hand-labeled images to provide the model with enough data to accurately make decisions. The solution must be quick. https://aws.amazon.com/rekognition/custom-labels-features/

Replies:

Comment: Sorry, never mind my answer it's actually CEF.

Replies:

Comment: not F, think carefully what is imbalanced data? what is its effect? does it only affect new product?

Comment: CEF is correct

Comment: D seems doing nothing with new product

Comment: CEF is correct in my opinion. D - The normalization and scaling would not provide the most dramatic improvement of the model. E - By using transfer learning, we are able to utilize a pre-trained model that can recognize a lot of the lower-level features and retain only the last few layers for this specific purpose.

Discussion for Question 158

Link: https://www.examtopics.com/discussions/amazon/view/75080-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 13 votes

Discussion

Comment: Isn't it A? the model doesn't classify C & D well.

Comment: the correct answer should be A, the model is clearly unable to tell C and D apart the reason why B is incorrect is subtle - there is holdout validation or cross-validation, but not holdout cross-validation; while I think it would be more reasonable to use CV with such a small dataset rather than holdout, the answer is mixing terms and therefore should be wrong also, the test set confusion matrix is still pretty comparable to the train set one, so I wouldn't say there is objective evidence to claim holdout is a wrong choice here

Comment: I would go for A as well.

Comment: I think option A is correct as C & D are behaving similarly.

Comment: I think the answer is D. A => C, D are similar in train but the testing results contradict that. There are many As and Bs for C

Comment: These results indicate that the model is overfitting for classes B and E, meaning that it is memorizing the specific features of these classes in the training data, but failing to capture the general features that are applicable to the test data. Overfitting is a common problem in machine learning, where the model performs well on the training data, but poorly on the test data3. Some possible causes of overfitting are: The model is too complex or has too many parameters for the given data. This makes the model flexible enough to fit the noise and outliers in the training data, but reduces its ability to generalize to new data

Comment: Actually, both A and D are true. It would be an easy one if we had to choose two answers. But we need to choose only one. So how to make sure that the person who created this question thought about A only? Also if we take a look into the test confusion matrix. We can see that the A class also missed with C class at the same rate as the C and D classes. I would even say that here the model is generally overfitted. I would go for B

Replies:

Comment: Also because of random peeking of test set entries, we got the wrong proportions of labels between train and test sets. So the answer can be even C

Comment: Letter A is correct. The model gets confused between (C) and (D) in training and testing.

Replies:

Comment: But on the test set it's even confused between A and C classes

Comment: Selected Answer: B Hold-out Hold-out is when you split up your dataset into a 'train' and 'test' set. The training set is what the model is trained on, and the test set is used to see how well that model performs on unseen data. A common split when using the hold-out method is using 80% of data for training and the remaining 20% of the data for testing. Hold-out Hold-out is when you split up your dataset into a 'train' and 'test' set. The training set is what the model is trained on, and the test set is used to see how well that model performs on unseen data. A common split when using the hold-out method is using 80% of data for training and the remaining 20% of the data for testing. Refere: https://medium.com/@eijaz/holdout-vs-cross-validation-in-machine-learning-7637112d3f8f

Comment: Model in unable to tell c&D

Comment: D - Training accuracies of B and E are higher than those of test, whereas A has similar accuracy in both. For C and D, test accuracy has actually improved.

Replies:

Comment: B and C has below 50% accuracy. D has 98% in train and 86% accuracy in test. And you are telling me, the take away is overfitting of D, Seriously???

Comment: I think the answer is A. The model doesn't perform well on class C and D in both training and testing dataset. I don't think B is relevant to the question(cross-validation is not mentioned in the question)

Comment: What means holdout cross validation. There should be holdout validation vs cross validation

Comment: B should be the correct answer

Discussion for Question 159

Link: https://www.examtopics.com/discussions/amazon/view/74806-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- BDE: 33 votes

Discussion

Comment: i will go for B, D and E. B and D for me are like doing partial regression and corr plot can actually tell you briefly how well the univariate is correlated with your target and i guess that also apply for D.. and E , feature importance ranking that's what feature selection strategy want from my POV. And for Data Binning is data enrichment just like augmentations , but then the question was saying they want to do feature selection over 1k+ variables which implies they actually care more about which variable(s) can contribute more on determining the price ?

Comment: B. Correlation plot with heat maps: This technique can be used to identify the relationship between each feature and the target variable (sales price). By creating a correlation plot with heat maps, the company can quickly visualize the strength and direction of the relationship between each feature and the target variable. D. Univariate selection: This technique can be used to select the features that have the strongest relationship with the target variable. It involves analyzing each feature independently and selecting the ones that have the highest correlation with the target variable. E. Feature importance with a tree-based classifier: This technique can be used to determine the most important features that contribute to the target variable. By using a tree-based classifier such as Random Forest or Gradient Boosting, the company can rank the importance of each feature and select the ones that have the highest importance.

Comment: For feature selection in machine learning, you can use the following techniques: B. Correlation plot with heat maps: Correlation analysis helps identify relationships between features and the target variable. A heat map can visually represent the correlation matrix, helping to identify highly correlated features. D. Univariate selection: Univariate selection methods evaluate the relationship between each feature and the target variable independently. Common techniques include statistical tests such as chi-squared tests, ANOVA, or mutual information. E. Feature importance with a tree-based classifier: Tree-based classifiers, such as decision trees or random forests, can provide feature importance scores. These scores help identify which features contribute the most to the predictive performance of the model.

Comment: A, C and F are not feature selection techniques.

Comment: BDE seem to be the only viable feature selection methods here

Comment: Those are the only ones for FS.

Comment: the most appropriate feature selection techniques for the company to determine the primary features contributing to the sales price are B (correlation plot with heat maps), D (univariate selection), and E (feature importance with a tree-based classifier).

Comment: BDE as stated here: https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e

Comment: BDE for me

Comment: throwing my weight behind B D E. Correlation with heatmaps help us eliminate multicollinearity, Univariate testing helps us see which ones are correlated with the target, same as feature importances of tree-based algorithms.

Comment: CDF for me

Discussion for Question 160

Link: https://www.examtopics.com/discussions/amazon/view/74920-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 29 votes

Discussion

Comment: Answer is C, CNN-QR and DeepAR accepts related time series data (weather data, number of people on property, etc.,)

Comment: Answer is C, CNN-QR and DeepAR accepts related time series data (weather data, number of people on property, etc.,). Classic forecasting methods, such as ARIMA or exponential smoothing (ETS), fit a single model to each individual time series. In contrast, DeepAR+ creates a global model (one model for all the time series) with the potential benefit of learning across time series. Source: https://aws.amazon.com/blogs/machine-learning/making-accurate-energy-consumption-predictions-with-amazon-forecast/

Comment: As per https://docs.aws.amazon.com/forecast/latest/dg/aws-forecast-choosing-recipes.html A. NO - no as powerful as NN B. NO - no as powerful as NN C. NO - works best with 100's of time series D. YES - best for strong seasonnability, expected for power

Comment: Based on this only CNN-QR can accept historical data https://www.examtopics.com/exams/amazon/aws-certified-machine-learning-specialty/view/32/

Comment: CNN-QR is a deep learning algorithm that can model complex relationships between the inputs and outputs, such as the weather and public holidays, with historical power consumption data. CNN-QR has been shown to be effective in generating accurate predictions in many different types of forecasting use cases, including demand forecasting. ETS (Exponential Smoothing) is a classical time series algorithm that is often used for forecasting. It can be effective for simple time series data that have regular patterns, but may not be sufficient to handle the complexity of the given data. ARIMA (Autoregressive Integrated Moving Average) is another classical time series algorithm that can model complex patterns in data. However, it may be difficult to use in cases where there are many different inputs and the relationships between the inputs and outputs are complex.

Comment: ARIMA & ES are both base time series algos that are available. DeeoAR+ & CNN-QR are refined and able to utilize external data as well to complement the time series data available

Comment: C, as explained here: https://docs.aws.amazon.com/forecast/latest/dg/aws-forecast-choosing-recipes.html

Comment: According to the link below, it is either ARIMA or DeepAR. So A is the answer here https://aws.amazon.com/blogs/machine-learning/making-accurate-energy-consumption-predictions-with-amazon-forecast/

Comment: Given the provided data, I would discard A and B. Amazon Forecast CNN-QR, Convolutional Neural Network - Quantile Regression, is a proprietary machine learning algorithm for forecasting scalar (one-dimensional) time series I would choose D, Prophet. https://docs.aws.amazon.com/forecast/latest/dg/aws-forecast-recipe-prophet.html How Prophet Works Prophet is especially useful for datasets that: Contain an extended time period (months or years) of detailed historical observations (hourly, daily, or weekly) Have multiple strong seasonalities Include previously known important, but irregular, events Have missing data points or large outliers Have non-linear growth trends that are approaching a limit Prophet is an additive regression model with a piecewise linear or logistic growth curve trend. It includes a yearly seasonal component modeled using Fourier series and a weekly seasonal component modeled using dummy variables.

Replies:

Comment: Prophet wont be able to use the additional data that is available in the question

Comment: Prophet doesn't accept historical-related time series, so it won't work here https://docs.aws.amazon.com/forecast/latest/dg/aws-forecast-choosing-recipes.html#comparing-algos

---

Discussion for Question 161

Link: https://www.examtopics.com/discussions/amazon/view/75078-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 15 votes

Discussion

Comment: Answer is C. https://aws.amazon.com/blogs/machine-learning/build-a-custom-vocabulary-to-enhance-speech-to-text-transcription-accuracy-with-amazon-transcribe/

Comment: -Creating a custom vocabulary file allows you to explicitly define the correct pronunciation of each product name. -Manually updating the custom vocabulary file based on these observations allows you to continuously improve the ASR system. - As new product names or variations emerge, you can easily add them to the custom vocabulary file without retraining the entire ASR model.

Comment: D was my initial choice however looking at the requirement "The company needs to ensure that everyone can update their customizations multiple times each hour." I changed my mind due to having to retrain the model with new vocabulary. C gives you the ability to update the vocabulary and have it take effect immediately

Comment: Answer is C D would required to build a model. It's well known the quantity of products, so it's not necessary.

Comment: the best approach to maximize transcription accuracy during the development phase is to use the audio transcripts to create a training dataset and build an Amazon Transcribe custom language model. Analyze the transcripts and update the training dataset with a manually corrected version of transcripts where product names are not being transcribed correctly. Create an updated custom language model.

Comment: Option D involves using the available audio transcripts to create a training dataset and building a custom language model with Amazon Transcribe. This approach provides a high degree of control over the transcription process and the ability to fine-tune the model to the specific vocabulary and pronunciation requirements of the company. Analyzing the transcripts and updating the training dataset with corrected versions is a crucial step in improving transcription accuracy. It enables the model to learn from mistakes and to incorporate the unique spelling and pronunciation of the 200 required product names.

Replies:

Comment: D is an ideal answer however, the question ask for "The company needs to ensure that everyone can update their customizations multiple times each hour". To retrain custom model each hour when we have changes, will be tedious and time consuming. I go with c, where we can ask everyone to just update the config file.

Comment: Thank you AjoseO for all these detailed explanations! They are very useful!

Replies:

Comment: say thank you to chat gpt

Comment: Why not D though?

Comment: I think C is correct.

Comment: A? any thought?

Discussion for Question 162

Link: https://www.examtopics.com/discussions/amazon/view/74919-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 21 votes

Discussion

Comment: B is wrong as it says it doesn't take advantage of GPUs

Comment: I believe answer should be C. a) Initial processing needs less cpu and memory so that can be done on a smaller instance. b) Second operation is memory intensive so instance type should be changed to R5 type instance.

Comment: I'd opt for C. A and B are wrong for obvious reasons. D sounds good but it doesn't have a ML instance and also it's just the development phase and we might not want to reserve an instance for too long.

Comment: Option A need only one instance all other options talks about 2 instances. so why can't it be A...

Comment: C Memory-optimized instances means provide a high memory In D they mention reserved instance. so it is costly

Comment: C is correct. Due that B is wrong, is not to use a GPU Instance based.

Comment: "Which solution will result in the MOST cost savings" Because of this, D is wrong: are you sure that allocating an instance for months / years for a 2h/day is cost saving? Correct is C

Comment: offers the best balance of cost savings and resource adequacy for both feature engineering and data preprocessing tasks.

Comment: It' C, as Reserved Instance no good for only 2 hours of daily work.

Comment: R instance with processes that uses lot of memory. Reserved instances for less cost

Replies:

> **Comment:** Selection of D is totally wrong, because you don't understand what "Reserved Instance" is!!! You cannot reserve a instance only for hours a day!!!! this is like apartment rent, can you just rent an apartment for nap time????

Comment: D over C because if the EC2 instance is being used consistently for the same two hours each day, customers could consider using a Reserved Instance with a term of 1 or 3 years and payment option that aligns with their usage pattern. This would provide significant cost savings compared to On-Demand pricing for those two hours each day.

Replies:

> **Comment:** is better not use everytime chatgpt, and read AWS documentation about instances.

Comment: B. C is wrong as ml.r5 is not stopped when not in use

Comment: IMO D is correct. The reserved instance option for an R5 instance, as in Option D, would provide the greatest cost savings, as reserved instances offer a discounted hourly rate in exchange for a one-time payment for a committed usage term.

Comment: I think C is correct. It only runs for 2 hours once a day, so RI is wasted. So I think D is wrong.

Replies:

> **Comment:** "Scheduled RIs: These are available to launch within the time windows you reserve. This option allows you to match your capacity reservation to a predictable recurring schedule that only requires a fraction of a day, a week, or a month." You have the option to reserve for a fraction of a day. Since the question specify precisely how long the job is, it makes it suitable. https://aws.amazon.com/ec2/pricing/reserved-instances/

Comment: 12-sep exam

Comment: I think it is D. Using RIs the customer can have the greatest cost savings, as stated by the question

Comment: I think the answer should be D: R5 instance is cheaper and appropriate for data processing under same memory size; reserved instance make it cheaper.

Replies:

> **Comment:** only use 2 hours per day, so no need for RI

Discussion for Question 163

Link: https://www.examtopics.com/discussions/amazon/view/75059-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- BE: 15 votes

Discussion

Comment: B, and E (frequency encoding)

Comment: BD？ any thought?

Replies:

> **Comment:** I think D cannot be because distances in RGB format are not representive of points. CIELAB correlates numerical color values consistently with human visual perception.

Comment: Using frequency encoding may help in some contexts but can introduce bias, especially if the frequency of a color is not related to the rental rate. This method does not leverage the actual differences between colors.

Comment: In this scenario, the specialist should use one-hot encoding and RGB encoding to allow the regression model to learn from the Wall_Color data. One-hot encoding is a technique used to convert categorical data into numerical data. It creates new columns that store one-hot representation of colors. For example, a variable named color has three categories: red, green, and blue. After one-hot encoding RGB encoding can capture the intensity and hue of a color, but it may also introduce correlation among the three columns. Therefore, using both one-hot encoding and RGB encoding can providemore information to the regression model than using either one alone.

Comment: Here we have a non-ordinal categorical variable to receive a numerical conversion for the model. Letter A is wrong as it is not an ordinal variable. Letter C is wrong as we are not going to retain any significant information for the model. The best solutions would be: Letter B and E. Letter D would be very interesting, but it would generate a problem of information fragmentation: most models consider the variables as being independent of each other, and these 3 columns by definition would not be independent.

Comment: B, and E (frequency encoding)

Comment: A+B make sense to me

Comment: B for sure D. This approach involves breaking down each color into its Red, Green, and Blue components and creating separate columns for each component. This allows the model to capture the information about the intensity of each color component, which can be useful in predicting the target variable. A, C, and E are not suitable for encoding color data in a way that can be used by a regression model. The integer transformation approach in option A arbitrarily assigns values to colors without any meaningful relationship between them. The approach in option C replaces the color names with their length, which does not provide any useful information for the model. Option E replaces each color name with its frequency in the training set, which does not capture any information about the color itself.

Comment: I think frequency encoding cannot be. What if some colors have same amount of frequency?

Comment: B. Add new columns that store one-hot representation of colors. One-hot encoding is a common approach to represent categorical variables as numerical values. This approach creates new binary variables for each category and assigns a value of 1 to the corresponding category and 0 to the others. In this case, the specialist can create three new binary variables, one for each color (Red, White, and Green) and use them as input to the regression model. E. Replace each color name by its training set frequency. Another approach to convert categorical variables into numerical ones is to replace each category with its frequency of occurrence in the training set. In this case, the specialist can replace the color names with their respective frequencies (1/3 for Red, 1/3 for White, and 1/3 for Green) to represent them numerically.

Replies:

> **Comment:** Frequency encoding is a feature engineering technique used to convert categorical variables into numerical ones by replacing each category with the frequency of its occurrence in the training set. This approach can be useful when dealing with high-cardinality categorical variables, which are categorical variables with a large number of distinct categories.

Comment: These are the only options preserving what "color" is. One-hot encoding is a default standard for any categorical data to be fed to a model that takes in numeric input. RGB format is a good numeric representation of any color by preserving its nature

Comment: A&B https://victorzhou.com/blog/one-hot/#:~:text=One-Hot%20Encoding%20takes%20a%20single%20integer%20and%20produces,of%20colors%20are%20possible%3A%20red%2C%20green%2C%20or%20blue.

Replies:

> **Comment:** B & E. It cannot be A because your URL specifically states that: "This is known as integer encoding. For Machine Learning, this encoding can be problematic - in this example, we're essentially saying “green” is the average of “red” and “blue”, which can lead to weird unexpected outcomes."

Comment: Frequency encoding

Comment: B and E

Comment: BE is correct. For e, please refer: https://medium.com/analytics-vidhya/different-type-of-feature-engineering-encoding-techniques-for-categorical-variable-encoding-

214363a016fb#:~:text=One%20Hot%20Encoding%3A%20%E2%80%94%20In%20this,slows%20down%20the%20learning%20significantly.

Discussion for Question 164

Link: https://www.examtopics.com/discussions/amazon/view/77501-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 11 votes

Discussion

Comment: answer : A, because the setting needs multi agents and is constrained with traffic light correlation.

Comment: A. The data scientist should obtain a correlated equilibrium policy by formulating this problem as a multi-agent reinforcement learning problem. In this scenario, where the traffic behavior at each light is correlated, a multi-agent reinforcement learning (MARL) approach is well-suited to model the problem. In MARL, multiple agents interact with each other and the environment, and their behavior is influenced by the behavior of other agents. This approach is particularly useful in modeling traffic systems, where the behavior of each vehicle is affected by the behavior of other vehicles and traffic lights. Formulating the problem as a MARL problem can help the data scientist obtain a correlated equilibrium policy, which can optimize traffic flow across multiple traffic lights by taking into account the correlations between them. By optimizing traffic flow across all traffic lights in a correlated way, it may be possible to reduce congestion and improve overall traffic efficiency.

Replies:

Comment: thank you chatgpt

Comment: i am wondering how is this actually implemented, i am learning deep RL right now

Comment: It's too complex problem for supervised or unsupervised. It's a multi-agent problem.

Comment: Answer is A

Comment: https://www.researchgate.net/publication/221456376_Multi-Agent_Reinforcement_Learning_for_Simulating_Pedestrian_Navigation

---

Discussion for Question 165

Link: https://www.examtopics.com/discussions/amazon/view/76486-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- D: 8 votes

Discussion

Comment: D: Option A, B & C don't make sense. D removes the stop words and help in count vectors

Comment: D The only solution that solves our problem: remove stopwords ASAP

Comment: D: Option A, B & C don't make sense. D removes the stop words and help in count vectors

Comment: D, ChatGPT confirm :)

Comment: Needs to remove stopwords and the rare worlds are feasible.

Comment: The stop words need to be removed. The rare words don't need to be removed because it has been found that they are feasible tags.

Comment: Why not A?

Replies:

Comment: check the requirement in question: "the generated model do not include the stopwords"

Discussion for Question 166

Link: https://www.examtopics.com/discussions/amazon/view/75399-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 26 votes

Discussion

Comment: https://aws.amazon.com/datasync/faqs/ Based on answer for the question - "How do I use AWS DataSync to migrate data to AWS?"

Comment: it's DataSync

Comment: All other options seem like they would require some manual coding to meet all requirements. DataSync appears as the best option as a result

Comment: Option C, using AWS DataSync, is the most appropriate solution. AWS DataSync is a service designed for data transfer between on-premises storage and AWS, and it provides the features the company needs:

Comment: C- DataSync is the answer

Comment: AWS DataSync is a service that can be used to transfer large amounts of data between on-premises storage and Amazon S3, EFS, or FSx for Windows File Server. DataSync is optimized for fast, automated, and secure transfers of large amounts of data, and it supports scheduling, monitoring, and data integrity validation. In this scenario, the company wants a solution that can transfer and automatically update data between the on-premises object storage and Amazon S3, with support for encryption, scheduling, monitoring, and data integrity validation. DataSync meets all of these requirements, as it can transfer data using secure network connections, schedule data transfers, verify data integrity, and encrypt data in transit and at rest.

Comment: A maybe? https://aws.amazon.com/datasync/faqs/

Replies:

Comment: I meant C

Discussion for Question 167

Link: https://www.examtopics.com/discussions/amazon/view/74997-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 32 votes
- D: 29 votes

Discussion

Comment: A : The data has label. So what we need to do is to enforce accuracy by reviewing low confidence ones internally

Replies:

Comment: Not A, bounding box should be a feature of Ground Truth. https://docs.aws.amazon.com/zh_cn/sagemaker/latest/dg/sms-bounding-box.html

Replies:

Comment: It's A. See https://aws.amazon.com/rekognition/custom-labels-features/. It says "The Rekognition Custom Labels console provides a visual interface to make labeling your images fast and simple. The interface allows you to apply a label to the entire image or to identify and label specific objects in images using bounding boxes with a simple click-and-drag interface." We are not using semantic segmentation, as it applies a label to every pixel. We don't want that, we want labels to bounding boxes.

Replies:

Comment: 1. you didn't understand what Rekognition is about. Rekognition is a CV model, not labeling tools. 2. you didn't read carefully about the document, just below your quote, it said "Alternately, if you have a large data set, you can use Amazon SageMaker Ground Truth to efficiently label your images at scale." Rekognition label function is for individual cases. 3. finally, you really sure this is an object detection, not pixel-level object classification? original model is object detection and it didn't work. semantic segmentation might be a solution and it is good for self-driving.

Comment: D; B is using MTurk which uses public workforce which violates the requirements that videos need to be kept private

Replies:

Comment: A quick google search on SageMaker ground truth will show you that you can indeed create your own private labelers workforce and send them labelling jobs through GroundTruth. "You have options to work with labelers inside and outside your organization. For example, you can send labeling jobs to your own labelers, or you can access a workforce of over 500,000 independent contractors who are already performing ML-related tasks through Amazon Mechanical Turk. If your data requires confidentiality or special skills, you can also use vendors that are pre-screened by AWS for quality and security procedures."

Replies:

Comment: Nevermind My bad. Just realized that you are referring particularly to Mechanical Turk being used as the labelers force for Ground Truth, which is what answer B referring to.

Comment: The Problem with Answer D though is that there is no "semantic segmentation labeling task" within the very limited list of Ground Truth type of jobs it offers. There is a video classification Job type, and there is Video Frame labeling job type, which includes "video frame object detection job" and "video frame object tracking job. However, there is no Semantic Segmentation Labeling Job"

Replies:

Comment: There is a Ground Truth for Semantic Segmentation labelling task - https://docs.aws.amazon.com/sagemaker/latest/dg/sms-semantic-segmentation.html. Therefore, D is correct.

Comment: https://docs.aws.amazon.com/sagemaker/latest/dg/sms-video.html

Comment: And Semantic segmentation is object classification done at the pixel level. Isnt that something only machines can do? Unless labelers are directing a machine to do the semantic segmentation for them, I think labelers are no use for it.

Comment: Agreed. Option A used Amazon Augmented AI (Amazon A2I), not a good way to review confidential data.

Comment: semantic segmentation may not be the right choice for labelling, instead Rekognition is ideal for this scenario and private workforce + A2I to validate the labelling.

Comment: While Amazon Rekognition Custom Labels with Amazon A2I could be used for object detection, semantic segmentation provides more detailed information about the spatial layout of objects in an image, making it potentially more suitable for tasks like demarcating safety lines.

Comment: Semantic segmentation will provide the precise pixel-level labeling required to demarcate the yellow safety line accurately, passengers, and trains. A private workforce will ensure that the video data remains confidential. As the original model can't correctly identify the line, semantic segmentation might offer the needed precision. So D is right.

Comment: Amazon Rekognition does not support creation of private workforce. Between A & D, D is the only option that allows its creation. Semantic segmnentation can easily identify the yellow line.

Comment: Given the requirements of the task and the need for confidentiality, the best approach would be: D. Use an Amazon SageMaker Ground Truth semantic segmentation labeling task with a private workforce. Semantic segmentation will provide the precise pixel-level labeling required to demarcate the yellow safety line accurately, passengers, and trains. A private workforce will ensure that the video data remains confidential.

Comment: D. Use an Amazon SageMaker Ground Truth semantic segmentation labeling task. Use a private workforce as the labeling workforce. Here's why this approach is suitable: Semantic Segmentation Labeling: Semantic segmentation involves labeling each pixel in the image, which is more granular than bounding boxes. This approach is ideal for accurately demarcating the yellow line, which might be difficult with just bounding boxes. It also allows for precise detection of passengers and trains. Private Workforce: Given the requirement for confidentiality, using a private workforce ensures that the data is handled by trusted, authorized personnel. This addresses the concern of keeping the video data confidential. Amazon SageMaker Ground Truth: This service provides tools for efficient and accurate labeling of image data, which is essential for training a robust object detection model.

Comment: A: D is too complicated. Option D suggests using SageMaker Ground Truth with a semantic segmentation labeling task and a private workforce. Semantic segmentation can be useful for delineating the yellow line clearly. However, it might be more complex than necessary for this scenario, and object detection might be more suitable. In my opinion, A is a better option.

Comment: Rekognition is not guaranteed no to use your data to improve their models. Similarly, mechanical turk will not keep data private. Only viable option is D

Comment: B and C excluded since use public workforce. D excluded since question is asking for "labeling approach for THIS model" it means you don't want to switch to a semantic segmentation problem. Therefore A is the correct answer

Comment: Given that the video data must remain confidential, options that use public workforces like Amazon Mechanical Turk or third-party AWS Marketplace vendors would not be suitable. option A relies on object detection (bounding boxes) and doesn't switch to semantic segmentation. semantic segmentation provides precise labels, especially when distinguishing between closely placed objects like the yellow line and the passengers.

Comment: A: Using Amazon Rekognition Custom Labels you can do the same of Ground Truth this answer is complete with all steps.

Comment: The company can use Amazon Rekognition Custom Labels to label the dataset and create a custom Amazon Rekognition object detection model. They can create a private workforce and use Amazon Augmented AI (Amazon A2I) to review the low-confidence predictions and retrain the custom Amazon Rekognition model. This approach will help the company improve the model as it allows them to train a custom model that is specific to their business needs. The custom model can be trained to detect the yellow line, passengers who cross the yellow line, and trains in the video feeds. The private workforce ensures that the video data remains confidential, while Amazon A2I helps to improve the accuracy of the model by reviewing low-confidence predictions and retraining the model. which make A is more suitable than D using the Sagemaker Ground Truth semantic segmentation.

Comment: The question asks for what labelling solution do you suggest, so based on that how can A be an answer? It is a solution that brings in a human review to the problem, while answer D is direct to the requirement

Comment: A. Leverage pre-trained Amazon Rekognition and has the tools for creating bounding boxes B. Images are already labeled, the problem is not enough data for good accuracy C. Use of AWS Marketplace vendor does not satisfy privacy requirements D. Same issue as B

Comment: We want to categorize the images for object detection while keeping the content confidential. That said, Letters B and C are false as they breach confidentiality. To generate labels, it is recommended to use Ground Truth. Letter A is wrong as it misuses AWS Rekognition! By elimination, alternative D. He spoke about the labeling process, he spoke about AWS Ground Truth.

Discussion for Question 168

Link: https://www.examtopics.com/discussions/amazon/view/75262-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- BC: 28 votes

Discussion

Comment: B,C should be the answer

Comment: BD should be the answer

Replies:

Comment: I change my mind to B and C now

Comment: PCA is sensitive to the variance of features, so it's a common practice to standardize (e.g., z-score normalization) or scale (e.g., min-max scaling) the features before applying PCA. If the features are on different scales, it can distort which principal components are viewed as the most important.

Comment: Well, first time that I go with the Suggested Answer. D - B is the way: We want to solve the base correlation problem. That said, Letters A - E don't solve this problem, so they're wrong. Letter C partially solves the problem, so it is wrong. As we want steps, the correct alternatives are: D (ensure that all variables are on the same scale) and B (apply PCA that removes all correlation from the base while keeping most of the information). Again, from my perspective (C) is vague and using (B) removes the necessity of drop highly correlated features.

Replies:

Comment: Agree. As a question asks us what steps to perform, then it is logical to say: "Scale features and apply PCA" we can't answer "remove a portion of correlated features and then apply PCA", or vice versa. as it doesn't make sense. It would make some sense if we were asked "What technics engineer can apply", or smth similar

Comment: the most effective steps to address the issue of high correlation among the features in the dataset are removing a portion of highly correlated features and applying principal component analysis (PCA) for dimensionality reduction. These steps will help improve the data quality and predictive performance of the model.

Comment: PCA is a widely used technique for reducing the dimensionality of high-dimensional datasets while retaining as much of the original variability as possible. It is particularly useful when dealing with highly correlated features. Removing a portion of highly correlated features can be another effective way to address the issue of high correlation. By removing some of the correlated features, the model can become less complex and less prone to overfitting.

Comment: MinMax scaling does nothing to fix the issue here

Comment: BC for me, minmax scaler cannot remove multicollinearity?

Discussion for Question 169

Link: https://www.examtopics.com/discussions/amazon/view/74998-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 27 votes

Discussion

Comment: C; Lambda execution time has hard limit of 15 mins which might not be enough for data processing

Replies:

Comment: but C requires some coding efforts

Replies:

Comment: I think C is correct, because pyspark are also a kind of Python, and it only require a little code change.

Comment: D is wrong as AWS Lambda has a maximum execution time of 15 minutes, which may not be sufficient for some of the scripts. C is right as it's serverless and not a lot of work.

Comment: Redshift is definitely going to require some effort to setup, lambda just won't cut it performance-wise if the EC2 instance can't. Guess whats left?

Comment: C seems the most correct but it misses the part of importing data in AWS

Comment: option c fit all requirements since, it provides the least development effort using AWS Glue, and convert the python to pyspark which provide the most performance. option D is not suitable because lambda function has a limitation of only 15 minutes ruining while the script needs 1 hour.

Comment: We want to eliminate server management and reduce development effort. That said, Letter A is wrong, as it brings effort to refactor code. Letter B is wrong as DynamoDB asks for server management. Letter D is wrong, because despite the services being serverless (Lambda and Step Functions), the maximum timeout of a Lambda function is 15 minutes, which would be less than the desired one (1 hour). Letter C is correct, even if there is a pure Python code conversion effort → PySpark, this is the solution that fits the requirements.

Comment: Option C is the best option because it allows you to use the existing Python scripts without having to convert them to a different language or framework. AWS Glue is a managed service that makes it easy to prepare data for analysis. PySpark is a Python library that allows you to use Spark to process data. This approach would address all of the requirements with the least development effort and would be able to handle large-scale data processing.

Comment: Overall, option C with AWS Glue and PySpark is the most efficient approach, as it requires the least amount of development effort while effectively addressing all the requirements, including moving away from EC2 maintenance and handling large-scale data processing.

Replies:

Comment: corrected to option c

Comment: The data pipeline involves cleaning, transforming, enriching, and compressing terabytes of data and storing the data in Amazon S3. AWS Glue is an ETL service that makes it easy to move data between data stores. The Glue job allows you to use PySpark scripts to perform ETL tasks. With AWS Glue, you do not need to provision and manage servers, which eliminates the need to maintain servers, as required by the company. Therefore, AWS Glue would address all of the company's requirements with the least development effort.

Comment: 1) eliminate the need to maintain servers - Lambda is serverless 2) the least development effort - python scripts do no need to be rewritten for Lambda function

Replies:

Comment: "with each script taking at least an hour" - lambda would be time-out during taking job.

Comment: C as for Redshift I need to build a new pipeline

Comment: Converting python scripts to pyspark is less coding effort than writing up SQL, which is somewhat limited in the types of transformations it can do. Lambda function responses are a deadend for reason already given (timeout)

Discussion for Question 170

Link: https://www.examtopics.com/discussions/amazon/view/74809-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- ABD: 37 votes

Discussion

Comment: ABD; Email exchange between customer and customer service would be valuable data source.

Comment: Everyone explained correct.

Comment: In conclusion, the data scientist should use customer service emails, social media posts, and publicly available customer reviews to augment the dataset of reviews for the analysis of customer feedback and identifying specific areas for improvement.

Comment: A: Emails exchanged by customers and the company's customer service agents can provide additional customer feedback and opinions about the products or services. This data can be used to improve the ML model. B: Social media posts containing the name of the company or its products can provide additional customer feedback and opinions about the products or services, which can be used to improve the ML model. D: A publicly available collection of customer reviews can be used to augment the existing dataset of reviews and increase the size of the dataset. This can help to improve the accuracy of the ML model.

Replies:

Comment: C: A publicly available collection of news articles and F: Instruction manuals for the company's products are not directly related to customer feedback and may not be relevant for improving the ML model in this context. E: Product sales revenue figures for the company can provide valuable insights into the company's financial performance, but this data is not directly related to customer feedback and may not be useful for improving the ML model in this context.

Comment: I think ABE or maybe(!) ABF A & B -> for sure C -> No glue how this should help D -> We have already reviews!? E -> Could help to find correlations between negative / positive reviews and sales F -> non-sense in the first moment and second as well, but maybe it could help to combine the information of Email, Social and Reviews to some problems.

Replies:

Comment: That's exactly the point though, we want more volume of reviews, or anything resembling reviews as is the case with email exchanges.

Comment: ABE makes more sense here.

Comment: Don't overthink conrad.

Comment: BDF - correct answer

Discussion for Question 171

Link: https://www.examtopics.com/discussions/amazon/view/74810-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- D: 20 votes

Discussion

Comment: while I agree that Sagemaker Experiments is the way to go, it only supports Training, Processing, and Transform jobs, so the right answer is to run the job as a processing job, hence D not B https://docs.aws.amazon.com/sagemaker/latest/dg/experiments-create.html#:~:text=CreateTrainingJob-,Processing,-Processor.run

Replies:

Comment: "Generally, you use load_run with no arguments to track metrics, parameters, and artifacts within a SageMaker training or processing job script." https://docs.aws.amazon.com/sagemaker/latest/dg/experiments-create.html

Comment: Run PySpark script in SageMaker processing job https://sagemaker.readthedocs.io/en/stable/amazon_sagemaker_processing.html

Comment: B - https://docs.aws.amazon.com/sagemaker/latest/dg/experiments.html

Replies:

Comment: But It doesn't describe glue job.

Comment: Pyspark -> AWS Glue

Comment: AWS Glue is a fully managed extract, transform, and load (ETL) service that is purpose-built for processing large datasets and executing PySpark scripts. It's more aligned with the task of running a PySpark script with complex window aggregation operations to prepare data for training and testing

Comment: D https://sagemaker-experiments.readthedocs.io/en/latest/tracker.html

Comment: A PySpark script can be run as a SageMaker processing job by using the SparkProcessor class. A SageMaker processing job can use Amazon SageMaker Experiments to track the input parameters, output metrics, and artifacts of each run. A SageMaker processing job can also use Amazon SageMaker Debugger to capture tensors and analyze the training behavior, but this is more useful for deep learning models than for data preparation tasks. Running the script as an AWS Glue job would not allow the ML specialist to use Amazon SageMaker Experiments or Amazon SageMaker Debugger, as these features are specific to SageMaker.

Comment: D: SageMaker Experiments automatically tracks the inputs, parameters, configurations, and results of your iterations as runs.

Comment: The PySpark script defined above is passed via via the submit_app parameter https://github.com/aws/amazon-sagemaker-examples/blob/main/sagemaker_processing/spark_distributed_data_processing/sagemaker-spark-processing.ipynb

Comment: Key metrics is the "key". Then D is not a correct answer

Replies:

Comment: what is the difference between key metrics and key parameteres? why we care about key metrics, because we can compare the key metrics of different parametes and then find impact of the number of features. so the key is "glue" or "SageMaker processing"

Comment: D looks the right answer

Comment: https://docs.aws.amazon.com/sagemaker/latest/dg/experiments.html ---- Use SageMaker Experiments to view, manage, analyze, and compare both custom experiments that you programmatically create and experiments automatically created from SageMaker jobs.

Replies:

Comment: "SageMaker jobs" not "Glue job", it is D!

Comment: B: Glue job goes with window aggregation operations

Comment: https://sagemaker-examples.readthedocs.io/en/latest/sagemaker_processing/spark_distributed_data_processing/sagemaker-spark-processing.html

Comment: here: https://aws.amazon.com/about-aws/whats-new/2018/10/aws-glue-now-supports-connecting-amazon-sagemaker-notebooks-to-development-endpoints/#:~:text=AWS%20Glue%20now%20supports%20connecting%20Amazon%20SageMaker%20notebooks%20to%20development%20endpoints,-Posted%20On%3A%20Oct&text=You%20can%20now%20create%20an,an%20AWS%20Glue%20development%20endpoint.

Discussion for Question 172

Link: https://www.examtopics.com/discussions/amazon/view/74877-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- D: 28 votes

Discussion

Comment: Should be D according to the following article https://aws.amazon.com/blogs/machine-learning/speed-up-training-on-amazon-sagemaker-using-amazon-efs-or-amazon-fsx-for-lustre-file-systems/

Comment: Using Amazon SageMaker for training, you can utilize an Amazon EFS as your data source as long as the data already resides in Amazon EFS before starting the training job. This option requires least integration work.

Comment: My God, the answer is not D!!! Using EFS for Lustre reduces the start-up time by eliminating the data download step of the training process and leveraging the various performance and throughput benefits of the file system to execute the training job faster. So, A IS the correct !!!

Replies:

Comment: "with the least number of steps and integration work required"

Comment: the management wants the data scientist to create and train a model with the least number of steps and integration work required, (this is the keyword) so there is no need to include more things than sagemaker and EFS which make option D is the most suitable

Comment: SageMaker Noteboob instances can take input data directly from below, 1. AWS S3 2. Elastic File System (EFS) 3. FSx for Lustre file system Since the question is only regarding less coding effort and does not concern high availability or high performance, Option D would be good

Replies:

> **Comment:** boobies

Comment: This option allows the data scientist to use the existing dataset in EFS without copying or moving it to another storage service. It also minimizes the number of steps and integration work required, as SageMaker supports EFS as a data source for training jobs. This option is also cost-effective and time-efficient, as it avoids additional charges and delays associated with data transfer and storage.

Comment: Less effort, then D

Comment: "When you create a training job, you specify the location of a training dataset and an input mode for accessing the dataset. For data location, Amazon SageMaker supports Amazon Simple Storage Service (Amazon S3), Amazon Elastic File System (Amazon EFS), and Amazon FSx for Lustre. " https://docs.aws.amazon.com/sagemaker/latest/dg/model-access-training-data.html

Comment: https://aws.amazon.com/blogs/machine-learning/mount-an-efs-file-system-to-an-amazon-sagemaker-notebook-with-lifecycle-configurations/

Comment: Amazon SageMaker now supports Amazon Elastic File System (Amazon EFS) and Amazon FSx for Lustre file systems as data sources for training machine learning models on SageMaker. then why not select D ??

Replies:

> **Comment:** Time constraints. A is the right answer for this question
>
> **Replies:**
>
> > **Comment:** A is not the right answer. It's D. - A requires this setup: EFS -> Lustre -> Sagemaker. - D requires this setup: EFS -> Sagemaker It's obviously not A.

Discussion for Question 173

Link: https://www.examtopics.com/discussions/amazon/view/76292-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 10 votes

Discussion

Comment: C is the correct answer.

Comment: Most accurately. C is a tempting answer. You will see the model degradation over time. You could see if it's slowly getting worse or was it sudden. B is more accurate. You will only have two histograms to compare, but you will easily see which direction the error move: Are we over or underestimating. In practice you would use both. I'm terms of the exam - most accurate - most added information, B gives more information than C. It's a preference though.

Replies:

> **Comment:** Although, following the docs it will be C: https://docs.aws.amazon.com/forecast/latest/dg/predictor-monitoring-results.html

Comment: D is the correct answer

Comment: Weekly MAE aggregates the error metrics over a larger time window, which can mask fluctuations and specific patterns in the model's performance on a daily basis. In situations where there are sudden changes or degradation in the model's accuracy within a week, this visualization might not capture those nuances effectively.

Comment: Degradation over time: line or scatter plot (options C, D) C is aggregrate weekly view and doesn't give any additional details. D compares the model's errors during 3 week period to the errors from before that period giving an accurate picture of anomalies

Comment: GPT: To accurately visualize the degradation of the model over time and understand the source of inaccuracies, the ML team should focus on comparing the model's performance before and after the reported period of inaccuracies. The most appropriate option is: D. Create a scatter plot of daily sales versus model error for the last 3 weeks. In addition, create a scatter plot of daily sales versus model error from before that period.

Replies:

> **Comment:** Claude 3 Sonnet: Based on the evidence from AWS documentation and best practices, Option B: Create a histogram of the model errors over the last 3 weeks. In addition, create a histogram of the model errors from before that period, is the most accurate approach for the ML team to visualize the model's degradation. Histograms of model errors directly visualize the distribution and patterns of the model's inaccuracies, which is crucial for understanding the source of the problem. By comparing the error distributions before and after the 3-week period, the ML team can identify any significant shifts or changes that may indicate the cause of the model's degradation. This approach aligns with AWS best practices for model monitoring and visualization, as recommended by the Amazon SageMaker Model Monitor documentation. It provides a clear and focused visualization of the model's performance, enabling the ML team to gain insights and take appropriate actions to address the inaccuracies.

Comment: The best option to visualize the model's degradation most accurately would be to compare the model's errors over the relevant periods. This directly addresses the issue of model accuracy and allows for a clear comparison of model performance before and after the reported period of inaccuracy. Therefore, the most appropriate approach would be: B. Create a histogram of the model errors over the last 3 weeks. In addition, create a histogram of the model errors from before that period. This approach will allow the ML team to see if the distribution of errors has changed recently, indicating a degradation in model performance.

Comment: Is there a reason to create weekly MAE plot, if the prediction is made on daily granularity?

Comment: this is the key sentence : At the end of each day, an AWS Glue job consolidates the input data that is used for the forecasting with the actual daily sales data and the predictions of the model and that is exactly what MAE do: mean absolute error (MAE) is a statistical measure of the difference between two continuous variables. It is calculated as the average of the absolute differences between the predicted and actual values so the answer is C

Comment: A. NO - Daily sales histogram does not help to see model error B. NO - Histogram of the model errors is good, but no point to have one for the first 3 weeks and another for older data C. YES - one chart of model errors is perfect D. NO - no point to have 2 charts again

Comment: C is correct.

Comment: option B with histograms of model errors for the specific time periods is the most accurate and appropriate visualization to understand the model's degradation and identify the reasons behind the inaccuracies in the daily sales forecasting.

Comment: C is the answer, line plots are good solutions for time series analysis.

Comment: C is correct. We could view the "Degradation" as a trend. Line charts are usually very helpful to show if there is any trend in the data over the period of time under analysis. Histogram is normally used to visualizing distributions in your data.

Comment: Should be A because it is daily forecasting and histograms before and after will show the comparable degradation.

Replies:

> **Comment:** It only states to plot daily sales.. how should that help with the error? You need a plot of the actual and predicted values or or the errors - def not A

Comment: B, I guess

Replies:

> **Comment:** No, it's neither A or B as they use a histogram which would plot the distribution of errors. It will not tell you anything about how the model degrades over time, as the histogram will have no time component. You need a line chart for this. So it's C.

Discussion for Question 174

Link: https://www.examtopics.com/discussions/amazon/view/74922-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- D: 20 votes

Discussion

Comment: Answer is D! K-means used for customer segmentation

Replies:

> **Comment:** well, both are used for customer segmentation Knn & kmeans but kmeans is for unsupervised learning and knn is for supervised learning. since we have the data it's better to use supervised learning in this case. Ref: https://rstudio-pubs-static.s3.amazonaws.com/599866_59be74824ca7482ba99dbc8466dc36a0.html#:~:text=The%20difference%20between%20the%20two,to%20predict%20the%20unlabelled%20data.

Comment: The answer is D. 1. "The current segmentation of consumers is unclear." so it is unsupervised learning. 2. Then K-means is for unsupervised learning.

Comment: Typical clustering problem - use K means

Comment: KNN is used to solve missing data in regression/supervised problems. Since the question says unknown segmentation, it is an unsupervised problem and K-Means is the right choice. So Option D it is.

Comment: D is the correct C is wrong because kNN stills a supervised algorithm

Comment: The XGBoost model is a supervised machine learning algorithm, which means it requires labeled data to learn from. However, the customers' current segmentation is unknown, so there are no labels to train or evaluate the model. The data scientist needs an unsupervised machine learning algorithm, which can discover patterns and clusters in unlabeled data. A k-means model is an example of an unsupervised machine learning algorithm that can partition the data into K groups based on similarity. By setting K = 5, the data scientist can obtain five customer segments based on age, income, and location.

Comment: KNN has no k parameter in its input. C is not the answer.

Replies:

> **Comment:** in K-means also there is no input parameter "K". What i mean to say here is in knn the k is nothing but "kNN classifier identifies the class of a data point using the majority voting principle. If k is set to 5, the classes of 5 nearest points are examined."

Comment: D The key work is that the classification is "unclear", therefore k-means

---

Discussion for Question 175

Link: https://www.examtopics.com/discussions/amazon/view/74994-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 21 votes

Discussion

Comment: agree it's A for me

Comment: A: it's overfitting so regularization is needed, need apply scaling on financial data fields as it's for regression problem; one hot encoding for city of the house field.

Comment: Option A is the most likely to improve the testing accuracy the most effectively because it uses appropriate preprocessing techniques for both categorical and numerical data and applies a regularization technique that can help in reducing overfitting, thereby potentially improving the model's generalization to unseen data.

Comment: I will go with B. (A) suggests applying regularization to the data. It doesn't make sense. (B) answer is well framed. At least it doesn't use the wrong formulation.

Replies:

> **Comment:** B also looks suspicious.

Comment: Use a one-hot encoder for the categorical fields in the dataset. Perform standardization on the financial fields in the dataset. Apply L1 regularization to the data.

Comment: Agree with A, but I think the answer is slightly inaccurate. L1 regularization within the model and to the loss function. As a result, some features will be removed in the model. The answer suggest L1 regularization is applied to the dataset directly.

Comment: Option A is the most appropriate approach to improve the testing accuracy of the model. One-hot encoding can effectively represent categorical variables in a numeric format that is suitable for machine learning models. Standardizing the financial fields can make the data more comparable and improve the model's performance. L1 regularization can help in feature selection and avoid overfitting by reducing the number of features.

Replies:

> **Comment:** How do you apply regularization to Data and not to the model params?

Comment: Why are most of the chosen answers by ExamTopics mostly obviously wrong? There is nothing like tokenisation of categorical variable and B should be obviously wrong.

Replies:

> **Comment:** When they were published (firtsly, they steal them by photo/camera) they didn't have chatgpt to see the answers, and of course, they don't have any ML specialist or time to resolve them.

Comment: 12-sep exam

Discussion for Question 176

Link: https://www.examtopics.com/discussions/amazon/view/75431-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- AC: 16 votes

Discussion

Comment: seq2seq and object2vec take care of more than just the words. Any response with blazingText is wrong because blazingText just uses a cbow (continuous bag of words), working only on individual words

Replies:

> **Comment:** “One of the well-known embedding techniques is Word2Vec, which provides embeddings for words.” “In addition to word embeddings, there are also use cases where we want to learn the embeddings of more general-purpose objects such as sentences, customers, and products. This is so we can build practical applications for information retrieval, product search, item matching, customer profiling based on similarity or as inputs for other supervised tasks. This is where Amazon SageMaker Object2Vec comes in.” https://aws.amazon.com/blogs/machine-learning/introduction-to-amazon-sagemaker-object2vec/

> **Comment:** This is wrong, maybe the response (and the question) is outdated because BlazingText now supports three diferent techniques

Comment: It should be B an D. The objective is to create a latent space/word embedding that puts similar words closer to each other for other purposes. Thus, we should use Sagemaker Blazing Text in unsupervised mode (Word2Vec mode). cbow, skip-grams, and batch skip-grams are the 3 algorithms for this. However, since we do not need to do the later part of E, E is not correct. The ans should be B and D.

Replies:

Comment: yeah, my initial thought was the same. But both B and D embed words, not sentences.

Comment: Best choices

Comment: To extract embedding vectors: Blazingtext Word2vec and Object2vec (B, C). Seq to seq: generate one sequence from another (A is out) Amazon SageMaker BlazingText algorithm in continuous bag-of-words (CBOW) mode does not capture word embeddings (D is out)

Comment: To capture word context and sequential QA information, the embedding vectors need to consider both the order and the meaning of the words in the text. Option B, Amazon SageMaker BlazingText algorithm in Skip-gram mode, is a valid option because it can learn word embeddings that capture the semantic similarity and syntactic relations between words based on their co-occurrence in a window of words. Skip-gram mode can also handle rare words better than continuous bag-of-words (CBOW) mode1. Option E, combination of the Amazon SageMaker BlazingText algorithm in Batch Skip-gram mode with a custom recurrent neural network (RNN), is another valid option because it can leverage the advantages of Skip-gram mode and also use an RNN to model the sequential nature of the text. An RNN can capture the temporal dependencies and long-term dependencies between words, which are important for QA tasks2

Comment: Considering the requirements, the two options that can produce the required embedding vectors that capture word context and sequential QA information are: C. Amazon SageMaker Object2Vec algorithm: Because it can learn to capture relationships in pairs of text, which could include the sequential nature of questions and answers. E. Combination of the Amazon SageMaker BlazingText algorithm in Batch Skip-gram mode with a custom recurrent neural network (RNN): This combination provides both context-aware word embeddings and the ability to capture sequential dependencies in text data.

Comment: C because Object2Vec is a neural network-based algorithm that can learn embeddings for a wide range of data types and tasks. E because If you want to capture word context and sequential information, especially in the context of natural language processing (NLP), it is advisable to use models that are specifically designed for sequence modeling, such as recurrent neural networks (RNNs) or more advanced models like long short-term memory networks (LSTMs) or transformers.

Comment: A. Amazon SageMaker seq2seq algorithm: Sequence-to-sequence (seq2seq) models are designed to convert sequences from one domain to sequences in another domain, often used in tasks like machine translation. They are capable of understanding the context and the sequence in which words appear, making them suitable for differentiating between questions and answers in a text. E. Combination of the Amazon SageMaker BlazingText algorithm in Batch Skip-gram mode with a custom recurrent neural network (RNN): This combination is promising. BlazingText in Skip-gram mode captures word context, and the recurrent neural network (RNN) is excellent for capturing sequential data, such as the flow in conversations or text. This combination should be effective at understanding both the context of individual words and the sequence of questions and answers.

Comment: One problem of Object2Vec is it takes two objects as input during training and loss minimizes the difference between embeddings of these two objects. I don't think we have some labels to pass to Object2Vec We might think that we have a QA which we can pass as two objects. But in the question we want embeddings to distinguish between Q & A, but this Object2Vec minimizes the difference. So I wouldn't tell it is for sure C https://aws.amazon.com/blogs/machine-learning/introduction-to-amazon-sagemaker-object2vec/

Comment: A . Seq2seq (sequence-to-sequence) models are designed to handle sequences. They are particularly well-suited for tasks like translating sentences from one language to another, but they can also be used for other tasks that involve sequences, such as converting questions to answers. Given that the embedding space must differentiate between questions and answers, a seq2seq model would be a good choice. E . BlazingText in Skip-gram mode can capture word context effectively. However, on its own, it might not capture the sequential information between questions and answers. By combining it with a custom RNN, the sequential nature of the sentences, especially in a QA setting, can be captured. RNNs are designed to work with sequences and can remember past information, making them suitable for this task.

Comment: A. NO - seq2seq not for word embeddings B. YES - BlazingText in Skip-gram works and can capture Q&A C. NO - object2vec not for word embeddings D. YES - BlazingText in CBOW works and can capture Q&A E. NO - no need for RNN

Comment: Answers should be B and C See the following for BlazingText with Skip-gram: https://arxiv.org/pdf/1604.04661.pdf (search skip-gram) Linked from this page https://docs.aws.amazon.com/sagemaker/latest/dg/blazingtext.html For Object2Vec see this page https://docs.aws.amazon.com/sagemaker/latest/dg/object2vec.html

Comment: C is confirmed but confused between A or E then lean to E

Comment: confusing

Comment: Seq2Seq will not generate an embedding vector, so A it's wrong from my POV. I go with B - C

Comment: IMO E for sure then either B or C Combination of the Amazon SageMaker BlazingText algorithm in Batch Skip-gram mode with a custom recurrent neural network (RNN) is a more sophisticated approach that can be used to capture sequential QA information. This is because RNNs are able to learn long-term dependencies between words.

Comment: I've never heard about using seq2seq to generate embedding vectors. B = In skip gram, the order of the words does matter C = It's made for these type of things, generate embeddings of "objects" (sentences or what have you).

Discussion for Question 177

Link: https://www.examtopics.com/discussions/amazon/view/74871-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- D: 26 votes

Discussion

Comment: I would say D. We shouldn't need Textract to extract columns from a database

Replies:

Comment: I think D because Textract doesn't support CSV but only PNG, JPEG, TIFF, and PDF formats

Comment: D because it does not require heavy machine learning expertise

Comment: A. NO - Object2Vec is unsupervised, it will create vector representations but not assign to a category the claims B. NO - we want a supervised method, LDA will create topics in an unsupervised way C. NO - again we want a supervised method D. YES - That is supervised; no need for ML skills, only basic API programming

Comment: C is wrong because the columns doesn't exist.

Comment: D. Export the database to a .csv file with two columns: claim_label and claim_text. Use Amazon Comprehend custom classification and the .csv file to train the custom classifier. Develop a service in the application to use the Amazon Comprehend API to process incoming claims, predict the labels, and route the claims to the appropriate queue.

Comment: Option D meets the requirements. The solution requires no ML expertise, and the small development team can use the Amazon Comprehend custom classification API to train a model to automatically detect claim categories. The company can export the database to a .csv file with two columns: claim_label and claim_text. Then, the development team can use the .csv file to train the custom classifier. Finally, the team can develop a service in the application to use the Amazon Comprehend API to process incoming claims, predict the labels, and route the claims to the appropriate queue. This solution is straightforward, does not require extensive expertise, and can be implemented quickly.

Comment: It should A because Object2Vec is meant for text classification. The problem is to categorize the text based on the content.

Comment: B. LDA is for topic modelling based on categories. Comprehend is for extracting the entities related to sentiments etc.

Replies:

Comment: But it says the solution should not require ML expertise. LDA requires ML expertise.

Comment: Comprehend can be used for custom classification of NLP too(https://aws.amazon.com/ko/comprehend/features/). LDA can find document topics and word distribution for topics, but it is necessary to manually link the topics with predefined customer category.

Comment: The firm needs a solution for their manual process and this means among others the routing of their client orders. To do that you will need textract

---

Discussion for Question 178

Link: https://www.examtopics.com/discussions/amazon/view/75200-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 19 votes

Discussion

Comment: "Choose logarithmic scaling when you are searching a range that spans several orders of magnitude. For example, if you are tuning a Tune a linear learner model model, and you specify a range of values between .0001 and 1.0 for the learning_rate hyperparameter, searching uniformly on a logarithmic scale gives you a better sample of the entire range than searching on a linear scale would, because searching on a linear scale would, on average, devote 90 percent of your training budget to only the values between .1 and 1.0, leaving only 10 percent of your training budget for the values between .0001 and .1." based on the above from this link https://docs.aws.amazon.com/sagemaker/latest/dg/automatic-model-tuning-define-ranges.html C is clearly the answer

Comment: I would choose C: https://docs.aws.amazon.com/sagemaker/latest/dg/automatic-model-tuning-define-ranges.html

Replies:

> **Comment:** But according to the doc you gave, "Logarithmic scaling works only for ranges that have only values greater than 0." I think choosing ScalingType=Linear is the best fit, but there's no such option.

Comment: C is the way.

Comment: not A since you choose reverse logarithmic scaling when you are searching a range that is highly sensitive to small changes that are very close to 1.

Comment: It should be B. In logarithmic parameters the min value is the maximum value. This is the reason that C is not correct

Comment: "Choose logarithmic scaling when you are searching a range that spans several orders of magnitude. For example, if you are tuning a Tune a linear learner model model, and you specify a range of values between .0001 and 1.0 for the learning_rate hyperparameter, searching uniformly on a logarithmic scale gives you a better sample of the entire range than searching on a linear scale would, because searching on a linear scale would, on average, devote 90 percent of your training budget to only the values between .1 and 1.0, leaving only 10 percent of your training budget for the values between .0001 and .1."

Comment: B looks better , because learning rates were splited up base on a previous experience (0.1 - 0,01) in this case we are changing the structure . On the other hand A and B only change scaletype and this means no real changes

---

Discussion for Question 179

Link: https://www.examtopics.com/discussions/amazon/view/74972-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 24 votes

Discussion

Comment: C; Using S3 for scalable training and SageMaker Neo for compiling model for edge devices

Comment: A. NO - SageMaker endpoint does not address low-connectivity for inference B. NO - Train on premises does not address scalability for training C. YES - maximize training scalability and works with low-connectivity D. NO - Train on premises does not address scalability for training

Comment: The company needs a solution that minimizes costs for compute infrastructure and that maximizes the scalability of resources for training --> S3 The solution also must facilitate the company's use of an ML model in the low-connectivity environments.---> Edge devices and AWS IOT Greengrass

Comment: Answer is c

Comment: Moving the training data to an Amazon S3 bucket and training and evaluating the model by using Amazon SageMaker will reduce the company's compute infrastructure costs and maximize the scalability of resources for training. Optimizing the model by using SageMaker Neo will further reduce costs by allowing the model to run on inexpensive edge devices. Setting up an edge device in the manufacturing facilities with AWS IoT Greengrass and deploying the model on the edge device will enable the company to use the ML model in the low-connectivity environments. This solution provides a complete end-to-end solution for the company's needs, from data storage to model deployment, while minimizing costs and providing scalability and offline capabilities.

Comment: C best satisfies the options of minimising cost, and taking care of lack of connectivity through edge deployment.

Comment: Same arguments as belie

Comment: C: is the correct answer. Upload 20 years of massive data to S3 for model training. Sagemaker for creating and training a model. Once ready, deploy at edge using IOT Greengrass (this takes care of poor internet connectivity issue which is not addressed by option A)

Discussion for Question 180

Link: https://www.examtopics.com/discussions/amazon/view/75022-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- AC: 31 votes
- AE: 5 votes

Discussion

Comment: AC - for me https://aws.amazon.com/machine-learning/elastic-inference/ https://aws.amazon.com/blogs/machine-learning/configuring-autoscaling-inference-endpoints-in-amazon-sagemaker/

Replies:

> **Comment:** Agree. The problem with E is that it mentioned "majority of its traffic between 8 AM and 6 PM on weekdays", therefore it is not cost effective during period from 6PM to 8AM weekdays and weekends. Whereas the auto-scaling(C) could save money during all the time.

Comment: AC looks correct

Comment: traffic between 8 AM and 6 PM on weekdays in a single time zone .--- Reconfigure the endpoint to use burstable instances. Configure the endpoint to automatically scale with the InvocationsPerInstance metric.

Comment: A. YES - EI provides GPU access (Sept 2023: now deprecated) B. NO - not best practice to scale (altough it might help ?) C. YES - you want to scale with traffic D. NO - not best practice to scale (altough it might help ?) E. NO - burstable instances is for inpredictable traffic, bursting for long period of time is not cost effective

Comment: Either AC or CE very confusing but C is confirmed and will go with E

Replies:

> **Comment:** Or since the question is really old so maybe AC is what needs the answer to be

Comment: A C for me

Comment: A - C for me E is costy

Comment: E is not correct. A C E are both effective method to optimize inference, but A is used for GPU, and E is used for CPU. since this question is about Tensorflow, it should be A and E is not effective.

Comment: A and C are the right answers

Comment: Interesting if option A will be relevant anymore, as AWS is discontinuing Elastic Inference starting Apr 15. https://docs.aws.amazon.com/sagemaker/latest/dg/ei.html. Wonder if they change the option to include Inf instane type

Comment: Why elastic inference given that GPU is not necessary? I guess C and E (IMO)

Comment: A - This is using Tensorflow which means Elastic inference can be used to save costs for GPU, thereby reducing the compute time. E - Since the load is not uniform, it will help to use burstable instances to operate above the threshold when the situation demands. C is not at all cost effective.

Comment: AC is the most correct

Comment: I'd go AE if it requests for cost minimum

Comment: AC is correct

Comment: AC - for me

Discussion for Question 181

Link: https://www.examtopics.com/discussions/amazon/view/74970-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- D: 18 votes

Discussion

Comment: D; A involves too much effort and management overhead.

Replies:

Comment: Agree. but A has feature engineering which is the problem of the current model... confusing

Replies:

Comment: AutoML also contains feature engineering/preprocessing tools.

Comment: we don't use Logistic Regression to predict price.

Comment: A. NO - Logistic regression model is for classification, not to predict numerical values B. NO - approach is the highest quality, but takes time C. NO - XGBoost is for classification D. YES - simplest option

Comment: D for me also

Comment: It's D as rest require more operation activities.

Comment: D is Correct: trick to eliminate is A can not as Logistic is classification algo which gives binary outcome.B &C seems a lot of work .

Comment: The problem is not a classification problem so A is incorrect as logistic regression is used for binary problems. D is the correct solution

Comment: The problem is not which model you chose but primitive level feature engineering therefore correct answer should be "B"

Comment: D. https://aws.amazon.com/sagemaker/autopilot/ Supports missing values, categorical features, etc. The simplest solution for this case

Comment: why not B?

Discussion for Question 182

Link: https://www.examtopics.com/discussions/amazon/view/75045-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- CD: 29 votes

Discussion

Comment: Sentiment analysis is the result of analysis, not feature engineering. I think this answer should be C & D.

Comment: A. NO - it is categorization of words and thus inferencing, not pre-processing B. NO - Coreferencing (eg. linking "He" to "Mark" seen in a previous sentence) is a complex task, not pre-processing C. YES - it consists of reducing words to their base form to reduce dimensionality D. YES - fast pre-processing task E. NO - it is not feature engineering, it is training

Comment: C and D for me

Comment: C for merge similar words D for remove not important words like "the, is, a"

Comment: 12-sep exam

Comment: https://towardsdatascience.com/text-analysis-feature-engineering-with-nlp-502d6ea9225d Sentiment analysis IS a part of feature engg in NLP.

Comment: sentiment analysis is not part of feature engineering

Comment: I agree ABE are not feature engineering

Comment: A, B, E are not feature engineering

Comment: why not C & D

Discussion for Question 183

Link: https://www.examtopics.com/discussions/amazon/view/75279-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- D: 12 votes

Discussion

Comment: https://ts.gluon.ai/tutorials/forecasting/quick_start_tutorial.html

Replies:

Comment: https://ts.gluon.ai/stable/tutorials/forecasting/quick_start_tutorial.html

Comment: C is correct based on this blog - https://aws.amazon.com/blogs/machine-learning/training-debugging-and-running-time-series-forecasting-models-with-the-gluonts-toolkit-on-amazon-sagemaker/

Comment: D for me

Comment: D: A well-calibrated model should have quantile coverage close to the desired coverage level (e.g., 90% quantile coverage should be close to 90%). If the quantile coverage is consistently off from the desired level, it may indicate the need to recalibrate the model or investigate the sources of uncertainty estimation errors.

Comment: https://apps.microsoft.com/store/detail/move-mouse/9NQ4QL59XLBF?hl=en-us≥=us

Comment: I think it is D

Comment: Thanks to ChatGPT Given the coverage score results, the data scientist can conclude that the distributional forecast related the test set is well calibrated. Specifically, when the model predicts quantiles, around % of the true values should fall within the 0.5 quantile range, and around90% of the true values should fall within the 0.9 quantile range., the GluonTS on Amazon SageMaker DeepAR model performance on the test set was concerning the coverage of the predicted quantiles.

Replies:

> **Comment:** My chatgpt is latest: The coverage of a distributional forecast at a given quantile is the fraction of observations that fall below the predicted quantile. In a well-calibrated forecast, the coverage score should be approximately equal to the quantile itself. Given the information: Coverage score is 0.489 at the 0.5 quantile. Coverage score is 0.889 at the 0.9 quantile. For a well-calibrated forecast: At the 0.5 quantile (or median), the coverage should be approximately 0.5. At the 0.9 quantile, the coverage should be approximately 0.9. The provided coverage scores closely match the quantiles, with slight deviations. Therefore, the correct conclusion is: Option D: The coverage scores indicate that the distributional forecast is correctly calibrated. These scores should be approximately equal to the quantile itself.

Comment: Scores should always fall below the quantile itself. Ref: https://d1.awsstatic.com/asset-repository/Amazon%20Forecast%20Technical%20Guide%20to%20Time-Series%20Forecasting%20Principles.pdf -- Pg 18

Replies:

> **Comment:** https://docs.aws.amazon.com/forecast/latest/dg/metrics.html#metrics-wQL A more concise doc.
>
> **Comment:** PDF Pg 23

Comment: C is correct

Discussion for Question 184

Link: https://www.examtopics.com/discussions/amazon/view/75149-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 16 votes

Discussion

Comment: Answer is C. B, D wrong because Kinesis data stream cannot write to S3 directly.

Comment: A. Not enough. C. Correct. Check https://docs.aws.amazon.com/firehose/latest/dev/writing-with-iot.html B, D: Wrong. KDS doesn't write directly to S3

Comment: A. NO - HTTP is not best protocol for IoT B. NO - No need to buffer write from Firehose to S3 with Kinesis/Kafka in the middle C. YES - Firehose is a good connector MQTT to S3 D. NO - Kinesis/Kafka cannot intake MQTT out-of-the-box, Firehose is the right connector

Comment: Answer is C

Comment: C is sorrect

Discussion for Question 185

Link: https://www.examtopics.com/discussions/amazon/view/74925-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 23 votes

Discussion

Comment: The answer is B. In multi-label mode, individual classes represent different categories, but these categories are not mutually exclusive while individual classes are mutually exclusive in multi-class mode

Comment: B - In simple language, it is tagging

Comment: A. NO - multi-class means more than binomial/2 classes possible targets, but still the document belongs to only 1 B. YES - multiple class can be assigned (eg. using SoftMax for different probabilities) C. NO - it is about assigning Entities to terms in the input documents, not classifying the documents D. NO - you need to customize the classes

Comment: Answer b

Comment: In multi-label classification, individual classes represent different categories, but these categories are somehow related and are not mutually exclusive. As a result, each document has at least one class assigned to it, but can have more. For example, a movie can simply be an action movie, or it can be an action movie, a science fiction movie, and a comedy, all at the same time. In multi-class classification, each document can have one and only one class assigned to it. The individual classes are mutually exclusive. For example, a movie can be classed as a documentary or as science fiction, but not both at the same time.

Comment: https://docs.aws.amazon.com/comprehend/latest/dg/prep-classifier-data-multi-label.html

Comment: 12-sep exam

Comment: B. Multi-label: https://docs.aws.amazon.com/comprehend/latest/dg/prep-classifier-data-multi-label.html

Comment: Written in the technical guide related but not mutually exclusive

---

Discussion for Question 186

Link: https://www.examtopics.com/discussions/amazon/view/74999-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- BCF: 19 votes

Discussion

Comment: BCF; This tutorial doc says so: https://docs.aws.amazon.com/glue/latest/dg/dev-endpoint-tutorial-sage.html

Comment: BDF is correct

Comment: when creating notebook why do we need Glue development endpoint? it should be D

Comment: A. YES - By requirement, the notebook must be in a VPC B. NO - Data is already in S3, we do not need to know it was made with AWS Glue C. NO - Data is already in S3 thanks to AWS Glue, no runtime relationship with SageMaker D. YES - need to create the Notebooks at some point E. NO - no need to decrypt, it is about ACL F. YES - notebooks need to be able to read S3

Replies:

> **Comment:** Sorry, B is true. Scripts must be accessible they say. Then A is wrong ?

Comment: Creating a SageMaker development endpoint in the data science team's VPC will allow the data science team to access the ETL scripts and the AWS Glue job from within their VPC. Creating an IAM policy and an IAM role for the SageMaker notebooks will allow the data science team to access the ETL scripts and the AWS Glue job with the appropriate permissions. Creating SageMaker notebooks by using the SageMaker console will allow the data science team to easily create and manage the SageMaker notebooks.

Comment: In the AWS Glue console, choose Dev endpoints to navigate to the development endpoints list. Select the check box next to the name of a development endpoint that you want to use, and on the Action menu, choose

Create SageMaker notebook. Fill out the Create and configure a notebook page as follows: Enter a notebook name. Under Attach to development endpoint, verify the development endpoint. Create or choose an AWS Identity and Access Management (IAM) role.

Comment: BCF is correct

Discussion for Question 187

Link: https://www.examtopics.com/discussions/amazon/view/89019-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 21 votes

Discussion

Comment: agreed with C

Comment: The answer was between C and D, but we are suppose to minimize use of Redshift cluster, answer is C. And B are too much effort, so not to be done as per constraints of question.

Comment: Simply the requirements are a full ETL process where data will be extracted from Redshift (E), then transformed by renaming, removing null values, or even separating the first column So (T), and finally load data to S3(L) all that with the least overhead, which make the AWS Glue ideal for these requirements

Comment: A. NO - AWS Glue (serverless) is a simpler option than EMR to run Spark jobs B. NO - Spark is a better option for datapipelines, it avoids the need for intermediary files C. YES - Spark and AWS Glue best combination D. NO - Amazon Redshift Spectrum is a "Lake House" architecture, meant to run SQL against against both DW & S3; here, we want to query only from the DW

Comment: The reason is that this solution can leverage the existing capabilities of AWS Glue, which is a fully managed service that can help users create, run, and manage ETL (extract, transform, and load) workflows. According to the web search results, AWS Glue can connect to various data sources and destinations, such as Amazon Redshift and Amazon S3, and use Apache Spark as the underlying processing engine. AWS Glue can also provide various built-in transforms that can perform common data manipulation operations, such as filtering, mapping, renaming, or joining. Moreover, AWS Glue can support scheduling and automation of ETL jobs using triggers or workflows.

Comment: agree with C

Comment: C Reason: we want to minimize infrastructure effort, so we should prioritize serverless solutions, we want something automated and minimize the load on the Redshift cluster. That said, Letter A is wrong as it uses a managed service (EMR) just like Letter B (EC2). Letter D brings Redshift Spectrum, however the base is not in S3, but Redshift! So, it's discarded this option, since we use this service to move data from S3 → Redshift using SQL. Letter C is correct.

Comment: https://docs.aws.amazon.com/glue/latest/dg/aws-glue-programming-python-transforms.html

---

Discussion for Question 188

Link: https://www.examtopics.com/discussions/amazon/view/88925-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 11 votes

Discussion

Comment: https://docs.aws.amazon.com/sagemaker/latest/dg/your-algorithms-training-algo-dockerfile.html

Comment: The /opt/ml directory is the default directory where SageMaker expects the training script and other related files to be located. The script at location above is triggered by setting environment variable SAGEMAKER_PROGRAM and *not* through an ENTRYPOINT in docker file

Comment: A. NO - There is no server here, we do training not inference B. YES C. NO - path to training data is externally provided, not hardcoded in the image D. NO - /opt/ml/train is the working directory of the ENTRYPOINT

Comment: Amazon SageMaker supports bringing custom training algorithms by using Docker containers, which are software packages that can contain all the dependencies and configurations needed to run an application. Dockerfile is a text file that contains the instructions for building a Docker image, which is a snapshot of a Docker container. ENTRYPOINT is an instruction in the Dockerfile that specifies the default executable or command that will run when the container is started. By specifying the training program in the ENTRYPOINT instruction, the ML specialist can ensure that Amazon SageMaker can run the training program automatically when it creates and runs a Docker container for the training job.

Comment: https://docs.aws.amazon.com/sagemaker/latest/dg/adapt-training-container.html In Step 2, it is mentioned to use this instruction on dockerfile: # Defines train.py as script entrypoint ENV SAGEMAKER_PROGRAM train.py

---

Discussion for Question 189

Link: https://www.examtopics.com/discussions/amazon/view/88731-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- AD: 19 votes

Discussion

Comment: AD are the right answer

Comment: A. YES - IP Insights works unsupervised on IP addresses; builtin algorithm B. NO - k-NN is unsupervised clustering, does not help with anomalities C. NO - Linear learner is supervised D. YES - Random Cut Forest (RCF) is unsupervised anomalities E. NO - XGBoost is supervised

Comment: IP Insights is an unsupervised learning algorithm that learns the usage patterns of IP addresses. It can capture associations between IP addresses and various entities, such as user IDs or account numbers. It can also identify anomalous events, such as a user attempting to log in from an unusual IP address, or an account that is creating resources from a suspicious IP address1. Random Cut Forest (RCF) is another unsupervised algorithm for detecting anomalous data points within a dataset. It can handle arbitrary-dimensional input and scale well with respect to number of features, data set size, and number of instances. It can detect anomalies such as unexpected spikes in time series data, breaks in periodicity, or unclassifiable data points2.

Comment: Can't be A, as we don't have data in the format expected for IP Insights algorithm(https://docs.aws.amazon.com/sagemaker/latest/dg/ip-insights-training-data-formats.html).

Replies:

> **Comment:** It expects CSV format and the question mentions data is in CSV format so IP Insights is correct

Comment: A and D are correct. A. IP Insights for Pattern recognition. D. Random Cut Forest (RCF) for Anomaly detection B,C,E are normally Supervised learning algorithm which are against the wordings "There is no label ..."

Comment: C is not part of the answer. IP insight because the data contain IP address. RCF because the data is unlabeled and anomaly is being detected for fraud.

Comment: AD are correct

Comment: apprently AD

Comment: AC is the correct answer to detect anomalies

Discussion for Question 190

Link: https://www.examtopics.com/discussions/amazon/view/89092-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- AC: 13 votes

Discussion

Comment: A and C seems fine

Comment: I think the answer is CD.

Replies:

Comment: Agree. The setting should be relevant to S3 and VPC, not the notebook.

Replies:

Comment: if notebook is not within vpc, then having s3 bucket policy to allow traffic only from vpc will block notebook to get data from s3. A C

Comment: A is wrong. SageMaker notebook does not need to have internet access disabled.

Comment: CD is right

Comment: AD is the answer While creating Sagemaker notebook instances we have to decide on the access (via VPC and/or direct internet). Here we will select access only from VPC. The same VPC should become a requirement to access S3 bucket via S3 bucket policy. C would have been fine but fails to mention creation of S3 access points and those access points can be restricted to VPC.

Comment: A&C as explained in this blog as well - https://aws.amazon.com/blogs/machine-learning/secure-amazon-s3-access-for-isolated-amazon-sagemaker-notebook-instances/

Comment: A. NO - Notebook must run in a VPC (SageMaker will provision an instance), but with a private subnet there is no need to disable internet traffic B. NO - VPN tunnel is to encrypt traffic with the Internet C. YES - Endpoint will prevent S3 traffic to flow over the internet D. YES - Create an S3 bucket policy that allows traffic from the VPC and denies traffic from the internet. E. NO - AWS Transit Gateway is for multiple VPCs

Comment: By configuring the SageMaker notebook instance to be launched with a VPC attached and internet access disabled, the data scientists can access the resources within the VPC, such as Amazon EFS or Amazon EC2 instances, without exposing them to the internet1. This also prevents the notebook instance from accessing any resources outside the VPC, such as Amazon S3, unless a VPC endpoint is configured2. By creating and configuring an S3 VPC endpoint and attaching it to the VPC, the data scientists can access the datasets stored in Amazon S3 from the notebook instance using private IP addresses. The S3 VPC endpoint is a gateway endpoint that routes the traffic between the VPC and Amazon S3 within the AWS network, without requiring an internet gateway or a NAT device3. This also enhances the security and performance of the data access1.

Comment: CD : Question is about make S3 Data not accessible from Internet & VPC Endpoint Only.

Replies:

Comment: S3 by default is not public, you don't have to deny traffic from internet. Just not make it public.

Comment: the requirements are about providing secure access from notebooks to S3, nothing else.

Comment: the answer is C,D. Firstly, the VPC need to connect with S3 through gateway endpoint, check "https://docs.aws.amazon.com/vpc/latest/privatelink/vpc-endpoints-s3.html" secondly, after connection is created, we need to define the policy from s3 side. restrict access to s3 only from specified VPC or VPC endpoint. "https://docs.aws.amazon.com/vpc/latest/privatelink/vpc-endpoints-s3.html#bucket-policies-s3" the confusion about A is tricky. Ideally, you need to create sagemaker in private subnet with no internet access. But I assume the question "access to the data from instances and services" only requires the process from obtaining data from s3, you don't need to specify the requirement about data egress from training service(even though disable internet connection from sagemaker is crucial)

Comment: A and C are correct. To disable direct internet access, you can specify a VPC for your notebook instance. By doing so, you prevent SageMaker from providing internet access to your notebook instance. As a result, the notebook instance can't train or host models unless your VPC has an interface endpoint (AWS PrivateLink) or a NAT gateway and your security groups allow outbound connections. https://docs.aws.amazon.com/sagemaker/latest/dg/appendix-notebook-and-internet-access.html https://docs.aws.amazon.com/sagemaker/latest/dg/notebook-interface-endpoint.html D is wrong. Bucket policy cant be used to deny internet access. It can only enforce access from VPC or VPC endpoint

Replies:

Comment: your statement "Bucket policy cant be used to deny internet access" is completely wrong, you can either specify "Allow" or "Deny" in bucket policy, check "https://docs.aws.amazon.com/vpc/latest/privatelink/vpc-endpoints-s3.html#bucket-policies-s3" You can create a bucket policy that restricts access to a specific endpoint by using the aws:sourceVpce condition key.

Comment: You can use Amazon S3 bucket policies to control access to buckets from specific virtual private cloud (VPC) endpoints, or specific VPCs. This section contains example bucket policies that can be used to control Amazon S3 bucket access from VPC endpoints. Notebook doesn't need to be created within Vpc

Replies:

Comment: The question requires to access to the data from instances and services used for training must not be transmitted over the internet. So the traffic has to go through the VPC endpoints, thus the notebook has to live in the VPC.

Discussion for Question 191

Link: https://www.examtopics.com/discussions/amazon/view/89093-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- AC: 11 votes
- AE: 8 votes

Discussion

Comment: I think the answer is BC

Comment: BC - https://docs.aws.amazon.com/forecast/latest/dg/aws-forecast-recipe-ets.html

Comment: A is for sure. C, watch the keywords mentioned in the question "too low or too high"!

Comment: Ans: AC A. Add Information About the Store's Sales Periods: This directly targets the issue of seasonality affecting the sales forecast. C. Apply Smoothing to Correct for Seasonal Variation: Smoothing techniques will help in handling the seasonal trends more effectively, which seems to be a major factor in the model's current performance issues.

Comment: both of A and C solves the seasonality issues

Comment: A&E for sure. For option E, refer to this blog - https://aws.amazon.com/blogs/machine-learning/prepare-time-series-data-with-amazon-sagemaker-data-wrangler/

Comment: A. YES - valuable contextual information B. NO - irrelevant to seasonal events C. YES - Removes noise and can help make patterns easier to identify D. NO - not point to loose precious information such as weekend days E. NO - 5% data loss is not a big deal, might as well drop them

Comment: Adding information about the store's sales periods to the dataset can help the model learn about patterns in sales that are specific to certain times of year. This can help the model make more accurate predictions around seasonal events. ---------- Smoothing can help correct for seasonal variation by removing some of the noise from the data. This can help the model make more accurate predictions ---------- None of the other options address the seasonal variation in my opinion

Comment: B&C were my top choices without looking at the key.

Comment: E for sure then either C or A. would go for A

Comment: Answer: AC Smoothing has different uses. Please find the definition Data smoothing can be defined as a statistical approach to eliminating outliers from datasets to make the patterns more noticeable.

Comment: ChatGPT say C + E

Comment: CE C. Apply smoothing to correct for seasonal variation: Seasonal variation can have a significant impact on sales data. By applying smoothing techniques such as moving averages or exponential smoothing, the ML specialist can reduce the noise and fluctuations caused by seasonal effects, allowing the model to capture the underlying patterns more effectively. E. Replace missing values in the dataset by using linear interpolation: Missing data can introduce biases and affect the accuracy of the model. Linear interpolation is a common technique for filling in missing values by estimating the missing data points based on the available data. By replacing the missing values, the ML specialist ensures that the model has a complete and representative dataset to learn from.

Comment: A to improve model in seasonal periods E to fill missing data

Comment: I would go for C and E C is quite obvious I think E Linear interpolation is a technique to fill the missing data https://towardsdatascience.com/4-techniques-to-handle-missing-values-in-time-series-data-c3568589b5a8

Comment: why not C and D?

Replies:

Comment: Maybe the sales event can last longer than a week?

Discussion for Question 192

Link: https://www.examtopics.com/discussions/amazon/view/88927-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 23 votes

Discussion

Comment: B is correct. IMO A - No, Random cut forest is for anomaly detection B - Yes, exactly was XGBoost is good for. Binary classification based on a variety of input features C - No, NTM is unsupervised. The problem states the table already has subscription status, therefore we need a supervised algorithm D - No, DeepAR is used for time-series data

Comment: Whether subscription status is binary or multi-class XGBoost can handle the problem in this case problem.

Comment: A. NO - Random Cut Forest (RCF) used for anomalities B. YES - XGBoost is good for classification C. NO - Neural Topic Model (NTM) is to find topics, not classify D. NO - that is for timeseries

Comment: XGboost

Comment: Letter B is correct as we have a supervised classification problem here.

Comment: XGboost for Binary classification

Comment: XGBoost is a popular and powerful algorithm for binary classification problems such as this one, where the goal is to predict a binary outcome (e.g. whether a customer subscribes or not). It is particularly effective when the dataset has a mix of numerical and categorical features.

Comment: The answer is B: A. Random Cut Forest (RCF): anomaly detection B. XGBoost: allows classification tasks like the use case in the question C. Neural Topic Model (NTM): topic modelling D. DeepAR forecasting: time series

Comment: I think the answer is B. It looks like no time serials condition, so it may be not suitable to A and D.

Comment: D. Refer to https://docs.aws.amazon.com/forecast/latest/dg/aws-forecast-recipe-deeparplus.html

Replies:

Comment: can you please let us know, where in question it states that date/time column is available. It is all about numerical or categorical columns and we need to predict subscription status which can be done by the use XG-BOOST

Replies:

Comment: you're right. there is no time series.

Discussion for Question 193

Link: https://www.examtopics.com/discussions/amazon/view/88928-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 12 votes

Discussion

Comment: It seems to be B

Comment: A. NO - you don't want temporary access B. YES - best practice C. NO - CloudHSM is overkill vs. KMS D. NO - transient keys are transient

Comment: AWS Key Management Service (AWS KMS) is a service that allows customers to create and manage encryption keys that can be used to encrypt data at rest in AWS services. AWS KMS provides a high level of security and control over the encryption keys, as well as integration with AWS CloudTrail to log key usage. By using customer managed keys in AWS KMS, the company can encrypt the storage volumes for all SageMaker instances, such as notebook instances, training instances, and endpoint instances. This can be done by specifying the KMS key ID when creating or updating the instances. The company can also encrypt the model artifacts and data in Amazon S3 by using the same or different KMS keys. This can be done by enabling server-side encryption with KMS keys when creating or updating the S3 buckets or objects.

Comment: Letra B. Normalmente, falou de criptografia na AWS, falou de KMS.

Comment: Why not "A"? I think using AWS STS to create temporary tokens is easier than create custom AWS KMS key.

Replies:

Comment: STS is for other purpose.

Comment: why no B?

Replies:

Comment: It should be B. This question is the same as Q143.

Discussion for Question 194

Link: https://www.examtopics.com/discussions/amazon/view/89135-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 18 votes

Discussion

Comment: I think the answer is B. The others are quite expensive and complicated.

Comment: A. NO - B is easier B. YES - works natively against S3 C. NO - no need to import S3 data to Redshift when Presto/Athena allows you to query directly D. NO - Redshift overkill

Comment: AWS Glue

Comment: We want to use SQL to explore the 500GB database saved in S3. We also want to minimize costs and have the least headache with the operation. Letters A - C - D mean managed services, hence: headache. Correct alternative is letter B.

Comment: B as "LEAST operational overhead"

Comment: The advantage of D is that Redshift, as a data warehouse, can handle large dataset(>1TB) and complex frequent query. In this example, 500GB dataset and infrequent query(I consider this just one-time ad-hoc query, just verify the data before training.) Athena would be a much better option.

Comment: The option that meets these requirements with the LEAST operational overhead is option B: Use AWS Glue to crawl the S3 bucket and create tables in the AWS Glue Data Catalog. Use Amazon Athena to explore the data. AWS Glue is a fully managed ETL service that can automatically discover and catalog metadata about data stored in various data stores, including Amazon S3. By using AWS Glue to crawl the S3 bucket, the data scientist can easily create tables in the AWS Glue Data Catalog, without needing to create or manage any infrastructure. Amazon Athena is an interactive query service that allows querying data stored in Amazon S3 using SQL. By using Amazon Athena, the data scientist can easily explore the data using SQL, without needing to set up any infrastructure.

Comment: Both Glue and Athena are serverless hence cost effective.

Comment: B is highly managed unlike other options.

Comment: It seems to be B

Discussion for Question 195

Link: https://www.examtopics.com/discussions/amazon/view/89137-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 16 votes

Discussion

Comment: Agreed with B https://aws.amazon.com/blogs/machine-learning/using-pipe-input-mode-for-amazon-sagemaker-algorithms/

Comment: Pipe input mode.

Comment: The correct solution would be to modify the training configuration to use Pipe input mode. This will allow the training data to stream directly into the training instance as it is being consumed, rather than first being downloaded from S3 into the instance's local storage. This can help reduce storage requirements and optimize performance, while also minimizing costs. Using more or larger instances may help with processing power, but it will not address the storage issue, and may even increase costs unnecessarily. Using Amazon EFS may also be an option, but it may come with additional costs and operational overhead.

Replies:

> **Comment:** the explanation of EFS is not correct. EFS is the default storage for sagemaker instance. That means, when you use the file mode, data is firstly copied to EFS and then fit model. So the issue" model is failing because of lack of storage" indicates EFS is not capable to store all s3 data. We have to use pipe mode to incrementally send data from s3 to EFS.

Comment: Pipe mode solves the problem without incurring extra storage cost. Data is streamed directly to the training algorithm without the need to be stored in the EBS volume.

Comment: I think the answer is B. D is incorrect because EFS is more expensive than S3. It looks like scaling up or out is no help for storage issue. Therefore A and C are not helpful.

Discussion for Question 196

Link: https://www.examtopics.com/discussions/amazon/view/88943-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 12 votes

Discussion

Comment: I think the answer is A. B: Snowcone has limit with 8 TB. C: is AWS on-premises solution, but the company wants to store these transcriptions in an Amazon S3 bucket in the AWS Cloud for model development. D: 100 Mbps cannot handle petabytes datasync.

Comment: A: is the Right Folks, please... OUTPOST is very complicated to implement, and the question is not talking about continue to do it after. We know that Snowball has limite of storage, but the ideia is not to send Petabytes of data to S3, just only the likelihood processed. so, the ideia is to use the Snowball as an Optimized machine to be able to process the data and send it to S3. The Snowball Edge Compute Optimized device provides 52 vCPUs, 208 GiB of memory, and an optional NVIDIA Tesla V100 GPU. For storage, the device provides 42 TB usable HDD capacity for Amazon S3 or Amazon EBS, as well as 7.68 TB of usable NVMe SSD capacity for EBS block volumes. Snowball Edge Compute Optimized devices run Amazon EC2 sbe-c and sbe-g instances, which are equivalent to C5, M5a, G3, and P3 instances.

Comment: A. NO - Snowball needs to be shipped back to AWS, that does not use Datasynch B. YES - that is an edge computing device C. NO - Too big admin overhead to have your local AWS Cloud D. NO - Too slow over the 100Mps connection

Comment: faster

Comment: A, as this is faster option.

Comment: A ==> Bring the compute closer to the data

Comment: The key is faster Transfer speeds of up to 100gb per second

Comment: It either A or C. But C is too complicated to setup. You order rack and AWS installs that plus you need enterprise support and the biggest reason it is not possible in this case is that it requires at least 1GB connection. The question clearly asks as soon as possible so A is the best choice in my opinion

Comment: The key is to deliver the transcripts to S3 as early as possible. Outpost order and provisioning takes months. I would go for A as its logical to do local inference and send transcripts to S3.

Comment: The model needs to run on-premises to don't be necessarily upload all the audio data to after run the model and this solution can use datasync to upload the results after too, then it's a good choice!

Comment: Outpost for use case when customer don't want to transfer their data out

Comment: It is A. C can't be the answer, as to transfer 1PB data, it may take 1,000 days under a 100 mbps network.

Comment: My vote goes to C, A. Snowball device only stores around 80 TB of data and uploading the newly transcribed data through datasync still goes through the slow connection between on-site and AWS. Outpost seem like the only feasible solution here that can satisfy both requriements

Comment: Correct Answer A. D is incorrect as it takes 1024 Days Approx to transfer Petabyte of data.

Comment: The best solution would be option A and D. Option A: Order and use an AWS Snowball Edge Compute Optimized device with NVIDIA Tesla modules to run the transcription algorithm. Use AWS DataSync to send the generated transcriptions to a transcription S3 bucket. This option allows to use a device that has the necessary GPU for running the transcription algorithm and then use the AWS DataSync to send the generated transcriptions to the S3 bucket. Option D: Use AWS DataSync to ingest the audio files to Amazon S3. Create an AWS Lambda function to run the transcription algorithm on the audio file as it is uploaded to Amazon S3. Configure the function to write the generated transcriptions to the transcriptions S3 bucket. This option allows to automatically transcribing the audio files as they are uploaded to S3. This means that the transcriptions are ready as soon as the audio files are uploaded and eliminates the need to transcribe the audio files separately.

Comment: ws.amazon.com/about-aws/whats-new/2020/07/aws-snowball-edge-compute-optimized-now-available-additional-aws-regions/

Comment: I guess C, given that "The company wants to store these transcriptions". Petabytes audio data can keep in on-prem.

Discussion for Question 197

Link: https://www.examtopics.com/discussions/amazon/view/89148-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 19 votes
- B: 19 votes

Discussion

Comment: B, but it was possible to use Kinesis Data Firehose directly, insted Kinesis Data Stream

Comment: I think the answer is B.

Comment: Moving window, and less components

Comment: C: Doesn't talk how to Store the data in Amazon S3

Comment: With Amazon Kinesis Data Analytics for Apache Flink, the ML specialist needs to manage the scaling and resource allocation for the Flink application, including determining the appropriate number of processing units (KPUs) and handling scaling based on the incoming data volume. This requires monitoring and adjusting resources as needed, adding to the operational overhead.

Comment: C is the correct answer. B is workable but is not good for small transformation required in question.

Comment: C Amazon Managed Service for Apache Flink was previously known as Amazon Kinesis Data Analytics for Apache Flink. it allows you to process and analyze streaming data providing the capability to perform transformations on the streaming data. B - no need of using an extra service aws glue

Comment: I would choose C. As we need to implement our detection on running window, and B only allows us to perform operations on the latest 10 minutes of data. If we choose B, we also need to decide how frequently to run the Glue job and it involves some orchestrator tools. C in other way works in real-time mode, and we don't need an orchestration tool to move the window. Based on this, I would go with C as it has less overhead

Comment: Would choose see as the transformation required is minimal which could be easily achieved with KDA (flink job)

Comment: Not sure between B & C A. NO - too many moving parts B. YES - clean & elegant C. YES - works as well in batch mode D. NO - MSK is outdated

Comment: https://aws.amazon.com/blogs/architecture/realtime-in-stream-inference-kinesis-sagemaker-flink/

Comment: Amazon Kinesis Data Streams is a fully managed real-time streaming service that can be used to ingest large amounts of data from multiple sources. This makes it a good choice for ingesting the event data from the podcast platform. Amazon Kinesis Data Analytics for Apache Flink is a fully managed service that can be used to process streaming data using Apache Flink. Apache Flink is a popular streaming processing framework that is known for its scalability and fault tolerance. This makes it a good choice for transforming the event data before inference.

Comment: It's B, ""LEAST operational overhead"", C is more operations overhead.

Replies:

Comment: No. it's not true. for running B we need some orchestrator to run the glue job frequently. but for C it is running constantly. So C doesn't have step with orchestration

Comment: Flink distributes the data across one or more stream partitions, and user-defined operators can transform the data stream.

Comment: B is the answer. least management overhead

Replies:

Comment: And for C you have to author and build your Apache Flink application. extra work

Replies:

Comment: And for Glue you need to write a SQL or spark script. Extra work. Ah, yes. and for B you need to create an orchestrator to run the ETL jobs frequently

Comment: Answer should be C. https://aws.amazon.com/blogs/architecture/realtime-in-stream-inference-kinesis-sagemaker-flink/

Replies:

Comment: Flink distributes the data across one or more stream partitions, and user-defined operators can transform the data stream.

Comment: It's C, this data requires small transformations who can be did with apache flink in kinesis data analytics. If wasn't this, then could be using Glue.

Discussion for Question 198

Link: https://www.examtopics.com/discussions/amazon/view/89150-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 19 votes

Discussion

Comment: yes, It seems to be 'B'

Replies:

Comment: agreed, shadow testing is supported on SageMaker. https://aws.amazon.com/cn/blogs/aws/new-for-amazon-sagemaker-perform-shadow-tests-to-compare-inference-performance-between-ml-model-variants/

Replies:

Comment: Not sure if the shadow testing fits into the purpose here. "The company has developed three versions of a machine learning (ML) model within Amazon SageMaker to classify document text."

Comment: B is correct, from within single endpoint, we can create multiple production variant. When lambda called, it should have been each target variant instead of production variant in the verbiage

Comment: C is fine B is not possible as it is single sagemaker endpoint (so we won't get prediction from all models for each document) D is wrong as we do not need three lambda functions A is also wrong as time gap is 3 seconds for which we should be running batch transform jobs

Comment: Will go with B

Comment: A. NO - you don't want to create a new job for each Lambda invokation B. YES - best practice C. NO - could work but does not leverage production variants which in-turn disable some built-in model performance evaluation features D. NO - more operationnal overhead to have 3 endpoints

Comment: Answer is B

Comment: Although C sounds like a better option but B is less operational overhead at least for short term.

Comment: It's B, you can use Invoke a Multi-Model Endpoint, when you call invoke_endpoint you need to provide which model filw to use. response1 = runtime_sagemaker_client.invoke_endpoint(EndpointName = "MAIN_ENDPOINT", TargetModel = "model1.tar.gz", Body = body) response2 = runtime_sagemaker_client.invoke_endpoint(EndpointName = "MAIN_ENDPOINT", TargetModel = "model2.tar.gz", Body = body) response3 = runtime_sagemaker_client.invoke_endpoint(EndpointName = "MAIN_ENDPOINT", TargetModel = "model3.tar.gz", Body = body) Ref: https://docs.aws.amazon.com/sagemaker/latest/dg/invoke-multi-model-endpoint.html

Comment: B - the reason is not shadow testing since it is not named and does not require client logic. The reason is that it is possible to target a model https://docs.aws.amazon.com/sagemaker/latest/dg/invoke-multi-model-endpoint.html

Comment: B it is which involves using single endpoint for multiple model versions

Comment: I think the answer should be C. As there is no production version of the model identified, all the 3 models need to be invoked.

Comment: C, prod variant is used for traffic routing. All model needs to be invoked.

Comment: C is correct

Comment: I think the answer is B.

Discussion for Question 199

Link: https://www.examtopics.com/discussions/amazon/view/89151-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- AE: 25 votes

Discussion

Comment: KNN can be used for dimensionality reduction through NCA (https://scikit-learn.org/stable/auto_examples/neighbors/plot_nca_dim_reduction.html#)

Comment: A. Correct B. Incorrect. MDS is Non-linear dimensionality reduction method. https://towardsdatascience.com/11-dimensionality-reduction-techniques-you-should-know-in-2021-dcb9500d388b C. Incorrect. This is a classification problem instead of Regression. D. Incorrect. K-means is for Clustering(Unsupervised learning). E. Correct.

Replies:

Comment: E : https://docs.aws.amazon.com/sagemaker/latest/dg/k-nearest-neighbors.html

Comment: Why "Non-linear dimensionality reduction method" is a problem? We can add non linear features as x^2 to a model to improve performance.

Comment: B is wrong as Multidimensional Scaling sits under the Unsupervised branch of Machine Learning algorithms

Comment: A. YES - F1 score is low. Reducing feature count could improve F1 score. B. NO - MDS is for visualization C. NO - regressor is to predict a numerical value, we want classification D. NO - K-means is clustering, we want classification E. YES - k-Means could work if a linear model is not best

Comment: A & E are correct

Comment: A and B are correct. But I understand E being correct as well.

Comment: https://docs.aws.amazon.com/sagemaker/latest/dg/k-nearest-neighbors.html

Comment: I think the answer is AB. k-means and k-nearest neighbors are not for reduce dimension.

Replies:

Comment: k-nn does. https://docs.aws.amazon.com/sagemaker/latest/dg/k-nearest-neighbors.html

---

Discussion for Question 200

Link: https://www.examtopics.com/discussions/amazon/view/88924-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 20 votes

Discussion

Comment: It's B https://aws.amazon.com/premiumsupport/knowledge-center/sagemaker-endpoint-latency/

Replies:

Comment: false. its C. in the link you shared under High ModelLatency, it states "If an endpoint is overused, it might cause higher model latency. You can add Auto scaling to an endpoint to dynamically increase and decrease the number of instances available for an instance."

Replies:

Comment: Autopilot is is not autoscaling in AWS. Autopilot is for model training, Autoscaling is during inference.

Comment: C is the most wrong solution.

Comment: A. NO - that is image processing so more CPU would only provide incremental improvement B. YES - that is image processing so GPU would provide a step change; supported by the built-in algorithm C. NO - Autopilot is for training, not inference D. NO - usually inference uses little memory

Comment: Attach an Amazon Elastic Inference ml.eia2.medium accelerator to the endpoint instance. Amazon Elastic Inference allows users to attach low-cost GPU-powered acceleration to Amazon EC2 and SageMaker instances or Amazon ECS tasks, to reduce the cost of running deep learning inference by up to 75%

Comment: B is not correct anymore. After April 15, 2023, new customers will not be able to launch instances with Amazon EI accelerators in Amazon SageMaker, Amazon ECS, or Amazon EC2. (https://docs.aws.amazon.com/sagemaker/latest/dg/ei.html)

Replies:

Comment: Changes in exams apply 6 months after the change as been applied (oct 2023)

Replies:

Comment: Freak!

Comment: It's B. Burstable instances only solves when lot of users are making inferences at the same time

Replies:

Comment: bro, isn't this exactly the question is asking for?

Comment: The ModelLatency metric shows that the model inference time is causing the latency issue. Amazon Elastic Inference is designed to speed up the inference process of a machine learning model without needing to deploy the model on a more powerful instance. By attaching an Elastic Inference accelerator to the endpoint instance, the ML specialist can offload the compute-intensive parts of the inference process to the accelerator, resulting in faster inference times and lower latency.

Comment: B - https://aws.amazon.com/premiumsupport/knowledge-center/sagemaker-endpoint-latency/

Comment: Elastic Inference accelerator (and AutoScaling but there is no autoscaling in the option). Be aware that Autopilot is is not autoscaling.

Discussion for Question 201

Link: https://www.examtopics.com/discussions/amazon/view/89262-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 13 votes

Discussion

Comment: A. NO - Must use SageMaker Debugger for visibility into model insights B. NO - Hyperparameters will most likely influence model accuracy but not response time C. YES - SageMaker Debugger is the right tool for model insights; filter (or "kernels") slides in CNN to identify specific features D. NO - SageMaker Model Monitor is for model performance

Comment: Pruning is a technique that reduces the complexity of convolutional neural networks (CNNs) by removing unimportant filters or neurons. This can lead to faster inference times and lower memory consumption, which are desirable for self-driving applications. Pruning can be done by ranking the filters based on some criteria, such as the norm of the weights, the activation outputs, or the Taylor expansion of the loss function123.

Comment: ChatGPT is an awesome tool, but please ML colleagues: study!

Replies:

> **Comment:** you are very right, about how awesome ChatGPT, but since we find it's answers over here, so some colleagues are trying to help us in proving why these could be the right answers without wasting our time to prove it. All the names over here are without any way of connection and most of the names are fictitious, so when the leave their answers, we don't know them but still we know the right answers with the right proof.

Comment: The company should use solution C. Use SageMaker Debugger for visibility into the training weights, gradients, biases, and activation outputs. Compute the filter ranks based on this information. Apply pruning to remove the low-ranking filters. Set the new weights. Run a new training job with the pruned model.

Comment: Same example here: https://aws.amazon.com/blogs/machine-learning/pruning-machine-learning-models-with-amazon-sagemaker-debugger-and-amazon-sagemaker-experiments/

Comment: Selected Answer: B To reduce the time required for performing inferences in autonomous cars, the automotive company should use SageMaker Debugger for visibility into the training weights, gradients, biases, and activation outputs. They can adjust the model hyperparameters and look for lower inference times. They can also use SageMaker Model Monitor for visibility into the ModelLatency metric and OverheadLatency metric of the model after the model is deployed. However, option C, which suggests computing the filter ranks based on the training outputs and applying pruning to remove the low-ranking filters, is not applicable for transfer learning models since the layers in the pre-trained model are already trained and cannot be changed. Therefore, the correct solution is B.

Replies:

> **Comment:** better not use chatgpt without knowing something of AWS, it will trick you

Comment: Even if a better machine could help, the problem is about the model, not about the general or the machine in specific.

Comment: Using SageMaker Debugger, the company can monitor the training process and evaluate the performance of the model by computing filter ranks based on information like weights, gradients, biases, and activation outputs. After identifying the low-ranking filters, the company can apply pruning to remove them and set new weights. By doing so, the company can reduce the model size and improve the inference time. Finally, a new training job with the pruned model can be run to verify the performance improvements Not D because Model Monitor is a tool for monitoring the performance of deployed models, and it does not provide any direct feedback or insights into the model training process or ways to improve model inference time. Therefore, while Model Monitor can be useful for monitoring the performance of deployed models, it is not the best choice for evaluating and improving the performance of the models during the training phase, which is what the question is asking for.

Comment: It's between C and D. But I think it's C. C. https://aws.amazon.com/blogs/machine-learning/pruning-machine-learning-models-with-amazon-sagemaker-debugger-and-amazon-sagemaker-experiments/ Everything is there. D: https://aws.amazon.com/premiumsupport/knowledge-center/sagemaker-endpoint-latency/ Here it says use Cloudwatch to view ModelLatency and OverheadLatency, not Model Monitor. I think Model Monitor is just for model performance i.e. drift, bias, accuracy etc.

Comment: The answer I guess D per below , they should have said Sagemaker model monitor using cloud watch https://docs.aws.amazon.com/sagemaker/latest/dg/model-monitor.html

Comment: C, https://aws.amazon.com/blogs/machine-learning/pruning-machine-learning-models-with-amazon-sagemaker-debugger-and-amazon-sagemaker-experiments/

Comment: I would say 'D as a more generic approach than C. The problem can be caused not just filters.

Comment: The answer is "c" as the question is asking for evaluate and improve the performance of the models? "https://docs.aws.amazon.com/sagemaker/latest/dg/debugger-visualization.html"

Comment: I think the answer is D.

Discussion for Question 202

Link: https://www.examtopics.com/discussions/amazon/view/89051-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- D: 18 votes

Discussion

Comment: It has to option D.

Comment: Should be D. TFRecord could be uploaded to S3 directly and be used as SageMaker's data source. https://sagemaker-examples.readthedocs.io/en/latest/sagemaker_batch_transform/working_with_tfrecords/working-with-tfrecords.html#Upload-dataset-to-S3

Replies:

> **Comment:** Then why not C
>
> ### Replies:
>
> > **Comment:** then how is local path a "scalable data storage solution" ? answer is D

Comment: A. NO - SageMaker can use TFRecods as-is in S3 B. NO - SageMaker can use TFRecods as-is in S3 C. NO - SageMaker must use S3 as input, it cannot read your local data D. YES - SageMaker can use TFRecods as-is in S3

Comment: SageMaker script mode allows you to use your existing TensorFlow training scripts without any modifications. You can use the same TFRecord data format that your model expects, and point the SageMaker training invocation to the S3 bucket where the data is stored. SageMaker will automatically download the data to the local path of the training instance and pass it as an argument to your train.py script. You don't need to reformat the data to protobuf format or rewrite your script to convert the data12.

Comment: This option allows the ML specialist to use the existing train.py script and TFRecord data without any changes, minimizing development overhead. By using SageMaker script mode, the specialist can run the existing TensorFlow script as-is, and by pointing the SageMaker training invocation to the S3 bucket containing the TFRecord data, the specialist can provide the training data to SageMaker without reformatting it.

Comment: This option leverages SageMaker's built-in support for the TensorFlow framework and script mode. The existing train.py script can be used without any modifications. SageMaker will automatically download the training data from the specified S3 location to the instance running the training job. This option saves development time by avoiding the need to rewrite the train.py script or reformat the training data.

Discussion for Question 203

Link: https://www.examtopics.com/discussions/amazon/view/89213-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- D: 18 votes

Discussion

Comment: https://docs.aws.amazon.com/sagemaker/latest/dg/model-managed-spot-training.html Managed spot training can optimize the cost of training models up to 90% over on-demand instances. SageMaker manages the Spot interruptions on your behalf. "Spot instances can be interrupted, causing jobs to take longer to start or finish. You can configure your managed spot training job to use checkpoints. SageMaker copies checkpoint data from a local path to Amazon S3. When the job is restarted, SageMaker copies the data from Amazon S3 back into the local path. The training job can then resume from the last checkpoint instead of restarting."

Comment: It has to be D. With Spot training we can reduce the cost and save the model weights with checkpoint enabled.

Replies:

Comment: agree. managed spot training is also cost effective

Comment: A. NO - D is simpler B. NO - D is simpler C. NO - D is simpler D. YES - works out-of-the-box

Comment: Managed spot training in Amazon SageMaker uses Amazon EC2 Spot instances to run training jobs, which can optimize the cost of training models by up to 90% over on-demand instances 1. SageMaker manages the Spot interruptions on the company's behalf 1. By enabling checkpointing, the company can ensure that if a Spot instance is interrupted, the training job can resume from the last checkpoint instead of restarting, avoiding loss of work and model retraining 1

Comment: Selected Answer: D Use managed spot training in Amazon SageMaker. Launch the training jobs with checkpointing enabled. Managed spot training in Amazon SageMaker can help minimize operational overhead and cost by using spot instances to perform the training. This can significantly reduce the cost of training, while still achieving the same accuracy. SageMaker provides built-in checkpointing capability, which allows saving model weights and progress to Amazon S3 periodically. This ensures that even if the spot instances are terminated, the training can resume from the last saved checkpoint. Additionally, SageMaker provides a managed service, so the ecommerce company does not need to worry about managing the infrastructure, and can focus on building and tuning their model.

Comment: Selected Answer: D The ML specialist should choose option D, which provides the training data to SageMaker with the least development overhead. This option involves putting the TFRecord data into an Amazon S3 bucket and pointing the SageMaker training invocation to the S3 bucket without reformatting the training data. Using SageMaker script mode is a convenient way to execute training scripts without any modification. Since the training script train.py already works with TFRecord data, it can be used as is without any changes. By storing the data in S3 and accessing it from there, the specialist can take advantage of SageMaker's built-in data distribution and parallelization capabilities, which can significantly speed up training. Rewriting the train.py script or using additional services like AWS Glue or Lambda would add unnecessary complexity and increase development overhead.

Comment: Managed spot training in Amazon SageMaker provides a cost-effective way to run large machine learning workloads. With managed spot training, the training jobs are executed using Amazon EC2 Spot instances, which can significantly reduce the cost of training. Additionally, by launching training jobs with checkpointing enabled, the work done up to the last checkpoint is saved to Amazon S3. This ensures that the training job can be resumed from the last checkpoint in case of instance failure or termination. This minimizes the risk of data loss and avoids the need for retraining the model from scratch. Using Amazon SageMaker also reduces the operational overhead required to set up and manage the training environment.

Discussion for Question 204

Link: https://www.examtopics.com/discussions/amazon/view/88838-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 16 votes

Discussion

Comment: A is correct. "If the statistical nature of the data that your model receives while in production drifts away from the nature of the baseline data it was trained on, the model begins to lose accuracy in its predictions. Amazon SageMaker Model Monitor uses rules to detect data drift and alerts you when it happens." https://docs.aws.amazon.com/sagemaker/latest/dg/model-monitor-data-quality.html

Comment: A. YES - Model Monitor can validate distribution of input data B. NO - a model quality baseline is for model performance eg. precision, F1 score, etc. C. NO - Model Monitor is the right tool D. NO - Model Monitor is the right tool

Comment: it's a problem of monitoring data distributions.

Comment: The reason for this choice is that Amazon SageMaker Model Monitor is a feature of Amazon SageMaker that allows you to monitor and analyze your machine learning models in production. Model Monitor can automatically detect data drift and other data quality issues by comparing your live data with a baseline dataset that you provide1. Model Monitor can also emit metrics and alerts when it detects violations of the constraints that you define or that it suggests based on the baseline2.

Comment: A is correct. The best solution to meet the requirements is to use Amazon SageMaker Model Monitor to create a data quality baseline. The ML team can set up a data quality baseline to detect when the input data to the model has drifted significantly from the historical distribution of the data. When data drift occurs, the Model Monitor emits a metric that can trigger an alarm in Amazon CloudWatch. The ML team can use this alarm to investigate and take corrective action. Option B is incorrect because model quality baseline monitors model performance, not the input data quality. Option C is incorrect because Amazon SageMaker Debugger is used to debug machine learning models and to identify problems with model training, not data quality. Option D is incorrect because Amazon CloudWatch does not provide any features to monitor data drift in the input data used for the machine learning model.

Comment: Data monitor https://docs.aws.amazon.com/sagemaker/latest/dg/model-monitor-data-quality.html Properties of independent variables changes due to seasonality, customer preferences Model monitor https://docs.aws.amazon.com/sagemaker/latest/dg/model-monitor-model-quality.html Concept of what is spam email, changes over time "The company's ML team determines that the inaccuracies are occurring because of a change in the value distributions of the model features." They know model features that is data for model input is changing so we monitor data https://pair-code.github.io/what-if-tool/learn/tutorials/features-overview/ a

Comment: Data monitor https://docs.aws.amazon.com/sagemaker/latest/dg/model-monitor-data-quality.html Properties of independent variables changes due to seasonality, customer preferences Model monitor https://docs.aws.amazon.com/sagemaker/latest/dg/model-monitor-model-quality.html Concept of what is spam email, changes over time "The company's ML team determines that the inaccuracies are occurring because of a change in the value distributions of the model features." They know model features that is data for model input is changing so we monitor data https://pair-code.github.io/what-if-tool/learn/tutorials/features-overview/ a

Comment: I think the answer is B. Data quality can be monitored via model monitor model quality baseline.

Comment: B. Since it's "a change in the value distributions of the model features".

Replies:

Comment: model features = data

Comment: What is the difference of ans A and B?

Replies:

Comment: model quality baseline vs data quality baseline

Discussion for Question 205

Link: https://www.examtopics.com/discussions/amazon/view/89266-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 18 votes
- B: 14 votes

Discussion

Comment: https://docs.aws.amazon.com/sagemaker/latest/dg/experiments.html

Comment: Autopilot is the faster. "Amazon SageMaker Autopilot experiments are now up to 2x faster in Hyperparameter Optimization training mode" . Refer to https://aws.amazon.com/about-aws/whats-new/2022/11/amazon-sagemaker-autopilot-experiments-2x-faster-hyperparameter-optimization-training-mode/?nc1=h_ls

Replies:

Comment: SageMaker Autopilot is designed to automatically build, train, and tune the best machine learning model based on a dataset, without the user needing to choose an algorithm. It's not designed to be used with custom container images.

Comment: It seems that Autopilot doesn't support image data (image classification), so B will be incorrect in this case https://docs.aws.amazon.com/sagemaker/latest/dg/autopilot-datasets-problem-types.html#autopilot-datasets

Comment: Option D stands out as the most effective approach because it leverages SageMaker's automatic model tuning capabilities for both the custom container image and the built-in image classification algorithm. This ensures:

Comment: D The question is talking about how to do HPO using AWS Sagemaker for a model in custom image. Experiment is not to do HPO because you need to input parameter manually. So D

Comment: Amazon sagemaker experiment is ideal for this.

Comment: D: By using SageMaker's automatic model tuning capability to tune both the custom container image model and the built-in image classification algorithm model simultaneously, it leverages the parallel processing capabilities of SageMaker. This approach allows for efficient utilization of compute resources and can potentially complete the tuning process for both models in a shorter amount of time compared to running separate tuning jobs sequentially. Additionally, option D aligns with the requirement of invoking all ML experiments and HPO jobs from scripts inside SageMaker Studio notebooks, as SageMaker's automatic model tuning can be initiated and managed through notebook scripts. While options B and C could potentially work, option D provides the most direct and efficient path to meeting the requirements in the least amount of time by leveraging SageMaker's parallel processing capabilities and avoiding potential development overhead or limitations associated with other approaches.

Comment: The best option to meet the requirements in the least amount of time is D. Use the automatic model tuning capability of SageMaker to tune the models of the custom container image and the built-in image classification algorithm at the same time. This approach directly utilizes SageMaker's built-in capabilities for HPO, applies to both custom containers and built-in algorithms, and avoids the inefficiencies associated with local mode or manual management of experiments. It's important to note that while the tuning jobs would not literally run "at the same time" in a single operation, this option represents the most efficient use of SageMaker's capabilities for both scenarios.

Comment: Should be C. We are looking at comparing 2 models here, where Sagemaker Experiments fits the bill. D is out because "Amazon SageMaker automatic model tuning (AMT), also known as hyperparameter tuning, finds the best version of a model by running many training jobs on your dataset." https://docs.aws.amazon.com/sagemaker/latest/dg/automatic-model-tuning.html

Comment: D can be done easily https://docs.aws.amazon.com/sagemaker/latest/dg/automatic-model-tuning.html "You can use SageMaker AMT with built-in algorithms, custom algorithms, or SageMaker pre-built containers for machine learning frameworks."

Comment: I will go with D https://docs.aws.amazon.com/sagemaker/latest/dg/automatic-model-tuning-how-it-works.html

Comment: Will go with B and Autopilot supports Image classification as per this link - https://docs.aws.amazon.com/sagemaker/latest/dg/autopilot-automate-model-development.html Autopilot currently supports the following problem types: Regression, binary, and multiclass classification with tabular data formatted as CSV or Parquet files in which each column contains a feature with a specific data type and each row contains an observation. The column data types accepted include numerical, categorical, text, and time series that consists of strings of comma-separated numbers. Text classification with data formatted as CSV or Parquet files in which a column provides the sentences to be classified, while another column should provide the corresponding class label. Image classification with images formats such as PNG, JPEG or a combination of both. Time-series forecasting with time-series data formatted as CSV or as Parquet files.

Replies:

Comment: SageMaker Autopilot is designed to automatically build, train, and tune the best machine learning model based on a dataset, without the user needing to choose an algorithm. It's not designed to be used with custom container images.

Comment: A. NO - try AMT (=Automatic Model Tuning) before using custom HPO scripts; further, no reason to use the local mode B. NO - Autopilot is not for HPO only, it will also select a model etc. C. NO - requires manual parameter setting for each experiments D. YES - AMT (=Automatic Model Tuning) work with custom containers

Comment: Answer B in my opinion. Key is autopilot in least amount of time and early stopping to switch over

Comment: Changing to option C

Comment: Going for Option D

Comment: It's a bit confusing but leaning towards D This option allows the ML specialist to use the automatic model tuning capability of SageMaker to tune both models simultaneously, saving time compared to tuning them sequentially. By selecting the model with the best objective metric value, the ML specialist can determine whether HPO with the SageMaker built-in image classification algorithm produces a better model than HPO with the custom container image. All of this can be done from scripts inside SageMaker Studio notebooks, meeting the requirements in the least amount of time.

Comment: Experiment to compare the runs. B is not feasible as you cannot use it to "tune" your custom model. Auto Pilot is meant to test your model with lots of other algorithms that, by the way, are not much "suitable" for image input.

Discussion for Question 206

Link: https://www.examtopics.com/discussions/amazon/view/89069-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- CE: 14 votes

Discussion

Comment: CE Option A: Launching multiple medium-sized instances in a distributed SageMaker Processing job and using the prebuilt Docker images for Apache Spark to query and plot the relevant data is a possible solution, but it may not be the most cost-effective solution as it requires spinning up multiple instances. Option B: Launching multiple medium-sized notebook instances with a PySpark kernel in distributed mode is another solution, but it may not be the most secure solution as the data would be stored on the instances and not in a centralized data repository. Option D: Using AWS Secrets Manager to store the Amazon Redshift credentials and launching a SageMaker extra-large notebook instance is a solution, but the block storage requirement that is slightly larger than 10 TB could be costly and may not be necessary.

Comment: C and E. No secure control is in option A.

Comment: As soon as I see, Download andpython client, I am worried about speed and efficiency. So I would say A and E

Comment: A. NO - SageMaker Processing job is a self-contained feature using the sagemaker.processing API; it does not rely on invoking Spark directly B. NO - you want to identify the relevant slice of data without having to download everything first C. YES - minimize data movement D. NO - you want to identify the relevant slice of data without having to download everything first E. YES - built-in tool specifically designed for that use case

Comment: E for sure but was a bit confused on A or C but based on the link would go for C https://aws.amazon.com/blogs/big-data/using-the-amazon-redshift-data-api-to-interact-from-an-amazon-sagemaker-jupyter-notebook/

Comment: e is obvious choice: c https://aws.amazon.com/blogs/big-data/using-the-amazon-redshift-data-api-to-interact-from-an-amazon-sagemaker-jupyter-notebook/

Comment: C & E seems right - https://docs.aws.amazon.com/sagemaker/latest/dg/data-wrangler.html

---

Discussion for Question 207

Link: https://www.examtopics.com/discussions/amazon/view/89267-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- D: 8 votes

Discussion

Comment: I don't see why everyone is voting for D. To fine tune BERT you should add a classifier on top of the [CLS] token representing the hidden state. So it's not clear to me what does the question mean with "last fully connected layer"

Comment: D is the right option since initializing the model with pretrained weights, the model can leverage the knowledge learned from a large corpus of text data, such as English Wikipedia text data, to improve its performance on a specific task, such as spam email classification . And Replacing the last fully connected layer with a classifier is necessary because the last layer of BERT is designed for predicting masked words in a sentence, which is different from the task of spam email classification

Comment: A. NO - the last fully connected layer will not do SoftMax classification B. YES - output of BERT (word embeddings) can be used as input of classification C. NO - random weights will discard previous transfer learning D. NO - we don't want to loose the word embeddings; "cut the head off" (replacing the last layer) is if we want to learn different classes than what the model was trained for, but here we want to augment

Replies:

Comment: You should consider that Stacking a classifier on top of the first output position and training it with labeled data is not recommended because it does not take advantage of the knowledge learned from pretraining on a large corpus of text data

Comment: D although was leaning towards B

Replies:

Comment: on a second thought going for B

Comment: Cut the Head Off

Comment: D seems correct

Comment: D is a best practice

Comment: Is B correct? - https://www.analyticsvidhya.com/blog/2020/07/transfer-learning-for-nlp-fine-tuning-bert-for-text-classification/ Freeze the entire architecture – We can even freeze all the layers of the model and attach a few neural network layers of our own and train this new model. Note that the weights of only the attached layers will be updated during model training.

Replies:

Comment: You would have two classifiers stacked, so your predictions would be based in the other classifier.

Comment: I think the answer is D.

Discussion for Question 208

Link: https://www.examtopics.com/discussions/amazon/view/88837-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 19 votes

Discussion

Comment: The correct answer is C. Use Contact Lens for Amazon Connect to process audio calls to produce transcripts, categorize calls, detect issues, and analyze sentiment. Contact Lens is a fully managed service that provides advanced analytics for customer service interactions in Amazon Connect. It includes call transcription, sentiment analysis, and issue detection, which meets all the requirements in the question. Using a single service like Contact Lens will reduce the complexity of integrating multiple AWS services and also minimize the need for custom model training. While Amazon Transcribe and Amazon Comprehend are also valuable AWS services, they are not designed specifically for customer service interactions and may require additional configuration and custom model training to meet the specific requirements listed in the question. You can see detail in https://www.udemy.com/course/aws-certified-machine-learning-specialty-2023/

Comment: I think C. Contact Lens can do all. https://aws.amazon.com/connect/contact-lens/

Comment: this is AWS Q answer Contact Lens for Amazon Connect is primarily designed to analyze conversations that occur within an Amazon Connect contact center. It provides real-time capabilities like sentiment analysis, issue detection and compliance monitoring for calls and chats handled by Amazon Connect agents. While it cannot directly analyze past recorded conversations outside of Amazon Connect, you could potentially use other AWS services like Amazon Transcribe to transcribe pre-recorded calls. Amazon Comprehend could then be used to analyze the transcripts for things like sentiment, topics, entities and key phrases.

Replies:

Comment: After further review and research. Contact Lens can indeed process past audio and chat records. Please ignore previous explanation. Chat GPT or AWS Q are not fully reliable as yet. Answer is indeed C

Comment: Based on the information provided and the goal of minimizing custom model training, Option C (Use Contact Lens for Amazon Connect to process audio calls to produce transcripts, categorize calls, detect issues, and analyze sentiment) is the best solution. Contact Lens for Amazon Connect is specifically designed to handle tasks related to call center operations, including all the features listed in the requirements, without the need for extensive custom model training.

Comment: "several years remaining on its contract" means no Amazon Connect.

Replies:

Comment: Amazon Connect is a pay-as-you-go cloud contact center. There are no required minimum monthly fees, long-term commitments, or upfront license charges, and pricing is not based on peak capacity, agent seats, or maintenance; you only pay for what you use.

Replies:

Comment: correct. But the contract with a legacy system will not disappear. So the company will need to pay for both Amazon Connect and Legacy system.

Comment: Amazon connect contact lens can fulfill all these requirements and on an enterprise level it provides contact center analytics and quality management capabilities that enable you to monitor, measure, and continuously improve contact quality and agent performance for a better overall customer experience Ref: https://docs.aws.amazon.com/connect/latest/adminguide/contact-lens.html https://aws.amazon.com/connect/contact-lens/

Comment: Answer is B.. no doubt

Comment: correct answer is C

Comment: It's B

Comment: Customer is not replacing telephony software with Amazon Connect hence c & d are ruled out. Answer B

Comment: B as they cannot use Amazon Connect due to several years on contract.

Comment: "several years remaining on its contract" means no Amazon Connect.

Comment: c https://docs.aws.amazon.com/connect/latest/adminguide/analyze-conversations.html

Comment: Company has several years remaining on its contract. Does this change the solution approach? That means they cant use Amazon Connect. So Contact Lens may not be applicable here. If that assumption is true, B should be the answer.

Replies:

Comment: I don't think it changes the solution approach. The question was about LEAST development effort, not lowest cost. Yes it will cost more to use Connect but dev cost is lower.

Comment: Having the same though. If the company needs to remain in the legacy telephony platform, the answer would be B. Otherwise, C.

Comment: Contact Lens is an out of the box feature for Amazon Connect that leverages Amazon Transcribe to generate call transcripts and Amazon Comprehend to apply natural language processing (NLP) on these transcripts, with no coding required. https://aws.amazon.com/connect/contact-lens/#:~:text=Contact%20Lens%20is%20an%20out,transcripts%2C%20with%20no%20coding%20required.

Comment: c seems to be correct

Comment: I think the answer is D. Using Amazon Comprehend to categorize calls, detect issues, and analyze sentiment.

Replies:

Comment: I correct myself. It should be C. tsangckl and akhjhjk are right.

Discussion for Question 209

Link: https://www.examtopics.com/discussions/amazon/view/98541-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 15 votes

Discussion

Comment: For time series data, it is important to split the dataset chronologically, with the training dataset containing the earlier dates and the validation dataset containing the later dates

Comment: A. YES - it is forecasting, so you want to predict the future and 20% of the data points after a date will do so B. NO - it is forecasting, we want to simulate an actual use case and not predict the past C. NO - there is data leakage as future datapoints are used in the predictions C. NO - there is data leakage as future datapoints are used in the predictions

Comment: Time series keep the order

Comment: A, As it's a time series problem.

Comment: It's a timeseries problem, then the splitting needs to be made by date.

Comment: Option A is the recommended approach where the training dataset contains historical prices that precede a certain date, and the validation dataset contains prices that occur after that date. This ensures that the model is trained on past data and evaluated on future data, which is more representative of real-world performance. Option D is NOT the recommended approach for time series data because it ignores the time aspect of the data. Randomly sampling data points without considering the time sequence can result in data leakage and poor model performance.

Comment: a since this is time series problem

Comment: a https://towardsdatascience.com/time-series-from-scratch-train-test-splits-and-evaluation-metrics-4fd654de1b37

Comment: Because it randomly selects data points for both the training and validation datasets, ensuring that the samples are representative of the entire dataset and reducing the chances of overfitting. By randomly sampling without replacement, the data scientist can avoid any biases in the selection of data points and ensure that the training and validation datasets are independent.

Replies:

Comment: For time series you should keep the order.

Comment: I think it should be A!

Discussion for Question 210

Link: https://www.examtopics.com/discussions/amazon/view/98527-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 17 votes

Discussion

Comment: A. NO - we want to leverage a prebuilt model for efficiency B. NO - PERSONALIZED_RANKING uses a predefined list of items as input C. YES - USER_PERSONALIZATION uses past user history as input D. NO - we want to leverage a prebuilt model for efficiency

Comment: Answer C

Comment: User personalization: Recommendations tailored to a user's profile, behavior, preferences, and history. This is most commonly used to boost customer engagement and satisfaction. It can also drive higher conversion rates. Personalized ranking: Items re-ranked in a category or search response based on user preference or history. This use case is used to surface relevant items or content to a specific user ensuring a better customer experience. Amazon Personalize supports re-ranking while optimizing for business priorities such as revenue, promotions, or trending items. https://aws.amazon.com/personalize/faqs/

Comment: B it,s the correct answer

Comment: C looks the right choice

Comment: Its C, User Personalizationis recommended for user interaction scenarios https://docs.aws.amazon.com/personalize/latest/dg/native-recipe-new-item-USER_PERSONALIZATION.html

Comment: It's C, User Personalization is recommended for self-user user case.

Comment: The User-Personalization (aws-user-personalization) recipe is optimized for all personalized recommendation scenarios. It predicts the items that a user will interact with based on Interactions, Items, and Users datasets. When recommending items, it uses automatic item exploration.

Comment: Option B: https://docs.aws.amazon.com/personalize/latest/dg/native-recipe-search.html "With Personalized-Ranking, you must manually create a new solution version (retrain the model) to reflect updates to your catalog and update the model with your user's most recent behavior." Option B has a disadvantage to update the catalog in retail company. So, Option C has the less effort to oprate than Option B

Comment: Option B is a better fit for the given requirements since it specifically mentions the need to filter out items that the user has previously purchased. The PERSONALIZED_RANKING recipe in Amazon Personalize is designed to provide personalized recommendations while allowing for exclusion of previously purchased items using a filter. In contrast, the USER_PERSONALIZATION recipe in option C is designed to provide personalized recommendations without the ability to filter out previously purchased items. Therefore, option B is the best choice for meeting the given requirements with the least development effort.

Comment: Answer is C https://docs.aws.amazon.com/personalize/latest/dg/native-recipe-new-item-USER_PERSONALIZATION.html

Comment: Answer is C https://docs.aws.amazon.com/personalize/latest

---

Discussion for Question 211

Link: https://www.examtopics.com/discussions/amazon/view/99688-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 10 votes

Discussion

Comment: A. NO - Incremental training not supported by XGBoost (https://docs.aws.amazon.com/sagemaker/latest/dg/incremental-training.html) B. NO - we don't want to change the objective and restart from scratch C. YES - warm start can leverage new data from production for further tuning D. NO - we don't want to start from the training from scratch or use F1 score as objective

Comment: Answer C

Comment: Given time constraint, I believe that C is the crrect one.

Comment: C is the correct answer because it uses the results from past HPO jobs and builds upon them to improve accuracy.

Comment: I go with C - warm start, A is not supported on XGBoost, and other options will start tuning from scratch and might be just as bad as the inital tuning job. We only have 1 day, so more tuning with existing job to inform the new trainging job is the only option here https://docs.aws.amazon.com/sagemaker/latest/dg/automatic-model-tuning-warm-start.html

Comment: C is the correct answer. You can't use Incremental training on Xgboost algorithm https://docs.aws.amazon.com/sagemaker/latest/dg/incremental-training.html

Replies:

Comment: It appears in 2023-April-3

Comment: Since ROC-AUC is presumed to be one of the best for a binary classification. Hence option B. Option A -- Incremental training is suited wherein the training dataset gets updated frequently.

Comment: https://docs.aws.amazon.com/sagemaker/latest/dg/automatic-model-tuning-warm-start.html https://francesca-donadoni.medium.com/training-an-xgboost-model-for-pricing-analysis-using-aws-sagemaker-55d777708e52 https://docs.aws.amazon.com/sagemaker/latest/dg/incremental-training.html C

Replies:

Comment: also it cannot be a: "Only three built-in algorithms currently support incremental training: Object Detection - MXNet, Image Classification - MXNet, and Semantic Segmentation Algorithm." from https://docs.aws.amazon.com/sagemaker/latest/dg/incremental-training.html

---

Discussion for Question 212

Link: https://www.examtopics.com/discussions/amazon/view/98756-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 7 votes

Discussion

Comment: B is right. Is very simple to create a conversion file JOB in AWS Glue, using just 3 workflow steps. WITH NO CODE.. CREATED AUTOMATICALLY BY GLUE (Scala or Python) (s3 - source data file) --> (Data Mapping) --> (target transformed data file)

Comment: A. NO - Crawler is to populate the data catalog B. YES - leverage serverless for distributed processing C. NO - Altough EMR can run Spark like Glue, it is not serverless D. NO - using the PySpark kernel will be single instance (running in the notebook)

Comment: Option B is better than option A because option A uses an AWS Glue crawler to convert the file format. A crawler is a component of AWS Glue that scans your data sources and infers the schema, format, partitioning, and other properties of your data. A crawler can create or update a table in the AWS Glue Data Catalog that points to your data source. However, a crawler cannot change the format of your data source itself. You still need to write a script or use a tool to convert your CSV files to Parquet files.

Comment: Option B. A - Glue crawler creates Glue Data Catalog from S3 buckets. It can be used to query by athena. C, D - not serverless and not generally used for etl.

Comment: AWS Glue is a fully-managed ETL service that makes it easy to move data between data stores. AWS Glue can be used to automate the conversion of CSV files to Parquet format with minimal effort. AWS Glue supports reading data from CSV files, transforming the data, and writing the transformed data to Parquet files. Option A is incorrect because AWS Glue crawler is used to infer the schema of data stored in S3 and create AWS Glue Data Catalog tables. Option C is incorrect because while Amazon EMR can be used to process large amounts of data and perform data conversions, it requires more operational effort than AWS Glue. Option D is incorrect because Amazon SageMaker is a machine learning service, and while it can be used for data processing, it is not the best option for simple data format conversion tasks.

Comment: in sagemaker notebook, you'd have to write python code but question is asking for something easy so i choose option b https://blog.searce.com/convert-csv-json-files-to-apache-parquet-using-aws-glue-a760d177b45f

Replies:

Comment: From you link, A(Glue crawler) Should be correct.

Replies:

Comment: crawler just creates the data catalog (schema), it does not actually converts the data to another format. As per details in that article, you are creating a job where source is schema created by crawler and destination is output s3 where we store formatted data.

Discussion for Question 213

Link: https://www.examtopics.com/discussions/amazon/view/99814-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 9 votes

Discussion

Comment: A. YES - Kinesis/Kafka acts as buffer for ingestion, Firehose provides good integration with Lambda (tranformation) & S3 (storage) B. NO - no point to save the data twice in S3 (raw and transformed) C. NO - since we do single-record transformation Glue/Spark is overkill D. NO - since we do single-record transformation Glue/Spark is overkill; further, we can reasonably expect devices to produce Kafka events but deploying a Firehose client API seem complicated

Comment: answer A

Comment: AWS Glue cannot get data from Kinesis Firehose, only from Kinesis Data Stream. It's not D. https://docs.aws.amazon.com/glue/latest/dg/add-job-streaming.html

Comment: It is D

Replies:

Comment: Glue can't read from Firehose. It's A.

Comment: It is C

Replies:

Comment: Option C uses AWS Glue, which can perform data transformation and load data into S3 buckets. However, Glue may not be the most efficient option for this use case, as it requires setting up a Glue job, which can introduce additional latency.

Replies:

Comment: Option A uses Amazon Kinesis data stream, which is optimized for high throughput, durable storage, and scalability.

Comment: Why not C?

Comment: Firehose can take just at a maximum of 5 minutes, then it's the best solution for transformations.

Comment: A general architecture for (near)real - time ingesting & processing data: Kinesis Data Streams - Kinesis Data Firehose - (If needs etl, lambda) - S3(Redshift, ...)

Comment: This solution provides a highly scalable and efficient way to ingest streaming data from devices with high throughput and durable storage by using Amazon Kinesis Data Streams and Amazon Kinesis Data Firehose. By configuring an AWS Lambda function to transform the data during the ingestion process, the solution also applies basic data transformation with low latency. Additionally, Amazon S3 provides highly durable and scalable storage for the transformed data, which can be easily accessed by downstream processes such as machine learning model training.

Comment: A: https://docs.aws.amazon.com/firehose/latest/dev/data-transformation.html

---

Discussion for Question 214

Link: https://www.examtopics.com/discussions/amazon/view/99689-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 10 votes

Discussion

Comment: A is correct, why not C : C require updating software in each of the 20,000 stores, which is operationally intensive. Moreover, the S3 File Gateway is designed for on-premises integration with S3.

Comment: Answer A

Comment: Firehose can use lambda functions to do data transformations!

Comment: A is the best option for this use case. By creating an AWS Lambda function that can transform the incoming records and enabling data transformation on the ingestion Kinesis Data Firehose delivery stream, the company can transform the data with minimal operational overhead. The Lambda function can be the invocation target for Kinesis Data Firehose, so that data is transformed as it is ingested. This approach is serverless and scalable, and it does not require the company to manage any additional infrastructure.

Comment: A: Lambda as invocation target just means that the function will invoke in response to the Firehose stream. See following: https://docs.aws.amazon.com/lambda/latest/dg/lambda-services.html https://docs.aws.amazon.com/lambda/latest/dg/services-kinesisfirehose.html note: invocationid in Firehose message event.

Comment: a - seems to be an easy to manage solution however the phrase "Use the Lambda function as the invocation target." confuses me a bit.

Replies:

Comment: well that is used by Kinesis Data Firehouse..

Discussion for Question 215

Link: https://www.examtopics.com/discussions/amazon/view/99696-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- BE: 15 votes

Discussion

Comment: It is clear from AWS docs. No doubt.

Comment: custom language model will be needed as custom vocab will just help with pronunciation, and requirement clearly states handling domain specific terminologies, which cannot be handled by custom vocab.

Comment: Option A - Amazon Transcribe with custom vocabularies, allows you to enhance the transcription accuracy by providing domain-specific terminology, names, and locations. Custom vocabularies enable you to train the transcription model to recognize specific words or phrases commonly used in the context of sports commentary. This would help ensure that the transcriptions accurately capture the specialized terminology and context of the commentary, meeting the requirements of the sports broadcasting company. Additionally, the option mentions supporting options to provide tuning data, which further enhances the flexibility and customization of the solution.

Comment: Could someone explain why A is not correct? As par AWS documentation: Use custom vocabularies to improve transcription accuracy for one or more specific words. These are generally domain-specific terms, such as brand names and acronyms, proper nouns, and words that Amazon Transcribe isn't rendering correctly. Custom vocabularies can be used with all supported languages.

Comment: A - Yes - Transcribe custom vocabs allow domain specific transforms B - No - Custom language models are typically used for fine-tuning for specific accents, dialects, or unique speech patterns. This question is about domain specific terminology C - No - building and tarining requires more overhead D - No - building and tarining requires more overhead E - Once transcribed in english, translate can perform laguage transformation

Comment: A. NO - Amazon Transcribe with custom vocabularies does not allow to take into account the broader context (https://docs.aws.amazon.com/transcribe/latest/dg/custom-vocabulary-create-list.html) B. YES - Custom language models are designed to improve transcription accuracy for domain-specific speech (https://docs.aws.amazon.com/transcribe/latest/dg/custom-language-models.html) C. NO - better to use built-in Translate service than base Seq2Seq D. NO - Hugging Face Speech2Text is custom model, use standard Transcribe E. YES - we need to translate English

Comment: custom language + Translate

Comment: Answer BE

Comment: BE : For Specific Language: Custom Language Model

Comment: Comentary context is custom language

Comment: It's BE, custom languages is for domain-specific speech like terminologies, custom vocabulary is for words liken nouns.

Replies:

Comment: https://docs.aws.amazon.com/transcribe/latest/dg/custom-language-models.html

Comment: Two sub-processes are needed: Speech to Text and Text to Text. We can consider Amazon Transcribe for Speech2Text. If we use custom language models or SageMaker, we would need to gather our own data to train or retrain models. for less effort, (a) option is better than (b) option. Then, Amazon Translate can be used to translate transcription to other language: (e) option

Replies:

Comment: option A cannot capture "commentary context". B and E should be correct.

Comment: b: because The transcriptions must be able to capture domain-specific terminology, names, and locations based on the "commentary context" e: managed service

---

Discussion for Question 216

Link: https://www.examtopics.com/discussions/amazon/view/98990-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 11 votes

Discussion

Comment: B - DeepAR _ Algo with Holidya featurization in Amazon Forecast --> works better than Option A as Option A may be suboptimal but totally doable. DeepAR shines in multiple correlated time series and meant for seasonal data patterns

Comment: A. YES - fully managed solution B. NO - DeepAR+ is more for multiple time series ("In many applications, however, you have many similar time series across a set of cross-sectional units"- https://docs.aws.amazon.com/forecast/latest/dg/aws-forecast-recipe-deeparplus.html) C. NO - SageMaker DeepAR is too low-level D. NO - Gluon is too low-level

Comment: Option B is a great choice but requires more development effort over A which is also a great choice. Since the question asked for Least Development I am going with A

Comment: Deep AR can understand seasonal effect

Comment: B is correct

Comment: DeepAR accepts exogenorous regressors different from ARIMA and can understand seasonal effects, ARIMA can't do it too.

Comment: It is deepAR

Comment: Option B is a good choice, as the DeepAR+ algorithm is specifically designed for forecasting in time series data with seasonality and long-term dependencies. However, it may require more development effort compared to the ARIMA algorithm.

Comment: Amazon Forecast is an AWS service that uses machine learning to build accurate time-series forecasts. It provides several built-in algorithms that support holiday featurization, and the DeepAR+ algorithm can handle the seasonality and correlation with other products with minimal development effort. With Amazon Forecast, the data scientist can easily configure the forecast horizon, select the appropriate forecast frequency, and configure the model training to incorporate the available historical data. Using Amazon SageMaker Processing to enrich the data with holiday information may require more development effort and does not offer the same level of automation and

integration as Amazon Forecast.

Comment: b https://docs.aws.amazon.com/forecast/latest/dg/holidays.html https://docs.aws.amazon.com/whitepapers/latest/time-series-forecasting-principles-with-amazon-forecast/appendix-a-faqs.html

---

Discussion for Question 217

Link: https://www.examtopics.com/discussions/amazon/view/97866-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 15 votes

Discussion

Comment: oversample the minority class

Comment: Undersampling not an option for already limited observations. SMOTE clearly MOST promising first action before trying to balance classes

Comment: Do we need to stratify the non-failure events by machine type?

Comment: B. Oversample the failure cases by using the Synthetic Minority Oversampling Technique (SMOTE). Since the number of failure cases is relatively small, oversampling the failure cases using techniques like SMOTE can help balance the class distribution and prevent the model from being biased towards the majority class. SMOTE creates synthetic samples for the minority class by interpolating new samples between existing ones. This will help improve the model's accuracy in predicting failure cases. Adjusting class weights (A) or undersampling (C, D) may not be as effective in this scenario.

Comment: The data provided is imbalanced, with only 100 failure cases out of 10,000 event samples. Therefore, it is important to address this imbalance to improve the accuracy of the predictive maintenance model.

Discussion for Question 218

Link: https://www.examtopics.com/discussions/amazon/view/98762-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 16 votes

Discussion

Comment: C. Train an Amazon SageMaker Neural Topic Model (NTM) model to generate the product categories. The task is to build a machine learning model to categorize documents for all the company's products. Among the given options, training an Amazon SageMaker Neural Topic Model (NTM) model would be the most efficient and effective solution. An NTM model can identify topics in text data and group similar documents into specific categories, making it a suitable model for document categorization. With an NTM model, the data scientist would not need to define product categories beforehand, as the model would automatically group similar documents into topics. This saves time and resources compared to the other options.

Replies:

Comment: thank you chatgpt

Replies:

Comment: Thanks to ChatGPT and also thanks Ajose O for saving our time looking for some evidence or a proof to the right answer Ajose you made some good work bringing this clarification for us, so Thank you so much, Gracias amigo :)

Comment: C. Train an Amazon SageMaker Neural Topic Model (NTM) model to generate the product categories. -- option doesn't talk about any classification activity

Comment: A. NO - no need to build a custom model B. NO - k-means is supervised model C. YES - unsupervised clustering algorithm D. NO - Blazing Text will do word embedding, not classification

Replies:

Comment: No, k-means is an unsupervised learning algorithm. I

Comment: Neural Topic Model (NTM) is one of the built-in algorithms of Amazon SageMaker that can perform topic modeling on text data. Topic modeling is a technique that can discover latent topics or themes from a collection of documents. Topic modeling can be used for document categorization by assigning each document to one or more topics based on its content.

Comment: Blazing Text is only for supervised problems.

Comment: Assign pre-defined categories to documents in a corpus: categorize books in a library into academic disciplines - BlazingText algorithm https://docs.aws.amazon.com/sagemaker/latest/dg/algos.html

Replies:

Comment: Reading it again - C

Comment: Option C is wrong because it suggests using a Neural Topic Model (NTM) to categorize documents. While NTM can be used to discover the underlying topics in a corpus of documents, it may not be the most efficient solution for categorizing documents for specific products. NTM is more suited for unsupervised learning problems where the goal is to discover the underlying themes or topics of the document corpus. In this scenario, the data scientist needs to categorize documents based on predefined product categories. Therefore, a supervised learning algorithm like a text classification model would be more suitable. Amazon SageMaker Blazing Text algorithm provides an efficient and scalable solution for text classification problems.

Replies:

Comment: "no predefined product categories" -> unsupervised learning, C.

Replies:

Comment: what is K-means?

Replies:

Comment: Good catch. The problem with B is the fact that is an incomplete question: "Tokenize the data and transform the data into tabular data" how are you going to do this conrad?

Comment: No predefined product category: topic modeling with NTM or LDA (Organize a set of documents into topics (not known in advance): tag a document as belonging to a medical category based on the terms used in the document.) Predefined product category: topic modeling with blazing text (categorize books in a library into academic disciplines) c

---

Discussion for Question 219

Link: https://www.examtopics.com/discussions/amazon/view/98763-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 7 votes

Discussion

Comment: Amazon Rekognition can handle large-scale and high-quality images with low latency and high accuracy. You can use Amazon Rekognition to process images from various sources, such as cameras, webcams, or media

files. You can also use Amazon Rekognition to process images in real time or in batch mode.

Comment: It's A, rekgonition has methods to do text detection.

Replies:

> **Comment:** https://aws.amazon.com/rekognition/

Comment: OCR : Amazon Rekognition

Comment: Rekognition's Text Detection feature allows you to easily extract text from images and videos without the need to create a custom model or perform complex training. It's a fully managed service that provides accurate and scalable text detection, recognition, and analysis capabilities. Additionally, Rekognition provides a simple API and SDKs for integrating text detection functionality into your applications. https://docs.aws.amazon.com/rekognition/latest/dg/text-detection.html?pg=ln&sec=ft

Comment: a - LEAST operational overhead

Discussion for Question 220

Link: https://www.examtopics.com/discussions/amazon/view/98764-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 15 votes

Discussion

Comment: Batch Transform can efficiently handle this workload by splitting the files into mini-batches and distributing them across multiple instances. Batch Transform can also scale down the instances when there are no files to process, so you only pay for the duration that the instances are actively processing files. Batch Transform is more cost-effective than Asynchronous Inference because Asynchronous Inference is designed for workloads with large payload sizes (up to 1GB) and long processing times (up to 15 minutes) that need near real-time responses. Asynchronous Inference queues incoming requests and processes them asynchronously, returning an output location as a response. Asynchronous Inference also autoscales the instance count to zero when there are no requests to process. However, Asynchronous Inference charges you for both request processing and request queuing time, which may be higher than Batch Transform for your use case.

Comment: A. YES - Batch Transform can pick up new files from S3 B. NO - no need for asynch, high-throughpt queues C. NO - Processing is not for model inference D. NO - no need for scaling through multiple endpoints

Comment: Key point is data is collected every hour. seems like a batch solution is most cost effective

Comment: I will go with A. The Async inference seems promising but the size of telemetry file is not known. As per https://docs.aws.amazon.com/sagemaker/latest/dg/inference-cost-optimization.html "Use batch inference for workloads for which you need inference for a large set of data for processes that happen offline (that is, you don't need a persistent endpoint). You pay for the instance for the duration of the batch inference job". As you pay for the batch job duration, cost should not be an issue with Batch transform. "Use asynchronous inference for asynchronous workloads that process up to 1 GB of data (such as text corpus, image, video, and audio) that are latency insensitive and cost sensitive. With asynchronous inference, you can control costs by specifying a fixed number of instances for the optimal processing rate instead of provisioning for the peak. You can also scale down to zero to save additional costs."

Comment: Based on the requirements and constraints given in the scenario, the MOST cost-effective solution for the company to use to run the model across the telemetry for all the devices is SageMaker Batch Transform. SageMaker Batch Transform is a cost-effective solution for performing offline inference, as it allows for large amounts of data to be processed at a lower cost compared to real-time inference. In this case, the telemetry data for each device is collected hourly and can be processed in batches using SageMaker Batch Transform. This can help to reduce the cost of inference, as the data is not being processed in real-time and can be processed offline.

Comment: B -- based on what Drcock87 said, as well as this: "Amazon SageMaker Asynchronous Inference is a new capability in SageMaker that queues incoming requests and processes them asynchronously. Compared to Batch Transform Asynchronous Inference provides immediate access to the results of the inference job rather than waiting for the job to complete"

Replies:

> **Comment:** I still think its A because: - "The 4-day telemetry of each device is collected in a separate file and is placed in an Amazon S3 bucket once every hour." Which means this is use-case where data is available upfront for inferencing. - Also, unlike async the batch transform does not keep an active endpoint all the time. async is similar to real-time inference, used when you need inference right-away; the question is not asking for real-time inference.

Comment: Real-Time Inference is suitable for workloads where payload sizes are up to 6MB and need to be processed with low latency requirements in the order of milliseconds or seconds. Serverless Inference: Serverless inference is ideal when you have intermittent or unpredictable traffic patterns. Batch transform is ideal for offline predictions on large batches of data that is available upfront. We are introducing Amazon SageMaker Asynchronous Inference, a new inference option in Amazon SageMaker that queues incoming requests and processes them asynchronously. This option is ideal for inferences with large payload sizes (up to 1GB) and/or long processing times (up to 15 minutes) that need to be processed as requests arrive. Asynchronous inference enables you to save on costs by autoscaling the instance count to zero when there are no requests to process, so you only pay when your endpoint is processing requests. a

Replies:

> **Comment:** Real-time inference is suitable for workloads where payload sizes are up to 6MB and need to be processed with low latency requirements in the order of milliseconds or seconds. Batch transform is ideal for offline predictions on large batches of data that is available upfront. The new asynchronous inference option is ideal for workloads where the request sizes are large (up to 1GB) and inference processing times are in the order of minutes (up to 15 minutes). Example workloads for asynchronous inference include running predictions on high resolution images generated from a mobile device at different intervals during the day and providing responses within minutes of receiving the request.

Discussion for Question 221

Link: https://www.examtopics.com/discussions/amazon/view/98201-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- D: 15 votes

Discussion

Comment: The Answer is D

Comment: Option D uses the elbow method, which is a popular and well-known method for determining the optimal value of k for k-means clustering1. It plots the sum of squared errors (SSE) for different values of k, and looks for the point where the SSE starts to decrease in a linear fashion. This point is called the elbow, and it indicates that adding more clusters does not improve the model significantly2.

Comment: The Sum of square shows variation within each cluster

Comment: D. Run the k-means clustering algorithm for a range of k. For each value of k, calculate the sum of squared errors (SSE). Plot a line chart of the SSE for each value of k. The optimal value of k is the point after which the curve starts decreasing in a linear fashion. The sum of squared errors (SSE) measures the total variation within each cluster, and the optimal value of k is typically the point where the SSE begins to level off or decrease sharply. Plotting the SSE against the number of clusters (k) allows the data scientist to identify the optimal number of clusters based on where the SSE curve starts decreasing linearly.

Comment: d https://towardsdatascience.com/explain-ml-in-a-simple-way-k-means-clustering-e925d019743b

---

Discussion for Question 222

Link: https://www.examtopics.com/discussions/amazon/view/98765-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- AB: 12 votes

Discussion

Comment: A. YES - fully automated pipeline B. YES - triggers the pipeline A as needed C. NO - email notification does not allow automation D. NO - manual steps required, not operationaly efficient E. NO - we need another step to trigger the Lambda

Comment: Option A uses SageMaker Pipelines to create an automated workflow that extracts fresh data, trains the model, and deploys a new version of the model. This option is operationally efficient because it eliminates the need for manual intervention and ensures that your model is always up to date with the latest data. You can also use SageMaker Pipelines to orchestrate your workflow using a graphical interface or a Python SDK1. Option B configures SageMaker Model Monitor with an accuracy threshold to check for model drift. Model drift occurs when the statistical properties of the target variable change over time, which can affect the performance of your model2.

Comment: https://aws.amazon.com/blogs/machine-learning/automate-model-retraining-with-amazon-sagemaker-pipelines-when-drift-is-detected/

Comment: Retrain the model when the accuracy is decreasing is the most recommended way to take of your models.

Comment: The MOST operationally efficient way for the data scientist to maintain the model's accuracy would be to choose options A and B: A. Use SageMaker Pipelines to create an automated workflow that extracts fresh data, trains the model, and deploys a new version of the model. Using SageMaker Pipelines allows the data scientist to automate the entire workflow from data extraction to model deployment. This ensures that the model is trained and deployed on the latest data automatically without the need for manual intervention. The data scientist can set up the pipeline to run on a schedule or trigger it based on certain events. B. Configure SageMaker Model Monitor with an accuracy threshold to check for model drift. Initiate an Amazon CloudWatch alarm when the threshold is exceeded. Connect the workflow in SageMaker Pipelines with the CloudWatch alarm to automatically initiate retraining.

Comment: Using SageMaker Pipelines to create an automated workflow that extracts fresh data, trains the model, and deploys a new version of the model is an efficient way to automate the process of model retraining and deployment. Configuring SageMaker Model Monitor with an accuracy threshold to check for model drift and initiating an Amazon CloudWatch alarm when the threshold is exceeded is an efficient way to monitor the accuracy of the deployed model and initiate retraining when necessary. This approach helps to maintain the accuracy of the model over time.

Comment: https://aws.amazon.com/blogs/machine-learning/automate-model-retraining-with-amazon-sagemaker-pipelines-when-drift-is-detected/

Comment: B first and then A

Discussion for Question 223

Link: https://www.examtopics.com/discussions/amazon/view/100146-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- BEF: 9 votes

Discussion

Comment: A. NO - reducing data will not help in a better model; the more the merrier :-) B. YES - It can address non-linearity in the full spectrum C. NO - reducing data will not help in a better model; the more the merrier :-) D. NO - residual is not constant when price > 50 E. YES - that can help make non-linear data linear F. YES - it can capture more complex relationships

Comment: Option E suggests that you examine the input data, and apply non-linear data transformations where appropriate. This option is helpful because it can reduce the non-linearity in your data and make it more suitable for a linear model. For example, you can apply a logarithmic, square root, or inverse transformation to your price variable and see if it improves the fit of your model1. You can also use the Box-Cox transformation, which is a method that automatically finds the best transformation for your data2. Option F suggests that you use a non-linear model instead of a linear model. This option is also helpful because it can capture the non-linear relationship between price and sales that is evident in your residual plot. Option B suggests that you create two different models for different sections of the data. This option is also helpful because it can account for the different behavior of your data at different price ranges.

Comment: The linear model y = ax + b works well for x < 50, but for x > 50 the residual increases linearly, meaning that the slope linear model increases, i.e., y = a'x + b' with a' != a. Offset will not help. Downsampling will not help either.

Comment: The linear model doesn't capture the data complexity

Replies:

Comment: Then, BEF

Replies:

Comment: It appears on 2023-April-03

Comment: As per wolfsong said

Comment: Two models , add a constant or in-put data transformation

Replies:

Comment: Bde should be the answer

Comment: The residual plot shows that the linear model is not fitting the data well, with a clear pattern indicating that the model is underfitting. To improve the accuracy of the predictions, the ML engineer should take the following actions: C. Downsample the data in sections where Price < 50: This could be an option since there seems to be a higher variance in the residuals in the region where Price < 50. D. Offset the input data by a constant value where Price > 50: This could be an option since there seems to be a systematic bias in the residuals in the region where Price > 50. E. Examine the input data, and apply non-linear data transformations where appropriate: This is necessary since the residual plot shows that the linear model is not capturing the non-linear relationships in the data.

Replies:

Comment: A good residual plot is a flat line at y = 0. So... - Not sure if D is right. If you offset by a constant value, you're just moving the plot up or down. You'd have to add a term like K*Price, where price > 50 and K > 0, for you to flatten that curve beyond Price > 50. - Also unsure about C. The variance looks fairly good for Price < 50 as it's mostly around zero which is what you want. The problem is the residual value at Price > 50 which goes way off. I'd go with B, E & F: E: obvious F: use non-linear model instead as it will remove the kink in the plot B: Not an answer I like, but if you can't use a nonlinear model, you need to use a piecewise-linear model that separates the data in two. Something like this: https://towardsdatascience.com/piecewise-linear-regression-model-what-is-it-and-when-can-we-use-it-93286cfee452

Comment: If D is the answer of this question, Isn't B the another answer too? Suppose that the initial linear model means y = aX + b, then D means y = a(X - C) + b --> y = aX + b' (when Price > 50) I think that D means we would use two different linear models for different sections (Price = 50) of the data.

Discussion for Question 224

Link: https://www.examtopics.com/discussions/amazon/view/98308-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 15 votes

Discussion

Comment: K-means unsupervised learning

Comment: It is A Memory Efficiency: K-nearest neighbors (k-NN) doesn't require storing a model with learned parameters, as it's an instance-based learning algorithm. It simply memorizes the training dataset. Therefore, it saves on memory costs compared to models with learned parameters like linear learners. Dimension Reduction: By employing dimension reduction techniques like Principal Component Analysis (PCA) in conjunction with k-NN, you can reduce the dimensionality of the dataset, which helps in saving memory costs. This makes k-NN with dimension reduction a suitable choice when memory efficiency is a concern. Similar Data Points: K-nearest neighbors naturally provides a measure of similarity between data points. Given a test data point, it finds the k nearest neighbors in the training data. This fulfills the requirement of being able to find similar data points for each test data point.

Comment: A. K-nearest neighbors (k-NN) with dimension reduction (KNN - Useful for Classification task for different types of veggies based on features + Dimensionality reductions like PCA can be applied prior to KNN to reduce no.of features in dataset , thereby saving memory costs during training and model deployment - also remove noise and data redundancy)

Comment: https://www.linkedin.com/advice/3/what-difference-between-knn-k-means-skills-computer-science-cx1hc

Comment: This is un unsupervised clustering problem not a classification one (A). k-means is a better choice from memory efficiency perspective.

Comment: While A is most voted comment, but knn is really high on memory usage as it stores the data points information to make predictions. Just voting for it because it mentions dimensionality reduction is obtuse. C is the next

most probable candidate that fits the bill on every account.

Comment: Should be A, as only A is can be used for classification, finding similar data points and dimensionality reduction

Comment: A. YES - K-NN will find the similar datapoints, and dimension reduction will save memory B. NO - Linear learner is for regression or classification, not finding similar data points C. NO - K-means is for unsupervised clustering, not find closest data ponits D. NO - Principal component analysis (PCA) with the algorithm mode set to random

Comment: C doesn't solve the "too many features" problem + It's well defined the vegetable classes. A is the way

Comment: KNN to reduce dimensionality which may help reduce memory utilisation.

Comment: It's A, needs to reduct the dimensionality of the dataset.

Comment: They wanna less feature

Comment: I will go with A. C is not valid, as K-means is a clustering algorithm that can group similar data points together. However, it does not perform classification, and it is not clear how it addresses the memory cost and similarity search requirements mentioned in the question.

Comment: It should be C, because it is unsupervised classification problem.

Replies:

Comment: Not sure that classification is unsupervised problem

Comment: option A suggests using the k-nearest neighbors (k-NN) algorithm with dimension reduction. The k-NN algorithm can be used for classification tasks and dimension reduction can help reduce memory costs. Additionally, k-NN can be used for finding similar data points. K-NN is a simple algorithm that works well with high-dimensional data and can find similar data points.

Replies:

Comment: agree. By reducing the number of dimensions, you may achieve comparable analysis results using less memory and in a shorter amount of time.

Comment: https://ealizadeh.com/blog/knn-and-kmeans/#:~:text=unsupervised%20learning%20algorithm.-,K%20in%20K%2DMeans%20refers%20to%20the%20number%20of%20clusters,using%20different%20values%20for%20K. KNN does use lot of memory because in lazy learning it stores (memorizes) the training dataset. AWS sagemaker however has an improved version of this algorithm. Because the questions does not mention we have labels, we cannot use supervised learning K-means : unsupervised and helps us to classify different vegetables based on their many features. This also find "similar data points for each test data point" c

Comment: "Training with the k-NN algorithm has three steps: sampling, dimension reduction, and index building. Sampling reduces the size of the initial dataset so that it fits into memory. For dimension reduction, the algorithm decreases the feature dimension of the data to reduce the footprint of the k-NN model in memory and inference latency." " The main objective of k-NN's training is to construct the index. The index enables efficient lookups of distances between points whose values or class labels have not yet been determined and the k nearest points to use for inference." https://docs.aws.amazon.com/sagemaker/latest/dg/k-nearest-neighbors.html

---

Discussion for Question 225

Link: https://www.examtopics.com/discussions/amazon/view/98200-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 11 votes

Discussion

Comment: It has to be C.

Comment: Option C uses the SageMaker Debugger vanishing_gradient and LowGPUUtilization built-in rules to detect issues and to launch the StopTrainingJob action if issues are detected. This option is the most efficient because it leverages the existing features of SageMaker Debugger to monitor and troubleshoot your training job without requiring any additional development effort. You can use the following steps to implement this option.

Comment: Answer is C The best option for the data scientist to identify and address training issues with the least development effort is option C: Use the SageMaker Debugger vanishing_gradient and LowGPUUtilization built-in rules to detect issues and to launch the StopTrainingJob action if issues are detected. SageMaker Debugger is a tool that helps to debug machine learning training processes. It provides several built-in rules that can detect and diagnose common issues that can occur during training. In this case, the data scientist suspects that the training is not converging and that resource utilization is not optimal. The vanishing_gradient and LowGPUUtilization rules can help to identify these issues.

Comment: C. Use the SageMaker Debugger vanishing_gradient and LowGPUUtilization built-in rules to detect issues and to launch the StopTrainingJob action if issues are detected. The SageMaker Debugger is a built-in tool that helps with debugging and profiling machine learning models trained in SageMaker. In this scenario, the data scientist suspects that there are issues with the training process, so using the SageMaker Debugger is the most appropriate solution. The vanishing_gradient and LowGPUUtilization built-in rules can detect common training issues such as a vanishing gradient problem or low GPU utilization, which could affect the training convergence and resource utilization. By launching the StopTrainingJob action if issues are detected, the training job can be stopped early, which can help to save resources and time. This approach requires the least development effort, as it is built-in to SageMaker and does not require the data scientist to create custom metrics or configure CloudWatch alarms.

Comment: https://docs.aws.amazon.com/sagemaker/latest/dg/debugger-built-in-rules.html

---

Discussion for Question 226

Link: https://www.examtopics.com/discussions/amazon/view/103014-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 13 votes

Discussion

Comment: As the model has been already trained and deployed, I will go with C. because B (SageMaker Debugger) is used at training time

Comment: C is right B and C are possible solutions, but the question requested the MOST operationally efficient manner

Comment: SageMaker Clarify provides tools to help ML modelers and developers understand model characteristics as a whole prior to deployment and to debug predictions provided by the model after it's deployed Option B is not recommended because retraining the model with SageMaker Debugger and configuring Debugger to calculate and collect Shapley values is time-consuming and may not be operationally efficient

Comment: It seems that both B and C are possible answers. SHAP baselines can be provided by both. The scenario says nothing about concern of bias, so perhaps Clarify is overkill? This post from AWS seem to be specifically addressing this case, and uses SageMaker Debugger. https://aws.amazon.com/blogs/machine-learning/ml-explainability-with-amazon-sagemaker-debugger/

Comment: Selected Answer: B Retrain the model by using SageMaker Debugger. Configure Debugger to calculate and collect Shapley values. Create a chart that shows features and SHapley Additive explanations (SHAP) values to explain how the features affect the model outcomes. While A, C, and D are all options for explaining the model's behavior, the most efficient way to meet the bank's requirements is to use SageMaker Debugger to calculate and collect S Shapley values for each prediction. This allows the data science team to easily explain why the model denied the promotion to certain customers. SageMaker Debugger also provides built-in integration with SageMaker Studio, which enables data scientists to visualize the Shapley values and other debugging information through a user-friendly interface.

Comment: I think it's D. Model monitor automatically integrated with Clarify

Comment: B. Retrain the model by using SageMaker Debugger. Configure Debugger to calculate and collect Shapley values. Create a chart that shows features and Shapley Additive explanations (SHAP) values to explain how the features affect the model outcomes would be the most operationally efficient way to meet the requirement of explaining why the model denies the promotion to some customers.

Comment: Clarify is the best solution. The key is training data https://www.amazonaws.cn/en/sagemaker/clarify/

Replies:

Comment: Amazon SageMaker Clarify provides machine learning developers with greater visibility into their training data and models so they can identify and limit bias and explain predictions. I think B is correct.

Comment: Its between B and C SageMaker Clarify is used to promote transparency and accountability in machine learning models. Thats what we are looking for why model denies promotion to some customers

Comment: "Explain individual model predictions Customers and internal stakeholders both want transparency into how models make their predictions. SageMaker Clarify integrates with SageMaker Experiments to show you the importance of each model input for a specific prediction. Results can be made available to customer-facing employees so that they have an understanding of the model's behavior when making decisions based on model predictions." https://www.amazonaws.cn/en/sagemaker/clarify/

Discussion for Question 227

Link: https://www.examtopics.com/discussions/amazon/view/103018-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- DEF: 8 votes

Discussion

Comment: It is leaning towards bias in data, rather than probability distribution.

Comment: BIAS in _data_ before training uses following matrices B - helps assess whether there is bias in how loan amounts are distributed among different categories C - ompares the proportions of positive (e.g., approved loans) and negative (e.g., rejected loans) outcomes across different facets (demographic groups) F - High total variation distance between between the predicted and observed labels suggests potential bias

Comment: Question asks for distributions. DEF are distributions. ABC are imbalances or disparities.

Comment: Since it is indicated in the official web site that D, E and F are used to determine how different the distributions for loan application outcomes are for different demographic groups https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-measure-data-bias.html

Comment: All valid answers ? They are listed as "pre-training bias" here https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-measure-data-bias.html

Replies:

> **Comment:** Jensen-Shannon divergence, Kullback-Leibler divergence, and Total variation distance, are used to measure differences between probability distributions, but they are not specifically pretraining bias metrics for checking bias distribution concerning categorical variables in this context.

Comment: confusing but lean towards B D and F

Replies:

> **Comment:** Confusing with DEF

Comment: All are valid answers, so definitely un scored question https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-measure-data-bias.html

Comment: answers "How different are the distributions for loan application outcomes for different demographic groups?"

Comment: https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-measure-data-bias.html

Comment: D. Jensen-Shannon divergence and E. Kullback-Leibler divergence are post-training bias metrics that measure the distance between two probability distributions. They are not pretraining bias metrics and cannot be used to check the bias distribution of the dataset. F. Total variation distance is a post-training bias metric that measures the difference between two probability distributions. It is not a pretraining bias metric and cannot be used to check the bias distribution of the dataset. Send a message...

Replies:

> **Comment:** The are all pretrainimg metrics https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-measure-data-bias.html
>
> ### Replies:
>
> > **Comment:** This is correct.... they are all valid answers. Seems this is one of the un-scored questions... those 15 that are used to calibrate or test possible future questions.

Comment: Agree with austinoy, answer should be DEF.

Comment: https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-measure-data-bias.html

Replies:

> **Comment:** based on the link shouldn't DEF be the answers?

Discussion for Question 228

Link: https://www.examtopics.com/discussions/amazon/view/103021-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 11 votes

Discussion

Comment: WQL is particularly useful when there are different costs for underpredicting and overpredicting. By setting the weight (τ) of the wQL function, you can automatically incorporate differing penalties for underpredicting and overpredicting.

Comment: A. NO - lot of data is available, better to use CNN-QR B. NO - lot of data is available, better to use CNN-QR C. YES - wQL is particularly useful when there are different costs for underpredicting and overpredicting (https://docs.aws.amazon.com/forecast/latest/dg/metrics.html#metrics-wQL) D. NO - WAPE will measure deviation, but no over vs. under forecasting

Comment: It should be A I think. CNN is for image analysis

Comment: Option C also suggests evaluating the model by using the Weighted Quantile Loss (wQL) metric at 0.75 (P75). This metric measures the accuracy of a model at a specified quantile, which is a point in the distribution of possible outcomes2. For example, P75 means that 75% of the outcomes are below that point, and 25% are above it. This metric is suitable for your use case because it can incorporate different costs for underpredicting and overpredicting2. Since the cost of running out of items in stock is much higher for your company than the cost of having excess inventory, you can set a high weight (τ) for the wQL function to penalize underpredictions more than overpredictions2. This way, you can optimize your model to produce prediction results that will maximize your company profit.

Comment: A retail company wants to use Amazon Forecast to predict daily stock levels of inventory. The cost of running out of items in stock is much higher for the company than the cost of having excess inventory. The company has millions of data samples for multiple years for thousands of items. The company's purchasing department needs to predict demand for 30-day cycles for each item to ensure that restocking occurs.

Replies:

> **Comment:** So, what is an argument for A here?

Comment: I'll go with, C? https://docs.aws.amazon.com/forecast/latest/dg/aws-forecast-algo-cnnqr.html https://docs.aws.amazon.com/forecast/latest/dg/metrics.html#metrics-wQL

Comment: https://docs.aws.amazon.com/forecast/latest/dg/metrics.html https://docs.aws.amazon.com/forecast/latest/dg/aws-forecast-recipe-arima.html https://docs.aws.amazon.com/forecast/latest/dg/aws-forecast-algo-cnnqr.html

---

Discussion for Question 229

Link: https://www.examtopics.com/discussions/amazon/view/103024-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- ACD: 20 votes

Discussion

Comment: A - Running the training jobs in a private VPC will ensure that the data is transmitted over an encrypted channel. Enabling inter-container traffic encryption will encrypt data that is transmitted between containers. This will help protect the data during the distributed training. C - Creating an S3 VPC endpoint will provide a secure and private connection between the VPC and the S3 bucket. Network routes, endpoint policies, and S3 bucket policies can be configured to further secure the data during the distributed training. D - Granting read-only access to SageMaker resources by using an IAM role will ensure that the data is only accessed by the necessary resources during the distributed training. This will help prevent unauthorized access to the data.

Comment: I'm not agree with E because assigns read-only access to Sagemaker

Comment: A and C are final, I think E is the third option where inbound traffic cannot access VPC resources.

Comment: Changing my options to ACD

Comment: Will go for A,C & F. Please check the keyword "Distributed" in question. In case of a distributed training, instances within a same security group are required to communicate with each other which is configured by allowing inbound traffic through security group. Check this section (Configure the VPC Security Group) in this document - https://docs.aws.amazon.com/sagemaker/latest/dg/train-vpc.html

Comment: A -->This ensures that the data is not exposed to the public internet and that all traffic between containers is encrypted C--> This ensures that all traffic between the Amazon SageMaker instances and the S3 bucket is kept within the VPC and is not exposed to the public internet. The endpoint policies and S3 bucket policies can be used to control access to the data D--> This ensures that only authorized users can access the SageMaker resources (option B) is not necessary as running jobs in a private VPC provides sufficient security creating a NAT gateway and assigning an Elastic IP address for the NAT gateway (option E) is not necessary as it does not provide any additional security benefits configuring an inbound rule to allow traffic from a security group that is associated with the training instances (option F) is not necessary as it does not provide any additional security benefits especially in the Prescence of the private endpoint

Comment: Reference is https://docs.aws.amazon.com/sagemaker/latest/dg/train-vpc.html A. YES - need a private VPC, inter-container traffic encryption optionnal but ok B. NO - no need for multple VPC C. YES - S3 VPC endpoint will prevent the traffic to flow through the internet D. YES - SageMaker resources (instances here) need to read the S3 files E. NO - NAT gateway is allow outbound traffic from a private subnet to Internet; not needed F. NO - The training instances does not need to receive inbound connections

Comment: ACD Is what I am going with ---------- It was a tough choice between D and F but when I look at Protecting the Data as the main point of the question I went with D (read only to S3)

Comment: think the best combination of steps for you are A, C, and D

Comment: Letra B está errada, pois torna o processo muito complexo. Letra A - C - D estão corretas. Letra F está errada, pois Inbound Rules não são relevantes para S3. Finalmente, Letra E é desnecessária.

Comment: Based on the context the inbound should be added to the data, which is stored in S3. Inbound rules are not relevant to S3. D should be correct instead of F.

Comment: A C and F look correct

Replies:

> **Comment:** Configure the VPC Security Group In distributed training, you must allow communication between the different containers in the same training job. To do that, configure a rule for your security group that allows inbound connections between members of the same security group. For EFA-enabled instances, ensure that both inbound and outbound connections allow all traffic from the same security group. For information, see Security Group Rules in the Amazon Virtual Private Cloud User Guide.

Comment: A,B pretty sure; D best guess. https://docs.aws.amazon.com/sagemaker/latest/dg/train-encrypt.html https://docs.aws.amazon.com/sagemaker/latest/dg/train-vpc.html

Replies:

> **Comment:** My best "guess" is ACF

Discussion for Question 230

Link: https://www.examtopics.com/discussions/amazon/view/103027-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 8 votes

Discussion

Comment: Just additional information, Elastic Inference is being deprecated and recommendation is use AWS Inferentia

Comment: Option A can help you meet your requirements most cost-effectively because it enables you to choose the instance type that is best suited to the overall compute and memory needs of your application, and then separately specify the amount of inference acceleration that you need. This reduces inference costs by up to 75% because you no longer need to over-provision GPU compute for inference1.

Comment: We want to improve the inference of the model. That said, Letter B - C does not solve this problem. Letter D solves it, but at a very high cost. Letter A is correct, as we solve the problem at the lowest possible cost.

Comment: Use Amazon Elastic Inference on the SageMaker hosted endpoint would be the most cost-effective solution for increasing throughput and decreasing latency. Amazon Elastic Inference is a service that allows you to attach GPU-powered inference acceleration to Amazon SageMaker hosted endpoints and EC2 instances. By attaching an Elastic Inference accelerator to the SageMaker endpoint, you can achieve better performance with lower costs than using a larger, more expensive instance type.

Comment: "cost efficient" therefore A based on slide 20: https://pages.awscloud.com/rs/112-TZM-766/images/AL-ML%20for%20Startups%20-%20Select%20the%20Right%20ML%20Instance.pdf

---

Discussion for Question 231

Link: https://www.examtopics.com/discussions/amazon/view/103028-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- ACE: 12 votes

Discussion

Comment: ACE is correct.

Comment: A. YES - AWS Lake Formation is fully managed & integrated B. NO - we want to use Lake Formation instead of raw Glue (Formation built on top of Glue) C. YES - Glue used in conjunction with Lake Formation (https://docs.aws.amazon.com/lake-formation/latest/dg/glue-features-lf.html) D. Apply granular access policies by using AWS Identity and Access Management (1AM). Configure server-side encryption on each data source. E. YES - AWS Lake Formation is fully managed & integrated F. NO - we want to use Lake Formation instead of raw Glue

Comment: Agree A, C, E

Comment: https://docs.aws.amazon.com/lake-formation/latest/dg/what-is-lake-formation.html

Replies:

> **Comment:** Lake Formation provides a single place to manage access controls for data in your data lake. You can define security policies that restrict access to data at the database, table, column, row, and cell levels. These policies apply to IAM users and roles, and to users and groups when federating through an external identity provider. You can use fine-grained controls to access data secured by Lake Formation within Amazon Redshift Spectrum, Athena, AWS Glue ETL, and Amazon EMR for Apache Spark. Whenever you create IAM identities, make sure to follow IAM best practices.

Comment: ACE looks good

Comment: I will choose ACE https://aws.amazon.com/blogs/big-data/build-secure-encrypted-data-lakes-with-aws-lake-formation/

Comment: ACE looks legit.

Comment: ACE looks correct

Comment: I'll go with, ACE?

Comment: A: https://docs.aws.amazon.com/lake-formation/latest/dg/what-is-lake-formation.html C: https://docs.aws.amazon.com/lake-formation/latest/dg/upgrade-glue-lake-formation.html D: https://docs.aws.amazon.com/lake-formation/latest/dg/what-is-lake-formation.html

Replies:

> **Comment:** Option D is incorrect and in fact, Server-side encryption would be discarded in favor of Lakeformation (Glue) encryption as per this AWS document - https://docs.aws.amazon.com/glue/latest/dg/encryption-security-configuration.html ACE are the correct options.

Discussion for Question 232

Link: https://www.examtopics.com/discussions/amazon/view/103029-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- DEF: 18 votes

Discussion

Comment: A: possible but unlikely for movie reviews B: wrong https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwi31N_10eX9AhWYQ0EAHXDFCAwQFnoECA8QAw&url=https%3A%2F%2Fdeepchecks.com%2Fquestion%2Fdoes-learning-rate-affect-overfitting%2F&usg=AOvVaw19RT-u_XyEe8FG_10R6aFC C: wrong because would increase complexity and potentially overfitting D: correct E: correct F: correct

Comment: Overfitting solutions must be regularization, dropout and adjusting learning rate. F is wrong, decreasing number of layers is not the top recommendations to solve overfitting, it may even cause underfitting.

Comment: A. NO B. NO - decreasing the learning rate make may increase accuracy thus increase overfitting C. NO - more complexity tend to increase overfitting D. YES - best practice E. YES - best practice, will reduce model complexity and thus increase generalization F. YES - best practice, will reduce model complexity and thus increase generalization

Comment: d and E for sure. i am a bit confused F and B

Comment: To improve the generalization of the deep learning sentiment analysis model and reduce overfitting, the following three solutions can be implemented: Add Dropout: Dropout is a regularization technique that randomly drops out (sets to zero) a certain percentage of nodes in the neural network during each training epoch. This helps to prevent overfitting and improve generalization. Add L1 and L2 Regularization: L1 and L2 regularization are techniques used to add a penalty to the loss function of the neural network, which helps to prevent overfitting. L1 regularization adds a penalty based on the absolute value of the weights, while L2 regularization adds a penalty based on the squared value of the weights. Decrease the number of layers in the network: Deep neural networks with too many layers can be prone to overfitting. Reducing the number of layers in the network can help to prevent overfitting and improve generalization.

Comment: We don't have to touch learning rate because the model is overfitting

Comment: DEF are correct

Discussion for Question 233

Link: https://www.examtopics.com/discussions/amazon/view/103030-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 11 votes

Discussion

Comment: A. YES - best to reduce feature count B. NO - L2 will reduce large weights and smooth features, to get rid of them C. NO - dropout is for NN D. NO - we are already converging, no need for more data

Comment: L1 regulization

Comment: A is correct

Comment: L1 shrinks the less important feature's coefficient to zero thus, removing some feature altogether. So, this works well for feature selection in case we have a huge number of features.

Comment: Yes L1 for feature reduction

Comment: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwim5YGd0-X9AhXPYcAKHSAlAkoQFnoECA0QAQ&url=https%3A%2F%2Fdocs.aws.amazon.com%2Fmachine-learning%2Flatest%2Fdg%2Fmodel-fit-underfitting-vs-overfitting.html&usg=AOvVaw2jwLt-J0jRSWeiDyjEzI_S

Discussion for Question 234

Link: https://www.examtopics.com/discussions/amazon/view/103032-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 11 votes

Discussion

Comment: SMOTE for minority class for unbalanced data

Comment: A. YES - we want to oversample the minority class = Fraud B. NO - we want more fraudulent cases C. NO - we want more fraudulent cases D. NO - we want more fraudulent cases

Comment: By applying SMOTE, you can balance the class distribution and increase the diversity of your data, which can help your model learn better and reduce overfitting1. You can use the imbalanced-learn library in Python to implement SMOTE on your data2.

Comment: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwjLkqfb1OX9AhXkQ0EAHYlVDq0QFnoECBMQAw&url=https%3A%2F%2Ftowardsdatascience.com%2F5-smote-techniques-for-oversampling-your-imbalance-data-b8155bdbe2b5&usg=AOvVaw1FdrxDEbLDjNhacXn3d-Tu

---

Discussion for Question 235

Link: https://www.examtopics.com/discussions/amazon/view/103033-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- D: 24 votes
- B: 11 votes

Discussion

Comment: When to use fast file mode: For larger datasets with larger files (more than 50 MB per file), the first option is to try fast file mode, which is more straightforward to use than FSx for Lustre because it doesn't require creating a file system, or connecting to a VPC. Fast file mode is ideal for large file containers (more than 150 MB), and might also do well with files more than 50 MB. https://docs.aws.amazon.com/sagemaker/latest/dg/model-access-training-data.html#model-access-training-data-best-practices

Comment: https://aws.amazon.com/about-aws/whats-new/2021/10/amazon-sagemaker-fast-file-mode/

Comment: Least setup is D, B could work but requires more setup!

Comment: D Amazon SageMaker now supports Fast File Mode for accessing data in training jobs. This enables high performance data access by streaming directly from Amazon S3 with no code changes from the existing File Mode. For example, training a K-Means clustering model on a 100GB dataset took 28 minutes with File Mode but only 5 minutes with Fast File Mode (82% decrease). https://aws.amazon.com/about-aws/whats-new/2021/10/amazon-sagemaker-fast-file-mode/

Comment: B. Yes Please, look this link (https://aws.amazon.com/blogs/aws/enhanced-amazon-s3-integration-for-amazon-fsx-for-lustre/)

Comment: A. NO - the files are too big and will fill the instance storage for no reason B. NO - Lustre create stripes for each file on different hard drives, maximizing throughput; our challenge is more about the volume of data to be made available on the training instance, not throughput C. NO - EFS support File semantic, but does not change any system property D. YES - FastFile allows training to start before the full file has been downloaded (like Pipe Mode) but does not require code change

Comment: changing to D

Comment: although D is very tempting but leaning towards B

Comment: https://aws.amazon.com/blogs/machine-learning/ensure-efficient-compute-resources-on-amazon-sagemaker/

Comment: When to use fast file mode For larger datasets with larger files (more than 50 MB per file), the first option is to try fast file mode, which is more straightforward to use than FSx for Lustre because it doesn't require creating a file system, or connecting to a VPC. Fast file mode is ideal for large file containers (more than 150 MB), and might also do well with files more than 50 MB. Because fast file mode provides a POSIX interface, it supports random reads (reading non-sequential byte-ranges). However, this is not the ideal use case, and your throughput might be lower than with the sequential reads. However, if you have a relatively large and computationally intensive ML model, fast file mode might still be able to saturate the effective bandwidth of the training pipeline and not result in an IO bottleneck.

Comment: Option D, FastFile mode, streams files on demand from S3 buckets to the training instance, which can be efficient for small datasets but may not be optimal for large datasets. Moreover, this solution does not provide a file system that is optimized for high performance, and it may require additional development effort to set up

Replies:

> **Comment:** B because we have 200TB https://saturncloud.io/blog/using-aws-sagemaker-input-modes-amazon-s3-efs-or-fsx/

Comment: Fast File Mode combines the ease of use of the existing File Mode with the performance of Pipe Mode. This provides convenient access to data as if it was downloaded locally, while offering the performance benefit of streaming the data directly from Amazon S3. No code change required or no lengthy setup

Comment: The solution that meets the requirements of the company is B, which involves creating an Amazon FSx for Lustre file system and linking it to the S3 buckets. Amazon FSx for Lustre is a fully managed, high-performance file system optimized for compute-intensive workloads, such as machine learning training. It is designed to provide low latencies and high throughput for processing large data sets, and it can directly access data from S3 buckets without any data movement or copying. This solution requires minimal setup and provides the shortest processing time since the data can be accessed in parallel by multiple instances.

Comment: I will go with D https://sagemaker.readthedocs.io/en/stable/api/utility/inputs.html

Comment: https://aws.amazon.com/blogs/machine-learning/speed-up-training-on-amazon-sagemaker-using-amazon-efs-or-amazon-fsx-for-lustre-file-systems/

Discussion for Question 236

Link: https://www.examtopics.com/discussions/amazon/view/103035-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 18 votes

Discussion

Comment: https://docs.aws.amazon.com/machine-learning/latest/dg/data-transformations-reference.html

Comment: C. Quantile binning: Quantile binning (or discretization) involves dividing a continuous variable into bins based on quantiles. This can be useful for handling skewed data by distributing the data more evenly across the bins. However, this method transforms the numerical feature into a categorical one, which might not be ideal for preserving the ordinal nature and the detailed variance of the 'duration' feature in a regression model. If the choice must be made from the given options, Option C (Quantile binning) might be the most suitable, albeit not ideal, as it can at least help in dealing with skewed distributions by distributing the data across bins more evenly. However, the data scientist should consider logarithmic or polynomial transformations for a more direct approach to addressing non-linearity.

Comment: A. NO - One-hot encoding is for featurization of categories B. NO - C. YES - Quantile binning can make data linear (https://docs.aws.amazon.com/machine-learning/latest/dg/data-transformations-reference.html#quantile-binning-transformation) D. NO - Normalization will recenter the data, not change the relationship

Comment: quantile binning

Comment: C is correct

Comment: C is the best answer I guess

Comment: the correct answer is C, Quantile binning. This transformation divides the data into quantiles (equal-sized intervals) based on the values of the feature (in this case, duration) and replaces the values with the bin number. This transformation can help capture non-linear relationships between features by creating more representative categories for skewed data. The transformed data can then be used to train a non-linear regression model, such as a polynomial regression, to better predict future book sales.

Discussion for Question 237

Link: https://www.examtopics.com/discussions/amazon/view/103036-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 13 votes

Discussion

Comment: For the Marketing and Human Resources department user groups, attach an IAM policy that provides access to only the folder that contains the non-sensitive datasets. Finance department user also need access to non-sensitive datasets.

Comment: I think attaching the policy is more flexible, in case this pattern needs to be repeated for another s3 bucket?

Comment: You cannot identify a user group as a principal in a policy (such as a resource-based policy) because groups relate to permissions, not authentication, and principals are authenticated IAM entities. an Amazon S3 bucket policy cannot have a user group as the principal directly. https://docs.aws.amazon.com/IAM/latest/UserGuide/reference_policies_elements_principal.html I stand corrected. I retract my previous answer.

Comment: D Use a bucket policy. User group cannot be a principal in IAM policy. adding each individual user to the policy is not practical

Comment: According to the AWS documentation, you cannot specify an IAM group as a principal in an S3 bucket policy. This is because groups relate to permissions, not authentication, and principals are authenticated IAM entities. You can only specify the following principals in a policy: AWS account and root user IAM user Federated user IAM role . If you want to grant permission to an IAM group, you can add the ARNs of all the IAM users in that group to the S3 bucket policy instead. so it is C to create 2 IAM roles and attach them to different groups you have

Replies:

Comment: REF: https://docs.aws.amazon.com/AmazonS3/latest/userguide/s3-bucket-user-policy-specifying-principal-intro.html

Comment: https://docs.aws.amazon.com/AmazonS3/latest/userguide/walkthrough1.html it does not show any option to use iam group based s3 bucket policy. (so D cannot be the right answer)

Comment: changing to C

Comment: Option D suggests creating a single S3 bucket that includes two folders to separate the sensitive datasets from the non-sensitive datasets. This option is helpful because it can simplify the data management and reduce the cost of using multiple S3 buckets. You can use a single S3 bucket to store all your datasets and use folders to organize them by their sensitivity level1. You can also use the Amazon S3 console or the AWS CLI to create and manage your folders2.

Comment: First it is more efficient to use one single bucket, S3 has limit of 100 buckets by default, answer C creates two policies while for answer D , it is done with one, and use Deny on the sensitive folder to the two groups not finance, and have an allow to the non sensitive, knowing that deny takes precendence

Comment: In S3 bucket Policy you CANNOT specify IAM Group as Principal, but you can specify IAM Users. So it's C.

Comment: Option C https://stackoverflow.com/questions/35944349/iam-aws-s3-to-restrict-to-a-specific-sub-folder https://aws.amazon.com/blogs/security/how-to-restrict-amazon-s3-bucket-access-to-a-specific-iam-role/

Comment: I will choose C

Comment: I will choose C

Comment: Both B and D look apparently correct but they are not because in s3 bucket policy , IAM Group cant be the principal. In other words you cant give access to a User group to s3 buckets using s3 bucket policy. It can only be an IAM user or role.https://stackoverflow.com/questions/30667678/s3-bucket-policy-how-to-allow-a-iam-group-from-another-account I would go for C

Comment: single bucket looks a better option. Ease of management and still secure

Replies:

Comment: Actually this is not possible. I will go for C

Replies:

Comment: https://stackoverflow.com/questions/30667678/s3-bucket-policy-how-to-allow-a-iam-group-from-another-account

Comment: Creating a single S3 bucket that includes two folders to separate the sensitive datasets from the non-sensitive datasets would be the best approach. The policy of the S3 bucket can be set to allow only the Finance department user group to access the folder that contains the sensitive datasets. The folder that contains non-sensitive datasets can be made available to all three department user groups. This approach will ensure that sensitive datasets are only accessible to users who need access to them.

Comment: I'll go with D

Replies:

Comment: I stand corrected - it's C

Discussion for Question 238

Link: https://www.examtopics.com/discussions/amazon/view/112449-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 6 votes

Discussion

Comment: A and B the only options to consider, A talks about Rekognition which is not well suited for video so B

Comment: Amazon Kinesis Video Streams is a fully managed service that makes it easy to ingest, store, and analyze streaming video data. The built-in HTTP live streaming (HLS) capability allows the security team to view the data in real time. Amazon Kinesis Video Streams is a pay-per-use service, so the company will only be charged for the amount of data that it ingests, stores, and analyzes.

Comment: Amazon Kinesis Video Streams is a fully managed service that makes it easy to ingest, store, and analyze streaming video data. The built-in HTTP live streaming (HLS) capability allows the security team to view the data in real time. Amazon Kinesis Video Streams is a pay-per-use service, so the company will only be charged for the amount of data that it ingests, stores, and analyzes.

Comment: It's B real time video ingestion = KVS (C and D are wrong) watch the footage = HLS (rekognition would be for ML, which is not required so A is wrong)

Comment: No any ML involved here, so it's B.

Comment: Selected B as per https://aws.amazon.com/about-aws/whats-new/2018/07/kinesis-video-adds-hls-support/

Discussion for Question 239

Link: https://www.examtopics.com/discussions/amazon/view/112450-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 9 votes

Discussion

Comment: Amazon Rekognition for image analysis and A21 private workforce for manual review

Comment: Amazon Rekognition is an image and video analysis service that can detect objects, scenes, and faces in images. The company can use Amazon Rekognition to automatically process images of plaques and identify defects that should cause a plaque to be rejected. Low-confidence predictions can be sent to an internal team of reviewers who are using Amazon A2I with a private workforce option for manual review. This will ensure that the plaques are thoroughly checked before being mailed to customers A is not suitable because Amazon Textract is a service that extracts text and data from scanned documents. It is not designed for image analysis. C is also not suitable because Amazon Transcribe is a service that converts speech to text. It is not designed for image analysis. D is not suitable because AWS Panorama is a computer vision service that runs on cameras and other edge devices. It is not designed for analyzing images stored in an S3 bucket

Comment: Amazon Rekognition is a service that can be used to detect objects, faces, and text in images. Amazon A2I is a service that can be used to automate manual tasks, such as reviewing images for defects. The private workforce option for Amazon A2I allows the engraving company to create a custom workforce of reviewers who are familiar with the types of defects that should cause a plaque to be rejected.

Comment: Keywords are image processing and internal users for review.

Discussion for Question 240

Link: https://www.examtopics.com/discussions/amazon/view/112451-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 7 votes

Discussion

Comment: From what I see, C is the only option that will meet the time constratins

Comment: A. NO - no need for intermediary S3 storage B. NO - Feature store does not have built-in transformations C. YES - https://aws.amazon.com/blogs/machine-learning/using-streaming-ingestion-with-amazon-sagemaker-feature-store-to-make-ml-backed-decisions-in-near-real-time/ D. NO - Computing a query time is expensive, you want it done once and cached

Comment: Amazon Kinesis Data Analytics is a fully managed service that makes it easy to process streaming data. Amazon Kinesis Data Analytics SQL is a feature of Amazon Kinesis Data Analytics that allows you to process streaming data using SQL. AWS Lambda is a serverless compute service that allows you to run code without provisioning or managing servers. SageMaker Feature Store is a managed service that makes it easy to store and manage features for machine learning models.

Replies:

> **Comment:** agree with all but the only question is how can we consume data directly using Kinesis Data Analytics? Don't we need Kinesis Data Stream or Firehose to consume the stream data?

Comment: The letter B is wrong as KDS does not have the ability to load (another service is needed for this, such as KDF). The letter D is wrong as it saves a variable that needs to be accessed quickly in an offline group in the Feature Store. Since the solution starts with KDS and we need the moving average results to be displayed in near real time, the letter C guarantees this: KDS → KDA → Lambda (triggered quickly) → SM FS. Letter A is wrong, as it does not guarantee near real-time feedback.

Comment: KDA provides facility for rolling averages and meet with realtime requirement

---

Discussion for Question 241

Link: https://www.examtopics.com/discussions/amazon/view/112452-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 9 votes

Discussion

Comment: A. YES - Quicksight has Z-score algorithm to detect outliers B. NO - Kinesis Data Streams cannot stream directly to Amazon S3. C. NO - RCF is overkill when QuickSight supports D. NO - Kinesis Data Streams cannot stream directly to Amazon S3.

Comment: both a and c are right.. quicker is a .. https://aws.amazon.com/quicksight/features-ml/

Comment: Amazon Kinesis Data Firehose is a fully managed service that makes it easy to stream data to Amazon S3. Amazon QuickSight ML Insights is a feature of Amazon QuickSight that allows you to detect outliers in your data using machine learning algorithms. Amazon QuickSight is a fully managed business intelligence service that allows you to visualize your data in dashboards.

Comment: Letters B - D are wrong, because KDS has no load power, that is, directly saving the files in any other service (you would need, for example, a KDF coupled to KDS). Letter A is correct, as QuickSight has tools to identify outliers. Letter C would be correct, but it requires more development to use something we already have in QuickSight.

Comment: It's should be Firehose, and then it's should be least development effort, SageMaker is complicated and require a lot of effort. so it's A.

Comment: Keywords Near real-time, visualization with minimal dev efforts

---

Discussion for Question 246

Link: https://www.examtopics.com/discussions/amazon/view/112456-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- D: 8 votes

Discussion

Comment: GetRecord API of Feature Store to retrieve the record

Comment: https://docs.aws.amazon.com/ko_kr/sagemaker/latest/APIReference/API_feature_store_BatchGetRecord.html

Comment: Why not A?

Comment: D. Using the SageMaker Feature Store GetRecord API with the record identifier1. This API allows customers to retrieve features from a single feature group and access one record per API call2. The record identifier is a unique value that identifies a record within a feature group1. The GetRecord API returns the latest version of the record by default1. This solution avoids the need to use additional queries or filters to find the latest record.

Comment: GetRecord API retrieves latest record whereas BatchGetRecrdbatch of records https://docs.aws.amazon.com/sagemaker/latest/APIReference/API_feature_store_GetRecord.html

---

Discussion for Question 247

Link: https://www.examtopics.com/discussions/amazon/view/112457-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- AE: 7 votes

Discussion

Comment: Create a New Endpoint Configuration: Option A: Create a new endpoint configuration that includes a production variant for each of the two models (the existing model and the new model). This allows traffic to be split between the two models, enabling comparative performance analysis. Update the Existing Endpoint: Option E: Update the existing endpoint to use the newly created endpoint configuration. This ensures that both models receive traffic according to the specified distribution, facilitating A/B testing on production data.

Comment: A. YES - we create a new configuration with 2 production variants (1) B. Create a new endpoint configuration that includes two target variants that point to different endpoints. C. NO - that would defeat A D. NO - that would defeat A E. YES - we modify the existing endpoint to now the new configuration (2)

Comment: A & E - supported by docs listed below.

Comment: A & E ....

Comment: Letter A - E are correct for doing the A/B test: We create an endpoint with production variant → We update the existing endpoint. C is wrong, because invalidate Letter A.

Comment: AE per https://docs.aws.amazon.com/sagemaker/latest/dg/serverless-endpoints-update.html This blog https://aws.amazon.com/blogs/machine-learning/a-b-testing-ml-models-in-production-using-amazon-sagemaker/ Say Deploy just because they deploy two models from the beginning.

Comment: https://docs.aws.amazon.com/sagemaker/latest/dg/serverless-endpoints-update.html E is a better choice than C as it's more specific... the endpoint must be updated to point to the new endpoint configuration.

Comment: https://aws.amazon.com/blogs/machine-learning/a-b-testing-ml-models-in-production-using-amazon-sagemaker/

Discussion for Question 248

Link: https://www.examtopics.com/discussions/amazon/view/112600-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 8 votes

Discussion

Comment: Data Wrangler can transform the date time to desire features

Comment: Amazon SageMaker Data Wrangler is a visual data preparation tool that makes it easy to clean, transform, and featurize data for machine learning. It provides a variety of built-in transformations, including the Featurize date/time transform, which can be used to generate the new variables from the timestamp variable. The other options require the data scientist to develop code, which can be more time-consuming and error-prone. Amazon EMR and AWS Glue are both batch processing services that can be used to run Python code. However, they require the data scientist to create and manage a cluster, which can be a significant operational overhead. Amazon SageMaker Processing is a serverless processing service that can also be used to run Python code. However, it is more expensive than Data Wrangler and does not provide the same level of visual tooling.

Comment: Letra C é a correta, pois o Data Wrangler permite low code para realizar esta tarefa e como queremos o menor operational overhead esta é a solução. Letra D também é possível, mas envolve desenvolvimento de código ficando mais complexa que a Letra C. Letra A requer subir um novo serviço e Letra B cai no mesmo cenário da Letra D (desenvolver código).

Comment: https://aws.amazon.com/blogs/machine-learning/prepare-time-series-data-with-amazon-sagemaker-data-wrangler/ "Featurize datetime time series transformation to add the month, day of the month, day of the year, week of the year, and quarter features to our dataset."

Discussion for Question 249

Link: https://www.examtopics.com/discussions/amazon/view/113070-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 10 votes

Discussion

Comment: We have a supervised classification problem + numerical features, clearly XGBoost is the best candidate

Comment: A. NO - it is not a foreacasting problem B. YES - it is a supervised classification problem C. NO - LDA is for topic modeling, sentiment analysis D. NO - ResNet (a type of CNN) is for image recognition

Comment: A bit confusing. Depends on sensor data. if it is images then D for sure but if it is tabular data then B

Comment: We have a tabular multiclass classification problem. Letter B is the correct solution for this.

Comment: Definitely B.

Comment: It's B

Discussion for Question 250

Link: https://www.examtopics.com/discussions/amazon/view/111690-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- D: 9 votes

Discussion

Comment: d - Cloudwatch - no brainer for diagnostics logs

Comment: CloudWatch

Comment: A. NO - Amazon S3 is for SageMaker input B. NO - EBS is for data store C. NO - CloudTrail is for access log/security D. YES - Amazon CloudWatch will grab errors

Comment: Agree D for logging cloud watch

Comment: Letter D is correct. Service for Logging (Cloudwatch), Service for Auditing and Access (CloudTrail), and Service for Storage (S3 and EBS).

Comment: D is correct.

Comment: All logs are captured in CloudWatch by default those can be exported to S3 if needed. https://repost.aws/knowledge-center/sagemaker-studio-custom-container

Comment: Amazon CloudWatch is a monitoring and logging service provided by AWS. It collects and stores log files and metrics from various AWS services, including Amazon SageMaker. CloudWatch allows you to gain visibility into your applications and infrastructure by providing a unified view of logs, metrics, and events.

Discussion for Question 251

Link: https://www.examtopics.com/discussions/amazon/view/111704-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- D: 14 votes

Discussion

Comment: Reinforcement learning allows the bot to continuously learn from its own experiences, adapt to changing market conditions, and optimize its decision-making process over time. It is well-suited for dynamic and uncertain environments like financial markets, where the optimal trading strategies may vary depending on various factors and trends.

Replies:

Comment: I like your answer conrad

Replies:

Comment: liked*

Comment: At first glance this is a problem where we have some sort of goal and an agent(.i.e. bot)

Comment: A. NO - there are no labelled data as input B. NO - it is not a clustering problem C. NO - there are no labelled data as input D. YES - it can learn over time what are the good and bad decisions

---

Discussion for Question 252

Link: https://www.examtopics.com/discussions/amazon/view/111706-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 11 votes

Discussion

Comment: A. NO - should use Spot instances B. YES - makes use of compiler and spot instances C. NO - that is to minimize training time, not reduced cost D. NO - MaxWaitTimeInSeconds is time to find a free spot instance, it is unrelated to MaxRuntimeInSeconds which is processing time once an istance has been acquired

Comment: managed spot

Comment: B is the only option which mentions spot instances... the question is to be most cost effective, so seems B is only viable option.

Replies:

Comment: I go with you

Comment: B. Turn on SageMaker Training Compiler by adding compiler_config=TrainingCompilerConfig() as a parameter. Turn on managed spot training by setting the use_spot_instances parameter to True. Pass the script to the estimator in the call to the TensorFlow fit() method.

Discussion for Question 253

Link: https://www.examtopics.com/discussions/amazon/view/111710-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 8 votes

Discussion

Comment: A. NO - CloudWatch cannot get filter ranks of filters. Run a new training job with the pruned model. B. NO - Ground Truth can help improve model performance, not reduce inference cost C. YES - Filter pruning based on ranks is specific to CNN and supported by SageMaker Debugger D. NO - learning rate impacts training, not inference

Comment: Gain insight into the training metrics with debugger and prune

Comment: Letter B uses AWS GT incorrectly, so it is wrong. Letter D is outside the scope of what the question asks, so it is wrong. Letter A falls into the same problem as Letter B. Letter C is correct.

Comment: SageMaker Model Monitor is for model drift, not for utilization metrics. SageMaker Debugger is the best choose. So it's C.

Comment: C. Use Amazon SageMaker Debugger to gain visibility into the training weights, gradients, biases, and activation outputs. Compute the filter ranks based on the training information. Apply pruning to remove the low-ranking filters. Set the new weights based on the pruned set of filters. Run a new training job with the pruned model.

Discussion for Question 254

Link: https://www.examtopics.com/discussions/amazon/view/111711-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- BCE: 13 votes

Discussion

Comment: BCE B. Decrease the amount of regularization that the model uses: Regularization is used to prevent overfitting, but if the fitted model has poor accuracy on both the training and test datasets, reducing the amount of regularization can help the model better capture the underlying patterns and improve its accuracy. C. Increase the number of training examples that the model uses: Increasing the number of training examples allows the model to learn from a larger and more diverse dataset, which can help improve its ability to generalize and make accurate predictions. E. Increase the number of model features that the model uses: Adding more relevant features to the model can enhance its ability to capture important patterns and relationships in the data, leading to improved accuracy.

Comment: since model is underfitting, reduce the regularization will allow to use the features more, large no. of training example meaning more learning. and increase features will help model understand the establish pattens better

Comment: Please someone corrects me if I am wrong but I don't see that the question mentions overfitting or underfitting. It tells that both training and test datasets have poor accuracy. For this reason, I wouldn't apply B and I would go with the general steps that would help me to improve model accuracy (CDE)

Comment: B and C obvious. Between E and F , I would start by increasing features before considering reduction

Comment: I will go with BCE, other options are for solving overfitting

Comment: A. NO - regularization will reduce overfitting, not accuracy B. YES - to much regularization will reduce complexity and thus decrease accuracy C. YES - the more data the merrier D. NO - test examples will no influence model performance E. YES - the more features the more there is to learn F. NO - as per E

Comment: B C and E

Comment: We have an underfitting problem here. To remedy this, we must follow the alternatives that increase the complexity of the model: Letter B - C - E.

Comment: @RRST summarized well.

Comment: We can decrease the overfitting by reducing the number of features, so how is F not an answer

Comment: The problem is stating the Underfitting scenario. So correct answers are ACE

Replies:

Comment: A would be to solve an overfitting problem.

Comment: A. Increase the amount of regularization that the model uses. C. Increase the number of training examples that the model uses. E. Increase the number of model features that the model uses.

Replies:

Comment: corrected to BCE . decrease regularization for under fitting and increase for overfitting

---

Discussion for Question 255

Link: https://www.examtopics.com/discussions/amazon/view/112460-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 9 votes

Discussion

Comment: Answer B - B. Use a deep convolutional neural network (CNN) classifier with the images as input. Include a softmax output layer that outputs the probability that an image contains a car. This approach leverages the powerful feature extraction capabilities of CNNs and uses a softmax output layer, which is most suitable for binary classification tasks like detecting a car's presence.

Comment: According to chat GPT it is A. when we say "linear output layer" in the context of binary classification, it might lead to confusion. The output itself for binary classification problems is not linear; instead, it's the result of

applying a sigmoid function to a linear combination of features extracted by the neural network. The term "linear" might be more accurately replaced with "sigmoid activation" for the output neuron to reflect its role in producing a probability.

Replies:

> **Comment:** Dont quite agree and Chat GPT not always perfect - https://ml4.me/a-deep-dive-into-convolutional-neural-network-architectures-with-tensorflow-in-sagemaker/

Comment: A. NO - linear output is not B. YES - CNN is good, softmax output is good C. NO - MLP is not good D. NO - MLP is not good

Comment: Softmax is particularly useful when you want the network to make a clear choice among multiple classes. The question doesn't even ask for probabilities, it just asks for a binary classification. Ideally sigmoid activation function is used for binary classification, but when there is only 1 class, softmax will work the same way as sigmoid. The answer is still B, but just because linear layer is used for regression and not for binary classification. And of course, CNNs are better than MLP for image classifications.

Comment: It is a wording puzzle. A should be right due to car or no car but in the answer it mentions probability. Hence, softmax is more appropriate despite it is not multi classification

Comment: For image classification problems we must use CNN, so we discard Letters C - D. Letter A is wrong, because linear output layer does not generate probability, but softmax. Letter B is correct.

Comment: As it's a binary classification problem (car vs. no car) I would argue a linear output layer makes more sense than softmax...

Replies:

> **Comment:** A linear layer will not generate a probability, so it's wrong my conrad.

Comment: B is right

Comment: Both MLP and CNN can process images, but CNN is more accurate and can be used for more complex images

---

Discussion for Question 256

Link: https://www.examtopics.com/discussions/amazon/view/112462-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- CD: 13 votes
- DE: 13 votes

Discussion

Comment: A. NO - we can't change the model for transfer learning B. NO - we can't change the model for transfer learning C. YES - lst file is how we give input to SageMaker (https://medium.com/@texasdave2/itty-bitty-lst-file-format-converter-for-machine-learning-image-classification-on-aws-sagemaker-b3828c7ba9cc) D. YES - obvious E. NO - there is no extra metadata we want to provide (https://docs.aws.amazon.com/sagemaker/latest/dg/augmented-manifest.html)

Comment: Create a .lst File (Option C): Explanation: The .lst file is a standard format used with Amazon SageMaker's image classification algorithm, which lists image files and their corresponding labels. This file is crucial for SageMaker to read and map the images correctly for training purposes. The .lst file needs to be uploaded to Amazon S3. Initiate Transfer Learning (Option D): Explanation: Transfer learning allows you to leverage pre-trained weights from existing models (like ImageNetV2) and fine-tune them using your own data. In this case, training with the 10,000 new labeled images helps the model recognize less common animal species. Transfer learning is more efficient since the model has already been trained on similar data.

Replies:

> **Comment:** Link for more information on the .lst requirement for training and validation channels: https://docs.aws.amazon.com/sagemaker/latest/dg/image-classification.html#IC-inputoutput

Comment: A - no need for full training only transfer learning B - no built in inception model in sagemaker and no transfer learning C - no data augmentation could introduce inaccuracies D - yes! Transfer learning E - yes, augmentation improves accuracy

Comment: We need to augment the existing images so E makes sense

Comment: A False - no transfer learning B False - no transfer learning C True - "If you use the Image format for training, specify train, validation, train_lst, and validation_lst channels as values for the InputDataConfig parameter of the CreateTrainingJob request. Specify the individual image data (.jpg or .png files) for the train and validation channels. Specify one .lst file in each of the train_lst and validation_lst channels. Set the content type for all four channels to application/x-image." D True - it uses transfer learning E False - "To include metadata with your dataset in a training job, use an augmented manifest file. " Here we don't have any metadata

Comment: I would go with E. as we have a hint in the question, that we need to use a Pipe mode, and E is used for pipe mode

Comment: D is obvious. Reason for option E is because they want to train model in Pipe mode and using an augmented manifest file in JSON Lines format enables trraining in Pipe mode as per this link - https://docs.aws.amazon.com/sagemaker/latest/dg/image-classification.html Refer to the section - Train with Augmented Manifest Image Format in this link for more details. Using an augmented manifest file is an alternative to preprocessing when you have labeled data. For training jobs using labeled data, you typically need to preprocess the dataset to combine input data with metadata before training. If your training dataset is large, preprocessing can be time consuming and expensive.

Comment: D is obvious. Keyword here is pipe mode. "The augmented manifest format enables you to do training in Pipe mode using image files without needing to create RecordIO files." Hence, E.

Comment: https://docs.aws.amazon.com/sagemaker/latest/dg/augmented-manifest.html

Comment: Letter A - B deviate from what is asked in the scope of the question. Correct alternatives are E - D. To understand that C is wrong, look here: https://docs.aws.amazon.com/sagemaker/latest/dg/image-classification.html TL;DR - .lst file is only for classification task

Comment: Augmented manifest format enables you to do training in Pipe mode using files without needing to create RecordIO files (.rec)

Comment: D+E https://docs.aws.amazon.com/sagemaker/latest/dg/augmented-manifest.html

Comment: C. Create a .lst file that contains a list of image files and corresponding class labels. Upload the .lst file to Amazon S3. D. Initiate transfer learning. Train the model by using the images of less common species. Details provided in this blog post: https://aws.amazon.com/blogs/machine-learning/classify-your-own-images-using-amazon-sagemaker/

Comment: Correct Ans DE

Discussion for Question 257

Link: https://www.examtopics.com/discussions/amazon/view/112605-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 11 votes

Discussion

Comment: Answer is A - "SageMaker Feature Store consists of an online and an offline mode for managing features. The online store is used for low-latency real-time inference use cases. The offline store is primarily used for batch predictions and model training." https://aws.amazon.com/blogs/machine-learning/speed-ml-development-using-sagemaker-feature-store-and-apache-iceberg-offline-store-compaction/

Comment: Best Choice: B. Amazon SageMaker Feature Store This option offers a centralized and efficient solution that meets all the requirements: Feature Storage: SageMaker Feature Store acts as a single repository for features used in both offline training and online inference. Online Store: Creating an online store within Feature Store eliminates the need for a separate S3 bucket for inference features, simplifying management. Feature History Tracking: Feature Store automatically tracks feature lineage and versions, allowing you to see changes and roll back if needed. Data Science Team Access: IAM roles can be created to grant data scientists access to search and explore features within Feature Groups in the store.

Comment: A meets all requirements, and looks like the easiest to setup

Comment: A. YES - online store will be faster for inference, offline store cheaper for batch B. NO - online store for offline will be too expensive C. NO - want to use Feature store D. NO - want to use Feature store

Comment: Amazon SageMaker Feature Store is a managed service that makes it easy to store and manage features for machine learning models. It provides a scalable and reliable way to store features, and it supports both online inference and offline model training. Creating separate online and offline stores in SageMaker Feature Store will allow the music streaming company to optimize the storage and performance of their features for each use case. The online store can be configured to be highly available and performant, while the offline store can be configured to be cost-effective and scalable.

Discussion for Question 258

Link: https://www.examtopics.com/discussions/amazon/view/112719-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 8 votes

Discussion

Comment: C. Semantic Segmentation allows for pixel-wise classification of the video frames, meaning it can precisely identify and isolate a visitor's hair by labeling each pixel in the image as belonging to hair (or other categories).Object detection uses bounding boxes, which would not effectively isolate hair, especially in cases where hair might not have clear boundaries.

Comment: Copied from ChatGPT: "Semantic segmentation provides a pixel-level classification of the image, meaning it labels each pixel in the image with the class of the object it belongs to. However, it does not inherently detect whether the object is present in the image. Instead, it assumes that the objects of interest are already present and segments the image accordingly." Since the input is video stream, Not all the frames(images) contain hair! Therefore I would go for A.

Comment: semantic segmentation identifies hair, and Resnet for type and color.

Comment: I was sure that it was option C but but when we want to select the option requiring the least amount of effort, it must be A. Hair color is detected by ResNet-50, not by semantic algorithm. So, object detection algorithms are generally easier to implement and fine-tune compared to semantic segmentation algorithms. They can accurately locate and extract specific objects, such as hair, from the video frames, simplifying the subsequent analysis. Additionally, ResNet-50 is a widely used pre-trained model for image classification tasks, making it relatively straightforward to apply for determining hair style and hair color

Comment: We need Semantic Segmentation to identify the hair style and color by pixel level mapping

Replies:

Comment: I mean to select "C"

Comment: CHAT GPT4= C

Replies:

Comment: Claude 3 Sonnet = C

Comment: I doubt that object detection will detect hair better than a semantic segmentation

Comment: Will go with A (Object detection) as semantic segmentation requires labelling every pixel in a picture which is more effort compared to object detection

Comment: C is the right. Semantic Segmentation = Pixel Level category assingment Restnet50 = used for image recognition and computer vision tasks

Comment: C The backbone is a network that produces reliable activation maps of image features. The decoder is a network that constructs the segmentation mask from the encoded activation maps. Amazon SageMaker semantic segmentation provides a choice of pre-trained or randomly initialized ResNet50 or ResNet101 as options for backbones. The backbones come with pre-trained artifacts that were originally trained on the ImageNet classification task. These are reliable pre-trained artifacts that users can use to fine-tune their FCN or PSP backbones for segmentation. Alternatively, users can initialize these networks from scratch. Decoders are never pre-trained. Semantic Segmentation algorithm is now available in Amazon SageMaker | AWS Machine Learning Blog https://aws.amazon.com/cn/blogs/machine-learning/semantic-segmentation-algorithm-is-now-available-in-amazon-sagemaker/

Comment: Letters B - D are wrong as they ultimately use a tabular classification model for an image problem, so we discard it. As we want a solution with the least effort, it is known that object detection requires less training effort than semantic segmentation, in addition to being able to keep the visitor's hair in the frame. Therefore, the correct alternative is Letter A.

Comment: Segmentation is too heavy

Comment: Using ResNet-50, you can determine hair style and hair color by passing the identified hair region from the semantic segmentation algorithm as input. ResNet-50 can classify the hair region into one of the 1000 categories from the ImageNet database, such as curly, straight, blonde, brunette, etc. Option C is the best option for your problem because it allows you to efficiently and accurately identify and classify a visitor's hair in video frames using two powerful deep learning algorithms: semantic segmentation and ResNet-50.

Comment: Xgboost not for Image detection

Comment: ChatGPT say it's A.

Comment: A Least Effort

Discussion for Question 259

Link: https://www.examtopics.com/discussions/amazon/view/112463-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 5 votes

Discussion

Comment: SageMaker Clarify can do the work

Comment: The model is trained already, so C. I imagine using a debugger in production is nuts

Comment: The solution that will meet these requirements with the least development effort is C. Using SageMaker Clarify to generate the explanation report, and attaching the report to the predicted results . This solution allows the financial services company to use SageMaker Clarify, a feature that provides machine learning (ML) model transparency and explainability, to generate the explanation report for each loan approval prediction. SageMaker Clarify can provide feature importance scores, which indicate how much each feature contributes to the prediction, and SHAP values, which measure how each feature affects the prediction compared to the average prediction . The company can use these metrics to generate and attach the explanation report that contains the reason for why the customer was approved or denied for a loan.

Comment: C, SageMaker Clarify can give explanation Why.

Comment: Sagemaker Clarity is better option

Discussion for Question 260

Link: https://www.examtopics.com/discussions/amazon/view/114475-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- D: 9 votes

Discussion

Comment: D makes more sence to me. Collaborative filtering takes into account other users preferences which is what we want to avoid because we do not want irrelevant promotions

Comment: C. Use the Neural Collaborative Filtering algorithm with a SageMaker batch inference job This solution uses the Neural Collaborative Filtering algorithm to leverage the latest techniques in recommendation systems, while

SageMaker's batch inference jobs provide efficient and cost-effective processing of recommendations in bulk. This aligns well with the company's weekly email campaigns and minimizes operational overhead.

Comment: I will go with D, as it is more operationally eficient

Comment: A. NO - Factorization Machines is classification B. NO - an endpoint needed be invoked C. YES - Collaborative Filtering is good for recommendations based on past activities, and a batch job will generate the fiel we want D. NO - Factorization Machines is classification

Comment: My choice D

Comment: factorization machine algorithm is used for regression or classification. Generating recommendation is neither. Use neural collaborative filtering and do batch inference to identify email addresses.

Comment: From Chat GPT The solution that will meet the requirements with the MOST operational efficiency is option C: Use the Neural Collaborative Filtering algorithm to build a model that can generate personalized offer recommendations for customers. Deploy a SageMaker batch inference job to generate offer recommendations. Feed the offer recommendations into the bulk email marketing system. By using the Neural Collaborative Filtering algorithm, the ML team can build a model that can provide personalized offer recommendations based on customer profiles and past accepted offers. Deploying a SageMaker batch inference job allows for efficient processing of a large batch of customer data to generate offer recommendations. These recommendations can then be fed directly into the bulk email marketing system, streamlining the process and improving operational efficiency.

Comment: : Use the Factorization Machines algorithm to build a model that can generate personalized offer recommendations for customers. Deploy a SageMaker batch inference job to generate offer recommendations. Feed the offer recommendations into the bulk email marketing system

Replies:

> **Comment:** , option D is more operationally efficient.

Comment: C batch predictions and collaborative

Comment: C is a better option for efficiency.

Discussion for Question 261

Link: https://www.examtopics.com/discussions/amazon/view/113104-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- D: 8 votes

Discussion

Comment: Amazon SageMaker's built-in image classification algorithm supports input data in RecordIO format for training. RecordIO is a binary file format that efficiently stores images and labels in a compact format, making it suitable for training deep learning models with large datasets. The im2rec utility tool provided by Apache MXNet can be used to generate RecordIO files from the manifest files (training.lst and validation.lst) containing image paths and labels. Using RecordIO files allows for efficient streaming of data during training, especially when combined with SageMaker's pipe mode, which can speed up the training process by reducing disk I/O.

Comment: A. NO - SageMaker requires RecordIO input B. NO - SageMaker requires RecordIO input C. NO - SageMaker requires RecordIO input D. YES - SageMaker requires RecordIO input

Comment: If they want to use the RecordIO content type for training in pipe mode, they should generate two RecordIO files, training.rec and validation.rec, from the manifest files by using the im2rec Apache MXNet utility tool1. They should upload the RecordIO files to the training S3 bucket. This corresponds to option D in the question.

Comment: https://aws.amazon.com/blogs/machine-learning/classify-your-own-images-using-amazon-sagemaker/

Comment: It's D. https://sagemaker-examples.readthedocs.io/en/latest/introduction_to_amazon_algorithms/imageclassification_caltech/Image-classification-fulltraining.html

Comment: should be D.he company wants to use the Amazon SageMaker image classification algorithm to train the model. SageMaker's image classification algorithm requires the data to be in the RecordIO format. Therefore, the company should use the im2rec utility tool, which is part of the Apache MXNet framework, to generate RecordIO files from the manifest files. The company should generate two RecordIO files, one for the training dataset (training.rec) and one for the validation dataset (validation.rec). These RecordIO files will contain the image data along with their corresponding labels.

Discussion for Question 262

Link: https://www.examtopics.com/discussions/amazon/view/114477-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 8 votes

Discussion

Comment: Its B. CloudTrail records API calls for AWS USERS in AWS accounts. It has nothing to do with some random users who submit pictures via an app. CloudTrail is NOT the answer.

Comment: Rekognition and cloud trail

Comment: A. NO - AWS Panorama is for edge computing B. NO - AWS Panorama is for edge computing C. YES - best practice D. NO - IP address cannot be captured from the image, but from internet traffic

Comment: Agree with C

Comment: C. Rekognition text detection is for text inside image and not for source ip.

Comment: C or D but CloudTrail detect IP only of API calls and not record API call for Invoke_Model.

Discussion for Question 263

Link: https://www.examtopics.com/discussions/amazon/view/114030-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- AC: 9 votes

Discussion

Comment: LDA and NTM are the only applicable options for topic modelling here

Comment: A. LDA is designed to discover abstract topics in a collection of documents. It is commonly used for topic modeling and is one of the most popular techniques for extracting topics from text data. C. NTM is also used in topic modeling. It uses deep learning to discover topics in a collection of documents, and it can produce similar results to LDA but potentially with better accuracy due to its neural network foundation. Incorrect choices: B. Random forest classifier is a classification algorithm. It is better suited for classification tasks based on labeled data. D. SVM is also a classification algorithm. It works well for binary classification problems. E. Linear regression is a regression algorithm used to predict continuous values. It's not suitable for topic modeling.

Comment: A. YES B. NO - for classification C. YES D. NO - for classification E. NO - for classification

Comment: both unsupervised learning algorithms that can discover abstract topics in a collection of text documents . These algorithms can help the data scientist to analyze the audit documents and provide a list of the top words for each category to help the auditors assess the relevance of the topic. LDA and NTM are different from other algorithms that are not suitable for this scenario, such as:

Comment: AC for topics

Comment: Although you can use both the Amazon SageMaker NTM and LDA algorithms for topic modeling, they are distinct algorithms and can be expected to produce different results on the same input data. A and C https://docs.aws.amazon.com/sagemaker/latest/dg/ntm.html https://docs.aws.amazon.com/sagemaker/latest/dg/lda.html

Discussion for Question 264

Link: https://www.examtopics.com/discussions/amazon/view/114427-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 9 votes

Discussion

Comment: A. Amazon Kendra is designed to search through various types of documents and provide relevant answers. B. Training a BiDAF network requires expertise in deep learning and natural language processing. It would require substantial effort in data preparation, model training, and integration. C. Amazon SageMaker Blazing Text is primarily used for text classification and word embeddings, not for extracting answers from company documents based on user queries. D. Amazon OpenSearch Service is a search and analytics engine, but it's not tailored for extracting precise answers from documents. The k-NN Query API is used for similarity searches and isn't inherently designed to answer questions based on document content.

Comment: Amazon Kendra is a managed search service that helps you find answers to your questions from your content. It uses natural language processing and machine learning to understand the meaning of your questions and match them to the most relevant content.

Comment: A is the correct. All the others are really hard.

Comment: For sure A.

Comment: A is correct Amazon Kendra is an intelligent search service powered by machine learning. It can be used to index and search through company documents, making it a suitable solution for the chatbot to base its answers on. Option A suggests indexing company documents using Amazon Kendra, which simplifies the process of searching and retrieving relevant information from the documentation. Integrating the chatbot with Amazon Kendra using the Kendra Query API operation allows the chatbot to send customer questions to Kendra and receive relevant answers based on the indexed documents. This solution requires minimal development effort as it leverages the built-in capabilities of Amazon Kendra and its integration with the chatbot. Option B, training a Bidirectional Attention Flow (BiDAF) network, and option C, training a SageMaker Blazing Text model, both involve training custom models, which would require significant development effort, including data preparation, model training, and deployment.

Discussion for Question 265

Link: https://www.examtopics.com/discussions/amazon/view/111820-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 8 votes

Discussion

Comment: https://aws.amazon.com/blogs/machine-learning/identify-rooftop-solar-panels-from-satellite-imagery-using-amazon-rekognition-custom-labels/

Comment: Least effort + no ML experience, so A

Comment: A. YES - Amazon Rekognition Custom Labels is better than other option like Face/Celebrity/etc.; Ground Truth Active Learning will require human labelling only when needed, works well with small internal team B. NO - missing Active Learning C. NO - SageMaker Object Detection is more complicated than labelling D. NO

Comment: Vote for A

Comment: SageMaker Ground Truth can use active learning to automate the labeling of the input data for certain built-in task types, such as object detection. Active learning is a machine learning technique that identifies data that should be labeled by human workers. This helps to reduce the cost and time that it takes to label the dataset compared to using only humans1. By setting up a private workforce, the internal team can use their own domain knowledge to label the data and ensure quality and consistency.

Comment: As we have an internal team working on this project, it is understood that they will do the labeling. Letter A is correct, as SageMaker Active Learning Feature allows you to streamline the team's efforts. Letters C - D are wrong as they use the wrong algorithm (object detection) and Letter B takes longer than Letter A.

Comment: Option D uses a public workforce to label the data. This means that the company can leverage a large pool of workers from Amazon Mechanical Turk, who are experienced and qualified in labeling tasks. The public workforce can provide more diverse and accurate labels than the internal team, who may have limited or biased perspectives. The public workforce can also complete the labeling task faster and more efficiently than the internal team, who may have other priorities or responsibilities.

Replies:

> **Comment:** changed to A. public workforce may need effort.

Comment: https://aws.amazon.com/blogs/machine-learning/identify-rooftop-solar-panels-from-satellite-imagery-using-amazon-rekognition-custom-labels/

Comment: https://docs.aws.amazon.com/sagemaker/latest/dg/sms-automated-labeling.html

Comment: It's A due to small team on the project and minimal effort from the team required. SageMaker Ground Truth active learning feature can speed up the labeling process for 8000 images.

Comment: B is correct

Comment: B. Set up a private workforce that consists of the internal team. Use the private workforce to label the data. Use Amazon Rekognition Custom Labels for model training and hosting. By setting up a private workforce consisting of the internal team and using Amazon Rekognition Custom Labels, the company can leverage the labeling capabilities of the internal team to label the data. Amazon Rekognition Custom Labels can then be used for model training and hosting. This option eliminates the need for additional complex steps such as active learning or object detection algorithm training, which may require more ML expertise and effort from the internal team. Instead, it relies on the simplicity and convenience of using Amazon Rekognition Custom Labels for model training and hosting, making it the least effort-intensive option for the team with no ML expertise or experience.

Comment: C is correct.

Discussion for Question 266

Link: https://www.examtopics.com/discussions/amazon/view/114478-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 7 votes

Discussion

Comment: https://docs.aws.amazon.com/databrew/latest/dg/personal-information-protection.html

Comment: option C is better than the other options because it can meet the company's requirements with the least development effort. Option C can leverage DataBrew's native capabilities to identify and handle PII data in a visual and intuitive way.

Comment: Answer C https://aws.amazon.com/blogs/big-data/introducing-pii-data-identification-and-handling-using-aws-glue-databrew/

Comment: C cause A require customization.

Discussion for Question 267

Link: https://www.examtopics.com/discussions/amazon/view/114479-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 12 votes
- D: 10 votes

Discussion

Comment: obvious. no need retrieval time to be short.

Comment: You cannot achieved durability in standard IA

Comment: Lowest cost - past 90 days in deep archive. Access only 4 times an year - standard IA - we need to be able to recover data in case of a zone failure

Comment: the question only mentions "the same durability" and "minimize cost". It didn't mention "availability" So D

Comment: If the company is required to "must" keep the files after 90 days then what happens if One Zone is hit by a disaster?! Therefore option "C" is the best solution here.

Comment: If the company can tolerate retrieval times of several hours and is looking for the lowest cost solution, Option C is the best choice over D, as S3 Glacier Deep Archive offers the lowest storage cost.

Comment: S3 Standard-IA : Designed for 99.9% availability over a given year S3 One Zone-IA : Designed for 99.5% availability over a given year https://aws.amazon.com/s3/storage-classes/?nc1=h_ls

Comment: Answer C Not A - S3 Glacier Flexible Retrieval is not needed. it cost more than Glacier Deep Archive. And the question doesnt mention to retrieve the file immediately when neede. Not B - Although One Zone-Infrequent Access is possible, the question doesnt mention thre is another copy of files somewhere else or there is no problem if some files are lost. Although the loss of data is very rare we use One Zone-Infrequent Access for data that can tolerate loss.

Comment: "Amazon S3 Standard, S3 Standard-IA, S3 One Zone-IA, and Amazon Glacier, are all designed for 99.999999999% durability." One Zone is as durable as Standard but cheaper than Standard IA. Glacier Deep Archive is cheaper than Flexible Retrieval. D is the cheapest (without compromising on durability).

Comment: D is the right option because S3 One Zone-Infrequent Access has the same durability (11 nines) as S3 Standard-IA with an added benefit of 20% less cost than S3 Standard-IA. Only downside is that One-Zone is less resilient than Standard-IA as data is not replicated to multiple AZs. This question is asking about durability not resiliency. Also, option D has the cheapest archival option (Glacier Deep Archive) which comes with 11 nines durability so, daily data files and archived files will have the same durability.

Comment: A. NO - can store in cheapest Glacier Deep Archive after 90 days B. NO - can store in cheapest Glacier Deep Archive after 90 days C. YES D. NO - need multiple zones for higher resiliency due to legal requirements

Comment: "Amazon S3 Standard, S3 Standard-IA, S3 One Zone-IA, and Amazon Glacier, are all designed for 99.999999999% durability." One Zone is as durable as Standard but cheaper than Standard IA. Glacier Deep Archive is cheaper than Flexible Retrieval. D is the cheapest (without compromising on durability).

Comment: Go for A

Comment: If we want to maintain the same durability, we cannot use One Zone, given the importance of maintaining these data for the quality of the model (thus, we discard B - D). As we want to minimize the cost and we need to access the data every 90 days, the correct alternative is Letter A as the minimum recovery time is 90 days for S3 Glacier Flexible Retrieval (in contrast, Deep Archive needs 180 days to be accessed).

Replies:

> **Comment:** Where the hell did you source your info? 180 days to be accessed?!!: - S3 Glacier Instant Retrieval for archiving data that might be needed once per quarter and needs to be restored quickly (milliseconds) - S3 Glacier Flexible Retrieval for archiving data that might infrequently need to be restored, once or twice per year, within a few hours - S3 Glacier Deep Archive for archiving long-term backup cycle data that might infrequently need to be restored within 12 hours https://docs.aws.amazon.com/prescriptive-guidance/latest/backup-recovery/amazon-s3-glacier.html

Comment: Indeed option C

Comment: C it is

Comment: For same storage Durability we should NOT use S3 One Zone-IA. For Glacier we will use Deep Archive for more cost saving. So it's C

Discussion for Question 268

Link: https://www.examtopics.com/discussions/amazon/view/114480-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- D: 6 votes
- B: 6 votes

Discussion

Comment: you can input metadata directly in sagemaker feature store, no need for dydb. https://docs.amazonaws.cn/en_us/sagemaker/latest/dg/feature-store-add-metadata.html

Comment: Option D is similar to option B but suggests using Amazon QuickSight for metadata analysis instead of SageMaker Studio. While QuickSight is a viable option for visualization, it may require additional configuration and setup compared to using SageMaker Studio, which is already integrated with SageMaker Feature Store.

Comment: On second thoughts, D (QuickSight) is also a possible option because feature store parquet files could be queried by Athena and Athena could be used with Quicksight without any development efforts. Would go with option D

Comment: B is the correct option because in option D, Quicksight is used which doesn't support parquet files. Sagemaker feature groups created in offline feature store use parquet to store feature values on S3. This question is talking about auditing which makes offline feature store an obvious choice. In order to use Quicksight, there is an additional step to convert feature store parquet file to a supported format (like CSV, JSON, etc.) and hence, it has more efforts compared to creating a dataframe in Data Wrangler and using it for visualizations

Comment: The answer should be D, both the sagemaker studio and the quicksight can analyze metadata but Amazon SageMaker Studio is a web-based, integrated development environment (IDE) for machine learning that provides all the tools you need to take your models from data preparation to experimentation to production. and the question is asking about a solution with the least development so it should be Quicksight

Comment: A. NO - Amazon SageMaker Feature Store cannot transform B. NO - SageMaker Studio require Python to analyze the metadata C. NO - custom algorithms are dev-intensive D. YES - use built-in functionnalities

Comment: Agree D

Comment: Por menor esforço de desenvolvimento descartamos Letra A (levantar um serviço gerenciado como DynamoDB normalmente não é a melhor solução), Letra B (não é performática), Letra C (mesmo motivo da A). Logo por eliminação, Letra D está correta.

Comment: This solution meets the requirements with the least development effort because it uses Amazon SageMaker Feature Store, which is a fully managed service that makes it easy to store and manage feature metadata. Amazon SageMaker Feature Store also provides built-in functionality for analyzing feature metadata, so there is no need to create custom algorithms or data flows.

Comment: https://aws.amazon.com/blogs/machine-learning/controlling-and-auditing-data-exploration-activities-with-amazon-sagemaker-studio-and-aws-lake-formation/ Studio supports Audit

Comment: I think it's B

Replies:

> **Comment:** Maybe It's D as QuickSight less development effort.
>
> #### Replies:
>
> > **Comment:** Agreed, should be D

Discussion for Question 269

Link: https://www.examtopics.com/discussions/amazon/view/114481-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 6 votes

Discussion

Comment: https://stackoverflow.com/questions/66692579/aws-sagemaker-permissionerror-access-denied-reading-data-from-s3-bucket 'C' is the correct answer.

Comment: Note that the question does not mention VPC, so we discard Letter B. Letter C appears to be correct, as they use IAM to grant permission (congruent with their role). SG is beyond the scope of KMS, so we discard Letter A. Letter D is incongruous.

Comment: Option C allows the ML specialist to assign an IAM role that provides S3 read access for the dataset to the SageMaker notebook. IAM is a service that helps users manage access to AWS resources. An IAM role is an entity that defines a set of permissions for making AWS service requests. The ML specialist can create an IAM role that has a policy that allows the notebook to read the dataset from the S3 bucket. The ML specialist can then attach the IAM role to the notebook when creating or updating it.

Comment: C it is

Comment: C is correct

Comment: 100% it's D.

Replies:

Comment: stop misleading people

Comment: It's not

Discussion for Question 270

Link: https://www.examtopics.com/discussions/amazon/view/114482-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: Easiest option would be C

Comment: A. NO - Kinesis Data Analytics can use only Firehose or Amazon Kinesis Data Streams as input (https://docs.aws.amazon.com/kinesisanalytics/latest/dev/how-it-works.html) B. NO - no need to save in S3, can do on-the-fly C. YES D. NO - AWS Lambda would be invoked on a per-record basis

Comment: why not B?

Comment: Option C also allows the ML specialist to use Amazon Kinesis Data Analytics to transform the most recent 10 minutes of data before inference. Kinesis Data Analytics is a fully managed service that enables users to analyze streaming data using SQL or Apache Flink. Kinesis Data Analytics can process streaming data in real time and generate insights, metrics, and alerts. Kinesis Data Analytics can also integrate with other AWS services, such as Lambda, S3, or SageMaker. The ML specialist can use Kinesis Data Analytics to apply SQL queries or Flink applications to transform the event data based on the 10-minute running window and prepare it for inference.

Comment: It's real-time and less operational overhead

Comment: I think it's C as we are using only 2 services and it's less operational effort.

Discussion for Question 271

Link: https://www.examtopics.com/discussions/amazon/view/114483-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 10 votes

Discussion

Comment: Option A allows the ML specialist to add class weights to the MLP's loss function, and then retrain. Class weights are a way of assigning different importance or penalties to different classes in a classification problem. Class weights can help balance the data distribution and reduce the bias towards the majority classes. Class weights can also help improve the recall metric, which is the ratio of true positives to the sum of true positives and false negatives. Recall measures how well the model can identify the relevant instances of a class, especially when the class is rare or unique. The ML specialist can use class weights to increase the importance or penalty of the target class of interest, and then retrain the MLP to improve its recall.

Comment: Option A - Add class weights to MLP's loss function - improve recall with the least amount of time and effort by making the model more sensitive to the underrepresented target class during training.

Comment: Apologies for the confusion but on second thoughts, A is the right answer as unique doesn't mean unknown and this is still a supervised learning problem. Adding weights to classes would even out the bias caused by unique class and improve recall as mentioned by other experts in this forum. Please ignore my previous comment. A is the correct option indeed.

Comment: Leaning towards C as the target class of interest is unique as compared to dataset (as given in this question). If the target class is unique / non-existing in data set then we are talking about unsupervised learning and k-means is a right fit so, option C seems to be more appropriate than option A. Adding weights may still not be able to solve the purpose as target class is not present in data set. It is almost an unlabeled data set if target class is unknown / unique as compared to existing classes in data set. Unlabeled data sets are better solved using unsupervised learning.

Comment: Agreed A

Comment: A as this is Faster solution.

Discussion for Question 272

Link: https://www.examtopics.com/discussions/amazon/view/114426-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: The key here is before training the model. If it is before training then we can do that using SageMaker Data Wrangler. After training and model is deployed for inference, we can use model monitor

Comment: A and C are correct. https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-data-bias-reports-ui.html

Comment: A. YES - SageMaker Clarify can be used to detect bias during prep (https://aws.amazon.com/sagemaker/clarify/) B. NO - SageMaker Ground Truth is not used in the solution C. NO - drift report to during inference D. YES - Configure SageMaker Data Wrangler to generate a bias report. E. NO - SageMaker Experiments is to compare model outputs

Comment: This combination meets all the requirements with the least operational overhead. You can use SageMaker Data Wrangler to ingest and clean your data in Amazon SageMaker Studio without writing any code. You can also use SageMaker Clarify to automatically detect potential bias in your data using predefined or custom metrics. You can then configure SageMaker Data Wrangler to generate a bias report that shows the results of the bias analysis in a visual and interactive way2.

Replies:

Comment: I think the combination of actions that will meet the requirements with the least operational overhead are A and D. Use SageMaker Clarify to automatically detect data bias and configure SageMaker Data

Wrangler to generate a bias report.

Comment: A: AWS Clarify used to generate Bias report. D: AWS Data Wrangler to generate Bias report

Comment: AD: https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-data-bias-reports-ui.html

Replies:

Comment: SageMaker Clarify is integrated with Amazon SageMaker Data Wrangler, which can help you identify bias during data preparation without having to write your own code. Data Wrangler provides an end-to-end solution to import, prepare, transform, featurize, and analyze data with Amazon SageMaker Studio.

Comment: I think it's A+E

Comment: AC is correct. A: AWS Clarify used to generate Bias report. C: AWS Data Wrangler to generate Bias report and is operationally efficient. https://catalog.us-east-1.prod.workshops.aws/workshops/1e224d5a-4273-444a-acec-28d44a5bfb28/en-US/data-preparation/amazon-sagemaker/data-wrangler

---

Discussion for Question 273

Link: https://www.examtopics.com/discussions/amazon/view/114425-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 8 votes

Discussion

Comment: Not D because Firehose is not best to aggregate, and lambda is not necessary.

Comment: A. NO - AWS Lambda is not per to aggregate, it works on a per-row basis B. NO - Kinesis Data Firehose is not best to aggregate C. YES - Amazon Kinesis Data Analytics can aggregate and stream to Firehose D. NO - Kinesis Data Firehose is not best to aggregate

Comment: Kinesis Data Firehose can invoke your Lambda function to transform incoming source data and deliver the transformed data to destinations. So the answer cannot be C, has to be D

Comment: C it is

Comment: changing back to C. very confusing

Comment: Changing to option D Option D is the best option because it allows the network security vendor to use Amazon Kinesis Data Firehose to read and aggregate the data hourly from Amazon Kinesis Data Streams, and use AWS Lambda to transform the data and store it in Amazon S3. This way, the network security vendor can leverage the benefits of both services: Amazon Kinesis Data Firehose can provide a simple and scalable way to ingest, buffer, compress, and batch the streaming data; AWS Lambda can provide a flexible and cost-effective way to perform custom logic on the data, such as selecting only 7 to 12 fields for Athena queries. This option meets the requirements with the least amount of customization to transform and store the ingested data

Comment: Letter B is wrong, as it brings the addition of yet another new service (EMR). Letter D is wrong, as we cannot directly use KDF to perform transformations (it's a load service only). Letter C is the most correct and fastest, as it uses the Kinesis family. Letter A is functional, as we can call Lambda via KDS, but it would involve more customization given the Lambda code to be built.

Comment: Changing my answer to C

Comment: This option meets all the requirements with the least amount of customization. You can use Amazon Kinesis Data Firehose to ingest streaming data from thousands of endpoints and configure it to buffer the data by size or time interval (such as 1 hour). You can use AWS Lambda to transform the data and select only the relevant fields before delivering it to Amazon S3. You can also configure Amazon Kinesis Data Firehose to convert the data to a columnar format such as Apache Parquet or Apache ORC, which are optimized for querying with Amazon Athena3.

Replies:

Comment: You can not transform data using KDF.

Comment: C it is

Comment: C is correct!

Comment: Vote for C.

Comment: C is correct.

Discussion for Question 274

Link: https://www.examtopics.com/discussions/amazon/view/128599-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: Ansewer is A: Given the scenario, One-Hot Encoding would be the most effective way to encode the categorical feature into a numerical feature

Comment: Similarity encoding is meant to encode high cardinality features having misspelled values by group them closer. In high cardinality, it's more efficient o create vectors of real numbers than creating insane number of one hot encoded columns

Comment: A - most effective. dataset seems small and manually fixing of spelling is possible since this is categorical

Comment: C- Data Wrangler similarity encoding on the column to create embeddings of vectors of real numbers.

Comment: The similarity encoder creates embeddings for columns with categorical data. An embedding is a mapping of discrete objects, such as words, to vectors of real numbers. It encodes similar strings to vectors containing similar values. For example, it creates very similar encodings for "California" and "Calfornia". https://docs.aws.amazon.com/sagemaker/latest/dg/data-wrangler-transform.html

Comment: a character-level Recurrent Neural Network (char-RNN) can be used to fix spelling mistakes in a column containing medication names.

Comment: Use similarity encoding when you have the following: 1. A large number of categorical variables 2. Noisy data

---

Discussion for Question 275

Link: https://www.examtopics.com/discussions/amazon/view/128589-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- AB: 5 votes

Discussion

Comment: A: By creating an IAM role in the development account that the integration and production accounts can assume, you establish a trust relationship between the accounts. You can attach IAM policies to the role that grant the necessary permissions to access the feature repository and S3 buckets. B: AWS Resource Access Manager (AWS RAM) enables resource sharing across AWS accounts. By sharing the feature repository associated with the S3 buckets using AWS RAM, you allow the integration and production accounts to access and reuse the features.

Comment: A - cross account access C - use STS to get credentials and assume IAM role from different account (https://docs.aws.amazon.com/STS/latest/APIReference/API_AssumeRole.html) B is incorrect - RAM can be used

only in AWS Organizations (which is not mentioned in question), also only Amazon S3 on Outposts can be shared using RAM

Comment: A & B Create an IAM role in the dev account and share the features repository using AWS RAM

Discussion for Question 276

Link: https://www.examtopics.com/discussions/amazon/view/128591-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: It's PCA. It's not MCA because all the values are numeric and not categorical.

Comment: need to reduce the dimension of features in order to enhance accuracy on train and test data since # of features are huge.

Comment: B. Use a principal component analysis (PCA) model. This is because PCA can help to reduce the number of variables while preserving the most important information, which can help to improve the accuracy of the model and reduce the processing time.

Discussion for Question 277

Link: https://www.examtopics.com/discussions/amazon/view/128608-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 8 votes

Discussion

Comment: Both L1 & L2 help with overfitting. However, L1 regularization does feature selection by reducing weights of rerelenat featurtes to zero - reducing dimensionality and removing noisy features as in this case. L2 on the other hand keeps all features including noisy ones.

Comment: If you suspect that some features are irrelevant, Option A (L1 Regularization) could be more effective as it can shrink some coefficients to zero, effectively performing feature selection be removing the noise. If you believe that most features are relevant but the model is too complex, Option D (L2 Regularization) is typically the better choice as it evenly shrinks all coefficients, thus reducing model complexity without eliminating features. In this case, option A would be ideal to get rid of the irrelevant noise.

Comment: It should be D as this is a overfitting problem. A might make the model oversimple in that case train acc will be bad. L2 is better than L1

Comment: A. L1 Regularization reduces the amount of noise in the model, https://docs.aws.amazon.com/machine-learning/latest/dg/training-parameters1.html

Comment: D L2 regularisation for overfitting and noise

Discussion for Question 278

Link: https://www.examtopics.com/discussions/amazon/view/128794-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- AC: 8 votes

Discussion

Comment: C. Use data augmentation to rotate and translate the labeled images. Data augmentation involves creating new training data by applying transformations such as rotation, translation, scaling, etc. This helps to increase the diversity of the training data and makes the model more robust without requiring additional labeled data. A. Use Amazon SageMaker Ground Truth to label the unlabeled images. Leveraging Amazon SageMaker Ground Truth can help in labeling the unlabeled images to expand the training dataset and reduce overfitting. Adding more labeled data can improve the generalization of the model and reduce overfitting.

Comment: A. Use Amazon SageMaker Ground Truth to label the unlabeled images C. Helps address the over-fitting problem

Comment: Ground Truth to label the unlabeled images and data augmentation to create multiple variations of the labeled images

Comment: Ground Truth to label the unlabelled images

Discussion for Question 279

Link: https://www.examtopics.com/discussions/amazon/view/128943-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 7 votes

Discussion

Comment: This solution offers the following advantages: Amazon SageMaker Data Wrangler provides a user-friendly interface to explore and experiment with feature transformations, making it efficient for the data science team to try out different options. SageMaker Data Wrangler templates for visualization can quickly generate visualizations for the resulting distribution of the dataset, streamlining the visualization process. Exporting the feature processing workflow to a SageMaker pipeline for automation automates the feature transformations efficiently within the SageMaker environment.

Comment: Amazon SageMaker Data Wrangler: provides preconfigured transformations that allow for easy exploration of feature transformations. This simplifies the experimentation process. SageMaker Data Wrangler templates for visualization: allows for visualizing the resulting distribution of the dataset, aiding in understanding the effects of feature transformations. Export the feature processing workflow to a SageMaker pipeline for automation: Once an appropriate set of feature transformations is identified, the workflow can be exported to a SageMaker pipeline for automation. This ensures reproducibility and scalability of the feature processing steps.

Comment: Data Wrangler is an amazing tool that take EDA to next level

---

Discussion for Question 280

Link: https://www.examtopics.com/discussions/amazon/view/128947-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 5 votes

Discussion

Comment: bastion host / is an outdated method

Comment: Since the connection is over IPSec VPN and internet access is prohibited, NAT gateway and Bastian hosts are unnecessary eliminating B, C, D. Also, traffic should not leave AWS network between services so

sagemaker notebook VPC endpoint is needed

Comment: A most effective solution

Comment: A - Never choose bastion host. Other answers don't make sense.

Comment: A has the least development cost comparing with B

Discussion for Question 281

Link: https://www.examtopics.com/discussions/amazon/view/128795-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 6 votes

Discussion

Comment: This method is the most efficient and scalable for processing a dataset of this size, significantly outperforming the other options in terms of processing time.

Comment: since there are million post, 15 minutes may not be enough so step function is needed and batchDetectSentiment is good way to go

Comment: https://docs.aws.amazon.com/comprehend/latest/APIReference/API_BatchDetectSentiment.html#API_BatchDetectSentiment_RequestParameters

Comment: It's B. Limit on BatchDetectSentiment is 25 documents. Other endpoints are for individual strings.

Comment: B. Use a combination of AWS Step Functions and an AWS Lambda function to call the BatchDetectSentiment API operation with batches of up to 25 posts at a time. Batch processing is generally more efficient for large datasets. The BatchDetectSentiment API operation allows you to process multiple items (up to 25) in a single call, which helps in reducing the overall processing time. Additionally, using AWS Step Functions to manage the workflow and AWS Lambda to handle the batch processing can make the implementation scalable and easier to manage.

Discussion for Question 282

Link: https://www.examtopics.com/discussions/amazon/view/128796-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- D: 12 votes
- B: 11 votes

Discussion

Comment: Could be B or D. The question calls out that 10% of the data is missing, which a lot. Smoothing would help as well. I'll go with D.

Comment: While both B & D will have effect on performace but MOST effect will be from B - smoothening of seasonal variations for forecasting. Linera interpolation may even have adverse effect on performance if relationship between variables in not linear.

Comment: smoothing is more important than missing data in this scenario

Replies:

Comment: After must consideration, I will change my answer to "D" First and foremost, solving missing data is more important. The question clearly states that 10% data are missing.

Comment: I'm going with B. 10% missing data on 10 years of data shouldn't matter too much, so D falls off. Seasonality introduces issues and should be fixed. A and C are wrong for obvious reasons.

Comment: Linear interpolation would help to handle 10% missing values

Comment: B. 10% of the days data missing out of 365*10 days

Comment: D. Based on the problem, we need to address missing data and not seasonal variance.

Comment: Answer: D

Comment: B. Apply smoothing to correct for seasonal variation. Smoothing techniques, such as using moving averages or other time series smoothing methods, can help in reducing noise and capturing the underlying patterns in the sales data. Seasonal variation is a common issue in time series data, especially in retail where sales may exhibit regular patterns based on seasons, holidays, or other recurring events.

Discussion for Question 283

Link: https://www.examtopics.com/discussions/amazon/view/128797-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 9 votes

Discussion

Comment: https://docs.aws.amazon.com/sagemaker/latest/dg/inference-recommender-recommendation-jobs.html

Comment: GPT+Claude 3: The Default job type (option A) involves SageMaker running a set of load tests on the recommended instance types, which can provide a quicker result as it takes less time to complete (within 45 minutes). On the other hand, the Advanced job type (option B) involves a custom load test where you have more control over the traffic pattern and requirements for latency and throughput. However, this option may take longer to complete (an average of 2 hours). Given the requirement for the least development effort, option A seems more suitable. It utilizes the Default job type, which is more automated and requires less manual configuration compared to the Advanced job type. Additionally, the shorter completion time aligns better with the goal of minimizing development effort.

Comment: Inference recommendations (Default job type) run a set of load tests on the recommended instance types. You can also load test for a serverless endpoint.. You only need to provide a model package Amazon Resource Name (ARN) to launch this type of recommendation job. Inference recommendation jobs complete within 45 minutes. Endpoint recommendations (Advanced job type) are based on a custom load test where you select your desired ML instances or a serverless endpoint, provide a custom traffic pattern, and provide requirements for latency and throughput based on your production requirements. This job takes an average of 2 hours to complete depending on the job duration set and the total number of inference configurations tested.

Comment: B. Traffic patterns are known.

Comment: It's either A or B. Advanced Job Type recommendations re based on a custom load test where you select your desired ML instances or a serverless endpoint, provide a custom traffic pattern, and provide requirements for latency and throughput based on your production requirements.

Comment: since traffic patterns are already known, it should be B.

Replies:

Comment: with Default job type, you only need to provide a model package Amazon Resource Name(ARN) to lunch this type of recommendation job, it does not support providing custom traffic patterns.

Comment: A. Register the model artifact and container to the SageMaker Model Registry. Use the SageMaker Inference Recommender Default job type. Provide the known traffic pattern for load testing to select the best instance type and configuration based on the workloads. Explanation: SageMaker Model Registry allows you to register and organize your trained models. The SageMaker Inference Recommender Default job type simplifies the process of selecting the best instance type and configuration based on the known traffic pattern. It automatically selects the best instance type for the model. Load testing with the known traffic pattern helps in understanding the actual workloads and selecting the most appropriate instance type and configuration. This approach leverages the capabilities provided by SageMaker without the need for additional infrastructure or open-source tools

Discussion for Question 284

Link: https://www.examtopics.com/discussions/amazon/view/128798-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- BE: 13 votes

Discussion

Comment: https://docs.aws.amazon.com/sagemaker/latest/dg/model-managed-spot-training.html

Comment: to pick up where you left off, checkpointing is important and 'sport instances' to save cost

Comment: Checkpoints & Spot instances

Comment: Spot instances and checkpoints.

Comment: Spot Instance are cheapest and can be used with Checkpoints.

Comment: To meet the requirements of reducing training costs and being cost-effective in an Amazon SageMaker environment, the company should consider the following combination of resources: E. Spot Instances: Spot Instances are spare EC2 instances that are available at a lower cost compared to On-Demand Instances. By using Spot Instances for training, the company can significantly reduce the cost of running SageMaker training jobs. B. Checkpoints: Checkpoints allow the model training process to save the model's current state during training. If the training job is interrupted (e.g., due to a Spot Instance termination), the model can resume from the last saved checkpoint rather than starting from scratch.

Discussion for Question 285

Link: https://www.examtopics.com/discussions/amazon/view/128799-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: B. Blazing Text for text classification.

Comment: BlazingText's implements a supervised multi-class, multi-label text classification algorithm.

Comment: B. Use the SageMaker BlazingText algorithm. Explanation: BlazingText for Text Classification: SageMaker BlazingText is designed for efficient and scalable text classification tasks. It supports multi-class classification, making it suitable for the scenario where user feedback needs to be classified into fixed categories. BlazingText uses a fast implementation of the Word2Vec algorithm, making it highly performant.

Discussion for Question 286

Link: https://www.examtopics.com/discussions/amazon/view/128698-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 9 votes
- A: 6 votes

Discussion

Comment: B 1.Data Cleaning: SageMaker Data Wrangler is designed for data preparation tasks, including handling missing values, duplicates, and rare values. It provides a visual interface to clean and transform tabular data efficiently. This addresses the data cleaning requirements mentioned in the question. 2.Model Training: Using the built-in SageMaker XGBoost algorithm is a common and effective choice for classification tasks like customer churn prediction. XGBoost is a powerful and widely used algorithm for binary classification problems.

Comment: B. Use SageMaker Data Wrangler to clean the data. Use the built-in SageMaker XGBoost algorithm to train a classification model. Explanation: SageMaker Data Wrangler: SageMaker Data Wrangler is designed for efficient data cleaning and preparation. It provides a visual interface that simplifies the process of cleaning tabular data, handling missing values, and addressing duplicate or rare values. Data Wrangler can generate the necessary preprocessing code automatically, reducing the development effort. SageMaker XGBoost (for Classification): XGBoost is a popular and powerful algorithm for classification tasks, including customer churn prediction. SageMaker provides a built-in XGBoost algorithm, making it easy to train a classification model without the need for extensive coding.

Comment: SageMaker Canvas is an excellent tool for those without ML expertise to build models, but it may not provide the detailed control needed for data cleaning and may not be as robust as Data Wrangler for complex cleaning tasks.

Comment: https://aws.amazon.com/tw/about-aws/whats-new/2022/05/amazon-sagemaker-canvas-adds-new-data-capabilities-usability-updates/

Comment: Answer A- Sagemaker Canvas + categorical model Reason : SageMaker Canvas: SageMaker Canvas is a no-code machine learning tool that allows users to perform data preparation, feature engineering, and model training with minimal technical expertise. It automatically handles tasks like data cleaning, including the removal of duplicates, filling missing values, and managing rare categories. Categorical Model: A categorical (classification) model is the correct type for churn prediction, as it aims to classify whether a customer will stop using the service (churn) or not. SageMaker Canvas provides user-friendly tools to build and evaluate this type of model.

Comment: While Amazon SageMaker Canvas can perform automatic data cleaning and preparation, it has certain limitations when it comes to handling complex data cleaning tasks. SageMaker Canvas is designed for building machine learning models with minimal code and effort, primarily targeting business analysts and non-technical users. It provides a guided user interface and automates many steps in the machine learning pipeline, including data cleaning and preparation. However, SageMaker Canvas has a set of built-in data cleaning and preparation operations, which may not be sufficient for handling all types of data quality issues or complex data transformations. If the data requires more advanced cleaning techniques or custom transformations, SageMaker Data Wrangler (option B) would be a better choice.

Comment: A is correct Canvas can do without writing single line of code

Comment: This can be done without code using SageMaker Canvas: https://aws.amazon.com/blogs/machine-learning/predicting-customer-churn-with-no-code-machine-learning-using-amazon-sagemaker-canvas/ Hence, A is right.

Comment: The best solution, meeting the requirements with the least development effort and correctly addressing the problem nature, is: A. Use SageMaker Canvas to automatically clean the data and to prepare a categorical model. This option leverages the simplicity and automatic features of SageMaker Canvas, ensuring minimal development effort while accurately targeting the need for a classification model in customer churn prediction.

Comment: See: https://aws.amazon.com/blogs/machine-learning/predicting-customer-churn-with-no-code-machine-learning-using-amazon-sagemaker-canvas/ Canvas also does no-code data cleaning and preparation. So, least development effort is Canvas.

Discussion for Question 287

Link: https://www.examtopics.com/discussions/amazon/view/128699-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- BE: 5 votes

Discussion

Comment: B. Use SageMaker Data Wrangler diagnostic visualization. Use principal components analysis (PCA) and singular value decomposition (SVD) to calculate singular values. PCA and SVD: These methods help identify multicollinearity by reducing the dataset's dimensionality, revealing relationships among variables. Multicollinear features often become evident through high correlations in principal components or singular values. C. Use the SageMaker Data Wrangler Quick Model visualization to quickly evaluate the dataset and to produce importance scores for each feature. Quick Model Visualization: This feature enables rapid evaluation of feature importance scores, which can

help detect multicollinearity by identifying features that may be overly correlated and thus less impactful independently.

Comment: B and E make sense

Comment: https://aws.amazon.com/about-aws/whats-new/2021/08/detect-multicollinearity-amazon-sagemaker-data-wrangler/

Comment: PCA and SVD calculate singular values, which indicate the contribution of each feature to the overall variance. Features with high singular values have less multicollinearity. LASSO regularization shrinks coefficient values of highly correlated features towards zero, highlighting potential multicollinearity through their relative sizes.

Comment: B. Use SageMaker Data Wrangler diagnostic visualization. Use principal components analysis (PCA) and singular value decomposition (SVD) to calculate singular values. PCA and SVD can help in identifying multicollinearity by analyzing the correlation structure of the variables. High condition numbers or small singular values may indicate multicollinearity issues. D. Use the SageMaker Data Wrangler Min Max Scaler transform to normalize the data. Normalizing the data using techniques like Min-Max scaling can mitigate the impact of multicollinearity. Normalization helps in bringing the features to a similar scale, reducing the sensitivity to differences in magnitudes.

Comment: B and E Explanation: Option B: Principal components analysis (PCA) and singular value decomposition (SVD) are techniques used to identify multicollinearity in a dataset. By visualizing the singular values, the data engineer can assess the level of multicollinearity present in the features. This approach is effective for detecting relationships among variables. Option E: LASSO (Least Absolute Shrinkage and Selection Operator) is a regularization technique that can be used to penalize certain coefficients and, in turn, highlight the most important features. By plotting the coefficient values from a LASSO model, the data engineer can identify variables that contribute the most to the model. This can be useful for identifying and mitigating multicollinearity.

Discussion for Question 288

Link: https://www.examtopics.com/discussions/amazon/view/133244-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- D: 8 votes

Discussion

Comment: ChatGPT

Comment: B Amazon QuickSight dashboard to display near real-time! order insights B provides the most efficient solution for near real-time access to new order information in QuickSight

Comment: Option C involves using an API call from QuickSight to access the data directly in Amazon DynamoDB. While this option can provide real-time access to the data, it requires direct integration between QuickSight and DynamoDB, which may involve additional development effort. Additionally, QuickSight's native integration with DynamoDB for real-time data access might be limited compared to its integration with data stored in Amazon S3. Therefore, while option C might offer real-time access, option D with Kinesis Data Firehose to S3 could be a more robust and scalable solution, especially considering the potential limitations of direct DynamoDB integration with QuickSight.

Comment: D is the best solution given options. if not directly, QuickSight can connect to DynamoDB via Athena using a connector https://aws.amazon.com/blogs/big-data/visualize-amazon-dynamodb-insights-in-amazon-quicksight-using-the-amazon-athena-dynamodb-connector-and-aws-glue/

Comment: Quicksight doesn't integrate with DynamoDB directly. It could use S3, Redshift, Aurora/RDS, Athena, IOT analytics and EC2 hosted databases as data sources. Glouw would work as well but Firehose (D) is the least delay option.

Comment: QuickSight provides the ability to connect to various data sources, including DynamoDB, to create visualizations and dashboards. QuickSight supports a direct connection to DynamoDB tables, allowing you to query and visualize data stored in DynamoDB in real-time. No need to consume a DynamoDB stream with firehose. C is right.

Replies:

> **Comment:** Quicksight doesn't integrate with DynamoDB directly

Comment: he best solution, considering the requirement for the least delay and the ability to handle continuous data flow efficiently, would be: D. Use Amazon Kinesis Data Firehose to export the data from Amazon DynamoDB to Amazon S3, and configure QuickSight to access the data in Amazon S3. This solution leverages the automatic, scalable streaming capture of Kinesis Data Firehose to move data into S3, where it can be readily accessed by QuickSight for analytics and visualization purposes. This approach balances the need for near real-time insights with the capabilities of AWS services to handle streaming data effectively.

Discussion for Question 289

Link: https://www.examtopics.com/discussions/amazon/view/133242-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 6 votes

Discussion

Comment: A - kinessis added unnecessary additional cost and complexity and will add to latency B - wrangler is better suited for daat prep & feature engineering no merging C - Serverless so cost effective, trigger happens immediately so within lambda 15 min window and glue is made for these use cases D - Costly setup and maintenace

Comment: hint is 30 to 60 minutes and lambda has 15 minutes. Plus Glue provides many built-in functionality to perform the merge process much easier

Comment: A is not needed as we don't need to add Kinesis, it has no purpose here. B is possible but DataWrangler is more expensive then C. C is serverless and cost optimized, so C is correct. D is obviously too expensive.

Comment: C. Use an S3 event on the AWS Data Exchange S3 bucket to invoke an AWS Lambda function. Program the Lambda function to run an AWS Glue job that will merge the existing business data with the Athena table. Write the results back to Amazon S3. This solution avoids the need for continuous data streams or provisioning a persistent database cluster, which can incur higher costs. AWS Lambda can trigger cost-effective, short-duration tasks, and AWS Glue is a managed ETL service that can handle the data transformation and merging efficiently. The integration with Amazon S3 and Athena also aligns with the existing data flow and tools.

Comment: fix c - > b The most cost-effective solution is to use an S3 event to trigger a Lambda function that uses SageMaker Data Wrangler to merge the data. This solution avoids the need to provision and manage any additional resources, such as Kinesis streams, Firehose delivery streams, Glue jobs, or Redshift clusters. SageMaker Data Wrangler provides a visual interface to import, prepare, transform, and analyze data from various sources, including AWS Data Exchange products. It can also export the data preparation workflow to a Python script that can be executed by a Lambda function. This solution can meet the time requirement of 30-60 minutes, depending on the size and complexity of the data. References: Using Amazon S3 Event Notifications Prepare ML Data with Amazon SageMaker Data Wrangler AWS Lambda Function

Comment: he most cost-effective and straightforward solution is C. Use an S3 event on the AWS Data Exchange S3 bucket to invoke an AWS Lambda function. Program the Lambda function to run an AWS Glue job that will merge the existing business data with the Athena table and write the results back to Amazon S3. This approach leverages the serverless architecture of AWS, minimizing operational overhead and cost while ensuring the transformations can be completed within the desired timeframe.

Discussion for Question 290

Link: https://www.examtopics.com/discussions/amazon/view/133245-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- DE: 6 votes

Discussion

Comment: D and E simple

Comment: agree 100%

Comment: Conclusion: The findings that indicate an ML-based solution is suitable for predictive maintenance in this scenario are: D. The historical sensor data from the cranes are available with high granularity for the last 3 years. E. The historical sensor data contains most common types of crane failures that the company wants to predict. These points suggest the availability of comprehensive and relevant data necessary for developing an effective ML model for predictive maintenance.

Discussion for Question 291

Link: https://www.examtopics.com/discussions/amazon/view/133249-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- CD: 7 votes

Discussion

Comment: Object detection for the count and Pose detection for posture

Comment: Object detection + pose detection will do.

Comment: agrees

Comment: C. Object Detection: This model can identify and locate multiple objects within an image frame. For the task of counting the number of students in a class, object detection models can recognize and count the number of people present. This is essential for understanding class size and ensuring that each student is accounted for in the analysis. D. Pose Estimation: Pose estimation models are designed to determine the positions and orientations of human bodies in images or videos. They can identify the location and angle of a person's arms, legs, and other body parts. This capability is crucial for analyzing whether a student is performing a yoga stretch correctly by comparing their pose to the desired alignment and form for each yoga pose.

Discussion for Question 292

Link: https://www.examtopics.com/discussions/amazon/view/133519-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: A is correct

Comment: Typical multivariant deployment workflow. - No additional endpoints required which eliminates A & C. - B: question isn't about autoscaling but traffic routing - D: textbook production variant deployment method

Comment: needs to use both product variant and target variant in this requirement

Comment: https://docs.aws.amazon.com/sagemaker/latest/dg/model-ab-testing.html

---

Discussion for Question 293

Link: https://www.examtopics.com/discussions/amazon/view/133250-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- BCE: 5 votes

Discussion

Comment: data scientist already was querying the athena plus invoking sagemaker endpoint issue would not solve. therefore, A is not a good choice

Comment: A: NO - not needed as user already has Athena access (he is already querying Athena by SQL) B: Yes - sagemaker:InvokeEndpoint permission is needed to invoke endpoint C: Yes - needed for IAM user context to read S3 bucket D: No - sagemaker:GetRecord has no relevance in this question E: Yes - used to call an external function, in this case, the ML function deployed on the SageMaker endpoint, within the Athena SQL query F: No - irrelevant

Comment: ABE C is wrong. Why sagemaker need access S3? Sagemaker receive data and request via the endpoint.

Comment: https://docs.aws.amazon.com/athena/latest/ug/querying-mlmodel.html https://docs.aws.amazon.com/athena/latest/ug/machine-learning-iam-access.html

Comment: The correct combination of actions to enable the data scientist's IAM user to invoke the SageMaker endpoint is B, C, and E, because they ensure that the IAM user has the necessary permissions, access, and syntax to query the ML model from Athena. These actions have the following benefits: B: Including a policy statement for the IAM user that allows the sagemaker:InvokeEndpoint action grants the IAM user the permission to call the SageMaker Runtime InvokeEndpoint API, which is used to get inferences from the model hosted at the endpoint1.

Discussion for Question 294

Link: https://www.examtopics.com/discussions/amazon/view/133253-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 5 votes

Discussion

Comment: The fact that the mode is lower than the median, and the median is lower than the mean, suggests that the data is positively skewed (i.e., has a long right tail). In such cases, a logarithmic transformation is often used to reduce skewness and make the data more symmetric. Therefore, the correct answer is B. Logarithmic transformation.

Comment: Explanation: A logarithmic transformation is a suitable data transformation for a linear regression model when the data has a skewed distribution, such as when the mode is lower than the median and the median is lower than the mean. A logarithmic transformation can reduce the skewness and make the data more symmetric and normally distributed, which are desirable properties for linear regression. A logarithmic transformation can also reduce the effect of outliers and heteroscedasticity (unequal variance) in the data. An exponential transformation would have the opposite effect of increasing the skewness and making the data more asymmetric. A polynomial transformation may not be able to capture the nonlinearity in the data and may introduce multicollinearity among the transformed variables. A sinusoidal transformation is not appropriate for data that does not have a periodic

---

Discussion for Question 295

Link: https://www.examtopics.com/discussions/amazon/view/133248-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 6 votes

Discussion

Comment: Should be D

Comment: For the outcome categories with partial information (3 or 4 out of 200 categories), supervised learning can be used to classify claims into those categories based on the available claim contents. For the remaining outcome categories without partial information, forecasting techniques using claim IDs and dates can be employed to predict the expected number of claims in each category every month.

Comment: While the argument for option C is valid in terms of using claim IDs and dates for forecasting, it does not address the scenario where partial information on claim contents is available for some outcome categories. By

ignoring this information, option C may miss an opportunity to improve the accuracy of predictions for those categories through classification techniques. Furthermore, various machine learning resources and best practices recommend combining different techniques, such as classification and forecasting, when dealing with complex datasets that contain both structured and unstructured data. This hybrid approach can often lead to more accurate and robust solutions.

Comment: A: No - not a classification problem B: No - Reinforcement learning does not apply to the situation - adding positive reinforcement/negative penalty to train the system does not apply C: Yes - leverages historical data (claim IDs and dates from the previous 3 years) to forecast future claim counts D: Not a classification problem C:

Comment: this is forecasting problem

Comment: C directly addresses the need to forecast the number of claims in each outcome category on a monthly basis, leveraging historical data patterns without the need for classifying individual claim records based on their content.

Discussion for Question 297

Link: https://www.examtopics.com/discussions/amazon/view/133265-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- D: 9 votes

Discussion

Comment: EventBridge provide least development effort because we can just configure and trigger based on dataset drops in S3. Lambda will require some development effort

Comment: D requires minimal effort as it involves configuring EventBridge to monitor the S3 bucket for new data uploads and automatically triggering the SageMaker pipeline to perform the transformations on the new data while leveraging native Eventbridge -> Pipeline integration.

Comment: https://docs.aws.amazon.com/sagemaker/latest/dg/automating-sagemaker-with-eventbridge.html

Comment: Answer: D Explanation: The best solution is to configure Amazon EventBridge to run a predefined SageMaker pipeline to perform the transformations when a new data is detected in the S3 bucket. This solution requires the least development effort because it leverages the native integration between EventBridge and SageMaker Pipelines, which allows you to trigger a pipeline execution based on an event rule. EventBridge can monitor the S3 bucket for new data uploads and invoke the pipeline that contains the same transformations and feature engineering steps that were defined in SageMaker Data Wrangler. The pipeline can then ingest the transformed data into the online feature store for training and inference.

Discussion for Question 298

Link: https://www.examtopics.com/discussions/amazon/view/133096-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 7 votes

Discussion

Comment: https://docs.aws.amazon.com/sagemaker/latest/dg/shadow-tests.html

Comment: Shadow deployment is new technique provided by sagemaker

Comment: Answer: C Explanation: The best solution for this scenario is to use shadow deployment, which is a technique that allows the company to run the new experimental model in parallel with the existing model, without exposing it to the end users. In shadow deployment, the company can route the same user requests to both models, but only return the responses from the existing model to the users. The responses from the new experimental model are logged and analyzed for quality and performance metrics, such as accuracy, latency, and resource consumption12. This way, the company can validate the new experimental model in a production environment, without affecting the current live traffic or user experience

Comment: Shadow deployment consists of releasing version B alongside version A, fork version A's incoming requests, and send them to version B without impacting production traffic. This is particularly useful to test production load on a new feature and measure model performance on a new version without impacting current live traffic. source: https://aws.amazon.com/blogs/machine-learning/deploy-shadow-ml-models-in-amazon-sagemaker/

Discussion for Question 299

Link: https://www.examtopics.com/discussions/amazon/view/133097-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 6 votes

Discussion

Comment: Why not D?

Replies:

> **Comment:** ChatGPT: While option D includes the use of Amazon SageMaker Model Monitor, it suggests using only the most recent data for incremental training. This could result in the loss of valuable information from older data, which might still be relevant. Incremental training should ideally update the model with new data while retaining useful insights from the entire dataset, not just the recent months.

Comment: Incremental training involves updating the model with new data over time. Amazon SageMaker Model Monitor is a suitable choice for monitoring model performance. It can detect drift and anomalies in real-time predictions and send notifications.

Comment: Option A makes more sense in this case

Comment: A cover all the requirements

Discussion for Question 300

Link: https://www.examtopics.com/discussions/amazon/view/133098-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- AD: 9 votes

Discussion

Comment: Correct AD

Comment: first, classify the student profiles. then use classification algorithm to run predictions

Comment: K-means is unsupervised, so not useful for clustering. For grouping, use GroundTruth. It's a classification problem. So, A and D are right.

Comment: This question is focusing on either yes/no type of response (binary). So I think Classification algorithm would work the best as compared to K-means which is solely responsible for clustering the data.

Comment: D. Use a classification algorithm to run predictions: This approach is suitable for binary outcomes, such as predicting whether a student will enroll ("enrolled") or not ("not enrolled"). A. Use Amazon SageMaker Ground

Truth to sort the data into two groups named "enrolled" or "not enrolled.": This service can help in labeling the dataset accurately, providing a strong foundation for training the classification model.

Comment: The data scientist should use Amazon SageMaker Ground Truth to sort the data into two groups named "enrolled" or "not enrolled." This will create a labeled dataset that can be used for supervised learning. The data scientist should then use a classification algorithm to run predictions on the test data. A classification algorithm is a suitable choice for predicting a binary outcome, such as enrollment status, based on the input features, such as academic performance. A classification IT Certification Guaranteed, The Easy Way! 163 algorithm will output a probability for each class label and assign the most likely label to each observation. References: Use Amazon SageMaker Ground Truth to Label Data Classification Algorithm in Machine Learning

Comment: It mentions combination of options. It is a classification problem and labels will be needed.

Discussion for Question 301

Link: https://www.examtopics.com/discussions/amazon/view/133266-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- CD: 9 votes

Discussion

Comment: https://docs.aws.amazon.com/sagemaker/latest/dg/distributed-training.html

Comment: CD Given the specific context of training a DeepAR forecasting model and the potential cost implications, the options B and D are generally more applicable and cost-effective approaches to decreasing training time. However, if cost is not a concern and the DeepAR algorithm can benefit significantly from GPU acceleration, then option C could be a valid approach as well.

Comment: C and D

Comment: CD is correct answer

Comment: The best approaches to decrease the training time of the model are C and D, because they can improve the computational efficiency and parallelization of the training process. These approaches have the following benefits: C: Replacing CPU-based EC2 instances with GPU-based EC2 instances can speed up the training of the DeepAR algorithm, as it can leverage the parallel processing power of GPUs to perform matrix operations and gradient computations faster than CPUs12. The DeepAR algorithm supports GPUbased EC2 instances such as ml.p2 and ml.p33. D: Using multiple training instances can also reduce the training time of the DeepAR algorithm, as it can distribute the workload across multiple nodes and perform data parallelism4. The DeepAR algorithm supports distributed training with multiple CPU-based or GPU-based EC2 instances3. The other options are not effective or relevant, because they have the following drawbacks:

Discussion for Question 302

Link: https://www.examtopics.com/discussions/amazon/view/133089-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: It is B

Comment: To maximize the probability of detecting an abnormality, the focus should be on high recall (the ability of the model to find all actual positives), especially in scenarios where missing an abnormality could have significant negative effects. Between the given options: B. Precision = 0.61 - Recall = 0.98 This option has the highest recall, meaning it is best at identifying actual abnormalities (label 1), which is crucial for minimizing the risk of undetected process abnormalities. Although precision is lower (indicating more false positives), in this context, ensuring abnormalities are detected (even at the cost of investigating more false alarms) is more critical.

Comment: if abnormality is not detected then higher cost. Therefore FN (false negatives) should be minimized. Which means recall should have the highest value

---

Discussion for Question 303

Link: https://www.examtopics.com/discussions/amazon/view/134338-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: Multi-model endpoints provide a scalable and cost-effective solution to deploying large numbers of models. https://docs.aws.amazon.com/sagemaker/latest/dg/multi-model-endpoints.html

Comment: By preparing a SageMaker Docker container based on the open-source multi-model server, the company can host all models in a single endpoint and dynamically select the appropriate model based on the city of each request. This approach optimizes resource utilization and avoids managing unnecessary resources, as opposed to having separate instances for each city

Comment: A multi-model endpoint in Amazon SageMaker is an endpoint that can host multiple machine learning models simultaneously. This allows you to deploy and manage multiple models on a single endpoint, reducing operational costs and simplifying deployment and management tasks. Each model is associated with a specific container image and can be invoked using a unique model name or endpoint name. This feature is useful when you have multiple models that need to be deployed together or when you want to reduce the number of endpoints that need to be managed.

Discussion for Question 304

Link: https://www.examtopics.com/discussions/amazon/view/133268-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: The topic is wierd. GPU is idle, which means CPU is not able to feed data to GPU in time, which means more CPU is needed. The current ratio is 12:1, and you need to increase the ratio. D is wrong because 6:1 is smaller than 12:1

Comment: A: No - removing GPU could significantly increase training time B: No - doesn't solve the issue of GPU under utilization C: No - doesn't solve the issue of GPU under utilization and may take longer D: Yes - Reducing CPU's should solve the issue of GPU underutilization withour causing training delays

Comment: ChatGPT D. Switch to an instance type that has a CPU:GPU ratio of 6:1. This solution aligns with reducing training costs without extending the duration of training jobs. By selecting an instance with a lower CPU:GPU ratio, the specialist can ensure more consistent utilization of the GPU, thereby reducing idle time and optimizing resource use without compromising training efficiency.

Replies:

> **Comment:** Is it from GPT 4? Because I got option B from it.
>
> **Replies:**
>
> > **Comment:** yes, GPT4o says the option D is the right answer

Comment: Switching to an instance type that has a CPU: GPU ratio of 6:1 will reduce the training costs by using fewer CPUs and GPUs, while maintaining the same level of performance. The GPU idle time indicates that the CPU is not able to feed the GPU with enough data, so reducing the CPU: GPU ratio will IT Certification Guaranteed, The Easy Way! 167 balance the workload and improve the GPU utilization. A lower CPU: GPU ratio also means less overhead for inter-process communication and synchronization between the CPU and GPU processes. References: Optimizing GPU utilization for AI/ML workloads on Amazon EC2 Analyze CPU vs. GPU Performance for AWS Machine Learning

---

Discussion for Question 305

Link: https://www.examtopics.com/discussions/amazon/view/133269-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 5 votes

Discussion

Comment: While SageMaker Data Wrangler (option C) is also a strong contender, DataBrew is slightly easier to use and requires even less implementation effort, especially for users who may not be as familiar with the SageMaker ecosystem.

Comment: Data Wrangler is better for ML work. Brew can be used as well

Comment: Data wrangler supports tight integration with Sagemaker and is better suited for this scenario since resampled data is used in further modelling. AWS Glue DataBrew is a data preparation service more for general purpose use.

Comment: Best for Data Wrangler

Comment: This is exactly what Data Wrangler is for

Comment: Answer: C Explanation: Amazon SageMaker Studio Data Wrangler is a visual data preparation tool that enables users to clean and normalize data without writing any code. Using Data Wrangler, the data scientist can easily import the time-series data from various sources, such as Amazon S3, Amazon Athena, or Amazon Redshift. Data Wrangler can automatically generate data insights and quality reports, which can help identify and fix missing values, outliers, and anomalies in the data. Data Wrangler also provides over 250 built-in transformations, such as resampling, interpolation, aggregation, and filtering, which can be applied to the data with a point-and-click interface. Data Wrangler can also export the prepared data to different destinations, such as Amazon S3, Amazon SageMaker Feature Store, or Amazon SageMaker Pipelines, for further modeling and analysis. D

Discussion for Question 306

Link: https://www.examtopics.com/discussions/amazon/view/133255-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- D: 5 votes

Discussion

Comment: It is the best way

Comment: D: visualization provides insights into regional sales trends over time and allows for comparisons between regions and the overall average.

Comment: D. Create an aggregated dataset by using the Pandas GroupBy function to get average sales for each year for each region. Create a bar plot, faceted by year, of average sales for each region. Add a horizontal line in each facet to represent average sales. This visualization allows the data scientist to compare yearly average sales across regions and see how each region's performance relates to the overall average, providing clear insights into trends and deviations.

Comment: Explanation: The best visualization for this task is to create a bar plot, faceted by year, of average sales for each region and add a horizontal line in each facet to represent average sales. This way, the data scientist can easily compare the yearly average sales for each region with the overall average sales and see the IT Certification Guaranteed, The Easy Way! 170 trends over time. The bar plot also allows the data scientist to see the relative performance of each region within each year and across years. The other options are less effective because they either do not show the yearly trends, do not show the overall average sales, or do not group the data by region. References: pandas.DataFrame.groupby - pandas 2.1.4 documentation pandas.DataFrame.plot.bar - pandas 2.1.4 documentation Matplotlib - Bar Plot - Online Tutorials Library

---

Discussion for Question 307

Link: https://www.examtopics.com/discussions/amazon/view/133256-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 6 votes

Discussion

Comment: https://docs.aws.amazon.com/sagemaker/latest/dg/data-wrangler-analyses.html#data-wrangler-time-series-anomaly-detection

Comment: Data Wrangler can do it all

Comment: Anomaly detection visualization feature in SageMaker Data Wrangler is designed to identify outliers in the dataset based on sensor data parameters such as temperature and pressure. By visually inspecting the anomalies, the ML specialist can easily identify and remove outliers using transformations within Data Wrangler data flows, minimizing operational overhead.

Comment: Anomaly detection visualization feature in SageMaker Data Wrangler is designed to identify outliers in the dataset based on sensor data parameters such as temperature and pressure. By visually inspecting the anomalies, the ML specialist can easily identify and remove outliers using transformations within Data Wrangler data flows, minimizing operational overhead.

Comment: Going with C

Comment: agree with C

Comment: Amazon SageMaker Data Wrangler is a tool that helps data scientists and ML developers to prepare data for ML. One of the features of Data Wrangler is the anomaly detection visualization, which uses an unsupervised ML algorithm to identify outliers in the dataset based on statistical properties. The ML specialist can use this feature to quickly explore the sensor data and find any anomalous values that may affect the model performance. The ML specialist can then add a transformation to a Data Wrangler data flow to remove the outliers from the dataset. The data flow can be exported as a script or a pipeline to automate the data preparation process. This option requires the least operational overhead compared to the other options. References: Amazon SageMaker Data Wrangler - Amazon Web Services (AWS) Anomaly Detection Visualization - Amazon SageMaker Transform Data - Amazon SageMaker

---

Discussion for Question 308

Link: https://www.examtopics.com/discussions/amazon/view/133088-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 6 votes

Discussion

Comment: With support for PCA in Data Wrangler, you can now easily reduce the dimensionality of a high dimensional data set in only a few clicks. You can access PCA by selecting Dimensionality Reduction from the “Add step” workflow. https://aws.amazon.com/about-aws/whats-new/2022/10/amazon-sagemaker-data-wrangler-reduce-dimensionality-pca/

Comment: PCA requires scaling => use min-max scaler

Comment: A: No - PCA requires feature scaling to remove dominance of high value variables B: Yes - Scaling addresses the issue of features with different ranges + PCA does feature reduction C: No - manual removal may lead to removal of important features D: No - manual removal may lead to removal of important features

Comment: Standard scaler is better for PCA

Comment: C performs a standard transformation and D removes variables with low correlations which will delete important features

---

Discussion for Question 309

Link: https://www.examtopics.com/discussions/amazon/view/133087-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- BD: 5 votes

Discussion

Comment: K means and PCA

Comment: K means and PCA

Comment: Classic clustering problem A: No - LDA is for topic modelling B: Yes - K-means is a clustering algorithm C: No - applies to images D: Yes - PCA makes sure only relevant features are selected E: No - FM is supervised regression/classification recommendation algorithm for sparse data

Comment: B. K-means: This algorithm is effective for clustering customers into distinct groups based on similarities across their features, which can reveal segments more likely to respond to marketing campaigns. D. Principal Component Analysis (PCA): Given the high dimensionality of the dataset, PCA can reduce the number of variables to a manageable size while retaining most of the variance, making the dataset more tractable for clustering algorithms like K-means.

Comment: k-means for clustering due to no comments regarding labels in the data and also PCA in order to reduce the amount of features

---

Discussion for Question 310

Link: https://www.examtopics.com/discussions/amazon/view/135608-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 6 votes

Discussion

Comment: https://aws.amazon.com/about-aws/whats-new/2022/04/amazon-sagemaker-data-wrangler-supports-random-sampling-stratified-sampling/

Comment: A. Balanced Class Representation. Stratified sampling divides the original dataset into strata (groups) based on the class labels. It then selects instances from each stratum in a proportional manner, ensuring that the class distribution in the training and validation datasets reflects the original class distribution. Improved Generalization. By having a balanced representation of all classes in the training and validation datasets, the model is exposed to a diverse range of instances during training. This helps the model learn the distinguishing features of each class more effectively, leading to better generalization performance on the validation dataset. Addressing Imbalanced Data. Stratified sampling directly addresses the issue of imbalanced data, which was identified as the root cause of the model's poor generalization performance on the validation dataset.

Comment: Stratified sampling

Comment: A: Yes - Stratified sampling ensures that each class is proportionally represented and mitigates the impact of class imbalance on model performance B: No - additional data about the majority classes does not solve class imbalance issue C: No - Does not solve class imbalance issue and may worsen the situation D: No - selecting data points at regular intervals does not solve class imbalance issue

---

Discussion for Question 311

Link: https://www.examtopics.com/discussions/amazon/view/135610-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- A: 10 votes
- B: 7 votes

Discussion

Comment: Amazon Athena does not natively support running Python code directly. Amazon Athena is primarily a serverless, interactive query service that allows you to analyze data in Amazon S3 using standard SQL. Use Apache Spark from within Amazon SageMaker. Amazon SageMaker allows you to run Jupyter notebooks and provides managed Apache Spark integration, which means you don't need to manage the underlying compute resources yourself. You can also use SageMaker to perform the analysis and pay only for the resources you consume during the execution of your queries.

Comment: A and not B also because of paying for queries that you run. Notebooks will continue to run and cost money

Comment: https://docs.aws.amazon.com/athena/latest/ug/notebooks-spark-working-with-notebooks.html

Comment: https://docs.aws.amazon.com/athena/latest/ug/notebooks-spark-editor.html

Comment: Correct Answer: A https://docs.aws.amazon.com/athena/latest/ug/notebooks-spark.html

Comment: It's A - Using Apache Spark on Amazon Athena https://aws-sdk-pandas.readthedocs.io/en/3.2.1/tutorials/041%20-%20Apache%20Spark%20on%20Amazon%20Athena.html

Comment: Serverless, Python, and Notebook are key elements for making the decision. It's B

Replies:

> **Comment:** I changed my mind, Athena supports spark. It's A

Comment: https://docs.amazonaws.cn/en_us/athena/latest/ug/notebooks-spark-getting-started.html

Comment: Just thinking out loud, how can it be not Redshift as well? The question also mentions pay for queries, and handle petabyte of data. Spark is an integration possible with Amazon Redshift, and Redshift has serverless version too. https://aws.amazon.com/blogs/aws/new-amazon-redshift-integration-with-apache-spark/

Comment: A: No - Athena does not support python code B: Yes - Sagemaker is serverless and SageMaker Processing allows you to run Spark jobs from a Jupyter notebook using Python. You only pay for resources used during processing jobs. C: No - involves managing the EMR cluster. You pay for running EC2 instances whether in use or not. D: No - Redshift can't run spark jobs and no native support for python/Jupiter notebooks

---

Discussion for Question 312

Link: https://www.examtopics.com/discussions/amazon/view/135611-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: Data Wrangler best tool for the job

Comment: Data wrangler for all you need when prepare data for ML.

Comment: A: No - more manual/operational overhead B: No - more manual/operational overhead C: No - Data transformation to parquet requires more unnecessary operational overhead D: Yes - least operational effort - wrangler has built-in identification of data quality issues and outliers

---

Discussion for Question 313

Link: https://www.examtopics.com/discussions/amazon/view/135612-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: Correct B. It seems there was a typographical error in the provided options. "validation'll" is not a valid metric name. It appears to be an error or a typo. B. Tune the csv_weight hyperparameter and the scale_pos_weight hyperparameter by using automatic model tuning (AMT). Optimize on {"HyperParameterTuningJobObjective": {"MetricName": "validation:f1", "Type": "Maximize"}}.

Comment: the ll in the option B is recall, there must be some bug make the system miss some charactor.

Comment: Given the imbalanced nature of the dataset where only 5% of customers return items, the focus should be on maximizing the model's ability to correctly identify the returned items, which corresponds to maximizing the recall or F1 score. Option C and D aim to optimize the F1 score, but option D specifies minimizing the F1 score, which is incorrect. C. Tune all possible hyperparameters by using automatic model tuning (AMT). Optimize on {"HyperParameterTuningJobObjective": {"MetricName": "validation:f1", "Type": "Maximize"}}.

Comment: B. Tune the csv_weight hyperparameter and the scale_pos_weight hyperparameter by using automatic model tuning (AMT). Optimize on {"HyperParameterTuningJobObjective": {"MetricName": "validation:recall", "Type": "Maximize"}}. The dataset is imbalanced, with only 5% of customers returning items (or the positive class). The goal is typically to capture as many instances of the minority class (returned items) as possible, even at the expense of some false positives. Option D might be incorrect, as the goal is to maximize the model's ability to capture instances of returned items, not minimize the F1 score.

Comment: A: No - tuning all hyperparemeters requires compute - not very cost effective B: No - Log Loss has to be not applicable to imbalanced dataset C: No - tuning all hyperparemeters requires compute - not very cost effective D: Yes - F1 metric combines both precision and recall which is more suitable for unbalanced datasets

Replies:

Comment: but D sais MINIMIZE, so its not correct

Discussion for Question 314

Link: https://www.examtopics.com/discussions/amazon/view/135662-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: (ChatGPT) Option A (Hyperband): Efficiently utilizes computational resources. Reduces computation time by early stopping unpromising training jobs. Allows for a broader search of hyperparameter space within a shorter time. Option C (Lower MaxNumberOfTrainingJobs): Reduces the total number of training jobs. Directly decreases computation time. Helps stay within the small compute budget.

Comment: https://docs.aws.amazon.com/sagemaker/latest/dg/automatic-model-tuning-considerations.html#automatic-model-tuning-num-hyperparameters

Comment: Correct AE

Comment: A for sure C will reduce the tuning time because we are limiting the no. of Max Training jobs. E would be find with object is not reducing time

Comment: A: Yes - hyperband tuning early stops bad performing tuning jobs and reallocates resources to better performing ones B: No - Increase in hyperparams will increase time taken C: Yes - Limiting number of tuning jobs causes system not to run through entire list of tuning jobs reducing time D: No - grid search is computationally expensive and will take longer E: No - will increase time taken

Comment: A and D surely. Grid search to tune the hyperparameters.Grid search algorithm is more efficient than exhaustive search and will speed up tuning the hyperparameters.

Comment: Option A: Use the Hyperband tuning strategy. The Hyperband tuning strategy is a resource-efficient and time-saving approach for hyperparameter tuning. It works by running a set of hyperparameter configurations for a small number of training iterations and eliminating the poorly performing configurations early on. This strategy can significantly reduce the overall computation time compared to traditional methods like grid search or random search, especially for large hyperparameter spaces or time-consuming models like neural networks. Option E: Set a lower value for the MaxParallelTrainingJobs parameter. The MaxParallelTrainingJobs parameter in Amazon SageMaker specifies the maximum number of concurrent training jobs to be run in parallel during the hyperparameter tuning process. By setting a lower value for this parameter, the data scientist can limit the amount of computational resources used simultaneously, potentially reducing the overall computation time and cost.

Replies:

Comment: On second thought A,C makes more sense. C. Set a lower value for the MaxNumberOfTrainingJobs parameter. - The MaxNumberOfTrainingJobs parameter specifies the maximum number of training jobs that can be created during the tuning job. - Setting a lower value for this parameter will limit the number of training jobs and potentially reduce the computation time. - However, it may also limit the exploration of the hyperparameter space and potentially lead to suboptimal results. - This option should be considered with caution and in conjunction with other strategies to ensure adequate hyperparameter exploration.

Discussion for Question 315

Link: https://www.examtopics.com/discussions/amazon/view/135663-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- C: 5 votes

Discussion

Comment: https://docs.aws.amazon.com/sagemaker/latest/dg/automatic-model-tuning.html

Comment: automated model Tuning will be the best solution here

Comment: Amazon SageMaker automatic model tuning (AMT) for sure

Comment: Automated model tuning minimizes operational overhead because it automates the entire process of hyperparameter tuning, including setting up and managing the training jobs, tracking performance metrics, and selecting the best model configuration

Comment: C. Use Amazon SageMaker automatic model tuning (AMT). Specify a range of values for each hyperparameter.

Discussion for Question 316

Link: https://www.examtopics.com/discussions/amazon/view/135664-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: The Quick model of Data Wrangler calculates feature importance for each feature using the Gini importance method. https://docs.aws.amazon.com/sagemaker/latest/dg/data-wrangler-analyses.html

Comment: Wrangler allows to calculate feature importance scores using Gini importance

Comment: A. Use SageMaker Data Wrangler to perform a Gini importance score analysis. By using the Gini importance score analysis in Data Wrangler, the ML developer can obtain importance scores for each feature of the sales dataset with minimal development effort, as it is a built-in functionality with a visual interface. This approach requires no coding or additional setup, making it the least effort-intensive solution compared to the other options involving custom coding or separate analyses.

Discussion for Question 317

Link: https://www.examtopics.com/discussions/amazon/view/135665-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: https://docs.aws.amazon.com/sagemaker/latest/APIReference/API_CreatePresignedDomainUrl.html

Comment: - Domain specific presigned URL's can be generated through dynamodb to route users to correct domain - central proxy app can authenticate existing users with company IDP -> retrieve presigned url for users dept from dynamodb -> redirect user to sagemake domain without needing direct auth on AWS

Comment: A. Use the SageMaker CreatePresignedDomainUrl API to generate a presigned URL for each domain according to the DynamoDB table. Pass the presigned URL to the proxy application. The AWS documentation mentions the CreatePresignedDomainUrl API, which generates a presigned URL that authenticates a user to a specified Amazon SageMaker Domain. By using this API, the company can generate presigned URLs for each department's SageMaker Studio domain based on the information stored in the DynamoDB table. These presigned URLs can then be passed to the central proxy application, which can authenticate users using the company's existing Identity Provider (IdP) and provide them with the appropriate presigned URL for their department's SageMaker Studio domain. When the user accesses the presigned URL, they will be automatically authenticated and routed to the corresponding SageMaker Studio domain.

Discussion for Question 318

Link: https://www.examtopics.com/discussions/amazon/view/135988-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: D If your model is overfitting the training data, it makes sense to take actions that reduce model flexibility. To reduce model flexibility, try the following: Feature selection: consider using fewer feature combinations, decrease n-grams size, and decrease the number of numeric attribute bins. Increase the amount of regularization used. https://docs.aws.amazon.com/machine-learning/latest/dg/model-fit-underfitting-vs-overfitting.html

Comment: Increase the value of L2 regularization

Comment: A- No - momentum is for SGD with momentum, N/A in this case B: No - reducing dropout may or may not help C: No - reducing learning rate may increase overfitting even further D: Yes - L2 regularization penalizes large weights in the model. Increasing can help prevent overfitting by encouraging smaller weights.

Discussion for Question 319

Link: https://www.examtopics.com/discussions/amazon/view/135990-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: Resample is needed. https://docs.aws.amazon.com/sagemaker/latest/dg/data-wrangler-transform.html#data-wrangler-resample-time-series

Comment: identify if data has any underlying patterns or trends should be the first step

Comment: - Daily aggregation is needed to forecast daily demand and also takes care of missing hourly values. - Seasonal Trend decomposition on the daily aggregated data helps in understanding the underlying patterns, trends, and seasonality, which is essential for determining whether an ARIMA model would be appropriate.

Discussion for Question 320

Link: https://www.examtopics.com/discussions/amazon/view/135991-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: https://docs.aws.amazon.com/sagemaker/latest/dg/inter-network-privacy.html

Comment: VPC interface Endpoints will do the trick.

Comment: - VPC Interface Endpoints allow notbook instances to communicate with sagemaker services without public internet traffic - Security groups allow outbound connections for training and hosting but block all other traffic

---

Discussion for Question 321

Link: https://www.examtopics.com/discussions/amazon/view/135994-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: https://docs.aws.amazon.com/sagemaker/latest/dg/automatic-model-tuning-warm-start.html

Comment: using warm start allows to reuse the results of previous tune run

Comment: A: Yes - Warm start allows you to reuse the results from a previously performed hyperparameter tuning job B: No - is related to saving intermediate model checkpoints during training C: No - won't directly impact the tuning job time D: No - would increase computational resources but won't necessarily reduce time

---

Discussion for Question 322

Link: https://www.examtopics.com/discussions/amazon/view/136022-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: TF-IDF

Comment: TF captures the importance of a term within an individual document, while IDF captures the importance of a term across the entire corpus. Multiplying TF by IDF gives higher weights to terms that are frequent within a document but rare across the entire corpus, thus highlighting terms that are both relevant and distinctive.

Discussion for Question 323

Link: https://www.examtopics.com/discussions/amazon/view/136023-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: two different AZ will provide highly available deployment

Comment: This solution meets the requirements because it provides high availability by deploying multiple instances across different subnets in a second Availability Zone. This approach ensures that if one Availability Zone goes down, the other can continue to serve requests, achieving the desired recovery time objective (RTO) of 5 minutes. This solution requires the least effort compared to the others because it doesn't involve managing resources across multiple regions or frequent backups, and it's more directly targeted at high availability compared to auto-scaling. Please note that while auto-scaling (option B) can help handle increased load, it doesn't directly address high availability in terms of uptime or recovery time objectives. Options A and D involve multiple regions, which can add complexity and may not be necessary for achieving the desired high availability and RTO.

Comment: A: No - Multi region setup unnecessary B: No - Auto scaling is for capacity not for failovers C: Yes - Multiple instances running in separate subnets in two different AZ's provide quick failover D: No - Backups are for DR, not production failover

---

Discussion for Question 324

Link: https://www.examtopics.com/discussions/amazon/view/136015-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: Text tract to extracts text comprehend to classify

Comment: -Amazon Textract automatically extracts text, handwriting, and data from scanned pdf and jpg documents and requires minimal setup and maintenance. - Comprehend can be used for document classification and requires minimum ongoing management

Comment: The most common use cases for Amazon Textract include: - Importing documents and forms into business applications - Creating smart search indexes - Building automated document processing workflows - Maintaining compliance in document archives - Extracting text for Natural Language Processing (NLP) - Extracting text for document classification https://aws.amazon.com/tw/textract/faqs/?nc1=h_ls

---

Discussion for Question 325

Link: https://www.examtopics.com/discussions/amazon/view/136025-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- DE: 5 votes

Discussion

Comment: recall and f1

Comment: We need both precision and recall. The F1 score incorporates both of these metrics. and , the true positive rate which is only the recall.

Comment: E. True positive rate (also known as Recall or Sensitivity): True positive rate measures the proportion of actual fraudulent transactions that are correctly identified by the classifier. Maximizing the true positive rate ensures that as many fraudulent transactions as possible are captured by the model, reducing the number of false negatives. D. F1 score: F1 score is the harmonic mean of precision and recall. It provides a balance between precision (the ability of the classifier to correctly identify positive cases) and recall (the ability of the classifier to capture all positive cases). Maximizing the F1 score ensures a good balance between capturing fraudulent transactions (high recall) and minimizing false positives (high precision).

Comment: Review the Confusion matrix. Depending on your selected model score threshold, you can see the simulated impact based on a sample of 100,000 events. The distribution of fraud and legitimate events simulates the fraud rate in your businesses. Use this information to find the right balance between true positive rate and false positive rate. https://docs.aws.amazon.com/frauddetector/latest/ug/training-performance-metrics.html

Comment: F1 Score True Positive Rate, Recall, Sensitivity are all same thing

Comment: Metrics such as specificity (A), accuracy (C), and F1 score (D) are also important but may not directly prioritize the detection of fraudulent transactions. Specificity focuses on the proportion of non-fraudulent transactions correctly identified, accuracy measures overall correctness, and F1 score balances precision and recall. While these metrics are useful for evaluating the overall performance of the classifier, they may not be the primary focus when the goal is to detect as many fraudulent transactions as possible. Therefore, the most suitable metrics for optimizing the classifier to detect fraudulent transactions are False positive rate (B) and True positive rate (E).

Comment: These metrics will help the data scientist optimize the classifier to detect as many fraudulent transactions as possible.

Comment: - Maximizing TPR ensures that as many fraudulent transactions as possible are captured. - F1 score balances precision and recall and is useful when the class distribution is imbalanced (as in credit card fraud detection)

Discussion for Question 326

Link: https://www.examtopics.com/discussions/amazon/view/136011-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: Cross Region Replication

Comment: S3: scalable versioned object storage CRR - Automatically replicates objects from source to destination buckets in a different AWS Regions

Comment: https://docs.aws.amazon.com/AmazonS3/latest/userguide/replication.html#crr-scenario

---

Discussion for Question 327

Link: https://www.examtopics.com/discussions/amazon/view/136010-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: deploy in same endpoint and control the percent of traffic via Production Variants

Comment: Production variants -> deploy multiple models behind a single endpoint to distribute traffic between different variants of the model without making changes to the applications that rely on the API.

Comment: https://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-model-validation.html

---

Discussion for Question 328

Link: https://www.examtopics.com/discussions/amazon/view/136002-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- B: 7 votes

Discussion

Comment: DeepAR is a supervised RNNs and is able to handle missing values directly within the model. Instead of pre-processing the data to impute missing values externally, DeepAR can work directly with missing values encoded as NaN.

Comment: target—An array of floating-point values or integers that represent the time series. You can encode missing values as null literals, or as "NaN" strings in JSON, or as nan floating-point values in Parquet.

Comment: https://docs.aws.amazon.com/sagemaker/latest/dg/deepar.html

---

Discussion for Question 329

Link: https://www.examtopics.com/discussions/amazon/view/135917-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: A. The AnalyzeDocument API action in Amazon Textract allows for the analysis of various features within a document, including signatures. By setting the FeatureTypes parameter to SIGNATURES, the law firm can instruct Textract to specifically focus on detecting signatures within the contracts. Additionally, Textract provides confidence scores for detected elements, including signatures. Therefore, by using this action and specifying the FeatureTypes parameter, the law firm can receive confidence scores for each page of each contract, facilitating the automation of signature detection.

Comment: AnalyzeDocument Signatures feature automatically detects and extracts signature images from the document and comes with a confidence score (ranging from 0 to 100) for each detected signature

Comment: https://docs.aws.amazon.com/textract/latest/dg/API_AnalyzeDocument.html

---

Discussion for Question 330

Link: https://www.examtopics.com/discussions/amazon/view/136026-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- ADE: 5 votes

Discussion

Comment: Detecting corrosion is too specialized a task for Rekognition - trying transfer learning would literally mean re-training. AWS blog explaining suitability of Object Detection/Semantic Segmentation (though they prefer the color classification approach using XGBoost): https://aws.amazon.com/blogs/machine-learning/rust-detection-using-machine-learning-on-aws/ The enhancement step may be needed to contrast enhance/sharpen images.

Comment: The combination of steps that would meet the requirements is indeed A, B, and E

Comment: This is object detection problem, which Rekognitation can do since they have 0.1% corrosion photos, augmentation is necessary

Comment: A. Use an object detection algorithm: This can help identify corrosion areas in a photo. E. Perform image augmentation on photos with corrosion: This can improve the model's ability to generalize by increasing the diversity of the training data. D. Use an XGBoost algorithm: This can classify the severity of the corrosion after the areas of corrosion have been identified. It's effective for multi-class classification problems.

Comment: A: Yes - train a model that identifies corrosion areas within the photos B: Yes - identify and label objects, including corrosion, in the images. C: No - Not for classification D: No - XGBoost doesn't work on images E: Yes - rotation, scaling, and flipping, can enhance the model's ability to generalize and improve its performance F: not required

Comment: Object detection algorithm can be trained to identify corrosion rather than too customize Rekognition

---

Discussion for Question 331

Link: https://www.examtopics.com/discussions/amazon/view/136031-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: B seems to be most cost effective

Comment: The most cost-effective solution is Option B: Use AWS Batch with Amazon EC2 Spot Instances and Amazon FSx for Lustre. This approach leverages the efficiency of AWS Batch, the cost benefits of Spot Instances, and the high-performance of FSx for Lustre, making it ideal for scoring a batch model. However, Spot Instances can be interrupted, so they're best for flexible workloads.

Comment: AWS Batch managed Amazon EC2 Spot Instances - cost effective! FSx for Lustre - large volume (2 TB) data

---

Discussion for Question 332

Link: https://www.examtopics.com/discussions/amazon/view/136027-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

- D: 5 votes

Discussion

Comment: Hyperband is he answer

Comment: The best technique for the data scientist's requirements is Hyperband. It's designed for a large number of experiments, stops low-performance models early, and allocates more resources to high-performance models. This reduces computational time compared to Grid Search, Random Search, and Bayesian Optimization which don't have these features.

Comment: Hyperband involves training multiple models with different hyperparameter configurations , eliminating poorly performing ones and allocating resources to promising ones.

Comment: . Hyperband: This technique is a bandit-based approach that allocates resources efficiently by running multiple configurations in parallel with varying durations. It eliminates poorly performing configurations early and focuses resources on promising ones, making it ideal for minimizing compute time.

Discussion for Question 334

Link: https://www.examtopics.com/discussions/amazon/view/147148-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: C is the best option in this case.

Discussion for Question 336

Link: https://www.examtopics.com/discussions/amazon/view/147146-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: The traffic is expected. Provisioned resouces have minimal cost.

Comment: By choosing serverless inference with provisioned concurrency, the media company can benefit from low latency during peak traffic periods while optimizing costs by only paying for the actual inference requests

---

Discussion for Question 338

Link: https://www.examtopics.com/discussions/amazon/view/147147-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: SMOTE is most effective.

Discussion for Question 342

Link: https://www.examtopics.com/discussions/amazon/view/147150-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: FastFile mode provides the best balance

Discussion for Question 347

Link: https://www.examtopics.com/discussions/amazon/view/147603-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: Should be D.

Discussion for Question 348

Link: https://www.examtopics.com/discussions/amazon/view/147689-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: C, A -X firehose can't store data.

Discussion for Question 349

Link: https://www.examtopics.com/discussions/amazon/view/147151-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: The correct steps are A,C,F

Discussion for Question 351

Link: https://www.examtopics.com/discussions/amazon/view/147155-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: This is target leakage

Discussion for Question 352

Link: https://www.examtopics.com/discussions/amazon/view/147037-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: its B, similarity encoding

Discussion for Question 353

Link: https://www.examtopics.com/discussions/amazon/view/147154-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: BC are correct options

Discussion for Question 354

Link: https://www.examtopics.com/discussions/amazon/view/147153-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: DE are correct.

Discussion for Question 355

Link: https://www.examtopics.com/discussions/amazon/view/147152-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: One-hot encoding

Discussion for Question 356

Link: https://www.examtopics.com/discussions/amazon/view/147156-exam-aws-certified-machine-learning-specialty-topic-1/

Most Voted

Discussion

Comment: B is the correct choice. Use SageMaker Canvas
