← Back to the **Course**

Level: Advanced

AWS Certified Machine Learning Specialty

Practice Test III

Completed on **Sat, 02 Jul 2022**

1st	59/65	90.77%
Attempt	Marks Obtained	Your Score

1h 9m 31s	PASS
Time Taken	Result

Domain wise Quiz Performance Report

Join us on Slack community

No.	Domain	Total Question	Correct	Incorrect	Unattempte
1	Data Engineering	13	9	4	0
2	Exploratory Data Analysis	15	13	2	0
3	Modeling	24	24	0	0
4	ML Implementation and Operations	13	13	0	0
Total	**All Domains**	**65**	**59**	**6**	**0**

Review the Answers

Filter By All Questions

Question 1

Correct

Domain: Modeling

You are a Machine Learning Specialist on a team that is designing a system to help improve sales for your auto parts division. You have clickstream data gathered from your user's activity on your product website. Your team has been tasked with using the large amount of clickstream information depicting user behavior and product preferences to build a recommendation engine similar to the Amazon.com feature that recommends products through the tagline of "users who bought this item also considered these items." Similarly, your team's task is to predict which products a given user may like based on the similarity between the given user and other users.

How should you and your team architect this solution?

A. Create a recommendation engine based on a neural combinative filtering model using TensorFlow and run it on SageMaker.

B. Create a recommendation engine based on model-based filtering using TensorFlow and run it on SageMaker.

C. Create a recommendation engine based on a neural collaborative filtering model using TensorFlow and run it on SageMaker. right

D. Create a recommendation engine based on content-based filtering using TensorFlow and run it on SageMaker.

Explanation:

Correct Answer: C

Option A is incorrect. There is no neural combinative filtering method used in recommendation engine models.

Option B is incorrect. The term model-based filtering is too generic. We are using a model to make our recommendations, but which type of model should we use?

Option C is correct. The famous Amazon.com recommendation engine is built using a neural collaborative filtering method. This method is optimized to find similarities in environments where you have large amounts of user actions that you can analyze.

Option D is incorrect. 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.

References:

Please see the article titled **BUILDING A RECOMMENDATION ENGINE WITH SPARK ML ON AMAZON EMR USING ZEPPELIN** (https://noise.getoto.net/2015/11/14/building-a-recommendation-engine-with-spark-ml-on-amazon-emr-using-zeppelin-2/),

The Wikipedia article titled **Collaborative filtering** (https://en.wikipedia.org/wiki/Collaborative_filtering),

The AWS Machine Learning blog titled **Building a customized recommender system in Amazon SageMaker** (https://aws.amazon.com/blogs/machine-learning/building-a-customized-

recommender-system-in-amazon-sagemaker/)

Ask our Experts

View Queries

Question 2

Correct

Domain: Data Engineering

Your company, a financial services firm, has asked your team to build an analytics and machine learning platform to analyze and forecast your company's trading operations using Athena, S3, and SageMaker Studio. The volume of data received on a daily basis is very high. The data, stored in S3, will be used as feature data for your machine learning model that uses the XGBoost SageMaker built-in algorithm. The source systems that stream data into your environment send their data in JSON format in real-time. Your team needs to transform the data in real-time to prepare it for your machine learning model. Before storing it on S3 for use in your SageMaker XGBoost algorithm-based model, how can you transform the data to prepare it for training?

- A. Use Kinesis Data Streams to ingest the JSON data from the source systems, then send the data to Kinesis Data Firehose, where you can leverage a Lambda function to convert the JSON to libsvm and then use a Kinesis Data Firehose transform to write the data to S3. right
- B. Use Apache Spark Structured Streaming in an EMR cluster to ingest the JSON data from the source systems, then run Apache Spark steps to convert the JSON data into x-recordio-protobuf.
- C. Use Kinesis Data Streams to ingest the JSON data from the source systems, then use a Glue ETL job to convert data from JSON into x-recordio.
- D. Use Apache Kafka Streams running on EC2 instances to ingest the JSON data from the source systems, then use the Kafka Connect S3 connector to serialize the data onto S3 as x-recordio.

Explanation:

Correct Answer: A

Option A is correct. This option satisfies the real-time requirement while also being the most efficient and requiring the least amount of effort for your team. Also, the XGBoost algorithm only supports the libsvm and CSV content types for training and inference.

Option B is incorrect. This option can meet your real-time requirement, but it is far more complex to set up and maintain for your team than using the Kinesis Data Streams and Kinesis Data Firehose option. Also, the XGBoost algorithm only supports the libsvm and CSV content types, not the x-recordio-protobuf content type for training and inference.

Option C is incorrect. This option is incorrect because Glue ETL jobs imply batch processing, which fails to meet your real-time requirement. Also, the XGBoost algorithm only supports the libsvm and

CSV content types, not the x-recordio content type for training and inference.

Option D is incorrect. This option is also incorrect because it is far more complex to set up and maintain for your team than using the Kinesis Data Streams and Kinesis Data Firehose option. Also, the XGBoost algorithm only supports the libsvm and CSV content types, not the x-recordio content type for training and inference.

References:

Please see the AWS blog titled **Archiving Amazon MSK Data to Amazon S3 with the Lenses.io S3 Kafka Connect Connector** (https://aws.amazon.com/blogs/apn/archiving-amazon-msk-data-to-amazon-s3-with-the-lenses-io-s3-kafka-connect-connector/),

The Amazon SageMaker developer guide titled **Prepare ML Data with Amazon SageMaker Data Wrangler** (https://docs.aws.amazon.com/sagemaker/latest/dg/data-wrangler.html),

The Amazon Kinesis Data Firehose developer guide titled **Converting Your Input Record Format in Kinesis Data Firehose** (https://docs.aws.amazon.com/firehose/latest/dev/record-format-conversion.html),

The Amazon SageMaker developer guide titled **Common Data Formats for Training** (https://docs.aws.amazon.com/sagemaker/latest/dg/cdf-training.html),

The Amazon SageMaker developer guide titled **XGBoost Algorithm** (https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost.html)

Ask our Experts

View Queries

Question 3

Correct

Domain: Modeling

You are a machine learning specialist working for a state government water safety department. The state needs to monitor water quality across all of its counties to ensure water contamination levels remain within acceptable thresholds. Your machine learning team is responsible for producing a forecasting report of water contaminants in parts per million for the next month, every month, across the state. Your team has daily data from the last year available as a starting point for your model.

Which SageMaker model will give you the best results for your monthly forecasting report?

A. Use multiple time series of the full previous year of data as your input to a SageMaker Linear Learner built-in algorithm with a predictor_type of the classifier.

B. Use a single time series of the full previous year of data as your input to a SageMaker Linear Learner built-in algorithm with a predictor_type of the regressor. right

C. Use a single time series of the full previous year of data as your input to a SageMaker Random Cut Forest (RCF) built-in algorithm.

D. Use multiple time series of the full previous year of data as your input to a SageMaker k-Nearest-Neighbors (kNN) built-in algorithm with a predictor_type of the classifier.

Explanation:

Correct Answer: B

Option A is incorrect. This problem requires a regression algorithm since we solve a real number (continuous) prediction value or label: water contaminants in parts per million. We are not producing a classification of unacceptable versus acceptable. Also, we should use a single time series for this type of regression, not multiple time series.

Option B is correct. This option is correct since the problem requires a regression algorithm because we solve a real number (continuous) prediction value or label: water contaminants in parts per million. The SageMaker Linear Learner algorithm is one of the go-to algorithms for regression on the Machine Learning exam.

Option C is incorrect. This option is incorrect because the Random Cut Forest algorithm is primarily used as an unsupervised algorithm for detecting anomalous data points within a data set. You would not use an RCF algorithm to solve a regression problem.

Option D is incorrect. This option is incorrect because while you can use the kNN algorithm for regression problems, this option states the use of a predictor_type of the classifier. This problem requires a regression algorithm. So you would need to use a predictor_type of the regressor. Also, we should use a single time series for this type of regression, not multiple time series.

References:

Please see the Amazon SageMaker developer guide titled **K-Nearest Neighbors (k-NN) Algorithm** (https://docs.aws.amazon.com/sagemaker/latest/dg/k-nearest-neighbors.html),

The Amazon SageMaker developer guide titled **Linear Learner Algorithm** (https://docs.aws.amazon.com/sagemaker/latest/dg/linear-learner.html),

The Amazon SageMaker developer guide titled **Random Cut Forest (RCF) Algorithm** (https://docs.aws.amazon.com/sagemaker/latest/dg/randomcutforest.html)

Ask our Experts

View Queries

Question 4

Correct

Domain: ML Implementation and Operations

You work as a machine learning specialist for a banking firm working in the credit card processing division. Your team builds a credit limit authorization model that needs to use a dataset containing personally identifiable information (PII), such as customer credit card information. How will your team ensure the PII data remains encrypted and the credit card information is not compromised?

A. Encrypt the data on S3 and SageMaker using KMS, obfuscate the credit card information from the customer data with a Glue ETL job. right

B. Encrypt the data using a SageMaker lifecycle configuration once the data is copied to the SageMaker instance in a VPC. Use the principal component analysis (PCA) algorithm to reduce the dimensionality of the credit card numbers.

C. Encrypt the data on the S3 bucket and Kinesis using an IAM policy, remove the customer credit card numbers and insert fabricated credit card numbers using a Glue ETL job.

D. Encrypt the data using a custom-coded encryption algorithm and store the data on a SageMaker instance in a VPC. Fabricate new credit card numbers using the SageMaker DeepAR built-in algorithm, replacing the customer credit card numbers.

Explanation:

Correct Answer: A

Option A is correct. Using KMS to encrypt the data is the best choice of the options given. KMS is a managed service and encrypts your data using the same key for S3 and SageMaker. Also, using a Glue ETL job, you can remove the credit card information from the dataset by dropping that column from the dataset in your ETL job.

Option B is incorrect. A lifecycle configuration is a script that runs when your notebook instance is created. Trying to use a lifecycle configuration to encrypt data is not an option that would work. Also, the Principal Component Analysis algorithm reduces the dimensionality of a dataset. One would not use PCA to obfuscate PII data.

Option C is incorrect. You can't encrypt data with an IAM policy alone. You need to combine your IAM policies with a service like KMS.

Option D is incorrect. This option is incorrect because writing your own encryption algorithm is counterproductive when you have a managed service in AWS, KMS, that will encrypt your data to the highest industry standards. Also, using a DeepAR algorithm to randomize the PII data further complicates this option.

References:

Please see the AWS Glue developer guide titled **Built-In Transforms** (https://docs.aws.amazon.com/glue/latest/dg/built-in-transforms.html),

The Amazon SageMaker developer guide titled **Principal Component Analysis (PCA) Algorithm** (https://docs.aws.amazon.com/sagemaker/latest/dg/pca.html),

The Amazon SageMaker developer guide titled **Customize a Notebook Instance Using a Lifecycle Configuration Script** (https://docs.aws.amazon.com/sagemaker/latest/dg/notebook-lifecycle-config.html),

The AWS Identity and Access Management user guide titled **Data protection in AWS Identity and Access Management** (https://docs.aws.amazon.com/IAM/latest/UserGuide/data-protection.html)

Ask our Experts

Ask our Experts

View Queries

Question 5

Correct

Domain: ML Implementation and Operations

You are a machine learning specialist at a rideshare startup company. Your team supports developers who are using SageMaker notebook instances in a private subnet of your corporate VPC. Your developers have important customer data stored on the SageMaker notebook instance's EBS volume, and so they wish to take a snapshot of that EBS volume. When your developers attempt to locate the SageMaker notebook EC2 instance and its EBS volume within your corporate VPC, they don't find them. What is the reason they don't see the instance or its EBS volume in the corporate VPC?

A. SageMaker notebook instances run on EKS containers running within an AWS service account.

B. SageMaker notebook instances run on EC2 instances running within an AWS service account. right

C. SageMaker notebook instances run on the container in the ECS service within your corporate account.

D. SageMaker notebook instances run on the EC2 instances within your corporate account, but they run outside of a VPC.

Explanation:

Correct Answer: B

Option A is incorrect. SageMaker notebook instances run in an AWS service account (not in customer accounts), but they run on EC2 instances, not in EKS containers.

Option B is correct. SageMaker runs its notebook instances on EC2 in an AWS service account, not in customer accounts. Therefore, you can't access the EC2 instance EBS volumes.

Option C is incorrect. SageMaker notebook instances run in an AWS service account (not in customer accounts), and they run on EC2 instances, not in ECS containers.

Option D is incorrect. SageMaker runs its notebook instances on EC2 in an AWS service account, not in customer accounts.

References:

Please see the Amazon SageMaker developer guide titled **Connect SageMaker Studio Notebooks to Resources in a VPC** (https://docs.aws.amazon.com/sagemaker/latest/dg/studio-notebooks-and-internet-access.html),

The **Amazon EC2 Auto Scaling FAQs page** (https://aws.amazon.com/ec2/autoscaling/faqs/),

The AWS Machine Learning blog titled **Understanding Amazon SageMaker notebook instance networking configurations and advanced routing options** (https://aws.amazon.com/blogs/machine-learning/understanding-amazon-sagemaker-notebook-instance-networking-configurations-and-advanced-routing-options/)

Ask our Experts

View Queries

Question 6

Correct

Domain: Modeling

You are part of a machine learning team in a financial services company that builds a model that will perform time series forecasting of stock price movement using SageMaker. You and your team have finished training the model, and you are now ready to performance test your endpoint to get the best parameters for configuring auto-scaling for your model variant. How can you most efficiently review the latency, memory utilization, and CPU utilization during the load test?

A. Stream the SageMaker model variant CloudWatch logs to ElasticSearch. Then visualize and query the log data in a Kibana dashboard.

B. Create custom CloudWatch logs containing the metrics you wish to monitor, then stream the SageMaker model variant logs to ElasticSearch and visualize/query the log data in a Kibana dashboard.

C. Create a CloudWatch dashboard to show a view of the latency, memory utilization, and CPU utilization metrics of the SageMaker model variant. right

D. Query the SageMaker model variant logs on S3 using Athena and leverage QuickSight to visualize the logs.

Explanation:

Correct Answer: C

Option A is incorrect. Using ElasticSearch and Kibana unnecessarily complicates the solution. A CloudWatch dashboard can show all of the metric data you need to evaluate your model variant.

Option B is incorrect. You don't need to create custom CloudWatch logs with the metrics you wish to monitor. All of the metrics (latency, memory utilization, and CPU utilization) you wish to view are generated by CloudWatch by default. Also, using ElasticSearch and Kibana unnecessarily complicates the solution.

Option C is correct. The simplest approach is to leverage the CloudWatch dashboard feature since it generates all of the metrics (latency, memory utilization, and CPU utilization) you wish to view by default.

Option D is incorrect. Using Athena and QuickSight unnecessarily complicates the solution. A CloudWatch dashboard can show all of the metric data you need to evaluate your model variant.

References:

Please see the Amazon SageMaker developer guide titled **Monitor Amazon SageMaker with Amazon CloudWatch** (https://docs.aws.amazon.com/sagemaker/latest/dg/monitoring-cloudwatch.html),

The Amazon CloudWatch user guide titled **Using Amazon CloudWatch Dashboards** (https://docs.aws.amazon.com/AmazonCloudWatch/latest/monitoring/CloudWatch_Dashboards.html),

The Amazon CloudWatch user guide titled **Creating a CloudWatch Dashboard** (https://docs.aws.amazon.com/AmazonCloudWatch/latest/monitoring/create_dashboard.html)

Ask our Experts

View Queries

Question 7

Correct

Domain: Exploratory Data Analysis

You work on a machine learning team at a manufacturing company that produces fire detection products. You are building a fire detection analytics model, the source data store of which has structured and unstructured data stored in an S3 bucket. You are in the data engineering and data analysis phase of the machine learning lifecycle. At this point, you need to use SQL to run queries on your source data to determine feature correlation and dimensionality. Which option allows you to query the data with the least amount of effort?

A. Run queries using Kinesis Data Analytics after using a Lambda function to transform the data.

B. Use the AWS Batch service to extract, transform, and load (ETL) the data and an ElasticSearch cluster to run the queries.

C. Use a Glue crawler to catalog the data and Athena to query the data on S3. right

D. Query the data in RDS after using Data Pipeline to transform the data and load it into RDS.

Explanation:

Correct Answer: C

Option A is incorrect. Transforming the data using a Lambda function and then querying the data in Kinesis Data Analytics requires you to write a Lambda function. This is more effort than crawling the data using a Glue crawler then using Athena to query the data directly on S3.

Option B is incorrect. Using AWS Batch to perform ETL on the data and then loading the data into an ElasticSearch cluster requires more effort than crawling the data using a Glue crawler, then using Athena to query the data directly on S3.

Option C is correct. Crawling the data using a Glue crawler and then querying the data directly on S3 using Athena only requires you to write your SQL code. You don't need to write a Lambda function or create an Aurora or RDS database.

Option D is incorrect. With this option, you have to create an RDS database and then load your data into the RDS instance using Data Pipeline. This requires more effort than crawling the data using a Glue crawler then using Athena to query the data directly on S3.

References:

Please see the AWS Data Pipeline developer guide titled **What is AWS Data Pipeline?** (https://docs.aws.amazon.com/datapipeline/latest/DeveloperGuide/what-is-datapipeline.html),

The AWS Glue developer guide titled **Defining Crawlers** (https://docs.aws.amazon.com/glue/latest/dg/add-crawler.html),

The AWS Batch user guide titled **What Is AWS Batch?** (https://docs.aws.amazon.com/batch/latest/userguide/what-is-batch.html),

The Amazon Kinesis Data Analytics for SQL Applications Developer Guide SQL developer guide titled **What Is Amazon Kinesis Data Analytics for SQL Applications?** (https://docs.aws.amazon.com/kinesisanalytics/latest/dev/what-is.html)

Ask our Experts

View Queries

Question 8

Correct

Domain: Exploratory Data Analysis

You work as a machine learning specialist for an audio processing and distribution company. You are currently working on a custom audio recommendation model for a criminal investigation application that recommends which audio file to use based on investigation details. The dataset you are attempting to use to train the model is extremely large, containing millions of data points. You are storing the dataset in an S3 bucket. You need to find an alternative to loading all of the data into a SageMaker notebook instance because it would take too long to load and exceed the 50 GB EBS volume attached to the notebook instance. Which approach should you select so that you can load all the data to train the model?

A. Split the training dataset using scikit-learn or pandas to create a subset of your training data. Load the subset of the training data into the SageMaker notebook and train the model in your notebook instance. Verify that the model trained accurately and that the model parameters

produce reasonable results. Use a Deep Learning AMI to start an EC2 instance and attach the S3 bucket to train the full dataset.

B. Split the training dataset using scikit-learn or pandas to create a subset of your training data. Use Glue to load your data into your SageMaker notebook, using your subset of the training data to verify that the model trained accurately and that the model parameters produce reasonable results. Next, run a training job using the entire dataset from the S3 bucket using Pipe input mode.

C. Use a Deep Learning AMI to start an EC2 instance and attach the S3 bucket. Split the training dataset using scikit-learn or pandas to create a subset of your training data. Train using the subset of the training data to verify the training code and hyperparameters. Use SageMaker to train using the full dataset.

○ D. Split the training dataset using scikit-learn or pandas to create a subset of your training data. Load the subset of the training data into the SageMaker notebook and train in your notebook. Verify that the model trained accurately and that the model parameters produce reasonable results. Run a SageMaker training job loading the complete dataset from the S3 bucket using Pipe input mode. right

Explanation:

Correct Answer: D

Option A is incorrect. Using a Deep Learning AMI for your EC2 instance will not help with loading the extremely large training dataset from S3.

Option B is incorrect. Glue cannot be used to load data into your SageMaker notebook to train a machine learning model. Glue could be used to place your data onto an S3 bucket that you could then use to load the data into your SageMaker notebook.

Option C is incorrect. This approach does not address the loading of the extremely large dataset onto the local storage of the EC2 instance.

Option D is correct. With this option, we are using the pipe input mode to stream our training data directly to our training instance instead of downloading it to our EBS storage first. This solves the problem of loading extremely large data into our notebook instance.

References:

Please see the AWS Machine Learning blog titled **Using Pipe input mode for Amazon SageMaker algorithms** (https://aws.amazon.com/blogs/machine-learning/using-pipe-input-mode-for-amazon-sagemaker-algorithms/)

The AWS Machine Learning blog titled **Accelerate model training using faster Pipe mode on Amazon SageMaker** (https://aws.amazon.com/blogs/machine-learning/accelerate-model-training-using-faster-pipe-mode-on-amazon-sagemaker/),

The AWS announcement titled **Amazon SageMaker Now Supports an Improved Pipe Mode Implementation** (https://aws.amazon.com/about-aws/whats-new/2018/10/amazon-sagemaker-now-supports-an-improved-pipe-mode-implementati/),

The Amazon Amazon SageMaker developer guide titled **Use Scikit-learn with Amazon SageMaker** (https://docs.aws.amazon.com/sagemaker/latest/dg/sklearn.html)

Ask our Experts

View Queries

Question 9

Correct

Domain: Data Engineering

You work as a machine learning specialist in a financial services company in their asset management division. You have completed a pilot of a Random Cut Forest algorithm-based model that you plan to use to find anomalies in your trading data. You have tested your model using a small data sample and are ready to implement your production model using SageMaker. The historical trading data that you will use for training is stored in an RDS Microsoft SQL Server database. How should you prepare your historical trading data to train your production model?

A. Use the DMS service to move the data to ElastiCache, then connect your SageMaker notebook with ElastiCache to load the data at low latency.

B. Use a Lambda function to load the data to DynamoDB tables, then connect your SageMaker notebook to DynamoDB to load your trading data.

C. Use the Data Pipeline service to move your trading data from the Microsoft SQL Server database to S3, then use the S3 bucket within your SageMaker notebook to load your trading data. right

D. Directly connect to the SQL database using Direct Connect from your SageMaker notebook and load the trading data.

Explanation:

Correct Answer: C

Option A is incorrect. You cannot load data directly from ElastiCache into a SageMaker notebook.

Option B is incorrect. You cannot load data into a SageMaker notebook directly from DynamoDB without first staging the data in S3.

Option C is correct. Loading your training data from S3 is the best approach for getting your data into your SageMaker notebook. Also, the Data Pipeline service is the preferred method of loading data from Microsoft SQL Server to S3.

Option D is incorrect. You cannot directly load data from an RDS instance into a SageMaker notebook instance without first staging the data in S3.

References:

Please see the AWS Re:Invent 2018 presentation titled **Train Models on Amazon SageMaker Using Data Not from Amazon S3 (AIM419) - AWS re:Invent 2018** (https://www.slideshare.net/AmazonWebServices/train-models-on-amazon-sagemaker-using-data-not-from-amazon-s3-aim419-aws-reinvent-2018),

The AWS SageMaker developer guide titled **Random Cut Forest (RCF) Algorithm** (https://docs.aws.amazon.com/sagemaker/latest/dg/randomcutforest.html),

The AWS Data Pipeline developer guide titled **What is AWS Data Pipeline?** (https://docs.aws.amazon.com/datapipeline/latest/DeveloperGuide/what-is-datapipeline.html),

The Amazon Amazon SageMaker developer guide titled **Download, Prepare, and Upload Training Data** (https://docs.aws.amazon.com/sagemaker/latest/dg/automatic-model-tuning-ex-data.html)

Ask our Experts

View Queries

Question 10

Correct

Domain: Modeling

You are a machine learning specialist for a retail clothing company. Your company receives a significant amount of its revenue from your retail website. Your marketing team wishes to implement a recommendation feature for your customers that uses online and in-store shopping patterns, user preferences, and overall fashion trends to give suggestions to your customers for items to purchase. Your dataset contains customer data such as demographics, prior visits, prior purchases, and location. Your task is to develop a machine learning model that uses the customer data to enhance the user's experience with your website while also giving informed recommendations. Which option best suits your task?

A. Build a model that uses the Random Cut Forest (RCF) algorithm with a training dataset of customer data to identify patterns in the customer data.

B. Build a model based on collaborative filtering that leverages implicit feedback derived from user activity, such as clicks and views to identify patterns in the customer data. right

C. Build a recurrent neural network (RNN) using the DeepAR algorithm with an initial weight decay coefficient that adds L2 regularization and a minimum of five layers to identify patterns in the customer data.

D. Build a model using the Neural Topic Model (NTM) algorithm using the Stochastic Gradient Descent (SGD) optimizer to identify patterns in the customer data.

Explanation:

Correct Answer: B

Option A is incorrect. The Random Cut Forest (RCF) algorithm is better suited to anomaly detection in your data. It is not a choice one would use for a recommendation engine.

Option B is correct. Collaborative filtering is the preferred approach for building an online recommendation engine that leverages customer behavior data.

Option C is incorrect. A Recurrent Neural Network (RNN) is best suited to forecasting scalar (one-dimensional) time series. You don't have time-series data. You have clickstream data from your website and customer purchasing patterns from in-store and online purchases.

Option D is incorrect. The Neural Topic Model algorithm is best suited for organizing a corpus of documents into topics that contain word groupings based on their statistical distribution. This type of document processing algorithm would not work as a recommendation engine for an online business.

References:

Please see the AWS SageMaker developer guide titled **Use Amazon SageMaker Built-in Algorithms** (https://docs.aws.amazon.com/sagemaker/latest/dg/algos.html),

The AWS Media blog titled **What's new in recommender systems** (https://aws.amazon.com/blogs/media/whats-new-in-recommender-systems/),

The AWS Machine Learning blog titled **Building a customized recommender system in Amazon SageMaker** (https://aws.amazon.com/blogs/machine-learning/building-a-customized-recommender-system-in-amazon-sagemaker/),

The AWS Database blog titled **Using collaborative filtering on Yelp data to build a recommendation system in Amazon Neptune** (https://aws.amazon.com/blogs/database/using-collaborative-filtering-on-yelp-data-to-build-a-recommendation-system-in-amazon-neptune/)

Ask our Experts

View Queries

Question 11

Correct

Domain: Modeling

You are a machine learning specialist at a music subscription company. Your company needs to understand its customer churn rate better. They plan to leverage this information to spend their marketing budget on attempting to retain customers that are likely to leave the service in the near future. Your task is to group your customer subscribers into categories based on which subscribers may or may not cancel their subscription in the near future (3 months). You have already performed data engineering and have subscriber data that is labeled. Which type of model is best suited to your task?

A. Linear regression

B. Clustering

C. Classification right

D. Unsupervised learning

Explanation:

Correct Answer: C

Option A is incorrect. The key to this scenario is that you have labeled data, and you are trying to group your customers into categories. The linear regression algorithm type is better suited to predicting a numeric/continuous value, such as estimating the value of a house.

Option B is incorrect. Clustering is best suited to grouping similar objects when you have unlabeled data. You have labeled your data.

Option C is correct. The classification algorithm type is best suited to grouping similar objects when you have labeled data. This is the case described in the scenario.

Option D is incorrect. The unsupervised learning algorithm type is best suited to clustering, dimension reduction, pattern recognition, and anomaly detection. These types of algorithms are used when you have unlabeled data. You have labeled your data.

References:

Please see the Amazon SageMaker developer guide titled **Use Amazon SageMaker Built-in Algorithms** (https://docs.aws.amazon.com/sagemaker/latest/dg/algos.html),

The Amazon SageMaker developer guide titled **Unsupervised Learning** (https://docs.aws.amazon.com/sagemaker/latest/dg/algos.html#algorithms-built-in-unsupervised-learning)

Ask our Experts

View Queries

Question 12

Correct

Domain: ML Implementation and Operations

You are a machine learning specialist at a government agency that processes citizen applications (online, mail, and in-person) for government documents such as driver's licenses and passports. Your machine learning team is responsible for using machine learning technology to determine fraudulent activity in the document application processes. You are preparing a subset of your agency data for model training. In order to use your data in your SageMaker notebook, you have stored your data in S3. By definition, your data contains Personally Identifiable Information (PII). In order to maintain the

required level of security, your data must be accessible only from within your VPC and cannot traverse the public internet. Which option best meets your security requirements?

A. Use a VPC endpoint and leverage a security group to restrict access to the VPC endpoint.

B. Use a VPC endpoint and leverage a Network Access Control List (NACL) to only allow traffic between the VPC endpoint and S3.

C. Use a VPC endpoint and leverage a bucket access policy to allow access to the S3 bucket from the VPC endpoint.

D. Use a VPC endpoint and leverage a bucket access policy to deny access to the S3 bucket from resources other than the VPC endpoint and the VPC. right

Explanation:

Correct Answer: D

Option A is incorrect. This option doesn't meet your requirements because it doesn't address access to the S3 bucket or the restriction of not traversing the internet.

Option B is incorrect. This option also doesn't meet your requirements because it doesn't address the restriction of not traversing the internet.

Option C is incorrect. We need to restrict access using a deny statement if our source (either the VPC endpoint or the VPC) is not equal to our VPC Endpoint or VPC. So we need to use a deny statement in our policy, not an allow statement. This option describes using an allow statement.

Option D is correct. You can control which VPCs or VPC endpoints have access to your buckets by using S3 bucket policies with deny statements. Also, when you use a VPC Endpoint, your traffic doesn't traverse the internet.

References:

Please see the Amazon Virtual Private Cloud AWS Privatelink documentation titled **Endpoints for Amazon S3** (https://docs.aws.amazon.com/vpc/latest/privatelink/vpc-endpoints-s3.html#vpc-endpoints-s3-bucket-policies),

The Amazon Simple Storage Service user guide titled **Controlling access from VPC endpoints with bucket policies** (https://docs.aws.amazon.com/AmazonS3/latest/userguide/example-bucket-policies-vpc-endpoint.html)

Ask our Experts

View Queries

Question 13 Correct

Domain: Data Engineering

You work as a machine learning specialist at a retail clothing chain. Your team builds a model that uses Kinesis Data Firehose to ingest transaction records from the chain's many (50,000) stores throughout the country. You are building your training data store using your streaming transactions received from Kinesis Data Firehose. The transaction records used for training require simple transformations, and you need to combine some attributes and drop other attributes. Also, you need to retrain the model daily. Which option will meet your requirements with the least effort?

- A. Have the Kinesis Data Firehose stream your transaction records to Kinesis Data Analytics, where you transform the transaction records and combine/drop attributes using Apache Flink and store the transformed records to S3. right
- B. Have the Kinesis Data Firehose stream your transaction records to S3. Launch an ECS cluster that runs the transformation logic as tasks that transform and combine/drop attributes on the data records on Amazon S3
- C. Have the Kinesis Data Firehose stream your transaction records to S3. Run an EMR cluster with Apache Hadoop and Apache Presto performing the transformation logic. Run the cluster every day to transform and combine attributes of the records on S3.
- D. Use Kinesis Data Streams instead of Kinesis Data Firehose to stream the transaction records to S3. Then transform and combine attributes using a Glue ETL job and store the results on S3.

Explanation:

Correct Answer: A

Option A is correct. With this option, you can transform and combine attributes on your records in Kinesis Data Analytics using Apache Flink using Flink's built-in operators. Using file sink integrations, Kinesis Data Analytics can store the transformed records to S3 for use in your model training. This option requires the least amount of effort on your part.

Option B is incorrect. This option requires the effort of writing transformation code to run on your ECS containers as well as administration effort to launch and maintain the ECS containers and tasks.

Option C is incorrect. This option requires significantly more effort due to the launching and maintaining of the EMR cluster. You would also have to create the Apache Hadoop map/reduce logic and Apache Presto scripts to do the transformations.

Option D is incorrect. Using Kinesis Data Streams instead of Kinesis Data Firehose would require you to write a Kinesis Client Library application and run it on EC2 instances, which you would have to launch and maintain to transform your data and then write the transformed data to S3. This would be significantly more effort than writing Apache Flink using Flink's built-in operators.

References:

Please see the **Amazon Kinesis Data Analytics FAQs** (refer to the question "What integrations are supported in a Kinesis Data Analytics for Apache Flink application?") (https://aws.amazon.com/kinesis/data-analytics/faqs/),

The Amazon Kinesis Data Analytics developer guide titled **Example: Writing to an Amazon S3 Bucket** (https://docs.aws.amazon.com/kinesisanalytics/latest/java/examples-s3.html),

The Amazon EMR page titled **Apache Hadoop on Amazon EMR** (https://aws.amazon.com/emr/features/hadoop/),

The Amazon EMR management guide titled **What is EMR?** (https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-what-is-emr.html)

Ask our Experts

View Queries

Question 14

Correct

Domain: ML Implementation and Operations

You work as a machine learning specialist for a financial services firm. Your firm contracts with market data generation services that deliver 5 TB of market activity record data every minute. To prepare this data for your machine learning models, your team queries the data using Athena. However, the queries perform poorly because they are operating on such a large data stream. You need to find a more performant option. Which file format for your market data records on S3 will give you the best performance?

A. TSV files

B. Compressed LZO files

C. Parquet files right

D. CSV files

Explanation:

Correct Answer: C

Option A is incorrect. The TSV file format uses a row-based file structure that uses tabs as an attribute separator. When Athena reads from these types of files, it must read the entire row for every row versus reading in a column when only the attribute in that column is needed for your query. Columnar-based file processing is much more efficient for queries of large datasets. Also, the TSV file format does not support the partitioning of your data.

Option B is incorrect. Compressed LZO Files do not support columnar processing nor partitioning. Therefore they will perform poorly when compared to columnar file formats like Parquet.

Option C is correct. The Parquet file format is a columnar-based format, and it supports partitioning. The other columnar-based file format supported by Athena is ORC. These columnar-based file formats outperform the tabular formats such as CSV and TSV when Athena works with very large datasets.

Option D is incorrect. The CSV file format uses a row-based file structure that uses commas as an attribute separator. When Athena reads from these types of files, it must read the entire row for every row versus reading in a column (columnar-based processing) when only the attribute in that column is needed for your query. Columnar-based file processing is much more efficient for queries of large datasets. Also, the CSV file format does not support the partitioning of your data.

References:

Please see the **Amazon Athena FAQs** (refer to the question "How do I improve the performance of my query?") (https://aws.amazon.com/athena/faqs/#:~:text=Amazon%20Athena%20supports%20a%20wide,%2C%20LZO%2C%20and%20GZIP%20formats.),

The AWS Big Data blog titled **Top 10 Performance Tuning Tips for Amazon Athena** (https://aws.amazon.com/blogs/big-data/top-10-performance-tuning-tips-for-amazon-athena/),

The Amazon Athena user guide titled **Compression Formats** (https://docs.aws.amazon.com/athena/latest/ug/compression-formats.html)

Ask our Experts

View Queries

Question 15

Correct

Domain: Exploratory Data Analysis

You are a machine learning specialist working for a large insurance company. You are building a machine learning model to predict the likelihood of an insured customer committing insurance fraud. Your training dataset has many attributes about the insured, the insurance policy, and their insurance claims. As its prediction, your model needs to produce a continuous value of the probability of fraud for any given customer claim. The feature set of your training data includes labeled outcomes for a set of 100,000 insurance claim observations. When you visualize the training dataset, you see that out of the 100,000 insurance claims, 24,350 claim records show the policy term length of 0 years. The remaining features for these observations show no anomalies. Which feature engineering option will give you the best dataset for your model training?

A. Use k-means clustering to impute the missing policy length features.

B. Use KNN to impute the missing policy length features. right

C. Populate the 0 policy length feature value with the mean or median value of the feature.

D. Drop the records from the dataset where policy length is 0.

Explanation:

Correct Answer: B

Option A is incorrect. The k-means algorithm is an unsupervised learning algorithm where we do not have labeled data. The k-means algorithm is used for clustering. This is not the best choice, nor is it a choice used by practicing machine learning specialists for feature imputation. Unsupervised learning using unlabeled data will give inferior results when compared to supervised learning with labeled data.

Option B is correct. The K Nearest Neighbor algorithm, when used for classification, is a supervised learning algorithm where we have labeled data. Using KNN, you can impute missing values using feature similarity to predict missing values based on the other non-missing values in the feature. This is a very common approach used by machine learning specialists to impute missing values.

Option C is incorrect. While it is common to replace missing feature values with the simple mean or median of the feature, this method is far less accurate than using the KNN approach to impute your missing values.

Option D is incorrect. Dropping the records with the missing values is another common approach for dealing with missing feature values. However, this approach reduces your feature set significantly in this scenario. You have missing features in approximately 24% of your training data. Dropping that many records will reduce the accuracy of your predictions.

References:

Please see the Amazon SageMaker developer guide titled **K-Nearest Neighbors (k-NN) Algorithm** (https://docs.aws.amazon.com/sagemaker/latest/dg/k-nearest-neighbors.html),

The Amazon SageMaker developer guide titled **K-Means Algorithm** (https://docs.aws.amazon.com/sagemaker/latest/dg/k-means.html),

The Towards Data Science article titled **6 Different Ways to Compensate for Missing Values In a Dataset (Data Imputation with examples)** (https://towardsdatascience.com/6-different-ways-to-compensate-for-missing-values-data-imputation-with-examples-6022d9ca0779)

Ask our Experts

View Queries

Question 16

Correct

Domain: Data Engineering

You work for a global consulting company as a machine learning specialist. You work with a team of data scientists that continually create datasets for your consultancy's analysis and trend prediction work using machine learning. You have been assigned the job of creating a data repository to store the large amount of training data generated by your data scientists for use in your machine learning models. You do not know how many new datasets your data scientists will create each day, so your solution must scale automatically, and your management team wants the storage solution to be cost-effective. Also, the data scientists and machine learning specialists must be able to query the data using SQL. Which option is the best solution to meet your requirements?

A. Have your data scientists store their new datasets in DynamoDB using global tables.

B. Have your data scientists store their new datasets as tables in a Redshift cluster using RA3 nodes with managed storage and Redshift Spectrum.

C. Have your data scientists store their new datasets as files in an EFS attached to EC2 instances instance.

D. Have your data scientists store their new datasets as files in S3. right

Explanation:

Correct Answer: D

Option A is incorrect. DynamoDB will not be the most cost-effective option when compared to S3. Also, there is no requirement suggesting the need for global data distribution. Also, when you need to use the data in your machine learning models, you will have to extract the data from DynamoDB and put it to S3.

Option B is incorrect. Redshift using RA3 node types with managed storage will give you very fast query access to your data, but it is not cost-effective when compared to S3. Also, when you need to use the data in your machine learning models, you can leverage Redshift Spectrum to extract the data from Redshift and put it to S3. But this adds another level of complexity and cost.

Option C is incorrect. Using EC2 instances and EFS volumes to store your data is very cost-ineffective when compared to using S3. Also, when you need to use the data in your machine learning models, you will have to move the data from the EC2 EFS volumes to S3.

Option D is correct. S3 is cost-effective, scales infinitely to any dataset volume and size, and it is where you need to have your data when you use it in your machine learning models.

References:

Please see the Amazon Machine Learning blog titled **Building secure machine learning environments with Amazon SageMaker** (https://aws.amazon.com/blogs/machine-learning/building-secure-machine-learning-environments-with-amazon-sagemaker/),

The Amazon Redshift products page titled **Amazon Redshift Pricing** (https://aws.amazon.com/redshift/pricing/),

The Amazon DynamoDB products page titled **Amazon DynamoDB pricing** (https://aws.amazon.com/dynamodb/pricing/)

Ask our Experts

View Queries

Question 17 Correct

Domain: Modeling

You are a machine learning specialist working for a social media software company. You have built and deployed a product recommendation model that recommends client products via social media posts in your company's social media app. When you first deployed the model, it generated great results with users clicking through and buying client products, thereby generating revenue for your social media company. Over time the product recommendations results have started to decline, and your users are clicking through to client product pages less. You had not changed your model from when you did your initial deployment. What is the best and most efficient option to use to improve your user click-through rate for your social media app over time?

- **A. Periodically retrain your model using your foundational training data from your initial deployment adding new data from new client products.** right

- **B. Periodically retrain your model from scratch using your foundational training data from your initial deployment, adding an L1 or L2 regularization value set to the higher range of the parameter to represent client product changes.**

- **C. Periodically update your model hyperparameters, setting the drift threshold to the higher range of the hyperparameter, to prevent model drift.**

- **D. Completely recreate your model as it no longer recognizes client product changes.**

Explanation:

Correct Answer: A

Option A is correct. Retraining your model with your initial training data plus new data representing new client products will keep the model in line with the product recommendation domain. This will allow your model to recognize new products while retaining the knowledge of the foundational product data.

Option B is incorrect. A regularization term will help prevent your model from overfitting, but it will not give your model the data it needs to recognize new product data.

Option C is incorrect. Model drift happens when you receive new data on which to train and your model. Changing your hyperparameters without retraining your model will not address your problem.

Option D is incorrect. Completely recreating your model is not necessary and definitely not the most efficient way to address your model performance problem. You will get better results more efficiently by adding new training data to your foundational training dataset and retraining your model.

References:

Please see the Amazon Machine Learning developer guide titled **Training Parameters** (https://docs.aws.amazon.com/machine-learning/latest/dg/training-parameters1.html),

The AWS News blog titled **Amazon SageMaker Model Monitor – Fully Managed Automatic Monitoring For Your Machine Learning Models** (https://aws.amazon.com/blogs/aws/amazon-sagemaker-model-monitor-fully-managed-automatic-monitoring-for-your-machine-learning-models/),

The Amazon SageMaker developer guide titled **How Hyperparameter Tuning Works** (https://docs.aws.amazon.com/sagemaker/latest/dg/automatic-model-tuning-how-it-works.html)

Ask our Experts

View Queries

Question 18

Correct

Domain: Data Engineering

You are a machine learning specialist working for an online retail shopping site. Your machine learning team is responsible for building out a machine learning environment using SageMaker Studio to make possible the running of models used to predict online sales and product pipeline optimization. Your team also needs to optimize the data ingestion solution into your data lake that is the primary source for your machine learning models. Your ingestion solution will also facilitate analytics (real-time and interactive analytics of historical data), clickstream analysis, as well as product recommendations. Which option best meets your team's requirements?

A. Use Athena as the data catalog of your data lake files, use Kinesis Data Streams and Kinesis Data Analytics for historical data insights, use Glue for clickstream analytics, and create personalized product recommendations.

B. Use Glue as the data catalog of your data lake files, use Kinesis Data Streams and Kinesis Data Analytics for historical data insights, use Kinesis Data Firehose to deliver your data to ElasticSearch for clickstream analytics, and leverage Kibana dashboards to create personalized product recommendations.

C. Use Athena as the data catalog of your data lake files, use Kinesis Data Streams and Kinesis Data Analytics to generate near-real-time data insights, use Kinesis Data Firehose for clickstream analytics, and use Glue to create personalized product recommendations.

D. Use Glue as the data catalog of your data lake files, use Kinesis Data Streams and Amazon Kinesis Data Analytics for real-time data insights, use Kinesis Data Firehose to deliver your data ElasticSearch for clickstream analytics, and use EMR to generate personalized product recommendations. right

Explanation:

Correct Answer: D

Option A is incorrect. You cannot use Athena as a data catalog. You either need to use the Glue data catalog or Apache Hive as the data catalog.

Option B is incorrect. The combination of Kinesis Data Streams and Kinesis Data Analytics is better suited to near-real time analytics than it is for historical data insights. Also, this option does not address your near-real time analytics requirement.

Option C is incorrect. You cannot use Athena as a data catalog. You either need to use the Glue data catalog or Apache Hive as the data catalog. Also, Kinesis Data Firehose alone cannot give you clickstream analysis.

Option D is correct. Glue is the correct choice for your data catalog, using the Glue data catalog. Kinesis Data Streams combined with Kinesis Data Analytics satisfies your near-real time analytics requirement. Kinesis Data Firehose to ElasticSearch satisfies your clickstream requirement, and EMR uses spark jobs to satisfy your recommendation requirement at scale.

References:

Please see the Amazon Kinesis Data Analytics for SQL Applications Developer Guide SQL developer guide titled **What Is Amazon Kinesis Data Analytics for SQL Applications?** (https://docs.aws.amazon.com/kinesisanalytics/latest/dev/what-is.html),

Amazon Kinesis Data Analytics for SQL Applications Developer Guide SQL developer guide titled **Amazon Kinesis Data Analytics for SQL Applications: How It Works** (https://docs.aws.amazon.com/kinesisanalytics/latest/dev/how-it-works.html),

The AWS Glue developer guide titled **Populating the AWS Glue Data Catalog** (https://docs.aws.amazon.com/glue/latest/dg/populate-data-catalog.html),

The AWS Quick Start reference article titled **Clickstream Analytics on AWS** (https://aws.amazon.com/quickstart/architecture/clickstream-analytics/)

Ask our Experts

View Queries

Question 19

Correct

Domain: Modeling

You work as a machine learning specialist at a government agency that creates an image recognition program to help detect missing persons by analyzing surveillance videos. You have built and are now training a deep learning model for your image classification. You see that it is overfitting the training data during your model training: your training accuracy is 99%, and your testing accuracy is 75%. Why is your model overfitting the training data, and how can you address the issue?

A. Optimization stopped before model training bounced out of a local minimum. Address the issue by increasing the epoch number.

B. The mini-batch size is too low. Address the issue by increasing the mini-batch size.

C. The model is not generalized. Address the issue by increasing the dropout rate at the flatten layer. right

D. Optimization is trapped at a local minimum during training. Address the issue by increasing the learning rate.

Explanation:

Correct Answer: C

Option A is incorrect. Increasing the epoch number will cause your model to train longer. But this alone won't allow your training to reach generalization.

Option B is incorrect. Increasing the mini-batch size results in models with poor generalization, therefore not allowing your training to reach generalization.

Option C is correct. Increasing the dropout rate in your deep learning model is proven to address the issue of overfitting. See the article "Dropout Regularization in Deep Learning Models With Keras" in the reference section.

Option D is incorrect. Increasing the learning rate alone won't bring your model to generalization. A learning rate that is too large will result in an unstable network. See the article "Understand the Impact of Learning Rate on Neural Network Performance" in the reference section.

References:

Please see the Machine Learning Mastery article titled **Dropout Regularization in Deep Learning Models With Keras** (https://machinelearningmastery.com/dropout-regularization-deep-learning-models-keras/),

The Machine Learning Mastery article titled **Understand the Impact of Learning Rate on Neural Network Performance** (https://machinelearningmastery.com/understand-the-dynamics-of-learning-rate-on-deep-learning-neural-networks/),

The AWS Glue developer guide titled **Populating the AWS Glue Data Catalog** (https://docs.aws.amazon.com/glue/latest/dg/populate-data-catalog.html),

The Amazon SageMaker developer guide titled **Object2Vec Hyperparameters** (particularly the descriptions of dropout and learning_rate) (https://docs.aws.amazon.com/sagemaker/latest/dg/object2vec-hyperparameters.html)

Ask our Experts

View Queries

Question 20

Correct

Domain: Exploratory Data Analysis

You work as a machine learning specialist at a mining and minerals company. Your company has asked you to build a model that predicts the efficacy of a given drilling site. Your model training dataset has a large number of features. For your modeling exploration, you have chosen to use regression models, such as linear regression and logistic regression. During exploratory data analysis, you notice a high correlation between many features that you believe will make your model unstable. How can you address the problem of having too many highly correlated features?

A. Use a Cramer's V correlation coefficient.

B. Use Principal Component Analysis to reduce the dimensionality of the dataset. right

C. Modify highly correlated features using vector multiplication.

D. Modify highly correlated features using a Spearman correlation coefficient.

Explanation:

Correct Answer: B

Option A is incorrect. A Cramer's V Correlation Coefficient varies between 0 and 1, where a value close to 0 means that there is very little association between the variables. A Cramer's V of close to 1 indicates a very strong association, but using this alone will not help you reduce the dimensionality of your feature set. You need to reduce the dimensionality of your feature set to eliminate the highly correlated features. Principal Component Analysis is the classic approach used to address the problem of highly correlated features in a feature set.

Option B is correct. Principal Component Analysis (PCA) is the classic approach used to address the problem of highly correlated features in a feature set. Using PCA, you reduce the number of features in your dataset by reducing the dimensionality of the dataset by finding the linearly independent feature set that best represents the overall feature while minimizing information loss.

Option C is incorrect. Vector multiplication will not help you address highly correlated features.

Option D is incorrect. The Spearman correlation coefficient is used with ordinal categorical variables or continuous variables to measure correlation. This technique alone will not help you address highly correlated features.

References:

Please see the Wikipedia article titled **Cramer's V** (https://en.wikipedia.org/wiki/Cram%C3%A9r%27s_V),

The Kaggle article titled **Step by Step PCA with Iris dataset** (https://www.kaggle.com/shrutimechlearn/step-by-step-pca-with-iris-dataset),

The Machine Learning Mastery article titled **Why One-Hot Encode Data in Machine Learning?** (https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/)

Ask our Experts

View Queries

Question 21 Correct

Domain: Modeling

You are a machine learning specialist working for a polling company where you have been given the assignment of creating a machine learning model that predicts voter turnout in various voting districts across the voting population. You have created a deep learning neural network model for your predictions. The model performs well on the training data, 99% accuracy, but it performs poorly on the test data, 70% accuracy. Which techniques can you leverage to address this overfitting situation with your deep learning model? (Select TWO)

A. Increase feature dimensionality.

B. Lower the dropout rate.

☐ C. Increase the dropout rate. right

☐ D. Increase L1 and/or L2 regularization amount. right

E. Decrease L1 and/or L2 regularization amount.

Explanation:

Correct Answers: C and D

Option A is incorrect. Your model is overfitting. You should decrease the feature combinations (dimensionality) using a technique such as Principal Component Analysis (PCA) which will help with overfitting. Increasing feature combinations, or increasing the feature dimensionality, would make the model have the opposite effect.

Option B is incorrect. With an overfitting model, you need to increase your dropout rate to help the model reach generalization. Decreasing dropout will have the opposite effect.

Option C is correct. Increasing the dropout rate in your deep learning model is a proven technique to address the issue of training data overfitting.

Option D is correct. Increasing regularization through L1 regularization, L2 regularization, or dropout helps lower the complexity of the model to help address overfitting.

Option E is incorrect. Decreasing regularization will increase the complexity of your model and will not help address overfitting.

References:

Please see the Machine Learning Mastery article titled **A Gentle Introduction to Dropout for Regularizing Deep Neural Networks** (https://machinelearningmastery.com/dropout-for-regularizing-deep-neural-networks/),

The Towards Data Science article titled **Regularization in Deep Learning — L1, L2, and Dropout** (https://towardsdatascience.com/regularization-in-deep-learning-l1-l2-and-dropout-377e75acc036),

The Amazon SageMaker developer guide titled **Principal Component Analysis (PCA) Algorithm** (https://docs.aws.amazon.com/sagemaker/latest/dg/pca.html)

Ask our Experts

View Queries

Question 22

Correct

Domain: ML Implementation and Operations

You work as a machine learning specialist for a company that is required to follow the Securities and Exchange Commission (SEC) regulations. One of the ways your company adheres to some of the SEC regulations is to apply a data security policy that does not allow the sending of your machine learning data over the internet. You are building a SageMaker environment to use for your team's machine learning models. Which is the best option to make the SageMaker service available in your company's AWS account without enabling direct internet access to your machine learning specialist's SageMaker notebook instances?

A. Use Transit Gateway between your corporate VPC and the Amazon VPC hosting SageMaker.

B. Connect directly to the SageMaker runtime (the Amazon VPC hosting SageMaker) through an interface endpoint in your corporate VPC. right

C. Route your SageMaker traffic through a network in your corporate data center.

D. Use a NAT gateway in your corporate VPC to connect to the Amazon VPC hosting SageMaker.

Explanation:

Correct Answer: B

Option A is incorrect. A Transit Gateway is a network transit hub used for creating a VPC connection between your own VPCs and an on-premise network. It is not used to communicate with AWS services like SageMaker. VPC interface and gateway endpoints are used to accomplish private VPC connections to services like SageMaker.

Option B is correct. Using an interface endpoint in your VPC, you can connect directly to the SageMaker API or the SageMaker runtime without ever connecting over the internet. Using a VPC interface endpoint, communication between your VPC and the SageMaker runtime is conducted securely within the AWS network.

Option C is incorrect. Routing your SageMaker machine learning data through your corporate data center network will not give you access to the Amazon VPC hosting SageMaker.

Option D is incorrect. If you use a NAT Gateway, you will send your traffic over the internet. The scenario explicitly states that you cannot send your machine learning data over the internet.

References:

Please see the Amazon Virtual Private Cloud AWS PrivateLink page titled **VPC endpoints** (https://docs.aws.amazon.com/vpc/latest/privatelink/vpc-endpoints.html),

The Amazon SageMaker development guide titled **Connect to SageMaker Through a VPC Interface Endpoint** (https://docs.aws.amazon.com/sagemaker/latest/dg/interface-vpc-endpoint.html),

The Amazon Virtual Private Cloud transit gateways guide titled **What is a Transit Gateway?** (https://docs.aws.amazon.com/vpc/latest/tgw/what-is-transit-gateway.html),

The Amazon Virtual Private Cloud VPC peering page titled **Unsupported VPC peering configurations** (https://docs.aws.amazon.com/vpc/latest/peering/invalid-peering-configurations.html)

Ask our Experts

View Queries

Question 23

Correct

Domain: Modeling

You work for a consumer electronics company as a machine learning specialist. Over time your company has built up a large set of labeled historical consumer electronic device sales data. You have been given the task of predicting how many memory components should be produced each quarter to satisfy the demand for your consumer electronic products. Which algorithm should you choose to get the best performing model to solve this prediction problem?

- A. Linear regression right
- B. Latent Dirichlet Allocation (LDA)
- C. Sequence-to-Sequence
- D. Logistic regression

Explanation:

Correct Answer: A

Option A is correct. When you solve a continuous number for your prediction (how many), you use linear regression. If you are solving for a binary prediction (yes/no), you use logistic regression.

Option B is incorrect. Latent Dirichlet Allocation (LDA) is not used for the prediction of continuous values. LDA is an approach used as an unsupervised learning algorithm that attempts to describe

a set of observations as a mixture of distinct categories. Using LDA, you discover a user-specified number of topics shared by documents within a text corpus.

Option C is incorrect. This option is incorrect because the Sequence-to-Sequence algorithm is primarily used as a supervised algorithm for language translation, text summarization, and speech-to-text. You would not use a Sequence-to-Sequence algorithm to solve a regression problem.

Option D is incorrect. When you solve a continuous number for your prediction (how many), you use linear regression. If you are solving for a binary prediction (yes/no), you use logistic regression.

References:

Please see the Amazon SageMaker developer guide titled **Latent Dirichlet Allocation (LDA) Algorithm** (https://docs.aws.amazon.com/sagemaker/latest/dg/lda.html),

The Amazon SageMaker developer guide titled **Linear Learner Algorithm** (https://docs.aws.amazon.com/sagemaker/latest/dg/linear-learner.html),

The Amazon SageMaker developer guide titled **Sequence-to-Sequence Algorithm** (https://docs.aws.amazon.com/sagemaker/latest/dg/seq-2-seq.html),

The Amazon Amazon Machine Learning developer guide titled **Regression Model Insights** (https://docs.aws.amazon.com/machine-learning/latest/dg/regression-model-insights.html)

Ask our Experts

View Queries

Question 24

Correct

Domain: Modeling

Your work for a company that performs seismic research for client firms that drill for petroleum. As a machine learning specialist, you have built a series of models that classify seismic waves to determine the seismic profile of a proposed drilling site. You need to select the best model to use in production. Which metric should you use to compare and evaluate your machine learning classification models against each other?

- A. Area Under the ROC Curve (AUC) right
- B. Mean square error (MSE)
- C. Mean Absolute Error (MAE)
- D. Recall

Explanation:

Correct Answer: A

Option A is correct. The area under the Receiver Operating Characteristic (ROC) curve is the most commonly used metric to compare classification models.

Option B is incorrect. The Mean Square Error (MSE) is commonly used to measure regression error. It finds the average squared error between the predicted and actual values. It is not used to compare classification models.

Option C is incorrect. The Mean Square Error is also commonly used to measure regression error. It finds the average absolute distance between the predicted and target values. It is not used to compare classification models.

Option D is incorrect. The recall metric is the percentage of results correctly classified by a model. This metric alone will not allow you to make a complete assessment and comparison of your models.

References:

Please see the Towards Data Science article titled **Metrics For Evaluating Machine Learning Classification Models** (https://towardsdatascience.com/metrics-for-evaluating-machine-learning-classification-models-python-example-59b905e079a5),

The Towards Data Science article titled **How to Evaluate a Classification Machine Learning Model** (https://towardsdatascience.com/how-to-evaluate-a-classification-machine-learning-model-d81901d491b1),

The Machine Learning Mastery article titled **Assessing and Comparing Classifier Performance with ROC Curves** (https://machinelearningmastery.com/assessing-comparing-classifier-performance-roc-curves-2/),

The Towards Data Science article titled **20 Popular Machine Learning Metrics. Part 1: Classification & Regression Evaluation Metrics** (https://towardsdatascience.com/20-popular-machine-learning-metrics-part-1-classification-regression-evaluation-metrics-1ca3e282a2ce),

The Data School article titled **Simple guide to confusion matrix terminology** (https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/),

The Medium article titled **Precision vs. Recall** (https://medium.com/@shrutisaxena0617/precision-vs-recall-386cf9f89488)

Ask our Experts

View Queries

Question 25

Correct Marked for review

Domain: Exploratory Data Analysis

You are working as a machine learning specialist for a car rental firm that wishes to use machine learning to optimize the cost per mile of its rental cars based on geographic region. You have a large

car rental database for use in your model training that has features such as rental car type, rental geographic region, miles driven, average regional gas price, etc. During your exploratory data analysis tasks, you notice that your feature seat contains outliers for the miles driven and average regional gas price. These outliers are likely to make your regression algorithm-based model unstable. What data preparation technique can you use to prepare your data for your model training?

A. Normalize your features to reduce the effect of the outlier data.

B. Use Quantile Binning of your features to reduce the effect of the outlier data.

C. Min-Max Scale your features to reduce the effect of the outlier data.

D. Standardize your features to reduce the effect of the outlier data. right

Explanation:

Correct Answer: D

Option A is incorrect. Both normalization and standardization are types of data scaling that machine learning specialists use to prepare their data for training and inference. Normalization rescales your data so that all values range from 0 to 1. Therefore, normalization doesn't handle outliers well.

Option B is incorrect. Quantile Binning is used to categorize feature values into bins. This technique would not reduce the effect of your outliers since the outliers would skew whichever bin in which they are placed.

Option C is incorrect. The Min-Max Scaling term is another name for normalizing your data. Normalization doesn't handle outliers as well as standardization.

Option D is correct. Standardization centers your feature values around the mean. So, it has no bounding range. Therefore, standardization handles outliers much better than normalization.

References:

Please see the article titled **Feature Scaling for Machine Learning: Understanding the Difference Between Normalization vs. Standardization** (https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/),

The Amazon Machine Learning developer guide titled **Data Transformations Reference** (https://docs.aws.amazon.com/machine-learning/latest/dg/data-transformations-reference.html)

Ask our Experts

View Queries

Question 26

Correct

Domain: Modeling

You work as a machine learning specialist for a hedge fund firm. Your traders trade in highly volatile securities and derivatives. In real-time, their trading activity must be monitored for an anomalous activity to keep the firm from entering into potentially very large high-risk transactions that could jeopardize the firm's valuation and collateral obligations with the Securities and Exchange Commission (SEC). Which of the following options best describes your optimal machine learning solution to this problem?

A. Use Kinesis Data Streams to gather the trading, valuation, and collateral data from your investment management systems, source the data from Kinesis Data Streams to Kinesis Data Analytics, use SQL to transform the data and write it to S3, use the SageMaker Random Cut Forest built-in algorithm to detect anomalous trading activity.

B. Use Kinesis Data Firehose to gather the trading, valuation, and collateral data from your investment management systems, source the data from Kinesis Data Firehose to Kinesis Data Analytics, use SQL to transform the data and write it to S3, use the SageMaker k-means built-in algorithm to detect anomalous trading activity.

C. Use Kinesis Data Streams to gather the trading, valuation, and collateral data from your investment management systems, source the data from Kinesis Data Streams to Kinesis Data Analytics, use Apache Flink to transform the data and write it to S3, use the SageMaker Random Cut Forest built-in algorithm to detect anomalous trading activity. right

D. Use a Glue ETL job to gather the trading, valuation, and collateral data from your investment management systems. Have the Glue ETL job transform the data and write it to S3, use the SageMaker Random Cut Forest built-in algorithm to detect anomalous trading activity.

Explanation:

Correct Answer: C

Option A is incorrect. Kinesis Data Streams will satisfy your real-time data gathering requirement. The SageMaker Random Cut Forest algorithm will satisfy your anomaly detection requirement. However, using SQL in your Kinesis Data Analytics application will require a Lambda function to write your data to S3.

Option B is incorrect. Kinesis Data Firehose will deliver your data in near real-time, not real-time like Kinesis Data Streams. Also, using SQL in your Kinesis Data Analytics application will require a Lambda function to write your data to S3. Finally, the k-means algorithm is not the best choice for anomaly detection. Random Cut Forest is the best algorithm for anomaly detection.

Option C is correct. Kinesis Data Streams will satisfy your real-time data gathering requirement. Kinesis Data Analytics running an Apache Flink application will allow you to transform your data and directly write it to S3. Finally, the Random Cut Forest algorithm is the best choice for anomaly detection.

Option D is incorrect. A Glue ETL job will not satisfy your real-time data delivery requirement. Glue ETL is used in batch applications.

References:

Please see the **US Securities and Exchange Commission compliance alert** dated July, 2008 (https://www.sec.gov/about/offices/ocie/complialert0708.htm),

The Amazon Kinesis Data Analytics developer guide titled **Example: Writing to an Amazon S3 Bucket** (https://docs.aws.amazon.com/kinesisanalytics/latest/java/examples-s3.html),

The Amazon SageMaker developer guide titled **Random Cut Forest (RCF) Algorithm** (https://docs.aws.amazon.com/sagemaker/latest/dg/randomcutforest.html)

Ask our Experts

View Queries

Question 27

Correct

Domain: Exploratory Data Analysis

You work as a machine learning specialist for an online news organization. Your company is implementing a crowd-sourced news feature that will allow subscribers to upload their own news stories complete with images. You have been tasked with building and deploying a machine learning model that alerts adult content in an image uploaded by a subscriber. Which option describes the best data preparation and implementation of your machine learning hosted inference?

A. Use the Semantic Segmentation SageMaker built-in algorithm and code your application to query the deployed endpoint providing an image in the image/png content type.

B. Use the Object2Vec SageMaker built-in algorithm and code your application to query the deployed endpoint providing an image in the application/jsonlines content type.

C. Use the Image Classification SageMaker built-in algorithm and code your application to query the deployed endpoint providing an image in the application/x-recordio content type.

D. Use the Image Classification SageMaker built-in algorithm and code your application to query the deployed endpoint providing an image in the application/x-image content type. right

Explanation:

Correct Answer: D

Option A is incorrect. The Semantic Segmentation SageMaker built-in algorithm is used to provide a fine-grained, pixel-level approach to developing computer vision applications. It would function poorly as a model to detect adult content in an image.

Option B is incorrect. The Object2Vec SageMaker built-in algorithm is used to compute the nearest neighbors of objects and visualize natural clusters of related objects in low-dimensional space. For example, common use cases include identifying duplicate support tickets or finding the correct routing based on the similarity of text in the tickets. It would function poorly as a model to detect adult content in an image.

Option C is incorrect. The Image Classification SageMaker built-in algorithm is the correct algorithm for using as an inference engine to detect adult content in an uploaded image. However, an Image Classification deployed endpoint doesn't take inference requests in the application/x-recordio content type.

Option D is correct. The Image Classification SageMaker built-in algorithm is the correct algorithm for using as an inference engine to detect adult content in an uploaded image. Also, an Image Classification deployed endpoint takes inference requests in the application/x-image content type.

References:

Please see the Amazon SageMaker developer guide titled **Use Amazon SageMaker Built-in Algorithms** (https://docs.aws.amazon.com/sagemaker/latest/dg/algos.html),

The Amazon SageMaker developer guide titled **Semantic Segmentation Algorithm** (https://docs.aws.amazon.com/sagemaker/latest/dg/semantic-segmentation.html),

The Amazon SageMaker developer guide titled **Object2Vec Algorithm** (https://docs.aws.amazon.com/sagemaker/latest/dg/object2vec.html),

The Amazon SageMaker developer guide titled **Image Classification Algorithm** (https://docs.aws.amazon.com/sagemaker/latest/dg/image-classification.html),

The Amazon SageMaker developer guide titled **Common Data Formats for Inference** (https://docs.aws.amazon.com/sagemaker/latest/dg/cdf-inference.html)

Ask our Experts

View Queries

Question 28

Correct Marked for review

Domain: Data Engineering

You work as a machine learning specialist for the airline traffic control agency of the federal government. Your machine learning team is responsible for producing the models that process all air traffic in-flight data to produce recommended flight paths for the aircraft currently aloft. The flight paths need to consider all of the prevailing conditions (weather, other flights in the path, etc.) that may affect an aircraft's flight path.

The data that your models need to process is massive in scale and requires large-scale data processing. How should you build the data transformation and feature engineering processing jobs so

that you can process all of the flight data in real-time?

A. Run Glue ETL distributed data processing jobs to perform the transformation and feature engineering on the flight data in real-time and save the data to S3 for your model training.

B. Use Kinesis Data Firehose to perform the transformation and feature engineering on the flight data in real-time and save the data to S3 for your model training.

C. Run Apache Spark Streaming data processing jobs to perform the transformation and feature engineering on the flight data in real-time and save the data to S3 for your model training. right

D. Use a Kinesis Data Analytics SQL application to perform the transformation and feature engineering on the flight data in real-time and save the data to S3 for your model training.

Explanation:

Correct Answer: C

Option A is incorrect. Glue ETL is used for batch processing. So it will not work in a real-time scenario.

Option B is incorrect. Kinesis Data Firehose is a near real-time processing service (it buffers your data as it processes it using the buffer size and buffer interval configuration settings). It will not work in a real-time scenario.

Option C is correct. Apache Spark Streaming is an analytics engine used for large-scale data processing that runs distributed data processing jobs. You can apply data transformations and extract features (feature engineering) using the Spark framework.

Option D is incorrect. Kinesis Data Analytics running a SQL application can't write directly to S3. Also, Kinesis Data Analytics cannot scale to the large-scale data processing capabilities that Apache Spark jobs can.

References:

Please see the Amazon SageMaker developer guide titled **Data Processing with Apache Spark** (https://docs.aws.amazon.com/sagemaker/latest/dg/use-spark-processing-container.html),

The Amazon SageMaker Examples GitHub repository titled **Distributed Data Processing using Apache Spark and SageMaker Processing** (https://github.com/aws/amazon-sagemaker-examples/blob/master/sagemaker_processing/spark_distributed_data_processing/sagemaker-spark-processing.ipynb),

The Amazon Kinesis Data Firehose developer guide titled **Configure Settings** (https://docs.aws.amazon.com/firehose/latest/dev/create-configure.html)

The Amazon Kinesis Data Analytics **FAQs** (https://aws.amazon.com/kinesis/data-analytics/faqs/)

Ask our Experts

View Queries

Question 29 Correct

Domain: ML Implementation and Operations

You work as a machine learning specialist for a political candidate that is mounting a campaign to get reelected in her US senate district. Your job is to build a machine learning model that allows the campaign to understand how to reach groups of similar counties by highlighting messages that resonate with those groups. The senate candidate has a limited budget. So, you need to build a cost-effective solution. Which machine learning services and features should you use to solve this problem?

A. Gather US anonymized census data from the US census on demographics by different US counties using the US Census Bureau Data API and stream it to Kinesis Data Streams. Write a Kinesis Client Library application to perform feature engineering on the data and write it to S3. Then use the Factorization Machine SageMaker built-in algorithm to produce the similar counties analysis to be used in the advertising for the grouped counties.

B. Gather US anonymized census data from the US census on demographics by different US counties using the US Census Bureau Data API and stream it to Kinesis Data Firehose. Use the Kinesis Data Firehose Lambda blueprints to create your Lambda function to use as transformations to perform feature engineering on the data and write it to S3. Then use the K-Means SageMaker built-in algorithm to produce the similar counties analysis to be used in the advertising for the grouped counties. right

C. Gather US anonymized census data from the US census on demographics by different US counties using the US Census Bureau Data API and process the data using a Glue ETL job. Have the Glue ETL job transformation the data to perform feature engineering on the data and write it to S3. Then use the K-Nearest Neighbors SageMaker built-in algorithm to produce the similar counties analysis to be used in the advertising for the grouped counties.

D. Gather US anonymized census data from the US census on demographics by different US counties using the US Census Bureau Data API and stream it to Kinesis Data Firehose. Use the Kinesis Data Firehose Lambda blueprints to create your Lambda function to use as transformations to perform feature engineering on the data and write it to S3. Then use the K-Nearest Neighbors SageMaker built-in algorithm to produce the similar counties analysis to be used in the advertising for the grouped counties.

Explanation:

Correct Answer: B

Option A is incorrect. Using Kinesis Data Streams and writing a Kinesis Data Streams Client Library application will be more costly than using Kinesis Data Firehose and its Lambda blueprint capability. Also, the Factorization Machines algorithm is not used for clustering groups as this scenario requires.

Option B is correct. Kinesis Data Firehose and its Lambda blueprints capability allow you to create the data gathering part of your machine learning solution at a lower cost than the other options. Also, the K-Means algorithm is the best algorithm to use for clustering of groups, as this scenario requires.

Option C is incorrect. You could use a Glue ETL job to gather the census data from the API. However, the K-Nearest Neighbors algorithm is not a good choice for the clustering of groups as this scenario requires.

Option D is incorrect. Kinesis Data Firehose and its Lambda blueprints capability allow you to create the data gathering part of your machine learning solution at a lower cost than the other options. However, the K-Nearest Neighbors algorithm is not a good choice for the clustering of groups as this scenario requires.

References:

Please see the Amazon SageMaker developer guide titled **K-Means Algorithm** (https://docs.aws.amazon.com/sagemaker/latest/dg/k-means.html),

The Amazon SageMaker developer guide titled **K-Nearest Neighbors (k-NN) Algorithm** (https://docs.aws.amazon.com/sagemaker/latest/dg/k-nearest-neighbors.html),

Amazon SageMaker Examples GitHub repository titled **Analyze US census data for population segmentation using Amazon SageMaker** (https://github.com/aws/amazon-sagemaker-examples/blob/master/introduction_to_applying_machine_learning/US-census_population_segmentation_PCA_Kmeans/sagemaker-countycensusclustering.ipynb),

The Amazon SageMaker developer guide titled **Factorization Machines Algorithm** (https://docs.aws.amazon.com/sagemaker/latest/dg/fact-machines.html),

The Amazon Kinesis Data Firehose developer guide titled **Amazon Kinesis Data Firehose Data Transformation** (https://docs.aws.amazon.com/firehose/latest/dev/data-transformation.html)

Ask our Experts

View Queries

Question 30

Correct

Domain: Modeling

You work as a machine learning specialist for a scientific instruments company. Your machine learning team has been assigned the task of developing a machine learning model that optimizes the electronic components in the production line to the current product development lifecycle. You have built your model using the XGBoost SageMaker built-in algorithm. You are now in the process of tuning the model hyperparameters. In your hyperparameter job, you have chosen the num_class, num_round, alpha, booster, early_stopping_rounds, min_child_weight, subsample, eta, and num_round hyperparameters to use in your optimization. When running your hyperparameter tuning job, you have noticed that the computational complexity of a hyperparameter tuning job is high. You

are getting suboptimal results. Which option should you implement to get better results from your hyperparameter tuning job?

A. Reduce the number of hyperparameters in your optimization by removing the alpha hyperparameter from your hyperparameter tuning job.

B. Reduce the number of hyperparameters in your optimization by removing the eta hyperparameter from your hyperparameter tuning job.

C. Adjust the range of values for each of the hyperparameters to a smaller range of values. right

D. Adjust the range of values for each of the hyperparameters to a larger range of values.

Explanation:

Correct Answer: C

Option A is incorrect. The alpha hyperparameter is one of the hyperparameters that have the greatest effect on optimizing the XGBoost evaluation metrics. Therefore, removing it will not help optimize your hyperparameter tuning job.

Option B is incorrect. The eta hyperparameter is one of the hyperparameters that have the greatest effect on optimizing the XGBoost evaluation metrics. Therefore, removing it will not help optimize your hyperparameter tuning job.

Option C is correct. Adjusting (limiting) the range of values that you search for each of the hyperparameters to a smaller range of values can significantly improve the success of hyperparameter optimization.

Option D is incorrect. Adjusting (expanding) the range of values that you search for each of the hyperparameters to a larger range of values will most likely make your hyperparameter optimization less optimal.

References:

Please see the Amazon SageMaker developer guide titled **How Hyperparameter Tuning Works** (https://docs.aws.amazon.com/sagemaker/latest/dg/automatic-model-tuning-how-it-works.html),

The Amazon SageMaker developer guide titled **Best Practices for Hyperparameter Tuning** (https://docs.aws.amazon.com/sagemaker/latest/dg/automatic-model-tuning-considerations.html),

The Amazon SageMaker developer guide titled **XGBoost Hyperparameters** (https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost_hyperparameters.html),

The Amazon SageMaker developer guide titled **Tune an XGBoost Model** (https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost-tuning.html),

The Amazon SageMaker developer guide titled **Example: Hyperparameter Tuning Job** (https://docs.aws.amazon.com/sagemaker/latest/dg/automatic-model-tuning-ex.html)

Ask our Experts

View Queries

Question 31

Correct Marked for review

Domain: Exploratory Data Analysis

You work as a machine learning specialist for a large lending agency that issues mortgage loans to the residential home-buying population. You and your machine learning team build a model to assess loan risk based on loan application data. You have not yet chosen which algorithm your model will use. Still, you need to sanitize your data to ensure your data is not biased by demographic disparities, such as having different distributions for loan application outcomes for different demographic groups. Which option is the most efficient approach to use to identify bias in your data prior to training your modeling using the data?

- **A. Run a SageMaker Clarify job using the Total Variation Distance (TVD) pretraining metric.** right
- **B. Run a SageMaker Clarify job using the Difference in Proportions of Outcomes (DPO) pretraining metric.**
- **C. Label your data using Ground Truth using Amazon Mechanical Turk with a custom labeling workflow.**
- **D. Label your data using Ground Truth using Amazon Mechanical Turk using automated data labeling.**

Explanation:

Correct Answer: A

Option A is correct. SageMaker Clarify allows you to identify bias during data preparation using attributes of interest, such as gender or demographic, and SageMaker Clarify runs a set of algorithms to detect the presence of bias in those attributes. The Total Variation Distance (TVD) metric measures the difference between distinct demographic distributions of the outcomes associated with different facets in a dataset, such as how different are the distributions for loan application outcomes for different demographics.

Option B is incorrect. There is no Difference in the Proportions of Outcomes (DPO) pretraining metric in SageMaker Clarify.

Option C is incorrect. Labeling your data using Ground Truth alone will not help you identify bias in your data, whether using a custom labeling workflow or automated data labeling.

Option D is incorrect. Labeling your data using Ground Truth alone will not help you identify bias in your data, whether using a custom labeling workflow or automated data labeling.

References:

Please see the Amazon SageMaker developer guide titled **Measure Pretraining Bias** (https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-measure-data-bias.html),

The Amazon SageMaker developer guide titled **Use Amazon SageMaker Ground Truth to Label Data** (https://docs.aws.amazon.com/sagemaker/latest/dg/sms.html),

The Amazon SageMaker developer guide titled **Generate Reports for Bias in Pretraining Data in SageMaker Studio** (https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-data-bias-reports-ui.html)

Ask our Experts

View Queries

Question 32

Incorrect Marked for review

Domain: Data Engineering

You work as a machine learning specialist for a financial services firm. Your machine learning team has been tasked with building a quantitative analysis model for your mutual fund portfolio managers in the firm's quant department. You have several financial data provider data sources that you need to use in your model. You are looking for the optimal data source platform to ingest data into your machine learning jupyter notebook environment. Which options are NOT a data source platform that you can use? (Select TWO)

A. Athena

B. Redshift

- [x] C. EMR wrong

D. DynamoDB right

- [x] E. RDS right

Explanation:

Correct Answers: D and E

Option A is incorrect. The most commonly used data source for SageMaker is an S3 bucket. However, you can also use Athena, EMR, and Redshift as data sources for SageMaker.

Option B is incorrect. The most commonly used data source for SageMaker is an S3 bucket. However, you can also use Athena, EMR, and Redshift as data sources for SageMaker.

Option C is incorrect. The most commonly used data source for SageMaker is an S3 bucket. However, you can also use Athena, EMR, and Redshift as data sources for SageMaker.

Option D is correct. DynamoDB is not a viable data source for ingesting data into your machine learning jupyter notebook environment.

Option E is correct. RDS is not a viable data source for ingesting data into your machine learning jupyter notebook environment.

References:

Please see the Amazon SageMaker Examples Read the Docs Data Ingestion guide titled **Get started with data ingestion** (https://sagemaker-examples.readthedocs.io/en/latest/ingest_data/index.html),

The Amazon SageMaker Examples Read the Docs Data Ingestion guide titled **Ingest data with Athena** (https://sagemaker-examples.readthedocs.io/en/latest/ingest_data/02_Ingest_data_with_Athena_v1.html),

The Amazon SageMaker Examples Read the Docs Data Ingestion guide titled **Ingest Data with EMR** (https://sagemaker-examples.readthedocs.io/en/latest/ingest_data/04_Ingest_data_with_EMR.html),

The Amazon SageMaker Examples Read the Docs Data Ingestion guide titled **Ingest data with Redshift** (https://sagemaker-examples.readthedocs.io/en/latest/ingest_data/03_Ingest_data_with_Redshift_v3.html)

Ask our Experts

View Queries

Question 33

Correct

Domain: ML Implementation and Operations

You work as a machine learning specialist for a banking firm where you are part of the machine learning team in the fraud department. Your team's latest assignment is to build and deploy a fraud prediction model based on the SageMaker Random Cut Forest built-in algorithm. You are working through your deployment steps manually using the SageMaker Studio before you automate the pipeline. You have created an MLOps project in SageMaker Studio and chosen the MLOps template for model building, training, and a deployment project template. You have cloned the model repo to your local SageMaker Studio environment, made your necessary pipeline changes, committed your code and MLOps has triggered a run of your pipeline. What are the steps that follow the MLOps triggering of your pipeline?

A. MLOps deploys your model to your production endpoint.

B. MLOps creates a new model version. You approve the new version, then MLOps deploys your new model version to your staging environment. From CodePipeline you choose your pipeline and

approve the DeployStaging stage which causes the MLOps system to deploy the model to the production endpoint. right

C. MLOps creates a new model version then MLOps deploys your new model version to your production endpoint.

D. MLOps creates a new model version. MLOps deploys your new model version to your staging environment. From CodePipeline you choose your pipeline and approve the DeployStaging stage which causes the MLOps system to deploy the model to the production endpoint.

Explanation:

Correct Answer: B

Option A is incorrect. You cannot deploy directly to production using the MLOps project flow. Deploying your SageMaker Studio project through the console requires you to approve the new version, deploy to a staging environment, and use CodePipeline to approve your DeployStaging stage.

Option B is correct. Deploying your SageMaker Studio project through the console requires you to approve the new version, deploy to a staging environment, and use CodePipeline to approve your DeployStaging stage.

Option C is incorrect. MLOps first deploys your new DeployStaging stage to your staging environment.

Option D is incorrect. You are required to approve your new version of your project before it deploys it to staging.

References:

Please see the Amazon SageMaker developer guide titled **What is a SageMaker Project?** (https://docs.aws.amazon.com/sagemaker/latest/dg/sagemaker-projects-whatis.html),

The Amazon SageMaker developer guide titled **SageMaker MLOps Project Walkthrough** (https://docs.aws.amazon.com/sagemaker/latest/dg/sagemaker-projects-walkthrough.html)

Ask our Experts

View Queries

Question 34

Correct

Domain: Data Engineering

You work as a machine learning specialist for a marketing consulting firm with a new client that requires marketing data to help them determine which marketing campaign will be the most productive for their new product. You and your machine learning team are using SageMaker Studio to

create a Data Flow in SageMaker Studio. You plan to use this Data Flow to prepare and visualize your data using SageMaker Data Wrangler. In SageMaker Data Wrangler, you are building your data preparation pipeline. The data you are using is gathered from marketing data providers. You have imported your source data from your S3 bucket, and you are now configuring your transform for your dataset. However, you have discovered that the built-in transforms provided by Data Wrangler do not meet your transformation needs. Which options are a viable approach to building your transform in Data Wrangler? (Select TWO)

A. Select a custom transform step and write your custom transform in the Python (Scikit) programming language.

B. Choose the built-in transform that most closely resembles the transform you need and customize that built-in transform.

- [x] C. Select a custom transform step and write your custom transform in the Python (Pandas) programming language. right

D. Select a custom transform step and write your custom transform in the Scala programming language.

- [x] E. Select a custom transform step and write your custom transform in the Python (PySpark) programming language. right

Explanation:

Correct Answers: C and E

Option A is incorrect. There are only three options for the language when creating a custom transform in SageMaker Data Wrangler: Python (PySpark), Python (Pandas), and SQL (PySpark SQL).

Option B is incorrect. You cannot customize the built-in SageMaker Data Wrangler transforms.

Option C is correct. Python (Pandas) is one of the available languages for creating a custom SageMaker Data Wrangler transform.

Option D is incorrect. There are only three options for the language when creating a custom transform in SageMaker Data Wrangler: Python (PySpark), Python (Pandas), and SQL (PySpark SQL).

Option E is correct. Python (PySpark) is one of the available languages for creating a custom SageMaker Data Wrangler transform.

References:

Please see the Amazon SageMaker developer guide titled **Use the Amazon SageMaker Studio Launcher** (https://docs.aws.amazon.com/sagemaker/latest/dg/studio-launcher.html),

The Amazon SageMaker developer guide titled **Create and Use a Data Wrangler Flow** (https://docs.aws.amazon.com/sagemaker/latest/dg/data-wrangler-data-flow.html)

Ask our Experts

View Queries

Question 35

Correct

Domain: Modeling

You work as a machine learning specialist for a small startup software company. You are the only machine learning specialist in the company. The founder of the company needs you to quickly build a machine learning model to test one of the team's minimum viable products with the intent of persevering or pivoting depending on the outcome of your model experiment. You have decided to use SageMaker Autopilot to create your experiment. You are creating your experiment in SageMaker Autopilot. You have selected the S3 bucket, data source, and target feature for which to make predictions. You are now ready to select the machine learning problem type and objective metric. Which are viable combinations for your selections?

A. **Problem type: Multiclass Classification; Objective: AUC**

B. **Problem type: Regression; Objective: F1**

C. **Problem type: Binary Classification; Objective: MSE**

D. **Problem type: Regression; Objective: MSE** right

Explanation:

Correct Answer: D

Option A is incorrect. When running a Multiclass Classification algorithm-based model, the only options that make sense in the SageMaker Autopilot available choices are Accuracy and F1macro.

Option B is incorrect. When running a Regression algorithm-based model, the only option that makes sense in the SageMaker Autopilot available choices is MSE.

Option C is incorrect. When running a Binary Classification algorithm-based model, the only options that make sense in the SageMaker Autopilot available choices are F1, Accuracy, and AUC.

Option D is correct. When running a Regression algorithm-based model, the option that makes sense in the SageMaker Autopilot available choices is MSE.

References:

Please see the Amazon SageMaker developer guide titled **Automate model development with Amazon SageMaker Autopilot** (https://docs.aws.amazon.com/sagemaker/latest/dg/autopilot-automate-model-development.html),

The AWS Getting Started Resource Center titled **Create a machine learning model automatically** (https://aws.amazon.com/getting-started/hands-on/create-machine-learning-model-automatically-sagemaker-autopilot/)

Ask our Experts

Ask our Experts

View Queries

Question 36 Correct Marked for review

Domain: Modeling

You work as a machine learning specialist for a social media software company. Your company produces social media game apps. Your machine learning team has been asked to produce a machine learning model to predict user purchase of apps similar to apps they have already purchased. You have created a model based on the SageMaker built-in XGBoost algorithm. You are now using hyperparameter tuning to get the best performing model for your problem. Which evaluation metrics and corresponding optimization direction should you choose for your automatic model tuning (a.k.a. hyperparameter tuning)? (Select TWO)

A. f1, minimize

B. map, minimize

☐ C. ndcg, maximize right

D. rmse, maximize

☐ E. mae, minimize right

Explanation:

Correct Answers: C and E

Option A is incorrect. XGBoost uses the f1 metric for model validation. However, you will want to maximize this metric.

Option B is incorrect. XGBoost uses the map (mean average precision) metric for model validation. However, you will want to maximize this metric.

Option C is correct. XGBoost uses the ndgc (Normalized Discounted Cumulative Gain) metric for model validation, and you will want to maximize this metric.

Option D is incorrect. XGBoost uses the rmse (Root mean square error) metric for model validation, and you will want to minimize this metric.

Option E is correct. XGBoost uses the mae (Mean Absolute Error) metric for model validation, and you will want to minimize this metric.

References:

Please see the Amazon SageMaker developer guide titled **Define Metrics** (https://docs.aws.amazon.com/sagemaker/latest/dg/automatic-model-tuning-define-metrics.html),

The Amazon SageMaker developer guide titled **Tune an XGBoost Model** (https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost-tuning.html)

Ask our Experts

View Queries

Question 37

Correct

Domain: Exploratory Data Analysis

You work as a machine learning specialist for an auto manufacturer. You are on a machine learning team that is responsible for analyzing the efficiency of potential electric car drive trains. These drivetrains have explicit energy storage requirements (regenerative braking) to help with efficiency when driving in cities. Your team is in the feature engineering phase of your model development. You need to produce visualizations to get an idea of which features are useful and which can be improved using dimensionality reduction. You have several data sources that you would like to visualize in your QuickSight environment. Which of your data sources cannot be directly used as data sources in QuickSight?

- A. DynamoDB right
- B. Snowflake
- C. Presto
- D. Teradata

Explanation:

Correct Answer: A

Option A is correct. DynamoDb is not supported as a direct data source for QuickSight. You need to use an intermediary service, such as Athena and its data source connectors, to make your DynamoDB table data available to QuickSight.

Option B is incorrect. Snowflake is a supported data source for QuickSight.

Option C is incorrect. Presto is a supported data source for QuickSight.

Option D is incorrect. Teradata is a supported data source for QuickSight.

References:

Please see the Amazon QuickSight user guide titled **Supported Data Sources** (https://docs.aws.amazon.com/quicksight/latest/user/supported-data-sources.html),

The article titled **GETTING TO THE HEART OF EVS: A CLOSE-UP LOOK AT THE ELECTRIC DRIVETRAIN** (https://www.innovativeautomation.com/the-electric-vehicle-drivetrain/),

The AWS Big Data blog titled **Accessing and visualizing data from multiple data sources with Amazon Athena and Amazon QuickSight,** (https://aws.amazon.com/blogs/big-data/accessing-and-visualizing-data-from-multiple-data-sources-with-amazon-athena-and-amazon-quicksight/)

Ask our Experts

View Queries

Question 38

Correct

Domain: ML Implementation and Operations

You work as a machine learning specialist for a research data streaming service that serves research reference content to subscribers. Your company's subscriber base is primarily made up of university research staff. However, your company occasionally produces research content that has broader appeal, and your service gets very big spikes in requests for streaming traffic. Your machine learning team has a critical component in the content delivery process. You have a recommendation engine model variant that processes inference requests for every content streaming request. When your model variant receives these spikes in inference requests, your company's streaming service suffers poor performance. You have decided to use SageMaker autoscaling to meet the varying demand for your model variant inference requests. Which type of scaling policy should you use in your SageMaker autoscaling implementation?

- A. target-tracking scaling right
- B. step scaling
- C. simple scaling
- D. scheduled scaling

Explanation:

Correct Answer: A

Option A is correct. AWS recommends that you use scaling policies for your autoscaling configuration because it is fully automated.

Option B is incorrect. AWS recommends that you use step scaling when you need an advanced configuration, such as specifying how many instances to deploy under certain circumstances. You don't have a specialized need like this, so you should use target-tracking scaling.

Option C is incorrect. SageMaker autoscaling doesn't have a simple scaling policy.

Option D is incorrect. SageMaker autoscaling doesn't have a scheduled scaling policy.

References:

Please see the Amazon SageMaker developer guide titled **Automatically Scale Amazon SageMaker Models** (https://docs.aws.amazon.com/sagemaker/latest/dg/endpoint-auto-scaling.html),

The Amazon SageMaker developer guide titled **Prerequisites** (https://docs.aws.amazon.com/sagemaker/latest/dg/endpoint-auto-scaling-prerequisites.html)

Ask our Experts

View Queries

Question 39

Correct

Domain: Modeling

You work as a machine learning specialist for a real estate company. Your company wishes to have you develop a model that predicts if a given property is in a “high value” neighborhood (properties with a median household value at or above $180,000). Your real estate agents will use this model to prioritize their sales work based on potential commission for any given property in their list of potential sales leads. Which option is the best approach to solve this problem?

A. Use SageMaker Linear Learner optimizing for a continuous objective, such as mean square error, cross-entropy loss, or absolute error to predict the median household value for each district.

B. Use SageMaker Linear Learner optimizing for a discrete objective suited for classification, such as F1, precision, recall, or accuracy to predict whether or not a district is "high value". right

C. Use SageMaker Linear Learner optimizing for a continuous objective, such as mean square error, cross-entropy loss, or absolute error to predict whether or not a district is "high value".

D. Use SageMaker Linear Learner optimizing for a continuous objective, such as F1, precision, recall, or accuracy to predict the median household value for each district.

Explanation:

Correct Answer: B

Option A is incorrect. We are solving a classification problem: predict if a given property is in a “high value” neighborhood. This is a discrete objective, which is suited for a classification solution, not a regression solution. The mean square error, cross-entropy loss, and absolute error objectives are used for regression problems.

Option B is correct. We are solving a classification problem: predict if a given property is in a “high value” neighborhood. Therefore, we will want to optimize using discrete objects such as F1, precision, recall, or accuracy.

Option C is incorrect. This option describes using continuous objectives (mean square error, cross-entropy loss, or absolute error) to solve a classification problem: predict whether or not a district is "high value."

Option D is incorrect. This option describes using discrete objectives (F1, precision, recall, or accuracy) to solve a regression problem: predict the median household value for each district.

References:

Please see the Kaggle challenge titled **Ethical ML: California Housing Classification** (https://www.kaggle.com/c/ethicalml-cahousing/overview),

The Amazon SageMaker developer guide titled **Linear Learner Algorithm** (https://docs.aws.amazon.com/sagemaker/latest/dg/linear-learner.html)

Ask our Experts

View Queries

Question 40

Correct

Domain: Exploratory Data Analysis

You work as a machine learning specialist for a financial services firm that specializes in risk analysis for other financial services firms. Your machine learning team has been tasked with building a model that categorizes a firm's foreign exchange risk for each of their portfolios. You have begun building your model using SageMaker Studio, and you are at the point in your data exploration where you need to know the importance of each of the features in your training dataset. Which option gives you the most efficient view of this feature comparison?

A. Use the SageMaker Data Wrangler target leakage visualization to show the importance score of each feature in a bar chart.

B. Use the SageMaker Clarify bias visualization to show the importance score of each feature in a table.

C. Use the SageMaker Data Wrangler bias visualization to show the importance score of each feature in a scatter chart.

D. Use the SageMaker Data Wrangler quick model visualization to show the importance score of each feature in a bar chart. right

Explanation:

Correct Answer: D

Option A is incorrect. The SageMaker Data Wrangler target leakage visualization shows when there is data in a machine learning training dataset that is strongly correlated with the target

label. This visualization will not give you the importance score of each feature.

Option B is incorrect. The SageMaker Clarify bias visualization helps you identify bias during data preparation. This visualization will not give you the importance score of each feature.

Option C is incorrect. The SageMaker Data Wrangler bias visualization helps you uncover potential biases in your data. This visualization will not give you the importance score of each feature.

Option D is correct. The SageMaker Data Wrangler target leakage visualization helps you evaluate your data by producing importance scores for each feature in your dataset.

References:

Please see the Amazon SageMaker developer guide titled **Analyze and Visualize** (https://docs.aws.amazon.com/sagemaker/latest/dg/data-wrangler-analyses.html),

The Amazon SageMaker developer guide titled **Generate Reports for Bias in Pretraining Data in SageMaker Studio** (https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-data-bias-reports-ui.html)

Ask our Experts

View Queries

Question 41

Correct Marked for review

Domain: Modeling

You work as a machine learning specialist for an alternative transportation ride-share company. Your company has scooters, electric longboards, and other electric personal transportation devices in several major cities across the US. Your machine learning team has been asked to produce a machine learning model that classifies device preference by trip duration for each of the available personal transportation devices you offer in each city. You have created a model based on the SageMaker built-in K-Means algorithm. You are now using hyperparameter tuning to get the best-performing model for your problem. Which evaluation metrics and corresponding optimization direction should you choose for your automatic model tuning (a.k.a. hyperparameter tuning)? (Select TWO)

A. msd, maximize

B. mse, minimize

- [x] C. ssd , minimize right

D. f1, maximize

- [x] E. msd, minimize right

Explanation:

Correct Answers: C and E

Option A is incorrect. K-Means uses the msd (Mean Squared Distances) metric for model validation. However, you will want to minimize this metric.

Option B is incorrect. K-Means does not use the mse (Mean Squared Error) metric for model validation.

Option C is correct. K-Means uses the ssd (Sum of the Squared Distances) metric for model validation, and you will want to minimize this metric.

Option D is incorrect. K-Means does not use the f1 (weighted average of precision and recall) metric for model validation.

Option E is correct. K-Means uses the msd (Mean Squared Distances) metric for model validation, and you will want to minimize this metric.

References:

Please see the Amazon SageMaker developer guide titled **Define Metrics** (https://docs.aws.amazon.com/sagemaker/latest/dg/automatic-model-tuning-define-metrics.html),

The Amazon SageMaker developer guide titled **Tune a K-Means Model** (https://docs.aws.amazon.com/sagemaker/latest/dg/k-means-tuning.html)

Ask our Experts

View Queries

Question 42

Incorrect Marked for review

Domain: Data Engineering

You work as a machine learning specialist for a cruise ship company. Due to new health restrictions, your company needs to only book their cruise ships at 50% capacity across all of their cruise offerings. To maximize profitability, you have been asked to create a model that gathers streaming data from various data sources such as weather services, census data, gross national product for various countries, spending habits across various countries, etc. You will use this data to build a model that uses clusters of data to predict cruise allocation. You need to perform feature engineering, such as feature transformations, on your streaming data and then load it into your company's MongoDB database. What is the most efficient solution for your scenario?

A. Stream your data sources via Kinesis Data Firehose to your MongoDB database, using a Lambda function to perform feature transformations. right

B. Stream your data sources via Kinesis Data Streams to your MongoDB database, using Kinesis Data Analytics to perform feature transformations. wrong

C. Stream your data sources via Kinesis Data Analytics (using Apache Flink to perform feature transformations) to your MongoDB database.

D. Stream your data sources via Kinesis Data Firehose to your MongoDB database, using a Glue ETL job to perform feature transformations.

Explanation:

Correct Answer: A

Option A is correct. You can stream your data sources via Kinesis Data Firehose, use a Lambda function that you write to perform feature transformations, then stream the transformed data to an HTTP endpoint for the MongoDB third-party service provider.

Option B is incorrect. Kinesis Data Analytics cannot write directly to a MongoDB HTTP endpoint.

Option C is incorrect. Kinesis Data Analytics cannot write directly to a MongoDB HTTP endpoint.

Option D is incorrect. This option will be much less efficient than option A because the Glue ETL job will have to write to S3, then you would have to write a script to load the data into MongoDB.

References:

Please see the Amazon Kinesis Data Firehose developer guide titled **What Is Amazon Kinesis Data Firehose?** (https://docs.aws.amazon.com/firehose/latest/dev/what-is-this-service.html),

The Amazon Kinesis Data Firehose developer guide titled **Amazon Kinesis Data Firehose Data Transformation** (https://docs.aws.amazon.com/firehose/latest/dev/data-transformation.html),

The Amazon Kinesis Data Analytics developer guide titled **Kinesis Data Analytics for Apache Flink: How It Works** (https://docs.aws.amazon.com/kinesisanalytics/latest/java/how-it-works.html),

The Amazon Kinesis Data Firehose developer guide titled **Using Amazon Kinesis Data Analytics** (https://docs.aws.amazon.com/firehose/latest/dev/data-analysis.html)

Ask our Experts

View Queries

Question 43

Correct

Domain: ML Implementation and Operations

You work as a machine learning specialist for a security firm that requires you to encrypt all of your machine learning infrastructure in transit and at rest. Your team is building a fraud detection algorithm using the Random Cut Forest SageMaker built-in algorithm. You and your teammates are using SageMaker notebook instances to build your model components. You need to customize the operating system of your notebook instances by installing custom libraries and setting specific operating system level configurations to meet your firm's security requirements. Your Chief Financial Officer wants to

keep the cost of running your SageMaker instances as low as possible. Therefore, you are required to manage the runtime of your SageMaker notebook instances, only having them running when they are actively in use. How can you meet your requirements most efficiently?

A. Stop your notebook instances at the end of each day and start them up again at the beginning of the next workday. Your customizations to the operating system will be maintained across the stopping and starting of your instances.

B. Keep your SageMaker notebook instances running until your team has completed building your Random Cut Forest model.

C. Use a lifecycle configuration to automate customizations of your notebook instances, stopping the instances at the end of each workday and starting the instances at the beginning of each workday. right

D. Terminate the SageMaker notebook instances at the end of each workday and recreate them at the start of each workday.

Explanation:

Correct Answer: C

Option A is incorrect. When you stop your SageMaker notebook instances, customizations to the operating system, such as installed custom libraries or operating system level settings, are lost.

Option B is incorrect. Keeping your SageMaker notebook instances running until your team has completed building your Random Cut Forest model will not meet your chief financial officer's requirement of keeping the cost of running your SageMaker instances as low as possible.

Option C is correct. To avoid losing the installation of custom libraries and setting specific operating system level configurations, you can use a lifecycle configuration to automate customizations of your notebook instances.

Option D is incorrect. When you terminate your SageMaker notebook instances, the snapshot and the ML storage volume are deleted, thereby deleting your installation of custom libraries and setting specific operating system level configurations.

References:

Please see the Amazon SageMaker developer guide titled **Random Cut Forest (RCF) Algorithm** (https://docs.aws.amazon.com/sagemaker/latest/dg/randomcutforest.html),

The Amazon SageMaker developer guide titled **Notebook instances and SageMaker jobs** (https://docs.aws.amazon.com/sagemaker/latest/dg/encryption-at-rest-nbi.html)

Ask our Experts

View Queries

Question 44

Incorrect

Domain: Exploratory Data Analysis

You work as a machine learning specialist for an online flight booking service that finds the lowest cost flights based on user input such as flight dates, origin, destination, number of layovers, and other factors. Your machine learning team gathers data from many sources, including airline flight databases, credit agencies, etc., to use in your model. You need to transform this data for your model training and in real-time for your model inference requests. What is the most efficient way to build these transformations into your model workflow?

- A. Use Apache Spark ML Streaming to deploy an inference pipeline that reuses the data transforms you developed for your training model. wrong
- B. Use Apache Spark ML Serving to deploy an inference pipeline that reuses the data transforms you developed for your training model. right
- C. Use Apache Spark MLlib to deploy an inference pipeline that reuses the data transforms you developed for your training model.
- D. Use a SageMaker lifecycle configuration to automate the reuse of the data transforms you developed for your training model.

Explanation:

Correct Answer: B

Option A is incorrect. Apache Spark Streaming supports real-time processing of streaming data. Apache Spark Streaming cannot be used to reuse the data transforms you developed for your training model in your inference requests.

Option B is correct. You can use SageMaker Spark ML Serving containers that reuse the data transformers developed for training models in your inference requests.

Option C is incorrect. Apache Spark MLlib is a machine learning library that runs algorithms designed to scale across clusters for classification, regression, clustering, collaborative filtering. Apache Spark MLlib cannot be used to reuse the data transforms you developed for your training model in your inference requests.

Option D is incorrect. SageMaker lifecycle configurations are used to install packages or sample notebooks on your notebook instances, configure notebook instance networking and security, or use a shell script to customize notebook instances. SageMaker lifecycle configurations cannot be used to reuse the data transforms you developed for your training model in your inference requests.

References:

Please see the Amazon Machine Learning Lens AWS Well-Architected Framework titled **Feature Engineering** (https://docs.aws.amazon.com/wellarchitected/latest/machine-learning-lens/feature-engineering.html),

The Amazon SageMaker developer guide titled **Deploy an Inference Pipeline** (https://docs.aws.amazon.com/sagemaker/latest/dg/inference-pipelines.html),

The Amazon SageMaker developer guide titled **Use SparkML Serving with Amazon SageMaker** (https://docs.aws.amazon.com/sagemaker/latest/dg/sparkml-serving.html),

The Amazon SageMaker developer guide titled **Customize a Notebook Instance Using a Lifecycle Configuration Script** (https://docs.aws.amazon.com/sagemaker/latest/dg/notebook-lifecycle-config.html),

The Apache Spark page titled **Streaming** (https://spark.apache.org/streaming/),

The Apache Spark page titled **MLlib** (https://spark.apache.org/mllib/)

Ask our Experts

View Queries

Question 45

Correct

Domain: Modeling

You work as a machine learning specialist for a computer hardware component producer. Your company produces individual components, such as processor chips, GPUs, etc. and assembled computer peripherals such as monitors, external disk drives, among others. You and your team have been tasked with building a machine learning model that predicts future product sales to improve supply chain management based on data from your semiconductor, transistor, and other base component suppliers as well as data from your sales department. After training your model, you now need to evaluate it to determine whether its performance and accuracy will allow you to use it to predict future product sales accurately. You have decided to perform an offline model evaluation of your model using your historical data. You have split your validation dataset into 10 parts. You then execute 10 training runs, which produces 10 models. You then aggregate the 10 models to get your final evaluation. Which model evaluation method are you using?

A. Holdout set validation

B. K-fold cross-validation right

C. Bayesian validation

D. Hierarchical validation

Explanation:

Correct Answer: B

Option A is incorrect. Holdout validation uses one validation dataset to use in model validation. Using this method, you evaluate your model using this one holdout set. But you have split your

validation dataset into 10 parts. So you will have 10 training runs in your validation process.

Option B is correct. You are using the k-fold cross-validation technique. Using this validation method, you split the example dataset into k parts, in your case, k = 10. You then run 10 training using each of the 10 example datasets. Finally, using this method, you aggregate the 10 run results to get your final evaluation.

Option C is incorrect. There is no bayesian validation method.

Option D is incorrect. There is no hierarchical validation method.

References:

Please see the Amazon SageMaker developer guide titled **Validate a Machine Learning Model** (https://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-model-validation.html),

The Towards Data Science article titled **Validating your Machine Learning Model** (https://towardsdatascience.com/validating-your-machine-learning-model-25b4c8643fb7)

Ask our Experts

View Queries

Question 46

Incorrect Marked for review

Domain: Data Engineering

You work as a machine learning specialist for a research department at a large university. Your team of machine learning specialists is responsible for all aspects of the machine learning lifecycle, including creating the data repositories used by your research scientists for their data science work. Your team has built a SageMaker infrastructure for your data scientists where you stream in data from many sources, such as satellite feeds, IoT devices like underwater sensors, and many others. You have recently implemented SageMaker Feature Store, and you are now implementing the ingestion of batch data from your streaming data sources. Which of the following options are viable approaches to streaming data into your SageMaker Feature Store? (Select TWO)

A. Stream your data sources through Apache Kafka into Feature Store. right

B. Stream your data sources through Kinesis Data Analytics and a Lambda function into Feature Store. right

☐ C. Stream your data sources through Apache Spark Streaming into Feature Store. wrong

D. Stream your data sources through Apache Spark ML Serving into Feature Store.

☐ E. Stream your data sources through Apache Flink into Feature Store. wrong

Explanation:

Correct Answers: A and B

Option A is correct. Apache Kafka can be used as a streaming data source where features are directly fed to the online feature store for feature creation.

Option B is correct. Kinesis Data Analytics, together with the use of a Lambda function, can be used as a streaming data source where features are directly fed to the online feature store for feature creation.

Option C is incorrect. Apache Spark Streaming is not supported as a direct streaming feed into SageMaker Feature Store.

Option D is incorrect. Apache Spark ML Serving is not supported as a direct streaming feed into SageMaker Feature Store.

Option E is incorrect. Apache Flink is not supported as a direct streaming feed into SageMaker Feature Store.

References:

Please see the Amazon SageMaker developer guide titled **Data Sources and Ingestion** (https://docs.aws.amazon.com/sagemaker/latest/dg/feature-store-ingest-data.html),

The AWS Machine Learning blog titled **Using streaming ingestion with Amazon SageMaker Feature Store to make ML-backed decisions in near-real time** (https://aws.amazon.com/blogs/machine-learning/using-streaming-ingestion-with-amazon-sagemaker-feature-store-to-make-ml-backed-decisions-in-near-real-time/)

Ask our Experts

View Queries

Question 47

Correct Marked for review

Domain: Modeling

You work as a machine learning specialist for a sports gambling website. Your machine learning team has been asked to create a football score prediction model that predicts the winner of a match, the score difference, and the shots-on-goal differential. You have collected historical football match data, and you have selected the SageMaker XGBoost built-in algorithm to use for your model. You are now ready to train your model using a SageMaker training job. Which of the following are NOT used by your SageMaker training job? (Select TWO)

A. URL of the S3 bucket where you have stored your training data

B. SageMaker managed ML compute instances that SageMaker will use for model training

- [ ] C. ECS path where the training code is stored right
- [ ] D. URL of the S3 bucket where you have stored your training code right

E. ECR path where the training code is stored

Explanation:

Correct Answers: C and D

Option A is incorrect. You must supply the URL of the S3 bucket where you have stored your training data to run your training job.

Option B is incorrect. You must supply the ML compute instances that SageMaker will use for model training.

Option C is correct. Your training code is referenced by an ECR path to your training code, not an ECS path.

Option D is correct. Your training code is referenced by an ECR path to your training code, not the URL of an S3 bucket.

Option E is incorrect. You must supply the ECR path where the training code is stored for your training job to use.

References:

Please see the Amazon SageMaker developer guide titled **Train a Model with Amazon SageMaker** (https://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-training.html),

The Amazon SageMaker developer guide titled **Docker Registry Paths for SageMaker Built-in Algorithms** (https://docs.aws.amazon.com/sagemaker/latest/dg/sagemaker-algo-docker-registry-paths.html)

Ask our Experts

View Queries

Question 48

Correct

Domain: ML Implementation and Operations

You work as a machine learning specialist for a mobile phone carrier network. [1] [2] Your team of machine learning specialists needs to produce a model that clusters network users by payment plan and real-time geographic region. Your management team wants to use the model to consider marketing offerings targeted to a customer's billing plan and geographic region in which they spend the most time. You have created a k-means model, and you need to test variations of the model on real-time customer data. Which option is considered the best practice for testing model variations as described by your use case?

A. Deploy the multiple variants of your model to the same SageMaker HTTPS endpoint, directing a percentage of traffic to each variant of the model, using an endpoint configuration that describes all variants of the model. right

B. Deploy the multiple variants of your model to multiple SageMaker HTTPS endpoints, directing a percentage of traffic to each variant of the model, using an endpoint configuration for each variant of the model that describes the model.

C. Deploy the multiple variants of your model to the same SageMaker HTTPS endpoint, directing a percentage of traffic to each variant of the model, using an endpoint configuration for each variant of the model that describes the model.

D. Deploy the multiple variants of your model to multiple SageMaker HTTPS endpoints, directing a percentage of traffic to each variant of the model, using an endpoint configuration that describes all variants of the model.

Explanation:

Correct Answer: A

Option A is correct. The best practice for your use case is to deploy the multiple variants of your model to a single SageMaker HTTPS endpoint. Where you can direct traffic to each variant as you see fit. You accomplish this by creating an endpoint configuration that describes all variants of the model.

Option B is incorrect. The best practice is to deploy the multiple variants of your model to a single SageMaker HTTPS endpoint, not multiple endpoints. Also, the best practice is to one endpoint configuration that describes all variants of the model, not multiple endpoint configurations.

Option C is incorrect. The best practice is to one endpoint configuration that describes all model variants, not multiple endpoint configurations.

Option D is incorrect. The best practice is to deploy the multiple variants of your model to a single SageMaker HTTPS endpoint, not multiple endpoints.

References:

Please see the Amazon SageMaker developer guide titled **Deploy a Model in Amazon SageMaker** (https://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-deployment.html#how-it-works-hosting),

The Amazon SageMaker developer guide titled **Step 5: Deploy the Model to Amazon EC2** (https://docs.aws.amazon.com/sagemaker/latest/dg/ex1-model-deployment.html#ex1-deploy-model)

Ask our Experts

View Queries

Question 49 Correct

Domain: Modeling

You work as a machine learning specialist for a city government agency in their urban housing department. You have been assigned the task of using a machine learning model to find the best housing location to place new public housing applicants. You have been asked to propose housing sites for new applicants based on the similarity of the applicant (such as applicant work location, number of people in the family group, applicant income range, etc.) to the other housing residents in the city. You have decided to use the SageMaker k-nearest neighbors built-in algorithm. You have produced a model variant and deployed it to an HTTPS endpoint. Based on your initial evaluation results, you would like to change the SageMaker endpoint by updating the ML compute instances of the existing variant to make them more powerful and add a new model variant. What is the best way to implement these changes?

A. Take your existing model variant out of service, make the changes to your endpoint configuration to change the ML instances, add the second model variant, and then bring your SageMaker HTTPS endpoint back into service.

B. Modify your existing model variant to change the ML instance types. Then, once you have your newly configured model variant performing appropriately, take your existing model variant out of service. Modify your endpoint configuration to add the second model variant and bring your SageMaker HTTPS endpoint back into service.

C. Create a new endpoint configuration that has the desired ML instance types and both model variants. Take your old endpoint configuration out of service. Deploy your new endpoint configuration into production.

○ D. Modify your SageMaker HTTPS endpoint without taking the model that is already deployed into production out of service. Change the existing model variant's ML instance type and add the new model variant. Do this by creating a new endpoint configuration and deploying the new endpoint configuration with the SageMaker UpdateEndpoint action. right

Explanation:

Correct Answer: D

Option A is incorrect. You don't need to disrupt your existing model variant serviced by your SageMaker HTTPS endpoint. You can deploy your new model variant and change the existing model variant's ML instances while keeping your endpoint in service.

Option B is incorrect. You don't have to implement your changes in piecemeal as described in this option. You don't need to disrupt your existing model variant serviced by your SageMaker HTTPS endpoint. You can deploy your new model variant and change the existing model variant's ML instances while keeping your endpoint in service.

Option C is incorrect. You will need to create a new endpoint configuration. However, you don't need to take your old endpoint configuration out of service before deploying your new endpoint configuration.

Option D is correct. To keep from disrupting your SageMaker HTTPS endpoint service, you can modify your SageMaker HTTPS endpoint without taking the model that is already deployed into production out of service.

References:

Please see the Amazon SageMaker developer guide titled **Deploy a Model in Amazon SageMaker** (https://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-deployment.html#how-it-works-hosting),

The Amazon SageMaker reference titled **UpdateEndpoint** (https://docs.aws.amazon.com/sagemaker/latest/APIReference/API_UpdateEndpoint.html),

The Statistics How To article titled **k-NN (k-Nearest Neighbor): Overview, Simple Example** (https://www.statisticshowto.com/k-nn-k-nearest-neighbor/)

Ask our Experts

View Queries

Question 50

Incorrect

Domain: Exploratory Data Analysis

You work as a machine learning specialist for a mobile phone operator where you need to build a machine learning model that predicts when a given customer is about to leave your phone service or churn. The inference data produced by your model will allow your marketing department to offer incentives to the customer to get them to stay with your service. Using data generated by customer activity with your service offering, you need to visualize the inference data in a dashboard. So your marketing department can quickly decide which customer churn candidates to offer additional incentives. How can you get your machine learning inference data into your dashboard visualization in the most efficient, performant manner?

- A. As your inference engine produces potential churn candidate data, write the data to S3. Use Athena to query the data and associate a QuickSight visualization data source with your Athena query results. wrong
- B. Create a JSON schema file that contains the metadata that QuickSight needs to process your model data, then use the Augment with SageMaker feature of QuickSight to visualize your customer churn data. right
- C. As your inference engine produces potential churn candidate data, write the data to S3. Use S3 Analytics to query the data and associate a QuickSight visualization data source with your S3 Analytics query results.
- D. As your inference engine produces potential churn candidate data, write the data to S3. Use Redshift Spectrum to query the data and associate a QuickSight visualization data source with your Redshift Spectrum query results.

Explanation:

Correct Answer: B

Option A is incorrect. Although you could use Athena and S3 to visualize your data as this option describes, it is not as efficient as using the Augment with SageMaker feature of QuickSight.

Option B is correct. You can use the Augment with SageMaker feature of QuickSight to integrate your SageMaker inference data into your QuickSight visualization. This is by far the most efficient option given.

Option C is incorrect. Although you could use S3 Analytics and S3 to visualize your data, as this option describes, it is not as efficient as using the Augment with SageMaker feature of QuickSight.

Option D is incorrect. Although you could use Redshift Spectrum and S3 to visualize your data, as this option describes, it is not as efficient as using the Augment with SageMaker feature of QuickSight. You would have to create a Redshift cluster for this option to work.

References:

Please see the AWS Machine Learning blog titled **Visualizing Amazon SageMaker machine learning predictions with Amazon QuickSight** (https://aws.amazon.com/blogs/machine-learning/making-machine-learning-predictions-in-amazon-quicksight-and-amazon-sagemaker/),

The AWS Samples GitHub repo titled **ML Predictions using Amazon QuickSight and Amazon SageMaker** (https://github.com/aws-samples/quicksight-sagemaker-integration-blog)

Ask our Experts

View Queries

Question 51

Correct

Domain: ML Implementation and Operations

You work as a machine learning specialist for a social media software company. Your company produces social media apps such as interactive games and photo-sharing communities. Your machine learning team has created a machine learning model that produces recommendations via advertising in your apps, such as showing advertising for skiing trips to a user who follows a ski resort in the photo-sharing app. The production variant that your team has deployed experiences very wild swings in traffic volume over the course of any given day. Also, since the app is relatively new to the mobile community, it receives no traffic for some time periods. You have set up SageMaker automatic scaling policy for your production variant instances. However, you have noticed that scaling-in does not happen when your traffic reduces to nothing for a period of time. Why might this happen?

- A. Scaling-in does not happen when there is no traffic. If traffic to your production variant becomes zero, SageMaker automatic scaling won't scale in. SageMaker doesn't emit metrics with a value of zero, so no CloudWatch events are triggered. right

B. You are using a Step Scaling policy, and you have set the number of instances to deploy to a too high level.

C. You are using a Step Scaling policy, and you have set the number of instances to deploy to a too low level.

D. You are using a Target Tracking policy and you have set the target value for your metric too low.

Explanation:

Correct Answer: A

Option A is correct. When your production variant doesn't receive any traffic, SageMaker does not emit any metrics. Therefore there is no target metric for CloudWatch to use as a trigger to initiate the scale as defined in your scaling policy.

Option B is incorrect. Even if you used a step scaling policy and set the number of instances to a high number, the policy won't be triggered by CloudWatch because SageMaker does not emit any metrics when traffic is 0 to your production variant.

Option C is incorrect. Even if you used a step scaling policy and set the number of instances to a low number, the policy won't be triggered by CloudWatch because SageMaker does not emit any metrics when traffic is 0 to your production variant.

Option D is incorrect. Even if you used a target tracking scaling policy and set the target value for your metric to a low value, the policy won't be triggered by CloudWatch because SageMaker does not emit any metrics when traffic is 0 to your production variant.

References:

Please see the AWS SageMaker developer guide titled **Prerequisites** (https://docs.aws.amazon.com/sagemaker/latest/dg/endpoint-auto-scaling-prerequisites.html),

The Application Auto Scaling user guide titled **Target tracking scaling policies for Application Auto Scaling** (https://docs.aws.amazon.com/autoscaling/application/userguide/application-auto-scaling-target-tracking.html),

The Application Auto Scaling user guide titled **Step scaling policies for Application Auto Scaling** (https://docs.aws.amazon.com/autoscaling/application/userguide/application-auto-scaling-step-scaling-policies.html)

Ask our Experts

View Queries

Question 52

Correct

Domain: Modeling

You work as a machine learning specialist for a book publishing firm. Your firm is releasing a new publication and would like to use a machine learning model to structure a marketing campaign for the new publication to decide whether to market to each of their registered customers or not. You and your machine learning team have developed a model using the XGBoost SageMaker built-in algorithm. You are now at the hyperparameter optimization stage, where you are trying to find the best version of your model by running several training jobs on your data using your XGBoost algorithm. How do you configure your hyperparameter tuning jobs to get a recommendation for the best values for your hyperparameters?

A. Set the eta, alpha, and min_child_weight to specific values, and set the max_depth to a range of values. Choose to minimize the area under the curve (auc) as your optimization metric.

B. Set ranges of values for the eta, alpha, and min_child_weight, and max_depth hyperparameters. Choose to maximize the area under the curve (auc) as your optimization metric. right

C. Set ranges of values for the eta, alpha, and min_child_weight, and max_depth hyperparameters. Choose to minimize the normalized discounted cumulative gain (ndcg) as your optimization metric.

D. Set ranges of values for the eta, alpha, and min_child_weight, and max_depth hyperparameters. Launch one training job. Choose to maximize the area under the curve (auc) as your optimization metric.

Explanation:

Correct Answer: B

Option A is incorrect. You do not want to restrict your hyperparameter tuning job by setting any of your tunable hyperparameters to specific values. Also, you will want to maximize the auc evaluation metric, not minimize it.

Option B is correct. Setting the values of your tunable hyperparameters to ranges of values allows your hyperparameter tuning jobs to use either bayesian or random search to find the best combination of values. Also, maximizing the auc optimization metric is a proven approach to reaching the optimal set of tunable hyperparameters.

Option C is incorrect. Setting the values of your tunable hyperparameters to ranges of values allows your hyperparameter tuning jobs to use either bayesian or random search to find the best combination of values. However, choosing to minimize ndcg as your optimization metric will result in a suboptimal result. The ndcg metric is supposed to be maximized, not minimized.

Option D is incorrect. Setting the values of your tunable hyperparameters to ranges of values allows your hyperparameter tuning jobs to use either bayesian or random search to find the best combination of values. However, running only one hyperparameter tuning job will not give you the optimal result. You need to run several training jobs to get to the best set of tunable hyperparameters in a reasonable amount of time.

References:

Please see the AWS SageMaker developer guide titled **Perform Automatic Model Tuning** (https://docs.aws.amazon.com/sagemaker/latest/dg/automatic-model-tuning.html),

The AWS SageMaker developer guide titled **Tune an XGBoost Model** (https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost-tuning.html),

The AWS SageMaker developer guide titled **How Hyperparameter Tuning Works** (https://docs.aws.amazon.com/sagemaker/latest/dg/automatic-model-tuning-how-it-works.html)

Ask our Experts

View Queries

Question 53

Correct

Domain: Data Engineering

You work as a machine learning specialist for the infectious disease testing department of a national government agency. Your machine learning team is responsible for creating a machine learning model that analyzes the daily test datasets for your country and produces daily predictions of trends of disease contraction and death rates. These projections are used throughout national and international news agencies to report on the daily projections of infectious disease progression. Since your model works on huge datasets on a daily basis, which of the following statements gives an accurate description of your inference processing?

A. You have set up a persistent endpoint to get predictions from your model using SageMaker batch transform.

B. You have set up a persistent endpoint to get predictions from your model using SageMaker hosting services.

C. You don't need a persistent endpoint. You use SageMaker batch transform to get inferences from your large datasets. right

D. You don't need a persistent endpoint. You use SageMaker hosting services to get inferences from your large datasets.

Explanation:

Correct Answer: C

Option A is incorrect. SageMaker batch transform does not use a persistent endpoint. You use SageMaker batch transform to get inferences from large datasets. Also, your process runs one per day, so a persistent endpoint does not make sense.

Option B is incorrect. You are processing large datasets on a daily basis. Therefore you should use SageMaker batch transform, not SageMaker hosting services. SageMaker hosting services are

used for real-time inference requests, not daily batch requests.

Option C is correct. Since you are using your endpoint to get inferences one per day from a large dataset, you don't need a persistent endpoint. Also, SageMaker batch transform is the best deployment option when getting inferences from an entire dataset.

Option D is incorrect. SageMaker hosting services need a persistent endpoint. Also, since you are processing large datasets on a daily basis, you should use SageMaker batch transform, not SageMaker hosting services.

References:

Please see the AWS SageMaker developer guide titled **Use Batch Transform** (https://docs.aws.amazon.com/sagemaker/latest/dg/batch-transform.html),

The AWS SageMaker developer guide titled **Deploy Models for Inference** (https://docs.aws.amazon.com/sagemaker/latest/dg/deploy-model.html)

Ask our Experts

View Queries

Question 54

Correct

Domain: Exploratory Data Analysis

You work as a machine learning specialist for a polling organization using US census data to predict whether a given polling respondent earns greater than $75,000. Your company will then sell the polling prediction data to candidates running for various political office positions across the country. You need to clean the polling data on which you wish to train your binary classification model. Specifically, you need to remove duplicate rows with erroneous data, transform the income column into a label column with two values, transform the age column to a categorical feature by binning the column, scale the capital gain and capital losses columns, and finally split the data into train and test datasets. Which of the options are the most efficient ways to achieve your data sanitizing and feature preparation? (Select TWO)

A. Create a SageMaker Processing job using a SageMaker Scala SDK with Processing container leveraging the pandas PDLearnProcessor package that performs your required preprocessing sanitizing and feature preparation tasks and then splits the data into the training and test datasets.

- [x] B. Create a SageMaker Processing job using a SageMaker Python SDK with Processing container leveraging the scikit-learn SKLearnProcessor package that performs your required preprocessing sanitizing and feature preparation tasks and then splits the data into the training and test datasets. right

C. Create a SageMaker Processing job using a SageMaker Python SDK with Data Wrangler container leveraging the scikit-learn SKLearnProcessor package that performs your required

preprocessing sanitizing and feature preparation tasks and then splits the data into the training and test datasets.

- [x] D. Create a SageMaker Processing job using a SageMaker Python SDK with Processing container leveraging the Spark PySparkProcessor package that performs your required preprocessing sanitizing and feature preparation tasks and then splits the data into the training and test datasets. right

E. Create a SageMaker Processing job using a SageMaker Python SDK with Processing container leveraging the SparkMLProcessor package that performs your required preprocessing sanitizing and feature preparation tasks and then splits the data into the training and test datasets.

Explanation:

Correct Answers: B and D

Option A is incorrect. There is no SageMaker Scala SDK. Also, there is no pandas PDLearnProcessing package. SageMaker Processing jobs can be written in Python, using the SageMaker Python SDK. You can leverage either the PySparkProcessor, SparkJarProcessor, or the SKLearnProcessor package to perform your preprocessing sanitizing and feature preparation tasks and also split your data into the training and test datasets.

Option B is correct. SageMaker Processing jobs can be written in Python, using the SageMaker Python SDK. You can leverage either the PySparkProcessor, SparkJarProcessor, or the SKLearnProcessor package to perform your preprocessing sanitizing and feature preparation tasks and also split your data into the training and test datasets.

Option C is incorrect. There is no Data Wrangler container in the SageMaker Processing Job containers.

Option D is correct. SageMaker Processing jobs can be written in Python, using the SageMaker Python SDK. You can leverage either the PySparkProcessor, SparkJarProcessor, or the SKLearnProcessor package to perform your preprocessing sanitizing and feature preparation tasks and also split your data into the training and test datasets.

Option E is incorrect. There is no SparkMLProcessor package in the SageMaker Processing service.

References:

Please see the AWS SageMaker developer guide titled **Data Processing with Apache Spark** (https://docs.aws.amazon.com/sagemaker/latest/dg/use-spark-processing-container.html),

The AWS Examples GitHub repository titled **Amazon SageMaker Processing jobs** (https://github.com/aws/amazon-sagemaker-examples/blob/master/sagemaker_processing/scikit_learn_data_processing_and_model_evaluation/scikit_learn_data_processing_and_model_evaluation.ipynb)

Ask our Experts

View Queries

Question 55 Correct

Domain: Modeling

You work as a machine learning specialist for a start-up software company that builds a mobile app that subscribers can use to identify various types of birds from pictures they take with their phone camera. You have a large set of unlabeled images of birds that you want to use as your training data for your image recognition application. Which option is the most efficient approach to creating a labeling job to build the training dataset for your mobile app?

A. Use the SageMaker k-means built-in algorithm to label your unlabeled images. Leverage a SageMaker Semantic Segmentation algorithm-based model to perform auto-annotation of your images.

B. Use SageMaker Data Wrangler to label your unlabeled images. Leverage a SageMaker Image Classification algorithm-based model to perform auto-annotation of your images.

C. Use SageMaker Ground Truth to label your unlabeled images, leveraging lambda functions to perform annotation consolidation and pre-labeling. Leverage a SageMaker Image Classification algorithm-based model to perform auto-annotation of your images. right

D. Use SageMaker Ground Truth to label your unlabeled images, leveraging Glue ETL jobs to perform annotation consolidation and pre-labeling. Leverage a SageMaker Object Detection algorithm-based model to perform auto-annotation of your images.

Explanation:

Correct Answer: C

Option A is incorrect. Using the k-means algorithm to label your unlabeled images would be far less efficient than using SageMaker Ground Truth. Also, the SageMaker Semantic Segmentation algorithm is not an efficient algorithm to use to auto-annotate your images.

Option B is incorrect. Using the Data Wrangler to label your unlabeled images would be far less efficient than using SageMaker Ground Truth.

Option C is correct. SageMaker Ground Truth is the preferred method of labeling unlabeled image data. Also, using lambda functions in your labeling job allows you to automate the annotation consolidation and pre-labeling tasks. Finally, the SageMaker Image Classification built-in algorithm is the best choice for the auto-annotation task.

Option D is incorrect. Glue ETL jobs cannot perform your annotation consolidation and pre-labeling tasks as efficiently as using lambda functions for these tasks in your labeling job.

References:

Please see the AWS SageMaker developer guide titled **Image Classification Algorithm** (https://docs.aws.amazon.com/sagemaker/latest/dg/image-classification.html),

The AWS Examples GitHub repository titled **From Unlabeled Data to a Deployed Machine Learning Model: A SageMaker Ground Truth Demonstration for Image Classification** (https://github.com/aws/amazon-sagemaker-examples/blob/master/ground_truth_labeling_jobs/from_unlabeled_data_to_deployed_machine_learning_model_ground_truth_demo_image_classification/from_unlabeled_data_to_deployed_machine_learning_model_ground_truth_demo_image_classification.ipynb),

The AWS SageMaker developer guide titled **Prepare ML Data with Amazon SageMaker Data Wrangler** (https://docs.aws.amazon.com/sagemaker/latest/dg/data-wrangler.html),

The AWS SageMaker developer guide titled **Object Detection Algorithm** (https://docs.aws.amazon.com/sagemaker/latest/dg/object-detection.html)

Ask our Experts

View Queries

Question 56 Incorrect Marked for review

Domain: Data Engineering

You work as a machine learning specialist for a home automation company that produces home automation devices such as automated door locks, security cameras, and alarm systems. Your machine learning team is building a new data repository for a device that your company will soon launch as a new product. The new device will generate large streams of IoT data using the MQTT protocol. You need to create a data repository for use in your machine learning models that will be used to produce future device usage predictions. Your management team will use these device usage predictions to inform their marketing campaigns. Which option is the most cost-effective configuration of AWS services to build your data repository? (Select TWO)

A. Receive the MQTT messages using a Kinesis Producer Library (KPL) application to transform your MQTT message data near real-time and write the MQTT messages to a Kinesis Data Streams partition. Configure your Kinesis Data Stream to write your transformed MQTT data directly to your S3 bucket that will be the data source of your machine learning model.

B. Receive your MQTT messages into IoT Core. Use an Apache Kafka IoT Core action to send your MQTT messages directly to Amazon Managed Streaming for Apache Kafka (Amazon MSK) to transform your MQTT data in near real-time. Then use an Msk-to-s3-feeder application to write your transformed MQTT data directly to your S3 bucket that will be the data source of your machine learning model. right

C. Receive your MQTT messages into IoT Core. Use a Kinesis Data Firehose IoT Core action to send your MQTT messages directly to a Kinesis Data Firehose to transform your MQTT data near real-time and write your transformed MQTT data directly to your S3 bucket that will be the data source of your machine learning model. right

- [ ] D. Receive your MQTT messages into IoT Core. Use an S3 IoT Core action to transform your MQTT message data and send the transformed MQTT messages directly to your S3 bucket that will be the data source of your machine learning model. wrong

E. Receive your MQTT messages into IoT Core. Use an HTTPS IoT Core action to transform your MQTT message data and send the transformed MQTT messages directly to your S3 bucket that will be the data source of your machine learning model.

Explanation:

Correct Answers: B and C

Option A is incorrect. You can use a Kinesis Data Streams producer application to transform your MQTT messages and then send the transformed MQTT messages to a Kinesis Data Stream. You cannot, however, write your MQTT message data directly to S3 from Kinesis Data Streams. You would need to create a Kinesis Data Streams consumer application to retrieve the transformed MQTT messages from the Kinesis Data Streams partition and then write the transformed MQTT messages to your S3 bucket. So this option does not include all of the work you would need to create your data repository.

Option B is correct. The Apache Kafka IoT Core action allows you to send your MQTT message data directly to Amazon MSK to transform your MQTT data in near real-time. Amazon MSK has an Msk-to-s3-feeder application that you can use to write your transformed MQTT data directly to your S3 bucket.

Option C is correct. The Kinesis Data Firehose IoT Core action allows you to use Kinesis Data Firehose to transform your MQTT message data in near real-time. Kinesis Data Firehose can then write your transformed MQTT data directly to your S3 bucket.

Option D is incorrect. You can use an S3 IoT Core action to write your MQTT messages directly to an S3 bucket. However, you cannot transform your messages without writing a lambda function (as an example, but there are other ways to do this) to trigger on the S3 create action to transform your MQTT message data. So this option does not include all of the work you would need to create your data repository.

Option E is incorrect. You can use an HTTPS IoT Core action to send your MQTT messages to an HTTPS endpoint. However, you cannot transform your messages using your HTTPS endpoint without writing an HTTPS service to process your MQTT message data and write it to S3. So this option does not include all of the work you would need to create your data repository.

References:

Please see the Amazon Kinesis Data Streams developer guide titled **Kinesis Data Streams Consumers** (https://docs.aws.amazon.com/streams/latest/dev/amazon-kinesis-consumers.html),

The AWS announcement titled **AWS IoT Core adds the ability to deliver data to Apache Kafka clusters** (https://aws.amazon.com/about-aws/whats-new/2020/12/aws-iot-core-adds-the-ability-to-deliver-data-to-apache-kafka-clusters/),

The AWS IoT Core developer guide titled **Apache Kafka** (https://docs.aws.amazon.com/iot/latest/developerguide/apache-kafka-rule-action.html),

The AWS IoT Core developer guide titled **Kinesis Data Firehose** (https://docs.aws.amazon.com/iot/latest/developerguide/kinesis-firehose-rule-action.html),

The AWS IoT Core developer guide titled **HTTPS** (https://docs.aws.amazon.com/iot/latest/developerguide/https-rule-action.html),

The AWS IoT Core developer guide titled **What is AWS IoT?** (https://docs.aws.amazon.com/iot/latest/developerguide/what-is-aws-iot.html),

The AWS IoT Core developer guide titled **Connecting to AWS IoT Core** (https://docs.aws.amazon.com/iot/latest/developerguide/connect-to-iot.html),

The Amazon Managed Streaming for Apache Kafka developer guide titled **What Is Amazon MSK?** (https://docs.aws.amazon.com/msk/latest/developerguide/what-is-msk.html)

Ask our Experts

View Queries

Question 57

Correct

Domain: Exploratory Data Analysis

You work as a machine learning specialist for a software company that offers real-time interactive sports viewing app for mobile phones and tablets. You gather real-time streaming sports statistics and game action data and use the streaming data to produce real-time analytics and active predictions of the likely outcome of the game. To produce your prediction, you need to use several machine learning models that use the real-time streaming data as their training and inference data sources. Since the real-time streaming game data is delivered from several different sources, the format and schema of the data need transformation and sanitation. Which option is the most efficient way to perform the feature engineering of your real-time streaming data for use in your training and inference requests?

A. Ingest the real-time streaming data using Kinesis Data Firehose using the kinesis-firehose-process-record Lambda blueprint for transformation. Stream the output of your Kinesis Data Firehose into SageMaker offline feature store FeatureGroup.

B. Ingest the real-time streaming data using Kinesis Data Firehose using the kinesis-process-record Lambda blueprint for transformation. Stream the output of your Kinesis Data Firehose into SageMaker offline feature store FeatureGroup.

C. Ingest the real-time streaming data using Kinesis Data Firehose using the kinesis-firehose-process-record Lambda blueprint for transformation. Stream the output of your Kinesis Data Firehose into SageMaker offline and online feature store FeatureGroups. right

D. Ingest the real-time streaming data using Kafka using the kinesis-process-record Lambda blueprint for transformation. Stream the output of your Kinesis Data Firehose into SageMaker

online feature store FeatureGroup.

Explanation:

Correct Answer: C

Option A is incorrect. You can ingest your streaming data using Kinesis Data Firehose and use the kinesis-firehose-process-record Lambda blueprint for transformation. However, you need to stream the output of your Kinesis Data Firehose into both the offline and online feature store FeatureGroups since you wish to train using your Feature Store groups and produce real-time inferences using your online Feature Store FeatureGroup.

Option B is incorrect. You can ingest your streaming data using Kinesis Data Firehose, and you could use the kinesis-process-record Lambda blueprint for transformation. However, you need to stream the output of your Kinesis Data Firehose into both the offline and online feature store FeatureGroups since you wish to train using your offline Feature Store groups and produce real-time inferences using your online Feature Store FeatureGroup.

Option C is correct. You can ingest your streaming data using Kinesis Data Firehose and use the kinesis-firehose-process-record Lambda blueprint for transformation. You will also want to stream the output of your Kinesis Data Firehose into both the offline and online feature store FeatureGroups since you wish to train using your offline Feature Store groups and produce real-time inferences using your online Feature Store FeatureGroup.

Option D is incorrect. You can ingest your streaming data using Kafka, and you could use the kinesis-process-record Lambda blueprint for transformation. However, you need to stream the output of your Kinesis Data Firehose into both the offline and online feature store FeatureGroups since you wish to train using your offline Feature Store groups and produce real-time inferences using your online Feature Store FeatureGroup.

References:

Please see the AWS Machine Learning blog titled **Understanding the key capabilities of Amazon SageMaker Feature Store** (https://aws.amazon.com/blogs/machine-learning/understanding-the-key-capabilities-of-amazon-sagemaker-feature-store/),

The Amazon SageMaker page titled **Amazon SageMaker Feature Store** (https://aws.amazon.com/sagemaker/feature-store/),

The Amazon SageMaker developer guide titled **Create, Store, and Share Features with Amazon SageMaker Feature Store** (https://docs.aws.amazon.com/sagemaker/latest/dg/feature-store.html),

The Amazon SageMaker developer guide titled **Create Feature Groups** (https://docs.aws.amazon.com/sagemaker/latest/dg/feature-store-create-feature-group.html),

The Amazon SageMaker Examples page titled **Fraud Detection with Amazon SageMaker FeatureStore** (https://sagemaker-examples.readthedocs.io/en/latest/sagemaker-featurestore/sagemaker_featurestore_fraud_detection_python_sdk.html),

The Amazon Kinesis Data Firehose developer guide titled **Amazon Kinesis Data Firehose Data Transformation** (https://docs.aws.amazon.com/firehose/latest/dev/data-transformation.html)

Ask our Experts

View Queries

Question 58

Correct

Domain: ML Implementation and Operations

You work as a machine learning specialist for a medical imaging company. You and your machine learning team have been assigned the task of building a model that predicts whether a breast mass image indicates a benign or malignant tumor. Your model will be used to help physicians quickly decide how to treat their patients using a verified diagnosis. Which option gives the appropriate machine learning services and features to train your model for your image diagnosis problem?

A. Specify the SageMaker role arn used to give learning and hosting access to your data by using the role = sagemaker.get_role() statement in your jupyter notebook. Load your data into a pandas dataframe. Split your data into 80% training, 10% validation and 10% testing. Set the predictor_type hyperparameter to binary_classifier. Then run your training job using the sagemaker.create_training_job statement in your jupyter notebook.

B. Specify the SageMaker role arn used to give learning and hosting access to your data by using the role = sagemaker.get_execution_role() statement in your jupyter notebook. Load your data into a pandas dataframe. Split your data into 80% training, 10% validation and 10% testing. Set the predictor_type hyperparameter to binary_classifier. Then run your training job using the sagemaker.create_training_job statement in your jupyter notebook.

C. Specify the SageMaker role arn used to give learning and hosting access to your data by using the role = sagemaker.get_execution_role() statement in your jupyter notebook. Load your data into a pandas dataframe. Split your data into 80% training, 10% validation and 10% testing. Set the predictor_type hyperparameter to the regressor type. Then run your training job using the sagemaker.create_training_job statement in your jupyter notebook. right

D. Specify the SageMaker role arn used to give learning and hosting access to your data by using the role = sagemaker.get_execution_role() statement in your jupyter notebook. Load your data into a pandas dataframe. Split your data into 80% training, 10% validation and 10% testing. Set the predictor_type hyperparameter to multiclass_classifier. Then run your training job using the sagemaker.create training_job statement in your jupyter notebook.

Explanation:

Correct Answer: C

Option A is incorrect. You specify the SageMaker role arn used to give learning and hosting access to your data by using the role = sagemaker.get_execution_role() statement, not the role =

sagemaker.get_role() statement. Also, since this problem is trying to predict whether an image mass is benign or malignant, you want to use a regression algorithm to give the probability that the mass is malignant, not a binary yes/no. Therefore, you should set the predictor_type hyperparameter to the value regressor, not binary_classifier.

Option B is incorrect. Since this problem is trying to predict whether an image mass is benign or malignant, you want to use a regression algorithm to give the probability that the mass is malignant, not a binary yes/no. Therefore, you should set the predictor_type hyperparameter to the value regressor, not binary_classifier.

Option C is correct. Since this problem is trying to predict whether an image mass is benign or malignant, you want to use a regression algorithm to give the probability that the mass is malignant, not a binary yes/no. Therefore, you should set the predictor_type hyperparameter to the regressor type. Also, it is correct to specify the SageMaker role arn used to give learning and hosting access to your data by using the role = sagemaker.get_execution_role() statement.

Option D is incorrect. Since this problem is trying to predict whether an image mass is benign or malignant, you want to use a regression algorithm to give the probability that the mass is malignant, not a classification across multiple classes. Therefore, you should set the predictor_type hyperparameter to a regressor, not a multiclass_classifier.

References:

Please see the AWS Amazon SageMaker Examples jupyter notebook titled **Breast Cancer Prediction** (https://github.com/aws/amazon-sagemaker-examples/blob/master/introduction_to_applying_machine_learning/breast_cancer_prediction/Breast%20Cancer%20Prediction.ipynb),

The Amazon SageMaker page titled **Linear Learner Algorithm** (https://docs.aws.amazon.com/sagemaker/latest/dg/linear-learner.html)

Ask our Experts

View Queries

Question 59

Correct

Domain: Modeling

You work as a machine learning specialist for a video game software company. You have been asked to produce a machine learning model that predicts whether a newly released game will eventually become a successful product that earns the company profits. Your data used for your model is product information and product ratings from social media. Your management team would like to use your model results to help them decide if a new game is worth investing in marketing dollars to promote the game further. Which model and objective will best match your model requirements?

A. XGboost, multi-softmax

B. XGboost, binary:logistic right

C. DeepAR, reg:logistic

D. Random Cut Forest, binary:logistic

Explanation:

Correct Answer: B

Option A is incorrect. XGBoost is a good choice for your algorithm, but the multi-softmax objective is used for multiclass classification. You are trying to predict whether your newly released game will eventually succeed in making your company money or not; a binary or logistic regression problem.

Option B is correct. XGBoost is a good choice for your algorithm, and the binary:logistic objective is the correct objective since it is used for binary classification problems. You are trying to predict whether your newly released game will eventually succeed in making your company money or not; a binary or logistic regression problem.

Option C is incorrect. The DeepAR algorithm is not the correct choice for your algorithm. The DeepAR algorithm is used with time-series data. You are using product information and product ratings from social media. Also, there is no reg:logistic objective for the DeepAR algorithm.

Option D is incorrect. The Random Cut Forest algorithm is an unsupervised algorithm used to detect anomalous data points in a data set. You would not try to use the Random Cut Forest algorithm to solve a logistic regression problem like predicting whether your newly released game will eventually succeed in making your company money or not.

References:

Please see the AWS Amazon SageMaker Examples jupyter notebook titled **Predicting Product Success When Review Data Is Available** (https://github.com/aws/amazon-sagemaker-examples/blob/master/introduction_to_applying_machine_learning/video_game_sales/video-game-sales-xgboost.ipynb),

The Amazon SageMaker developer guide page titled **XGBoost Hyperparameters** (https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost_hyperparameters.html),

The Amazon SageMaker developer guide page titled **Random Cut Forest (RCF) Algorithm** (https://docs.aws.amazon.com/sagemaker/latest/dg/randomcutforest.html),

The Amazon SageMaker developer guide page titled **DeepAR Forecasting Algorithm** (https://docs.aws.amazon.com/sagemaker/latest/dg/deepar.html),

The Amazon SageMaker GitHub repository titled **XGBoost Parameters** (https://github.com/dmlc/xgboost/blob/master/doc/parameter.rst#learning-task-parameters),

The Wikipedia page titled **Logistic regression** (https://en.wikipedia.org/wiki/Logistic_regression#:~:text=Logistic%20regression%20is%20a%20statistical,a%20form%20of%20binary%20regression)

Ask our Experts

View Queries

Question 60 Correct

Domain: Data Engineering

You work as a machine learning specialist for a financial services organization. Your machine learning team is responsible for building models that predict index fund tracking errors for the various funds managed by your mutual fund portfolio management department. You need to ingest data into your data lake for use in your machine learning models. The required securities pricing data come from varying sources that deliver the data you need to use in your model inferences in near real-time. You need to perform data transformation, such as compression, of the data before writing it to your S3 data lake. Which option gives you the most efficient solution for ingesting the data into your data lake?

A. Ingest the pricing data using a Kinesis Data Analytics application where you use Apache Flink to compress your data into the GZIP format and write it to your S3 data lake.

B. Ingest the pricing data into Kinesis Data Streams using a Kinesis Producer Library (KPL) application running on EC2 instances; use a Kinesis Client Library (KCL) application to compress your data into the GZIP format and write it to your S3 data lake.

C. Ingest the pricing data using Kinesis Data Firehose where you use a Lambda function to compress your data into the GZIP format and have the Lambda function write the data to your S3 data lake.

D. Ingest the pricing data using Kinesis Data Firehose where you use a Lambda function to compress your data into the GZIP format; Kinesis Data Firehose writes the data to your S3 data lake. right

Explanation:

Correct Answer: D

Option A is incorrect. Kinesis Data Analytics needs to be fed the streaming data by either Kinesis Data Streams or Kinesis Data Firehose. Kinesis Data Analytics cannot ingest data directly. Also, Apache Flink can write your data to S3 using the streaming file sink, but it writes in the AVRO and Parquet formats, not GZIP.

Option B is incorrect. The solution described in this option will technically work. However, it is much less efficient than using Kinesis Data Firehose to ingest, compress using Lambda, and write your data to S3.

Option C is incorrect. You can ingest your pricing data using Kinesis Data Firehose and use lambda to compress your data into the GZIP format. However, you should leverage the Kinesis Data Firehose capability to write your data directly to your S3 bucket. This is more efficient than writing your own code in your Lambda function to write the data to S3.

Option D is correct. Ingesting the data using Kinesis Data Firehose, using Lambda to compress the data into the GZIP format, and then having Kinesis Data Firehose write your data to S3 is a very common example of using Kinesis Data Firehose for a very efficient data ingestion solution.

References:

Please see the Building Big Data Storage Solutions (Data Lakes) for Maximum Flexibility AWS Whitepaper titled **Data Ingestion Methods** (https://docs.aws.amazon.com/whitepapers/latest/building-data-lakes/data-ingestion-methods.html),

The Investopedia page titled **Tracking Error** (https://www.investopedia.com/terms/t/trackingerror.asp#:~:text=Tracking%20error%20is%20the%20difference,and%20its%20corresponding%20risk%20level.),

The Apache Flink developer guide titled **Streaming File Sink** (https://ci.apache.org/projects/flink/flink-docs-stable/dev/connectors/streamfile_sink.html),

The Amazon Kinesis Data Streams product page titled **Getting started with Amazon Kinesis Data Streams** (https://aws.amazon.com/kinesis/data-streams/getting-started/),

The Amazon Kinesis Data Analytics for SQL Applications Developer Guide SQL developer guide titled **Amazon Kinesis Data Analytics for SQL Applications: How It Works** (https://docs.aws.amazon.com/kinesisanalytics/latest/dev/how-it-works.html)

Ask our Experts

View Queries

Question 61

Correct

Domain: ML Implementation and Operations

You work as a machine learning specialist for a data mining department of a large bank. Your department is responsible for leveraging the bank's huge data lake to gain insights and make predictions for your marketing and risk departments. Your team's latest project, an XGBoost prediction model, is ready for production deployment. However, you want to run some additional batch predictions using a batch inference job to make sure your model can handle the production prediction workload. In your SageMaker notebook, how do you extend your estimator to read input data in batch from a specified S3 bucket and make predictions?

A. Extend the estimator to a Transformer object. right

B. Extend the estimator to a Predictor object.

C. Extend the estimator to a MultiDataModel object.

D. Extend the estimator to a BatchPredictor object.

Explanation:

Correct Answer: A

Option A is correct. You can extend your estimator to a transformer object, which is derived from the SageMaker Transformer class. The batch transformer reads input data from a specified S3 bucket and makes predictions.

Option B is incorrect. The Predictor object makes prediction requests to an Amazon SageMaker endpoint. However, it is not the SageMaker API used to perform batch predictions.

Option C is incorrect. The MultiDataModel object is used to deploy multiple models to the same endpoint, not to make batch predictions.

Option D is incorrect. There is no BatchPredictor SageMaker API.

References:

Please see the Amazon SageMaker developer guide titled **Deploy a Model in Amazon SageMaker** (https://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-deployment.html),

The Amazon SageMaker developer guide titled **Step 5: Deploy the Model to Amazon EC2** (https://docs.aws.amazon.com/sagemaker/latest/dg/ex1-model-deployment.html#ex1-deploy-model), a

The SageMaker API page titled **Transformer** (https://sagemaker.readthedocs.io/en/stable/api/inference/transformer.html),

The SageMaker API page titled **Predictors** (https://sagemaker.readthedocs.io/en/stable/api/inference/predictors.html),

The SageMaker API page titled **MultiDataModel** (https://sagemaker.readthedocs.io/en/stable/api/inference/multi_data_model.html),

The SageMaker API page titled **Inference APIs** (https://sagemaker.readthedocs.io/en/stable/api/inference/index.html)

Ask our Experts

View Queries

Question 62

Correct

Domain: Exploratory Data Analysis

You work as a machine learning specialist for the sales department of a large web retailer that needs to gain insight into their sales patterns. They need a way to use a visualization to show their sales data in near-real time so that they can quickly recognize higher-than-expected sales of specific products. This will help your product operations quickly meet high demands. Which option is a viable, efficient solution to your problem?

A. Use Kinesis Data Streams to stream your data to S3. Then run a Random Cut Forest SagaMaker model on the data continuously, using the output as source data to visualization in QuickSight.

B. Use Kinesis Data Firehose to stream your data to S3. Then run a Random Cut Forest SagaMaker model on the data continuously, using the output as source data to visualization in QuickSight.

C. Use Kinesis Data Firehose to stream your data to S3. Use QuickSight ML Insights to use the output as source data to visualization in QuickSight. right

D. Use Kinesis Data Firehose to stream your data to a Kinesis Data Analytics application that runs a Random Cut Forest SagaMaker model on the data continuously, writing the output to S3 which is then used as source data to visualization in QuickSight.

Explanation:

Correct Answer: C

Option A is incorrect. Kinesis Data Streams cannot stream your data directly to S3. Also, running your own SageMaker Random Cut Forest model against your data is much less efficient than using the QuickSight ML Insights integrated Random Cut Forest capability.

Option B is incorrect. While Kinesis Data Firehose can stream your data directly to S3, running your own SageMaker Random Cut Forest model against your data is much less efficient than using the QuickSight ML Insights integrated Random Cut Forest capability.

Option C is correct. Streaming your data directly to S3 using Kinesis Data Firehose is very efficient. Also, using QuickSight's integrated ML Insights Random Cut Forest capability requires far less development and coding effort than the other options.

Option D is incorrect. Kinesis Data Analytics has a Random Cut Forest capability that you can use to detect your sales outliers. However, you would still have to build your visualization in QuickSight. The option of using ML Insights directly within QuickSight allows you to run your anomaly detection and visualize your data more quickly.

References:

Please see the Amazon QuickSight user guide titled **Working with ML Insights** (https://docs.aws.amazon.com/quicksight/latest/user/making-data-driven-decisions-with-ml-in-quicksight.html),

The Amazon QuickSight user guide titled **Detecting Outliers with ML-Powered Anomaly Detection** (https://docs.aws.amazon.com/quicksight/latest/user/anomaly-detection.html),

The AWS Machine Learning blog titled **Visualizing Amazon SageMaker machine learning predictions with Amazon QuickSight** (https://aws.amazon.com/blogs/machine-learning/making-machine-learning-predictions-in-amazon-quicksight-and-amazon-sagemaker/),

The Amazon Kinesis Data Analytics SQL reference titled **RANDOM_CUT_FOREST** (https://docs.aws.amazon.com/kinesisanalytics/latest/sqlref/sqlrf-random-cut-forest.html)

Ask our Experts

Ask our Experts

View Queries

Question 63

Correct

Domain: Modeling

You work as a machine learning specialist for a security company that uses video feeds to identify criminal activity in a client company's retail environments. You are building a convolutional neural network model using TensorFlow to use for your video classification. The model is expected to classify video scenes as criminal, such as theft, or benign. You have thousands of hours of video on which to train your model. Therefore, you plan to leverage hyperparameter tuning to run multiple training jobs using different hyperparameter combinations. The goal is to find the model with the best training result. You are writing your hyperparameter tuning job in your SageMaker jupyter notebook. When you create your HyperparameterTuner object in your python code, which parameters do you pass in the method? (Select TWO)

- [x] A. TensorFlow estimator right
- [ ] B. Training steps
- [ ] C. Evaluation steps
- [x] D. Hyperparameter ranges right
- [ ] E. Instance type

Explanation:

Correct Answers: A and D

Option A is correct. When you create the HyperparameterTuner object in your python code, you need to specify the estimator you are using, in this case, TensorFlow, the ranges for your hyperparameters, the objective metric you wish to solve to, and resource configuration details, such as the number of training jobs to run in total and how many training jobs can be run in parallel.

Option B is incorrect. You specify the training steps when you create your TensorFlow estimator in your python code.

Option C is incorrect. You specify the evaluation steps when you create your TensorFlow estimator in your python code.

Option D is correct. When you create the HyperparameterTuner object in your python code, you need to specify the estimator you are using, in this case, TensorFlow, the ranges for your hyperparameters, the objective metric you wish to solve to, and resource configuration details, such as the number of training jobs to run in total and how many training jobs can be run in parallel.

Option E is incorrect. You specify the instance type when you create your TensorFlow estimator in your python code.

References:

Please see the Amazon SageMaker page titled **Machine learning for every developer and data scientist** (https://aws.amazon.com/machine-learning/accelerate-amazon-sagemaker/#Train_machine_learning_models),

The Amazon SageMaker Examples titled **Hyperparameter Tuning using SageMaker Tensorflow Container** (https://github.com/aws/amazon-sagemaker-examples/blob/master/hyperparameter_tuning/tensorflow_mnist/hpo_tensorflow_mnist.ipynb),

The Medium article titled **Video Classification using CNNs** (https://medium.com/analytics-vidhya/video-classification-using-cnns-db18bd2b7e72)

Ask our Experts

View Queries

Question 64

Correct

Domain: Modeling

You work as a machine learning specialist for a social media software company that produces games for mobile devices. Your company has a new game that they believe will generate a large following very quickly. You need to build a model to predict whether users will purchase additional game features via in-app purchases. You have a large dataset to use for your training, and you need to find the best hyperparameters by using a hyperparameter tuning job. You have configured the training jobs the hyperparameter tuning job will run by defining an estimator and objective. You want to run your training jobs in a highly parallel manner because you want to complete your hyperparameter tuning quickly. Also, you know that the order of magnitude is more important than the absolute value for your hyperparameter values. For example, a change from 1 to 2 is expected to have a much bigger impact than a change from 100 to 101. Which scaling type and search type combination should you use for your hyperparameter tuning job?

A. Logarithmic scaling and Bayesian search

B. Logarithmic scaling and Random search right

C. Linear scaling and Bayesian search

D. Linear scaling and Random search

Explanation:

Correct Answer: B

Option A is incorrect. When an order of magnitude is more important than the absolute value for your hyperparameters in your tuning, you should use logarithmic scaling. However, when you wish to run many training jobs in parallel, you should use a Random search strategy, not a Bayesian search strategy.

Option B is correct. When an order of magnitude is more important than the absolute value for your hyperparameters in your tuning, you should use logarithmic scaling. When you wish to run many training jobs in parallel, you should use a Random search strategy.

Option C is incorrect. When an order of magnitude is more important than the absolute value for your hyperparameters in your tuning, you should use logarithmic scaling, not linear scaling. When you wish to run many training jobs in parallel, you should use a Random search strategy, not a Bayesian search strategy.

Option D is incorrect. When an order of magnitude is more important than the absolute value for your hyperparameters in your tuning, you should use logarithmic scaling, not linear scaling.

References:

Please see the Amazon SageMaker developer guide titled **How Hyperparameter Tuning Works** (https://docs.aws.amazon.com/sagemaker/latest/dg/automatic-model-tuning-how-it-works.html),

The Amazon SageMaker Examples titled **Random search and hyperparameter scaling with SageMaker XGBoost and Automatic Model Tuning** (https://github.com/aws/amazon-sagemaker-examples/blob/master/hyperparameter_tuning/xgboost_random_log/hpo_xgboost_random_log.ipynb)

Ask our Experts

View Queries

Question 65

Correct

Domain: Exploratory Data Analysis

You work as a machine learning specialist for a scientific research lab that analyzes fossils found in geological research digs worldwide. You are currently working on a project that is analyzing bones of ancient mammals found during archaeological excavations in Africa. The data that the archaeologists provide to you for each specimen is exact for measurements, density, skeleton structural component, etc. Your SageMaker Linear Learner model predicts the age of the specimen based on the collected data. This data needs to be sanitized and categorized before you can feed it into your inference engine to get the estimated age. You are using the SageMaker built-in Scikit-learn library to do your data preprocessing. You need to transform categorical values such as skeletal components (femur, skull, rib cage, etc.) into numerical values. You also need to replace missing values with meaningful estimates. Which Scikit-learn library methods should you use to perform these data preprocessing tasks? (Select TWO)?

A. Normalizer

B. StandardScaler

- [x] C. SimpleImputer right

D. Binarizer

- [x] E. OneHotEncoder right

Explanation:

Correct Answers: C and E

Option A is incorrect. The Scikit-learn Normalizer normalizes values to a unit norm. You need to transform categorical values into numerical representations, and you need to replace missing values.

Option B is incorrect. The Scikit-learn Standardizer standardizes values to a unit norm. You need to transform categorical values into numerical representations, and you need to replace missing values.

Option C is correct. The SimpleImputer completes or estimates missing values. This is one of the two sanitation tasks you need to perform.

Option D is incorrect. The Scikit-learn Binarizer sets feature values to 0 or 1 according to a threshold. You need to transform categorical values into numerical values that can represent many different categories, and you need to replace missing values.

Option E is correct. The OneHotEncoder encodes categorical features into a one-hot numeric array with each entry in the array representing a category. There are as many entries in the array as there are categories in the feature. The 'one' in a given array element represents a categorical value numerically.

References:

Please see the AWS Machine Learning blog titled **Preprocess input data before making predictions using Amazon SageMaker inference pipelines and Scikit-learn** (https://aws.amazon.com/blogs/machine-learning/preprocess-input-data-before-making-predictions-using-amazon-sagemaker-inference-pipelines-and-scikit-learn/),

The Amazon SageMaker Examples titled **Inference Pipeline with Scikit-learn and Linear Learner** (https://github.com/aws/amazon-sagemaker-examples/blob/master/sagemaker-python-sdk/scikit_learn_inference_pipeline/Inference%20Pipeline%20with%20Scikit-learn%20and%20Linear%20Learner.ipynb),

Amazon SageMaker developer guide titled **Use Scikit-learn with Amazon SageMaker** (https://docs.aws.amazon.com/sagemaker/latest/dg/sklearn.html),

Scikit-learn API page titled **sklearn.preprocessing.OneHotEncoder** (https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html),

Scikit-learn API page titled **sklearn.impute.SimpleImputer** (https://scikit-learn.org/stable/modules/generated/sklearn.impute.SimpleImputer.html),

Scikit-learn API page titled **API Reference** (https://scikit-learn.org/stable/modules/classes.html)

Ask our Experts

View Queries

Finish Review