NO.1 A Machine Learning Specialist is working with multiple data sources containing billions of records that need to be joined. What feature engineering and model development approach should the Specialist take with a dataset this large?

A. Use an Amazon SageMaker notebook for both feature engineering and model development

B. Use an Amazon SageMaker notebook for feature engineering and Amazon ML for model development

C. Use Amazon EMR for feature engineering and Amazon SageMaker SDK for model development

D. Use Amazon ML for both feature engineering and model development.

Answer: C

Explanation:

Amazon EMR is a service that can process large amounts of data efficiently and cost-effectively. It can run distributed frameworks such as Apache Spark, which can perform feature engineering on big data. Amazon SageMaker SDK is a Python library that can interact with Amazon SageMaker service to train and deploy machine learning models. It can also use Amazon EMR as a data source for training data. References:

Amazon EMR

Amazon SageMaker SDK

NO.2 A Machine Learning Specialist is building a supervised model that will evaluate customers' satisfaction with their mobile phone service based on recent usage The model's output should infer whether or not a customer is likely to switch to a competitor in the next 30 days Which of the following modeling techniques should the Specialist use1?

A. Time-series prediction

B. Anomaly detection

C. Binary classification

D. Regression

Answer: C

Explanation:

The modeling technique that the Machine Learning Specialist should use is binary classification. Binary classification is a type of supervised learning that predicts whether an input belongs to one of two possible classes. In this case, the input is the customer's recent usage data and the output is whether or not the customer is likely to switch to a competitor in the next 30 days. This is a binary outcome, either yes or no, so binary classification is suitable for this problem. The other options are not appropriate for this problem. Time-series prediction is a type of supervised learning that forecasts future values based on past and present data. Anomaly detection is a type of unsupervised learning that identifies outliers or abnormal patterns in the data. Regression is a type of supervised learning that estimates a continuous numerical value based on the input features. References: Binary Classification, Time Series Prediction, Anomaly Detection, Regression

NO.3 A Machine Learning Specialist is assigned a TensorFlow project using Amazon SageMaker for training, and needs to continue working for an extended period with no Wi-Fi access.
Which approach should the Specialist use to continue working?

A. Install Python 3 and boto3 on their laptop and continue the code development using that environment.

B. Download the TensorFlow Docker container used in Amazon SageMaker from GitHub to their local environment, and use the Amazon SageMaker Python SDK to test the code.

C. Download TensorFlow from tensorflow.org to emulate the TensorFlow kernel in the SageMaker environment.

D. Download the SageMaker notebook to their local environment then install Jupyter Notebooks on their laptop and continue the development in a local notebook.

Answer: B

Explanation:

Amazon SageMaker is a fully managed service that enables developers and data scientists to quickly and easily build, train, and deploy machine learning models at any scale. SageMaker provides a variety of tools and frameworks to support the entire machine learning workflow, from data preparation to model deployment.

One of the tools that SageMaker offers is the Amazon SageMaker Python SDK, which is a high-level library that simplifies the interaction with SageMaker APIs and services. The SageMaker Python SDK allows you to write code in Python and use popular frameworks such as TensorFlow, PyTorch, MXNet, and more. You can use the SageMaker Python SDK to create and manage SageMaker resources such as notebook instances, training jobs, endpoints, and feature store.

If you need to continue working on a TensorFlow project using SageMaker for training without Wi-Fi access, the best approach is to download the TensorFlow Docker container used in SageMaker from GitHub to your local environment, and use the SageMaker Python SDK to test the code. This way, you can ensure that your code is compatible with the SageMaker environment and avoid any potential issues when you upload your code to SageMaker and start the training job. You can also use the same code to deploy your model to a SageMaker endpoint when you have Wi-Fi access again.

To download the TensorFlow Docker container used in SageMaker, you can visit the SageMaker Docker GitHub repository and follow the instructions to build the image locally. You can also use the SageMaker Studio Image Build CLI to automate the process of building and pushing the Docker image to Amazon Elastic Container Registry (Amazon ECR). To use the SageMaker Python SDK to test the code, you can install the SDK on your local machine by following the installation guide. You can also refer to the TensorFlow documentation for more details on how to use the SageMaker Python SDK with TensorFlow.

References:

SageMaker Docker GitHub repository

SageMaker Studio Image Build CLI

SageMaker Python SDK installation guide

SageMaker Python SDK TensorFlow documentation

NO.4 A Data Scientist wants to gain real-time insights into a data stream of GZIP files. Which solution would allow the use of SQL to query the stream with the LEAST latency?

A. Amazon Kinesis Data Analytics with an AWS Lambda function to transform the data.

B. AWS Glue with a custom ETL script to transform the data.

C. An Amazon Kinesis Client Library to transform the data and save it to an Amazon ES cluster.

D. Amazon Kinesis Data Firehose to transform the data and put it into an Amazon S3 bucket.

Answer: A

Explanation:

Amazon Kinesis Data Analytics is a service that enables you to analyze streaming data in real time using SQL or Apache Flink applications. You can use Kinesis Data Analytics to process and gain insights

from data streams such as web logs, clickstreams, IoT data, and more.
To use SQL to query a data stream of GZIP files, you need to first transform the data into a format that Kinesis Data Analytics can understand, such as JSON, CSV, or Apache Parquet. You can use an AWS Lambda function to perform this transformation and send the output to a Kinesis data stream that is connected to your Kinesis Data Analytics application. This way, you can use SQL to query the stream with the least latency, as Lambda functions are triggered in near real time by the incoming data and Kinesis Data Analytics can process the data as soon as it arrives.
The other options are not optimal for this scenario, as they introduce more latency or complexity. AWS Glue is a serverless data integration service that can perform ETL (extract, transform, and load) tasks on data sources, but it is not designed for real-time streaming data analysis. An Amazon Kinesis Client Library is a Java library that enables you to build custom applications that process data from Kinesis data streams, but it requires more coding and configuration than using a Lambda function. Amazon Kinesis Data Firehose is a service that can deliver streaming data to destinations such as Amazon S3, Amazon Redshift, Amazon OpenSearch Service, and Splunk, but it does not support SQL queries on the data.
References:
What Is Amazon Kinesis Data Analytics for SQL Applications?
Using AWS Lambda with Amazon Kinesis Data Streams
Using AWS Lambda with Amazon Kinesis Data Firehose

NO.5 A Machine Learning Specialist must build out a process to query a dataset on Amazon S3 using Amazon Athena The dataset contains more than 800.000 records stored as plaintext CSV files Each record contains 200 columns and is approximately 1 5 MB in size Most queries will span 5 to 10 columns only How should the Machine Learning Specialist transform the dataset to minimize query runtime?

A. Convert the records to Apache Parquet format

B. Convert the records to JSON format

C. Convert the records to GZIP CSV format

D. Convert the records to XML format

Answer: A

Explanation:
To optimize the query performance of Athena, one of the best practices is to convert the data into a columnar format, such as Apache Parquet or Apache ORC. Columnar formats store data by columns rather than by rows, which allows Athena to scan only the columns that are relevant to the query, reducing the amount of data read and improving the query speed. Columnar formats also support compression and encoding schemes that can reduce the storage space and the data scanned per query, further enhancing the performance and reducing the cost.
In contrast, plaintext CSV files store data by rows, which means that Athena has to scan the entire row even if only a few columns are needed for the query. This increases the amount of data read and the query latency. Moreover, plaintext CSV files do not support compression or encoding, which means that they take up more storage space and incur higher query costs.
Therefore, the Machine Learning Specialist should transform the dataset to Apache Parquet format to minimize query runtime.
References:
Top 10 Performance Tuning Tips for Amazon Athena
Columnar Storage Formats

Using compressions will reduce the amount of data scanned by Amazon Athena, and also reduce your S3 bucket storage. It's a Win-Win for your AWS bill. Supported formats: GZIP, LZO, SNAPPY (Parquet) and ZLIB.

NO.6 A Machine Learning Specialist is developing a daily ETL workflow containing multiple ETL jobs
The workflow consists of the following processes
* Start the workflow as soon as data is uploaded to Amazon S3
* When all the datasets are available in Amazon S3, start an ETL job to join the uploaded datasets with multiple terabyte-sized datasets already stored in Amazon S3
* Store the results of joining datasets in Amazon S3
* If one of the jobs fails, send a notification to the Administrator
Which configuration will meet these requirements?

A. Use AWS Lambda to trigger an AWS Step Functions workflow to wait for dataset uploads to complete in Amazon S3. Use AWS Glue to join the datasets Use an Amazon CloudWatch alarm to send an SNS notification to the Administrator in the case of a failure

B. Develop the ETL workflow using AWS Lambda to start an Amazon SageMaker notebook instance Use a lifecycle configuration script to join the datasets and persist the results in Amazon S3 Use an Amazon CloudWatch alarm to send an SNS notification to the Administrator in the case of a failure

C. Develop the ETL workflow using AWS Batch to trigger the start of ETL jobs when data is uploaded to Amazon S3 Use AWS Glue to join the datasets in Amazon S3 Use an Amazon CloudWatch alarm to send an SNS notification to the Administrator in the case of a failure

D. Use AWS Lambda to chain other Lambda functions to read and join the datasets in Amazon S3 as soon as the data is uploaded to Amazon S3 Use an Amazon CloudWatch alarm to send an SNS notification to the Administrator in the case of a failure

Answer: A

Explanation:
To develop a daily ETL workflow containing multiple ETL jobs that can start as soon as data is uploaded to Amazon S3, the best configuration is to use AWS Lambda to trigger an AWS Step Functions workflow to wait for dataset uploads to complete in Amazon S3. Use AWS Glue to join the datasets. Use an Amazon CloudWatch alarm to send an SNS notification to the Administrator in the case of a failure.
AWS Lambda is a serverless compute service that lets you run code without provisioning or managing servers. You can use Lambda to create functions that respond to events such as data uploads to Amazon S3. You can also use Lambda to invoke other AWS services such as AWS Step Functions and AWS Glue.
AWS Step Functions is a service that lets you coordinate multiple AWS services into serverless workflows. You can use Step Functions to create a state machine that defines the sequence and logic of your ETL workflow. You can also use Step Functions to handle errors and retries, and to monitor the execution status of your workflow.
AWS Glue is a serverless data integration service that makes it easy to discover, prepare, and combine data for analytics. You can use Glue to create and run ETL jobs that can join data from multiple sources in Amazon S3. You can also use Glue to catalog your data and make it searchable and queryable.
Amazon CloudWatch is a service that monitors your AWS resources and applications. You can use CloudWatch to create alarms that trigger actions when a metric or a log event meets a specified

threshold. You can also use CloudWatch to send notifications to Amazon Simple Notification Service (SNS) topics, which can then deliver the notifications to subscribers such as email addresses or phone numbers.
Therefore, by using these services together, you can achieve the following benefits:
You can start the ETL workflow as soon as data is uploaded to Amazon S3 by using Lambda functions to trigger Step Functions workflows.
You can wait for all the datasets to be available in Amazon S3 by using Step Functions to poll the S3 buckets and check the data completeness.
You can join the datasets with terabyte-sized datasets in Amazon S3 by using Glue ETL jobs that can scale and parallelize the data processing.
You can store the results of joining datasets in Amazon S3 by using Glue ETL jobs to write the output to S3 buckets.
You can send a notification to the Administrator if one of the jobs fails by using CloudWatch alarms to monitor the Step Functions or Glue metrics and send SNS notifications in case of a failure.

NO.7 An agency collects census information within a country to determine healthcare and social program needs by province and city. The census form collects responses for approximately 500 questions from each citizen Which combination of algorithms would provide the appropriate insights? (Select TWO)

A. The factorization machines (FM) algorithm

B. The Latent Dirichlet Allocation (LDA) algorithm

C. The principal component analysis (PCA) algorithm

D. The k-means algorithm

E. The Random Cut Forest (RCF) algorithm

Answer: C,D

Explanation:
The agency wants to analyze the census data for population segmentation, which is a type of unsupervised learning problem that aims to group similar data points together based on their attributes. The agency can use a combination of algorithms that can perform dimensionality reduction and clustering on the data to achieve this goal.
Dimensionality reduction is a technique that reduces the number of features or variables in a dataset while preserving the essential information and relationships. Dimensionality reduction can help improve the efficiency and performance of clustering algorithms, as well as facilitate data visualization and interpretation. One of the most common algorithms for dimensionality reduction is principal component analysis (PCA), which transforms the original features into a new set of orthogonal features called principal components that capture the maximum variance in the data. PCA can help reduce the noise and redundancy in the data and reveal the underlying structure and patterns.
Clustering is a technique that partitions the data into groups or clusters based on their similarity or distance. Clustering can help discover the natural segments or categories in the data and understand their characteristics and differences. One of the most popular algorithms for clustering is k-means, which assigns each data point to one of k clusters based on the nearest mean or centroid. K-means can handle large and high-dimensional datasets and produce compact and spherical clusters.
Therefore, the combination of algorithms that would provide the appropriate insights for population segmentation are PCA and k-means. The agency can use PCA to reduce the dimensionality of the census data from 500 features to a smaller number of principal components that capture most of the

variation in the data. Then, the agency can use k-means to cluster the data based on the principal components and identify the segments of the population that share similar characteristics.
References:
Amazon SageMaker Principal Component Analysis (PCA)
Amazon SageMaker K-Means Algorithm

NO.8 A large consumer goods manufacturer has the following products on sale
* 34 different toothpaste variants
* 48 different toothbrush variants
* 43 different mouthwash variants
The entire sales history of all these products is available in Amazon S3 Currently, the company is using custom-built autoregressive integrated moving average (ARIMA) models to forecast demand for these products The company wants to predict the demand for a new product that will soon be launched Which solution should a Machine Learning Specialist apply?

A. Train a custom ARIMA model to forecast demand for the new product.

B. Train an Amazon SageMaker DeepAR algorithm to forecast demand for the new product

C. Train an Amazon SageMaker k-means clustering algorithm to forecast demand for the new product.

D. Train a custom XGBoost model to forecast demand for the new product

Answer: B
Explanation:
The company wants to predict the demand for a new product that will soon be launched, based on the sales history of similar products. This is a time series forecasting problem, which requires a machine learning algorithm that can learn from historical data and generate future predictions.
One of the most suitable solutions for this problem is to use the Amazon SageMaker DeepAR algorithm, which is a supervised learning algorithm for forecasting scalar time series using recurrent neural networks (RNN). DeepAR can handle multiple related time series, such as the sales of different products, and learn a global model that captures the common patterns and trends across the time series. DeepAR can also generate probabilistic forecasts that provide confidence intervals and quantify the uncertainty of the predictions.
DeepAR can outperform traditional forecasting methods, such as ARIMA, especially when the dataset contains hundreds or thousands of related time series. DeepAR can also use the trained model to forecast the demand for new products that are similar to the ones it has been trained on, by using the categorical features that encode the product attributes. For example, the company can use the product type, brand, flavor, size, and price as categorical features to group the products and learn the typical behavior for each group.
Therefore, the Machine Learning Specialist should apply the Amazon SageMaker DeepAR algorithm to forecast the demand for the new product, by using the sales history of the existing products as the training dataset, and the product attributes as the categorical features.
References:
DeepAR Forecasting Algorithm - Amazon SageMaker
Now available in Amazon SageMaker: DeepAR algorithm for more accurate time series forecasting

NO.9 A Data Scientist needs to migrate an existing on-premises ETL process to the cloud The current process runs at regular time intervals and uses PySpark to combine and format multiple large data sources into a single consolidated output for downstream processing The Data Scientist has been

given the following requirements for the cloud solution
* Combine multiple data sources
* Reuse existing PySpark logic
* Run the solution on the existing schedule
* Minimize the number of servers that will need to be managed

Which architecture should the Data Scientist use to build this solution?

A. Write the raw data to Amazon S3 Schedule an AWS Lambda function to submit a Spark step to a persistent Amazon EMR cluster based on the existing schedule Use the existing PySpark logic to run the ETL job on the EMR cluster Output the results to a "processed" location m Amazon S3 that is accessible tor downstream use

B. Write the raw data to Amazon S3 Create an AWS Glue ETL job to perform the ETL processing against the input data Write the ETL job in PySpark to leverage the existing logic Create a new AWS Glue trigger to trigger the ETL job based on the existing schedule Configure the output target of the ETL job to write to a "processed" location in Amazon S3 that is accessible for downstream use.

C. Write the raw data to Amazon S3 Schedule an AWS Lambda function to run on the existing schedule and process the input data from Amazon S3 Write the Lambda logic in Python and implement the existing PySpartc logic to perform the ETL process Have the Lambda function output the results to a "processed" location in Amazon S3 that is accessible for downstream use

D. Use Amazon Kinesis Data Analytics to stream the input data and perform realtime SQL queries against the stream to carry out the required transformations within the stream Deliver the output results to a "processed" location in Amazon S3 that is accessible for downstream use

Answer: B

Explanation:

The Data Scientist needs to migrate an existing on-premises ETL process to the cloud, using a solution that can combine multiple data sources, reuse existing PySpark logic, run on the existing schedule, and minimize the number of servers that need to be managed. The best architecture for this scenario is to use AWS Glue, which is a serverless data integration service that can create and run ETL jobs on AWS.

AWS Glue can perform the following tasks to meet the requirements:

Combine multiple data sources: AWS Glue can access data from various sources, such as Amazon S3, Amazon RDS, Amazon Redshift, Amazon DynamoDB, and more. AWS Glue can also crawl the data sources and discover their schemas, formats, and partitions, and store them in the AWS Glue Data Catalog, which is a centralized metadata repository for all the data assets.

Reuse existing PySpark logic: AWS Glue supports writing ETL scripts in Python or Scala, using Apache Spark as the underlying execution engine. AWS Glue provides a library of built-in transformations and connectors that can simplify the ETL code. The Data Scientist can write the ETL job in PySpark and leverage the existing logic to perform the data processing.

Run the solution on the existing schedule: AWS Glue can create triggers that can start ETL jobs based on a schedule, an event, or a condition. The Data Scientist can create a new AWS Glue trigger to run the ETL job based on the existing schedule, using a cron expression or a relative time interval.

Minimize the number of servers that need to be managed: AWS Glue is a serverless service, which means that it automatically provisions, configures, scales, and manages the compute resources required to run the ETL jobs. The Data Scientist does not need to worry about setting up, maintaining, or monitoring any servers or clusters for the ETL process.

Therefore, the Data Scientist should use the following architecture to build the cloud solution:

Write the raw data to Amazon S3: The Data Scientist can use any method to upload the raw data from the on-premises sources to Amazon S3, such as AWS DataSync, AWS Storage Gateway, AWS Snowball, or AWS Direct Connect. Amazon S3 is a durable, scalable, and secure object storage service that can store any amount and type of data.
Create an AWS Glue ETL job to perform the ETL processing against the input data: The Data Scientist can use the AWS Glue console, AWS Glue API, AWS SDK, or AWS CLI to create and configure an AWS Glue ETL job. The Data Scientist can specify the input and output data sources, the IAM role, the security configuration, the job parameters, and the PySpark script location. The Data Scientist can also use the AWS Glue Studio, which is a graphical interface that can help design, run, and monitor ETL jobs visually.
Write the ETL job in PySpark to leverage the existing logic: The Data Scientist can use a code editor of their choice to write the ETL script in PySpark, using the existing logic to transform the data. The Data Scientist can also use the AWS Glue script editor, which is an integrated development environment (IDE) that can help write, debug, and test the ETL code. The Data Scientist can store the ETL script in Amazon S3 or GitHub, and reference it in the AWS Glue ETL job configuration.
Create a new AWS Glue trigger to trigger the ETL job based on the existing schedule: The Data Scientist can use the AWS Glue console, AWS Glue API, AWS SDK, or AWS CLI to create and configure an AWS Glue trigger. The Data Scientist can specify the name, type, and schedule of the trigger, and associate it with the AWS Glue ETL job. The trigger will start the ETL job according to the defined schedule.
Configure the output target of the ETL job to write to a "processed" location in Amazon S3 that is accessible for downstream use: The Data Scientist can specify the output location of the ETL job in the PySpark script, using the AWS Glue DynamicFrame or Spark DataFrame APIs. The Data Scientist can write the output data to a "processed" location in Amazon S3, using a format such as Parquet, ORC, JSON, or CSV, that is suitable for downstream processing.
References:
What Is AWS Glue?
AWS Glue Components
AWS Glue Studio
AWS Glue Triggers

NO.10 A large company has developed a B1 application that generates reports and dashboards using data collected from various operational metrics The company wants to provide executives with an enhanced experience so they can use natural language to get data from the reports The company wants the executives to be able ask questions using written and spoken interlaces Which combination of services can be used to build this conversational interface? (Select THREE)

A. Alexa for Business

B. Amazon Connect

C. Amazon Lex

D. Amazon Poly

E. Amazon Comprehend

F. Amazon Transcribe

Answer: C,E,F

Explanation:
To build a conversational interface that can use natural language to get data from the reports, the

company can use a combination of services that can handle both written and spoken inputs, understand the user's intent and query, and extract the relevant information from the reports. The services that can be used for this purpose are:

Amazon Lex: A service for building conversational interfaces into any application using voice and text. Amazon Lex can create chatbots that can interact with users using natural language, and integrate with other AWS services such as Amazon Connect, Amazon Comprehend, and Amazon Transcribe. Amazon Lex can also use lambda functions to implement the business logic and fulfill the user's requests.

Amazon Comprehend: A service for natural language processing and text analytics. Amazon Comprehend can analyze text and speech inputs and extract insights such as entities, key phrases, sentiment, syntax, and topics. Amazon Comprehend can also use custom classifiers and entity recognizers to identify specific terms and concepts that are relevant to the domain of the reports.

Amazon Transcribe: A service for speech-to-text conversion. Amazon Transcribe can transcribe audio inputs into text outputs, and add punctuation and formatting. Amazon Transcribe can also use custom vocabularies and language models to improve the accuracy and quality of the transcription for the specific domain of the reports.

Therefore, the company can use the following architecture to build the conversational interface:

Use Amazon Lex to create a chatbot that can accept both written and spoken inputs from the executives. The chatbot can use intents, utterances, and slots to capture the user's query and parameters, such as the report name, date, metric, or filter.

Use Amazon Transcribe to convert the spoken inputs into text outputs, and pass them to Amazon Lex. Amazon Transcribe can use a custom vocabulary and language model to recognize the terms and concepts related to the reports.

Use Amazon Comprehend to analyze the text inputs and outputs, and extract the relevant information from the reports. Amazon Comprehend can use a custom classifier and entity recognizer to identify the report name, date, metric, or filter from the user's query, and the corresponding data from the reports.

Use a lambda function to implement the business logic and fulfillment of the user's query, such as retrieving the data from the reports, performing calculations or aggregations, and formatting the response. The lambda function can also handle errors and validations, and provide feedback to the user.

Use Amazon Lex to return the response to the user, either in text or speech format, depending on the user's preference.

References:

What Is Amazon Lex?

What Is Amazon Comprehend?

What Is Amazon Transcribe?

NO.11 A Machine Learning Specialist is applying a linear least squares regression model to a dataset with 1 000 records and 50 features Prior to training, the ML Specialist notices that two features are perfectly linearly dependent Why could this be an issue for the linear least squares regression model ?

A. It could cause the backpropagation algorithm to fail during training

B. It could create a singular matrix during optimization which fails to define a unique solution

C. It could modify the loss function during optimization causing it to fail during training

D. It could introduce non-linear dependencies within the data which could invalidate the linear assumptions of the model

Answer: B

Explanation:

Linear least squares regression is a method of fitting a linear model to a set of data by minimizing the sum of squared errors between the observed and predicted values. The solution of the linear least squares problem can be obtained by solving the normal equations, which are given by ATAx=ATb, where A is the matrix of explanatory variables, b is the vector of response variables, and x is the vector of unknown coefficients.

However, if the matrix A has two features that are perfectly linearly dependent, then the matrix ATA will be singular, meaning that it does not have a unique inverse. This implies that the normal equations do not have a unique solution, and the linear least squares problem is ill-posed. In other words, there are infinitely many values of x that can satisfy the normal equations, and the linear model is not identifiable.

This can be an issue for the linear least squares regression model, as it can lead to instability, inconsistency, and poor generalization of the model. It can also cause numerical difficulties when trying to solve the normal equations using computational methods, such as matrix inversion or decomposition. Therefore, it is advisable to avoid or remove the linearly dependent features from the matrix A before applying the linear least squares regression model.

References:

Linear least squares (mathematics)

Linear Regression in Matrix Form

Singular Matrix Problem

NO.12 A Machine Learning Specialist uploads a dataset to an Amazon S3 bucket protected with server-side encryption using AWS KMS.

How should the ML Specialist define the Amazon SageMaker notebook instance so it can read the same dataset from Amazon S3?

A. Define security group(s) to allow all HTTP inbound/outbound traffic and assign those security group(s) to the Amazon SageMaker notebook instance.

B. Configure the Amazon SageMaker notebook instance to have access to the VPC. Grant permission in the KMS key policy to the notebook's KMS role.

C. Assign an IAM role to the Amazon SageMaker notebook with S3 read access to the dataset. Grant permission in the KMS key policy to that role.

D. Assign the same KMS key used to encrypt data in Amazon S3 to the Amazon SageMaker notebook instance.

Answer: C

Explanation:

To read data from an Amazon S3 bucket that is protected with server-side encryption using AWS KMS, the Amazon SageMaker notebook instance needs to have an IAM role that has permission to access the S3 bucket and the KMS key. The IAM role is an identity that defines the permissions for the notebook instance to interact with other AWS services. The IAM role can be assigned to the notebook instance when it is created or updated later.

The KMS key policy is a document that specifies who can use and manage the KMS key. The KMS key policy can grant permission to the IAM role of the notebook instance to decrypt the data in the S3 bucket. The KMS key policy can also grant permission to other principals, such as AWS accounts, IAM

users, or IAM roles, to use the KMS key for encryption and decryption operations.
Therefore, the Machine Learning Specialist should assign an IAM role to the Amazon SageMaker notebook with S3 read access to the dataset. Grant permission in the KMS key policy to that role. This way, the notebook instance can use the IAM role credentials to access the S3 bucket and the KMS key, and read the encrypted data from the S3 bucket.
References:
Create an IAM Role to Grant Permissions to Your Notebook Instance
Using Key Policies in AWS KMS

NO.13 A web-based company wants to improve its conversion rate on its landing page Using a large historical dataset of customer visits, the company has repeatedly trained a multi-class deep learning network algorithm on Amazon SageMaker However there is an overfitting problem training data shows 90% accuracy in predictions, while test data shows 70% accuracy only The company needs to boost the generalization of its model before deploying it into production to maximize conversions of visits to purchases Which action is recommended to provide the HIGHEST accuracy model for the company's test and validation data?

A. Increase the randomization of training data in the mini-batches used in training.

B. Allocate a higher proportion of the overall data to the training dataset

C. Apply L1 or L2 regularization and dropouts to the training.

D. Reduce the number of layers and units (or neurons) from the deep learning network.

Answer: C

Explanation:
Regularization and dropouts are techniques that can help reduce overfitting in deep learning models. Overfitting occurs when the model learns too much from the training data and fails to generalize well to new data. Regularization adds a penalty term to the loss function that penalizes the model for having large or complex weights. This prevents the model from memorizing the noise or irrelevant features in the training data. L1 and L2 are two types of regularization that differ in how they calculate the penalty term. L1 regularization uses the absolute value of the weights, while L2 regularization uses the square of the weights. Dropouts are another technique that randomly drops out some units or neurons from the network during training. This creates a thinner network that is less prone to overfitting. Dropouts also act as a form of ensemble learning, where multiple sub-models are combined to produce a better prediction. By applying regularization and dropouts to the training, the web-based company can improve the generalization and accuracy of its deep learning model on the test and validation data. References:
Regularization: A video that explains the concept and benefits of regularization in deep learning.
Dropout: A video that demonstrates how dropout works and why it helps reduce overfitting.

NO.14 A Data Scientist is building a model to predict customer churn using a dataset of 100 continuous numerical features. The Marketing team has not provided any insight about which features are relevant for churn prediction. The Marketing team wants to interpret the model and see the direct impact of relevant features on the model outcome. While training a logistic regression model, the Data Scientist observes that there is a wide gap between the training and validation set accuracy.
Which methods can the Data Scientist use to improve the model performance and satisfy the Marketing team's needs? (Choose two.)

A. Add L1 regularization to the classifier

B. Add features to the dataset

C. Perform recursive feature elimination

D. Perform t-distributed stochastic neighbor embedding (t-SNE)

E. Perform linear discriminant analysis

Answer: A,C

Explanation:

The Data Scientist is building a model to predict customer churn using a dataset of 100 continuous numerical features. The Marketing team wants to interpret the model and see the direct impact of relevant features on the model outcome. However, the Data Scientist observes that there is a wide gap between the training and validation set accuracy, which indicates that the model is overfitting the data and generalizing poorly to new data.

To improve the model performance and satisfy the Marketing team's needs, the Data Scientist can use the following methods:

Add L1 regularization to the classifier: L1 regularization is a technique that adds a penalty term to the loss function of the logistic regression model, proportional to the sum of the absolute values of the coefficients. L1 regularization can help reduce overfitting by shrinking the coefficients of the less important features to zero, effectively performing feature selection. This can simplify the model and make it more interpretable, as well as improve the validation accuracy.

Perform recursive feature elimination: Recursive feature elimination (RFE) is a feature selection technique that involves training a model on a subset of the features, and then iteratively removing the least important features one by one until the desired number of features is reached. The idea behind RFE is to determine the contribution of each feature to the model by measuring how well the model performs when that feature is removed. The features that are most important to the model will have the greatest impact on performance when they are removed. RFE can help improve the model performance by eliminating the irrelevant or redundant features that may cause noise or multicollinearity in the data. RFE can also help the Marketing team understand the direct impact of the relevant features on the model outcome, as the remaining features will have the highest weights in the model.

References:

Regularization for Logistic Regression

Recursive Feature Elimination

NO.15 An aircraft engine manufacturing company is measuring 200 performance metrics in a time-series. Engineers want to detect critical manufacturing defects in near-real time during testing. All of the data needs to be stored for offline analysis.

What approach would be the MOST effective to perform near-real time defect detection?

A. Use AWS IoT Analytics for ingestion, storage, and further analysis. Use Jupyter notebooks from within AWS IoT Analytics to carry out analysis for anomalies.

B. Use Amazon S3 for ingestion, storage, and further analysis. Use an Amazon EMR cluster to carry out Apache Spark ML k-means clustering to determine anomalies.

C. Use Amazon S3 for ingestion, storage, and further analysis. Use the Amazon SageMaker Random Cut Forest (RCF) algorithm to determine anomalies.

D. Use Amazon Kinesis Data Firehose for ingestion and Amazon Kinesis Data Analytics Random Cut Forest (RCF) to perform anomaly detection. Use Kinesis Data Firehose to store data in Amazon S3 for further

Answer: D

analysis.

Explanation:

The company wants to perform near-real time defect detection on a time-series of 200 performance metrics, and store all the data for offline analysis. The best approach for this scenario is to use Amazon Kinesis Data Firehose for ingestion and Amazon Kinesis Data Analytics Random Cut Forest (RCF) to perform anomaly detection. Use Kinesis Data Firehose to store data in Amazon S3 for further analysis.

Amazon Kinesis Data Firehose is a service that can capture, transform, and deliver streaming data to destinations such as Amazon S3, Amazon Redshift, Amazon OpenSearch Service, and Splunk. Kinesis Data Firehose can handle any amount and frequency of data, and automatically scale to match the throughput. Kinesis Data Firehose can also compress, encrypt, and batch the data before delivering it to the destination, reducing the storage cost and enhancing the security.

Amazon Kinesis Data Analytics is a service that can analyze streaming data in real time using SQL or Apache Flink applications. Kinesis Data Analytics can use built-in functions and algorithms to perform various analytics tasks, such as aggregations, joins, filters, windows, and anomaly detection. One of the built-in algorithms that Kinesis Data Analytics supports is Random Cut Forest (RCF), which is a supervised learning algorithm for forecasting scalar time series using recurrent neural networks. RCF can detect anomalies in streaming data by assigning an anomaly score to each data point, based on how distant it is from the rest of the data. RCF can handle multiple related time series, such as the performance metrics of the aircraft engine, and learn a global model that captures the common patterns and trends across the time series.

Therefore, the company can use the following architecture to build the near-real time defect detection solution:

Use Amazon Kinesis Data Firehose for ingestion: The company can use Kinesis Data Firehose to capture the streaming data from the aircraft engine testing, and deliver it to two destinations: Amazon S3 and Amazon Kinesis Data Analytics. The company can configure the Kinesis Data Firehose delivery stream to specify the source, the buffer size and interval, the compression and encryption options, the error handling and retry logic, and the destination details.

Use Amazon Kinesis Data Analytics Random Cut Forest (RCF) to perform anomaly detection: The company can use Kinesis Data Analytics to create a SQL application that can read the streaming data from the Kinesis Data Firehose delivery stream, and apply the RCF algorithm to detect anomalies. The company can use the RANDOM_CUT_FOREST or RANDOM_CUT_FOREST_WITH_EXPLANATION functions to compute the anomaly scores and attributions for each data point, and use the WHERE clause to filter out the normal data points. The company can also use the CURSOR function to specify the input stream, and the PUMP function to write the output stream to another destination, such as Amazon Kinesis Data Streams or AWS Lambda.

Use Kinesis Data Firehose to store data in Amazon S3 for further analysis: The company can use Kinesis Data Firehose to store the raw and processed data in Amazon S3 for offline analysis. The company can use the S3 destination of the Kinesis Data Firehose delivery stream to store the raw data, and use another Kinesis Data Firehose delivery stream to store the output of the Kinesis Data Analytics application. The company can also use AWS Glue or Amazon Athena to catalog, query, and analyze the data in Amazon S3.

References:

What Is Amazon Kinesis Data Firehose?

What Is Amazon Kinesis Data Analytics for SQL Applications?
DeepAR Forecasting Algorithm - Amazon SageMaker

NO.16 A Machine Learning team runs its own training algorithm on Amazon SageMaker. The training algorithm requires external assets. The team needs to submit both its own algorithm code and algorithm-specific parameters to Amazon SageMaker.
What combination of services should the team use to build a custom algorithm in Amazon SageMaker?
(Choose two.)

A. AWS Secrets Manager

B. AWS CodeStar

C. Amazon ECR

D. Amazon ECS

E. Amazon S3

Answer: C,E

Explanation:
The Machine Learning team wants to use its own training algorithm on Amazon SageMaker, and submit both its own algorithm code and algorithm-specific parameters. The best combination of services to build a custom algorithm in Amazon SageMaker are Amazon ECR and Amazon S3.
Amazon ECR is a fully managed container registry service that allows you to store, manage, and deploy Docker container images. You can use Amazon ECR to create a Docker image that contains your training algorithm code and any dependencies or libraries that it requires. You can also use Amazon ECR to push, pull, and manage your Docker images securely and reliably.
Amazon S3 is a durable, scalable, and secure object storage service that can store any amount and type of data. You can use Amazon S3 to store your training data, model artifacts, and algorithm-specific parameters. You can also use Amazon S3 to access your data and parameters from your training algorithm code, and to write your model output to a specified location.
Therefore, the Machine Learning team can use the following steps to build a custom algorithm in Amazon SageMaker:
Write the training algorithm code in Python, using the Amazon SageMaker Python SDK or the Amazon SageMaker Containers library to interact with the Amazon SageMaker service. The code should be able to read the input data and parameters from Amazon S3, and write the model output to Amazon S3.
Create a Dockerfile that defines the base image, the dependencies, the environment variables, and the commands to run the training algorithm code. The Dockerfile should also expose the ports that Amazon SageMaker uses to communicate with the container.
Build the Docker image using the Dockerfile, and tag it with a meaningful name and version.
Push the Docker image to Amazon ECR, and note the registry path of the image.
Upload the training data, model artifacts, and algorithm-specific parameters to Amazon S3, and note the S3 URIs of the objects.
Create an Amazon SageMaker training job, using the Amazon SageMaker Python SDK or the AWS CLI. Specify the registry path of the Docker image, the S3 URIs of the input and output data, the algorithm-specific parameters, and other configuration options, such as the instance type, the number of instances, the IAM role, and the hyperparameters.
Monitor the status and logs of the training job, and retrieve the model output from Amazon S3.
References:

Use Your Own Training Algorithms
Amazon ECR - Amazon Web Services
Amazon S3 - Amazon Web Services

NO.17 A company uses a long short-term memory (LSTM) model to evaluate the risk factors of a particular energy sector. The model reviews multi-page text documents to analyze each sentence of the text and categorize it as either a potential risk or no risk. The model is not performing well, even though the Data Scientist has experimented with many different network structures and tuned the corresponding hyperparameters.
Which approach will provide the MAXIMUM performance boost?

A. Initialize the words by term frequency-inverse document frequency (TF-IDF) vectors pretrained on a large collection of news articles related to the energy sector.

B. Use gated recurrent units (GRUs) instead of LSTM and run the training process until the validation loss stops decreasing.

C. Reduce the learning rate and run the training process until the training loss stops decreasing.

D. Initialize the words by word2vec embeddings pretrained on a large collection of news articles related to the energy sector.

Answer: D
Explanation:
Initializing the words by word2vec embeddings pretrained on a large collection of news articles related to the energy sector will provide the maximum performance boost for the LSTM model. Word2vec is a technique that learns distributed representations of words based on their co-occurrence in a large corpus of text. These representations capture semantic and syntactic similarities between words, which can help the LSTM model better understand the meaning and context of the sentences in the text documents. Using word2vec embeddings that are pretrained on a relevant domain (energy sector) can further improve the performance by reducing the vocabulary mismatch and increasing the coverage of the words in the text documents. References:
AWS Machine Learning Specialty Exam Guide
AWS Machine Learning Training - Text Classification with TF-IDF, LSTM, BERT: a comparison of performance AWS Machine Learning Training - Machine Learning - Exam Preparation Path

NO.18 A Machine Learning Specialist previously trained a logistic regression model using scikit-learn on a local machine, and the Specialist now wants to deploy it to production for inference only.
What steps should be taken to ensure Amazon SageMaker can host a model that was trained locally?

A. Build the Docker image with the inference code. Tag the Docker image with the registry hostname and upload it to Amazon ECR.

B. Serialize the trained model so the format is compressed for deployment. Tag the Docker image with the registry hostname and upload it to Amazon S3.

C. Serialize the trained model so the format is compressed for deployment. Build the image and upload it to Docker Hub.

D. Build the Docker image with the inference code. Configure Docker Hub and upload the image to Amazon ECR.

Answer: A
Explanation:
To deploy a model that was trained locally to Amazon SageMaker, the steps are:

Build the Docker image with the inference code. The inference code should include the model loading, data preprocessing, prediction, and postprocessing logic. The Docker image should also include the dependencies and libraries required by the inference code and the model.
Tag the Docker image with the registry hostname and upload it to Amazon ECR. Amazon ECR is a fully managed container registry that makes it easy to store, manage, and deploy container images. The registry hostname is the Amazon ECR registry URI for your account and Region. You can use the AWS CLI or the Amazon ECR console to tag and push the Docker image to Amazon ECR.
Create a SageMaker model entity that points to the Docker image in Amazon ECR and the model artifacts in Amazon S3. The model entity is a logical representation of the model that contains the information needed to deploy the model for inference. The model artifacts are the files generated by the model training process, such as the model parameters and weights. You can use the AWS CLI, the SageMaker Python SDK, or the SageMaker console to create the model entity.
Create an endpoint configuration that specifies the instance type and number of instances to use for hosting the model. The endpoint configuration also defines the production variants, which are the different versions of the model that you want to deploy. You can use the AWS CLI, the SageMaker Python SDK, or the SageMaker console to create the endpoint configuration.
Create an endpoint that uses the endpoint configuration to deploy the model. The endpoint is a web service that exposes an HTTP API for inference requests. You can use the AWS CLI, the SageMaker Python SDK, or the SageMaker console to create the endpoint.
References:
AWS Machine Learning Specialty Exam Guide
AWS Machine Learning Training - Deploy a Model on Amazon SageMaker
AWS Machine Learning Training - Use Your Own Inference Code with Amazon SageMaker Hosting Services

NO.19 A trucking company is collecting live image data from its fleet of trucks across the globe. The data is growing rapidly and approximately 100 GB of new data is generated every day. The company wants to explore machine learning uses cases while ensuring the data is only accessible to specific IAM users.
Which storage option provides the most processing flexibility and will allow access control with IAM?

A. Use a database, such as Amazon DynamoDB, to store the images, and set the IAM policies to restrict access to only the desired IAM users.

B. Use an Amazon S3-backed data lake to store the raw images, and set up the permissions using bucket policies.

C. Setup up Amazon EMR with Hadoop Distributed File System (HDFS) to store the files, and restrict access to the EMR instances using IAM policies.

D. Configure Amazon EFS with IAM policies to make the data available to Amazon EC2 instances owned by the IAM users.

Answer: B
Explanation:
The best storage option for the trucking company is to use an Amazon S3-backed data lake to store the raw images, and set up the permissions using bucket policies. A data lake is a centralized repository that allows you to store all your structured and unstructured data at any scale. Amazon S3 is the ideal choice for building a data lake because it offers high durability, scalability, availability, and security. You can store any type of data in Amazon S3, such as images, videos, audio, text, etc. You can also use AWS services such as Amazon Rekognition, Amazon SageMaker, and Amazon EMR to

analyze and process the data in the data lake. To ensure the data is only accessible to specific IAM users, you can use bucket policies to grant or deny access to the S3 buckets based on the IAM user's identity or role. Bucket policies are JSON documents that specify the permissions for the bucket and the objects in it. You can use conditions to restrict access based on various factors, such as IP address, time, source, etc. By using bucket policies, you can control who can access the data in the data lake and what actions they can perform on it.
References:
AWS Machine Learning Specialty Exam Guide
AWS Machine Learning Training - Build a Data Lake Foundation with Amazon S3 AWS Machine Learning Training - Using Bucket Policies and User Policies

NO.20 A credit card company wants to build a credit scoring model to help predict whether a new credit card applicant will default on a credit card payment. The company has collected data from a large number of sources with thousands of raw attributes. Early experiments to train a classification model revealed that many attributes are highly correlated, the large number of features slows down the training speed significantly, and that there are some overfitting issues.
The Data Scientist on this project would like to speed up the model training time without losing a lot of information from the original dataset.
Which feature engineering technique should the Data Scientist use to meet the objectives?

A. Run self-correlation on all features and remove highly correlated features

B. Normalize all numerical values to be between 0 and 1

C. Use an autoencoder or principal component analysis (PCA) to replace original features with new features

D. Cluster raw data using k-means and use sample data from each cluster to build a new dataset

Answer: C
Explanation:
The best feature engineering technique to speed up the model training time without losing a lot of information from the original dataset is to use an autoencoder or principal component analysis (PCA) to replace original features with new features. An autoencoder is a type of neural network that learns a compressed representation of the input data, called the latent space, by minimizing the reconstruction error between the input and the output. PCA is a statistical technique that reduces the dimensionality of the data by finding a set of orthogonal axes, called the principal components, that capture the maximum variance of the data. Both techniques can help reduce the number of features and remove the noise and redundancy in the data, which can improve the model performance and speed up the training process. References:
AWS Machine Learning Specialty Exam Guide
AWS Machine Learning Training - Dimensionality Reduction for Machine Learning AWS Machine Learning Training - Deep Learning with Amazon SageMaker

NO.21 A Data Scientist is training a multilayer perception (MLP) on a dataset with multiple classes. The target class of interest is unique compared to the other classes within the dataset, but it does not achieve and acceptable ecall metric. The Data Scientist has already tried varying the number and size of the MLP's hidden layers, which has not significantly improved the results. A solution to improve recall must be implemented as quickly as possible.
Which techniques should be used to meet these requirements?

A. Gather more data using Amazon Mechanical Turk and then retrain

B. Train an anomaly detection model instead of an MLP

C. Train an XGBoost model instead of an MLP

D. Add class weights to the MLP's loss function and then retrain

Answer: D

Explanation:

The best technique to improve the recall of the MLP for the target class of interest is to add class weights to the MLP's loss function and then retrain. Class weights are a way of assigning different importance to each class in the dataset, such that the model will pay more attention to the classes with higher weights. This can help mitigate the class imbalance problem, where the model tends to favor the majority class and ignore the minority class. By increasing the weight of the target class of interest, the model will try to reduce the false negatives and increase the true positives, which will improve the recall metric. Adding class weights to the loss function is also a quick and easy solution, as it does not require gathering more data, changing the model architecture, or switching to a different algorithm.

References:

AWS Machine Learning Specialty Exam Guide

AWS Machine Learning Training - Deep Learning with Amazon SageMaker

AWS Machine Learning Training - Class Imbalance and Weighted Loss Functions

NO.22 A Machine Learning Specialist works for a credit card processing company and needs to predict which transactions may be fraudulent in near-real time. Specifically, the Specialist must train a model that returns the probability that a given transaction may fraudulent.

How should the Specialist frame this business problem?

A. Streaming classification

B. Binary classification

C. Multi-category classification

D. Regression classification

Answer: B

Explanation:

The business problem of predicting whether a new credit card applicant will default on a credit card payment can be framed as a binary classification problem. Binary classification is the task of predicting a discrete class label output for an example, where the class label can only take one of two possible values. In this case, the class label can be either "default" or "no default", indicating whether the applicant will or will not default on a credit card payment. A binary classification model can return the probability that a given applicant belongs to each class, and then assign the applicant to the class with the highest probability. For example, if the model predicts that an applicant has a 0.8 probability of defaulting and a 0.2 probability of not defaulting, then the model will classify the applicant as "default". Binary classification is suitable for this problem because the outcome of interest is categorical and binary, and the model needs to return the probability of each outcome.

References:

AWS Machine Learning Specialty Exam Guide

AWS Machine Learning Training - Classification vs Regression in Machine Learning

NO.23 A real estate company wants to create a machine learning model for predicting housing

prices based on a historical dataset. The dataset contains 32 features.
Which model will meet the business requirement?

A. Logistic regression

B. Linear regression

C. K-means

D. Principal component analysis (PCA)

Answer: B

Explanation:
The best model for predicting housing prices based on a historical dataset with 32 features is linear regression. Linear regression is a supervised learning algorithm that fits a linear relationship between a dependent variable (housing price) and one or more independent variables (features). Linear regression can handle multiple features and output a continuous value for the housing price. Linear regression can also return the coefficients of the features, which indicate how each feature affects the housing price. Linear regression is suitable for this problem because the outcome of interest is numerical and continuous, and the model needs to capture the linear relationship between the features and the outcome.
References:
AWS Machine Learning Specialty Exam Guide
AWS Machine Learning Training - Regression vs Classification in Machine Learning AWS Machine Learning Training - Linear Regression with Amazon SageMaker

NO.24 A Machine Learning Specialist was given a dataset consisting of unlabeled data The Specialist must create a model that can help the team classify the data into different buckets What model should be used to complete this work?

A. K-means clustering

B. Random Cut Forest (RCF)

C. XGBoost

D. BlazingText

Answer: A

Explanation:
K-means clustering is a machine learning technique that can be used to classify unlabeled data into different groups based on their similarity. It is an unsupervised learning method, which means it does not require any prior knowledge or labels for the data. K-means clustering works by randomly assigning data points to a number of clusters, then iteratively updating the cluster centers and reassigning the data points until the clusters are stable. The result is a partition of the data into distinct and homogeneous groups. K-means clustering can be useful for exploratory data analysis, data compression, anomaly detection, and feature extraction. References:
K-Means Clustering: A tutorial on how to use K-means clustering with Amazon SageMaker.
Unsupervised Learning: A video that explains the concept and applications of unsupervised learning.

NO.25 A Machine Learning Specialist wants to bring a custom algorithm to Amazon SageMaker. The Specialist implements the algorithm in a Docker container supported by Amazon SageMaker.
How should the Specialist package the Docker container so that Amazon SageMaker can launch the training correctly?

A. Modify the bash_profile file in the container and add a bash command to start the training program

B. Use CMD config in the Dockerfile to add the training program as a CMD of the image

C. Configure the training program as an ENTRYPOINT named train

D. Copy the training program to directory /opt/ml/train

Answer: C

Explanation:
To use a custom algorithm in Amazon SageMaker, the Docker container image must have an executable file named train that acts as the ENTRYPOINT for the container. This file is responsible for running the training code and communicating with the Amazon SageMaker service. The train file must be in the PATH of the container and have execute permissions. The other options are not valid ways to package the Docker container for Amazon SageMaker. References:
Use Docker containers to build models - Amazon SageMaker
Create a container with your own algorithms and models - Amazon SageMaker

NO.26 A Data Scientist needs to analyze employment dat
a. The dataset contains approximately 10 million
observations on people across 10 different features. During the preliminary analysis, the Data Scientist notices that income and age distributions are not normal. While income levels shows a right skew as expected, with fewer individuals having a higher income, the age distribution also show a right skew, with fewer older individuals participating in the workforce.
Which feature transformations can the Data Scientist apply to fix the incorrectly skewed data? (Choose two.)

A. Cross-validation

B. Numerical value binning

C. High-degree polynomial transformation

D. Logarithmic transformation

E. One hot encoding

Answer: B,D

Explanation:
To fix the incorrectly skewed data, the Data Scientist can apply two feature transformations: numerical value binning and logarithmic transformation. Numerical value binning is a technique that groups continuous values into discrete bins or categories. This can help reduce the skewness of the data by creating more balanced frequency distributions. Logarithmic transformation is a technique that applies the natural logarithm function to each value in the data. This can help reduce the right skewness of the data by compressing the large values and expanding the small values. Both of these transformations can make the data more suitable for machine learning algorithms that assume normality of the data. References:
Data Transformation - Amazon SageMaker
Transforming Skewed Data for Machine Learning

NO.27 A Machine Learning Specialist is given a structured dataset on the shopping habits of a company's customer base. The dataset contains thousands of columns of data and hundreds of numerical columns for each customer. The Specialist wants to identify whether there are natural groupings for these columns across all customers and visualize the results as quickly as possible.
What approach should the Specialist take to accomplish these tasks?

A. Embed the numerical features using the t-distributed stochastic neighbor embedding (t-SNE) algorithm and create a scatter plot.

B. Run k-means using the Euclidean distance measure for different values of k and create an elbow plot.

C. Embed the numerical features using the t-distributed stochastic neighbor embedding (t-SNE) algorithm and create a line graph.

D. Run k-means using the Euclidean distance measure for different values of k and create box plots for each numerical column within each cluster.

Answer: A

Explanation:

The best approach to identify and visualize the natural groupings for the numerical columns across all customers is to embed the numerical features using the t-distributed stochastic neighbor embedding (t-SNE) algorithm and create a scatter plot. t-SNE is a dimensionality reduction technique that can project high-dimensional data into a lower-dimensional space, while preserving the local structure and distances of the data points. A scatter plot can then show the clusters of data points in the reduced space, where each point represents a customer and the color indicates the cluster membership. This approach can help the Specialist quickly explore the patterns and similarities among the customers based on their numerical features.

The other options are not as effective or efficient as the t-SNE approach. Running k-means for different values of k and creating an elbow plot can help determine the optimal number of clusters, but it does not provide a visual representation of the clusters or the customers. Embedding the numerical features using t-SNE and creating a line graph does not make sense, as a line graph is used to show the change of a variable over time, not the distribution of data points in a space. Running k-means for different values of k and creating box plots for each numerical column within each cluster can provide some insights into the statistics of each cluster, but it is very time-consuming and cumbersome to create and compare thousands of box plots. References:

Dimensionality Reduction - Amazon SageMaker

Visualize high dimensional data using t-SNE - Amazon SageMaker

NO.28 A Machine Learning Specialist is planning to create a long-running Amazon EMR cluster. The EMR cluster will have 1 master node, 10 core nodes, and 20 task nodes. To save on costs, the Specialist will use Spot Instances in the EMR cluster.

Which nodes should the Specialist launch on Spot Instances?

A. Master node

B. Any of the core nodes

C. Any of the task nodes

D. Both core and task nodes

Answer: C

Explanation:

The best option for using Spot Instances in a long-running Amazon EMR cluster is to use them for the task nodes. Task nodes are optional nodes that are used to increase the processing power of the cluster. They do not store any data and can be added or removed without affecting the cluster's operation. Therefore, they are more resilient to interruptions caused by Spot Instance termination. Using Spot Instances for the master node or the core nodes is not recommended, as they store important data and metadata for the cluster. If they are terminated, the cluster may fail or lose data.

References:
Amazon EMR on EC2 Spot Instances
Instance purchasing options - Amazon EMR