AWS Certified ML Specialty Exam Questions

Question 1

Question:
An ecommerce company wants to train a large image classification model with 10,000 classes. The company runs multiple model training iterations and needs to minimize operational overhead and cost. The company also needs to avoid loss of work and model retraining.
Which solution will meet these requirements?

Options:

- A. Create the training jobs as AWS Batch jobs that use Amazon EC2 Spot Instances in a managed compute environment.
- B. Use Amazon EC2 Spot Instances to run the training jobs. Use a Spot Instance interruption notice to save a snapshot of the model to Amazon S3 before an instance is terminated.
- C. Use AWS Lambda to run the training jobs. Save model weights to Amazon S3.
- D. Use managed spot training in Amazon SageMaker. Launch the training jobs with checkpointing enabled.

Answer: **D**

Explanation:

- Amazon SageMaker's managed spot training minimizes costs and reduces operational overhead by using Spot Instances.
- It also supports **checkpointing**, which saves model progress to Amazon S3. This ensures the training job can resume without loss of work in case of Spot Instance interruptions.
- Other options, such as Lambda, are not suitable for long-running tasks like model training. Spot Instances without managed SageMaker training require manual checkpointing.

Question 2

Question:
A data scientist is developing a pipeline to ingest streaming web traffic data. The data scientist needs to implement a process to identify unusual web traffic patterns. The solution must:

1. Calculate an anomaly score for each web traffic entry.

2. Adapt unusual event identification to changing web patterns over time.

Which approach should the data scientist implement to meet these requirements?

Options:

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

Answer: **D**

Explanation:

- Amazon RCF is specifically designed for anomaly detection. Using **Amazon Managed Service for Apache Flink** with **RCF SQL extensions** provides real-time anomaly detection.
- Sliding windows are better suited for dynamic event streams as they allow for flexible time-based anomaly detection.
- Other options (A and B) rely on Lambda-based enrichment, which introduces latency and lacks adaptability over time. Option C does not utilize the specialized RCF algorithm.

Question 3

Question:
A company wants to conduct targeted marketing to sell solar panels to homeowners. 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 working on the project. The internal team has no ML expertise or experience.
Which solution will meet these requirements with the **LEAST** amount of effort from the internal team?

Options:

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

Answer: **D**

Explanation:

- A **public workforce** reduces effort for the internal team as Amazon provides trained and scalable workforces.
- SageMaker Ground Truth with a public workforce simplifies labeling for the company, and the SageMaker Object Detection algorithm ensures effective training and inference.
- Private workforces require setup and additional management from the team, increasing their effort.

Question 4

Question:
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 stored in Amazon S3. 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?

Options:

- 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.

Answer: **B**

Explanation:

- Using Amazon Mechanical Turk with **SageMaker Ground Truth** is the least effort approach for labeling. Amazon provides built-in task types and manifests for image classification.
- Options like private workforces (C) and building custom tools (A, D) increase development complexity and effort.

Question 5

Question:
A data scientist is designing a repository that will contain many images of vehicles. The repository must scale automatically to store new images every day. The repository must support versioning of the images. The data scientist must implement a solution that maintains multiple immediately accessible copies of the data in different AWS Regions.
Which solution will meet these requirements?

Options:

- A. Amazon S3 with S3 Cross-Region Replication (CRR)
- B. Amazon Elastic Block Store (Amazon EBS) with snapshots that are shared in a secondary Region
- C. Amazon Elastic File System (Amazon EFS) Standard storage that is configured with Regional availability
- D. AWS Storage Gateway Volume Gateway

Answer: **A**

Explanation:

- **Amazon S3 with Cross-Region Replication (CRR)** is designed to replicate data automatically across multiple AWS Regions. It supports **versioning**, ensuring previous versions of images are retained.
- EBS snapshots are not suitable for frequent updates or multi-region replication.
- EFS is region-specific and does not support automatic cross-region replication.
- Storage Gateway is for hybrid storage use cases, not for scalable cloud-native solutions.

Question 6

Question:
A bank has collected customer data for 10 years in CSV format. The bank stores the data on an on-premises server. A data science team wants to use Amazon SageMaker to build and train a machine learning model to predict churn probability. The team will use the historical data. The data scientists want to perform data transformations quickly and to generate data insights before training the model.
Which solution will meet these requirements with the **LEAST** development effort?

Options:

- A. Upload the data into the SageMaker Data Wrangler console directly. Perform data transformations and generate insights within Data Wrangler.
- B. Upload the data into an Amazon S3 bucket. Allow SageMaker to access the data that is in the bucket. Import the data from the S3 bucket into SageMaker Data Wrangler. Perform data transformations and generate insights within Data Wrangler.
- C. Upload the data into the SageMaker Data Wrangler console directly. Allow SageMaker and Amazon QuickSight to access the data that is in an Amazon S3 bucket. Perform data transformations in Data Wrangler and save the transformed data into a second S3 bucket. Use QuickSight to generate data insights.
- D. Upload the data into an Amazon S3 bucket. Allow SageMaker to access the data that is in the bucket. Import the data from the bucket into SageMaker Data Wrangler. Perform data transformations in Data Wrangler. Save the data into a second S3 bucket. Use a SageMaker Studio notebook to generate data insights.

Answer: **A**

Explanation:

- Directly uploading data into SageMaker **Data Wrangler** provides an easy-to-use, interactive environment for **data preparation, transformation, and analysis**, reducing development effort.
- Other options involve extra steps, like using S3 and SageMaker notebooks, which add unnecessary complexity.

Question 7

Question:
A company uses sensors on devices such as motor engines and factory machines to measure parameters, temperature, and pressure. The company wants to use the sensor data to predict equipment malfunctions and reduce service outages.
A machine learning (ML) specialist needs to gather the sensor data to train a model to predict device malfunctions. The ML specialist must ensure that the data does not contain outliers before training the model.
How can the ML specialist meet these requirements with the **LEAST operational overhead**?

Options:

- A. Load the data into an Amazon SageMaker Studio notebook. Calculate the first and third quartile. Use a SageMaker Data Wrangler data flow to remove only values that are outside of those quartiles.
- B. Use an Amazon SageMaker Data Wrangler bias report to find outliers in the dataset. Use a Data Wrangler data flow to remove outliers based on the bias report.
- C. Use an Amazon SageMaker Data Wrangler anomaly detection visualization to find outliers in the dataset. Add a transformation to a Data Wrangler data flow to remove outliers.
- D. Use Amazon Lookout for Equipment to find and remove outliers from the dataset.

Answer: **C**

Explanation:

- SageMaker **Data Wrangler** has built-in anomaly detection tools to visualize and detect outliers with minimal operational overhead.
- Using Lookout for Equipment (Option D) is overkill and requires setting up an ML model for anomaly detection, which is not necessary in this case.

Question 8

Question:
A gaming company has launched an online game where people can start playing for free, but they need to pay if they choose to use certain features. The company needs to build an automated system to predict whether or not a new user will become a paid user within 1 year. The company has gathered a labeled dataset from 1 million users.
The training dataset consists of 1,000 positive samples (users who became paid users) and 999,000 negative samples (users who did not). Using this dataset for training, the data science

team trained a random forest model that converged with over 99% accuracy on the training set. However, the prediction results on a test dataset were not satisfactory.
Which of the following approaches should the data science team take to mitigate this issue? (Select **TWO**.)

Options:

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

Answer: **C, D**

Explanation:

- **Option C**: Duplicating positive samples and adding noise addresses the class imbalance by generating more positive samples, improving the model's ability to predict minority classes.
- **Option D**: Changing the cost function to penalize false negatives helps in prioritizing the correct identification of positive users.

Question 9

Question:
A machine learning (ML) specialist is running an Amazon SageMaker hyperparameter optimization job for a model that is based on the XGBoost algorithm. The ML specialist selects Root Mean Square Error (RMSE) as the objective evaluation metric.
The ML specialist discovers that the model is overfitting and cannot generalize well on the validation data. The ML specialist decides to resolve the model overfitting by using SageMaker automatic model tuning (AMT).
Which solution will meet this requirement?

Options:

- A. Configure SageMaker AMT to use a static range of hyperparameter values.
- B. Configure SageMaker AMT to increase the number of parallel training jobs.
- C. Configure SageMaker AMT to stop training jobs early.
- D. Configure SageMaker AMT to run the training jobs with a warm start.

Answer: **C**

Explanation:

- SageMaker **Automatic Model Tuning** supports early stopping, which can help prevent overfitting by stopping poorly performing models during training.
- Increasing parallel training jobs (Option B) or warm starts (Option D) do not address overfitting directly.
- Static hyperparameter ranges (Option A) might limit the exploration of optimal values, reducing tuning effectiveness.

Question 10

Question:
A company distributes an online multiple-choice survey to several thousand people. Respondents to the survey can select multiple options for each question.
A machine learning (ML) engineer needs to comprehensively represent every response from all respondents in a dataset. The ML engineer will use the dataset to train a logistic regression model.
Which solution will meet these requirements?

Options:

- A. Perform one-hot encoding on every possible option for each question of the survey.
- B. Perform binning on all the answers each respondent selected for each question.
- C. Use Amazon Mechanical Turk to create categorical labels for each set of possible responses.
- D. Use Amazon Textract to create numeric features for each set of possible responses.

Answer: **A**

Explanation:

- **One-hot encoding** is the standard approach to represent categorical data in a format suitable for machine learning models, ensuring each option is treated as an independent feature.
- Other options, like binning or Textract, are unsuitable for this structured categorical data.

Question 11

Question:
A manufacturing company wants to monitor its devices for anomalous behavior. A data scientist

has trained an Amazon SageMaker scikit-learn model that classifies a device as normal or anomalous based on its 4-day telemetry. The 4-day telemetry of each device is collected in a separate file and is placed in an Amazon S3 bucket once every hour. The total time to run the model across the telemetry for all devices is 5 minutes.
What is the **MOST** cost-effective solution for the company to use to run the model across the telemetry for all the devices?

Options:

- A. SageMaker Batch Transform
- B. SageMaker Asynchronous Inference
- C. SageMaker Processing
- D. A SageMaker multi-container endpoint

Answer: **A**

Explanation:

- **Batch Transform** is the most cost-effective option when predictions need to be run on a batch of data at regular intervals.
- Asynchronous Inference (Option B) and multi-container endpoints (Option D) are better suited for real-time or near real-time use cases, which are unnecessary in this scenario.
- SageMaker Processing (Option C) is used for pre-processing, not inference.

Question 12

Question:
A media company is building a computer vision model to analyze images that are shared 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?

Options:

- A. Instead of File mode, configure the SageMaker training job to use Pipe mode. Ingest the data from a pipe.
- B. Instead of File mode, configure the SageMaker training job to use FastFile mode with no other changes.

- C. Instead of On-Demand Instances, configure the SageMaker training job to use Spot Instances. Make no other changes.
- D. Instead of On-Demand Instances, configure the SageMaker training job to use Spot Instances. Implement model checkpoints.

Answer: **D**

Explanation:

- Using **Spot Instances** reduces costs, and implementing model checkpoints ensures progress is saved in case of interruptions.
- Pipe mode (Option A) and FastFile mode (Option B) reduce data ingestion time but do not directly lower costs.
- Spot Instances without model checkpoints (Option C) risk losing training progress if interrupted.

Question 13

Question:
A manufacturing company needs to identify returned smartphones that have been damaged by moisture. The company has an automated process that produces 2,000 diagnostic values for each phone. The database contains more than five million phone evaluations. The evaluation process is consistent, and there are no missing values in the data. A machine learning (ML) specialist has trained an Amazon SageMaker linear learner ML model to classify phones as moisture damaged or not moisture damaged by using all available features. The model's F1 score is 0.6.
Which changes in model training would **MOST** likely improve the model's F1 score? (Select **TWO**.)

Options:

- A. Continue to use the SageMaker linear learner algorithm. Reduce the number of features with the SageMaker principal component analysis (PCA) algorithm.
- B. Continue to use the SageMaker linear learner algorithm. Reduce the number of features with the scikit-learn multi-dimensional scaling (MDS) algorithm.
- C. Continue to use the SageMaker linear learner algorithm. Set the predictor type to regressor.
- D. Use the SageMaker k-means algorithm with k of less than 1,000 to train the model.
- E. Use the SageMaker k-nearest neighbors (k-NN) algorithm. Set a dimension reduction target of less than 1,000 to train the model.

Answer: **A, E**

Explanation:

- **PCA** (Option A) and **dimension reduction with k-NN** (Option E) are effective for high-dimensional data. Reducing dimensionality helps eliminate irrelevant features and improves the model's focus on significant ones.
- The other options do not directly address high-dimensional data or improve F1 score performance.

Question 14

Question:
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 use a machine learning (ML) model to determine when additional information should be requested.
The company has several terabytes of data from its existing e-commerce 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?

Options:

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

Answer: **B**

Explanation:

- **Amazon SageMaker IP Insights** is specifically designed for detecting unusual access patterns in log files based on IP addresses.
- Factorization Machines (Option A) and Object2Vec (Option D) are more suited for recommendation and text similarity tasks, respectively.
- Ground Truth labeling (Option C) is unnecessary because IP Insights can operate without pre-labeled data.

Question 15

Question:
A media company wants to deploy a machine learning (ML) model that uses Amazon SageMaker to recommend new articles to the company's readers. The company's readers are primarily located in a single city.
The company notices that the heaviest reader traffic predictably occurs early in the morning, after lunch, and again after work hours. There is very little traffic at other times of day. The media company needs to minimize the time required to deliver recommendations to its readers. The expected amount of data that the API call will return for inference is less than 4 MB.
Which solution will meet these requirements in the **MOST** cost-effective way?

Options:

- A. Real-time inference with auto scaling
- B. Serverless inference with provisioned concurrency
- C. Asynchronous inference
- D. A batch transform task

Answer: **B**

Explanation:

- **Serverless inference** is cost-effective for workloads with variable demand, as you only pay for what you use. Provisioned concurrency ensures low latency during peak traffic times.
- Real-time inference (Option A) can be more expensive due to always-on resources.
- Asynchronous inference (Option C) and batch transform (Option D) are not ideal for low-latency requirements.

Question 16

Question:
An online delivery company wants to choose the fastest courier for each delivery at the moment an order is placed. The company wants to implement this feature for existing users and new users of its application. Data scientists have trained separate models with XGBoost for existing users and new users, and the models are stored in Amazon S3. There is one model for each city where the company operates.
Operation engineers are hosting these models in Amazon EC2 for responding to requests in real time. However, the existing instances are under low utilization in CPU and memory.

Which solution will enable the company to achieve its goal with the **LEAST** operational overhead?

Options:

- A. Create an Amazon SageMaker notebook instance for pulling all the models from Amazon S3 using the boto3 library. Remove the existing instances and use the notebook to perform a SageMaker batch transform for performing inferences offline for all the cities. Store the results in different files in Amazon S3. Point the web client to the files.
- B. Prepare an Amazon SageMaker Docker container based on the open-source multi-model server. Remove the existing instances and create a multi-model endpoint in SageMaker instead, pointing to the S3 bucket containing all the models. Invoke the endpoint from the web client at runtime, specifying the TargetModel parameter according to the city of each request.
- C. Keep only a single EC2 instance for hosting all the models. Install a model server in the instance and load each model by pulling it from Amazon S3. Integrate the instance with the web client using Amazon API Gateway for responding to the requests in real time.
- D. Prepare a Docker container based on the prebuilt images in Amazon SageMaker. Replace the existing instances with separate SageMaker endpoints, one for each city where the company operates. Invoke the endpoints from the web client, specifying the URL and EndpointName parameter according to the city of each request.

Answer: **B**

Explanation:

- **Multi-model endpoints in SageMaker** are designed for hosting multiple models efficiently with low overhead. They allow you to invoke specific models dynamically based on the TargetModel parameter, reducing operational complexity.
- Other options either increase overhead or do not efficiently handle real-time dynamic requests for multiple models.

Question 17

Question:
A data scientist for a medical diagnostic testing company has developed a machine learning (ML) model to identify patients who have a specific disease. The dataset that the scientist used to train the model is imbalanced. The dataset contains a large number of healthy patients and only a small number of patients who have the disease. The model should consider that patients who are incorrectly identified as positive for the disease will increase costs for the company.
Which metric will **MOST** accurately evaluate the performance of this model?

Options:

- A. Recall
- B. F1 score
- C. Accuracy
- D. Precision

Answer: **D**

Explanation:

- **Precision** is the most appropriate metric when false positives need to be minimized because it measures the proportion of true positives among all predicted positives.
- Recall (Option A) focuses on capturing true positives, which may not align with cost concerns.
- F1 score (Option B) balances precision and recall but does not specifically address minimizing false positives.
- Accuracy (Option C) is not suitable for imbalanced datasets.

Question 18

Question:
A financial services company wants to automate its loan approval process by building a machine learning (ML) model. Each loan data point contains credit history from a third-party data source and demographic information about the customer. Each loan approval prediction must come with a report that contains an explanation for why the customer was approved for a loan or was denied for a loan. The company will use Amazon SageMaker to build the model.
Which solution will meet these requirements with the **LEAST** development effort?

Options:

- A. Use SageMaker Model Debugger to automatically debug the predictions, generate the explanation, and attach the explanation report.
- B. Use AWS Lambda to provide feature importance and partial dependence plots. Use the plots to generate and attach the explanation report.
- C. Use SageMaker Clarify to generate the explanation report. Attach the report to the predicted results.
- D. Use custom Amazon CloudWatch metrics to generate the explanation report. Attach the report to the predicted results.

Answer: **C**

Explanation:

- **SageMaker Clarify** is specifically designed for generating explainability reports for machine learning models. It provides insights into feature importance and bias detection, making it the best option for generating explanation reports with minimal development effort.
- Other options, such as Lambda (Option B) or CloudWatch (Option D), require significant custom development to achieve similar functionality.

Question 19

Question:
A company is setting up a mechanism for data scientists and engineers from different departments to access an Amazon SageMaker Studio domain. Each department has a unique SageMaker Studio domain.
The company wants to build a central proxy application that data scientists can log in to by using their corporate credentials. The proxy application will authenticate users by using the company's existing Identity provider (IdP). The application will then route users to the appropriate SageMaker Studio domain.
The company plans to maintain a table in Amazon DynamoDB that contains SageMaker domains for each department.
How should the company meet these requirements?

Options:

- A. Use the SageMaker CreatePresignedDomainUrl API to generate a presigned URL for each domain according to the DynamoDB table. Pass the presigned URL to the proxy application.
- B. Use the SageMaker CreateHumanTaskUi API to generate a UI URL. Pass the URL to the proxy application.
- C. Use the Amazon SageMaker ListHumanTaskUis API to list all UI URLs. Pass the appropriate URL to the DynamoDB table so that the proxy application can use the URL.
- D. Use the SageMaker CreatePresignedNotebookInstanceUrl API to generate a presigned URL. Pass the presigned URL to the proxy application.

Answer: **A**

Explanation:

- The **CreatePresignedDomainUrl API** generates presigned URLs that allow access to specific SageMaker Studio domains. This is the best approach for implementing role-based access to SageMaker domains while integrating with a proxy application.
- Other options, such as HumanTaskUi APIs (Options B and C), are irrelevant to this use case as they are designed for labeling jobs.

Question 20

Question:
A healthcare company is planning to use neural networks to classify their X-ray images into normal and abnormal classes. The labeled data is divided into a training set of 1,000 images and a test set of 200 images. The initial training of a neural network model with 50 hidden layers yielded 99% accuracy on the training set but only 56% accuracy on the test set.
What changes should the specialist consider to solve this issue? (**Select THREE**)

Options:

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

Answer: **B, D, F**

Explanation:

- **B (Choose a lower number of layers)**: Reducing the number of layers can prevent overfitting, especially in scenarios where the dataset size is small.
- **D (Enable dropout)**: Dropout helps prevent overfitting by randomly deactivating certain neurons during training.
- **F (Enable early stopping)**: Early stopping halts training when validation loss stops improving, further mitigating overfitting.
- Increasing layers (Option A) or including test data in the training set (Option E) would not address the problem and could lead to poor generalization.

Question 21

Question:
An agriculture company wants to improve crop yield forecasting for the upcoming season by using crop yields from the last three seasons. The company wants to compare the performance of its new scikit-learn model to the benchmark.
A data scientist needs to package the code into a container that computes both the new model forecast and the benchmark. The data scientist wants AWS to be responsible for the operational maintenance of the container.
Which solution will meet these requirements?

Options:

- A. Package the code as the training script for an Amazon SageMaker scikit-learn container.
- B. Package the code into a custom-built container. Push the container to Amazon Elastic Container Registry (Amazon ECR).
- C. Package the code into a custom-built container. Push the container to AWS Fargate.
- D. Package the code by extending an Amazon SageMaker scikit-learn container.

Answer: **D**

Explanation:

- Extending an **Amazon SageMaker scikit-learn container** allows the data scientist to use a prebuilt container that includes the necessary dependencies while also letting AWS handle the operational maintenance.
- Custom-built containers (Options B and C) increase operational effort.
- Using a training script (Option A) is not appropriate for this scenario, as it requires manual container setup.

Question 22

Question:
An ecommerce company has developed an XGBoost model in Amazon SageMaker to predict whether a customer will return a purchased item. The dataset is imbalanced, as only 5% of customers return items.
A data scientist must find the hyperparameters to capture as many instances of returned items as possible. The company has a small budget for compute.
How should the data scientist meet these requirements **MOST cost-effectively**?

Options:

- A. Tune all possible hyperparameters by using automatic model tuning (AMT). Optimize on `{"HyperParameterTuningJobObjective": {"MetricName": "validation:accuracy", "Type": "Maximize"}}`.
- B. Tune the `csv_weight` hyperparameter and the `scale_pos_weight` hyperparameter by using automatic model tuning (AMT). Optimize on `{"HyperParameterTuningJobObjective": {"MetricName": "validation:f1", "Type": "Maximize"}}`.
- C. Tune all possible hyperparameters by using automatic model tuning (AMT). Optimize on `{"HyperParameterTuningJobObjective": {"MetricName": "validation:f1", "Type": "Maximize"}}`.

- D. Tune the `csv_weight` hyperparameter and the `scale_pos_weight` hyperparameter by using automatic model tuning (AMT). Optimize on `{"HyperParameterTuningJobObjective": {"MetricName": "validation:f1", "Type": "Minimize"}}`.

Answer: **B**

Explanation:

- Tuning **only `csv_weight` and `scale_pos_weight`** hyperparameters addresses the imbalance problem by focusing on the weights of the positive class (returns) without incurring high costs.
- Optimizing for **F1 score** is suitable for imbalanced datasets as it balances precision and recall.
- Tuning all hyperparameters (Options A and C) would be costlier and unnecessary.
- Option D is incorrect because minimizing F1 does not make sense for this scenario.

Question 23

Question:
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 10x10 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?

Options:

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

- D. Prepare the images in Apache Parquet format and upload them to Amazon S3. Use Amazon SageMaker to train, test, and validate the model using an image classification algorithm to categorize images into various weed classes.

Answer: **C**

Explanation:

- The **object-detection single-shot multibox detector (SSD) algorithm** is appropriate for detecting specific types of weeds and their locations.
- Preparing the images in **RecordIO format** is recommended for image-related tasks in SageMaker, as it supports efficient storage and processing of image data.
- Apache Parquet (Options B and D) is not optimized for image data.

Question 24

Question:
A trucking company is collecting live image data from its fleet of trucks across the globe. The data is growing rapidly and approximately 100 GB of new data is generated every day. The company wants to explore machine learning use cases while ensuring the data is only accessible to specific IAM users.
Which storage option provides the most processing flexibility and will allow access control with IAM?

Options:

- A. Use a database, such as Amazon DynamoDB, to store the images, and set the IAM policies to restrict access to only the desired IAM users.
- B. Use an Amazon S3-backed data lake to store the raw images, and set up the permissions using bucket policies.
- C. Set up Amazon EMR with Hadoop Distributed File System (HDFS) to store the files, and restrict access to the EMR instances using IAM policies.
- D. Configure Amazon EFS with IAM policies to make the data available to Amazon EC2 instances owned by the IAM users.

Answer: **B**

Explanation:

- An **Amazon S3-backed data lake** is the most flexible option for storing large amounts of data, as it provides scalability, supports various ML workflows, and integrates with AWS IAM for fine-grained access control.
- DynamoDB (Option A) is not suitable for storing large image files.

- EMR (Option C) and EFS (Option D) are more complex and less efficient for handling this use case.

Question 25

Question:
A machine learning (ML) specialist is developing a model for a company. The model will classify and predict sequences of objects that are displayed in a video. The ML specialist decides to use a hybrid architecture that consists of a convolutional neural network (CNN) followed by a classifier three-layer recurrent neural network (RNN).
The company developed a similar model previously but trained the model to classify a different set of objects. The ML specialist wants to save time by using the previously trained model and adapting the model for the current use case and set of objects.
Which combination of steps will accomplish this goal with the **LEAST amount of effort**? (Select **TWO**)

Options:

- A. Reinitialize the weights of the entire CNN. Retrain the CNN on the classification task by using the new set of objects.
- B. Reinitialize the weights of the entire network. Retrain the entire network on the prediction task by using the new set of objects.
- C. Reinitialize the weights of the entire RNN. Retrain the entire model on the prediction task by using the new set of objects.
- D. Reinitialize the weights of the last fully connected layer of the CNN. Retrain the CNN on the classification task by using the new set of objects.
- E. Reinitialize the weights of the last layer of the RNN. Retrain the entire model on the prediction task by using the new set of objects.

Answer: **D, E**

Explanation:

- **D**: Retraining only the last fully connected layer of the CNN is efficient because lower CNN layers often learn generic features that can be reused.
- **E**: Retraining only the last layer of the RNN ensures that the model is fine-tuned for the new dataset without retraining the entire architecture.
- Options A, B, and C require retraining the entire model, which is unnecessary and time-consuming.

Question 26

Question:
A machine learning (ML) model for predicting house prices is performing poorly on validation data. The validation accuracy is low, and the model frequently predicts outliers. The feature set includes numeric columns for the number of rooms, the lot size, and other continuous variables.
What preprocessing step should the data scientist take to improve the model's accuracy?

Options:

- A. Apply feature binning to the numeric columns.
- B. Apply one-hot encoding to all the numeric columns.
- C. Normalize the problematic features.
- D. Perform a grid search over the numeric features.
- E. Remove the problematic features from the training dataset.

Answer: **C**

Explanation:

- **Normalization** scales numeric features to the same range, ensuring they contribute equally to the model's learning process. This reduces bias caused by disproportionate feature values.
- Binning (Option A) may lead to information loss, while one-hot encoding (Option B) is not applicable to numeric data.
- Removing features (Option E) discards potentially valuable information.

Question 27

Question:
A manufacturing company wants to create a pipeline for processing real-time telemetry data from IoT devices to predict potential machine failures. The company wants to use Amazon SageMaker for model training and inference. The processed data must also be stored in Amazon S3 for future analysis.
Which solution will meet these requirements?

Options:

- A. Use Amazon Kinesis Data Streams to ingest real-time telemetry data. Use Amazon Kinesis Data Analytics to process the data. Invoke a SageMaker endpoint for inference and store the processed data in Amazon S3.
- B. Use Amazon S3 to store the raw telemetry data. Use AWS Glue to perform ETL. Use SageMaker for inference and store the processed data in an S3 bucket.
- C. Use AWS IoT Core to ingest real-time telemetry data. Use AWS Lambda to invoke a SageMaker endpoint for inference. Use Amazon S3 to store the processed data.

- D. Use Amazon Kinesis Data Firehose to ingest real-time telemetry data. Use AWS Glue to process the data. Use SageMaker for inference and store the results in an Amazon Redshift cluster.

Answer: **A**

Explanation:

- **Amazon Kinesis Data Streams** provides real-time data ingestion, and **Kinesis Data Analytics** enables real-time processing for telemetry data.
- SageMaker endpoints allow inference, and processed results can be directly stored in S3 for analysis.
- AWS Glue (Options B and D) is better for batch ETL processing, not real-time data. IoT Core (Option C) adds unnecessary complexity.

Question 28

Question:
A company's on-premises storage system produces 1 TB of log data every week. The data needs to be encrypted and transferred to Amazon S3 for analysis. The transfer process must run weekly, and the solution must support incremental updates to the data during the transfer process.
Which solution will meet these requirements?

Options:

- A. Use the S3 sync command with encryption enabled to transfer the data to S3 every week.
- B. Use AWS Transfer Family to transfer the data to S3 by using FTPS.
- C. Use AWS DataSync to make an initial copy of the data to S3. Schedule subsequent incremental transfers to S3.
- D. Use AWS Snowball Edge to transfer the data to S3 every week.

Answer: **C**

Explanation:

- **AWS DataSync** is the best solution for transferring large datasets with incremental updates. It supports encryption, scheduling, and validating data integrity.
- S3 sync (Option A) requires manual execution and lacks validation features. Transfer Family (Option B) and Snowball Edge (Option D) are not ideal for regular incremental updates.

Question 29

Question:
A data scientist is preparing a dataset for regression modeling. The dataset includes a categorical feature, `Wall_Color`, which has three unique values: `Red`, `Green`, and `Blue`.
What preprocessing steps should the data scientist take to prepare the `Wall_Color` feature?

Options:

- A. Replace each color with integers: Red = 1, Green = 2, Blue = 3.
- B. Add new columns that store one-hot representation of colors.
- C. Replace the color names with their string lengths.
- D. Create three columns to encode the color in RGB format.
- E. Replace each color name with its training set frequency.

Answer: **B, D**

Explanation:

- **B (One-hot encoding)**: Converts the categorical feature into independent binary variables, avoiding unintended ordinal relationships.
- **D (RGB encoding)**: Encodes the colors as numerical RGB values, preserving useful color information.
- Using integers (Option A) introduces ordinal bias, and string length or frequency encoding (Options C and E) is not meaningful for this context.

Question 30

Question:
A media company needs to detect whether user-submitted images contain their company logo. The company plans to use an ML model for image classification.
Which algorithm should the company choose for this task?

Options:

- A. Principal Component Analysis (PCA)
- B. Recurrent Neural Network (RNN)
- C. Convolutional Neural Network (CNN)
- D. K-nearest neighbors (k-NN)

Answer: **C**

Explanation:

- **CNNs** are specifically designed for image-related tasks, as they excel at extracting spatial features from images.
- PCA (Option A) is for dimensionality reduction, RNNs (Option B) are for sequential data, and k-NN (Option D) is not optimized for large-scale image classification.

Question 31

Question:
A data scientist notices low accuracy for both training and test datasets when evaluating an ML model.
What actions should the data scientist take to improve the model's performance? (Select **TWO**.)

Options:

- A. Increase the number of hidden layers in the model.
- B. Increase the number of training examples.
- C. Decrease the learning rate of the model.
- D. Add new domain-specific features to the training dataset.
- E. Decrease the number of features used in the training dataset.

Answer: **B, D**

Explanation:

- **B (Increase training examples)**: Adding more data reduces underfitting and improves model performance.
- **D (Add domain-specific features)**: Enhances the dataset by providing more relevant information for the model to learn from.
- Increasing hidden layers (Option A) or decreasing features (Option E) may worsen underfitting, while adjusting the learning rate (Option C) does not address the root issue.

Question 32

Question:
A data scientist trained a machine learning (ML) model on an imbalanced dataset where only 2% of the samples represent the positive class. The model is consistently predicting the negative class for most samples.
Which techniques should the data scientist use to improve the model's performance? (Select **TWO**.)

Options:

- A. Apply Synthetic Minority Oversampling Technique (SMOTE) to the training dataset.
- B. Use anomaly detection algorithms instead of binary classification algorithms.
- C. Use class weighting to adjust for the class imbalance.
- D. Reduce the training dataset size by undersampling the negative class.
- E. Increase the size of the training dataset by duplicating the positive class samples.

Answer: **A, C**

Explanation:

- **A (SMOTE)**: Generates synthetic samples for the minority class to balance the dataset, improving model performance.
- **C (Class weighting)**: Assigns higher weights to the positive class, encouraging the model to focus on minority class predictions.
- Options B, D, and E are less effective: anomaly detection is not appropriate for classification, undersampling reduces data, and duplicating samples can lead to overfitting.

Question 33

Question:
A machine learning (ML) specialist is hosting a recommendation model as a SageMaker endpoint for a video streaming application. The company requires the endpoint to have high availability with a recovery time objective (RTO) of 5 minutes.
What should the ML specialist do to meet this requirement?

Options:

- A. Deploy multiple instances for each endpoint in a VPC that spans at least two Regions.
- B. Use the SageMaker auto-scaling feature for the hosted recommendation models.
- C. Deploy multiple instances for each production endpoint in a VPC that spans at least two subnets that are in a second Availability Zone.
- D. Frequently generate backups of the production recommendation model. Deploy the backups in a second Region.

Answer: **C**

Explanation:

- Deploying **multiple instances across multiple Availability Zones (AZs)** ensures high availability and meets the RTO requirement of 5 minutes.
- Cross-region deployments (Option A) are unnecessary for this RTO. Auto-scaling (Option B) optimizes load handling but does not address AZ failure. Backups (Option D) are better suited for disaster recovery, not high availability.

Question 34

Question:
A company wants to use a chatbot to answer common customer questions by searching the company's documentation. The chatbot must integrate with the documentation storage solution. The solution should require the **least development effort**.
Which solution meets these requirements?

Options:

- A. Use Amazon Kendra to index the documents. Integrate the chatbot with the Amazon Kendra Query API to answer customer questions.
- B. Use AWS Lambda to run queries on an Amazon S3 bucket containing the documents. Integrate the chatbot with the Lambda function.
- C. Store the documents in an Amazon OpenSearch Service cluster. Integrate the chatbot with the OpenSearch k-Nearest Neighbors (k-NN) query API to retrieve answers.
- D. Use Amazon Textract to extract text from the documents. Use a SageMaker endpoint to run inference and retrieve answers.

Answer: **A**

Explanation:

- **Amazon Kendra** is designed for document indexing and question answering. Its Query API simplifies integration with chatbots, minimizing development effort.
- Other options involve custom development or tools (e.g., OpenSearch, Textract) that are less efficient for this use case.

Question 35

Question:
A company collects data from IoT devices across multiple factories. The data includes temperature, humidity, and pressure readings. The company wants to predict when devices are likely to fail. The company also wants to store the transformed data for future analysis.
Which solution should the company implement?

Options:

- A. Use Amazon Kinesis Data Streams to ingest the data. Use Kinesis Data Analytics to preprocess the data and invoke a SageMaker endpoint for inference. Store the transformed data in Amazon S3.

- B. Use AWS IoT Core to ingest the data. Use AWS Lambda to preprocess the data and invoke a SageMaker endpoint for inference. Store the transformed data in Amazon DynamoDB.
- C. Use Amazon Kinesis Data Firehose to ingest the data. Use AWS Glue to preprocess the data and invoke a SageMaker endpoint for inference. Store the transformed data in Amazon Redshift.
- D. Use AWS IoT Analytics to ingest the data, preprocess the data, and run ML inference. Store the transformed data in Amazon S3.

Answer: **A**

Explanation:

- **Kinesis Data Streams and Kinesis Data Analytics** provide a scalable, real-time pipeline for ingesting and transforming IoT data. The transformed data is stored in **Amazon S3**, and SageMaker endpoints handle inference.
- IoT Core (Option B) and Firehose with Glue (Option C) are less suitable for real-time pipelines. IoT Analytics (Option D) is not ideal for complex transformations or predictive modeling.

Question 36

Question:
A data scientist notices that a machine learning (ML) model trained on a traffic sign classification dataset has high accuracy on the training set but low accuracy on the test set. The dataset contains labeled images of traffic signs, and the test dataset includes images captured in various lighting conditions.
What actions should the data scientist take to improve the model? (Select **TWO**.)

Options:

- A. Add more labeled images that include a variety of lighting conditions to the training dataset.
- B. Use image preprocessing techniques to normalize the lighting in the dataset.
- C. Increase the number of training epochs.
- D. Use dropout layers in the model architecture.
- E. Increase the learning rate of the model.

Answer: **A, B**

Explanation:

- **A (Add more labeled images)**: Adding diverse training data improves the model's ability to generalize across varying conditions.

- **B (Normalize lighting)**: Preprocessing helps address variations caused by lighting, reducing input data inconsistencies.
- Increasing epochs (Option C) risks overfitting, dropout (Option D) is not directly relevant to data issues, and increasing the learning rate (Option E) can destabilize training.

Question 37

Question:
A data scientist needs to identify data quality issues in a CSV dataset stored in an Amazon S3 bucket. The scientist must find missing values, invalid values, and outliers in the dataset. The solution should require the **least operational effort**.
What should the data scientist do?

Options:

- A. Create an AWS Glue job to transform the data into Apache Parquet format. Use an AWS Glue crawler to populate a Data Catalog, then run Athena queries to identify issues.
- B. Leave the dataset in CSV format. Use an AWS Glue crawler and Amazon Athena queries to identify issues.
- C. Import the dataset into Amazon SageMaker Data Wrangler. Use the Data Quality and Insights Report to detect missing and invalid values and outliers.
- D. Use the S3 Select feature to query the data for missing and invalid values.

Answer: **C**

Explanation:

- SageMaker **Data Wrangler** provides a built-in **Data Quality and Insights Report**, making it the simplest and most efficient tool for detecting missing values, invalid values, and outliers.
- Options A and B require more manual effort (e.g., writing Athena queries). S3 Select (Option D) has limited functionality for complex data quality checks.

Question 38

Question:
A company needs to deploy a machine learning (ML) model to an endpoint that can process inference requests in real time. The endpoint must dynamically load and use different models depending on the incoming request. The solution must minimize operational overhead.
Which solution will meet these requirements?

Options:

- A. Deploy each model as a separate Amazon SageMaker endpoint. Use an Amazon API Gateway to route requests to the appropriate endpoint.
- B. Use Amazon SageMaker multi-model endpoints to host all the models. Specify the TargetModel parameter in the request to select the appropriate model.
- C. Use Amazon EC2 instances with a custom application to dynamically load models. Deploy the application behind an Application Load Balancer.
- D. Use Amazon Lambda to load models dynamically based on the request.

Answer: **B**

Explanation:

- **SageMaker multi-model endpoints** are purpose-built for hosting multiple models on a single endpoint, dynamically loading the required model at runtime with the **TargetModel** parameter.
- Deploying separate endpoints (Option A) increases costs. Options C and D require custom implementations, increasing operational overhead.

Question 39

Question:
A healthcare company is using Amazon SageMaker to train a binary classification model that predicts whether patients have a certain condition. The dataset is highly imbalanced, with only 1% of the records being positive cases. The company wants to minimize false negatives to ensure patients with the condition are not missed.
Which evaluation metric should the company use?

Options:

- A. Precision
- B. Recall
- C. F1 Score
- D. Accuracy

Answer: **B**

Explanation:

- **Recall** is the most relevant metric when minimizing false negatives because it measures the proportion of true positives correctly identified.

- Precision (Option A) focuses on minimizing false positives, which is not the priority. F1 Score (Option C) balances precision and recall but does not emphasize minimizing false negatives. Accuracy (Option D) is not useful for imbalanced datasets.

Question 40

Question:
A transportation company wants to improve its vehicle routing process by using machine learning. The company collects GPS data from its vehicles and stores it in Amazon S3. A data scientist needs to preprocess the data, train a model using SageMaker, and schedule periodic retraining jobs based on new data.
Which solution will meet these requirements?

Options:

- A. Use AWS Glue for preprocessing. Use a SageMaker training job for training. Schedule the retraining job with Amazon EventBridge.
- B. Use AWS Glue for preprocessing. Use a SageMaker training job for training. Schedule the retraining job with AWS Lambda.
- C. Use SageMaker Processing for preprocessing. Use SageMaker pipelines for training and retraining.
- D. Use AWS Lambda for preprocessing. Use a SageMaker notebook instance for training. Schedule the retraining job with Amazon EventBridge.

Answer: **C**

Explanation:

- **SageMaker Processing** simplifies data preprocessing, and **SageMaker Pipelines** automate training and retraining workflows with built-in scheduling capabilities.
- Glue and Lambda (Options A, B, and D) add unnecessary complexity compared to SageMaker's integrated capabilities.

Question 41

Question:
A data scientist is training a model using the XGBoost algorithm in Amazon SageMaker. The model is performing poorly on the validation dataset. The data scientist suspects that the features may not be informative enough for the task.
What should the data scientist do to improve the model's performance?

Options:

- A. Perform feature engineering to create more domain-specific features.
- B. Increase the number of boosting rounds.
- C. Reduce the learning rate of the model.
- D. Use a different algorithm, such as linear regression.

Answer: **A**

Explanation:

- **Feature engineering** improves the dataset by adding more meaningful and domain-specific features, which enhances the model's ability to learn patterns.
- Increasing boosting rounds (Option B) or reducing the learning rate (Option C) does not address the root issue of inadequate features. Linear regression (Option D) is not suitable for complex relationships.

Question 42

Question:
A company has a dataset containing customer purchase behavior, including time-series data and demographic attributes. The company wants to predict which customers are likely to make a purchase within the next 30 days.
Which algorithm should the company use?

Options:

- A. Amazon SageMaker Random Cut Forest
- B. Amazon SageMaker DeepAR
- C. Amazon SageMaker XGBoost
- D. Amazon SageMaker k-Nearest Neighbors (k-NN)

Answer: **B**

Explanation:

- **DeepAR** is designed for time-series forecasting and can handle datasets with additional covariates (e.g., demographic attributes).
- Random Cut Forest (Option A) is for anomaly detection. XGBoost (Option C) is not optimized for sequential data, and k-NN (Option D) is not suitable for time-series tasks.

Question 43

Question:
A financial company is building an ML model to detect fraudulent transactions. The company has a large historical dataset of transactions, with only 2% labeled as fraudulent. The company wants to maximize the F1 score.
What should the company do to handle the imbalanced dataset?

Options:

- A. Oversample the minority class using SMOTE.
- B. Use anomaly detection instead of binary classification.
- C. Use class weighting to adjust for the imbalance.
- D. Apply both SMOTE and class weighting.

Answer: **D**

Explanation:

- Combining **SMOTE** (oversampling) with **class weighting** allows the model to learn from balanced data while still accounting for the importance of the minority class during training.
- Anomaly detection (Option B) is inappropriate for binary classification tasks.

Question 44

Question:
A social media company wants to detect and block inappropriate content uploaded by users. The company has a large dataset of labeled images and plans to train a machine learning (ML) model using Amazon SageMaker. The company wants to automatically preprocess and augment the dataset to improve the model's generalization.
What is the **MOST** efficient way to preprocess and augment the dataset?

Options:

- A. Use Amazon SageMaker Processing with a custom script to preprocess and augment the dataset.
- B. Use AWS Glue to preprocess the dataset and AWS Lambda to augment the dataset.
- C. Use Amazon SageMaker Data Wrangler to preprocess the dataset and create a SageMaker Processing job for augmentation.
- D. Use SageMaker built-in data augmentation tools during the model training process.

Answer: **D**

Explanation:

- **SageMaker's built-in data augmentation tools** streamline the process by applying standard transformations during training, eliminating the need for separate preprocessing and augmentation jobs.
- Options A, B, and C introduce additional complexity and overhead without adding significant benefits for this use case.

Question 45

Question:
A company is building an ML pipeline for a video streaming platform to recommend content to users. The pipeline must:

1. Continuously update with new user data.
2. Re-train the model weekly with the latest data.
3. Minimize operational overhead.
   Which solution meets these requirements?

Options:

- A. Use SageMaker Pipelines to orchestrate the pipeline. Schedule weekly retraining jobs using Amazon EventBridge.
- B. Use AWS Glue to preprocess the data. Use Lambda to trigger SageMaker training jobs and retrain the model weekly.
- C. Use Amazon Kinesis Data Streams to preprocess the data. Trigger SageMaker training jobs with Kinesis Data Analytics.
- D. Use SageMaker Model Monitor to retrain the model weekly with a cron schedule.

Answer: **A**

Explanation:

- **SageMaker Pipelines** automates the ML pipeline with built-in scheduling and orchestration capabilities, minimizing operational overhead.
- Lambda (Option B) and Kinesis (Option C) require custom implementations for scheduling and retraining. Model Monitor (Option D) is for monitoring model drift, not scheduling retraining.

Question 46

Question:
A company needs to classify customer complaints into different categories. The dataset contains labeled complaints in English. The company wants to use Amazon SageMaker built-in

algorithms for training.
Which algorithm should the company use?

Options:

- A. SageMaker k-means
- B. SageMaker XGBoost
- C. SageMaker BlazingText
- D. SageMaker Random Cut Forest

Answer: **C**

Explanation:

- **BlazingText** is specifically designed for text classification and natural language processing tasks.
- k-means (Option A) is for clustering, XGBoost (Option B) is not optimized for text data, and Random Cut Forest (Option D) is for anomaly detection.

Question 47

Question:
A media company wants to deploy a machine learning model for sentiment analysis. The company expects spikes in traffic during newsworthy events and wants to minimize costs during periods of low traffic.
Which solution will meet these requirements?

Options:

- A. Deploy the model using SageMaker real-time inference with auto-scaling enabled.
- B. Deploy the model using SageMaker serverless inference.
- C. Deploy the model using SageMaker asynchronous inference.
- D. Deploy the model using SageMaker batch transform.

Answer: **B**

Explanation:

- **Serverless inference** is cost-effective for workloads with variable traffic. It automatically scales based on demand and reduces costs during low traffic periods.
- Real-time inference (Option A) is costlier due to provisioned instances. Asynchronous inference (Option C) is for non-real-time requests, and batch transform (Option D) is unsuitable for real-time sentiment analysis.

Question 48

Question:
A healthcare company has trained a SageMaker model to predict whether a patient has a certain disease. The company needs to monitor the model in production to ensure it continues to perform well as new data becomes available.
Which solution will meet this requirement?

Options:

- A. Use SageMaker Model Monitor to detect data quality issues and model drift.
- B. Use SageMaker Clarify to detect model bias and fairness issues.
- C. Use Amazon CloudWatch to monitor endpoint latency and error rates.
- D. Use SageMaker Pipelines to automate retraining of the model weekly.

Answer: **A**

Explanation:

- **SageMaker Model Monitor** detects data quality issues and model drift by continuously analyzing incoming data and model predictions in production.
- SageMaker Clarify (Option B) is for detecting bias, CloudWatch (Option C) monitors infrastructure metrics, and Pipelines (Option D) does not provide monitoring capabilities.

Question 49

Question:
A retail company wants to forecast future sales of its products using historical sales data. The data includes additional features like promotions, holidays, and weather. The company plans to use Amazon SageMaker.
Which algorithm is the **BEST** choice for this task?

Options:

- A. SageMaker XGBoost
- B. SageMaker DeepAR
- C. SageMaker Random Cut Forest
- D. SageMaker BlazingText

Answer: **B**

Explanation:

- **DeepAR** is specifically designed for time-series forecasting with covariates, making it ideal for forecasting sales based on historical data and additional features.
- XGBoost (Option A) is not optimized for sequential data. Random Cut Forest (Option C) is for anomaly detection, and BlazingText (Option D) is for text-based tasks.

Question 50

Question:
A company is building a fraud detection model. The training dataset contains labeled examples of fraudulent and non-fraudulent transactions. The dataset is highly imbalanced, with only 1% of transactions being fraudulent. The company wants the model to prioritize identifying fraudulent transactions, even if some non-fraudulent transactions are flagged incorrectly.
Which evaluation metric should the company optimize?

Options:

- A. Precision
- B. Recall
- C. Accuracy
- D. F1 Score

Answer: **B**

Explanation:

- **Recall** is the most appropriate metric for prioritizing the identification of fraudulent transactions. It measures the proportion of true positives correctly identified.
- Precision (Option A) focuses on minimizing false positives, which is not the goal. Accuracy (Option C) is not useful for imbalanced datasets. F1 Score (Option D) balances precision and recall but does not specifically prioritize one over the other.

Question 51

Question:
A logistics company is analyzing GPS data from its delivery trucks to predict late deliveries. The data contains routes, timestamps, and historical delivery status. The model needs to handle sequential dependencies in the data.
Which algorithm should the company use?

Options:

- A. SageMaker XGBoost

- B. SageMaker Linear Learner
- C. SageMaker BlazingText
- D. SageMaker Seq2Seq

Answer: **D**

Explanation:

- **Seq2Seq (Sequence-to-Sequence)** is designed to handle sequential dependencies, making it suitable for time-series or sequential data like GPS and delivery timestamps.
- XGBoost (Option A) and Linear Learner (Option B) do not consider sequence relationships, while BlazingText (Option C) is for NLP tasks.

Question 52

Question:
A company is developing an ML model to classify whether customers will purchase a product. The company wants to integrate this model into its customer-facing application. Predictions must occur in real time with minimal latency.
Which SageMaker deployment option should the company choose?

Options:

- A. Batch Transform
- B. Asynchronous Inference
- C. Real-Time Inference
- D. Multi-Model Endpoints

Answer: **C**

Explanation:

- **Real-Time Inference** ensures minimal latency for predictions, making it the best option for integrating with customer-facing applications.
- Batch Transform (Option A) and Asynchronous Inference (Option B) are for non-real-time tasks. Multi-Model Endpoints (Option D) are suitable for hosting multiple models but not specifically for low-latency requirements.

Question 53

Question:
A healthcare company is building a model to predict whether patients are likely to respond to a

treatment. The dataset is small and includes many features, some of which are likely irrelevant.
What preprocessing step should the company take to improve the model?

Options:

- A. Apply Principal Component Analysis (PCA).
- B. Apply one-hot encoding to all features.
- C. Perform feature selection based on correlation.
- D. Normalize all features to the same scale.

Answer: **C**

Explanation:

- **Feature selection** removes irrelevant features, improving the model's ability to generalize. Correlation-based selection is effective when the dataset is small and likely includes redundant features.
- PCA (Option A) reduces dimensionality but may discard important information. One-hot encoding (Option B) is unnecessary for non-categorical features.

Question 54

Question:
A company wants to monitor and detect anomalies in its manufacturing equipment. The data consists of numerical sensor readings from thousands of machines. The company plans to use an unsupervised learning approach.
Which SageMaker algorithm should the company use?

Options:

- A. Random Cut Forest
- B. XGBoost
- C. BlazingText
- D. Neural Topic Model

Answer: **A**

Explanation:

- **Random Cut Forest** is ideal for anomaly detection in numerical datasets and does not require labeled data.
- XGBoost (Option B) is for supervised learning, BlazingText (Option C) is for NLP, and Neural Topic Model (Option D) is for topic discovery.

Question 55

Question:
A company wants to build a recommendation system for its online store. The system should recommend products to users based on their purchase history and browsing behavior.
Which SageMaker algorithm is best suited for this task?

Options:

- A. XGBoost
- B. Factorization Machines
- C. Random Cut Forest
- D. Neural Topic Model

Answer: **B**

Explanation:

- **Factorization Machines** are designed for recommendation systems, as they capture interactions between features such as user preferences and item characteristics.
- XGBoost (Option A) is not specialized for recommendation systems.

Question 56

Question:
A data scientist needs to process 500,000 customer reviews and identify key topics discussed. The company wants to use SageMaker for training.
Which algorithm should the data scientist choose?

Options:

- A. BlazingText
- B. Neural Topic Model (NTM)
- C. XGBoost
- D. DeepAR

Answer: **B**

Explanation:

- **Neural Topic Model (NTM)** is designed for unsupervised topic modeling, making it ideal for extracting key topics from large text datasets.

- BlazingText (Option A) is for NLP tasks like classification.

Question 57

Question:
A retail company wants to use machine learning to identify anomalies in its sales data. The dataset contains thousands of daily sales records, and the company does not have labeled data.
Which SageMaker algorithm should the company use?

Options:

- A. BlazingText
- B. DeepAR
- C. XGBoost
- D. Random Cut Forest

Answer: **D**

Explanation:

- **Random Cut Forest** is specifically designed for anomaly detection in numerical data without requiring labels.
- DeepAR (Option B) is for time-series forecasting, and BlazingText (Option A) is for NLP tasks.

Question 58

Question:
A company is deploying a SageMaker endpoint with a trained model. The endpoint is experiencing high latency during peak hours. The company needs to reduce latency while minimizing costs.
Which solution should the company implement?

Options:

- A. Enable auto-scaling for the endpoint.
- B. Deploy the model on Spot Instances.
- C. Use multi-model endpoints.
- D. Use SageMaker asynchronous inference.

Answer: **A**

Explanation:

- **Auto-scaling** adjusts the number of instances dynamically based on traffic, ensuring low latency during peak hours while minimizing costs during off-peak times.
- Spot Instances (Option B) are cost-effective but not suitable for latency-sensitive tasks.

Question 59

Question:
A data scientist is building an ML model using a highly imbalanced dataset. The scientist wants to ensure that the model achieves a balance between precision and recall.
Which evaluation metric should the scientist optimize?

Options:

- A. Accuracy
- B. Precision
- C. Recall
- D. F1 Score

Answer: **D**

Explanation:

- **F1 Score** balances precision and recall, making it suitable for imbalanced datasets where both metrics are important.
- Accuracy (Option A) is not meaningful for imbalanced datasets.

Question 60

Question:
A manufacturing company wants to predict equipment failure using telemetry data. The company needs to forecast future equipment states and requires the ability to incorporate additional covariates like weather and workload.
Which SageMaker algorithm is best suited for this task?

Options:

- A. DeepAR
- B. XGBoost
- C. Neural Topic Model
- D. BlazingText

Answer: **A**

Explanation:

- **DeepAR** is designed for time-series forecasting and supports covariates, making it ideal for predicting equipment failure based on telemetry data.
- XGBoost (Option B) and other options are not optimized for sequential forecasting.

Question 61

Question:
A financial institution wants to identify anomalous transactions in real-time. The dataset includes numerical transaction features and is unlabelled. The company requires a low-latency solution that can integrate directly with its transaction processing system.
Which SageMaker algorithm should the company use?

Options:

- A. XGBoost
- B. Neural Topic Model
- C. Random Cut Forest
- D. BlazingText

Answer: **C**

Explanation:

- **Random Cut Forest** is ideal for real-time anomaly detection in numerical datasets without labels. It integrates well with transaction processing systems.
- XGBoost (Option A) is for supervised learning, and Neural Topic Model (Option B) and BlazingText (Option D) are not suitable for numerical anomaly detection.

Question 62

Question:
A retail company wants to recommend products to customers based on their browsing and purchasing behavior. The company's dataset includes information about user-product interactions. The model should consider both user preferences and product attributes.
Which SageMaker algorithm is the **BEST** choice for this task?

Options:

- A. Factorization Machines
- B. XGBoost
- C. Neural Topic Model
- D. DeepAR

Answer: **A**

Explanation:

- **Factorization Machines** are specifically designed for recommendation systems, as they model interactions between users and items efficiently.
- XGBoost (Option B) is not optimized for collaborative filtering.

Question 63

Question:
A company is training a machine learning (ML) model on a large dataset stored in Amazon S3. The training job is interrupted when using Spot Instances, resulting in the loss of progress. The company wants to ensure that training jobs can resume from the last checkpoint without restarting.
What should the company do?

Options:

- A. Use SageMaker training jobs with checkpointing enabled. Store checkpoint files in Amazon S3.
- B. Use SageMaker Batch Transform with checkpointing enabled. Store checkpoint files in Amazon DynamoDB.
- C. Use SageMaker real-time inference and configure auto-scaling to handle interruptions.
- D. Use AWS Glue ETL jobs to preprocess data and store intermediate results.

Answer: **A**

Explanation:

- **SageMaker training jobs with checkpointing** allow the training process to resume from the last saved state. Checkpoint files are stored in S3.
- Batch Transform (Option B) is for inference, not training.

Question 64

Question:
A machine learning (ML) specialist is training a binary classification model using SageMaker. The dataset contains an imbalanced class distribution, where only 5% of samples are positive. The ML specialist wants the model to focus on the positive class.
What should the ML specialist do?

Options:

- A. Use class weighting during training.
- B. Use the Neural Topic Model (NTM) algorithm.
- C. Oversample the positive class using SMOTE.
- D. Combine class weighting and oversampling techniques.

Answer: **D**

Explanation:

- **Combining class weighting and SMOTE** addresses class imbalance effectively. Class weighting prioritizes the positive class, while SMOTE increases positive samples.

Question 65

Question:
A company is training an image classification model using Amazon SageMaker. The training data consists of 100,000 images stored in Amazon S3. The company wants to speed up data ingestion during training.
What is the **BEST** solution?

Options:

- A. Use SageMaker training in Pipe mode.
- B. Use SageMaker training in File mode.
- C. Use SageMaker Processing to preprocess the data.
- D. Use AWS Glue to convert images to Apache Parquet format.

Answer: **A**

Explanation:

- **Pipe mode** streams data directly to the training job, reducing data loading time compared to File mode.
- File mode (Option B) downloads data before training, increasing time.

Question 66

Question:
A healthcare company needs to automate the task of summarizing patient notes. The company has labeled datasets of patient notes and summaries. Which SageMaker algorithm is the **MOST** suitable for this task?

Options:

- A. Seq2Seq
- B. BlazingText
- C. XGBoost
- D. Random Cut Forest

Answer: **A**

Explanation:

- **Seq2Seq** is designed for sequence transformation tasks, such as text summarization.
- BlazingText (Option B) and XGBoost (Option C) are not suitable for summarization.

Question 67

Question:
A company is building a predictive maintenance system using telemetry data from sensors. The company wants to use a time-series forecasting algorithm that supports multiple related time-series.
Which SageMaker algorithm is **BEST** for this use case?

Options:

- A. DeepAR
- B. Neural Topic Model
- C. Random Cut Forest
- D. BlazingText

Answer: **A**

Explanation:

- **DeepAR** supports multiple related time-series and is ideal for predictive maintenance tasks.
- Other algorithms are not designed for forecasting.

Question 68

Question:
A company is training an ML model on SageMaker to classify user-generated content as safe or inappropriate. The company wants to reduce false negatives to ensure no inappropriate content is missed.
Which evaluation metric should the company prioritize?

Options:

- A. Precision
- B. Recall
- C. F1 Score
- D. Accuracy

Answer: **B**

Explanation:

- **Recall** focuses on minimizing false negatives, ensuring inappropriate content is not missed.
- Precision (Option A) focuses on minimizing false positives.

Question 69

Question:
A company collects telemetry data from IoT devices and stores it in Amazon S3. The company wants to build a pipeline to preprocess the data and train a SageMaker model automatically.
Which AWS service should the company use to orchestrate the pipeline?

Options:

- A. AWS Glue
- B. SageMaker Pipelines
- C. AWS Step Functions
- D. Amazon Kinesis Data Analytics

Answer: **B**

Explanation:

- **SageMaker Pipelines** is designed for ML workflows, including preprocessing and training.
- Step Functions (Option C) are more general-purpose but less tailored for ML tasks.

Question 70

Question:
A company wants to detect fraud in credit card transactions using a SageMaker model. The dataset contains numerical features and has a highly imbalanced distribution of fraudulent vs. non-fraudulent transactions.
Which techniques should the company use? (Select **TWO**.)

Options:

- A. Use SMOTE to oversample the minority class.
- B. Use class weighting during training.
- C. Use Principal Component Analysis (PCA).
- D. Use anomaly detection instead of classification.
- E. Use the SageMaker k-means algorithm.

Answer: **A, B**

Explanation:

- **SMOTE** increases positive samples, and **class weighting** ensures the model focuses on the minority class.
- PCA (Option C) is for dimensionality reduction.
