Number: MLS-C01
Passing Score: 800
Time Limit: 120
File Version: 21.0

Exam Code: MLS-C01

Exam Name: AWS Certified Machine Learning - Specialty

Exam A

QUESTION 1
A Data Science team is designing a dataset repository where it will store a large amount of training data commonly used in its machine learning models. As Data Scientists may create an arbitrary number of new datasets every day the solution has to scale automatically and be cost-effective. Also, it must be possible to explore the data using SQL.
Which storage scheme is MOST adapted to this scenario?

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

Correct Answer: A
Section:
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.
References:
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 2
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

Correct Answer: C
Section:

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 rest3Specifying 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 needed4Executing 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 3
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.

Correct Answer: C
Section:
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 company's cost is lower than if the model had more false negatives and fewer false positives.

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

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

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

A. Shuffle all interaction data. Split off the last 10% of the interaction data for the test set.
B. Identify the most recent 10% of interactions for each user. Split off these interactions for the test set.
C. Identify the 10% of users with the least interaction data. Split off all interaction data from these users for the test set.
D. Randomly select 10% of the users. Split off all interaction data from these users for the test set.

Correct Answer: D
Section:
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 .References:
Amazon SageMaker Developer Guide: Train and Test Datasets
Amazon Personalize Developer Guide: Preparing and Importing Data

QUESTION 5
A financial services company wants to adopt Amazon SageMaker as its default data science environment. The company's data scientists run machine learning (ML) models on confidential financial data. The company is worried about data egress and wants an ML engineer to secure the environment.
Which mechanisms can the ML engineer use to control data egress from SageMaker? (Choose three.)

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

Correct Answer: A, D, F
Section:
Explanation:
Use AWS Key Management Service (AWS KMS) to manage encryption keys. 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. References: 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 6
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 data. 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

Correct Answer: C
Section:
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.
References:
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 7
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.

Correct Answer: C
Section:
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.
References:
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 8
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.

Correct Answer: B
Section:
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.
References:
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 9
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

Correct Answer: C
Section:
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
References:
1:Amazon SageMaker Notebook Instances - Amazon SageMaker
2:CreateNotebookInstanceLifecycleConfig - Amazon SageMaker
3:Create a Notebook Instance - Amazon SageMaker
4:UpdateNotebookInstanceLifecycleConfig - Amazon SageMaker

QUESTION 10
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 claim's record in a central database. The claim's record includes the claim's category.
The company has no data science team or experience in the field of machine learning (ML). The company's 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.

Correct Answer: D
Section:
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.
References:
Amazon Comprehend Custom Classification
Amazon SageMaker Object2Vec
Amazon SageMaker Latent Dirichlet Allocation
Amazon Textract

QUESTION 11
A machine learning (ML) specialist is using Amazon SageMaker hyperparameter optimization (HPO) to improve a model's 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:

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:
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.

Correct Answer: C
Section:
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 sub-intervals, which may miss some combinations of hyperparameters.References:
How Hyperparameter Tuning Works - Amazon SageMaker
Tuning Hyperparameters - Amazon SageMaker

QUESTION 12
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 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 company's 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.

Correct Answer: C
Section:
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 low-connectivity 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 low-connectivity 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 company's ability to leverage the cloud's 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.
References:
1: Amazon S3
2: Amazon SageMaker
3: SageMaker Neo
4: AWS IoT Greengrass

QUESTION 13
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.

Correct Answer: A, C
Section:
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.
References:
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 14
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.

Correct Answer: C
Section:
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 media.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.
References:
1: Amazon Rekognition Celebrity Recognition
2: AWS CloudTrail Overview
3: AWS Panorama Overview
4: AWS Panorama Device SDK
5: Amazon Rekognition Text Detection

QUESTION 15
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.

Correct Answer: A

Section:
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.
References:
1: AWS Lambda
2: Amazon Kinesis Data Firehose
3: Amazon EMR
4: Amazon S3 File Gateway
5: Amazon EC2

QUESTION 16
A company wants to segment a large group of customers into subgroups based on shared characteristics. The company's data scientist is planning to use the Amazon SageMaker built-in k-means 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.

Correct Answer: D
Section:
Explanation:
The solution D is the best data visualization approach to determine the optimal value of k for the k-means 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 two-dimensional 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.
References:
1: How to Determine the Optimal K for K-Means?
2: Principal Component Analysis
3: t-Distributed Stochastic Neighbor Embedding

QUESTION 17
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.

Correct Answer: A
Section:
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.References:
AWS Certified Machine Learning - Specialty Exam Guide
AWS Training - Machine Learning on AWS
AWS Whitepaper - An Overview of Machine Learning on AWS

QUESTION 18
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.

Correct Answer: A
Section:
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.
References:
1: Amazon SageMaker Data Wrangler
2: Amazon SageMaker Pipelines
3: AWS Lambda
4: AWS Glue Studio
5: AWS Step Functions

QUESTION 19
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.

Correct Answer: A
Section:
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.
References:
1: Amazon SageMaker Ground Truth
2: Amazon Rekognition Custom Labels
3: Amazon SageMaker Object Detection
4: Amazon Mechanical Turk

QUESTION 20
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.

Correct Answer: C
Section:
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.
References:
1: Managed Spot Training - Amazon SageMaker
2: Pipe Mode - Amazon SageMaker
3: FastFile Mode - Amazon SageMaker
4: Checkpoints - Amazon SageMaker

QUESTION 21
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.

Correct Answer: C
Section:
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.
References:
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 22
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.

Correct Answer: C
Section:
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 fine-tuning 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.
References:
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 23
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

Correct Answer: B
Section:
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.
References:
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 24
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

Correct Answer: A, C
Section:
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.
References:
1: Latent Dirichlet Allocation
2: Neural Topic Modeling
3: Random Forest Classifier
4: Linear Support Vector Machine
5: Linear Regression

QUESTION 25
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.

Correct Answer: B
Section:
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.
References:
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 26
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.

Correct Answer: A
Section:
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.
References:
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 27
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.

Correct Answer: D
Section:
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 line-counting 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 restaurant's 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.
References:
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 28
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.

Correct Answer: D
Section:
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 multi-model 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
References:
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 29
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.

Correct Answer: C
Section:
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.
References:
1:UpdateEndpointWeightsAndCapacities - Amazon SageMaker
2:InvokeEndpoint - Amazon SageMaker
3:CreateEndpointConfig - Amazon SageMaker
4:Application Load Balancer - Elastic Load Balancing

QUESTION 30
A company offers an online shopping service to its customers. The company wants to enhance the site's 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.

Correct Answer: B
Section:
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.
References:
IP Insights - Amazon SageMaker
Factorization Machines Algorithm - Amazon SageMaker
Object2Vec Algorithm - Amazon SageMaker

QUESTION 31
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.

Correct Answer: C
Section:
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 32
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:CreatePresignedNotebookInstanceUrl and sagemaker:DescribeNotebookInstance 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.

Correct Answer: A
Section:
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.
References:
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 33
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 data. 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.

Correct Answer: C
Section:
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.
References:
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 34
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.

Correct Answer: D
Section:
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.References:AWS Glue,Amazon SageMaker

QUESTION 35
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.

Correct Answer: A
Section:
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-to-sequence 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.
References:
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 36
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.

Correct Answer: D
Section:
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 ''comedy'' with synonyms ''funny'', ''fun'', and ''humor'', then any of these words entered by the user will be resolved to ''comedy'' 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 free-form 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.
References:
Custom slot type - Amazon Lex
Using Synonyms - Amazon Lex
Built-in Slot Types - Amazon Lex

QUESTION 37
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.

Correct Answer: D
Section:
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.References:
AWS IoT Greengrass
Deploying Machine Learning Models to Edge Devices
AWS Certified Machine Learning - Specialty Exam Guide

QUESTION 38
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.

Correct Answer: C
Section:
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.References:
Amazon SageMaker Roles
Accessing Amazon S3 from a SageMaker Notebook Instance

QUESTION 39
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 media. 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.

Correct Answer: D
Section:
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.
References:
Amazon Comprehend
Amazon CloudWatch

QUESTION 40
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.

Correct Answer: A
Section:
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.References:
Amazon SageMaker Studio
Amazon SageMaker Model Monitor
SHAP library

QUESTION 41
A data science team is planning to build a natural language processing (NLP) application. The application's 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.

Correct Answer: D
Section:
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.
References:
Amazon Comprehend
AWS Deep Learning Containers
Amazon SageMaker

QUESTION 42
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.

Correct Answer: A
Section:
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.
References:
Dimensionality Reduction with Amazon SageMaker
Amazon SageMaker PCA Algorithm

QUESTION 43
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.

Correct Answer: C, E
Section:
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-to-sequence 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.References:
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 44

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.

Correct Answer: B
Section:
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.References:
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 45
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.

Correct Answer: C, D
Section:
Explanation:
Request 10 years of sales data, which would provide 200 MB of training data for the model. 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 model's 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. References: Amazon Forecast Forecasting: Principles and Practice

QUESTION 46
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

Correct Answer: C
Section:
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.References:Amazon Forecast Algorithms,Amazon Forecast CNN-QR

QUESTION 47
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.

Correct Answer: C
Section:
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.
References:
Amazon Transcribe -- Custom Vocabulary
Amazon Transcribe -- Custom Language Models

[Amazon Lex -- Limits]

QUESTION 48
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.

Correct Answer: C
Section:
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 general-purpose 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.
References:
Amazon SageMaker Pricing
Manage Notebook Instances - Amazon SageMaker
Amazon EC2 Pricing - Reserved Instances

QUESTION 49
A Machine Learning Specialist is designing a system for improving sales for a company. The objective is to use the large amount of information the company has on users' behavior and product preferences to predict which

products users would like based on the users' similarity to other users.
What should the Specialist do to meet this objective?

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

Correct Answer: B
Section:
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 50
A Data Engineer needs to build a model using a dataset containing customer credit card information.
How can the Data Engineer ensure the data remains encrypted and the credit card information is secure?

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

Correct Answer: D
Section:
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 real-time 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 51

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

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

Correct Answer: C
Section:
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 instance's 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 52
A manufacturing company has structured and unstructured data stored in an Amazon S3 bucket. A Machine Learning Specialist wants to use SQL to run queries on this data.
Which solution requires the LEAST effort to be able to query this data?

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

Correct Answer: B
Section:
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 format1Amazon 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 tables45Using 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 53

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

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

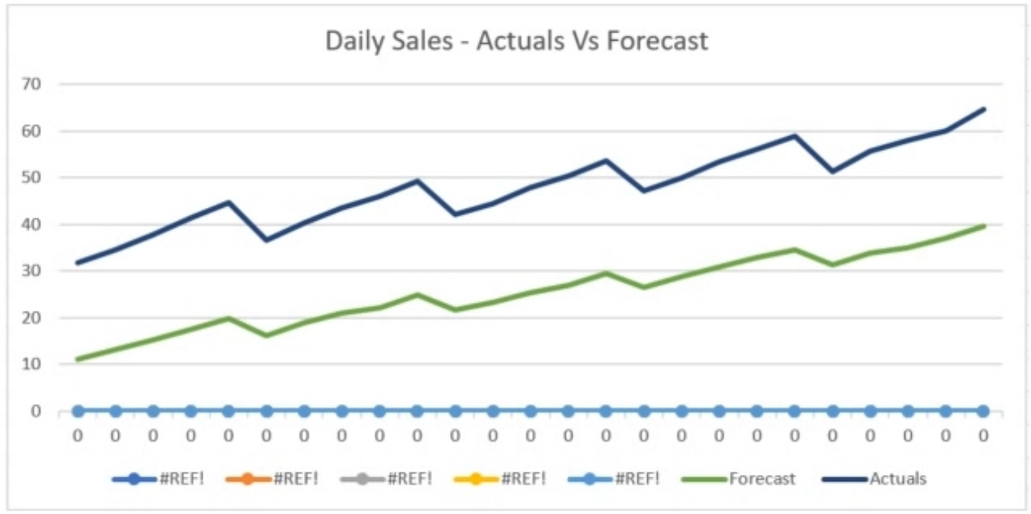
Correct Answer: C
Section:
Explanation:
Collaborative filtering is a machine learning technique that recommends products or services to users based on the ratings or preferences of other users. This technique is well-suited for identifying customer shopping patterns and preferences because it takes into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

QUESTION 54
A Machine Learning Specialist is working with a large company to leverage machine learning within its products. The company wants to group its customers into categories based on which customers will and will not churn within the next 6 months. The company has labeled the data available to the Specialist.
Which machine learning model type should the Specialist use to accomplish this task?

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

Correct Answer: B
Section:
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?

QUESTION 55
The displayed graph is from a foresting model for testing a time series.

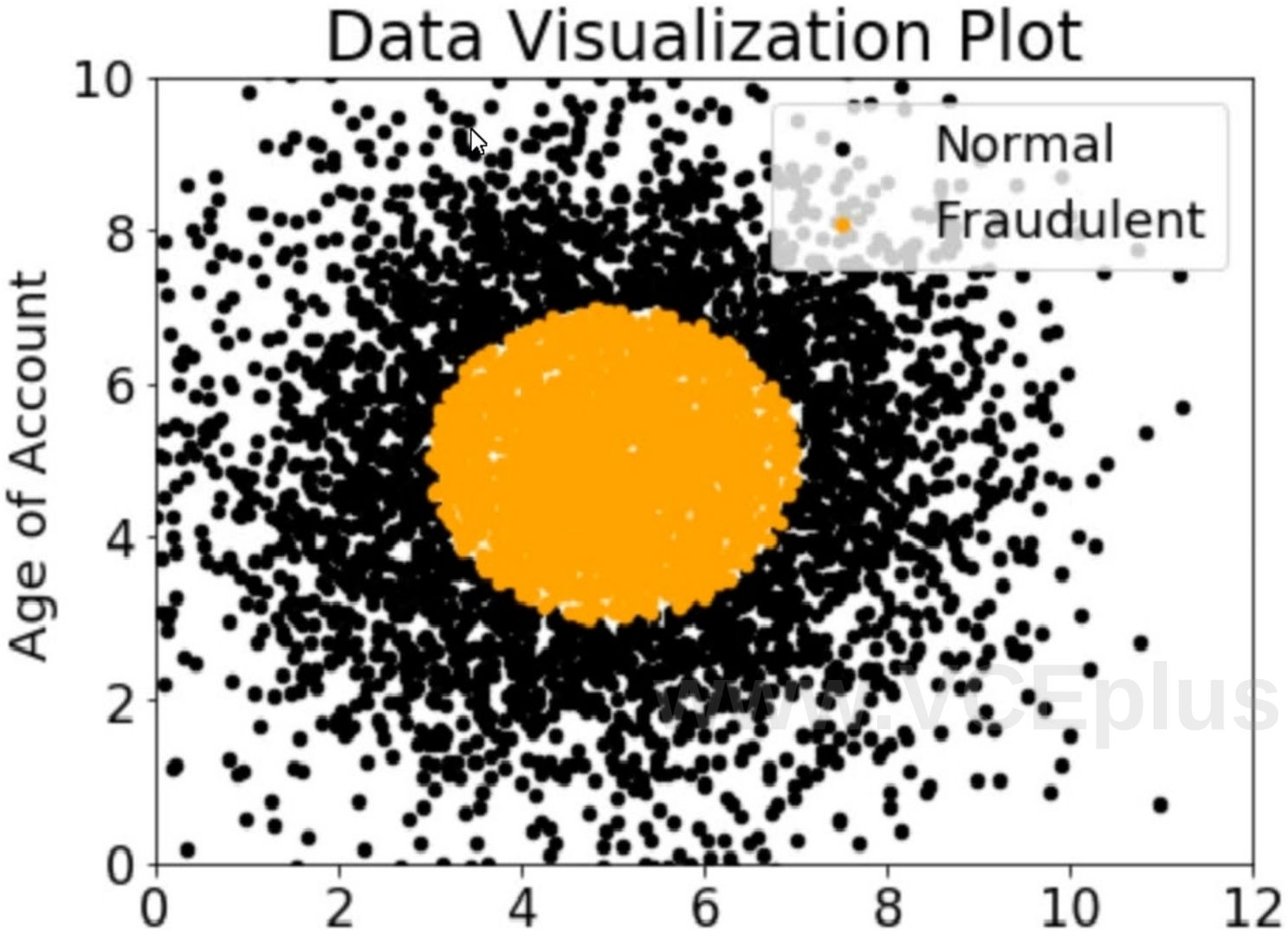


Considering the graph only, which conclusion should a Machine Learning Specialist make about the behavior of the model?

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.

Correct Answer: D
Section:

QUESTION 56
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

Correct Answer: C
Section:
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 high-dimensional 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
References:1: https://docs.aws.amazon.com/sagemaker/latest/dg/svm.html

QUESTION 57
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

Correct Answer: A
Section:
Explanation:
The best services to use for building a data ingestion solution for the company's 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.
References:
AWS Glue - Fully Managed ETL Service
Amazon Kinesis - Data Streaming Service
Amazon OpenSearch Service - Managed OpenSearch Service
Amazon EMR - Managed Hadoop Framework

QUESTION 58
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.

Correct Answer: C, D
Section:
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.
References:
Use Your Own Algorithms or Models with Amazon SageMaker
Use the SageMaker TensorFlow Serving Container
TensorFlow Hub

QUESTION 59
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.

Correct Answer: B
Section:
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.
References:

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 60
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.

Correct Answer: B
Section:
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.
References:
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 61
A Machine Learning Specialist is implementing a full Bayesian network on a dataset that describes public transit in New York City. One of the random variables is discrete, and represents the number of minutes New Yorkers wait for a bus given that the buses cycle every 10 minutes, with a mean of 3 minutes.
Which prior probability distribution should the ML Specialist use for this variable?

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

Correct Answer: A
Section:
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.
References:
1: Poisson Distribution - Amazon SageMaker
2: Poisson Distribution - Wikipedia

QUESTION 62
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.

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

QUESTION 63
A Machine Learning Specialist has completed a proof of concept for a company using a small data sample and now the Specialist is ready to implement an end-to-end solution in AWS using Amazon SageMaker The historical training data is stored in Amazon RDS
Which approach should the Specialist use for training a model using that data?

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

Correct Answer: B
Section:
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.References:
Amazon SageMaker
AWS Data Pipeline

QUESTION 64
Which of the following metrics should a Machine Learning Specialist generally use to compare/evaluate machine learning classification models against each other?

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

Correct Answer: D
Section:
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.References:
AWS Machine Learning Specialty Exam Guide
AWS Machine Learning Specialty Sample Questions

QUESTION 65
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)

Correct Answer: C
Section:
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.
References:
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 66
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)

Correct Answer: B
Section:
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.
References:
XGBoost Algorithm
Use XGBoost for Binary Classification with Amazon SageMaker
Seq2seq Algorithm
K-means Algorithm
[Random Cut Forest Algorithm]

QUESTION 67
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

Correct Answer: B
Section:
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.References:
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 68
An online reseller has a large, multi-column dataset with one column missing 30% of its data A Machine Learning Specialist believes that certain columns in the dataset could be used to reconstruct the missing data.
Which reconstruction approach should the Specialist use to preserve the integrity of the dataset?

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

Correct Answer: C
Section:
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.References:
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 69
A Machine Learning Specialist discover the following statistics while experimenting on a model.

Experiment 1
Baseline model:
Train error = 5%
Test error = 16%

Experiment 2
The Specialist added more layers and neurons to the model and received the following results:
Train error = 5.2%
Test error = 15.7%

Experiment 3
The Specialist reverted back to the original number of neurons from Experiment 1 and implemented regularization in the neural network, which yielded the following results:
Train error = 4.7%
Test error = 9.5%

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

Correct Answer: A
Section:
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.References:
Bias Variance Tradeoff - Clearly Explained - Machine Learning Plus
The Bias-Variance Trade-off in Machine Learning - Stack Abuse

QUESTION 70
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

Correct Answer: C
Section:
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.References:
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 71
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

Correct Answer: B, C
Section:
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.References:
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 72
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

Correct Answer: B, E
Section:
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 model's objective function, which is typically based on the model's 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 model's parameters multiplied by a regularization parameter (lambda) to the model's objective function. L2 regularization can shrink the model's 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 model's 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 73
The Chief Editor for a product catalog wants the Research and Development team to build a machine learning system that can be used to detect whether or not individuals in a collection of images are wearing the company's retail brand The team has a set of training data
Which machine learning algorithm should the researchers use that BEST meets their requirements?

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

Correct Answer: D
Section:
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 pre-trained models that have been trained on large-scale image datasets, such as ImageNet, and fine-tune them for specific tasks, such as detecting the company's 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, Scikit-learn, 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 74

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.

Correct Answer: D
Section:
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 tree-based 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 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.

Correct Answer: A
Section:
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

Correct Answer: A
Section:
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 Spanish1Amazon 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 English2Amazon 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

Correct Answer: B
Section:
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 container's 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 pizza. The Specialist is trying to build the optimal model with an ideal classification threshold.
What model evaluation technique should the Specialist use to understand how different classification thresholds will impact the model's performance?

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

Correct Answer: A
Section:
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 model's 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 embedding's pre-trained on a large corpus.

Correct Answer: D
Section:
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.

QUESTION 80
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

Correct Answer: C
Section:
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.
References:
: What Is Amazon VPC? - Amazon Virtual Private Cloud
: Subnet Routing - Amazon Virtual Private Cloud
: VPC Endpoints - Amazon Virtual Private Cloud

QUESTION 81
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.

Correct Answer: B, C, F
Section:
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.
References:
1: Regularization for Deep Learning - Amazon SageMaker
2: Dropout - Amazon SageMaker
3: Feature Engineering - Amazon SageMaker

QUESTION 82
A Data Scientist needs to create a serverless ingestion and analytics solution for high-velocity, real-time streaming data.
The ingestion process must buffer and convert incoming records from JSON to a query-optimized, columnar format without data loss. The output datastore must be highly available, and Analysts must be able to run SQL queries against the data and connect to existing business intelligence dashboards.
Which solution should the Data Scientist build to satisfy the requirements?

A. Create a schema in the AWS Glue Data Catalog of the incoming data format. Use an Amazon Kinesis Data Firehose delivery stream to stream the data and transform the data to Apache Parquet or ORC format using the AWS Glue Data Catalog before delivering to Amazon S3. Have the Analysts query the data directly from Amazon S3 using Amazon Athena, and connect to 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.

Correct Answer: A
Section:
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 cost-effective queries4.
References:
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 83
A company is setting up an Amazon SageMaker environment. The corporate data security policy does not allow communication over the internet.
How can the company enable the Amazon SageMaker service without enabling direct internet access to Amazon SageMaker notebook instances?

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

Correct Answer: C
Section:
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.
References:
1: Connect to SageMaker Within your VPC - Amazon SageMaker

QUESTION 84
An office security agency conducted a successful pilot using 100 cameras installed at key locations within the main office. Images from the cameras were uploaded to Amazon S3 and tagged using Amazon Rekognition, and the results were stored in Amazon ES. The agency is now looking to expand the pilot into a full production system using thousands of video cameras in its office locations globally. The goal is to identify activities performed by non-employees in real time.
Which solution should the agency consider?

A. Use a proxy server at each local office and for each camera, and stream the RTSP feed to a unique Amazon Kinesis Video Streams video stream. On each stream, use Amazon Rekognition Video and create a stream processor to detect faces from a collection of known employees, and alert when non-employees are detected.
B. Use a proxy server at each local office and for each camera, and stream the RTSP feed to a unique Amazon Kinesis Video Streams video stream. On each stream, use Amazon Rekognition Image to detect faces from a collection of known employees and alert when non-employees are detected.
C. Install AWS DeepLens cameras and use the DeepLens_Kinesis_Video module to stream video to Amazon Kinesis Video Streams for each camera. On each stream, use Amazon Rekognition Video and create a stream processor to detect faces from a collection on each stream, and alert when 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.

Correct Answer: A
Section:
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.
References:
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 85
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.

Correct Answer: A
Section:
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.
References:
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 86
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.

Correct Answer: B
Section:
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.
References:
1: Horovod (machine learning) - Wikipedia
2: Home - Horovod
3: Amazon SageMaker -- Machine Learning Service -- AWS
4: Use Horovod with Amazon SageMaker - Amazon SageMaker

QUESTION 87
A Machine Learning Specialist is required to build a supervised image-recognition model to identify a cat. The ML Specialist performs some tests and records the following results for a neural network-based image classifier:
Total number of images available = 1,000 Test set images = 100 (constant test set)
The ML Specialist notices that, in over 75% of the misclassified images, the cats were held upside down by their owners.
Which techniques can be used by the ML Specialist to improve this specific test error?

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

Correct Answer: A
Section:
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.
References:
1: Image Augmentation - Amazon SageMaker

QUESTION 88
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 10 grids. The company also has a large training dataset that consists of annotated images of popular weed classes like broadleaf and non-broadleaf docks.
The company wants to build a weed detection model that will detect specific types of weeds and the location of each type within the field. Once the model is ready, it will be hosted on Amazon SageMaker endpoints. The model will perform real-time inferencing using the images captured by the cameras.
Which approach should a Machine Learning Specialist take to obtain accurate predictions?

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

Correct Answer: C
Section:
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.
References:
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 89
A manufacturer is operating a large number of factories with a complex supply chain relationship where unexpected downtime of a machine can cause production to stop at several factories. A data scientist wants to analyze sensor data from the factories to identify equipment in need of preemptive maintenance and then dispatch a service team to prevent unplanned downtime. The sensor readings from a single machine can include up to 200 data points including temperatures, voltages, vibrations, RPMs, and pressure readings.
To collect this sensor data, the manufacturer deployed Wi-Fi and LANs across the factories. Even though many factory locations do not have reliable or high-speed internet connectivity, the manufacturer would like to maintain near-real-time inference capabilities.
Which deployment architecture for the model will address these business requirements?

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

Correct Answer: B
Section:
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-real-time 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.
References:
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 90
A Machine Learning Specialist is designing a scalable data storage solution for Amazon SageMaker. There is an existing TensorFlow-based model implemented as a train.py script that relies on static training data that is currently stored as TFRecords.
Which method of providing training data to Amazon SageMaker would meet the business requirements with the LEAST development overhead?

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

Correct Answer: B
Section:
Explanation:
Amazon SageMaker script mode is a feature that allows users to use training scripts similar to those they would use outside SageMaker with SageMaker's 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.
References:
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 91
The chief editor for a product catalog wants the research and development team to build a machine learning system that can be used to detect whether or not individuals in a collection of images are wearing the company's retail brand. The team has a set of training data.
Which machine learning algorithm should the researchers use that BEST meets their requirements?

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

Correct Answer: D
Section:
Explanation:
The problem of detecting whether or not individuals in a collection of images are wearing the company's 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.
References:
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 92
A retail company is using Amazon Personalize to provide personalized product recommendations for its customers during a marketing campaign. The company sees a significant increase in sales of recommended items to existing customers immediately after deploying a new solution version, but these sales decrease a short time after deployment. Only historical data from before the marketing campaign is available for training.
How should a data scientist adjust the solution?

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

Correct Answer: A
Section:
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 user's 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 user's 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.
References:
Recording events - Amazon Personalize
Using real-time events - Amazon Personalize

QUESTION 93
A machine learning (ML) specialist wants to secure calls to the Amazon SageMaker Service API. The specialist has configured Amazon VPC with a VPC interface endpoint for the Amazon SageMaker Service API and is attempting to secure traffic from specific sets of instances and IAM users. The VPC is configured with a single public subnet.
Which combination of steps should the ML specialist take to secure the traffic? (Choose two.)

A. Add a VPC endpoint policy to allow access to the IAM users.

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

Correct Answer: C, E
Section:
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.
References:
Security groups for your VPC - Amazon Virtual Private Cloud
Connect to SageMaker Within your VPC - Amazon SageMaker
Network ACLs - Amazon Virtual Private Cloud

QUESTION 94
An e commerce company wants to launch a new cloud-based product recommendation feature for its web application. Due to data localization regulations, any sensitive data must not leave its on-premises data center, and the product recommendation model must be trained and tested using nonsensitive data only. Data transfer to the cloud must use IPsec. The web application is hosted on premises with a PostgreSQL database that contains all the data. The company wants the data to be uploaded securely to Amazon S3 each day for model retraining.
How should a machine learning specialist meet these requirements?

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

Correct Answer: C
Section:
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.
References:
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 95
A logistics company needs a forecast model to predict next month's inventory requirements for a single item in 10 warehouses. A machine learning specialist uses Amazon Forecast to develop a forecast model from 3 years of monthly data. There is no missing data. The specialist selects the DeepAR+ algorithm to train a predictor. The predictor means absolute percentage error (MAPE) is much larger than the MAPE produced by the current human forecasters.
Which changes to the CreatePredictor API call could improve the MAPE? (Choose two.)

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

Correct Answer: A, D
Section:
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 month's 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.
References:
CreatePredictor - Amazon Forecast
HPOConfig - Amazon Forecast

QUESTION 96
A data scientist wants to use Amazon Forecast to build a forecasting model for inventory demand for a retail company. The company has provided a dataset of historic inventory demand for its products as a .csv file stored in an Amazon S3 bucket. The table below shows a sample of the dataset.

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

How should the data scientist transform the data?

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

Correct Answer: A
Section:
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 .References:
How Amazon Forecast Works - Amazon Forecast
Choosing Datasets - Amazon Forecast

QUESTION 97
A machine learning specialist is running an Amazon SageMaker endpoint using the built-in object detection algorithm on a P3 instance for real-time predictions in a company's production application. When evaluating the model's resource utilization, the specialist notices that the model is using only a fraction of the GPU.
Which architecture changes would ensure that provisioned resources are being utilized effectively?

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

Correct Answer: B
Section:
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 you're working with large batches of data, don't 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.
References:
Amazon Elastic Inference - Amazon SageMaker
Batch Transform - Amazon SageMaker
Amazon EC2 P3 Instances
Amazon EC2 P3dn Instances
Amazon Elastic Container Service

QUESTION 98
A data scientist uses an Amazon SageMaker notebook instance to conduct data exploration and analysis. This requires certain Python packages that are not natively available on Amazon SageMaker to be installed on the notebook instance.
How can a machine learning specialist ensure that required packages are automatically available on the notebook instance for the data scientist to use?

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

Correct Answer: D
Section:

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.
References:
Customize a notebook instance using a lifecycle configuration script - Amazon SageMaker
AWS Systems Manager Automation - AWS Systems Manager
Conda environments - Amazon SageMaker

QUESTION 99
A data scientist needs to identify fraudulent user accounts for a company's ecommerce platform. The company wants the ability to determine if a newly created account is associated with a previously known fraudulent user. The data scientist is using AWS Glue to cleanse the company's application logs during ingestion.
Which strategy will allow the data scientist to identify fraudulent accounts?

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

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

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

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

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

Correct Answer: B, D
Section:
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.
References:
XGBoost Parameters - Amazon Machine Learning
Using XGBoost with Amazon SageMaker - AWS Machine Learning Blog

QUESTION 101
A data scientist has developed a machine learning translation model for English to Japanese by using Amazon SageMaker's built-in seq2seq algorithm with 500,000 aligned sentence pairs. While testing with sample sentences, the data scientist finds that the translation quality is reasonable for an example as short as five words. However, the quality becomes unacceptable if the sentence is 100 words long.
Which action will resolve the problem?

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

Correct Answer: C
Section:
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.
References:
Sequence-to-Sequence Algorithm - Amazon SageMaker
Attention Mechanism - Sockeye Documentation

QUESTION 102
A financial company is trying to detect credit card fraud. The company observed that, on average, 2% of credit card transactions were fraudulent. A data scientist trained a classifier on a year's worth of credit card transactions data. The model needs to identify the fraudulent transactions (positives) from the regular ones (negatives). The company's goal is to accurately capture as many positives as possible.
Which metrics should the data scientist use to optimize the model? (Choose two.)

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

Correct Answer: D, E
Section:
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 company's goal.
References:
Metrics for Imbalanced Classification in Python - Machine Learning Mastery
Precision-Recall - scikit-learn

QUESTION 103
A machine learning specialist is developing a proof of concept for government users whose primary concern is security. The specialist is using Amazon SageMaker to train a convolutional neural network (CNN) model for a photo classifier application. The specialist wants to protect the data so that it cannot be accessed and transferred to a remote host by malicious code accidentally installed on the training container.
Which action will provide the MOST secure protection?

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

Correct Answer: D
Section:
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.

References:
Run Training and Inference Containers in Internet-Free Mode - Amazon SageMaker

QUESTION 104
A medical imaging company wants to train a computer vision model to detect areas of concern on patients' CT scans. The company has a large collection of unlabeled CT scans that are linked to each patient and stored in an Amazon S3 bucket. The scans must be accessible to authorized users only. A machine learning engineer needs to build a labeling pipeline.
Which set of steps should the engineer take to build the labeling pipeline with the LEAST effort?

A. Create a workforce with AWS Identity and Access Management (IAM). Build a labeling tool on Amazon EC2 Queue images for labeling by using Amazon Simple Queue Service (Amazon SQS). Write the labeling instructions.
B. Create an Amazon Mechanical Turk workforce and manifest file. Create a labeling job by using the built-in image classification task type in Amazon SageMaker Ground Truth. Write the labeling instructions.
C. Create a private workforce and manifest file. Create a labeling job by using the built-in bounding box task type in Amazon SageMaker Ground Truth. Write the labeling instructions.
D. Create a workforce with Amazon Cognito. Build a labeling web application with AWS Amplify. Build a labeling workflow backend using AWS Lambda. Write the labeling instructions.

Correct Answer: C
Section:
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.
References:
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 105
A company is using Amazon Textract to extract textual data from thousands of scanned text-heavy legal documents daily. The company uses this information to process loan applications automatically. Some of the documents fail business validation and are returned to human reviewers, who investigate the errors. This activity increases the time to process the loan applications.
What should the company do to reduce the processing time of loan applications?

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

Correct Answer: C
Section:
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 Rekognition's 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.
References:
Amazon Augmented AI
Amazon SageMaker Ground Truth
Amazon Textract Operations
Amazon Rekognition

QUESTION 106
A company ingests machine learning (ML) data from web advertising clicks into an Amazon S3 data lake. Click data is added to an Amazon Kinesis data stream by using the Kinesis Producer Library (KPL). The data is loaded into the S3 data lake from the data stream by using an Amazon Kinesis Data Firehose delivery stream. As the data volume increases, an ML specialist notices that the rate of data ingested into Amazon S3 is relatively constant. There also is an increasing backlog of data for Kinesis Data Streams and Kinesis Data Firehose to ingest.
Which next step is MOST likely to improve the data ingestion rate into Amazon S3?

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

Correct Answer: C
Section:
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.
References:
Shard - Amazon Kinesis Data Streams
Scaling Amazon Kinesis Data Streams with AWS CloudFormation - AWS Big Data Blog

QUESTION 107
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.

Correct Answer: B, D
Section:
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.
References:
Feature Engineering for Categorical Data
How to Perform Feature Selection with Categorical Data

QUESTION 108
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.

Correct Answer: A
Section:
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 environment1.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 agents3. 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 agents4.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 actions5.
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.
References:
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 109
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.

Correct Answer: D
Section:
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 counts1.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 words1. By using this function, the data scientist can remove the stop words such as ''a,'' ''an,'' and ''the'' 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 documents2.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 topic3. 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 data5. 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 text2.
Option C is not suitable because the SageMaker built-in Object Detection algorithm is used to detect and localize objects in images6. 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).
References:
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 110
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 company's data currently resides on premises and is 40 in size.
The company wants a solution that can transfer and automatically update data between the on-premises 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.

Correct Answer: C
Section:
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 services1.AWS DataSync can copy data between on-premises object storage and Amazon S3, and also supports encryption, scheduling, monitoring, and data integrity validation1.
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 AWS2.The DataSync agent can be deployed as a virtual machine (VM) on your existing hypervisor, or as an Amazon EC2 instance in your AWS account2.
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 transfer3.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 prefixes3.
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 locations4.A sync task transfers only the data that has changed or that doesn't exist in the destination, and also deletes any files or objects from the destination that were deleted from the source since the last sync4.
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.
References:
Data Transfer Service - AWS DataSync
Deploying a DataSync Agent
Creating a Task
Syncing Data with AWS DataSync

QUESTION 111
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.

Correct Answer: D
Section:

QUESTION 112
A data engineer at a bank is evaluating a new tabular dataset that includes customer data. 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.

Correct Answer: B, C
Section:
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.References: =
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 113
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.

Correct Answer: C
Section:
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 analytics1.AWS Glue can run Python and Scala scripts to process data from various sources, such as

Amazon S3, Amazon DynamoDB, Amazon Redshift, and more2.AWS Glue also provides a serverless Apache Spark environment to run ETL jobs, eliminating the need to provision and manage servers3.
PySpark is the Python API for Apache Spark, a unified analytics engine for large-scale data processing4.PySpark can perform various data transformations and manipulations on structured and unstructured data, such as cleaning, enriching, and compressing5.PySpark can also leverage the distributed computing power of Spark to handle terabytes of data efficiently and scalably6.
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 job7.Moreover, the company can use the AWS Glue Data Catalog to store and manage the metadata of the data sources and targets8.
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 company's 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 company's 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 company's use case.
References:
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 114
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 company's 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 company's products

Correct Answer: A, B, D

Section:
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 company's 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.
References:
Detect sentiment from customer reviews using Amazon Comprehend | AWS Machine Learning Blog
How to Apply Machine Learning to Customer Feedback

QUESTION 115
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.

Correct Answer: D
Section:
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.
References:
Amazon SageMaker Experiments
Process Data and Evaluate Models

QUESTION 116
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.

Correct Answer: D
Section:
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.References:
Access Training Data - Amazon SageMaker
Mount an EFS file system to an Amazon SageMaker notebook (with lifecycle configurations) | AWS Machine Learning Blog

QUESTION 117
A retail company uses a machine learning (ML) model for daily sales forecasting. The company's 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 company's 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.

Correct Answer: B
Section:
Explanation:
The best way to visualize the model's 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 model's performance.References:
Histogram - Wikipedia
Model error - Wikipedia
SageMaker Model Monitor - visualizing monitoring results

QUESTION 118
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.

Correct Answer: D
Section:
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.References:
What is XGBoost? - Amazon SageMaker
What is Clustering? - Amazon SageMaker
K-Means Algorithm - Amazon SageMaker