



- Expert Verified, Online, **Free**.

[Custom View Settings](#)

Topic 1 - Single Topic

Question #1

Topic 1

A large mobile network operating company is building a machine learning model to predict customers who are likely to unsubscribe from the service. The company plans to offer an incentive for these customers as the cost of churn is far greater than the cost of the incentive.

The model produces the following confusion matrix after evaluating on a test dataset of 100 customers:

n= 100	PREDICTED CHURN	
	Yes	No
ACTUAL Churn Yes	10	4
Actual No	10	76

Based on the model evaluation results, why is this a viable model for production?

- 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.
- B. The precision of the model is 86%, which is less than the accuracy of the model.
- C. The model is 86% accurate and the cost incurred by the company as a result of false positives is less than the false negatives.
- D. The precision of the model is 86%, which is greater than the accuracy of the model.

Correct Answer: A

🗨️ 👤 **joep21** Highly Voted 👍 4 months, 4 weeks ago

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 negative (predicted not churn, actual churn). Incentive cost < churn cost.

upvoted 11 times

🗨️ 👤 **felbuch** 4 months, 3 weeks ago

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.

upvoted 6 times

🗨️ 👤 **ExamTaker177** 3 months ago

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.

upvoted 4 times

🗨️ 👤 **AShahine21** Most Recent 🕒 1 month ago

It should be C, as the Accuracy is 86% $(TP+TN)/(Total) = 86/100=86\%$

And FN is less than FP

upvoted 1 times

🗨️ 👤 **Bala1212081** 4 weeks ago

As you said FN is less than FP then in this context it should be A, but how it could be C?

upvoted 1 times

🗨️ 👤 **mhd911** 2 months, 3 weeks ago

Answer is C

Accuracy = $(10+76) / 100 = 86\%$

FN = 4

FP = 10

the cost of churn is far greater than the cost of the incentive.

FN plenty is greater than FP

FP plenty is less than FN

since FN is less than FP then it is a viable model.

M Moftah

upvoted 2 times

🗨️ 👤 **mhd911** 2 months, 3 weeks ago

Accuracy = $(10+76) / 100 = 86\%$

FN = 4

FP = 10

the cost of churn is far greater than the cost of the incentive.

FN plenty is greater than FP

FP plenty is less than FN

since FN is less than FP then it is a viable model.

M Moftah

upvoted 1 times

🗨️ 👤 **Sanj** 4 months, 3 weeks ago

Answer is C

upvoted 2 times

Question #2

Topic 1

A Machine Learning Specialist is designing a system for improving sales for a company. The objective is to use the large amount of information the company has on users' behavior and product preferences to predict which products users would like based on the users' similarity to other users.

What should the Specialist do to meet this objective?

- A. Build a content-based filtering recommendation engine with Apache Spark ML on Amazon EMR
- B. Build a collaborative filtering recommendation engine with Apache Spark ML on Amazon EMR.
- C. Build a model-based filtering recommendation engine with Apache Spark ML on Amazon EMR
- D. Build a combinative filtering recommendation engine with Apache Spark ML on Amazon EMR

Correct Answer: B

Many developers want to implement the famous Amazon model that was used to power the "People who bought this also bought these items" feature on

Amazon.com. This model is based on a method called Collaborative Filtering. It takes items such as movies, books, and products that were rated highly by a set of users and recommending them to other users who also gave them high ratings. This method works well in domains where explicit ratings or implicit user actions can be gathered and analyzed.

Reference:

<https://aws.amazon.com/blogs/big-data/building-a-recommendation-engine-with-spark-ml-on-amazon-emr-using-zeppelin/>

  **mlyu** Highly Voted 1 year, 6 months ago

B

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

upvoted 10 times

  **cybe001** Highly Voted 1 year, 5 months ago

B is correct

<https://aws.amazon.com/blogs/big-data/building-a-recommendation-engine-with-spark-ml-on-amazon-emr-using-zeppelin/>

upvoted 5 times

  **kalyanvarma** Most Recent 6 months, 1 week ago

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

upvoted 5 times

  **roytruong** 1 year, 1 month ago

go for B

upvoted 2 times

Question #3

Topic 1

A Mobile Network Operator is building an analytics platform to analyze and optimize a company's operations using Amazon Athena and Amazon S3.

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.

Which solution takes the LEAST effort to implement?

- A. Ingest .CSV data using Apache Kafka Streams on Amazon EC2 instances and use Kafka Connect S3 to serialize data as Parquet
- B. Ingest .CSV data from Amazon Kinesis Data Streams and use Amazon Glue to convert data into Parquet.
- C. Ingest .CSV data using Apache Spark Structured Streaming in an Amazon EMR cluster and use Apache Spark to convert data into Parquet.
- D. Ingest .CSV data from Amazon Kinesis Data Streams and use Amazon Kinesis Data Firehose to convert data into Parquet.

Correct Answer: B

🗨️ **DonaldCMLIN** Highly Voted 1 year, 7 months ago

Answer is B

<https://github.com/ecloudvalley/Building-a-Data-Lake-with-AWS-Glue-and-Amazon-S3>

upvoted 13 times

🗨️ **OmarSaadEldien** 6 months, 1 week ago

the Approve Of B

<https://aws.amazon.com/blogs/aws/new-serverless-streaming-etl-with-aws-glue/>

upvoted 3 times

🗨️ **Antriksh** 1 year ago

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

upvoted 2 times

🗨️ **zzeng** 11 months, 2 weeks ago

AWS Glue can do it now (2020 May)

<https://aws.amazon.com/jp/blogs/news/new-serverless-streaming-etl-with-aws-glue/>

upvoted 5 times

🗨️ **scuzzy2010** 11 months ago

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 say data must be converted to parquet in real time, it only says the csv data is received as a stream in real time.

upvoted 1 times

🗨️ **GeeBeeEI** 11 months ago

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



upvoted 2 times

  **vetal** Highly Voted 1 year, 6 months ago

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/search/convert-csv-json-files-to-apache-parquet-using-aws-glue-a760d177b45f>

upvoted 9 times

  **zzeng** 11 months, 2 weeks ago

<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

upvoted 4 times

  **StelSen** Most Recent 1 month, 3 weeks ago

The Answer is B. LINK: <https://aws.amazon.com/glue/faqs/>

(Search 'real' in this FAQ and you will get an answer. Answer D required one more component called Lambda. So straightaway rejected)

upvoted 1 times

  **Vita_Rasta84444** 3 months ago

D. Glue cannot work with live streams, while only Firehose could ingest the data to S3 and make transformation from csv to parquet.

upvoted 1 times

  **ahquiceno** 5 months ago

Answer is D go to: <https://docs.aws.amazon.com/firehose/latest/dev/record-format-conversion.html>

upvoted 1 times

  **harmanbirstudy** 5 months, 4 weeks ago

A-- cannot be answer as Apache kafka S3 cannot write in parquet

B-- seems like a good answer but if the question is old, then at that time Glue did not had compatibility with Kinesis data streams

C--- cannot be the answer as Spark structured streaming need to read data from somewhere like Kafka topics, we cannot publish data directly to it and like other three options mention Apache kafka/Kinesis in it so this means it should have been in this one also if this was the correct answer.

D-- Seems like a good answer, but it needs lambda to convert CSV -- JSON and then Firehose's inbuilt ability to convert JSON to Parquet before storing to S3


Now the final analysis -- :)

-- The comments on this question started some time in December 2019/January 2020 and the Glue capability to consume real time Kinesis data streams was announced much later on April 27, 2020 which means this answer must be incorrect before this date

--Hence the likely choice is D, even though firehose Lambda thing is not mentioned in it specifically

Any counter comment ???

upvoted 3 times

  **cnethers** 4 months, 4 weeks ago

Very good points made, I would agree that the timing of the question needs to be taken into account. When the question was first posted Glue could not take real-time data and transform it, so that would make B a no starter and D the Answer however if you consider that there is now integration of Kinesis and Glue you can now do real-time processing so B is now possible and less effort than D. On balance, the question may not even be asked nowadays but it does demonstrate how answers can change over time. Do your research and make sure you understand what capabilities a service has.

upvoted 1 times

  **mrsimoes** 10 months ago

I think that the answer is B.

Please check:

<https://aws.amazon.com/about-aws/whats-new/2020/04/aws-glue-now-supports-serverless-streaming-etl/#:~:text=AWS%20Glue%20now%20supports%20streaming%20ETL.&text=Streaming%20ETL%20jobs%20in%20AWS,warehouses%20or%20other%20data%20stores.>

%20or%20other%20data%20stores.

upvoted 2 times

🗨️ **syu31svc** 10 months, 1 week ago
 Answer is B as per link:
<https://aws.amazon.com/blogs/big-data/build-a-data-lake-foundation-with-aws-glue-and-amazon-s3/>
 upvoted 1 times

🗨️ **vivky247** 10 months, 3 weeks ago
 where do I find the link to other AWS Machine Learning questions ?
 upvoted 1 times

🗨️ **dikers** 10 months, 3 weeks ago
 Answer is D
https://docs.aws.amazon.com/zh_cn/firehose/latest/dev/record-format-conversion.html

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. Parquet and ORC are columnar data formats that save space and enable faster queries compared to row-oriented format like JSON.

upvoted 1 times

🗨️ **GeeBeeEI** 11 months ago
 Kinesis Firehose can only convert JSON to Parquet or ORC. It cannot convert CSV. For CSV you need a lambda transformation <https://docs.aws.amazon.com/firehose/latest/dev/record-format-conversion.html> hence D is out
 According to <http://blogs.quovantis.com/how-to-convert-csv-to-parquet-files/> Apache Spark can convert csv to parquet but results will be binary format, you will not be able to read it --- this is true for all parquet files..... C is possible. Remember its streaming data
 Glue can convert CSV/JSON to Parquet see <https://medium.com/search/convert-csv-json-files-to-apache-parquet-using-aws-glue-a760d177b45f> glue also supports streaming data coming from Kinesis data stream see <https://aws.amazon.com/glue/faqs/>
 Since the competition is between Spark and Glue, see <https://stackshare.io/stackups/aws-glue-vs-spark> , Glue is EASIER to implement (a requirement in the question)
 upvoted 1 times

🗨️ **Urban_Life** 1 year ago
 I think the answer will be B. Two reasons : 1. The Q's itself talks about Athena and S3.
 2. I've done that in my project (It's proven concept)
 upvoted 1 times

🗨️ **Antriksh** 1 year ago
 The answer would be D.
 Here is my explanation:
 Option A clearly requires maximum amount of efforts
 B & C rules out because you cannot read directly through apache kafka, you need a streaming queue like Kinesis. Additionally Glue is not a suitable option for streaming jobs (Though they have introduced capability of streaming within Glue 7 days back from the day of this comm
 So finally option D is the right option, where, the data would be streamed through firehose, it converts JSON to Parquet format with LEAST effort. I believe there would be a lambda blueprint to convert CSV to JSON.
 upvoted 1 times

🗨️ **ardisch** 1 year ago
 Sorry, Answer is D
https://aws.amazon.com/about-aws/whats-new/2018/05/stream_real_time_data_in_apache_parquet_or_orc_format_using_firehose/
 upvoted 1 times

🗨️ **ardisch** 1 year ago
 Answer is C, you need to save data so S3 and Firehose can convert data to Parquet format automatically.
https://aws.amazon.com/about-aws/whats-new/2018/05/stream_real_time_data_in_apache_parquet_or_orc_format_using_firehose/
 upvoted 1 times

🗨️ 👤 **mawsman** 1 year, 1 month ago

This question is made to make you fail.

- Least effort would indicate a managed service so that technically eliminates A and C.
- CSV to Parquet would eliminate Firehose which would be my choice because it would deliver to S3 directly. You could do it with a Lambda function to pre-process, but because it doesn't say that then we can eliminate D.
- Glue could do CSV to Parquet it but in that case you need to trigger the glue job somehow and it falls out of "real-time" scope and it does say that the glue job would deliver to S3. So that eliminates B.

The answer is actually C because this exact use case is demonstrated in this AWS blog where an EMR cluster with a pyspark script transfers data from CSV to Parquet:

<https://aws.amazon.com/blogs/big-data/analyzing-data-in-s3-using-amazon-athena/>

It's the only way to do it in real time and satisfy all the conditions. Whoever wrote this question just wants you to fail the exam. This isn't testing our knowledge, this is testing if you read the aws blogs.

upvoted 4 times

🗨️ 👤 **roytruong** 1 year, 1 month ago

B is possible, this link shows that glue can connect with streaming source easily like kinesis data stream or kafka.

<https://docs.aws.amazon.com/glue/latest/dg/add-job-streaming.html>

upvoted 1 times

Question #4

Topic 1

A city wants to monitor its air quality to address the consequences of air pollution. A Machine Learning Specialist needs to forecast the air quality in parts per million of contaminants for the next 2 days in the city. As this is a prototype, only daily data from the last year is available.

Which model is MOST likely to provide the best results in Amazon SageMaker?

- A. Use the Amazon SageMaker k-Nearest-Neighbors (kNN) algorithm on the single time series consisting of the full year of data with a predictor_type of regressor.
- B. Use Amazon SageMaker Random Cut Forest (RCF) on the single time series consisting of the full year of data.
- C. Use the Amazon SageMaker Linear Learner algorithm on the single time series consisting of the full year of data with a predictor_type of regressor.
- D. Use the Amazon SageMaker Linear Learner algorithm on the single time series consisting of the full year of data with a predictor_type of classifier.

Correct Answer: C

Reference:

<https://aws.amazon.com/blogs/machine-learning/build-a-model-to-predict-the-impact-of-weather-on-urban-air-quality-using-amazon-sagemaker/?ref=Welcome.AI>

🗲️ **ozan11** Highly Voted 👍 1 year, 5 months ago
answer should be C
upvoted 9 times

🗲️ **syu31svc** Most Recent 🕒 10 months, 1 week ago
If it's about forecasting then answer is C 100%; regression
upvoted 2 times

🗲️ **roytruong** 1 year, 1 month ago
go for C
upvoted 3 times

🗲️ **VB** 1 year, 4 months ago
C should be the answer.. D is not because it's not a classifier problem (..forecast the air quality in parts per million..) it should be regression
upvoted 2 times

🗲️ **rajs** 1 year, 4 months ago
It's a Regression problem so A)KNN B)RCF D) can be eliminated leaving us with the answer C
upvoted 3 times

🗲️ **Nahe** 6 months, 1 week ago
K-NN can also be used for regression. But it is not preferred here as we have very limited data of 2-3 days
upvoted 2 times

Question #5

Topic 1

A Data Engineer needs to build a model using a dataset containing customer credit card information

How can the Data Engineer ensure the data remains encrypted and the credit card information is secure?

- A. Use a custom encryption algorithm to encrypt the data and store the data on an Amazon SageMaker instance in a VPC. Use the SageMaker DeepAR algorithm to randomize the credit card numbers.
- B. Use an IAM policy to encrypt the data on the Amazon S3 bucket and Amazon Kinesis to automatically discard credit card numbers and insert fake credit card numbers.
- C. Use an Amazon SageMaker launch configuration to encrypt the data once it is copied to the SageMaker instance in a VPC. Use the SageMaker principal component analysis (PCA) algorithm to reduce the length of the credit card numbers.
- D. Use AWS KMS to encrypt the data on Amazon S3 and Amazon SageMaker, and redact the credit card numbers from the customer data with AWS Glue.

Correct Answer: D

  **vetal** Highly Voted 1 year, 6 months ago



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

upvoted 23 times

  **WWODIN** 1 year, 6 months ago

should be D

upvoted 8 times

  **zzeng** 11 months, 2 weeks ago

Should be D.

Use Glue to do ETL to Hash the card number

upvoted 5 times

  **Antriksh** 1 year ago

Answer would be D

upvoted 6 times

  **cybe001** Highly Voted 1 year, 5 months ago

D is correct

upvoted 7 times

  **jerto97** Most Recent 2 weeks, 3 days ago

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 not a good option. Hashing should be a safe option to keep it in the learning path

upvoted 1 times

  **cloud_trail** 5 months ago

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.

upvoted 1 times

- 🗨️ 👤 **cnethers** 5 months, 1 week ago
Can Glue do redaction?
upvoted 1 times
- 🗨️ 👤 **cloud_trail** 5 months ago
Just have the Glue job remove the credit card column.
upvoted 1 times
- 🗨️ 👤 **syu31svc** 10 months, 1 week ago
Encryption on AWS can be done using KMS so D is the answer
upvoted 1 times
- 🗨️ 👤 **roytruong** 1 year, 1 month ago
D is correct
upvoted 1 times
- 🗨️ 👤 **PRC** 1 year, 3 months ago
D..KMS fully managed and other options are too whacky..
upvoted 4 times
- 🗨️ 👤 **AKT** 1 year, 4 months ago
D is correct
upvoted 1 times
- 🗨️ 👤 **bhavesh0124** 1 year, 4 months ago
Ans D is correct
upvoted 2 times
- 🗨️ 👤 **satya_alapati** 1 year, 4 months ago
is this really a viable reference? so misleading.
upvoted 2 times
- 🗨️ 👤 **JayK** 1 year, 5 months ago

Question #6

Topic 1

A Machine Learning Specialist is using an Amazon SageMaker notebook instance in a private subnet of a corporate VPC. The ML Specialist has important data stored on the Amazon SageMaker notebook instance's Amazon EBS volume, and needs to take a snapshot of that EBS volume. However, the ML Specialist cannot find the Amazon SageMaker notebook instance's EBS volume or Amazon EC2 instance within the VPC. Why is the ML Specialist not seeing the instance visible in the VPC?

- A. Amazon SageMaker notebook instances are based on the EC2 instances within the customer account, but they run outside of VPCs.
- B. Amazon SageMaker notebook instances are based on the Amazon ECS service within customer accounts.
- C. Amazon SageMaker notebook instances are based on EC2 instances running within AWS service accounts.
- D. Amazon SageMaker notebook instances are based on AWS ECS instances running within AWS service accounts.

Correct Answer: C

Reference:

<https://docs.aws.amazon.com/sagemaker/latest/dg/gs-setup-working-env.html>

- 🗨️ 👤 **dhs227** Highly Voted 👍 1 year, 2 months ago

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/>

upvoted 13 times

  **mawsmann** 1 year, 1 month ago

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.

upvoted 9 times

  **scuzzy2010** 11 months ago

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.

upvoted 6 times

  **cloud_trail** 5 months ago

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.

upvoted 1 times

  **mlyu** Highly Voted 1 year, 5 months ago

I think the answer should be C

upvoted 9 times

  **AShahine21** Most Recent 1 month ago


I will go with C, as "but they run outside of VPCs" is wrong

upvoted 1 times

  **cloud_trail** 5 months ago



Answer C. I've gone back and forth on this but reading the info at the link below, the customer notebook instance has to run either in a customer managed VPC or the Sagemaker service managed VPC. Either way, it's in a VPC. <https://aws.amazon.com/blogs/machine-learning/understanding-amazon-sagemaker-notebook-instance-networking-configurations-and-advanced-routing-options/>

upvoted 3 times

  **Joe_Zhang** 5 months, 1 week ago


the answer should be C: see this link: <https://docs.aws.amazon.com/sagemaker/latest/dg/studio-notebooks-and-internet-access.html>

upvoted 1 times

  **crispogioele** 5 months, 3 weeks ago

I think the answer is A. They are EC2 instance and they are attached to the customer VPC through an ENI (as shown in this link <https://aws.amazon.com/blogs/machine-learning/understanding-amazon-sagemaker-notebook-instance-networking-configurations-and-advanced-routing-options/>)

upvoted 2 times

  **cnethers** 5 months, 1 week ago

I would agree that A is the answer. That is the same ref article I have looked at

upvoted 1 times

  **harmanbirstudy** 5 months, 3 weeks ago

C is correct as it is Fully managed service so you cannot have OS level access to its resources.

A-- cannot be correct because it says Sagemakers EC2 instances run with VPC which is not correct , as with VPC means governed by its network policies incase of Sagemaker

upvoted 2 times

  **Achievement** 11 months, 2 weeks ago

C should be correct since the EC2 instance is not under customer manage

upvoted 2 times

🗨️ 👤 **AdityaB** 1 year, 1 month ago

A is the right answer .

upvoted 5 times

🗨️ 👤 **roytruong** 1 year, 1 month ago

go for C

upvoted 3 times

🗨️ 👤 **cybe001** 1 year, 5 months ago

C looks correct, all sagemaker runs on EC2

<https://aws.amazon.com/sagemaker/pricing/instance-types/>

upvoted 2 times

Question #7

Topic 1

A Machine Learning Specialist is building a model that will perform time series forecasting using Amazon SageMaker. The Specialist has finished training the model and is now planning to perform load testing on the endpoint so they can configure Auto Scaling for the model variant. Which approach will allow the Specialist to review the latency, memory utilization, and CPU utilization during the load test?

- A. Review SageMaker logs that have been written to Amazon S3 by leveraging Amazon Athena and Amazon QuickSight to visualize logs as they are being produced.
- B. Generate an Amazon CloudWatch dashboard to create a single view for the latency, memory utilization, and CPU utilization metrics that are outputted by Amazon SageMaker.
- C. Build custom Amazon CloudWatch Logs and then leverage Amazon ES and Kibana to query and visualize the log data as it is generated by Amazon SageMaker.
- D. Send Amazon CloudWatch Logs that were generated by Amazon SageMaker to Amazon ES and use Kibana to query and visualize the log data.

Correct Answer: B

Reference:

<https://docs.aws.amazon.com/sagemaker/latest/dg/monitoring-cloudwatch.html>

  **mlyu** Highly Voted 1 year, 5 months ago

Agreed. Ans is B
upvoted 8 times

  **WillNguyen22** Most Recent 10 months ago

answer is B
upvoted 1 times

  **syu31svc** 10 months, 1 week ago

Answer is B 100%; very straightforward method
upvoted 1 times

  **scuzzy2010** 11 months ago

B is correct. Don't need to use Kibana or QuickSight.
upvoted 1 times

  **roytruong** 1 year, 1 month ago

ans is B
upvoted 3 times

  **cybe001** 1 year, 5 months ago

B is correct
upvoted 3 times

Question #8



Topic 1



A manufacturing company has structured and unstructured data stored in an Amazon S3 bucket. A Machine Learning Specialist wants to use SQL to run queries on this data.

Which solution requires the LEAST effort to be able to query this data?



- A. Use AWS Data Pipeline to transform the data and Amazon RDS to run queries.
- B. Use AWS Glue to catalogue the data and Amazon Athena to run queries.
- C. Use AWS Batch to run ETL on the data and Amazon Aurora to run the queries.
- D. Use AWS Lambda to transform the data and Amazon Kinesis Data Analytics to run queries.



Correct Answer: B



  **cybe001** Highly Voted 1 year, 5 months ago
B is correct
upvoted 13 times



  **dhs227** Highly Voted 1 year, 2 months ago
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 ad-hoc queries.
upvoted 5 times

  **gcpwhiz** Most Recent 2 months, 2 weeks ago
If AWS asks the question of querying unstructured data in an efficient manner, it is almost always Athena
upvoted 1 times

  **cloud_trail** 5 months ago
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.
upvoted 2 times

  **Willnguyen22** 10 months ago
answer is B
upvoted 1 times

  **syu31svc** 10 months, 1 week ago
SQL on S3 is Athena so answer is B for sure
upvoted 1 times

  **roytruong** 1 year, 1 month ago
B is right
upvoted 2 times

  **Jayraam** 1 year, 2 months ago



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/>

upvoted 1 times

  **PRC** 1 year, 3 months ago



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

upvoted 1 times

  **Urban_Life** 1 year ago



May I know why you are taking complex route?

upvoted 8 times

  **BigEv** 1 year, 5 months ago


Can Glue Crawler process unstructured data?

upvoted 1 times

  **ExamTaker123456789** 1 year ago

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 log files.

upvoted 1 times

  **mawsman** 1 year, 1 month ago

<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.

Question #9

Topic 1

A Machine Learning Specialist is developing a custom video recommendation model for an application. The dataset used to train this model is very large with millions of data points and is hosted in an Amazon S3 bucket. The Specialist wants to avoid loading all of this data onto an Amazon SageMaker notebook instance because it would take hours to move and will exceed the attached 5 GB Amazon EBS volume on the notebook instance.

Which approach allows the Specialist to use all the data to train the model?

- A. Load a smaller subset of the data into the SageMaker notebook and train locally. Confirm that the training code is executing and the model parameters seem reasonable. Initiate a SageMaker training job using the full dataset from the S3 bucket using Pipe input mode.
- B. Launch an Amazon EC2 instance with an AWS Deep Learning AMI and attach the S3 bucket to the instance. Train on a small amount of the data to verify the training code and hyperparameters. Go back to Amazon SageMaker and train using the full dataset
- C. Use AWS Glue to train a model using a small subset of the data to confirm that the data will be compatible with Amazon SageMaker. Initiate a SageMaker training job using the full dataset from the S3 bucket using Pipe input mode.
- D. Load a smaller subset of the data into the SageMaker notebook and train locally. Confirm that the training code is executing and the model parameters seem reasonable. Launch an Amazon EC2 instance with an AWS Deep Learning AMI and attach the S3 bucket to train the full dataset.

Correct Answer: A

- 🗨️ **JayK** Highly Voted 1 year, 5 months ago
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
upvoted 18 times
- 🗨️ **liangfb** Highly Voted 1 year, 6 months ago
Answer is A.
upvoted 12 times
- 🗨️ **cloud_trail** Most Recent 5 months ago
Gotta be A. You need to use Pipe mode but Glue cannot train a model.
upvoted 2 times
- 🗨️ **bobdylan1** 6 months ago
AAAAAAAAAAa
upvoted 1 times
- 🗨️ **WillNguyen22** 10 months ago
ans is A
upvoted 1 times
- 🗨️ **GeeBeeEI** 11 months ago
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 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/>
upvoted 3 times
- 🗨️ **roytruong** 1 year, 1 month ago
go for A
upvoted 2 times
- 🗨️ **PRC** 1 year, 3 months ago
Agree with A.
upvoted 2 times
- 🗨️ **cybe001** 1 year, 5 months ago
A is correct
upvoted 4 times
- 🗨️ **PhilipAWS** 1 year, 6 months ago
Answer is D not A
upvoted 2 times
- 🗨️ **hailiang** 9 months, 2 weeks ago
oh, come on
upvoted 7 times
- 🗨️ **cmm103** 1 year, 7 months ago
<https://aws.amazon.com/blogs/machine-learning/using-pipe-input-mode-for-amazon-sagemaker-algorithms/>
upvoted 9 times

Question #10

Topic 1

A Machine Learning Specialist has completed a proof of concept for a company using a small data sample, and now the Specialist is ready to implement an end-to-end solution in AWS using Amazon SageMaker. The historical training data is stored in Amazon RDS. Which approach should the Specialist use for training a model using that data?

- A. Write a direct connection to the SQL database within the notebook and pull data in
- B. Push the data from Microsoft SQL Server to Amazon S3 using an AWS Data Pipeline and provide the S3 location within the notebook.
- C. Move the data to Amazon DynamoDB and set up a connection to DynamoDB within the notebook to pull data in.
- D. Move the data to Amazon ElastiCache using AWS DMS and set up a connection within the notebook to pull data in for fast access.

Correct Answer: B

  **JayK** Highly Voted 1 year, 5 months ago



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

upvoted 15 times

  **jasonsunbao** 1 year, 5 months ago

I agree. As from the ML developer guide I just read, it is the MySQL RDS that can be used as SQL datasource.

upvoted 2 times

  **mlyu** 1 year, 5 months ago

<https://www.slideshare.net/AmazonWebServices/train-models-on-amazon-sagemaker-using-data-not-from-amazon-s3-aim419-aws-reinvent-2018>

upvoted 2 times

  **HaiHN** 8 months ago

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 answer should be B.

upvoted 1 times

  **SophieSu** Most Recent 4 months, 1 week ago

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."

upvoted 1 times

🗨️ **cnethers** 5 months, 1 week ago

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/>

upvoted 2 times

🗨️ **gcpwhiz** 2 months, 2 weeks ago

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

upvoted 1 times

🗨️ **scuzzy2010** 11 months ago

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

upvoted 1 times

🗨️ **roytruong** 1 year, 1 month ago

B is right

upvoted 2 times

🗨️ **bhavesh0124** 1 year, 4 months ago

B it is

upvoted 1 times

Question #11

Topic 1

A Machine Learning Specialist receives customer data for an online shopping website. The data includes demographics, past visits, and locality information. The

Specialist must develop a machine learning approach to identify the customer shopping patterns, preferences, and trends to enhance the website for better service and smart recommendations.

Which solution should the Specialist recommend?

- A. Latent Dirichlet Allocation (LDA) for the given collection of discrete data to identify patterns in the customer database.
- B. A neural network with a minimum of three layers and random initial weights to identify patterns in the customer database.
- C. Collaborative filtering based on user interactions and correlations to identify patterns in the customer database.
- D. Random Cut Forest (RCF) over random subsamples to identify patterns in the customer database.

Correct Answer: C

🗨️ **WWODIN** Highly Voted 1 year, 6 months ago

answer should be C

Collaborative filtering is for recommendation, LDA is for topic modeling

upvoted 13 times

  **syu31svc** Most Recent 10 months ago



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

upvoted 4 times

  **roytruong** 1 year, 1 month ago

this is C

upvoted 1 times

  **sdsfsdsf** 1 year, 3 months ago

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)

upvoted 3 times

  **cybe001** 1 year, 5 months ago

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

upvoted 3 times

  **cybe001** 1 year, 5 months ago

Collaborative filtering is appropriate

upvoted 3 times

  **DonaldCMLIN** 1 year, 7 months ago

Answer A might be more suitable than other

https://docs.aws.amazon.com/zh_tw/sagemaker/latest/dg/lda-how-it-works.html

upvoted 2 times

  **C10ud9** 1 year ago


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

upvoted 1 times

  **C10ud9** 1 year ago

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

upvoted 1 times

  **rsimham** 1 year, 6 months ago

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/>

upvoted 9 times

Question #12

Topic 1

A Machine Learning Specialist is working with a large company to leverage machine learning within its products. The company wants to group its customers into categories based on which customers will and will not churn within the next 6 months. The company has labeled the data available to the Specialist.

Which machine learning model type should the Specialist use to accomplish this task?

- A. Linear regression
- B. Classification
- C. Clustering
- D. Reinforcement learning

Correct Answer: B

The goal of classification is to determine to which class or category a data point (customer in our case) belongs to. For classification problems, data scientists would use historical data with predefined target variables AKA labels (churner/non-churner) "" answers that need to be predicted "" to train an algorithm. With classification, businesses can answer the following questions:

- ☞ Will this customer churn or not?
- ☞ Will a customer renew their subscription?
- ☞ Will a user downgrade a pricing plan?
- ☞ Are there any signs of unusual customer behavior?

Reference:

<https://www.kdnuggets.com/2019/05/churn-prediction-machine-learning.html>

🗲️ 👤 **rsimham** Highly Voted 👍 1 year, 6 months ago
B seems to be okay
upvoted 8 times

🗲️ 👤 **FabG** Most Recent ⌚ 7 months, 3 weeks ago
B - it's a Binary Classification problem. Will the customer churn: Yes or No
upvoted 2 times

🗲️ 👤 **syu31svc** 10 months, 1 week ago
100% is B since it is about labelled data
upvoted 1 times

🗲️ 👤 **ejj** 11 months, 4 weeks ago
i think the key is "the company has labeled the data" so this is classification, so it's B
upvoted 1 times

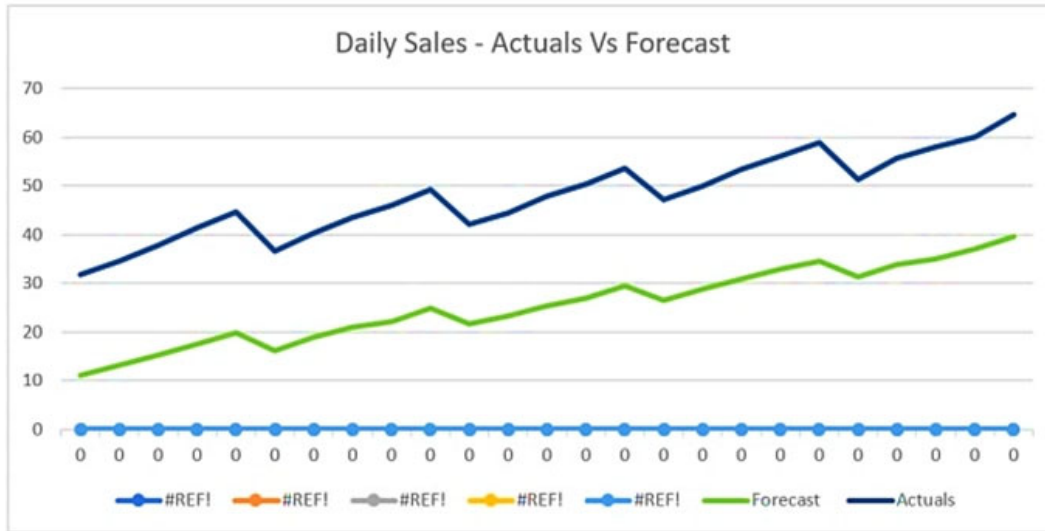
🗲️ 👤 **roytruong** 1 year, 1 month ago
B is okey
upvoted 2 times

🗲️ 👤 **cybe001** 1 year, 5 months ago
B is correct
upvoted 3 times

Question #13

Topic 1

The displayed graph is from a forecasting model for testing a time series.



Considering the graph only, which conclusion should a Machine Learning Specialist make about the behavior of the model?

- A. The model predicts both the trend and the seasonality well
- B. The model predicts the trend well, but not the seasonality.
- C. The model predicts the seasonality well, but not the trend.
- D. The model does not predict the trend or the seasonality well.

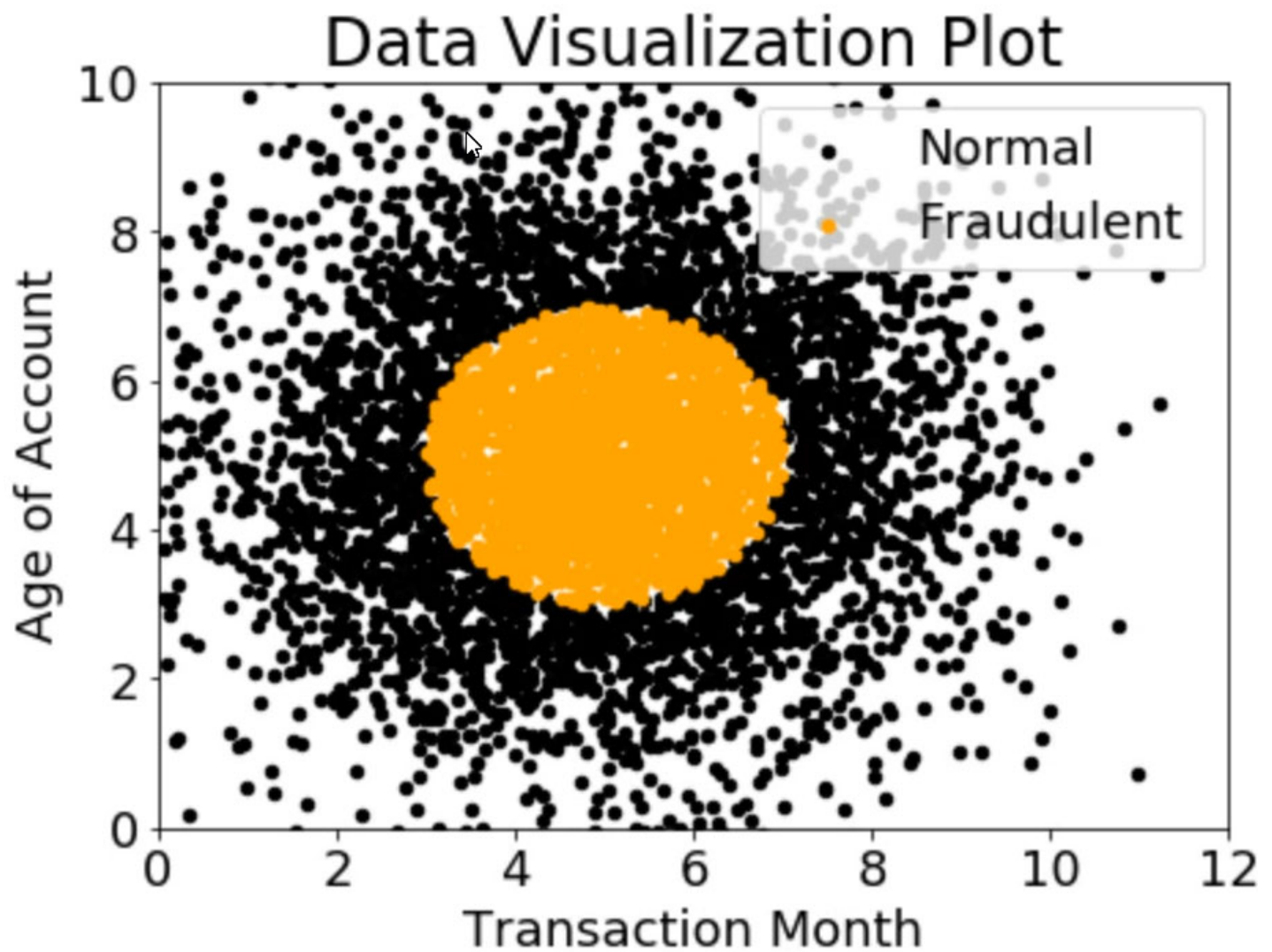
Correct Answer: A

- 🗨️ 👤 **btsql** 4 weeks ago
C is correct answer
upvoted 1 times
- 🗨️ 👤 **btsql** 3 weeks, 6 days ago
A is correct answer. Not C
upvoted 1 times
- 🗨️ 👤 **Kuntazulu** 1 month ago
The trend is up, so isn't it correctly predicted? And the seasonality is also in sync, the amplitude is wrong.
upvoted 1 times
- 🗨️ 👤 **jeetss1** 1 month, 2 weeks ago
A is correct answer.
Please Refer: <https://machinelearningmastery.com/decompose-time-series-data-trend-seasonality/>
upvoted 3 times
- 🗨️ 👤 **georschi** 2 months, 1 week ago
A is right. trend and seasonality are fine, level is the one the model gets wrong
upvoted 1 times
- 🗨️ 👤 **NotAnMLProfessional** 3 months, 2 weeks ago
Should be C
upvoted 1 times
- 🗨️ 👤 **ashlash** 4 months, 1 week ago
Should be A
upvoted 2 times

Question #14

Topic 1

A company wants to classify user behavior as either fraudulent or normal. Based on internal research, a Machine Learning Specialist would like to build a binary classifier based on two features: age of account and transaction month. The class distribution for these features is illustrated in the figure provided.



Based on this information, which model would have the HIGHEST accuracy?

- A. Long short-term memory (LSTM) model with scaled exponential linear unit (SELU)
- B. Logistic regression
- C. Support vector machine (SVM) with non-linear kernel
- D. Single perceptron with tanh activation function

Correct Answer: C

 **mizuakari** Highly Voted 2 months, 4 weeks ago

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.

upvoted 6 times

Question #15

Topic 1

A Machine Learning Specialist at a company sensitive to security is preparing a dataset for model training. The dataset is stored in Amazon S3 and contains

Personally Identifiable Information (PII).

The dataset:

- ☞ Must be accessible from a VPC only.
- ☞ Must not traverse the public internet.

How can these requirements be satisfied?

- A. Create a VPC endpoint and apply a bucket access policy that restricts access to the given VPC endpoint and the VPC.
- B. Create a VPC endpoint and apply a bucket access policy that allows access from the given VPC endpoint and an Amazon EC2 instance.
- C. Create a VPC endpoint and use Network Access Control Lists (NACLs) to allow traffic between only the given VPC endpoint and an Amazon EC2 instance.
- D. Create a VPC endpoint and use security groups to restrict access to the given VPC endpoint and an Amazon EC2 instance

Correct Answer: A

 **rajs** Highly Voted 1 year, 4 months ago

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

upvoted 15 times

 **sdsfsdsf** Highly Voted 1 year, 3 months ago

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.

upvoted 6 times

🗨️ 👤 **achiko** Most Recent 3 months, 1 week ago

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>

upvoted 1 times

🗨️ 👤 **senseikimoji** 7 months ago

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

upvoted 1 times

🗨️ 👤 **cloud_trail** 5 months ago

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.

upvoted 1 times

🗨️ 👤 **yeetusdeleetus** 8 months ago

I also vote A.

upvoted 1 times

🗨️ 👤 **Thai_Xuan** 9 months ago

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>

upvoted 2 times

🗨️ 👤 **tff** 1 year ago

agree with A

upvoted 1 times

🗨️ 👤 **morara** 1 year ago

A is the answer. No need to bring in EC2 coz the requirements defined are between S3 and the VPC.

upvoted 1 times

🗨️ 👤 **roytruong** 1 year, 1 month ago

It is A

upvoted 1 times

🗨️ 👤 **ac427** 1 year, 3 months ago

Instances in the VPC access the VPC EndPoint, not the VPC itself. So A is poorly worded like lots of AWS questions.

upvoted 1 times

🗨️ 👤 **grandgale** 1 year, 4 months ago

A THE VPC access the s3 through the VPC endpoint without internet traffic.

upvoted 4 times

🗨️ 👤 **JayK** 1 year, 5 months ago

I stand corrected. The answer is B as A is talking about "restrict access to the VPC endpoint" which is not what we want

upvoted 3 times

🗨️ 👤 **JayK** 1 year, 6 months ago

Should be A. As the question is talking about Amazon S3 and A refers to the bucket policies that belong to S3. There is no mention of EC2 the question

upvoted 5 times

🗨️ 👤 **BigEv** 1 year, 5 months ago

I guess it talks about the ml instance spin up by SageMaker.

Question #16

Topic 1

During mini-batch training of a neural network for a classification problem, a Data Scientist notices that training accuracy oscillates. What is the MOST likely cause of this issue?

- A. The class distribution in the dataset is imbalanced.
- B. Dataset shuffling is disabled.
- C. The batch size is too big.
- D. The learning rate is very high.

Correct Answer: D

Reference:

<https://towardsdatascience.com/deep-learning-personal-notes-part-1-lesson-2-8946fe970b95>

  **gaku1016** Highly Voted 1 year, 4 months ago

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.

upvoted 8 times

  **ozan11** Highly Voted 1 year, 5 months ago

maybe D ?

Question #17

Topic 1

An employee found a video clip with audio on a company's social media feed. The language used in the video is Spanish. English is the employee's first language, and they do not understand Spanish. The employee wants to do a sentiment analysis.

What combination of services is the MOST efficient to accomplish the task?

- A. Amazon Transcribe, Amazon Translate, and Amazon Comprehend
- B. Amazon Transcribe, Amazon Comprehend, and Amazon SageMaker seq2seq
- C. Amazon Transcribe, Amazon Translate, and Amazon SageMaker Neural Topic Model (NTM)
- D. Amazon Transcribe, Amazon Translate and Amazon SageMaker BlazingText

Correct Answer: A

  **DonaldCMLIN** Highly Voted 1 year, 7 months ago



the MOST efficient means to you don't need to coding, building infra

All of services are managed by AWS is good,

Transcribe, Amazon Translate, and Amazon Comprehend

Answer is A

upvoted 30 times

  **WWODIN** 1 year, 6 months ago

Agree, Answer is A

upvoted 6 times

  **Phong** Highly Voted 1 year, 4 months ago

Definitely A.

upvoted 5 times

  **harmanbirstudy** Most Recent 5 months, 3 weeks ago

The Question/Answer 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 not help the employee as he doesn't know Spanish. So the translate 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 :)

upvoted 1 times

  **senseikimoji** 7 months ago

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

upvoted 3 times

- 🗲️ 👤 **ybad** 7 months, 1 week ago
A
because Comprehend can provide sentiment analysis
upvoted 2 times
- 🗲️ 👤 **FastTrack** 9 months, 3 weeks ago
A,
<https://aws.amazon.com/getting-started/hands-on/analyze-sentiment-comprehend/>
upvoted 4 times
- 🗲️ 👤 **syu31svc** 10 months, 1 week ago
Amazon Comprehend is needed for sure; answer is A 100%
upvoted 3 times
- 🗲️ 👤 **hans1234** 11 months ago
You actually do not need Translate. Comprehend can do sentiment analysis on spanish language. But in this case it is A.
upvoted 2 times
- 🗲️ 👤 **Aanish** 6 months, 1 week ago
You are right but since the employee doesn't know Spanish it would be good if he has an insight into how well the sentiments were captured by comprehend.
upvoted 2 times
- 🗲️ 👤 **Antriksh** 1 year ago
These answers are clearly misleading. The correct answer is A
upvoted 3 times
- 🗲️ 👤 **roytruong** 1 year, 1 month ago
it's A
upvoted 3 times
- 🗲️ 👤 **AKT** 1 year, 4 months ago
Answer is A.
upvoted 3 times
- 🗲️ 👤 **grandgale** 1 year, 4 months ago
should be D
<https://docs.aws.amazon.com/sagemaker/latest/dg/blazingtext.html>
upvoted 1 times
- 🗲️ 👤 **oji** 11 months, 4 weeks ago
A is more efficient because if using sagemaker we still need to build a model first, but amazon comprehend is pre-build model so A is more efficient. CMIIW
upvoted 6 times
- 🗲️ 👤 **h_sahu** 7 months, 2 weeks ago
It shouldn't be D as per me. Blazing text is used for classification. This can indirectly help in sentiment analysis no doubt. I believe A is answer. Because, transcribe will be used to convert speech to text, translate will be used to translate those texts received from transcript and Comprehend is a complete NLP tool from amazon out-of-the-box which is capable of sentiment analysis.
upvoted 1 times

A Machine Learning Specialist is packaging a custom ResNet model into a Docker container so the company can leverage Amazon SageMaker for training. The

Specialist is using Amazon EC2 P3 instances to train the model and needs to properly configure the Docker container to leverage the NVIDIA GPUs.

What does the Specialist need to do?

- A. Bundle the NVIDIA drivers with the Docker image.
- B. Build the Docker container to be NVIDIA-Docker compatible.
- C. Organize the Docker container's file structure to execute on GPU instances.
- D. Set the GPU flag in the Amazon SageMaker CreateTrainingJob request body.

Correct Answer: A

  **vetal** Highly Voted 1 year, 6 months ago

<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

upvoted 24 times

  **devsean** 1 year, 4 months ago

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

Question #19

Topic 1

A Machine Learning Specialist is building a logistic regression model that will predict whether or not a person will order a pizza. The Specialist is trying to build the optimal model with an ideal classification threshold.

What model evaluation technique should the Specialist use to understand how different classification thresholds will impact the model's performance?

- A. Receiver operating characteristic (ROC) curve
- B. Misclassification rate
- C. Root Mean Square Error (RMSE)
- D. L1 norm

Correct Answer: A

Reference:

<https://docs.aws.amazon.com/machine-learning/latest/dg/binary-model-insights.html>

  **rsimham** Highly Voted 1 year, 6 months ago

Ans. A is correct
upvoted 14 times

  **AKT** Highly Voted 1 year, 4 months ago



Answer is A.
An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds
upvoted 6 times

  **syu31svc** Most Recent 10 months, 1 week ago

Question is about classification so confusion matrix would come into mind; A is the answer
upvoted 1 times

  **GeeBeeEI** 11 months ago

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:
 $FPR = FP / (FP + TN)$
upvoted 2 times

  **hans1234** 11 months, 1 week ago

It is A.
upvoted 1 times

  **roytruong** 1 year, 1 month ago

obviously A
upvoted 1 times

Question #20

Topic 1

An interactive online dictionary wants to add a widget that displays words used in similar contexts. A Machine Learning Specialist is asked to provide word features for the downstream nearest neighbor model powering the widget.
What should the Specialist do to meet these requirements?

- A. Create one-hot word encoding vectors.
- B. Produce a set of synonyms for every word using Amazon Mechanical Turk.
- C. Create word embedding vectors that store edit distance with every other word.
- D. Download word embeddings pre-trained on a large corpus.

Correct Answer: A

Reference:

<https://aws.amazon.com/blogs/machine-learning/amazon-sagemaker-object2vec-adds-new-features-that-support-automatic-negative-sampling-and-speed-up-training/>

  **JayK** Highly Voted 1 year, 5 months ago

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 dictionary is unique and does not have a pre-trained word embedding.

Thus I strongly feel the answer is D

upvoted 16 times

  **cybe001** Highly Voted 1 year, 5 months ago

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

upvoted 11 times

  **engomaradel** Most Recent 6 months, 3 weeks ago



D for sure

upvoted 2 times

  **yeetusdeleetus** 8 months ago

Definitely D.

upvoted 2 times

  **weslleylc** 8 months, 2 weeks ago

A)It requires that document text be cleaned and prepared such that each word is one-hot encoded.

Ref:<https://machinelearningmastery.com/what-are-word-embeddings/>

upvoted 1 times

  **syu31svc** 10 months, 1 week ago

I don't see how one-hot encoding works; I would say D 100%

B & C are definitely wrong

upvoted 2 times

  **GeeBeeEI** 11 months ago

took a look at <https://medium.com/@athif.shaffy/one-hot-encoding-of-text-b69124bef0a7> it appears this is useful at the point of running the model. See: <https://arxiv.org/abs/1705.08488> Word embeddings are dense, low-dimensional vector representations of words that are commonly used as input features in a variety of natural language processing (NLP) tasks [1]. In contrast to symbolic one-hot or hierarchical clustering-based representations, real-valued embedding vectors easily reflect varying degrees of similarity between words, and significant reduce sparsity in linear algebra operations.

upvoted 1 times

  **GeeBeeEI** 11 months ago

Appears D is the correct answer....

upvoted 2 times

  **Achievement** 11 months, 2 weeks ago

I think should be A, we need to provide features for next step

upvoted 1 times

  **andreyjh** 1 year ago



Why not A since it is required to pass words as features to a nearest neighbor model?

upvoted 3 times

  **roytruong** 1 year, 1 month ago

absolutely D

upvoted 3 times

  **AKT** 1 year, 4 months ago

answer is D

upvoted 2 times

- 🗨️ 👤 **WWODIN** 1 year, 6 months ago
Seems D?
upvoted 5 times
- 🗨️ 👤 **WWODIN** 1 year, 6 months ago
sorry, seems A is a necessary first step
upvoted 2 times
- 🗨️ 👤 **WWODIN** 1 year, 5 months ago
sorry again, should be C or D, but more towards D
upvoted 2 times
- 🗨️ 👤 **HaiHN** 8 months, 2 weeks ago
C is irrelevant as it only concern about the "edit distance", not the meaning of the word for "used in similar context"

Correct answer should be D
upvoted 2 times
- 🗨️ 👤 **cloud_trail** 5 months ago
Correct. Edit distance or Levenshtein Distance concerns spelling and is used by spell checkers. Has nothing to do with context is the obvious answer. There's no need to create your own embedding when you can just download pre-trained ones.
upvoted 3 times
- 🗨️ 👤 **rsimham** 1 year, 6 months ago
C sounds to be right for me, not sure.
<https://medium.com/@athif.shaffy/one-hot-encoding-of-text-b69124bef0a7>
upvoted 2 times
- 🗨️ 👤 **vetal** 1 year, 6 months ago
Why not D?
upvoted 4 times

Question #21

Topic 1

A Machine Learning Specialist is configuring Amazon SageMaker so multiple Data Scientists can access notebooks, train models, and deploy endpoints. To ensure the best operational performance, the Specialist needs to be able to track how often the Scientists are deploying models, GPU and CPU utilization on the deployed SageMaker endpoints, and all errors that are generated when an endpoint is invoked.

Which services are integrated with Amazon SageMaker to track this information? (Choose two.)

- A. AWS CloudTrail
- B. AWS Health
- C. AWS Trusted Advisor
- D. Amazon CloudWatch
- E. AWS Config

Correct Answer: AD

Reference:

<https://aws.amazon.com/sagemaker/faqs/>

  **rsimham** Highly Voted 1 year, 6 months ago

AD is correct

upvoted 16 times

  **eji** Most Recent 11 months, 4 weeks ago

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

upvoted 2 times

  **Urban_Life** 1 year ago

absolutely

upvoted 1 times

  **roytruong** 1 year, 1 month ago

cloudtrail and cloudwatch, no thinking

upvoted 4 times

Question #22

Topic 1

A retail chain has been ingesting purchasing records from its network of 20,000 stores to Amazon S3 using Amazon Kinesis Data Firehose. To support training an improved machine learning model, training records will require new but simple transformations, and some attributes will be combined. The model needs to be retrained daily.

Given the large number of stores and the legacy data ingestion, which change will require the LEAST amount of development effort?

- A. Require that the stores to switch to capturing their data locally on AWS Storage Gateway for loading into Amazon S3, then use AWS Glue to do the transformation.

- B. Deploy an Amazon EMR cluster running Apache Spark with the transformation logic, and have the cluster run each day on the accumulating records in Amazon S3, outputting new/transformed records to Amazon S3.
- C. Spin up a fleet of Amazon EC2 instances with the transformation logic, have them transform the data records accumulating on Amazon S3, and output the transformed records to Amazon S3.
- D. Insert an Amazon Kinesis Data Analytics stream downstream of the Kinesis Data Firehose stream that transforms raw record attributes into simple transformed values using SQL.

Correct Answer: D

  **cybe001**  1 year, 5 months ago

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>

upvoted 24 times

  **kakalotka** 1 month, 3 weeks ago

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

upvoted 1 times

  **mawsmann** 1 year, 1 month ago

Best explanation here, kudos.

upvoted 3 times

  **PRC**  1 year, 3 months ago

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

upvoted 6 times

  **cloud_trail**  5 months ago

I go with D. A tough question, though. And C are definitely out. The 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.

upvoted 1 times

  **harmanbirstudy** 5 months, 3 weeks ago



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 to 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>

upvoted 2 times

  **gamaX** 6 months, 4 weeks ago

It seems that LEAST development effort:

<https://aws.amazon.com/fr/blogs/big-data/preprocessing-data-in-amazon-kinesis-analytics-with-aws-lambda/>

and GREATEST 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/>

upvoted 1 times

🗨️ 👤 **HaiHN** 8 months, 2 weeks ago
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."

upvoted 1 times

🗨️ 👤 **h_sahu** 7 months, 1 week ago

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 along with on-premise applications to move data to S3. Then Glue can be used to transform the data.

upvoted 1 times

🗨️ 👤 **cloud_trail** 5 months ago

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

upvoted 2 times

🗨️ 👤 **hans1234** 11 months, 1 week ago

It is D.

upvoted 1 times

🗨️ 👤 **dikers** 1 year ago

I think the answer is D, because it requires the LEAST amount of development effort.

upvoted 1 times

🗨️ 👤 **roytruong** 1 year, 1 month ago

It's D, Kinesis Analytics can easily connect with Firehose

upvoted 2 times

🗨️ 👤 **dreemswang** 1 year, 2 months ago

Why not A. It seems good to me

upvoted 1 times

🗨️ 👤 **ExamTaker123456789** 1 year ago

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

upvoted 1 times

🗨️ 👤 **devsean** 1 year, 4 months ago

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.

upvoted 2 times

🗨️ 👤 **hailiang** 9 months, 2 weeks ago

It's D, because with KDA you can transform the data with SQL while with EMR you need to write code, considering the requirement of "low development effort", so D

upvoted 2 times

🗨️ 👤 **devsean** 1 year, 4 months ago

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.

upvoted 4 times

🗨️ 👤 **HaiHN** 8 months, 2 weeks ago

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/>

upvoted 1 times

🗨️ 👤 **am7** 1 year, 5 months ago

can be B

upvoted 1 times

🗨️ 👤 **vetal** 1 year, 6 months ago

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

upvoted 2 times

Question #23

Topic 1

A Machine Learning Specialist is building a convolutional neural network (CNN) that will classify 10 types of animals. The Specialist has built a series of layers in a neural network that will take an input image of an animal, pass it through a series of convolutional and pooling layers, and then finally pass it through a dense and fully connected layer with 10 nodes. The Specialist would like to get an output from the neural network that is a probability distribution of how likely it is that the input image belongs to each of the 10 classes.

Which function will produce the desired output?

- A. Dropout
- B. Smooth L1 loss
- C. Softmax
- D. Rectified linear units (ReLU)

Correct Answer: D

Reference:

<https://towardsdatascience.com/building-a-convolutional-neural-network-cnn-in-keras-329fbbadc5f5>

🗨️ **DonaldCMLIN** Highly Voted 1 year, 7 months ago

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>

upvoted 20 times

🗨️ **rsimham** Highly Voted 1 year, 6 months ago

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

upvoted 11 times

🗨️ **takahirokoyama** Most Recent 4 months, 4 weeks ago

Absolute C.

upvoted 3 times

🗨️ **cloud_trail** 5 months ago

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

upvoted 2 times

🗨️ **felbuch** 5 months, 1 week ago

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)

upvoted 3 times

🗨️ **MOMoez** 6 months, 1 week ago

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.

upvoted 1 times

🗨️ **ybad** 7 months, 1 week ago

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>

upvoted 1 times

🗨️ **yeetusdeleetus** 8 months ago

Definitely C

upvoted 1 times

🗨️ **bids** 10 months, 1 week ago

Definitely softmax.

upvoted 1 times

- 🗨️ **hans1234** 11 months, 1 week ago
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.
upvoted 1 times
- 🗨️ **hans1234** 11 months ago
Yes, softmax returns indeed a vector of probabilities.
upvoted 1 times
- 🗨️ **roytruong** 1 year, 1 month ago
C, everyone with basic knowledge in neural network can easily see that
upvoted 2 times
- 🗨️ **deep_n** 1 year, 1 month ago
Not even debatable!
C!!!!
upvoted 3 times
- 🗨️ **jasonsunbao** 1 year, 5 months ago
It is softmax. ReLu is used for activation, not for classification
upvoted 1 times
- 🗨️ **oji** 11 months, 4 weeks ago
softmax is also activation function
upvoted 1 times
- 🗨️ **BigEv** 1 year, 5 months ago
Ans is C
Pls Check this out
<https://github.com/Kulbear/deep-learning-nano-foundation/wiki/ReLU-and-Softmax-Activation-Functions>
upvoted 1 times

Question #24

Topic 1

A Machine Learning Specialist trained a regression model, but the first iteration needs optimizing. The Specialist needs to understand whether the model is more frequently overestimating or underestimating the target.

What option can the Specialist use to determine whether it is overestimating or underestimating the target value?

A. Root Mean Square Error (RMSE)

- B. Residual plots
- C. Area under the curve
- D. Confusion matrix

Correct Answer: B

🗲️ **vetal** Highly Voted 1 year, 6 months ago

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

upvoted 20 times

🗲️ **rsimham** Highly Voted 1 year, 6 months ago

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

upvoted 11 times

🗲️ **felbuch** Most Recent 5 months, 1 week ago

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.

upvoted 3 times

🗲️ **cnethers** 4 months, 4 weeks ago

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 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.

upvoted 1 times

🗲️ **cnethers** 4 months, 3 weeks ago

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

upvoted 1 times

🗲️ **Thai_Xuan** 9 months ago

residual plot

<https://docs.aws.amazon.com/machine-learning/latest/dg/regression-model-insights.html>

upvoted 2 times

🗲️ **syu31svc** 10 months ago

[https://stattrek.com/statistics/dictionary.aspx?](https://stattrek.com/statistics/dictionary.aspx?definition=residual%20plot#:~:text=A%20residual%20plot%20is%20a,nonlinear%20model%20is%20more%20appropriate.)

definition=residual%20plot#:~:text=A%20residual%20plot%20is%20a,nonlinear%20model%20is%20more%20appropriate.

Answer is B

upvoted 1 times


🗲️ **Antriksh** 1 year ago

without a second thought residual plot

upvoted 2 times

- 🗳️ 👤 **qururu** 1 year ago
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](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)
upvoted 1 times
- 🗳️ 👤 **C10ud9** 1 year ago
Residual plot it is Option B
upvoted 1 times
- 🗳️ 👤 **roytruong** 1 year, 1 month ago
Residual plot
upvoted 2 times
- 🗳️ 👤 **deep_n** 1 year, 1 month ago
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
upvoted 4 times
- 🗳️ 👤 **swagy** 1 year, 2 months ago
It should be B. The residual plot will be give whether the target value is overestimated or underestimated.
upvoted 1 times
- 🗳️ 👤 **Jayraam** 1 year, 2 months ago
Answer is C.

<https://www.youtube.com/watch?v=MrjWcywVEiU>
upvoted 2 times
- 🗳️ 👤 **ExamTaker123456789** 1 year ago
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.
upvoted 2 times
- 🗳️ 👤 **cloud_trail** 5 months ago
Good grief, AUC is used for classification not regression.
upvoted 1 times
- 🗳️ 👤 **PRC** 1 year, 3 months ago
B..Residual helps to find out whether the model is underestimating or overestimating
upvoted 3 times
- 🗳️ 👤 **AKT** 1 year, 4 months ago
answer is B
upvoted 2 times
- 🗳️ 👤 **Phong** 1 year, 4 months ago
Go for B. Residual spots can handle this problem
upvoted 4 times
- 🗳️ 👤 **jasonsunbao** 1 year, 5 months ago
It is B residual error for sure
upvoted 5 times

 **BigEv** 1 year, 5 months ago

I choose B:

Root Mean Square Error (RMSE) is the standard deviation of the residuals.

Since the question is about whether the model is more frequently overestimating or underestimating the target, you should plot the residual

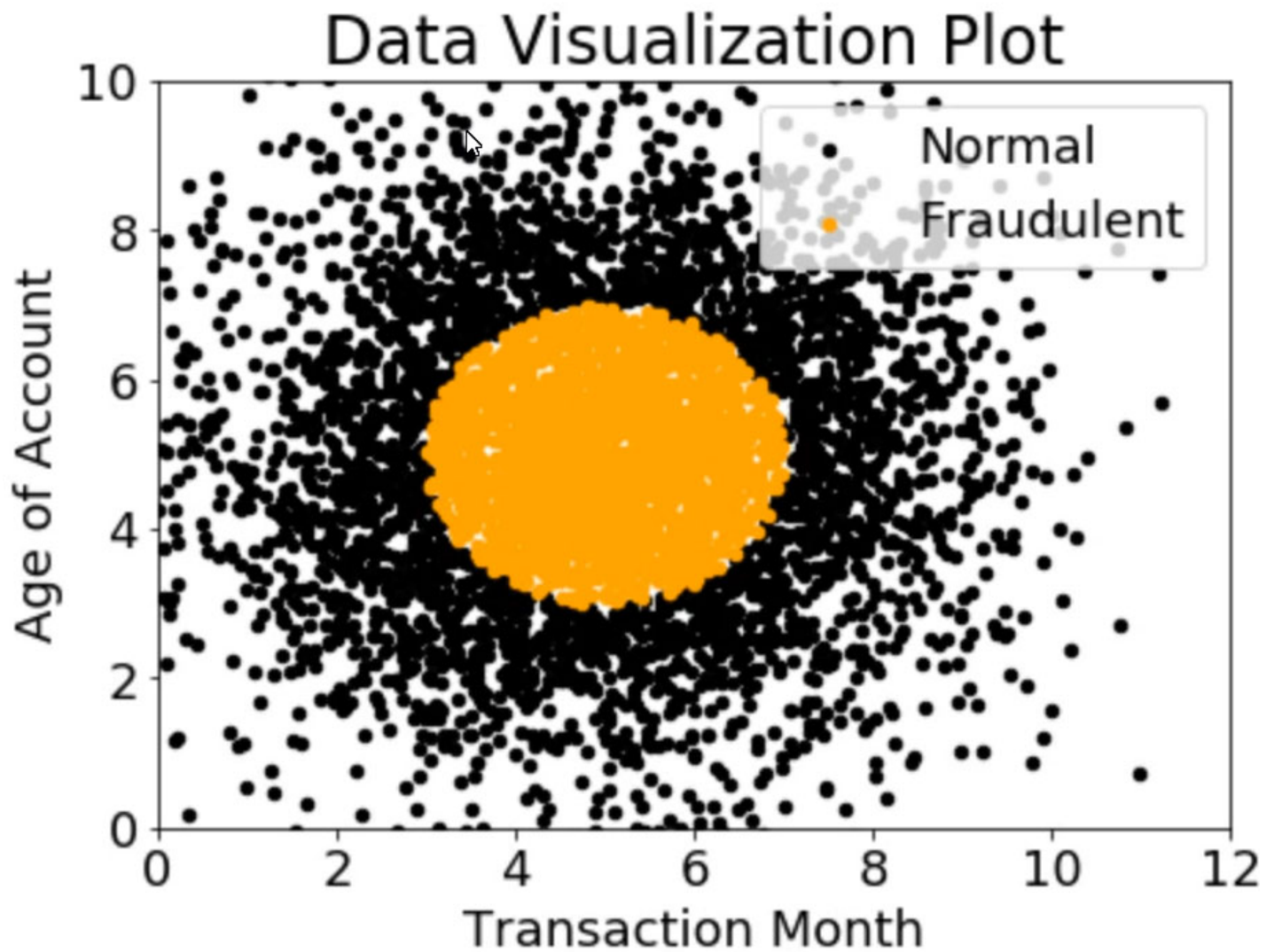
<https://www.statisticshowto.datasciencecentral.com/rmse/>

upvoted 6 times

Question #25

Topic 1

A company wants to classify user behavior as either fraudulent or normal. Based on internal research, a Machine Learning Specialist would like to build a binary classifier based on two features: age of account and transaction month. The class distribution for these features is illustrated in the figure provided.



Based on this information, which model would have the HIGHEST recall with respect to the fraudulent class?

- A. Decision tree
- B. Linear support vector machine (SVM)
- C. Naive Bayesian classifier
- D. Single Perceptron with sigmoidal activation function

Correct Answer: C

INNN 6 days, 1 hour ago












Answer is correct. It has to be C. Check out there:

<https://stackoverflow.com/questions/21468469/logistic-regression-and-naive-bayes-for-this-dataset>

The variance in yellow is smaller and bigger in black. NB can make a perfect prediction.

On the other hand, tree models are known for inefficiency on round decision boundary, since you have to get thousand small rectangles to approximate the round shape.

upvoted 1 times

-  **blubb** 1 month ago
 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.
 upvoted 2 times
-  **yummytaco** 1 month, 2 weeks ago
 I think it's decision tree since it's a non linear and Naive Bayes is a linear classifier based on:
<https://www.cs.cornell.edu/courses/cs4780/2018fa/lectures/lecturenote05.html>
 upvoted 2 times
-  **Vita_Rasta84444** 3 months ago
 It is B. Only SVM could create the most non-linear bound between fraudulent and non fraudulent and hence achieve the highest recall.
 upvoted 1 times
-  **jerto97** 2 weeks, 3 days ago
 No, you'd need a kernel function to classify that boundary
 upvoted 1 times
-  **gcpwhiz** 2 months, 2 weeks ago
 B is linear SVM. To use SVM here you need non-linear w hyperplane
 upvoted 4 times
-  **achiko** 3 months, 1 week ago
 why not a decision tree? its also non linear
 upvoted 3 times
-  **SophieSu** 4 months, 1 week ago
 The point of this question is "Non-Linear" clustering. Naive Bayes classifier in general is not linear.
 upvoted 3 times
-  **cnethers** 4 months, 4 weeks ago
 Real time Prediction: Naive Bayes is an eager learning classifier and it is sure fast. Thus, it could be used for making predictions in real time
 Multi class Prediction: This algorithm is also well known for multi class prediction feature. Here we can predict the probability of multiple classes of target variable.
 Text classification/ Spam Filtering/ Sentiment Analysis: Naive Bayes classifiers mostly used in text classification (due to better result in multi class problems and independence rule) have higher success rate as compared to other algorithms. As a result, it is widely used in Spam filtering (identify spam e-mail) and Sentiment Analysis (in social media analysis, to identify positive and negative customer sentiments)
 upvoted 1 times
-  **cnethers** 4 months, 4 weeks ago
 Naive Bayes Algorithms
 If categorical variable has a category (in test data set), which was not observed in training data set, then model will assign a 0 (zero) probab and will be unable to make a prediction. This is often known as "Zero Frequency". To solve this, we can use the smoothing technique. One the simplest smoothing techniques is called Laplace estimation.
 On the other side naive Bayes is also known as a bad estimator, so the probability outputs from predict_proba are not to be taken too seriously.
 Another limitation of Naive Bayes is the assumption of independent predictors. In real life, it is almost impossible that we get a set of predictors which are completely independent.
 upvoted 1 times
-  **cnethers** 4 months, 4 weeks ago
 If SVM was set to use non linear kernel then I would have to argue that would be the best option for a binary classification issue
<https://uk.mathworks.com/discovery/support-vector-machine.html>
 upvoted 2 times
-  **garud** 1 month, 2 weeks ago
 So what is your final verdict as far as the answer goes?

A Machine Learning Specialist kicks off a hyperparameter tuning job for a tree-based ensemble model using Amazon SageMaker with Area Under the ROC Curve

(AUC) as the objective metric. This workflow will eventually be deployed in a pipeline that retrains and tunes hyperparameters each night to model click-through on data that goes stale every 24 hours.

With the goal of decreasing the amount of time it takes to train these models, and ultimately to decrease costs, the Specialist wants to reconfigure the input hyperparameter range(s).

Which visualization will accomplish this?

- A. A histogram showing whether the most important input feature is Gaussian.
- B. A scatter plot with points colored by target variable that uses t-Distributed Stochastic Neighbor Embedding (t-SNE) to visualize the large number of input variables in an easier-to-read dimension.
- C. A scatter plot showing the performance of the objective metric over each training iteration.
- D. A scatter plot showing the correlation between maximum tree depth and the objective metric.

Correct Answer: B

🗨️ **cloud_trail** Highly Voted 5 months ago

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 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 only one of them.

upvoted 11 times

🗨️ **cnethers** 4 months, 4 weeks ago

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.

upvoted 2 times

🗨️ **heihei** Highly Voted 1 year, 6 months ago

B doesn't make sense

I think it's D

upvoted 8 times

🗨️ **SophieSu** Most Recent 4 months, 1 week ago

Focus on 2 points: Tree Model and Hyperparameter tuning. Definitely D

upvoted 2 times

🗨️ **harmanbirstudy** 5 months, 3 weeks ago

I think its A ..

<https://www.analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning-xgboost-with-codes-python/>

upvoted 1 times

🗨️ 👤 **ybad** 7 months, 1 week ago

B is right...

it actually does make sense to have B, it is used for dimensional reductions, but it does this by properly understanding the parameters of the model and visualizing the multi dimensional data into clusters, thus understanding what the model is doing to optimize itself,

Also, reducing the dimensionality of a model allows it to train faster, thus reducing the cost, which works for the required goal...

https://en.wikipedia.org/wiki/T-distributed_stochastic_neighbor_embedding

upvoted 1 times

🗨️ 👤 **yeetusdeleetus** 8 months ago

Operator wants to decrease hyperparameter ranges.

Only D involves visualizing a hyperparameter, which could be used to change the ranges given in grid search of, e.g., 90% of the given values are always bad.

upvoted 1 times

🗨️ 👤 **Thai_Xuan** 9 months ago

D

<https://medium.com/all-things-ai/in-depth-parameter-tuning-for-random-forest-d67bb7e920d>

upvoted 1 times

🗨️ 👤 **Thai_Xuan** 9 months ago

D, as is seen here

<https://medium.com/all-things-ai/in-depth-parameter-tuning-for-gradient-boosting-3363992e9bae>

upvoted 1 times

🗨️ 👤 **syu31svc** 10 months ago

A is wrong for sure; makes no sense at all

t-Distributed Stochastic Neighbor Embedding (t-SNE) is a technique for dimensionality reduction that is particularly well suited for the visualization of high-dimensional datasets so B is wrong.

Between C and D, I would say D as the assumption is that it would save time to set the number of training iterations and see what the max depth should be

upvoted 1 times

🗨️ 👤 **roytruong** 1 year, 1 month ago



go for D

upvoted 1 times

🗨️ 👤 **zxl** 1 year, 1 month ago

Answer is C here. With the plot, you can identify a group of iterations with highest AUC.

upvoted 3 times

  **rajs** 1 year, 4 months ago

B does not answer the question though what you are saying is correct

Question is asking how the operator can minimize the training time....the operator can control the Hyperparameter Range & Objective a scatter plot will give an idea on both and since D covers both aspects

D is the answer

upvoted 6 times

  **georgeZ** 1 year, 3 months ago

Why not C? Checking and optimize the epoch number I will reduce the training time more efficiently.

upvoted 2 times

  **cloud_trail** 5 months ago

The question specifically asks how to reconfigure hyperparameter range, not reduce training time. Always read the question very

Question #27

Topic 1

A Machine Learning Specialist is creating a new natural language processing application that processes a dataset comprised of 1 million sentences. The aim is to then run Word2Vec to generate embeddings of the sentences and enable different types of predictions.

Here is an example from the dataset:

"The quck BROWN FOX jumps over the lazy dog."

Which of the following are the operations the Specialist needs to perform to correctly sanitize and prepare the data in a repeatable manner?

(Choose three.)

- A. Perform part-of-speech tagging and keep the action verb and the nouns only.
- B. Normalize all words by making the sentence lowercase.
- C. Remove stop words using an English stopword dictionary.
- D. Correct the typography on "quck" to "quick."
- E. One-hot encode all words in the sentence.
- F. Tokenize the sentence into words.

Correct Answer: BCF

  **ozan11**  1 year, 5 months ago

B C F should be correct.

upvoted 24 times

  **BigEv**  1 year, 5 months ago

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>

upvoted 16 times

  **SophieSu**  4 months, 1 week ago

BCF correct. D is not correct (Pay attention to "in a repeatable manner" in the question.)

upvoted 1 times

- 🗳️ 👤 **cloud_trail** 5 months ago
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 have a typo.
upvoted 1 times
- 🗳️ 👤 **harmanbirstudy** 5 months, 3 weeks ago
I saw this exact question on "whizlabs" practice exam and correct options were B/C/F
upvoted 1 times
- 🗳️ 👤 **GeeBeeEI** 11 months ago
<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
upvoted 5 times
- 🗳️ 👤 **cnethers** 5 months, 1 week ago
BCF is 100% correct
upvoted 2 times
- 🗳️ 👤 **Antriksh** 1 year ago
Correct answers are B, C and F
upvoted 2 times
- 🗳️ 👤 **TuanAnh** 1 year, 1 month ago
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
upvoted 4 times
- 🗳️ 👤 **TuanAnh** 1 year, 1 month ago
sorry E: word2vec can use CBOW or skipgram, so no need to have one-hot decoding here
upvoted 4 times
- 🗳️ 👤 **PRC** 1 year, 3 months ago
BCF is correct
upvoted 2 times
- 🗳️ 👤 **AKT** 1 year, 4 months ago
B, C F correct
upvoted 2 times
- 🗳️ 👤 **Phong** 1 year, 4 months ago
B, C, and F are correct answers. I have done this question many times in many practice tests.
upvoted 12 times
- 🗳️ 👤 **tap123** 1 year, 5 months ago
B, C, F are my choice. D is also possible but not as widely used as others.
upvoted 3 times

  **cybe001** 1 year, 5 months ago

Why C is not included in the answer? ABCD, all are correct answers
upvoted 1 times

  **halfway** 1 year ago

"choose three"
upvoted 1 times

Question #28

Topic 1

A company is using Amazon Polly to translate plaintext documents to speech for automated company announcements. However, company acronyms are being mispronounced in the current documents.

How should a Machine Learning Specialist address this issue for future documents?

- A. Convert current documents to SSML with pronunciation tags.
- B. Create an appropriate pronunciation lexicon.
- C. Output speech marks to guide in pronunciation.
- D. Use Amazon Lex to preprocess the text files for pronunciation

Correct Answer: A

Reference:

<https://docs.aws.amazon.com/polly/latest/dg/ssml.html>

  **VB**  1 year, 3 months ago

SSML is specific to that particular document, like W3C can 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.

upvoted 20 times

  **cloud_trail** 5 months ago

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.

upvoted 2 times

  **cybe001**  1 year, 5 months ago

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/>



upvoted 12 times

  **AShahine21** Most Recent 1 month ago

It should be B based on: <https://docs.aws.amazon.com/polly/latest/dg/managing-lexicons.html>

1. it will solve the current problem
2. it will be helpful for Future documents

upvoted 2 times



  **j2_me** 2 months, 1 week ago

<https://gigaom.com/2017/04/24/give-your-products-the-power-of-speech-using-amazon-polly/>

A is the correct answer , reasons

- 1) convert to SSML document
- 2) Associate custom pronunciation lexicons to SSML

upvoted 1 times

  **gcpwhiz** 2 months, 1 week ago



Answer is B: <https://docs.aws.amazon.com/polly/latest/dg/managing-lexicons.html>

upvoted 2 times

  **SophieSu** 4 months, 1 week ago

Acronyms - use SSML; Foreign accent such as Indian Accent - use lexicons. The point of this question is absolutely all about acronyms.

upvoted 1 times

  **seanLu** 4 months, 1 week ago

Is the solution provided by this website the actual solution? Because if previous discussions are true, then the A option is not correct. I found several cases where provided solutions do not match with the discussion.



upvoted 1 times

  **harmanbirstudy** 5 months, 3 weeks ago

Lexicon and SSML can both achieve this result ... but the last statement "address this issue for future documents" is the reason we should choose lexicon as SSML tags will only be applied in the current doc like `_{W3C}` , but if you want to make it work for future document so it has to be to lexicon.

SO ANSWER is B.

upvoted 5 times

  **cnethers** 4 months, 2 weeks ago

100% Agree with B (good rational)

Some docs:

""https://docs.amazonaws.cn/en_us/polly/latest/dg/managing-lexicons.html""

""https://docs.amazonaws.cn/en_us/polly/latest/dg/ssml.html""

upvoted 1 times

  **MennaMN** 10 months ago

I believe it's B as using lexicons enables doing customization to specific words/acronyms to how they will be pronounced instead.

Example: "World Wide Web Consortium" instead of "W3C"

upvoted 1 times

  **syu31svc** 10 months, 1 week ago

I would say B

Pronunciation lexicons enable you to customize the pronunciation of words

From <https://docs.aws.amazon.com/polly/latest/dg/managing-lexicons.html>

upvoted 2 times

  **scuzzy2010** 10 months, 4 weeks ago

I think answer should be B. From "<https://docs.aws.amazon.com/polly/latest/dg/managing-lexicons.html>" - Common words are sometimes stylized with numbers taking the place of letters, as with "g3t sm4rt" (get smart). Humans can read these words correctly. However, a Text-to-Speech (TTS) engine reads the text literally, pronouncing the name exactly as it is spelled. This is where you can leverage lexicons to customize the synthesized speech by using Amazon Polly. In this example, you can specify an alias (get smart) for the word "g3t sm4rt" in a lexicon.

upvoted 2 times

🗨️ **hans1234** 11 months ago

It is B.

A SSML document includes the text/document itself and adds pronunciation tags with additional info.

A lexicon is an additional consistent document that can be added to new text documents.

upvoted 2 times

🗨️ **Urban_Life** 12 months ago

Can some one explain why the answer is A? I also think answer will be B.

upvoted 1 times

🗨️ **Antriksh** 1 year ago

B is correct answer

upvoted 4 times

🗨️ **C10ud9** 1 year ago

business use case is to use plain text documents, so pronunciation lexicons(B) should solve the issue. converting documents of current and future documents to SSML documents with pronunciation tags works too, but not is not correctly given in the choices. The choices refer to only current documents.

upvoted 2 times

🗨️ **roytruong** 1 year, 1 month ago

this is A

upvoted 1 times

🗨️ **Phong** 1 year, 4 months ago

Go for A.

upvoted 3 times

Question #29

Topic 1

An insurance company is developing a new device for vehicles that uses a camera to observe drivers' behavior and alert them when they appear distracted. The company created approximately 10,000 training images in a controlled environment that a Machine Learning Specialist will use to train and evaluate machine learning models.

During the model evaluation, the Specialist notices that the training error rate diminishes faster as the number of epochs increases and the model is not accurately inferring on the unseen test images.

Which of the following should be used to resolve this issue? (Choose two.)


- A. Add vanishing gradient to the model.
- B. Perform data augmentation on the training data.
- C. Make the neural network architecture complex.
- D. Use gradient checking in the model.
- E. Add L2 regularization to the model.

Correct Answer: BE

🗲️  **vetal** Highly Voted 🏆 1 year, 6 months ago

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.

upvoted 20 times

🗲️  **rajs** 1 year, 4 months ago

Agreed 100%

upvoted 5 times

🗲️  **jasonsunbao** 1 year, 5 months ago

agree on BE

upvoted 3 times

🗲️  **benson2021** Most Recent 🕒 2 months, 3 weeks ago

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 augmentation
4. Use regularization
5. Use dropouts

upvoted 1 times

🗲️  **engomaradel** 6 months, 3 weeks ago

B & E is the correct ans

upvoted 1 times

🗲️  **roytruong** 1 year, 1 month ago


BE is exact

upvoted 3 times

🗲️  **stamarpadar** 1 year, 1 month ago

BE are the correct answers

upvoted 4 times

🗲️  **VB** 1 year, 2 months ago

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

upvoted 3 times

🗲️  **C10ud9** 1 year ago

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. would say B, E are correct

upvoted 2 times

🗲️  **VB** 1 year, 2 months ago

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

upvoted 2 times

🗲️  **hughhughhugh** 1 year, 2 months ago

why not because of vanishing gradient?

upvoted 1 times

🗲️  **PRC** 1 year, 3 months ago

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

upvoted 1 times

🗨️ 👤 **WWODIN** 1 year, 5 months ago
agree BE
upvoted 3 times

Question #30

Topic 1

When submitting Amazon SageMaker training jobs using one of the built-in algorithms, which common parameters MUST be specified? (Choose three.)

- A. The training channel identifying the location of training data on an Amazon S3 bucket.
- B. The validation channel identifying the location of validation data on an Amazon S3 bucket.
- C. The IAM role that Amazon SageMaker can assume to perform tasks on behalf of the users.
- D. Hyperparameters in a JSON array as documented for the algorithm used.
- E. The Amazon EC2 instance class specifying whether training will be run using CPU or GPU.
- F. The output path specifying where on an Amazon S3 bucket the trained model will persist.

Correct Answer: AEF

🗨️ 👤 **DonaldCMLIN** Highly Voted 🏆 1 year, 7 months ago
THE ANSWER SHOULD BE CEF
IAM ROLE, INSTANCE TYPE, OUTPUT PATH
upvoted 14 times

🗨️ 👤 **HaiHN** 8 months, 2 weeks ago
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.ipyn](https://github.com/aws/amazon-sagemaker-examples/blob/master/introduction_to_amazon_algorithms/semantic_segmentation_pascalvoc/semantic_segmentation_pascalvoc.ipyn#)
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)
```

upvoted 3 times

VB Highly Voted 1 year, 3 months ago

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.

upvoted 9 times

cloud_trail 5 months ago

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.

upvoted 2 times

cloud_trail 5 months ago

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.

upvoted 1 times

CloudGuru_ZA 1 month, 4 weeks ago

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

upvoted 1 times

AShahine21 Most Recent 1 month ago

After re-double checking the documentation (https://docs.aws.amazon.com/sagemaker/latest/APIReference/API_CreateTrainingJob.html): I will go with C E F

A is not required, we can read more under InputDataConfig

upvoted 1 times

orangechickencombo 1 month, 3 weeks ago

https://docs.aws.amazon.com/sagemaker/latest/APIReference/API_CreateTrainingJob.html

Cmd + F

search for 'Required: Yes'

CEF

upvoted 1 times

gcpwhiz 2 months, 1 week ago

This is a very simple question, one of the first items provided in the Developer Guide. Stop looking for alternative answers and minute detail when the answer is simple. There are 4 items needed for training jobs:

1. The URL of the Amazon Simple Storage Service (Amazon S3) bucket where you've stored the training data.
2. The compute resources that you want SageMaker to use for model training. Compute resources are ML compute instances that are managed by SageMaker.
3. The URL of the S3 bucket where you want to store the output of the job.
4. The Amazon Elastic Container Registry path where the training code is stored.

<https://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-training.html>

upvoted 1 times

tmdl 2 months, 2 weeks ago

CEF

<https://docs.aws.amazon.com/sagemaker/latest/dg/sagemaker-mkt-algo-train.html>

upvoted 1 times

🗨️ **sonalev419** 3 months ago

CEF (CANT BE A or B BECAUSE INPUTS CAN BE ON LOCAL MODE inputpath = FILE://
<https://sagemaker.readthedocs.io/en/stable/api/training/estimators.html#sagemaker.estimator.EstimatorBase>
upvoted 1 times

🗨️ **exploringaws** 3 months, 2 weeks ago

I think it is about parameters required to "running a training job" vs "creating a training job". Question is focusing on "Submitting a training", so answer should be A, E, F based on the AWS documentation.
upvoted 2 times

🗨️ **cnethers** 4 months, 2 weeks ago

When creating a training job you need:

- image_uri
- role
- train_instance_count
- train_instance_type
- train_volume_size
- output_path
- sagemaker_session
- rules

<https://docs.aws.amazon.com/sagemaker/latest/dg/ex1-train-model.html#ex1-train-model-sdk>

Based on that and the options available the ANSWER = C (Role) E (Instance Type) F (output path)

Some people feel training data and or validation data is required, this is not defined in the training job it is defined in the fitting of the algo
upvoted 2 times

🗨️ **scuzzy2010** 4 months, 4 weeks ago

A,C,E & F can all be correct, BUT for E (EC2), it says "specifying whether training will be run using CPU or GPU" - when we select the machine size, we don't specify whether to use CPU or GPU, so I think that option is out. Answer should be ACF
upvoted 1 times

🗨️ **Joe_Zhang** 5 months ago

FOR SAGEMAKER, THE TRAINING JOB DATA MUST FROM S3. SO AEF
upvoted 2 times

🗨️ **DzR** 5 months, 3 weeks ago

The answer is ACF according to the following documentation: The URL of the Amazon Simple Storage Service (Amazon S3) bucket where you've stored the training data.

The compute resources that you want SageMaker to use for model training. Compute resources are ML compute instances that are managed by SageMaker.

The URL of the S3 bucket where you want to store the output of the job.

The Amazon Elastic Container Registry path where the training code is stored.

upvoted 2 times

🗨️ **MennaMN** 10 months ago

I think it's "A,E,F" as the documentation discuss ARN of IAM not IAM itself. Maybe wordings matter in the choices we make.

This is the link:

https://docs.aws.amazon.com/sagemaker/latest/APIReference/API_CreateTrainingJob.html#API_CreateTrainingJob_RequestParameters
upvoted 1 times

🗨️ 👤 **bidds** 10 months, 1 week ago

This link states that the answer is AEF.

<https://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-training.html>

upvoted 1 times

🗨️ 👤 **bidds** 10 months, 1 week ago

Then again, this seems to contradict that:

[https://docs.aws.amazon.com/sagemaker/latest/APIReference/API_CreateTrainingJob.html#API_CreateTrainingJob_RequestParameter:](https://docs.aws.amazon.com/sagemaker/latest/APIReference/API_CreateTrainingJob.html#API_CreateTrainingJob_RequestParameter)

upvoted 1 times

🗨️ 👤 **Antriksh** 1 year ago

C E and F

upvoted 2 times

🗨️ 👤 **roytruong** 1 year, 1 month ago

CEF, see this https://docs.aws.amazon.com/sagemaker/latest/APIReference/API_CreateTrainingJob.html

upvoted 3 times

🗨️ 👤 **PRC** 1 year, 3 months ago

CEF is correct

upvoted 1 times

Question #31

Topic 1

A monitoring service generates 1 TB of scale metrics record data every minute. A Research team performs queries on this data using Amazon Athena. The queries run slowly due to the large volume of data, and the team requires better performance.

How should the records be stored in Amazon S3 to improve query performance?

- A. CSV files
- B. Parquet files
- C. Compressed JSON
- D. RecordIO

Correct Answer: B

 **gaku1016** Highly Voted 1 year, 4 months ago

Answer is B. Athena is best in Parquet format.

upvoted 15 times

 **emailtorajivk** Highly Voted 1 year, 3 months ago

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

upvoted 10 times

 **benson2021** Most Recent 2 months, 3 weeks ago

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?

upvoted 1 times

Question #32

Topic 1

Machine Learning Specialist is working with a media company to perform classification on popular articles from the company's website. The company is using random forests to classify how popular an article will be before it is published. A sample of the data being used is below.

Article_Title	Author	Top_Keywords	Day_Of_Week	URL_of_Article	Page_Views
Building a Big Data Platform	Jane Doe	Big Data, Spark, Hadoop	Tuesday	http://examplecorp.com/data_platform.html	1300456
Getting Started with Deep Learning	John Doe	Deep Learning, Machine Learning, Spark	Tuesday	http://examplecorp.com/started_deep_learning.html	1230661
MXNet ML Guide	Jane Doe	Machine Learning, MXNet, Logistic Regression	Thursday	http://examplecorp.com/mxnet_guide.html	937291
Intro to NoSQL	Mary Major	NoSQL, Operations	Monday	http://examplecorp.com/nosql_intro_guide.html	407812


Databases		Database			
-----------	--	----------	--	--	--

Given the dataset, the Specialist wants to convert the Day_Of_Week column to binary values.
What technique should be used to convert this column to binary values?

- A. Binarization
- B. One-hot encoding
- C. Tokenization
- D. Normalization transformation

Correct Answer: B

 **omar_bahrain** Highly Voted 3 months, 3 weeks ago
I choose b
upvoted 5 times

 **Juka3lj** Most Recent 2 months ago
Correct answer is B.
Example:
Mon | Tue | Wed
1 0 0
0 1 0
upvoted 2 times

Question #33

Topic 1

A gaming company has launched an online game where people can start playing for free, but they need to pay if they choose to use certain features. The company needs to build an automated system to predict whether or not a new user will become a paid user within 1 year. The company has gathered a labeled dataset from 1 million users.

The training dataset consists of 1,000 positive samples (from users who ended up paying within 1 year) and 999,000 negative samples (from users who did not use any paid features). Each data sample consists of 200 features including user age, device, location, and play patterns. Using this dataset for training, the Data Science team trained a random forest model that converged with over 99% accuracy on the training set. However, the prediction results on a test dataset were not satisfactory

Which of the following approaches should the Data Science team take to mitigate this issue? (Choose two.)

- A. Add more deep trees to the random forest to enable the model to learn more features.

- B. Include a copy of the samples in the test dataset in the training dataset.
- C. Generate more positive samples by duplicating the positive samples and adding a small amount of noise to the duplicated data.
- D. Change the cost function so that false negatives have a higher impact on the cost value than false positives.
- E. Change the cost function so that false positives have a higher impact on the cost value than false negatives.

Correct Answer: CD

🗲️ 👤 **Phong** Highly Voted 👍 1 year, 5 months ago

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

upvoted 19 times

🗲️ 👤 **Phong** Highly Voted 👍 1 year, 5 months ago

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

upvoted 16 times

🗲️ 👤 **cloud_trail** Most Recent 🕒 5 months ago

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.

upvoted 3 times

🗲️ 👤 **felbuch** 5 months, 1 week ago

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.

upvoted 3 times

🗲️ 👤 **engomaradel** 6 months, 3 weeks ago

CD 100%

upvoted 1 times

🗲️ 👤 **ybad** 7 months, 1 week ago

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..

upvoted 1 times

🗲️ 👤 **Omar_Cascudo** 8 months ago

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.

upvoted 1 times

-  **bidbs** 10 months, 1 week ago
 CD are correct for sure.
 upvoted 3 times
-  **hans1234** 11 months, 1 week ago
 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
 upvoted 2 times
-  **Wira** 1 year ago
 CD

 although i am worried about the noise being introduced as it could skew the data nevertheless no better answer is given
 upvoted 2 times
-  **aws_razor** 1 year, 1 month ago
 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.
 upvoted 3 times
-  **roytruong** 1 year, 1 month ago
 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 cost function
 upvoted 3 times
-  **wuha5086** 1 year, 2 months ago
 CD, FN are valuable players, we should care more on FN
 upvoted 6 times
-  **VB** 1 year, 3 months ago
 Is my assumption right here?


 ACTUAL

 P PAY NPAY
 R -----
 E PAY TP FP
 D
 I NPAY FN TN
 C -----
 upvoted 1 times
-  **AKT** 1 year, 4 months ago
 C,D correct
 upvoted 5 times
-  **rajs** 1 year, 4 months ago
 #1 Imbalanced data - needs balancing
 Agreed with C

 # Imbalanced Cost of FP & FN
 We have a higher probability of FN and lower probability of FP
 To balance that increase the sensitivity/weight towards FP

 So the do E

 Answer is CE
 upvoted 6 times

 **JayK** 1 year, 4 months ago

CD: should take care of False Negatives over False Positives
upvoted 9 times

Question #34

Topic 1

A Data Scientist is developing a machine learning model to predict future patient outcomes based on information collected about each patient and their treatment plans. The model should output a continuous value as its prediction. The data available includes labeled outcomes for a set of 4,000 patients. The study was conducted on a group of individuals over the age of 65 who have a particular disease that is known to worsen with age.

Initial models have performed poorly. While reviewing the underlying data, the Data Scientist notices that, out of 4,000 patient observations, there are 450 where the patient age has been input as 0. The other features for these observations appear normal compared to the rest of the sample population

How should the Data Scientist correct this issue?

- A. Drop all records from the dataset where age has been set to 0.
- B. Replace the age field value for records with a value of 0 with the mean or median value from the dataset
- C. Drop the age feature from the dataset and train the model using the rest of the features.
- D. Use k-means clustering to handle missing features

Correct Answer: B

 **vetal** Highly Voted 1 year, 6 months ago

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?

upvoted 12 times

🗨️ 👤 **rajs** Highly Voted 1 year, 4 months ago

Dropping the Age feature is a NOT AT ALL 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 this means that using Unsupervised Learning we can make pretty good prediction of "Age"

So the answer is D Use K-Means clustering

upvoted 12 times

🗨️ 👤 **L2007** 1 year, 4 months ago

<https://www.displayr.com/5-ways-deal-missing-data-cluster-analysis/>

B is correct

upvoted 2 times

🗨️ 👤 **hero67** Most Recent 1 week, 3 days ago

Since the missing values are considerable fraction from the whole data set, yet they are crucial to the model (disease is highly correlated to those aged than 65) then there is a need for imputation. It seems odd to use clustering technique to do such a job. Hence, the only correct answer is impute using mean or median values. On the other hand, imputing using K-nearest neighbors or KNNs would be a viable solution but not K-means clustering.

upvoted 1 times

🗨️ 👤 **abdohanfi** 1 month ago

i think k means clustering is unsupervised and he said that the data is labeled so doesn't need unsupervised algorithm so i think the mean or median for the 0s are good option

upvoted 1 times

🗨️ 👤 **Vita_Rasta84444** 3 months ago

I think it is D. If we just impute the median we will lose the important information, and age is one of the predictors. With K-means, we can make clusters and then impute the cluster age mean for the cluster.

upvoted 1 times

🗨️ 👤 **omar_bahrain** 4 months, 1 week ago

searching the net, I have noticed that even K-Means Clustering can be also used for imputing missing data so D is still a good candidate

upvoted 3 times

🗨️ 👤 **cloud_trail** 5 months ago

Wow there are a lot of bad comments for this one. Age is a critically important column so you cannot just dump the whole column. Eliminate C. k-means is not used for imputation. KNN is. D is a deception option and incorrect. KNN imputation would have been the best option but not listed here. Because age is a critical predictor as stated in the question, you just cannot assign the mean or median to it. That will ruin your model. Eliminate B. A is the best option listed. Dropping 11% of your training examples is not optimal but it is acceptable when you still have thousands of samples.

upvoted 2 times

🗨️ 👤 **DzR** 5 months, 3 weeks ago

It will be my first time to handle missing values with K-Means, so I will go with B, choosing between the mean and the median looking at the distribution of the data.

upvoted 1 times

🗨️ 👤 **dd1024** 6 months ago

Should be B. Using K-mean is OK, but D said "handle missing feature". those 450 samples have incorrect input feature instead of "missing feature", so D is a misleading option.

upvoted 2 times

  **yeetusdeleetus** 8 months ago

Correct answer is B.

The fact that they mentioned the data with missing values appears normal compared to the rest of the population (ie, normally distributed) means that the mean will not introduce very significant bias. If anything it will introduce noise, not bias.

Using k-means is clever but would not solve the problem unless you also imputed values, which the answer does not state.



I personally would prefer to drop rows with missing values and I think it gives better results, and if the resulting dataset is too small you should gather more data. But, in the context of the question where they mention the data with missing values is 'normal' compared to the remainder of data, the mean is what we are pointed to.

upvoted 2 times

  **Omar_Cascudo** 8 months ago



B and D are correct paths to follow. Clustering the patients, as disease gets worst with age, and assign the mean or median age of the cluster should not introduce bias. I go with D.

upvoted 1 times

  **weslleylc** 8 months, 2 weeks ago

Drop rows is better than filling the wrong values. Since age is the most important feature, using missing data techniques here can be dangerous. It is better to avoid input noise in your most important variable.

upvoted 4 times

  **Tunermania** 9 months, 2 weeks ago



Since there have been a huge suggestion of the answer being B or C, I will like to analyse the solution from a different angle:

From the question it's stated that "The other features for these observations appear normal compared to the rest of the sample population" which logically mean, only the age feature is having missing values, if I am right.

Option D states that "Use k-means clustering to handle missing <features>". I want us to notice the "s" added to the feature in option D. This contradicts the statement in the question about age feature being the only one that has missing values.



Hence, option B should be the right answer.

upvoted 1 times

  **lightblue** 6 months, 2 weeks ago



what about the "w" added to 'right' - in your comment

upvoted 1 times

  **Th3Dud3** 9 months, 2 weeks ago

D. Use k-means clustering to handle missing features ---> Nothing specific about AGE

upvoted 2 times

  **WillNguyen22** 9 months, 3 weeks ago

B is the best choice. Because with A, B is definitely not suitable for this problem. D if we only use K-Means we only cluster the age and then what next? we need a step after clustering. So D is not enough for this problem



upvoted 1 times

  **syu31svc** 10 months ago

C and D are wrong; k-means clustering does not handle missing values in any way

Between A and B, B is the better choice since not most of the values are missing for it to be dropped.

upvoted 1 times

  **yddmj** 10 months, 1 week ago

Question #35

Topic 1

A Data Science team is designing a dataset repository where it will store a large amount of training data commonly used in its machine learning models. As Data

Scientists may create an arbitrary number of new datasets every day, the solution has to scale automatically and be cost-effective. Also, it must be possible to explore the data using SQL.

Which storage scheme is MOST adapted to this scenario?

- A. Store datasets as files in Amazon S3.
- B. Store datasets as files in an Amazon EBS volume attached to an Amazon EC2 instance.
- C. Store datasets as tables in a multi-node Amazon Redshift cluster.
- D. Store datasets as global tables in Amazon DynamoDB.

Correct Answer: A

🗉 👤 **rsimham** Highly Voted 👍 1 year, 6 months ago

Ans: A (S3) is most cost effective

upvoted 11 times

🗉 👤 **sonalev419** Most Recent ⌚ 3 months ago

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

upvoted 3 times

🗉 👤 **harmanbirstudy** 5 months, 3 weeks ago

"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.

Answer is A

upvoted 2 times

🗉 👤 **Stephen_C** 6 months ago

Amazon Redshift is not cost-effective.

upvoted 1 times

🗉 👤 **syu31svc** 9 months, 2 weeks ago

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."

upvoted 2 times

🗉 👤 **HaiHN** 8 months, 2 weeks ago

Data warehouse is not needed. For exploring data using SQL, you can use Athena

upvoted 4 times

🗉 👤 **kwangje** 6 months, 1 week ago

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.

upvoted 1 times

🗉 👤 **roytruong** 1 year, 1 month ago

s3 is right

upvoted 1 times

🗉 👤 **cybe001** 1 year, 5 months ago

A, S3 is most cost effective

upvoted 2 times

Question #36

Topic 1

A Machine Learning Specialist deployed a model that provides product recommendations on a company's website. Initially, the model was performing very well and resulted in customers buying more products on average. However, within the past few months, the Specialist has noticed that the effect of product recommendations has diminished and customers are starting to return to their original habits of spending less. The Specialist is unsure of what happened, as the model has not changed from its initial deployment over a year ago.

Which method should the Specialist try to improve model performance?

- A. The model needs to be completely re-engineered because it is unable to handle product inventory changes.
- B. The model's hyperparameters should be periodically updated to prevent drift.
- C. The model should be periodically retrained from scratch using the original data while adding a regularization term to handle product inventory changes
- D. The model should be periodically retrained using the original training data plus new data as product inventory changes.

Correct Answer: D

  **rsimham** Highly Voted 1 year, 6 months ago



Ans: D

upvoted 18 times

  **cloud_trail** Most Recent 5 months ago

Incremental training. D.

upvoted 1 times

  **gamaX** 7 months, 1 week ago

Periodically Re-Fit

D

upvoted 1 times

  **ejj** 11 months, 3 weeks ago

agree with D

upvoted 3 times

  **C10ud9** 1 year ago

D is correct



upvoted 3 times

Question #37

Topic 1

A Machine Learning Specialist working for an online fashion company wants to build a data ingestion solution for the company's Amazon S3-based data lake.

The Specialist wants to create a set of ingestion mechanisms that will enable future capabilities comprised of:

-  Real-time analytics
-  Interactive analytics of historical data

- ⇒ Clickstream analytics
- ⇒ Product recommendations

Which services should the Specialist use?

- A. AWS Glue as the data catalog; Amazon Kinesis Data Streams and Amazon Kinesis Data Analytics for real-time data insights; Amazon Kinesis Data Firehose for delivery to Amazon ES for clickstream analytics; Amazon EMR to generate personalized product recommendations
- B. Amazon Athena as the data catalog; Amazon Kinesis Data Streams and Amazon Kinesis Data Analytics for near-real-time data insights; Amazon Kinesis Data Firehose for clickstream analytics; AWS Glue to generate personalized product recommendations
- C. AWS Glue as the data catalog; Amazon Kinesis Data Streams and Amazon Kinesis Data Analytics for historical data insights; Amazon Kinesis Data Firehose for delivery to Amazon ES for clickstream analytics; Amazon EMR to generate personalized product recommendations
- D. Amazon Athena as the data catalog; Amazon Kinesis Data Streams and Amazon Kinesis Data Analytics for historical data insights; Amazon DynamoDB streams for clickstream analytics; AWS Glue to generate personalized product recommendations

Correct Answer: A

🗄️ 👤 **rsimham** Highly Voted 1 year, 6 months ago
 Ans: A seems to be reasonable
 upvoted 25 times

🗄️ 👤 **cybe001** Highly Voted 1 year, 5 months ago
 A looks correct but it is missing for "Interactive analytics of historical data"
 upvoted 7 times

🗄️ 👤 **ejj** 11 months, 3 weeks ago
 but C is missing for "real-time analytics"
 upvoted 1 times

🗄️ 👤 **ejj** 11 months, 3 weeks ago
 and also C is saying historical data analytics for Kinesis Data analytics which is real-time analytics not historical, so the answer might C but the answer is A
 upvoted 1 times

🗄️ 👤 **gamaX** Most Recent 3 months, 3 weeks ago
 A or C
<https://aws.amazon.com/blogs/big-data/retaining-data-streams-up-to-one-year-with-amazon-kinesis-data-streams/>
 upvoted 1 times

🗄️ 👤 **harmanbirstudy** 5 months, 3 weeks ago
 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 ther 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 S >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
 upvoted 3 times

🗨️ **ybad** 7 months, 1 week ago

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 money no?

upvoted 2 times

🗨️ **Th3Dud3** 9 months, 2 weeks ago

Answer = A

upvoted 4 times

🗨️ **C10ud9** 1 year ago

A it is

upvoted 2 times

🗨️ **roytruong** 1 year, 1 month ago

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

upvoted 3 times

🗨️ **BigPlums** 1 year, 2 months ago

Only C and D mention interactive analytics of historical data.

Glue won't provide personalised recommendation so it is C

upvoted 1 times

🗨️ **BigEv** 1 year, 5 months ago

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

upvoted 2 times

🗨️ **JayK** 1 year, 4 months ago

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

upvoted 6 times

🗨️ **ComPah** 1 year, 6 months ago

Looks like C Amazon ES has Kibana which supports click stream

upvoted 2 times

🗨️ **ComPah** 1 year, 6 months ago

A is Correct

upvoted 4 times

Question #38

Topic 1

A company is observing low accuracy while training on the default built-in image classification algorithm in Amazon SageMaker. The Data Science team wants to use an Inception neural network architecture instead of a ResNet architecture.

Which of the following will accomplish this? (Choose two.)

- A. Customize the built-in image classification algorithm to use Inception and use this for model training.
- B. Create a support case with the SageMaker team to change the default image classification algorithm to Inception.
- C. Bundle a Docker container with TensorFlow Estimator loaded with an Inception network and use this for model training.

D. Use custom code in Amazon SageMaker with TensorFlow Estimator to load the model with an Inception network, and use this for model training.

E. Download and apt-get install the inception network code into an Amazon EC2 instance and use this instance as a Jupyter notebook in Amazon SageMaker.

Correct Answer: CD

🗲️ 👤 **DonaldCMLIN** Highly Voted 👍 1 year, 7 months ago

You might be spent a lot of money for ask AWS A.CHANGE built-in image OR B.Create a support case.

The effectual 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

upvoted 20 times

🗲️ 👤 **Phong** Highly Voted 👍 1 year, 4 months ago

I will go for C & D

upvoted 8 times

🗲️ 👤 **ahquiceno** Most Recent 🕒 5 months ago

Answers AD go to: <https://docs.aws.amazon.com/sagemaker/latest/dg/docker-containers.html>

upvoted 2 times

🗲️ 👤 **ybad** 7 months, 1 week ago

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

upvoted 1 times

🗲️ 👤 **jaydec** 11 months, 3 weeks ago

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

upvoted 1 times

🗲️ 👤 **Antriksh** 1 year ago

C and D are correct

upvoted 4 times

🗲️ 👤 **DonaldCMLIN** 1 year, 7 months ago

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

upvoted 1 times

Question #39

Topic 1

A Machine Learning Specialist built an image classification deep learning model. However, the Specialist ran into an overfitting problem in which the training and testing accuracies were 99% and 75%, respectively.

How should the Specialist address this issue and what is the reason behind it?

- A. The learning rate should be increased because the optimization process was trapped at a local minimum.
- B. The dropout rate at the flatten layer should be increased because the model is not generalized enough.
- C. The dimensionality of dense layer next to the flatten layer should be increased because the model is not complex enough.
- D. The epoch number should be increased because the optimization process was terminated before it reached the global minimum.

Correct Answer: D

Reference:

https://www.tensorflow.org/tutorials/keras/overfit_and_underfit

  **DonaldCMLIN** Highly Voted 1 year, 7 months ago

DROPOUT HELPS PREVENT OVERFITTING

<https://keras.io/layers/core/#dropout>

THE BEAUTIFUL ANSWER SHOULD BE B.

upvoted 31 times

  **rsimham** 1 year, 6 months ago



agree. it should be B

upvoted 6 times

  **eganilovic** Most Recent 2 months ago



Correct answer is B! Number of epochs should be decreased.

upvoted 2 times

  **gcpwhiz** 2 months, 2 weeks ago



The reference article listed by the website also says dropout is the correct answer...go with B

upvoted 2 times

  **Vita_Rasta84444** 3 months ago

B is true answer. More epochs will not stop overfitting.

upvoted 2 times

  **littlewat** 3 months, 2 weeks ago

B!!!!!!!!!!!!!!

upvoted 2 times

  **DzR** 5 months, 3 weeks ago

I will go for b because the problem is the model is failing to generalize and hence dropout will solve the problem.

upvoted 1 times



  **harmanbirstudy** 5 months, 3 weeks ago

It cannot be because with 99% accuracy the testing does need epochs and nowhere it says about epoch getting Early Terminated in Question, so we cannot just assume it.

Among the rest of the list only "Increasing Dropout" helps preventing Overfitting.

ANSWER is B

upvoted 1 times

  **ybad** 7 months, 1 week ago

Agree with all it should be B, only solution to reduce overfitting,

increasing epochs would actually cause it to overfit more....

upvoted 1 times

- 🗨️ **syu31svc** 10 months ago
<https://kharshit.github.io/blog/2018/05/04/dropout-prevent-overfitting>
Answer is B 100%
upvoted 4 times
- 🗨️ **hans1234** 11 months ago
It is not A, because the training error is at 99% versus 75% for test, which means it is not at a local minimum, but has overfitting problems. The answer is B.
upvoted 3 times
- 🗨️ **mrsimoes** 10 months ago
I agree with dropout, the problem is the second part of the answer. The model to have 99% accuracy in the training means it is generalizing well in the train set.
upvoted 1 times
- 🗨️ **algorithmish** 11 months, 3 weeks ago
it is ok to use dropout before flatten layer. Answer is B
https://github.com/keras-team/keras/blob/master/examples/mnist_cnn.py
upvoted 2 times
- 🗨️ **andrey1h** 1 year ago
My answer is A, since dropout is used on fully connected (dense) layers not flatten.
upvoted 2 times
- 🗨️ **[Removed]** 1 year, 3 months ago
epoch number is utilized only in training phase, but there is a 99% of accuracy in this phase it wouldn't get better accuracy, I think increasing number of epoch does not improve the result in test
upvoted 2 times
- 🗨️ **Phong** 1 year, 4 months ago
go for B
upvoted 3 times
- 🗨️ **tap123** 1 year, 4 months ago
Why not A? B is slightly confusing as flatten and dropout are separate layer types in TensorFlow.
upvoted 1 times
- 🗨️ **Antriksh** 1 year ago
with dropout you are reducing over dependency on neurons thereby making model more generic and less complex. Over depending on neuron makes the network learn the model. This can cause model to overfit. Dropping neuron prevents that.
Hence correct answer is dropout
upvoted 2 times
- 🗨️ **vetal** 1 year, 6 months ago
What is interesting in the suggested link the best options are regularization and dropout :)
https://www.tensorflow.org/tutorials/keras/overfit_and_underfit
upvoted 3 times
- 🗨️ **SabinaMystique** 8 months, 4 weeks ago
and even say this "Dropout is one of the most effective and most commonly used regularization techniques for neural networks"
upvoted 1 times

🗨️ **DonaldCMLIN** 1 year, 7 months ago

NOT D.

Although increase epoch could improve overfitting, but not in status of early stopping

B.INCREASE DROPOUT FOR not generalize well

<https://developer.ibm.com/articles/image-recognition-challenge-with-tensorflow-and-keras-pt2/>

upvoted 4 times

🗨️ **ahquiceno** 5 months ago

100% B. The model need to be less complex, it only memorized the info, needs a "trepanning" :-).

upvoted 1 times

🗨️ **zzeng** 11 months, 2 weeks ago

https://keras.io/api/layers/regularization_layers/dropout/

The Dropout layer randomly sets input units to 0 with a frequency of rate at each step during training time, which helps prevent overfittir

upvoted 1 times

Question #40

Topic 1

A Machine Learning team uses Amazon SageMaker to train an Apache MXNet handwritten digit classifier model using a research dataset. The team wants to receive a notification when the model is overfitting. Auditors want to view the Amazon SageMaker log activity report to ensure there are no unauthorized API calls.

What should the Machine Learning team do to address the requirements with the least amount of code and fewest steps?

- A. Implement an AWS Lambda function to log Amazon SageMaker API calls to Amazon S3. Add code to push a custom metric to Amazon CloudWatch. Create an alarm in CloudWatch with Amazon SNS to receive a notification when the model is overfitting.
- B. Use AWS CloudTrail to log Amazon SageMaker API calls to Amazon S3. Add code to push a custom metric to Amazon CloudWatch. Create an alarm in CloudWatch with Amazon SNS to receive a notification when the model is overfitting.
- C. Implement an AWS Lambda function to log Amazon SageMaker API calls to AWS CloudTrail. Add code to push a custom metric to Amazon CloudWatch. Create an alarm in CloudWatch with Amazon SNS to receive a notification when the model is overfitting.
- D. Use AWS CloudTrail to log Amazon SageMaker API calls to Amazon S3. Set up Amazon SNS to receive a notification when the model is overfitting

Correct Answer: B

🗨️ **DonaldCMLIN** Highly Voted 👍 1 year, 7 months ago

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>

upvoted 27 times

- 🗲️ 👤 **rajs** Highly Voted 1 year, 4 months ago
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
upvoted 7 times
- 🗲️ 👤 **sonalev419** Most Recent 3 months ago
B. cloudwatch + metrics from sagemaker + sns https://docs.aws.amazon.com/fr_fr/sagemaker/latest/dg/training-metrics.html#define-train-metrics
upvoted 1 times
- 🗲️ 👤 **ybad** 7 months, 1 week ago
B requires the least amount of code and satisfies all conditions
upvoted 2 times
- 🗲️ 👤 **tochiebby** 9 months, 3 weeks ago
What does this line do?
- "Add code to push a custom metric to Amazon CloudWatch"
upvoted 1 times
- 🗲️ 👤 **Omar_Cascudo** 8 months ago
It creates a metric for overfitting (accuracy of training data and accuracy of test data).
upvoted 2 times
- 🗲️ 👤 **jonclem** 11 months, 2 weeks ago
Its not B. Why would you use CloudTrail?
- Having used Lambda for API calls I'm inclined to agree with the original answer, C.
upvoted 1 times
- 🗲️ 👤 **fhuadeen** 9 months, 2 weeks ago
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?
upvoted 2 times
- 🗲️ 👤 **Pja1** 10 months, 2 weeks ago
<https://docs.aws.amazon.com/sagemaker/latest/dg/logging-using-cloudtrail.html>
upvoted 3 times
- 🗲️ 👤 **Antriksh** 1 year ago
B it is
upvoted 2 times
- 🗲️ 👤 **C10ud9** 1 year ago
B it is
upvoted 1 times
- 🗲️ 👤 **lisiuyiu** 1 year, 2 months ago
Agree on B
upvoted 2 times
- 🗲️ 👤 **ac427** 1 year, 3 months ago
ALL AWS Service's API calls are logged to CloudTrail automatically.
upvoted 3 times

Question #41

Topic 1



A Machine Learning Specialist is building a prediction model for a large number of features using linear models, such as linear regression and logistic regression.



During exploratory data analysis, the Specialist observes that many features are highly correlated with each other. This may make the model unstable.



What should be done to reduce the impact of having such a large number of features?



- A. Perform one-hot encoding on highly correlated features.
- B. Use matrix multiplication on highly correlated features.
- C. Create a new feature space using principal component analysis (PCA)
- D. Apply the Pearson correlation coefficient.

Correct Answer: C

  **cybe001** Highly Voted 1 year, 5 months ago
C is correct
upvoted 10 times

  **syu31svc** Most Recent 10 months ago
You want to reduce features/dimension so PCA is the answer
upvoted 2 times

  **Urban_Life** 12 months ago
Of course, it's PCA.
upvoted 1 times

  **C10ud9** 1 year ago
PCA is the solution. So, answer is C
upvoted 1 times

Question #42

Topic 1

















A Machine Learning Specialist is implementing a full Bayesian network on a dataset that describes public transit in New York City. One of the random variables is discrete, and represents the number of minutes New Yorkers wait for a bus given that the buses cycle every 10 minutes, with a mean of 3 minutes.

Which prior probability distribution should the ML Specialist use for this variable?



- A. Poisson distribution
- B. Uniform distribution



- C. Normal distribution
- D. Binomial distribution

Correct Answer: D

-   **Compah** Highly Voted 1 year, 5 months ago
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
upvoted 36 times
-   **DScode** Highly Voted 1 year ago
definitely Poisson distribution. No two ways about it. this a poisson process.
upvoted 6 times
-   **tmld** Most Recent 2 months, 2 weeks ago
A!!!!!!!
upvoted 1 times
-   **littlewat** 3 months, 2 weeks ago
A!!!!!!!
upvoted 2 times
-   **takahirokoyama** 4 months, 4 weeks ago
Completely A.
Check this site.(<https://jakevdp.github.io/blog/2018/09/13/waiting-time-paradox/>)
upvoted 2 times
-   **DzR** 5 months, 3 weeks ago
When ever you are given the mean occurrence of things in an interval, the best distribution to model such an event is Poisson. A is the correct answer.
upvoted 2 times
-   **harmanbirstudy** 5 months, 3 weeks ago
Answer is A, I am no expert here but if you just google " buses cycle every 10 minutes, with a mean of 3 minutes " , resulting page talk about Poisson distribution
upvoted 2 times
-   **shoshi** 6 months, 3 weeks ago
Hi, I passed the exam today, this was my question. good luck all
upvoted 2 times
-   **harmanbirstudy** 5 months, 3 weeks ago
I have mine tomorrow .. what was your reply ?
upvoted 1 times
-   **bm25** 6 months, 2 weeks ago
what was the right response for this question?
upvoted 4 times
-   **Thai_Xuan** 7 months, 3 weeks ago
uniform distribution. Note the word "every". No randomness in bus arrivals is assumed here.
<https://stats.stackexchange.com/questions/122722/please-explain-the-waiting-paradox>
https://www.reddit.com/r/askscience/comments/59qkk7/how_does_the_wait_time_paradox_work/
Don't rely on the phrase "prior probability".
upvoted 2 times

- 🗨️ **sebtac** 9 months, 4 weeks ago
Answer: A. it is a classical Poisson Example
upvoted 1 times
- 🗨️ **syu31svc** 10 months ago
The Poisson distribution is used to model the number of events occurring within a given time interval.
So answer is A
upvoted 2 times
- 🗨️ **GeeBeeEI** 11 months ago
C is the answer -- Normal Check <https://www.healthknowledge.org.uk/public-health-textbook/research-methods/1b-statistical-methods/statistical-distributions>
upvoted 1 times
- 🗨️ **ybad** 7 months, 1 week ago
Normal distribution is a continuous probability distribution, thats why poisson works better, since it is discrete!
upvoted 3 times
- 🗨️ **Antriksh** 1 year ago
Poisson distribution it is
upvoted 3 times
- 🗨️ **deep_n** 1 year, 1 month ago
leaning towards A
<https://www.analyticsvidhya.com/blog/2017/09/6-probability-distributions-data-science/>
upvoted 3 times
- 🗨️ **swagy** 1 year, 2 months ago
Ans: A
<https://brilliant.org/wiki/poisson-distribution/>
upvoted 6 times
- 🗨️ **cybe001** 1 year, 5 months ago
Answer is D.
Binomial distribution: A binomial random variable is the number of successes in n trials of a random experiment. A random variable x is said to follow binomial distribution when, the random variable can have only two outcomes(success and failure).
<https://towardsdatascience.com/probability-and-statistics-explained-in-the-context-of-deep-learning-ed1509b2eb3f>
upvoted 1 times
- 🗨️ **cybe001** 1 year, 5 months ago
Since the variable is discrete, all the possible values can be grouped into two
upvoted 1 times
- 🗨️ **VB** 1 year, 3 months ago
I don't think the answer is just yes or no.. it is supposed to be a wait time from 0 to 10 minutes, with an average of 3 min. I think the answer should be A. Poisson Distr.
upvoted 7 times
- 🗨️ **felbuch** 5 months, 1 week ago
It is not a Poisson distribution, because if it were, there would be no maximum time a person could stand waiting for the bus. But since buses go about every 10 minutes, then no person will wait for more than 10 minutes. So it's a Binomial Distr.
upvoted 1 times

  **DonaldCMLIN** 1 year, 7 months ago
SINCE "prior probability distribution", THE ANSWER B.
upvoted 4 times

  **gaow** 8 months, 3 weeks ago
But why the Suggested Answer by you for this question is D?
upvoted 1 times

Question #43

Topic 1




A Data Science team within a large company uses Amazon SageMaker notebooks to access data stored in Amazon S3 buckets. The IT Security team is concerned that internet-enabled notebook instances create a security vulnerability where malicious code running on the instances could compromise data privacy.

The company mandates that all instances stay within a secured VPC with no internet access, and data communication traffic must stay within the AWS network.




How should the Data Science team configure the notebook instance placement to meet these requirements?




- A. Associate the Amazon SageMaker notebook with a private subnet in a VPC. Place the Amazon SageMaker endpoint and S3 buckets within the same VPC.
- B. Associate the Amazon SageMaker notebook with a private subnet in a VPC. Use IAM policies to grant access to Amazon S3 and Amazon SageMaker.
- C. Associate the Amazon SageMaker notebook with a private subnet in a VPC. Ensure the VPC has S3 VPC endpoints and Amazon SageMaker VPC endpoints attached to it.
- D. Associate the Amazon SageMaker notebook with a private subnet in a VPC. Ensure the VPC has a NAT gateway and an associated security group allowing only outbound connections to Amazon S3 and Amazon SageMaker.

Correct Answer: C

  **DonaldCMLIN**  1 year, 7 months ago
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
upvoted 19 times

  **BigEv**  1 year, 5 months ago
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>
upvoted 13 times

  **StelSen**  1 month, 3 weeks ago
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.
upvoted 1 times

🗳️ 👤 **cloud_trail** 5 months ago

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%20public%20resources>

upvoted 2 times

🗳️ 👤 **harmanbirstudy** 5 months, 3 weeks ago

"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 its sole purpose is to create route from VPC endpoint to S3, without going over the Internet.

upvoted 2 times

🗳️ 👤 **scuzzy2010** 7 months, 1 week ago

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>

upvoted 1 times

🗳️ 👤 **v24143** 7 months, 2 weeks ago

D is correct

upvoted 1 times

🗳️ 👤 **krakow1234** 7 months, 3 weeks ago

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

upvoted 1 times

🗳️ 👤 **Potato_Noodle** 8 months, 3 weeks ago

NAT is the way that a VPC connects to internet and other AWS service when there is NO INTERNET ACCESS FOR VPC. Thus the answer is

upvoted 1 times

🗳️ 👤 **Th3Dud3** 9 months, 2 weeks ago

"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!

upvoted 2 times

🗳️ 👤 **yeetusdeleetus** 7 months, 4 weeks ago

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 to go to other AWS services, which is the best solution for this question.

upvoted 2 times

🗳️ 👤 **Th3Dud3** 9 months, 3 weeks ago

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.

upvoted 3 times

- 🗨️ 👤 **sebtac** 9 months, 4 weeks ago
what is the difference between A & C? are both answers OK?
upvoted 1 times
- 🗨️ 👤 **syu31svc** 10 months ago
I would say the answer is C
<https://docs.aws.amazon.com/vpc/latest/userguide/vpc-endpoints.html>
<https://docs.aws.amazon.com/sagemaker/latest/dg/interface-vpc-endpoint.html>
upvoted 1 times
- 🗨️ 👤 **AShahine21** 1 month ago
S3 buckets don't live in a VPC
upvoted 1 times
- 🗨️ 👤 **GeeBeeEI** 11 months ago
See <https://docs.aws.amazon.com/sagemaker/latest/dg/appendix-notebook-and-internet-access.html>
A is wrong, S3 buckets don't live in a VPC
B is wrong, you don't use IAM policies to grant access to SageMaker
C may be right depending on what "it" is --- attached to it If by it, this is talking about the notebook VPC, then yes
D talks of security group and NAT gateway which was mentioned in the link. D is definitely correct based on link
upvoted 2 times
- 🗨️ 👤 **haison8x** 11 months, 1 week ago
I think the answer is D, there are no such thing as S3 VPC endpoints and Amazon SageMaker
upvoted 1 times
- 🗨️ 👤 **Thai_Xuan** 7 months, 3 weeks ago
take a look at <https://docs.aws.amazon.com/sagemaker/latest/dg/train-vpc.html#train-vpc-s3>
upvoted 1 times
- 🗨️ 👤 **tff** 12 months ago
Answer C
<https://docs.aws.amazon.com/sagemaker/latest/dg/host-vpc.html>
upvoted 3 times
- 🗨️ 👤 **Wira** 1 year ago
C
very obvious one
<https://docs.aws.amazon.com/sagemaker/latest/dg/interface-vpc-endpoint.html>
upvoted 2 times

Question #44

Topic 1



A Machine Learning Specialist has created a deep learning neural network model that performs well on the training data but performs poorly on the test data.



Which of the following methods should the Specialist consider using to correct this? (Choose three.)



- A. Decrease regularization.
- B. Increase regularization.



- C. Increase dropout.
- D. Decrease dropout.
- E. Increase feature combinations.
- F. Decrease feature combinations.



Correct Answer: BCF



  **cybe001** Highly Voted 1 year, 5 months ago
Yes, answer is BCF
upvoted 10 times



  **Phong** Highly Voted 1 year, 4 months ago
Go for BCF
upvoted 8 times



  **ahquiceno** Most Recent 4 months, 2 weeks ago
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 feature with high correlation.
<https://towardsdatascience.com/dealing-with-highly-dimensional-data-using-principal-component-analysis-pca-fea1ca817fe6>
upvoted 1 times

  **cloud_trail** 5 months ago
B/C/F Easy peasy.
upvoted 1 times

  **apnu** 5 months, 2 weeks ago
BCF 100%
upvoted 1 times

  **obaidur** 9 months, 1 week ago
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>
upvoted 2 times

  **fhuadeen** 9 months, 2 weeks ago
It is BCE:
you increase regularisation to reduce overfitting
you increase dropout to reduce overfitting
and feature combination is the combination of the strength of multiple complementary features to yield a more powerful feature (which means you are reducing number of features). Reducing number of features and using features with stronger information helps to reduce overfitting
upvoted 1 times

  **sebtac** 9 months, 4 weeks ago
BCF

it is C not D as higher value = higher ratio of 0-valued weights = higher regularization = less overfitting
upvoted 1 times

- 🗳️ 👤 **syu31svc** 10 months ago
BCF for sure
upvoted 2 times
- 🗳️ 👤 **algorithmish** 11 months, 3 weeks ago
overfitting.
B - increase regularisation helps to generalize
C - increase dropout (probability of a node to be turned off) also helps to generalize
F - decreasing feature combinations (removing features like mean or variance, artificial features) helps to generalize
upvoted 4 times
- 🗳️ 👤 **ejj** 11 months, 3 weeks ago
i think BCE , E because if we increase features combination, we also decrease the number of feature
upvoted 2 times
- 🗳️ 👤 **tff** 12 months ago
B, C, F for sure
upvoted 1 times
- 🗳️ 👤 **DScode** 1 year ago
BCF is the combination to go.
for drop-out value of 1 means no output (or 100% filtering out). Features should definitely decrease as more features tends to overfit.
upvoted 2 times
- 🗳️ 👤 **mrpwny2** 1 year, 1 month ago
BDF
Reg & Dropout go in the opposite direction. Increase in regularization & decrease in dropout do the same thing. F less features is better. E is wrong as combining the features doesnt imply you are dropping the single ones
upvoted 1 times
- 🗳️ 👤 **VB** 1 year, 3 months ago
I thin B,D,E are correct.
B. Increase regularization. -> regularization can be increased to avoid overfitting
D. Decrease dropout. -> dropout values go from 0 (all removed, no output) to 1 (all allowed), so, decreasing the dropout value to 0.9 or 0.8 0.7 meaning more and more nodes are removed..so, this option is correct.
E. Increase feature combinations. -> combining MORE features to reduce the total number of features or bring newer combined features.

I think these are the right answers.
upvoted 4 times
- 🗳️ 👤 **VB** 1 year, 2 months ago
BCF ..
upvoted 3 times
- 🗳️ 👤 **rajs** 1 year, 4 months ago
Agreed with B & F

Let me explain why D not C
Dropout =1 For no dropout whereas Dropout = 0 is for ignoring the entire layer so to make dropout effective you need to decrease its value hence the correct answer is D
upvoted 5 times
- 🗳️ 👤 **devsean** 1 year, 4 months ago
BCE is correct. I believe that they mean combining the features rather than just adding new combinations of the features. I.e. X, Y goes to XY not X, Y, XY
upvoted 6 times


A Data Scientist needs to create a serverless ingestion and analytics solution for high-velocity, real-time streaming data.

The ingestion process must buffer and convert incoming records from JSON to a query-optimized, columnar format without data loss. The output datastore must be highly available, and Analysts must be able to run SQL queries against the data and connect to existing business intelligence dashboards.

Which solution should the Data Scientist build to satisfy the requirements?


- A. Create a schema in the AWS Glue Data Catalog of the incoming data format. Use an Amazon Kinesis Data Firehose delivery stream to stream the data and transform the data to Apache Parquet or ORC format using the AWS Glue Data Catalog before delivering to Amazon S3. Have the Analysts query the data directly from Amazon S3 using Amazon Athena, and connect to BI tools using the Athena Java Database Connectivity (JDBC) connector.
- B. Write each JSON record to a staging location in Amazon S3. Use the S3 Put event to trigger an AWS Lambda function that transforms the data into Apache Parquet or ORC format and writes the data to a processed data location in Amazon S3. Have the Analysts query the data directly from Amazon S3 using Amazon Athena, and connect to BI tools using the Athena Java Database Connectivity (JDBC) connector.
- C. Write each JSON record to a staging location in Amazon S3. Use the S3 Put event to trigger an AWS Lambda function that transforms the data into Apache Parquet or ORC format and inserts it into an Amazon RDS PostgreSQL database. Have the Analysts query and run dashboards from the RDS database.
- D. Use Amazon Kinesis Data Analytics to ingest the streaming data and perform real-time SQL queries to convert the records to Apache Parquet before delivering to Amazon S3. Have the Analysts query the data directly from Amazon S3 using Amazon Athena and connect to BI tools using the Athena Java Database Connectivity (JDBC) connector.

Correct Answer: A

 **DonaldCMLIN** Highly Voted 1 year, 7 months ago
Kinesis Data Analytics NO PARQUET FORMAT,
BESIDES THAT JSON NO NEED TO STORE IN S3.
RDS ISN'T serverless ingestion and analytics solution

ANSWER IS A.
upvoted 21 times

 **georgeZ** Highly Voted 1 year, 3 months ago
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/>
upvoted 10 times

 **harmanbirstudy** Most Recent 5 months, 3 weeks ago
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 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-a-amazon-redshift/>
upvoted 4 times

- 🗨️ **lightblue** 6 months, 1 week ago
<https://aws.amazon.com/blogs/aws/new-serverless-streaming-etl-with-aws-glue/>
A is the answer imo
upvoted 1 times
- 🗨️ **oMARKOo** 9 months ago
Should be A.
<https://docs.aws.amazon.com/firehose/latest/dev/record-format-conversion.html>
it states that " Use AWS Glue to create a schema in the AWS Glue Data Catalog. Kinesis Data Firehose then references that schema and us it to interpret your input data. "
upvoted 1 times
- 🗨️ **williamsuning** 9 months, 1 week ago
I was initially about to choose A, but I think the answer A would be better if using Kinesis Data stream before firehose, since data passed to firehose has to be loss-tolerant during transfer and conflicts with "must buffer" during ingestion process.
upvoted 1 times
- 🗨️ **syu31svc** 10 months ago
Answer is A
Following links support it:
<https://docs.aws.amazon.com/athena/latest/ug/connect-with-jdbc.html>
https://aws.amazon.com/about-aws/whats-new/2018/05/stream_real_time_data_in_apache_parquet_or_orc_format_using_firehose/
upvoted 2 times
- 🗨️ **Urban_Life** 12 months ago
Answer 'A' is confusing. Still don't understand rational behind answer 'B'. Answer C & D is not even close. Anybody else has 2nd opinion? Either A or B?
upvoted 2 times
- 🗨️ **Achievement** 11 months, 2 weeks ago
because it's streaming data, so not choose B?
upvoted 1 times
- 🗨️ **Sadhna** 1 year, 1 month ago
I think A is correct answer
upvoted 2 times
- 🗨️ **dhs227** 1 year, 2 months ago
The correct answer HAS TO be A. Other choices are just too wacky
upvoted 4 times
- 🗨️ **PRC** 1 year, 3 months ago
A it is...
upvoted 2 times
- 🗨️ **NNR** 1 year, 5 months ago
But Firehose does not interact with Glue for delivery to S3.. Firehose can convert JSON to ORC itself and deliver in S3 for downstream consumption... Shouldn't be B
upvoted 3 times
- 🗨️ **sdsfsdsf** 1 year, 3 months ago
Yeah, according to: <https://aws.amazon.com/glue/faqs/>
Glue Data Catalog does not work with Firehose (it works with Athena, Redshift, EMR). I'd say B is the solution
upvoted 1 times
- 🗨️ **lightblue** 6 months, 1 week ago
It is not required for glue to be attached to firehose in anyway. after data is in s3 glue will come into picture.
upvoted 1 times

Question #46

Topic 1

An online reseller has a large, multi-column dataset with one column missing 30% of its data. A Machine Learning Specialist believes that certain columns in the dataset could be used to reconstruct the missing data.

Which reconstruction approach should the Specialist use to preserve the integrity of the dataset?

- A. Listwise deletion
- B. Last observation carried forward
- C. Multiple imputation
- D. Mean substitution

Correct Answer: C

Reference:

<https://worldwidescience.org/topicpages/i/imputing+missing+values.html>

 **rsimham** Highly Voted 1 year, 6 months ago

C looks correct since multiple imputation can be performed based on the related variable as given in the question
upvoted 19 times

 **harmanbirstudy** Most Recent 5 months, 3 weeks ago

Multiple Imputation by Chained Equations or MICE, as per udemy this is always the best answer of all
upvoted 2 times

 **syu31svc** 10 months ago

If it's handling missing data then imputation comes into play
Answer is C 100%
upvoted 1 times

 **Wira** 1 year ago

<https://www.countants.com/blogs/heres-how-you-can-configure-automatic-imputation-of-missing-data/> C
upvoted 1 times

 **roytruong** 1 year, 1 month ago

it's C
upvoted 1 times

 **dhs227** 1 year, 2 months ago

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>
upvoted 2 times

Question #47

Topic 1

A company is setting up an Amazon SageMaker environment. The corporate data security policy does not allow communication over the internet. How can the company enable the Amazon SageMaker service without enabling direct internet access to Amazon SageMaker notebook instances?

- A. Create a NAT gateway within the corporate VPC.
- B. Route Amazon SageMaker traffic through an on-premises network.
- C. Create Amazon SageMaker VPC interface endpoints within the corporate VPC.
- D. Create VPC peering with Amazon VPC hosting Amazon SageMaker.

Correct Answer: A

Reference:


<https://docs.aws.amazon.com/sagemaker/latest/dg/sagemaker-dg.pdf>
(46)

  **DonaldCMLIN** Highly Voted 1 year, 7 months ago

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
upvoted 19 times

  **rsimham** 1 year, 6 months ago

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

upvoted 2 times

  **Selectron** 1 year, 1 month ago

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.

upvoted 4 times

  **rsimham** 1 year, 6 months ago

A may the right answer

upvoted 1 times

  **tap123** Highly Voted 1 year, 4 months ago

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>

upvoted 11 times

  **gcpwhiz** Most Recent 2 months, 2 weeks ago

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 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.

upvoted 1 times

  **scuzzy2010** 4 months, 4 weeks ago

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."

upvoted 1 times

  **cloud_trail** 5 months ago



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.

upvoted 2 times

  **Omar_Cascudo** 8 months ago

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.

upvoted 2 times

  **weslleylc** 8 months, 2 weeks ago

Notebooks are Internet-enabled by default. If disabled, your VPC needs an interface endpoint (PrivateLink) or NAT Gateway, and allow outbound connections, for training and hosting to work. A doesn't have an outbound connection allowed the correct answer must be C.



upvoted 2 times

  **syu31svc** 10 months ago

Answer is C

<https://docs.aws.amazon.com/sagemaker/latest/dg/interface-vpc-endpoint.html>

upvoted 1 times

  **jonclem** 11 months, 2 weeks ago

I'd be inclined to say A myself. Having Architected and built environments using PrivateLink (VPC Endpoints) the fundamental reason behind them is to keep your infra accessible in a private capacity.

So, say you are connecting from corp office to the Cloud and don't want the end-user exposing any data over the public internet, you would utilise an Endpoint connection.

@Donald... take your finger off CAPS dude ! You can make your comment without "shouting" !

upvoted 4 times

  **ardisch** 1 year ago

The right answer is C

<https://docs.aws.amazon.com/sagemaker/latest/dg/sagemaker-dg.pdf> (927)

upvoted 1 times

- 🗳️ 👤 **kumarvn** 1 year ago
Both A & C are correct
upvoted 1 times
- 🗳️ 👤 **kumarvn** 1 year ago
Notebooks are internet enabled by default. If disabled, VPC needs an interface endpoint or NAT gateway, and allow outbound connection for training and hosting to work.
upvoted 1 times
- 🗳️ 👤 **C10ud9** 1 year ago
Has to be C
upvoted 2 times
- 🗳️ 👤 **roytruong** 1 year, 1 month ago
C is correct, NAT used when company need to access internet from VPC, VPC interface endpoint allow instances in VPC access to the other AWS service, see: <https://docs.aws.amazon.com/vpc/latest/userguide/vpce-interface.html>
upvoted 1 times
- 🗳️ 👤 **aws_razor** 1 year, 1 month ago
Answer should be A
To have VPC without Direct Internet access.
Notebook needs to be in either be in a private subnet with a NAT or to access the internet back through a virtual private gateway.
If you just use VPC with Interface Gateway in that case Traffic to Interface VPC Endpoints will still go through your VPC.
BUT Traffic to Gateway VPC Endpoints like Amazon S3 and DynamoDB will go through the public internet AND SageMaker provides a network interface that allows the notebook to communicate with the internet through a VPC managed by Amazon SageMaker(Public Internet)
upvoted 3 times
- 🗳️ 👤 **dhs227** 1 year, 2 months ago
The correct answer HAS TO BE C

It says no communication over internet, period. So NAT is out.
upvoted 2 times
- 🗳️ 👤 **Phong** 1 year, 4 months ago
The question doesn't mention Hosting so NAT seems not suitable. C is better.
upvoted 3 times
- 🗳️ 👤 **devsean** 1 year, 4 months ago
A is correct.
"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 connections

From here on page 212 (pdf page 220), bullet "g":
<https://docs.aws.amazon.com/sagemaker/latest/dg/sagemaker-dg.pdf>
upvoted 3 times
- 🗳️ 👤 **georgeZ** 1 year, 3 months ago
but you need to allow outbound connections when using NAT.
upvoted 1 times

Question #48

Topic 1

A Machine Learning Specialist is training a model to identify the make and model of vehicles in images. The Specialist wants to use transfer learning and an existing model trained on images of general objects. The Specialist collated a large custom dataset of pictures containing different vehicle makes and models.

What should the Specialist do to initialize the model to re-train it with the custom data?

- A. Initialize the model with random weights in all layers including the last fully connected layer.
- B. Initialize the model with pre-trained weights in all layers and replace the last fully connected layer.
- C. Initialize the model with random weights in all layers and replace the last fully connected layer.
- D. Initialize the model with pre-trained weights in all layers including the last fully connected layer.

Correct Answer: B

 **rsimham** Highly Voted 1 year, 6 months ago

Ans B sounds correct
upvoted 19 times

 **DzR** Most Recent 5 months, 3 weeks ago

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.
upvoted 1 times

 **bobdylan1** 6 months ago

B is correct
upvoted 1 times

 **sebtac** 9 months, 4 weeks ago

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.
upvoted 1 times

 **syu31svc** 10 months ago

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 type. So answer is B
upvoted 2 times

🗨️ 👤 **scuzzy2010** 10 months, 4 weeks ago

I feel the answer is "C" because it was not trained on images of "similar" objects, it was trained on images of "general" objects. If it were trained on similar objects, then answer would be "B".

upvoted 1 times

🗨️ 👤 **cloud_trail** 5 months ago

If you do option C, you're not doing transfer learning at all. You might just as well start over with your own architecture.

upvoted 2 times

🗨️ 👤 **yeetusdeleetus** 7 months, 4 weeks ago

Transfer learning doesn't work like that. All images are 'somewhat' similar, so a pretrained neural network that was trained on different type of images is still better than random weights. B.

upvoted 1 times

🗨️ 👤 **tff** 12 months ago

<https://www.hackerearth.com/practice/machine-learning/transfer-learning/transfer-learning-intro/tutorial/>

answer is B.

Read the scenario:

The target dataset is large and similar to the base training dataset.

Since the target dataset is large, we have more confidence that we won't overfit if we try to fine-tune through the full network. Therefore, we

- 1.Remove the last fully connected layer and replace with the layer matching the number of classes in the target dataset;
- 2.Randomly initialize the weights in the new fully connected layer;
- 3.Initialize the rest of the weights using the pre-trained weights, i.e., unfreeze the layers of the pre-trained network;
- 4.Retrain the entire neural network;

So, first look at 3, and then at 1.

upvoted 3 times

🗨️ 👤 **fw** 12 months ago

From what I read in <https://kharshit.github.io/blog/2018/08/10/transfer-learning>

Choice seems to be C.

Case III: Large dataset, Similar data

remove the last fully connected layer and replace with a layer matching the number of classes in the new data set

randomly initialize the weights in the new fully connected layer

initialize the rest of the weights using the pre-trained weights

re-train the entire neural network

Case IV: Large dataset, Different data

remove the last fully connected layer and replace with a layer matching the number of classes in the new data set

retrain the network from scratch with randomly initialized weights

alternatively, you could just use the same strategy as the "large and similar" data case

upvoted 1 times

🗨️ 👤 **C10ud9** 1 year ago

B is correct, replace only the last layer in transfer learning

upvoted 1 times

🗨️ 👤 **roytruong** 1 year, 1 month ago

B is right

upvoted 2 times

Topic 1

An office security agency conducted a successful pilot using 100 cameras installed at key locations within the main office. Images from the cameras were uploaded to Amazon S3 and tagged using Amazon Rekognition, and the results were stored in Amazon ES. The agency is now looking to expand the pilot into a full production system using thousands of video cameras in its office locations globally. The goal is to identify activities performed by non-employees in real time

Which solution should the agency consider?

- 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.
- B. 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 Image to detect faces from a collection of known employees and alert when non-employees are detected.
- C. Install AWS DeepLens cameras and use the DeepLens_Kinesis_Video module to stream video to Amazon Kinesis Video Streams for each camera. On each stream, use Amazon Rekognition Video and create a stream processor to detect faces from a collection on each stream, and alert when non-employees are detected.
- D. Install AWS DeepLens cameras and use the DeepLens_Kinesis_Video module to stream video to Amazon Kinesis Video Streams for each camera. On each stream, run an AWS Lambda function to capture image fragments and then call Amazon Rekognition Image to detect faces from a collection of known employees, and alert when non-employees are detected.

Correct Answer: D

Reference:

<https://aws.amazon.com/blogs/machine-learning/video-analytics-in-the-cloud-and-at-the-edge-with-aws-deeplens-and-kinesis-video-streams/>

 **scuzzy2010** Highly Voted 10 months, 4 weeks ago

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. "

upvoted 10 times

🗨️ 👤 **WWODIN** Highly Voted 👍 1 year, 5 months ago

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?

upvoted 9 times

🗨️ 👤 **sdsfsdsf** 1 year, 3 months ago

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.

upvoted 8 times

🗨️ 👤 **cybe001** 1 year, 5 months ago

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.

upvoted 4 times

🗨️ 👤 **scuzzy2010** 10 months, 4 weeks ago

DeepLens is for developers only, it is not available as a commercial product.

upvoted 3 times

🗨️ 👤 **StelSen** Most Recent ⌚ 1 month, 3 weeks ago

I will go with 'A'. Because setting up DeepLens will cost S\$304.59. Whereas setting up proxy server using Raspberry PI (production equivalent) will take around S\$50 + Camera may be \$50. So overall solution is for A might be \$100. Company already did a pilot with some existing camera. So why intro DeepLens now. Use cost effective solution.

upvoted 2 times

🗨️ 👤 **srinu3054** 3 months, 1 week ago

Its C!

<https://docs.aws.amazon.com/rekognition/latest/dg/streaming-video.html>

upvoted 2 times

🗨️ 👤 **SophieSu** 4 months, 1 week ago

D is correct. Because you DEFINITELY need the Lambda Function to "CALL" the Amazon Rekognition to start the job (detect faces)!

upvoted 1 times

🗨️ 👤 **achiko** 3 months, 1 week ago

C. You don't need Lambda between video stream and rekognition

upvoted 2 times

🗨️ 👤 **SophieSu** 4 months, 1 week ago

D is the correct answer. The point is that the lambda function is necessary before the Amazon Rekognition. See "<https://docs.aws.amazon.com/rekognition/latest/dg/stored-video-lambda.html>". "You can use Lambda functions with Amazon Rekognition Video operations. For example, the following diagram shows a website that uses a Lambda function to automatically start analysis of a video when it's uploaded to an Amazon S3 bucket. When the Lambda function is triggered, it calls StartLabelDetection to start detecting labels in the uploaded video. For information about using Lambda to process event notifications from an Amazon S3 bucket, see Using AWS Lambda with Amazon S3 Events."

upvoted 1 times

🗨️ 👤 **Aashi22** 4 months, 1 week ago

Confusion b/w option C & D. Answer should be option D as per link <https://aws.amazon.com/blogs/machine-learning/video-analytics-in-the-cloud-and-at-the-edge-with-aws-deeplens-and-kinesis-video-streams/>

upvoted 2 times

🗨️ 👤 **scuzzy2010** 4 months, 4 weeks ago

Answer is "A". C and D are out because DeepLens is NOT a commercial product, it's sold only to developers.

A mentions 'stream processor', and as mentioned here is specifically used to detect and recognise faces in streaming video ->

(https://docs.aws.amazon.com/rekognition/latest/dg/API_CreateStreamProcessor.html) "Amazon Rekognition stream processor that you can use to detect and recognize faces in a streaming video."

upvoted 4 times

🗨️ **cloud_trail** 5 months ago

Option D. I think that the goal for the company is to do the facial recognition at the edge, which is what DeepLens does. The pilot program cameras were not ML enabled so they had to upload all all images from all cameras to S3 and then use Rekognition. Not practical with 100 of "dumb" cameras. DL can do object detection and the call Rekognition to identify non-employees and the send alerts.

upvoted 1 times

🗨️ **harmanbirstudy** 5 months, 3 weeks ago

Deep lens cannot be used at large scale so it cannot be C or D

--Between A and B , A mentions Amazon Rekognition Video which is used for detection on video data.

-- Hence ANSWER is A

upvoted 2 times

🗨️ **Potato_Noodle** 8 months, 3 weeks ago

The pilot program is done by using image, I wonder what would happen all of the sudden the solution is now done by using video, imagine cost incurred by processing and storing these info, so D looks more sensible

upvoted 3 times

🗨️ **oMARKOo** 9 months ago

Regarding the official page for deeplans, deeplans has capability for face and object detection. This event could trigger the lambda function which will call built-in Amazon Recognition model trained on pictures of employees.

I think that D is the right answer.

upvoted 1 times

🗨️ **Th3Dud3** 9 months, 3 weeks ago

Hmmm. >>>Use a proxy server at each local office and for each camera<<< one server per camera? Wow!

upvoted 1 times

🗨️ **ml_sexpert** 11 months, 3 weeks ago

Installing thousands of AWS DeepLens cameras globally is laughable. C & D can be immediately discredited.

upvoted 1 times

🗨️ **scuzzy2010** 10 months, 4 weeks ago

Especially if it's not even sold for production use. It is purely a developers tool.

upvoted 1 times

🗨️ **C10ud9** 1 year ago

Agree C

upvoted 2 times

🗨️ **dhs227** 1 year, 2 months ago

The answer HAS TO BE C.

A and B are eliminated to begin with. Deploying thousands of camera globally while managing who knows how many proxy servers at each location is just ridiculous.

Between C and D, you know better. Fully managed service is better in this use case.

upvoted 1 times

Question #50

Topic 1

A Marketing Manager at a pet insurance company plans to launch a targeted marketing campaign on social media to acquire new customers.

Currently, the company has the following data in Amazon Aurora:

- ☞ Profiles for all past and existing customers
- ☞ Profiles for all past and existing insured pets
- ☞ Policy-level information
- ☞ Premiums received
- ☞ Claims paid

What steps should be taken to implement a machine learning model to identify potential new customers on social media?

- A. Use regression on customer profile data to understand key characteristics of consumer segments. Find similar profiles on social media
- B. Use clustering on customer profile data to understand key characteristics of consumer segments. Find similar profiles on social media
- C. Use a recommendation engine on customer profile data to understand key characteristics of consumer segments. Find similar profiles on social media.
- D. Use a decision tree classifier engine on customer profile data to understand key characteristics of consumer segments. Find similar profiles on social media.

Correct Answer: C

 **DonaldCMLIN** Highly Voted 1 year, 7 months ago

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 are not used.

<https://docs.aws.amazon.com/sagemaker/latest/dg/algorithm-kmeans-tech-notes.html>

THE ANSWER COULD BE B. clustering on customer profile data to understand key characteristic


upvoted 19 times

 **haison8x** 11 months ago

<https://towardsdatascience.com/customer-segmentation-with-machine-learning-a0ac8c3d4d84>

B

upvoted 2 times

 **rsimham** 1 year, 6 months ago

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

upvoted 10 times

 **Vita_Rasta84444** Most Recent 2 months, 2 weeks ago

I agree, the Answer is B. We do not have a new uncategorized customers, on the system, but we should find similar customers on the web which we can do by clustering, making profile of customer with insured pet.

upvoted 1 times

 **astonm13** 4 months, 2 weeks ago

Considering the quote "to understand characteristics of customer segments", I would say recommender engine does not fit to it. I would go for the clustering (answer B). It is not supervised, cause we don't have info on those people who were targeted but refused to become our customers.

upvoted 1 times

 **ss001maryam** 4 months, 3 weeks ago

Answer is B. recommendation engine is correct if we want to make decisions on current customers, here we want to make decision on new customers as well. So we need clustering.

upvoted 1 times

🗨️ 👤 **cloud_trail** 5 months ago

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 you are comparing them to.

upvoted 2 times

🗨️ 👤 **harmanbirstudy** 5 months, 3 weeks ago

Recommendation Engine would make sense if we are recommending someone coming on our website and we recommend something there. Here we need to go to social media, where we cannot apply our own machine learning model, which means we need to understand the data before targeting new customers on the social media ads.

Hence among the list only option that makes sense is Clustering, so Answer is B

upvoted 2 times

🗨️ 👤 **R_cool** 6 months ago

what is the real answer

upvoted 1 times

🗨️ 👤 **syu31svc** 10 months ago

Answer is B; recommendation engine does not identify new customers as it suggests products to be given based on profiles

upvoted 2 times

🗨️ 👤 **GeeBeeEI** 11 months ago

C is not correct at all. A recommendation engine would probably be a factorization machine. This is a grouping problem, a clustering problem.....it is not a recommendation engine! B is the correct answer!!!

upvoted 3 times

🗨️ 👤 **roytruong** 1 year, 1 month ago

Go for C, find new user on social media that is similar with our current user, it's called user-user collaborative filtering

upvoted 2 times

🗨️ 👤 **DScode** 1 year ago

sorry, but any kind of collaborative filtering needs an interaction data, which is not present, and for new user, collaborative filtering suffers from cold start problems since they don't have data beforehand. This is a pure case of unsupervised market segmentation problem, since we do not have any labels. Thus, by far and wide, the answer should be Clustering, i.e. option B.

upvoted 8 times

🗨️ 👤 **lightblue** 6 months, 1 week ago

we have labels - past customers and present customers

upvoted 1 times

🗨️ 👤 **lightblue** 6 months, 1 week ago

D can be the answer imo

upvoted 1 times

🗨️ 👤 **hughhughhugh** 1 year, 2 months ago



why not D? marketing campaigns mainly use classifiers

upvoted 4 times

🗨️ 👤 **mirik** 1 year, 2 months ago

what is the difference between old and new customer? The premiums received. So, here we can assume premiums as labels and have supervised learning. Please correct me if I'm mistaken.

upvoted 1 times

  **rickywck** 1 year, 3 months ago

I would pick C but honest speaking, the question content is not entirely clear.

Very often when we launch a marketing campaign, there is specific target(s) in mind, e.g. selling specific type of insurance in this case, and usually we then analyze the profile of existing customers whose have that product and then try to target those with similar profile. This wha most recommendation engine does.

unvoted 2 times

Question #51



Topic 1

A manufacturing company has a large set of labeled historical sales data. The manufacturer would like to predict how many units of a particular part should be produced each quarter.

Which machine learning approach should be used to solve this problem?

- A. Logistic regression
- B. Random Cut Forest (RCF)
- C. Principal component analysis (PCA)
- D. Linear regression



Correct Answer: B



  **DonaldCMLIN** Highly Voted 1 year, 7 months ago
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.


upvoted 34 times

  **rsimham** 1 year, 6 months ago
agree. RCF is mostly used for anomaly detection or separate outliers
upvoted 7 times

  **syu31svc** Highly Voted 10 months ago
Amazon SageMaker Random Cut Forest (RCF) is an unsupervised algorithm for detecting anomalous data points within a data set
Answer is D 100%
upvoted 6 times

  **AWS__Newbie** Most Recent 4 weeks, 1 day ago
Where does the answer come from? RCF is for anomaly detection. How can it be the right answer?
upvoted 1 times

  **littlewat** 3 months, 2 weeks ago
D!!!!!!
upvoted 2 times

  **astonm13** 4 months, 2 weeks ago

Question #52

Topic 1

A financial services company is building a robust serverless data lake on Amazon S3. The data lake should be flexible and meet the following requirements:

- ☞ Support querying old and new data on Amazon S3 through Amazon Athena and Amazon Redshift Spectrum.
- ☞ Support event-driven ETL pipelines
- ☞ Provide a quick and easy way to understand metadata

Which approach meets these requirements?

- A. Use an AWS Glue crawler to crawl S3 data, an AWS Lambda function to trigger an AWS Glue ETL job, and an AWS Glue Data catalog to search and discover metadata.
- B. Use an AWS Glue crawler to crawl S3 data, an AWS Lambda function to trigger an AWS Batch job, and an external Apache Hive metastore to search and discover metadata.
- C. Use an AWS Glue crawler to crawl S3 data, an Amazon CloudWatch alarm to trigger an AWS Batch job, and an AWS Glue Data Catalog to search and discover metadata.
- D. Use an AWS Glue crawler to crawl S3 data, an Amazon CloudWatch alarm to trigger an AWS Glue ETL job, and an external Apache Hive metastore to search and discover metadata.

Correct Answer: A

🗨️ **DonaldCMLIN** Highly Voted 1 year, 7 months ago

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 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.

upvoted 26 times

🗨️ **rsimham** 1 year, 6 months ago

I am thinking about Answer C, because events can be triggered by cloudwatch w/Glue metastore

upvoted 1 times

🗨️ **qwerty456** 11 months, 3 weeks ago

you can't schedule AWS Batch with CloudWatch

upvoted 1 times

🗨️ **kalyanvarma** 6 months, 1 week ago

We can schedule batch with cloud watch events.

upvoted 1 times

🗨️ **qwerty456** 11 months, 3 weeks ago

srr, looks like you can apart from Cron, the argument should be AWS Batch aren't SERVERLESS

upvoted 1 times

🗨️ **ComPah** 1 year, 6 months ago

if we use Flexible as key word ..Using Lambda might be a constraint

upvoted 3 times

🗨️ **cybe001** Highly Voted 1 year, 5 months ago

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.

upvoted 13 times

🗨️ **syu31svc** Most Recent 10 months ago

Answer is A 100%

upvoted 1 times

🗨️ **halfway** 1 year ago

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

upvoted 2 times

🗨️ **PRC** 1 year, 3 months ago

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

upvoted 3 times

Topic 1

A company's Machine Learning Specialist needs to improve the training speed of a time-series forecasting model using TensorFlow. The training is currently implemented on a single-GPU machine and takes approximately 23 hours to complete. The training needs to be run daily. The model accuracy is acceptable, but the company anticipates a continuous increase in the size of the training data and a need to update the model on an hourly, rather than a daily, basis. The company also wants to minimize coding effort and infrastructure changes. What should the Machine Learning Specialist do to the training solution to allow it to scale for future demand?

- A. Do not change the TensorFlow code. Change the machine to one with a more powerful GPU to speed up the training.
- B. Change the TensorFlow code to implement a Horovod distributed framework supported by Amazon SageMaker. Parallelize the training to as many machines as needed to achieve the business goals.
- C. Switch to using a built-in AWS SageMaker DeepAR model. Parallelize the training to as many machines as needed to achieve the business goals.
- D. Move the training to Amazon EMR and distribute the workload to as many machines as needed to achieve the business goals.

Correct Answer: B

  **JayK** Highly Voted 1 year, 5 months ago

the answer is B. using Horovod distribution results in less coding effort
upvoted 19 times

  **cybe001** Highly Voted 1 year, 5 months ago

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.
upvoted 7 times

  **cloud_trail** Most Recent 5 months ago

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.
upvoted 3 times

  **RaniaSayed** 5 months ago

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/>
upvoted 1 times

  **harmanbirstudy** 5 months, 3 weeks ago

Seen similar question on udemy/whizlab , its always Horovod when Tensorflow needs scaling. ANSWER is B
upvoted 4 times

🗨️ **yeetusdeleetus** 7 months, 4 weeks ago

"Minimize code and infrastructure changes" and "training set size increasing continuously" are in conflict.

The latter is necessary, while the former is more optional. This rules out A, because it will not scale to continuously-increasing dataset size daily training.

C requires a total rewrite, which does not minimize coding effort.

This leaves B and D, and the question will be which is easier to implement. EMR is not, especially, known for being easy to implement, whereas Horovod + SageMaker specifically are. Also SageMaker is a named AWS service and is extremely expensive, so of course that's what Amazon recommends. :)

upvoted 1 times

🗨️ **sebtac** 9 months, 4 weeks ago

correct answer is B; the expected probably is C as AWS wants us to use their own solutions :)

upvoted 2 times

🗨️ **syu31svc** 10 months ago

I would say that the answer is A and this is just my reasoning:

Minimize infrastructure changes would mean saving costs. Having to increase the number of machines to "as many as possible" would increase costs. Also, from <https://d1.awsstatic.com/whitepapers/aws-power-ml-at-scale.pdf>:

If the training times on a GPU card are insufficient for your business needs, we recommend that you try a more powerful GPU before moving to multiple GPUs

upvoted 1 times

🗨️ **mawsmann** 1 year, 1 month ago

Possible B or C because of scaling and future needs. But Switching TensorFlow to DeepAT for time-series forecasting would be more effort than re-coding to horovod. I would choose B.

upvoted 3 times

🗨️ **deep_n** 1 year, 1 month ago

Correct answer is B

upvoted 2 times

🗨️ **VB** 1 year, 3 months ago

The question does not talk about running that application in AWS.. the question is very general ML related..and "...The company also wants minimize coding effort and infrastructure changes...". So, can answer be A ?

upvoted 2 times

🗨️ **Erso** 1 year, 2 months ago

Correct point of view! With B you have to change the code and implement infrastructure change and the question says "The company wants to minimize coding effort and infrastructure changes"...I'll go with A

upvoted 1 times

🗨️ **BigPlums** 1 year, 2 months ago

True but does it "allow it to scale for future demand" ?

upvoted 1 times

🗨️ 👤 **tap123** 1 year, 4 months ago

Why not A?

upvoted 2 times

🗨️ 👤 **Littlefishfish** 1 year ago

I opt A.

<https://aws.amazon.com/blogs/machine-learning/launching-tensorflow-distributed-training-easily-with-horovod-or-parameter-servers-in-amazon-sagemaker/>

"Before moving to distributed training in a cluster, make sure that you have first tried scaling up on a single machine with multiple GPUs. Communication between multiple GPUs on a single machine is faster than communicating across a network between multiple machines. For more details, see the AWS whitepaper *Power Machine Learning at Scale*."

upvoted 4 times

🗨️ 👤 **lightblue** 6 months, 1 week ago

not sure if they mean multiple GPU by this -> 'Change the machine to one with a more powerful GPU'

upvoted 2 times

🗨️ 👤 **rsimham** 1 year, 6 months ago

I think Answer is C.

<https://aws.amazon.com/blogs/machine-learning/now-available-in-amazon-sagemaker-deepar-algorithm-for-more-accurate-time-series-forecasting/>

upvoted 1 times

🗨️ 👤 **ComPah** 1 year, 6 months ago

Looks like C if you take minimize coding as key word

upvoted 1 times

🗨️ 👤 **ComPah** 1 year, 5 months ago

Its B from Horovod github page

The primary motivation for this project is to make it easy to take a single-GPU TensorFlow program and successfully train it on many GPUs faster.

Question #54

Topic 1

Which of the following metrics should a Machine Learning Specialist generally use to compare/evaluate machine learning classification models against each other?

- A. Recall
- B. Misclassification rate
- C. Mean absolute percentage error (MAPE)
- D. Area Under the ROC Curve (AUC)

Correct Answer: D

🗨️ 👤 **DonaldCMLIN** Highly Voted 🍌 1 year, 7 months ago

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.

upvoted 22 times

- 🗨️ 👤 **devsean** 1 year, 4 months ago
AUC is to determine hyperparams in a single model, not compare different models.
upvoted 5 times
- 🗨️ 👤 **DScode** 1 year ago
Not might be, but should be D
upvoted 2 times
- 🗨️ 👤 **cloud_trail** Most Recent 5 months ago
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.
upvoted 2 times
- 🗨️ 👤 **harmanbirstudy** 5 months, 3 weeks ago
AUC/ROC work well with special case of Binary Classification not in general
upvoted 2 times
- 🗨️ 👤 **MohamedSharaf** 5 months, 1 week ago
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.
upvoted 3 times
- 🗨️ 👤 **harmanbirstudy** 5 months, 3 weeks ago
Its Accuracy, Precision, Recall and F1 score, there is no mention of AUC/ROC for comparing models in many articles, so ANSWER is A
upvoted 1 times
- 🗨️ 👤 **Thai_Xuan** 8 months, 4 weeks ago
D. AUC is scale- and threshold-invariant, enabling it compare models.
<https://towardsdatascience.com/how-to-evaluate-a-classification-machine-learning-model-d81901d491b1>
upvoted 1 times
- 🗨️ 👤 **johnny_chick** 12 months ago
Actually A, B and D seem to be correct
upvoted 1 times
- 🗨️ 👤 **deep_n** 1 year, 1 month ago
Probably D
<https://towardsdatascience.com/metrics-for-evaluating-machine-learning-classification-models-python-example-59b905e079a5>
upvoted 1 times
- 🗨️ 👤 **hughhughhugh** 1 year, 2 months ago
why not B?
upvoted 1 times
- 🗨️ 👤 **PRC** 1 year, 3 months ago
Answer should be D..ROC is used to determine the diagnostic capability of classification model varying on threshold
upvoted 3 times

  **Hypermasterd** 1 year, 3 months ago

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

upvoted 4 times

  **oMARKOo** 9 months ago

Actually AUC could be generalized for multi-class problem.
<https://www.datascienceblog.net/post/machine-learning/performance-measures-multi-class-problems/>

upvoted 1 times

  **sebas10** 1 year ago

Could be, you mean in a multiclass classification problem. But in that context recall directly can't be compared 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 ?

upvoted 2 times

Question #55

Topic 1

A company is running a machine learning prediction service that generates 100 TB of predictions every day. A Machine Learning Specialist must generate a visualization of the daily precision-recall curve from the predictions, and forward a read-only version to the Business team.

Which solution requires the LEAST coding effort?

- A. Run a daily Amazon EMR workflow to generate precision-recall data, and save the results in Amazon S3. Give the Business team read-only access to S3.
- B. Generate daily precision-recall data in Amazon QuickSight, and publish the results in a dashboard shared with the Business team.
- C. Run a daily Amazon EMR workflow to generate precision-recall data, and save the results in Amazon S3. Visualize the arrays in Amazon QuickSight, and publish them in a dashboard shared with the Business team.
- D. Generate daily precision-recall data in Amazon ES, and publish the results in a dashboard shared with the Business team.

Correct Answer: C

🗨️ 👤 **rsimham** Highly Voted 👍 1 year, 6 months ago

Ans C is reasonable
upvoted 13 times

🗨️ 👤 **cloud_trail** Most Recent 🕒 5 months ago

Agree with C. Quicksight cannot handle 100TB each day.
upvoted 2 times

🗨️ 👤 **Wira** 1 year ago

can someone explain why we need EMR here? Quicksight is capable of running calculations..
upvoted 1 times

🗨️ 👤 **Urban_Life** 12 months ago

Large volume.....that's why EMR is good. you can't even proceed with EC2 instance, however, Quicksight is light weight BI tool. From my own project experience, we move away from Quicksight to Tableau in middle of the deployment.
upvoted 4 times

🗨️ 👤 **oief2oi3fj23ogi23g** 12 months ago

100TB data to much for quick sight alone
upvoted 4 times

🗨️ 👤 **PRC** 1 year, 3 months ago

C is correct...100 TB daily can be handled by EMR and Quicksight (no coding) is the right solution for providing read only access to business
upvoted 2 times

🗨️ 👤 **rajs** 1 year, 4 months ago

Agree with C

BUT

B would have been more appropriate if CURVE was specified instead of data

Generate daily precision-recall CURVE (instead of data) in Amazon QuickSight, and publish the results in a dashboard shared with the Business team

upvoted 2 times

🗨️ 👤 **qwerty456** 11 months, 3 weeks ago

B won't solve the "daily scheduled" requirement
upvoted 2 times

Question #56


Topic 1



A Machine Learning Specialist is preparing data for training on Amazon SageMaker. The Specialist is using one of the SageMaker built-in algorithms for the training. The dataset is stored in .CSV format and is transformed into a numpy.array, which appears to be negatively affecting the speed of the training.



What should the Specialist do to optimize the data for training on SageMaker?



- A. Use the SageMaker batch transform feature to transform the training data into a DataFrame.
- B. Use AWS Glue to compress the data into the Apache Parquet format.
- C. Transform the dataset into the RecordIO protobuf format.
- D. Use the SageMaker hyperparameter optimization feature to automatically optimize the data.

Correct Answer: C

  **rsimham** Highly Voted 1 year, 6 months ago
C is okay
upvoted 12 times

  **stamarpadar** Highly Voted 1 year, 4 months ago
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>
upvoted 7 times

  **C10ud9** Most Recent 1 year ago
C is the best
upvoted 3 times

  **PRC** 1 year, 3 months ago
Agree with C
upvoted 3 times

Question #57

Topic 1

A Machine Learning Specialist is required to build a supervised image-recognition model to identify a cat. The ML Specialist performs some tests and records the following results for a neural network-based image classifier:

Total number of images available = 1,000

Test set images = 100 (constant test set)

The ML Specialist notices that, in over 75% of the misclassified images, the cats were held upside down by their owners.

Which techniques can be used by the ML Specialist to improve this specific test error?

- A. Increase the training data by adding variation in rotation for training images.
- B. Increase the number of epochs for model training

- C. Increase the number of layers for the neural network.
- D. Increase the dropout rate for the second-to-last layer.

Correct Answer: B

🗨️ 👤 **DonaldCMLIN** Highly Voted 👍 1 year, 7 months ago

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](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

upvoted 33 times

🗨️ 👤 **tap123** 1 year, 4 months ago

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.

upvoted 1 times

🗨️ 👤 **mawsmann** 1 year, 1 month ago

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

upvoted 2 times

🗨️ 👤 **Urban_Life** 12 months ago

What's your answer B?

upvoted 1 times

🗨️ 👤 **VB** 1 year, 3 months ago

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

upvoted 1 times

🗨️ 👤 **rsimham** 1 year, 6 months ago

agree with A

upvoted 8 times

🗨️ 👤 **eganilovic** Most Recent 🕒 2 months ago

The answer is A!

upvoted 2 times

🗨️ 👤 **tmld** 2 months, 2 weeks ago

My answer is A

upvoted 1 times

🗨️ 👤 **ahquiceno** 5 months ago

Answer is A, model need variation of data that CNN need learn and then predict, in this case data augmentation is the recommended step.

upvoted 1 times

- 🗨️ **wesleylc** 8 months, 2 weeks ago
Since the test set is constant, to modify your training data based on what you have saw on your test set will overfitting your model. Answer probably B.
upvoted 2 times
- 🗨️ **syu31svc** 10 months ago
The ML Specialist notices that, in over 75% of the misclassified images, the cats were held upside down by their owners. Option A would solve this problem
upvoted 2 times
- 🗨️ **Urban_Life** 12 months ago
I think the answer is A. If you add more training data with rotated image, it will mitigate the problem. I think B comes next, after answer A is solving the problem. Agree?
upvoted 1 times
- 🗨️ **Antriksh** 1 year ago
correct answer is A
upvoted 1 times
- 🗨️ **qururu** 1 year ago
The answer is A. This is a no-brainer if you have DL experience with image recognition. Data augmentation is a must have.
upvoted 2 times
- 🗨️ **C10ud9** 1 year ago
A it is
upvoted 2 times
- 🗨️ **deep_n** 1 year, 1 month ago
by far A is correct
upvoted 2 times
- 🗨️ **AKT** 1 year, 3 months ago
Answer is A. you need to do data augmentation of training images to get better results.
upvoted 1 times
- 🗨️ **Phong** 1 year, 5 months ago
It can't be A because "the cats were held upside down by their owners". The question wants to says that the owners did the data augmentation manually.
It doesn't say anything about the train accuracy so B or C is still ok.
If the train accuracy is low too, B is okay.
If B can't help, I will do C.
B came first so B is my choice.
upvoted 1 times
- 🗨️ **Phong** 1 year, 5 months ago
It can't be A because "the cats were held upside down by their owners". The question wants to says that the owners did the data augmentation manually.
It doesn't say anything about the train accuracy so B or C is still ok.
If the train accuracy is low too, B is okay.
If B can't help, I will do C.
B came first so C is my choice.
upvoted 1 times
- 🗨️ **Walz** 1 year, 4 months ago
Look more carefully. The upside down cats were in the test set not the training set. So you still need rotated examples to train on...
upvoted 6 times

Question #58

Topic 1

A Machine Learning Specialist needs to be able to ingest streaming data and store it in Apache Parquet files for exploration and analysis. Which of the following services would both ingest and store this data in the correct format?

- A. AWS DMS
- B. Amazon Kinesis Data Streams
- C. Amazon Kinesis Data Firehose
- D. Amazon Kinesis Data Analytics



Correct Answer: C

  **JayK** Highly Voted 1 year, 5 months ago



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
upvoted 35 times

  **eganilovic** Most Recent 2 months ago

Firehose
upvoted 3 times

  **Thai_Xuan** 8 months, 2 weeks ago

B
<https://github.com/ravsau/aws-exam-prep/issues/10>
upvoted 2 times

  **weslleylc** 8 months, 2 weeks ago

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.
upvoted 1 times

  **In** 9 months, 1 week ago

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/
upvoted 2 times

🗨️ **GeeBeeEI** 11 months ago

It appears all agree that the answer is between Firehose and Analytics. Kinesis Firehose is used for ingestion. Both firehose and analytics c store, only firehose can ingest. <https://docs.aws.amazon.com/firehose/latest/dev/record-format-conversion.html> shows firehose can store parquet to S3

upvoted 1 times

🗨️ **GeeBeeEI** 11 months ago

It appears all agree that the answer is between Firehose and Analytics. Data Streams handle stuff like event data, clickstream etc. Its no 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

upvoted 1 times

🗨️ **Urban_Life** 12 months ago

Think just like this -- batch process Glue ETL and Streaming process Firehose ETLcovert to parquet or any other format.

upvoted 1 times

🗨️ **CMMC** 1 year ago

C for Firehose

upvoted 2 times

🗨️ **Erso** 1 year, 2 months ago

Just in case https://acloud.guru/forums/aws-certified-big-data-specialty/discussion/-KhI3MgPEo-FY5rfgl3J/what_is_difference_between_k

upvoted 2 times

🗨️ **BigEv** 1 year, 5 months ago

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

upvoted 3 times

🗨️ **rsimham** 1 year, 6 months ago

I would go with B. Kinesis data streams stores data, while Firehose not.

upvoted 2 times

🗨️ **cloud_trail** 5 months ago

It's the other way around. Firehouses stores data; data streams does not.

upvoted 1 times

Question #59

Topic 1

A Data Scientist is developing a machine learning model to classify whether a financial transaction is fraudulent. The labeled data available for training consists of

100,000 non-fraudulent observations and 1,000 fraudulent observations.

The Data Scientist applies the XGBoost algorithm to the data, resulting in the following confusion matrix when the trained model is applied to a previously unseen validation dataset. The accuracy of the model is 99.1%, but the Data Scientist has been asked to reduce the number of false negatives.

Predicted	0	1
Actual	0 99,966	1 34
	1 877	123

Which combination of steps should the Data Scientist take to reduce the number of false positive predictions by the model? (Choose two.)

A. Change the XGBoost eval_metric parameter to optimize based on rmse instead of error.

- B. Increase the XGBoost scale_pos_weight parameter to adjust the balance of positive and negative weights.
- C. Increase the XGBoost max_depth parameter because the model is currently underfitting the data.
- D. Change the XGBoost eval_metric parameter to optimize based on AUC instead of error.
- E. Decrease the XGBoost max_depth parameter because the model is currently overfitting the data.

Correct Answer: DE

🗲️ 👤 **Paul_NoName** Highly Voted 👍 5 months ago
B and D
upvoted 7 times

🗲️ 👤 **btsql** Most Recent ⌚ 3 weeks, 6 days ago
i also think B & D.
can't guess overfitting or underfitting
upvoted 1 times

🗲️ 👤 **tmlid** 2 months, 2 weeks ago
I agree with B&D
upvoted 1 times

🗲️ 👤 **Vita_Rasta84444** 3 months ago
B and D
upvoted 1 times

🗲️ 👤 **cnethers** 4 months, 4 weeks ago
There is no mention of validation or test datasets, so you can't assume that the training is overfitting or underfitting.
So that rules out Option C, D.
That leave A, B, D
Due to the fact this is a classification problem that rules out the use of RMSE so A is not an option.
That leave B & D
ROC and AUC are classification metrics used to understand how a classification model is performing so D is 100%
upvoted 4 times

Question #60

Topic 1

A Machine Learning Specialist is assigned a TensorFlow project using Amazon SageMaker for training, and needs to continue working for an extended period with no Wi-Fi access.

Which approach should the Specialist use to continue working?



- A. Install Python 3 and boto3 on their laptop and continue the code development using that environment.
- B. Download the TensorFlow Docker container used in Amazon SageMaker from GitHub to their local environment, and use the Amazon SageMaker Python SDK to test the code.
- C. Download TensorFlow from tensorflow.org to emulate the TensorFlow kernel in the SageMaker environment.
- D. Download the SageMaker notebook to their local environment, then install Jupyter Notebooks on their laptop and continue the development in a local notebook.

Correct Answer: B

  **DonaldCMLIN** Highly Voted 1 year, 7 months ago
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://sagemaker.readthedocs.io/en/stable/overview.html#local-mode)
[HTTPS://GITHUB.COM/AWSLABS/AMAZON-SAGEMAKER-EXAMPLES/BLOB/MASTER/SAGEMAKER-PYTHON-SDK/TENSORFLOW_DISTIBUTED_MNIST/TENSORFLOW_LOCAL_MODE_MNIST.IPYNB](https://github.com/awslabs/amazon-sagemaker-examples/blob/master/sagemaker-python-sdk/tensorflow_distributed_mnist/tensorflow_local_mode_mnist.ipynb)
upvoted 27 times

  **CMMC** Most Recent 1 year ago
Agreed for B
upvoted 2 times

  **VB** 1 year, 3 months ago
<https://aws.amazon.com/blogs/machine-learning/use-the-amazon-sagemaker-local-mode-to-train-on-your-notebook-instance/>
B
upvoted 3 times

Question #61

Topic 1

A Machine Learning Specialist is working with a large cybersecurity company that manages security events in real time for companies around the world. The cybersecurity company wants to design a solution that will allow it to use machine learning to score malicious events as anomalies on the data as it is being ingested. The company also wants to be able to save the results in its data lake for later processing and analysis.

What is the MOST efficient way to accomplish these tasks?


- A. Ingest the data using Amazon Kinesis Data Firehose, and use Amazon Kinesis Data Analytics Random Cut Forest (RCF) for anomaly detection. Then use Kinesis Data Firehose to stream the results to Amazon S3.
- B. Ingest the data into Apache Spark Streaming using Amazon EMR, and use Spark MLlib with k-means to perform anomaly detection. Then

store the results in an Apache Hadoop Distributed File System (HDFS) using Amazon EMR with a replication factor of three as the data lake.


C. Ingest the data and store it in Amazon S3. Use AWS Batch along with the AWS Deep Learning AMIs to train a k-means model using TensorFlow on the data in Amazon S3.


D. Ingest the data and store it in Amazon S3. Have an AWS Glue job that is triggered on demand transform the new data. Then use the built-in Random Cut Forest (RCF) model within Amazon SageMaker to detect anomalies in the data.


Correct Answer: B


 **DonaldCMLIN** Highly Voted 1 year, 7 months ago
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-detect>
upvoted 34 times


 **JayK** Highly Voted 1 year, 5 months ago
Answer is A. As the word anomaly talks about Random Cut Forest in the exam and that can be done in a cost effective manner using Kines Data Analytics
upvoted 9 times


 **btsql** Most Recent 3 weeks, 6 days ago
B is possible But A is correct i think
upvoted 1 times

 **yeetusdeleetus** 7 months, 4 weeks ago
Most efficient, streaming = A
upvoted 2 times


 **KK3** 8 months ago
The answer is B since the requirement was real time and KFH is near real time i.e it has 60 sec latency
upvoted 1 times


 **NotAnMLProfessional** 3 months, 2 weeks ago
"as it is being ingested" is not necessarily realtime.. it could be near realtime :)
upvoted 1 times

 **fhuadeen** 9 months, 4 weeks ago
A doesn't seem correct actually. There is no service in AWS that is called "Amazon Kinesis Data Analytics Random Cut Forest (RCF)" for anomaly detection. Pay close attention to that phrase, there is nothing separating them, it is like a full name for one service.
upvoted 2 times

 **Predicare** 9 months ago
There is actually. Please check

https://www.google.com/url?sa=t&source=web&rct=j&url=https://docs.aws.amazon.com/kinesisanalytics/latest/sqlref/sqlrf-random-cut-forest.html&ved=2ahUKEwj_x-_pzovsAhXFURUIHYIvCssQFjAAegQIDRAC&usg=AOvVaw2a7_hUr7fxkO40fd_2eF5P
upvoted 1 times

 **syu31svc** 10 months ago
Answer is A 100%; stream the results to S3 would answer the part on "save the results in its data lake" from the question
upvoted 2 times

 **PRC** 1 year, 3 months ago
A is the answer..Anomaly RCF, Real time Kinesis Analytics..Data Lake - S3 via Firehose
upvoted 3 times

🗨️ 👤 **Phong** 1 year, 4 months ago

A is the best suitable answer
upvoted 4 times

🗨️ 👤 **devsean** 1 year, 4 months ago

It's A. B doesn't put the data into the datalake in the end, which was one of the requirements.
upvoted 4 times

🗨️ 👤 **vetal** 1 year, 6 months ago

It depends on what "most efficient" means. The simplest solution is A - and it still supports all the requirements.
upvoted 5 times

🗨️ 👤 **ComPah** 1 year, 6 months ago

Doesn't Key Word Large means it needs distributed architecture SPARK
upvoted 1 times

Question #62

Topic 1

A Data Scientist wants to gain real-time insights into a data stream of GZIP files.
Which solution would allow the use of SQL to query the stream with the LEAST latency?

- A. Amazon Kinesis Data Analytics with an AWS Lambda function to transform the data.
- B. AWS Glue with a custom ETL script to transform the data.
- C. An Amazon Kinesis Client Library to transform the data and save it to an Amazon ES cluster.
- D. Amazon Kinesis Data Firehose to transform the data and put it into an Amazon S3 bucket.

Correct Answer: A

Reference:

<https://aws.amazon.com/big-data/real-time-analytics-featured-partners/>

  **cybe001** Highly Voted 1 year, 5 months ago

A is correct. Kinesis Data Analytics can use lambda 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/>

upvoted 30 times

  **VB** Highly Voted 1 year, 3 months ago

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



..."

upvoted 6 times

  **HalloSpencer** Most Recent 11 months, 3 weeks ago



what about "LEAST latency"?

upvoted 3 times

  **Erso** 1 year, 2 months ago


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 :)

upvoted 2 times

  **JayK** 1 year, 5 months ago

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

upvoted 1 times

  **SophieSu** 4 months, 1 week ago

so you need lambda

upvoted 1 times

  **am7** 1 year, 5 months ago

But we need to run SQL on real time stream data.

upvoted 1 times

Question #63

Topic 1

A retail company intends to use machine learning to categorize new products. A labeled dataset of current products was provided to the Data Science team. The dataset includes 1,200 products. The labeled dataset has 15 features for each product such as title dimensions, weight, and price. Each product is labeled as belonging to one of six categories such as books, games, electronics, and movies.

Which model should be used for categorizing new products using the provided dataset for training?

- A. AnXGBoost model where the objective parameter is set to multi:softmax
- B. A deep convolutional neural network (CNN) with a softmax activation function for the last layer
- C. A regression forest where the number of trees is set equal to the number of product categories

D. A DeepAR forecasting model based on a recurrent neural network (RNN)

Correct Answer: B

  **rsimham** Highly Voted 1 year, 6 months ago

Ans: A XGBoost multi class classification. <https://medium.com/@gabrielziegler3/multiclass-multilabel-classification-with-xgboost-66195e4d9f2d>

CNN is used for image classificaiton problems

upvoted 22 times

  **JayK** Highly Voted 1 year, 5 months ago

Answer is A. This a classification problem thus XGBoost and the fact that there are six categories SOFTMAX is the right activation function

upvoted 8 times

  **syu31svc** Most Recent 10 months ago

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 classificatic
Recurrent neural networks (RNN) are a class of neural networks that is powerful for modeling sequence data such as time series or natural language

upvoted 3 times

  **Antriksh** 1 year ago



A is correct. XGBoost with softmax

upvoted 1 times

  **PRC** 1 year, 3 months ago

A should be the answer. Multi-classification problem and hence XGBoost...CNN is for image classification

upvoted 1 times

  **Phong** 1 year, 4 months ago

CNN is mostly used for images. A is the answer

upvoted 3 times

Question #64

Topic 1

A Data Scientist is working on an application that performs sentiment analysis. The validation accuracy is poor, and the Data Scientist thinks that the cause may be a rich vocabulary and a low average frequency of words in the dataset.

Which tool should be used to improve the validation accuracy?

- A. Amazon Comprehend syntax analysis and entity detection
- B. Amazon SageMaker BlazingText cbow mode
- C. Natural Language Toolkit (NLTK) stemming and stop word removal
- D. Scikit-learn term frequency-inverse document frequency (TF-IDF) vectorizer


Correct Answer: D

Reference:

<https://monkeylearn.com/sentiment-analysis/>

  **tap123**  1 year, 4 months ago

D is correct. Amazon Comprehend syntax analysis \neq Amazon Comprehend sentiment analysis. You need to read choices very carefully.
upvoted 15 times

  **mawsmann** 1 year, 1 month ago

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 carrying words infrequent. We're not looking to replace the sentiment analysis tool with Comprehend.
upvoted 1 times

  **DonaldCMLIN**  1 year, 7 months ago

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, A LANGUAGE DETECTION APIS SO YOU CAN EASILY INTEGRATE NATURAL LANGUAGE PROCESSING INTO YOUR APPLICATIONS.

[HTTPS://AWS.AMAZON.COM/COMPREHEND/FEATURES/?NC1=H_LS](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.

upvoted 11 times

  **ComPah** 1 year, 6 months ago

Agree Also Keyword is TOOL rest are frameworks
upvoted 2 times

  **srinu3054**  3 months, 1 week ago

Its B!! Blazing text has out-of-vocabulary (OOV) feature which can embed the non vocabulary words.
<https://docs.aws.amazon.com/sagemaker/latest/dg/blazingtext.html>
upvoted 4 times

🗨️ **mlpcr** 3 months, 1 week ago

I think Answer is C. Problem statement - "validation accuracy is poor, and ... cause may be a rich vocabulary and a low average frequency words in the dataset."

So to improve frequency of useful/meaningful words in given text we need to remove noisy words and for other purpose so stemming.

upvoted 2 times

🗨️ **cloud_trail** 5 months ago

I go with B. As another commenter said, go with the Amazon service over non-Amazon. Syntax and entity detection does not address the problem. Word embedding will, by providing context to words. Stemming and stop word removal won't help. TF-IDF might help a little but nearly as much as Word2Vec, which is option B.

upvoted 6 times

🗨️ **yeetusdeleetus** 7 months, 4 weeks ago

I'm a vote for B.

The other solutions do not directly address the stated problem. Word2vec, which is what BlazingText is, does.

upvoted 5 times

🗨️ **scuzzy2010** 4 months, 4 weeks ago

I vote B too -> "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"

upvoted 4 times

🗨️ **williamsuning** 9 months ago

validation document has rich vocabulary, it sounds to me creating difficulty and sparsity in classification. Especially, one-hot encoding/ tf-idf doesn't/less take into accounts the semantic of words. So word embedding helps to reduce the sparsity caused by rich vocabulary, and more importantly the semantic relationships between words are reflected in representation.

So I choose B, one more bonus for B is it also applies AWS Sagemaker features:

<https://docs.aws.amazon.com/sagemaker/latest/dg/blazingtext.html>

any answer better than this?

upvoted 7 times

🗨️ **williamsuning** 9 months ago

I think people/answers here focus more on keyword "frequency" - but less focus on the problem caused from the rich vocabulary and low avg freq

upvoted 1 times

🗨️ **yddmj** 10 months, 1 week ago

It mentioned such NLP techniques: syntax analysis, entity detection, cbow, stemming, stop word removal and TF-IDF. TF-IDF is the only one related to frequency of words.

It seems TF-IDF weighted method improve the performance.

<https://appliedmachinelearning.blog/2017/02/12/sentiment-analysis-using-tf-idf-weighting-python-scikit-learn/>

I vote D

upvoted 4 times

🗨️ **cloud_trail** 5 months ago

All that TF-IDF does is tell you what you already know: that have many low-frequency words. So what? That doesn't address the problem.

upvoted 1 times

🗨️ **trphng** 8 months ago

Agree. Question has pointed out: the cause may be a rich vocabulary and a low average frequency of words in the dataset

upvoted 1 times

  **GeeBeeEI** 11 months ago

I dont think the use of the term tool implies Comprehend, after all BlazingText may be considered a tool. The problem is a low average frequency of words in the dataset. Even if you use Comprehend, it will not increase the frequency of words. If however, you can make it possible to get a value that increases proportionally to the number of times a word appears in the document, then you are working on the frequency. A vectorizer is a tool, not a framework. Also I do agree with the concept that AWS would prefer to promote its service like Comprehend, but the question is more generic than that. The option on Comprehend is syntax anaysis (Syntax analysis —enabling custom to analyze text using tokenization and Parts of Speech (PoS).) The request is for frequency

upvoted 4 times

  **GeeBeeEI** 11 months ago

Based on the above and <https://medium.com/analytics-vidhya/fundamentals-of-bag-of-words-and-tf-idf-9846d301ff22>
Look out for

- tf-idf or TFIDF, short for term frequency-inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. The tf-idf value increases proportionally to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the word, which helps to adjust for the fact that some words appear more frequently in general. tf-idf is one of the most popular term-weighting schemes today; 83% of text-based recommender systems in digital libraries use tf-idf.

I go with D

upvoted 2 times

  **Urban_Life** 12 months ago

From my last couple of cert prep- I've seen that Amazon service always comes 1st rather go with right solution. Therefore, I will go with A.

upvoted 3 times

Question #65

Topic 1

Machine Learning Specialist is building a model to predict future employment rates based on a wide range of economic factors. While exploring the data, the

Specialist notices that the magnitude of the input features vary greatly. The Specialist does not want variables with a larger magnitude to dominate the model.


What should the Specialist do to prepare the data for model training?

- A. Apply quantile binning to group the data into categorical bins to keep any relationships in the data by replacing the magnitude with distribution.
- B. Apply the Cartesian product transformation to create new combinations of fields that are independent of the magnitude.
- C. Apply normalization to ensure each field will have a mean of 0 and a variance of 1 to remove any significant magnitude.
- D. Apply the orthogonal sparse bigram (OSB) transformation to apply a fixed-size sliding window to generate new features of a similar magnitude.

Correct Answer: C

Reference:

<https://docs.aws.amazon.com/machine-learning/latest/dg/data-transformations-reference.html>

  **rsimham**  1 year, 6 months ago

Ans: C; Normalization is correct

upvoted 23 times

🗨️ 👤 **gcpwhiz** 2 months, 1 week ago

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

upvoted 1 times

🗨️ 👤 **gcpwhiz** 2 months, 1 week ago

nevermind ignore this

upvoted 1 times

🗨️ 👤 **Phong** Highly Voted 🍌 1 year, 4 months ago

Guys, I passed the exam today. It is a tough one but there are many questions here. Good luck everyone! Thank examtopics

upvoted 10 times

🗨️ 👤 **haison8x** 11 months, 1 week ago

Hi Phong!

Please add my skype: haison8x

upvoted 1 times

🗨️ 👤 **grandgale** Most Recent 🕒 1 year, 4 months ago

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!

upvoted 8 times

🗨️ 👤 **joker34** 1 year, 3 months ago

So there are 40+ other questions on the exam that aren't included in Examtopics?

upvoted 2 times

🗨️ 👤 **nez15** 1 year, 6 months ago

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

upvoted 4 times

🗨️ 👤 **VB** 1 year, 3 months ago

<https://aws.amazon.com/blogs/machine-learning/forecasting-time-series-with-dynamic-deep-learning-on-aws/>

Answer: B

upvoted 1 times

🗨️ 👤 **nez15** 1 year, 6 months ago

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

upvoted 5 times

🗨️ 👤 **VB** 1 year, 3 months ago

<https://aws.amazon.com/blogs/machine-learning/analyze-us-census-data-for-population-segmentation-using-amazon-sagemaker/>

Answer: C and D

upvoted 3 times

🗨️ 👤 **cybe001** 1 year, 5 months ago

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.

upvoted 2 times

🗨️ 👤 **jasonsunbao** 1 year, 5 months ago

FM is mainly used in recommendation system to find hidden variables between two known variables to find correlation between two variables.

upvoted 1 times

🗨️ 👤 **nez15** 1 year, 6 months ago

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

upvoted 6 times

  **nez15** 1 year, 6 months ago



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?

upvoted 3 times

  **nez15** 1 year, 6 months ago

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

upvoted 10 times

Question #66

Topic 1

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

Using compressions will reduce the amount of data scanned by Amazon Athena, and also reduce your S3 bucket storage. It's a Win-Win for your AWS bill.



Supported formats: GZIP, LZ0, SNAPPY (Parquet) and ZLIB.

Reference:

<https://www.cloudforecast.io/blog/using-parquet-on-athena-to-save-money-on-aws/>

  **Erso** Highly Voted 1 year, 2 months ago

Answer A seems correct...
upvoted 5 times

  **Erso** 1 year, 2 months ago

sorry, the link <https://aws.amazon.com/blogs/big-data/prepare-data-for-model-training-and-invoke-machine-learning-models-with-amazon-athena/>
upvoted 1 times

  **sonalev419** Most Recent 3 months ago

A (Most queries will span 5 to 10 columns only)
upvoted 3 times

Question #67

Topic 1

A Machine Learning Specialist is developing a daily ETL workflow containing multiple ETL jobs. The workflow consists of the following processes:
"ç Start the workflow as soon as data is uploaded to Amazon S3. "ç 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.

"ç Store the results of joining datasets in Amazon S3.

"ç If one of the jobs fails, send a notification to the Administrator.

Which configuration will meet these requirements?

- 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

Reference:

<https://aws.amazon.com/step-functions/use-cases/>

  **HaiHN** Highly Voted 8 months, 1 week ago

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 meet 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>

upvoted 10 times

  **cloud_trail** 5 months ago

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.

upvoted 2 times

  **scuzzy2010** 7 months, 1 week ago

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.

upvoted 1 times

  **johnvik** Most Recent 2 months, 4 weeks ago

<https://d1.awsstatic.com/r2018/a/product-page-diagram-aws-step-functions-use-case-aws-glue.bc69d97a332c2dd29abb724dd747fd82ae110352.png>

upvoted 1 times

Question #68

Topic 1

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? (Choose 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

The PCA and K-means algorithms are useful in collection of data using census form.

  **hans1234** Highly Voted 11 months ago

<https://aws.amazon.com/blogs/machine-learning/analyze-us-census-data-for-population-segmentation-using-amazon-sagemaker/>

Answer: C and D

upvoted 8 times

  **HaiHN** Most Recent 8 months, 2 weeks ago

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 anomaly in data

upvoted 4 times

  **ac427** 1 year, 3 months ago

This is the same question as Topic 2 Q3

upvoted 1 times

Question #69

Topic 1

A large consumer goods manufacturer has the following products on sale: "ç 34 different toothpaste variants "ç 48 different toothbrush variants "ç 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

The Amazon SageMaker DeepAR forecasting algorithm is a supervised learning algorithm for forecasting scalar (one-dimensional) time series using recurrent neural networks (RNN). Classical forecasting methods, such as autoregressive integrated moving average (ARIMA) or exponential smoothing (ETS), fit a single model to each individual time series. They then use that model to extrapolate the time series into the future.

Reference:

<https://docs.aws.amazon.com/sagemaker/latest/dg/deepar.html>

  **HaiHN** Highly Voted 8 months, 2 weeks ago

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 the trained model to generate forecasts for new time series that are similar to the ones it has been trained on."

upvoted 5 times

  **hans1234** Most Recent 11 months ago

It is B

upvoted 2 times

  **ac427** 1 year, 3 months ago

This is the same question as Topic 2 Q4

upvoted 1 times

Question #70

Topic 1

A Machine Learning Specialist uploads a dataset to an Amazon S3 bucket protected with server-side encryption using AWS KMS.

How should the ML Specialist define the Amazon SageMaker notebook instance so it can read the same dataset from Amazon S3?

- A. Define security group(s) to allow all HTTP inbound/outbound traffic and assign those security group(s) to the Amazon SageMaker notebook

instance.

- B. Configure the Amazon SageMaker notebook instance to have access to the VPC. Grant permission in the KMS key policy to the notebook's KMS role.
- 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.
- D. Assign the same KMS key used to encrypt data in Amazon S3 to the Amazon SageMaker notebook instance.

Correct Answer: D

Reference:

<https://docs.aws.amazon.com/sagemaker/latest/dg/encryption-at-rest.html>

  **seanLu** Highly Voted 4 months, 1 week ago

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 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 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

upvoted 8 times



  **mona_mansour** Most Recent 1 month, 3 weeks ago

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 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 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.

upvoted 1 times

  **askaron** 4 months, 2 weeks ago

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.



upvoted 4 times

  **ahquiceno** 4 months, 2 weeks ago

C is correct go to: <https://docs.aws.amazon.com/sagemaker/latest/dg/sagemaker-roles.html>

<https://docs.aws.amazon.com/glue/latest/dg/create-an-iam-role-sagemaker-notebook.html>.

upvoted 4 times

  **joep21** 4 months, 3 weeks ago

D seems reasonable based on the explanation: <https://docs.aws.amazon.com/sagemaker/latest/dg/encryption-at-rest.html>

upvoted 1 times

  **takahirokoyama** 4 months, 3 weeks ago

I think so.

upvoted 1 times

Question #71

Topic 1

A Data Scientist needs to migrate an existing on-premises ETL process to the cloud. The current process runs at regular time intervals and uses PySpark to combine and format multiple large data sources into a single consolidated output for downstream processing.

The Data Scientist has been given the following requirements to the cloud solution:

- ☞ Combine multiple data sources.
- ☞ Reuse existing PySpark logic.
- ☞ Run the solution on the existing schedule.
- ☞ Minimize the number of servers that will need to be managed.

Which architecture should the Data Scientist use to build this solution?

- A. Write the raw data to Amazon S3. Schedule an AWS Lambda function to submit a Spark step to a persistent Amazon EMR cluster based on the existing schedule. Use the existing PySpark logic to run the ETL job on the EMR cluster. Output the results to a "processed" location in Amazon S3 that is accessible for downstream use.
- B. Write the raw data to Amazon S3. Create an AWS Glue ETL job to perform the ETL processing against the input data. Write the ETL job in PySpark to leverage the existing logic. Create a new AWS Glue trigger to trigger the ETL job based on the existing schedule. Configure the output target of the ETL job to write to a "processed" location in Amazon S3 that is accessible for downstream use.
- C. Write the raw data to Amazon S3. Schedule an AWS Lambda function to run on the existing schedule and process the input data from Amazon S3. Write the Lambda logic in Python and implement the existing PySpark logic to perform the ETL process. Have the Lambda function output the results to a "processed" location in Amazon S3 that is accessible for downstream use.
- D. Use Amazon Kinesis Data Analytics to stream the input data and perform real-time SQL queries against the stream to carry out the required transformations within the stream. Deliver the output results to a "processed" location in Amazon S3 that is accessible for downstream use.

Correct Answer: D

🗨️ 👤 **Paul_NoName** Highly Voted 4 months, 4 weeks ago

B it is .

upvoted 9 times

🗨️ 👤 **joep21** 4 months, 3 weeks ago

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-medicare.html>

upvoted 3 times

🗨️ 👤 **gcpwhiz** Most Recent 2 months, 2 weeks ago

Answer is B as they specifically ask about reusing existing PySpark, which can be done with Glue

upvoted 2 times

🗨️ 👤 **Aashi22** 4 months ago

https://docs.aws.amazon.com/glue/latest/dg/creating_running_workflows.html

upvoted 1 times

🗨️ 👤 **SophieSu** 4 months, 1 week ago

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

upvoted 1 times

🗨️ 👤 **astonm13** 4 months, 2 weeks ago

It is B. ! "Minimize number of servers to be managed". B is a Serverless solution which fulfils other requirements!

upvoted 2 times

🗨️ 👤 **cnethers** 4 months, 3 weeks ago

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 as there is no direct integration of Pyspark in Glue so there are some assumptions being made.

On Balance, I am leaning towards answer A

upvoted 1 times

🗨️ 👤 **Joe_Zhang** 5 months ago

The answer should be A.

upvoted 1 times

Question #72

Topic 1

A Data Scientist is building a model to predict customer churn using a dataset of 100 continuous numerical features. The Marketing team has not provided any insight about which features are relevant for churn prediction. The Marketing team wants to interpret the model and see the direct impact of relevant features on the model outcome. While training a logistic regression model, the Data Scientist observes that there is a wide gap

between the training and validation set accuracy.

Which methods can the Data Scientist use to improve the model performance and satisfy the Marketing team's needs? (Choose two.)

- A. Add L1 regularization to the classifier
- B. Add features to the dataset
- C. Perform recursive feature elimination
- D. Perform t-distributed stochastic neighbor embedding (t-SNE)
- E. Perform linear discriminant analysis

Correct Answer: *BE*

 **SophieSu** Highly Voted 4 months, 1 week ago

Key Words:

1. 100 continuous numerical features – too many features
2. No feature selection has been done
3. Easy interpretation - direct relationship between X and Y are preferred
4. gap between the training and validation set accuracy – overfitting

A: Correct. L1 regularization = feature selection/dimensionality reduction, solves overfitting, interpretation is easy, direct relationships between x and y

B: Wrong. More features, Overfitting will be worse.

C: Correct. Recursive feature elimination=feature selection/dimensionality reduction, solves overfitting, interpretation is easy, direct relationships between x and y

D: Perform t-distributed stochastic neighbor embedding (t-SNE)= Amazon's favorite dimensionality reduction technique, frequently shows up in the questions. However, same as PCA, less interpretable. You won't be able to see the direct impact of relevant features on the model outcome.

E: Wrong. If you have more than two classes then Linear Discriminant Analysis is the preferred linear classification technique.

My answer is A & C.

upvoted 11 times

 **DSJingguo** Most Recent 2 months ago

The answers should be AC.

This is a tricky question, I will show you why AC.

1. "The Marketing team wants to interpret the model and see the direct impact of relevant features on the model outcome" = The original features who impact the target. (this point is important, because the transformed features have not useful means)
2. "there is a wide gap between the training and validation set accuracy" = Model overfitting

To resolve the two problems:

A. (Correct) Regularization could prevent overfitting

C. (Correct) Recursive feature elimination could eliminate the least important features (dimensionality reduction)

E. (Wrong) Linear Discriminant Analysis could help reduce dimensionality but transform features also, you could not recognize transformed features.

Vote to help others see this post

upvoted 1 times

 **DSJingguo** 2 months ago

supplement: meantime L1 Regularization could do feature selection also

upvoted 1 times

🗳️ 👤 **ShirleyX** 2 months, 1 week ago

From the question, my understand is that we need to give answer of these 2 problems:

- #1. improve the model performance
- #2. satisfy the Marketing team's needs

for #1, the model is overfitting, so A is the solution to solve overfitting

for #2, the question said "The Marketing team wants to interpret the model and see the direct impact of relevant features on the model outcome", so C - feature selection based on feature importance.

so I think the right answer will be A&C

upvoted 1 times

🗳️ 👤 **gcpwhiz** 2 months, 1 week ago

You have overfitting problem and 100 features. Adding more features is a backwards approach.

upvoted 1 times

🗳️ 👤 **Vita_Rasta84444** 3 months ago

It is A & E. A will reduce the number of features and hold the most important ones, while LDA will create the one function composed of variables that have the highest impact on differentiating churns and no churns, which will help the interpretability. I am sure, A and E.

upvoted 2 times

🗳️ 👤 **achiko** 3 months, 1 week ago

A & C, A for accuracy C for showing feature importance

upvoted 2 times

🗳️ 👤 **xpada001** 3 months, 2 weeks ago

This is a churn prediction, which is known to be a supervised classification problem. Also, the question states "While training a logistic regression model...", and we all know logistic regression model is a supervised model. so I believe it is a supervised classification problem. In this case I voted A&C.

upvoted 3 times

🗳️ 👤 **astonm13** 4 months, 2 weeks ago

Churn prediction is a supervised classification problem. The discrepancy between train and validation accuracy indicates overfitting. To avoid overfitting use regularisation. I would say A & C

upvoted 3 times

🗳️ 👤 **cnethers** 4 months, 3 weeks ago

Summary

A due to the fact it's regression approach requires a supervised learning approach,

B is an option

C due to the fact RFE is a regression / classification approach it does not fit.

D is out because it's a visualization technique which does not help improve performance

E is an option as it looks at feature selection

Answer

B & E

upvoted 3 times

🗳️ 👤 **cnethers** 4 months, 3 weeks ago

Option (A) L1 regularization is a form of regression, that constrains/ regularizes or shrinks the coefficient estimates towards zero. In other words, this technique discourages learning a more complex or flexible model, so as to avoid the risk of overfitting.

Option (B) Add features to the dataset. This is a valid approach to improving accuracy


Option (C) Recursive Feature Elimination, or RFE for short, is a feature selection algorithm. It's an efficient approach for eliminating features from a training dataset for feature selection.

RFE can be used for classification and regression predictive modeling. This means RFE is a supervised learning approach which rules it out

Option (D) t-SNE is a machine learning algorithm for visualization, so this does not help improve accuracy.

Option (E) Linear discriminant analysis (LDA) is commonly used for dimensionality reduction much in the same way principal component analysis (PCA) is used.

upvoted 1 times

 **cnethers** 4 months, 3 weeks ago

Based on the fact you have a continuous dataset and there is no labeled data to help, regression can be excluded which leaves reduction of dimensionality which is an unsupervised learning approach. So focused efforts to an unsupervised algo that reduce in answers. The question calls out that there is a "gap between the training and validation set accuracy".

If accuracy needs to be improved then valid methods are:

- Add more data
- Treat missing & outlier values through data imputation
- Feature Engineering such as normalization, one hot encoding, etc.
- Feature selection (PCA / LDA)
- Multiple Algos
- Algo Tuning (hyperparameters)
- Ensemble methods (Bagging or Boosting)

upvoted 1 times

Question #73

Topic 1

An aircraft engine manufacturing company is measuring 200 performance metrics in a time-series. Engineers want to detect critical manufacturing defects in near-real time during testing. All of the data needs to be stored for offline analysis.

What approach would be the MOST effective to perform near-real time defect detection?

- A. Use AWS IoT Analytics for ingestion, storage, and further analysis. Use Jupyter notebooks from within AWS IoT Analytics to carry out analysis for anomalies.
- B. Use Amazon S3 for ingestion, storage, and further analysis. Use an Amazon EMR cluster to carry out Apache Spark ML k-means clustering to determine anomalies.
- C. Use Amazon S3 for ingestion, storage, and further analysis. Use the Amazon SageMaker Random Cut Forest (RCF) algorithm to determine anomalies.
- D. Use Amazon Kinesis Data Firehose for ingestion and Amazon Kinesis Data Analytics Random Cut Forest (RCF) to perform anomaly detection. Use Kinesis Data Firehose to store data in Amazon S3 for further analysis.

Correct Answer: B

🗨️ 👤 **Joe_Zhang** Highly Voted 5 months ago

D near-real time

upvoted 13 times

🗨️ 👤 **tmlid** Most Recent 2 months, 2 weeks ago

S3 is an object storage, not for data ingestion. B&C are wrong. I go with D

upvoted 3 times

🗨️ 👤 **Vita_Rasta84444** 3 months ago

Answer is D

upvoted 2 times

🗨️ 👤 **astonm13** 4 months, 2 weeks ago

D is correct!

upvoted 2 times

🗨️ 👤 **cnethers** 4 months, 3 weeks ago

Glad we are all in agreement D is the correct answer

upvoted 4 times

🗨️ 👤 **takahirokoyama** 4 months, 3 weeks ago

Ans. is D.

Question #74

Topic 1

A Machine Learning team runs its own training algorithm on Amazon SageMaker. The training algorithm requires external assets. The team needs to submit both its own algorithm code and algorithm-specific parameters to Amazon SageMaker.

What combination of services should the team use to build a custom algorithm in Amazon SageMaker? (Choose two.)

- A. AWS Secrets Manager
- B. AWS CodeStar
- C. Amazon ECR
- D. Amazon ECS
- E. Amazon S3

Correct Answer: CE

🗨️ 👤 **Paul_NoName** Highly Voted 4 months, 4 weeks ago

CE is the right answer. ECR uses ECS internally while using SGM.

upvoted 5 times

🗨️ 👤 **joep21** 4 months, 3 weeks ago

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."

upvoted 2 times

🗨️ 👤 **randomnamer** Most Recent 3 months ago

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.

upvoted 1 times

🗨️ 👤 **SophieSu** 4 months ago

CE IS THE CORRECT ANSWER 100%

upvoted 3 times

🗨️ 👤 **cnethers** 4 months, 3 weeks ago

<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

upvoted 3 times

🗨️ 👤 **ahquiceno** 4 months, 4 weeks ago

For me CD. needs storage and create a custom docker using ECR to store it.

upvoted 1 times

🗨️ 👤 **ahquiceno** 3 months, 3 weeks ago

Sorry, CE is correct.

upvoted 1 times

🗨️ 👤 **gcpwhiz** 2 months, 2 weeks ago

Sagemaker will spin up the instances needed with the right image. No need to use ECS. CE is right

upvoted 1 times

Question #75

Topic 1

A Machine Learning Specialist wants to determine the appropriate SageMakerVariantInvocationsPerInstance setting for an endpoint automatic scaling configuration. The Specialist has performed a load test on a single instance and determined that peak requests per second (RPS) without service degradation is about 20 RPS. As this is the first deployment, the Specialist intends to set the invocation safety factor to 0.5. Based on the stated parameters and given that the invocations per instance setting is measured on a per-minute basis, what should the Specialist set as the SageMakerVariantInvocationsPerInstance setting?

- A. 10
- B. 30
- C. 600
- D. 2,400

Correct Answer: C

 **Paul_NoName** Highly Voted 4 months, 4 weeks ago

C is correct .

$\text{SageMakerVariantInvocationsPerInstance} = (\text{MAX_RPS} * \text{SAFETY_FACTOR}) * 60$

AWS recommended Saf_fac = 0.5

upvoted 6 times

 **ahquiceno** Most Recent 4 months, 4 weeks ago

Answer C: $\text{SageMakerVariantInvocationsPerInstance} = (\text{MAX_RPS} * \text{SAFETY_FACTOR}) * 60$

<https://docs.aws.amazon.com/sagemaker/latest/dg/endpoint-scaling-loadtest.html>

upvoted 3 times

Question #76

Topic 1

A company uses a long short-term memory (LSTM) model to evaluate the risk factors of a particular energy sector. The model reviews multi-page text documents to analyze each sentence of the text and categorize it as either a potential risk or no risk. The model is not performing well, even though the Data Scientist has experimented with many different network structures and tuned the corresponding hyperparameters. Which approach will provide the MAXIMUM performance boost?

- A. Initialize the words by term frequency-inverse document frequency (TF-IDF) vectors pretrained on a large collection of news articles related to the energy sector.
- B. Use gated recurrent units (GRUs) instead of LSTM and run the training process until the validation loss stops decreasing.
- C. Reduce the learning rate and run the training process until the training loss stops decreasing.
- D. Initialize the words by word2vec embeddings pretrained on a large collection of news articles related to the energy sector.

Correct Answer: C

🗨️ 👤 **jiadong** Highly Voted 👍 4 months, 4 weeks ago

I think the right answer is D

upvoted 13 times

🗨️ 👤 **ChanduPatil** Most Recent 🕒 2 weeks, 2 days ago

why not B??

upvoted 1 times

🗨️ 👤 **jkreddy** 2 months ago

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

upvoted 1 times

🗨️ 👤 **YJ4219** 5 days, 14 hours ago

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.

upvoted 1 times

🗨️ 👤 **tmld** 2 months, 2 weeks ago

I go for C

upvoted 1 times

🗨️ 👤 **SophieSu** 4 months, 1 week ago

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

upvoted 3 times

🗨️ 👤 **astonm13** 4 months, 2 weeks ago

Agree. D seems more reasonable, as word2vec provides content of the sentences which is very important for evaluation of the risk.

upvoted 2 times

Question #77

Topic 1

A Machine Learning Specialist needs to move and transform data in preparation for training. Some of the data needs to be processed in near-real time, and other data can be moved hourly. There are existing Amazon EMR MapReduce jobs to clean and feature engineering to perform on the data.

Which of the following services can feed data to the MapReduce jobs? (Choose two.)

- A. AWS DMS
- B. Amazon Kinesis
- C. AWS Data Pipeline
- D. Amazon Athena
- E. Amazon ES

Correct Answer: AE

🗲️ 👤 **Joe_Zhang** Highly Voted 👍 5 months ago
should be BC
upvoted 11 times

🗲️ 👤 **joep21** 4 months, 3 weeks ago
Agreed, AWS Example: <https://docs.aws.amazon.com/datapipeline/latest/DeveloperGuide/what-is-datapipeline.html>
upvoted 4 times

🗲️ 👤 **Vita_Rasta84444** Most Recent ⌚ 3 months ago
Answer is B and C
upvoted 1 times

🗲️ 👤 **astonm13** 4 months, 2 weeks ago
Answer is for sure BC
upvoted 1 times

🗲️ 👤 **takahirokoyama** 4 months, 4 weeks ago
Ans is BC.
(<https://aws.amazon.com/jp/emr/?whats-new-cards.sort-by=item.additionalFields.postDateTime&whats-new-cards.sort-order=desc>)
upvoted 2 times

Question #78

Topic 1

A Machine Learning Specialist previously trained a logistic regression model using scikit-learn on a local machine, and the Specialist now wants to deploy it to production for inference only.

What steps should be taken to ensure Amazon SageMaker can host a model that was trained locally?

- A. Build the Docker image with the inference code. Tag the Docker image with the registry hostname and upload it to Amazon ECR.
- B. Serialize the trained model so the format is compressed for deployment. Tag the Docker image with the registry hostname and upload it to Amazon S3.
- C. Serialize the trained model so the format is compressed for deployment. Build the image and upload it to Docker Hub.
- D. Build the Docker image with the inference code. Configure Docker Hub and upload the image to Amazon ECR.

Correct Answer: D

 **arulrajayaraj** Highly Voted 4 months, 4 weeks ago

Ans : A Refer the below :

<https://sagemaker-workshop.com/custom/containers.html>

upvoted 5 times

 **Paul_NoName** Highly Voted 4 months, 4 weeks ago

A

<https://sagemaker-workshop.com/custom/containers.html>

upvoted 5 times

 **astonm13** Most Recent 4 months, 2 weeks ago

Answer is A.

upvoted 3 times

 **cnethers** 4 months, 3 weeks ago

Docker Hub is a repository so ANS D makes no sense. Option A is the way to go.

upvoted 2 times

Question #79

Topic 1

A trucking company is collecting live image data from its fleet of trucks across the globe. The data is growing rapidly and approximately 100 GB of new data is generated every day. The company wants to explore machine learning uses cases while ensuring the data is only accessible to specific IAM users.

Which storage option provides the most processing flexibility and will allow access control with IAM?

- A. Use a database, such as Amazon DynamoDB, to store the images, and set the IAM policies to restrict access to only the desired IAM users.
- B. Use an Amazon S3-backed data lake to store the raw images, and set up the permissions using bucket policies.
- C. Setup up Amazon EMR with Hadoop Distributed File System (HDFS) to store the files, and restrict access to the EMR instances using IAM policies.
- D. Configure Amazon EFS with IAM policies to make the data available to Amazon EC2 instances owned by the IAM users.

Correct Answer: C

  **Paul_NoName** Highly Voted 4 months, 4 weeks ago


B is the right answer

upvoted 8 times

  **Vita_Rasta84444** Most Recent 3 months ago

B is the right answer

upvoted 2 times

  **srinu3054** 3 months, 1 week ago



S3 is the easy, scalable and secure option to store the image data.

upvoted 1 times

  **astonm13** 4 months, 2 weeks ago

B is the right answer

upvoted 1 times

  **zzaibis** 4 months, 3 weeks ago

B is an appropriate choice

upvoted 2 times

Question #80

Topic 1

A credit card company wants to build a credit scoring model to help predict whether a new credit card applicant will default on a credit card payment. The company has collected data from a large number of sources with thousands of raw attributes. Early experiments to train a classification model revealed that many attributes are highly correlated, the large number of features slows down the training speed significantly, and that there are some overfitting issues.

The Data Scientist on this project would like to speed up the model training time without losing a lot of information from the original dataset.

Which feature engineering technique should the Data Scientist use to meet the objectives?

- A. Run self-correlation on all features and remove highly correlated features
- B. Normalize all numerical values to be between 0 and 1
- C. Use an autoencoder or principal component analysis (PCA) to replace original features with new features
- D. Cluster raw data using k-means and use sample data from each cluster to build a new dataset

Correct Answer: B

  **ahquiceno** Highly Voted 4 months, 4 weeks ago

Answer C. Need reduce the features preserving the information on it this is achieve using PCA.
upvoted 8 times

  **Vita_Rasta84444** Most Recent 3 months ago



Answer is C
upvoted 1 times

  **SophieSu** 4 months ago

C is the correct answer. PCA reduces the dimensionality, solves the overfitting, at the mean time does not cause information loss.
upvoted 1 times

  **astonm13** 4 months, 2 weeks ago



Answer is C
upvoted 2 times

  **joep21** 4 months, 3 weeks ago

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.
upvoted 3 times

  **SophieSu** 4 months ago

If you REMOVE highly correlated features(that means in pairs), the model lost a lot of information.
upvoted 1 times

  **wolf90** 4 months, 4 weeks ago

I think it should be C?
upvoted 2 times

  **Paul_NoName** 4 months, 4 weeks ago

C seems to be correct.
upvoted 3 times

  **takahirokoyama** 4 months, 4 weeks ago

Ans. is C.
upvoted 2 times

Question #81

Topic 1

A Data Scientist is training a multilayer perception (MLP) 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. 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.

Which techniques should be used to meet these requirements?

- A. Gather more data using Amazon Mechanical Turk and then retrain
- B. Train an anomaly detection model instead of an MLP
- C. Train an XGBoost model instead of an MLP
- D. Add class weights to the MLP's loss function and then retrain

Correct Answer: C

  **joep21**  4 months, 3 weeks ago



For me answer is D, adjust to higher weight for class of interest: <https://androidkt.com/set-class-weight-for-imbalance-dataset-in-keras/>. If data may/may not be available and a data labeling job will take time.

upvoted 11 times

  **ksarda11**  3 months, 3 weeks ago

In case of the quickest possible way, D seems fine. For XGBoost, it will take a bit of time to code again

upvoted 2 times

  **ahquiceno** 4 months, 4 weeks ago

For me Answer A. Why not other model instead of xgBoost, the model needs more labeled data to be trained and learn more positive examples

upvoted 2 times

  **SophieSu** 4 months ago

A is incorrect. Even if you hire Amazon Mechanical Turk, you won't have more data. This question is NOT asking about "labeling".

upvoted 1 times

Question #82

Topic 1

A Machine Learning Specialist works for a credit card processing company and needs to predict which transactions may be fraudulent in near-real time.



Specifically, the Specialist must train a model that returns the probability that a given transaction may be fraudulent.



How should the Specialist frame this business problem?



- A. Streaming classification
- B. Binary classification
- C. Multi-category classification
- D. Regression classification



Correct Answer: C



  **ahquiceno** Highly Voted 4 months, 4 weeks ago
 Answer B.
 upvoted 10 times



  **SophieSu** Highly Voted 4 months, 1 week ago
 B IS NOT CORRECT! Return the probability. Not the 1 or 0. D IS THE CORRECT ANSWER.
 upvoted 7 times



  **srinu3054** 3 months, 1 week ago
 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.
 upvoted 4 times



  **seanLu** 3 months, 4 weeks ago
 Logistic regression will give the probability, and logistic regression is a binary classification algorithm.
<https://machinelearningmastery.com/logistic-regression-for-machine-learning/>
 upvoted 2 times

  **AjithkumarSL** Most Recent 3 weeks, 6 days ago
 Looks like Answer is D, The model need to predict the probability and Regression can do that. If the second sentence of the question was there, then binary makes sense, It would have an Yes or No Thing.
 upvoted 1 times

  **MrCarter** 4 weeks, 1 day ago
 Should D say Logistic Regression? That would make it the best choice for sure
 upvoted 1 times

  **StelSen** 1 month, 3 weeks ago
 Answer is B: Understand the inputs and outputs for Classification and Regression from here:
<https://machinelearningmastery.com/classification-versus-regression-in-machine-learning/#:~:text=Fundamentally%2C%20classification%20is%20about%20predicting,is%20about%20predicting%20a%20quantity.&text=at%20classification%20is%20the%20problem,quantity%20output%20for%20an%20example.>
 upvoted 1 times

  **cnethers** 4 months, 3 weeks ago
 this is a classic Binary Classification so B is the ANS
 upvoted 5 times

  **Paul_NoName** 4 months, 4 weeks ago
 D seems to be right.
 upvoted 2 times

Question #83


Topic 1

A real estate company wants to create a machine learning model for predicting housing prices based on a historical dataset. The dataset contains 32 features.

Which model will meet the business requirement?

- A. Logistic regression
- B. Linear regression
- C. K-means
- D. Principal component analysis (PCA)

Correct Answer: B

 **SophieSu** Highly Voted 4 months ago
B is the correct answer.
upvoted 5 times

 **ahquiceno** Most Recent 4 months, 4 weeks ago
Answer B.
upvoted 3 times

Question #84

Topic 1

A Machine Learning Specialist is applying a linear least squares regression model to a dataset with 1,000 records and 50 features. Prior to training, the ML

Specialist notices that two features are perfectly linearly dependent.

Why could this be an issue for the linear least squares regression model?



- A. It could cause the backpropagation algorithm to fail during training
- B. It could create a singular matrix during optimization, which fails to define a unique solution
- C. It could modify the loss function during optimization, causing it to fail during training
- D. It could introduce non-linear dependencies within the data, which could invalidate the linear assumptions of the model

Correct Answer: C

  **Paul_NoName** Highly Voted 4 months, 4 weeks ago

B is correct answer .

upvoted 6 times

  **pravv** 4 months, 2 weeks ago

why B is the correct answer and not C?

upvoted 1 times

  **SophieSu** 4 months ago

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.

upvoted 1 times

  **jerto97** Most Recent 2 weeks, 1 day ago

B. See the multicollinearity problem in wikipedia <https://en.wikipedia.org/wiki/Multicollinearity> (second paragraph)

upvoted 1 times

  **takahirokoyama** 4 months, 4 weeks ago

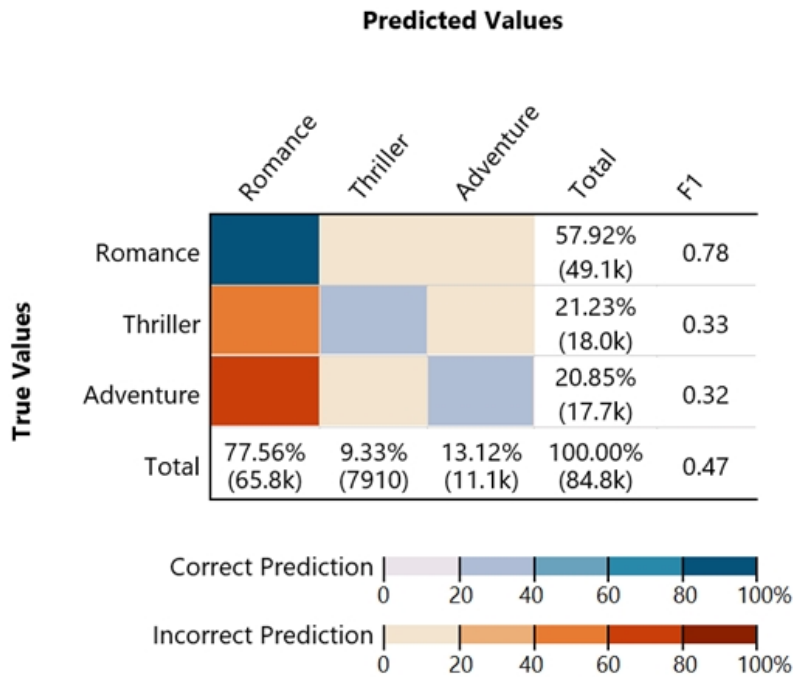
This issue is overfitting.

upvoted 1 times

Question #85

Topic 1

Given the following confusion matrix for a movie classification model, what is the true class frequency for Romance and the predicted class frequency for Adventure?



- A. The true class frequency for Romance is 77.56% and the predicted class frequency for Adventure is 20.85%
- B. The true class frequency for Romance is 57.92% and the predicted class frequency for Adventure is 13.12%
- C. The true class frequency for Romance is 0.78 and the predicted class frequency for Adventure is (0.47-0.32)
- D. The true class frequency for Romance is $77.56\% \times 0.78$ and the predicted class frequency for Adventure is $20.85\% \times 0.32$

Correct Answer: B

SophieSu Highly Voted 4 months ago

B is the correct answer. Straightforward!
upvoted 8 times

Juka3lj Most Recent 2 months, 1 week ago

B is correct
upvoted 1 times

NotAnMLProfessional 3 months, 2 weeks ago

A seems to be correct
upvoted 1 times

cnethers 4 months, 3 weeks ago

<https://docs.aws.amazon.com/machine-learning/latest/dg/multiclass-model-insights.html>
upvoted 4 times

Question #86

Topic 1

A Machine Learning Specialist wants to bring a custom algorithm to Amazon SageMaker. The Specialist implements the algorithm in a Docker container supported by Amazon SageMaker.

How should the Specialist package the Docker container so that Amazon SageMaker can launch the training correctly?

- A. Modify the `bash_profile` file in the container and add a bash command to start the training program
- B. Use `CMD` config in the Dockerfile to add the training program as a CMD of the image
- C. Configure the training program as an `ENTRYPOINT` named `train`
- D. Copy the training program to directory `/opt/ml/train`

Correct Answer: B

 **Paul_NoName** Highly Voted 4 months, 4 weeks ago

C seems correct as per documentations.

upvoted 10 times

 **joep21** Highly Voted 4 months, 3 weeks ago

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"

upvoted 6 times

 **Juka3lj** Most Recent 2 months, 1 week ago

C is correct

upvoted 1 times

 **Aashi22** 4 months ago

option C https://github.com/awsdocs/amazon-sagemaker-developer-guide/blob/master/doc_source/your-algorithms-training-algo-dockerfile.md

upvoted 1 times

Question #87

Topic 1

A Data Scientist needs to analyze employment data. The dataset contains approximately 10 million observations on people across 10 different features. During the preliminary analysis, the Data Scientist notices that income and age distributions are not normal. While income levels shows a right skew as expected, with fewer individuals having a higher income, the age distribution also show a right skew, with fewer older individuals participating in the workforce.

Which feature transformations can the Data Scientist apply to fix the incorrectly skewed data? (Choose two.)

- A. Cross-validation
- B. Numerical value binning
- C. High-degree polynomial transformation

D. Logarithmic transformation

E. One hot encoding

Correct Answer: AB

🗨️ **seanLu** Highly Voted 4 months, 1 week ago

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>

upvoted 10 times

🗨️ **OmarSaadEldien** 1 month ago

Agree with B &D

B binning for age

D for make income in normal dist

upvoted 1 times

🗨️ **omar_bahrain** 3 months, 3 weeks ago

agree B&D. both are strategies to eliminate the effect of skewing

upvoted 3 times

🗨️ **Joe_Zhang** Highly Voted 5 months ago

SHOULD BE C,D

upvoted 5 times

🗨️ **Juka3lj** Most Recent 2 months, 1 week ago

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}

upvoted 4 times

🗨️ **Vita_Rasta84444** 3 months ago

should be c and d

upvoted 4 times

🗨️ **achiko** 3 months, 1 week ago

polynomial transformations can also be used for skewed data. <https://machinelearningmastery.com/polynomial-features-transforms-for-machine-learning/>

upvoted 3 times

🗨️ **jiadong** 4 months, 3 weeks ago

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>

upvoted 4 times

Question #88

Topic 1

A web-based company wants to improve its conversion rate on its landing page. Using a large historical dataset of customer visits, the company has repeatedly trained a multi-class deep learning network algorithm on Amazon SageMaker. However, there is an overfitting problem: training data shows 90% accuracy in predictions, while test data shows 70% accuracy only.

The company needs to boost the generalization of its model before deploying it into production to maximize conversions of visits to purchases. Which action is recommended to provide the HIGHEST accuracy model for the company's test and validation data?

- A. Increase the randomization of training data in the mini-batches used in training
- B. Allocate a higher proportion of the overall data to the training dataset
- C. Apply L1 or L2 regularization and dropouts to the training
- D. Reduce the number of layers and units (or neurons) from the deep learning network

Correct Answer: D

🗨️ **joep21** Highly Voted 4 months, 3 weeks ago

I would answer C, add regularization to mitigate overfitting: <https://docs.aws.amazon.com/machine-learning/latest/dg/model-fit-underfitting-vs-overfitting.html>

upvoted 5 times

🗨️ **Vita_Rasta84444** Most Recent 3 months ago

C is the answer

upvoted 1 times

🗨️ **SophieSu** 4 months ago

A is the BEST answer. There are several manners in which we can reduce overfitting in deep learning models. The best option is to get more training data. Unfortunately, in real-world situations, you often do not have this possibility due to time, budget or technical constraints. Another way to reduce overfitting is to lower the capacity of the model to memorize the training data. As such, the model will need to focus the relevant patterns in the training data, which results in better generalization.

Pay attention that the questions is asking "Deep Learning" model, usually the regularization technique is to use early stop or dropout.

D (reduce layers) is not appropriate. Use Dropout of neurons instead.

upvoted 1 times

🗨️ **q333** 4 months ago

C is the answer. If this is a ComputerVision problem augmentation can help and we may consider A an option. However in analyzing customer historic data, there is no easy way to increase randomization in training. If you go deep into modelling and coding. When you build model with tensorflow/pytorch, most of the time the trainloader is already sampling in data in random manner (with shuffle enable). What we usually do to reduce overfitting is by adding dropout.

upvoted 2 times

🗨️ **astonm13** 4 months, 2 weeks ago

C should be correct

upvoted 1 times

🗨️ **asdfkljh** 4 months, 4 weeks ago

Is it C?

upvoted 1 times

🗨️ **takahirokoyama** 4 months, 4 weeks ago

Ans. is C.

upvoted 1 times

Question #89

Topic 1

A Machine Learning Specialist is given a structured dataset on the shopping habits of a company's customer base. The dataset contains thousands of columns of data and hundreds of numerical columns for each customer. The Specialist wants to identify whether there are natural groupings for these columns across all customers and visualize the results as quickly as possible.

What approach should the Specialist take to accomplish these tasks?

- A. Embed the numerical features using the t-distributed stochastic neighbor embedding (t-SNE) algorithm and create a scatter plot.
- B. Run k-means using the Euclidean distance measure for different values of k and create an elbow plot.
- C. Embed the numerical features using the t-distributed stochastic neighbor embedding (t-SNE) algorithm and create a line graph.
- D. Run k-means using the Euclidean distance measure for different values of k and create box plots for each numerical column within each cluster.

Correct Answer: B

  **ac71** Highly Voted 4 months, 4 weeks ago

A is correct. tSNE can do segmentation or grouping as well. Refer: <https://towardsdatascience.com/an-introduction-to-t-sne-with-python-example-5a3a293108d1>

upvoted 9 times

  **SophieSu** Highly Voted 4 months ago

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 use to determine the optimal # of groups=K).



upvoted 7 times

  **orangechickencombo** Most Recent 1 month, 3 weeks ago

B.



tSNE need to have grouped labels for visualization. Here looking for whether there is group or not

upvoted 1 times

  **Juka3lj** 2 months, 1 week ago

A is correct answer

upvoted 1 times

  **cnethers** 4 months, 3 weeks ago

B looks like a good answer to me

upvoted 2 times

  **cnethers** 4 months, 3 weeks ago

Based on the fact following fact :

<https://machinelearningmastery.com/clustering-algorithms-with-python/>

Examples of Clustering Algorithms

- Library Installation
- Clustering Dataset
- Affinity Propagation
- Agglomerative Clustering
- BIRCH
- DBSCAN
- K-Means
- Mini-Batch K-Means
- Mean Shift
- OPTICS
- Spectral Clustering
- Gaussian Mixture Model

+

<https://predictivehacks.com/k-means-elbow-method-code-for-python/>

The Elbow method is a very popular technique and the idea is to run k-means clustering for a range of clusters k (let's say from 1 to 10) and each value, we are calculating the sum of squared distances from each point to its assigned center(distortions).

upvoted 3 times

Question #90

Topic 1

A Machine Learning Specialist is planning to create a long-running Amazon EMR cluster. The EMR cluster will have 1 master node, 10 core nodes, and 20 task nodes. To save on costs, the Specialist will use Spot Instances in the EMR cluster.

Which nodes should the Specialist launch on Spot Instances?

- A. Master node
- B. Any of the core nodes
- C. Any of the task nodes
- D. Both core and task nodes



Correct Answer: A

  **joep21** Highly Voted 4 months, 4 weeks ago

Answer is C. <https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-plan-instances-guidelines.html>
upvoted 8 times

  **benson2021** Most Recent 2 months, 2 weeks ago

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.
upvoted 2 times

  **xpada001** 3 months, 2 weeks ago

Answer should be C.
upvoted 2 times

  **SophieSu** 4 months ago

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."
upvoted 2 times

  **ac71** 4 months, 4 weeks ago

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
upvoted 3 times

Question #91

Topic 1

A manufacturer of car engines collects data from cars as they are being driven. The data collected includes timestamp, engine temperature, rotations per minute (RPM), and other sensor readings. The company wants to predict when an engine is going to have a problem, so it can notify drivers in advance to get engine maintenance. The engine data is loaded into a data lake for training.

Which is the MOST suitable predictive model that can be deployed into production?



- A. Add labels over time to indicate which engine faults occur at what time in the future to turn this into a supervised learning problem. Use a recurrent neural network (RNN) to train the model to recognize when an engine might need maintenance for a certain fault.
- B. This data requires an unsupervised learning algorithm. Use Amazon SageMaker k-means to cluster the data.
- C. Add labels over time to indicate which engine faults occur at what time in the future to turn this into a supervised learning problem. Use a convolutional neural network (CNN) to train the model to recognize when an engine might need maintenance for a certain fault.
- D. This data is already formulated as a time series. Use Amazon SageMaker seq2seq to model the time series.

Correct Answer: B

  **ac71** Highly Voted 4 months, 4 weeks ago

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 d here. Hence, A seems appropriate.

upvoted 21 times

  **astonm13** 4 months, 2 weeks ago

Agree, the answer is A

upvoted 4 times

  **Vita_Rasta84444** Most Recent 3 months ago

It is A

upvoted 1 times

Question #92

Topic 1

A company wants to predict the sale prices of houses based on available historical sales data. The target variable in the company's dataset is the sale price. The features include parameters such as the lot size, living area measurements, non-living area measurements, number of bedrooms, number of bathrooms, year built, and postal code. The company wants to use multi-variable linear regression to predict house sale prices.

Which step should a machine learning specialist take to remove features that are irrelevant for the analysis and reduce the model's complexity?

- A. Plot a histogram of the features and compute their standard deviation. Remove features with high variance.
- B. Plot a histogram of the features and compute their standard deviation. Remove features with low variance.
- 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.


Correct Answer: D

  **ahquiceno** Highly Voted 4 months, 4 weeks ago

Answer B. Is not the best solution 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.


upvoted 8 times

  **YJ4219** Most Recent 5 days, 12 hours ago

I think the answer is D, because the correlation between each feature and the target, if this feature has low variance (as B suggests) this correlation calculation will be low and thus will be removed.

In short I think D is a more general and more accurate answer.

upvoted 1 times

  **Vita_Rasta84444** 3 months ago



It is D. We want to exclude the values with low predictive power. It is although more probable that variables with low variance have the lower predictive power, but it is not necessary so.

upvoted 1 times

  **SophieSu** 4 months ago

B is the correct answer. Variation is information. No variation = Constant.

upvoted 2 times

  **cnethers** 4 months, 3 weeks ago



If features are linearly dependent they should be removed as there is no benefit in having them, for that reason I choose option D

upvoted 4 times

  **jerto97** 1 week, 6 days ago

Consider feature A is uncorrelated, and B is uncorrelated with the output. BUT It might be that the correlation is with A + B. You cannot remove it like that, that's the point of multivariable regression

upvoted 1 times

  **achiko** 3 months, 1 week ago

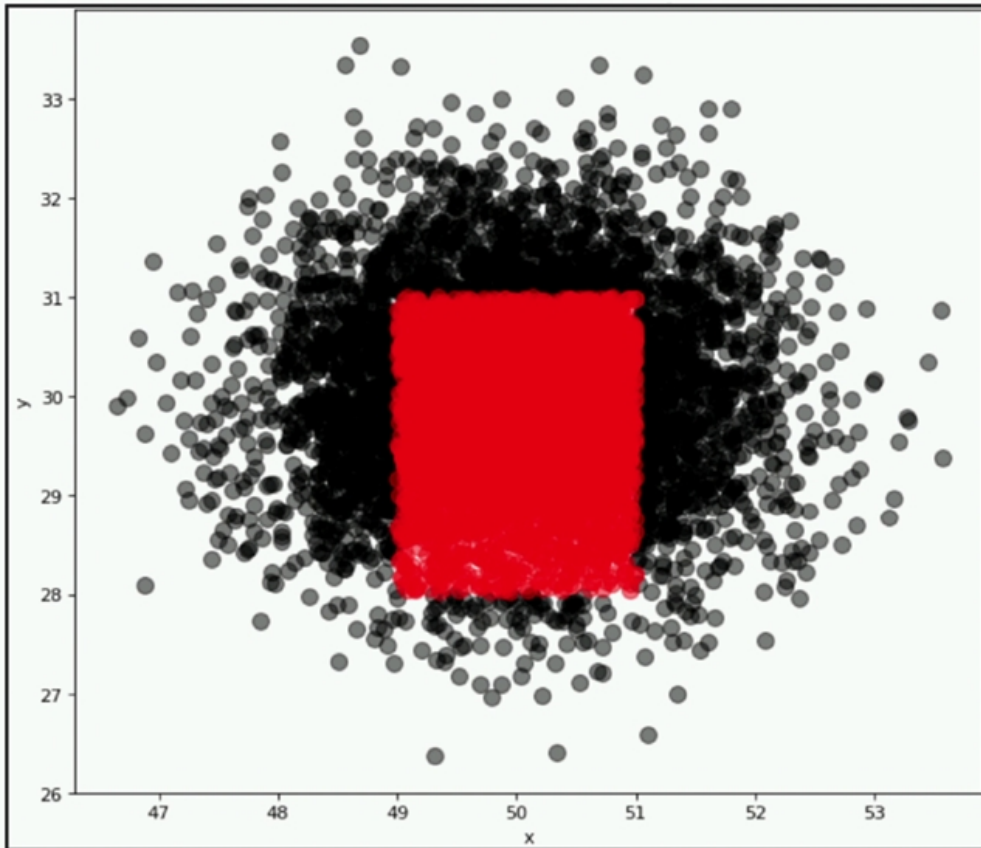
D says low correlation score against target variable. B is the answer

upvoted 3 times

Question #93

Topic 1

A company wants to classify user behavior as either fraudulent or normal. Based on internal research, a machine learning specialist will build a binary classifier based on two features: age of account, denoted by x , and transaction month, denoted by y . The class distributions are illustrated in the provided figure. The positive class is portrayed in red, while the negative class is portrayed in black.



Which model would have the HIGHEST accuracy?

- A. Linear support vector machine (SVM)
- B. Decision tree
- C. Support vector machine (SVM) with a radial basis function kernel
- D. Single perceptron with a Tanh activation function

Correct Answer: C

🗨️ 👤 **joep21** Highly Voted 4 months, 4 weeks ago

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

upvoted 7 times

🗨️ 👤 **MrCarter** 4 weeks, 1 day ago

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

upvoted 1 times

🗨️ 👤 **SophieSu** Highly Voted 4 months ago

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 able to approximate the rectangular boundary shape.

upvoted 5 times

🗨️ 👤 **MrCarter** Most Recent 4 weeks, 1 day ago

Answer is definitely C.

upvoted 1 times

🗨️ 👤 **gcpwhiz** 2 months, 1 week ago

answer is decision tree. Decision tree produces square boundary, SVM with radial function produces a circular boundary.

upvoted 2 times

🗨️ 👤 **Vita_Rasta84444** 3 months ago

Question #94



Topic 1





















A health care company is planning to use neural networks to classify their X-ray images into normal and abnormal classes. The labeled data is divided into a training set of 1,000 images and a test set of 200 images. The initial training of a neural network model with 50 hidden layers yielded 99% accuracy on the training set, but only 55% accuracy on the test set.

What changes should the Specialist consider to solve this issue? (Choose three.)

- A. Choose a higher number of layers
- B. Choose a lower number of layers
- C. Choose a smaller learning rate
- D. Enable dropout
- E. Include all the images from the test set in the training set
- F. Enable early stopping

Correct Answer: ADE

-   **cnethers** Highly Voted 4 months, 3 weeks ago
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
upvoted 9 times
-   **johnvik** Most Recent 2 months, 4 weeks ago
choose smaller learning rate c, d, f,
upvoted 1 times
-   **johnvik** 2 months, 4 weeks ago
ignore answer is correct BDF
upvoted 1 times
-   **Vita_Rasta84444** 3 months ago
BDF!!!
upvoted 2 times
-   **SophieSu** 4 months, 1 week ago
BDF !!!
upvoted 3 times
-   **astonm13** 4 months, 2 weeks ago
It is supposed to be BDF
upvoted 1 times
-   **joep21** 4 months, 3 weeks ago
I would choose B, D and E.
upvoted 1 times
-   **joep21** 4 months, 2 weeks ago
I meant to say B, D and F. Ignore above.
upvoted 3 times
-   **asdfkjh** 4 months, 4 weeks ago
is it BDF?
upvoted 1 times
-   **takahirokoyama** 4 months, 4 weeks ago
Ans. is B,,C,D.
Because this problem is overfitting.
upvoted 1 times
-   **ahquiceno** 4 months, 4 weeks ago
Answers BDF
upvoted 1 times

Question #95

Topic 1

This graph shows the training and validation loss against the epochs for a neural network.

The network being trained is as follows:

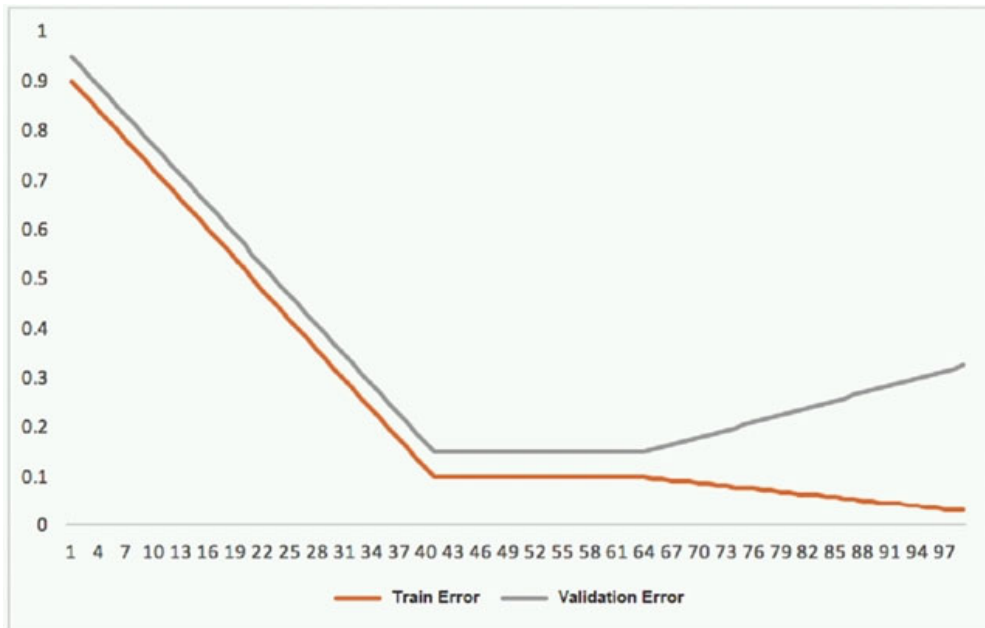
☞ Two dense layers, one output neuron

100 neurons in each layer

100 epochs

Random initialization of weights

▪



Which technique can be used to improve model performance in terms of accuracy in the validation set?

- A. Early stopping
- B. Random initialization of weights with appropriate seed
- C. Increasing the number of epochs
- D. Adding another layer with the 100 neurons

Correct Answer: C

 **ahquiceno** Highly Voted 4 months, 4 weeks ago

Answer A.

upvoted 5 times

 **chrisabc** Most Recent 2 weeks, 1 day ago


Early Stopping can improve the model?

upvoted 1 times

 **eganilovic** 2 months ago


The answer is Early Stopping. Stopp the training before accuracy start do decrease.

upvoted 2 times

 **StelSen** 1 month, 3 weeks ago

Appreciates your explanation. Cheers

upvoted 1 times

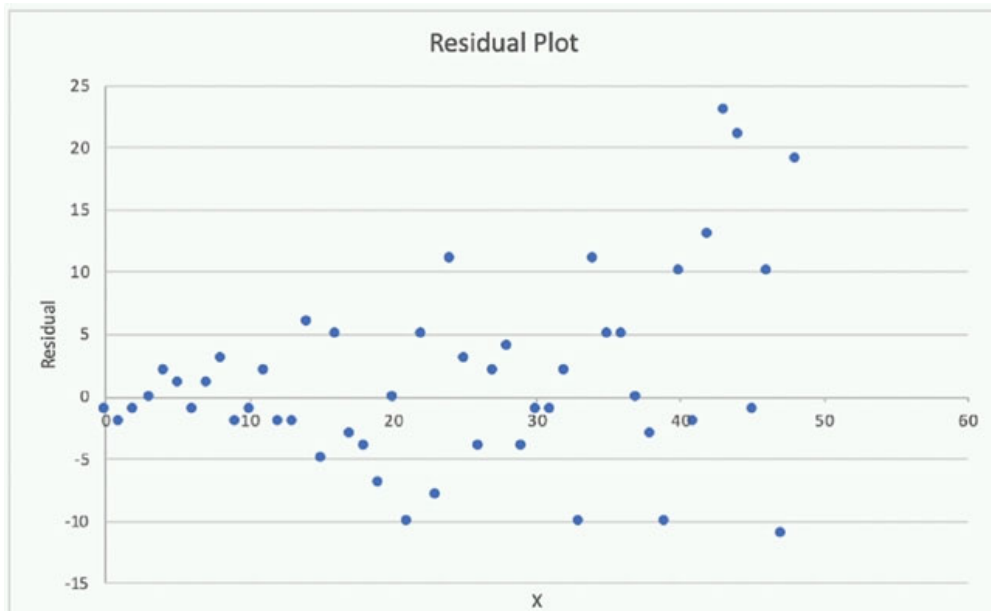
 **Vita_Rasta84444** 3 months ago

A is the answer

Question #96

Topic 1

A Machine Learning Specialist is attempting to build a linear regression model.



Given the displayed residual plot only, what is the MOST likely problem with the model?

- A. Linear regression is inappropriate. The residuals do not have constant variance.
- B. Linear regression is inappropriate. The underlying data has outliers.
- C. Linear regression is appropriate. The residuals have a zero mean.
- D. Linear regression is appropriate. The residuals have constant variance.

Correct Answer: D

🗲️ 👤 **joep21** Highly Voted 4 months, 3 weeks ago

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>
upvoted 5 times

🗲️ 👤 **takahirokoyama** Highly Voted 4 months, 4 weeks ago

Ans. is A.
High-degree polynomial transformation.
upvoted 5 times

🗲️ 👤 **TrekkingMachine** 4 months, 3 weeks ago

I think so too.
upvoted 1 times

🗲️ 👤 **yummytaco** Most Recent 1 month, 1 week ago

Do not have content variance
<https://stats.stackexchange.com/questions/52089/what-does-having-constant-variance-in-a-linear-regression-model-mean>
upvoted 1 times

🗲️ 👤 **Vita_Rasta84444** 3 months ago

Answer is A. As x raises, the residuals become higher and higher...
upvoted 1 times

🗲️ 👤 **cnethers** 4 months, 3 weeks ago

Some Good Reading https://www.andrew.cmu.edu/user/achoulde/94842/homework/regression_diagnostics.html

Ans is A
upvoted 4 times

Question #97

Topic 1

A large company has developed a BI application that generates reports and dashboards using data collected from various operational metrics. The company wants to provide executives with an enhanced experience so they can use natural language to get data from the reports. The company wants the executives to be able ask questions using written and spoken interfaces. Which combination of services can be used to build this conversational interface? (Choose three.)

A. Alexa for Business

- B. Amazon Connect
- C. Amazon Lex
- D. Amazon Polly
- E. Amazon Comprehend
- F. Amazon Transcribe

Correct Answer: BEF

  **eganilovic** Highly Voted 2 months ago

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

upvoted 7 times

  **eganilovic** 2 months ago

*C - Lex

So C,D,F

upvoted 5 times

  **weelz** 1 week, 1 day ago

I second that, the keyword here is "conversational interface". so, no conversation without Amazon Lex

upvoted 1 times

  **astonm13** Highly Voted 4 months, 2 weeks ago

C - voice and text interface

E - understanding

F - Speech to text



upvoted 6 times

  **jerto97** Most Recent 1 week, 6 days ago

People seem to think the interface needs to speak back, but I do not think this is the case. Person talk --> graph returned


I'd go for lex, comprehend, and speech to text

upvoted 1 times

  **gcpwhiz** 2 months, 1 week ago

Why do you need C and F? Lex already does speech to text, you don't need transcribe.

upvoted 1 times

  **Vita_Rasta84444** 3 months ago

CEF !!!

upvoted 1 times

  **SophieSu** 4 months ago

C, E and F

upvoted 1 times

  **scuzzy2010** 4 months, 1 week ago

C (building conversational interfaces into any application using voice and text.)

D (text to speech - app can answer verbally)

E (natural language processing (NLP) service that uses machine learning to find insights and relationships in text - can be used with Lex)

upvoted 3 times

🗨️ 👤 **jdstone** 4 months, 2 weeks ago

C (written questions)
E (sentiment and understanding)
F (speech to text)
upvoted 2 times

🗨️ 👤 **cnethers** 4 months, 3 weeks ago

D. Amazon Polly Text to Speech
E. Amazon Comprehend Derive sentiment and understanding
F. Amazon Transcribe Speech to text
upvoted 1 times

🗨️ 👤 **joep21** 4 months, 3 weeks ago

I would answer A (spoken questions), C(written questions), and E (NLP to parse reports).
upvoted 2 times

Question #98

Topic 1

A machine learning specialist works for a fruit processing company and needs to build a system that categorizes apples into three types. The specialist has collected a dataset that contains 150 images for each type of apple and applied transfer learning on a neural network that was pretrained on ImageNet with this dataset.

The company requires at least 85% accuracy to make use of the model.

After an exhaustive grid search, the optimal hyperparameters produced the following:

- 🗨️ 68% accuracy on the training set
- 🗨️ 67% accuracy on the validation set

What can the machine learning specialist do to improve the system's accuracy?

- A. Upload the model to an Amazon SageMaker notebook instance and use the Amazon SageMaker HPO feature to optimize the model's hyperparameters.
- B. Add more data to the training set and retrain the model using transfer learning to reduce the bias.
- C. Use a neural network model with more layers that are pretrained on ImageNet and apply transfer learning to increase the variance.
- D. Train a new model using the current neural network architecture.

Correct Answer: B

🗨️ 👤 **Seyi_0** 4 months ago

Correct, the answer is B, this is an underfitting problem!
upvoted 4 times

🗨️ 👤 **SophieSu** 4 months, 1 week ago

B. underfitting, reduce the bias
upvoted 4 times

Question #99

Topic 1

A company uses camera images of the tops of items displayed on store shelves to determine which items were removed and which ones still remain. After several hours of data labeling, the company has a total of 1,000 hand-labeled images covering 10 distinct items. The training results were poor.

Which machine learning approach fulfills the company's long-term needs?

- A. Convert the images to grayscale and retrain the model
- B. Reduce the number of distinct items from 10 to 2, build the model, and iterate
- C. Attach different colored labels to each item, take the images again, and build the model
- D. Augment training data for each item using image variants like inversions and translations, build the model, and iterate.

Correct Answer: A

  **joep21** Highly Voted 4 months, 4 weeks ago



D. Data augmentation would help with small dataset size. <https://medium.com/snowdog-labs/data-augmentation-techniques-and-pitfalls-c-small-datasets-e5a657fc404f>

upvoted 8 times

  **Vita_Rasta84444** Most Recent 3 months ago

I would choose D

upvoted 1 times

  **astonm13** 4 months, 2 weeks ago

I would choose D

upvoted 1 times

Question #100

Topic 1

A Data Scientist is developing a binary classifier to predict whether a patient has a particular disease on a series of test results. The Data Scientist has data on

400 patients randomly selected from the population. The disease is seen in 3% of the population.

Which cross-validation strategy should the Data Scientist adopt?

- A. A k-fold cross-validation strategy with $k=5$
- B. A stratified k-fold cross-validation strategy with $k=5$
- C. A k-fold cross-validation strategy with $k=5$ and 3 repeats
- D. An 80/20 stratified split between training and validation

Correct Answer: B

 **scuzzy2010** Highly Voted 4 months, 1 week ago


B - stratified k-fold cross-validation will enforce the class distribution in each split of the data to match the distribution in the complete train dataset.

upvoted 5 times

 **AWS__Newbie** Most Recent 2 weeks, 3 days ago

Why $K=5$?

upvoted 1 times

 **Vita_Rasta84444** 3 months ago

Yes, B...

upvoted 1 times

 **SophieSu** 4 months ago

B is the correct answer. Use Stratified k-Fold Cross-Validation for Imbalanced Classification. Stratified train/test splits is an option too. But question is specifically asking "cross-validation" strategy.

upvoted 4 times

Question #101

Topic 1

A technology startup is using complex deep neural networks and GPU compute to recommend the company's products to its existing customers based upon each customer's habits and interactions. The solution currently pulls each dataset from an Amazon S3 bucket before loading the data into a TensorFlow model pulled from the company's Git repository that runs locally. This job then runs for several hours while continually outputting its progress to the same S3 bucket. The job can be paused, restarted, and continued at any time in the event of a failure, and is run from a central queue.

Senior managers are concerned about the complexity of the solution's resource management and the costs involved in repeating the process regularly. They ask for the workload to be automated so it runs once a week, starting Monday and completing by the close of business Friday.


Which architecture should be used to scale the solution at the lowest cost?

- A. Implement the solution using AWS Deep Learning Containers and run the container as a job using AWS Batch on a GPU-compatible Spot

Instance

- B. Implement the solution using a low-cost GPU-compatible Amazon EC2 instance and use the AWS Instance Scheduler to schedule the task
- C. Implement the solution using AWS Deep Learning Containers, run the workload using AWS Fargate running on Spot Instances, and then schedule the task using the built-in task scheduler
- D. Implement the solution using Amazon ECS running on Spot Instances and schedule the task using the ECS service scheduler

Correct Answer: C

  **SophieSu** 4 months, 1 week ago

D is correct.

upvoted 1 times

  **SophieSu** 4 months ago



D is not correct. ECS is responsible for managing the lifecycle and placement of tasks. However, ECS does not run or execute your container. ECS only provides the control plane to manage tasks.

A is correct.

"You can set up compute environments that use a particular type of EC2 instance, a particular model such as c5.2xlarge or m5.10xlarge simply specify that you want to use the newest instance types. You can also specify the minimum, desired, and maximum number of vCPUs for the environment, along with the amount you are willing to pay for a Spot Instance as a percentage of the On-Demand Instance price and a target set of VPC subnets. AWS Batch will efficiently launch, manage, and terminate compute types as needed. You can also manage your own compute environments. In this case you are responsible for setting up and scaling the instances in an Amazon ECS cluster that AWS Batch creates for you. "

<https://docs.aws.amazon.com/batch/latest/userguide/what-is-batch.html>



upvoted 8 times

  **jdstone** 4 months, 2 weeks ago

Answer is A

<https://aws.amazon.com/blogs/compute/gpu-workloads-on-aws-batch/>

upvoted 4 times

  **Juka3lj** 3 months, 2 weeks ago

Makes most sense

upvoted 1 times

  **astonm13** 4 months, 2 weeks ago

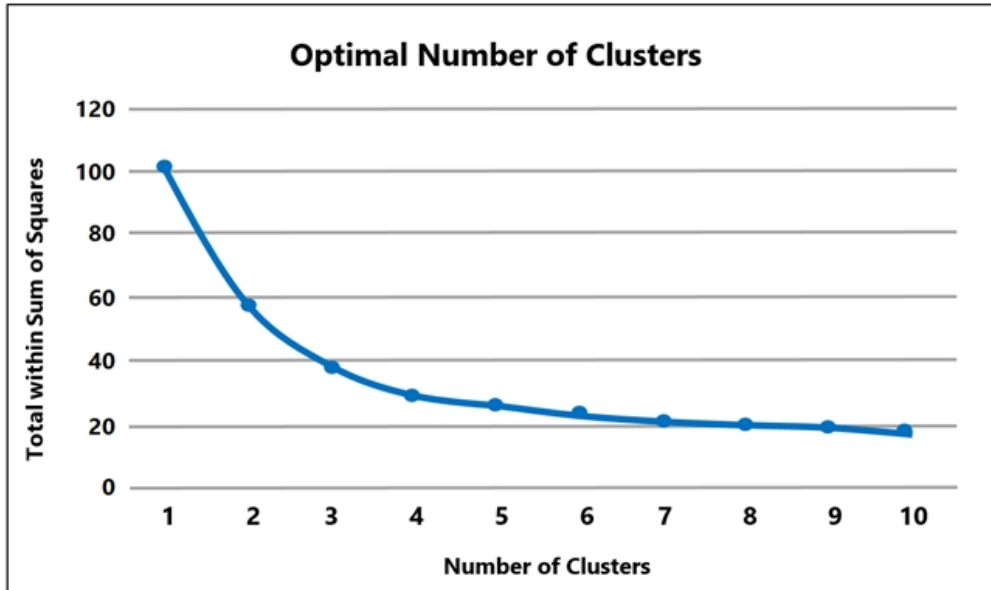
I would go for D. As far as I know Fargate does not support GPU computing.

upvoted 4 times

Question #102

Topic 1

A Machine Learning Specialist prepared the following graph displaying the results of k-means for $k = [1..10]$:



Considering the graph, what is a reasonable selection for the optimal choice of k ?

- A. 1
- B. 4
- C. 7
- D. 10

Correct Answer: C

ksrivastavaSumit Highly Voted 4 months, 3 weeks ago
B seems correct based on the elbow method
upvoted 8 times

Juka3lj 3 months, 2 weeks ago
I agree, most likely B.
upvoted 2 times

Vita_Rasta84444 Most Recent 2 months, 2 weeks ago
number 4 is the elbow of the hand, B is correct
upvoted 2 times

cnethers 4 months, 3 weeks ago
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 percentile
upvoted 2 times



Question #103



Topic 1



A media company with a very large archive of unlabeled images, text, audio, and video footage wishes to index its assets to allow rapid identification of relevant content by the Research team. The company wants to use machine learning to accelerate the efforts of its in-house researchers who have limited machine learning expertise.



Which is the FASTEST route to index the assets?



- A. Use Amazon Rekognition, Amazon Comprehend, and Amazon Transcribe to tag data into distinct categories/classes.
- B. Create a set of Amazon Mechanical Turk Human Intelligence Tasks to label all footage.
- 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.
- D. Use the AWS Deep Learning AMI and Amazon EC2 GPU instances to create custom models for audio transcription and topic modeling, and use object detection to tag data into distinct categories/classes.



Correct Answer: A



  **SophieSu** Highly Voted 4 months, 1 week ago
A. Fastest route must Amazon Services.
upvoted 9 times



  **AShahine21** 1 month ago
Amazon Mechanical Turk is an Amazon service
upvoted 1 times

  **YJ4219** Most Recent 5 days, 10 hours ago
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
upvoted 1 times

  **AjithkumarSL** 3 weeks, 5 days ago
Would go for A
upvoted 1 times

  **AShahine21** 1 month ago
I will go with B
upvoted 1 times

  **Juka3lj** 3 months, 2 weeks ago
Correct answer is A
upvoted 2 times

  **ksrivastavaSumit** 4 months, 3 weeks ago
B. as no one in-house is an expert and It probably is the fastest way to get there
upvoted 2 times

Question #104

Topic 1

A Machine Learning Specialist is working for an online retailer that wants to run analytics on every customer visit, processed through a machine learning pipeline.

The data needs to be ingested by Amazon Kinesis Data Streams at up to 100 transactions per second, and the JSON data blob is 100 KB in size. What is the MINIMUM number of shards in Kinesis Data Streams the Specialist should use to successfully ingest this data?

- A. 1 shards
- B. 10 shards
- C. 100 shards
- D. 1,000 shards

Correct Answer: B

  **joep21** Highly Voted 4 months, 3 weeks ago

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 $100 \text{ KB} * 100 = 10 \text{ MB}$ (10 shards required)
upvoted 11 times

  **benson2021** Most Recent 2 months, 3 weeks ago

Reference: <https://docs.aws.amazon.com/streams/latest/dev/service-sizes-and-limits.html>
upvoted 2 times

Question #105

Topic 1

A Machine Learning Specialist is deciding between building a naive Bayesian model or a full Bayesian network for a classification problem. The Specialist computes the Pearson correlation coefficients between each feature and finds that their absolute values range between 0.1 to 0.95. Which model describes the underlying data in this situation?


- A. A naive Bayesian model, since the features are all conditionally independent.
- B. A full Bayesian network, since the features are all conditionally independent.
- C. A naive Bayesian model, since some of the features are statistically dependent.
- D. A full Bayesian network, since some of the features are statistically dependent.

Correct Answer: C

 **joep21** Highly Voted 4 months, 3 weeks ago

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>

upvoted 10 times

 **Juka3lj** 3 months, 2 weeks ago

I agree, makes moste sense

upvoted 1 times

 **SophieSu** Highly Voted 4 months, 1 week ago


D. Naive bayes - features are independent given the class.

upvoted 6 times

 **Vita_Rasta84444** Most Recent 3 months ago

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.

upvoted 2 times

 **astonm13** 4 months, 2 weeks ago

I would say, B. Naive Bayes assumes conditional independence and not statistical

upvoted 2 times

 **abdohanfi** 1 month ago

you mean (a) naive bayes not (b)

upvoted 1 times

 **cnethers** 4 months, 3 weeks ago

This is also a good source of information to help build your understanding <https://www.simplypsychology.org/correlation.html>

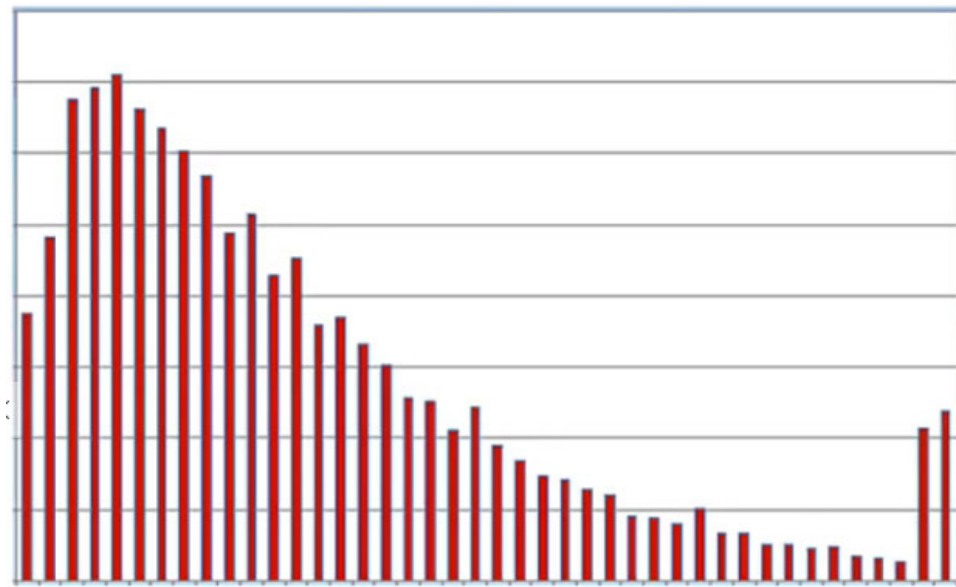
upvoted 1 times

Question #106

Topic 1

A Data Scientist is building a linear regression model and will use resulting p-values to evaluate the statistical significance of each coefficient. Upon inspection of the dataset, the Data Scientist discovers that most of the features are normally distributed. The plot of one feature in the

dataset is shown in the graphic.



What transformation should the Data Scientist apply to satisfy the statistical assumptions of the linear regression model?

- A. Exponential transformation
- B. Logarithmic transformation
- C. Polynomial transformation
- D. Sinusoidal transformation

Correct Answer: A

astonm13 Highly Voted 4 months, 2 weeks ago

I would say B. Logarithmic transformation converts skewed distributions towards normal
upvoted 10 times

YJ4219 Most Recent 5 days, 10 hours ago

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 calculation 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."
upvoted 1 times

konradL 2 months, 1 week ago

I would also go for B, as Log transformation is often mentioned, when we are talking about right (positive) skewness.
upvoted 2 times

Question #107

Topic 1

A Machine Learning Specialist is assigned to a Fraud Detection team and must tune an XGBoost model, which is working appropriately for test

data. However, with unknown data, it is not working as expected. The existing parameters are provided as follows.

```
param = {  
    'eta': 0.05, # the training step for each iteration  
    'silent': 1, # logging mode - quiet  
    'n_estimators': 2000,  
    'max_depth': 30,  
    'min_child_weight': 3,  
    'gamma': 0,  
    'subsample': 0.8,  
    'objective': 'multi:softprob', # error evaluation for multiclass training  
    'num_class': 201} # the number of classes that exist in this dataset  
num_round = 60 # the number of training iterations
```

Which parameter tuning guidelines should the Specialist follow to avoid overfitting?

- A. Increase the max_depth parameter value.
- B. Lower the max_depth parameter value.
- C. Update the objective to binary:logistic.
- D. Lower the min_child_weight parameter value.

Correct Answer: B

  **SophieSu** Highly Voted 4 months ago

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.

upvoted 8 times

  **cnethers** Most Recent 4 months, 3 weeks ago

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



upvoted 2 times

  **arulrajayaraj** 4 months, 3 weeks ago

Ans : B , Lower values avoid over-fitting.



No for D - Larger values avoid over-fitting.

upvoted 5 times

  **cnethers** 4 months, 3 weeks ago


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

upvoted 2 times

  **cnethers** 4 months, 3 weeks ago

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 a 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.

upvoted 1 times

  **cnethers** 4 months, 3 weeks ago

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.

upvoted 2 times

Question #108

Topic 1

A data scientist is developing a pipeline to ingest streaming web traffic data. The data scientist needs to implement a process to identify unusual web traffic patterns as part of the pipeline. The patterns will be used downstream for alerting and incident response. The data scientist has access to unlabeled historic data to use, if needed.

The solution needs to do the following:

☞ Calculate an anomaly score for each web traffic entry.

Adapt unusual event identification to changing web patterns over time.

▪

Which approach should the data scientist implement to meet these requirements?

- A. Use historic web traffic data to train an anomaly detection model using the Amazon SageMaker Random Cut Forest (RCF) built-in model. Use an Amazon Kinesis Data Stream to process the incoming web traffic data. Attach a preprocessing AWS Lambda function to perform data enrichment by calling the RCF model to calculate the anomaly score for each record.
- B. Use historic web traffic data to train an anomaly detection model using the Amazon SageMaker built-in XGBoost model. Use an Amazon Kinesis Data Stream to process the incoming web traffic data. Attach a preprocessing AWS Lambda function to perform data enrichment by calling the XGBoost model to calculate the anomaly score for each record.
- C. Collect the streaming data using Amazon Kinesis Data Firehose. Map the delivery stream as an input source for Amazon Kinesis Data Analytics. Write a SQL query to run in real time against the streaming data with the k-Nearest Neighbors (kNN) SQL extension to calculate anomaly scores for each record using a tumbling window.
- D. Collect the streaming data using Amazon Kinesis Data Firehose. Map the delivery stream as an input source for Amazon Kinesis Data Analytics. Write a SQL query to run in real time against the streaming data with the Amazon Random Cut Forest (RCF) SQL extension to calculate anomaly scores for each record using a sliding window.

Correct Answer: A

🗨️ **jiadong** Highly Voted 4 months, 2 weeks ago

I think the answer is D - RCF works together with Data Analytics, and sliding window helped on new information
upvoted 9 times

🗨️ **SophieSu** 4 months ago

better to say "RCF is a built-in algorithm/function in Kinesis Data Analytics"
upvoted 1 times

🗨️ **gbrnq** Most Recent 2 months, 1 week ago

"Adapt unusual event identification to changing web patterns over time." -> option A does not satisfy this, only mentions build the model once
upvoted 1 times

🗨️ **randomnamer** 3 months ago

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.
upvoted 2 times

🗨️ **seanLu** 4 months ago

A and D both seem to work. 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.
upvoted 1 times

🗨️ **astonm13** 4 months, 2 weeks ago

It is definitely D
upvoted 1 times

Question #109

Topic 1

A Data Scientist received a set of insurance records, each consisting of a record ID, the final outcome among 200 categories, and the date of the final outcome.

Some partial information on claim contents is also provided, but only for a few of the 200 categories. For each outcome category, there are hundreds of records distributed over the past 3 years. The Data Scientist wants to predict how many claims to expect in each category from month to month, a few months in advance.

What type of machine learning model should be used?

- A. Classification month-to-month using supervised learning of the 200 categories based on claim contents.
- B. Reinforcement learning using claim IDs and timestamps where the agent will identify how many claims in each category to expect from month to month.
- C. Forecasting using claim IDs and timestamps to identify how many claims in each category to expect from month to month.
- D. Classification with supervised learning of the categories for which partial information on claim contents is provided, and forecasting using claim IDs and timestamps for all other categories.

Correct Answer: D

  **JBX2010** Highly Voted 4 months, 2 weeks ago

I think it should be C as the final outcome among 200 categories is already known. No need to build a classification model. It's pure forecast problem.

upvoted 10 times

  **abdohanfi** 1 month ago

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

upvoted 1 times

  **randomnamer** Most Recent 3 months ago

It is true that the final outcome is known. But C does not use the partial information from the 200 categories. Reinforcement learning current state of the art in stock prediction and other time series. Why waste valuable information? For me it's B.


upvoted 1 times

  **SophieSu** 4 months ago

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.

upvoted 3 times

  **cnethers** 4 months, 3 weeks ago

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."

upvoted 1 times



Question #110

Topic 1

A company that promotes healthy sleep patterns by providing cloud-connected devices currently hosts a sleep tracking application on AWS. The application collects device usage information from device users. The company's Data Science team is building a machine learning model to predict if and when a user will stop utilizing the company's devices. Predictions from this model are used by a downstream application that determines the best approach for contacting users.

The Data Science team is building multiple versions of the machine learning model to evaluate each version against the company's business goals. To measure long-term effectiveness, the team wants to run multiple versions of the model in parallel for long periods of time, with the ability to control the portion of inferences served by the models.

Which solution satisfies these requirements with MINIMAL effort?

- A. Build and host multiple models in Amazon SageMaker. Create multiple Amazon SageMaker endpoints, one for each model. Programmatically control invoking different models for inference at the application layer.
- B. Build and host multiple models in Amazon SageMaker. Create an Amazon SageMaker endpoint configuration with multiple production variants. Programmatically control the portion of the inferences served by the multiple models by updating the endpoint configuration.
- C. Build and host multiple models in Amazon SageMaker Neo to take into account different types of medical devices. Programmatically control which model is invoked for inference based on the medical device type.
- D. Build and host multiple models in Amazon SageMaker. Create a single endpoint that accesses multiple models. Use Amazon SageMaker batch transform to control invoking the different models through the single endpoint.

Correct Answer: D

 **SophieSu** Highly Voted 4 months ago

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.

upvoted 7 times

 **joep21** Most Recent 4 months, 2 weeks ago

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>

upvoted 3 times

Question #111

Topic 1

An agricultural company is interested in using machine learning to detect specific types of weeds in a 100-acre grassland field. Currently, the

company uses tractor-mounted cameras to capture multiple images of the field as 10 Å– 10 grids. The company also has a large training dataset that consists of annotated images of popular weed classes like broadleaf and non-broadleaf docks.

The company wants to build a weed detection model that will detect specific types of weeds and the location of each type within the field. Once the model is ready, it will be hosted on Amazon SageMaker endpoints. The model will perform real-time inferencing using the images captured by the cameras.

Which approach should a Machine Learning Specialist take to obtain accurate predictions?

- A. Prepare the images in RecordIO format and upload them to Amazon S3. Use Amazon SageMaker to train, test, and validate the model using an image classification algorithm to categorize images into various weed classes.
- B. Prepare the images in Apache Parquet format and upload them to Amazon S3. Use Amazon SageMaker to train, test, and validate the model using an object- detection single-shot multibox detector (SSD) algorithm.
- C. Prepare the images in RecordIO format and upload them to Amazon S3. Use Amazon SageMaker to train, test, and validate the model using an object- detection single-shot multibox detector (SSD) algorithm.
- D. Prepare the images in Apache Parquet format and upload them to Amazon S3. Use Amazon SageMaker to train, test, and validate the model using an image classification algorithm to categorize images into various weed classes.

Correct Answer: C

 **SophieSu** Highly Voted 4 months ago

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.

upvoted 9 times

Question #112

Topic 1

A manufacturer is operating a large number of factories with a complex supply chain relationship where unexpected downtime of a machine can cause production to stop at several factories. A data scientist wants to analyze sensor data from the factories to identify equipment in need of preemptive maintenance and then dispatch a service team to prevent unplanned downtime. The sensor readings from a single machine can include up to 200 data points including temperatures, voltages, vibrations, RPMs, and pressure readings.

To collect this sensor data, the manufacturer deployed Wi-Fi and LANs across the factories. Even though many factory locations do not have reliable or high-speed internet connectivity, the manufacturer would like to maintain near-real-time inference capabilities.

Which deployment architecture for the model will address these business requirements?

- A. Deploy the model in Amazon SageMaker. Run sensor data through this model to predict which machines need maintenance.
- B. Deploy the model on AWS IoT Greengrass in each factory. Run sensor data through this model to infer which machines need maintenance.
- C. Deploy the model to an Amazon SageMaker batch transformation job. Generate inferences in a daily batch report to identify machines that need maintenance.
- D. Deploy the model in Amazon SageMaker and use an IoT rule to write data to an Amazon DynamoDB table. Consume a DynamoDB stream from the table with an AWS Lambda function to invoke the endpoint.

Correct Answer: A

  **joep21** Highly Voted 4 months, 3 weeks ago

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-ic-services/>

<https://aws.amazon.com/blogs/iot/using-aws-iot-for-predictive-maintenance/>

upvoted 8 times

  **Vita_Rasta84444** Most Recent 3 months ago

I would choose B because IoT reduce latency because they work on local machine

upvoted 1 times

  **SophieSu** 4 months ago

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

upvoted 4 times

  **astonm13** 4 months, 2 weeks ago

I would choose B

upvoted 1 times

Question #113

Topic 1

A Machine Learning Specialist is designing a scalable data storage solution for Amazon SageMaker. There is an existing TensorFlow-based model implemented as a train.py script that relies on static training data that is currently stored as TFRecords.

Which method of providing training data to Amazon SageMaker would meet the business requirements with the LEAST development overhead?

- A. Use Amazon SageMaker script mode and use train.py unchanged. Point the Amazon SageMaker training invocation to the local path of the data without reformatting the training data.
- B. Use Amazon SageMaker script mode and use train.py unchanged. Put the TFRecord data into an Amazon S3 bucket. Point the Amazon SageMaker training invocation to the S3 bucket without reformatting the training data.
- C. Rewrite the train.py script to add a section that converts TFRecords to protobuf and ingests the protobuf data instead of TFRecords.
- D. Prepare the data in the format accepted by Amazon SageMaker. Use AWS Glue or AWS Lambda to reformat and store the data in an Amazon S3 bucket.

Correct Answer: D

  **joep21** Highly Voted 4 months, 3 weeks ago

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>

upvoted 10 times

  **SophieSu** Highly Voted 4 months ago

B is my answer.

Reading Data

```
filenames = ["s3://bucketname/path/to/file1.tfrecord",
```

```
"s3://bucketname/path/to/file2.tfrecord"]
```

```
dataset = tf.data.TFRecordDataset(filenames)
```

upvoted 6 times

  **cnethers** Most Recent 4 months, 3 weeks ago

Unfortunately 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 create the TensorFlow estimator, an S3 location under the default training job bucket is used. Distributed training with parameter servers requires 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
```


Because of the pre-rec Ans A and B are an easy disqualification.

There is no need to change the training format so option C is a red herring

Ans is D

Not the most obvious answer

upvoted 3 times

  **SophieSu** 4 months, 1 week ago

according to your explanation, the correct answer should be B

upvoted 2 times

Question #114

Topic 1

The chief editor for a product catalog wants the research and development team to build a machine learning system that can be used to detect whether or not individuals in a collection of images are wearing the company's retail brand. The team has a set of training data.

Which machine learning algorithm should the researchers use that BEST meets their requirements?

- A. Latent Dirichlet Allocation (LDA)
- B. Recurrent neural network (RNN)
- C. K-means
- D. Convolutional neural network (CNN)

Correct Answer: D



SophieSu Highly Voted 4 months, 1 week ago

D. CNN - image

upvoted 11 times

Question #115

Topic 1

A retail company is using Amazon Personalize to provide personalized product recommendations for its customers during a marketing campaign. The company sees a significant increase in sales of recommended items to existing customers immediately after deploying a new solution version, but these sales decrease a short time after deployment. Only historical data from before the marketing campaign is available for training. How should a data scientist adjust the solution?

- A. Use the event tracker in Amazon Personalize to include real-time user interactions.
- B. Add user metadata and use the HRNN-Metadata recipe in Amazon Personalize.
- C. Implement a new solution using the built-in factorization machines (FM) algorithm in Amazon SageMaker.
- D. Add event type and event value fields to the interactions dataset in Amazon Personalize.

Correct Answer: D

  **SophieSu** Highly Voted  4 months, 1 week ago

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 new event data.

upvoted 15 times

  **AjithkumarSL** 3 weeks, 5 days ago

I agree.. A is the right choice.. The model need the real time data to adjust to create recommendations..

upvoted 1 times

Question #116

Topic 1

A machine learning (ML) specialist wants to secure calls to the Amazon SageMaker Service API. The specialist has configured Amazon VPC with a VPC interface endpoint for the Amazon SageMaker Service API and is attempting to secure traffic from specific sets of instances and IAM users. The VPC is configured with a single public subnet.

Which combination of steps should the ML specialist take to secure the traffic? (Choose two.)

- A. Add a VPC endpoint policy to allow access to the IAM users.
- B. Modify the users' IAM policy to allow access to Amazon SageMaker Service API calls only.
- C. Modify the security group on the endpoint network interface to restrict access to the instances.
- D. Modify the ACL on the endpoint network interface to restrict access to the instances.
- E. Add a SageMaker Runtime VPC endpoint interface to the VPC.

Correct Answer: AC

Reference:

<https://aws.amazon.com/blogs/machine-learning/private-package-installation-in-amazon-sagemaker-running-in-internet-free-mode/>

 **mona_mansour** Highly Voted 1 month, 3 weeks ago

A&C...><https://aws.amazon.com/blogs/machine-learning/securing-all-amazon-sagemaker-api-calls-with-aws-privatelink/>
upvoted 6 times

 **cnethers** Most Recent 4 months, 3 weeks ago

<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>
upvoted 1 times

Question #117

Topic 1

An e commerce company wants to launch a new cloud-based product recommendation feature for its web application. Due to data localization regulations, any sensitive data must not leave its on-premises data center, and the product recommendation model must be trained and tested using nonsensitive data only. Data transfer to the cloud must use IPsec. The web application is hosted on premises with a PostgreSQL database that contains all the data. The company wants the data to be uploaded securely to Amazon S3 each day for model retraining.

How should a machine learning specialist meet these requirements?

- 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.
- B. Create an AWS Glue job to connect to the PostgreSQL DB instance. Ingest all data through an AWS Site-to-Site VPN connection into Amazon S3 while removing sensitive data using a PySpark job.
- C. 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 into Amazon S3.

D. Use PostgreSQL logical replication to replicate all data to PostgreSQL in Amazon EC2 through AWS Direct Connect with a VPN connection.
Use AWS Glue to move data from Amazon EC2 to Amazon S3.

Correct Answer: C

Reference:

https://docs.aws.amazon.com/dms/latest/userguide/CHAP_Source.PostgreSQL.html

  **cnethers** Highly Voted 4 months, 3 weeks ago

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



Answer C only makes a 443 (SSL) connection so does not meet the IPsec requirement

upvoted 9 times

  **ksrivastavaSumit** Highly Voted 4 months, 3 weeks ago

A? IPsec needs to be covered as well

upvoted 5 times

  **StelSen** 1 month, 2 weeks ago

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.

upvoted 1 times



  **granmastak** Most Recent 4 days, 5 hours ago

This is a very tricky question since DMS over VPN started to be supported Nov 2020

<https://aws.amazon.com/about-aws/whats-new/2020/11/now-privately-connect-to-aws-database-migration-service-from-amazon-virtual-private-cloud/>

So both A&C can work. If this question is older then A is the right answer, there have been a bit over 6 months since November 2020 so I would expect this question to become a bit clearer if it stays on the exam in its current form



upvoted 1 times

  **StelSen** 1 month, 2 weeks ago

I was bit confused between C & A. But chose A. <https://aws.amazon.com/blogs/big-data/how-to-access-and-analyze-on-premises-data-stores-using-aws-glue/>

Because one of the requirement is to go through via IPsec. So, VPN required. DMS is not IPsec based although it's secured. AWS Glue supports on-premise

upvoted 1 times

  **joep21** 4 months, 3 weeks ago

C it is, explanation is good.

upvoted 2 times

Question #118

Topic 1

A logistics company needs a forecast model to predict next month's inventory requirements for a single item in 10 warehouses. A machine learning specialist uses

Amazon Forecast to develop a forecast model from 3 years of monthly data. There is no missing data. The specialist selects the DeepAR+ algorithm to train a predictor. The predictor means absolute percentage error (MAPE) is much larger than the MAPE produced by the current human forecasters.

Which changes to the CreatePredictor API call could improve the MAPE? (Choose two.)

- A. Set PerformAutoML to true.
- B. Set ForecastHorizon to 4.
- C. Set ForecastFrequency to W for weekly.
- D. Set PerformHPO to true.
- E. Set FeaturizationMethodName to filling.

Correct Answer: CD

Reference:

<https://docs.aws.amazon.com/forecast/latest/dg/forecast.dg.pdf>



  **scuzzy2010**  3 months, 3 weeks ago

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 not necessarily both.

A - If you want Amazon Forecast to evaluate each algorithm and choose the one that minimizes the objective function, set PerformAutoML true.

D - The following algorithms support HPO: - > DeepAR+.

upvoted 5 times

  **Oscaaaaar** 3 months, 3 weeks ago

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>

upvoted 1 times



  **mona_mansour** Most Recent 1 month, 3 weeks ago

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 metric 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.

upvoted 2 times

  **Vita_Rasta84444** 3 months ago

It is A and D, there are no weekly data, they have only monthly data and can not switch horizon to 4

upvoted 4 times

  **SophieSu** 4 months ago

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.

upvoted 2 times

  **seanLu** 4 months ago

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.

upvoted 5 times

  **SophieSu** 4 months ago

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.

upvoted 2 times

Question #119

Topic 1

A data scientist wants to use Amazon Forecast to build a forecasting model for inventory demand for a retail company. The company has provided a dataset of historic inventory demand for its products as a .csv file stored in an Amazon S3 bucket. The table below shows a sample of the dataset.

timestamp	item_id	demand	category	lead_time
2019-12-14	uni_000736	120	hardware	90
2020-01-31	uni_003429	98	hardware	30
2020-03-04	uni_000211	234	accessories	10

How should the data scientist transform the data?

- A. Use ETL jobs in AWS Glue to separate the dataset into a target time series dataset and an item metadata dataset. Upload both datasets as .csv files to Amazon S3.
- B. Use a Jupyter notebook in Amazon SageMaker to separate the dataset into a related time series dataset and an item metadata dataset. Upload both datasets as tables in Amazon Aurora.
- C. Use AWS Batch jobs to separate the dataset into a target time series dataset, a related time series dataset, and an item metadata dataset. Upload them directly to Forecast from a local machine.
- D. Use a Jupyter notebook in Amazon SageMaker to transform the data into the optimized protobuf recordIO format. Upload the dataset in this format to Amazon S3.

Correct Answer: B

  **joep21** Highly Voted 4 months, 3 weeks ago

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>

upvoted 12 times



  **DSJingguo** Most Recent 2 months ago

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."

upvoted 1 times

  **achiko** 3 months, 1 week ago

lead time belongs to related time series, as its not a target variable

upvoted 1 times

Question #120

Topic 1

A machine learning specialist is running an Amazon SageMaker endpoint using the built-in object detection algorithm on a P3 instance for real-time predictions in a company's production application. When evaluating the model's resource utilization, the specialist notices that the model is using only a fraction of the GPU.

Which architecture changes would ensure that provisioned resources are being utilized effectively?

- A. Redeploy the model as a batch transform job on an M5 instance.
- B. Redeploy the model on an M5 instance. Attach Amazon Elastic Inference to the instance.
- C. Redeploy the model on a P3dn instance.
- D. Deploy the model onto an Amazon Elastic Container Service (Amazon ECS) cluster using a P3 instance.



Correct Answer: D

  **joep21** Highly Voted 4 months, 3 weeks ago

B is correct. Redeploy with CPU and add elastic inference to reduce costs. See: <https://aws.amazon.com/machine-learning/elastic-inference/>
upvoted 12 times

  **mona_mansour** Most Recent 1 month, 3 weeks ago

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.
upvoted 2 times

  **Vita_Rasta84444** 3 months ago

B is correct
upvoted 1 times

Question #121

Topic 1

A data scientist uses an Amazon SageMaker notebook instance to conduct data exploration and analysis. This requires certain Python packages that are not natively available on Amazon SageMaker to be installed on the notebook instance.

How can a machine learning specialist ensure that required packages are automatically available on the notebook instance for the data scientist to use?

- A. Install AWS Systems Manager Agent on the underlying Amazon EC2 instance and use Systems Manager Automation to execute the package installation commands.
- B. Create a Jupyter notebook file (.ipynb) with cells containing the package installation commands to execute and place the file under the /etc/init directory of each Amazon SageMaker notebook instance.
- C. Use the conda package manager from within the Jupyter notebook console to apply the necessary conda packages to the default kernel of the notebook.
- D. Create an Amazon SageMaker lifecycle configuration with package installation commands and assign the lifecycle configuration to the

notebook instance.

Correct Answer: B

Reference:

<https://towardsdatascience.com/automating-aws-sagemaker-notebooks-2dec62bc2c84>

🗲️ 👤 **joep21** Highly Voted 👍 4 months, 3 weeks ago

I would select D. See AWS documentation: <https://docs.aws.amazon.com/sagemaker/latest/dg/nbi-add-external.html>
upvoted 8 times

🗲️ 👤 **abdohanfi** Most Recent 🕒 1 month ago

based on the reference given under the answer its D not B
upvoted 1 times

🗲️ 👤 **mona_mansour** 1 month, 3 weeks ago

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.

upvoted 1 times

🗲️ 👤 **mona_mansour** 1 month, 3 weeks ago

NOT B ...>/etc/init contains configuration files used by Upstart.
ANS...>D

upvoted 1 times

🗲️ 👤 **astonm13** 4 months, 2 weeks ago

Its for sure D
upvoted 1 times

🗲️ 👤 **ksrivastavaSumit** 4 months, 3 weeks ago

D <https://docs.aws.amazon.com/sagemaker/latest/dg/nbi-add-external.html>
upvoted 3 times

Question #122

Topic 1

A data scientist needs to identify fraudulent user accounts for a company's ecommerce platform. The company wants the ability to determine if a newly created account is associated with a previously known fraudulent user. The data scientist is using AWS Glue to cleanse the company's application logs during ingestion.

Which strategy will allow the data scientist to identify fraudulent accounts?

- A. Execute the built-in FindDuplicates Amazon Athena query.
- B. Create a FindMatches machine learning transform in AWS Glue.
- C. Create an AWS Glue crawler to infer duplicate accounts in the source data.
- D. Search for duplicate accounts in the AWS Glue Data Catalog.

Correct Answer: B


Reference:

<https://docs.aws.amazon.com/glue/latest/dg/machine-learning.html>

  **mona_mansour** 1 month, 3 weeks ago

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.



upvoted 3 times

  **omar_bahrain** 4 months, 2 weeks ago

Agree. Please refer to:

<https://aws.amazon.com/blogs/big-data/integrate-and-deduplicate-datasets-using-aws-lake-formation-findmatches/>

upvoted 1 times

  **joep21** 4 months, 3 weeks ago

B it is. Reasonable explanation.

upvoted 4 times

Question #123

Topic 1

A Data Scientist is developing a machine learning model to classify whether a financial transaction is fraudulent. The labeled data available for training consists of

100,000 non-fraudulent observations and 1,000 fraudulent observations.





The Data Scientist applies the XGBoost algorithm to the data, resulting in the following confusion matrix when the trained model is applied to a previously unseen validation dataset. The accuracy of the model is 99.1%, but the Data Scientist needs to reduce the number of false negatives.

Predicted	0	1
Actual	0 99,966	34
	1 877	123

Which combination of steps should the Data Scientist take to reduce the number of false negative predictions by the model? (Choose two.)

- A. Change the XGBoost eval_metric parameter to optimize based on Root Mean Square Error (RMSE).
- B. Increase the XGBoost scale_pos_weight parameter to adjust the balance of positive and negative weights.
- C. Increase the XGBoost max_depth parameter because the model is currently underfitting the data.
- D. Change the XGBoost eval_metric parameter to optimize based on Area Under the ROC Curve (AUC).
- E. Decrease the XGBoost max_depth parameter because the model is currently overfitting the data.

Correct Answer: BD

-  **StelSen** 1 month, 2 weeks ago
Same as Qstn-59. B&D is answer
upvoted 2 times
-  **mona_mansour** 1 month, 3 weeks ago
my ANS is B & D
upvoted 2 times
-  **takahirokoyama** 4 months, 3 weeks ago
D,E. Model clearly shows over-fitting.
upvoted 2 times
-  **SophieSu** 4 months, 1 week ago
No. The target is to reduce the FN. Answers are BD.
upvoted 12 times

Question #124

Topic 1

A data scientist has developed a machine learning translation model for English to Japanese by using Amazon SageMaker's built-in seq2seq algorithm with

500,000 aligned sentence pairs. While testing with sample sentences, the data scientist finds that the translation quality is reasonable for an example as short as five words. However, the quality becomes unacceptable if the sentence is 100 words long.

Which action will resolve the problem?

- A. Change preprocessing to use n-grams.
- B. Add more nodes to the recurrent neural network (RNN) than the largest sentence's word count.
- C. Adjust hyperparameters related to the attention mechanism.
- D. Choose a different weight initialization type.

Correct Answer: B

  **cnethers** Highly Voted 4 months, 3 weeks ago

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.



upvoted 9 times

  **Juka3lj** Most Recent 2 months ago

c is correct

<https://docs.aws.amazon.com/sagemaker/latest/dg/seq-2-seq-howitworks.html>

upvoted 1 times

  **rajesriv** 4 months, 3 weeks ago

I believe the answer is C

<https://docs.aws.amazon.com/sagemaker/latest/dg/seq-2-seq-howitworks.html>

upvoted 3 times

Question #125

Topic 1

A financial company is trying to detect credit card fraud. The company observed that, on average, 2% of credit card transactions were fraudulent. A data scientist trained a classifier on a year's worth of credit card transactions data. The model needs to identify the fraudulent transactions (positives) from the regular ones

(negatives). The company's goal is to accurately capture as many positives as possible.

Which metrics should the data scientist use to optimize the model? (Choose two.)

- A. Specificity
- B. False positive rate
- C. Accuracy
- D. Area under the precision-recall curve
- E. True positive rate

Correct Answer: *AB*

 **littlewat** Highly Voted 3 months, 1 week ago

D, E is the answer. we need to make the recall rate(not precision) high.
upvoted 12 times

 **joep21** Highly Voted 4 months, 3 weeks ago

To maximize detection of fraud in real-world, imbalanced datasets, D and E should always be applied.

https://en.wikipedia.org/wiki/Sensitivity_and_specificity

Question #126

Topic 1

A machine learning specialist is developing a proof of concept for government users whose primary concern is security. The specialist is using Amazon

SageMaker to train a convolutional neural network (CNN) model for a photo classifier application. The specialist wants to protect the data so that it cannot be accessed and transferred to a remote host by malicious code accidentally installed on the training container.

Which action will provide the MOST secure protection?

- A. Remove Amazon S3 access permissions from the SageMaker execution role.
- B. Encrypt the weights of the CNN model.
- C. Encrypt the training and validation dataset.
- D. Enable network isolation for training jobs.

Correct Answer: D

🗨️ 👤 **achiko** Highly Voted 👍 3 months, 1 week ago

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.

upvoted 5 times

🗨️ 👤 **omar_bahrain** Highly Voted 👍 3 months, 3 weeks ago

most likely it is C.

<https://docs.aws.amazon.com/sagemaker/latest/dg/data-protection.html>

upvoted 5 times

🗨️ 👤 **AShahine21** Most Recent 🕒 1 month, 2 weeks ago

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."

upvoted 2 times

🗨️ 👤 **benson2021** 2 months, 2 weeks ago

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.

upvoted 2 times

🗨️ 👤 **Vita_Rasta84444** 3 months ago

I would choose C

upvoted 1 times

Question #127

Topic 1

A medical imaging company wants to train a computer vision model to detect areas of concern on patients' CT scans. The company has a large collection of unlabeled CT scans that are linked to each patient and stored in an Amazon S3 bucket. The scans must be accessible to authorized users only. A machine learning engineer needs to build a labeling pipeline.

Which set of steps should the engineer take to build the labeling pipeline with the LEAST effort?

- A. Create a workforce with AWS Identity and Access Management (IAM). Build a labeling tool on Amazon EC2 Queue images for labeling by using Amazon Simple Queue Service (Amazon SQS). Write the labeling instructions.
- B. Create an Amazon Mechanical Turk workforce and manifest file. Create a labeling job by using the built-in image classification task type in Amazon SageMaker Ground Truth. Write the labeling instructions.
- C. Create a private workforce and manifest file. Create a labeling job by using the built-in bounding box task type in Amazon SageMaker Ground Truth. Write the labeling instructions.
- D. Create a workforce with Amazon Cognito. Build a labeling web application with AWS Amplify. Build a labeling workflow backend using AWS Lambda. Write the labeling instructions.

Correct Answer: B

  **joep21** Highly Voted 4 months, 3 weeks ago

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>

upvoted 8 times

  **CharlesChiang** 3 months, 1 week ago



Agree C

upvoted 1 times

  **astonm13** 4 months, 2 weeks ago

Yes it is C

upvoted 1 times

  **cnethers** 4 months, 3 weeks ago

agree C

upvoted 1 times

  **benson2021** Most Recent 2 months, 2 weeks ago

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>

upvoted 2 times

Question #128

Topic 1

A company is using Amazon Textract to extract textual data from thousands of scanned text-heavy legal documents daily. The company uses this information to process loan applications automatically. Some of the documents fail business validation and are returned to human reviewers, who investigate the errors. This activity increases the time to process the loan applications.

What should the company do to reduce the processing time of loan applications?

- A. Configure Amazon Textract to route low-confidence predictions to Amazon SageMaker Ground Truth. Perform a manual review on those words before performing a business validation.
- B. Use an Amazon Textract synchronous operation instead of an asynchronous operation.
- C. Configure Amazon Textract to route low-confidence predictions to Amazon Augmented AI (Amazon A2I). Perform a manual review on those words before performing a business validation.
- D. Use Amazon Rekognition's feature to detect text in an image to extract the data from scanned images. Use this information to process the loan applications.

Correct Answer: C

  **joep21** Highly Voted 4 months, 3 weeks ago

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/>

upvoted 10 times

  **Juka3lj** Most Recent 2 months ago

correct is C

upvoted 1 times

Question #129

Topic 1

A company ingests machine learning (ML) data from web advertising clicks into an Amazon S3 data lake. Click data is added to an Amazon Kinesis data stream by using the Kinesis Producer Library (KPL). The data is loaded into the S3 data lake from the data stream by using an Amazon Kinesis Data Firehose delivery stream. As the data volume increases, an ML specialist notices that the rate of data ingested into Amazon S3 is relatively constant. There also is an increasing backlog of data for Kinesis Data Streams and Kinesis Data Firehose to ingest.

Which next step is MOST likely to improve the data ingestion rate into Amazon S3?

- A. Increase the number of S3 prefixes for the delivery stream to write to.
- B. Decrease the retention period for the data stream.
- C. Increase the number of shards for the data stream.
- D. Add more consumers using the Kinesis Client Library (KCL).

Correct Answer: C



SophieSu Highly Voted 4 months ago

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
- upvoted 7 times

Question #130

Topic 1

A data scientist must build a custom recommendation model in Amazon SageMaker for an online retail company. Due to the nature of the company's products, customers buy only 4-5 products every 5-10 years. So, the company relies on a steady stream of new customers. When a new customer signs up, the company collects data on the customer's preferences. Below is a sample of the data available to the data scientist.

timestamp	user_id	product_id	preference_1	...	preference_10
2020-03-04	90	25	0	...	0.374
2020-03-04	90	61	0	...	0.374
2020-02-21	203	56	1	...	0.098

How should the data scientist split the dataset into a training and test set for this use case?

- A. Shuffle all interaction data. Split off the last 10% of the interaction data for the test set.
- B. Identify the most recent 10% of interactions for each user. Split off these interactions for the test set.
- C. Identify the 10% of users with the least interaction data. Split off all interaction data from these users for the test set.
- D. Randomly select 10% of the users. Split off all interaction data from these users for the test set.

Correct Answer: D

 **joep21** Highly Voted 4 months, 3 weeks ago

I would select B, straight from this AWS example: <https://aws.amazon.com/blogs/machine-learning/building-a-customized-recommender-system-in-amazon-sagemaker/>

upvoted 12 times

 **Juka3lj** Most Recent 2 months ago

B is correct, takes all users into consideration

D is wrong because it takes only 10% of total number of users.

upvoted 1 times

Question #131

Topic 1

A financial services company wants to adopt Amazon SageMaker as its default data science environment. The company's data scientists run machine learning

(ML) models on confidential financial data. The company is worried about data egress and wants an ML engineer to secure the environment. Which mechanisms can the ML engineer use to control data egress from SageMaker? (Choose three.)


- A. Connect to SageMaker by using a VPC interface endpoint powered by AWS PrivateLink.
- B. Use SCPs to restrict access to SageMaker.
- C. Disable root access on the SageMaker notebook instances.
- D. Enable network isolation for training jobs and models.
- E. Restrict notebook presigned URLs to specific IPs used by the company.
- F. Protect data with encryption at rest and in transit. Use AWS Key Management Service (AWS KMS) to manage encryption keys.

Correct Answer: BDF

  **SophieSu**  4 months ago

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.

upvoted 9 times

  **scuzzy2010** 3 months, 3 weeks ago

I agree with ADF. SCP is to control access to a service, it's not related to securing data.

upvoted 3 times

  **cnethers**  4 months, 3 weeks ago

B. Use SCPs to restrict access to SageMaker. (policy to restrict data access)

D. Enable network isolation for training jobs and models. (protecting the data through isolation)

F. Protect data with encryption at rest and in transit. Use AWS Key Management Service (AWS KMS) to manage encryption keys. (Protecting the data through encryption)

The ask is to protect the data not who can access the data, subtle difference

upvoted 6 times

  **btsql**  3 weeks, 4 days ago

ADF is my answer. SCP is just secure copy.

upvoted 2 times

  **DSJingguo** 2 months ago

ADF

F: Protect data with encryption at rest and in transit

A: Internet Traffic Privacy

D: Infrastructure Security

<https://docs.aws.amazon.com/sagemaker/latest/dg/data-protection.html>

upvoted 3 times

  **DonRichy** 4 months ago

Answer is ADE, check below:



<https://aws.amazon.com/blogs/machine-learning/millennium-management-secure-machine-learning-using-amazon-sagemaker/>

upvoted 1 times

  **DonRichy** 4 months ago



Sorry , its BDE and not ADE

upvoted 2 times

  **astonm13** 4 months, 2 weeks ago

I think it is ADE

upvoted 4 times

  **jiadong** 4 months, 3 weeks ago

bde

<https://aws.amazon.com/blogs/machine-learning/millennium-management-secure-machine-learning-using-amazon-sagemaker/>

upvoted 3 times