QUESTION 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. Reference:
Amazon EMR
Amazon SageMaker SDK

QUESTION 2
A Machine Learning Specialist has completed a proof of concept for a company using a small data sample and now the Specialist is ready to implement an end-to-end solution in AWS using Amazon SageMaker The historical training data is stored in Amazon RDS
Which approach should the Specialist use for training a model using that data?

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

Answer: B

Explanation:
Pushing the data from Microsoft SQL Server to Amazon S3 using an AWS Data Pipeline and providing the S3 location within the notebook is the best approach for training a model using the data stored in Amazon RDS. This is because Amazon SageMaker can directly access data from Amazon S3 and train models on it. AWS Data Pipeline is a service that can automate the movement and transformation of data between different AWS services. It can also use Amazon RDS as a data source and Amazon S3 as a data destination. This way, the data can be transferred efficiently and securely without writing any

code within the notebook.
Reference:
Amazon SageMaker
AWS Data Pipeline

QUESTION 3
Which of the following metrics should a Machine Learning Specialist generally use to compare/evaluate machine learning classification models against each other?

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

Answer: D

Explanation:
Area Under the ROC Curve (AUC) is a metric that measures the performance of a binary classifier across all possible thresholds. It is also known as the probability that a randomly chosen positive example will be ranked higher than a randomly chosen negative example by the classifier. AUC is a good metric to compare different classification models because it is independent of the class distribution and the decision threshold. It also captures both the sensitivity (true positive rate) and the specificity (true negative rate) of the model. Reference:
AWS Machine Learning Specialty Exam Guide
AWS Machine Learning Specialty Sample Questions

QUESTION 4
A Machine Learning Specialist is using Amazon Sage Maker to host a model for a highly available customer-facing application.
The Specialist has trained a new version of the model, validated it with historical data, and now wants to deploy it to production To limit any risk of a negative customer experience, the Specialist wants to be able to monitor the model and roll it back, if needed
What is the SIMPLEST approach with the LEAST risk to deploy the model and roll it back, if needed?

A. Create a SageMaker endpoint and configuration for the new model version. Redirect production traffic to the new endpoint by updating the client configuration. Revert traffic to the last version if the model does not perform as expected.
B. Create a SageMaker endpoint and configuration for the new model version. Redirect production traffic to the new endpoint by using a load balancer Revert traffic to the last version if the model does not perform as expected.
C. Update the existing SageMaker endpoint to use a new configuration that is weighted to send 5% of the traffic to the new variant. Revert traffic to the last version by resetting the weights if the model

does not perform as expected.
D. Update the existing SageMaker endpoint to use a new configuration that is weighted to send 100% of the traffic to the new variant Revert traffic to the last version by resetting the weights if the model does not perform as expected.

Answer: C

Explanation:
Updating the existing SageMaker endpoint to use a new configuration that is weighted to send 5% of the traffic to the new variant is the simplest approach with the least risk to deploy the model and roll it back, if needed. This is because SageMaker supports A/B testing, which allows the Specialist to compare the performance of different model variants by sending a portion of the traffic to each variant. The Specialist can monitor the metrics of each variant and adjust the weights accordingly. If the new variant does not perform as expected, the Specialist can revert traffic to the last version by resetting the weights to 100% for the old variant and 0% for the new variant. This way, the Specialist can deploy the model without affecting the customer experience and roll it back easily if needed. Reference:
Amazon SageMaker
Deploying models to Amazon SageMaker hosting services

QUESTION 5
A manufacturing company has a large set of labeled historical sales data The manufacturer would like to predict how many units of a particular part should be produced each quarter Which machine learning approach should be used to solve this problem?

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

Answer: D

Explanation:
Linear regression is a machine learning approach that can be used to solve this problem. Linear regression is a supervised learning technique that can model the relationship between one or more input variables (features) and an output variable (target). In this case, the input variables could be the historical sales data of the part, such as the quarter, the demand, the price, the inventory, etc. The output variable could be the number of units to be produced for the part. Linear regression can learn the coefficients (weights) of the input variables that best fit the output variable, and then use them to make predictions for new data. Linear regression is suitable for problems that involve continuous and numeric output variables, such as predicting house prices, stock prices, or sales volumes. Reference:

QUESTION 6
A manufacturing company has structured and unstructured data stored in an Amazon S3 bucket A Machine Learning Specialist wants to use SQL to run queries on this dat
a. Which solution requires the LEAST effort to be able to query this data?

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

Answer: B

Explanation:
AWS Glue is a serverless data integration service that can catalogue, clean, enrich, and move data between various data stores. Amazon Athena is an interactive query service that can run SQL queries on data stored in Amazon S3. By using AWS Glue to catalogue the data and Amazon Athena to run queries, the Machine Learning Specialist can leverage the existing data in Amazon S3 without any additional data transformation or loading. This solution requires the least effort compared to the other options, which involve more complex and costly data processing and storage services. Reference: AWS Glue, Amazon Athena

---

QUESTION 7
A Machine Learning Specialist is packaging a custom ResNet model into a Docker container so the company can leverage Amazon SageMaker for training The Specialist is using Amazon EC2 P3 instances to train the model and needs to properly configure the Docker container to leverage the NVIDIA GPUs
What does the Specialist need to do1?

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

Answer: B

Explanation:
To leverage the NVIDIA GPUs on Amazon EC2 P3 instances, the Machine Learning Specialist needs to build the Docker container to be NVIDIA-Docker compatible. NVIDIA-Docker is a tool that enables GPU-accelerated containers to run on Docker. It automatically configures the container to access the

NVIDIA drivers and libraries on the host system. The Specialist does not need to bundle the NVIDIA drivers with the Docker image, as they are already installed on the EC2 P3 instances. The Specialist does not need to organize the Docker containers file structure to execute on GPU instances, as this is not relevant for GPU compatibility. The Specialist does not need to set the GPU flag in the Amazon SageMaker Create TrainingJob request body, as this is only required for using Elastic Inference accelerators, not EC2 P3 instances. Reference: NVIDIA-Docker, Using GPU-Accelerated Containers, Using Elastic Inference in Amazon SageMaker

---

QUESTION 8
A large JSON dataset for a project has been uploaded to a private Amazon S3 bucket The Machine Learning Specialist wants to securely access and explore the data from an Amazon SageMaker notebook instance A new VPC was created and assigned to the Specialist
How can the privacy and integrity of the data stored in Amazon S3 be maintained while granting access to the Specialist for analysis?

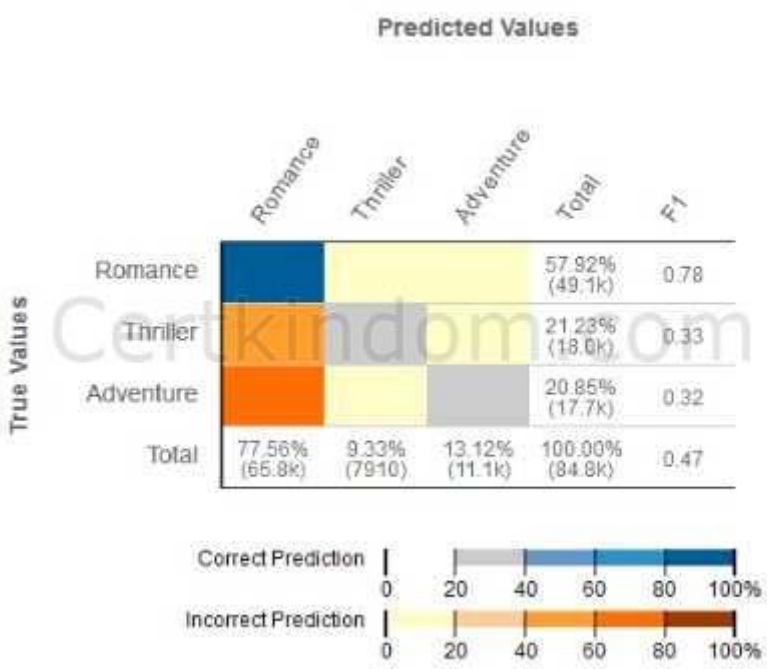
A. Launch the SageMaker notebook instance within the VPC with SageMaker-provided internet access enabled Use an S3 ACL to open read privileges to the everyone group
B. Launch the SageMaker notebook instance within the VPC and create an S3 VPC endpoint for the notebook to access the data Copy the JSON dataset from Amazon S3 into the ML storage volume on the SageMaker notebook instance and work against the local dataset
C. Launch the SageMaker notebook instance within the VPC and create an S3 VPC endpoint for the notebook to access the data Define a custom S3 bucket policy to only allow requests from your VPC to access the S3 bucket
D. Launch the SageMaker notebook instance within the VPC with SageMaker-provided internet access enabled. Generate an S3 pre-signed URL for access to data in the bucket

Answer: C

Explanation:
The best way to maintain the privacy and integrity of the data stored in Amazon S3 is to use a combination of VPC endpoints and S3 bucket policies. A VPC endpoint allows the SageMaker notebook instance to access the S3 bucket without going through the public internet. A bucket policy allows the S3 bucket owner to specify which VPCs or VPC endpoints can access the bucket. This way, the data is protected from unauthorized access and tampering. The other options are either insecure (A and D) or inefficient (B). Reference: Using Amazon S3 VPC Endpoints, Using Bucket Policies and User Policies

---

QUESTION 9
Given the following confusion matrix for a movie classification model, what is the true class frequency for Romance and the predicted class frequency for Adventure?



A. The true class frequency for Romance is 77.56% and the predicted class frequency for Adventure is 20 85%
B. The true class frequency for Romance is 57.92% and the predicted class frequency for Adventure is 1312%
C. The true class frequency for Romance is 0 78 and the predicted class frequency for Adventure is (0 47 - 0.32).
D. The true class frequency for Romance is 77.56% * 0.78 and the predicted class frequency for Adventure is 20 85% ' 0.32

Answer: B

Explanation:
The true class frequency for Romance is the percentage of movies that are actually Romance out of all the movies. This can be calculated by dividing the sum of the true values for Romance by the total number of movies. The predicted class frequency for Adventure is the percentage of movies that are predicted to be Adventure out of all the movies. This can be calculated by dividing the sum of the predicted values for Adventure by the total number of movies. Based on the confusion matrix, the true class frequency for Romance is 57.92% and the predicted class frequency for Adventure is 13.12%. Reference: Confusion Matrix, Classification Metrics

---

QUESTION 10
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 customers 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. Reference: Binary Classification, Time Series Prediction, Anomaly Detection, Regression

---

QUESTION 11
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. Reference:
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.

QUESTION 12
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. Reference:
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.

QUESTION 13
A retail company intends to use machine learning to categorize new products A labeled dataset of current products was provided to the Data Science team The dataset includes 1 200 products The labeled dataset has 15 features for each product such as title dimensions, weight, and price Each product is labeled as belonging to one of six categories such as books, games, electronics, and movies.

Which model should be used for categorizing new products using the provided dataset for training?

A. An XGBoost model where the objective parameter is set to multi: softmax
B. A deep convolutional neural network (CNN) with a softmax activation function for the last layer
C. A regression forest where the number of trees is set equal to the number of product categories
D. A DeepAR forecasting model based on a recurrent neural network (RNN)

Answer: A

Explanation:
XGBoost is a machine learning framework that can be used for classification, regression, ranking, and other tasks. It is based on the gradient boosting algorithm, which builds an ensemble of weak learners (usually decision trees) to produce a strong learner. XGBoost has several advantages over other algorithms, such as scalability, parallelization, regularization, and sparsity handling. For categorizing new products using the provided dataset, an XGBoost model would be a suitable choice, because it can handle multiple features and multiple classes efficiently and accurately. To train an XGBoost model for multi-class classification, the objective parameter should be set to multi: softmax, which means that the model will output a probability distribution over the classes and predict the class with the highest probability. Alternatively, the objective parameter can be set to multi: softprob, which means that the model will output the raw probability of each class instead of the predicted class label. This can be useful for evaluating the model performance or for post-processing the predictions. Reference:
XGBoost: A tutorial on how to use XGBoost with Amazon SageMaker.
XGBoost Parameters: A reference guide for the parameters of XGBoost.

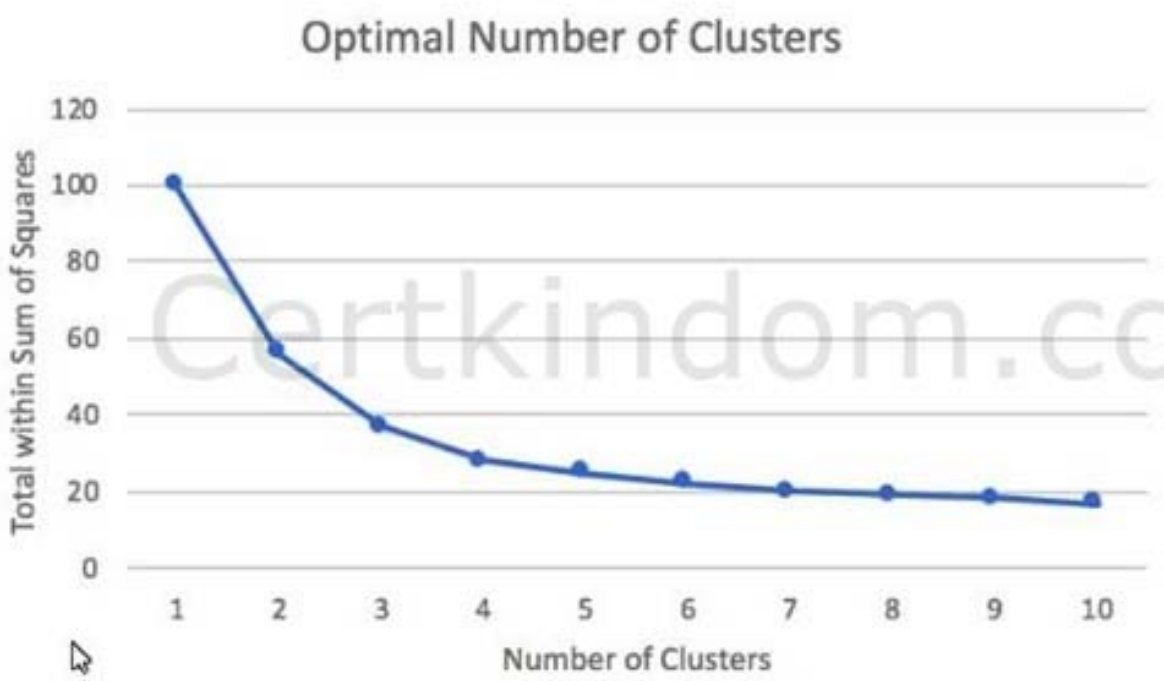
QUESTION 14
A Machine Learning Specialist is building a model to predict future employment rates based on a wide range of economic factors While exploring the data, the Specialist notices that the magnitude of the input features vary greatly The Specialist does not want variables with a larger magnitude to dominate the model
What should the Specialist do to prepare the data for model training'?

A. Apply quantile binning to group the data into categorical bins to keep any relationships in the data by replacing the magnitude with distribution
B. Apply the Cartesian product transformation to create new combinations of fields that are independent of the magnitude
C. Apply normalization to ensure each field will have a mean of 0 and a variance of 1 to remove any significant magnitude
D. Apply the orthogonal sparse Diagram (OSB) transformation to apply a fixed-size sliding window to generate new features of a similar magnitude.

Answer: C

Explanation:
Normalization is a data preprocessing technique that can be used to scale the input features to a common range, such as [-1, 1] or [0, 1]. Normalization can help reduce the effect of outliers, improve the convergence of gradient-based algorithms, and prevent variables with a larger magnitude from dominating the model. One common method of normalization is standardization, which transforms each feature to have a mean of 0 and a variance of 1. This can be done by subtracting the mean and dividing by the standard deviation of each feature. Standardization can be useful for models that assume the input features are normally distributed, such as linear regression, logistic regression, and support vector machines. Reference:
Data normalization and standardization: A video that explains the concept and benefits of data normalization and standardization.
Standardize or Normalize?: A blog post that compares different methods of scaling the input features.

QUESTION 15
A Machine Learning Specialist prepared the following graph displaying the results of k-means for k = [1:10]



Considering the graph, what is a reasonable selection for the optimal choice of k?

A. 1
B. 4

C. 7
D. 10

Answer: B

Explanation:
The elbow method is a technique that we use to determine the number of centroids (k) to use in a kmeans clustering algorithm. In this method, we plot the within-cluster sum of squares (WCSS) against the number of clusters (k) and look for the point where the curve bends sharply. This point is called the elbow point and it indicates that adding more clusters does not improve the model significantly. The graph in the question shows that the elbow point is at k = 4, which means that 4 is a reasonable choice for the optimal number of clusters. Reference:
Elbow Method for optimal value of k in KMeans: A tutorial on how to use the elbow method with Amazon SageMaker.
K-Means Clustering: A video that explains the concept and benefits of k-means clustering.

QUESTION 16
A company is using Amazon Polly to translate plaintext documents to speech for automated company announcements However company acronyms are being mispronounced in the current documents How should a Machine Learning Specialist address this issue for future documents?

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

Answer: B

Explanation:
A pronunciation lexicon is a file that defines how words or phrases should be pronounced by Amazon Polly. A lexicon can help customize the speech output for words that are uncommon, foreign, or have multiple pronunciations. A lexicon must conform to the Pronunciation Lexicon Specification (PLS) standard and can be stored in an AWS region using the Amazon Polly API. To use a lexicon for synthesizing speech, the lexicon name must be specified in the <speak> SSML tag. For example, the following lexicon defines how to pronounce the acronym W3C:
<lexicon version=œ1.0 xmlns=œhttp://www.w3.org5/01/pronunciation-lexicon alphabet=œipa xml:lang=œen-US > <lexeme> <grapheme>W3C</grapheme> <alias>World Wide Web Consortium</alias> </lexeme> </lexicon>
To use this lexicon, the text input must include the following SSML tag:
<speak version=œ1.1 xmlns=œhttp://www.w3.org1/synthesis xml:lang=œen-US > <voice name=œJoanna > <lexicon name=œw3c_lexicon /> The <say-as interpret-as=œcharacters >W3C</sayas>

is an international community that develops open standards to ensure the long-term growth of the Web. </voice> </speak>
Reference:
Customize pronunciation using lexicons in Amazon Polly: A blog post that explains how to use lexicons for creating custom pronunciations.
Managing Lexicons: A documentation page that describes how to store and retrieve lexicons using the Amazon Polly API.

QUESTION 17
A Machine Learning Specialist is using Apache Spark for pre-processing training data As part of the Spark pipeline, the Specialist wants to use Amazon SageMaker for training a model and hosting it Which of the following would the Specialist do to integrate the Spark application with SageMaker? (Select THREE)

A. Download the AWS SDK for the Spark environment
B. Install the SageMaker Spark library in the Spark environment.
C. Use the appropriate estimator from the SageMaker Spark Library to train a model.
D. Compress the training data into a ZIP file and upload it to a pre-defined Amazon S3 bucket.
E. Use the sageMakerModel. transform method to get inferences from the model hosted in SageMaker
F. Convert the DataFrame object to a CSV file, and use the CSV file as input for obtaining inferences from SageMaker.

Answer: B, C, E

Explanation:
The SageMaker Spark library is a library that enables Apache Spark applications to integrate with Amazon SageMaker for training and hosting machine learning models. The library provides several features, such as:
Estimators: Classes that allow Spark users to train Amazon SageMaker models and host them on Amazon SageMaker endpoints using the Spark MLlib Pipelines API. The library supports various builtin algorithms, such as linear learner, XGBoost, K-means, etc., as well as custom algorithms using Docker containers.
Model classes: Classes that wrap Amazon SageMaker models in a Spark MLlib Model abstraction. This allows Spark users to use Amazon SageMaker endpoints for inference within Spark applications.
Data sources: Classes that allow Spark users to read data from Amazon S3 using the Spark Data Sources API. The library supports various data formats, such as CSV, LibSVM, RecordIO, etc.
To integrate the Spark application with SageMaker, the Machine Learning Specialist should do the following:
Install the SageMaker Spark library in the Spark environment. This can be done by using Maven, pip, or downloading the JAR file from GitHub.

Use the appropriate estimator from the SageMaker Spark Library to train a model. For example, to train a linear learner model, the Specialist can use the following code:

Python

```
from sagemaker_pyspark import IAMRole, S3DataPath
from sagemaker_pyspark.algorithms import LinearLearnerSageMakerEstimator

# Create an IAM role for SageMaker
role = IAMRole("arn:aws:iam::account-id:role/role-name")

# Create an estimator for linear learner
linear_learner_estimator = LinearLearnerSageMakerEstimator(
    trainingInstanceType="ml.m4.xlarge",
    trainingInstanceCount=1,
    endpointInstanceType="ml.m4.xlarge",
    endpointInitialInstanceCount=1,
    sagemakerRole=role)

# Set the hyperparameters
linear_learner_estimator.setFeatureDim(10)
linear_learner_estimator.setMiniBatchSize(200)
linear_learner_estimator.setPredictorType("regressor")

# Train the model using a DataFrame
model = linear_learner_estimator.fit(trainingData)
```


Use the sageMakerModel. transform method to get inferences from the model hosted in SageMaker. For example, to get predictions for a test DataFrame, the Specialist can use the following code:

Reference:
[SageMaker Spark]: A documentation page that introduces the SageMaker Spark library and its features.
[SageMaker Spark GitHub Repository]: A GitHub repository that contains the source code, examples, and installation instructions for the SageMaker Spark library.

QUESTION 18
A Machine Learning Specialist is working with a large cybersecurily company that manages security events in real time for companies around the world The cybersecurity company wants to design a solution that will allow it to use machine learning to score malicious events as anomalies on the data as it is being ingested The company also wants be able to save the results in its data lake for later processing and analysis
What is the MOST efficient way to accomplish these tasks'?

A. Ingest the data using Amazon Kinesis Data Firehose, and use Amazon Kinesis Data Analytics Random Cut Forest (RCF) for anomaly detection Then use Kinesis Data Firehose to stream the results to Amazon S3
B. Ingest the data into Apache Spark Streaming using Amazon EMR. and use Spark MLlib with kmeans to perform anomaly detection Then store the results in an Apache Hadoop Distributed File System (HDFS) using Amazon EMR with a replication factor of three as the data lake
C. Ingest the data and store it in Amazon S3 Use AWS Batch along with the AWS Deep Learning AMIs to train a k-means model using TensorFlow on the data in Amazon S3.
D. Ingest the data and store it in Amazon S3. Have an AWS Glue job that is triggered on demand transform the new data Then use the built-in Random Cut Forest (RCF) model within Amazon SageMaker to detect anomalies in the data

Answer: A

Explanation:
Amazon Kinesis Data Firehose is a fully managed service that can capture, transform, and load

streaming data into AWS data stores, such as Amazon S3, Amazon Redshift, Amazon Elasticsearch Service, and Splunk. It can also invoke AWS Lambda functions to perform custom transformations on the data. Amazon Kinesis Data Analytics is a service that can analyze streaming data in real time using SQL or Apache Flink applications. It can also use machine learning algorithms, such as Random Cut Forest (RCF), to perform anomaly detection on streaming data. RCF is an unsupervised learning algorithm that assigns an anomaly score to each data point based on how different it is from the rest of the data. By using Kinesis Data Firehose and Kinesis Data Analytics, the cybersecurity company can ingest the data in real time, score the malicious events as anomalies, and stream the results to Amazon S3, which can serve as a data lake for later processing and analysis. This is the most efficient way to accomplish these tasks, as it does not require any additional infrastructure, coding, or training.
Reference:
Amazon Kinesis Data Firehose - Amazon Web Services
Amazon Kinesis Data Analytics - Amazon Web Services
Anomaly Detection with Amazon Kinesis Data Analytics - Amazon Web Services
[AWS Certified Machine Learning - Specialty Sample Questions]

QUESTION 19
A Machine Learning Specialist works for a credit card processing company and needs to predict which transactions may be fraudulent in near-real time. Specifically, the Specialist must train a model that returns the probability that a given transaction may be fraudulent
How should the Specialist frame this business problem'?

A. Streaming classification
B. Binary classification
C. Multi-category classification
D. Regression classification

Answer: B

Explanation:
Binary classification is a type of supervised learning problem where the goal is to predict a categorical label that has only two possible values, such as Yes or No, True or False, Positive or Negative. In this case, the label is whether a transaction is fraudulent or not, which is a binary outcome. Binary classification can be used to estimate the probability of an observation belonging to a certain class, such as the probability of a transaction being fraudulent. This can help the business to make decisions based on the risk level of each transaction. Reference:
Binary Classification - Amazon Machine Learning
AWS Certified Machine Learning - Specialty Sample Questions

QUESTION 20
Amazon Connect has recently been tolled out across a company as a contact call center The solution

has been configured to store voice call recordings on Amazon S3
The content of the voice calls are being analyzed for the incidents being discussed by the call operators Amazon Transcribe is being used to convert the audio to text, and the output is stored on Amazon S3
Which approach will provide the information required for further analysis?

A. Use Amazon Comprehend with the transcribed files to build the key topics
B. Use Amazon Translate with the transcribed files to train and build a model for the key topics
C. Use the AWS Deep Learning AMI with Gluon Semantic Segmentation on the transcribed files to train and build a model for the key topics
D. Use the Amazon SageMaker k-Nearest-Neighbors (kNN) algorithm on the transcribed files to generate a word embeddings dictionary for the key topics

Answer: A

Explanation:
Amazon Comprehend is a natural language processing (NLP) service that uses machine learning to find insights and relationships in text. It can analyze text documents and identify the key topics, entities, sentiments, languages, and more. In this case, Amazon Comprehend can be used with the transcribed files from Amazon Transcribe to extract the main topics that are being discussed by the call operators. This can help to understand the common issues and concerns of the customers, and provide insights for further analysis and improvement. Reference:
Amazon Comprehend - Amazon Web Services
AWS Certified Machine Learning - Specialty Sample Questions

QUESTION 21
A Machine Learning Specialist is building a prediction model for a large number of features using linear models, such as linear regression and logistic regression During exploratory data analysis the Specialist observes that many features are highly correlated with each other This may make the model unstable
What should be done to reduce the impact of having such a large number of features?

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

Answer: C

Explanation:
Principal component analysis (PCA) is an unsupervised machine learning algorithm that attempts to reduce the dimensionality (number of features) within a dataset while still retaining as much

information as possible. This is done by finding a new set of features called components, which are composites of the original features that are uncorrelated with one another. They are also constrained so that the first component accounts for the largest possible variability in the data, the second component the second most variability, and so on. By using PCA, the impact of having a large number of features that are highly correlated with each other can be reduced, as the new feature space will have fewer dimensions and less redundancy. This can make the linear models more stable and less prone to overfitting. Reference:
Principal Component Analysis (PCA) Algorithm - Amazon SageMaker
Perform a large-scale principal component analysis faster using Amazon SageMaker | AWS Machine Learning Blog
Machine Learning- Prinicipal Component Analysis | i2tutorials

---

QUESTION 22
A Machine Learning Specialist wants to determine the appropriate SageMaker Variant Invocations Per Instance setting for an endpoint automatic scaling configuration. The Specialist has performed a load test on a single instance and determined that peak requests per second (RPS) without service degradation is about 20 RPS As this is the first deployment, the Specialist intends to set the invocation safety factor to 0 5
Based on the stated parameters and given that the invocations per instance setting is measured on a per-minute basis, what should the Specialist set as the sageMaker variant invocations Per instance setting?

A. 10
B. 30
C. 600
D. 2,400

Answer: C

Explanation:
The SageMaker Variant Invocations Per Instance setting is the target value for the average number of invocations per instance per minute for the model variant. It is used by the automatic scaling policy to add or remove instances to keep the metric close to the specified value. To determine this value, the following equation can be used in combination with load testing:
SageMakerVariantInvocationsPerInstance = (MAX_RPS * SAFETY_FACTOR) * 60
Where MAX_RPS is the maximum requests per second that the model variant can handle without service degradation, SAFETY_FACTOR is a factor that ensures that the clients do not exceed the maximum RPS, and 60 is the conversion factor from seconds to minutes. In this case, the given parameters are:
MAX_RPS = 20 SAFETY_FACTOR = 0.5
Plugging these values into the equation, we get:
SageMakerVariantInvocationsPerInstance = (20 * 0.5) * 60 SageMakerVariantInvocationsPerInstance

= 600
Therefore, the Specialist should set the SageMaker Variant Invocations Per Instance setting to 600.
Reference:
Load testing your auto scaling configuration - Amazon SageMaker
Configure model auto scaling with the console - Amazon SageMaker

---

QUESTION 23
A Machine Learning Specialist deployed a model that provides product recommendations on a company's website Initially, the model was performing very well and resulted in customers buying more products on average However within the past few months the Specialist has noticed that the effect of product recommendations has diminished and customers are starting to return to their original habits of spending less The Specialist is unsure of what happened, as the model has not changed from its initial deployment over a year ago
Which method should the Specialist try to improve model performance?

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

Answer: D

Explanation:
The problem that the Machine Learning Specialist is facing is likely due to concept drift, which is a phenomenon where the statistical properties of the target variable change over time, making the model less accurate and relevant. Concept drift can occur due to various reasons, such as changes in customer preferences, market trends, product inventory, seasonality, etc. In this case, the product recommendations model may have become outdated as the product inventory changed over time, making the recommendations less appealing to the customers. To address this issue, the model should be periodically retrained using the original training data plus new data as product inventory changes. This way, the model can learn from the latest data and adapt to the changing customer behavior and preferences. Retraining the model from scratch using the original data while adding a regularization term may not be sufficient, as it does not account for the new data. Updating the models hyperparameters may not help either, as it does not address the underlying data distribution change. Re-engineering the model completely may not be necessary, as the model may still be valid and useful with periodic retraining.
Reference:
Concept Drift - Amazon SageMaker

Detecting and Handling Concept Drift - Amazon SageMaker
Machine Learning Concepts - Amazon Machine Learning

QUESTION 24
A manufacturer of car engines collects data from cars as they are being driven The data collected includes timestamp, engine temperature, rotations per minute (RPM), and other sensor readings The company wants to predict when an engine is going to have a problem so it can notify drivers in advance to get engine maintenance The engine data is loaded into a data lake for training
Which is the MOST suitable predictive model that can be deployed into production'?

A. Add labels over time to indicate which engine faults occur at what time in the future to turn this into a supervised learning problem Use a recurrent neural network (RNN) to train the model to recognize when an engine might need maintenance for a certain fault.
B. This data requires an unsupervised learning algorithm Use Amazon SageMaker k-means to cluster the data
C. Add labels over time to indicate which engine faults occur at what time in the future to turn this into a supervised learning problem Use a convolutional neural network (CNN) to train the model to recognize when an engine might need maintenance for a certain fault.
D. This data is already formulated as a time series Use Amazon SageMaker seq2seq to model the time series.

Answer: A

Explanation:
A recurrent neural network (RNN) is a type of neural network that can process sequential data, such as time series, by maintaining a hidden state that captures the temporal dependencies between the inputs. RNNs are well suited for predicting future events based on past observations, such as forecasting engine failures based on sensor readings. To train an RNN model, the data needs to be labeled with the target variable, which in this case is the type and time of the engine fault. This makes the problem a supervised learning problem, where the goal is to learn a mapping from the input sequence (sensor readings) to the output sequence (engine faults). By using an RNN model, the manufacturer can leverage the temporal information in the data and detect patterns that indicate when an engine might need maintenance for a certain fault.
Reference:
Recurrent Neural Networks - Amazon SageMaker
Use Amazon SageMaker Built-in Algorithms or Pre-trained Models
Recurrent Neural Network Definition | DeepAI
What are Recurrent Neural Networks? An Ultimate Guide for Newbies!
Lee and Carter go Machine Learning: Recurrent Neural Networks - SSRN

QUESTION 25
A Data Scientist is working on an application that performs sentiment analysis. The validation

accuracy is poor and the Data Scientist thinks that the cause may be a rich vocabulary and a low average frequency of words in the dataset
Which tool should be used to improve the validation accuracy?

A. Amazon Comprehend syntax analysts and entity detection
B. Amazon SageMaker BlazingText allow mode
C. Natural Language Toolkit (NLTK) stemming and stop word removal
D. Scikit-learn term frequency-inverse document frequency (TF-IDF) vectorizers

Answer: D

Explanation:
Term frequency-inverse document frequency (TF-IDF) is a technique that assigns a weight to each word in a document based on how important it is to the meaning of the document. The term frequency (TF) measures how often a word appears in a document, while the inverse document frequency (IDF) measures how rare a word is across a collection of documents. The TF-IDF weight is the product of the TF and IDF values, and it is high for words that are frequent in a specific document but rare in the overall corpus. TF-IDF can help improve the validation accuracy of a sentiment analysis model by reducing the impact of common words that have little or no sentiment value, such as œthe , œa , œand , etc. Scikit-learn is a popular Python library for machine learning that provides a TF-IDF vectorizer class that can transform a collection of text documents into a matrix of TF-IDF features. By using this tool, the Data Scientist can create a more informative and discriminative feature representation for the sentiment analysis task.
Reference:
TfidfVectorizer - scikit-learn
Text feature extraction - scikit-learn
TF-IDF for Beginners | by Jana Schmidt | Towards Data Science
Sentiment Analysis: Concept, Analysis and Applications | by Susan Li | Towards Data Science

---

QUESTION 26
A Machine Learning Specialist is developing recommendation engine for a photography blog Given a picture, the recommendation engine should show a picture that captures similar objects The Specialist would like to create a numerical representation feature to perform nearest-neighbor searches
What actions would allow the Specialist to get relevant numerical representations?

A. Reduce image resolution and use reduced resolution pixel values as features
B. Use Amazon Mechanical Turk to label image content and create a one-hot representation indicating the presence of specific labels
C. Run images through a neural network pie-trained on ImageNet, and collect the feature vectors from the penultimate layer
D. Average colors by channel to obtain three-dimensional representations of images.

Answer: C

Explanation:
A neural network pre-trained on ImageNet is a deep learning model that has been trained on a large dataset of images containing 1000 classes of objects. The model can learn to extract high-level features from the images that capture the semantic and visual information of the objects. The penultimate layer of the model is the layer before the final output layer, and it contains a feature vector that represents the input image in a lower-dimensional space. By running images through a pre-trained neural network and collecting the feature vectors from the penultimate layer, the Specialist can obtain relevant numerical representations that can be used for nearest-neighbor searches. The feature vectors can capture the similarity between images based on the presence and appearance of similar objects, and they can be compared using distance metrics such as Euclidean distance or cosine similarity. This approach can enable the recommendation engine to show a picture that captures similar objects to a given picture.
Reference:
ImageNet - Wikipedia
How to use a pre-trained neural network to extract features from images | by Rishabh Anand | Analytics Vidhya | Medium
Image Similarity using Deep Ranking | by Aditya Oke | Towards Data Science

---

QUESTION 27
A gaming company has launched an online game where people can start playing for free but they need to pay if they choose to use certain features The company needs to build an automated system to predict whether or not a new user will become a paid user within 1 year The company has gathered a labeled dataset from 1 million users
The training dataset consists of 1.000 positive samples (from users who ended up paying within 1 year) and 999.000 negative samples (from users who did not use any paid features) Each data sample consists of 200 features including user age, device, location, and play patterns
Using this dataset for training, the Data Science team trained a random forest model that converged with over 99% accuracy on the training set However, the prediction results on a test dataset were not satisfactory.
Which of the following approaches should the Data Science team take to mitigate this issue? (Select TWO.)

A. Add more deep trees to the random forest to enable the model to learn more features.
B. indicate a copy of the samples in the test database in the training dataset
C. Generate more positive samples by duplicating the positive samples and adding a small amount of noise to the duplicated data.
D. Change the cost function so that false negatives have a higher impact on the cost value than false positives
E. Change the cost function so that false positives have a higher impact on the cost value than false

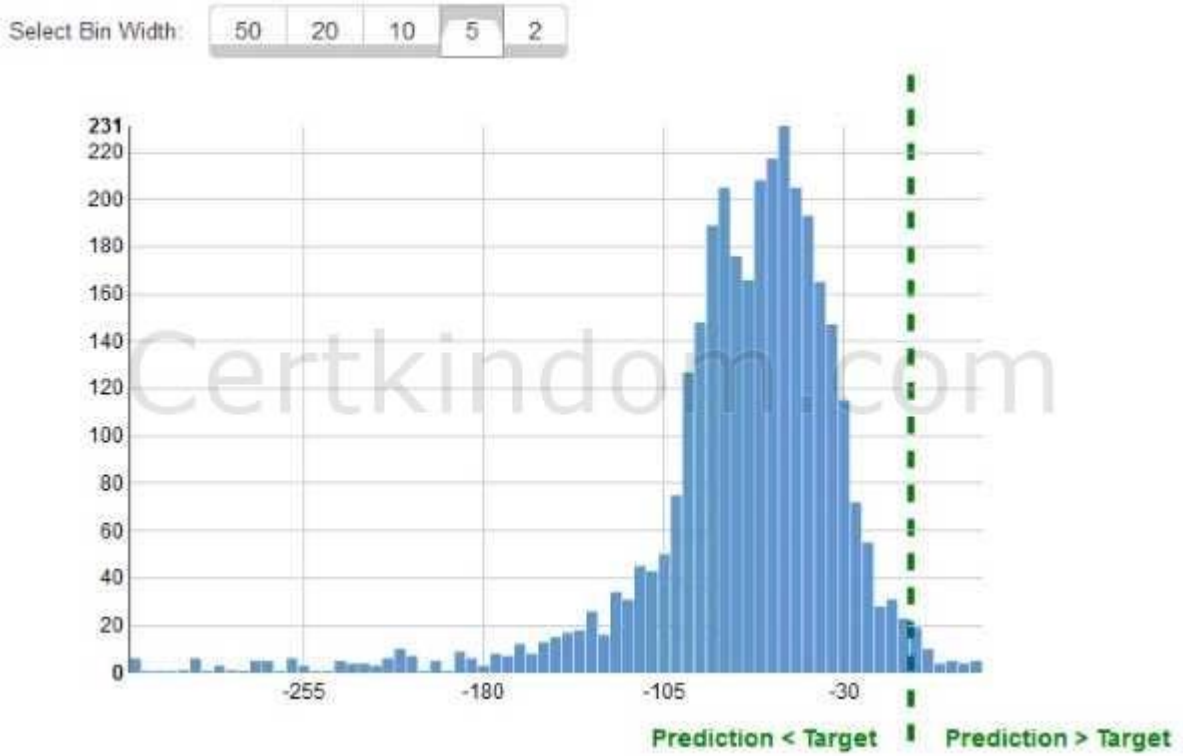
negatives

Answer: C, D

Explanation:
The Data Science team is facing a problem of imbalanced data, where the positive class (paid users) is much less frequent than the negative class (non-paid users). This can cause the random forest model to be biased towards the majority class and have poor performance on the minority class. To mitigate this issue, the Data Science team can try the following approaches:
C) Generate more positive samples by duplicating the positive samples and adding a small amount of noise to the duplicated data. This is a technique called data augmentation, which can help increase the size and diversity of the training data for the minority class. This can help the random forest model learn more features and patterns from the positive class and reduce the imbalance ratio.
D) Change the cost function so that false negatives have a higher impact on the cost value than false positives. This is a technique called cost-sensitive learning, which can assign different weights or costs to different classes or errors. By assigning a higher cost to false negatives (predicting non-paid when the user is actually paid), the random forest model can be more sensitive to the minority class and try to minimize the misclassification of the positive class.
Reference:
Bagging and Random Forest for Imbalanced Classification
Surviving in a Random Forest with Imbalanced Datasets
machine learning - random forest for imbalanced data? - Cross Validated
Biased Random Forest For Dealing With the Class Imbalance Problem

---

QUESTION 28
While reviewing the histogram for residuals on regression evaluation data a Machine Learning Specialist notices that the residuals do not form a zero-centered bell shape as shown What does this mean?



A. The model might have prediction errors over a range of target values.
B. The dataset cannot be accurately represented using the regression model
C. There are too many variables in the model
D. The model is predicting its target values perfectly.

Answer: A

Explanation:
Residuals are the differences between the actual and predicted values of the target variable in a regression model. A histogram of residuals is a graphical tool that can help evaluate the performance and assumptions of the model. Ideally, the histogram of residuals should have a zero-centered bell shape, which indicates that the residuals are normally distributed with a mean of zero and a constant variance. This means that the model has captured the true relationship between the input and output variables, and that the errors are random and unbiased. However, if the histogram of residuals does not have a zero-centered bell shape, as shown in the image, this means that the model might have prediction errors over a range of target values. This is because the residuals do not form a symmetrical and homogeneous distribution around zero, which implies that the model has some systematic bias or heteroscedasticity. This can affect the accuracy and validity of the model, and indicate that the model needs to be improved or modified.
Reference:
Residual Analysis in Regression - Statistics By Jim

How to Check Residual Plots for Regression Analysis - dummies
Histogram of Residuals - Statistics How To

QUESTION 29
During mini-batch training of a neural network for a classification problem, a Data Scientist notices that training accuracy oscillates What is the MOST likely cause of this issue?

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

Answer: D

Explanation:
Mini-batch gradient descent is a variant of gradient descent that updates the model parameters using a subset of the training data (called a mini-batch) at each iteration. The learning rate is a hyperparameter that controls how much the model parameters change in response to the gradient. If the learning rate is very high, the model parameters may overshoot the optimal values and oscillate around the minimum of the cost function. This can cause the training accuracy to fluctuate and prevent the model from converging to a stable solution. To avoid this issue, the learning rate should be chosen carefully, such as by using a learning rate decay schedule or an adaptive learning rate algorithm1. Alternatively, the batch size can be increased to reduce the variance of the gradient estimates2. However, the batch size should not be too big, as this can slow down the training process and reduce the generalization ability of the model3. Dataset shuffling and class distribution are not likely to cause oscillations in training accuracy, as they do not affect the gradient updates directly. Dataset shuffling can help avoid getting stuck in local minima and improve the convergence speed of mini-batch gradient descent4. Class distribution can affect the performance and fairness of the model, especially if the dataset is imbalanced, but it does not necessarily cause fluctuations in training accuracy.

QUESTION 30
A Machine Learning Specialist observes several performance problems with the training portion of a machine learning solution on Amazon SageMaker The solution uses a large training dataset 2 TB in size and is using the SageMaker k-means algorithm The observed issues include the unacceptable length of time it takes before the training job launches and poor I/O throughput while training the model
What should the Specialist do to address the performance issues with the current solution?

A. Use the SageMaker batch transform feature
B. Compress the training data into Apache Parquet format.
C. Ensure that the input mode for the training job is set to Pipe.

D. Copy the training dataset to an Amazon EFS volume mounted on the SageMaker instance.

Answer: C

Explanation:
The input mode for the training job determines how the training data is transferred from Amazon S3 to the SageMaker instance. There are two input modes: File and Pipe. File mode copies the entire training dataset from S3 to the local file system of the instance before starting the training job. This can cause a long delay before the training job launches, especially if the dataset is large. Pipe mode streams the data from S3 to the instance as the training job runs. This can reduce the startup time and improve the I/O throughput, as the data is read in smaller batches. Therefore, to address the performance issues with the current solution, the Specialist should ensure that the input mode for the training job is set to Pipe. This can be done by using the SageMaker Python SDK and setting the input_mode parameter to Pipe when creating the estimator or the fit method12. Alternatively, this can be done by using the AWS CLI and setting the InputMode parameter to Pipe when creating the training job3.
Reference:
Access Training Data - Amazon SageMaker
Choosing Data Input Mode Using the SageMaker Python SDK - Amazon SageMaker
CreateTrainingJob - Amazon SageMaker Service

QUESTION 31
A Machine Learning Specialist is building a convolutional neural network (CNN) that will classify 10 types of animals. The Specialist has built a series of layers in a neural network that will take an input image of an animal, pass it through a series of convolutional and pooling layers, and then finally pass it through a dense and fully connected layer with 10 nodes The Specialist would like to get an output from the neural network that is a probability distribution of how likely it is that the input image belongs to each of the 10 classes
Which function will produce the desired output?

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

Answer: C

Explanation:
The softmax function is a function that can transform a vector of arbitrary real values into a vector of real values in the range (0,1) that sum to 1. This means that the softmax function can produce a valid probability distribution over multiple classes. The softmax function is often used as the activation function of the output layer in a neural network, especially for multi-class classification problems.

The softmax function can assign higher probabilities to the classes with higher scores, which allows the network to make predictions based on the most likely class. In this case, the Machine Learning Specialist wants to get an output from the neural network that is a probability distribution of how likely it is that the input image belongs to each of the 10 classes of animals. Therefore, the softmax function is the most suitable function to produce the desired output.
Reference:
Softmax Activation Function for Deep Learning: A Complete Guide
What is Softmax in Machine Learning? - reason.town
machine learning - Why is the softmax function often used as activation ¦
Multi-Class Neural Networks: Softmax | Machine Learning | Google for ¦

QUESTION 32
A Machine Learning Specialist is building a model that will perform time series forecasting using Amazon SageMaker The Specialist has finished training the model and is now planning to perform load testing on the endpoint so they can configure Auto Scaling for the model variant
Which approach will allow the Specialist to review the latency, memory utilization, and CPU utilization during the load test"?

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

Answer: B

Explanation:
Amazon CloudWatch is a service that can monitor and collect various metrics and logs from AWS resources, such as Amazon SageMaker. Amazon CloudWatch can also generate dashboards to create a single view for the metrics and logs that are of interest. By using Amazon CloudWatch, the Machine Learning Specialist can review the latency, memory utilization, and CPU utilization during the load test, as these are some of the metrics that are outputted by Amazon SageMaker. The Specialist can create a custom dashboard that displays these metrics in different widgets, such as graphs, tables, or text. The dashboard can also be configured to refresh automatically and show the latest data as the load test is running. This approach will allow the Specialist to monitor the performance and resource utilization of the model variant and adjust the Auto Scaling configuration accordingly.
Reference:
[Monitoring Amazon SageMaker with Amazon CloudWatch - Amazon SageMaker]

[Using Amazon CloudWatch Dashboards - Amazon CloudWatch]
[Create a CloudWatch Dashboard - Amazon CloudWatch]

QUESTION 33
An Amazon SageMaker notebook instance is launched into Amazon VPC The SageMaker notebook references data contained in an Amazon S3 bucket in another account The bucket is encrypted using SSE-KMS The instance returns an access denied error when trying to access data in Amazon S3.
Which of the following are required to access the bucket and avoid the access denied error? (Select THREE)

A. An AWS KMS key policy that allows access to the customer master key (CMK)
B. A SageMaker notebook security group that allows access to Amazon S3
C. An 1AM role that allows access to the specific S3 bucket
D. A permissive S3 bucket policy
E. An S3 bucket owner that matches the notebook owner
F. A SegaMaker notebook subnet ACL that allow traffic to Amazon S3.

Answer: A, B, C

Explanation:
To access an Amazon S3 bucket in another account that is encrypted using SSE-KMS, the following are required:
A) An AWS KMS key policy that allows access to the customer master key (CMK). The CMK is the encryption key that is used to encrypt and decrypt the data in the S3 bucket. The KMS key policy defines who can use and manage the CMK. To allow access to the CMK from another account, the key policy must include a statement that grants the necessary permissions (such as kms:Decrypt) to the principal from the other account (such as the SageMaker notebook IAM role).
B) A SageMaker notebook security group that allows access to Amazon S3. A security group is a virtual firewall that controls the inbound and outbound traffic for the SageMaker notebook instance. To allow the notebook instance to access the S3 bucket, the security group must have a rule that allows outbound traffic to the S3 endpoint on port 443 (HTTPS).
C) An IAM role that allows access to the specific S3 bucket. An IAM role is an identity that can be assumed by the SageMaker notebook instance to access AWS resources. The IAM role must have a policy that grants the necessary permissions (such as s3:GetObject) to access the specific S3 bucket. The policy must also include a condition that allows access to the CMK in the other account.
The following are not required or correct:
D) A permissive S3 bucket policy. A bucket policy is a resource-based policy that defines who can access the S3 bucket and what actions they can perform. A permissive bucket policy is not required and not recommended, as it can expose the bucket to unauthorized access. A bucket policy should follow the principle of least privilege and grant the minimum permissions necessary to the specific principals that need access.
E) An S3 bucket owner that matches the notebook owner. The S3 bucket owner and the notebook

owner do not need to match, as long as the bucket owner grants cross-account access to the notebook owner through the KMS key policy and the bucket policy (if applicable).
F) A SegaMaker notebook subnet ACL that allow traffic to Amazon S3. A subnet ACL is a network access control list that acts as an optional layer of security for the SageMaker notebook instances subnet. A subnet ACL is not required to access the S3 bucket, as the security group is sufficient to control the traffic. However, if a subnet ACL is used, it must not block the traffic to the S3 endpoint.

QUESTION 34
A monitoring service generates 1 TB of scale metrics record data every minute A Research team performs queries on this data using Amazon Athena The queries run slowly due to the large volume of data, and the team requires better performance
How should the records be stored in Amazon S3 to improve query performance?

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

Answer: B

Explanation:
Parquet is a columnar storage format that can store data in a compressed and efficient way. Parquet files can improve query performance by reducing the amount of data that needs to be scanned, as only the relevant columns are read from the files. Parquet files can also support predicate pushdown, which means that the filtering conditions are applied at the storage level, further reducing the data that needs to be processed. Parquet files are compatible with Amazon Athena, which can leverage the benefits of the columnar format and provide faster and cheaper queries. Therefore, the records should be stored in Parquet files in Amazon S3 to improve query performance.
Reference:
Columnar Storage Formats - Amazon Athena
Parquet SerDe - Amazon Athena
Optimizing Amazon Athena Queries - Amazon Athena
Parquet - Apache Software Foundation

QUESTION 35
A Machine Learning Specialist needs to create a data repository to hold a large amount of time-based training data for a new model. In the source system, new files are added every hour Throughout a single 24-hour period, the volume of hourly updates will change significantly. The Specialist always wants to train on the last 24 hours of the data
Which type of data repository is the MOST cost-effective solution?

A. An Amazon EBS-backed Amazon EC2 instance with hourly directories

B. An Amazon RDS database with hourly table partitions
C. An Amazon S3 data lake with hourly object prefixes
D. An Amazon EMR cluster with hourly hive partitions on Amazon EBS volumes

Answer: C

Explanation:
An Amazon S3 data lake is a cost-effective solution for storing and analyzing large amounts of timebased training data for a new model. Amazon S3 is a highly scalable, durable, and secure object storage service that can store any amount of data in any format. Amazon S3 also offers low-cost storage classes, such as S3 Standard-IA and S3 One Zone-IA, that can reduce the storage costs for infrequently accessed data. By using hourly object prefixes, the Machine Learning Specialist can organize the data into logical partitions based on the time of ingestion. This can enable efficient data access and management, as well as support incremental updates and deletes. The Specialist can also use Amazon S3 lifecycle policies to automatically transition the data to lower-cost storage classes or delete the data after a certain period of time. This way, the Specialist can always train on the last 24 hours of the data and optimize the storage costs.
Reference:
What is a data lake? - Amazon Web Services
Amazon S3 Storage Classes - Amazon Simple Storage Service
Managing your storage lifecycle - Amazon Simple Storage Service
Best Practices Design Patterns: Optimizing Amazon S3 Performance

---

QUESTION 36
A retail chain has been ingesting purchasing records from its network of 20,000 stores to Amazon S3 using Amazon Kinesis Data Firehose To support training an improved machine learning model, training records will require new but simple transformations, and some attributes will be combined The model needs lo be retrained daily
Given the large number of stores and the legacy data ingestion, which change will require the LEAST amount of development effort?

A. Require that the stores to switch to capturing their data locally on AWS Storage Gateway for loading into Amazon S3 then use AWS Glue to do the transformation
B. Deploy an Amazon EMR cluster running Apache Spark with the transformation logic, and have the cluster run each day on the accumulating records in Amazon S3, outputting new/transformed records to Amazon S3
C. Spin up a fleet of Amazon EC2 instances with the transformation logic, have them transform the data records accumulating on Amazon S3, and output the transformed records to Amazon S3.
D. Insert an Amazon Kinesis Data Analytics stream downstream of the Kinesis Data Firehouse stream that transforms raw record attributes into simple transformed values using SQL.

Answer: D

Explanation:
Amazon Kinesis Data Analytics is a service that can analyze streaming data in real time using SQL or Apache Flink applications. It can also use machine learning algorithms, such as Random Cut Forest (RCF), to perform anomaly detection on streaming data. By inserting a Kinesis Data Analytics stream downstream of the Kinesis Data Firehose stream, the retail chain can transform the raw record attributes into simple transformed values using SQL queries. This can be done without changing the existing data ingestion process or deploying additional resources. The transformed records can then be outputted to another Kinesis Data Firehose stream that delivers them to Amazon S3 for training the machine learning model. This approach will require the least amount of development effort, as it leverages the existing Kinesis Data Firehose stream and the built-in SQL capabilities of Kinesis Data Analytics.
Reference:
Amazon Kinesis Data Analytics - Amazon Web Services
Anomaly Detection with Amazon Kinesis Data Analytics - Amazon Web Services
Amazon Kinesis Data Firehose - Amazon Web Services
Amazon S3 - Amazon Web Services

QUESTION 37
A city wants to monitor its air quality to address the consequences of air pollution A Machine Learning Specialist needs to forecast the air quality in parts per million of contaminates for the next 2 days in the city as this is a prototype, only daily data from the last year is available
Which model is MOST likely to provide the best results in Amazon SageMaker?

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

Answer: A

Explanation:
The Amazon SageMaker k-Nearest-Neighbors (kNN) algorithm is a supervised learning algorithm that can perform both classification and regression tasks. It can also handle time series data, such as the

air quality data in this case. The kNN algorithm works by finding the k most similar instances in the training data to a given query instance, and then predicting the output based on the average or majority of the outputs of the k nearest neighbors. The kNN algorithm can be configured to use different distance metrics, such as Euclidean or cosine, to measure the similarity between instances. To use the kNN algorithm on the single time series consisting of the full year of data, the Machine Learning Specialist needs to set the predictor_type parameter to regressor, as the output variable (air quality in parts per million of contaminates) is a continuous value. The kNN algorithm can then forecast the air quality for the next 2 days by finding the k most similar days in the past year and averaging their air quality values.
Reference:
Amazon SageMaker k-Nearest-Neighbors (kNN) Algorithm - Amazon SageMaker
Time Series Forecasting using k-Nearest Neighbors (kNN) in Python | by ¦
Time Series Forecasting with k-Nearest Neighbors | by Nishant Malik ¦

QUESTION 38
For the given confusion matrix, what is the recall and precision of the model?

A. Recall = 0.92 Precision = 0.84
B. Recall = 0.84 Precision = 0.8
C. Recall = 0.92 Precision = 0.8
D. Recall = 0.8 Precision = 0.92

Answer: C

Explanation:
Recall and precision are two metrics that can be used to evaluate the performance of a classification model. Recall is the ratio of true positives to the total number of actual positives, which measures how well the model can identify all the relevant cases. Precision is the ratio of true positives to the total number of predicted positives, which measures how accurate the model is when it makes a positive prediction. Based on the confusion matrix in the image, we can calculate the recall and precision as follows:
Recall = TP / (TP + FN) = 12 / (12 + 1) = 0.92
Precision = TP / (TP + FP) = 12 / (12 + 3) = 0.8
Where TP is the number of true positives, FN is the number of false negatives, and FP is the number of false positives. Therefore, the recall and precision of the model are 0.92 and 0.8, respectively.

QUESTION 39
A Machine Learning Specialist is working with a media company to perform classification on popular articles from the company's website. The company is using random forests to classify how popular an article will be before it is published A sample of the data being used is below.
Given the dataset, the Specialist wants to convert the Day-Of_Week column to binary values.
What technique should be used to convert this column to binary values.

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

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

Answer: B

Explanation:
One-hot encoding is a technique that can be used to convert a categorical variable, such as the Day-Of_Week column, to binary values. One-hot encoding creates a new binary column for each unique value in the original column, and assigns a value of 1 to the column that corresponds to the value in the original column, and 0 to the rest. For example, if the original column has values Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, and Sunday, one-hot encoding will create seven new columns, each representing one day of the week. If the value in the original column is Tuesday, then the column for Tuesday will have a value of 1, and the other columns will have a value of 0. One-hot encoding can help improve the performance of machine learning models, as it eliminates the ordinal relationship between the values and creates a more informative and sparse representation of the data.

Reference:
One-Hot Encoding - Amazon SageMaker
One-Hot Encoding: A Simple Guide for Beginners | by Jana Schmidt ¦
One-Hot Encoding in Machine Learning | by Nishant Malik | Towards ¦

QUESTION 40
A company has raw user and transaction data stored in AmazonS3 a MySQL database, and Amazon RedShift A Data Scientist needs to perform an analysis by joining the three datasets from Amazon S3, MySQL, and Amazon RedShift, and then calculating the average-of a few selected columns from the joined data
Which AWS service should the Data Scientist use?

A. Amazon Athena
B. Amazon Redshift Spectrum
C. AWS Glue
D. Amazon QuickSight

Answer: A

Explanation:
Amazon Athena is a serverless interactive query service that can analyze data in Amazon S3 using standard SQL. Amazon Athena can also query data from other sources, such as MySQL and Amazon Redshift, by using federated queries. Federated queries allow Amazon Athena to run SQL queries across data sources, such as relational and non-relational databases, data warehouses, and data lakes. By using Amazon Athena, the Data Scientist can perform an analysis by joining the three datasets from Amazon S3, MySQL, and Amazon Redshift, and then calculating the average of a few selected columns from the joined data. Amazon Athena can also integrate with other AWS services, such as AWS Glue and Amazon QuickSight, to provide additional features, such as data cataloging and visualization.
Reference:
What is Amazon Athena? - Amazon Athena
Federated Query Overview - Amazon Athena
Querying Data from Amazon S3 - Amazon Athena
Querying Data from MySQL - Amazon Athena
[Querying Data from Amazon Redshift - Amazon Athena]

QUESTION 41
A Mobile Network Operator is building an analytics platform to analyze and optimize a company's operations using Amazon Athena and Amazon S3
The source systems send data in CSV format in real lime The Data Engineering team wants to transform the data to the Apache Parquet format before storing it on Amazon S3
Which solution takes the LEAST effort to implement?

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

Answer: D

Explanation:
Amazon Kinesis Data Streams is a service that can capture, store, and process streaming data in real time. Amazon Kinesis Data Firehose is a service that can deliver streaming data to various destinations, such as Amazon S3, Amazon Redshift, or Amazon Elasticsearch Service. Amazon Kinesis Data Firehose can also transform the data before delivering it, such as converting the data format, compressing the data, or encrypting the data. One of the supported data formats that Amazon Kinesis Data Firehose can convert to is Apache Parquet, which is a columnar storage format that can improve the performance and cost-efficiency of analytics queries. By using Amazon Kinesis Data Streams and Amazon Kinesis Data Firehose, the Mobile Network Operator can ingest the .CSV data from the source systems and use Amazon Kinesis Data Firehose to convert the data into Parquet before storing it on Amazon S3. This solution takes the least effort to implement, as it does not require any additional resources, such as Amazon EC2 instances, Amazon EMR clusters, or Amazon Glue jobs. The solution can also leverage the built-in features of Amazon Kinesis Data Firehose, such as data buffering, batching, retry, and error handling.
Reference:
Amazon Kinesis Data Streams - Amazon Web Services
Amazon Kinesis Data Firehose - Amazon Web Services
Data Transformation - Amazon Kinesis Data Firehose
Apache Parquet - Amazon Athena

QUESTION 42
An e-commerce company needs a customized training model to classify images of its shirts and pants products The company needs a proof of concept in 2 to 3 days with good accuracy Which compute choice should the Machine Learning Specialist select to train and achieve good accuracy on the model quickly?

A. m5 4xlarge (general purpose)

B. r5.2xlarge (memory optimized)
C. p3.2xlarge (GPU accelerated computing)
D. p3 8xlarge (GPU accelerated computing)

Answer: C

Explanation:
Image classification is a machine learning task that involves assigning labels to images based on their content. Image classification can be performed using various algorithms, such as convolutional neural networks (CNNs), which are a type of deep learning model that can learn to extract high-level features from images. To train a customized image classification model, the e-commerce company needs a compute choice that can support the high computational demands of deep learning and provide good accuracy on the model quickly. A GPU accelerated computing instance, such as p3.2xlarge, is a suitable choice for this task, as it can leverage the parallel processing power of GPUs to speed up the training process and reduce the training time. A p3.2xlarge instance has one NVIDIA Tesla V100 GPU, which can provide up to 125 teraflops of mixed-precision performance and 16 GB of GPU memory. A p3.2xlarge instance can also use various deep learning frameworks, such as TensorFlow, PyTorch, MXNet, etc., to build and train the image classification model. A p3.2xlarge instance is also more cost-effective than a p3.8xlarge instance, which has four NVIDIA Tesla V100 GPUs, as the latter may not be necessary for a proof of concept with a small dataset. Therefore, the Machine Learning Specialist should select p3.2xlarge as the compute choice to train and achieve good accuracy on the model quickly.
Reference:
Amazon EC2 P3 Instances - Amazon Web Services
Image Classification - Amazon SageMaker
Convolutional Neural Networks - Amazon SageMaker
Deep Learning AMIs - Amazon Web Services

QUESTION 43
A Marketing Manager at a pet insurance company plans to launch a targeted marketing campaign on social media to acquire new customers Currently, the company has the following data in Amazon Aurora
Profiles for all past and existing customers
Profiles for all past and existing insured pets
Policy-level information
Premiums received
Claims paid
What steps should be taken to implement a machine learning model to identify potential new customers on social media?

A. Use regression on customer profile data to understand key characteristics of consumer segments
Find similar profiles on social media.

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

Answer: B

Explanation:
Clustering is a machine learning technique that can group data points into clusters based on their similarity or proximity. Clustering can help discover the underlying structure and patterns in the data, as well as identify outliers or anomalies. Clustering can also be used for customer segmentation, which is the process of dividing customers into groups based on their characteristics, behaviors, preferences, or needs. Customer segmentation can help understand the key features and needs of different customer segments, as well as design and implement targeted marketing campaigns for each segment. In this case, the Marketing Manager at a pet insurance company plans to launch a targeted marketing campaign on social media to acquire new customers. To do this, the Manager can use clustering on customer profile data to understand the key characteristics of consumer segments, such as their demographics, pet types, policy preferences, premiums paid, claims made, etc. The Manager can then find similar profiles on social media, such as Facebook, Twitter, Instagram, etc., by using the cluster features as filters or keywords. The Manager can then target these potential new customers with personalized and relevant ads or offers that match their segments needs and interests. This way, the Manager can implement a machine learning model to identify potential new customers on social media.

QUESTION 44
A company is running an Amazon SageMaker training job that will access data stored in its Amazon S3 bucket A compliance policy requires that the data never be transmitted across the internet How should the company set up the job?

A. Launch the notebook instances in a public subnet and access the data through the public S3 endpoint
B. Launch the notebook instances in a private subnet and access the data through a NAT gateway
C. Launch the notebook instances in a public subnet and access the data through a NAT gateway
D. Launch the notebook instances in a private subnet and access the data through an S3 VPC endpoint.

Answer: D

Explanation:
A private subnet is a subnet that does not have a route to the internet gateway, which means that

the resources in the private subnet cannot access the internet or be accessed from the internet. An S3 VPC endpoint is a gateway endpoint that allows the resources in the VPC to access the S3 service without going through the internet. By launching the notebook instances in a private subnet and accessing the data through an S3 VPC endpoint, the company can set up the job in a secure and compliant way, as the data never leaves the AWS network and is not exposed to the internet. This can also improve the performance and reliability of the data transfer, as the traffic does not depend on the internet bandwidth or availability.
Reference:
Amazon VPC Endpoints - Amazon Virtual Private Cloud
Endpoints for Amazon S3 - Amazon Virtual Private Cloud
Connect to SageMaker Within your VPC - Amazon SageMaker
Working with VPCs and Subnets - Amazon Virtual Private Cloud

QUESTION 45
A Machine Learning Specialist is preparing data for training on Amazon SageMaker The Specialist is transformed into a numpy .array, which appears to be negatively affecting the speed of the training What should the Specialist do to optimize the data for training on SageMaker'?

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

Answer: C

Explanation:
The Recordio protobuf format is a binary data format that is optimized for training on SageMaker. It allows faster data loading and lower memory usage compared to other formats such as CSV or numpy arrays. The Recordio protobuf format also supports features such as sparse input, variablelength input, and label embedding. To use the Recordio protobuf format, the data needs to be serialized and deserialized using the appropriate libraries. Some of the built-in algorithms in SageMaker support the Recordio protobuf format as a content type for training and inference. Reference:
Common Data Formats for Training
Using RecordIO Format
Content Types Supported by Built-in Algorithms

QUESTION 46
A Machine Learning Specialist is training a model to identify the make and model of vehicles in images The Specialist wants to use transfer learning and an existing model trained on images of general objects The Specialist collated a large custom dataset of pictures containing different vehicle makes and models.

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

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

Answer: B

Explanation:
Transfer learning is a technique that allows us to use a model trained for a certain task as a starting point for a machine learning model for a different task. For image classification, a common practice is to use a pre-trained model that was trained on a large and general dataset, such as ImageNet, and then customize it for the specific task. One way to customize the model is to replace the last fully connected layer, which is responsible for the final classification, with a new layer that has the same number of units as the number of classes in the new task. This way, the model can leverage the features learned by the previous layers, which are generic and useful for many image recognition tasks, and learn to map them to the new classes. The new layer can be initialized with random weights, and the rest of the model can be initialized with the pre-trained weights. This method is also known as feature extraction, as it extracts meaningful features from the pre-trained model and uses them for the new task. Reference:
Transfer learning and fine-tuning
Deep transfer learning for image classification: a survey

QUESTION 47
A Machine Learning Specialist is developing a custom video recommendation model for an application The dataset used to train this model is very large with millions of data points and is hosted in an Amazon S3 bucket The Specialist wants to avoid loading all of this data onto an Amazon SageMaker notebook instance because it would take hours to move and will exceed the attached 5 GB Amazon EBS volume on the notebook instance.
Which approach allows the Specialist to use all the data to train the model?

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

compatible
with Amazon SageMaker. Initiate a SageMaker training job using the full dataset from the S3 bucket using
Pipe input mode.
D. Load a smaller subset of the data into the SageMaker notebook and train locally. Confirm that the training
code is executing and the model parameters seem reasonable. Launch an Amazon EC2 instance with an
AWS Deep Learning AMI and attach the S3 bucket to train the full dataset.

Answer: A

Explanation:
Pipe input mode is a feature of Amazon SageMaker that allows streaming large datasets from Amazon S3 directly to the training algorithm without downloading them to the local disk. This reduces the startup time, disk space, and cost of training jobs. Pipe input mode is supported by most of the built-in algorithms and can also be used with custom training algorithms. To use Pipe input mode, the data needs to be in a binary format such as protobuf recordIO or TFRecord. The training code needs to use the PipeModeDataset class to read the data from the named pipe provided by SageMaker. To verify that the training code and the model parameters are working as expected, it is recommended to train locally on a smaller subset of the data before launching a full-scale training job on SageMaker. This approach is faster and more efficient than the other options, which involve either downloading the full dataset to an EC2 instance or using AWS Glue, which is not designed for training machine learning models. Reference:
Using Pipe input mode for Amazon SageMaker algorithms
Using Pipe Mode with Your Own Algorithms
PipeModeDataset Class

QUESTION 48
A Machine Learning Specialist is creating a new natural language processing application that processes a dataset comprised of 1 million sentences The aim is to then run Word2Vec to generate embeddings of the sentences and enable different types of predictions
Here is an example from the dataset
"The quck BROWN FOX jumps over the lazy dog "
Which of the following are the operations the Specialist needs to perform to correctly sanitize and prepare the data in a repeatable manner? (Select THREE)

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

F. Tokenize the sentence into words.

Answer: B, C, F

Explanation:
To prepare the data for Word2Vec, the Specialist needs to perform some preprocessing steps that can help reduce the noise and complexity of the data, as well as improve the quality of the embeddings. Some of the common preprocessing steps for Word2Vec are:
Normalizing all words by making the sentence lowercase: This can help reduce the vocabulary size and treat words with different capitalizations as the same word. For example, œFox and œfox should be considered as the same word, not two different words.
Removing stop words using an English stopword dictionary: Stop words are words that are very common and do not carry much semantic meaning, such as œthe , œa , œand , etc. Removing them can help focus on the words that are more relevant and informative for the task.
Tokenizing the sentence into words: Tokenization is the process of splitting a sentence into smaller units, such as words or subwords. This is necessary for Word2Vec, as it operates on the word level and requires a list of words as input.
The other options are not necessary or appropriate for Word2Vec:
Performing part-of-speech tagging and keeping the action verb and the nouns only: Part-of-speech tagging is the process of assigning a grammatical category to each word, such as noun, verb, adjective, etc. This can be useful for some natural language processing tasks, but not for Word2Vec, as it can lose some important information and context by discarding other words.
Correcting the typography on œquck to œquick : Typo correction can be helpful for some tasks, but not for Word2Vec, as it can introduce errors and inconsistencies in the data. For example, if the typo is intentional or part of a dialect, correcting it can change the meaning or style of the sentence. Moreover, Word2Vec can learn to handle typos and variations in spelling by learning similar embeddings for them.
One-hot encoding all words in the sentence: One-hot encoding is a way of representing words as vectors of 0s and 1s, where only one element is 1 and the rest are 0. The index of the 1 element corresponds to the words position in the vocabulary. For example, if the vocabulary is [œcat , œdog , œfox], then œcat can be encoded as [1, 0, 0], œdog as [0, 1, 0], and œfox as [0, 0, 1]. This can be useful for some machine learning models, but not for Word2Vec, as it does not capture the semantic similarity and relationship between words. Word2Vec aims to learn dense and low-dimensional embeddings for words, where similar words have similar vectors.

QUESTION 49
This graph shows the training and validation loss against the epochs for a neural network
The network being trained is as follows
Two dense layers one output neuron
100 neurons in each layer
100 epochs
Random initialization of weights

Which technique can be used to improve model performance in terms of accuracy in the validation set?

A. Early stopping
B. Random initialization of weights with appropriate seed
C. Increasing the number of epochs
D. Adding another layer with the 100 neurons

Answer: A

Explanation:
Early stopping is a technique that can be used to prevent overfitting and improve model performance on the validation set. Overfitting occurs when the model learns the training data too well and fails to generalize to new and unseen data. This can be seen in the graph, where the training loss keeps decreasing, but the validation loss starts to increase after some point. This means that the model is fitting the noise and patterns in the training data that are not relevant for the validation

data. Early stopping is a way of stopping the training process before the model overfits the training data. It works by monitoring the validation loss and stopping the training when the validation loss stops decreasing or starts increasing. This way, the model is saved at the point where it has the best performance on the validation set. Early stopping can also save time and resources by reducing the number of epochs needed for training. Reference:
Early Stopping
How to Stop Training Deep Neural Networks At the Right Time Using Early Stopping

---

QUESTION 50
A manufacturing company asks its Machine Learning Specialist to develop a model that classifies defective parts into one of eight defect types. The company has provided roughly 100000 images per defect type for training During the injial training of the image classification model the Specialist notices that the validation accuracy is 80%, while the training accuracy is 90% It is known that human-level performance for this type of image classification is around 90%
What should the Specialist consider to fix this issue1?

A. A longer training time
B. Making the network larger
C. Using a different optimizer
D. Using some form of regularization

Answer: D

Explanation:
Regularization is a technique that can be used to prevent overfitting and improve model performance on unseen data. Overfitting occurs when the model learns the training data too well and fails to generalize to new and unseen data. This can be seen in the question, where the validation accuracy is lower than the training accuracy, and both are lower than the human-level performance. Regularization is a way of adding some constraints or penalties to the model to reduce its complexity and prevent it from memorizing the training data. Some common forms of regularization for image classification are:
Weight decay: Adding a term to the loss function that penalizes large weights in the model. This can help reduce the variance and noise in the model and make it more robust to small changes in the input.
Dropout: Randomly dropping out some units or connections in the model during training. This can help reduce the co-dependency among the units and make the model more resilient to missing or corrupted features.
Data augmentation: Artificially increasing the size and diversity of the training data by applying random transformations, such as cropping, flipping, rotating, scaling, etc. This can help the model learn more invariant and generalizable features and reduce the risk of overfitting to specific patterns in the training data.

The other options are not likely to fix the issue of overfitting, and may even worsen it:
A longer training time: This can lead to more overfitting, as the model will have more chances to fit the noise and details in the training data that are not relevant for the validation data.
Making the network larger: This can increase the model capacity and complexity, which can also lead to more overfitting, as the model will have more parameters to learn and adjust to the training data.
Using a different optimizer: This can affect the speed and stability of the training process, but not necessarily the generalization ability of the model. The choice of optimizer depends on the characteristics of the data and the model, and there is no guarantee that a different optimizer will prevent overfitting.
Reference:
Regularization (machine learning)
Image Classification: Regularization
How to Reduce Overfitting With Dropout Regularization in Keras

---

QUESTION 51
Example Corp has an annual sale event from October to December. The company has sequential sales data from the past 15 years and wants to use Amazon ML to predict the sales for this year's upcoming event.
Which method should Example Corp use to split the data into a training dataset and evaluation dataset?

A. Pre-split the data before uploading to Amazon S3
B. Have Amazon ML split the data randomly.
C. Have Amazon ML split the data sequentially.
D. Perform custom cross-validation on the data

Answer: C

Explanation:
A sequential split is a method of splitting data into training and evaluation datasets while preserving the order of the data records. This method is useful when the data has a temporal or sequential structure, and the order of the data matters for the prediction task. For example, if the data contains sales data for different months or years, and the goal is to predict the sales for the next month or year, a sequential split can ensure that the training data comes from the earlier period and the evaluation data comes from the later period. This can help avoid data leakage, which occurs when the training data contains information from the future that is not available at the time of prediction. A sequential split can also help evaluate the model performance on the most recent data, which may be more relevant and representative of the future data.
In this question, Example Corp has sequential sales data from the past 15 years and wants to use Amazon ML to predict the sales for this years upcoming annual sale event. A sequential split is the most appropriate method for splitting the data, as it can preserve the order of the data and prevent data leakage. For example, Example Corp can use the data from the first 14 years as the training

dataset, and the data from the last year as the evaluation dataset. This way, the model can learn from the historical data and be tested on the most recent data.
Amazon ML provides an option to split the data sequentially when creating the training and evaluation datasources. To use this option, Example Corp can specify the percentage of the data to use for training and evaluation, and Amazon ML will use the first part of the data for training and the remaining part of the data for evaluation. For more information, see Splitting Your Data - Amazon Machine Learning.

QUESTION 52
A company is running a machine learning prediction service that generates 100 TB of predictions every day A Machine Learning Specialist must generate a visualization of the daily precision-recall curve from the predictions, and forward a read-only version to the Business team.
Which solution requires the LEAST coding effort?

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

Answer: C

Explanation:
A precision-recall curve is a plot that shows the trade-off between the precision and recall of a binary classifier as the decision threshold is varied. It is a useful tool for evaluating and comparing the performance of different models. To generate a precision-recall curve, the following steps are needed:
Calculate the precision and recall values for different threshold values using the predictions and the true labels of the data.
Plot the precision values on the y-axis and the recall values on the x-axis for each threshold value.
Optionally, calculate the area under the curve (AUC) as a summary metric of the model performance.
Among the four options, option C requires the least coding effort to generate and share a visualization of the daily precision-recall curve from the predictions. This option involves the following steps:
Run a daily Amazon EMR workflow to generate precision-recall data: Amazon EMR is a service that allows running big data frameworks, such as Apache Spark, on a managed cluster of EC2 instances. Amazon EMR can handle large-scale data processing and analysis, such as calculating the precision and recall values for different threshold values from 100 TB of predictions. Amazon EMR supports

various languages, such as Python, Scala, and R, for writing the code to perform the calculations. Amazon EMR also supports scheduling workflows using Apache Airflow or AWS Step Functions, which can automate the daily execution of the code.
Save the results in Amazon S3: Amazon S3 is a service that provides scalable, durable, and secure object storage. Amazon S3 can store the precision-recall data generated by Amazon EMR in a costeffective and accessible way. Amazon S3 supports various data formats, such as CSV, JSON, or Parquet, for storing the data. Amazon S3 also integrates with other AWS services, such as Amazon QuickSight, for further processing and visualization of the data.
Visualize the arrays in Amazon QuickSight: Amazon QuickSight is a service that provides fast, easy-touse, and interactive business intelligence and data visualization. Amazon QuickSight can connect to Amazon S3 as a data source and import the precision-recall data into a dataset. Amazon QuickSight can then create a line chart to plot the precision-recall curve from the dataset. Amazon QuickSight also supports calculating the AUC and adding it as an annotation to the chart.
Publish them in a dashboard shared with the Business team: Amazon QuickSight allows creating and publishing dashboards that contain one or more visualizations from the datasets. Amazon QuickSight also allows sharing the dashboards with other users or groups within the same AWS account or across different AWS accounts. The Business team can access the dashboard with read-only permissions and view the daily precision-recall curve from the predictions.
The other options require more coding effort than option C for the following reasons:
Option A: This option requires writing code to plot the precision-recall curve from the data stored in Amazon S3, as well as creating a mechanism to share the plot with the Business team. This can involve using additional libraries or tools, such as matplotlib, seaborn, or plotly, for creating the plot, and using email, web, or cloud services, such as AWS Lambda or Amazon SNS, for sharing the plot.
Option B: This option requires transforming the predictions into a format that Amazon QuickSight can recognize and import as a data source, such as CSV, JSON, or Parquet. This can involve writing code to process and convert the predictions, as well as uploading them to a storage service, such as Amazon S3 or Amazon Redshift, that Amazon QuickSight can connect to.
Option D: This option requires writing code to generate precision-recall data in Amazon ES, as well as creating a dashboard to visualize the data. Amazon ES is a service that provides a fully managed Elasticsearch cluster, which is mainly used for search and analytics purposes. Amazon ES is not designed for generating precision-recall data, and it requires using a specific data format, such as JSON, for storing the data. Amazon ES also requires using a tool, such as Kibana, for creating and sharing the dashboard, which can involve additional configuration and customization steps.
Reference:
Precision-Recall
What Is Amazon EMR?
What Is Amazon S3?
[What Is Amazon QuickSight?]
[What Is Amazon Elasticsearch Service?]

QUESTION 53
A Machine Learning Specialist has built a model using Amazon SageMaker built-in algorithms and is

not getting expected accurate results The Specialist wants to use hyperparameter optimization to increase the model's accuracy
Which method is the MOST repeatable and requires the LEAST amount of effort to achieve this?

A. Launch multiple training jobs in parallel with different hyperparameters
B. Create an AWS Step Functions workflow that monitors the accuracy in Amazon CloudWatch Logs and relaunches the training job with a defined list of hyperparameters
C. Create a hyperparameter tuning job and set the accuracy as an objective metric.
D. Create a random walk in the parameter space to iterate through a range of values that should be used for each individual hyperparameter

Answer: C

Explanation:
A hyperparameter tuning job is a feature of Amazon SageMaker that allows automatically finding the best combination of hyperparameters for a machine learning model. Hyperparameters are high-level parameters that influence the learning process and the performance of the model, such as the learning rate, the number of layers, the regularization factor, etc. A hyperparameter tuning job works by launching multiple training jobs with different hyperparameters, evaluating the results using an objective metric, and choosing the next set of hyperparameters to try based on a search strategy. The objective metric is a measure of the quality of the model, such as accuracy, precision, recall, etc. The search strategy is a method of exploring the hyperparameter space, such as random search, grid search, or Bayesian optimization.
Among the four options, option C is the most repeatable and requires the least amount of effort to use hyperparameter optimization to increase the models accuracy. This option involves the following steps:
Create a hyperparameter tuning job: Amazon SageMaker provides an easy-to-use interface for creating a hyperparameter tuning job, either through the AWS Management Console, the AWS CLI, or the AWS SDKs. To create a hyperparameter tuning job, the Machine Learning Specialist needs to specify the following information:
The name and type of the algorithm to use, either a built-in algorithm or a custom algorithm.
The ranges and types of the hyperparameters to tune, such as categorical, continuous, or integer.
The name and type of the objective metric to optimize, such as accuracy, and whether to maximize or minimize it.
The resource limits for the tuning job, such as the maximum number of training jobs and the maximum parallel training jobs.
The input data channels and the output data location for the training jobs.
The configuration of the training instances, such as the instance type, the instance count, the volume size, etc.
Set the accuracy as an objective metric: To use accuracy as an objective metric, the Machine Learning Specialist needs to ensure that the training algorithm writes the accuracy value to a file

called metric_definitions in JSON format and prints it to stdout or stderr. For example, the file can contain the following content:

```
[
  {
    "Name": "accuracy",
    "Regex": "#quality_metric: host=\\S+, epoch=\\S+, accuracy=(\\S+)"
  }
]
```

This means that the training algorithm prints a line like this:

```
#quality_metric: host=algo-1, epoch=12, accuracy=0.8765
```

Amazon SageMaker reads the accuracy value from the line and uses it to evaluate and compare the training jobs.
The other options are not as repeatable and require more effort than option C for the following reasons:
Option A: This option requires manually launching multiple training jobs in parallel with different hyperparameters, which can be tedious and error-prone. It also requires manually monitoring and comparing the results of the training jobs, which can be time-consuming and subjective.
Option B: This option requires writing code to create an AWS Step Functions workflow that monitors the accuracy in Amazon CloudWatch Logs and relaunches the training job with a defined list of hyperparameters, which can be complex and challenging. It also requires maintaining and updating the list of hyperparameters, which can be inefficient and suboptimal.
Option D: This option requires writing code to create a random walk in the parameter space to iterate through a range of values that should be used for each individual hyperparameter, which can be unreliable and unpredictable. It also requires defining and implementing a stopping criterion, which can be arbitrary and inconsistent.
Reference:
Automatic Model Tuning - Amazon SageMaker
Define Metrics to Monitor Model Performance

QUESTION 54
IT leadership wants Jo transition a company's existing machine learning data storage environment to AWS as a temporary ad hoc solution The company currently uses a custom software process that heavily leverages SOL as a query language and exclusively stores generated csv documents for machine learning
The ideal state for the company would be a solution that allows it to continue to use the current

workforce of SQL experts The solution must also support the storage of csv and JSON files, and be able to query over semi-structured data The following are high priorities for the company:
Solution simplicity
Fast development time
Low cost
High flexibility
What technologies meet the company's requirements?

A. Amazon S3 and Amazon Athena
B. Amazon Redshift and AWS Glue
C. Amazon DynamoDB and DynamoDB Accelerator (DAX)
D. Amazon RDS and Amazon ES

Answer: A

Explanation:
Amazon S3 and Amazon Athena are technologies that meet the companys requirements for a temporary ad hoc solution for machine learning data storage and query. Amazon S3 and Amazon Athena have the following features and benefits:
Amazon S3 is a service that provides scalable, durable, and secure object storage for any type of data. Amazon S3 can store csv and JSON files, as well as other formats, and can handle large volumes of data with high availability and performance. Amazon S3 also integrates with other AWS services, such as Amazon Athena, for further processing and analysis of the data.
Amazon Athena is a service that allows querying data stored in Amazon S3 using standard SQL. Amazon Athena can query over semi-structured data, such as JSON, as well as structured data, such as csv, without requiring any loading or transformation. Amazon Athena is serverless, meaning that there is no infrastructure to manage and users only pay for the queries they run. Amazon Athena also supports the use of AWS Glue Data Catalog, which is a centralized metadata repository that can store and manage the schema and partition information of the data in Amazon S3.
Using Amazon S3 and Amazon Athena, the company can achieve the following high priorities:
Solution simplicity: Amazon S3 and Amazon Athena are easy to use and require minimal configuration and maintenance. The company can simply upload the csv and JSON files to Amazon S3 and use Amazon Athena to query them using SQL. The company does not need to worry about provisioning, scaling, or managing any servers or clusters.
Fast development time: Amazon S3 and Amazon Athena can enable the company to quickly access and analyze the data without any data preparation or loading. The company can use the existing workforce of SQL experts to write and run queries on Amazon Athena and get results in seconds or minutes.
Low cost: Amazon S3 and Amazon Athena are cost-effective and offer pay-as-you-go pricing models. Amazon S3 charges based on the amount of storage used and the number of requests made. Amazon Athena charges based on the amount of data scanned by the queries. The company can also reduce the costs by using compression, encryption, and partitioning techniques to optimize the data storage

and query performance.
High flexibility: Amazon S3 and Amazon Athena are flexible and can support various data types, formats, and sources. The company can store and query any type of data in Amazon S3, such as csv, JSON, Parquet, ORC, etc. The company can also query data from multiple sources in Amazon S3, such as data lakes, data warehouses, log files, etc.
The other options are not as suitable as option A for the companys requirements for the following reasons:
Option B: Amazon Redshift and AWS Glue are technologies that can be used for data warehousing and data integration, but they are not ideal for a temporary ad hoc solution. Amazon Redshift is a service that provides a fully managed, petabyte-scale data warehouse that can run complex analytical queries using SQL. AWS Glue is a service that provides a fully managed extract, transform, and load (ETL) service that can prepare and load data for analytics. However, using Amazon Redshift and AWS Glue would require more effort and cost than using Amazon S3 and Amazon Athena. The company would need to load the data from Amazon S3 to Amazon Redshift using AWS Glue, which can take time and incur additional charges. The company would also need to manage the capacity and performance of the Amazon Redshift cluster, which can be complex and expensive.
Option C: Amazon DynamoDB and DynamoDB Accelerator (DAX) are technologies that can be used for fast and scalable NoSQL database and caching, but they are not suitable for the companys data storage and query needs. Amazon DynamoDB is a service that provides a fully managed, key-value and document database that can deliver single-digit millisecond performance at any scale. DynamoDB Accelerator (DAX) is a service that provides a fully managed, in-memory cache for DynamoDB that can improve the read performance by up to 10 times. However, using Amazon DynamoDB and DAX would not allow the company to continue to use SQL as a query language, as Amazon DynamoDB does not support SQL. The company would need to use the DynamoDB API or the AWS SDKs to access and query the data, which can require more coding and learning effort. The company would also need to transform the csv and JSON files into DynamoDB items, which can involve additional processing and complexity.
Option D: Amazon RDS and Amazon ES are technologies that can be used for relational database and search and analytics, but they are not optimal for the companys data storage and query scenario. Amazon RDS is a service that provides a fully managed, relational database that supports various database engines, such as MySQL, PostgreSQL, Oracle, etc. Amazon ES is a service that provides a fully managed, Elasticsearch cluster, which is mainly used for search and analytics purposes. However, using Amazon RDS and Amazon ES would not be as simple and cost-effective as using Amazon S3 and Amazon Athena. The company would need to load the data from Amazon S3 to Amazon RDS, which can take time and incur additional charges. The company would also need to manage the capacity and performance of the Amazon RDS and Amazon ES clusters, which can be complex and expensive. Moreover, Amazon RDS and Amazon ES are not designed to handle semistructured data, such as JSON, as well as Amazon S3 and Amazon Athena.
Reference:
Amazon S3
Amazon Athena
Amazon Redshift

AWS Glue
Amazon DynamoDB
[DynamoDB Accelerator (DAX)]
[Amazon RDS]
[Amazon ES]

QUESTION 55
A Machine Learning Specialist is working for a credit card processing company and receives an unbalanced dataset containing credit card transactions. It contains 99,000 valid transactions and 1,000 fraudulent transactions The Specialist is asked to score a model that was run against the dataset The Specialist has been advised that identifying valid transactions is equally as important as identifying fraudulent transactions
What metric is BEST suited to score the model?

A. Precision
B. Recall
C. Area Under the ROC Curve (AUC)
D. Root Mean Square Error (RMSE)

Answer: C

Explanation:
Area Under the ROC Curve (AUC) is a metric that is best suited to score the model for the given scenario. AUC is a measure of the performance of a binary classifier, such as a model that predicts whether a credit card transaction is valid or fraudulent. AUC is calculated based on the Receiver Operating Characteristic (ROC) curve, which is a plot that shows the trade-off between the true positive rate (TPR) and the false positive rate (FPR) of the classifier as the decision threshold is varied. The TPR, also known as recall or sensitivity, is the proportion of actual positive cases (fraudulent transactions) that are correctly predicted as positive by the classifier. The FPR, also known as the fall-out, is the proportion of actual negative cases (valid transactions) that are incorrectly predicted as positive by the classifier. The ROC curve illustrates how well the classifier can distinguish between the two classes, regardless of the class distribution or the error costs. A perfect classifier would have a TPR of 1 and an FPR of 0 for all thresholds, resulting in a ROC curve that goes from the bottom left to the top left and then to the top right of the plot. A random classifier would have a TPR and an FPR that are equal for all thresholds, resulting in a ROC curve that goes from the bottom left to the top right of the plot along the diagonal line. AUC is the area under the ROC curve, and it ranges from 0 to 1. A higher AUC indicates a better classifier, as it means that the classifier has a higher TPR and a lower FPR for all thresholds. AUC is a useful metric for imbalanced classification problems, such as the credit card transaction dataset, because it is insensitive to the class imbalance and the error costs. AUC can capture the overall performance of the classifier across all possible scenarios, and it can be used to compare different classifiers based on their ROC curves.
The other options are not as suitable as AUC for the given scenario for the following reasons:

Precision: Precision is the proportion of predicted positive cases (fraudulent transactions) that are actually positive. Precision is a useful metric when the cost of a false positive is high, such as in spam detection or medical diagnosis. However, precision is not a good metric for imbalanced classification problems, because it can be misleadingly high when the positive class is rare. For example, a classifier that predicts all transactions as valid would have a precision of 0, but a very high accuracy of 99%. Precision is also dependent on the decision threshold and the error costs, which may vary for different scenarios.
Recall: Recall is the same as the TPR, and it is the proportion of actual positive cases (fraudulent transactions) that are correctly predicted as positive by the classifier. Recall is a useful metric when the cost of a false negative is high, such as in fraud detection or cancer diagnosis. However, recall is not a good metric for imbalanced classification problems, because it can be misleadingly low when the positive class is rare. For example, a classifier that predicts all transactions as fraudulent would have a recall of 1, but a very low accuracy of 1%. Recall is also dependent on the decision threshold and the error costs, which may vary for different scenarios.
Root Mean Square Error (RMSE): RMSE is a metric that measures the average difference between the predicted and the actual values. RMSE is a useful metric for regression problems, where the goal is to predict a continuous value, such as the price of a house or the temperature of a city. However, RMSE is not a good metric for classification problems, where the goal is to predict a discrete value, such as the class label of a transaction. RMSE is not meaningful for classification problems, because it does not capture the accuracy or the error costs of the predictions.
Reference:
ROC Curve and AUC
How and When to Use ROC Curves and Precision-Recall Curves for Classification in Python
Precision-Recall
Root Mean Squared Error

QUESTION 56
A bank's Machine Learning team is developing an approach for credit card fraud detection The company has a large dataset of historical data labeled as fraudulent The goal is to build a model to take the information from new transactions and predict whether each transaction is fraudulent or not
Which built-in Amazon SageMaker machine learning algorithm should be used for modeling this problem?

A. Seq2seq
B. XGBoost
C. K-means
D. Random Cut Forest (RCF)

Answer: B

Explanation:

XGBoost is a built-in Amazon SageMaker machine learning algorithm that should be used for modeling the credit card fraud detection problem. XGBoost is an algorithm that implements a scalable and distributed gradient boosting framework, which is a popular and effective technique for supervised learning problems. Gradient boosting is a method of combining multiple weak learners, such as decision trees, into a strong learner, by iteratively fitting new models to the residual errors of the previous models and adding them to the ensemble. XGBoost can handle various types of data, such as numerical, categorical, or text, and can perform both regression and classification tasks. XGBoost also supports various features and optimizations, such as regularization, missing value handling, parallelization, and cross-validation, that can improve the performance and efficiency of the algorithm.

XGBoost is suitable for the credit card fraud detection problem for the following reasons:

The problem is a binary classification problem, where the goal is to predict whether a transaction is fraudulent or not, based on the information from new transactions. XGBoost can perform binary classification by using a logistic regression objective function and outputting the probability of the positive class (fraudulent) for each transaction.

The problem involves a large and imbalanced dataset of historical data labeled as fraudulent. XGBoost can handle large-scale and imbalanced data by using distributed and parallel computing, as well as techniques such as weighted sampling, class weighting, or stratified sampling, to balance the classes and reduce the bias towards the majority class (non-fraudulent).

The problem requires a high accuracy and precision for detecting fraudulent transactions, as well as a low false positive rate for avoiding false alarms. XGBoost can achieve high accuracy and precision by using gradient boosting, which can learn complex and non-linear patterns from the data and reduce the variance and overfitting of the model. XGBoost can also achieve a low false positive rate by using regularization, which can reduce the complexity and noise of the model and prevent it from fitting spurious signals in the data.

The other options are not as suitable as XGBoost for the credit card fraud detection problem for the following reasons:

Seq2seq: Seq2seq is an algorithm that implements a sequence-to-sequence model, which is a type of neural network model that can map an input sequence to an output sequence. Seq2seq is mainly used for natural language processing tasks, such as machine translation, text summarization, or dialogue generation. Seq2seq is not suitable for the credit card fraud detection problem, because the problem is not a sequence-to-sequence task, but a binary classification task. The input and output of the problem are not sequences of words or tokens, but vectors of features and labels.

K-means: K-means is an algorithm that implements a clustering technique, which is a type of unsupervised learning method that can group similar data points into clusters. K-means is mainly used for exploratory data analysis, dimensionality reduction, or anomaly detection. K-means is not suitable for the credit card fraud detection problem, because the problem is not a clustering task, but a classification task. The problem requires using the labeled data to train a model that can predict the labels of new data, not finding the optimal number of clusters or the cluster memberships of the data.

Random Cut Forest (RCF): RCF is an algorithm that implements an anomaly detection technique, which is a type of unsupervised learning method that can identify data points that deviate from the

normal behavior or distribution of the data. RCF is mainly used for detecting outliers, frauds, or faults in the data. RCF is not suitable for the credit card fraud detection problem, because the problem is not an anomaly detection task, but a classification task. The problem requires using the labeled data to train a model that can predict the labels of new data, not finding the anomaly scores or the anomalous data points in the data.
Reference:
XGBoost Algorithm
Use XGBoost for Binary Classification with Amazon SageMaker
Seq2seq Algorithm
K-means Algorithm
[Random Cut Forest Algorithm]

QUESTION 57
While working on a neural network project, a Machine Learning Specialist discovers thai some features in the data have very high magnitude resulting in this data being weighted more in the cost function What should the Specialist do to ensure better convergence during backpropagation?

A. Dimensionality reduction
B. Data normalization
C. Model regulanzation
D. Data augmentation for the minority class

Answer: B

Explanation:
Data normalization is a data preprocessing technique that scales the features to a common range, such as [0, 1] or [-1, 1]. This helps reduce the impact of features with high magnitude on the cost function and improves the convergence during backpropagation. Data normalization can be done using different methods, such as min-max scaling, z-score standardization, or unit vector normalization. Data normalization is different from dimensionality reduction, which reduces the number of features; model regularization, which adds a penalty term to the cost function to prevent overfitting; and data augmentation, which increases the amount of data by creating synthetic samples. Reference:
Data processing options for AI/ML | AWS Machine Learning Blog
Data preprocessing - Machine Learning Lens
How to Normalize Data Using scikit-learn in Python
Normalization | Machine Learning | Google for Developers

QUESTION 58
An online reseller has a large, multi-column dataset with one column missing 30% of its data A Machine Learning Specialist believes that certain columns in the dataset could be used to reconstruct the missing data.

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

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

Answer: C

Explanation:
Multiple imputation is a technique that uses machine learning to generate multiple plausible values for each missing value in a dataset, based on the observed data and the relationships among the variables. Multiple imputation preserves the integrity of the dataset by accounting for the uncertainty and variability of the missing data, and avoids the bias and loss of information that may result from other methods, such as listwise deletion, last observation carried forward, or mean substitution. Multiple imputation can improve the accuracy and validity of statistical analysis and machine learning models that use the imputed dataset. Reference:
Managing missing values in your target and related datasets with automated imputation support in Amazon Forecast
Imputation by feature importance (IBFI): A methodology to impute missing data in large datasets
Multiple Imputation by Chained Equations (MICE) Explained

QUESTION 59
A Machine Learning Specialist discover the following statistics while experimenting on a model.

What can the Specialist from the experiments?

A. The model In Experiment 1 had a high variance error lhat was reduced in Experiment 3 by regularization Experiment 2 shows that there is minimal bias error in Experiment 1
B. The model in Experiment 1 had a high bias error that was reduced in Experiment 3 by regularization Experiment 2 shows that there is minimal variance error in Experiment 1

C. The model in Experiment 1 had a high bias error and a high variance error that were reduced in Experiment 3 by regularization Experiment 2 shows thai high bias cannot be reduced by increasing layers and neurons in the model
D. The model in Experiment 1 had a high random noise error that was reduced in Experiment 3 by regularization Experiment 2 shows that random noise cannot be reduced by increasing layers and neurons in the model

Answer: A

Explanation:
The model in Experiment 1 had a high variance error because it performed well on the training data (train error = 5%) but poorly on the test data (test error = 8%). This indicates that the model was overfitting the training data and not generalizing well to new data. The model in Experiment 3 had a lower variance error because it performed similarly on the training data (train error = 5.1%) and the test data (test error = 5.4%). This indicates that the model was more robust and less sensitive to the fluctuations in the training data. The model in Experiment 3 achieved this improvement by implementing regularization, which is a technique that reduces the complexity of the model and prevents overfitting by adding a penalty term to the loss function. The model in Experiment 2 had a minimal bias error because it performed similarly on the training data (train error = 5.2%) and the test data (test error = 5.7%) as the model in Experiment 1. This indicates that the model was not underfitting the data and capturing the true relationship between the input and output variables.
The model in Experiment 2 increased the number of layers and neurons in the model, which is a way to increase the complexity and flexibility of the model. However, this did not improve the performance of the model, as the variance error remained high. This shows that increasing the complexity of the model is not always the best way to reduce the bias error, and may even increase the variance error if the model becomes too complex for the data. Reference:
Bias Variance Tradeoff - Clearly Explained - Machine Learning Plus
The Bias-Variance Trade-off in Machine Learning - Stack Abuse

---

QUESTION 60
A Machine Learning Specialist needs to be able to ingest streaming data and store it in Apache Parquet files for exploration and analysis. Which of the following services would both ingest and store this data in the correct format?

A. AWSDMS
B. Amazon Kinesis Data Streams
C. Amazon Kinesis Data Firehose
D. Amazon Kinesis Data Analytics

Answer: C

Explanation:
Amazon Kinesis Data Firehose is a service that can ingest streaming data and store it in various destinations, including Amazon S3, Amazon Redshift, Amazon Elasticsearch Service, and Splunk. Amazon Kinesis Data Firehose can also convert the incoming data to Apache Parquet or Apache ORC format before storing it in Amazon S3. This can reduce the storage cost and improve the performance of analytical queries on the data. Amazon Kinesis Data Firehose supports various data sources, such as Amazon Kinesis Data Streams, Amazon Managed Streaming for Apache Kafka, AWS IoT, and custom applications. Amazon Kinesis Data Firehose can also apply data transformation and compression using AWS Lambda functions.
AWSDMS is not a valid service name. AWS Database Migration Service (AWS DMS) is a service that can migrate data from various sources to various targets, but it does not support streaming data or Parquet format.
Amazon Kinesis Data Streams is a service that can ingest and process streaming data in real time, but it does not store the data in any destination. Amazon Kinesis Data Streams can be integrated with Amazon Kinesis Data Firehose to store the data in Parquet format.
Amazon Kinesis Data Analytics is a service that can analyze streaming data using SQL or Apache Flink, but it does not store the data in any destination. Amazon Kinesis Data Analytics can be integrated with Amazon Kinesis Data Firehose to store the data in Parquet format. Reference:
Amazon Kinesis Data Firehose - Amazon Web Services
What Is Amazon Kinesis Data Firehose? - Amazon Kinesis Data Firehose
Amazon Kinesis Data Firehose FAQs - Amazon Web Services

QUESTION 61
A Machine Learning Specialist needs to move and transform data in preparation for training Some of the data needs to be processed in near-real time and other data can be moved hourly There are existing Amazon EMR MapReduce jobs to clean and feature engineering to perform on the data
Which of the following services can feed data to the MapReduce jobs? (Select TWO)

A. AWSDMS
B. Amazon Kinesis
C. AWS Data Pipeline
D. Amazon Athena
E. Amazon ES

Answer: B, C

Explanation:
Amazon Kinesis and AWS Data Pipeline are two services that can feed data to the Amazon EMR MapReduce jobs. Amazon Kinesis is a service that can ingest, process, and analyze streaming data in real time. Amazon Kinesis can be integrated with Amazon EMR to run MapReduce jobs on streaming data sources, such as web logs, social media, IoT devices, and clickstreams. Amazon Kinesis can handle data that needs to be processed in near-real time, such as for anomaly detection, fraud

detection, or dashboarding. AWS Data Pipeline is a service that can orchestrate and automate data movement and transformation across various AWS services and on-premises data sources. AWS Data Pipeline can be integrated with Amazon EMR to run MapReduce jobs on batch data sources, such as Amazon S3, Amazon RDS, Amazon DynamoDB, and Amazon Redshift. AWS Data Pipeline can handle data that can be moved hourly, such as for data warehousing, reporting, or machine learning.
AWSDMS is not a valid service name. AWS Database Migration Service (AWS DMS) is a service that can migrate data from various sources to various targets, but it does not support streaming data or MapReduce jobs.
Amazon Athena is a service that can query data stored in Amazon S3 using standard SQL, but it does not feed data to Amazon EMR or run MapReduce jobs.
Amazon ES is a service that provides a fully managed Elasticsearch cluster, which can be used for search, analytics, and visualization, but it does not feed data to Amazon EMR or run MapReduce jobs. Reference:
Using Amazon Kinesis with Amazon EMR - Amazon EMR
AWS Data Pipeline - Amazon Web Services
Using AWS Data Pipeline to Run Amazon EMR Jobs - AWS Data Pipeline

---

QUESTION 62
An insurance company is developing a new device for vehicles that uses a camera to observe drivers' behavior and alert them when they appear distracted The company created approximately 10,000 training images in a controlled environment that a Machine Learning Specialist will use to train and evaluate machine learning models
During the model evaluation the Specialist notices that the training error rate diminishes faster as the number of epochs increases and the model is not accurately inferring on the unseen test images
Which of the following should be used to resolve this issue? (Select TWO)

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

Answer: B, E

Explanation:
The issue described in the question is a sign of overfitting, which is a common problem in machine learning when the model learns the noise and details of the training data too well and fails to generalize to new and unseen data. Overfitting can result in a low training error rate but a high test error rate, which indicates poor performance and validity of the model. There are several techniques that can be used to prevent or reduce overfitting, such as data augmentation and regularization.
Data augmentation is a technique that applies various transformations to the original training data, such as rotation, scaling, cropping, flipping, adding noise, changing brightness, etc., to create new

and diverse data samples. Data augmentation can increase the size and diversity of the training data, which can help the model learn more features and patterns and reduce the variance of the model. Data augmentation is especially useful for image data, as it can simulate different scenarios and perspectives that the model may encounter in real life. For example, in the question, the device uses a camera to observe drivers behavior, so data augmentation can help the model deal with different lighting conditions, angles, distances, etc. Data augmentation can be done using various libraries and frameworks, such as TensorFlow, PyTorch, Keras, OpenCV, etc12
Regularization is a technique that adds a penalty term to the models objective function, which is typically based on the models parameters. Regularization can reduce the complexity and flexibility of the model, which can prevent overfitting by avoiding learning the noise and details of the training data. Regularization can also improve the stability and robustness of the model, as it can reduce the sensitivity of the model to small fluctuations in the data. There are different types of regularization, such as L1, L2, dropout, etc., but they all have the same goal of reducing overfitting. L2 regularization, also known as weight decay or ridge regression, is one of the most common and effective regularization techniques. L2 regularization adds the squared norm of the models parameters multiplied by a regularization parameter (lambda) to the models objective function. L2 regularization can shrink the models parameters towards zero, which can reduce the variance of the model and improve the generalization ability of the model. L2 regularization can be implemented using various libraries and frameworks, such as TensorFlow, PyTorch, Keras, Scikit-learn, etc34
The other options are not valid or relevant for resolving the issue of overfitting. Adding vanishing gradient to the model is not a technique, but a problem that occurs when the gradient of the models objective function becomes very small and the model stops learning. Making the neural network architecture complex is not a solution, but a possible cause of overfitting, as a complex model can have more parameters and more flexibility to fit the training data too well. Using gradient checking in the model is not a technique, but a debugging method that verifies the correctness of the gradient computation in the model. Gradient checking is not related to overfitting, but to the implementation of the model.

QUESTION 63
The Chief Editor for a product catalog wants the Research and Development team to build a machine learning system that can be used to detect whether or not individuals in a collection of images are wearing the company's retail brand The team has a set of training data
Which machine learning algorithm should the researchers use that BEST meets their requirements?

A. Latent Dirichlet Allocation (LDA)
B. Recurrent neural network (RNN)
C. K-means
D. Convolutional neural network (CNN)

Answer: D

Explanation:

A convolutional neural network (CNN) is a type of machine learning algorithm that is suitable for image classification tasks. A CNN consists of multiple layers that can extract features from images and learn to recognize patterns and objects. A CNN can also use transfer learning to leverage pretrained models that have been trained on large-scale image datasets, such as ImageNet, and finetune them for specific tasks, such as detecting the companys retail brand. A CNN can achieve high accuracy and performance for image classification problems, as it can handle complex and diverse images and reduce the dimensionality and noise of the input data. A CNN can be implemented using various frameworks and libraries, such as TensorFlow, PyTorch, Keras, MXNet, etc12
The other options are not valid or relevant for the image classification task. Latent Dirichlet Allocation (LDA) is a type of machine learning algorithm that is suitable for topic modeling tasks. LDA can discover the hidden topics and their proportions in a collection of text documents, such as news articles, tweets, reviews, etc. LDA is not applicable for image data, as it requires textual input and output. LDA can be implemented using various frameworks and libraries, such as Gensim, Scikitlearn, Mallet, etc34
Recurrent neural network (RNN) is a type of machine learning algorithm that is suitable for sequential data tasks. RNN can process and generate data that has temporal or sequential dependencies, such as natural language, speech, audio, video, etc. RNN is not optimal for image data, as it does not capture the spatial features and relationships of the pixels. RNN can be implemented using various frameworks and libraries, such as TensorFlow, PyTorch, Keras, MXNet, etc.
K-means is a type of machine learning algorithm that is suitable for clustering tasks. K-means can partition a set of data points into a predefined number of clusters, based on the similarity and distance between the data points. K-means is not suitable for image classification tasks, as it does not learn to label the images or detect the objects of interest. K-means can be implemented using various frameworks and libraries, such as Scikit-learn, TensorFlow, PyTorch, etc.

QUESTION 64
A Machine Learning Specialist kicks off a hyperparameter tuning job for a tree-based ensemble model using Amazon SageMaker with Area Under the ROC Curve (AUC) as the objective metric This workflow will eventually be deployed in a pipeline that retrains and tunes hyperparameters each night to model click-through on data that goes stale every 24 hours
With the goal of decreasing the amount of time it takes to train these models, and ultimately to decrease costs, the Specialist wants to reconfigure the input hyperparameter range(s)
Which visualization will accomplish this?

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

Answer: D

Explanation:
A scatter plot showing the correlation between maximum tree depth and the objective metric is a visualization that can help the Machine Learning Specialist reconfigure the input hyperparameter range(s) for the tree-based ensemble model. A scatter plot is a type of graph that displays the relationship between two variables using dots, where each dot represents one observation. A scatter plot can show the direction, strength, and shape of the correlation between the variables, as well as any outliers or clusters. In this case, the scatter plot can show how the maximum tree depth, which is a hyperparameter that controls the complexity and depth of the decision trees in the ensemble model, affects the AUC, which is the objective metric that measures the performance of the model in terms of the trade-off between true positive rate and false positive rate. By looking at the scatter plot, the Machine Learning Specialist can see if there is a positive, negative, or no correlation between the maximum tree depth and the AUC, and how strong or weak the correlation is. The Machine Learning Specialist can also see if there is an optimal value or range of values for the maximum tree depth that maximizes the AUC, or if there is a point of diminishing returns or overfitting where increasing the maximum tree depth does not improve or even worsens the AUC. Based on the scatter plot, the Machine Learning Specialist can reconfigure the input hyperparameter range(s) for the maximum tree depth to focus on the values that yield the best AUC, and avoid the values that result in poor AUC. This can decrease the amount of time and cost it takes to train the model, as the hyperparameter tuning job can explore fewer and more promising combinations of values. A scatter plot can be created using various tools and libraries, such as Matplotlib, Seaborn, Plotly, etc12
The other options are not valid or relevant for reconfiguring the input hyperparameter range(s) for the tree-based ensemble model. A histogram showing whether the most important input feature is Gaussian is a visualization that can help the Machine Learning Specialist understand the distribution and shape of the input data, but not the hyperparameters. A histogram is a type of graph that displays the frequency or count of values in a single variable using bars, where each bar represents a bin or interval of values. A histogram can show if the variable is symmetric, skewed, or multimodal, and if it follows a normal or Gaussian distribution, which is a bell-shaped curve that is often assumed by many machine learning algorithms. In this case, the histogram can show if the most important input feature, which is a variable that has the most influence or predictive power on the output variable, is Gaussian or not. However, this does not help the Machine Learning Specialist reconfigure the input hyperparameter range(s) for the tree-based ensemble model, as the input feature is not a hyperparameter that can be tuned or optimized. A histogram can be created using various tools and libraries, such as Matplotlib, Seaborn, Plotly, etc34
A scatter plot with points colored by target variable that uses t-Distributed Stochastic Neighbor Embedding (t-SNE) to visualize the large number of input variables in an easier-to-read dimension is a visualization that can help the Machine Learning Specialist understand the structure and clustering of the input data, but not the hyperparameters. t-SNE is a technique that can reduce the dimensionality of high-dimensional data, such as images, text, or gene expression, and project it onto a lower-dimensional space, such as two or three dimensions, while preserving the local similarities and distances between the data points. t-SNE can help visualize and explore the patterns

and relationships in the data, such as the clusters, outliers, or separability of the classes. In this case, the scatter plot can show how the input variables, which are the features or predictors of the output variable, are mapped onto a two-dimensional space using t-SNE, and how the points are colored by the target variable, which is the output or response variable that the model tries to predict. However, this does not help the Machine Learning Specialist reconfigure the input hyperparameter range(s) for the tree-based ensemble model, as the input variables and the target variable are not hyperparameters that can be tuned or optimized. A scatter plot with t-SNE can be created using various tools and libraries, such as Scikit-learn, TensorFlow, PyTorch, etc5
A scatter plot showing the performance of the objective metric over each training iteration is a visualization that can help the Machine Learning Specialist understand the learning curve and convergence of the model, but not the hyperparameters. A scatter plot is a type of graph that displays the relationship between two variables using dots, where each dot represents one observation. A scatter plot can show the direction, strength, and shape of the correlation between the variables, as well as any outliers or clusters. In this case, the scatter plot can show how the objective metric, which is the performance measure that the model tries to optimize, changes over each training iteration, which is the number of times that the model updates its parameters using a batch of data. A scatter plot can show if the objective metric improves, worsens, or stagnates over time, and if the model converges to a stable value or oscillates or diverges. However, this does not help the Machine Learning Specialist reconfigure the input hyperparameter range(s) for the treebased ensemble model, as the objective metric and the training iteration are not hyperparameters that can be tuned or optimized. A scatter plot can be created using various tools and libraries, such as Matplotlib, Seaborn, Plotly, etc.

QUESTION 65
A Machine Learning Specialist is configuring automatic model tuning in Amazon SageMaker
When using the hyperparameter optimization feature, which of the following guidelines should be followed to improve optimization?
Choose the maximum number of hyperparameters supported by

A. Amazon SageMaker to search the largest number of combinations possible
B. Specify a very large hyperparameter range to allow Amazon SageMaker to cover every possible value.
C. Use log-scaled hyperparameters to allow the hyperparameter space to be searched as quickly as possible
D. Execute only one hyperparameter tuning job at a time and improve tuning through successive rounds of experiments

Answer: C

Explanation:
Using log-scaled hyperparameters is a guideline that can improve the automatic model tuning in Amazon SageMaker. Log-scaled hyperparameters are hyperparameters that have values that span

several orders of magnitude, such as learning rate, regularization parameter, or number of hidden units. Log-scaled hyperparameters can be specified by using a log-uniform distribution, which assigns equal probability to each order of magnitude within a range. For example, a log-uniform distribution between 0.001 and 1000 can sample values such as 0.001, 0.01, 0.1, 1, 10, 100, or 1000 with equal probability. Using log-scaled hyperparameters can allow the hyperparameter optimization feature to search the hyperparameter space more efficiently and effectively, as it can explore different scales of values and avoid sampling values that are too small or too large. Using log-scaled hyperparameters can also help avoid numerical issues, such as underflow or overflow, that may occur when using linear-scaled hyperparameters. Using log-scaled hyperparameters can be done by setting the ScalingType parameter to Logarithmic when defining the hyperparameter ranges in Amazon SageMaker12

The other options are not valid or relevant guidelines for improving the automatic model tuning in Amazon SageMaker. Choosing the maximum number of hyperparameters supported by Amazon SageMaker to search the largest number of combinations possible is not a good practice, as it can increase the time and cost of the tuning job and make it harder to find the optimal values. Amazon SageMaker supports up to 20 hyperparameters for tuning, but it is recommended to choose only the most important and influential hyperparameters for the model and algorithm, and use default or fixed values for the rest3 Specifying a very large hyperparameter range to allow Amazon SageMaker to cover every possible value is not a good practice, as it can result in sampling values that are irrelevant or impractical for the model and algorithm, and waste the tuning budget. It is recommended to specify a reasonable and realistic hyperparameter range based on the prior knowledge and experience of the model and algorithm, and use the results of the tuning job to refine the range if needed4 Executing only one hyperparameter tuning job at a time and improving tuning through successive rounds of experiments is not a good practice, as it can limit the exploration and exploitation of the hyperparameter space and make the tuning process slower and less efficient. It is recommended to use parallelism and concurrency to run multiple training jobs simultaneously and leverage the Bayesian optimization algorithm that Amazon SageMaker uses to guide the search for the best hyperparameter values5

QUESTION 66

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

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

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

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

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

Answer: C

Explanation:
Based on the model evaluation results, this is a viable model for production because the model is 86% accurate and the cost incurred by the company as a result of false positives is less than the false negatives. The accuracy of the model is the proportion of correct predictions out of the total predictions, which can be calculated by adding the true positives and true negatives and dividing by the total number of observations. In this case, the accuracy of the model is (10 + 76) / 100 = 0.86, which means that the model correctly predicted 86% of the customers churn status. The cost incurred by the company as a result of false positives and false negatives is the loss or damage that the company suffers when the model makes incorrect predictions. A false positive is when the model predicts that a customer will churn, but the customer actually does not churn. A false negative is when the model predicts that a customer will not churn, but the customer actually churns. In this case, the cost of a false positive is the incentive that the company offers to the customer who is predicted to churn, which is a relatively low cost. The cost of a false negative is the revenue that the company loses when the customer churns, which is a relatively high cost. Therefore, the cost of a false positive is less than the cost of a false negative, and the company would prefer to have more false positives than false negatives. The model has 10 false positives and 4 false negatives, which means that the companys cost is lower than if the model had more false negatives and fewer false positives.

QUESTION 67
A Machine Learning Specialist is designing a system for improving sales for a company. The objective is to use the large amount of information the company has on users' behavior and product preferences to predict which products users would like based on the users' similarity to other users. What should the Specialist do to meet this objective?

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

Answer: B

Explanation:

A collaborative filtering recommendation engine is a type of machine learning system that can improve sales for a company by using the large amount of information the company has on users behavior and product preferences to predict which products users would like based on the users similarity to other users. A collaborative filtering recommendation engine works by finding the users who have similar ratings or preferences for the products, and then recommending the products that the similar users have liked but the target user has not seen or rated. A collaborative filtering recommendation engine can leverage the collective wisdom of the users and discover the hidden patterns and associations among the products and the users. A collaborative filtering recommendation engine can be implemented using Apache Spark ML on Amazon EMR, which are two services that can handle large-scale data processing and machine learning tasks. Apache Spark ML is a library that provides various tools and algorithms for machine learning, such as classification, regression, clustering, recommendation, etc. Apache Spark ML can run on Amazon EMR, which is a service that provides a managed cluster platform that simplifies running big data frameworks, such as Apache Spark, on AWS. Apache Spark ML on Amazon EMR can build a collaborative filtering recommendation engine using the Alternating Least Squares (ALS) algorithm, which is a matrix factorization technique that can learn the latent factors that represent the users and the products, and then use them to predict the ratings or preferences of the users for the products. Apache Spark ML on Amazon EMR can also support both explicit feedback, such as ratings or reviews, and implicit feedback, such as views or clicks, for building a collaborative filtering recommendation engine12

QUESTION 68
A Data Engineer needs to build a model using a dataset containing customer credit card information. How can the Data Engineer ensure the data remains encrypted and the credit card information is secure?

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

Answer: D

Explanation:

AWS KMS is a service that provides encryption and key management for data stored in AWS services and applications. AWS KMS can generate and manage encryption keys that are used to encrypt and decrypt data at rest and in transit. AWS KMS can also integrate with other AWS services, such as Amazon S3 and Amazon SageMaker, to enable encryption of data using the keys stored in AWS KMS. Amazon S3 is a service that provides object storage for data in the cloud. Amazon S3 can use AWS KMS to encrypt data at rest using server-side encryption with AWS KMS-managed keys (SSE-KMS). Amazon SageMaker is a service that provides a platform for building, training, and deploying machine learning models. Amazon SageMaker can use AWS KMS to encrypt data at rest on the SageMaker instances and volumes, as well as data in transit between SageMaker and other AWS services. AWS Glue is a service that provides a serverless data integration platform for data preparation and transformation. AWS Glue can use AWS KMS to encrypt data at rest on the Glue Data Catalog and Glue ETL jobs. AWS Glue can also use built-in or custom classifiers to identify and redact sensitive data, such as credit card numbers, from the customer data1234
The other options are not valid or secure ways to encrypt the data and protect the credit card information. Using a custom encryption algorithm to encrypt the data and store the data on an Amazon SageMaker instance in a VPC is not a good practice, as custom encryption algorithms are not recommended for security and may have flaws or vulnerabilities. Using the SageMaker DeepAR algorithm to randomize the credit card numbers is not a good practice, as DeepAR is a forecasting algorithm that is not designed for data anonymization or encryption. Using an IAM policy to encrypt the data on the Amazon S3 bucket and Amazon Kinesis to automatically discard credit card numbers and insert fake credit card numbers is not a good practice, as IAM policies are not meant for data encryption, but for access control and authorization. Amazon Kinesis is a service that provides realtime data streaming and processing, but it does not have the capability to automatically discard or insert data values. Using an Amazon SageMaker launch configuration to encrypt the data once it is copied to the SageMaker instance in a VPC is not a good practice, as launch configurations are not meant for data encryption, but for specifying the instance type, security group, and user data for the SageMaker instance. Using the SageMaker principal component analysis (PCA) algorithm to reduce the length of the credit card numbers is not a good practice, as PCA is a dimensionality reduction algorithm that is not designed for data anonymization or encryption.

QUESTION 69
A Machine Learning Specialist is using an Amazon SageMaker notebook instance in a private subnet of a corporate VPC. The ML Specialist has important data stored on the Amazon SageMaker notebook instance's Amazon EBS volume, and needs to take a snapshot of that EBS volume. However the ML Specialist cannot find the Amazon SageMaker notebook instance's EBS volume or Amazon EC2 instance within the VPC.
Why is the ML Specialist not seeing the instance visible in the VPC?

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

accounts.
C. Amazon SageMaker notebook instances are based on EC2 instances running within AWS service accounts.
D. Amazon SageMaker notebook instances are based on AWS ECS instances running within AWS service
accounts.

Answer: C

Explanation:
Amazon SageMaker notebook instances are fully managed environments that provide an integrated Jupyter notebook interface for data exploration, analysis, and machine learning. Amazon SageMaker notebook instances are based on EC2 instances that run within AWS service accounts, not within customer accounts. This means that the ML Specialist cannot find the Amazon SageMaker notebook instances EC2 instance or EBS volume within the VPC, as they are not visible or accessible to the customer. However, the ML Specialist can still take a snapshot of the EBS volume by using the Amazon SageMaker console or API. The ML Specialist can also use VPC interface endpoints to securely connect the Amazon SageMaker notebook instance to the resources within the VPC, such as Amazon S3 buckets, Amazon EFS file systems, or Amazon RDS databases

---

QUESTION 70
A manufacturing company has structured and unstructured data stored in an Amazon S3 bucket. A Machine Learning Specialist wants to use SQL to run queries on this data.
Which solution requires the LEAST effort to be able to query this data?

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

Answer: B

Explanation:
Using AWS Glue to catalogue the data and Amazon Athena to run queries is the solution that requires the least effort to be able to query the data stored in an Amazon S3 bucket using SQL. AWS Glue is a service that provides a serverless data integration platform for data preparation and transformation. AWS Glue can automatically discover, crawl, and catalogue the data stored in various sources, such as Amazon S3, Amazon RDS, Amazon Redshift, etc. AWS Glue can also use AWS KMS to encrypt the data at rest on the Glue Data Catalog and Glue ETL jobs. AWS Glue can handle both structured and unstructured data, and support various data formats, such as CSV, JSON, Parquet, etc. AWS Glue can also use built-in or custom classifiers to identify and parse the data schema and format1 Amazon Athena is a service that provides an interactive query engine that can run SQL

queries directly on data stored in Amazon S3. Amazon Athena can integrate with AWS Glue to use the Glue Data Catalog as a central metadata repository for the data sources and tables. Amazon Athena can also use AWS KMS to encrypt the data at rest on Amazon S3 and the query results. Amazon Athena can query both structured and unstructured data, and support various data formats, such as CSV, JSON, Parquet, etc. Amazon Athena can also use partitions and compression to optimize the query performance and reduce the query cost23
The other options are not valid or require more effort to query the data stored in an Amazon S3 bucket using SQL. Using AWS Data Pipeline to transform the data and Amazon RDS to run queries is not a good option, as it involves moving the data from Amazon S3 to Amazon RDS, which can incur additional time and cost. AWS Data Pipeline is a service that can orchestrate and automate data movement and transformation across various AWS services and on-premises data sources. AWS Data Pipeline can be integrated with Amazon EMR to run ETL jobs on the data stored in Amazon S3. Amazon RDS is a service that provides a managed relational database service that can run various database engines, such as MySQL, PostgreSQL, Oracle, etc. Amazon RDS can use AWS KMS to encrypt the data at rest and in transit. Amazon RDS can run SQL queries on the data stored in the database tables45 Using AWS Batch to run ETL on the data and Amazon Aurora to run the queries is not a good option, as it also involves moving the data from Amazon S3 to Amazon Aurora, which can incur additional time and cost. AWS Batch is a service that can run batch computing workloads on AWS. AWS Batch can be integrated with AWS Lambda to trigger ETL jobs on the data stored in Amazon S3. Amazon Aurora is a service that provides a compatible and scalable relational database engine that can run MySQL or PostgreSQL. Amazon Aurora can use AWS KMS to encrypt the data at rest and in transit. Amazon Aurora can run SQL queries on the data stored in the database tables. Using AWS Lambda to transform the data and Amazon Kinesis Data Analytics to run queries is not a good option, as it is not suitable for querying data stored in Amazon S3 using SQL. AWS Lambda is a service that can run serverless functions on AWS. AWS Lambda can be integrated with Amazon S3 to trigger data transformation functions on the data stored in Amazon S3. Amazon Kinesis Data Analytics is a service that can analyze streaming data using SQL or Apache Flink. Amazon Kinesis Data Analytics can be integrated with Amazon Kinesis Data Streams or Amazon Kinesis Data Firehose to ingest streaming data sources, such as web logs, social media, IoT devices, etc. Amazon Kinesis Data Analytics is not designed for querying data stored in Amazon S3 using SQL.

QUESTION 71
A Machine Learning Specialist receives customer data for an online shopping website. The data includes demographics, past visits, and locality information. The Specialist must develop a machine learning approach to identify the customer shopping patterns, preferences and trends to enhance the website for better service and smart recommendations.
Which solution should the Specialist recommend?

A. Latent Dirichlet Allocation (LDA) for the given collection of discrete data to identify patterns in the customer database.
B. A neural network with a minimum of three layers and random initial weights to identify patterns in the customer database

C. Collaborative filtering based on user interactions and correlations to identify patterns in the customer database
D. Random Cut Forest (RCF) over random subsamples to identify patterns in the customer database

Answer: C

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

QUESTION 72
A Machine Learning Specialist is working with a large company to leverage machine learning within its products. The company wants to group its customers into categories based on which customers will and will not churn within the next 6 months. The company has labeled the data available to the Specialist.
Which machine learning model type should the Specialist use to accomplish this task?

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

Answer: B

Explanation:
The goal of classification is to determine to which class or category a data point (customer in our case) belongs to. For classification problems, data scientists would use historical data with predefined target variables AKA labels (churner/non-churner) " answers that need to be predicted " to train an algorithm. With classification, businesses can answer the following questions:
Will this customer churn or not?
Will a customer renew their subscription?
Will a user downgrade a pricing plan?
Are there any signs of unusual customer behavior?
Reference: https://www.kdnuggets.com9/05/churn-prediction-machine-learning.html

---

QUESTION 73
The displayed graph is from a foresting model for testing a time series.

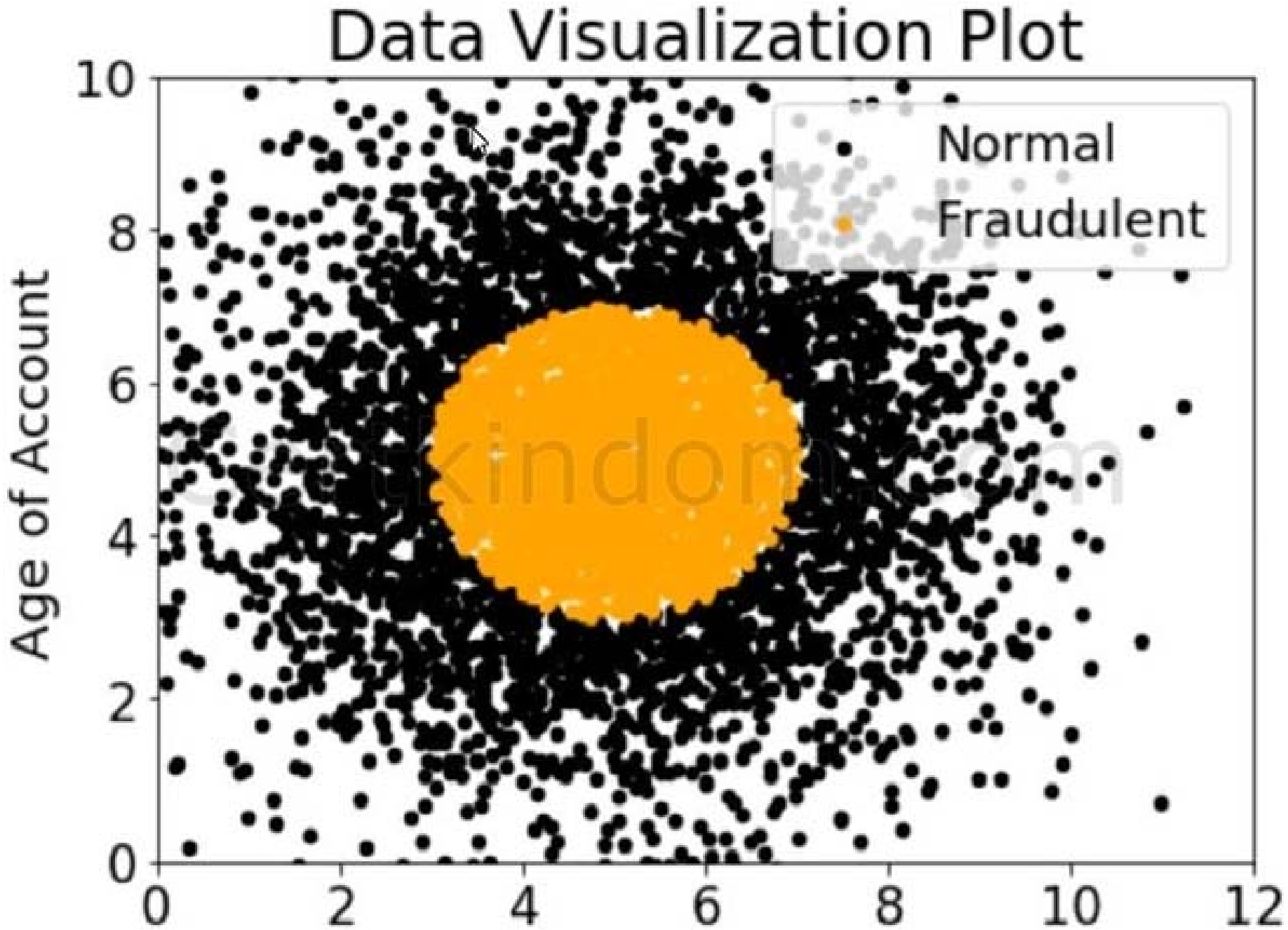
Considering the graph only, which conclusion should a Machine Learning Specialist make about the behavior of the model?

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

Answer: D

Explanation:

QUESTION 74
A company wants to classify user behavior as either fraudulent or normal. Based on internal research, a Machine Learning Specialist would like to build a binary classifier based on two features: age of account and transaction month. The class distribution for these features is illustrated in the figure provided.



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

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

Answer: C

Explanation:
Based on the figure provided, the data is not linearly separable. Therefore, a non-linear model such as SVM with a non-linear kernel would be the best choice. SVMs are particularly effective in highdimensional spaces and are versatile in that they can be used for both linear and non-linear data. Additionally, SVMs have a high level of accuracy and are less prone to overfitting1
Reference: 1: https://docs.aws.amazon.com/sagemaker/latest/dg/svm.html

---

QUESTION 75
A Machine Learning Specialist at a company sensitive to security is preparing a dataset for model training. The dataset is stored in Amazon S3 and contains Personally Identifiable Information (Pll).
The dataset:
* Must be accessible from a VPC only.
* Must not traverse the public internet.
How can these requirements be satisfied?

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

Answer: A

Explanation:
A VPC endpoint is a logical device that enables private connections between a VPC and supported AWS services. A VPC endpoint can be either a gateway endpoint or an interface endpoint. A gateway endpoint is a gateway that is a target for a specified route in the route table, used for traffic destined to a supported AWS service. An interface endpoint is an elastic network interface with a private IP address that serves as an entry point for traffic destined to a supported service1
In this case, the Machine Learning Specialist can create a gateway endpoint for Amazon S3, which is a supported service for gateway endpoints. A gateway endpoint for Amazon S3 enables the VPC to access Amazon S3 privately, without requiring an internet gateway, NAT device, VPN connection, or AWS Direct Connect connection. The traffic between the VPC and Amazon S3 does not leave the Amazon network2
To restrict access to the dataset stored in Amazon S3, the Machine Learning Specialist can apply a

bucket access policy that allows access only from the given VPC endpoint and the VPC. A bucket access policy is a resource-based policy that defines who can access a bucket and what actions they can perform. A bucket access policy can use various conditions to control access, such as the source IP address, the source VPC, the source VPC endpoint, etc. In this case, the Machine Learning Specialist can use the aws:sourceVpce condition to specify the ID of the VPC endpoint, and the aws:sourceVpc condition to specify the ID of the VPC. This way, only the requests that originate from the VPC endpoint or the VPC can access the bucket that contains the dataset34
The other options are not valid or secure ways to satisfy the requirements. Creating a VPC endpoint and applying a bucket access policy that allows access from the given VPC endpoint and an Amazon EC2 instance is not a good option, as it does not restrict access to the VPC. An Amazon EC2 instance is a virtual server that runs in the AWS cloud. An Amazon EC2 instance can have a public IP address or a private IP address, depending on the network configuration. Allowing access from an Amazon EC2 instance does not guarantee that the instance is in the same VPC as the VPC endpoint, and may expose the dataset to unauthorized access. Creating a VPC endpoint and using Network Access Control Lists (NACLs) to allow traffic between only the given VPC endpoint and an Amazon EC2 instance is not a good option, as it does not restrict access to the VPC. NACLs are stateless firewalls that can control inbound and outbound traffic at the subnet level. NACLs can use rules to allow or deny traffic based on the protocol, port, and source or destination IP address. However, NACLs do not support VPC endpoints as a source or destination, and cannot filter traffic based on the VPC endpoint ID or the VPC ID. Therefore, using NACLs does not guarantee that the traffic is from the VPC endpoint or the VPC, and may expose the dataset to unauthorized access. Creating a VPC endpoint and using security groups to restrict access to the given VPC endpoint and an Amazon EC2 instance is not a good option, as it does not restrict access to the VPC. Security groups are stateful firewalls that can control inbound and outbound traffic at the instance level. Security groups can use rules to allow or deny traffic based on the protocol, port, and source or destination. However, security groups do not support VPC endpoints as a source or destination, and cannot filter traffic based on the VPC endpoint ID or the VPC ID. Therefore, using security groups does not guarantee that the traffic is from the VPC endpoint or the VPC, and may expose the dataset to unauthorized access.

QUESTION 76
An employee found a video clip with audio on a company's social media feed. The language used in the video is Spanish. English is the employee's first language, and they do not understand Spanish. The employee wants to do a sentiment analysis.
What combination of services is the MOST efficient to accomplish the task?

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

Answer: A

Explanation:
Amazon Transcribe, Amazon Translate, and Amazon Comprehend are the most efficient combination of services to accomplish the task of sentiment analysis on a video clip with audio in Spanish. Amazon Transcribe is a service that can convert speech to text using deep learning. Amazon Transcribe can transcribe audio from various sources, such as video files, audio files, or streaming audio. Amazon Transcribe can also recognize multiple speakers, different languages, accents, dialects, and custom vocabularies. In this case, Amazon Transcribe can transcribe the audio from the video clip in Spanish to text in Spanish1 Amazon Translate is a service that can translate text from one language to another using neural machine translation. Amazon Translate can translate text from various sources, such as documents, web pages, chat messages, etc. Amazon Translate can also support multiple languages, domains, and styles. In this case, Amazon Translate can translate the text from Spanish to English2 Amazon Comprehend is a service that can analyze and derive insights from text using natural language processing. Amazon Comprehend can perform various tasks, such as sentiment analysis, entity recognition, key phrase extraction, topic modeling, etc. Amazon Comprehend can also support multiple languages and domains. In this case, Amazon Comprehend can perform sentiment analysis on the text in English and determine whether the feedback is positive, negative, neutral, or mixed3
The other options are not valid or efficient for accomplishing the task of sentiment analysis on a video clip with audio in Spanish. Amazon Comprehend, Amazon SageMaker seq2seq, and Amazon SageMaker Neural Topic Model (NTM) are not a good combination, as they do not include a service that can transcribe speech to text, which is a necessary step for processing the audio from the video clip. Amazon Comprehend, Amazon Translate, and Amazon SageMaker BlazingText are not a good combination, as they do not include a service that can perform sentiment analysis, which is the main goal of the task. Amazon SageMaker BlazingText is a service that can train and deploy text classification and word embedding models using deep learning. Amazon SageMaker BlazingText can perform tasks such as text classification, named entity recognition, part-of-speech tagging, etc., but not sentiment analysis4

QUESTION 77
A Machine Learning Specialist is packaging a custom ResNet model into a Docker container so the company can leverage Amazon SageMaker for training. The Specialist is using Amazon EC2 P3 instances to train the model and needs to properly configure the Docker container to leverage the NVIDIA GPUs.
What does the Specialist need to do?

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

Answer: B

Explanation:
To leverage the NVIDIA GPUs on Amazon EC2 P3 instances for training a custom ResNet model using Amazon SageMaker, the Machine Learning Specialist needs to build the Docker container to be NVIDIA-Docker compatible. NVIDIA-Docker is a tool that enables GPU-accelerated containers to run on Docker. NVIDIA-Docker can automatically configure the Docker container with the necessary drivers, libraries, and environment variables to access the NVIDIA GPUs. NVIDIA-Docker can also isolate the GPU resources and ensure that each container has exclusive access to a GPU.
To build a Docker container that is NVIDIA-Docker compatible, the Machine Learning Specialist needs to follow these steps:
Install the NVIDIA Container Toolkit on the host machine that runs Docker. This toolkit includes the NVIDIA Container Runtime, which is a modified version of the Docker runtime that supports GPU hardware.
Use the base image provided by NVIDIA as the first line of the Dockerfile. The base image contains the NVIDIA drivers and CUDA toolkit that are required for GPU-accelerated applications. The base image can be specified as FROM nvcr.io/nvidia/cuda:tag, where tag is the version of CUDA and the operating system.
Install the required dependencies and frameworks for the ResNet model, such as PyTorch, torchvision, etc., in the Dockerfile.
Copy the ResNet model code and any other necessary files to the Docker container in the Dockerfile.
Build the Docker image using the docker build command.
Push the Docker image to a repository, such as Amazon Elastic Container Registry (Amazon ECR), using the docker push command.
Specify the Docker image URI and the instance type (ml.p3.xlarge) in the Amazon SageMaker CreateTrainingJob request body.
The other options are not valid or sufficient for building a Docker container that can leverage the NVIDIA GPUs on Amazon EC2 P3 instances. Bundling the NVIDIA drivers with the Docker image is not a good option, as it can cause driver conflicts and compatibility issues with the host machine and the NVIDIA GPUs. Organizing the Docker containers file structure to execute on GPU instances is not a good option, as it does not ensure that the Docker container can access the NVIDIA GPUs and the CUDA toolkit. Setting the GPU flag in the Amazon SageMaker CreateTrainingJob request body is not a good option, as it does not apply to custom Docker containers, but only to built-in algorithms and frameworks that support GPU instances.

QUESTION 78
A Machine Learning Specialist is building a logistic regression model that will predict whether or not a person will order a pizz
a. The Specialist is trying to build the optimal model with an ideal classification threshold.
What model evaluation technique should the Specialist use to understand how different classification thresholds will impact the model's performance?

A. Receiver operating characteristic (ROC) curve
B. Misclassification rate

C. Root Mean Square Error (RM&)
D. L1 norm

Answer: A

Explanation:
A receiver operating characteristic (ROC) curve is a model evaluation technique that can be used to understand how different classification thresholds will impact the models performance. A ROC curve plots the true positive rate (TPR) against the false positive rate (FPR) for various values of the classification threshold. The TPR, also known as sensitivity or recall, is the proportion of positive instances that are correctly classified as positive. The FPR, also known as the fall-out, is the proportion of negative instances that are incorrectly classified as positive. A ROC curve can show the trade-off between the TPR and the FPR for different thresholds, and help the Machine Learning Specialist to select the optimal threshold that maximizes the TPR and minimizes the FPR. A ROC curve can also be used to compare the performance of different models by calculating the area under the curve (AUC), which is a measure of how well the model can distinguish between the positive and negative classes. A higher AUC indicates a better model

---

QUESTION 79
An interactive online dictionary wants to add a widget that displays words used in similar contexts. A Machine Learning Specialist is asked to provide word features for the downstream nearest neighbor model powering the widget.
What should the Specialist do to meet these requirements?

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

Answer: D

Explanation:
Word embeddings are a type of dense representation of words, which encode semantic meaning in a vector form. These embeddings are typically pre-trained on a large corpus of text data, such as a large set of books, news articles, or web pages, and capture the context in which words are used. Word embeddings can be used as features for a nearest neighbor model, which can be used to find words used in similar contexts. Downloading pre-trained word embeddings is a good way to get started quickly and leverage the strengths of these representations, which have been optimized on a large amount of data. This is likely to result in more accurate and reliable features than other options like one-hot encoding, edit distance, or using Amazon Mechanical Turk to produce synonyms.
Reference: https://aws.amazon.com/blogs/machine-learning/amazon-sagemaker-object2vec-addsnew-features-that-support-automatic-negative-sampling-and-speed-up-training/

QUESTION 80
A Machine Learning Specialist is configuring Amazon SageMaker so multiple Data Scientists can access notebooks, train models, and deploy endpoints. To ensure the best operational performance, the Specialist needs to be able to track how often the Scientists are deploying models, GPU and CPU utilization on the deployed SageMaker endpoints, and all errors that are generated when an endpoint is invoked.
Which services are integrated with Amazon SageMaker to track this information? (Select TWO.)

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

Answer: A, D

Explanation:
The services that are integrated with Amazon SageMaker to track the information that the Machine Learning Specialist needs are AWS CloudTrail and Amazon CloudWatch. AWS CloudTrail is a service that records the API calls and events for AWS services, including Amazon SageMaker. AWS CloudTrail can track the actions performed by the Data Scientists, such as creating notebooks, training models, and deploying endpoints. AWS CloudTrail can also provide information such as the identity of the user, the time of the action, the parameters used, and the response elements returned. AWS CloudTrail can help the Machine Learning Specialist to monitor the usage and activity of Amazon SageMaker, as well as to audit and troubleshoot any issues1 Amazon CloudWatch is a service that collects and analyzes the metrics and logs for AWS services, including Amazon SageMaker. Amazon CloudWatch can track the performance and utilization of the Amazon SageMaker endpoints, such as the CPU and GPU utilization, the inference latency, the number of invocations, etc. Amazon CloudWatch can also track the errors and alarms that are generated when an endpoint is invoked, such as the model errors, the throttling errors, the HTTP errors, etc. Amazon CloudWatch can help the Machine Learning Specialist to optimize the operational performance and reliability of Amazon SageMaker, as well as to set up notifications and actions based on the metrics and logs

---

QUESTION 81
A Machine Learning Specialist trained a regression model, but the first iteration needs optimizing. The Specialist needs to understand whether the model is more frequently overestimating or underestimating the target.
What option can the Specialist use to determine whether it is overestimating or underestimating the target value?

A. Root Mean Square Error (RMSE)

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

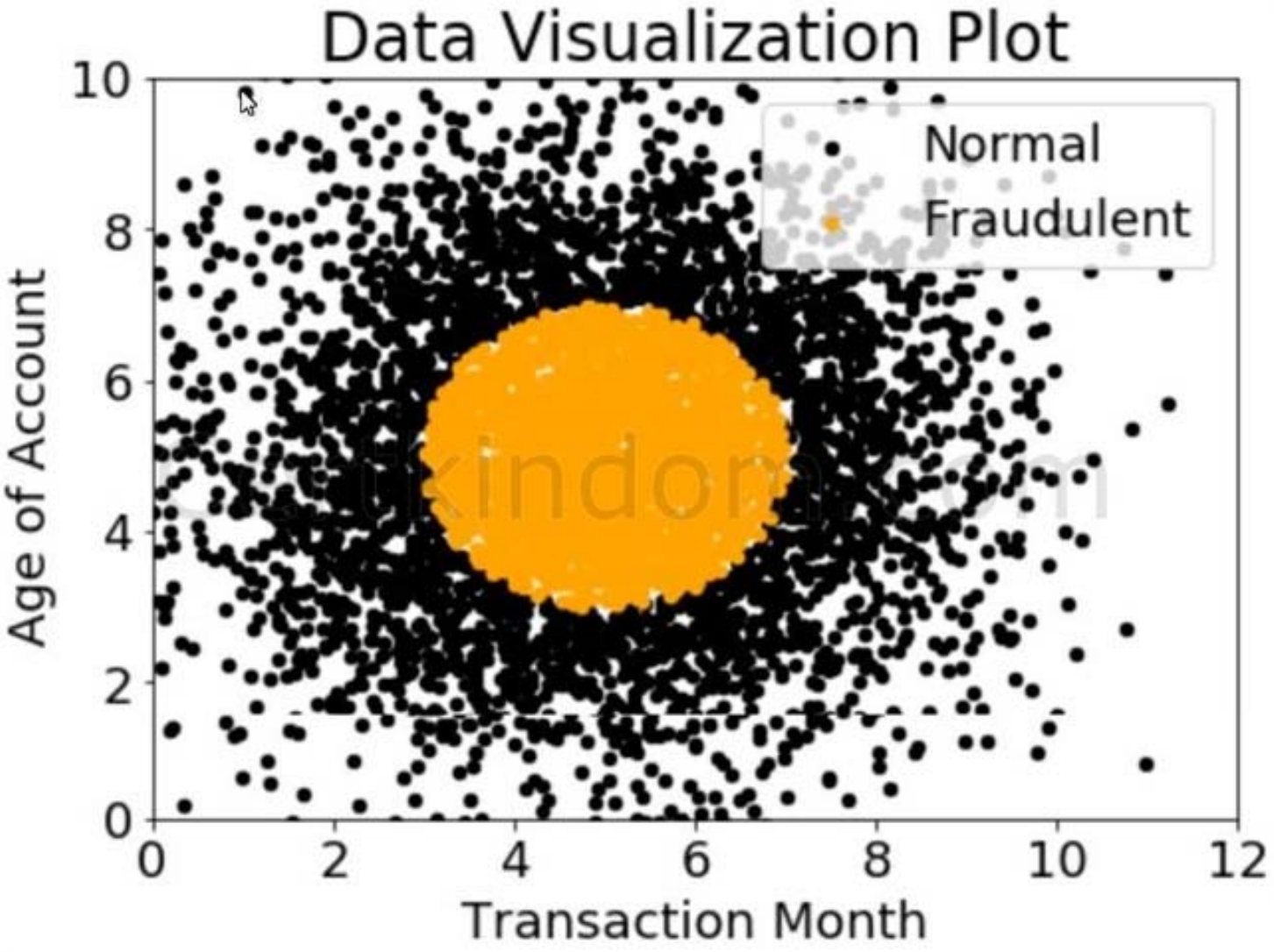
Answer: B

Explanation:
Residual plots are a model evaluation technique that can be used to understand whether a regression model is more frequently overestimating or underestimating the target. Residual plots are graphs that plot the residuals (the difference between the actual and predicted values) against the predicted values or other variables. Residual plots can help the Machine Learning Specialist to identify the patterns and trends in the residuals, such as the direction, shape, and distribution. Residual plots can also reveal the presence of outliers, heteroscedasticity, non-linearity, or other problems in the model12
To determine whether the model is overestimating or underestimating the target, the Machine Learning Specialist can use a residual plot that plots the residuals against the predicted values. This type of residual plot is also known as a prediction error plot. A prediction error plot can show the magnitude and direction of the errors made by the model. If the model is overestimating the target, the residuals will be negative, and the points will be below the zero line. If the model is underestimating the target, the residuals will be positive, and the points will be above the zero line. If the model is accurate, the residuals will be close to zero, and the points will be scattered around the zero line. A prediction error plot can also show the variance and bias of the model. If the model has high variance, the residuals will have a large spread, and the points will be far from the zero line. If the model has high bias, the residuals will have a systematic pattern, such as a curve or a slope, and the points will not be randomly distributed around the zero line. A prediction error plot can help the Machine Learning Specialist to optimize the model by adjusting the complexity, features, or parameters of the model34
The other options are not valid or suitable for determining whether the model is overestimating or underestimating the target. Root Mean Square Error (RMSE) is a model evaluation metric that measures the average magnitude of the errors made by the model. RMSE is the square root of the mean of the squared residuals. RMSE can indicate the overall accuracy and performance of the model, but it cannot show the direction or distribution of the errors. RMSE can also be influenced by outliers or extreme values, and it may not be comparable across different models or datasets5 Area under the curve (AUC) is a model evaluation metric that measures the ability of the model to distinguish between the positive and negative classes. AUC is the area under the receiver operating characteristic (ROC) curve, which plots the true positive rate against the false positive rate for various classification thresholds. AUC can indicate the overall quality and performance of the model, but it is only applicable for binary classification models, not regression models. AUC cannot show the magnitude or direction of the errors made by the model. Confusion matrix is a model evaluation technique that summarizes the number of correct and incorrect predictions made by the model for each class. A confusion matrix is a table that shows the counts of true positives, false positives, true negatives, and false negatives for each class. A confusion matrix can indicate the accuracy, precision,

recall, and F1-score of the model for each class, but it is only applicable for classification models, not regression models. A confusion matrix cannot show the magnitude or direction of the errors made by the model.

QUESTION 82

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



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

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

Answer: A

Explanation:
Based on the figure provided, a decision tree would have the highest recall with respect to the fraudulent class. Recall is a model evaluation metric that measures the proportion of actual positive instances that are correctly classified by the model. Recall is calculated as follows:
Recall = True Positives / (True Positives + False Negatives)
A decision tree is a type of machine learning model that can perform classification tasks by splitting the data into smaller and purer subsets based on a series of rules or conditions. A decision tree can handle both linear and non-linear data, and can capture complex patterns and interactions among the features. A decision tree can also be easily visualized and interpreted1
In this case, the data is not linearly separable, and has a clear pattern of seasonality. The fraudulent class forms a large circle in the center of the plot, while the normal class is scattered around the edges. A decision tree can use the transaction month and the age of account as the splitting criteria, and create a circular boundary that separates the fraudulent class from the normal class. A decision tree can achieve a high recall for the fraudulent class, as it can correctly identify most of the black dots as positive instances, and minimize the number of false negatives. A decision tree can also adjust the depth and complexity of the tree to balance the trade-off between recall and precision23
The other options are not valid or suitable for achieving a high recall for the fraudulent class. A linear support vector machine (SVM) is a type of machine learning model that can perform classification tasks by finding a linear hyperplane that maximizes the margin between the classes. A linear SVM can handle linearly separable data, but not non-linear data. A linear SVM cannot capture the circular pattern of the fraudulent class, and may misclassify many of the black dots as negative instances, resulting in a low recall4 A naive Bayesian classifier is a type of machine learning model that can perform classification tasks by applying the Bayes theorem and assuming conditional independence among the features. A naive Bayesian classifier can handle both linear and non-linear data, and can incorporate prior knowledge and probabilities into the model. However, a naive Bayesian classifier may not perform well when the features are correlated or dependent, as in this case. A naive Bayesian classifier may not capture the circular pattern of the fraudulent class, and may misclassify many of the black dots as negative instances, resulting in a low recall5 A single perceptron with sigmoidal activation function is a type of machine learning model that can perform classification tasks by applying a weighted linear combination of the features and a non-linear activation function. A single perceptron with sigmoidal activation function can handle linearly separable data, but not non-linear data. A single perceptron with sigmoidal activation function cannot capture the circular pattern of the fraudulent class, and may misclassify many of the black dots as negative instances, resulting in a low recall.

QUESTION 83
When submitting Amazon SageMaker training jobs using one of the built-in algorithms, which common parameters MUST be specified? (Select THREE.)

A. The training channel identifying the location of training data on an Amazon S3 bucket.
B. The validation channel identifying the location of validation data on an Amazon S3 bucket.
C. The 1AM role that Amazon SageMaker can assume to perform tasks on behalf of the users.
D. Hyperparameters in a JSON array as documented for the algorithm used.
E. The Amazon EC2 instance class specifying whether training will be run using CPU or GPU.
F. The output path specifying where on an Amazon S3 bucket the trained model will persist.

Answer: A, C, F

Explanation:
When submitting Amazon SageMaker training jobs using one of the built-in algorithms, the common parameters that must be specified are:
The training channel identifying the location of training data on an Amazon S3 bucket. This parameter tells SageMaker where to find the input data for the algorithm and what format it is in. For example, TrainingInputMode: File means that the input data is in files stored in S3.
The IAM role that Amazon SageMaker can assume to perform tasks on behalf of the users. This parameter grants SageMaker the necessary permissions to access the S3 buckets, ECR repositories, and other AWS resources needed for the training job. For example, RoleArn: arn:aws:iam::123456789012:role/service-role/AmazonSageMaker-ExecutionRole-20200303T150948 means that SageMaker will use the specified role to run the training job.
The output path specifying where on an Amazon S3 bucket the trained model will persist. This parameter tells SageMaker where to save the model artifacts, such as the model weights and parameters, after the training job is completed. For example, OutputDataConfig: {S3OutputPath: s3://my-bucket/my-training-job} means that SageMaker will store the model artifacts in the specified S3 location.
The validation channel identifying the location of validation data on an Amazon S3 bucket is an optional parameter that can be used to provide a separate dataset for evaluating the model performance during the training process. This parameter is not required for all algorithms and can be omitted if the validation data is not available or not needed.
The hyperparameters in a JSON array as documented for the algorithm used is another optional parameter that can be used to customize the behavior and performance of the algorithm. This parameter is specific to each algorithm and can be used to tune the model accuracy, speed, complexity, and other aspects. For example, HyperParameters: {num_round: "10", objective: "binary:logistic"} means that the XGBoost algorithm will use 10 boosting rounds and the logistic loss function for binary classification.
The Amazon EC2 instance class specifying whether training will be run using CPU or GPU is not a parameter that is specified when submitting a training job using a built-in algorithm. Instead, this

parameter is specified when creating a training instance, which is a containerized environment that runs the training code and algorithm. For example, ResourceConfig: {InstanceType: ml.m5.xlarge, InstanceCount: 1, VolumeSizeInGB: 10} means that SageMaker will use one m5.xlarge instance with 10 GB of storage for the training instance.
Reference:
Train a Model with Amazon SageMaker
Use Amazon SageMaker Built-in Algorithms or Pre-trained Models
CreateTrainingJob - Amazon SageMaker Service

QUESTION 84
A Data Scientist is developing a machine learning model to predict future patient outcomes based on information collected about each patient and their treatment plans. The model should output a continuous value as its prediction. The data available includes labeled outcomes for a set of 4,000 patients. The study was conducted on a group of individuals over the age of 65 who have a particular disease that is known to worsen with age.
Initial models have performed poorly. While reviewing the underlying data, the Data Scientist notices that, out of 4,000 patient observations, there are 450 where the patient age has been input as 0. The other features for these observations appear normal compared to the rest of the sample population.
How should the Data Scientist correct this issue?

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

Answer: B

Explanation:
The best way to handle the missing values in the patient age feature is to replace them with the mean or median value from the dataset. This is a common technique for imputing missing values that preserves the overall distribution of the data and avoids introducing bias or reducing the sample size. Dropping the records or the feature would result in losing valuable information and reducing the accuracy of the model. Using k-means clustering would not be appropriate for handling missing values in a single feature, as it is a method for grouping similar data points based on multiple features.
Reference:
Effective Strategies to Handle Missing Values in Data Analysis
How To Handle Missing Values In Machine Learning Data With Weka
How to handle missing values in Python - Machine Learning Plus

---

QUESTION 85
A Data Science team is designing a dataset repository where it will store a large amount of training data commonly used in its machine learning models. As Data Scientists may create an arbitrary number of new datasets every day the solution has to scale automatically and be cost-effective. Also, it must be possible to explore the data using SQL.
Which storage scheme is MOST adapted to this scenario?

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

Answer: A

Explanation:
The best storage scheme for this scenario is to store datasets as files in Amazon S3. Amazon S3 is a scalable, cost-effective, and durable object storage service that can store any amount and type of data. Amazon S3 also supports querying data using SQL with Amazon Athena, a serverless interactive query service that can analyze data directly in S3. This way, the Data Science team can easily explore and analyze their datasets without having to load them into a database or a compute instance.
The other options are not as suitable for this scenario because:
Storing datasets as files in an Amazon EBS volume attached to an Amazon EC2 instance would limit the scalability and availability of the data, as EBS volumes are only accessible within a single availability zone and have a maximum size of 16 TiB. Also, EBS volumes are more expensive than S3 buckets and require provisioning and managing EC2 instances.
Storing datasets as tables in a multi-node Amazon Redshift cluster would incur higher costs and complexity than using S3 and Athena. Amazon Redshift is a data warehouse service that is optimized for analytical queries over structured or semi-structured data. However, it requires setting up and maintaining a cluster of nodes, loading data into tables, and choosing the right distribution and sort keys for optimal performance. Moreover, Amazon Redshift charges for both storage and compute, while S3 and Athena only charge for the amount of data stored and scanned, respectively.
Storing datasets as global tables in Amazon DynamoDB would not be feasible for large amounts of data, as DynamoDB is a key-value and document database service that is designed for fast and consistent performance at any scale. However, DynamoDB has a limit of 400 KB per item and 25 GB per partition key value, which may not be enough for storing large datasets. Also, DynamoDB does not support SQL queries natively, and would require using a service like Amazon EMR or AWS Glue to run SQL queries over DynamoDB data.
Reference:
Amazon S3 - Cloud Object Storage
Amazon Athena " Interactive SQL Queries for Data in Amazon S3
Amazon EBS - Amazon Elastic Block Store (EBS)

Amazon Redshift " Data Warehouse Solution - AWS
Amazon DynamoDB " NoSQL Cloud Database Service

QUESTION 86
A Machine Learning Specialist working for an online fashion company wants to build a data ingestion solution for the company's Amazon S3-based data lake.
The Specialist wants to create a set of ingestion mechanisms that will enable future capabilities comprised of:
Real-time analytics
Interactive analytics of historical data
Clickstream analytics
Product recommendations
Which services should the Specialist use?

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

Answer: A

Explanation:
The best services to use for building a data ingestion solution for the companys Amazon S3-based data lake are:
AWS Glue as the data catalog: AWS Glue is a fully managed extract, transform, and load (ETL) service that can discover, crawl, and catalog data from various sources and formats, and make it available for analysis. AWS Glue can also generate ETL code in Python or Scala to transform, enrich, and join data using AWS Glue Data Catalog as the metadata repository. AWS Glue Data Catalog is a central metadata store that integrates with Amazon Athena, Amazon EMR, and Amazon Redshift Spectrum, allowing users to create a unified view of their data across various sources and formats.
Amazon Kinesis Data Streams and Amazon Kinesis Data Analytics for real-time data insights: Amazon Kinesis Data Streams is a service that enables users to collect, process, and analyze real-time streaming data at any scale. Users can create data streams that can capture data from various sources, such as web and mobile applications, IoT devices, and social media platforms. Amazon

Kinesis Data Analytics is a service that allows users to analyze streaming data using standard SQL queries or Apache Flink applications. Users can create real-time dashboards, metrics, and alerts based on the streaming data analysis results.
Amazon Kinesis Data Firehose for delivery to Amazon ES for clickstream analytics: Amazon Kinesis Data Firehose is a service that enables users to load streaming data into data lakes, data stores, and analytics services. Users can configure Kinesis Data Firehose to automatically deliver data to various destinations, such as Amazon S3, Amazon Redshift, Amazon OpenSearch Service, and third-party solutions. For clickstream analytics, users can use Kinesis Data Firehose to deliver data to Amazon OpenSearch Service, a fully managed service that offers search and analytics capabilities for log data. Users can use Amazon OpenSearch Service to perform interactive analysis and visualization of clickstream data using Kibana, an open-source tool that is integrated with Amazon OpenSearch Service.
Amazon EMR to generate personalized product recommendations: Amazon EMR is a service that enables users to run distributed data processing frameworks, such as Apache Spark, Apache Hadoop, and Apache Hive, on scalable clusters of EC2 instances. Users can use Amazon EMR to perform advanced analytics, such as machine learning, on large and complex datasets stored in Amazon S3 or other sources. For product recommendations, users can use Amazon EMR to run Spark MLlib, a library that provides scalable machine learning algorithms, such as collaborative filtering, to generate personalized recommendations based on user behavior and preferences.
Reference:
AWS Glue - Fully Managed ETL Service
Amazon Kinesis - Data Streaming Service
Amazon OpenSearch Service - Managed OpenSearch Service
Amazon EMR - Managed Hadoop Framework

QUESTION 87
A company is observing low accuracy while training on the default built-in image classification algorithm in Amazon SageMaker. The Data Science team wants to use an Inception neural network architecture instead of a ResNet architecture.
Which of the following will accomplish this? (Select TWO.)

A. Customize the built-in image classification algorithm to use Inception and use this for model training.
B. Create a support case with the SageMaker team to change the default image classification algorithm to Inception.
C. Bundle a Docker container with TensorFlow Estimator loaded with an Inception network and use this for model training.
D. Use custom code in Amazon SageMaker with TensorFlow Estimator to load the model with an Inception network and use this for model training.
E. Download and apt-get install the inception network code into an Amazon EC2 instance and use this instance as a Jupyter notebook in Amazon SageMaker.

Answer: C, D

Explanation:
The best options to use an Inception neural network architecture instead of a ResNet architecture for image classification in Amazon SageMaker are:
Bundle a Docker container with TensorFlow Estimator loaded with an Inception network and use this for model training. This option allows users to customize the training environment and use any TensorFlow model they want. Users can create a Docker image that contains the TensorFlow Estimator API and the Inception model from the TensorFlow Hub, and push it to Amazon ECR. Then, users can use the SageMaker Estimator class to train the model using the custom Docker image and the training data from Amazon S3.
Use custom code in Amazon SageMaker with TensorFlow Estimator to load the model with an Inception network and use this for model training. This option allows users to use the built-in TensorFlow container provided by SageMaker and write custom code to load and train the Inception model. Users can use the TensorFlow Estimator class to specify the custom code and the training data from Amazon S3. The custom code can use the TensorFlow Hub module to load the Inception model and fine-tune it on the training data.
The other options are not feasible for this scenario because:
Customize the built-in image classification algorithm to use Inception and use this for model training. This option is not possible because the built-in image classification algorithm in SageMaker does not support customizing the neural network architecture. The built-in algorithm only supports ResNet models with different depths and widths.
Create a support case with the SageMaker team to change the default image classification algorithm to Inception. This option is not realistic because the SageMaker team does not provide such a service. Users cannot request the SageMaker team to change the default algorithm or add new algorithms to the built-in ones.
Download and apt-get install the inception network code into an Amazon EC2 instance and use this instance as a Jupyter notebook in Amazon SageMaker. This option is not advisable because it does not leverage the benefits of SageMaker, such as managed training and deployment, distributed training, and automatic model tuning. Users would have to manually install and configure the Inception network code and the TensorFlow framework on the EC2 instance, and run the training and inference code on the same instance, which may not be optimal for performance and scalability.
Reference:
Use Your Own Algorithms or Models with Amazon SageMaker
Use the SageMaker TensorFlow Serving Container
TensorFlow Hub

QUESTION 88
A Machine Learning Specialist built an image classification deep learning model. However the Specialist ran into an overfitting problem in which the training and testing accuracies were 99% and 75%r respectively.
How should the Specialist address this issue and what is the reason behind it?

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

Answer: B

Explanation:
The best way to address the overfitting problem in image classification is to increase the dropout rate at the flatten layer because the model is not generalized enough. Dropout is a regularization technique that randomly drops out some units from the neural network during training, reducing the co-adaptation of features and preventing overfitting. The flatten layer is the layer that converts the output of the convolutional layers into a one-dimensional vector that can be fed into the dense layers. Increasing the dropout rate at the flatten layer means that more features from the convolutional layers will be ignored, forcing the model to learn more robust and generalizable representations from the remaining features.
The other options are not correct for this scenario because:
Increasing the learning rate would not help with the overfitting problem, as it would make the optimization process more unstable and prone to overshooting the global minimum. A high learning rate can also cause the model to diverge or oscillate around the optimal solution, resulting in poor performance and accuracy.
Increasing the dimensionality of the dense layer next to the flatten layer would not help with the overfitting problem, as it would make the model more complex and increase the number of parameters to be learned. A more complex model can fit the training data better, but it can also memorize the noise and irrelevant details in the data, leading to overfitting and poor generalization.
Increasing the epoch number would not help with the overfitting problem, as it would make the model train longer and more likely to overfit the training data. A high epoch number can cause the model to converge to the global minimum, but it can also cause the model to over-optimize the training data and lose the ability to generalize to new data.
Reference:
Dropout: A Simple Way to Prevent Neural Networks from Overfitting
How to Reduce Overfitting With Dropout Regularization in Keras
How to Control the Stability of Training Neural Networks With the Learning Rate
How to Choose the Number of Hidden Layers and Nodes in a Feedforward Neural Network?
How to decide the optimal number of epochs to train a neural network?

QUESTION 89
A Machine Learning team uses Amazon SageMaker to train an Apache MXNet handwritten digit classifier model using a research dataset. The team wants to receive a notification when the model is overfitting. Auditors want to view the Amazon SageMaker log activity report to ensure there are no unauthorized API calls.
What should the Machine Learning team do to address the requirements with the least amount of code and fewest steps?

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

Answer: B

Explanation:
To log Amazon SageMaker API calls, the team can use AWS CloudTrail, which is a service that provides a record of actions taken by a user, role, or an AWS service in SageMaker1. CloudTrail captures all API calls for SageMaker, with the exception of InvokeEndpoint and InvokeEndpointAsync, as events1. The calls captured include calls from the SageMaker console and code calls to the SageMaker API operations1. The team can create a trail to enable continuous delivery of CloudTrail events to an Amazon S3 bucket, and configure other AWS services to further analyze and act upon the event data collected in CloudTrail logs1. The auditors can view the CloudTrail log activity report in the CloudTrail console or download the log files from the S3 bucket1.
To receive a notification when the model is overfitting, the team can add code to push a custom metric to Amazon CloudWatch, which is a service that provides monitoring and observability for AWS resources and applications2. The team can use the MXNet metric API to define and compute the custom metric, such as the validation accuracy or the validation loss, and use the boto3 CloudWatch client to put the metric data to CloudWatch3 . The team can then create an alarm in CloudWatch with Amazon SNS to receive a notification when the custom metric crosses a threshold that indicates overfitting . For example, the team can set the alarm to trigger when the validation loss increases for a certain number of consecutive periods, which means the model is learning the noise in the training data and not generalizing well to the validation data.
Reference:
1: Log Amazon SageMaker API Calls with AWS CloudTrail - Amazon SageMaker

2: What Is Amazon CloudWatch? - Amazon CloudWatch
3: Metric API ” Apache MXNet documentation
: CloudWatch ” Boto 3 Docs 1.20.21 documentation
: Creating Amazon CloudWatch Alarms - Amazon CloudWatch
: What is Amazon Simple Notification Service? - Amazon Simple Notification Service
: Overfitting and Underfitting - Machine Learning Crash Course

---

QUESTION 90
A Machine Learning Specialist is implementing a full Bayesian network on a dataset that describes public transit in New York City. One of the random variables is discrete, and represents the number of minutes New Yorkers wait for a bus given that the buses cycle every 10 minutes, with a mean of 3 minutes.
Which prior probability distribution should the ML Specialist use for this variable?

A. Poisson distribution ,
B. Uniform distribution
C. Normal distribution
D. Binomial distribution

Answer: A

Explanation:
The prior probability distribution for the discrete random variable that represents the number of minutes New Yorkers wait for a bus is a Poisson distribution. A Poisson distribution is suitable for modeling the number of events that occur in a fixed interval of time or space, given a known average rate of occurrence. In this case, the event is waiting for a bus, the interval is 10 minutes, and the average rate is 3 minutes. The Poisson distribution can capture the variability of the waiting time, which can range from 0 to 10 minutes, with different probabilities.
Reference:
1: Poisson Distribution - Amazon SageMaker
2: Poisson Distribution - Wikipedia

QUESTION 91
A Data Science team within a large company uses Amazon SageMaker notebooks to access data stored in Amazon S3 buckets. The IT Security team is concerned that internet-enabled notebook instances create a security vulnerability where malicious code running on the instances could compromise data privacy. The company mandates that all instances stay within a secured VPC with no internet access, and data communication traffic must stay within the AWS network.
How should the Data Science team configure the notebook instance placement to meet these requirements?

A. Associate the Amazon SageMaker notebook with a private subnet in a VPC. Place the Amazon

SageMaker endpoint and S3 buckets within the same VPC.
B. Associate the Amazon SageMaker notebook with a private subnet in a VPC. Use 1AM policies to grant access to Amazon S3 and Amazon SageMaker.
C. Associate the Amazon SageMaker notebook with a private subnet in a VPC. Ensure the VPC has S3 VPC endpoints and Amazon SageMaker VPC endpoints attached to it.
D. Associate the Amazon SageMaker notebook with a private subnet in a VPC. Ensure the VPC has a NAT gateway and an associated security group allowing only outbound connections to Amazon S3 and Amazon SageMaker

Answer: C

Explanation:
To configure the notebook instance placement to meet the requirements, the Data Science team should associate the Amazon SageMaker notebook with a private subnet in a VPC. A VPC is a virtual network that is logically isolated from other networks in AWS. A private subnet is a subnet that has no internet gateway attached to it, and therefore cannot communicate with the internet. By placing the notebook instance in a private subnet, the team can ensure that it stays within a secured VPC with no internet access.
However, to access data stored in Amazon S3 buckets and other AWS services, the team needs to ensure that the VPC has S3 VPC endpoints and Amazon SageMaker VPC endpoints attached to it. A VPC endpoint is a gateway that enables private connections between the VPC and supported AWS services. A VPC endpoint does not require an internet gateway, a NAT device, or a VPN connection, and ensures that the traffic between the VPC and the AWS service does not leave the AWS network. By using VPC endpoints, the team can access Amazon S3 and Amazon SageMaker from the notebook instance without compromising data privacy or security.
Reference:
: What Is Amazon VPC? - Amazon Virtual Private Cloud
: Subnet Routing - Amazon Virtual Private Cloud
: VPC Endpoints - Amazon Virtual Private Cloud

QUESTION 92
A Machine Learning Specialist has created a deep learning neural network model that performs well on the training data but performs poorly on the test data.
Which of the following methods should the Specialist consider using to correct this? (Select THREE.)

A. Decrease regularization.
B. Increase regularization.
C. Increase dropout.
D. Decrease dropout.
E. Increase feature combinations.
F. Decrease feature combinations.

Answer: B, C, F

Explanation:
The problem of poor performance on the test data is a sign of overfitting, which means the model has learned the training data too well and failed to generalize to new and unseen data. To correct this, the Machine Learning Specialist should consider using methods that reduce the complexity of the model and increase its ability to generalize. Some of these methods are:
Increase regularization: Regularization is a technique that adds a penalty term to the loss function of the model, which reduces the magnitude of the model weights and prevents overfitting. There are different types of regularization, such as L1, L2, and elastic net, that apply different penalties to the weights1.
Increase dropout: Dropout is a technique that randomly drops out some units or connections in the neural network during training, which reduces the co-dependency of the units and prevents overfitting. Dropout can be applied to different layers of the network, and the dropout rate can be tuned to control the amount of dropout2.
Decrease feature combinations: Feature combinations are the interactions between different input features that can be used to create new features for the model. However, too many feature combinations can increase the complexity of the model and cause overfitting. Therefore, the Specialist should decrease the number of feature combinations and select only the most relevant and informative ones for the model3.
Reference:
1: Regularization for Deep Learning - Amazon SageMaker
2: Dropout - Amazon SageMaker
3: Feature Engineering - Amazon SageMaker

QUESTION 93
A Data Scientist needs to create a serverless ingestion and analytics solution for high-velocity, realtime streaming data.
The ingestion process must buffer and convert incoming records from JSON to a query-optimized, columnar format without data loss. The output datastore must be highly available, and Analysts must be able to run SQL queries against the data and connect to existing business intelligence dashboards.
Which solution should the Data Scientist build to satisfy the requirements?

A. Create a schema in the AWS Glue Data Catalog of the incoming data format. Use an Amazon Kinesis Data Firehose delivery stream to stream the data and transform the data to Apache Parquet or ORC format using the AWS Glue Data Catalog before delivering to Amazon S3. Have the Analysts query the data directly from Amazon S3 using Amazon Athena, and connect to Bl tools using the Athena Java Database Connectivity (JDBC) connector.
B. Write each JSON record to a staging location in Amazon S3. Use the S3 Put event to trigger an AWS Lambda function that transforms the data into Apache Parquet or ORC format and writes the data to a processed data location in Amazon S3. Have the Analysts query the data directly from Amazon S3 using Amazon Athena, and connect to Bl tools using the Athena Java Database

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

Answer: A

Explanation:
To create a serverless ingestion and analytics solution for high-velocity, real-time streaming data, the Data Scientist should use the following AWS services:
AWS Glue Data Catalog: This is a managed service that acts as a central metadata repository for data assets across AWS and on-premises data sources. The Data Scientist can use AWS Glue Data Catalog to create a schema of the incoming data format, which defines the structure, format, and data types of the JSON records. The schema can be used by other AWS services to understand and process the data1.
Amazon Kinesis Data Firehose: This is a fully managed service that delivers real-time streaming data to destinations such as Amazon S3, Amazon Redshift, Amazon Elasticsearch Service, and Splunk. The Data Scientist can use Amazon Kinesis Data Firehose to stream the data from the source and transform the data to a query-optimized, columnar format such as Apache Parquet or ORC using the AWS Glue Data Catalog before delivering to Amazon S3. This enables efficient compression, partitioning, and fast analytics on the data2.
Amazon S3: This is an object storage service that offers high durability, availability, and scalability. The Data Scientist can use Amazon S3 as the output datastore for the transformed data, which can be organized into buckets and prefixes according to the desired partitioning scheme. Amazon S3 also integrates with other AWS services such as Amazon Athena, Amazon EMR, and Amazon Redshift Spectrum for analytics3.
Amazon Athena: This is a serverless interactive query service that allows users to analyze data in Amazon S3 using standard SQL. The Data Scientist can use Amazon Athena to run SQL queries against the data in Amazon S3 and connect to existing business intelligence dashboards using the Athena Java Database Connectivity (JDBC) connector. Amazon Athena leverages the AWS Glue Data Catalog to access the schema information and supports formats such as Parquet and ORC for fast and costeffective queries4.
Reference:
1: What Is the AWS Glue Data Catalog? - AWS Glue
2: What Is Amazon Kinesis Data Firehose? - Amazon Kinesis Data Firehose
3: What Is Amazon S3? - Amazon Simple Storage Service
4: What Is Amazon Athena? - Amazon Athena

QUESTION 94
A company is setting up an Amazon SageMaker environment. The corporate data security policy does not allow communication over the internet.
How can the company enable the Amazon SageMaker service without enabling direct internet access to Amazon SageMaker notebook instances?

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

Answer: C

Explanation:
To enable the Amazon SageMaker service without enabling direct internet access to Amazon SageMaker notebook instances, the company should create Amazon SageMaker VPC interface endpoints within the corporate VPC. A VPC interface endpoint is a gateway that enables private connections between the VPC and supported AWS services without requiring an internet gateway, a NAT device, a VPN connection, or an AWS Direct Connect connection. The instances in the VPC do not need to connect to the public internet in order to communicate with the Amazon SageMaker service. The VPC interface endpoint connects the VPC directly to the Amazon SageMaker service using AWS PrivateLink, which ensures that the traffic between the VPC and the service does not leave the AWS network1.
Reference:
1: Connect to SageMaker Within your VPC - Amazon SageMaker

QUESTION 95
An office security agency conducted a successful pilot using 100 cameras installed at key locations within the main office. Images from the cameras were uploaded to Amazon S3 and tagged using Amazon Rekognition, and the results were stored in Amazon ES. The agency is now looking to expand the pilot into a full production system using thousands of video cameras in its office locations globally. The goal is to identify activities performed by non-employees in real time.
Which solution should the agency consider?

A. Use a proxy server at each local office and for each camera, and stream the RTSP feed to a unique Amazon Kinesis Video Streams video stream. On each stream, use Amazon Rekognition Video and create
a stream processor to detect faces from a collection of known employees, and alert when nonemployees are detected.
B. Use a proxy server at each local office and for each camera, and stream the RTSP feed to a unique Amazon Kinesis Video Streams video stream. On each stream, use Amazon Rekognition Image to

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

Answer: A

Explanation:
The solution that the agency should consider is to use a proxy server at each local office and for each camera, and stream the RTSP feed to a unique Amazon Kinesis Video Streams video stream. On each stream, use Amazon Rekognition Video and create a stream processor to detect faces from a collection of known employees, and alert when non-employees are detected.
This solution has the following advantages:
It can handle thousands of video cameras in real time, as Amazon Kinesis Video Streams can scale elastically to support any number of producers and consumers1.
It can leverage the Amazon Rekognition Video API, which is designed and optimized for video analysis, and can detect faces in challenging conditions such as low lighting, occlusions, and different poses2.
It can use a stream processor, which is a feature of Amazon Rekognition Video that allows you to create a persistent application that analyzes streaming video and stores the results in a Kinesis data stream3. The stream processor can compare the detected faces with a collection of known employees, which is a container for persisting faces that you want to search for in the input video stream4. The stream processor can also send notifications to Amazon Simple Notification Service (Amazon SNS) when non-employees are detected, which can trigger downstream actions such as sending alerts or storing the events in Amazon Elasticsearch Service (Amazon ES)3.
Reference:
1: What Is Amazon Kinesis Video Streams? - Amazon Kinesis Video Streams
2: Detecting and Analyzing Faces - Amazon Rekognition
3: Using Amazon Rekognition Video Stream Processor - Amazon Rekognition
4: Working with Stored Faces - Amazon Rekognition

QUESTION 96
A financial services company is building a robust serverless data lake on Amazon S3. The data lake should be flexible and meet the following requirements:

* Support querying old and new data on Amazon S3 through Amazon Athena and Amazon Redshift Spectrum.
* Support event-driven ETL pipelines.
* Provide a quick and easy way to understand metadata.
Which approach meets trfese requirements?

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

Answer: A

Explanation:
To build a robust serverless data lake on Amazon S3 that meets the requirements, the financial services company should use the following AWS services:
AWS Glue crawler: This is a service that connects to a data store, progresses through a prioritized list of classifiers to determine the schema for the data, and then creates metadata tables in the AWS Glue Data Catalog1. The company can use an AWS Glue crawler to crawl the S3 data and infer the schema, format, and partition structure of the data. The crawler can also detect schema changes and update the metadata tables accordingly. This enables the company to support querying old and new data on Amazon S3 through Amazon Athena and Amazon Redshift Spectrum, which are serverless interactive query services that use the AWS Glue Data Catalog as a central location for storing and retrieving table metadata23.
AWS Lambda function: This is a service that lets you run code without provisioning or managing servers. You pay only for the compute time you consume - there is no charge when your code is not running. You can also use AWS Lambda to create event-driven ETL pipelines, by triggering other AWS services based on events such as object creation or deletion in S3 buckets4. The company can use an AWS Lambda function to trigger an AWS Glue ETL job, which is a serverless way to extract, transform, and load data for analytics. The AWS Glue ETL job can perform various data processing tasks, such as converting data formats, filtering, aggregating, joining, and more.
AWS Glue Data Catalog: This is a managed service that acts as a central metadata repository for data assets across AWS and on-premises data sources. The AWS Glue Data Catalog provides a uniform repository where disparate systems can store and find metadata to keep track of data in data silos, and use that metadata to query and transform the data. The company can use the AWS Glue Data Catalog to search and discover metadata, such as table definitions, schemas, and partitions. The AWS Glue Data Catalog also integrates with Amazon Athena, Amazon Redshift Spectrum, Amazon EMR, and AWS Glue ETL jobs, providing a consistent view of the data across different query and analysis

services.
Reference:
1: What Is a Crawler? - AWS Glue
2: What Is Amazon Athena? - Amazon Athena
3: Amazon Redshift Spectrum - Amazon Redshift
4: What is AWS Lambda? - AWS Lambda
: AWS Glue ETL Jobs - AWS Glue
: What Is the AWS Glue Data Catalog? - AWS Glue

QUESTION 97
A company's Machine Learning Specialist needs to improve the training speed of a time-series forecasting model using TensorFlow. The training is currently implemented on a single-GPU machine and takes approximately 23 hours to complete. The training needs to be run daily.
The model accuracy js acceptable, but the company anticipates a continuous increase in the size of the training data and a need to update the model on an hourly, rather than a daily, basis. The company also wants to minimize coding effort and infrastructure changes
What should the Machine Learning Specialist do to the training solution to allow it to scale for future demand?

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

Answer: B

Explanation:
To improve the training speed of a time-series forecasting model using TensorFlow, the Machine Learning Specialist should change the TensorFlow code to implement a Horovod distributed framework supported by Amazon SageMaker. Horovod is a free and open-source software framework for distributed deep learning training using TensorFlow, Keras, PyTorch, and Apache MXNet1. Horovod can scale up to hundreds of GPUs with upwards of 90% scaling efficiency2. Horovod is easy to use, as it requires only a few lines of Python code to modify an existing training script2. Horovod is also portable, as it runs the same for TensorFlow, Keras, PyTorch, and MXNet; on premise, in the cloud, and on Apache Spark2.
Amazon SageMaker is a fully managed service that provides every developer and data scientist with the ability to build, train, and deploy machine learning models quickly3. Amazon SageMaker

supports Horovod as a built-in distributed training framework, which means that the Machine Learning Specialist does not need to install or configure Horovod separately4. Amazon SageMaker also provides a number of features and tools to simplify and optimize the distributed training process, such as automatic scaling, debugging, profiling, and monitoring4. By using Amazon SageMaker, the Machine Learning Specialist can parallelize the training to as many machines as needed to achieve the business goals, while minimizing coding effort and infrastructure changes.
Reference:
1: Horovod (machine learning) - Wikipedia
2: Home - Horovod
3: Amazon SageMaker " Machine Learning Service " AWS
4: Use Horovod with Amazon SageMaker - Amazon SageMaker

QUESTION 98
A Machine Learning Specialist is required to build a supervised image-recognition model to identify a cat. The ML Specialist performs some tests and records the following results for a neural networkbased image classifier:
Total number of images available = 1,000 Test set images = 100 (constant test set)
The ML Specialist notices that, in over 75% of the misclassified images, the cats were held upside down by their owners.
Which techniques can be used by the ML Specialist to improve this specific test error?

A. Increase the training data by adding variation in rotation for training images.
B. Increase the number of epochs for model training.
C. Increase the number of layers for the neural network.
D. Increase the dropout rate for the second-to-last layer.

Answer: A

Explanation:
To improve the test error for the image classifier, the Machine Learning Specialist should use the technique of increasing the training data by adding variation in rotation for training images. This technique is called data augmentation, which is a way of artificially expanding the size and diversity of the training dataset by applying various transformations to the original images, such as rotation, flipping, cropping, scaling, etc. Data augmentation can help the model learn more robust features that are invariant to the orientation, position, and size of the objects in the images. This can improve the generalization ability of the model and reduce the test error, especially for cases where the images are not well-aligned or have different perspectives1.
Reference:
1: Image Augmentation - Amazon SageMaker

QUESTION 99
A Data Scientist is developing a machine learning model to classify whether a financial transaction is

fraudulent. The labeled data available for training consists of 100,000 non-fraudulent observations and 1,000 fraudulent observations.
The Data Scientist applies the XGBoost algorithm to the data, resulting in the following confusion matrix when the trained model is applied to a previously unseen validation dataset. The accuracy of the model is 99.1%, but the Data Scientist has been asked to reduce the number of false negatives.

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

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

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

Answer: B, D

Explanation:
The XGBoost algorithm is a popular machine learning technique for classification problems. It is based on the idea of boosting, which is to combine many weak learners (decision trees) into a strong learner (ensemble model).
The XGBoost algorithm can handle imbalanced data by using the scale_pos_weight parameter, which controls the balance of positive and negative weights in the objective function. A typical value to consider is the ratio of negative cases to positive cases in the data. By increasing this parameter, the algorithm will pay more attention to the minority class (positive) and reduce the number of false negatives.
The XGBoost algorithm can also use different evaluation metrics to optimize the model performance. The default metric is error, which is the misclassification rate. However, this metric can be misleading for imbalanced data, as it does not account for the different costs of false positives and false negatives. A better metric to use is AUC, which is the area under the receiver operating characteristic (ROC) curve. The ROC curve plots the true positive rate against the false positive rate for different threshold values. The AUC measures how well the model can distinguish between the two classes, regardless of the threshold. By changing the eval_metric parameter to AUC, the algorithm will try to

maximize the AUC score and reduce the number of false negatives.
Therefore, the combination of steps that should be taken to reduce the number of false negatives are to increase the scale_pos_weight parameter and change the eval_metric parameter to AUC.
Reference:
XGBoost Parameters
XGBoost for Imbalanced Classification

QUESTION 100
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.
Reference:
SageMaker Docker GitHub repository
SageMaker Studio Image Build CLI
SageMaker Python SDK installation guide
SageMaker Python SDK TensorFlow documentation

QUESTION 101
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.
Reference:
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

QUESTION 102
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:
Amazon Athena is an interactive query service that allows you to analyze data stored in Amazon S3 using standard SQL. Athena is serverless, so you only pay for the queries that you run and there is no infrastructure to manage.
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.
Reference:
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. Its a Win-Win for your AWS bill. Supported formats: GZIP, LZO, SNAPPY (Parquet) and ZLIB.
Reference: https://www.cloudforecast.io/blog/using-parquet-on-athena-to-save-money-on-aws/

QUESTION 103
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.

QUESTION 104
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.
Reference:
Amazon SageMaker Principal Component Analysis (PCA)
Amazon SageMaker K-Means Algorithm

QUESTION 105
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.
Reference:
DeepAR Forecasting Algorithm - Amazon SageMaker
Now available in Amazon SageMaker: DeepAR algorithm for more accurate time series forecasting

---

QUESTION 106
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 œprocessed 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 œprocessed location in Amazon S3, using a format such as Parquet, ORC, JSON, or CSV, that is suitable for downstream processing.
Reference:
What Is AWS Glue?
AWS Glue Components
AWS Glue Studio
AWS Glue Triggers

QUESTION 107
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 users 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 users 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 users 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 users query, and the corresponding data from the reports.
Use a lambda function to implement the business logic and fulfillment of the users 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 users preference.
Reference:
What Is Amazon Lex?
What Is Amazon Comprehend?
What Is Amazon Transcribe?

QUESTION 108
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.
Reference:
Linear least squares (mathematics)
Linear Regression in Matrix Form
Singular Matrix Problem

QUESTION 109
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. Đ¡onfigure the Amazon SageMaker notebook instance to have access to the VPC. Grant permission in the
KMS key policy to the notebooks 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.
Reference:
Create an IAM Role to Grant Permissions to Your Notebook Instance
Using Key Policies in AWS KMS

QUESTION 110
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 teams
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 teams 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.
Reference:
Regularization for Logistic Regression
Recursive Feature Elimination

QUESTION 111
An aircraft engine manufacturing company is measuring 200 performance metrics in a time-series. Engineers
want to detect critical manufacturing defects in near-real time during testing. All of the data needs to be stored
for offline analysis.
What approach would be the MOST effective to perform near-real time defect detection?

A. Use AWS IoT Analytics for ingestion, storage, and further analysis. Use Jupyter notebooks from within
AWS IoT Analytics to carry out analysis for anomalies.
B. Use Amazon S3 for ingestion, storage, and further analysis. Use an Amazon EMR cluster to carry out
Apache Spark ML k-means clustering to determine anomalies.
C. Use Amazon S3 for ingestion, storage, and further analysis. Use the Amazon SageMaker Random

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

Answer: D

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.
Reference:
What Is Amazon Kinesis Data Firehose?
What Is Amazon Kinesis Data Analytics for SQL Applications?
DeepAR Forecasting Algorithm - Amazon SageMaker

QUESTION 112
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 algorithmspecific 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 algorithmspecific 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.
Reference:
Use Your Own Training Algorithms
Amazon ECR - Amazon Web Services
Amazon S3 - Amazon Web Services

QUESTION 113
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 cooccurrence 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. Reference:
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

---

QUESTION 114
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.
Reference:
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

---

QUESTION 115
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 users 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.
Reference:
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

---

QUESTION 116
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. Reference:
AWS Machine Learning Specialty Exam Guide
AWS Machine Learning Training - Dimensionality Reduction for Machine Learning
AWS Machine Learning Training - Deep Learning with Amazon SageMaker

---

QUESTION 117
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 MLPs 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 MLPs 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 MLPs 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.
Reference:
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

---

QUESTION 118
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 œdefault or œno 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 œdefault . 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.
Reference:
AWS Machine Learning Specialty Exam Guide
AWS Machine Learning Training - Classification vs Regression in Machine Learning

---

QUESTION 119
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.
Reference:
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

---

QUESTION 120
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. Reference:
Use Docker containers to build models - Amazon SageMaker
Create a container with your own algorithms and models - Amazon SageMaker

---

QUESTION 121
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. Reference:
Data Transformation - Amazon SageMaker
Transforming Skewed Data for Machine Learning

QUESTION 122
A Machine Learning Specialist is given a structured dataset on the shopping habits of a companys 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 kmeans 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. Reference:
Dimensionality Reduction - Amazon SageMaker
Visualize high dimensional data using t-SNE - Amazon SageMaker

---

QUESTION 123
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 clusters 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. Reference:
Amazon EMR on EC2 Spot Instances
Instance purchasing options - Amazon EMR

QUESTION 124
A company wants to predict the sale prices of houses based on available historical sales dat
a. The target
variable in the companys dataset is the sale price. The features include parameters such as the lot size, living
area measurements, non-living area measurements, number of bedrooms, number of bathrooms, year built,
and postal code. The company wants to use multi-variable linear regression to predict house sale prices.
Which step should a machine learning specialist take to remove features that are irrelevant for the analysis
and reduce the models complexity?

A. Plot a histogram of the features and compute their standard deviation. Remove features with high variance.
B. Plot a histogram of the features and compute their standard deviation. Remove features with low variance.
C. Build a heatmap showing the correlation of the dataset against itself. Remove features with low mutual correlation scores.
D. Run a correlation check of all features against the target variable. Remove features with low target variable correlation scores.

Answer: D

Explanation:
Feature selection is the process of reducing the number of input variables to those that are most

relevant for predicting the target variable. One way to do this is to run a correlation check of all features against the target variable and remove features with low target variable correlation scores. This means that these features have little or no linear relationship with the target variable and are not useful for the prediction. This can reduce the models complexity and improve its performance. Reference:
Feature engineering - Machine Learning Lens
Feature Selection For Machine Learning in Python

QUESTION 125
A health care company is planning to use neural networks to classify their X-ray images into normal and abnormal classes. The labeled data is divided into a training set of 1,000 images and a test set of 200 images. The initial training of a neural network model with 50 hidden layers yielded 99% accuracy on the training set, but only 55% accuracy on the test set.
What changes should the Specialist consider to solve this issue? (Choose three.)

A. Choose a higher number of layers
B. Choose a lower number of layers
C. Choose a smaller learning rate
D. Enable dropout
E. Include all the images from the test set in the training set
F. Enable early stopping

Answer: B, D, F

Explanation:
The problem described in the question is a case of overfitting, where the neural network model performs well on the training data but poorly on the test data. This means that the model has learned the noise and specific patterns of the training data, but cannot generalize to new and unseen data. To solve this issue, the Specialist should consider the following changes:
Choose a lower number of layers: Reducing the number of layers can reduce the complexity and capacity of the neural network model, making it less prone to overfitting. A model with 50 hidden layers is likely too deep for the given data size and task. A simpler model with fewer layers can learn the essential features of the data without memorizing the noise.
Enable dropout: Dropout is a regularization technique that randomly drops out some units in the neural network during training. This prevents the units from co-adapting too much and forces the model to learn more robust features. Dropout can improve the generalization and test performance of the model by reducing overfitting.
Enable early stopping: Early stopping is another regularization technique that monitors the validation error during training and stops the training process when the validation error stops decreasing or starts increasing. This prevents the model from overtraining on the training data and reduces overfitting.
Reference:

Deep Learning - Machine Learning Lens
How to Avoid Overfitting in Deep Learning Neural Networks
How to Identify Overfitting Machine Learning Models in Scikit-Learn

---

QUESTION 126
A Machine Learning Specialist is attempting to build a linear regression model.
Given the displayed residual plot only, what is the MOST likely problem with the model?

A. Linear regression is inappropriate. The residuals do not have constant variance.
B. Linear regression is inappropriate. The underlying data has outliers.
C. Linear regression is appropriate. The residuals have a zero mean.
D. Linear regression is appropriate. The residuals have constant variance.

Answer: A

Explanation:
A residual plot is a type of plot that displays the values of a predictor variable in a regression model along the x-axis and the values of the residuals along the y-axis. This plot is used to assess whether or not the residuals in a regression model are normally distributed and whether or not they exhibit heteroscedasticity. Heteroscedasticity means that the variance of the residuals is not constant across different values of the predictor variable. This violates one of the assumptions of linear regression and can lead to biased estimates and unreliable predictions. The displayed residual plot shows a clear pattern of heteroscedasticity, as the residuals spread out as the fitted values increase. This indicates that linear regression is inappropriate for this data and a different model should be used. Reference:
Regression - Amazon Machine Learning
How to Create a Residual Plot by Hand
How to Create a Residual Plot in Python

QUESTION 127
A machine learning specialist works for a fruit processing company and needs to build a system that categorizes apples into three types. The specialist has collected a dataset that contains 150 images for each type of apple and applied transfer learning on a neural network that was pretrained on ImageNet with this dataset.
The company requires at least 85% accuracy to make use of the model.
After an exhaustive grid search, the optimal hyperparameters produced the following:
68% accuracy on the training set
67% accuracy on the validation set
What can the machine learning specialist do to improve the systems accuracy?

A. Upload the model to an Amazon SageMaker notebook instance and use the Amazon SageMaker HPO feature to optimize the models hyperparameters.

B. Add more data to the training set and retrain the model using transfer learning to reduce the bias.
C. Use a neural network model with more layers that are pretrained on ImageNet and apply transfer learning to increase the variance.
D. Train a new model using the current neural network architecture.

Answer: B

Explanation:
The problem described in the question is a case of underfitting, where the neural network model performs poorly on both the training and validation sets. This means that the model has not learned the features of the data well enough and has high bias. To solve this issue, the machine learning specialist should consider the following change:
Add more data to the training set and retrain the model using transfer learning to reduce the bias:
Adding more data to the training set can help the model learn more patterns and variations in the data and improve its performance. Transfer learning can also help the model leverage the knowledge from the pre-trained network and adapt it to the new data. This can reduce the bias and increase the accuracy of the model.
Reference:
Transfer learning for TensorFlow image classification models in Amazon SageMaker
Transfer learning for custom labels using a TensorFlow container and œbring your own algorithm in Amazon SageMaker
Machine Learning Concepts - AWS Training and Certification

QUESTION 128
A company uses camera images of the tops of items displayed on store shelves to determine which items
were removed and which ones still remain. After several hours of data labeling, the company has a total of
1,000 hand-labeled images covering 10 distinct items. The training results were poor.
Which machine learning approach fulfills the companys long-term needs?

A. Convert the images to grayscale and retrain the model
B. Reduce the number of distinct items from 10 to 2, build the model, and iterate
C. Attach different colored labels to each item, take the images again, and build the model
D. Augment training data for each item using image variants like inversions and translations, build the model, and iterate.

Answer: D

Explanation:
Data augmentation is a technique that can increase the size and diversity of the training data by applying various transformations to the original images, such as inversions, translations, rotations,

scaling, cropping, flipping, and color variations. Data augmentation can help improve the performance and generalization of image classification models by reducing overfitting and introducing more variability to the data. Data augmentation is especially useful when the original data is limited or imbalanced, as in the case of the companys problem. By augmenting the training data for each item using image variants, the company can build a more robust and accurate model that can recognize the items on the store shelves from different angles, positions, and lighting conditions. The company can also iterate on the model by adding more data or fine-tuning the hyperparameters to achieve better results.
Reference:
Build high performing image classification models using Amazon SageMaker JumpStart
The Effectiveness of Data Augmentation in Image Classification using Deep Learning
Data augmentation for improving deep learning in image classification problem
Class-Adaptive Data Augmentation for Image Classification

QUESTION 129
A Data Scientist is developing a binary classifier to predict whether a patient has a particular disease on a series of test results. The Data Scientist has data on 400 patients randomly selected from the population. The disease is seen in 3% of the population.
Which cross-validation strategy should the Data Scientist adopt?

A. A k-fold cross-validation strategy with k=5
B. A stratified k-fold cross-validation strategy with k=5
C. A k-fold cross-validation strategy with k=5 and 3 repeats
D. An 80 stratified split between training and validation

Answer: B

Explanation:
A stratified k-fold cross-validation strategy is a technique that preserves the class distribution in each fold. This is important for imbalanced datasets, such as the one in the question, where the disease is seen in only 3% of the population. If a random k-fold cross-validation strategy is used, some folds may have no positive cases or very few, which would lead to poor estimates of the model performance. A stratified k-fold cross-validation strategy ensures that each fold has the same proportion of positive and negative cases as the whole dataset, which makes the evaluation more reliable and robust. A k-fold cross-validation strategy with k=5 and 3 repeats is also a possible option, but it is more computationally expensive and may not be necessary if the stratification is done properly. An 80 stratified split between training and validation is another option, but it uses less data for training and validation than k-fold cross-validation, which may result in higher variance and lower accuracy of the estimates. Reference:
AWS Machine Learning Specialty Certification Exam Guide
AWS Machine Learning Training: Model Evaluation
How to Fix k-Fold Cross-Validation for Imbalanced Classification

QUESTION 130
A technology startup is using complex deep neural networks and GPU compute to recommend the companys products to its existing customers based upon each customers habits and interactions. The solution currently pulls each dataset from an Amazon S3 bucket before loading the data into a TensorFlow model pulled from the companys Git repository that runs locally. This job then runs for several hours while continually outputting its progress to the same S3 bucket. The job can be paused, restarted, and continued at any time in the event of a failure, and is run from a central queue. Senior managers are concerned about the complexity of the solutions resource management and the costs involved in repeating the process regularly. They ask for the workload to be automated so it runs once a week, starting Monday and completing by the close of business Friday.
Which architecture should be used to scale the solution at the lowest cost?

A. Implement the solution using AWS Deep Learning Containers and run the container as a job using AWS Batch on a GPU-compatible Spot Instance
B. Implement the solution using a low-cost GPU-compatible Amazon EC2 instance and use the AWS Instance Scheduler to schedule the task
C. Implement the solution using AWS Deep Learning Containers, run the workload using AWS Fargate running on Spot Instances, and then schedule the task using the built-in task scheduler
D. Implement the solution using Amazon ECS running on Spot Instances and schedule the task using the ECS service scheduler

Answer: A

Explanation:
The best architecture to scale the solution at the lowest cost is to implement the solution using AWS Deep Learning Containers and run the container as a job using AWS Batch on a GPU-compatible Spot Instance. This option has the following advantages:
AWS Deep Learning Containers: These are Docker images that are pre-installed and optimized with popular deep learning frameworks such as TensorFlow, PyTorch, and MXNet. They can be easily deployed on Amazon EC2, Amazon ECS, Amazon EKS, and AWS Fargate. They can also be integrated with AWS Batch to run containerized batch jobs. Using AWS Deep Learning Containers can simplify the setup and configuration of the deep learning environment and reduce the complexity of the resource management.
AWS Batch: This is a fully managed service that enables you to run batch computing workloads on AWS. You can define compute environments, job queues, and job definitions to run your batch jobs. You can also use AWS Batch to automatically provision compute resources based on the requirements of the batch jobs. You can specify the type and quantity of the compute resources, such as GPU instances, and the maximum price you are willing to pay for them. You can also use AWS Batch to monitor the status and progress of your batch jobs and handle any failures or interruptions.
GPU-compatible Spot Instance: This is an Amazon EC2 instance that uses a spare compute capacity that is available at a lower price than the On-Demand price. You can use Spot Instances to run your

deep learning training jobs at a lower cost, as long as you are flexible about when your instances run and how long they run. You can also use Spot Instances with AWS Batch to automatically launch and terminate instances based on the availability and price of the Spot capacity. You can also use Spot Instances with Amazon EBS volumes to store your datasets, checkpoints, and logs, and attach them to your instances when they are launched. This way, you can preserve your data and resume your training even if your instances are interrupted.
Reference:
AWS Deep Learning Containers
AWS Batch
Amazon EC2 Spot Instances
Using Amazon EBS Volumes with Amazon EC2 Spot Instances

QUESTION 131
A media company with a very large archive of unlabeled images, text, audio, and video footage wishes to index its assets to allow rapid identification of relevant content by the Research team. The company wants to use machine learning to accelerate the efforts of its in-house researchers who have limited machine learning expertise.
Which is the FASTEST route to index the assets?

A. Use Amazon Rekognition, Amazon Comprehend, and Amazon Transcribe to tag data into distinct categories/classes.
B. Create a set of Amazon Mechanical Turk Human Intelligence Tasks to label all footage.
C. Use Amazon Transcribe to convert speech to text. Use the Amazon SageMaker Neural Topic Model (NTM) and Object Detection algorithms to tag data into distinct categories/classes.
D. Use the AWS Deep Learning AMI and Amazon EC2 GPU instances to create custom models for audio transcription and topic modeling, and use object detection to tag data into distinct categories/classes.

Answer: A

Explanation:
Amazon Rekognition, Amazon Comprehend, and Amazon Transcribe are AWS machine learning services that can analyze and extract metadata from images, text, audio, and video content. These services are easy to use, scalable, and do not require any machine learning expertise. They can help the media company to quickly index its assets and enable rapid identification of relevant content by the research team. Using these services is the fastest route to index the assets, compared to the other options that involve human intervention, custom model development, or additional steps. Reference:
AWS Media Intelligence Solutions
AWS Machine Learning Services
The Best Services For Running Machine Learning Models On AWS

QUESTION 132
A Machine Learning Specialist is working for an online retailer that wants to run analytics on every customer visit, processed through a machine learning pipeline. The data needs to be ingested by Amazon Kinesis Data Streams at up to 100 transactions per second, and the JSON data blob is 100 KB in size.
What is the MINIMUM number of shards in Kinesis Data Streams the Specialist should use to successfully ingest this data?

A. 1 shards
B. 10 shards
C. 100 shards
D. 1,000 shards

Answer: A

Explanation:
According to the Amazon Kinesis Data Streams documentation, the maximum size of data blob (the data payload before Base64-encoding) per record is 1 MB. The maximum number of records that can be sent to a shard per second is 1,000. Therefore, the maximum throughput of a shard is 1 MB/sec for input and 2 MB/sec for output. In this case, the input throughput is 100 transactions per second * 100 KB per transaction = 10 MB/sec. Therefore, the minimum number of shards required is 10 MB/sec / 1 MB/sec = 10 shards. However, the question asks for the minimum number of shards in Kinesis Data Streams, not the minimum number of shards per stream. A Kinesis Data Streams account can have multiple streams, each with its own number of shards. Therefore, the minimum number of shards in Kinesis Data Streams is 1, which is the minimum number of shards per stream. Reference:
Amazon Kinesis Data Streams Terminology and Concepts
Amazon Kinesis Data Streams Limits

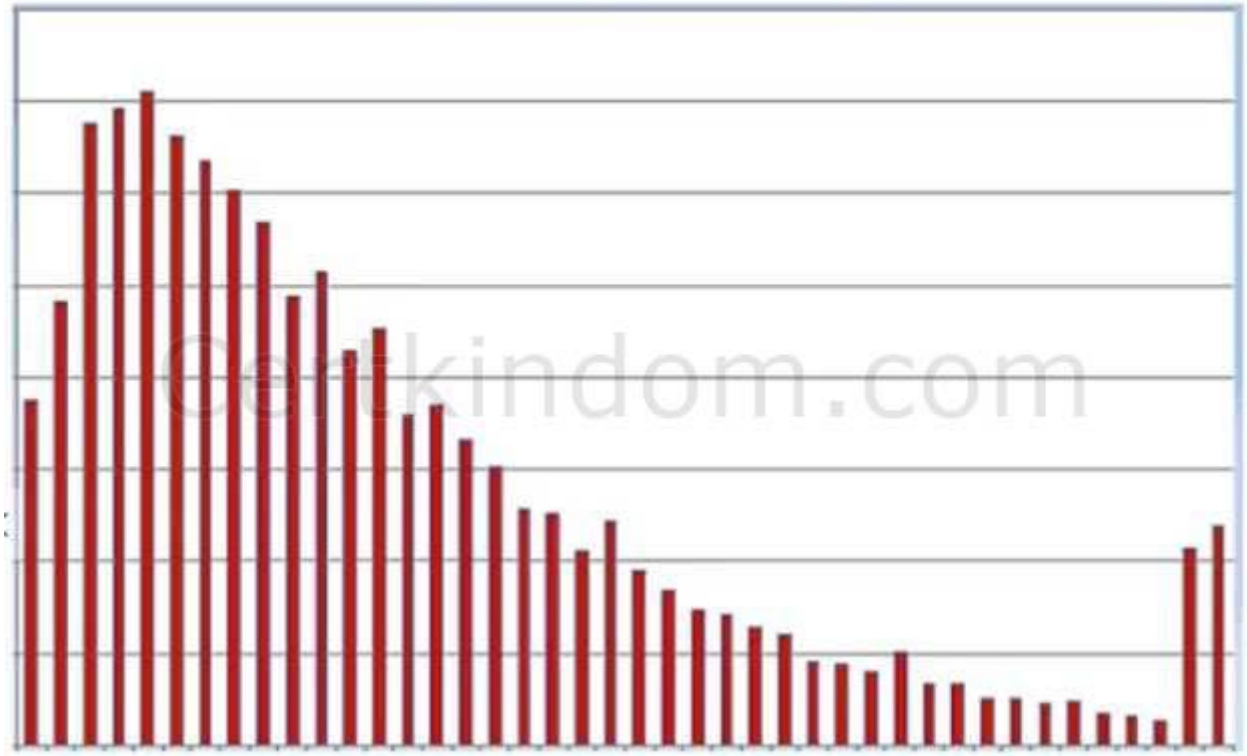
QUESTION 133
A Machine Learning Specialist is deciding between building a naive Bayesian model or a full Bayesian network for a classification problem. The Specialist computes the Pearson correlation coefficients between each feature and finds that their absolute values range between 0.1 to 0.95.
Which model describes the underlying data in this situation?

A. A naive Bayesian model, since the features are all conditionally independent.
B. A full Bayesian network, since the features are all conditionally independent.
C. A naive Bayesian model, since some of the features are statistically dependent.
D. A full Bayesian network, since some of the features are statistically dependent.

Answer: D

Explanation:
A naive Bayesian model assumes that the features are conditionally independent given the class label. This means that the joint probability of the features and the class can be factorized as the product of the class prior and the feature likelihoods. A full Bayesian network, on the other hand, does not make this assumption and allows for modeling arbitrary dependencies between the features and the class using a directed acyclic graph. In this case, the joint probability of the features and the class is given by the product of the conditional probabilities of each node given its parents in the graph. If the features are statistically dependent, meaning that their correlation coefficients are not close to zero, then a naive Bayesian model would not capture these dependencies and would likely perform worse than a full Bayesian network that can account for them. Therefore, a full Bayesian network describes the underlying data better in this situation. Reference:
Naive Bayes and Text Classification I
Bayesian Networks

QUESTION 134
A Data Scientist is building a linear regression model and will use resulting p-values to evaluate the statistical significance of each coefficient. Upon inspection of the dataset, the Data Scientist discovers that most of the features are normally distributed. The plot of one feature in the dataset is shown in the graphic.



What transformation should the Data Scientist apply to satisfy the statistical assumptions of the

linear
regression model?

A. Exponential transformation
B. Logarithmic transformation
C. Polynomial transformation
D. Sinusoidal transformation

Answer: B

Explanation:
The plot in the graphic shows a right-skewed distribution, which violates the assumption of normality for linear regression. To correct this, the Data Scientist should apply a logarithmic transformation to the feature. This will help to make the distribution more symmetric and closer to a normal distribution, which is a key assumption for linear regression. Reference:
Linear Regression
Linear Regression with Amazon Machine Learning
Machine Learning on AWS

QUESTION 135
A Machine Learning Specialist is assigned to a Fraud Detection team and must tune an XGBoost model, which is working appropriately for test dat
a. However, with unknown data, it is not working as expected. The existing parameters are provided as follows.

```
param = {
      'eta': 0.05,  # the training step for each iteration
      'silent': 1,  # logging mode - quiet
      'n_estimators':2000,
      'max_depth':30,
      'min_child_weight': 3,
      'gamma': 0,
      'subsample': 0.8,
      'objective': 'multi:softprob',  # error evaluation for multiclass training
      'num_class': 201}  # the number of classes that exist in this dataset
num_round = 60  # the number of training iterations
```


Which parameter tuning guidelines should the Specialist follow to avoid overfitting?

A. Increase the max_depth parameter value.

B. Lower the max_depth parameter value.
C. Update the objective to binary:logistic.
D. Lower the min_child_weight parameter value.

Answer: B

Explanation:
Overfitting occurs when a model performs well on the training data but poorly on the test data. This is often because the model has learned the training data too well and is not able to generalize to new data. To avoid overfitting, the Machine Learning Specialist should lower the max_depth parameter value. This will reduce the complexity of the model and make it less likely to overfit. According to the XGBoost documentation1, the max_depth parameter controls the maximum depth of a tree and lower values can help prevent overfitting. The documentation also suggests other ways to control overfitting, such as adding randomness, using regularization, and using early stopping1. Reference: XGBoost Parameters

---

QUESTION 136
A data scientist is developing a pipeline to ingest streaming web traffic dat
a. The data scientist needs to implement a process to identify unusual web traffic patterns as part of the pipeline. The patterns will be used downstream for alerting and incident response. The data scientist has access to unlabeled historic data to use, if needed.
The solution needs to do the following:
Calculate an anomaly score for each web traffic entry.
Adapt unusual event identification to changing web patterns over time.
Which approach should the data scientist implement to meet these requirements?

A. Use historic web traffic data to train an anomaly detection model using the Amazon SageMaker Random Cut Forest (RCF) built-in model. Use an Amazon Kinesis Data Stream to process the incoming web traffic data. Attach a preprocessing AWS Lambda function to perform data enrichment by calling the RCF model to calculate the anomaly score for each record.
B. Use historic web traffic data to train an anomaly detection model using the Amazon SageMaker built-in XGBoost model. Use an Amazon Kinesis Data Stream to process the incoming web traffic data. Attach a preprocessing AWS Lambda function to perform data enrichment by calling the XGBoost model to calculate the anomaly score for each record.
C. Collect the streaming data using Amazon Kinesis Data Firehose. Map the delivery stream as an input source for Amazon Kinesis Data Analytics. Write a SQL query to run in real time against the streaming data with the k-Nearest Neighbors (kNN) SQL extension to calculate anomaly scores for each record using a tumbling window.
D. Collect the streaming data using Amazon Kinesis Data Firehose. Map the delivery stream as an input source for Amazon Kinesis Data Analytics. Write a SQL query to run in real time against the streaming data with the Amazon Random Cut Forest (RCF) SQL extension to calculate anomaly scores

for each record using a sliding window.

Answer: D

Explanation:
Amazon Kinesis Data Analytics is a service that allows users to analyze streaming data in real time using SQL queries. Amazon Random Cut Forest (RCF) is a SQL extension that enables anomaly detection on streaming data. RCF is an unsupervised machine learning algorithm that assigns an anomaly score to each data point based on how different it is from the rest of the data. A sliding window is a type of window that moves along with the data stream, so that the anomaly detection model can adapt to changing patterns over time. A tumbling window is a type of window that has a fixed size and does not overlap with other windows, so that the anomaly detection model is based on a fixed period of time. Therefore, option D is the best approach to meet the requirements of the question, as it uses RCF to calculate anomaly scores for each web traffic entry and uses a sliding window to adapt to changing web patterns over time.
Option A is incorrect because Amazon SageMaker Random Cut Forest (RCF) is a built-in model that can be used to train and deploy anomaly detection models on batch or streaming data, but it requires more steps and resources than using the RCF SQL extension in Amazon Kinesis Data Analytics. Option B is incorrect because Amazon SageMaker XGBoost is a built-in model that can be used for supervised learning tasks such as classification and regression, but not for unsupervised learning tasks such as anomaly detection. Option C is incorrect because k-Nearest Neighbors (kNN) is a SQL extension that can be used for classification and regression tasks on streaming data, but not for anomaly detection. Moreover, using a tumbling window would not allow the anomaly detection model to adapt to changing web patterns over time.
Reference:
Using CloudWatch anomaly detection
Anomaly Detection With CloudWatch
Performing Real-time Anomaly Detection using AWS
What Is AWS Anomaly Detection? (And Is There A Better Option?)

QUESTION 137
A Data Scientist received a set of insurance records, each consisting of a record ID, the final outcome among 200 categories, and the date of the final outcome. Some partial information on claim contents is also provided, but only for a few of the 200 categories. For each outcome category, there are hundreds of records distributed over the past 3 years. The Data Scientist wants to predict how many claims to expect in each category from month to month, a few months in advance.
What type of machine learning model should be used?

A. Classification month-to-month using supervised learning of the 200 categories based on claim contents.
B. Reinforcement learning using claim IDs and timestamps where the agent will identify how many claims in each category to expect from month to month.

C. Forecasting using claim IDs and timestamps to identify how many claims in each category to expect from month to month.
D. Classification with supervised learning of the categories for which partial information on claim contents is provided, and forecasting using claim IDs and timestamps for all other categories.

Answer: C

Explanation:
: Forecasting is a type of machine learning model that predicts future values of a target variable based on historical data and other features. Forecasting is suitable for problems that involve timeseries data, such as the number of claims in each category from month to month. Forecasting can handle multiple categories of the target variable, as well as missing or partial information on some features. Therefore, option C is the best choice for the given problem.
Option A is incorrect because classification is a type of machine learning model that assigns a label to an input based on predefined categories. Classification is not suitable for predicting continuous or numerical values, such as the number of claims in each category from month to month. Moreover, classification requires sufficient and complete information on the features that are relevant to the target variable, which is not the case for the given problem. Option B is incorrect because reinforcement learning is a type of machine learning model that learns from its own actions and rewards in an interactive environment. Reinforcement learning is not suitable for problems that involve historical data and do not require an agent to take actions. Option D is incorrect because it combines two different types of machine learning models, which is unnecessary and inefficient. Moreover, classification is not suitable for predicting the number of claims in some categories, as explained in option A.
Reference:
Forecasting | AWS Solutions for Machine Learning (AI/ML) | AWS Solutions Library
Time Series Forecasting Service " Amazon Forecast " Amazon Web Services
Amazon Forecast: Guide to Predicting Future Outcomes - Onica
Amazon Launches What-If Analyses for Machine Learning Forecasting ¦

QUESTION 138
A company that promotes healthy sleep patterns by providing cloud-connected devices currently hosts a sleep tracking application on AWS. The application collects device usage information from device users. The company's Data Science team is building a machine learning model to predict if and when a user will stop utilizing the company's devices. Predictions from this model are used by a downstream application that determines the best approach for contacting users.
The Data Science team is building multiple versions of the machine learning model to evaluate each version against the companys business goals. To measure long-term effectiveness, the team wants to run multiple versions of the model in parallel for long periods of time, with the ability to control the portion of inferences served by the models.
Which solution satisfies these requirements with MINIMAL effort?

A. Build and host multiple models in Amazon SageMaker. Create multiple Amazon SageMaker endpoints, one for each model. Programmatically control invoking different models for inference at the application layer.
B. Build and host multiple models in Amazon SageMaker. Create an Amazon SageMaker endpoint configuration with multiple production variants. Programmatically control the portion of the inferences served by the multiple models by updating the endpoint configuration.
C. Build and host multiple models in Amazon SageMaker Neo to take into account different types of medical devices. Programmatically control which model is invoked for inference based on the medical device type.
D. Build and host multiple models in Amazon SageMaker. Create a single endpoint that accesses multiple models. Use Amazon SageMaker batch transform to control invoking the different models through the single endpoint.

Answer: B

Explanation:
Amazon SageMaker is a service that allows users to build, train, and deploy ML models on AWS. Amazon SageMaker endpoints are scalable and secure web services that can be used to perform real-time inference on ML models. An endpoint configuration defines the models that are deployed and the resources that are used by the endpoint. An endpoint configuration can have multiple production variants, each representing a different version or variant of a model. Users can specify the portion of the inferences served by each production variant using the initialVariantWeight parameter. Users can also programmatically update the endpoint configuration to change the portion of the inferences served by each production variant using the UpdateEndpointWeightsAndCapacities API. Therefore, option B is the best solution to satisfy the requirements with minimal effort.
Option A is incorrect because creating multiple endpoints for each model would incur more cost and complexity than using a single endpoint with multiple production variants. Moreover, controlling the invocation of different models at the application layer would require more custom logic and coordination than using the UpdateEndpointWeightsAndCapacities API. Option C is incorrect because Amazon SageMaker Neo is a service that allows users to optimize ML models for different hardware platforms, such as edge devices. It is not relevant to the problem of running multiple versions of a model in parallel for long periods of time. Option D is incorrect because Amazon SageMaker batch transform is a service that allows users to perform asynchronous inference on large datasets. It is not suitable for the problem of performing real-time inference on streaming data from device users.
Reference:
Deploying models to Amazon SageMaker hosting services - Amazon SageMaker
Update an Amazon SageMaker endpoint to accommodate new models - Amazon SageMaker
UpdateEndpointWeightsAndCapacities - Amazon SageMaker

QUESTION 139
An agricultural company is interested in using machine learning to detect specific types of weeds in a

100-acre grassland field. Currently, the company uses tractor-mounted cameras to capture multiple images of the field as 10 Ã— 10 grids. The company also has a large training dataset that consists of annotated images of popular weed classes like broadleaf and non-broadleaf docks.
The company wants to build a weed detection model that will detect specific types of weeds and the location of each type within the field. Once the model is ready, it will be hosted on Amazon SageMaker endpoints. The model will perform real-time inferencing using the images captured by the cameras.
Which approach should a Machine Learning Specialist take to obtain accurate predictions?

A. Prepare the images in RecordIO format and upload them to Amazon S3. Use Amazon SageMaker to train, test, and validate the model using an image classification algorithm to categorize images into various weed classes.
B. Prepare the images in Apache Parquet format and upload them to Amazon S3. Use Amazon SageMaker to train, test, and validate the model using an object-detection single-shot multibox detector (SSD) algorithm.
C. Prepare the images in RecordIO format and upload them to Amazon S3. Use Amazon SageMaker to train, test, and validate the model using an object-detection single-shot multibox detector (SSD) algorithm.
D. Prepare the images in Apache Parquet format and upload them to Amazon S3. Use Amazon SageMaker to train, test, and validate the model using an image classification algorithm to categorize images into various weed classes.

Answer: C

Explanation:
The problem of detecting specific types of weeds and their location within the field is an example of object detection, which is a type of machine learning model that identifies and localizes objects in an image. Amazon SageMaker provides a built-in object detection algorithm that uses a single-shot multibox detector (SSD) to perform real-time inference on streaming images. The SSD algorithm can handle multiple objects of varying sizes and scales in an image, and generate bounding boxes and scores for each object category. Therefore, option C is the best approach to obtain accurate predictions.
Option A is incorrect because image classification is a type of machine learning model that assigns a label to an image based on predefined categories. Image classification is not suitable for localizing objects within an image, as it does not provide bounding boxes or scores for each object. Option B is incorrect because Apache Parquet is a columnar storage format that is optimized for analytical queries. Apache Parquet is not suitable for storing images, as it does not preserve the spatial information of the pixels. Option D is incorrect because it combines the wrong format (Apache Parquet) and the wrong algorithm (image classification) for the given problem, as explained in options A and B.
Reference:
Object Detection algorithm now available in Amazon SageMaker

Image classification and object detection using Amazon Rekognition Custom Labels and Amazon SageMaker JumpStart
Object Detection with Amazon SageMaker - W3Schools
aws-samples/amazon-sagemaker-tensorflow-object-detection-api

QUESTION 140
A manufacturer is operating a large number of factories with a complex supply chain relationship where unexpected downtime of a machine can cause production to stop at several factories. A data scientist wants to analyze sensor data from the factories to identify equipment in need of preemptive maintenance and then dispatch a service team to prevent unplanned downtime. The sensor readings from a single machine can include up to 200 data points including temperatures, voltages, vibrations, RPMs, and pressure readings.
To collect this sensor data, the manufacturer deployed Wi-Fi and LANs across the factories. Even though many factory locations do not have reliable or high-speed internet connectivity, the manufacturer would like to maintain near-real-time inference capabilities.
Which deployment architecture for the model will address these business requirements?

A. Deploy the model in Amazon SageMaker. Run sensor data through this model to predict which machines need maintenance.
B. Deploy the model on AWS IoT Greengrass in each factory. Run sensor data through this model to infer which machines need maintenance.
C. Deploy the model to an Amazon SageMaker batch transformation job. Generate inferences in a daily batch report to identify machines that need maintenance.
D. Deploy the model in Amazon SageMaker and use an IoT rule to write data to an Amazon DynamoDB table. Consume a DynamoDB stream from the table with an AWS Lambda function to invoke the endpoint.

Answer: B

Explanation:
AWS IoT Greengrass is a service that extends AWS to edge devices, such as sensors and machines, so they can act locally on the data they generate, while still using the cloud for management, analytics, and durable storage. AWS IoT Greengrass enables local device messaging, secure data transfer, and local computing using AWS Lambda functions and machine learning models. AWS IoT Greengrass can run machine learning inference locally on devices using models that are created and trained in the cloud. This allows devices to respond quickly to local events, even when they are offline or have intermittent connectivity. Therefore, option B is the best deployment architecture for the model to address the business requirements of the manufacturer.
Option A is incorrect because deploying the model in Amazon SageMaker would require sending the sensor data to the cloud for inference, which would not work well for factory locations that do not have reliable or high-speed internet connectivity. Moreover, this option would not provide near-realtime inference capabilities, as there would be latency and bandwidth issues involved in transferring

the data to and from the cloud. Option C is incorrect because deploying the model to an Amazon SageMaker batch transformation job would not provide near-real-time inference capabilities, as batch transformation is an asynchronous process that operates on large datasets. Batch transformation is not suitable for streaming data that requires low-latency responses. Option D is incorrect because deploying the model in Amazon SageMaker and using an IoT rule to write data to an Amazon DynamoDB table would also require sending the sensor data to the cloud for inference, which would have the same drawbacks as option
A. Moreover, this option would introduce
additional complexity and cost by involving multiple services, such as IoT Core, DynamoDB, and Lambda.
Reference:
AWS Greengrass Machine Learning Inference - Amazon Web Services
Machine learning components - AWS IoT Greengrass
What is AWS Greengrass? | AWS IoT Core | Onica
GitHub - aws-samples/aws-greengrass-ml-deployment-sample
AWS IoT Greengrass Architecture and Its Benefits | Quick Guide - XenonStack

---

QUESTION 141
A Machine Learning Specialist is designing a scalable data storage solution for Amazon SageMaker. There is an existing TensorFlow-based model implemented as a train.py script that relies on static training data that is currently stored as TFRecords.
Which method of providing training data to Amazon SageMaker would meet the business requirements with the LEAST development overhead?

A. Use Amazon SageMaker script mode and use train.py unchanged. Point the Amazon SageMaker training invocation to the local path of the data without reformatting the training data.
B. Use Amazon SageMaker script mode and use train.py unchanged. Put the TFRecord data into an Amazon S3 bucket. Point the Amazon SageMaker training invocation to the S3 bucket without reformatting the training data.
C. Rewrite the train.py script to add a section that converts TFRecords to protobuf and ingests the protobuf data instead of TFRecords.
D. Prepare the data in the format accepted by Amazon SageMaker. Use AWS Glue or AWS Lambda to reformat and store the data in an Amazon S3 bucket.

Answer: B

Explanation:
Amazon SageMaker script mode is a feature that allows users to use training scripts similar to those they would use outside SageMaker with SageMakers prebuilt containers for various frameworks such as TensorFlow. Script mode supports reading data from Amazon S3 buckets without requiring any changes to the training script. Therefore, option B is the best method of providing training data to Amazon SageMaker that would meet the business requirements with the least development

overhead.
Option A is incorrect because using a local path of the data would not be scalable or reliable, as it would depend on the availability and capacity of the local storage. Moreover, using a local path of the data would not leverage the benefits of Amazon S3, such as durability, security, and performance. Option C is incorrect because rewriting the train.py script to convert TFRecords to protobuf would require additional development effort and complexity, as well as introduce potential errors and inconsistencies in the data format. Option D is incorrect because preparing the data in the format accepted by Amazon SageMaker would also require additional development effort and complexity, as well as involve using additional services such as AWS Glue or AWS Lambda, which would increase the cost and maintenance of the solution.
Reference:
Bring your own model with Amazon SageMaker script mode
GitHub - aws-samples/amazon-sagemaker-script-mode
Deep Dive on TensorFlow training with Amazon SageMaker and Amazon S3
amazon-sagemaker-script-mode/generate_cifar10_tfrecords.py at master

---

QUESTION 142
The chief editor for a product catalog wants the research and development team to build a machine learning system that can be used to detect whether or not individuals in a collection of images are wearing the company's retail brand. The team has a set of training data.
Which machine learning algorithm should the researchers use that BEST meets their requirements?

A. Latent Dirichlet Allocation (LDA)
B. Recurrent neural network (RNN)
C. K-means
D. Convolutional neural network (CNN)

Answer: D

Explanation:
The problem of detecting whether or not individuals in a collection of images are wearing the companys retail brand is an example of image recognition, which is a type of machine learning task that identifies and classifies objects in an image. Convolutional neural networks (CNNs) are a type of machine learning algorithm that are well-suited for image recognition, as they can learn to extract features from images and handle variations in size, shape, color, and orientation of the objects. CNNs consist of multiple layers that perform convolution, pooling, and activation operations on the input images, resulting in a high-level representation that can be used for classification or detection. Therefore, option D is the best choice for the machine learning algorithm that meets the requirements of the chief editor.
Option A is incorrect because latent Dirichlet allocation (LDA) is a type of machine learning algorithm that is used for topic modeling, which is a task that discovers the hidden themes or topics in a collection of text documents. LDA is not suitable for image recognition, as it does not preserve the

spatial information of the pixels. Option B is incorrect because recurrent neural networks (RNNs) are a type of machine learning algorithm that are used for sequential data, such as text, speech, or time series. RNNs can learn from the temporal dependencies and patterns in the input data, and generate outputs that depend on the previous states. RNNs are not suitable for image recognition, as they do not capture the spatial dependencies and patterns in the input images. Option C is incorrect because k-means is a type of machine learning algorithm that is used for clustering, which is a task that groups similar data points together based on their features. K-means is not suitable for image recognition, as it does not perform classification or detection of the objects in the images.
Reference:
Image Recognition Software - ML Image & Video Analysis - Amazon ¦
Image classification and object detection using Amazon Rekognition ¦
AWS Amazon Rekognition - Deep Learning Face and Image Recognition ¦
GitHub - awslabs/aws-ai-solution-kit: Machine Learning APIs for common ¦
Meet iNaturalist, an AWS-powered nature app that helps you identify ¦

QUESTION 143
A retail company is using Amazon Personalize to provide personalized product recommendations for its customers during a marketing campaign. The company sees a significant increase in sales of recommended items to existing customers immediately after deploying a new solution version, but these sales decrease a short time after deployment. Only historical data from before the marketing campaign is available for training.
How should a data scientist adjust the solution?

A. Use the event tracker in Amazon Personalize to include real-time user interactions.
B. Add user metadata and use the HRNN-Metadata recipe in Amazon Personalize.
C. Implement a new solution using the built-in factorization machines (FM) algorithm in Amazon SageMaker.
D. Add event type and event value fields to the interactions dataset in Amazon Personalize.

Answer: A

Explanation:
The best option is to use the event tracker in Amazon Personalize to include real-time user interactions. This will allow the model to learn from the feedback of the customers during the marketing campaign and adjust the recommendations accordingly. The event tracker can capture click-through, add-to-cart, purchase, and other types of events that indicate the users preferences. By using the event tracker, the company can improve the relevance and freshness of the recommendations and avoid the decrease in sales.
The other options are not as effective as using the event tracker. Adding user metadata and using the HRNN-Metadata recipe in Amazon Personalize can help capture the users attributes and preferences, but it will not reflect the changes in user behavior during the marketing campaign. Implementing a new solution using the built-in factorization machines (FM) algorithm in Amazon

SageMaker can also provide personalized recommendations, but it will require more time and effort to train and deploy the model. Adding event type and event value fields to the interactions dataset in Amazon Personalize can help capture the importance and context of each interaction, but it will not update the model with the latest user feedback.
Reference:
Recording events - Amazon Personalize
Using real-time events - Amazon Personalize

QUESTION 144
A machine learning (ML) specialist wants to secure calls to the Amazon SageMaker Service API. The specialist has configured Amazon VPC with a VPC interface endpoint for the Amazon SageMaker Service API and is attempting to secure traffic from specific sets of instances and IAM users. The VPC is configured with a single public subnet.
Which combination of steps should the ML specialist take to secure the traffic? (Choose two.)

A. Add a VPC endpoint policy to allow access to the IAM users.
B. Modify the users' IAM policy to allow access to Amazon SageMaker Service API calls only.
C. Modify the security group on the endpoint network interface to restrict access to the instances.
D. Modify the ACL on the endpoint network interface to restrict access to the instances.
E. Add a SageMaker Runtime VPC endpoint interface to the VPC.

Answer: C, E

Explanation:
To secure calls to the Amazon SageMaker Service API, the ML specialist should take the following steps:
Modify the security group on the endpoint network interface to restrict access to the instances. This will allow the ML specialist to control which instances in the VPC can communicate with the VPC interface endpoint for the Amazon SageMaker Service API. The security group can specify inbound and outbound rules based on the instance IDs, IP addresses, or CIDR blocks1.
Add a SageMaker Runtime VPC endpoint interface to the VPC. This will allow the ML specialist to invoke the SageMaker endpoints from within the VPC without using the public internet. The SageMaker Runtime VPC endpoint interface connects the VPC directly to the SageMaker Runtime using AWS PrivateLink2.
The other options are not as effective or necessary as the steps above. Adding a VPC endpoint policy to allow access to the IAM users is not required, as the IAM users can already access the Amazon SageMaker Service API through the VPC interface endpoint. Modifying the users IAM policy to allow access to Amazon SageMaker Service API calls only is not sufficient, as it does not prevent unauthorized instances from accessing the VPC interface endpoint. Modifying the ACL on the endpoint network interface to restrict access to the instances is not possible, as network ACLs are associated with subnets, not network interfaces3.

Reference:
Security groups for your VPC - Amazon Virtual Private Cloud
Connect to SageMaker Within your VPC - Amazon SageMaker
Network ACLs - Amazon Virtual Private Cloud

QUESTION 145
An e commerce company wants to launch a new cloud-based product recommendation feature for its web application. Due to data localization regulations, any sensitive data must not leave its onpremises data center, and the product recommendation model must be trained and tested using nonsensitive data only. Data transfer to the cloud must use IPsec. The web application is hosted on premises with a PostgreSQL database that contains all the dat
a. The company wants the data to be uploaded securely to Amazon S3 each day for model retraining. How should a machine learning specialist meet these requirements?

A. Create an AWS Glue job to connect to the PostgreSQL DB instance. Ingest tables without sensitive data through an AWS Site-to-Site VPN connection directly into Amazon S3.
B. Create an AWS Glue job to connect to the PostgreSQL DB instance. Ingest all data through an AWS Site- to-Site VPN connection into Amazon S3 while removing sensitive data using a PySpark job.
C. Use AWS Database Migration Service (AWS DMS) with table mapping to select PostgreSQL tables with no sensitive data through an SSL connection. Replicate data directly into Amazon S3.
D. Use PostgreSQL logical replication to replicate all data to PostgreSQL in Amazon EC2 through AWS Direct Connect with a VPN connection. Use AWS Glue to move data from Amazon EC2 to Amazon S3.

Answer: C

Explanation:
The best option is to use AWS Database Migration Service (AWS DMS) with table mapping to select PostgreSQL tables with no sensitive data through an SSL connection. Replicate data directly into Amazon S3. This option meets the following requirements:
It ensures that only nonsensitive data is transferred to the cloud by using table mapping to filter out the tables that contain sensitive data1.
It uses IPsec to secure the data transfer by enabling SSL encryption for the AWS DMS endpoint2.
It uploads the data to Amazon S3 each day for model retraining by using the ongoing replication feature of AWS DMS3.
The other options are not as effective or feasible as the option above. Creating an AWS Glue job to connect to the PostgreSQL DB instance and ingest data through an AWS Site-to-Site VPN connection directly into Amazon S3 is possible, but it requires more steps and resources than using AWS DMS. Also, it does not specify how to filter out the sensitive data from the tables. Creating an AWS Glue job to connect to the PostgreSQL DB instance and ingest all data through an AWS Site-to-Site VPN connection into Amazon S3 while removing sensitive data using a PySpark job is also possible, but it is more complex and error-prone than using AWS DMS. Also, it does not use IPsec as required. Using

PostgreSQL logical replication to replicate all data to PostgreSQL in Amazon EC2 through AWS Direct Connect with a VPN connection, and then using AWS Glue to move data from Amazon EC2 to Amazon S3 is not feasible, because PostgreSQL logical replication does not support replicating only a subset of data4. Also, it involves unnecessary data movement and additional costs.
Reference:
Table mapping - AWS Database Migration Service
Using SSL to encrypt a connection to a DB instance - AWS Database Migration Service
Ongoing replication - AWS Database Migration Service
Logical replication - PostgreSQL

QUESTION 146
A logistics company needs a forecast model to predict next month's inventory requirements for a single item in 10 warehouses. A machine learning specialist uses Amazon Forecast to develop a forecast model from 3 years of monthly dat
a. There is no missing data. The specialist selects the DeepAR+ algorithm to train a predictor. The predictor means absolute percentage error (MAPE) is much larger than the MAPE produced by the current human forecasters.
Which changes to the CreatePredictor API call could improve the MAPE? (Choose two.)

A. Set PerformAutoML to true.
B. Set ForecastHorizon to 4.
C. Set ForecastFrequency to W for weekly.
D. Set PerformHPO to true.
E. Set FeaturizationMethodName to filling.

Answer: A, D

Explanation:
The MAPE of the predictor could be improved by making the following changes to the CreatePredictor API call:
Set PerformAutoML to true. This will allow Amazon Forecast to automatically evaluate different algorithms and choose the one that minimizes the objective function, which is the mean of the weighted losses over the forecast types. By default, these are the p10, p50, and p90 quantile losses1. This option can help find a better algorithm than DeepAR+ for the given data.
Set PerformHPO to true. This will enable hyperparameter optimization (HPO), which is the process of finding the optimal values for the algorithm-specific parameters that affect the quality of the forecasts. HPO can improve the accuracy of the predictor by tuning the hyperparameters based on the training data2.
The other options are not likely to improve the MAPE of the predictor. Setting ForecastHorizon to 4 will reduce the number of time steps that the model predicts, which may not match the business requirement of predicting next months inventory. Setting ForecastFrequency to W for weekly will change the granularity of the forecasts, which may not be appropriate for the monthly data. Setting

FeaturizationMethodName to filling will not have any effect, since there is no missing data in the dataset.
Reference:
CreatePredictor - Amazon Forecast
HPOConfig - Amazon Forecast

QUESTION 147
A data scientist wants to use Amazon Forecast to build a forecasting model for inventory demand for a retail company. The company has provided a dataset of historic inventory demand for its products as a .csv file stored in an Amazon S3 bucket. The table below shows a sample of the dataset.

timestamp	item_id	demand	category	lead_time
2019-12-14	uni_000736	120	hardware	90
2020-01-31	uni_003429	98	hardware	30
2020-03-04	uni_000211	234	accessories	10

How should the data scientist transform the data?

A. Use ETL jobs in AWS Glue to separate the dataset into a target time series dataset and an item metadata dataset. Upload both datasets as .csv files to Amazon S3.
B. Use a Jupyter notebook in Amazon SageMaker to separate the dataset into a related time series dataset and an item metadata dataset. Upload both datasets as tables in Amazon Aurora.
C. Use AWS Batch jobs to separate the dataset into a target time series dataset, a related time series dataset, and an item metadata dataset. Upload them directly to Forecast from a local machine.
D. Use a Jupyter notebook in Amazon SageMaker to transform the data into the optimized protobuf recordIO format. Upload the dataset in this format to Amazon S3.

Answer: A

Explanation:
Amazon Forecast requires the input data to be in a specific format. The data scientist should use ETL jobs in AWS Glue to separate the dataset into a target time series dataset and an item metadata dataset. The target time series dataset should contain the timestamp, item_id, and demand columns, while the item metadata dataset should contain the item_id, category, and lead_time columns. Both datasets should be uploaded as .csv files to Amazon S3 . Reference:
How Amazon Forecast Works - Amazon Forecast
Choosing Datasets - Amazon Forecast

QUESTION 148
A machine learning specialist is running an Amazon SageMaker endpoint using the built-in object

detection algorithm on a P3 instance for real-time predictions in a company's production application. When evaluating the model's resource utilization, the specialist notices that the model is using only a fraction of the GPU.
Which architecture changes would ensure that provisioned resources are being utilized effectively?

A. Redeploy the model as a batch transform job on an M5 instance.
B. Redeploy the model on an M5 instance. Attach Amazon Elastic Inference to the instance.
C. Redeploy the model on a P3dn instance.
D. Deploy the model onto an Amazon Elastic Container Service (Amazon ECS) cluster using a P3 instance.

Answer: B

Explanation:
The best way to ensure that provisioned resources are being utilized effectively is to redeploy the model on an M5 instance and attach Amazon Elastic Inference to the instance. Amazon Elastic Inference allows you to attach low-cost GPU-powered acceleration to Amazon EC2 and Amazon SageMaker instances to reduce the cost of running deep learning inference by up to 75%. By using Amazon Elastic Inference, you can choose the instance type that is best suited to the overall CPU and memory needs of your application, and then separately configure the amount of inference acceleration that you need with no code changes. This way, you can avoid wasting GPU resources and pay only for what you use.
Option A is incorrect because a batch transform job is not suitable for real-time predictions. Batch transform is a high-performance and cost-effective feature for generating inferences using your trained models. Batch transform manages all of the compute resources required to get inferences. Batch transform is ideal for scenarios where youre working with large batches of data, dont need sub-second latency, or need to process data that is stored in Amazon S3.
Option C is incorrect because redeploying the model on a P3dn instance would not improve the resource utilization. P3dn instances are designed for distributed machine learning and high performance computing applications that need high network throughput and packet rate performance. They are not optimized for inference workloads.
Option D is incorrect because deploying the model onto an Amazon ECS cluster using a P3 instance would not ensure that provisioned resources are being utilized effectively. Amazon ECS is a fully managed container orchestration service that allows you to run and scale containerized applications on AWS. However, using Amazon ECS would not address the issue of underutilized GPU resources. In fact, it might introduce additional overhead and complexity in managing the cluster.
Reference:
Amazon Elastic Inference - Amazon SageMaker
Batch Transform - Amazon SageMaker
Amazon EC2 P3 Instances
Amazon EC2 P3dn Instances
Amazon Elastic Container Service

QUESTION 149
A data scientist uses an Amazon SageMaker notebook instance to conduct data exploration and analysis. This requires certain Python packages that are not natively available on Amazon SageMaker to be installed on the notebook instance.
How can a machine learning specialist ensure that required packages are automatically available on the notebook instance for the data scientist to use?

A. Install AWS Systems Manager Agent on the underlying Amazon EC2 instance and use Systems Manager Automation to execute the package installation commands.
B. Create a Jupyter notebook file (.ipynb) with cells containing the package installation commands to execute and place the file under the /etc/init directory of each Amazon SageMaker notebook instance.
C. Use the conda package manager from within the Jupyter notebook console to apply the necessary conda packages to the default kernel of the notebook.
D. Create an Amazon SageMaker lifecycle configuration with package installation commands and assign the lifecycle configuration to the notebook instance.

Answer: D

Explanation:
The best way to ensure that required packages are automatically available on the notebook instance for the data scientist to use is to create an Amazon SageMaker lifecycle configuration with package installation commands and assign the lifecycle configuration to the notebook instance. A lifecycle configuration is a shell script that runs when you create or start a notebook instance. You can use a lifecycle configuration to customize the notebook instance by installing libraries, changing environment variables, or downloading datasets. You can also use a lifecycle configuration to automate the installation of custom Python packages that are not natively available on Amazon SageMaker.
Option A is incorrect because installing AWS Systems Manager Agent on the underlying Amazon EC2 instance and using Systems Manager Automation to execute the package installation commands is not a recommended way to customize the notebook instance. Systems Manager Automation is a feature that lets you safely automate common and repetitive IT operations and tasks across AWS resources. However, using Systems Manager Automation would require additional permissions and configurations, and it would not guarantee that the packages are installed before the notebook instance is ready to use.
Option B is incorrect because creating a Jupyter notebook file (.ipynb) with cells containing the package installation commands to execute and placing the file under the /etc/init directory of each Amazon SageMaker notebook instance is not a valid way to customize the notebook instance. The /etc/init directory is used to store scripts that are executed during the boot process of the operating system, not the Jupyter notebook application. Moreover, a Jupyter notebook file is not a shell script that can be executed by the operating system.

Option C is incorrect because using the conda package manager from within the Jupyter notebook console to apply the necessary conda packages to the default kernel of the notebook is not an automatic way to customize the notebook instance. This option would require the data scientist to manually run the conda commands every time they create or start a new notebook instance. This would not be efficient or convenient for the data scientist.
Reference:
Customize a notebook instance using a lifecycle configuration script - Amazon SageMaker
AWS Systems Manager Automation - AWS Systems Manager
Conda environments - Amazon SageMaker

QUESTION 150
A data scientist needs to identify fraudulent user accounts for a company's ecommerce platform. The company wants the ability to determine if a newly created account is associated with a previously known fraudulent user. The data scientist is using AWS Glue to cleanse the company's application logs during ingestion.
Which strategy will allow the data scientist to identify fraudulent accounts?

A. Execute the built-in FindDuplicates Amazon Athena query.
B. Create a FindMatches machine learning transform in AWS Glue.
C. Create an AWS Glue crawler to infer duplicate accounts in the source data.
D. Search for duplicate accounts in the AWS Glue Data Catalog.

Answer: B

Explanation:
The best strategy to identify fraudulent accounts is to create a FindMatches machine learning transform in AWS Glue. The FindMatches transform enables you to identify duplicate or matching records in your dataset, even when the records do not have a common unique identifier and no fields match exactly. This can help you improve fraud detection by finding accounts that are associated with a previously known fraudulent user. You can teach the FindMatches transform your definition of a œduplicate or a œmatch through examples, and it will use machine learning to identify other potential duplicates or matches in your dataset. You can then use the FindMatches transform in your AWS Glue ETL jobs to cleanse your data.
Option A is incorrect because there is no built-in FindDuplicates Amazon Athena query. Amazon Athena is an interactive query service that makes it easy to analyze data in Amazon S3 using standard SQL. However, Amazon Athena does not provide a predefined query to find duplicate records in a dataset. You would have to write your own SQL query to perform this task, which might not be as effective or accurate as using the FindMatches transform.
Option C is incorrect because creating an AWS Glue crawler to infer duplicate accounts in the source data is not a valid strategy. An AWS Glue crawler is a program that connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in the AWS Glue Data Catalog. A crawler does not perform any data

cleansing or record matching tasks.
Option D is incorrect because searching for duplicate accounts in the AWS Glue Data Catalog is not a feasible strategy. The AWS Glue Data Catalog is a central repository to store structural and operational metadata for your data assets. The Data Catalog does not store the actual data, but rather the metadata that describes where the data is located, how it is formatted, and what it contains. Therefore, you cannot search for duplicate records in the Data Catalog.
Reference:
Record matching with AWS Lake Formation FindMatches - AWS Glue
Amazon Athena " Interactive SQL Queries for Data in Amazon S3
AWS Glue Crawlers - AWS Glue
AWS Glue Data Catalog - AWS Glue

QUESTION 151
A Data Scientist is developing a machine learning model to classify whether a financial transaction is fraudulent. The labeled data available for training consists of 100,000 non-fraudulent observations and 1,000 fraudulent observations.
The Data Scientist applies the XGBoost algorithm to the data, resulting in the following confusion matrix when the trained model is applied to a previously unseen validation dataset. The accuracy of the model is 99.1%, but the Data Scientist needs to reduce the number of false negatives.

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

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

A. Change the XGBoost eval_metric parameter to optimize based on Root Mean Square Error (RMSE).
B. Increase the XGBoost scale_pos_weight parameter to adjust the balance of positive and negative weights.
C. Increase the XGBoost max_depth parameter because the model is currently underfitting the data.
D. Change the XGBoost eval_metric parameter to optimize based on Area Under the ROC Curve (AUC).
E. Decrease the XGBoost max_depth parameter because the model is currently overfitting the data.

Answer: B, D

Explanation:
The Data Scientist should increase the XGBoost scale_pos_weight parameter to adjust the balance of positive and negative weights and change the XGBoost eval_metric parameter to optimize based on

Area Under the ROC Curve (AUC). This will help reduce the number of false negative predictions by the model.
The scale_pos_weight parameter controls the balance of positive and negative weights in the XGBoost algorithm. It is useful for imbalanced classification problems, such as fraud detection, where the number of positive examples (fraudulent transactions) is much smaller than the number of negative examples (non-fraudulent transactions). By increasing the scale_pos_weight parameter, the Data Scientist can assign more weight to the positive class and make the model more sensitive to detecting fraudulent transactions.
The eval_metric parameter specifies the metric that is used to measure the performance of the model during training and validation. The default metric for binary classification problems is the error rate, which is the fraction of incorrect predictions. However, the error rate is not a good metric for imbalanced classification problems, because it does not take into account the cost of different types of errors. For example, in fraud detection, a false negative (failing to detect a fraudulent transaction) is more costly than a false positive (flagging a non-fraudulent transaction as fraudulent). Therefore, the Data Scientist should use a metric that reflects the trade-off between the true positive rate (TPR) and the false positive rate (FPR), such as the Area Under the ROC Curve (AUC). The AUC is a measure of how well the model can distinguish between the positive and negative classes, regardless of the classification threshold. A higher AUC means that the model can achieve a higher TPR with a lower FPR, which is desirable for fraud detection.
Reference:
XGBoost Parameters - Amazon Machine Learning
Using XGBoost with Amazon SageMaker - AWS Machine Learning Blog

---

QUESTION 152
A data scientist has developed a machine learning translation model for English to Japanese by using Amazon SageMaker's built-in seq2seq algorithm with 500,000 aligned sentence pairs. While testing with sample sentences, the data scientist finds that the translation quality is reasonable for an example as short as five words. However, the quality becomes unacceptable if the sentence is 100 words long.
Which action will resolve the problem?

A. Change preprocessing to use n-grams.
B. Add more nodes to the recurrent neural network (RNN) than the largest sentence's word count.
C. Adjust hyperparameters related to the attention mechanism.
D. Choose a different weight initialization type.

Answer: C

Explanation:
The data scientist should adjust hyperparameters related to the attention mechanism to resolve the

problem. The attention mechanism is a technique that allows the decoder to focus on different parts of the input sequence when generating the output sequence. It helps the model cope with long input sequences and improve the translation quality. The Amazon SageMaker seq2seq algorithm supports different types of attention mechanisms, such as dot, general, concat, and mlp. The data scientist can use the hyperparameter attention_type to choose the type of attention mechanism. The data scientist can also use the hyperparameter attention_coverage_type to enable coverage, which is a mechanism that penalizes the model for attending to the same input positions repeatedly. By adjusting these hyperparameters, the data scientist can fine-tune the attention mechanism and reduce the number of false negative predictions by the model.
Reference:
Sequence-to-Sequence Algorithm - Amazon SageMaker
Attention Mechanism - Sockeye Documentation

QUESTION 153
A financial company is trying to detect credit card fraud. The company observed that, on average, 2% of credit card transactions were fraudulent. A data scientist trained a classifier on a year's worth of credit card transactions dat
a. The model needs to identify the fraudulent transactions (positives) from the regular ones (negatives). The company's goal is to accurately capture as many positives as possible.
Which metrics should the data scientist use to optimize the model? (Choose two.)

A. Specificity
B. False positive rate
C. Accuracy
D. Area under the precision-recall curve
E. True positive rate

Answer: D, E

Explanation:
The data scientist should use the area under the precision-recall curve and the true positive rate to optimize the model. These metrics are suitable for imbalanced classification problems, such as credit card fraud detection, where the positive class (fraudulent transactions) is much rarer than the negative class (non-fraudulent transactions).
The area under the precision-recall curve (AUPRC) is a measure of how well the model can identify the positive class among all the predicted positives. Precision is the fraction of predicted positives that are actually positive, and recall is the fraction of actual positives that are correctly predicted. A higher AUPRC means that the model can achieve a higher precision with a higher recall, which is desirable for fraud detection.
The true positive rate (TPR) is another name for recall. It is also known as sensitivity or hit rate. It measures the proportion of actual positives that are correctly identified by the model. A higher TPR means that the model can capture more positives, which is the companys goal.

Reference:
Metrics for Imbalanced Classification in Python - Machine Learning Mastery
Precision-Recall - scikit-learn

QUESTION 154
A machine learning specialist is developing a proof of concept for government users whose primary concern is security. The specialist is using Amazon SageMaker to train a convolutional neural network (CNN) model for a photo classifier application. The specialist wants to protect the data so that it cannot be accessed and transferred to a remote host by malicious code accidentally installed on the training container.
Which action will provide the MOST secure protection?

A. Remove Amazon S3 access permissions from the SageMaker execution role.
B. Encrypt the weights of the CNN model.
C. Encrypt the training and validation dataset.
D. Enable network isolation for training jobs.

Answer: D

Explanation:
The most secure action to protect the data from being accessed and transferred to a remote host by malicious code accidentally installed on the training container is to enable network isolation for training jobs. Network isolation is a feature that allows you to run training and inference containers in internet-free mode, which blocks any outbound network calls from the containers, even to other AWS services such as Amazon S3. Additionally, no AWS credentials are made available to the container runtime environment. This way, you can prevent unauthorized access to your data and resources by malicious code or users. You can enable network isolation by setting the EnableNetworkIsolation parameter to True when you call CreateTrainingJob, CreateHyperParameterTuningJob, or CreateModel.
Reference:
Run Training and Inference Containers in Internet-Free Mode - Amazon SageMaker

---

QUESTION 155
A medical imaging company wants to train a computer vision model to detect areas of concern on patients' CT scans. The company has a large collection of unlabeled CT scans that are linked to each patient and stored in an Amazon S3 bucket. The scans must be accessible to authorized users only. A machine learning engineer needs to build a labeling pipeline.
Which set of steps should the engineer take to build the labeling pipeline with the LEAST effort?

A. Create a workforce with AWS Identity and Access Management (IAM). Build a labeling tool on Amazon EC2 Queue images for labeling by using Amazon Simple Queue Service (Amazon SQS). Write the labeling instructions.

B. Create an Amazon Mechanical Turk workforce and manifest file. Create a labeling job by using the built-in image classification task type in Amazon SageMaker Ground Truth. Write the labeling instructions.
C. Create a private workforce and manifest file. Create a labeling job by using the built-in bounding box task type in Amazon SageMaker Ground Truth. Write the labeling instructions.
D. Create a workforce with Amazon Cognito. Build a labeling web application with AWS Amplify. Build a labeling workflow backend using AWS Lambd
a. Write the labeling instructions.

Answer: C

Explanation:
The engineer should create a private workforce and manifest file, and then create a labeling job by using the built-in bounding box task type in Amazon SageMaker Ground Truth. This will allow the engineer to build the labeling pipeline with the least effort.
A private workforce is a group of workers that you manage and who have access to your labeling tasks. You can use a private workforce to label sensitive data that requires confidentiality, such as medical images. You can create a private workforce by using Amazon Cognito and inviting workers by email. You can also use AWS Single Sign-On or your own authentication system to manage your private workforce.
A manifest file is a JSON file that lists the Amazon S3 locations of your input data. You can use a manifest file to specify the data objects that you want to label in your labeling job. You can create a manifest file by using the AWS CLI, the AWS SDK, or the Amazon SageMaker console.
A labeling job is a process that sends your input data to workers for labeling. You can use the Amazon SageMaker console to create a labeling job and choose from several built-in task types, such as image classification, text classification, semantic segmentation, and bounding box. A bounding box task type allows workers to draw boxes around objects in an image and assign labels to them. This is suitable for object detection tasks, such as identifying areas of concern on CT scans.
Reference:
Create and Manage Workforces - Amazon SageMaker
Use Input and Output Data - Amazon SageMaker
Create a Labeling Job - Amazon SageMaker
Bounding Box Task Type - Amazon SageMaker

QUESTION 156
A company is using Amazon Textract to extract textual data from thousands of scanned text-heavy legal documents daily. The company uses this information to process loan applications automatically. Some of the documents fail business validation and are returned to human reviewers, who investigate the errors. This activity increases the time to process the loan applications.
What should the company do to reduce the processing time of loan applications?

A. Configure Amazon Textract to route low-confidence predictions to Amazon SageMaker

Ground Truth. Perform a manual review on those words before performing a business validation.
B. Use an Amazon Textract synchronous operation instead of an asynchronous operation.
C. Configure Amazon Textract to route low-confidence predictions to Amazon Augmented AI (Amazon A2I). Perform a manual review on those words before performing a business validation.
D. Use Amazon Rekognition's feature to detect text in an image to extract the data from scanned images. Use this information to process the loan applications.

Answer: C

Explanation:
The company should configure Amazon Textract to route low-confidence predictions to Amazon Augmented AI (Amazon A2I). Amazon A2I is a service that allows you to implement human review of machine learning (ML) predictions. It also comes integrated with some of the Artificial Intelligence (AI) services such as Amazon Textract. By using Amazon A2I, the company can perform a manual review on those words that have low confidence scores before performing a business validation. This will help reduce the processing time of loan applications by avoiding errors and rework.
Option A is incorrect because Amazon SageMaker Ground Truth is not a suitable service for human review of Amazon Textract predictions. Amazon SageMaker Ground Truth is a service that helps you build highly accurate training datasets for machine learning. It allows you to label your own data or use a workforce of human labelers. However, it does not provide an easy way to integrate with Amazon Textract and route low-confidence predictions for human review.
Option B is incorrect because using an Amazon Textract synchronous operation instead of an asynchronous operation will not reduce the processing time of loan applications. A synchronous operation is a request-response operation that returns the results immediately. An asynchronous operation is a start-and-check operation that returns a job identifier that you can use to check the status and results later. The choice of operation depends on the size and complexity of the document, not on the confidence of the predictions.
Option D is incorrect because using Amazon Rekognitions feature to detect text in an image to extract the data from scanned images is not a better alternative than using Amazon Textract. Amazon Rekognition is a service that provides computer vision capabilities, such as face recognition, object detection, and scene analysis. It can also detect text in an image, but it does not provide the same level of accuracy and functionality as Amazon Textract. Amazon Textract can not only detect text, but also extract data from tables and forms, and understand the layout and structure of the document.
Reference:
Amazon Augmented AI
Amazon SageMaker Ground Truth
Amazon Textract Operations
Amazon Rekognition

QUESTION 157
A company ingests machine learning (ML) data from web advertising clicks into an Amazon S3 data lake. Click data is added to an Amazon Kinesis data stream by using the Kinesis Producer Library

(KPL). The data is loaded into the S3 data lake from the data stream by using an Amazon Kinesis Data Firehose delivery stream. As the data volume increases, an ML specialist notices that the rate of data ingested into Amazon S3 is relatively constant. There also is an increasing backlog of data for Kinesis Data Streams and Kinesis Data Firehose to ingest.
Which next step is MOST likely to improve the data ingestion rate into Amazon S3?

A. Increase the number of S3 prefixes for the delivery stream to write to.
B. Decrease the retention period for the data stream.
C. Increase the number of shards for the data stream.
D. Add more consumers using the Kinesis Client Library (KCL).

Answer: C

Explanation:
The data ingestion rate into Amazon S3 can be improved by increasing the number of shards for the data stream. A shard is the base throughput unit of a Kinesis data stream. One shard provides 1 MB/second data input and 2 MB/second data output. Increasing the number of shards increases the data ingestion capacity of the stream. This can help reduce the backlog of data for Kinesis Data Streams and Kinesis Data Firehose to ingest.
Reference:
Shard - Amazon Kinesis Data Streams
Scaling Amazon Kinesis Data Streams with AWS CloudFormation - AWS Big Data Blog

---

QUESTION 158
A data scientist must build a custom recommendation model in Amazon SageMaker for an online retail company. Due to the nature of the company's products, customers buy only 4-5 products every 5-10 years. So, the company relies on a steady stream of new customers. When a new customer signs up, the company collects data on the customer's preferences. Below is a sample of the data available to the data scientist.

timestamp	user_id	product_id	preference_1	...	preference_10
2020-03-04	90	25	0	...	0.374
2020-03-04	90	61	0	...	0.374
2020-02-21	203	56	1	...	0.098

How should the data scientist split the dataset into a training and test set for this use case?

A. Shuffle all interaction dat
a. Split off the last 10% of the interaction data for the test set.
B. Identify the most recent 10% of interactions for each user. Split off these interactions for the test set.

C. Identify the 10% of users with the least interaction data. Split off all interaction data from these users for the test set.
D. Randomly select 10% of the users. Split off all interaction data from these users for the test set.

Answer: D

Explanation:
The best way to split the dataset into a training and test set for this use case is to randomly select 10% of the users and split off all interaction data from these users for the test set. This is because the company relies on a steady stream of new customers, so the test set should reflect the behavior of new customers who have not been seen by the model before. The other options are not suitable because they either mix old and new customers in the test set (A and B), or they bias the test set towards users with less interaction data Â©. Reference:
Amazon SageMaker Developer Guide: Train and Test Datasets
Amazon Personalize Developer Guide: Preparing and Importing Data

---

QUESTION 159
A financial services company wants to adopt Amazon SageMaker as its default data science environment. The company's data scientists run machine learning (ML) models on confidential financial dat
a. The company is worried about data egress and wants an ML engineer to secure the environment. Which mechanisms can the ML engineer use to control data egress from SageMaker? (Choose three.)

A. Connect to SageMaker by using a VPC interface endpoint powered by AWS PrivateLink.
B. Use SCPs to restrict access to SageMaker.
C. Disable root access on the SageMaker notebook instances.
D. Enable network isolation for training jobs and models.
E. Restrict notebook presigned URLs to specific IPs used by the company.
F. Protect data with encryption at rest and in transit. Use AWS Key Management Service (AWS KMS) to manage encryption keys.

Answer: A, D, F
To control data egress from SageMaker, the ML engineer can use the following mechanisms:
Connect to SageMaker by using a VPC interface endpoint powered by AWS PrivateLink. This allows the ML engineer to access SageMaker services and resources without exposing the traffic to the public internet. This reduces the risk of data leakage and unauthorized access1
Enable network isolation for training jobs and models. This prevents the training jobs and models from accessing the internet or other AWS services. This ensures that the data used for training and inference is not exposed to external sources2
Protect data with encryption at rest and in transit. Use AWS Key Management Service (AWS KMS) to

manage encryption keys. This enables the ML engineer to encrypt the data stored in Amazon S3 buckets, SageMaker notebook instances, and SageMaker endpoints. It also allows the ML engineer to encrypt the data in transit between SageMaker and other AWS services. This helps protect the data from unauthorized access and tampering3

The other options are not effective in controlling data egress from SageMaker:

Use SCPs to restrict access to SageMaker. SCPs are used to define the maximum permissions for an organization or organizational unit (OU) in AWS Organizations. They do not control the data egress from SageMaker, but rather the access to SageMaker itself4

Disable root access on the SageMaker notebook instances. This prevents the users from installing additional packages or libraries on the notebook instances. It does not prevent the data from being transferred out of the notebook instances.

Restrict notebook presigned URLs to specific IPs used by the company. This limits the access to the notebook instances from certain IP addresses. It does not prevent the data from being transferred out of the notebook instances.

Reference:

1: Amazon SageMaker Interface VPC Endpoints (AWS PrivateLink) - Amazon SageMaker
2: Network Isolation - Amazon SageMaker
3: Encrypt Data at Rest and in Transit - Amazon SageMaker
4: Using Service Control Policies - AWS Organizations
: Disable Root Access - Amazon SageMaker
: Create a Presigned Notebook Instance URL - Amazon SageMaker

---

QUESTION 160

A company needs to quickly make sense of a large amount of data and gain insight from it. The data is in different formats, the schemas change frequently, and new data sources are added regularly. The company wants to use AWS services to explore multiple data sources, suggest schemas, and enrich and transform the dat

a. The solution should require the least possible coding effort for the data flows and the least possible infrastructure management.

Which combination of AWS services will meet these requirements?

A.

Amazon EMR for data discovery, enrichment, and transformation
Amazon Athena for querying and analyzing the results in Amazon S3 using standard SQL
Amazon QuickSight for reporting and getting insights

B.

Amazon Kinesis Data Analytics for data ingestion
Amazon EMR for data discovery, enrichment, and transformation
Amazon Redshift for querying and analyzing the results in Amazon S3

C.

AWS Glue for data discovery, enrichment, and transformation
Amazon Athena for querying and analyzing the results in Amazon S3 using standard SQL
Amazon QuickSight for reporting and getting insights

D.
AWS Data Pipeline for data transfer
AWS Step Functions for orchestrating AWS Lambda jobs for data discovery, enrichment, and transformation
Amazon Athena for querying and analyzing the results in Amazon S3 using standard SQL
Amazon QuickSight for reporting and getting insights

Answer: C

Explanation:
The best combination of AWS services to meet the requirements of data discovery, enrichment, transformation, querying, analysis, and reporting with the least coding and infrastructure management is AWS Glue, Amazon Athena, and Amazon QuickSight. These services are:
AWS Glue for data discovery, enrichment, and transformation. AWS Glue is a serverless data integration service that automatically crawls, catalogs, and prepares data from various sources and formats. It also provides a visual interface called AWS Glue DataBrew that allows users to apply over 250 transformations to clean, normalize, and enrich data without writing code1
Amazon Athena for querying and analyzing the results in Amazon S3 using standard SQL. Amazon Athena is a serverless interactive query service that allows users to analyze data in Amazon S3 using standard SQL. It supports a variety of data formats, such as CSV, JSON, ORC, Parquet, and Avro. It also integrates with AWS Glue Data Catalog to provide a unified view of the data sources and schemas2
Amazon QuickSight for reporting and getting insights. Amazon QuickSight is a serverless business intelligence service that allows users to create and share interactive dashboards and reports. It also provides ML-powered features, such as anomaly detection, forecasting, and natural language queries, to help users discover hidden insights from their data3
The other options are not suitable because they either require more coding effort, more infrastructure management, or do not support the desired use cases. For example:
Option A uses Amazon EMR for data discovery, enrichment, and transformation. Amazon EMR is a managed cluster platform that runs Apache Spark, Apache Hive, and other open-source frameworks for big data processing. It requires users to write code in languages such as Python, Scala, or SQL to perform data integration tasks. It also requires users to provision, configure, and scale the clusters according to their needs4
Option B uses Amazon Kinesis Data Analytics for data ingestion. Amazon Kinesis Data Analytics is a service that allows users to process streaming data in real time using SQL or Apache Flink. It is not suitable for data discovery, enrichment, and transformation, which are typically batch-oriented tasks. It also requires users to write code to define the data processing logic and the output destination5
Option D uses AWS Data Pipeline for data transfer and AWS Step Functions for orchestrating AWS Lambda jobs for data discovery, enrichment, and transformation. AWS Data Pipeline is a service that helps users move data between AWS services and on-premises data sources. AWS Step Functions is a service that helps users coordinate multiple AWS services into workflows. AWS Lambda is a service that lets users run code without provisioning or managing servers. These services require users to

write code to define the data sources, destinations, transformations, and workflows. They also require users to manage the scalability, performance, and reliability of the data pipelines.
Reference:
1: AWS Glue - Data Integration Service - Amazon Web Services
2: Amazon Athena " Interactive SQL Query Service - AWS
3: Amazon QuickSight - Business Intelligence Service - AWS
4: Amazon EMR - Amazon Web Services
5: Amazon Kinesis Data Analytics - Amazon Web Services
: AWS Data Pipeline - Amazon Web Services
: AWS Step Functions - Amazon Web Services
: AWS Lambda - Amazon Web Services

QUESTION 161
A company is converting a large number of unstructured paper receipts into images. The company wants to create a model based on natural language processing (NLP) to find relevant entities such as date, location, and notes, as well as some custom entities such as receipt numbers.
The company is using optical character recognition (OCR) to extract text for data labeling. However, documents are in different structures and formats, and the company is facing challenges with setting up the manual workflows for each document type. Additionally, the company trained a named entity recognition (NER) model for custom entity detection using a small sample size. This model has a very low confidence score and will require retraining with a large dataset.
Which solution for text extraction and entity detection will require the LEAST amount of effort?

A. Extract text from receipt images by using Amazon Textract. Use the Amazon SageMaker BlazingText algorithm to train on the text for entities and custom entities.
B. Extract text from receipt images by using a deep learning OCR model from the AWS Marketplace. Use the NER deep learning model to extract entities.
C. Extract text from receipt images by using Amazon Textract. Use Amazon Comprehend for entity detection, and use Amazon Comprehend custom entity recognition for custom entity detection.
D. Extract text from receipt images by using a deep learning OCR model from the AWS Marketplace. Use Amazon Comprehend for entity detection, and use Amazon Comprehend custom entity recognition for custom entity detection.

Answer: C

Explanation:
The best solution for text extraction and entity detection with the least amount of effort is to use Amazon Textract and Amazon Comprehend. These services are:
Amazon Textract for text extraction from receipt images. Amazon Textract is a machine learning service that can automatically extract text and data from scanned documents. It can handle different structures and formats of documents, such as PDF, TIFF, PNG, and JPEG, without any preprocessing steps. It can also extract key-value pairs and tables from documents1

Amazon Comprehend for entity detection and custom entity detection. Amazon Comprehend is a natural language processing service that can identify entities, such as dates, locations, and notes, from unstructured text. It can also detect custom entities, such as receipt numbers, by using a custom entity recognizer that can be trained with a small amount of labeled data2
The other options are not suitable because they either require more effort for text extraction, entity detection, or custom entity detection. For example:
Option A uses the Amazon SageMaker BlazingText algorithm to train on the text for entities and custom entities. BlazingText is a supervised learning algorithm that can perform text classification and word2vec. It requires users to provide a large amount of labeled data, preprocess the data into a specific format, and tune the hyperparameters of the model3
Option B uses a deep learning OCR model from the AWS Marketplace and a NER deep learning model for text extraction and entity detection. These models are pre-trained and may not be suitable for the specific use case of receipt processing. They also require users to deploy and manage the models on Amazon SageMaker or Amazon EC2 instances4
Option D uses a deep learning OCR model from the AWS Marketplace for text extraction. This model has the same drawbacks as option B. It also requires users to integrate the model output with Amazon Comprehend for entity detection and custom entity detection.
Reference:
1: Amazon Textract " Extract text and data from documents
2: Amazon Comprehend " Natural Language Processing (NLP) and Machine Learning (ML)
3: BlazingText - Amazon SageMaker
4: AWS Marketplace: OCR

QUESTION 162
A company is building a predictive maintenance model based on machine learning (ML). The data is stored in a fully private Amazon S3 bucket that is encrypted at rest with AWS Key Management Service (AWS KMS) CMKs. An ML specialist must run data preprocessing by using an Amazon SageMaker Processing job that is triggered from code in an Amazon SageMaker notebook. The job should read data from Amazon S3, process it, and upload it back to the same S3 bucket. The preprocessing code is stored in a container image in Amazon Elastic Container Registry (Amazon ECR). The ML specialist needs to grant permissions to ensure a smooth data preprocessing workflow.
Which set of actions should the ML specialist take to meet these requirements?

A. Create an IAM role that has permissions to create Amazon SageMaker Processing jobs, S3 read and write access to the relevant S3 bucket, and appropriate KMS and ECR permissions. Attach the role to the SageMaker notebook instance. Create an Amazon SageMaker Processing job from the notebook.
B. Create an IAM role that has permissions to create Amazon SageMaker Processing jobs. Attach the role to the SageMaker notebook instance. Create an Amazon SageMaker Processing job with an IAM role that has read and write permissions to the relevant S3 bucket, and appropriate KMS and ECR permissions.
C. Create an IAM role that has permissions to create Amazon SageMaker Processing jobs and to

access Amazon ECR. Attach the role to the SageMaker notebook instance. Set up both an S3 endpoint and a KMS endpoint in the default VPC. Create Amazon SageMaker Processing jobs from the notebook.
D. Create an IAM role that has permissions to create Amazon SageMaker Processing jobs. Attach the role to the SageMaker notebook instance. Set up an S3 endpoint in the default VPC. Create Amazon SageMaker Processing jobs with the access key and secret key of the IAM user with appropriate KMS and ECR permissions.

Answer: B

Explanation:
The correct solution for granting permissions for data preprocessing is to use the following steps:
Create an IAM role that has permissions to create Amazon SageMaker Processing jobs. Attach the role to the SageMaker notebook instance. This role allows the ML specialist to run Processing jobs from the notebook code1
Create an Amazon SageMaker Processing job with an IAM role that has read and write permissions to the relevant S3 bucket, and appropriate KMS and ECR permissions. This role allows the Processing job to access the data in the encrypted S3 bucket, decrypt it with the KMS CMK, and pull the container image from ECR23
The other options are incorrect because they either miss some permissions or use unnecessary steps. For example:
Option A uses a single IAM role for both the notebook instance and the Processing job. This role may have more permissions than necessary for the notebook instance, which violates the principle of least privilege4
Option C sets up both an S3 endpoint and a KMS endpoint in the default VPC. These endpoints are not required for the Processing job to access the data in the encrypted S3 bucket. They are only needed if the Processing job runs in network isolation mode, which is not specified in the question.
Option D uses the access key and secret key of the IAM user with appropriate KMS and ECR permissions. This is not a secure way to pass credentials to the Processing job. It also requires the ML specialist to manage the IAM user and the keys.
Reference:
1: Create an Amazon SageMaker Notebook Instance - Amazon SageMaker
2: Create a Processing Job - Amazon SageMaker
3: Use AWS KMS"Managed Encryption Keys - Amazon Simple Storage Service
4: IAM Best Practices - AWS Identity and Access Management
: Network Isolation - Amazon SageMaker
: Understanding and Getting Your Security Credentials - AWS General Reference

---

QUESTION 163
A data scientist has been running an Amazon SageMaker notebook instance for a few weeks. During this time, a new version of Jupyter Notebook was released along with additional software updates. The security team mandates that all running SageMaker notebook instances use the latest security

and software updates provided by SageMaker.
How can the data scientist meet these requirements?

A. Call the CreateNotebookInstanceLifecycleConfig API operation
B. Create a new SageMaker notebook instance and mount the Amazon Elastic Block Store (Amazon EBS) volume from the original instance
C. Stop and then restart the SageMaker notebook instance
D. Call the UpdateNotebookInstanceLifecycleConfig API operation

Answer: C

Explanation:
The correct solution for updating the software on a SageMaker notebook instance is to stop and then restart the notebook instance. This will automatically apply the latest security and software updates provided by SageMaker1
The other options are incorrect because they either do not update the software or require unnecessary steps. For example:
Option A calls the CreateNotebookInstanceLifecycleConfig API operation. This operation creates a lifecycle configuration, which is a set of shell scripts that run when a notebook instance is created or started. A lifecycle configuration can be used to customize the notebook instance, such as installing additional libraries or packages. However, it does not update the software on the notebook instance2
Option B creates a new SageMaker notebook instance and mounts the Amazon Elastic Block Store (Amazon EBS) volume from the original instance. This option will create a new notebook instance with the latest software, but it will also incur additional costs and require manual steps to transfer the data and settings from the original instance3
Option D calls the UpdateNotebookInstanceLifecycleConfig API operation. This operation updates an existing lifecycle configuration. As explained in option A, a lifecycle configuration does not update the software on the notebook instance4
Reference:
1: Amazon SageMaker Notebook Instances - Amazon SageMaker
2: CreateNotebookInstanceLifecycleConfig - Amazon SageMaker
3: Create a Notebook Instance - Amazon SageMaker
4: UpdateNotebookInstanceLifecycleConfig - Amazon SageMaker

QUESTION 164
A library is developing an automatic book-borrowing system that uses Amazon Rekognition. Images of library members faces are stored in an Amazon S3 bucket. When members borrow books, the Amazon Rekognition CompareFaces API operation compares real faces against the stored faces in Amazon S3.
The library needs to improve security by making sure that images are encrypted at rest. Also, when the images are used with Amazon Rekognition. they need to be encrypted in transit. The library also must ensure that the images are not used to improve Amazon Rekognition as a service.

How should a machine learning specialist architect the solution to satisfy these requirements?

A. Enable server-side encryption on the S3 bucket. Submit an AWS Support ticket to opt out of allowing images to be used for improving the service, and follow the process provided by AWS Support.
B. Switch to using an Amazon Rekognition collection to store the images. Use the IndexFaces and SearchFacesByImage API operations instead of the CompareFaces API operation.
C. Switch to using the AWS GovCloud (US) Region for Amazon S3 to store images and for Amazon Rekognition to compare faces. Set up a VPN connection and only call the Amazon Rekognition API operations through the VPN.
D. Enable client-side encryption on the S3 bucket. Set up a VPN connection and only call the Amazon Rekognition API operations through the VPN.

Answer: A

Explanation:
The best solution for encrypting images at rest and in transit, and opting out of data usage for service improvement, is to use the following steps:
Enable server-side encryption on the S3 bucket. This will encrypt the images stored in the bucket using AWS Key Management Service (AWS KMS) customer master keys (CMKs). This will protect the data at rest from unauthorized access1
Submit an AWS Support ticket to opt out of allowing images to be used for improving the service, and follow the process provided by AWS Support. This will prevent AWS from storing or using the images processed by Amazon Rekognition for service development or enhancement purposes. This will protect the data privacy and ownership2
Use HTTPS to call the Amazon Rekognition CompareFaces API operation. This will encrypt the data in transit between the client and the server using SSL/TLS protocols. This will protect the data from interception or tampering3
The other options are incorrect because they either do not encrypt the images at rest or in transit, or do not opt out of data usage for service improvement. For example:
Option B switches to using an Amazon Rekognition collection to store the images. A collection is a container for storing face vectors that are calculated by Amazon Rekognition. It does not encrypt the images at rest or in transit, and it does not opt out of data usage for service improvement. It also requires changing the API operations from CompareFaces to IndexFaces and SearchFacesByImage, which may not have the same functionality or performance4
Option C switches to using the AWS GovCloud (US) Region for Amazon S3 and Amazon Rekognition. The AWS GovCloud (US) Region is an isolated AWS Region designed to host sensitive data and regulated workloads in the cloud. It does not automatically encrypt the images at rest or in transit, and it does not opt out of data usage for service improvement. It also requires migrating the data and the application to a different Region, which may incur additional costs and complexity5
Option D enables client-side encryption on the S3 bucket. This means that the client is responsible for encrypting and decrypting the images before uploading or downloading them from the bucket.

This adds extra overhead and complexity to the client application, and it does not encrypt the data in transit when calling the Amazon Rekognition API. It also does not opt out of data usage for service improvement.
Reference:
1: Protecting Data Using Server-Side Encryption with AWS KMS"Managed Keys (SSE-KMS) - Amazon Simple Storage Service
2: Opting Out of Content Storage and Use for Service Improvements - Amazon Rekognition
3: HTTPS - Wikipedia
4: Working with Stored Faces - Amazon Rekognition
5: AWS GovCloud (US) - Amazon Web Services
: Protecting Data Using Client-Side Encryption - Amazon Simple Storage Service

---

QUESTION 165
A company is building a line-counting application for use in a quick-service restaurant. The company wants to use video cameras pointed at the line of customers at a given register to measure how many people are in line and deliver notifications to managers if the line grows too long. The restaurant locations have limited bandwidth for connections to external services and cannot accommodate multiple video streams without impacting other operations.
Which solution should a machine learning specialist implement to meet these requirements?

A. Install cameras compatible with Amazon Kinesis Video Streams to stream the data to AWS over the restaurant's existing internet connection. Write an AWS Lambda function to take an image and send it to Amazon Rekognition to count the number of faces in the image. Send an Amazon Simple Notification Service (Amazon SNS) notification if the line is too long.
B. Deploy AWS DeepLens cameras in the restaurant to capture video. Enable Amazon Rekognition on the AWS DeepLens device, and use it to trigger a local AWS Lambda function when a person is recognized. Use the Lambda function to send an Amazon Simple Notification Service (Amazon SNS) notification if the line is too long.
C. Build a custom model in Amazon SageMaker to recognize the number of people in an image. Install cameras compatible with Amazon Kinesis Video Streams in the restaurant. Write an AWS Lambda function to take an image. Use the SageMaker endpoint to call the model to count people. Send an Amazon Simple Notification Service (Amazon SNS) notification if the line is too long.
D. Build a custom model in Amazon SageMaker to recognize the number of people in an image. Deploy AWS DeepLens cameras in the restaurant. Deploy the model to the cameras. Deploy an AWS Lambda function to the cameras to use the model to count people and send an Amazon Simple Notification Service (Amazon SNS) notification if the line is too long.

Answer: D

Explanation:
The best solution for building a line-counting application for use in a quick-service restaurant is to use the following steps:

Build a custom model in Amazon SageMaker to recognize the number of people in an image. Amazon SageMaker is a fully managed service that provides tools and workflows for building, training, and deploying machine learning models. A custom model can be tailored to the specific use case of linecounting and achieve higher accuracy than a generic model1
Deploy AWS DeepLens cameras in the restaurant to capture video. AWS DeepLens is a wireless video camera that integrates with Amazon SageMaker and AWS Lambda. It can run machine learning inference locally on the device without requiring internet connectivity or streaming video to the cloud. This reduces the bandwidth consumption and latency of the application2
Deploy the model to the cameras. AWS DeepLens allows users to deploy trained models from Amazon SageMaker to the cameras with a few clicks. The cameras can then use the model to process the video frames and count the number of people in each frame2
Deploy an AWS Lambda function to the cameras to use the model to count people and send an Amazon Simple Notification Service (Amazon SNS) notification if the line is too long. AWS Lambda is a serverless computing service that lets users run code without provisioning or managing servers. AWS DeepLens supports running Lambda functions on the device to perform actions based on the inference results. Amazon SNS is a service that enables users to send notifications to subscribers via email, SMS, or mobile push23
The other options are incorrect because they either require internet connectivity or streaming video to the cloud, which may impact the bandwidth and performance of the application. For example:
Option A uses Amazon Kinesis Video Streams to stream the data to AWS over the restaurants existing internet connection. Amazon Kinesis Video Streams is a service that enables users to capture, process, and store video streams for analytics and machine learning. However, this option requires streaming multiple video streams to the cloud, which may consume a lot of bandwidth and cause network congestion. It also requires internet connectivity, which may not be reliable or available in some locations4
Option B uses Amazon Rekognition on the AWS DeepLens device. Amazon Rekognition is a service that provides computer vision capabilities, such as face detection, face recognition, and object detection. However, this option requires calling the Amazon Rekognition API over the internet, which may introduce latency and require bandwidth. It also uses a generic face detection model, which may not be optimized for the line-counting use case.
Option C uses Amazon SageMaker to build a custom model and an Amazon SageMaker endpoint to call the model. Amazon SageMaker endpoints are hosted web services that allow users to perform inference on their models. However, this option requires sending the images to the endpoint over the internet, which may consume bandwidth and introduce latency. It also requires internet connectivity, which may not be reliable or available in some locations.
Reference:
1: Amazon SageMaker " Machine Learning Service - AWS
2: AWS DeepLens - Deep learning enabled video camera - AWS
3: Amazon Simple Notification Service (SNS) - AWS
4: Amazon Kinesis Video Streams - Amazon Web Services
: Amazon Rekognition " Video and Image - AWS
: Deploy a Model - Amazon SageMaker

QUESTION 166
A company has set up and deployed its machine learning (ML) model into production with an endpoint using Amazon SageMaker hosting services. The ML team has configured automatic scaling for its SageMaker instances to support workload changes. During testing, the team notices that additional instances are being launched before the new instances are ready. This behavior needs to change as soon as possible.
How can the ML team solve this issue?

A. Decrease the cooldown period for the scale-in activity. Increase the configured maximum capacity of instances.
B. Replace the current endpoint with a multi-model endpoint using SageMaker.
C. Set up Amazon API Gateway and AWS Lambda to trigger the SageMaker inference endpoint.
D. Increase the cooldown period for the scale-out activity.

Answer: D

Explanation:
: The correct solution for changing the scaling behavior of the SageMaker instances is to increase the cooldown period for the scale-out activity. The cooldown period is the amount of time, in seconds, after a scaling activity completes before another scaling activity can start. By increasing the cooldown period for the scale-out activity, the ML team can ensure that the new instances are ready before launching additional instances. This will prevent over-scaling and reduce costs1
The other options are incorrect because they either do not solve the issue or require unnecessary steps. For example:
Option A decreases the cooldown period for the scale-in activity and increases the configured maximum capacity of instances. This option does not address the issue of launching additional instances before the new instances are ready. It may also cause under-scaling and performance degradation.
Option B replaces the current endpoint with a multi-model endpoint using SageMaker. A multimodel endpoint is an endpoint that can host multiple models using a single endpoint. It does not affect the scaling behavior of the SageMaker instances. It also requires creating a new endpoint and updating the application code to use it2
Option C sets up Amazon API Gateway and AWS Lambda to trigger the SageMaker inference endpoint. Amazon API Gateway is a service that allows users to create, publish, maintain, monitor, and secure APIs. AWS Lambda is a service that lets users run code without provisioning or managing servers. These services do not affect the scaling behavior of the SageMaker instances. They also require creating and configuring additional resources and services34
Reference:
1: Automatic Scaling - Amazon SageMaker
2: Create a Multi-Model Endpoint - Amazon SageMaker

3: Amazon API Gateway - Amazon Web Services
4: AWS Lambda - Amazon Web Services

QUESTION 167
A telecommunications company is developing a mobile app for its customers. The company is using an Amazon SageMaker hosted endpoint for machine learning model inferences.
Developers want to introduce a new version of the model for a limited number of users who subscribed to a preview feature of the app. After the new version of the model is tested as a preview, developers will evaluate its accuracy. If a new version of the model has better accuracy, developers need to be able to gradually release the new version for all users over a fixed period of time.
How can the company implement the testing model with the LEAST amount of operational overhead?

A. Update the ProductionVariant data type with the new version of the model by using the CreateEndpointConfig operation with the InitialVariantWeight parameter set to 0. Specify the TargetVariant parameter for InvokeEndpoint calls for users who subscribed to the preview feature. When the new version of the model is ready for release, gradually increase InitialVariantWeight until all users have the updated version.
B. Configure two SageMaker hosted endpoints that serve the different versions of the model. Create an Application Load Balancer (ALB) to route traffic to both endpoints based on the TargetVariant query string parameter. Reconfigure the app to send the TargetVariant query string parameter for users who subscribed to the preview feature. When the new version of the model is ready for release, change the ALB's routing algorithm to weighted until all users have the updated version.
C. Update the DesiredWeightsAndCapacity data type with the new version of the model by using the UpdateEndpointWeightsAndCapacities operation with the DesiredWeight parameter set to 0. Specify the TargetVariant parameter for InvokeEndpoint calls for users who subscribed to the preview feature. When the new version of the model is ready for release, gradually increase DesiredWeight until all users have the updated version.
D. Configure two SageMaker hosted endpoints that serve the different versions of the model. Create an Amazon Route 53 record that is configured with a simple routing policy and that points to the current version of the model. Configure the mobile app to use the endpoint URL for users who subscribed to the preview feature and to use the Route 53 record for other users. When the new version of the model is ready for release, add a new model version endpoint to Route 53, and switch the policy to weighted until all users have the updated version.

Answer: C

Explanation:
The best solution for implementing the testing model with the least amount of operational overhead is to use the following steps:
Update the DesiredWeightsAndCapacity data type with the new version of the model by using the UpdateEndpointWeightsAndCapacities operation with the DesiredWeight parameter set to 0. This

operation allows the developers to update the variant weights and capacities of an existing SageMaker endpoint without deleting and recreating the endpoint. Setting the DesiredWeight parameter to 0 means that the new version of the model will not receive any traffic initially1
Specify the TargetVariant parameter for InvokeEndpoint calls for users who subscribed to the preview feature. This parameter allows the developers to override the variant weights and direct a request to a specific variant. This way, the developers can test the new version of the model for a limited number of users who opted in for the preview feature2
When the new version of the model is ready for release, gradually increase DesiredWeight until all users have the updated version. This operation allows the developers to perform a gradual rollout of the new version of the model and monitor its performance and accuracy. The developers can adjust the variant weights and capacities as needed until the new version of the model serves all the traffic1
The other options are incorrect because they either require more operational overhead or do not support the desired use cases. For example:
Option A uses the CreateEndpointConfig operation with the InitialVariantWeight parameter set to 0. This operation creates a new endpoint configuration, which requires deleting and recreating the endpoint to apply the changes. This adds extra overhead and downtime for the endpoint. It also does not support the gradual rollout of the new version of the model3
Option B uses two SageMaker hosted endpoints that serve the different versions of the model and an Application Load Balancer (ALB) to route traffic to both endpoints based on the TargetVariant query string parameter. This option requires creating and managing additional resources and services, such as the second endpoint and the ALB. It also requires changing the app code to send the query string parameter for the preview feature4
Option D uses the access key and secret key of the IAM user with appropriate KMS and ECR permissions. This is not a secure way to pass credentials to the Processing job. It also requires the ML specialist to manage the IAM user and the keys.
Reference:
1: UpdateEndpointWeightsAndCapacities - Amazon SageMaker
2: InvokeEndpoint - Amazon SageMaker
3: CreateEndpointConfig - Amazon SageMaker
4: Application Load Balancer - Elastic Load Balancing

QUESTION 168
A company offers an online shopping service to its customers. The company wants to enhance the sites security by requesting additional information when customers access the site from locations that are different from their normal location. The company wants to update the process to call a machine learning (ML) model to determine when additional information should be requested.
The company has several terabytes of data from its existing ecommerce web servers containing the source IP addresses for each request made to the web server. For authenticated requests, the records also contain the login name of the requesting user.
Which approach should an ML specialist take to implement the new security feature in the web application?

A. Use Amazon SageMaker Ground Truth to label each record as either a successful or failed access attempt. Use Amazon SageMaker to train a binary classification model using the factorization machines (FM) algorithm.
B. Use Amazon SageMaker to train a model using the IP Insights algorithm. Schedule updates and retraining of the model using new log data nightly.
C. Use Amazon SageMaker Ground Truth to label each record as either a successful or failed access attempt. Use Amazon SageMaker to train a binary classification model using the IP Insights algorithm.
D. Use Amazon SageMaker to train a model using the Object2Vec algorithm. Schedule updates and retraining of the model using new log data nightly.

Answer: B

Explanation:
The IP Insights algorithm is designed to capture associations between entities and IP addresses, and can be used to identify anomalous IP usage patterns. The algorithm can learn from historical data that contains pairs of entities and IP addresses, and can return a score that indicates how likely the pair is to occur. The company can use this algorithm to train a model that can detect when a customer is accessing the site from a different location than usual, and request additional information accordingly. The company can also schedule updates and retraining of the model using new log data nightly to keep the model up to date with the latest IP usage patterns.
The other options are not suitable for this use case because:
Option A: The factorization machines (FM) algorithm is a general-purpose supervised learning algorithm that can be used for both classification and regression tasks. However, it is not optimized for capturing associations between entities and IP addresses, and would require labeling each record as either a successful or failed access attempt, which is a costly and time-consuming process.
Option C: The IP Insights algorithm is a good choice for this use case, but it does not require labeling each record as either a successful or failed access attempt. The algorithm is unsupervised and can learn from the historical data without labels. Labeling the data would be unnecessary and wasteful.
Option D: The Object2Vec algorithm is a general-purpose neural embedding algorithm that can learn low-dimensional dense embeddings of high-dimensional objects. However, it is not designed to capture associations between entities and IP addresses, and would require a different input format than the one provided by the company. The Object2Vec algorithm expects pairs of objects and their relationship labels or scores as inputs, while the company has data containing the source IP addresses and the login names of the requesting users.
Reference:
IP Insights - Amazon SageMaker
Factorization Machines Algorithm - Amazon SageMaker
Object2Vec Algorithm - Amazon SageMaker

QUESTION 169
A retail company wants to combine its customer orders with the product description data from its product catalog. The structure and format of the records in each dataset is different. A data analyst tried to use a spreadsheet to combine the datasets, but the effort resulted in duplicate records and records that were not properly combined. The company needs a solution that it can use to combine similar records from the two datasets and remove any duplicates.
Which solution will meet these requirements?

A. Use an AWS Lambda function to process the data. Use two arrays to compare equal strings in the fields from the two datasets and remove any duplicates.
B. Create AWS Glue crawlers for reading and populating the AWS Glue Data Catalog. Call the AWS Glue SearchTables API operation to perform a fuzzy-matching search on the two datasets, and cleanse the data accordingly.
C. Create AWS Glue crawlers for reading and populating the AWS Glue Data Catalog. Use the FindMatches transform to cleanse the data.
D. Create an AWS Lake Formation custom transform. Run a transformation for matching products from the Lake Formation console to cleanse the data automatically.

Answer: C

Explanation:
The FindMatches transform is a machine learning transform that can identify and match similar records from different datasets, even when the records do not have a common unique identifier or exact field values. The FindMatches transform can also remove duplicate records from a single dataset. The FindMatches transform can be used with AWS Glue crawlers and jobs to process the data from various sources and store it in a data lake. The FindMatches transform can be created and managed using the AWS Glue console, API, or AWS Glue Studio.
The other options are not suitable for this use case because:
Option A: Using an AWS Lambda function to process the data and compare equal strings in the fields from the two datasets is not an efficient or scalable solution. It would require writing custom code and handling the data loading and cleansing logic. It would also not account for variations or inconsistencies in the field values, such as spelling errors, abbreviations, or missing data.
Option B: The AWS Glue SearchTables API operation is used to search for tables in the AWS Glue Data Catalog based on a set of criteria. It is not a machine learning transform that can match records across different datasets or remove duplicates. It would also require writing custom code to invoke the API and process the results.
Option D: AWS Lake Formation does not provide a custom transform feature. It provides predefined blueprints for common data ingestion scenarios, such as database snapshot, incremental database, and log file. These blueprints do not support matching records across different datasets or removing duplicates.

QUESTION 170
A company provisions Amazon SageMaker notebook instances for its data science team and creates Amazon VPC interface endpoints to ensure communication between the VPC and the notebook instances. All connections to the Amazon SageMaker API are contained entirely and securely using the AWS network. However, the data science team realizes that individuals outside the VPC can still connect to the notebook instances across the internet.
Which set of actions should the data science team take to fix the issue?

A. Modify the notebook instances' security group to allow traffic only from the CIDR ranges of the VPC. Apply this security group to all of the notebook instances' VPC interfaces.
B. Create an IAM policy that allows the sagemaker:CreatePresignedNotebooklnstanceUrl and sagemaker:DescribeNotebooklnstance actions from only the VPC endpoints. Apply this policy to all IAM users, groups, and roles used to access the notebook instances.
C. Add a NAT gateway to the VPC. Convert all of the subnets where the Amazon SageMaker notebook instances are hosted to private subnets. Stop and start all of the notebook instances to reassign only private IP addresses.
D. Change the network ACL of the subnet the notebook is hosted in to restrict access to anyone outside the VPC.

Answer: A

Explanation:
The issue is that the notebook instances security group allows inbound traffic from any source IP address, which means that anyone with the authorized URL can access the notebook instances over the internet. To fix this issue, the data science team should modify the security group to allow traffic only from the CIDR ranges of the VPC, which are the IP addresses assigned to the resources within the VPC. This way, only the VPC interface endpoints and the resources within the VPC can communicate with the notebook instances. The data science team should apply this security group to all of the notebook instances VPC interfaces, which are the network interfaces that connect the notebook instances to the VPC.
The other options are not correct because:
Option B: Creating an IAM policy that allows the sagemaker:CreatePresignedNotebookInstanceUrl and sagemaker:DescribeNotebookInstance actions from only the VPC endpoints does not prevent individuals outside the VPC from accessing the notebook instances. These actions are used to generate and retrieve the authorized URL for the notebook instances, but they do not control who can use the URL to access the notebook instances. The URL can still be shared or leaked to unauthorized users, who can then access the notebook instances over the internet.
Option C: Adding a NAT gateway to the VPC and converting the subnets where the notebook instances are hosted to private subnets does not solve the issue either. A NAT gateway is used to enable outbound internet access from a private subnet, but it does not affect inbound internet access. The notebook instances can still be accessed over the internet if their security group allows inbound traffic from any source IP address. Moreover, stopping and starting the notebook instances

to reassign only private IP addresses is not necessary, because the notebook instances already have private IP addresses assigned by the VPC interface endpoints.
Option D: Changing the network ACL of the subnet the notebook is hosted in to restrict access to anyone outside the VPC is not a good practice, because network ACLs are stateless and apply to the entire subnet. This means that the data science team would have to specify both the inbound and outbound rules for each IP address range that they want to allow or deny. This can be cumbersome and error-prone, especially if the VPC has multiple subnets and resources. It is better to use security groups, which are stateful and apply to individual resources, to control the access to the notebook instances.
Reference:
Connect to SageMaker Within your VPC - Amazon SageMaker
Security Groups for Your VPC - Amazon Virtual Private Cloud
VPC Interface Endpoints - Amazon Virtual Private Cloud

QUESTION 171
A company will use Amazon SageMaker to train and host a machine learning (ML) model for a marketing campaign. The majority of data is sensitive customer dat
a. The data must be encrypted at rest. The company wants AWS to maintain the root of trust for the master keys and wants encryption key usage to be logged.
Which implementation will meet these requirements?

A. Use encryption keys that are stored in AWS Cloud HSM to encrypt the ML data volumes, and to encrypt the model artifacts and data in Amazon S3.
B. Use SageMaker built-in transient keys to encrypt the ML data volumes. Enable default encryption for new Amazon Elastic Block Store (Amazon EBS) volumes.
C. Use customer managed keys in AWS Key Management Service (AWS KMS) to encrypt the ML data volumes, and to encrypt the model artifacts and data in Amazon S3.
D. Use AWS Security Token Service (AWS STS) to create temporary tokens to encrypt the ML storage volumes, and to encrypt the model artifacts and data in Amazon S3.

Answer: C

Explanation:
Amazon SageMaker supports encryption at rest for the ML storage volumes, the model artifacts, and the data in Amazon S3 using AWS Key Management Service (AWS KMS). AWS KMS is a service that allows customers to create and manage encryption keys that can be used to encrypt data. AWS KMS also provides an audit trail of key usage by logging key events to AWS CloudTrail. Customers can use either AWS managed keys or customer managed keys to encrypt their data. AWS managed keys are created and managed by AWS on behalf of the customer, while customer managed keys are created and managed by the customer. Customer managed keys offer more control and flexibility over the key policies, permissions, and rotation. Therefore, to meet the requirements of the company, the best option is to use customer managed keys in AWS KMS to encrypt the ML data volumes, and to

encrypt the model artifacts and data in Amazon S3.
The other options are not correct because:
Option A: AWS Cloud HSM is a service that provides hardware security modules (HSMs) to store and use encryption keys. AWS Cloud HSM is not integrated with Amazon SageMaker, and cannot be used to encrypt the ML data volumes, the model artifacts, or the data in Amazon S3. AWS Cloud HSM is more suitable for customers who need to meet strict compliance requirements or who need direct control over the HSMs.
Option B: SageMaker built-in transient keys are temporary keys that are used to encrypt the ML data volumes and are discarded immediately after encryption. These keys do not provide persistent encryption or logging of key usage. Enabling default encryption for new Amazon Elastic Block Store (Amazon EBS) volumes does not affect the ML data volumes, which are encrypted separately by SageMaker. Moreover, this option does not address the encryption of the model artifacts and data in Amazon S3.
Option D: AWS Security Token Service (AWS STS) is a service that provides temporary credentials to access AWS resources. AWS STS does not provide encryption keys or encryption services. AWS STS cannot be used to encrypt the ML storage volumes, the model artifacts, or the data in Amazon S3.
Reference:
Protect Data at Rest Using Encryption - Amazon SageMaker
What is AWS Key Management Service? - AWS Key Management Service
What is AWS CloudHSM? - AWS CloudHSM
What is AWS Security Token Service? - AWS Security Token Service

---

QUESTION 172
A machine learning specialist stores IoT soil sensor data in Amazon DynamoDB table and stores weather event data as JSON files in Amazon S3. The dataset in DynamoDB is 10 GB in size and the dataset in Amazon S3 is 5 GB in size. The specialist wants to train a model on this data to help predict soil moisture levels as a function of weather events using Amazon SageMaker.
Which solution will accomplish the necessary transformation to train the Amazon SageMaker model with the LEAST amount of administrative overhead?

A. Launch an Amazon EMR cluster. Create an Apache Hive external table for the DynamoDB table and S3 data. Join the Hive tables and write the results out to Amazon S3.
B. Crawl the data using AWS Glue crawlers. Write an AWS Glue ETL job that merges the two tables and writes the output to an Amazon Redshift cluster.
C. Enable Amazon DynamoDB Streams on the sensor table. Write an AWS Lambda function that consumes the stream and appends the results to the existing weather files in Amazon S3.
D. Crawl the data using AWS Glue crawlers. Write an AWS Glue ETL job that merges the two tables and writes the output in CSV format to Amazon S3.

Answer: D

Explanation:

The solution that will accomplish the necessary transformation to train the Amazon SageMaker model with the least amount of administrative overhead is to crawl the data using AWS Glue crawlers, write an AWS Glue ETL job that merges the two tables and writes the output in CSV format to Amazon S3. This solution leverages the serverless capabilities of AWS Glue to automatically discover the schema of the data sources, and to perform the data integration and transformation without requiring any cluster management or configuration. The output in CSV format is compatible with Amazon SageMaker and can be easily loaded into a training job. Reference: AWS Glue, Amazon SageMaker

---

QUESTION 173
A company sells thousands of products on a public website and wants to automatically identify products with potential durability problems. The company has 1.000 reviews with date, star rating, review text, review summary, and customer email fields, but many reviews are incomplete and have empty fields. Each review has already been labeled with the correct durability result.
A machine learning specialist must train a model to identify reviews expressing concerns over product durability. The first model needs to be trained and ready to review in 2 days.
What is the MOST direct approach to solve this problem within 2 days?

A. Train a custom classifier by using Amazon Comprehend.
B. Build a recurrent neural network (RNN) in Amazon SageMaker by using Gluon and Apache MXNet.
C. Train a built-in BlazingText model using Word2Vec mode in Amazon SageMaker.
D. Use a built-in seq2seq model in Amazon SageMaker.

Answer: A

Explanation:
The most direct approach to solve this problem within 2 days is to train a custom classifier by using Amazon Comprehend. Amazon Comprehend is a natural language processing (NLP) service that can analyze text and extract insights such as sentiment, entities, topics, and syntax. Amazon Comprehend also provides a custom classification feature that allows users to create and train a custom text classifier using their own labeled data. The custom classifier can then be used to categorize any text document into one or more custom classes. For this use case, the custom classifier can be trained to identify reviews that express concerns over product durability as a class, and use the star rating, review text, and review summary fields as input features. The custom classifier can be created and trained using the Amazon Comprehend console or API, and does not require any coding or machine learning expertise. The training process is fully managed and scalable, and can handle large and complex datasets. The custom classifier can be trained and ready to review in 2 days or less, depending on the size and quality of the dataset.
The other options are not the most direct approaches because:
Option B: Building a recurrent neural network (RNN) in Amazon SageMaker by using Gluon and Apache MXNet is a more complex and time-consuming approach that requires coding and machine learning skills. RNNs are a type of deep learning models that can process sequential data, such as

text, and learn long-term dependencies between tokens. Gluon is a high-level API for MXNet that simplifies the development of deep learning models. Amazon SageMaker is a fully managed service that provides tools and frameworks for building, training, and deploying machine learning models. However, to use this approach, the machine learning specialist would have to write custom code to preprocess the data, define the RNN architecture, train the model, and evaluate the results. This would likely take more than 2 days and involve more administrative overhead.
Option C: Training a built-in BlazingText model using Word2Vec mode in Amazon SageMaker is not a suitable approach for text classification. BlazingText is a built-in algorithm in Amazon SageMaker that provides highly optimized implementations of the Word2Vec and text classification algorithms. The Word2Vec algorithm is useful for generating word embeddings, which are dense vector representations of words that capture their semantic and syntactic similarities. However, word embeddings alone are not sufficient for text classification, as they do not account for the context and structure of the text documents. To use this approach, the machine learning specialist would have to combine the word embeddings with another classifier model, such as a logistic regression or a neural network, which would add more complexity and time to the solution.
Option D: Using a built-in seq2seq model in Amazon SageMaker is not a relevant approach for text classification. Seq2seq is a built-in algorithm in Amazon SageMaker that provides a sequence-tosequence framework for neural machine translation based on MXNet. Seq2seq is a supervised learning algorithm that can generate an output sequence of tokens given an input sequence of tokens, such as translating a sentence from one language to another. However, seq2seq is not designed for text classification, which requires assigning a label or a category to a text document, not generating another text sequence. To use this approach, the machine learning specialist would have to modify the seq2seq algorithm to fit the text classification task, which would be challenging and inefficient.
Reference:
Custom Classification - Amazon Comprehend
Build a Text Classification Model with Amazon Comprehend - AWS Machine Learning Blog
Recurrent Neural Networks - Gluon API
BlazingText Algorithm - Amazon SageMaker
Sequence-to-Sequence Algorithm - Amazon SageMaker

QUESTION 174
A company that runs an online library is implementing a chatbot using Amazon Lex to provide book recommendations based on category. This intent is fulfilled by an AWS Lambda function that queries an Amazon DynamoDB table for a list of book titles, given a particular category. For testing, there are only three categories implemented as the custom slot types: "comedy," "adventure, and "documentary.
A machine learning (ML) specialist notices that sometimes the request cannot be fulfilled because Amazon Lex cannot understand the category spoken by users with utterances such as "funny," "fun," and "humor." The ML specialist needs to fix the problem without changing the Lambda code or data in DynamoDB.
How should the ML specialist fix the problem?

A. Add the unrecognized words in the enumeration values list as new values in the slot type.
B. Create a new custom slot type, add the unrecognized words to this slot type as enumeration values, and use this slot type for the slot.
C. Use the AMAZON.SearchQuery built-in slot types for custom searches in the database.
D. Add the unrecognized words as synonyms in the custom slot type.

Answer: D

Explanation:
The best way to fix the problem without changing the Lambda code or data in DynamoDB is to add the unrecognized words as synonyms in the custom slot type. This way, Amazon Lex can resolve the synonyms to the corresponding slot values and pass them to the Lambda function. For example, if the slot type has a value œcomedy with synonyms œfunny , œfun , and œhumor , then any of these words entered by the user will be resolved to œcomedy and the Lambda function can query the DynamoDB table for the book titles in that category. Adding synonyms to the custom slot type can be done easily using the Amazon Lex console or API, and does not require any code changes.
The other options are not correct because:
Option A: Adding the unrecognized words in the enumeration values list as new values in the slot type would not fix the problem, because the Lambda function and the DynamoDB table are not aware of these new values. The Lambda function would not be able to query the DynamoDB table for the book titles in the new categories, and the request would still fail. Moreover, adding new values to the slot type would increase the complexity and maintenance of the chatbot, as the Lambda function and the DynamoDB table would have to be updated accordingly.
Option B: Creating a new custom slot type, adding the unrecognized words to this slot type as enumeration values, and using this slot type for the slot would also not fix the problem, for the same reasons as option
A. The Lambda function and the DynamoDB table would not be able to handle the new slot type and its values, and the request would still fail. Furthermore, creating a new slot type would require more effort and time than adding synonyms to the existing slot type.
Option C: Using the AMAZON.SearchQuery built-in slot types for custom searches in the database is not a suitable approach for this use case. The AMAZON.SearchQuery slot type is used to capture freeform user input that corresponds to a search query. However, this slot type does not perform any validation or resolution of the user input, and passes the raw input to the Lambda function. This means that the Lambda function would have to handle the logic of parsing and matching the user input to the DynamoDB table, which would require changing the Lambda code and adding more complexity to the solution.
Reference:
Custom slot type - Amazon Lex
Using Synonyms - Amazon Lex
Built-in Slot Types - Amazon Lex

QUESTION 175
A manufacturing company uses machine learning (ML) models to detect quality issues. The models use images that are taken of the company's product at the end of each production step. The company has thousands of machines at the production site that generate one image per second on average.
The company ran a successful pilot with a single manufacturing machine. For the pilot, ML specialists used an industrial PC that ran AWS IoT Greengrass with a long-running AWS Lambda function that uploaded the images to Amazon S3. The uploaded images invoked a Lambda function that was written in Python to perform inference by using an Amazon SageMaker endpoint that ran a custom model. The inference results were forwarded back to a web service that was hosted at the production site to prevent faulty products from being shipped.
The company scaled the solution out to all manufacturing machines by installing similarly configured industrial PCs on each production machine. However, latency for predictions increased beyond acceptable limits. Analysis shows that the internet connection is at its capacity limit.
How can the company resolve this issue MOST cost-effectively?

A. Set up a 10 Gbps AWS Direct Connect connection between the production site and the nearest AWS Region. Use the Direct Connect connection to upload the images. Increase the size of the instances and the number of instances that are used by the SageMaker endpoint.
B. Extend the long-running Lambda function that runs on AWS IoT Greengrass to compress the images and upload the compressed files to Amazon S3. Decompress the files by using a separate Lambda function that invokes the existing Lambda function to run the inference pipeline.
C. Use auto scaling for SageMaker. Set up an AWS Direct Connect connection between the production site and the nearest AWS Region. Use the Direct Connect connection to upload the images.
D. Deploy the Lambda function and the ML models onto the AWS IoT Greengrass core that is running on the industrial PCs that are installed on each machine. Extend the long-running Lambda function that runs on AWS IoT Greengrass to invoke the Lambda function with the captured images and run the inference on the edge component that forwards the results directly to the web service.

Answer: D

Explanation:
The best option is to deploy the Lambda function and the ML models onto the AWS IoT Greengrass core that is running on the industrial PCs that are installed on each machine. This way, the inference can be performed locally on the edge devices, without the need to upload the images to Amazon S3 and invoke the SageMaker endpoint. This will reduce the latency and the network bandwidth consumption. The long-running Lambda function can be extended to invoke the Lambda function with the captured images and run the inference on the edge component that forwards the results directly to the web service. This will also simplify the architecture and eliminate the dependency on the internet connection.
Option A is not cost-effective, as it requires setting up a 10 Gbps AWS Direct Connect connection and increasing the size and number of instances for the SageMaker endpoint. This will increase the

operational costs and complexity.
Option B is not optimal, as it still requires uploading the images to Amazon S3 and invoking the SageMaker endpoint. Compressing and decompressing the images will add additional processing overhead and latency.
Option C is not sufficient, as it still requires uploading the images to Amazon S3 and invoking the SageMaker endpoint. Auto scaling for SageMaker will help to handle the increased workload, but it will not reduce the latency or the network bandwidth consumption. Setting up an AWS Direct Connect connection will improve the network performance, but it will also increase the operational costs and complexity. Reference:
AWS IoT Greengrass
Deploying Machine Learning Models to Edge Devices
AWS Certified Machine Learning - Specialty Exam Guide

QUESTION 176
A data scientist is using an Amazon SageMaker notebook instance and needs to securely access data stored in a specific Amazon S3 bucket.
How should the data scientist accomplish this?

A. Add an S3 bucket policy allowing GetObject, PutObject, and ListBucket permissions to the Amazon SageMaker notebook ARN as principal.
B. Encrypt the objects in the S3 bucket with a custom AWS Key Management Service (AWS KMS) key that only the notebook owner has access to.
C. Attach the policy to the IAM role associated with the notebook that allows GetObject, PutObject, and ListBucket operations to the specific S3 bucket.
D. Use a script in a lifecycle configuration to configure the AWS CLI on the instance with an access key ID and secret.

Answer: C

Explanation:
The best way to securely access data stored in a specific Amazon S3 bucket from an Amazon SageMaker notebook instance is to attach a policy to the IAM role associated with the notebook that allows GetObject, PutObject, and ListBucket operations to the specific S3 bucket. This way, the notebook can use the AWS SDK or CLI to access the S3 bucket without exposing any credentials or requiring any additional configuration. This is also the recommended approach by AWS for granting access to S3 from SageMaker. Reference:
Amazon SageMaker Roles
Accessing Amazon S3 from a SageMaker Notebook Instance

QUESTION 177
A company is launching a new product and needs to build a mechanism to monitor comments about the company and its new product on social medi

a. The company needs to be able to evaluate the sentiment expressed in social media posts, and visualize trends and configure alarms based on various thresholds.
The company needs to implement this solution quickly, and wants to minimize the infrastructure and data science resources needed to evaluate the messages. The company already has a solution in place to collect posts and store them within an Amazon S3 bucket.
What services should the data science team use to deliver this solution?

A. Train a model in Amazon SageMaker by using the BlazingText algorithm to detect sentiment in the corpus of social media posts. Expose an endpoint that can be called by AWS Lambda. Trigger a Lambda function when posts are added to the S3 bucket to invoke the endpoint and record the sentiment in an Amazon DynamoDB table and in a custom Amazon CloudWatch metric. Use CloudWatch alarms to notify analysts of trends.
B. Train a model in Amazon SageMaker by using the semantic segmentation algorithm to model the semantic content in the corpus of social media posts. Expose an endpoint that can be called by AWS Lambda. Trigger a Lambda function when objects are added to the S3 bucket to invoke the endpoint and record the sentiment in an Amazon DynamoDB table. Schedule a second Lambda function to query recently added records and send an Amazon Simple Notification Service (Amazon SNS) notification to notify analysts of trends.
C. Trigger an AWS Lambda function when social media posts are added to the S3 bucket. Call Amazon Comprehend for each post to capture the sentiment in the message and record the sentiment in an Amazon DynamoDB table. Schedule a second Lambda function to query recently added records and send an Amazon Simple Notification Service (Amazon SNS) notification to notify analysts of trends.
D. Trigger an AWS Lambda function when social media posts are added to the S3 bucket. Call Amazon Comprehend for each post to capture the sentiment in the message and record the sentiment in a custom Amazon CloudWatch metric and in S3. Use CloudWatch alarms to notify analysts of trends.

Answer: D

Explanation:
The solution that uses Amazon Comprehend and Amazon CloudWatch is the most suitable for the given scenario. Amazon Comprehend is a natural language processing (NLP) service that can analyze text and extract insights such as sentiment, entities, topics, and syntax. Amazon CloudWatch is a monitoring and observability service that can collect and track metrics, create dashboards, and set alarms based on various thresholds. By using these services, the data science team can quickly and easily implement a solution to monitor the sentiment of social media posts without requiring much infrastructure or data science resources. The solution also meets the requirements of storing the sentiment in both S3 and CloudWatch, and using CloudWatch alarms to notify analysts of trends.
Reference:
Amazon Comprehend
Amazon CloudWatch

QUESTION 178
A bank wants to launch a low-rate credit promotion. The bank is located in a town that recently experienced economic hardship. Only some of the bank's customers were affected by the crisis, so the bank's credit team must identify which customers to target with the promotion. However, the credit team wants to make sure that loyal customers' full credit history is considered when the decision is made.
The bank's data science team developed a model that classifies account transactions and understands credit eligibility. The data science team used the XGBoost algorithm to train the model. The team used 7 years of bank transaction historical data for training and hyperparameter tuning over the course of several days.
The accuracy of the model is sufficient, but the credit team is struggling to explain accurately why the model denies credit to some customers. The credit team has almost no skill in data science.
What should the data science team do to address this issue in the MOST operationally efficient manner?

A. Use Amazon SageMaker Studio to rebuild the model. Create a notebook that uses the XGBoost training container to perform model training. Deploy the model at an endpoint. Enable Amazon SageMaker Model Monitor to store inferences. Use the inferences to create Shapley values that help explain model behavior. Create a chart that shows features and SHapley Additive exPlanations (SHAP) values to explain to the credit team how the features affect the model outcomes.
B. Use Amazon SageMaker Studio to rebuild the model. Create a notebook that uses the XGBoost training container to perform model training. Activate Amazon SageMaker Debugger, and configure it to calculate and collect Shapley values. Create a chart that shows features and SHapley Additive exPlanations (SHAP) values to explain to the credit team how the features affect the model outcomes.
C. Create an Amazon SageMaker notebook instance. Use the notebook instance and the XGBoost library to locally retrain the model. Use the plot_importance() method in the Python XGBoost interface to create a feature importance chart. Use that chart to explain to the credit team how the features affect the model outcomes.
D. Use Amazon SageMaker Studio to rebuild the model. Create a notebook that uses the XGBoost training container to perform model training. Deploy the model at an endpoint. Use Amazon SageMaker Processing to post-analyze the model and create a feature importance explainability chart automatically for the credit team.

Answer: A

Explanation:
The best option is to use Amazon SageMaker Studio to rebuild the model and deploy it at an endpoint. Then, use Amazon SageMaker Model Monitor to store inferences and use the inferences to create Shapley values that help explain model behavior. Shapley values are a way of attributing the contribution of each feature to the model output. They can help the credit team understand why the model makes certain decisions and how the features affect the model outcomes. A chart that shows

features and SHapley Additive exPlanations (SHAP) values can be created using the SHAP library in Python. This option is the most operationally efficient because it leverages the existing XGBoost training container and the built-in capabilities of Amazon SageMaker Model Monitor and SHAP library. Reference:
Amazon SageMaker Studio
Amazon SageMaker Model Monitor
SHAP library

QUESTION 179
A data science team is planning to build a natural language processing (NLP) application. The applications text preprocessing stage will include part-of-speech tagging and key phase extraction. The preprocessed text will be input to a custom classification algorithm that the data science team has already written and trained using Apache MXNet.
Which solution can the team build MOST quickly to meet these requirements?

A. Use Amazon Comprehend for the part-of-speech tagging, key phase extraction, and classification tasks.
B. Use an NLP library in Amazon SageMaker for the part-of-speech tagging. Use Amazon Comprehend for the key phase extraction. Use AWS Deep Learning Containers with Amazon SageMaker to build the custom classifier.
C. Use Amazon Comprehend for the part-of-speech tagging and key phase extraction tasks. Use Amazon SageMaker built-in Latent Dirichlet Allocation (LDA) algorithm to build the custom classifier.
D. Use Amazon Comprehend for the part-of-speech tagging and key phase extraction tasks. Use AWS Deep Learning Containers with Amazon SageMaker to build the custom classifier.

Answer: D

Explanation:
Amazon Comprehend is a natural language processing (NLP) service that can perform part-of-speech tagging and key phrase extraction tasks. AWS Deep Learning Containers are Docker images that are pre-installed with popular deep learning frameworks such as Apache MXNet. Amazon SageMaker is a fully managed service that can help build, train, and deploy machine learning models. Using Amazon Comprehend for the text preprocessing tasks and AWS Deep Learning Containers with Amazon SageMaker to build the custom classifier is the solution that can be built most quickly to meet the requirements.
Reference:
Amazon Comprehend
AWS Deep Learning Containers
Amazon SageMaker

QUESTION 180
A machine learning (ML) specialist must develop a classification model for a financial services

company. A domain expert provides the dataset, which is tabular with 10,000 rows and 1,020 features. During exploratory data analysis, the specialist finds no missing values and a small percentage of duplicate rows. There are correlation scores of > 0.9 for 200 feature pairs. The mean value of each feature is similar to its 50th percentile.
Which feature engineering strategy should the ML specialist use with Amazon SageMaker?

A. Apply dimensionality reduction by using the principal component analysis (PCA) algorithm.
B. Drop the features with low correlation scores by using a Jupyter notebook.
C. Apply anomaly detection by using the Random Cut Forest (RCF) algorithm.
D. Concatenate the features with high correlation scores by using a Jupyter notebook.

Answer: A

Explanation:
The best feature engineering strategy for this scenario is to apply dimensionality reduction by using the principal component analysis (PCA) algorithm. PCA is a technique that transforms a large set of correlated features into a smaller set of uncorrelated features called principal components. This can help reduce the complexity and noise in the data, improve the performance and interpretability of the model, and avoid overfitting. Amazon SageMaker provides a built-in PCA algorithm that can be used to perform dimensionality reduction on tabular data. The ML specialist can use Amazon SageMaker to train and deploy the PCA model, and then use the output of the PCA model as the input for the classification model.
Reference:
Dimensionality Reduction with Amazon SageMaker
Amazon SageMaker PCA Algorithm

QUESTION 181
A machine learning specialist needs to analyze comments on a news website with users across the globe. The specialist must find the most discussed topics in the comments that are in either English or Spanish.
What steps could be used to accomplish this task? (Choose two.)

A. Use an Amazon SageMaker BlazingText algorithm to find the topics independently from language. Proceed with the analysis.
B. Use an Amazon SageMaker seq2seq algorithm to translate from Spanish to English, if necessary. Use a SageMaker Latent Dirichlet Allocation (LDA) algorithm to find the topics.
C. Use Amazon Translate to translate from Spanish to English, if necessary. Use Amazon Comprehend topic modeling to find the topics.
D. Use Amazon Translate to translate from Spanish to English, if necessary. Use Amazon Lex to extract topics form the content.
E. Use Amazon Translate to translate from Spanish to English, if necessary. Use Amazon SageMaker Neural Topic Model (NTM) to find the topics.

Answer: C, E

Explanation:
To find the most discussed topics in the comments that are in either English or Spanish, the machine learning specialist needs to perform two steps: first, translate the comments from Spanish to English if necessary, and second, apply a topic modeling algorithm to the comments. The following options are valid ways to accomplish these steps using AWS services:
Option C: Use Amazon Translate to translate from Spanish to English, if necessary. Use Amazon Comprehend topic modeling to find the topics. Amazon Translate is a neural machine translation service that delivers fast, high-quality, and affordable language translation. Amazon Comprehend is a natural language processing (NLP) service that uses machine learning to find insights and relationships in text. Amazon Comprehend topic modeling is a feature that automatically organizes a collection of text documents into topics that contain commonly used words and phrases.
Option E: Use Amazon Translate to translate from Spanish to English, if necessary. Use Amazon SageMaker Neural Topic Model (NTM) to find the topics. Amazon SageMaker is a fully managed service that provides every developer and data scientist with the ability to build, train, and deploy machine learning (ML) models quickly. Amazon SageMaker Neural Topic Model (NTM) is an unsupervised learning algorithm that is used to organize a corpus of documents into topics that contain word groupings based on their statistical distribution.
The other options are not valid because:
Option A: Amazon SageMaker BlazingText algorithm is not a topic modeling algorithm, but a text classification and word embedding algorithm. It cannot find the topics independently from language, as different languages have different word distributions and semantics.
Option B: Amazon SageMaker seq2seq algorithm is not a translation algorithm, but a sequence-tosequence learning algorithm that can be used for tasks such as summarization, chatbot, and question answering. Amazon SageMaker Latent Dirichlet Allocation (LDA) algorithm is a topic modeling algorithm, but it requires the input documents to be in the same language and preprocessed into a bag-of-words format.
Option D: Amazon Lex is not a topic modeling algorithm, but a service for building conversational interfaces into any application using voice and text. It cannot extract topics from the content, but only intents and slots based on a predefined bot configuration. Reference:
Amazon Translate
Amazon Comprehend
Amazon SageMaker
Amazon SageMaker Neural Topic Model (NTM) Algorithm
Amazon SageMaker BlazingText
Amazon SageMaker Seq2Seq
Amazon SageMaker Latent Dirichlet Allocation (LDA) Algorithm
Amazon Lex

QUESTION 182
A machine learning (ML) specialist is administering a production Amazon SageMaker endpoint with model monitoring configured. Amazon SageMaker Model Monitor detects violations on the SageMaker endpoint, so the ML specialist retrains the model with the latest dataset. This dataset is statistically representative of the current production traffic. The ML specialist notices that even after deploying the new SageMaker model and running the first monitoring job, the SageMaker endpoint still has violations.
What should the ML specialist do to resolve the violations?

A. Manually trigger the monitoring job to re-evaluate the SageMaker endpoint traffic sample.
B. Run the Model Monitor baseline job again on the new training set. Configure Model Monitor to use the new baseline.
C. Delete the endpoint and recreate it with the original configuration.
D. Retrain the model again by using a combination of the original training set and the new training set.

Answer: B

Explanation:
The ML specialist should run the Model Monitor baseline job again on the new training set and configure Model Monitor to use the new baseline. This is because the baseline job computes the statistics and constraints for the data quality and model quality metrics, which are used to detect violations. If the training set changes, the baseline job should be updated accordingly to reflect the new distribution of the data and the model performance. Otherwise, the old baseline may not be representative of the current production traffic and may cause false alarms or miss violations. Reference:
Monitor data and model quality - Amazon SageMaker
Detecting and analyzing incorrect model predictions with Amazon SageMaker Model Monitor and Debugger | AWS Machine Learning Blog

---

QUESTION 183
A company supplies wholesale clothing to thousands of retail stores. A data scientist must create a model that predicts the daily sales volume for each item for each store. The data scientist discovers that more than half of the stores have been in business for less than 6 months. Sales data is highly consistent from week to week. Daily data from the database has been aggregated weekly, and weeks with no sales are omitted from the current dataset. Five years (100 MB) of sales data is available in Amazon S3.
Which factors will adversely impact the performance of the forecast model to be developed, and which actions should the data scientist take to mitigate them? (Choose two.)

A. Detecting seasonality for the majority of stores will be an issue. Request categorical data to relate new stores with similar stores that have more historical data.

B. The sales data does not have enough variance. Request external sales data from other industries to improve the model's ability to generalize.
C. Sales data is aggregated by week. Request daily sales data from the source database to enable building a daily model.
D. The sales data is missing zero entries for item sales. Request that item sales data from the source database include zero entries to enable building the model.
E. Only 100 MB of sales data is available in Amazon S3. Request 10 years of sales data, which would provide 200 MB of training data for the model.

Answer: C,D
The factors that will adversely impact the performance of the forecast model are:
Sales data is aggregated by week. This will reduce the granularity and resolution of the data, and make it harder to capture the daily patterns and variations in sales volume. The data scientist should request daily sales data from the source database to enable building a daily model, which will be more accurate and useful for the prediction task.
Sales data is missing zero entries for item sales. This will introduce bias and incompleteness in the data, and make it difficult to account for the items that have no demand or are out of stock. The data scientist should request that item sales data from the source database include zero entries to enable building the model, which will be more robust and realistic.
The other options are not valid because:
Detecting seasonality for the majority of stores will not be an issue, as sales data is highly consistent from week to week. Requesting categorical data to relate new stores with similar stores that have more historical data may not improve the model performance significantly, and may introduce unnecessary complexity and noise.
The sales data does not need to have more variance, as it reflects the actual demand and behavior of the customers. Requesting external sales data from other industries will not improve the models ability to generalize, but may introduce irrelevant and misleading information.
Only 100 MB of sales data is not a problem, as it is sufficient to train a forecast model with Amazon S3 and Amazon Forecast. Requesting 10 years of sales data will not provide much benefit, as it may contain outdated and obsolete information that does not reflect the current market trends and customer preferences.
Reference:
Amazon Forecast
Forecasting: Principles and Practice

QUESTION 184
An ecommerce company is automating the categorization of its products based on images. A data scientist has trained a computer vision model using the Amazon SageMaker image classification algorithm. The images for each product are classified according to specific product lines. The accuracy of the model is too low when categorizing new products. All of the product images have the same dimensions and are stored within an Amazon S3 bucket. The company wants to improve the model so it can be used for new products as soon as possible.

Which steps would improve the accuracy of the solution? (Choose three.)

A. Use the SageMaker semantic segmentation algorithm to train a new model to achieve improved accuracy.
B. Use the Amazon Rekognition DetectLabels API to classify the products in the dataset.
C. Augment the images in the dataset. Use open-source libraries to crop, resize, flip, rotate, and adjust the brightness and contrast of the images.
D. Use a SageMaker notebook to implement the normalization of pixels and scaling of the images. Store the new dataset in Amazon S3.
E. Use Amazon Rekognition Custom Labels to train a new model.
F. Check whether there are class imbalances in the product categories, and apply oversampling or undersampling as required. Store the new dataset in Amazon S3.

Answer: C, E, F

Explanation:
Option C is correct because augmenting the images in the dataset can help the model learn more features and generalize better to new products. Image augmentation is a common technique to increase the diversity and size of the training data.
Option E is correct because Amazon Rekognition Custom Labels can train a custom model to detect specific objects and scenes that are relevant to the business use case. It can also leverage the existing models from Amazon Rekognition that are trained on tens of millions of images across many categories.
Option F is correct because class imbalance can affect the performance and accuracy of the model, as it can cause the model to be biased towards the majority class and ignore the minority class. Applying oversampling or undersampling can help balance the classes and improve the models ability to learn from the data.
Option A is incorrect because the semantic segmentation algorithm is used to assign a label to every pixel in an image, not to classify the whole image into a category. Semantic segmentation is useful for applications such as autonomous driving, medical imaging, and satellite imagery analysis.
Option B is incorrect because the DetectLabels API is a general-purpose image analysis service that can detect objects, scenes, and concepts in an image, but it cannot be customized to the specific product lines of the ecommerce company. The DetectLabels API is based on the pre-trained models from Amazon Rekognition, which may not cover all the categories that the company needs.
Option D is incorrect because normalizing the pixels and scaling the images are preprocessing steps that should be done before training the model, not after. These steps can help improve the models convergence and performance, but they are not sufficient to increase the accuracy of the model on new products.
Reference:
: Image Augmentation - Amazon SageMaker
: Amazon Rekognition Custom Labels Features
: [Handling Imbalanced Datasets in Machine Learning]

: [Semantic Segmentation - Amazon SageMaker]
: [DetectLabels - Amazon Rekognition]
: [Image Classification - MXNet - Amazon SageMaker]
: [https://towardsdatascience.com/handling-imbalanced-datasets-in-machine-learning-7a0e84220f28]
: [https://docs.aws.amazon.com/sagemaker/latest/dg/semantic-segmentation.html]
: [https://docs.aws.amazon.com/rekognition/latest/dg/API_DetectLabels.html]
: [https://docs.aws.amazon.com/sagemaker/latest/dg/image-classification.html]
: [https://towardsdatascience.com/handling-imbalanced-datasets-in-machine-learning-7a0e84220f28]
: [https://docs.aws.amazon.com/sagemaker/latest/dg/semantic-segmentation.html]
: [https://docs.aws.amazon.com/rekognition/latest/dg/API_DetectLabels.html]
: [https://docs.aws.amazon.com/sagemaker/latest/dg/image-classification.html]
: [https://towardsdatascience.com/handling-imbalanced-datasets-in-machine-learning-7a0e84220f28]
: [https://docs.aws.amazon.com/sagemaker/latest/dg/semantic-segmentation.html]
: [https://docs.aws.amazon.com/rekognition/latest/dg/API_DetectLabels.html]
: [https://docs.aws.amazon.com/sagemaker/latest/dg/image-classification.html]

QUESTION 185

A data scientist is training a text classification model by using the Amazon SageMaker built-in BlazingText algorithm. There are 5 classes in the dataset, with 300 samples for category A, 292 samples for category B, 240 samples for category C, 258 samples for category D, and 310 samples for category E.

The data scientist shuffles the data and splits off 10% for testing. After training the model, the data scientist generates confusion matrices for the training and test sets.

Training data confusion matrix

		Predicted class					
		A	**B**	**C**	**D**	**E**	**Total**
True class	**A**	270	0	0	0	0	270
	B	1	260	0	0	2	263
	C	0	0	111	100	5	216
	D	4	3	132	92	1	232
	E	0	0	2	3	274	279
	Total	275	263	245	195	282	1260

Test data confusion matrix

		Predicted class					
		A	**B**	**C**	**D**	**E**	**Total**
True class	**A**	9	1	0	0	0	10
	B	2	25	0	2	0	29
	C	10	2	11	10	1	34
	D	1	0	12	14	0	27
	E	9	1	4	1	25	40
	Total	31	29	27	27	26	140

What could the data scientist conclude form these results?

A. Classes C and D are too similar.
B. The dataset is too small for holdout cross-validation.
C. The data distribution is skewed.
D. The model is overfitting for classes B and E.

Answer: D

Explanation:
A confusion matrix is a matrix that summarizes the performance of a machine learning model on a set of test data. It displays the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) produced by the model on the test data1. For multi-class classification, the matrix shape will be equal to the number of classes i.e for n classes it will be nXn1. The diagonal values represent the number of correct predictions for each class, and the off-diagonal values represent the number of incorrect predictions for each class1.
The BlazingText algorithm is a proprietary machine learning algorithm for forecasting time series using causal convolutional neural networks (CNNs). BlazingText works best with large datasets containing hundreds of time series. It accepts item metadata, and is the only Forecast algorithm that accepts related time series data without future values2.
From the confusion matrices for the training and test sets, we can observe the following:
The model has a high accuracy on the training set, as most of the diagonal values are high and the off-diagonal values are low. This means that the model is able to learn the patterns and features of the training data well.
However, the model has a lower accuracy on the test set, as some of the diagonal values are lower and some of the off-diagonal values are higher. This means that the model is not able to generalize well to the unseen data and makes more errors.
The model has a particularly high error rate for classes B and E on the test set, as the values of M_22 and M_55 are much lower than the values of M_12, M_21, M_15, M_25, M_51, and M_52. This means that the model is confusing classes B and E with other classes more often than it should.
The model has a relatively low error rate for classes A, C, and D on the test set, as the values of M_11, M_33, and M_44 are high and the values of M_13, M_14, M_23, M_24, M_31, M_32, M_34, M_41, M_42, and M_43 are low. This means that the model is able to distinguish classes A, C, and D from other classes well.
These results indicate that the model is overfitting for classes B and E, meaning that it is memorizing the specific features of these classes in the training data, but failing to capture the general features that are applicable to the test data. Overfitting is a common problem in machine learning, where the model performs well on the training data, but poorly on the test data3. Some possible causes of overfitting are:
The model is too complex or has too many parameters for the given data. This makes the model flexible enough to fit the noise and outliers in the training data, but reduces its ability to generalize to new data.
The data is too small or not representative of the population. This makes the model learn from a

limited or biased sample of data, but fails to capture the variability and diversity of the population. The data is imbalanced or skewed. This makes the model learn from a disproportionate or uneven distribution of data, but fails to account for the minority or rare classes.
Some possible solutions to prevent or reduce overfitting are:
Simplify the model or use regularization techniques. This reduces the complexity or the number of parameters of the model, and prevents it from fitting the noise and outliers in the data. Regularization techniques, such as L1 or L2 regularization, add a penalty term to the loss function of the model, which shrinks the weights of the model and reduces overfitting3.
Increase the size or diversity of the data. This provides more information and examples for the model to learn from, and increases its ability to generalize to new data. Data augmentation techniques, such as rotation, flipping, cropping, or noise addition, can generate new data from the existing data by applying some transformations3.
Balance or resample the data. This adjusts the distribution or the frequency of the data, and ensures that the model learns from all classes equally. Resampling techniques, such as oversampling or undersampling, can create a balanced dataset by increasing or decreasing the number of samples for each class3.
Reference:
Confusion Matrix in Machine Learning - GeeksforGeeks
BlazingText algorithm - Amazon SageMaker
Overfitting and Underfitting in Machine Learning - GeeksforGeeks

---

QUESTION 186
A company that manufactures mobile devices wants to determine and calibrate the appropriate sales price for its devices. The company is collecting the relevant data and is determining data features that it can use to train machine learning (ML) models. There are more than 1,000 features, and the company wants to determine the primary features that contribute to the sales price.
Which techniques should the company use for feature selection? (Choose three.)

A. Data scaling with standardization and normalization
B. Correlation plot with heat maps
C. Data binning
D. Univariate selection
E. Feature importance with a tree-based classifier
F. Data augmentation

Answer: B, D, E

Explanation:
Feature selection is the process of selecting a subset of extracted features that are relevant and contribute to minimizing the error rate of a trained model. Some techniques for feature selection are:

Correlation plot with heat maps: This technique visualizes the correlation between features using a color-coded matrix. Features that are highly correlated with each other or with the target variable can be identified and removed to reduce redundancy and noise.
Univariate selection: This technique evaluates each feature individually based on a statistical test, such as chi-square, ANOVA, or mutual information, and selects the features that have the highest scores or p-values. This technique is simple and fast, but it does not consider the interactions between features.
Feature importance with a tree-based classifier: This technique uses a tree-based classifier, such as random forest or gradient boosting, to rank the features based on their importance in splitting the nodes. Features that have low importance scores can be dropped from the model. This technique can capture the non-linear relationships and interactions between features.
The other options are not techniques for feature selection, but rather for feature engineering, which is the process of creating, transforming, or extracting features from the original data. Feature engineering can improve the performance and interpretability of the model, but it does not reduce the number of features.
Data scaling with standardization and normalization: This technique transforms the features to have a common scale, such as zero mean and unit variance, or a range between 0 and 1. This technique can help some algorithms, such as k-means or logistic regression, to converge faster and avoid numerical instability, but it does not change the number of features.
Data binning: This technique groups the continuous features into discrete bins or categories based on some criteria, such as equal width, equal frequency, or clustering. This technique can reduce the noise and outliers in the data, and also create ordinal or nominal features that can be used for some algorithms, such as decision trees or naive Bayes, but it does not reduce the number of features.
Data augmentation: This technique generates new data from the existing data by applying some transformations, such as rotation, flipping, cropping, or noise addition. This technique can increase the size and diversity of the data, and help prevent overfitting, but it does not reduce the number of features.
Reference:
Feature engineering - Machine Learning Lens
Amazon SageMaker Autopilot now provides feature selection and the ability to change data types while creating an AutoML experiment
Feature Selection in Machine Learning | Baeldung on Computer Science
Feature Selection in Machine Learning: An easy Introduction

QUESTION 187
A power company wants to forecast future energy consumption for its customers in residential properties and commercial business properties. Historical power consumption data for the last 10 years is available. A team of data scientists who performed the initial data analysis and feature selection will include the historical power consumption data and data such as weather, number of individuals on the property, and public holidays.
The data scientists are using Amazon Forecast to generate the forecasts.
Which algorithm in Forecast should the data scientists use to meet these requirements?

A. Autoregressive Integrated Moving Average (AIRMA)
B. Exponential Smoothing (ETS)
C. Convolutional Neural Network - Quantile Regression (CNN-QR)
D. Prophet

Answer: C

Explanation:
CNN-QR is a proprietary machine learning algorithm for forecasting time series using causal convolutional neural networks (CNNs). CNN-QR works best with large datasets containing hundreds of time series. It accepts item metadata, and is the only Forecast algorithm that accepts related time series data without future values. In this case, the power company has historical power consumption data for the last 10 years, which is a large dataset with multiple time series. The data also includes related data such as weather, number of individuals on the property, and public holidays, which can be used as item metadata or related time series data. Therefore, CNN-QR is the most suitable algorithm for this scenario. Reference: Amazon Forecast Algorithms, Amazon Forecast CNN-QR

QUESTION 188
A company wants to use automatic speech recognition (ASR) to transcribe messages that are less than 60 seconds long from a voicemail-style application. The company requires the correct identification of 200 unique product names, some of which have unique spellings or pronunciations. The company has 4,000 words of Amazon SageMaker Ground Truth voicemail transcripts it can use to customize the chosen ASR model. The company needs to ensure that everyone can update their customizations multiple times each hour.
Which approach will maximize transcription accuracy during the development phase?

A. Use a voice-driven Amazon Lex bot to perform the ASR customization. Create customer slots within the bot that specifically identify each of the required product names. Use the Amazon Lex synonym mechanism to provide additional variations of each product name as mis-transcriptions are identified in development.
B. Use Amazon Transcribe to perform the ASR customization. Analyze the word confidence scores in the transcript, and automatically create or update a custom vocabulary file with any word that has a confidence score below an acceptable threshold value. Use this updated custom vocabulary file in all future transcription tasks.
C. Create a custom vocabulary file containing each product name with phonetic pronunciations, and use it with Amazon Transcribe to perform the ASR customization. Analyze the transcripts and manually update the custom vocabulary file to include updated or additional entries for those names that are not being correctly identified.
D. Use the audio transcripts to create a training dataset and build an Amazon Transcribe custom language model. Analyze the transcripts and update the training dataset with a manually corrected version of transcripts where product names are not being transcribed correctly. Create an updated

custom language model.

Answer: C

Explanation:
The best approach to maximize transcription accuracy during the development phase is to create a custom vocabulary file containing each product name with phonetic pronunciations, and use it with Amazon Transcribe to perform the ASR customization. A custom vocabulary is a list of words and phrases that are likely to appear in your audio input, along with optional information about how to pronounce them. By using a custom vocabulary, you can improve the transcription accuracy of domain-specific terms, such as product names, that may not be recognized by the general vocabulary of Amazon Transcribe. You can also analyze the transcripts and manually update the custom vocabulary file to include updated or additional entries for those names that are not being correctly identified.
The other options are not as effective as option C for the following reasons:
Option A is not suitable because Amazon Lex is a service for building conversational interfaces, not for transcribing voicemail messages. Amazon Lex also has a limit of 100 slots per bot, which is not enough to accommodate the 200 unique product names required by the company.
Option B is not optimal because it relies on the word confidence scores in the transcript, which may not be accurate enough to identify all the mis-transcribed product names. Moreover, automatically creating or updating a custom vocabulary file may introduce errors or inconsistencies in the pronunciation or display of the words.
Option D is not feasible because it requires a large amount of training data to build a custom language model. The company only has 4,000 words of Amazon SageMaker Ground Truth voicemail transcripts, which is not enough to train a robust and reliable custom language model. Additionally, creating and updating a custom language model is a time-consuming and resource-intensive process, which may not be suitable for the development phase where frequent changes are expected.
Reference:
Amazon Transcribe " Custom Vocabulary
Amazon Transcribe " Custom Language Models
[Amazon Lex " Limits]

QUESTION 189
A company is building a demand forecasting model based on machine learning (ML). In the development stage, an ML specialist uses an Amazon SageMaker notebook to perform feature engineering during work hours that consumes low amounts of CPU and memory resources. A data engineer uses the same notebook to perform data preprocessing once a day on average that requires very high memory and completes in only 2 hours. The data preprocessing is not configured to use GPU. All the processes are running well on an ml.m5.4xlarge notebook instance.
The company receives an AWS Budgets alert that the billing for this month exceeds the allocated budget.
Which solution will result in the MOST cost savings?

A. Change the notebook instance type to a memory optimized instance with the same vCPU number as the ml.m5.4xlarge instance has. Stop the notebook when it is not in use. Run both data preprocessing and feature engineering development on that instance.
B. Keep the notebook instance type and size the same. Stop the notebook when it is not in use. Run data preprocessing on a P3 instance type with the same memory as the ml.m5.4xlarge instance by using Amazon SageMaker Processing.
C. Change the notebook instance type to a smaller general-purpose instance. Stop the notebook when it is not in use. Run data preprocessing on an ml. r5 instance with the same memory size as the ml.m5.4xlarge instance by using Amazon SageMaker Processing.
D. Change the notebook instance type to a smaller general-purpose instance. Stop the notebook when it is not in use. Run data preprocessing on an R5 instance with the same memory size as the ml.m5.4xlarge instance by using the Reserved Instance option.

Answer: C

Explanation:
The best solution to reduce the cost of the notebook instance and the data preprocessing job is to change the notebook instance type to a smaller general-purpose instance, stop the notebook when it is not in use, and run data preprocessing on an ml.r5 instance with the same memory size as the ml.m5.4xlarge instance by using Amazon SageMaker Processing. This solution will result in the most cost savings because:
Changing the notebook instance type to a smaller general-purpose instance will reduce the hourly cost of running the notebook, since the feature engineering development does not require high CPU and memory resources. For example, an ml.t3.medium instance costs $0.0464 per hour, while an ml.m5.4xlarge instance costs $0.888 per hour1.
Stopping the notebook when it is not in use will also reduce the cost, since the notebook will only incur charges when it is running. For example, if the notebook is used for 8 hours per day, 5 days per week, then stopping it when it is not in use will save about 76% of the monthly cost compared to leaving it running all the time2.
Running data preprocessing on an ml.r5 instance with the same memory size as the ml.m5.4xlarge instance by using Amazon SageMaker Processing will reduce the cost of the data preprocessing job, since the ml.r5 instance is optimized for memory-intensive workloads and has a lower cost per GB of memory than the ml.m5 instance. For example, an ml.r5.4xlarge instance has 128 GB of memory and costs $1.008 per hour, while an ml.m5.4xlarge instance has 64 GB of memory and costs $0.888 per hour1. Therefore, the ml.r5.4xlarge instance can process the same amount of data in half the time and at a lower cost than the ml.m5.4xlarge instance. Moreover, using Amazon SageMaker Processing will allow the data preprocessing job to run on a separate, fully managed infrastructure that can be scaled up or down as needed, without affecting the notebook instance.
The other options are not as effective as option C for the following reasons:
Option A is not optimal because changing the notebook instance type to a memory optimized instance with the same vCPU number as the ml.m5.4xlarge instance has will not reduce the cost of

the notebook, since the memory optimized instances have a higher cost per vCPU than the generalpurpose instances. For example, an ml.r5.4xlarge instance has 16 vCPUs and costs $1.008 per hour, while an ml.m5.4xlarge instance has 16 vCPUs and costs $0.888 per hour1. Moreover, running both data preprocessing and feature engineering development on the same instance will not take advantage of the scalability and flexibility of Amazon SageMaker Processing.
Option B is not suitable because running data preprocessing on a P3 instance type with the same memory as the ml.m5.4xlarge instance by using Amazon SageMaker Processing will not reduce the cost of the data preprocessing job, since the P3 instance type is optimized for GPU-based workloads and has a higher cost per GB of memory than the ml.m5 or ml.r5 instance types. For example, an ml.p3.2xlarge instance has 61 GB of memory and costs $3.06 per hour, while an ml.m5.4xlarge instance has 64 GB of memory and costs $0.888 per hour1. Moreover, the data preprocessing job does not require GPU, so using a P3 instance type will be wasteful and inefficient.
Option D is not feasible because running data preprocessing on an R5 instance with the same memory size as the ml.m5.4xlarge instance by using the Reserved Instance option will not reduce the cost of the data preprocessing job, since the Reserved Instance option requires a commitment to a consistent amount of usage for a period of 1 or 3 years3. However, the data preprocessing job only runs once a day on average and completes in only 2 hours, so it does not have a consistent or predictable usage pattern. Therefore, using the Reserved Instance option will not provide any cost savings and may incur additional charges for unused capacity.
Reference:
Amazon SageMaker Pricing
Manage Notebook Instances - Amazon SageMaker
Amazon EC2 Pricing - Reserved Instances

QUESTION 190
A machine learning specialist is developing a regression model to predict rental rates from rental listings. A variable named Wall_Color represents the most prominent exterior wall color of the property. The following is the sample data, excluding all other variables:

Property_ID	Wall_Color
1000	Red
1001	White
1002	Green

The specialist chose a model that needs numerical input data.
Which feature engineering approaches should the specialist use to allow the regression model to learn from the Wall_Color data? (Choose two.)

A. Apply integer transformation and set Red = 1, White = 5, and Green = 10.
B. Add new columns that store one-hot representation of colors.
C. Replace the color name string by its length.
D. Create three columns to encode the color in RGB format.

E. Replace each color name by its training set frequency.

Answer: B, D

Explanation:
In this scenario, the specialist should use one-hot encoding and RGB encoding to allow the regression model to learn from the Wall_Color data. One-hot encoding is a technique used to convert categorical data into numerical data. It creates new columns that store one-hot representation of colors. For example, a variable named color has three categories: red, green, and blue. After one-hot encoding, the new variables should be like this:

color_red	color_green	color_blue
1	0	0
0	1	0
0	0	1

One-hot encoding can capture the presence or absence of a color, but it cannot capture the intensity or hue of a color. RGB encoding is a technique used to represent colors in a digital image. It creates three columns to encode the color in RGB format. For example, a variable named color has three categories: red, green, and blue. After RGB encoding, the new variables should be like this:

color_R	color_G	color_B
255	0	0
0	255	0
0	0	255

RGB encoding can capture the intensity and hue of a color, but it may also introduce correlation among the three columns. Therefore, using both one-hot encoding and RGB encoding can provide more information to the regression model than using either one alone.
Reference:
Feature Engineering for Categorical Data
How to Perform Feature Selection with Categorical Data

QUESTION 191
A data scientist is working on a public sector project for an urban traffic system. While studying the traffic patterns, it is clear to the data scientist that the traffic behavior at each light is correlated, subject to a small stochastic error term. The data scientist must model the traffic behavior to analyze the traffic patterns and reduce congestion.
How will the data scientist MOST effectively model the problem?

A. The data scientist should obtain a correlated equilibrium policy by formulating this problem as a multi-agent reinforcement learning problem.
B. The data scientist should obtain the optimal equilibrium policy by formulating this problem as a single-agent reinforcement learning problem.
C. Rather than finding an equilibrium policy, the data scientist should obtain accurate predictors of traffic flow by using historical data through a supervised learning approach.
D. Rather than finding an equilibrium policy, the data scientist should obtain accurate predictors of traffic flow by using unlabeled simulated data representing the new traffic patterns in the city and applying an unsupervised learning approach.

Answer: A

Explanation:
The data scientist should obtain a correlated equilibrium policy by formulating this problem as a multi-agent reinforcement learning problem. This is because:
Multi-agent reinforcement learning (MARL) is a subfield of reinforcement learning that deals with learning and coordination of multiple agents that interact with each other and the environment 1. MARL can be applied to problems that involve distributed decision making, such as traffic signal control, where each traffic light can be modeled as an agent that observes the traffic state and chooses an action (e.g., changing the signal phase) to optimize a reward function (e.g., minimizing the delay or congestion) 2.
A correlated equilibrium is a solution concept in game theory that generalizes the notion of Nash equilibrium. It is a probability distribution over the joint actions of the agents that satisfies the following condition: no agent can improve its expected payoff by deviating from the distribution, given that it knows the distribution and the actions of the other agents 3. A correlated equilibrium can capture the correlation among the agents actions, which is useful for modeling the traffic behavior at each light that is subject to a small stochastic error term.
A correlated equilibrium policy is a policy that induces a correlated equilibrium in a MARL setting. It can be obtained by using various methods, such as policy gradient, actor-critic, or Q-learning algorithms, that can learn from the feedback of the environment and the communication among the agents 4. A correlated equilibrium policy can achieve a better performance than a Nash equilibrium policy, which assumes that the agents act independently and ignore the correlation among their actions 5.
Therefore, by obtaining a correlated equilibrium policy by formulating this problem as a MARL problem, the data scientist can most effectively model the traffic behavior and reduce congestion.

Reference:
Multi-Agent Reinforcement Learning
Multi-Agent Reinforcement Learning for Traffic Signal Control: A Survey
Correlated Equilibrium
Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments
Correlated Q-Learning

QUESTION 192
A data scientist is using the Amazon SageMaker Neural Topic Model (NTM) algorithm to build a model that recommends tags from blog posts. The raw blog post data is stored in an Amazon S3 bucket in JSON format. During model evaluation, the data scientist discovered that the model recommends certain stopwords such as "a," "an, and "the" as tags to certain blog posts, along with a few rare words that are present only in certain blog entries. After a few iterations of tag review with the content team, the data scientist notices that the rare words are unusual but feasible. The data scientist also must ensure that the tag recommendations of the generated model do not include the stopwords.
What should the data scientist do to meet these requirements?

A. Use the Amazon Comprehend entity recognition API operations. Remove the detected words from the blog post data. Replace the blog post data source in the S3 bucket.
B. Run the SageMaker built-in principal component analysis (PCA) algorithm with the blog post data from the S3 bucket as the data source. Replace the blog post data in the S3 bucket with the results of the training job.
C. Use the SageMaker built-in Object Detection algorithm instead of the NTM algorithm for the training job to process the blog post data.
D. Remove the stop words from the blog post data by using the Count Vectorizer function in the scikit-learn library. Replace the blog post data in the S3 bucket with the results of the vectorizer.

Answer: D

Explanation:
The data scientist should remove the stop words from the blog post data by using the Count Vectorizer function in the scikit-learn library, and replace the blog post data in the S3 bucket with the results of the vectorizer. This is because:
The Count Vectorizer function is a tool that can convert a collection of text documents to a matrix of token counts 1. It also enables the pre-processing of text data prior to generating the vector representation, such as removing accents, converting to lowercase, and filtering out stop words 1. By using this function, the data scientist can remove the stop words such as œa, œan, and œthe from the blog post data, and obtain a numerical representation of the text that can be used as input for the NTM algorithm.
The NTM algorithm is a neural network-based topic modeling technique that can learn latent topics from a corpus of documents 2. It can be used to recommend tags from blog posts by finding the most

probable topics for each document, and ranking the words associated with each topic 3. However, the NTM algorithm does not perform any text pre-processing by itself, so it relies on the quality of the input data. Therefore, the data scientist should replace the blog post data in the S3 bucket with the results of the vectorizer, to ensure that the NTM algorithm does not include the stop words in the tag recommendations.
The other options are not suitable for the following reasons:
Option A is not relevant because the Amazon Comprehend entity recognition API operations are used to detect and extract named entities from text, such as people, places, organizations, dates, etc4. This is not the same as removing stop words, which are common words that do not carry much meaning or information. Moreover, removing the detected entities from the blog post data may reduce the quality and diversity of the tag recommendations, as some entities may be relevant and useful as tags.
Option B is not optimal because the SageMaker built-in principal component analysis (PCA) algorithm is used to reduce the dimensionality of a dataset by finding the most important features that capture the maximum amount of variance in the data 5. This is not the same as removing stop words, which are words that have low variance and high frequency in the data. Moreover, replacing the blog post data in the S3 bucket with the results of the PCA algorithm may not be compatible with the input format expected by the NTM algorithm, which requires a bag-of-words representation of the text 2.
Option C is not suitable because the SageMaker built-in Object Detection algorithm is used to detect and localize objects in images 6. This is not related to the task of recommending tags from blog posts, which are text documents. Moreover, using the Object Detection algorithm instead of the NTM algorithm would require a different type of input data (images instead of text), and a different type of output data (bounding boxes and labels instead of topics and words).
Reference:
Neural Topic Model (NTM) Algorithm
Introduction to the Amazon SageMaker Neural Topic Model
Amazon Comprehend - Entity Recognition
sklearn.feature_extraction.text.CountVectorizer
Principal Component Analysis (PCA) Algorithm
Object Detection Algorithm

QUESTION 193
A company wants to create a data repository in the AWS Cloud for machine learning (ML) projects. The company wants to use AWS to perform complete ML lifecycles and wants to use Amazon S3 for the data storage. All of the companys data currently resides on premises and is 40 Đ¢Đ’ in size.
The company wants a solution that can transfer and automatically update data between the onpremises object storage and Amazon S3. The solution must support encryption, scheduling, monitoring, and data integrity validation.
Which solution meets these requirements?

A. Use the S3 sync command to compare the source S3 bucket and the destination S3 bucket.

Determine which source files do not exist in the destination S3 bucket and which source files were modified.
B. Use AWS Transfer for FTPS to transfer the files from the on-premises storage to Amazon S3.
C. Use AWS DataSync to make an initial copy of the entire dataset. Schedule subsequent incremental transfers of changing data until the final cutover from on premises to AWS.
D. Use S3 Batch Operations to pull data periodically from the on-premises storage. Enable S3 Versioning on the S3 bucket to protect against accidental overwrites.

Answer: C

Explanation:
The best solution to meet the requirements of the company is to use AWS DataSync to make an initial copy of the entire dataset, and schedule subsequent incremental transfers of changing data until the final cutover from on premises to AWS. This is because:
AWS DataSync is an online data movement and discovery service that simplifies data migration and helps you quickly, easily, and securely transfer your file or object data to, from, and between AWS storage services 1. AWS DataSync can copy data between on-premises object storage and Amazon S3, and also supports encryption, scheduling, monitoring, and data integrity validation 1.
AWS DataSync can make an initial copy of the entire dataset by using a DataSync agent, which is a software appliance that connects to your on-premises storage and manages the data transfer to AWS 2. The DataSync agent can be deployed as a virtual machine (VM) on your existing hypervisor, or as an Amazon EC2 instance in your AWS account 2.
AWS DataSync can schedule subsequent incremental transfers of changing data by using a task, which is a configuration that specifies the source and destination locations, the options for the transfer, and the schedule for the transfer 3. You can create a task to run once or on a recurring schedule, and you can also use filters to include or exclude specific files or objects based on their names or prefixes 3.
AWS DataSync can perform the final cutover from on premises to AWS by using a sync task, which is a type of task that synchronizes the data in the source and destination locations 4. A sync task transfers only the data that has changed or that doesnt exist in the destination, and also deletes any files or objects from the destination that were deleted from the source since the last sync 4.
Therefore, by using AWS DataSync, the company can create a data repository in the AWS Cloud for machine learning projects, and use Amazon S3 for the data storage, while meeting the requirements of encryption, scheduling, monitoring, and data integrity validation.
Reference:
Data Transfer Service - AWS DataSync
Deploying a DataSync Agent
Creating a Task
Syncing Data with AWS DataSync

QUESTION 194
A company has video feeds and images of a subway train station. The company wants to create a

deep learning model that will alert the station manager if any passenger crosses the yellow safety line when there is no train in the station. The alert will be based on the video feeds. The company wants the model to detect the yellow line, the passengers who cross the yellow line, and the trains in the video feeds. This task requires labeling. The video data must remain confidential.
A data scientist creates a bounding box to label the sample data and uses an object detection model. However, the object detection model cannot clearly demarcate the yellow line, the passengers who cross the yellow line, and the trains.
Which labeling approach will help the company improve this model?

A. Use Amazon Rekognition Custom Labels to label the dataset and create a custom Amazon Rekognition object detection model. Create a private workforce. Use Amazon Augmented AI (Amazon A2I) to review the low-confidence predictions and retrain the custom Amazon Rekognition model.
B. Use an Amazon SageMaker Ground Truth object detection labeling task. Use Amazon Mechanical Turk as the labeling workforce.
C. Use Amazon Rekognition Custom Labels to label the dataset and create a custom Amazon Rekognition object detection model. Create a workforce with a third-party AWS Marketplace vendor. Use Amazon Augmented AI (Amazon A2I) to review the low-confidence predictions and retrain the custom Amazon Rekognition model.
D. Use an Amazon SageMaker Ground Truth semantic segmentation labeling task. Use a private workforce as the labeling workforce.

Answer: D

Explanation:

QUESTION 195
A data engineer at a bank is evaluating a new tabular dataset that includes customer dat
a. The data engineer will use the customer data to create a new model to predict customer behavior. After creating a correlation matrix for the variables, the data engineer notices that many of the 100 features are highly correlated with each other.
Which steps should the data engineer take to address this issue? (Choose two.)

A. Use a linear-based algorithm to train the model.
B. Apply principal component analysis (PCA).
C. Remove a portion of highly correlated features from the dataset.
D. Apply min-max feature scaling to the dataset.
E. Apply one-hot encoding category-based variables.

Answer: B, C

Explanation:

B) Apply principal component analysis (PCA): PCA is a technique that reduces the dimensionality of a dataset by transforming the original features into a smaller set of new features that capture most of the variance in the data. PCA can help address the issue of multicollinearity, which occurs when some features are highly correlated with each other and can cause problems for some machine learning algorithms. By applying PCA, the data engineer can reduce the number of features and remove the redundancy in the data.
C) Remove a portion of highly correlated features from the dataset: Another way to deal with multicollinearity is to manually remove some of the features that are highly correlated with each other. This can help simplify the model and avoid overfitting. The data engineer can use the correlation matrix to identify the features that have a high correlation coefficient (e.g., above 0.8 or below -0.8) and remove one of them from the dataset. Reference: =
Principal Component Analysis: This is a document from AWS that explains what PCA is, how it works, and how to use it with Amazon SageMaker.
Multicollinearity: This is a document from AWS that describes what multicollinearity is, how to detect it, and how to deal with it.

QUESTION 196
A company is building a new version of a recommendation engine. Machine learning (ML) specialists need to keep adding new data from users to improve personalized recommendations. The ML specialists gather data from the users interactions on the platform and from sources such as external websites and social media.
The pipeline cleans, transforms, enriches, and compresses terabytes of data daily, and this data is stored in Amazon S3. A set of Python scripts was coded to do the job and is stored in a large Amazon EC2 instance. The whole process takes more than 20 hours to finish, with each script taking at least an hour. The company wants to move the scripts out of Amazon EC2 into a more managed solution that will eliminate the need to maintain servers.
Which approach will address all of these requirements with the LEAST development effort?

A. Load the data into an Amazon Redshift cluster. Execute the pipeline by using SQL. Store the results in Amazon S3.
B. Load the data into Amazon DynamoDB. Convert the scripts to an AWS Lambda function. Execute the pipeline by triggering Lambda executions. Store the results in Amazon S3.
C. Create an AWS Glue job. Convert the scripts to PySpark. Execute the pipeline. Store the results in Amazon S3.
D. Create a set of individual AWS Lambda functions to execute each of the scripts. Build a step function by using the AWS Step Functions Data Science SDK. Store the results in Amazon S3.

Answer: C

Explanation:
The best approach to address all of the requirements with the least development effort is to create an AWS Glue job, convert the scripts to PySpark, execute the pipeline, and store the results in

Amazon S3. This is because:
AWS Glue is a fully managed extract, transform, and load (ETL) service that makes it easy to prepare and load data for analytics 1. AWS Glue can run Python and Scala scripts to process data from various sources, such as Amazon S3, Amazon DynamoDB, Amazon Redshift, and more 2. AWS Glue also provides a serverless Apache Spark environment to run ETL jobs, eliminating the need to provision and manage servers 3.
PySpark is the Python API for Apache Spark, a unified analytics engine for large-scale data processing 4. PySpark can perform various data transformations and manipulations on structured and unstructured data, such as cleaning, enriching, and compressing 5. PySpark can also leverage the distributed computing power of Spark to handle terabytes of data efficiently and scalably 6.
By creating an AWS Glue job and converting the scripts to PySpark, the company can move the scripts out of Amazon EC2 into a more managed solution that will eliminate the need to maintain servers. The company can also reduce the development effort by using the AWS Glue console, AWS SDK, or AWS CLI to create and run the job 7. Moreover, the company can use the AWS Glue Data Catalog to store and manage the metadata of the data sources and targets 8.
The other options are not as suitable as option C for the following reasons:
Option A is not optimal because loading the data into an Amazon Redshift cluster and executing the pipeline by using SQL will incur additional costs and complexity for the company. Amazon Redshift is a fully managed data warehouse service that enables fast and scalable analysis of structured data . However, it is not designed for ETL purposes, such as cleaning, transforming, enriching, and compressing data. Moreover, using SQL to perform these tasks may not be as expressive and flexible as using Python scripts. Furthermore, the company will have to provision and configure the Amazon Redshift cluster, and load and unload the data from Amazon S3, which will increase the development effort and time.
Option B is not feasible because loading the data into Amazon DynamoDB and converting the scripts to an AWS Lambda function will not work for the companys use case. Amazon DynamoDB is a fully managed key-value and document database service that provides fast and consistent performance at any scale . However, it is not suitable for storing and processing terabytes of data daily, as it has limits on the size and throughput of each table and item . Moreover, using AWS Lambda to execute the pipeline will not be efficient or cost-effective, as Lambda has limits on the memory, CPU, and execution time of each function . Therefore, using Amazon DynamoDB and AWS Lambda will not meet the companys requirements for processing large amounts of data quickly and reliably.
Option D is not relevant because creating a set of individual AWS Lambda functions to execute each of the scripts and building a step function by using the AWS Step Functions Data Science SDK will not address the main issue of moving the scripts out of Amazon EC2. AWS Step Functions is a fully managed service that lets you coordinate multiple AWS services into serverless workflows . The AWS Step Functions Data Science SDK is an open source library that allows data scientists to easily create workflows that process and publish machine learning models using Amazon SageMaker and AWS Step Functions . However, these services and tools are not designed for ETL purposes, such as cleaning, transforming, enriching, and compressing data. Moreover, as mentioned in option B, using AWS Lambda to execute the scripts will not be efficient or cost-effective for the companys use case.
Reference:

What Is AWS Glue?
AWS Glue Components
AWS Glue Serverless Spark ETL
PySpark - Overview
PySpark - RDD
PySpark - SparkContext
Adding Jobs in AWS Glue
Populating the AWS Glue Data Catalog
[What Is Amazon Redshift?]
[What Is Amazon DynamoDB?]
[Service, Account, and Table Quotas in DynamoDB]
[AWS Lambda quotas]
[What Is AWS Step Functions?]
[AWS Step Functions Data Science SDK for Python]

QUESTION 197
A retail company is selling products through a global online marketplace. The company wants to use machine learning (ML) to analyze customer feedback and identify specific areas for improvement. A developer has built a tool that collects customer reviews from the online marketplace and stores them in an Amazon S3 bucket. This process yields a dataset of 40 reviews. A data scientist building the ML models must identify additional sources of data to increase the size of the dataset.
Which data sources should the data scientist use to augment the dataset of reviews? (Choose three.)

A. Emails exchanged by customers and the companys customer service agents
B. Social media posts containing the name of the company or its products
C. A publicly available collection of news articles
D. A publicly available collection of customer reviews
E. Product sales revenue figures for the company
F. Instruction manuals for the companys products

Answer: A, B, D

Explanation:
The data sources that the data scientist should use to augment the dataset of reviews are those that contain relevant and diverse customer feedback about the company or its products. Emails exchanged by customers and the companys customer service agents can provide valuable insights into the issues and complaints that customers have, as well as the solutions and responses that the company offers. Social media posts containing the name of the company or its products can capture the opinions and sentiments of customers and potential customers, as well as their reactions to marketing campaigns and product launches. A publicly available collection of customer reviews can provide a large and varied sample of feedback from different online platforms and marketplaces, which can help to generalize the ML models and avoid bias.

Reference:
Detect sentiment from customer reviews using Amazon Comprehend | AWS Machine Learning Blog
How to Apply Machine Learning to Customer Feedback

QUESTION 198
A machine learning (ML) specialist wants to create a data preparation job that uses a PySpark script with complex window aggregation operations to create data for training and testing. The ML specialist needs to evaluate the impact of the number of features and the sample count on model performance.
Which approach should the ML specialist use to determine the ideal data transformations for the model?

A. Add an Amazon SageMaker Debugger hook to the script to capture key metrics. Run the script as an AWS Glue job.
B. Add an Amazon SageMaker Experiments tracker to the script to capture key metrics. Run the script as an AWS Glue job.
C. Add an Amazon SageMaker Debugger hook to the script to capture key parameters. Run the script as a SageMaker processing job.
D. Add an Amazon SageMaker Experiments tracker to the script to capture key parameters. Run the script as a SageMaker processing job.

Answer: D

Explanation:
Amazon SageMaker Experiments is a service that helps track, compare, and evaluate different iterations of ML models. It can be used to capture key parameters such as the number of features and the sample count from a PySpark script that runs as a SageMaker processing job. A SageMaker processing job is a flexible and scalable way to run data processing workloads on AWS, such as feature engineering, data validation, model evaluation, and model interpretation.
Reference:
Amazon SageMaker Experiments
Process Data and Evaluate Models

QUESTION 199
A data scientist has a dataset of machine part images stored in Amazon Elastic File System (Amazon EFS). The data scientist needs to use Amazon SageMaker to create and train an image classification machine learning model based on this dataset. Because of budget and time constraints, management wants the data scientist to create and train a model with the least number of steps and integration work required.
How should the data scientist meet these requirements?

A. Mount the EFS file system to a SageMaker notebook and run a script that copies the data to an

Amazon FSx for Lustre file system. Run the SageMaker training job with the FSx for Lustre file system as the data source.
B. Launch a transient Amazon EMR cluster. Configure steps to mount the EFS file system and copy the data to an Amazon S3 bucket by using S3DistCp. Run the SageMaker training job with Amazon S3 as the data source.
C. Mount the EFS file system to an Amazon EC2 instance and use the AWS CLI to copy the data to an Amazon S3 bucket. Run the SageMaker training job with Amazon S3 as the data source.
D. Run a SageMaker training job with an EFS file system as the data source.

Answer: D

Explanation:
The simplest and fastest way to use the EFS dataset for SageMaker training is to run a SageMaker training job with an EFS file system as the data source. This option does not require any data copying or additional integration steps. SageMaker supports EFS as a data source for training jobs, and it can mount the EFS file system to the training container using the FileSystemConfig parameter. This way, the training script can access the data files as if they were on the local disk of the training instance. Reference:
Access Training Data - Amazon SageMaker
Mount an EFS file system to an Amazon SageMaker notebook (with lifecycle configurations) | AWS Machine Learning Blog

---

QUESTION 200
A retail company uses a machine learning (ML) model for daily sales forecasting. The companys brand manager reports that the model has provided inaccurate results for the past 3 weeks.
At the end of each day, an AWS Glue job consolidates the input data that is used for the forecasting with the actual daily sales data and the predictions of the model. The AWS Glue job stores the data in Amazon S3. The companys ML team is using an Amazon SageMaker Studio notebook to gain an understanding about the source of the model's inaccuracies.
What should the ML team do on the SageMaker Studio notebook to visualize the model's degradation MOST accurately?

A. Create a histogram of the daily sales over the last 3 weeks. In addition, create a histogram of the daily sales from before that period.
B. Create a histogram of the model errors over the last 3 weeks. In addition, create a histogram of the model errors from before that period.
C. Create a line chart with the weekly mean absolute error (MAE) of the model.
D. Create a scatter plot of daily sales versus model error for the last 3 weeks. In addition, create a scatter plot of daily sales versus model error from before that period.

Answer: B

Explanation:
The best way to visualize the models degradation is to create a histogram of the model errors over the last 3 weeks and compare it with a histogram of the model errors from before that period. A histogram is a graphical representation of the distribution of numerical data. It shows how often each value or range of values occurs in the data. A model error is the difference between the actual value and the predicted value. A high model error indicates a poor fit of the model to the data. By comparing the histograms of the model errors, the ML team can see if there is a significant change in the shape, spread, or center of the distribution. This can indicate if the model is underfitting, overfitting, or drifting from the data. A line chart or a scatter plot would not be as effective as a histogram for this purpose, because they do not show the distribution of the errors. A line chart would only show the trend of the errors over time, which may not capture the variability or outliers. A scatter plot would only show the relationship between the errors and another variable, such as daily sales, which may not be relevant or informative for the models performance. Reference:
Histogram - Wikipedia
Model error - Wikipedia
SageMaker Model Monitor - visualizing monitoring results

QUESTION 201
An ecommerce company sends a weekly email newsletter to all of its customers. Management has hired a team of writers to create additional targeted content. A data scientist needs to identify five customer segments based on age, income, and location. The customers current segmentation is unknown. The data scientist previously built an XGBoost model to predict the likelihood of a customer responding to an email based on age, income, and location.
Why does the XGBoost model NOT meet the current requirements, and how can this be fixed?

A. The XGBoost model provides a true/false binary output. Apply principal component analysis (PCA) with five feature dimensions to predict a segment.
B. The XGBoost model provides a true/false binary output. Increase the number of classes the XGBoost model predicts to five classes to predict a segment.
C. The XGBoost model is a supervised machine learning algorithm. Train a k-Nearest-Neighbors (kNN) model with K = 5 on the same dataset to predict a segment.
D. The XGBoost model is a supervised machine learning algorithm. Train a k-means model with K = 5 on the same dataset to predict a segment.

Answer: D

Explanation:
The XGBoost model is a supervised machine learning algorithm, which means it requires labeled data to learn from. The customers current segmentation is unknown, so there is no label to train the XGBoost model on. Moreover, the XGBoost model is designed for classification or regression tasks, not for clustering. Clustering is a type of unsupervised machine learning, which means it does not require labeled data. Clustering algorithms try to find natural groups or clusters in the data based on

their similarity or distance. A common clustering algorithm is k-means, which partitions the data into K clusters, where each data point belongs to the cluster with the nearest mean. To meet the current requirements, the data scientist should train a k-means model with K = 5 on the same dataset to predict a segment for each customer. This way, the data scientist can identify five customer segments based on age, income, and location, without needing any labels. Reference:
What is XGBoost? - Amazon SageMaker
What is Clustering? - Amazon SageMaker
K-Means Algorithm - Amazon SageMaker

QUESTION 202
A global financial company is using machine learning to automate its loan approval process. The company has a dataset of customer information. The dataset contains some categorical fields, such as customer location by city and housing status. The dataset also includes financial fields in different units, such as account balances in US dollars and monthly interest in US cents.
The companys data scientists are using a gradient boosting regression model to infer the credit score for each customer. The model has a training accuracy of 99% and a testing accuracy of 75%. The data scientists want to improve the models testing accuracy.
Which process will improve the testing accuracy the MOST?

A. Use a one-hot encoder for the categorical fields in the dataset. Perform standardization on the financial fields in the dataset. Apply L1 regularization to the data.
B. Use tokenization of the categorical fields in the dataset. Perform binning on the financial fields in the dataset. Remove the outliers in the data by using the z-score.
C. Use a label encoder for the categorical fields in the dataset. Perform L1 regularization on the financial fields in the dataset. Apply L2 regularization to the data.
D. Use a logarithm transformation on the categorical fields in the dataset. Perform binning on the financial fields in the dataset. Use imputation to populate missing values in the dataset.

Answer: A

Explanation:
The question is about improving the testing accuracy of a gradient boosting regression model. The testing accuracy is much lower than the training accuracy, which indicates that the model is overfitting the training data. To reduce overfitting, the following steps are recommended:
Use a one-hot encoder for the categorical fields in the dataset. This will create binary features for each category and avoid imposing an ordinal relationship among them. This can help the model learn the patterns better and generalize to unseen data.
Perform standardization on the financial fields in the dataset. This will scale the features to have zero mean and unit variance, which can improve the convergence and performance of the model. This can also help the model handle features with different units and ranges.
Apply L1 regularization to the data. This will add a penalty term to the loss function that is proportional to the absolute value of the coefficients. This can help the model reduce the complexity

and select the most relevant features by shrinking the coefficients of less important features to zero.
Reference:
1: AWS Machine Learning Specialty Exam Guide
2: AWS Machine Learning Specialty Course
3: AWS Machine Learning Blog

QUESTION 203
A machine learning (ML) specialist needs to extract embedding vectors from a text series. The goal is to provide a ready-to-ingest feature space for a data scientist to develop downstream ML predictive models. The text consists of curated sentences in English. Many sentences use similar words but in different contexts. There are questions and answers among the sentences, and the embedding space must differentiate between them.
Which options can produce the required embedding vectors that capture word context and sequential QA information? (Choose two.)

A. Amazon SageMaker seq2seq algorithm
B. Amazon SageMaker BlazingText algorithm in Skip-gram mode
C. Amazon SageMaker Object2Vec algorithm
D. Amazon SageMaker BlazingText algorithm in continuous bag-of-words (CBOW) mode
E. Combination of the Amazon SageMaker BlazingText algorithm in Batch Skip-gram mode with a custom recurrent neural network (RNN)

Answer: B, E

Explanation:
To capture word context and sequential QA information, the embedding vectors need to consider both the order and the meaning of the words in the text.
Option B, Amazon SageMaker BlazingText algorithm in Skip-gram mode, is a valid option because it can learn word embeddings that capture the semantic similarity and syntactic relations between words based on their co-occurrence in a window of words. Skip-gram mode can also handle rare words better than continuous bag-of-words (CBOW) mode1.
Option E, combination of the Amazon SageMaker BlazingText algorithm in Batch Skip-gram mode with a custom recurrent neural network (RNN), is another valid option because it can leverage the advantages of Skip-gram mode and also use an RNN to model the sequential nature of the text. An RNN can capture the temporal dependencies and long-term dependencies between words, which are important for QA tasks2.
Option A, Amazon SageMaker seq2seq algorithm, is not a valid option because it is designed for sequence-to-sequence tasks such as machine translation, summarization, or chatbots. It does not produce embedding vectors for text series, but rather generates an output sequence given an input sequence3.
Option C, Amazon SageMaker Object2Vec algorithm, is not a valid option because it is designed for learning embeddings for pairs of objects, such as text-image, text-text, or image-image. It does not

produce embedding vectors for text series, but rather learns a similarity function between pairs of objects4.
Option D, Amazon SageMaker BlazingText algorithm in continuous bag-of-words (CBOW) mode, is not a valid option because it does not capture word context as well as Skip-gram mode. CBOW mode predicts a word given its surrounding words, while Skip-gram mode predicts the surrounding words given a word. CBOW mode is faster and more suitable for frequent words, but Skip-gram mode can learn more meaningful embeddings for rare words1.
Reference:
1: Amazon SageMaker BlazingText
2: Recurrent Neural Networks (RNNs)
3: Amazon SageMaker Seq2Seq
4: Amazon SageMaker Object2Vec

QUESTION 204
A retail company wants to update its customer support system. The company wants to implement automatic routing of customer claims to different queues to prioritize the claims by category.
Currently, an operator manually performs the category assignment and routing. After the operator classifies and routes the claim, the company stores the claims record in a central database. The claims record includes the claims category.
The company has no data science team or experience in the field of machine learning (ML). The companys small development team needs a solution that requires no ML expertise.
Which solution meets these requirements?

A. Export the database to a .csv file with two columns: claim_label and claim_text. Use the Amazon SageMaker Object2Vec algorithm and the .csv file to train a model. Use SageMaker to deploy the model to an inference endpoint. Develop a service in the application to use the inference endpoint to process incoming claims, predict the labels, and route the claims to the appropriate queue.
B. Export the database to a .csv file with one column: claim_text. Use the Amazon SageMaker Latent Dirichlet Allocation (LDA) algorithm and the .csv file to train a model. Use the LDA algorithm to detect labels automatically. Use SageMaker to deploy the model to an inference endpoint. Develop a service in the application to use the inference endpoint to process incoming claims, predict the labels, and route the claims to the appropriate queue.
C. Use Amazon Textract to process the database and automatically detect two columns: claim_label and claim_text. Use Amazon Comprehend custom classification and the extracted information to train the custom classifier. Develop a service in the application to use the Amazon Comprehend API to process incoming claims, predict the labels, and route the claims to the appropriate queue.
D. Export the database to a .csv file with two columns: claim_label and claim_text. Use Amazon Comprehend custom classification and the .csv file to train the custom classifier. Develop a service in the application to use the Amazon Comprehend API to process incoming claims, predict the labels, and route the claims to the appropriate queue.

Answer: D

Explanation:
Amazon Comprehend is a natural language processing (NLP) service that can analyze text and extract insights such as sentiment, entities, topics, and language. Amazon Comprehend also provides custom classification and custom entity recognition features that allow users to train their own models using their own data and labels. For the scenario of routing customer claims to different queues based on categories, Amazon Comprehend custom classification is a suitable solution. The custom classifier can be trained using a .csv file that contains the claim text and the claim label as columns. The custom classifier can then be used to process incoming claims and predict the labels using the Amazon Comprehend API. The predicted labels can be used to route the claims to the appropriate queue. This solution does not require any machine learning expertise or model deployment, and it can be easily integrated with the existing application.
The other options are not suitable because:
Option A: Amazon SageMaker Object2Vec is an algorithm that can learn embeddings of objects such as words, sentences, or documents. It can be used for tasks such as text classification, sentiment analysis, or recommendation systems. However, using this algorithm requires machine learning expertise and model deployment using SageMaker, which are not available for the company.
Option B: Amazon SageMaker Latent Dirichlet Allocation (LDA) is an algorithm that can discover the topics or themes in a collection of documents. It can be used for tasks such as topic modeling, document clustering, or text summarization. However, using this algorithm requires machine learning expertise and model deployment using SageMaker, which are not available for the company. Moreover, LDA does not provide labels for the topics, but rather a distribution of words for each topic, which may not match the existing categories of the claims.
Option C: Amazon Textract is a service that can extract text and data from scanned documents or images. It can be used for tasks such as document analysis, data extraction, or form processing. However, using this service is unnecessary and inefficient for the scenario, since the company already has the claim text and label in a database. Moreover, Amazon Textract does not provide custom classification features, so it cannot be used to train a custom classifier using the existing data and labels.
Reference:
Amazon Comprehend Custom Classification
Amazon SageMaker Object2Vec
Amazon SageMaker Latent Dirichlet Allocation
Amazon Textract

QUESTION 205
A machine learning (ML) specialist is using Amazon SageMaker hyperparameter optimization (HPO) to improve a models accuracy. The learning rate parameter is specified in the following HPO configuration:

```
{
    "Name": "learning_rate",
    "MaxValue" : "0.0001",
    "MinValue": "0.1"
}
```

During the results analysis, the ML specialist determines that most of the training jobs had a learning rate between 0.01 and 0.1. The best result had a learning rate of less than 0.01. Training jobs need to run regularly over a changing dataset. The ML specialist needs to find a tuning mechanism that uses different learning rates more evenly from the provided range between MinValue and MaxValue.
Which solution provides the MOST accurate result?
A.Modify the HPO configuration as follows:

```
{
    "Name": "learning_rate",
    "MaxValue" : "0.0001",
    "MinValue": "0.1"
    "ScalingType": "ReverseLogarithmic"
}
```

Select the most accurate hyperparameter configuration form this HPO job.
B.
Run three different HPO jobs that use different learning rates form the following intervals for MinValue and MaxValue while using the same number of training jobs for each HPO job:
[0.01, 0.1]
[0.001, 0.01]
[0.0001, 0.001]
Select the most accurate hyperparameter configuration form these three HPO jobs.
C.Modify the HPO configuration as follows:

```
{
    "Name": "learning_rate",
    "MaxValue" : "0.0001",
    "MinValue": "0.1"
    "ScalingType": "Logarithmic"
}
```

Select the most accurate hyperparameter configuration form this training job.

D.
Run three different HPO jobs that use different learning rates form the following intervals for MinValue and MaxValue. Divide the number of training jobs for each HPO job by three:
[0.01, 0.1]
[0.001, 0.01]
[0.0001, 0.001]
Select the most accurate hyperparameter configuration form these three HPO jobs.

Answer: C

Explanation:
The solution C modifies the HPO configuration to use a logarithmic scale for the learning rate parameter. This means that the values of the learning rate are sampled from a log-uniform distribution, which gives more weight to smaller values. This can help to explore the lower end of the range more evenly and find the optimal learning rate more efficiently. The other solutions either use a linear scale, which may not sample enough values from the lower end, or divide the range into subintervals, which may miss some combinations of hyperparameters. Reference:
How Hyperparameter Tuning Works - Amazon SageMaker
Tuning Hyperparameters - Amazon SageMaker

QUESTION 206
A manufacturing company wants to use machine learning (ML) to automate quality control in its facilities. The facilities are in remote locations and have limited internet connectivity. The company has 20 Đ¢Đ’ of training data that consists of labeled images of defective product parts. The training data is in the corporate on-premises data center.
The company will use this data to train a model for real-time defect detection in new parts as the parts move on a conveyor belt in the facilities. The company needs a solution that minimizes costs for compute infrastructure and that maximizes the scalability of resources for training. The solution also must facilitate the companys use of an ML model in the low-connectivity environments.
Which solution will meet these requirements?

A. Move the training data to an Amazon S3 bucket. Train and evaluate the model by using Amazon SageMaker. Optimize the model by using SageMaker Neo. Deploy the model on a SageMaker hosting services endpoint.
B. Train and evaluate the model on premises. Upload the model to an Amazon S3 bucket. Deploy the model on an Amazon SageMaker hosting services endpoint.
C. Move the training data to an Amazon S3 bucket. Train and evaluate the model by using Amazon SageMaker. Optimize the model by using SageMaker Neo. Set up an edge device in the manufacturing facilities with AWS IoT Greengrass. Deploy the model on the edge device.
D. Train the model on premises. Upload the model to an Amazon S3 bucket. Set up an edge device in the manufacturing facilities with AWS IoT Greengrass. Deploy the model on the edge device.

Answer: C

Explanation:
The solution C meets the requirements because it minimizes costs for compute infrastructure, maximizes the scalability of resources for training, and facilitates the use of an ML model in lowconnectivity environments. The solution C involves the following steps:
Move the training data to an Amazon S3 bucket. This will enable the company to store the large amount of data in a durable, scalable, and cost-effective way. It will also allow the company to access the data from the cloud for training and evaluation purposes1.
Train and evaluate the model by using Amazon SageMaker. This will enable the company to use a fully managed service that provides various features and tools for building, training, tuning, and deploying ML models. Amazon SageMaker can handle large-scale data processing and distributed training, and it can leverage the power of AWS compute resources such as Amazon EC2, Amazon EKS, and AWS Fargate2.
Optimize the model by using SageMaker Neo. This will enable the company to reduce the size of the model and improve its performance and efficiency. SageMaker Neo can compile the model into an executable that can run on various hardware platforms, such as CPUs, GPUs, and edge devices3.
Set up an edge device in the manufacturing facilities with AWS IoT Greengrass. This will enable the company to deploy the model on a local device that can run inference in real time, even in lowconnectivity environments. AWS IoT Greengrass can extend AWS cloud capabilities to the edge, and it can securely communicate with the cloud for updates and synchronization4.
Deploy the model on the edge device. This will enable the company to automate quality control in its facilities by using the model to detect defects in new parts as they move on a conveyor belt. The model can run inference locally on the edge device without requiring internet connectivity, and it can send the results to the cloud when the connection is available4.
The other options are not suitable because:
Option A: Deploying the model on a SageMaker hosting services endpoint will not facilitate the use of the model in low-connectivity environments, as it will require internet access to perform inference. Moreover, it may incur higher costs for hosting and data transfer than deploying the model on an edge device.
Option B: Training and evaluating the model on premises will not minimize costs for compute infrastructure, as it will require the company to maintain and upgrade its own hardware and software. Moreover, it will not maximize the scalability of resources for training, as it will limit the companys ability to leverage the clouds elasticity and flexibility.
Option D: Training the model on premises will not minimize costs for compute infrastructure, nor maximize the scalability of resources for training, for the same reasons as option B.
Reference:
1: Amazon S3
2: Amazon SageMaker
3: SageMaker Neo
4: AWS IoT Greengrass

QUESTION 207
A company has an ecommerce website with a product recommendation engine built in TensorFlow. The recommendation engine endpoint is hosted by Amazon SageMaker. Three compute-optimized instances support the expected peak load of the website.
Response times on the product recommendation page are increasing at the beginning of each month. Some users are encountering errors. The website receives the majority of its traffic between 8 AM and 6 PM on weekdays in a single time zone.
Which of the following options are the MOST effective in solving the issue while keeping costs to a minimum? (Choose two.)

A. Configure the endpoint to use Amazon Elastic Inference (EI) accelerators.
B. Create a new endpoint configuration with two production variants.
C. Configure the endpoint to automatically scale with the Invocations Per Instance metric.
D. Deploy a second instance pool to support a blue/green deployment of models.
E. Reconfigure the endpoint to use burstable instances.

Answer: A, C

Explanation:
The solution A and C are the most effective in solving the issue while keeping costs to a minimum. The solution A and C involve the following steps:
Configure the endpoint to use Amazon Elastic Inference (EI) accelerators. This will enable the company to reduce the cost and latency of running TensorFlow inference on SageMaker. Amazon EI provides GPU-powered acceleration for deep learning models without requiring the use of GPU instances. Amazon EI can attach to any SageMaker instance type and provide the right amount of acceleration based on the workload1.
Configure the endpoint to automatically scale with the Invocations Per Instance metric. This will enable the company to adjust the number of instances based on the demand and traffic patterns of the website. The Invocations Per Instance metric measures the average number of requests that each instance processes over a period of time. By using this metric, the company can scale out the endpoint when the load increases and scale in when the load decreases. This can improve the response time and availability of the product recommendation engine2.
The other options are not suitable because:
Option B: Creating a new endpoint configuration with two production variants will not solve the issue of increasing response time and errors. Production variants are used to split the traffic between different models or versions of the same model. They can be useful for testing, updating, or A/B testing models. However, they do not provide any scaling or acceleration benefits for the inference workload3.
Option D: Deploying a second instance pool to support a blue/green deployment of models will not solve the issue of increasing response time and errors. Blue/green deployment is a technique for updating models without downtime or disruption. It involves creating a new endpoint configuration

with a different instance pool and model version, and then shifting the traffic from the old endpoint to the new endpoint gradually. However, this technique does not provide any scaling or acceleration benefits for the inference workload4.
Option E: Reconfiguring the endpoint to use burstable instances will not solve the issue of increasing response time and errors. Burstable instances are instances that provide a baseline level of CPU performance with the ability to burst above the baseline when needed. They can be useful for workloads that have moderate CPU utilization and occasional spikes. However, they are not suitable for workloads that have high and consistent CPU utilization, such as the product recommendation engine. Moreover, burstable instances may incur additional charges when they exceed their CPU credits5.
Reference:
1: Amazon Elastic Inference
2: How to Scale Amazon SageMaker Endpoints
3: Deploying Models to Amazon SageMaker Hosting Services
4: Updating Models in Amazon SageMaker Hosting Services
5: Burstable Performance Instances

QUESTION 208
A real-estate company is launching a new product that predicts the prices of new houses. The historical data for the properties and prices is stored in .csv format in an Amazon S3 bucket. The data has a header, some categorical fields, and some missing values. The companys data scientists have used Python with a common open-source library to fill the missing values with zeros. The data scientists have dropped all of the categorical fields and have trained a model by using the opensource linear regression algorithm with the default parameters.
The accuracy of the predictions with the current model is below 50%. The company wants to improve the model performance and launch the new product as soon as possible.
Which solution will meet these requirements with the LEAST operational overhead?

A. Create a service-linked role for Amazon Elastic Container Service (Amazon ECS) with access to the S3 bucket. Create an ECS cluster that is based on an AWS Deep Learning Containers image. Write the code to perform the feature engineering. Train a logistic regression model for predicting the price, pointing to the bucket with the dataset. Wait for the training job to complete. Perform the inferences.
B. Create an Amazon SageMaker notebook with a new IAM role that is associated with the notebook. Pull the dataset from the S3 bucket. Explore different combinations of feature engineering transformations, regression algorithms, and hyperparameters. Compare all the results in the notebook, and deploy the most accurate configuration in an endpoint for predictions.
C. Create an IAM role with access to Amazon S3, Amazon SageMaker, and AWS Lambda. Create a training job with the SageMaker built-in XGBoost model pointing to the bucket with the dataset. Specify the price as the target feature. Wait for the job to complete. Load the model artifact to a Lambda function for inference on prices of new houses.
D. Create an IAM role for Amazon SageMaker with access to the S3 bucket. Create a SageMaker

AutoML job with SageMaker Autopilot pointing to the bucket with the dataset. Specify the price as the target attribute. Wait for the job to complete. Deploy the best model for predictions.

Answer: D

Explanation:
The solution D meets the requirements with the least operational overhead because it uses Amazon SageMaker Autopilot, which is a fully managed service that automates the end-to-end process of building, training, and deploying machine learning models. Amazon SageMaker Autopilot can handle data preprocessing, feature engineering, algorithm selection, hyperparameter tuning, and model deployment. The company only needs to create an IAM role for Amazon SageMaker with access to the S3 bucket, create a SageMaker AutoML job pointing to the bucket with the dataset, specify the price as the target attribute, and wait for the job to complete. Amazon SageMaker Autopilot will generate a list of candidate models with different configurations and performance metrics, and the company can deploy the best model for predictions1.
The other options are not suitable because:
Option A: Creating a service-linked role for Amazon Elastic Container Service (Amazon ECS) with access to the S3 bucket, creating an ECS cluster based on an AWS Deep Learning Containers image, writing the code to perform the feature engineering, training a logistic regression model for predicting the price, and performing the inferences will incur more operational overhead than using Amazon SageMaker Autopilot. The company will have to manage the ECS cluster, the container image, the code, the model, and the inference endpoint. Moreover, logistic regression may not be the best algorithm for predicting the price, as it is more suitable for binary classification tasks2.
Option B: Creating an Amazon SageMaker notebook with a new IAM role that is associated with the notebook, pulling the dataset from the S3 bucket, exploring different combinations of feature engineering transformations, regression algorithms, and hyperparameters, comparing all the results in the notebook, and deploying the most accurate configuration in an endpoint for predictions will incur more operational overhead than using Amazon SageMaker Autopilot. The company will have to write the code for the feature engineering, the model training, the model evaluation, and the model deployment. The company will also have to manually compare the results and select the best configuration3.
Option C: Creating an IAM role with access to Amazon S3, Amazon SageMaker, and AWS Lambda, creating a training job with the SageMaker built-in XGBoost model pointing to the bucket with the dataset, specifying the price as the target feature, loading the model artifact to a Lambda function for inference on prices of new houses will incur more operational overhead than using Amazon SageMaker Autopilot. The company will have to create and manage the Lambda function, the model artifact, and the inference endpoint. Moreover, XGBoost may not be the best algorithm for predicting the price, as it is more suitable for classification and ranking tasks4.
Reference:
1: Amazon SageMaker Autopilot
2: Amazon Elastic Container Service

3: Amazon SageMaker Notebook Instances
4: Amazon SageMaker XGBoost Algorithm

QUESTION UESTION NO. 181
A data scientist is training a large PyTorch model by using Amazon SageMaker. It takes 10 hours on average to train the model on GPU instances. The data scientist suspects that training is not converging and that
resource utilization is not optimal.
What should the data scientist do to identify and address training issues with the LEAST development effort?
Use CPU utilization metrics that are captured in Amazon CloudWatch. Configure a CloudWatch alarm to stop the training job early if low CPU utilization occurs.
Use high-resolution custom metrics that are captured in Amazon CloudWatch. Configure an AWS Lambda function to analyze the metrics and to stop the training job early if issues are detected.
Use the SageMaker Debugger vanishing_gradient and LowGPUUtilization built-in rules to detect issues and to launch the StopTrainingJob action if issues are detected.
Use the SageMaker Debugger confusion and feature_importance_overweight built-in rules to detect issues and to launch the StopTrainingJob action if issues are detected.

Answer: C

Explanation:
The solution C is the best option to identify and address training issues with the least development effort. The solution C involves the following steps:
Use the SageMaker Debugger vanishing_gradient and LowGPUUtilization built-in rules to detect issues. SageMaker Debugger is a feature of Amazon SageMaker that allows data scientists to monitor, analyze, and debug machine learning models during training. SageMaker Debugger provides a set of built-in rules that can automatically detect common issues and anomalies in model training, such as vanishing or exploding gradients, overfitting, underfitting, low GPU utilization, and more1. The data scientist can use the vanishing_gradient rule to check if the gradients are becoming too small and causing the training to not converge. The data scientist can also use the LowGPUUtilization rule to check if the GPU resources are underutilized and causing the training to be inefficient2.
Launch the StopTrainingJob action if issues are detected. SageMaker Debugger can also take actions based on the status of the rules. One of the actions is StopTrainingJob, which can terminate the training job if a rule is in an error state. This can help the data scientist to save time and money by stopping the training early if issues are detected3.
The other options are not suitable because:
Option A: Using CPU utilization metrics that are captured in Amazon CloudWatch and configuring a CloudWatch alarm to stop the training job early if low CPU utilization occurs will not identify and address training issues effectively. CPU utilization is not a good indicator of model training performance, especially for GPU instances. Moreover, CloudWatch alarms can only trigger actions

based on simple thresholds, not complex rules or conditions4.
Option B: Using high-resolution custom metrics that are captured in Amazon CloudWatch and configuring an AWS Lambda function to analyze the metrics and to stop the training job early if issues are detected will incur more development effort than using SageMaker Debugger. The data scientist will have to write the code for capturing, sending, and analyzing the custom metrics, as well as for invoking the Lambda function and stopping the training job. Moreover, this solution may not be able to detect all the issues that SageMaker Debugger can5.
Option D: Using the SageMaker Debugger confusion and feature_importance_overweight built-in rules and launching the StopTrainingJob action if issues are detected will not identify and address training issues effectively. The confusion rule is used to monitor the confusion matrix of a classification model, which is not relevant for a regression model that predicts prices. The feature_importance_overweight rule is used to check if some features have too much weight in the model, which may not be related to the convergence or resource utilization issues2.
Reference:
1: Amazon SageMaker Debugger
2: Built-in Rules for Amazon SageMaker Debugger
3: Actions for Amazon SageMaker Debugger
4: Amazon CloudWatch Alarms
5: Amazon CloudWatch Custom Metrics

QUESTION 209
A company needs to deploy a chatbot to answer common questions from customers. The chatbot must base its answers on company documentation.
Which solution will meet these requirements with the LEAST development effort?

A. Index company documents by using Amazon Kendra. Integrate the chatbot with Amazon Kendra by using the Amazon Kendra Query API operation to answer customer questions.
B. Train a Bidirectional Attention Flow (BiDAF) network based on past customer questions and company documents. Deploy the model as a real-time Amazon SageMaker endpoint. Integrate the model with the chatbot by using the SageMaker Runtime InvokeEndpoint API operation to answer customer questions.
C. Train an Amazon SageMaker BlazingText model based on past customer questions and company documents. Deploy the model as a real-time SageMaker endpoint. Integrate the model with the chatbot by using the SageMaker Runtime InvokeEndpoint API operation to answer customer questions.
D. Index company documents by using Amazon OpenSearch Service. Integrate the chatbot with OpenSearch Service by using the OpenSearch Service k-nearest neighbors (k-NN) Query API operation to answer customer questions.

Answer: A

Explanation:

The solution A will meet the requirements with the least development effort because it uses Amazon Kendra, which is a highly accurate and easy to use intelligent search service powered by machine learning. Amazon Kendra can index company documents from various sources and formats, such as PDF, HTML, Word, and more. Amazon Kendra can also integrate with chatbots by using the Amazon Kendra Query API operation, which can understand natural language questions and provide relevant answers from the indexed documents. Amazon Kendra can also provide additional information, such as document excerpts, links, and FAQs, to enhance the chatbot experience1.
The other options are not suitable because:
Option B: Training a Bidirectional Attention Flow (BiDAF) network based on past customer questions and company documents, deploying the model as a real-time Amazon SageMaker endpoint, and integrating the model with the chatbot by using the SageMaker Runtime InvokeEndpoint API operation will incur more development effort than using Amazon Kendra. The company will have to write the code for the BiDAF network, which is a complex deep learning model for question answering. The company will also have to manage the SageMaker endpoint, the model artifact, and the inference logic2.
Option C: Training an Amazon SageMaker BlazingText model based on past customer questions and company documents, deploying the model as a real-time SageMaker endpoint, and integrating the model with the chatbot by using the SageMaker Runtime InvokeEndpoint API operation will incur more development effort than using Amazon Kendra. The company will have to write the code for the BlazingText model, which is a fast and scalable text classification and word embedding algorithm. The company will also have to manage the SageMaker endpoint, the model artifact, and the inference logic3.
Option D: Indexing company documents by using Amazon OpenSearch Service and integrating the chatbot with OpenSearch Service by using the OpenSearch Service k-nearest neighbors (k-NN) Query API operation will not meet the requirements effectively. Amazon OpenSearch Service is a fully managed service that provides fast and scalable search and analytics capabilities. However, it is not designed for natural language question answering, and it may not provide accurate or relevant answers for the chatbot. Moreover, the k-NN Query API operation is used to find the most similar documents or vectors based on a distance function, not to find the best answers based on a natural language query4.
Reference:
1: Amazon Kendra
2: Bidirectional Attention Flow for Machine Comprehension
3: Amazon SageMaker BlazingText
4: Amazon OpenSearch Service

QUESTION 210
A company ingests machine learning (ML) data from web advertising clicks into an Amazon S3 data lake. Click data is added to an Amazon Kinesis data stream by using the Kinesis Producer Library (KPL). The data is loaded into the S3 data lake from the data stream by using an Amazon Kinesis Data Firehose delivery stream. As the data volume increases, an ML specialist notices that the rate of data ingested into Amazon S3 is relatively constant. There also is an increasing backlog of data for Kinesis

Data Streams and Kinesis Data Firehose to ingest.
Which next step is MOST likely to improve the data ingestion rate into Amazon S3?

A. Increase the number of S3 prefixes for the delivery stream to write to.
B. Decrease the retention period for the data stream.
C. Increase the number of shards for the data stream.
D. Add more consumers using the Kinesis Client Library (KCL).

Answer: C

Explanation:
The solution C is the most likely to improve the data ingestion rate into Amazon S3 because it increases the number of shards for the data stream. The number of shards determines the throughput capacity of the data stream, which affects the rate of data ingestion. Each shard can support up to 1 MB per second of data input and 2 MB per second of data output. By increasing the number of shards, the company can increase the data ingestion rate proportionally. The company can use the UpdateShardCount API operation to modify the number of shards in the data stream1.
The other options are not likely to improve the data ingestion rate into Amazon S3 because:
Option A: Increasing the number of S3 prefixes for the delivery stream to write to will not affect the data ingestion rate, as it only changes the way the data is organized in the S3 bucket. The number of S3 prefixes can help to optimize the performance of downstream applications that read the data from S3, but it does not impact the performance of Kinesis Data Firehose2.
Option B: Decreasing the retention period for the data stream will not affect the data ingestion rate, as it only changes the amount of time the data is stored in the data stream. The retention period can help to manage the data availability and durability, but it does not impact the throughput capacity of the data stream3.
Option D: Adding more consumers using the Kinesis Client Library (KCL) will not affect the data ingestion rate, as it only changes the way the data is processed by downstream applications. The consumers can help to scale the data processing and handle failures, but they do not impact the data ingestion into S3 by Kinesis Data Firehose4.
Reference:
1: Resharding - Amazon Kinesis Data Streams
2: Amazon S3 Prefixes - Amazon Kinesis Data Firehose
3: Data Retention - Amazon Kinesis Data Streams
4: Developing Consumers Using the Kinesis Client Library - Amazon Kinesis Data Streams

---

QUESTION 211
A manufacturing company has a production line with sensors that collect hundreds of quality metrics. The company has stored sensor data and manual inspection results in a data lake for several months. To automate quality control, the machine learning team must build an automated mechanism that determines whether the produced goods are good quality, replacement market quality, or scrap quality based on the manual inspection results.

Which modeling approach will deliver the MOST accurate prediction of product quality?

A. Amazon SageMaker DeepAR forecasting algorithm
B. Amazon SageMaker XGBoost algorithm
C. Amazon SageMaker Latent Dirichlet Allocation (LDA) algorithm
D. A convolutional neural network (CNN) and ResNet

Answer: D

Explanation:
A convolutional neural network (CNN) is a type of deep learning model that can learn to extract features from images and perform tasks such as classification, segmentation, and detection1. ResNet is a popular CNN architecture that uses residual connections to overcome the problem of vanishing gradients and enable very deep networks2. For the task of predicting product quality based on sensor data, a CNN and ResNet approach can leverage the spatial structure of the data and learn complex patterns that distinguish different quality levels.
Reference:
Convolutional Neural Networks (CNNs / ConvNets)
PyTorch ResNet: The Basics and a Quick Tutorial

QUESTION 212
A media company wants to create a solution that identifies celebrities in pictures that users upload. The company also wants to identify the IP address and the timestamp details from the users so the company can prevent users from uploading pictures from unauthorized locations.
Which solution will meet these requirements with LEAST development effort?

A. Use AWS Panorama to identify celebrities in the pictures. Use AWS CloudTrail to capture IP address and timestamp details.
B. Use AWS Panorama to identify celebrities in the pictures. Make calls to the AWS Panorama Device SDK to capture IP address and timestamp details.
C. Use Amazon Rekognition to identify celebrities in the pictures. Use AWS CloudTrail to capture IP address and timestamp details.
D. Use Amazon Rekognition to identify celebrities in the pictures. Use the text detection feature to capture IP address and timestamp details.

Answer: C

Explanation:
The solution C will meet the requirements with the least development effort because it uses Amazon Rekognition and AWS CloudTrail, which are fully managed services that can provide the desired functionality. The solution C involves the following steps:
Use Amazon Rekognition to identify celebrities in the pictures. Amazon Rekognition is a service that

can analyze images and videos and extract insights such as faces, objects, scenes, emotions, and more. Amazon Rekognition also provides a feature called Celebrity Recognition, which can recognize thousands of celebrities across a number of categories, such as politics, sports, entertainment, and medi
a. Amazon Rekognition can return the name, face, and confidence score of the recognized celebrities, as well as additional information such as URLs and biographies1.
Use AWS CloudTrail to capture IP address and timestamp details. AWS CloudTrail is a service that can record the API calls and events made by or on behalf of AWS accounts. AWS CloudTrail can provide information such as the source IP address, the user identity, the request parameters, and the response elements of the API calls. AWS CloudTrail can also deliver the event records to an Amazon S3 bucket or an Amazon CloudWatch Logs group for further analysis and auditing2.
The other options are not suitable because:
Option A: Using AWS Panorama to identify celebrities in the pictures and using AWS CloudTrail to capture IP address and timestamp details will not meet the requirements effectively. AWS Panorama is a service that can extend computer vision to the edge, where it can run inference on video streams from cameras and other devices. AWS Panorama is not designed for identifying celebrities in pictures, and it may not provide accurate or relevant results. Moreover, AWS Panorama requires the use of an AWS Panorama Appliance or a compatible device, which may incur additional costs and complexity3.
Option B: Using AWS Panorama to identify celebrities in the pictures and making calls to the AWS Panorama Device SDK to capture IP address and timestamp details will not meet the requirements effectively, for the same reasons as option
A. Additionally, making calls to the AWS Panorama Device
SDK will require more development effort than using AWS CloudTrail, as it will involve writing custom code and handling errors and exceptions4.
Option D: Using Amazon Rekognition to identify celebrities in the pictures and using the text detection feature to capture IP address and timestamp details will not meet the requirements effectively. The text detection feature of Amazon Rekognition is used to detect and recognize text in images and videos, such as street names, captions, product names, and license plates. It is not suitable for capturing IP address and timestamp details, as these are not part of the pictures that users upload. Moreover, the text detection feature may not be accurate or reliable, as it depends on the quality and clarity of the text in the images and videos5.
Reference:
1: Amazon Rekognition Celebrity Recognition
2: AWS CloudTrail Overview
3: AWS Panorama Overview
4: AWS Panorama Device SDK
5: Amazon Rekognition Text Detection

QUESTION 213
A retail company is ingesting purchasing records from its network of 20,000 stores to Amazon S3 by using Amazon Kinesis Data Firehose. The company uses a small, server-based application in each

store to send the data to AWS over the internet. The company uses this data to train a machine learning model that is retrained each day. The company's data science team has identified existing attributes on these records that could be combined to create an improved model.
Which change will create the required transformed records with the LEAST operational overhead?

A. Create an AWS Lambda function that can transform the incoming records. Enable data transformation on the ingestion Kinesis Data Firehose delivery stream. Use the Lambda function as the invocation target.
B. Deploy an Amazon EMR cluster that runs Apache Spark and includes the transformation logic. Use Amazon EventBridge (Amazon CloudWatch Events) to schedule an AWS Lambda function to launch the cluster each day and transform the records that accumulate in Amazon S3. Deliver the transformed records to Amazon S3.
C. Deploy an Amazon S3 File Gateway in the stores. Update the in-store software to deliver data to the S3 File Gateway. Use a scheduled daily AWS Glue job to transform the data that the S3 File Gateway delivers to Amazon S3.
D. Launch a fleet of Amazon EC2 instances that include the transformation logic. Configure the EC2 instances with a daily cron job to transform the records that accumulate in Amazon S3. Deliver the transformed records to Amazon S3.

Answer: A

Explanation:
The solution A will create the required transformed records with the least operational overhead because it uses AWS Lambda and Amazon Kinesis Data Firehose, which are fully managed services that can provide the desired functionality. The solution A involves the following steps:
Create an AWS Lambda function that can transform the incoming records. AWS Lambda is a service that can run code without provisioning or managing servers. AWS Lambda can execute the transformation logic on the purchasing records and add the new attributes to the records1.
Enable data transformation on the ingestion Kinesis Data Firehose delivery stream. Use the Lambda function as the invocation target. Amazon Kinesis Data Firehose is a service that can capture, transform, and load streaming data into AWS data stores. Amazon Kinesis Data Firehose can enable data transformation and invoke the Lambda function to process the incoming records before delivering them to Amazon S3. This can reduce the operational overhead of managing the transformation process and the data storage2.
The other options are not suitable because:
Option B: Deploying an Amazon EMR cluster that runs Apache Spark and includes the transformation logic, using Amazon EventBridge (Amazon CloudWatch Events) to schedule an AWS Lambda function to launch the cluster each day and transform the records that accumulate in Amazon S3, and delivering the transformed records to Amazon S3 will incur more operational overhead than using AWS Lambda and Amazon Kinesis Data Firehose. The company will have to manage the Amazon EMR cluster, the Apache Spark application, the AWS Lambda function, and the Amazon EventBridge rule. Moreover, this solution will introduce a delay in the transformation process, as it will run only

once a day3.
Option C: Deploying an Amazon S3 File Gateway in the stores, updating the in-store software to deliver data to the S3 File Gateway, and using a scheduled daily AWS Glue job to transform the data that the S3 File Gateway delivers to Amazon S3 will incur more operational overhead than using AWS Lambda and Amazon Kinesis Data Firehose. The company will have to manage the S3 File Gateway, the in-store software, and the AWS Glue job. Moreover, this solution will introduce a delay in the transformation process, as it will run only once a day4.
Option D: Launching a fleet of Amazon EC2 instances that include the transformation logic, configuring the EC2 instances with a daily cron job to transform the records that accumulate in Amazon S3, and delivering the transformed records to Amazon S3 will incur more operational overhead than using AWS Lambda and Amazon Kinesis Data Firehose. The company will have to manage the EC2 instances, the transformation code, and the cron job. Moreover, this solution will introduce a delay in the transformation process, as it will run only once a day5.
Reference:
1: AWS Lambda
2: Amazon Kinesis Data Firehose
3: Amazon EMR
4: Amazon S3 File Gateway
5: Amazon EC2

QUESTION 214
A company wants to segment a large group of customers into subgroups based on shared characteristics. The companys data scientist is planning to use the Amazon SageMaker built-in kmeans clustering algorithm for this task. The data scientist needs to determine the optimal number of subgroups (k) to use.
Which data visualization approach will MOST accurately determine the optimal value of k?

A. Calculate the principal component analysis (PCA) components. Run the k-means clustering algorithm for a range of k by using only the first two PCA components. For each value of k, create a scatter plot with a different color for each cluster. The optimal value of k is the value where the clusters start to look reasonably separated.
B. Calculate the principal component analysis (PCA) components. Create a line plot of the number of components against the explained variance. The optimal value of k is the number of PCA components after which the curve starts decreasing in a linear fashion.
C. Create a t-distributed stochastic neighbor embedding (t-SNE) plot for a range of perplexity values. The optimal value of k is the value of perplexity, where the clusters start to look reasonably separated.
D. Run the k-means clustering algorithm for a range of k. For each value of k, calculate the sum of squared errors (SSE). Plot a line chart of the SSE for each value of k. The optimal value of k is the point after which the curve starts decreasing in a linear fashion.

Answer: D

Explanation:
The solution D is the best data visualization approach to determine the optimal value of k for the kmeans clustering algorithm. The solution D involves the following steps:
Run the k-means clustering algorithm for a range of k. For each value of k, calculate the sum of squared errors (SSE). The SSE is a measure of how well the clusters fit the data. It is calculated by summing the squared distances of each data point to its closest cluster center. A lower SSE indicates a better fit, but it will always decrease as the number of clusters increases. Therefore, the goal is to find the smallest value of k that still has a low SSE1.
Plot a line chart of the SSE for each value of k. The line chart will show how the SSE changes as the value of k increases. Typically, the line chart will have a shape of an elbow, where the SSE drops rapidly at first and then levels off. The optimal value of k is the point after which the curve starts decreasing in a linear fashion. This point is also known as the elbow point, and it represents the balance between the number of clusters and the SSE1.
The other options are not suitable because:
Option A: Calculating the principal component analysis (PCA) components, running the k-means clustering algorithm for a range of k by using only the first two PCA components, and creating a scatter plot with a different color for each cluster will not accurately determine the optimal value of k. PCA is a technique that reduces the dimensionality of the data by transforming it into a new set of features that capture the most variance in the data. However, PCA may not preserve the original structure and distances of the data, and it may lose some information in the process. Therefore, running the k-means clustering algorithm on the PCA components may not reflect the true clusters in the data. Moreover, using only the first two PCA components may not capture enough variance to represent the data well. Furthermore, creating a scatter plot may not be reliable, as it depends on the subjective judgment of the data scientist to decide when the clusters look reasonably separated2.
Option B: Calculating the PCA components and creating a line plot of the number of components against the explained variance will not determine the optimal value of k. This approach is used to determine the optimal number of PCA components to use for dimensionality reduction, not for clustering. The explained variance is the ratio of the variance of each PCA component to the total variance of the data. The optimal number of PCA components is the point where adding more components does not significantly increase the explained variance. However, this number may not correspond to the optimal number of clusters, as PCA and k-means clustering have different objectives and assumptions2.
Option C: Creating a t-distributed stochastic neighbor embedding (t-SNE) plot for a range of perplexity values will not determine the optimal value of k. t-SNE is a technique that reduces the dimensionality of the data by embedding it into a lower-dimensional space, such as a twodimensional plane. t-SNE preserves the local structure and distances of the data, and it can reveal clusters and patterns in the data. However, t-SNE does not assign labels or centroids to the clusters, and it does not provide a measure of how well the clusters fit the data. Therefore, t-SNE cannot determine the optimal number of clusters, as it only visualizes the data. Moreover, t-SNE depends on the perplexity parameter, which is a measure of how many neighbors each point considers. The

perplexity parameter can affect the shape and size of the clusters, and there is no optimal value for it. Therefore, creating a t-SNE plot for a range of perplexity values may not be consistent or reliable3.
Reference:
1: How to Determine the Optimal K for K-Means?
2: Principal Component Analysis
3: t-Distributed Stochastic Neighbor Embedding

QUESTION 215
A car company is developing a machine learning solution to detect whether a car is present in an image. The image dataset consists of one million images. Each image in the dataset is 200 pixels in height by 200 pixels in width. Each image is labeled as either having a car or not having a car.
Which architecture is MOST likely to produce a model that detects whether a car is present in an image with the highest accuracy?

A. Use a deep convolutional neural network (CNN) classifier with the images as input. Include a linear output layer that outputs the probability that an image contains a car.
B. Use a deep convolutional neural network (CNN) classifier with the images as input. Include a softmax output layer that outputs the probability that an image contains a car.
C. Use a deep multilayer perceptron (MLP) classifier with the images as input. Include a linear output layer that outputs the probability that an image contains a car.
D. Use a deep multilayer perceptron (MLP) classifier with the images as input. Include a softmax output layer that outputs the probability that an image contains a car.

Answer: A

Explanation:
A deep convolutional neural network (CNN) classifier is a suitable architecture for image classification tasks, as it can learn features from the images and reduce the dimensionality of the input. A linear output layer that outputs the probability that an image contains a car is appropriate for a binary classification problem, as it can produce a single scalar value between 0 and 1. A softmax output layer is more suitable for a multi-class classification problem, as it can produce a vector of probabilities that sum up to 1. A deep multilayer perceptron (MLP) classifier is not as effective as a CNN for image classification, as it does not exploit the spatial structure of the images and requires a large number of parameters to process the high-dimensional input. Reference:
AWS Certified Machine Learning - Specialty Exam Guide
AWS Training - Machine Learning on AWS
AWS Whitepaper - An Overview of Machine Learning on AWS

QUESTION 216
A data science team is working with a tabular dataset that the team stores in Amazon S3. The team wants to experiment with different feature transformations such as categorical feature encoding. Then the team wants to visualize the resulting distribution of the dataset. After the team finds an

appropriate set of feature transformations, the team wants to automate the workflow for feature transformations.
Which solution will meet these requirements with the MOST operational efficiency?

A. Use Amazon SageMaker Data Wrangler preconfigured transformations to explore feature transformations. Use SageMaker Data Wrangler templates for visualization. Export the feature processing workflow to a SageMaker pipeline for automation.
B. Use an Amazon SageMaker notebook instance to experiment with different feature transformations. Save the transformations to Amazon S3. Use Amazon QuickSight for visualization. Package the feature processing steps into an AWS Lambda function for automation.
C. Use AWS Glue Studio with custom code to experiment with different feature transformations. Save the transformations to Amazon S3. Use Amazon QuickSight for visualization. Package the feature processing steps into an AWS Lambda function for automation.
D. Use Amazon SageMaker Data Wrangler preconfigured transformations to experiment with different feature transformations. Save the transformations to Amazon S3. Use Amazon QuickSight for visualzation. Package each feature transformation step into a separate AWS Lambda function. Use AWS Step Functions for workflow automation.

Answer: A

Explanation:
The solution A will meet the requirements with the most operational efficiency because it uses Amazon SageMaker Data Wrangler, which is a service that simplifies the process of data preparation and feature engineering for machine learning. The solution A involves the following steps:
Use Amazon SageMaker Data Wrangler preconfigured transformations to explore feature transformations. Amazon SageMaker Data Wrangler provides a visual interface that allows data scientists to apply various transformations to their tabular data, such as encoding categorical features, scaling numerical features, imputing missing values, and more. Amazon SageMaker Data Wrangler also supports custom transformations using Python code or SQL queries1.
Use SageMaker Data Wrangler templates for visualization. Amazon SageMaker Data Wrangler also provides a set of templates that can generate visualizations of the data, such as histograms, scatter plots, box plots, and more. These visualizations can help data scientists to understand the distribution and characteristics of the data, and to compare the effects of different feature transformations1.
Export the feature processing workflow to a SageMaker pipeline for automation. Amazon SageMaker Data Wrangler can export the feature processing workflow as a SageMaker pipeline, which is a service that orchestrates and automates machine learning workflows. A SageMaker pipeline can run the feature processing steps as a preprocessing step, and then feed the output to a training step or an inference step. This can reduce the operational overhead of managing the feature processing workflow and ensure its consistency and reproducibility2.
The other options are not suitable because:
Option B: Using an Amazon SageMaker notebook instance to experiment with different feature

transformations, saving the transformations to Amazon S3, using Amazon QuickSight for visualization, and packaging the feature processing steps into an AWS Lambda function for automation will incur more operational overhead than using Amazon SageMaker Data Wrangler. The data scientist will have to write the code for the feature transformations, the data storage, the data visualization, and the Lambda function. Moreover, AWS Lambda has limitations on the execution time, memory size, and package size, which may not be sufficient for complex feature processing tasks3.

Option C: Using AWS Glue Studio with custom code to experiment with different feature transformations, saving the transformations to Amazon S3, using Amazon QuickSight for visualization, and packaging the feature processing steps into an AWS Lambda function for automation will incur more operational overhead than using Amazon SageMaker Data Wrangler. AWS Glue Studio is a visual interface that allows data engineers to create and run extract, transform, and load (ETL) jobs on AWS Glue. However, AWS Glue Studio does not provide preconfigured transformations or templates for feature engineering or data visualization. The data scientist will have to write custom code for these tasks, as well as for the Lambda function. Moreover, AWS Glue Studio is not integrated with SageMaker pipelines, and it may not be optimized for machine learning workflows4.

Option D: Using Amazon SageMaker Data Wrangler preconfigured transformations to experiment with different feature transformations, saving the transformations to Amazon S3, using Amazon QuickSight for visualization, packaging each feature transformation step into a separate AWS Lambda function, and using AWS Step Functions for workflow automation will incur more operational overhead than using Amazon SageMaker Data Wrangler. The data scientist will have to create and manage multiple AWS Lambda functions and AWS Step Functions, which can increase the complexity and cost of the solution. Moreover, AWS Lambda and AWS Step Functions may not be compatible with SageMaker pipelines, and they may not be optimized for machine learning workflows5.

Reference:

1: Amazon SageMaker Data Wrangler
2: Amazon SageMaker Pipelines
3: AWS Lambda
4: AWS Glue Studio
5: AWS Step Functions

QUESTION 217

A company wants to conduct targeted marketing to sell solar panels to homeowners. The company wants to use machine learning (ML) technologies to identify which houses already have solar panels. The company has collected 8,000 satellite images as training data and will use Amazon SageMaker Ground Truth to label the data.

The company has a small internal team that is working on the project. The internal team has no ML expertise and no ML experience.

Which solution will meet these requirements with the LEAST amount of effort from the internal team?

A. Set up a private workforce that consists of the internal team. Use the private workforce and the SageMaker Ground Truth active learning feature to label the data. Use Amazon Rekognition Custom Labels for model training and hosting.
B. Set up a private workforce that consists of the internal team. Use the private workforce to label the data. Use Amazon Rekognition Custom Labels for model training and hosting.
C. Set up a private workforce that consists of the internal team. Use the private workforce and the SageMaker Ground Truth active learning feature to label the data. Use the SageMaker Object Detection algorithm to train a model. Use SageMaker batch transform for inference.
D. Set up a public workforce. Use the public workforce to label the data. Use the SageMaker Object Detection algorithm to train a model. Use SageMaker batch transform for inference.

Answer: A

Explanation:
The solution A will meet the requirements with the least amount of effort from the internal team because it uses Amazon SageMaker Ground Truth and Amazon Rekognition Custom Labels, which are fully managed services that can provide the desired functionality. The solution A involves the following steps:
Set up a private workforce that consists of the internal team. Use the private workforce and the SageMaker Ground Truth active learning feature to label the data. Amazon SageMaker Ground Truth is a service that can create high-quality training datasets for machine learning by using human labelers. A private workforce is a group of labelers that the company can manage and control. The internal team can use the private workforce to label the satellite images as having solar panels or not. The SageMaker Ground Truth active learning feature can reduce the labeling effort by using a machine learning model to automatically label the easy examples and only send the difficult ones to the human labelers1.
Use Amazon Rekognition Custom Labels for model training and hosting. Amazon Rekognition Custom Labels is a service that can train and deploy custom machine learning models for image analysis. Amazon Rekognition Custom Labels can use the labeled data from SageMaker Ground Truth to train a model that can detect solar panels in satellite images. Amazon Rekognition Custom Labels can also host the model and provide an API endpoint for inference2.
The other options are not suitable because:
Option B: Setting up a private workforce that consists of the internal team, using the private workforce to label the data, and using Amazon Rekognition Custom Labels for model training and hosting will incur more effort from the internal team than using SageMaker Ground Truth active learning feature. The internal team will have to label all the images manually, without the assistance of the machine learning model that can automate some of the labeling tasks1.
Option C: Setting up a private workforce that consists of the internal team, using the private workforce and the SageMaker Ground Truth active learning feature to label the data, using the SageMaker Object Detection algorithm to train a model, and using SageMaker batch transform for inference will incur more operational overhead than using Amazon Rekognition Custom Labels. The company will have to manage the SageMaker training job, the model artifact, and the batch

transform job. Moreover, SageMaker batch transform is not suitable for real-time inference, as it processes the data in batches and stores the results in Amazon S33.
Option D: Setting up a public workforce, using the public workforce to label the data, using the SageMaker Object Detection algorithm to train a model, and using SageMaker batch transform for inference will incur more operational overhead and cost than using a private workforce and Amazon Rekognition Custom Labels. A public workforce is a group of labelers from Amazon Mechanical Turk, a crowdsourcing marketplace. The company will have to pay the public workforce for each labeling task, and it may not have full control over the quality and security of the labeled data. The company will also have to manage the SageMaker training job, the model artifact, and the batch transform job, as explained in option C4.
Reference:
1: Amazon SageMaker Ground Truth
2: Amazon Rekognition Custom Labels
3: Amazon SageMaker Object Detection
4: Amazon Mechanical Turk

QUESTION 218
A media company is building a computer vision model to analyze images that are on social media. The model consists of CNNs that the company trained by using images that the company stores in Amazon S3. The company used an Amazon SageMaker training job in File mode with a single Amazon EC2 On-Demand Instance.
Every day, the company updates the model by using about 10,000 images that the company has collected in the last 24 hours. The company configures training with only one epoch. The company wants to speed up training and lower costs without the need to make any code changes.
Which solution will meet these requirements?

A. Instead of File mode, configure the SageMaker training job to use Pipe mode. Ingest the data from a pipe.
B. Instead Of File mode, configure the SageMaker training job to use FastFile mode with no Other changes.
C. Instead Of On-Demand Instances, configure the SageMaker training job to use Spot Instances. Make no Other changes.
D. Instead Of On-Demand Instances, configure the SageMaker training job to use Spot Instances. Implement model checkpoints.

Answer: C

Explanation:
The solution C will meet the requirements because it uses Amazon SageMaker Spot Instances, which are unused EC2 instances that are available at up to 90% discount compared to On-Demand prices. Amazon SageMaker Spot Instances can speed up training and lower costs by taking advantage of the spare EC2 capacity. The company does not need to make any code changes to use Spot Instances, as

it can simply enable the managed spot training option in the SageMaker training job configuration. The company also does not need to implement model checkpoints, as it is using only one epoch for training, which means the model will not resume from a previous state1.
The other options are not suitable because:
Option A: Configuring the SageMaker training job to use Pipe mode instead of File mode will not speed up training or lower costs significantly. Pipe mode is a data ingestion mode that streams data directly from S3 to the training algorithm, without copying the data to the local storage of the training instance. Pipe mode can reduce the startup time of the training job and the disk space usage, but it does not affect the computation time or the instance price. Moreover, Pipe mode may require some code changes to handle the streaming data, depending on the training algorithm2.
Option B: Configuring the SageMaker training job to use FastFile mode instead of File mode will not speed up training or lower costs significantly. FastFile mode is a data ingestion mode that copies data from S3 to the local storage of the training instance in parallel with the training process. FastFile mode can reduce the startup time of the training job and the disk space usage, but it does not affect the computation time or the instance price. Moreover, FastFile mode is only available for distributed training jobs that use multiple instances, which is not the case for the company3.
Option D: Configuring the SageMaker training job to use Spot Instances and implementing model checkpoints will not meet the requirements without the need to make any code changes. Model checkpoints are a feature that allows the training job to save the model state periodically to S3, and resume from the latest checkpoint if the training job is interrupted. Model checkpoints can help to avoid losing the training progress and ensure the model convergence, but they require some code changes to implement the checkpointing logic and the resuming logic4.
Reference:
1: Managed Spot Training - Amazon SageMaker
2: Pipe Mode - Amazon SageMaker
3: FastFile Mode - Amazon SageMaker
4: Checkpoints - Amazon SageMaker

QUESTION 219
A data scientist is working on a forecast problem by using a dataset that consists of .csv files that are stored in Amazon S3. The files contain a timestamp variable in the following format:
March 1st, 2020, 08:14pm -
There is a hypothesis about seasonal differences in the dependent variable. This number could be higher or lower for weekdays because some days and hours present varying values, so the day of the week, month, or hour could be an important factor. As a result, the data scientist needs to transform the timestamp into weekdays, month, and day as three separate variables to conduct an analysis.
Which solution requires the LEAST operational overhead to create a new dataset with the added features?

A. Create an Amazon EMR cluster. Develop PySpark code that can read the timestamp variable as a string, transform and create the new variables, and save the dataset as a new file in Amazon S3.
B. Create a processing job in Amazon SageMaker. Develop Python code that can read the timestamp

variable as a string, transform and create the new variables, and save the dataset as a new file in Amazon S3.
C. Create a new flow in Amazon SageMaker Data Wrangler. Import the S3 file, use the Featurize date/time transform to generate the new variables, and save the dataset as a new file in Amazon S3.
D. Create an AWS Glue job. Develop code that can read the timestamp variable as a string, transform and create the new variables, and save the dataset as a new file in Amazon S3.

Answer: C

Explanation:
The solution C will create a new dataset with the added features with the least operational overhead because it uses Amazon SageMaker Data Wrangler, which is a service that simplifies the process of data preparation and feature engineering for machine learning. The solution C involves the following steps:
Create a new flow in Amazon SageMaker Data Wrangler. A flow is a visual representation of the data preparation steps that can be applied to one or more datasets. The data scientist can create a new flow in the Amazon SageMaker Studio interface and import the S3 file as a data source1.
Use the Featurize date/time transform to generate the new variables. Amazon SageMaker Data Wrangler provides a set of preconfigured transformations that can be applied to the data with a few clicks. The Featurize date/time transform can parse a date/time column and generate new columns for the year, month, day, hour, minute, second, day of week, and day of year. The data scientist can use this transform to create the new variables from the timestamp variable2.
Save the dataset as a new file in Amazon S3. Amazon SageMaker Data Wrangler can export the transformed dataset as a new file in Amazon S3, or as a feature store in Amazon SageMaker Feature Store. The data scientist can choose the output format and location of the new file3.
The other options are not suitable because:
Option A: Creating an Amazon EMR cluster and developing PySpark code that can read the timestamp variable as a string, transform and create the new variables, and save the dataset as a new file in Amazon S3 will incur more operational overhead than using Amazon SageMaker Data Wrangler. The data scientist will have to manage the Amazon EMR cluster, the PySpark application, and the data storage. Moreover, the data scientist will have to write custom code for the date/time parsing and feature generation, which may require more development effort and testing4.
Option B: Creating a processing job in Amazon SageMaker and developing Python code that can read the timestamp variable as a string, transform and create the new variables, and save the dataset as a new file in Amazon S3 will incur more operational overhead than using Amazon SageMaker Data Wrangler. The data scientist will have to manage the processing job, the Python code, and the data storage. Moreover, the data scientist will have to write custom code for the date/time parsing and feature generation, which may require more development effort and testing5.
Option D: Creating an AWS Glue job and developing code that can read the timestamp variable as a string, transform and create the new variables, and save the dataset as a new file in Amazon S3 will incur more operational overhead than using Amazon SageMaker Data Wrangler. The data scientist will have to manage the AWS Glue job, the code, and the data storage. Moreover, the data scientist

will have to write custom code for the date/time parsing and feature generation, which may require more development effort and testing6.
Reference:
1: Amazon SageMaker Data Wrangler
2: Featurize Date/Time - Amazon SageMaker Data Wrangler
3: Exporting Data - Amazon SageMaker Data Wrangler
4: Amazon EMR
5: Processing Jobs - Amazon SageMaker
6: AWS Glue

QUESTION 220
An automotive company uses computer vision in its autonomous cars. The company trained its object detection models successfully by using transfer learning from a convolutional neural network (CNN). The company trained the models by using PyTorch through the Amazon SageMaker SDK. The vehicles have limited hardware and compute power. The company wants to optimize the model to reduce memory, battery, and hardware consumption without a significant sacrifice in accuracy. Which solution will improve the computational efficiency of the models?

A. Use Amazon CloudWatch metrics to gain visibility into the SageMaker training weights, gradients, biases, and activation outputs. Compute the filter ranks based on the training information. Apply pruning to remove the low-ranking filters. Set new weights based on the pruned set of filters. Run a new training job with the pruned model.
B. Use Amazon SageMaker Ground Truth to build and run data labeling workflows. Collect a larger labeled dataset with the labelling workflows. Run a new training job that uses the new labeled data with previous training data.
C. Use Amazon SageMaker Debugger to gain visibility into the training weights, gradients, biases, and activation outputs. Compute the filter ranks based on the training information. Apply pruning to remove the low-ranking filters. Set the new weights based on the pruned set of filters. Run a new training job with the pruned model.
D. Use Amazon SageMaker Model Monitor to gain visibility into the ModelLatency metric and OverheadLatency metric of the model after the company deploys the model. Increase the model learning rate. Run a new training job.

Answer: C

Explanation:
The solution C will improve the computational efficiency of the models because it uses Amazon SageMaker Debugger and pruning, which are techniques that can reduce the size and complexity of the convolutional neural network (CNN) models. The solution C involves the following steps:
Use Amazon SageMaker Debugger to gain visibility into the training weights, gradients, biases, and activation outputs. Amazon SageMaker Debugger is a service that can capture and analyze the tensors that are emitted during the training process of machine learning models. Amazon SageMaker

Debugger can provide insights into the model performance, quality, and convergence. Amazon SageMaker Debugger can also help to identify and diagnose issues such as overfitting, underfitting, vanishing gradients, and exploding gradients1.
Compute the filter ranks based on the training information. Filter ranking is a technique that can measure the importance of each filter in a convolutional layer based on some criterion, such as the average percentage of zero activations or the L1-norm of the filter weights. Filter ranking can help to identify the filters that have little or no contribution to the model output, and thus can be removed without affecting the model accuracy2.
Apply pruning to remove the low-ranking filters. Pruning is a technique that can reduce the size and complexity of a neural network by removing the redundant or irrelevant parts of the network, such as neurons, connections, or filters. Pruning can help to improve the computational efficiency, memory usage, and inference speed of the model, as well as to prevent overfitting and improve generalization3.
Set the new weights based on the pruned set of filters. After pruning, the model will have a smaller and simpler architecture, with fewer filters in each convolutional layer. The new weights of the model can be set based on the pruned set of filters, either by initializing them randomly or by finetuning them from the original weights4.
Run a new training job with the pruned model. The pruned model can be trained again with the same or a different dataset, using the same or a different framework or algorithm. The new training job can use the same or a different configuration of Amazon SageMaker, such as the instance type, the hyperparameters, or the data ingestion mode. The new training job can also use Amazon SageMaker Debugger to monitor and analyze the training process and the model quality5.
The other options are not suitable because:
Option A: Using Amazon CloudWatch metrics to gain visibility into the SageMaker training weights, gradients, biases, and activation outputs will not be as effective as using Amazon SageMaker Debugger. Amazon CloudWatch is a service that can monitor and observe the operational health and performance of AWS resources and applications. Amazon CloudWatch can provide metrics, alarms, dashboards, and logs for various AWS services, including Amazon SageMaker. However, Amazon CloudWatch does not provide the same level of granularity and detail as Amazon SageMaker Debugger for the tensors that are emitted during the training process of machine learning models. Amazon CloudWatch metrics are mainly focused on the resource utilization and the training progress, not on the model performance, quality, and convergence6.
Option B: Using Amazon SageMaker Ground Truth to build and run data labeling workflows and collecting a larger labeled dataset with the labeling workflows will not improve the computational efficiency of the models. Amazon SageMaker Ground Truth is a service that can create high-quality training datasets for machine learning by using human labelers. A larger labeled dataset can help to improve the model accuracy and generalization, but it will not reduce the memory, battery, and hardware consumption of the model. Moreover, a larger labeled dataset may increase the training time and cost of the model7.
Option D: Using Amazon SageMaker Model Monitor to gain visibility into the ModelLatency metric and OverheadLatency metric of the model after the company deploys the model and increasing the model learning rate will not improve the computational efficiency of the models. Amazon SageMaker

Model Monitor is a service that can monitor and analyze the quality and performance of machine learning models that are deployed on Amazon SageMaker endpoints. The ModelLatency metric and the OverheadLatency metric can measure the inference latency of the model and the endpoint, respectively. However, these metrics do not provide any information about the training weights, gradients, biases, and activation outputs of the model, which are needed for pruning. Moreover, increasing the model learning rate will not reduce the size and complexity of the model, but it may affect the model convergence and accuracy.
Reference:
1: Amazon SageMaker Debugger
2: Pruning Convolutional Neural Networks for Resource Efficient Inference
3: Pruning Neural Networks: A Survey
4: Learning both Weights and Connections for Efficient Neural Networks
5: Amazon SageMaker Training Jobs
6: Amazon CloudWatch Metrics for Amazon SageMaker
7: Amazon SageMaker Ground Truth
: Amazon SageMaker Model Monitor

QUESTION 221
A chemical company has developed several machine learning (ML) solutions to identify chemical process abnormalities. The time series values of independent variables and the labels are available for the past 2 years and are sufficient to accurately model the problem.
The regular operation label is marked as 0. The abnormal operation label is marked as 1 . Process abnormalities have a significant negative effect on the companys profits. The company must avoid these abnormalities.
Which metrics will indicate an ML solution that will provide the GREATEST probability of detecting an abnormality?

A. Precision = 0.91
Recall = 0.6
B. Precision = 0.61
Recall = 0.98
C. Precision = 0.7
Recall = 0.9
D. Precision = 0.98
Recall = 0.8

Answer: B

Explanation:
The metrics that will indicate an ML solution that will provide the greatest probability of detecting an abnormality are precision and recall. Precision is the ratio of true positives (TP) to the total number of predicted positives (TP + FP), where FP is false positives. Recall is the ratio of true positives (TP) to

the total number of actual positives (TP + FN), where FN is false negatives. A high precision means that the ML solution has a low rate of false alarms, while a high recall means that the ML solution has a high rate of true detections. For the chemical company, the goal is to avoid process abnormalities, which are marked as 1 in the labels. Therefore, the company needs an ML solution that has a high recall for the positive class, meaning that it can detect most of the abnormalities and minimize the false negatives. Among the four options, option B has the highest recall for the positive class, which is 0.98. This means that the ML solution can detect 98% of the abnormalities and miss only 2%. Option B also has a reasonable precision for the positive class, which is 0.61. This means that the ML solution has a false alarm rate of 39%, which may be acceptable for the company, depending on the cost and benefit analysis. The other options have lower recall for the positive class, which means that they have higher false negative rates, which can be more detrimental for the company than false positive rates.
Reference:
1: AWS Certified Machine Learning - Specialty Exam Guide
2: AWS Training - Machine Learning on AWS
3: AWS Whitepaper - An Overview of Machine Learning on AWS
4: Precision and recall

QUESTION 222
A pharmaceutical company performs periodic audits of clinical trial sites to quickly resolve critical findings. The company stores audit documents in text format. Auditors have requested help from a data science team to quickly analyze the documents. The auditors need to discover the 10 main topics within the documents to prioritize and distribute the review work among the auditing team members. Documents that describe adverse events must receive the highest priority.
A data scientist will use statistical modeling to discover abstract topics and to provide a list of the top words for each category to help the auditors assess the relevance of the topic.
Which algorithms are best suited to this scenario? (Choose two.)

A. Latent Dirichlet allocation (LDA)
B. Random Forest classifier
C. Neural topic modeling (NTM)
D. Linear support vector machine
E. Linear regression

Answer: A, C

Explanation:
The algorithms that are best suited to this scenario are latent Dirichlet allocation (LDA) and neural topic modeling (NTM), as they are both unsupervised learning methods that can discover abstract topics from a collection of text documents. LDA and NTM can provide a list of the top words for each topic, as well as the topic distribution for each document, which can help the auditors assess the relevance and priority of the topic12.

The other options are not suitable because:
Option B: A random forest classifier is a supervised learning method that can perform classification or regression tasks by using an ensemble of decision trees. A random forest classifier is not suitable for discovering abstract topics from text documents, as it requires labeled data and predefined classes3.
Option D: A linear support vector machine is a supervised learning method that can perform classification or regression tasks by using a linear function that separates the data into different classes. A linear support vector machine is not suitable for discovering abstract topics from text documents, as it requires labeled data and predefined classes4.
Option E: A linear regression is a supervised learning method that can perform regression tasks by using a linear function that models the relationship between a dependent variable and one or more independent variables. A linear regression is not suitable for discovering abstract topics from text documents, as it requires labeled data and a continuous output variable5.
Reference:
1: Latent Dirichlet Allocation
2: Neural Topic Modeling
3: Random Forest Classifier
4: Linear Support Vector Machine
5: Linear Regression

QUESTION 223
A company wants to predict the classification of documents that are created from an application. New documents are saved to an Amazon S3 bucket every 3 seconds. The company has developed three versions of a machine learning (ML) model within Amazon SageMaker to classify document text. The company wants to deploy these three versions to predict the classification of each document.
Which approach will meet these requirements with the LEAST operational overhead?

A. Configure an S3 event notification that invokes an AWS Lambda function when new documents are created. Configure the Lambda function to create three SageMaker batch transform jobs, one batch transform job for each model for each document.
B. Deploy all the models to a single SageMaker endpoint. Treat each model as a production variant. Configure an S3 event notification that invokes an AWS Lambda function when new documents are created. Configure the Lambda function to call each production variant and return the results of each model.
C. Deploy each model to its own SageMaker endpoint Configure an S3 event notification that invokes an AWS Lambda function when new documents are created. Configure the Lambda function to call each endpoint and return the results of each model.
D. Deploy each model to its own SageMaker endpoint. Create three AWS Lambda functions. Configure each Lambda function to call a different endpoint and return the results. Configure three S3 event notifications to invoke the Lambda functions when new documents are created.

Answer: B

Explanation:
The approach that will meet the requirements with the least operational overhead is to deploy all the models to a single SageMaker endpoint, treat each model as a production variant, configure an S3 event notification that invokes an AWS Lambda function when new documents are created, and configure the Lambda function to call each production variant and return the results of each model. This approach involves the following steps:
Deploy all the models to a single SageMaker endpoint. Amazon SageMaker is a service that can build, train, and deploy machine learning models. Amazon SageMaker can deploy multiple models to a single endpoint, which is a web service that can serve predictions from the models. Each model can be treated as a production variant, which is a version of the model that runs on one or more instances. Amazon SageMaker can distribute the traffic among the production variants according to the specified weights1.
Treat each model as a production variant. Amazon SageMaker can deploy multiple models to a single endpoint, which is a web service that can serve predictions from the models. Each model can be treated as a production variant, which is a version of the model that runs on one or more instances. Amazon SageMaker can distribute the traffic among the production variants according to the specified weights1.
Configure an S3 event notification that invokes an AWS Lambda function when new documents are created. Amazon S3 is a service that can store and retrieve any amount of data. Amazon S3 can send event notifications when certain actions occur on the objects in a bucket, such as object creation, deletion, or modification. Amazon S3 can invoke an AWS Lambda function as a destination for the event notifications. AWS Lambda is a service that can run code without provisioning or managing servers2.
Configure the Lambda function to call each production variant and return the results of each model. AWS Lambda can execute the code that can call the SageMaker endpoint and specify the production variant to invoke. AWS Lambda can use the AWS SDK or the SageMaker Runtime API to send requests to the endpoint and receive the predictions from the models. AWS Lambda can return the results of each model as a response to the event notification3.
The other options are not suitable because:
Option A: Configuring an S3 event notification that invokes an AWS Lambda function when new documents are created, configuring the Lambda function to create three SageMaker batch transform jobs, one batch transform job for each model for each document, will incur more operational overhead than using a single SageMaker endpoint. Amazon SageMaker batch transform is a service that can process large datasets in batches and store the predictions in Amazon S3. Amazon SageMaker batch transform is not suitable for real-time inference, as it introduces a delay between the request and the response. Moreover, creating three batch transform jobs for each document will increase the complexity and cost of the solution4.
Option C: Deploying each model to its own SageMaker endpoint, configuring an S3 event notification that invokes an AWS Lambda function when new documents are created, configuring the Lambda function to call each endpoint and return the results of each model, will incur more operational

overhead than using a single SageMaker endpoint. Deploying each model to its own endpoint will increase the number of resources and endpoints to manage and monitor. Moreover, calling each endpoint separately will increase the latency and network traffic of the solution5.
Option D: Deploying each model to its own SageMaker endpoint, creating three AWS Lambda functions, configuring each Lambda function to call a different endpoint and return the results, configuring three S3 event notifications to invoke the Lambda functions when new documents are created, will incur more operational overhead than using a single SageMaker endpoint and a single Lambda function. Deploying each model to its own endpoint will increase the number of resources and endpoints to manage and monitor. Creating three Lambda functions will increase the complexity and cost of the solution. Configuring three S3 event notifications will increase the number of triggers and destinations to manage and monitor6.
Reference:
1: Deploying Multiple Models to a Single Endpoint - Amazon SageMaker
2: Configuring Amazon S3 Event Notifications - Amazon Simple Storage Service
3: Invoke an Endpoint - Amazon SageMaker
4: Get Inferences for an Entire Dataset with Batch Transform - Amazon SageMaker
5: Deploy a Model - Amazon SageMaker
6: AWS Lambda

QUESTION 224
A company wants to detect credit card fraud. The company has observed that an average of 2% of credit card transactions are fraudulent. A data scientist trains a classifier on a year's worth of credit card transaction dat
a. The classifier needs to identify the fraudulent transactions. The company wants to accurately capture as many fraudulent transactions as possible.
Which metrics should the data scientist use to optimize the classifier? (Select TWO.)

A. Specificity
B. False positive rate
C. Accuracy
D. Fl score
E. True positive rate

Answer: D, E

Explanation:
The F1 score is a measure of the harmonic mean of precision and recall, which are both important for fraud detection. Precision is the ratio of true positives to all predicted positives, and recall is the ratio of true positives to all actual positives. A high F1 score indicates that the classifier can correctly identify fraudulent transactions and avoid false negatives. The true positive rate is another name for recall, and it measures the proportion of fraudulent transactions that are correctly detected by the classifier. A high true positive rate means that the classifier can capture as many fraudulent

transactions as possible.
Reference:
Fraud Detection Using Machine Learning | Implementations | AWS Solutions
Detect fraudulent transactions using machine learning with Amazon SageMaker | AWS Machine Learning Blog
1. Introduction ” Reproducible Machine Learning for Credit Card Fraud Detection

---

QUESTION 225
Each morning, a data scientist at a rental car company creates insights about the previous days rental car reservation demands. The company needs to automate this process by streaming the data to Amazon S3 in near real time. The solution must detect high-demand rental cars at each of the companys locations. The solution also must create a visualization dashboard that automatically refreshes with the most recent data.
Which solution will meet these requirements with the LEAST development time?

A. Use Amazon Kinesis Data Firehose to stream the reservation data directly to Amazon S3. Detect high-demand outliers by using Amazon QuickSight ML Insights. Visualize the data in QuickSight.
B. Use Amazon Kinesis Data Streams to stream the reservation data directly to Amazon S3. Detect high-demand outliers by using the Random Cut Forest (RCF) trained model in Amazon SageMaker. Visualize the data in Amazon QuickSight.
C. Use Amazon Kinesis Data Firehose to stream the reservation data directly to Amazon S3. Detect high-demand outliers by using the Random Cut Forest (RCF) trained model in Amazon SageMaker. Visualize the data in Amazon QuickSight.
D. Use Amazon Kinesis Data Streams to stream the reservation data directly to Amazon S3. Detect high-demand outliers by using Amazon QuickSight ML Insights. Visualize the data in QuickSight.

Answer: A

Explanation:
The solution that will meet the requirements with the least development time is to use Amazon Kinesis Data Firehose to stream the reservation data directly to Amazon S3, detect high-demand outliers by using Amazon QuickSight ML Insights, and visualize the data in QuickSight. This solution does not require any custom development or ML domain expertise, as it leverages the built-in features of QuickSight ML Insights to automatically run anomaly detection and generate insights on the streaming data. QuickSight ML Insights can also create a visualization dashboard that automatically refreshes with the most recent data, and allows the data scientist to explore the outliers and their key drivers. Reference:
1: Simplify and automate anomaly detection in streaming data with Amazon Lookout for Metrics | AWS Machine Learning Blog
2: Detecting outliers with ML-powered anomaly detection - Amazon QuickSight
3: Real-time Outlier Detection Over Streaming Data - IEEE Xplore
4: Towards a deep learning-based outlier detection ¦ - Journal of Big Data

QUESTION 226
A network security vendor needs to ingest telemetry data from thousands of endpoints that run all over the world. The data is transmitted every 30 seconds in the form of records that contain 50 fields. Each record is up to 1 KB in size. The security vendor uses Amazon Kinesis Data Streams to ingest the dat
a. The vendor requires hourly summaries of the records that Kinesis Data Streams ingests. The vendor will use Amazon Athena to query the records and to generate the summaries. The Athena queries will target 7 to 12 of the available data fields.
Which solution will meet these requirements with the LEAST amount of customization to transform and store the ingested data?

A. Use AWS Lambda to read and aggregate the data hourly. Transform the data and store it in Amazon S3 by using Amazon Kinesis Data Firehose.
B. Use Amazon Kinesis Data Firehose to read and aggregate the data hourly. Transform the data and store it in Amazon S3 by using a short-lived Amazon EMR cluster.
C. Use Amazon Kinesis Data Analytics to read and aggregate the data hourly. Transform the data and store it in Amazon S3 by using Amazon Kinesis Data Firehose.
D. Use Amazon Kinesis Data Firehose to read and aggregate the data hourly. Transform the data and store it in Amazon S3 by using AWS Lambda.

Answer: C

Explanation:
The solution that will meet the requirements with the least amount of customization to transform and store the ingested data is to use Amazon Kinesis Data Analytics to read and aggregate the data hourly, transform the data and store it in Amazon S3 by using Amazon Kinesis Data Firehose. This solution leverages the built-in features of Kinesis Data Analytics to perform SQL queries on streaming data and generate hourly summaries. Kinesis Data Analytics can also output the transformed data to Kinesis Data Firehose, which can then deliver the data to S3 in a specified format and partitioning scheme. This solution does not require any custom code or additional infrastructure to process the data. The other solutions either require more customization (such as using Lambda or EMR) or do not meet the requirement of aggregating the data hourly (such as using Lambda to read the data from Kinesis Data Streams). Reference:
1: Boosting Resiliency with an ML-based Telemetry Analytics Architecture | AWS Architecture Blog
2: AWS Cloud Data Ingestion Patterns and Practices
3: IoT ingestion and Machine Learning analytics pipeline with AWS IoT ¦
4: AWS IoT Data Ingestion Simplified 101: The Complete Guide - Hevo Data

QUESTION 227
A machine learning (ML) specialist uploads 5 TB of data to an Amazon SageMaker Studio environment. The ML specialist performs initial data cleansing. Before the ML specialist begins to

train a model, the ML specialist needs to create and view an analysis report that details potential bias in the uploaded data.
Which combination of actions will meet these requirements with the LEAST operational overhead? (Choose two.)

A. Use SageMaker Clarify to automatically detect data bias
B. Turn on the bias detection option in SageMaker Ground Truth to automatically analyze data features.
C. Use SageMaker Model Monitor to generate a bias drift report.
D. Configure SageMaker Data Wrangler to generate a bias report.
E. Use SageMaker Experiments to perform a data check

Answer: A, D

Explanation:
The combination of actions that will meet the requirements with the least operational overhead is to use SageMaker Clarify to automatically detect data bias and to configure SageMaker Data Wrangler to generate a bias report. SageMaker Clarify is a feature of Amazon SageMaker that provides machine learning (ML) developers with tools to gain greater insights into their ML training data and models. SageMaker Clarify can detect potential bias during data preparation, after model training, and in your deployed model. For instance, you can check for bias related to age in your dataset or in your trained model and receive a detailed report that quantifies different types of potential bias1. SageMaker Data Wrangler is another feature of Amazon SageMaker that enables you to prepare data for machine learning (ML) quickly and easily. You can use SageMaker Data Wrangler to identify potential bias during data preparation without having to write your own code. You specify input features, such as gender or age, and SageMaker Data Wrangler runs an analysis job to detect potential bias in those features. SageMaker Data Wrangler then provides a visual report with a description of the metrics and measurements of potential bias so that you can identify steps to remediate the bias2. The other actions either require more customization (such as using SageMaker Model Monitor or SageMaker Experiments) or do not meet the requirement of detecting data bias (such as using SageMaker Ground Truth). Reference:
1: Bias Detection and Model Explainability " Amazon Web Services
2: Amazon SageMaker Data Wrangler " Amazon Web Services

QUESTION 228
A medical device company is building a machine learning (ML) model to predict the likelihood of device recall based on customer data that the company collects from a plain text survey. One of the survey questions asks which medications the customer is taking. The data for this field contains the names of medications that customers enter manually. Customers misspell some of the medication names. The column that contains the medication name data gives a categorical feature with high cardinality but redundancy.
What is the MOST effective way to encode this categorical feature into a numeric feature?

A. Spell check the column. Use Amazon SageMaker one-hot encoding on the column to transform a categorical feature to a numerical feature.
B. Fix the spelling in the column by using char-RNN. Use Amazon SageMaker Data Wrangler one-hot encoding to transform a categorical feature to a numerical feature.
C. Use Amazon SageMaker Data Wrangler similarity encoding on the column to create embeddings Of vectors Of real numbers.
D. Use Amazon SageMaker Data Wrangler ordinal encoding on the column to encode categories into an integer between O and the total number Of categories in the column.

Answer: C

Explanation:
The most effective way to encode this categorical feature into a numeric feature is to use Amazon SageMaker Data Wrangler similarity encoding on the column to create embeddings of vectors of real numbers. Similarity encoding is a technique that transforms categorical features into numerical features by computing the similarity between the categories. Similarity encoding can handle high cardinality and redundancy in categorical features, as it can group similar categories together based on their string similarity. For example, if the column contains the values œaspirin , œasprin , and œibuprofen , similarity encoding will assign a high similarity score to œaspirin and œasprin , and a low similarity score to œibuprofen . Similarity encoding can also create embeddings of vectors of real numbers, which can be used as input for machine learning models. Amazon SageMaker Data Wrangler is a feature of Amazon SageMaker that enables you to prepare data for machine learning quickly and easily. You can use SageMaker Data Wrangler to apply similarity encoding to a column of categorical data, and generate embeddings of vectors of real numbers that capture the similarity between the categories1. The other options are either less effective or more complex to implement. Spell checking the column and using one-hot encoding would require additional steps and resources, and may not capture all the misspellings or redundancies. One-hot encoding would also create a large number of features, which could increase the dimensionality and sparsity of the dat
a. Ordinal encoding would assign an arbitrary order to the categories, which could introduce bias or noise in the data. Reference:
1: Amazon SageMaker Data Wrangler " Amazon Web Services

QUESTION 229
A manufacturing company wants to create a machine learning (ML) model to predict when equipment is likely to fail. A data science team already constructed a deep learning model by using TensorFlow and a custom Python script in a local environment. The company wants to use Amazon SageMaker to train the model.
Which TensorFlow estimator configuration will train the model MOST cost-effectively?

A. Turn on SageMaker Training Compiler by adding compiler_config=TrainingCompilerConfig() as a parameter. Pass the script to the estimator in the call to the TensorFlow fit() method.

B. Turn on SageMaker Training Compiler by adding compiler_config=TrainingCompilerConfig() as a parameter. Turn on managed spot training by setting the use_spot_instances parameter to True. Pass the script to the estimator in the call to the TensorFlow fit() method.
C. Adjust the training script to use distributed data parallelism. Specify appropriate values for the distribution parameter. Pass the script to the estimator in the call to the TensorFlow fit() method.
D. Turn on SageMaker Training Compiler by adding compiler_config=TrainingCompilerConfig() as a parameter. Set the MaxWaitTimeInSeconds parameter to be equal to the MaxRuntimeInSeconds parameter. Pass the script to the estimator in the call to the TensorFlow fit() method.

Answer: B

Explanation:
The TensorFlow estimator configuration that will train the model most cost-effectively is to turn on SageMaker Training Compiler by adding compiler_config=TrainingCompilerConfig() as a parameter, turn on managed spot training by setting the use_spot_instances parameter to True, and pass the script to the estimator in the call to the TensorFlow fit() method. This configuration will optimize the model for the target hardware platform, reduce the training cost by using Amazon EC2 Spot Instances, and use the custom Python script without any modification.
SageMaker Training Compiler is a feature of Amazon SageMaker that enables you to optimize your TensorFlow, PyTorch, and MXNet models for inference on a variety of target hardware platforms. SageMaker Training Compiler can improve the inference performance and reduce the inference cost of your models by applying various compilation techniques, such as operator fusion, quantization, pruning, and graph optimization. You can enable SageMaker Training Compiler by adding compiler_config=TrainingCompilerConfig() as a parameter to the TensorFlow estimator constructor1.
Managed spot training is another feature of Amazon SageMaker that enables you to use Amazon EC2 Spot Instances for training your machine learning models. Amazon EC2 Spot Instances let you take advantage of unused EC2 capacity in the AWS Cloud. Spot Instances are available at up to a 90% discount compared to On-Demand prices. You can use Spot Instances for various fault-tolerant and flexible applications. You can enable managed spot training by setting the use_spot_instances parameter to True and specifying the max_wait and max_run parameters in the TensorFlow estimator constructor2.
The TensorFlow estimator is a class in the SageMaker Python SDK that allows you to train and deploy TensorFlow models on SageMaker. You can use the TensorFlow estimator to run your own Python script on SageMaker, without any modification. You can pass the script to the estimator in the call to the TensorFlow fit() method, along with the location of your input data. The fit() method starts a SageMaker training job and runs your script as the entry point in the training containers3.
The other options are either less cost-effective or more complex to implement. Adjusting the training script to use distributed data parallelism would require modifying the script and specifying appropriate values for the distribution parameter, which could increase the development time and complexity. Setting the MaxWaitTimeInSeconds parameter to be equal to the MaxRuntimeInSeconds parameter would not reduce the cost, as it would only specify the maximum duration of the training job, regardless of the instance type.

Reference:
1: Optimize TensorFlow, PyTorch, and MXNet models for deployment using Amazon SageMaker Training Compiler | AWS Machine Learning Blog
2: Managed Spot Training: Save Up to 90% On Your Amazon SageMaker Training Jobs | AWS Machine Learning Blog
3: sagemaker.tensorflow ” sagemaker 2.66.0 documentation

QUESTION 230
A company is creating an application to identify, count, and classify animal images that are uploaded to the companys website. The company is using the Amazon SageMaker image classification algorithm with an ImageNetV2 convolutional neural network (CNN). The solution works well for most animal images but does not recognize many animal species that are less common.
The company obtains 10,000 labeled images of less common animal species and stores the images in Amazon S3. A machine learning (ML) engineer needs to incorporate the images into the model by using Pipe mode in SageMaker.
Which combination of steps should the ML engineer take to train the model? (Choose two.)

A. Use a ResNet model. Initiate full training mode by initializing the network with random weights.
B. Use an Inception model that is available with the SageMaker image classification algorithm.
C. Create a .lst file that contains a list of image files and corresponding class labels. Upload the .lst file to Amazon S3.
D. Initiate transfer learning. Train the model by using the images of less common species.
E. Use an augmented manifest file in JSON Lines format.

Answer: C, D

Explanation:
The combination of steps that the ML engineer should take to train the model are to create a .lst file that contains a list of image files and corresponding class labels, upload the .lst file to Amazon S3, and initiate transfer learning by training the model using the images of less common species. This approach will allow the ML engineer to leverage the existing ImageNetV2 CNN model and fine-tune it with the new data using Pipe mode in SageMaker.
A .lst file is a text file that contains a list of image files and corresponding class labels, separated by tabs. The .lst file format is required for using the SageMaker image classification algorithm with Pipe mode. Pipe mode is a feature of SageMaker that enables streaming data directly from Amazon S3 to the training instances, without downloading the data first. Pipe mode can reduce the startup time, improve the I/O throughput, and enable training on large datasets that exceed the disk size limit. To use Pipe mode, the ML engineer needs to upload the .lst file to Amazon S3 and specify the S3 path as the input data channel for the training job1.
Transfer learning is a technique that enables reusing a pre-trained model for a new task by finetuning the model parameters with new data. Transfer learning can save time and computational resources, as well as improve the performance of the model, especially when the new task is similar

to the original task. The SageMaker image classification algorithm supports transfer learning by allowing the ML engineer to specify the number of output classes and the number of layers to be retrained. The ML engineer can use the existing ImageNetV2 CNN model, which is trained on 1,000 classes of common objects, and fine-tune it with the new data of less common animal species, which is a similar task2.
The other options are either less effective or not supported by the SageMaker image classification algorithm. Using a ResNet model and initiating full training mode would require training the model from scratch, which would take more time and resources than transfer learning. Using an Inception model is not possible, as the SageMaker image classification algorithm only supports ResNet and ImageNetV2 models. Using an augmented manifest file in JSON Lines format is not compatible with Pipe mode, as Pipe mode only supports .lst files for image classification1.
Reference:
1: Using Pipe input mode for Amazon SageMaker algorithms | AWS Machine Learning Blog
2: Image Classification Algorithm - Amazon SageMaker

QUESTION 231
A credit card company wants to identify fraudulent transactions in real time. A data scientist builds a machine learning model for this purpose. The transactional data is captured and stored in Amazon S3. The historic data is already labeled with two classes: fraud (positive) and fair transactions (negative). The data scientist removes all the missing data and builds a classifier by using the XGBoost algorithm in Amazon SageMaker. The model produces the following results:
True positive rate (TPR): 0.700
False negative rate (FNR): 0.300
True negative rate (TNR): 0.977
False positive rate (FPR): 0.023
Overall accuracy: 0.949
Which solution should the data scientist use to improve the performance of the model?

A. Apply the Synthetic Minority Oversampling Technique (SMOTE) on the minority class in the training dataset. Retrain the model with the updated training data.
B. Apply the Synthetic Minority Oversampling Technique (SMOTE) on the majority class in the training dataset. Retrain the model with the updated training data.
C. Undersample the minority class.
D. Oversample the majority class.

Answer: A

Explanation:
The solution that the data scientist should use to improve the performance of the model is to apply the Synthetic Minority Oversampling Technique (SMOTE) on the minority class in the training dataset, and retrain the model with the updated training data. This solution can address the problem of class imbalance in the dataset, which can affect the models ability to learn from the rare but

important positive class (fraud).
Class imbalance is a common issue in machine learning, especially for classification tasks. It occurs when one class (usually the positive or target class) is significantly underrepresented in the dataset compared to the other class (usually the negative or non-target class). For example, in the credit card fraud detection problem, the positive class (fraud) is much less frequent than the negative class (fair transactions). This can cause the model to be biased towards the majority class, and fail to capture the characteristics and patterns of the minority class. As a result, the model may have a high overall accuracy, but a low recall or true positive rate for the minority class, which means it misses many fraudulent transactions.
SMOTE is a technique that can help mitigate the class imbalance problem by generating synthetic samples for the minority class. SMOTE works by finding the k-nearest neighbors of each minority class instance, and randomly creating new instances along the line segments connecting them. This way, SMOTE can increase the number and diversity of the minority class instances, without duplicating or losing any information. By applying SMOTE on the minority class in the training dataset, the data scientist can balance the classes and improve the models performance on the positive class1.
The other options are either ineffective or counterproductive. Applying SMOTE on the majority class would not balance the classes, but increase the imbalance and the size of the dataset.
Undersampling the minority class would reduce the number of instances available for the model to learn from, and potentially lose some important information. Oversampling the majority class would also increase the imbalance and the size of the dataset, and introduce redundancy and overfitting.
Reference:
1: SMOTE for Imbalanced Classification with Python - Machine Learning Mastery

---

QUESTION 232
A company processes millions of orders every day. The company uses Amazon DynamoDB tables to store order information. When customers submit new orders, the new orders are immediately added to the DynamoDB tables. New orders arrive in the DynamoDB tables continuously.
A data scientist must build a peak-time prediction solution. The data scientist must also create an Amazon OuickSight dashboard to display near real-lime order insights. The data scientist needs to build a solution that will give QuickSight access to the data as soon as new order information arrives.
Which solution will meet these requirements with the LEAST delay between when a new order is processed and when QuickSight can access the new order information?

A. Use AWS Glue to export the data from Amazon DynamoDB to Amazon S3. Configure OuickSight to access the data in Amazon S3.
B. Use Amazon Kinesis Data Streams to export the data from Amazon DynamoDB to Amazon S3. Configure OuickSight to access the data in Amazon S3.
C. Use an API call from OuickSight to access the data that is in Amazon DynamoDB directly
D. Use Amazon Kinesis Data Firehose to export the data from Amazon DynamoDB to Amazon S3. Configure OuickSight to access the data in Amazon S3.

Answer: B

Explanation:
The best solution for this scenario is to use Amazon Kinesis Data Streams to export the data from Amazon DynamoDB to Amazon S3, and then configure QuickSight to access the data in Amazon S3. This solution has the following advantages:
It allows near real-time data ingestion from DynamoDB to S3 using Kinesis Data Streams, which can capture and process data continuously and at scale1.
It enables QuickSight to access the data in S3 using the Athena connector, which supports federated queries to multiple data sources, including Kinesis Data Streams2.
It avoids the need to create and manage a Lambda function or a Glue crawler, which are required for the other solutions.
The other solutions have the following drawbacks:
Using AWS Glue to export the data from DynamoDB to S3 introduces additional latency and complexity, as Glue is a batch-oriented service that requires scheduling and configuration3.
Using an API call from QuickSight to access the data in DynamoDB directly is not possible, as QuickSight does not support direct querying of DynamoDB4.
Using Kinesis Data Firehose to export the data from DynamoDB to S3 is less efficient and flexible than using Kinesis Data Streams, as Firehose does not support custom data processing or transformation, and has a minimum buffer interval of 60 seconds5.
Reference:
1: Amazon Kinesis Data Streams - Amazon Web Services
2: Visualize Amazon DynamoDB insights in Amazon QuickSight using the Amazon Athena DynamoDB connector and AWS Glue | AWS Big Data Blog
3: AWS Glue - Amazon Web Services
4: Visualising your Amazon DynamoDB data with Amazon QuickSight - DEV Community
5: Amazon Kinesis Data Firehose - Amazon Web Services

QUESTION 233
A retail company wants to build a recommendation system for the company's website. The system needs to provide recommendations for existing users and needs to base those recommendations on each user's past browsing history. The system also must filter out any items that the user previously purchased.
Which solution will meet these requirements with the LEAST development effort?

A. Train a model by using a user-based collaborative filtering algorithm on Amazon SageMaker. Host the model on a SageMaker real-time endpoint. Configure an Amazon API Gateway API and an AWS Lambda function to handle real-time inference requests that the web application sends. Exclude the items that the user previously purchased from the results before sending the results back to the web application.
B. Use an Amazon Personalize PERSONALIZED_RANKING recipe to train a model. Create a real-time filter to exclude items that the user previously purchased. Create and deploy a campaign on Amazon

Personalize. Use the GetPersonalizedRanking API operation to get the real-time recommendations.
C. Use an Amazon Personalize USER_ PERSONAL IZATION recipe to train a model Create a real-time filter to exclude items that the user previously purchased. Create and deploy a campaign on Amazon Personalize. Use the GetRecommendations API operation to get the real-time recommendations.
D. Train a neural collaborative filtering model on Amazon SageMaker by using GPU instances. Host the model on a SageMaker real-time endpoint. Configure an Amazon API Gateway API and an AWS Lambda function to handle real-time inference requests that the web application sends. Exclude the items that the user previously purchased from the results before sending the results back to the web application.

Answer: C

Explanation:
Amazon Personalize is a fully managed machine learning service that makes it easy for developers to create personalized user experiences at scale. It uses the same recommender system technology that Amazon uses to create its own personalized recommendations. Amazon Personalize provides several pre-built recipes that can be used to train models for different use cases. The USER_PERSONALIZATION recipe is designed to provide personalized recommendations for existing users based on their past interactions with items. The PERSONALIZED_RANKING recipe is designed to re-rank a list of items for a user based on their preferences. The USER_PERSONALIZATION recipe is more suitable for this use case because it can generate recommendations for each user without requiring a list of candidate items. To filter out the items that the user previously purchased, a realtime filter can be created and applied to the campaign. A real-time filter is a dynamic filter that uses the latest interaction data to exclude items from the recommendations. By using Amazon Personalize, the development effort is minimized because it handles the data processing, model training, and deployment automatically. The web application can use the GetRecommendations API operation to get the real-time recommendations from the campaign. Reference:
Amazon Personalize
What is Amazon Personalize?
USER_PERSONALIZATION recipe
PERSONALIZED_RANKING recipe
Filtering recommendations
GetRecommendations API operation

QUESTION 234
A data engineer is preparing a dataset that a retail company will use to predict the number of visitors to stores. The data engineer created an Amazon S3 bucket. The engineer subscribed the S3 bucket to an AWS Data Exchange data product for general economic indicators. The data engineer wants to join the economic indicator data to an existing table in Amazon Athena to merge with the business dat
a. All these transformations must finish running in 30-60 minutes.
Which solution will meet these requirements MOST cost-effectively?

A. Configure the AWS Data Exchange product as a producer for an Amazon Kinesis data stream. Use an Amazon Kinesis Data Firehose delivery stream to transfer the data to Amazon S3 Run an AWS Glue job that will merge the existing business data with the Athena table. Write the result set back to Amazon S3.
B. Use an S3 event on the AWS Data Exchange S3 bucket to invoke an AWS Lambda function. Program the Lambda function to use Amazon SageMaker Data Wrangler to merge the existing business data with the Athena table. Write the result set back to Amazon S3.
C. Use an S3 event on the AWS Data Exchange S3 bucket to invoke an AWS Lambda Function Program the Lambda function to run an AWS Glue job that will merge the existing business data with the Athena table Write the results back to Amazon S3.
D. Provision an Amazon Redshift cluster. Subscribe to the AWS Data Exchange product and use the product to create an Amazon Redshift Table Merge the data in Amazon Redshift. Write the results back to Amazon S3.

Answer: B

Explanation:
The most cost-effective solution is to use an S3 event to trigger a Lambda function that uses SageMaker Data Wrangler to merge the data. This solution avoids the need to provision and manage any additional resources, such as Kinesis streams, Firehose delivery streams, Glue jobs, or Redshift clusters. SageMaker Data Wrangler provides a visual interface to import, prepare, transform, and analyze data from various sources, including AWS Data Exchange products. It can also export the data preparation workflow to a Python script that can be executed by a Lambda function. This solution can meet the time requirement of 30-60 minutes, depending on the size and complexity of the data.
Reference:
Using Amazon S3 Event Notifications
Prepare ML Data with Amazon SageMaker Data Wrangler
AWS Lambda Function

QUESTION 235
A social media company wants to develop a machine learning (ML) model to detect Inappropriate or offensive content in images. The company has collected a large dataset of labeled images and plans to use the built-in Amazon SageMaker image classification algorithm to train the model. The company also intends to use SageMaker pipe mode to speed up the training.
...company splits the dataset into training, validation, and testing datasets. The company stores the training and validation images in folders that are named Training and Validation, respectively. The folder ...ain subfolders that correspond to the names of the dataset classes. The company resizes the images to the same sue and generates two input manifest files named training.1st and validation.1st, for the ..ing dataset and the validation dataset. respectively. Finally, the company creates two separate Amazon S3 buckets for uploads of the training dataset and the validation dataset.
...h additional data preparation steps should the company take before uploading the files to Amazon S3?

A. Generate two Apache Parquet files, training.parquet and validation.parquet. by reading the images into a Pandas data frame and storing the data frame as a Parquet file. Upload the Parquet files to the training S3 bucket
B. Compress the training and validation directories by using the Snappy compression library Upload the manifest and compressed files to the training S3 bucket
C. Compress the training and validation directories by using the gzip compression library. Upload the manifest and compressed files to the training S3 bucket.
D. Generate two RecordIO files, training rec and validation.rec. from the manifest files by using the im2rec Apache MXNet utility tool. Upload the RecordlO files to the training S3 bucket.

Answer: D

Explanation:
The SageMaker image classification algorithm supports both RecordIO and image content types for training in file mode, and supports the RecordIO content type for training in pipe mode1. However, the algorithm also supports training in pipe mode using the image files without creating RecordIO files, by using the augmented manifest format2. In this case, the company should generate

---

QUESTION 236
A company operates large cranes at a busy port. The company plans to use machine learning (ML) for predictive maintenance of the cranes to avoid unexpected breakdowns and to improve productivity. The company already uses sensor data from each crane to monitor the health of the cranes in real time. The sensor data includes rotation speed, tension, energy consumption, vibration, pressure, and ¦perature for each crane. The company contracts AWS ML experts to implement an ML solution.
Which potential findings would indicate that an ML-based solution is suitable for this scenario? (Select TWO.)

A. The historical sensor data does not include a significant number of data points and attributes for certain time periods.
B. The historical sensor data shows that simple rule-based thresholds can predict crane failures.
C. The historical sensor data contains failure data for only one type of crane model that is in operation and lacks failure data of most other types of crane that are in operation.
D. The historical sensor data from the cranes are available with high granularity for the last 3 years.
E. The historical sensor data contains most common types of crane failures that the company wants to predict.

Answer: DE

Explanation:
The best indicators that an ML-based solution is suitable for this scenario are D and E, because they imply that the historical sensor data is sufficient and relevant for building a predictive maintenance

model. This model can use machine learning techniques such as regression, classification, or anomaly detection to learn from the past data and forecast future failures or issues12. Having high granularity and diversity of data can improve the accuracy and generalization of the model, as well as enable the detection of complex patterns and relationships that are not captured by simple rulebased thresholds3.
The other options are not good indicators that an ML-based solution is suitable, because they suggest that the historical sensor data is incomplete, inconsistent, or inadequate for building a predictive maintenance model. These options would require additional data collection, preprocessing, or augmentation to overcome the data quality issues and ensure that the model can handle different scenarios and types of cranes4 .
Reference:
1: Machine Learning Techniques for Predictive Maintenance
2: A Guide to Predictive Maintenance & Machine Learning
3: Machine Learning for Predictive Maintenance: Reinventing Asset Upkeep
4: Predictive Maintenance with Machine Learning: A Complete Guide
: [Machine Learning for Predictive Maintenance - AWS Online Tech Talks]

QUESTION 237
A company wants to create an artificial intelligence (Al) yoga instructor that can lead large classes of students. The company needs to create a feature that can accurately count the number of students who are in a class. The company also needs a feature that can differentiate students who are performing a yoga stretch correctly from students who are performing a stretch incorrectly.
...etermine whether students are performing a stretch correctly, the solution needs to measure the location and angle of each student's arms and legs A data scientist must use Amazon SageMaker to ...ss video footage of a yoga class by extracting image frames and applying computer vision models.
Which combination of models will meet these requirements with the LEAST effort? (Select TWO.)

A. Image Classification
B. Optical Character Recognition (OCR)
C. Object Detection
D. Pose estimation
E. Image Generative Adversarial Networks (GANs)

Answer: CD

Explanation:
To count the number of students who are in a class, the solution needs to detect and locate each student in the video frame. Object detection is a computer vision model that can identify and locate multiple objects in an image. To differentiate students who are performing a stretch correctly from students who are performing a stretch incorrectly, the solution needs to measure the location and angle of each students arms and legs. Pose estimation is a computer vision model that can estimate the pose of a person by detecting the position and orientation of key body parts. Image classification,

OCR, and image GANs are not relevant for this use case. Reference:
Object Detection: A computer vision technique that identifies and locates objects within an image or video.
Pose Estimation: A computer vision technique that estimates the pose of a person by detecting the position and orientation of key body parts.
Amazon SageMaker: A fully managed service that provides every developer and data scientist with the ability to build, train, and deploy machine learning (ML) models quickly.

QUESTION 238
A wildlife research company has a set of images of lions and cheetahs. The company created a dataset of the images. The company labeled each image with a binary label that indicates whether an image contains a lion or cheetah. The company wants to train a model to identify whether new images contain a lion or cheetah.
.... Dh Amazon SageMaker algorithm will meet this requirement?

A. XGBoost
B. Image Classification - TensorFlow
C. Object Detection - TensorFlow
D. Semantic segmentation - MXNet

Answer: B

Explanation:
The best Amazon SageMaker algorithm for this task is Image Classification - TensorFlow. This algorithm is a supervised learning algorithm that supports transfer learning with many pretrained models from the TensorFlow Hub. Transfer learning allows the company to fine-tune one of the available pretrained models on their own dataset, even if a large amount of image data is not available. The image classification algorithm takes an image as input and outputs a probability for each provided class label. The company can choose from a variety of models, such as MobileNet, ResNet, or Inception, depending on their accuracy and speed requirements. The algorithm also supports distributed training, data augmentation, and hyperparameter tuning.
Reference:
Image Classification - TensorFlow - Amazon SageMaker
Amazon SageMaker Provides New Built-in TensorFlow Image Classification Algorithm
Image Classification with ResNet :: Amazon SageMaker Workshop
Image classification on Amazon SageMaker | by Julien Simon - Medium

---

QUESTION 239
An ecommerce company has used Amazon SageMaker to deploy a factorization machines (FM) model to suggest products for customers. The company's data science team has developed two new models by using the TensorFlow and PyTorch deep learning frameworks. The company needs to use A/B testing to evaluate the new models against the deployed model.

...required A/B testing setup is as follows:
Send 70% of traffic to the FM model, 15% of traffic to the TensorFlow model, and 15% of traffic to the Py Torch model.
For customers who are from Europe, send all traffic to the TensorFlow model
..sh architecture can the company use to implement the required A/B testing setup?

A. Create two new SageMaker endpoints for the TensorFlow and PyTorch models in addition to the existing SageMaker endpoint. Create an Application Load Balancer Create a target group for each endpoint. Configure listener rules and add weight to the target groups. To send traffic to the TensorFlow model for customers who are from Europe, create an additional listener rule to forward traffic to the TensorFlow target group.
B. Create two production variants for the TensorFlow and PyTorch models. Create an auto scaling policy and configure the desired A/B weights to direct traffic to each production variant Update the existing SageMaker endpoint with the auto scaling policy. To send traffic to the TensorFlow model for customers who are from Europe, set the TargetVariant header in the request to point to the variant name of the TensorFlow model.
C. Create two new SageMaker endpoints for the TensorFlow and PyTorch models in addition to the existing SageMaker endpoint. Create a Network Load Balancer. Create a target group for each endpoint. Configure listener rules and add weight to the target groups. To send traffic to the TensorFlow model for customers who are from Europe, create an additional listener rule to forward traffic to the TensorFlow target group.
D. Create two production variants for the TensorFlow and PyTorch models. Specify the weight for each production variant in the SageMaker endpoint configuration. Update the existing SageMaker endpoint with the new configuration. To send traffic to the TensorFlow model for customers who are from Europe, set the TargetVariant header in the request to point to the variant name of the TensorFlow model.

Answer: D

Explanation:
The correct answer is D because it allows the company to use the existing SageMaker endpoint and leverage the built-in functionality of production variants for A/B testing. Production variants can be used to test ML models that have been trained using different training datasets, algorithms, and ML frameworks; test how they perform on different instance types; or a combination of all of the above1. By specifying the weight for each production variant in the endpoint configuration, the company can control how much traffic to send to each variant. By setting the TargetVariant header in the request, the company can invoke a specific variant directly for each request2. This enables the company to implement the required A/B testing setup without creating additional endpoints or load balancers.
Reference:
1: Production variants - Amazon SageMaker
2: A/B Testing ML models in production using Amazon SageMaker | AWS Machine Learning Blog

QUESTION 240
A data scientist stores financial datasets in Amazon S3. The data scientist uses Amazon Athena to query the datasets by using SQL.
The data scientist uses Amazon SageMaker to deploy a machine learning (ML) model. The data scientist wants to obtain inferences from the model at the SageMaker endpoint However, when the data ¦. ntist attempts to invoke the SageMaker endpoint, the data scientist receives SOL statement failures The data scientist's 1AM user is currently unable to invoke the SageMaker endpoint
Which combination of actions will give the data scientist's 1AM user the ability to invoke the SageMaker endpoint? (Select THREE.)

A. Attach the AmazonAthenaFullAccess AWS managed policy to the user identity.
B. Include a policy statement for the data scientist's 1AM user that allows the 1AM user to perform the sagemaker: lnvokeEndpoint action,
C. Include an inline policy for the data scientists 1AM user that allows SageMaker to read S3 objects
D. Include a policy statement for the data scientist's 1AM user that allows the 1AM user to perform the sagemakerGetRecord action.
E. Include the SQL statement "USING EXTERNAL FUNCTION ml_function_name" in the Athena SQL query.
F. Perform a user remapping in SageMaker to map the 1AM user to another 1AM user that is on the hosted endpoint.

Answer: BCE

Explanation:
The correct combination of actions to enable the data scientists IAM user to invoke the SageMaker endpoint is B, C, and E, because they ensure that the IAM user has the necessary permissions, access, and syntax to query the ML model from Athena. These actions have the following benefits:
B: Including a policy statement for the IAM user that allows the sagemaker:InvokeEndpoint action grants the IAM user the permission to call the SageMaker Runtime InvokeEndpoint API, which is used to get inferences from the model hosted at the endpoint1.
C: Including an inline policy for the IAM user that allows SageMaker to read S3 objects enables the IAM user to access the data stored in S3, which is the source of the Athena queries2.
E: Including the SQL statement œUSING EXTERNAL FUNCTION ml_function_name in the Athena SQL query allows the IAM user to invoke the ML model as an external function from Athena, which is a feature that enables querying ML models from SQL statements3.
The other options are not correct or necessary, because they have the following drawbacks:
A: Attaching the AmazonAthenaFullAccess AWS managed policy to the user identity is not sufficient, because it does not grant the IAM user the permission to invoke the SageMaker endpoint, which is required to query the ML model4.
D: Including a policy statement for the IAM user that allows the IAM user to perform the sagemaker:GetRecord action is not relevant, because this action is used to retrieve a single record

from a feature group, which is not the case in this scenario5.
F: Performing a user remapping in SageMaker to map the IAM user to another IAM user that is on the hosted endpoint is not applicable, because this feature is only available for multi-model endpoints, which are not used in this scenario.
Reference:
1: InvokeEndpoint - Amazon SageMaker
2: Querying Data in Amazon S3 from Amazon Athena - Amazon Athena
3: Querying machine learning models from Amazon Athena using Amazon SageMaker | AWS Machine Learning Blog
4: AmazonAthenaFullAccess - AWS Identity and Access Management
5: GetRecord - Amazon SageMaker Feature Store Runtime
: [Invoke a Multi-Model Endpoint - Amazon SageMaker]

QUESTION 241
A company is using Amazon SageMaker to build a machine learning (ML) model to predict customer churn based on customer call transcripts. Audio files from customer calls are located in an onpremises VoIP system that has petabytes of recorded calls. The on-premises infrastructure has highvelocity networking and connects to the company's AWS infrastructure through a VPN connection over a 100 Mbps connection.
The company has an algorithm for transcribing customer calls that requires GPUs for inference. The company wants to store these transcriptions in an Amazon S3 bucket in the AWS Cloud for model development.
Which solution should an ML specialist use to deliver the transcriptions to the S3 bucket as quickly as possible?

A. Order and use an AWS Snowball Edge Compute Optimized device with an NVIDIA Tesla module to run the transcription algorithm. Use AWS DataSync to send the resulting transcriptions to the transcription S3 bucket.
B. Order and use an AWS Snowcone device with Amazon EC2 Inf1 instances to run the transcription algorithm Use AWS DataSync to send the resulting transcriptions to the transcription S3 bucket
C. Order and use AWS Outposts to run the transcription algorithm on GPU-based Amazon EC2 instances. Store the resulting transcriptions in the transcription S3 bucket.
D. Use AWS DataSync to ingest the audio files to Amazon S3. Create an AWS Lambda function to run the transcription algorithm on the audio files when they are uploaded to Amazon S3. Configure the function to write the resulting transcriptions to the transcription S3 bucket.

Answer: A

Explanation:
The company needs to transcribe petabytes of audio files from an on-premises VoIP system to an S3 bucket in the AWS Cloud. The transcription algorithm requires GPUs for inference, which are not available on the on-premises system. The VPN connection over a 100 Mbps connection is not

sufficient to transfer the large amount of data quickly. Therefore, the company should use an AWS Snowball Edge Compute Optimized device with an NVIDIA Tesla module to run the transcription algorithm locally and leverage the GPU power. The device can store up to 42 TB of data and can be shipped back to AWS for data ingestion. The company can use AWS DataSync to send the resulting transcriptions to the transcription S3 bucket in the AWS Cloud. This solution minimizes the network bandwidth and latency issues and enables faster data processing and transfer.
Option B is incorrect because AWS Snowcone is a small, portable, rugged, and secure edge computing and data transfer device that can store up to 8 TB of data. It is not suitable for processing petabytes of data and does not support GPU-based instances.
Option C is incorrect because AWS Outposts is a service that extends AWS infrastructure, services, APIs, and tools to virtually any data center, co-location space, or on-premises facility. It is not designed for data transfer and ingestion, and it would require additional infrastructure and maintenance costs.
Option D is incorrect because AWS DataSync is a service that makes it easy to move large amounts of data to and from AWS over the internet or AWS Direct Connect. However, using DataSync to ingest the audio files to S3 would still be limited by the network bandwidth and latency. Moreover, running the transcription algorithm on AWS Lambda would incur additional costs and complexity, and it would not leverage the GPU power that the algorithm requires.
Reference:
AWS Snowball Edge Compute Optimized
AWS DataSync
AWS Snowcone
AWS Outposts
AWS Lambda

QUESTION 242
A data scientist is building a linear regression model. The scientist inspects the dataset and notices that the mode of the distribution is lower than the median, and the median is lower than the mean. Which data transformation will give the data scientist the ability to apply a linear regression model?

A. Exponential transformation
B. Logarithmic transformation
C. Polynomial transformation
D. Sinusoidal transformation

Answer: B

Explanation:
A logarithmic transformation is a suitable data transformation for a linear regression model when the data has a skewed distribution, such as when the mode is lower than the median and the median is lower than the mean. A logarithmic transformation can reduce the skewness and make the data more symmetric and normally distributed, which are desirable properties for linear regression. A

logarithmic transformation can also reduce the effect of outliers and heteroscedasticity (unequal variance) in the data. An exponential transformation would have the opposite effect of increasing the skewness and making the data more asymmetric. A polynomial transformation may not be able to capture the nonlinearity in the data and may introduce multicollinearity among the transformed variables. A sinusoidal transformation is not appropriate for data that does not have a periodic pattern.
Reference:
Data Transformation - Scaler Topics
Linear Regression - GeeksforGeeks
Linear Regression - Scribbr

QUESTION 243
A company is planning a marketing campaign to promote a new product to existing customers. The company has data (or past promotions that are similar. The company decides to try an experiment to send a more expensive marketing package to a smaller number of customers. The company wants to target the marketing campaign to customers who are most likely to buy the new product. The experiment requires that at least 90% of the customers who are likely to purchase the new product receive the marketing materials.
...company trains a model by using the linear learner algorithm in Amazon SageMaker. The model has a recall score of 80% and a precision of 75%.
...should the company retrain the model to meet these requirements?

A. Set the target_recall hyperparameter to 90% Set the binaryclassrfier model_selection_critena hyperparameter to recall_at_target_precision.
B. Set the targetprecision hyperparameter to 90%. Set the binary classifier model selection criteria hyperparameter to precision at_jarget recall.
C. Use 90% of the historical data for training Set the number of epochs to 20.
D. Set the normalize_jabel hyperparameter to true. Set the number of classes to 2.

Answer: A

Explanation:
The best way to retrain the model to meet the requirements is to set the target_recall hyperparameter to 90% and set the binary_classifier_model_selection_criteria hyperparameter to recall_at_target_precision. This will instruct the linear learner algorithm to optimize the model for a high recall score, while maintaining a reasonable precision score. Recall is the proportion of actual positives that were identified correctly, which is important for the companys goal of reaching at least 90% of the customers who are likely to buy the new product1. Precision is the proportion of positive identifications that were actually correct, which is also relevant for the companys budget and efficiency2. By setting the target_recall to 90%, the algorithm will try to achieve a recall score of at least 90%, and by setting the binary_classifier_model_selection_criteria to recall_at_target_precision, the algorithm will select the model that has the highest recall score

among those that have a precision score equal to or higher than the target precision3. The target precision is automatically set to the median of the precision scores of all the models trained in parallel4.
The other options are not correct or optimal, because they have the following drawbacks:
B: Setting the target_precision hyperparameter to 90% and setting the binary_classifier_model_selection_criteria hyperparameter to precision_at_target_recall will optimize the model for a high precision score, while maintaining a reasonable recall score. However, this is not aligned with the companys goal of reaching at least 90% of the customers who are likely to buy the new product, as precision does not reflect how well the model identifies the actual positives1. Moreover, setting the target_precision to 90% might be too high and unrealistic for the dataset, as the current precision score is only 75%4.
C: Using 90% of the historical data for training and setting the number of epochs to 20 will not necessarily improve the recall score of the model, as it does not change the optimization objective or the model selection criteria. Moreover, using more data for training might reduce the amount of data available for validation, which is needed for selecting the best model among the ones trained in parallel3. The number of epochs is also not a decisive factor for the recall score, as it depends on the learning rate, the optimizer, and the convergence of the algorithm5.
D: Setting the normalize_label hyperparameter to true and setting the number of classes to 2 will not affect the recall score of the model, as these are irrelevant hyperparameters for binary classification problems. The normalize_label hyperparameter is only applicable for regression problems, as it controls whether the label is normalized to have zero mean and unit variance3. The number of classes hyperparameter is only applicable for multiclass classification problems, as it specifies the number of output classes3.
Reference:
1: Classification: Precision and Recall | Machine Learning | Google for Developers
2: Precision and recall - Wikipedia
3: Linear Learner Algorithm - Amazon SageMaker
4: How linear learner works - Amazon SageMaker
5: Getting hands-on with Amazon SageMaker Linear Learner - Pluralsight

---

QUESTION 244
A data scientist receives a collection of insurance claim records. Each record includes a claim ID. the final outcome of the insurance claim, and the date of the final outcome.
The final outcome of each claim is a selection from among 200 outcome categories. Some claim records include only partial information. However, incomplete claim records include only 3 or 4 outcome ...gones from among the 200 available outcome categories. The collection includes hundreds of records for each outcome category. The records are from the previous 3 years.
The data scientist must create a solution to predict the number of claims that will be in each outcome category every month, several months in advance.
Which solution will meet these requirements?

A. Perform classification every month by using supervised learning of the 20X3 outcome categories

based on claim contents.
B. Perform reinforcement learning by using claim IDs and dates Instruct the insurance agents who submit the claim records to estimate the expected number of claims in each outcome category every month
C. Perform forecasting by using claim IDs and dates to identify the expected number ot claims in each outcome category every month.
D. Perform classification by using supervised learning of the outcome categories for which partial information on claim contents is provided. Perform forecasting by using claim IDs and dates for all other outcome categories.

Answer: C

Explanation:
The best solution for this scenario is to perform forecasting by using claim IDs and dates to identify the expected number of claims in each outcome category every month. This solution has the following advantages:
It leverages the historical data of claim outcomes and dates to capture the temporal patterns and trends of the claims in each category1.
It does not require the claim contents or any other features to make predictions, which simplifies the data preparation and reduces the impact of missing or incomplete data2.
It can handle the high cardinality of the outcome categories, as forecasting models can output multiple values for each time point3.
It can provide predictions for several months in advance, which is useful for planning and budgeting purposes4.
The other solutions have the following drawbacks:
A: Performing classification every month by using supervised learning of the 200 outcome categories based on claim contents is not suitable, because it assumes that the claim contents are available and complete for all the records, which is not the case in this scenario2. Moreover, classification models usually output a single label for each input, which is not adequate for predicting the number of claims in each category3. Additionally, classification models do not account for the temporal aspect of the data, which is important for forecasting1.
B: Performing reinforcement learning by using claim IDs and dates and instructing the insurance agents who submit the claim records to estimate the expected number of claims in each outcome category every month is not feasible, because it requires a feedback loop between the model and the agents, which might not be available or reliable in this scenario5. Furthermore, reinforcement learning is more suitable for sequential decision making problems, where the model learns from its actions and rewards, rather than forecasting problems, where the model learns from historical data and outputs future values6.
D: Performing classification by using supervised learning of the outcome categories for which partial information on claim contents is provided and performing forecasting by using claim IDs and dates for all other outcome categories is not optimal, because it combines two different methods that might not be consistent or compatible with each other7. Also, this solution suffers from the same

limitations as solution A, such as the dependency on claim contents, the inability to handle multiple outputs, and the ignorance of temporal patterns123.
Reference:
1: Time Series Forecasting - Amazon SageMaker
2: Handling Missing Data for Machine Learning | AWS Machine Learning Blog
3: Forecasting vs Classification: Whats the Difference? | DataRobot
4: Amazon Forecast " Time Series Forecasting Made Easy | AWS News Blog
5: Reinforcement Learning - Amazon SageMaker
6: What is Reinforcement Learning? The Complete Guide | Edureka
7: Combining Machine Learning Models | by Will Koehrsen | Towards Data Science

---

QUESTION 245
A retail company stores 100 GB of daily transactional data in Amazon S3 at periodic intervals. The company wants to identify the schema of the transactional dat
a. The company also wants to perform transformations on the transactional data that is in Amazon S3.
The company wants to use a machine learning (ML) approach to detect fraud in the transformed data.
Which combination of solutions will meet these requirements with the LEAST operational overhead? {Select THREE.)

A. Use Amazon Athena to scan the data and identify the schema.
B. Use AWS Glue crawlers to scan the data and identify the schema.
C. Use Amazon Redshift to store procedures to perform data transformations
D. Use AWS Glue workflows and AWS Glue jobs to perform data transformations.
E. Use Amazon Redshift ML to train a model to detect fraud.
F. Use Amazon Fraud Detector to train a model to detect fraud.

Answer: BDF

Explanation:
To meet the requirements with the least operational overhead, the company should use AWS Glue crawlers, AWS Glue workflows and jobs, and Amazon Fraud Detector. AWS Glue crawlers can scan the data in Amazon S3 and identify the schema, which is then stored in the AWS Glue Data Catalog. AWS Glue workflows and jobs can perform data transformations on the data in Amazon S3 using serverless Spark or Python scripts. Amazon Fraud Detector can train a model to detect fraud using the transformed data and the companys historical fraud labels, and then generate fraud predictions using a simple API call.
Option A is incorrect because Amazon Athena is a serverless query service that can analyze data in Amazon S3 using standard SQL, but it does not perform data transformations or fraud detection.
Option C is incorrect because Amazon Redshift is a cloud data warehouse that can store and query data using SQL, but it requires provisioning and managing clusters, which adds operational overhead.

Moreover, Amazon Redshift does not provide a built-in fraud detection capability.
Option E is incorrect because Amazon Redshift ML is a feature that allows users to create, train, and deploy machine learning models using SQL commands in Amazon Redshift. However, using Amazon Redshift ML would require loading the data from Amazon S3 to Amazon Redshift, which adds complexity and cost. Also, Amazon Redshift ML does not support fraud detection as a use case.
Reference:
AWS Glue Crawlers
AWS Glue Workflows and Jobs
Amazon Fraud Detector

QUESTION 246
A data scientist uses Amazon SageMaker Data Wrangler to define and perform transformations and feature engineering on historical dat
a. The data scientist saves the transformations to SageMaker Feature Store.
The historical data is periodically uploaded to an Amazon S3 bucket. The data scientist needs to transform the new historic data and add it to the online feature store The data scientist needs to prepare the .....historic data for training and inference by using native integrations.
Which solution will meet these requirements with the LEAST development effort?

A. Use AWS Lambda to run a predefined SageMaker pipeline to perform the transformations on each new dataset that arrives in the S3 bucket.
B. Run an AWS Step Functions step and a predefined SageMaker pipeline to perform the transformations on each new dalaset that arrives in the S3 bucket
C. Use Apache Airflow to orchestrate a set of predefined transformations on each new dataset that arrives in the S3 bucket.
D. Configure Amazon EventBridge to run a predefined SageMaker pipeline to perform the transformations when a new data is detected in the S3 bucket.

Answer: D

Explanation:
The best solution is to configure Amazon EventBridge to run a predefined SageMaker pipeline to perform the transformations when a new data is detected in the S3 bucket. This solution requires the least development effort because it leverages the native integration between EventBridge and SageMaker Pipelines, which allows you to trigger a pipeline execution based on an event rule. EventBridge can monitor the S3 bucket for new data uploads and invoke the pipeline that contains the same transformations and feature engineering steps that were defined in SageMaker Data Wrangler. The pipeline can then ingest the transformed data into the online feature store for training and inference.
The other solutions are less optimal because they require more development effort and additional services. Using AWS Lambda or AWS Step Functions would require writing custom code to invoke the SageMaker pipeline and handle any errors or retries. Using Apache Airflow would require setting up

and maintaining an Airflow server and DAGs, as well as integrating with the SageMaker API.
Reference:
Amazon EventBridge and Amazon SageMaker Pipelines integration
Create a pipeline using a JSON specification
Ingest data into a feature group

QUESTION 247
A data scientist at a financial services company used Amazon SageMaker to train and deploy a model that predicts loan defaults. The model analyzes new loan applications and predicts the risk of loan default. To train the model, the data scientist manually extracted loan data from a database. The data scientist performed the model training and deployment steps in a Jupyter notebook that is hosted on SageMaker Studio notebooks. The model's prediction accuracy is decreasing over time.
Which combination of slept in the MOST operationally efficient way for the data scientist to maintain the model's accuracy? (Select TWO.)

A. Use SageMaker Pipelines to create an automated workflow that extracts fresh data, trains the model, and deploys a new version of the model.
B. Configure SageMaker Model Monitor with an accuracy threshold to check for model drift. Initiate an Amazon CloudWatch alarm when the threshold is exceeded. Connect the workflow in SageMaker Pipelines with the CloudWatch alarm to automatically initiate retraining.
C. Store the model predictions in Amazon S3 Create a daily SageMaker Processing job that reads the predictions from Amazon S3, checks for changes in model prediction accuracy, and sends an email notification if a significant change is detected.
D. Rerun the steps in the Jupyter notebook that is hosted on SageMaker Studio notebooks to retrain the model and redeploy a new version of the model.
E. Export the training and deployment code from the SageMaker Studio notebooks into a Python script. Package the script into an Amazon Elastic Container Service (Amazon ECS) task that an AWS Lambda function can initiate.

Answer: AB

Explanation:
Option A is correct because SageMaker Pipelines is a service that enables you to create and manage automated workflows for your machine learning projects. You can use SageMaker Pipelines to orchestrate the steps of data extraction, model training, and model deployment in a repeatable and scalable way1.
Option B is correct because SageMaker Model Monitor is a service that monitors the quality of your models in production and alerts you when there are deviations in the model quality. You can use SageMaker Model Monitor to set an accuracy threshold for your model and configure a CloudWatch alarm that triggers when the threshold is exceeded. You can then connect the alarm to the workflow in SageMaker Pipelines to automatically initiate retraining and deployment of a new version of the model2.

Option C is incorrect because it is not the most operationally efficient way to maintain the models accuracy. Creating a daily SageMaker Processing job that reads the predictions from Amazon S3 and checks for changes in model prediction accuracy is a manual and time-consuming process. It also requires you to write custom code to perform the data analysis and send the email notification. Moreover, it does not automatically retrain and deploy the model when the accuracy drops.
Option D is incorrect because it is not the most operationally efficient way to maintain the models accuracy. Rerunning the steps in the Jupyter notebook that is hosted on SageMaker Studio notebooks to retrain the model and redeploy a new version of the model is a manual and error-prone process. It also requires you to monitor the models performance and initiate the retraining and deployment steps yourself. Moreover, it does not leverage the benefits of SageMaker Pipelines and SageMaker Model Monitor to automate and streamline the workflow.
Option E is incorrect because it is not the most operationally efficient way to maintain the models accuracy. Exporting the training and deployment code from the SageMaker Studio notebooks into a Python script and packaging the script into an Amazon ECS task that an AWS Lambda function can initiate is a complex and cumbersome process. It also requires you to manage the infrastructure and resources for the Amazon ECS task and the AWS Lambda function. Moreover, it does not leverage the benefits of SageMaker Pipelines and SageMaker Model Monitor to automate and streamline the workflow.
Reference:
1: SageMaker Pipelines - Amazon SageMaker
2: Monitor data and model quality - Amazon SageMaker

QUESTION 248
An insurance company developed a new experimental machine learning (ML) model to replace an existing model that is in production. The company must validate the quality of predictions from the new experimental model in a production environment before the company uses the new experimental model to serve general user requests.
Which one model can serve user requests at a time. The company must measure the performance of the new experimental model without affecting the current live traffic
Which solution will meet these requirements?

A. A/B testing
B. Canary release
C. Shadow deployment
D. Blue/green deployment

Answer: C

Explanation:
The best solution for this scenario is to use shadow deployment, which is a technique that allows the company to run the new experimental model in parallel with the existing model, without exposing it to the end users. In shadow deployment, the company can route the same user requests to both

models, but only return the responses from the existing model to the users. The responses from the new experimental model are logged and analyzed for quality and performance metrics, such as accuracy, latency, and resource consumption12. This way, the company can validate the new experimental model in a production environment, without affecting the current live traffic or user experience.
The other solutions are not suitable, because they have the following drawbacks:
A: A/B testing is a technique that involves splitting the user traffic between two or more models, and comparing their outcomes based on predefined metrics. However, this technique exposes the new experimental model to a portion of the end users, which might affect their experience if the model is not reliable or consistent with the existing model3.
B: Canary release is a technique that involves gradually rolling out the new experimental model to a small subset of users, and monitoring its performance and feedback. However, this technique also exposes the new experimental model to some end users, and requires careful selection and segmentation of the user groups4.
D: Blue/green deployment is a technique that involves switching the user traffic from the existing model (blue) to the new experimental model (green) at once, after testing and verifying the new model in a separate environment. However, this technique does not allow the company to validate the new experimental model in a production environment, and might cause service disruption or inconsistency if the new model is not compatible or stable5.
Reference:
1: Shadow Deployment: A Safe Way to Test in Production | LaunchDarkly Blog
2: Shadow Deployment: A Safe Way to Test in Production | LaunchDarkly Blog
3: A/B Testing for Machine Learning Models | AWS Machine Learning Blog
4: Canary Releases for Machine Learning Models | AWS Machine Learning Blog
5: Blue-Green Deployments for Machine Learning Models | AWS Machine Learning Blog

---

QUESTION 249
An ecommerce company wants to use machine learning (ML) to monitor fraudulent transactions on its website. The company is using Amazon SageMaker to research, train, deploy, and monitor the ML models.
The historical transactions data is in a .csv file that is stored in Amazon S3 The data contains features such as the user's IP address, navigation time, average time on each page, and the number of clicks for ....session. There is no label in the data to indicate if a transaction is anomalous.
Which models should the company use in combination to detect anomalous transactions? (Select TWO.)

A. IP Insights
B. K-nearest neighbors (k-NN)
C. Linear learner with a logistic function
D. Random Cut Forest (RCF)
E. XGBoost

Answer: DE

Explanation:
To detect anomalous transactions, the company can use a combination of Random Cut Forest (RCF) and XGBoost models. RCF is an unsupervised algorithm that can detect outliers in the data by measuring the depth of each data point in a collection of random decision trees. XGBoost is a supervised algorithm that can learn from the labeled data points generated by RCF and classify them as normal or anomalous. RCF can also provide anomaly scores that can be used as features for XGBoost to improve the accuracy of the classification. Reference:
1: Amazon SageMaker Random Cut Forest
2: Amazon SageMaker XGBoost Algorithm
3: Anomaly Detection with Amazon SageMaker Random Cut Forest and Amazon SageMaker XGBoost

---

QUESTION 250
A finance company needs to forecast the price of a commodity. The company has compiled a dataset of historical daily prices. A data scientist must train various forecasting models on 80% of the dataset and must validate the efficacy of those models on the remaining 20% of the dataset.
What should the data scientist split the dataset into a training dataset and a validation dataset to compare model performance?

A. Pick a date so that 80% to the data points precede the date Assign that group of data points as the training dataset. Assign all the remaining data points to the validation dataset.
B. Pick a date so that 80% of the data points occur after the date. Assign that group of data points as the training dataset. Assign all the remaining data points to the validation dataset.
C. Starting from the earliest date in the dataset. pick eight data points for the training dataset and two data points for the validation dataset. Repeat this stratified sampling until no data points remain.
D. Sample data points randomly without replacement so that 80% of the data points are in the training dataset. Assign all the remaining data points to the validation dataset.

Answer: A

Explanation:
A Comprehensive The best way to split the dataset into a training dataset and a validation dataset is to pick a date so that 80% of the data points precede the date and assign that group of data points as the training dataset. This method preserves the temporal order of the data and ensures that the validation dataset reflects the most recent trends and patterns in the commodity price. This is important for forecasting models that rely on time series analysis and sequential data. The other methods would either introduce bias or lose information by ignoring the temporal structure of the data.
Reference:
Time Series Forecasting - Amazon SageMaker

Time Series Splitting - scikit-learn
Time Series Forecasting - Towards Data Science

QUESTION 251
A manufacturing company needs to identify returned smartphones that have been damaged by moisture. The company has an automated process that produces 2.000 diagnostic values for each phone. The database contains more than five million phone evaluations. The evaluation process is consistent, and there are no missing values in the dat
a. A machine learning (ML) specialist has trained an Amazon SageMaker linear learner ML model to classify phones as moisture damaged or not moisture damaged by using all available features. The model's F1 score is 0.6.
What changes in model training would MOST likely improve the model's F1 score? (Select TWO.)

A. Continue to use the SageMaker linear learner algorithm. Reduce the number of features with the SageMaker principal component analysis (PCA) algorithm.
B. Continue to use the SageMaker linear learner algorithm. Reduce the number of features with the scikit-iearn multi-dimensional scaling (MDS) algorithm.
C. Continue to use the SageMaker linear learner algorithm. Set the predictor type to regressor.
D. Use the SageMaker k-means algorithm with k of less than 1.000 to train the model
E. Use the SageMaker k-nearest neighbors (k-NN) algorithm. Set a dimension reduction target of less than 1,000 to train the model.

Answer: AE

Explanation:
Option A is correct because reducing the number of features with the SageMaker PCA algorithm can help remove noise and redundancy from the data, and improve the models performance. PCA is a dimensionality reduction technique that transforms the original features into a smaller set of linearly uncorrelated features called principal components. The SageMaker linear learner algorithm supports PCA as a built-in feature transformation option.
Option E is correct because using the SageMaker k-NN algorithm with a dimension reduction target of less than 1,000 can help the model learn from the similarity of the data points, and improve the models performance. k-NN is a non-parametric algorithm that classifies an input based on the majority vote of its k nearest neighbors in the feature space. The SageMaker k-NN algorithm supports dimension reduction as a built-in feature transformation option.
Option B is incorrect because using the scikit-learn MDS algorithm to reduce the number of features is not a feasible option, as MDS is a computationally expensive technique that does not scale well to large datasets. MDS is a dimensionality reduction technique that tries to preserve the pairwise distances between the original data points in a lower-dimensional space.
Option C is incorrect because setting the predictor type to regressor would change the models objective from classification to regression, which is not suitable for the given problem. A regressor model would output a continuous value instead of a binary label for each phone.

Option D is incorrect because using the SageMaker k-means algorithm with k of less than 1,000 would not help the model classify the phones, as k-means is a clustering algorithm that groups the data points into k clusters based on their similarity, without using any labels. A clustering model would not output a binary label for each phone.
Reference:
Amazon SageMaker Linear Learner Algorithm
Amazon SageMaker K-Nearest Neighbors (k-NN) Algorithm
[Principal Component Analysis - Scikit-learn]
[Multidimensional Scaling - Scikit-learn]

QUESTION 252
A company deployed a machine learning (ML) model on the company website to predict real estate prices. Several months after deployment, an ML engineer notices that the accuracy of the model has gradually decreased.
The ML engineer needs to improve the accuracy of the model. The engineer also needs to receive notifications for any future performance issues.
Which solution will meet these requirements?

A. Perform incremental training to update the model. Activate Amazon SageMaker Model Monitor to detect model performance issues and to send notifications.
B. Use Amazon SageMaker Model Governance. Configure Model Governance to automatically adjust model hyper para meters. Create a performance threshold alarm in Amazon CloudWatch to send notifications.
C. Use Amazon SageMaker Debugger with appropriate thresholds. Configure Debugger to send Amazon CloudWatch alarms to alert the team Retrain the model by using only data from the previous several months.
D. Use only data from the previous several months to perform incremental training to update the model. Use Amazon SageMaker Model Monitor to detect model performance issues and to send notifications.

Answer: A

Explanation:
The best solution to improve the accuracy of the model and receive notifications for any future performance issues is to perform incremental training to update the model and activate Amazon SageMaker Model Monitor to detect model performance issues and to send notifications.
Incremental training is a technique that allows you to update an existing model with new data without retraining the entire model from scratch. This can save time and resources, and help the model adapt to changing data patterns. Amazon SageMaker Model Monitor is a feature that continuously monitors the quality of machine learning models in production and notifies you when there are deviations in the model quality, such as data drift and anomalies. You can set up alerts that trigger actions, such as sending notifications to Amazon Simple Notification Service (Amazon SNS)

topics, when certain conditions are met.
Option B is incorrect because Amazon SageMaker Model Governance is a set of tools that help you implement ML responsibly by simplifying access control and enhancing transparency. It does not provide a mechanism to automatically adjust model hyperparameters or improve model accuracy.
Option C is incorrect because Amazon SageMaker Debugger is a feature that helps you debug and optimize your model training process by capturing relevant data and providing real-time analysis. However, using Debugger alone does not update the model or monitor its performance in production. Also, retraining the model by using only data from the previous several months may not capture the full range of data variability and may introduce bias or overfitting.
Option D is incorrect because using only data from the previous several months to perform incremental training may not be sufficient to improve the model accuracy, as explained above. Moreover, this option does not specify how to activate Amazon SageMaker Model Monitor or configure the alerts and notifications.
Reference:
Incremental training
Amazon SageMaker Model Monitor
Amazon SageMaker Model Governance
Amazon SageMaker Debugger

QUESTION 253
A university wants to develop a targeted recruitment strategy to increase new student enrollment. A data scientist gathers information about the academic performance history of students. The data scientist wants to use the data to build student profiles. The university will use the profiles to direct resources to recruit students who are likely to enroll in the university.
Which combination of steps should the data scientist take to predict whether a particular student applicant is likely to enroll in the university? (Select TWO)

A. Use Amazon SageMaker Ground Truth to sort the data into two groups named "enrolled" or "not enrolled."
B. Use a forecasting algorithm to run predictions.
C. Use a regression algorithm to run predictions.
D. Use a classification algorithm to run predictions
E. Use the built-in Amazon SageMaker k-means algorithm to cluster the data into two groups named "enrolled" or "not enrolled."

Answer: AD

Explanation:
The data scientist should use Amazon SageMaker Ground Truth to sort the data into two groups named œenrolled or œnot enrolled. This will create a labeled dataset that can be used for supervised learning. The data scientist should then use a classification algorithm to run predictions on the test data. A classification algorithm is a suitable choice for predicting a binary outcome, such as

enrollment status, based on the input features, such as academic performance. A classification algorithm will output a probability for each class label and assign the most likely label to each observation.
Reference:
Use Amazon SageMaker Ground Truth to Label Data
Classification Algorithm in Machine Learning

QUESTION 254
A company's machine learning (ML) specialist is building a computer vision model to classify 10 different traffic signs. The company has stored 100 images of each class in Amazon S3, and the company has another 10.000 unlabeled images. All the images come from dash cameras and are a size of 224 pixels * 224 pixels. After several training runs, the model is overfitting on the training data.
Which actions should the ML specialist take to address this problem? (Select TWO.)

A. Use Amazon SageMaker Ground Truth to label the unlabeled images
B. Use image preprocessing to transform the images into grayscale images.
C. Use data augmentation to rotate and translate the labeled images.
D. Replace the activation of the last layer with a sigmoid.
E. Use the Amazon SageMaker k-nearest neighbors (k-NN) algorithm to label the unlabeled images.

Answer: CE

Explanation:
Data augmentation is a technique to increase the size and diversity of the training data by applying random transformations such as rotation, translation, scaling, flipping, etc. This can help reduce overfitting and improve the generalization of the model. Data augmentation can be done using the Amazon SageMaker image classification algorithm, which supports various augmentation options such as horizontal_flip, vertical_flip, rotate, brightness, contrast, etc1
The Amazon SageMaker k-nearest neighbors (k-NN) algorithm is a supervised learning algorithm that can be used to label unlabeled data based on the similarity to the labeled data. The k-NN algorithm assigns a label to an unlabeled instance by finding the k closest labeled instances in the feature space and taking a majority vote among their labels. This can help increase the size and diversity of the training data and reduce overfitting. The k-NN algorithm can be used with the Amazon SageMaker image classification algorithm by extracting features from the images using a pre-trained model and then applying the k-NN algorithm on the feature vectors2
Using Amazon SageMaker Ground Truth to label the unlabeled images is not a good option because it is a manual and costly process that requires human annotators. Moreover, it does not address the issue of overfitting on the existing labeled data.
Using image preprocessing to transform the images into grayscale images is not a good option because it reduces the amount of information and variation in the images, which can degrade the performance of the model. Moreover, it does not address the issue of overfitting on the existing

labeled data.
Replacing the activation of the last layer with a sigmoid is not a good option because it is not suitable for a multi-class classification problem. A sigmoid activation function outputs a value between 0 and 1, which can be interpreted as a probability of belonging to a single class. However, for a multi-class classification problem, the output should be a vector of probabilities that sum up to 1, which can be achieved by using a softmax activation function.
Reference:
1: Image classification algorithm - Amazon SageMaker
2: k-nearest neighbors (k-NN) algorithm - Amazon SageMaker

QUESTION 255
A machine learning (ML) specialist is using the Amazon SageMaker DeepAR forecasting algorithm to train a model on CPU-based Amazon EC2 On-Demand instances. The model currently takes multiple hours to train. The ML specialist wants to decrease the training time of the model.
Which approaches will meet this requirement7 (SELECT TWO)

A. Replace On-Demand Instances with Spot Instances
B. Configure model auto scaling dynamically to adjust the number of instances automatically.
C. Replace CPU-based EC2 instances with GPU-based EC2 instances.
D. Use multiple training instances.
E. Use a pre-trained version of the model. Run incremental training.

Answer: CD

Explanation:
The best approaches to decrease the training time of the model are C and D, because they can improve the computational efficiency and parallelization of the training process. These approaches have the following benefits:
C: Replacing CPU-based EC2 instances with GPU-based EC2 instances can speed up the training of the DeepAR algorithm, as it can leverage the parallel processing power of GPUs to perform matrix operations and gradient computations faster than CPUs12. The DeepAR algorithm supports GPUbased EC2 instances such as ml.p2 and ml.p33.
D: Using multiple training instances can also reduce the training time of the DeepAR algorithm, as it can distribute the workload across multiple nodes and perform data parallelism4. The DeepAR algorithm supports distributed training with multiple CPU-based or GPU-based EC2 instances3.
The other options are not effective or relevant, because they have the following drawbacks:
A: Replacing On-Demand Instances with Spot Instances can reduce the cost of the training, but not necessarily the time, as Spot Instances are subject to interruption and availability5. Moreover, the DeepAR algorithm does not support checkpointing, which means that the training cannot resume from the last saved state if the Spot Instance is terminated3.
B: Configuring model auto scaling dynamically to adjust the number of instances automatically is not applicable, as this feature is only available for inference endpoints, not for training jobs6.

E: Using a pre-trained version of the model and running incremental training is not possible, as the DeepAR algorithm does not support incremental training or transfer learning3. The DeepAR algorithm requires a full retraining of the model whenever new data is added or the hyperparameters are changed7.
Reference:
1: GPU vs CPU: What Matters Most for Machine Learning? | by Louis (Whats AI) Bouchard | Towards Data Science
2: How GPUs Accelerate Machine Learning Training | NVIDIA Developer Blog
3: DeepAR Forecasting Algorithm - Amazon SageMaker
4: Distributed Training - Amazon SageMaker
5: Managed Spot Training - Amazon SageMaker
6: Automatic Scaling - Amazon SageMaker
7: How the DeepAR Algorithm Works - Amazon SageMaker

QUESTION 256
An engraving company wants to automate its quality control process for plaques. The company performs the process before mailing each customized plaque to a customer. The company has created an Amazon S3 bucket that contains images of defects that should cause a plaque to be rejected. Low-confidence predictions must be sent to an internal team of reviewers who are using Amazon Augmented Al (Amazon A2I).
Which solution will meet these requirements?

A. Use Amazon Textract for automatic processing. Use Amazon A2I with Amazon Mechanical Turk for manual review.
B. Use Amazon Rekognition for automatic processing. Use Amazon A2I with a private workforce option for manual review.
C. Use Amazon Transcribe for automatic processing. Use Amazon A2I with a private workforce option for manual review.
D. Use AWS Panorama for automatic processing Use Amazon A2I with Amazon Mechanical Turk for manual review

Answer: B

Explanation:
Amazon Rekognition is a service that provides computer vision capabilities for image and video analysis, such as object, scene, and activity detection, face and text recognition, and custom label detection. Amazon Rekognition can be used to automate the quality control process for plaques by comparing the images of the plaques with the images of defects in the Amazon S3 bucket and returning a confidence score for each defect. Amazon A2I is a service that enables human review of machine learning predictions, such as low-confidence predictions from Amazon Rekognition. Amazon A2I can be integrated with a private workforce option, which allows the engraving company to use its own internal team of reviewers to manually inspect the plaques that are flagged by Amazon

Rekognition. This solution meets the requirements of automating the quality control process, sending low-confidence predictions to an internal team of reviewers, and using Amazon A2I for manual review. Reference:
1: Amazon Rekognition documentation
2: Amazon A2I documentation
3: Amazon Rekognition Custom Labels documentation
4: Amazon A2I Private Workforce documentation

QUESTION 257
An online delivery company wants to choose the fastest courier for each delivery at the moment an order is placed. The company wants to implement this feature for existing users and new users of its application. Data scientists have trained separate models with XGBoost for this purpose, and the models are stored in Amazon S3. There is one model fof each city where the company operates.
The engineers are hosting these models in Amazon EC2 for responding to the web client requests, with one instance for each model, but the instances have only a 5% utilization in CPU and memory, ...operation engineers want to avoid managing unnecessary resources.
Which solution will enable the company to achieve its goal with the LEAST operational overhead?

A. Create an Amazon SageMaker notebook instance for pulling all the models from Amazon S3 using the boto3 library. Remove the existing instances and use the notebook to perform a SageMaker batch transform for performing inferences offline for all the possible users in all the cities. Store the results in different files in Amazon S3. Point the web client to the files.
B. Prepare an Amazon SageMaker Docker container based on the open-source multi-model server. Remove the existing instances and create a multi-model endpoint in SageMaker instead, pointing to the S3 bucket containing all the models Invoke the endpoint from the web client at runtime, specifying the TargetModel parameter according to the city of each request.
C. Keep only a single EC2 instance for hosting all the models. Install a model server in the instance and load each model by pulling it from Amazon S3. Integrate the instance with the web client using Amazon API Gateway for responding to the requests in real time, specifying the target resource according to the city of each request.
D. Prepare a Docker container based on the prebuilt images in Amazon SageMaker. Replace the existing instances with separate SageMaker endpoints. one for each city where the company operates. Invoke the endpoints from the web client, specifying the URL and EndpomtName parameter according to the city of each request.

Answer: B

Explanation:
The best solution for this scenario is to use a multi-model endpoint in Amazon SageMaker, which allows hosting multiple models on the same endpoint and invoking them dynamically at runtime. This way, the company can reduce the operational overhead of managing multiple EC2 instances and model servers, and leverage the scalability, security, and performance of SageMaker hosting

services. By using a multi-model endpoint, the company can also save on hosting costs by improving endpoint utilization and paying only for the models that are loaded in memory and the API calls that are made. To use a multi-model endpoint, the company needs to prepare a Docker container based on the open-source multi-model server, which is a framework-agnostic library that supports loading and serving multiple models from Amazon S3. The company can then create a multi-model endpoint in SageMaker, pointing to the S3 bucket containing all the models, and invoke the endpoint from the web client at runtime, specifying the TargetModel parameter according to the city of each request. This solution also enables the company to add or remove models from the S3 bucket without redeploying the endpoint, and to use different versions of the same model for different cities if needed. Reference:
Use Docker containers to build models
Host multiple models in one container behind one endpoint
Multi-model endpoints using Scikit Learn
Multi-model endpoints using XGBoost

QUESTION 258
A company builds computer-vision models that use deep learning for the autonomous vehicle industry. A machine learning (ML) specialist uses an Amazon EC2 instance that has a CPU: GPU ratio of 12:1 to train the models.
The ML specialist examines the instance metric logs and notices that the GPU is idle half of the time
The ML specialist must reduce training costs without increasing the duration of the training jobs.
Which solution will meet these requirements?

A. Switch to an instance type that has only CPUs.
B. Use a heterogeneous cluster that has two different instances groups.
C. Use memory-optimized EC2 Spot Instances for the training jobs.
D. Switch to an instance type that has a CPU GPU ratio of 6:1.

Answer: D

Explanation:
Switching to an instance type that has a CPU: GPU ratio of 6:1 will reduce the training costs by using fewer CPUs and GPUs, while maintaining the same level of performance. The GPU idle time indicates that the CPU is not able to feed the GPU with enough data, so reducing the CPU: GPU ratio will balance the workload and improve the GPU utilization. A lower CPU: GPU ratio also means less overhead for inter-process communication and synchronization between the CPU and GPU processes. Reference:
Optimizing GPU utilization for AI/ML workloads on Amazon EC2
Analyze CPU vs. GPU Performance for AWS Machine Learning

QUESTION 259
A company is building a new supervised classification model in an AWS environment. The company's

data science team notices that the dataset has a large quantity of variables Ail the variables are numeric. The model accuracy for training and validation is low. The model's processing time is affected by high latency The data science team needs to increase the accuracy of the model and decrease the processing.
How it should the data science team do to meet these requirements?

A. Create new features and interaction variables.
B. Use a principal component analysis (PCA) model.
C. Apply normalization on the feature set.
D. Use a multiple correspondence analysis (MCA) model

Answer: B

Explanation:
The best way to meet the requirements is to use a principal component analysis (PCA) model, which is a technique that reduces the dimensionality of the dataset by transforming the original variables into a smaller set of new variables, called principal components, that capture most of the variance and information in the data1. This technique has the following advantages:
It can increase the accuracy of the model by removing noise, redundancy, and multicollinearity from the data, and by enhancing the interpretability and generalization of the model23.
It can decrease the processing time of the model by reducing the number of features and the computational complexity of the model, and by improving the convergence and stability of the model45.
It is suitable for numeric variables, as it relies on the covariance or correlation matrix of the data, and it can handle a large quantity of variables, as it can extract the most relevant ones16.
The other options are not effective or appropriate, because they have the following drawbacks:
A: Creating new features and interaction variables can increase the accuracy of the model by capturing more complex and nonlinear relationships in the data, but it can also increase the processing time of the model by adding more features and increasing the computational complexity of the model7. Moreover, it can introduce more noise, redundancy, and multicollinearity in the data, which can degrade the performance and interpretability of the model8.
C: Applying normalization on the feature set can increase the accuracy of the model by scaling the features to a common range and avoiding the dominance of some features over others, but it can also decrease the processing time of the model by reducing the numerical instability and improving the convergence of the model . However, normalization alone is not enough to address the high dimensionality and high latency issues of the dataset, as it does not reduce the number of features or the variance in the data.
D: Using a multiple correspondence analysis (MCA) model is not suitable for numeric variables, as it is a technique that reduces the dimensionality of the dataset by transforming the original categorical variables into a smaller set of new variables, called factors, that capture most of the inertia and information in the data. MCA is similar to PCA, but it is designed for nominal or ordinal variables, not for continuous or interval variables.

Reference:
1: Principal Component Analysis - Amazon SageMaker
2: How to Use PCA for Data Visualization and Improved Performance in Machine Learning | by Pratik Shukla | Towards Data Science
3: Principal Component Analysis (PCA) for Feature Selection and some of its Pitfalls | by Nagesh Singh Chauhan | Towards Data Science
4: How to Reduce Dimensionality with PCA and Train a Support Vector Machine in Python | by James Briggs | Towards Data Science
5: Dimensionality Reduction and Its Applications | by Aniruddha Bhandari | Towards Data Science
6: Principal Component Analysis (PCA) in Python | by Susan Li | Towards Data Science
7: Feature Engineering for Machine Learning | by Dipanjan (DJ) Sarkar | Towards Data Science
8: Feature Engineering " How to Engineer Features and How to Get Good at It | by Parul Pandey | Towards Data Science
: [Feature Scaling for Machine Learning: Understanding the Difference Between Normalization vs. Standardization | by Benjamin Obi Tayo Ph.D. | Towards Data Science]
: [Why, How and When to Scale your Features | by George Seif | Towards Data Science]
: [Normalization vs Dimensionality Reduction | by Saurabh Annadate | Towards Data Science]
: [Multiple Correspondence Analysis - Amazon SageMaker]
: [Multiple Correspondence Analysis (MCA) | by Raul Eulogio | Towards Data Science]

QUESTION 260
A company wants to forecast the daily price of newly launched products based on 3 years of data for older product prices, sales, and rebates. The time-series data has irregular timestamps and is missing some values.
Data scientist must build a dataset to replace the missing values. The data scientist needs a solution that resamptes the data daily and exports the data for further modeling.
Which solution will meet these requirements with the LEAST implementation effort?

A. Use Amazon EMR Serveriess with PySpark.
B. Use AWS Glue DataBrew.
C. Use Amazon SageMaker Studio Data Wrangler.
D. Use Amazon SageMaker Studio Notebook with Pandas.

Answer: C

Explanation:
Amazon SageMaker Studio Data Wrangler is a visual data preparation tool that enables users to clean and normalize data without writing any code. Using Data Wrangler, the data scientist can easily import the time-series data from various sources, such as Amazon S3, Amazon Athena, or Amazon Redshift. Data Wrangler can automatically generate data insights and quality reports, which can help identify and fix missing values, outliers, and anomalies in the data. Data Wrangler also provides over 250 built-in transformations, such as resampling, interpolation, aggregation, and filtering, which can

be applied to the data with a point-and-click interface. Data Wrangler can also export the prepared data to different destinations, such as Amazon S3, Amazon SageMaker Feature Store, or Amazon SageMaker Pipelines, for further modeling and analysis. Data Wrangler is integrated with Amazon SageMaker Studio, a web-based IDE for machine learning, which makes it easy to access and use the tool. Data Wrangler is a serverless and fully managed service, which means the data scientist does not need to provision, configure, or manage any infrastructure or clusters.
Option A is incorrect because Amazon EMR Serverless is a serverless option for running big data analytics applications using open-source frameworks, such as Apache Spark. However, using Amazon EMR Serverless would require the data scientist to write PySpark code to perform the data preparation tasks, such as resampling, imputation, and aggregation. This would require more implementation effort than using Data Wrangler, which provides a visual and code-free interface for data preparation.
Option B is incorrect because AWS Glue DataBrew is another visual data preparation tool that can be used to clean and normalize data without writing code. However, DataBrew does not support timeseries data as a data type, and does not provide built-in transformations for resampling, interpolation, or aggregation of time-series data. Therefore, using DataBrew would not meet the requirements of the use case.
Option D is incorrect because using Amazon SageMaker Studio Notebook with Pandas would also require the data scientist to write Python code to perform the data preparation tasks. Pandas is a popular Python library for data analysis and manipulation, which supports time-series data and provides various methods for resampling, interpolation, and aggregation. However, using Pandas would require more implementation effort than using Data Wrangler, which provides a visual and code-free interface for data preparation.
Reference:
1: Amazon SageMaker Data Wrangler documentation
2: Amazon EMR Serverless documentation
3: AWS Glue DataBrew documentation
4: Pandas documentation

QUESTION 261
A data scientist is building a forecasting model for a retail company by using the most recent 5 years of sales records that are stored in a data warehouse. The dataset contains sales records for each of the company's stores across five commercial regions The data scientist creates a working dataset with StorelD. Region. Date, and Sales Amount as columns. The data scientist wants to analyze yearly average sales for each region. The scientist also wants to compare how each region performed compared to average sales across all commercial regions.
Which visualization will help the data scientist better understand the data trend?

A. Create an aggregated dataset by using the Pandas GroupBy function to get average sales for each year for each store. Create a bar plot, faceted by year, of average sales for each store. Add an extra bar in each facet to represent average sales.
B. Create an aggregated dataset by using the Pandas GroupBy function to get average sales for each

year for each store. Create a bar plot, colored by region and faceted by year, of average sales for each store. Add a horizontal line in each facet to represent average sales.
C. Create an aggregated dataset by using the Pandas GroupBy function to get average sales for each year for each region Create a bar plot of average sales for each region. Add an extra bar in each facet to represent average sales.
D. Create an aggregated dataset by using the Pandas GroupBy function to get average sales for each year for each region Create a bar plot, faceted by year, of average sales for each region Add a horizontal line in each facet to represent average sales.

Answer: D

Explanation:
The best visualization for this task is to create a bar plot, faceted by year, of average sales for each region and add a horizontal line in each facet to represent average sales. This way, the data scientist can easily compare the yearly average sales for each region with the overall average sales and see the trends over time. The bar plot also allows the data scientist to see the relative performance of each region within each year and across years. The other options are less effective because they either do not show the yearly trends, do not show the overall average sales, or do not group the data by region.
Reference:
pandas.DataFrame.groupby ” pandas 2.1.4 documentation
pandas.DataFrame.plot.bar ” pandas 2.1.4 documentation
Matplotlib - Bar Plot - Online Tutorials Library

QUESTION 262
A company uses sensors on devices such as motor engines and factory machines to measure parameters, temperature and pressure. The company wants to use the sensor data to predict equipment malfunctions and reduce services outages.
The Machine learning (ML) specialist needs to gather the sensors data to train a model to predict device malfunctions The ML spoctafst must ensure that the data does not contain outliers before training the ..el.
What can the ML specialist meet these requirements with the LEAST operational overhead?

A. Load the data into an Amazon SagcMaker Studio notebook. Calculate the first and third quartile Use a SageMaker Data Wrangler data (low to remove only values that are outside of those quartiles.
B. Use an Amazon SageMaker Data Wrangler bias report to find outliers in the dataset Use a Data Wrangler data flow to remove outliers based on the bias report.
C. Use an Amazon SageMaker Data Wrangler anomaly detection visualization to find outliers in the dataset. Add a transformation to a Data Wrangler data flow to remove outliers.
D. Use Amazon Lookout for Equipment to find and remove outliers from the dataset.

Answer: C

Explanation:
Amazon SageMaker Data Wrangler is a tool that helps data scientists and ML developers to prepare data for ML. One of the features of Data Wrangler is the anomaly detection visualization, which uses an unsupervised ML algorithm to identify outliers in the dataset based on statistical properties. The ML specialist can use this feature to quickly explore the sensor data and find any anomalous values that may affect the model performance. The ML specialist can then add a transformation to a Data Wrangler data flow to remove the outliers from the dataset. The data flow can be exported as a script or a pipeline to automate the data preparation process. This option requires the least operational overhead compared to the other options.
Reference:
Amazon SageMaker Data Wrangler - Amazon Web Services (AWS)
Anomaly Detection Visualization - Amazon SageMaker
Transform Data - Amazon SageMaker

QUESTION 263
A data engineer needs to provide a team of data scientists with the appropriate dataset to run machine learning training jobs. The data will be stored in Amazon S3. The data engineer is obtaining the data from an Amazon Redshift database and is using join queries to extract a single tabular dataset. A portion of the schema is as follows:
...traction Timestamp (Timeslamp)
...JName(Varchar)
...JNo (Varchar)
Th data engineer must provide the data so that any row with a CardNo value of NULL is removed. Also, the TransactionTimestamp column must be separated into a TransactionDate column and a isactionTime column Finally, the CardName column must be renamed to NameOnCard.
The data will be extracted on a monthly basis and will be loaded into an S3 bucket. The solution must minimize the effort that is needed to set up infrastructure for the ingestion and transformation. The solution must be automated and must minimize the load on the Amazon Redshift cluster
Which solution meets these requirements?

A. Set up an Amazon EMR cluster Create an Apache Spark job to read the data from the Amazon Redshift cluster and transform the data. Load the data into the S3 bucket. Schedule the job to run monthly.
B. Set up an Amazon EC2 instance with a SQL client tool, such as SQL Workbench/J. to query the data from the Amazon Redshift cluster directly. Export the resulting dataset into a We. Upload the file into the S3 bucket. Perform these tasks monthly.
C. Set up an AWS Glue job that has the Amazon Redshift cluster as the source and the S3 bucket as the destination Use the built-in transforms Filter, Map. and RenameField to perform the required transformations. Schedule the job to run monthly.
D. Use Amazon Redshift Spectrum to run a query that writes the data directly to the S3 bucket. Create an AWS Lambda function to run the query monthly

Answer: C

Explanation:
The best solution for this scenario is to set up an AWS Glue job that has the Amazon Redshift cluster as the source and the S3 bucket as the destination, and use the built-in transforms Filter, Map, and RenameField to perform the required transformations. This solution has the following advantages:
It minimizes the effort that is needed to set up infrastructure for the ingestion and transformation, as AWS Glue is a fully managed service that provides a serverless Apache Spark environment, a graphical interface to define data sources and targets, and a code generation feature to create and edit scripts1.
It automates the extraction and transformation process, as AWS Glue can schedule the job to run monthly, and handle the connection, authentication, and configuration of the Amazon Redshift cluster and the S3 bucket2.
It minimizes the load on the Amazon Redshift cluster, as AWS Glue can read the data from the cluster in parallel and use a JDBC connection that supports SSL encryption3.
It performs the required transformations, as AWS Glue can use the built-in transforms Filter, Map, and RenameField to remove the rows with NULL values, split the timestamp column into date and time columns, and rename the card name column, respectively4.
The other solutions are not optimal or suitable, because they have the following drawbacks:
A: Setting up an Amazon EMR cluster and creating an Apache Spark job to read the data from the Amazon Redshift cluster and transform the data is not the most efficient or convenient solution, as it requires more effort and resources to provision, configure, and manage the EMR cluster, and to write and maintain the Spark code5.
B: Setting up an Amazon EC2 instance with a SQL client tool to query the data from the Amazon Redshift cluster directly and export the resulting dataset into a CSV file is not a scalable or reliable solution, as it depends on the availability and performance of the EC2 instance, and the manual execution and upload of the SQL queries and the CSV file6.
D: Using Amazon Redshift Spectrum to run a query that writes the data directly to the S3 bucket and creating an AWS Lambda function to run the query monthly is not a feasible solution, as Amazon Redshift Spectrum does not support writing data to external tables or S3 buckets, only reading data from them7.
Reference:
1: What Is AWS Glue? - AWS Glue
2: Populating the Data Catalog - AWS Glue
3: Best Practices When Using AWS Glue with Amazon Redshift - AWS Glue
4: Built-In Transforms - AWS Glue
5: What Is Amazon EMR? - Amazon EMR
6: Amazon EC2 - Amazon Web Services (AWS)
7: Using Amazon Redshift Spectrum to Query External Data - Amazon Redshift

QUESTION 264
A data scientist obtains a tabular dataset that contains 150 correlated features with different ranges to build a regression model. The data scientist needs to achieve more efficient model training by implementing a solution that minimizes impact on the model's performance. The data scientist decides to perform a principal component analysis (PCA) preprocessing step to reduce the number of features to a smaller set of independent features before the data scientist uses the new features in the regression model.
Which preprocessing step will meet these requirements?

A. Use the Amazon SageMaker built-in algorithm for PCA on the dataset to transform the data
B. Load the data into Amazon SageMaker Data Wrangler. Scale the data with a Min Max Scaler transformation step Use the SageMaker built-in algorithm for PCA on the scaled dataset to transform the data.
C. Reduce the dimensionality of the dataset by removing the features that have the highest correlation Load the data into Amazon SageMaker Data Wrangler Perform a Standard Scaler transformation step to scale the data Use the SageMaker built-in algorithm for PCA on the scaled dataset to transform the data
D. Reduce the dimensionality of the dataset by removing the features that have the lowest correlation. Load the data into Amazon SageMaker Data Wrangler. Perform a Min Max Scaler transformation step to scale the data. Use the SageMaker built-in algorithm for PCA on the scaled dataset to transform the data.

Answer: B

Explanation:
Principal component analysis (PCA) is a technique for reducing the dimensionality of datasets, increasing interpretability but at the same time minimizing information loss. It does so by creating new uncorrelated variables that successively maximize variance. PCA is useful when dealing with datasets that have a large number of correlated features. However, PCA is sensitive to the scale of the features, so it is important to standardize or normalize the data before applying PC
A. Amazon
SageMaker provides a built-in algorithm for PCA that can be used to transform the data into a lowerdimensional representation. Amazon SageMaker Data Wrangler is a tool that allows data scientists to visually explore, clean, and prepare data for machine learning. Data Wrangler provides various transformation steps that can be applied to the data, such as scaling, encoding, imputing, etc. Data Wrangler also integrates with SageMaker built-in algorithms, such as PCA, to enable feature engineering and dimensionality reduction. Therefore, option B is the correct answer, as it involves scaling the data with a Min Max Scaler transformation step, which rescales the data to a range of [0, 1], and then using the SageMaker built-in algorithm for PCA on the scaled dataset to transform the data. Option A is incorrect, as it does not involve scaling the data before applying PCA, which can affect the results of the dimensionality reduction. Option C is incorrect, as it involves removing the features that have the highest correlation, which can lead to information loss and reduce the

performance of the regression model. Option D is incorrect, as it involves removing the features that have the lowest correlation, which can also lead to information loss and reduce the performance of the regression model. Reference:
Principal Component Analysis (PCA) - Amazon SageMaker
Scale data with a Min Max Scaler - Amazon SageMaker Data Wrangler
Use Amazon SageMaker built-in algorithms - Amazon SageMaker Data Wrangler

---

QUESTION 265
A financial services company wants to automate its loan approval process by building a machine learning (ML) model. Each loan data point contains credit history from a third-party data source and demographic information about the customer. Each loan approval prediction must come with a report that contains an explanation for why the customer was approved for a loan or was denied for a loan. The company will use Amazon SageMaker to build the model.
Which solution will meet these requirements with the LEAST development effort?

A. Use SageMaker Model Debugger to automatically debug the predictions, generate the explanation, and attach the explanation report.
B. Use AWS Lambda to provide feature importance and partial dependence plots. Use the plots to generate and attach the explanation report.
C. Use SageMaker Clarify to generate the explanation report. Attach the report to the predicted results.
D. Use custom Amazon Cloud Watch metrics to generate the explanation report. Attach the report to the predicted results.

Answer: C

Explanation:
The best solution for this scenario is to use SageMaker Clarify to generate the explanation report and attach it to the predicted results. SageMaker Clarify provides tools to help explain how machine learning (ML) models make predictions using a model-agnostic feature attribution approach based on SHAP values. It can also detect and measure potential bias in the data and the model. SageMaker Clarify can generate explanation reports during data preparation, model training, and model deployment. The reports include metrics, graphs, and examples that help understand the model behavior and predictions. The reports can be attached to the predicted results using the SageMaker SDK or the SageMaker API.
The other solutions are less optimal because they require more development effort and additional services. Using SageMaker Model Debugger would require modifying the training script to save the model output tensors and writing custom rules to debug and explain the predictions. Using AWS Lambda would require writing code to invoke the ML model, compute the feature importance and partial dependence plots, and generate and attach the explanation report. Using custom Amazon CloudWatch metrics would require writing code to publish the metrics, create dashboards, and generate and attach the explanation report.

Reference:
Bias Detection and Model Explainability - Amazon SageMaker Clarify - AWS
Amazon SageMaker Clarify Model Explainability
Amazon SageMaker Clarify: Machine Learning Bias Detection and Explainability
GitHub - aws/amazon-sagemaker-clarify: Fairness Aware Machine Learning

---

QUESTION 266
An obtain relator collects the following data on customer orders: demographics, behaviors, location, shipment progress, and delivery time. A data scientist joins all the collected datasets. The result is a single dataset that includes 980 variables.
The data scientist must develop a machine learning (ML) model to identify groups of customers who are likely to respond to a marketing campaign.
Which combination of algorithms should the data scientist use to meet this requirement? (Select TWO.)

A. Latent Dirichlet Allocation (LDA)
B. K-means
C. Se mantic feg mentation
D. Principal component analysis (PCA)
E. Factorization machines (FM)

Answer: BD

Explanation:
The data scientist should use K-means and principal component analysis (PCA) to meet this requirement. K-means is a clustering algorithm that can group customers based on their similarity in the feature space. PCA is a dimensionality reduction technique that can transform the original 980 variables into a smaller set of uncorrelated variables that capture most of the variance in the data. This can help reduce the computational cost and noise in the data, and improve the performance of the clustering algorithm.
Reference:
Clustering - Amazon SageMaker
Dimensionality Reduction - Amazon SageMaker

QUESTION 267
A companys data scientist has trained a new machine learning model that performs better on test data than the companys existing model performs in the production environment. The data scientist wants to replace the existing model that runs on an Amazon SageMaker endpoint in the production environment. However, the company is concerned that the new model might not work well on the production environment data.
The data scientist needs to perform A/B testing in the production environment to evaluate whether the new model performs well on production environment data.

Which combination of steps must the data scientist take to perform the A/B testing? (Choose two.)

A. Create a new endpoint configuration that includes a production variant for each of the two models.
B. Create a new endpoint configuration that includes two target variants that point to different endpoints.
C. Deploy the new model to the existing endpoint.
D. Update the existing endpoint to activate the new model.
E. Update the existing endpoint to use the new endpoint configuration.

Answer: A, E

Explanation:
The combination of steps that the data scientist must take to perform the A/B testing are to create a new endpoint configuration that includes a production variant for each of the two models, and update the existing endpoint to use the new endpoint configuration. This approach will allow the data scientist to deploy both models on the same endpoint and split the inference traffic between them based on a specified distribution.
Amazon SageMaker is a fully managed service that provides developers and data scientists the ability to quickly build, train, and deploy machine learning models. Amazon SageMaker supports A/B testing on machine learning models by allowing the data scientist to run multiple production variants on an endpoint. A production variant is a version of a model that is deployed on an endpoint. Each production variant has a name, a machine learning model, an instance type, an initial instance count, and an initial weight. The initial weight determines the percentage of inference requests that the variant will handle. For example, if there are two variants with weights of 0.5 and 0.5, each variant will handle 50% of the requests. The data scientist can use production variants to test models that have been trained using different training datasets, algorithms, and machine learning frameworks; test how they perform on different instance types; or a combination of all of the above1.
To perform A/B testing on machine learning models, the data scientist needs to create a new endpoint configuration that includes a production variant for each of the two models. An endpoint configuration is a collection of settings that define the properties of an endpoint, such as the name, the production variants, and the data capture configuration. The data scientist can use the Amazon SageMaker console, the AWS CLI, or the AWS SDKs to create a new endpoint configuration. The data scientist needs to specify the name, model name, instance type, initial instance count, and initial variant weight for each production variant in the endpoint configuration2.
After creating the new endpoint configuration, the data scientist needs to update the existing endpoint to use the new endpoint configuration. Updating an endpoint is the process of deploying a new endpoint configuration to an existing endpoint. Updating an endpoint does not affect the availability or scalability of the endpoint, as Amazon SageMaker creates a new endpoint instance with the new configuration and switches the DNS record to point to the new instance when it is ready. The data scientist can use the Amazon SageMaker console, the AWS CLI, or the AWS SDKs to update an endpoint. The data scientist needs to specify the name of the endpoint and the name of

the new endpoint configuration to update the endpoint3.
The other options are either incorrect or unnecessary. Creating a new endpoint configuration that includes two target variants that point to different endpoints is not possible, as target variants are only used to invoke a specific variant on an endpoint, not to define an endpoint configuration. Deploying the new model to the existing endpoint would replace the existing model, not run it sideby-side with the new model. Updating the existing endpoint to activate the new model is not a valid operation, as there is no activation parameter for an endpoint.
Reference:
1: A/B Testing ML models in production using Amazon SageMaker | AWS Machine Learning Blog
2: Create an Endpoint Configuration - Amazon SageMaker
3: Update an Endpoint - Amazon SageMaker

QUESTION 268
An online store is predicting future book sales by using a linear regression model that is based on past sales data. The data includes duration, a numerical feature that represents the number of days that a book has been listed in the online store. A data scientist performs an exploratory data analysis and discovers that the relationship between book sales and duration is skewed and non-linear.
Which data transformation step should the data scientist take to improve the predictions of the model?

A. One-hot encoding
B. Cartesian product transformation
C. Quantile binning
D. Normalization

Answer: C

Explanation:
Quantile binning is a data transformation technique that can be used to handle skewed and nonlinear numerical features. It divides the range of a feature into equal-sized bins based on the percentiles of the data. Each bin is assigned a numerical value that represents the midpoint of the bin. This way, the feature values are transformed into a more uniform distribution that can improve the performance of linear models. Quantile binning can also reduce the impact of outliers and noise in the data.
One-hot encoding, Cartesian product transformation, and normalization are not suitable for this scenario. One-hot encoding is used to transform categorical features into binary features. Cartesian product transformation is used to create new features by combining existing features. Normalization is used to scale numerical features to a standard range, but it does not change the shape of the distribution. Reference:
Data Transformations for Machine Learning
Quantile Binning Transformation

QUESTION 269
A beauty supply store wants to understand some characteristics of visitors to the store. The store has security video recordings from the past several years. The store wants to generate a report of hourly visitors from the recordings. The report should group visitors by hair style and hair color.
Which solution will meet these requirements with the LEAST amount of effort?

A. Use an object detection algorithm to identify a visitors hair in video frames. Pass the identified hair to an ResNet-50 algorithm to determine hair style and hair color.
B. Use an object detection algorithm to identify a visitors hair in video frames. Pass the identified hair to an XGBoost algorithm to determine hair style and hair color.
C. Use a semantic segmentation algorithm to identify a visitors hair in video frames. Pass the identified hair to an ResNet-50 algorithm to determine hair style and hair color.
D. Use a semantic segmentation algorithm to identify a visitors hair in video frames. Pass the identified hair to an XGBoost algorithm to determine hair style and hair.

Answer: C

Explanation:
The solution that will meet the requirements with the least amount of effort is to use a semantic segmentation algorithm to identify a visitors hair in video frames, and pass the identified hair to an ResNet-50 algorithm to determine hair style and hair color. This solution can leverage the existing Amazon SageMaker algorithms and frameworks to perform the tasks of hair segmentation and classification.
Semantic segmentation is a computer vision technique that assigns a class label to every pixel in an image, such that pixels with the same label share certain characteristics. Semantic segmentation can be used to identify and isolate different objects or regions in an image, such as a visitors hair in a video frame. Amazon SageMaker provides a built-in semantic segmentation algorithm that can train and deploy models for semantic segmentation tasks. The algorithm supports three state-of-the-art network architectures: Fully Convolutional Network (FCN), Pyramid Scene Parsing Network (PSP), and DeepLab v3. The algorithm can also use pre-trained or randomly initialized ResNet-50 or ResNet-101 as the backbone network. The algorithm can be trained using P2/P3 type Amazon EC2 instances in single machine configurations1.
ResNet-50 is a convolutional neural network that is 50 layers deep and can classify images into 1000 object categories. ResNet-50 is trained on more than a million images from the ImageNet database and can achieve high accuracy on various image recognition tasks. ResNet-50 can be used to determine hair style and hair color from the segmented hair regions in the video frames. Amazon SageMaker provides a built-in image classification algorithm that can use ResNet-50 as the network architecture. The algorithm can also perform transfer learning by fine-tuning the pre-trained ResNet-50 model with new data. The algorithm can be trained using P2/P3 type Amazon EC2 instances in single or multiple machine configurations2.
The other options are either less effective or more complex to implement. Using an object detection algorithm to identify a visitors hair in video frames would not segment the hair at the pixel level, but

only draw bounding boxes around the hair regions. This could result in inaccurate or incomplete hair segmentation, especially if the hair is occluded or has irregular shapes. Using an XGBoost algorithm to determine hair style and hair color would require transforming the segmented hair images into numerical features, which could lose some information or introduce noise. XGBoost is also not designed for image classification tasks, and may not achieve high accuracy or performance.
Reference:
1: Semantic Segmentation Algorithm - Amazon SageMaker
2: Image Classification Algorithm - Amazon SageMaker

QUESTION 270
A company wants to predict stock market price trends. The company stores stock market data each business day in Amazon S3 in Apache Parquet format. The company stores 20 GB of data each day for each stock code.
A data engineer must use Apache Spark to perform batch preprocessing data transformations quickly so the company can complete prediction jobs before the stock market opens the next day. The company plans to track more stock market codes and needs a way to scale the preprocessing data transformations.
Which AWS service or feature will meet these requirements with the LEAST development effort over time?

A. AWS Glue jobs
B. Amazon EMR cluster
C. Amazon Athena
D. AWS Lambda

Answer: A

Explanation:
AWS Glue jobs is the AWS service or feature that will meet the requirements with the least development effort over time. AWS Glue jobs is a fully managed service that enables data engineers to run Apache Spark applications on a serverless Spark environment. AWS Glue jobs can perform batch preprocessing data transformations on large datasets stored in Amazon S3, such as converting data formats, filtering data, joining data, and aggregating data. AWS Glue jobs can also scale the Spark environment automatically based on the data volume and processing needs, without requiring any infrastructure provisioning or management. AWS Glue jobs can reduce the development effort and time by providing a graphical interface to create and monitor Spark applications, as well as a code generation feature that can generate Scala or Python code based on the data sources and targets. AWS Glue jobs can also integrate with other AWS services, such as Amazon Athena, Amazon EMR, and Amazon SageMaker, to enable further data analysis and machine learning tasks1.
The other options are either more complex or less scalable than AWS Glue jobs. Amazon EMR cluster is a managed service that enables data engineers to run Apache Spark applications on a cluster of Amazon EC2 instances. However, Amazon EMR cluster requires more development effort and time

than AWS Glue jobs, as it involves setting up, configuring, and managing the cluster, as well as writing and deploying the Spark code. Amazon EMR cluster also does not scale automatically, but requires manual or scheduled resizing of the cluster based on the data volume and processing needs2. Amazon Athena is a serverless interactive query service that enables data engineers to analyze data stored in Amazon S3 using standard SQL. However, Amazon Athena is not suitable for performing complex data transformations, such as joining data from multiple sources, aggregating data, or applying custom logic. Amazon Athena is also not designed for running Spark applications, but only supports SQL queries3. AWS Lambda is a serverless compute service that enables data engineers to run code without provisioning or managing servers. However, AWS Lambda is not optimized for running Spark applications, as it has limitations on the execution time, memory size, and concurrency of the functions. AWS Lambda is also not integrated with Amazon S3, and requires additional steps to read and write data from S3 buckets.
Reference:
1: AWS Glue - Fully Managed ETL Service - Amazon Web Services
2: Amazon EMR - Amazon Web Services
3: Amazon Athena " Interactive SQL Queries for Data in Amazon S3
[4]: AWS Lambda " Serverless Compute - Amazon Web Services

QUESTION 271
A company wants to enhance audits for its machine learning (ML) systems. The auditing system must be able to perform metadata analysis on the features that the ML models use. The audit solution must generate a report that analyzes the metadata. The solution also must be able to set the data sensitivity and authorship of features.
Which solution will meet these requirements with the LEAST development effort?

A. Use Amazon SageMaker Feature Store to select the features. Create a data flow to perform feature-level metadata analysis. Create an Amazon DynamoDB table to store feature-level metadata. Use Amazon QuickSight to analyze the metadata.
B. Use Amazon SageMaker Feature Store to set feature groups for the current features that the ML models use. Assign the required metadata for each feature. Use SageMaker Studio to analyze the metadata.
C. Use Amazon SageMaker Features Store to apply custom algorithms to analyze the feature-level metadata that the company requires. Create an Amazon DynamoDB table to store feature-level metadata. Use Amazon QuickSight to analyze the metadata.
D. Use Amazon SageMaker Feature Store to set feature groups for the current features that the ML models use. Assign the required metadata for each feature. Use Amazon QuickSight to analyze the metadata.

Answer: D

Explanation:
The solution that will meet the requirements with the least development effort is to use Amazon

SageMaker Feature Store to set feature groups for the current features that the ML models use, assign the required metadata for each feature, and use Amazon QuickSight to analyze the metadata. This solution can leverage the existing AWS services and features to perform feature-level metadata analysis and reporting.
Amazon SageMaker Feature Store is a fully managed, purpose-built repository to store, update, search, and share machine learning (ML) features. The service provides feature management capabilities such as enabling easy feature reuse, low latency serving, time travel, and ensuring consistency between features used in training and inference workflows. A feature group is a logical grouping of ML features whose organization and structure is defined by a feature group schema. A feature group schema consists of a list of feature definitions, each of which specifies the name, type, and metadata of a feature. The metadata can include information such as data sensitivity, authorship, description, and parameters. The metadata can help make features discoverable, understandable, and traceable. Amazon SageMaker Feature Store allows users to set feature groups for the current features that the ML models use, and assign the required metadata for each feature using the AWS SDK for Python (Boto3), AWS Command Line Interface (AWS CLI), or Amazon SageMaker Studio1.
Amazon QuickSight is a fully managed, serverless business intelligence service that makes it easy to create and publish interactive dashboards that include ML insights. Amazon QuickSight can connect to various data sources, such as Amazon S3, Amazon Athena, Amazon Redshift, and Amazon SageMaker Feature Store, and analyze the data using standard SQL or built-in ML-powered analytics. Amazon QuickSight can also create rich visualizations and reports that can be accessed from any device, and securely shared with anyone inside or outside an organization. Amazon QuickSight can be used to analyze the metadata of the features stored in Amazon SageMaker Feature Store, and generate a report that summarizes the metadata analysis2.
The other options are either more complex or less effective than the proposed solution. Using Amazon SageMaker Data Wrangler to select the features and create a data flow to perform featurelevel metadata analysis would require additional steps and resources, and may not capture all the metadata attributes that the company requires. Creating an Amazon DynamoDB table to store feature-level metadata would introduce redundancy and inconsistency, as the metadata is already stored in Amazon SageMaker Feature Store. Using SageMaker Studio to analyze the metadata would not generate a report that can be easily shared and accessed by the company.
Reference:
1: Amazon SageMaker Feature Store " Amazon Web Services
2: Amazon QuickSight " Business Intelligence Service - Amazon Web Services

---

QUESTION 272
A machine learning (ML) engineer has created a feature repository in Amazon SageMaker Feature Store for the company. The company has AWS accounts for development, integration, and production. The company hosts a feature store in the development account. The company uses Amazon S3 buckets to store feature values offline. The company wants to share features and to allow the integration account and the production account to reuse the features that are in the feature repository.
Which combination of steps will meet these requirements? (Select TWO.)

A. Create an IAM role in the development account that the integration account and production account can assume. Attach IAM policies to the role that allow access to the feature repository and the S3 buckets.
B. Share the feature repository that is associated the S3 buckets from the development account to the integration account and the production account by using AWS Resource Access Manager (AWS RAM).
C. Use AWS Security Token Service (AWS STS) from the integration account and the production account to retrieve credentials for the development account.
D. Set up S3 replication between the development S3 buckets and the integration and production S3 buckets.
E. Create an AWS PrivateLink endpoint in the development account for SageMaker.

Answer: A, B

Explanation:
The combination of steps that will meet the requirements are to create an IAM role in the development account that the integration account and production account can assume, attach IAM policies to the role that allow access to the feature repository and the S3 buckets, and share the feature repository that is associated with the S3 buckets from the development account to the integration account and the production account by using AWS Resource Access Manager (AWS RAM). This approach will enable cross-account access and sharing of the features stored in Amazon SageMaker Feature Store and Amazon S3.
Amazon SageMaker Feature Store is a fully managed, purpose-built repository to store, update, search, and share curated data used in training and prediction workflows. The service provides feature management capabilities such as enabling easy feature reuse, low latency serving, time travel, and ensuring consistency between features used in training and inference workflows. A feature group is a logical grouping of ML features whose organization and structure is defined by a feature group schema. A feature group schema consists of a list of feature definitions, each of which specifies the name, type, and metadata of a feature. Amazon SageMaker Feature Store stores the features in both an online store and an offline store. The online store is a low-latency, highthroughput store that is optimized for real-time inference. The offline store is a historical store that is backed by an Amazon S3 bucket and is optimized for batch processing and model training1.
AWS Identity and Access Management (IAM) is a web service that helps you securely control access to AWS resources for your users. You use IAM to control who can use your AWS resources (authentication) and what resources they can use and in what ways (authorization). An IAM role is an IAM identity that you can create in your account that has specific permissions. You can use an IAM role to delegate access to users, applications, or services that dont normally have access to your AWS resources. For example, you can create an IAM role in your development account that allows the integration account and the production account to assume the role and access the resources in the development account. You can attach IAM policies to the role that specify the permissions for the feature repository and the S3 buckets. You can also use IAM conditions to restrict the access based on the source account, IP address, or other factors2.
AWS Resource Access Manager (AWS RAM) is a service that enables you to easily and securely share AWS resources with any AWS account or within your AWS Organization. You can share AWS resources that you own with other accounts using resource shares. A resource share is an entity that defines

the resources that you want to share, and the principals that you want to share with. For example, you can share the feature repository that is associated with the S3 buckets from the development account to the integration account and the production account by creating a resource share in AWS RAM. You can specify the feature group ARN and the S3 bucket ARN as the resources, and the integration account ID and the production account ID as the principals. You can also use IAM policies to further control the access to the shared resources3.

The other options are either incorrect or unnecessary. Using AWS Security Token Service (AWS STS) from the integration account and the production account to retrieve credentials for the development account is not required, as the IAM role in the development account can provide temporary security credentials for the cross-account access. Setting up S3 replication between the development S3 buckets and the integration and production S3 buckets would introduce redundancy and inconsistency, as the S3 buckets are already shared through AWS RAM. Creating an AWS PrivateLink endpoint in the development account for SageMaker is not relevant, as it is used to securely connect to SageMaker services from a VPC, not from another account.

Reference:

1: Amazon SageMaker Feature Store " Amazon Web Services
2: What Is IAM? - AWS Identity and Access Management
3: What Is AWS Resource Access Manager? - AWS Resource Access Manager
3: What Is AWS Resource Access Manager? - AWS Resource Access Manager