Question: 41
**Main Topic :** Machine Learning
**Sub Topic :** **Perform feature engineering**

**Domain:** Exploratory Data Analysis

**Question text**:
You work for a language translation software company. Your company needs to move from traditional translation software to a machine learning model based approach that produces the translations accurately. One of your first tasks is to take text given in the form of a document and use a histogram to measure the occurrence of individual words in the document for use in document classification.

Which of the following text feature engineering techniques is the best solution for this task?

A) Orthogonal Sparse Bigram (OSB)
B) Term Frequency-Inverse Document Frequency (tf-idf)
C) Bag-of-Words
D) N-Gram

**Answer:** C

**Explanation:**
Option A is incorrect. The Orthogonal Sparse Bigram natural language processing algorithm creates groups of words and outputs the pairs of words that includes the first word. You are trying to measure the occurrence of individual words.

Option B is incorrect. Term Frequency-Inverse Document Frequency determines how important a word is in a document by giving weights to words that are common and less common in the document. You are not trying to determine the importance of the words in your document, just the count of the individual words.

Option C is correct. The Bag-of-Words natural language processing algorithm creates tokens of the input document text and outputs a statistical depiction of the text. The statistical depiction, such as a histogram, shows the count of each word in the document.

Option D is incorrect. The N-Gram natural language processing algorithm is used to find multi-word phrases in the text of a document. You are not trying to find multi-word phrases, you are just trying to find the count of the individual words.

**Reference:**
Please see the article titled Introduction to Natural Language Processing for Text

Question: 42
**Main Topic :** Machine Learning
**Sub Topic :** **Perform feature engineering**

**Domain:** Exploratory Data Analysis

**Question text**:
You work for a marketing firm that wants to analyze Twitter user stream data to find popular subjects among users who buy products produced by the firm's clients. You need to analyze the streamed text to find important or relevant repeated common words and phrases and correlate this data to client products. You'll then include these topics in your client product marketing material.

Which of the following text feature engineering techniques is the best solution for this task?

A) Orthogonal Sparse Bigram (OSB)
B) Term Frequency-Inverse Document Frequency (tf-idf)
C) Bag-of-Words
D) N-Gram

**Answer:** C

**Explanation:**
Option A is incorrect. The Orthogonal Sparse Bigram natural language processing algorithm creates groups of words and outputs the pairs of words that includes the first word. You are trying to determine how important a word is in a document by finding relevant repeated common words.

Option B is correct. Term Frequency-Inverse Document Frequency determines how important a word is in a document by giving weights to words that are common and less common in the document. You can use this information to select the most important repeated phrases in the user's tweets in your client marketing material.

Option C is incorrect. The Bag-of-Words natural language processing algorithm creates tokens of the input document text and outputs a statistical depiction of the text. The statistical depiction, such as a histogram, shows the count of each word in the document. You are looking for relevant common repeated phrases, not individual words.

Option D is incorrect. The N-Gram natural language processing algorithm is used to find multi-word phrases in the text of a document. However, it does not weight common words or phrases. You need the weighting aspect of the tf-idf algorithm to find the relevant, important repeated phrases used in the tweets.

**Reference:**
Please see the article titled Introduction to Natural Language Processing for Text

Question: 43
**Main Topic :** Machine Learning
**Sub Topic :** **Select the appropriate model(s) for a given machine learning problem**

**Domain:** Modeling

**Question text**:
You work as a machine learning specialist for major phone and internet provider. Your customer support department needs to upgrade their phone response systems to reduce the number of human service representatives needed to handle their dramatically increasing call volume. You senior management team has leveraged off-shore call center services to reduce costs, but they now want to take advantage of voice recognition to automate many of the most frequent support call types, such as “I forgot my password”, or “my internet is down.”

Your management team has assigned you to the team that will implement the machine learning model behind the voice recognition system. Which SageMaker built-in algorithm is the best choice for this problem?

A) Sequence-to-Sequence
B) K-Means
C) Semantic Segmentation
D) Neural Topic Model (NTM)

**Answer:** B

**Explanation:**
Option A is correct. The Sequence-to-Sequence algorithm takes audio as input data and generates a sequence of tokens, such as the words in the audio. This can then be used to provide automated responses to users’ requests.

Option B is incorrect. The K-Means algorithm is used to find groups within data where the members of the group are similar to each other but different from members of other groups. This algorithm will not help you encode speech audio streams.

Option C is incorrect. The semantic segmentation algorithm is used to develop computer vision applications. You are trying to solve a speech recognition problem, so this algorithm would not work for this problem.

Option D is incorrect. The Neural Topic Model algorithm is used to group documents into topics using the statistical distribution of words within the documents. You are trying to solve a speech recognition problem, so this algorithm would not work for this problem.

**Reference:**
Please see the SageMaker developer guide titled Using Amazon SageMaker Built-in Algorithms

Question: 44
**Main Topic :** Machine Learning
**Sub Topic :** **Deploy and operationalize machine learning solutions**

**Domain:** Machine Learning Implementation and Operations

**Question text**:
You work as a machine learning specialist for an eyewear manufacturing plant where you have used XGBoost to train a model that uses assembly line image data to categorize contact lenses as malformed or correctly formed. You have engineered your data and used CSV as your Training ContentType. You are now ready to deploy your model using the Amazon SageMaker hosting service.

Assuming you used the default configuration settings, which of the following are true statements about your hosted model? (Select THREE)

A) The training instance class is GPU
B) The algorithm is not parallelizable for distributed training
C) The training data target value should be in the first column of the CSV with no header
D) The training data target value should be in the last column of the CSV with no header
E) The inference data target value should be in the first column of the CSV with no header
F) The inference CSV data has no label column
G) The training instance class is CPU

**Answers:** C, F, G

**Explanation:**
Option A is incorrect. The SageMaker XGBoost currently only supports a CPU instance type for training. (See the Amazon SageMaker developer guide titled XGBoost Algorithm particularly the EC2 Instance Recommendation for the XGBoost Algorithm section)

Option B is incorrect. The XGBoost algorithm is parallelizable and therefore can be deployed on multiple instances for distributed training. (See the Amazon SageMaker developer guide titled Common Parameters for Built-in Algorithms)

Option C is correct. From the Amazon SageMaker developer guide titled XGBoost Algorithm "For CSV training, the algorithm assumes that the target variable is in the first column and that the CSV does not have a header record"

Option D is incorrect. From the Amazon SageMaker developer guide titled Common Data Formats for Training "Amazon SageMaker requires that a CSV file doesn't have a header record and that the target variable is in the first column"

Option E is incorrect. From the Amazon SageMaker developer guide titled XGBoost Algorithm "For CSV inference, the algorithm assumes that CSV input does not have the label column"

Option F is correct. From the Amazon SageMaker developer guide titled XGBoost Algorithm "For CSV inference, the algorithm assumes that CSV input does not have the label column"

Option G is correct. The SageMaker XGBoost currently only supports a CPU instance type for training. (See the Amazon SageMaker developer guide titled XGBoost Algorithm particularly the EC2 Instance Recommendation for the XGBoost Algorithm section)

**Reference:**
Please see the Amazon SageMaker developer guide titled Deploy a Model on Amazon SageMaker Hosting Services for an overview of the deployment of a SageMaker model.

Question: 45
**Main Topic :** Machine Learning
**Sub Topic :** **Perform hyperparameter optimization**

**Domain:** Modeling

**Question text**:
You work as a machine learning specialist for the highway toll collection division of the regional state area. The toll collection division uses cameras to identify car license plates as the cars pass through the various toll gates on the state highways. You are on the team that is using SageMaker Image Classification machine learning to read and classify license plates by state and then identify the actual license plate number.

Very rarely, cars pass through the toll gates with plates from foreign countries, for example Great Britain, or Mexico. The outliers must not adversely affect your model's predictions. Which hyperparameter should you set, and to what value, to ensure your model is not adversely impacted by these outliers?

A) feature_dim set to 5
B) feature_dim set to 1

C) sample_size set to 10
D) sample_size set to 100
E) learning_rate set to 0.1
F) learning_rate set to 0.75

**Answer:** E

**Explanation:**
Option A is incorrect. The feature_dim hyperparameter is a setting on the K-Means and K-Nearest Neighbors algorithms, not the Image Classification algorithm.

Option B is incorrect. The feature_dim hyperparameter is a setting on the K-Means and K-Nearest Neighbors algorithms, not the Image Classification algorithm.

Option C is incorrect. The sample_size hyperparameter is a setting on the K-Nearest Neighbors algorithm, not the Image Classification algorithm.

Option D is incorrect. The sample_size hyperparameter is a setting on the K-Nearest Neighbors algorithm, not the Image Classification algorithm.

Option E is correct. The learning_rate hyperparameter governs how quickly the model adapts to new or changing data. Valid values range from 0.0 to 1.0. Setting this hyperparameter to a low value, such as 0.1, will make the model learn more slowly and be less sensitive to outliers. This is what you want, you want your model to not be adversely impacted by outlier data.

Option F is incorrect. The learning_rate hyperparameter governs how quickly the model adapts to new or changing data. Valid values range from 0.0 to 1.0. Setting this hyperparameter to a high value, such as 0.75, will make the model learn more quickly but be sensitive to outliers. This is not what you want, you want your model to not be adversely impacted by outlier data.

**Reference:**
Please see the Amazon SageMaker developer guide titled Image Classification Hyperparameters, and the Amazon SageMaker developer guide titled Use Amazon SageMaker Built-in Algorithms

Question: 46
**Main Topic :** Machine Learning
**Sub Topic :** **Deploy and operationalize machine learning solutions**

**Domain:** Machine Learning Implementation and Operations

**Question text**:
You work as a machine learning specialist for a major oil refinery company. Your company needs to do complex analysis on its crude and oil chemical compound structures. You have selected an algorithm for your machine learning model that is not one of the SageMaker built-in algorithms. You have created your model using CreateModel and you have created your HTTPS endpoint. Your docker container running your model is now ready to receive inference requests for real-time inferences. When SageMaker returns the inference result from a client's request which of the following are true? (Select TWO)

A) To receive inference requests your inference container must have a web server running on port 8080
B) Your inference container must accept `GET` requests to the `/invocations` endpoint
C) Your inference container must accept `PUT` requests to the `/inferences` endpoint
D) Amazon SageMaker strips all `POST` headers except those supported by `InvokeEndpoint`. Amazon SageMaker might add additional headers. Your inference container must be able to safely ignore these additional headers
E) Your inference container must accept `POST` requests to the `/inferences` endpoint
F) Your inference container must accept `POST` requests to the `/invocations` endpoint

**Answers:** A, D, F

**Explanation:**
Option A is correct. From the Amazon SageMaker developer guide titled Use You Own Inference Code with Hosting Services "To receive inference requests, the container must have a web server listening on port 8080"

Option B is incorrect. The inference container must accept `POST` requests to the `/invocations` endpoint. (See the Amazon SageMaker developer guide titled Use You Own Inference Code with Hosting Services)

Option C is incorrect. The inference container must accept `POST` requests to the `/invocations` endpoint. (See the Amazon SageMaker developer guide titled Use You Own Inference Code with Hosting Services)Maker developer guide titled Deploy a Model on Amazon SageMaker Hosting Services)

Option D is correct. From the Amazon SageMaker developer guide titled Use You Own Inference Code with Hosting Services "Amazon SageMaker strips all POST headers except those supported by InvokeEndpoint. Amazon SageMaker might add additional headers. Inference containers must be able to safely ignore these additional headers."

Option E is incorrect. The inference container must accept `POST` requests to the `/invocations` endpoint. (See the Amazon SageMaker developer guide titled Use You Own Inference Code with Hosting Services)

Option F is correct. The inference container must accept `POST` requests to the `/invocations` endpoint. (See the Amazon SageMaker developer guide titled Use You Own Inference Code with Hosting Services)

**Reference:**
Please see the Amazon SageMaker developer guide titled Deploy a Model

Question: 47
**Main Topic :** Machine Learning
**Sub Topic :** **Frame business problems as machine learning problems**

**Domain:** Modeling

**Question text**:
You work as a machine learning specialist for a personal care product manufacturer. You are creating a binary classification model that you want to use to predict whether a customer is likely to positively respond to toothbrush and toothpaste sample mailed to their house. Since your company incurs expenses for the products and the shipping when sending samples, you only want to send your samples to customers who you believe have a high probability of buying your products. When analyzing if a customer will follow up with a purchase, which outcome will you want to minimize in your confusion matrix to save costs?

A) True Negative
B) False Negative
C) False Affirmative
D) True Positive
E) False Positive

**Answer:** E

**Explanation:**
Option A is incorrect. True Negatives are definitely not an outcome you want to minimize because you definitely don't want to send samples to customers who will not respond.

Option B is incorrect. You don't need to limit False Negatives as much as false positives, since False Negatives only omit customers with a higher probability of following up. Not sending a sample to these customers won't save costs.

Option C is incorrect. The terms used in a confusion matrix are: True Positive, False Negative, True Negative, and False Positive.

Option D is incorrect. True Positives are the ones to which you want to send your samples.

Option E is correct. You use a confusion matrix, or table, to describe the performance of a classification model on a set of test data when you know the true values. It's called a confusion matrix because it shows when one class is mislabeled (or confused) as another. For example, when the observation is negative but the model prediction is positive (a False Positive). To reduce the number of mailings to customers who probably won't follow up with a purchase, you want to limit False Positives.

**Reference:**
Please see the Wikipedia article titled Confusion Matrix

Question: 48
**Main Topic :** Machine Learning
**Sub Topic :** **Train machine learning models**

**Domain:** Modeling

**Question text**:
You work as a machine learning specialist for a clothing manufacturer. You have built a linear regression model using SageMaker's built-in linear learner algorithm to predict sales for a given year. Your training dataset observations are based on a number of features such as marketing dollars spent, number of active stores, traffic per store, online traffic to the company website, overall market indicators, etc. You have decided to use the k-fold method of cross-validation to assess how the results of your model will generalize beyond your training data.

Which of these will indicate that you don't have biased training data?

A) The variance of the estimate increases as you increase k
B) You shouldn't have to worry about bias because your error function removes bias in the data
C) Every k-fold cross-validation round increases the training error rate
D) Every k-fold cross-validation round has a very similar error rate to the rate of all the other rounds
E) You would not normally use k-fold with linear regression models

**Answer:** D

**Explanation:**
Option A is incorrect. When using k-fold for cross-validation the variance of the estimate is reduced as you increase k. So a 10-fold cross-validation should have lower variance than a 5-fold cross-validation.

Option B is incorrect. The k-fold error function just gives you the error rate of the cross-validation round, it doesn't resolve bias.

Option C is incorrect. The goal of k-fold cross validation is to produce relatively equal error rates for each round (indicating proper randomization of the data) not to reduce the error rate for each round.

Option D is correct. If you have relatively equal error rates for all k-fold rounds it is an indication that you have properly randomized your test data, therefore reducing the chance of bias.

Option E is incorrect. The k-fold cross-validation technique is commonly used with linear regression analysis.

**Reference:**
Please see the Amazon Machine Learning developer guide titled Evaluating ML Models, and the Amazon Machine Learning developer guide titled Cross Validation

Question: 49
**Main Topic :** Machine Learning
**Sub Topic :** **Select the appropriate model for a given machine learning problem**

**Domain:** Modeling

**Question text**:
You work as a machine learning specialist for the National Oceanic and Atmospheric Administration (NOAA Research). NOAA has developed a great white shark detection program to help warn shore populations when the sharks are in the area of a populated beach. You have the assignment to use your machine learning expertise to decide where to place 10 high tech shark detection sensors on the oceanic floor as part of a pilot to determine if the NOAA should invest broadly in these sensors, which are very expensive. You have great white sightings data from around the globe gathered over the past several years to use as your model training and test data. The model dataset contains several useful features such as the longitude and latitude of each sighting.

You have decided to use an unsupervised learning algorithm that attempts to find discrete groupings within the data. Specifically, you want to find similarities in the longitude and latitude and find groupings of these. You need to produce 10 longitude and latitude pairs to determine where to place the sensors.

Which algorithm can you use in SageMaker that best suits this task?

A) Linear Learner

B) Neural Topic Model
C) K-Means
D) Random Cut Forest
E) Semantic Segmentation
F) XGBoost

**Answer:** C

**Explanation:**
Option A is incorrect. From the Amazon SageMaker developer guide titled Linear Learner Algorithm "*Linear models* are supervised learning algorithms used for solving either classification or regression problems." But you are trying to solve a data clustering problem so you can find the ten best clustered sightings in order to determine where to place your shark detection sensors.

Option B is incorrect. From the Amazon SageMaker developer guide titled Neural Topic Model (NTM) Algorithm "Amazon SageMaker 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." So this algorithm is used for natural language processing, not data clustering.

Option C is correct. The k-means algorithm is a clustering algorithm. From the Amazon SageMaker developer guide titled K-Means Algorithm "K-means is an unsupervised learning algorithm. It attempts to find discrete groupings within data, where members of a group are as similar as possible to one another and as different as possible from members of other groups." By setting the k hyperparameter to 10, this algorithm will allow you to find the 10 best groupings of shark sightings around the world.

Option D is incorrect. From the Amazon SageMaker developer guide titled Random Cut Forest (RCF) Algorithm "Amazon SageMaker Random Cut Forest (RCF) is an unsupervised algorithm for detecting anomalous data points within a data set." But you are trying to solve a data clustering problem so you can find the ten best clustered sightings in order to determine where to place your shark detection sensors.

Option E is incorrect. From the Amazon SageMaker developer guide titled Semantic Segmentation Algorithm "The Amazon SageMaker semantic segmentation algorithm provides a fine-grained, pixel-level approach to developing computer vision applications." So the Semantic Segmentation algorithm is used for computer vision applications, but you are trying to solve a data clustering problem.

Option F is incorrect. The XGBoost algorithm is a gradient boosting algorithm. From the Amazon SageMaker developer guide titled XGBoost Algorithm "gradient boosting is a supervised learning algorithm that attempts to accurately predict a target variable by combining

an ensemble of estimates from a set of simpler, weaker models." You are not trying to predict a target value, you are trying to find discrete groupings in your dataset.

**Reference:**
Please see the Amazon SageMaker developer guide titled Use Amazon SageMaker Built-in Algorithms

Question: 50
**Main Topic :** Machine Learning
**Sub Topic : Select the appropriate model for a given machine learning problem**

**Domain:** Modeling

**Question text**:
You work as a machine learning specialist for a sports analytics company. Your company has been contracted by the Major League Baseball Association to perform real-time analytics on baseball statistics as baseball plays unfold live on national television. Your first assignment is to predict the outcome of situational set plays (such as stolen bases or pitch results) as they are about to unfold. Therefore, your model must deliver its predictions in close to real-time.

You have decided to use a SageMaker built-in algorithm. You have looked at classical forecasting methods like autoregressive integrated moving average (ARIMA) and exponential smoothing (ETS) which use one model for each time series in your data. However, you have many time series over which to train.

Based on your performance requirements and your training requirements, which SageMaker built-in algorithm should you use?

A) Linear Learner
B) Neural Topic Model
C) K-Means
D) Random Cut Forest
E) DeepAR Forecasting
F) XGBoost

**Answer:** E

**Explanation:**
Option A is incorrect. From the Amazon SageMaker developer guide titled Linear Learner Algorithm "*Linear models* are supervised learning algorithms used for solving either classification or regression problems." But you are trying to solve a one-dimensional time series problem so you can extrapolate baseball play time series into the future.

Option B is incorrect. From the Amazon SageMaker developer guide titled Neural Topic Model (NTM) Algorithm "Amazon SageMaker 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." So this algorithm is used for natural language processing, not time series problems.

Option C is incorrect. The k-means algorithm is a clustering algorithm. From the Amazon SageMaker developer guide titled K-Means Algorithm "K-means is an unsupervised learning algorithm. It attempts to find discrete groupings within data, where members of a group are as similar as possible to one another and as different as possible from members of other groups." You are trying to solve a one-dimensional time series problem so you can extrapolate play time series into the future, not a data clustering problem.

Option D is incorrect. From the Amazon SageMaker developer guide titled Random Cut Forest (RCF) Algorithm "Amazon SageMaker Random Cut Forest (RCF) is an unsupervised algorithm for detecting anomalous data points within a data set." But you are trying to solve a one-dimensional time series problem so you can extrapolate baseball play time series into the future.

Option E is correct. From the Amazon SageMaker developer guide titled DeepAR Forecasting Algorithm "... you have many similar time series across a set of cross-sectional units. For example, you might have time series groupings for demand for different products, server loads, and requests for webpages. For this type of application, you can benefit from training a single model jointly over all of the time series. DeepAR takes this approach. When your dataset contains hundreds of related time series, DeepAR outperforms the standard ARIMA and ETS methods. You can also use the trained model to generate forecasts for new time series that are similar to the ones it has been trained on." Also from the same developer guide "The training input for the DeepAR algorithm is one or, preferably, more target time series that have been generated by the same process or similar processes. Based on this input dataset, the algorithm trains a model that learns an approximation of this process/processes and uses it to predict how the target time series evolves." So the DeepAR algorithm is used for one-dimensional time series problems for complex analysis like baseball play prediction.

Option F is incorrect. The XGBoost algorithm is a gradient boosting algorithm. From the Amazon SageMaker developer guide titled XGBoost Algorithm "gradient boosting is a supervised learning algorithm that attempts to accurately predict a target variable by combining an ensemble of estimates from a set of simpler, weaker models." You are not trying to predict a target value, you are trying to solve a one-dimensional time series problem.

**Reference:**
Please see the Amazon SageMaker developer guide titled Use Amazon SageMaker Built-in Algorithms, the AWS Machine Learning Blog titled Now Available in Amazon SageMaker:

DeepAR algorithm for more accurate time series forecasting, and the AWS StatCast AI page titled See how AI on AWS gives baseball fans new insights into the game

Question: 51
**Main Topic :** Machine Learning
**Sub Topic : Build machine learning solutions for performance, availability, scalability, resiliency, and fault tolerance**

**Domain:** Machine Learning Implementation and Operations

**Question text**:
You work as a machine learning specialist for a flight data company. Your company has a contract with the US National Defence to produce real-time prediction capabilities for fighter jet flight assist software. Due to the nature of the use case, the implementation of the algorithm you choose for your machine learning model must be able to perform predictions in as close to real-time as possible.

You are in the development stages and have chosen to use the DeepAR SageMaker built-in deep learning model. You are setting up your jupyter notebook instance in SageMaker. Which of the following jupyter notebook settings will allow you to test and evaluate production performance when you are building your models?

A) Notebook instance type
B) Lifecycle configuration
C) Volume size
D) Elastic inference
E) Primary container

**Answer:** E

**Explanation:**
Option A is incorrect. This is the type of EC2 instance on which your notebook will run. This won't help you understand production performance.

Option B is incorrect. The lifecycle configuration allows you to customize your notebook environment with default scripts and plugins. Default jupyter notebook scripts and plugins won't give you any insight into production performance.

Option C is incorrect. The volume size is just the size of the jupyter instance in GBs. This won't give you any insight into production performance.

Option D is correct. From the Amazon SageMaker developer guide titled Amazon SageMaker Elastic Inference (EI) "By using Amazon Elastic Inference (EI), you can speed up the throughput and decrease the latency of getting real-time inferences from your deep learning models … You can also add an EI accelerator to an Amazon SageMaker notebook instance so that you can test and evaluate inference performance when you are building your models" Therefore, while you are in the development stage using jupyter notebooks, Elastic Inference allows you to gain insight into the production performance of your model once it is deployed.

Option E is incorrect. From the Amazon SageMaker developer guide titled CreateModel "... you name the model and describe a primary container. For the primary container, you specify the docker image containing inference code, artifacts (from prior training), and custom environment map that the inference code uses when you deploy the model for predictions.
Use this API to create a model if you want to use Amazon SageMaker hosting services or run a batch transform job." So the primary container is a parameter used in the CreateModel request when you are creating a model in SageMaker. It is not used when setting up your jupyter notebook.

**Reference:**
Please see the Amazon SageMaker developer guide titled Amazon SageMaker Elastic Inference (EI), the AWS FAQ titled Amazon Elastic Inference FAQs, and the AWS Machine Learning blog titled Optimizing costs in Amazon Elastic Inference with TensorFlow

Question: 52
**Main Topic :** Machine Learning
**Sub Topic :** **Build machine learning solutions for performance, availability, scalability, resiliency, and fault tolerance**

**Domain:** Machine Learning Implementation and Operations

**Question text**:
You work as a machine learning specialist for a polling research company. You have national polling data for the last 10 presidential elections that you have engineered, randomized, partitioned into various training and test datasets, and stored on S3. You have selected a SageMaker built-in algorithm to use for your model. Your training datasets are very large. As you repeatedly run your training job with different large datasets you find your training is taking a very long time.

How can you improve the performance of your training runs? (Select TWO)

A) Use the protobuf recordIO format
B) Convert your data to XML and use file mode to load your data to the EBS training instance volumes

C) Use pipe mode to stream the training data directly to your EBS training instance volumes
D) Convert your data to CSV and use file mode to load your data to the EBS training instance volumes
E) Change your Elastic Inference accelerator type to a larger instance type

**Answers:** A, C

**Explanation:**
Option A is correct. The protobuf recordIO format, used for training data, is the optimal way to load data into your model for training. (See the Amazon SageMaker developer guide titled Common Data Formats for Training)

Option B is incorrect. XML is not a supported data format for training in SageMaker. (See the Amazon SageMaker developer guide titled Common Data Formats for Training)

Option C is correct. When you use the protobuf recordIO format you can also take advantage of pipe mode when training your model. Pipe mode, used together with the protobuf recordIO format, gives you the best data load performance by streaming your data directly from S3 to your EBS volumes used byu your training instance. (See the Amazon SageMaker developer guide titled Common Data Formats for Training)

Option D is incorrect. When you use the CSV format and file mode all of your data is loaded from S3 to the EBS volumes used by your training instance. This is much less efficient from a performance perspective than streaming the training data directly from S3 to your EBS volumes used by your training instance. (See the Amazon SageMaker developer guide titled Common Data Formats for Training)

Option E is incorrect. Elastic Inference is used to speed up the throughput of retrieving real-time inferences from models deployed as SageMaker hosted models. Elastic Inference accelerators accelerate your inference calls, they aren't used while training. (See the Amazon SageMaker developer guide titled Amazon SageMaker Elastic Inference (EI))

**Reference:**
Please see the Amazon SageMaker developer guide titled Common Data Formats for Built-in Algorithms and the AWS FAQ titled Amazon Elastic Inference FAQs

Question: 53
**Main Topic :** Machine Learning
**Sub Topic :** **Identify and implement a data ingestion solution**

**Domain:** Data Engineering

**Question text**:
You work for a financial services company where you have a large Hadoop cluster hosting a data lake in your on premises data center. Your department has loaded your data lake with financial services operational data from your corporate actions, order management, cash management, reconciliations, and trade management systems. Your investment management operations team now wants to use data from the data lake to build financial prediction models. You want to use data from the Hadoop cluster in your machine learning training jobs. Your Hadoop cluster has Hive, Spark, Sqoop, and Flume installed.

How can you most effectively load data from your Hadoop cluster into you SageMaker model for training?

A) Use the distcp utility to copy your dataset from your hadoop platform to the S3 bucket where your SageMaker training job can use it
B) Use the HadoopActivity command with AWS Data Pipeline to move your dataset from your hadoop platform to the S3 bucket where your SageMaker training job can use it
C) Use the SageMaker Spark library using the data frames in your Spark clusters to train your model
D) Use the Sqoop export command to export your dataset from your Hadoop cluster to the S3 bucket where your SageMaker training job can use it

**Answer:** C

**Explanation:**
Option A is incorrect. The Hadoop distcp utility is used for inter/intra cluster data movement. It is not an efficient method to get data into your SageMaker training instance. (See the Apache Hadoop distcp guide)

Option B is incorrect. The HadoopActivity command is used to run a job on a cluster. You would have to write the job to extract and load the data onto S3. This would not be the most efficient method of the options listed. (See AWS Data Pipeline developer guide titled HadoopActivity)

Option C is correct. The SageMaker Spark library that makes it so you can easily train models using data frames in your Spark clusters. This is the most efficient method of the options listed. (See the Amazon SageMaker developer guide titled Use Apache Spark with Amazon SageMaker)

Option D is incorrect. The Sqoop export command is used for exporting files from HDFS to an RDBMS. This would not help you load your data into your SageMaker training instance. (See the Sqoop User Guide)

**Reference:**
Please see the Amazon SageMaker developer guide titled Use Machine Learning Frameworks with Amazon SageMaker

Question: 54
**Main Topic :** Machine Learning
**Sub Topic :** **Sanitize and prepare data for modeling**

**Domain:** Exploratory Data Analysis

**Question text**:
You are working for a consulting firm in their machine learning practice. Your current client is a sports equipment manufacturer. You are building a linear regression model to predict ski and snowboard sales based on the daily snowfall in various regions around the country.

After you have cleaned your CSV data, which of the following tasks would you perform next?

A) Use the scikit-learn cross_validate method to evaluate the estimation precision of your model
B) Load your data into a pandas DataFrame and remove header rows and any superfluous features
C) Use one-hot encoding to convert categorical values, such as 'region of the country' to numerical values
D) Randomize your data using a shuffling technique

**Answer:** D

**Explanation:**
Option A is incorrect. The scikit-learn cross_validate method is used to evaluate your model's precision while tuning the model's hyperparameters. (See Scikit-Learn user guide titled cross_validate)

Option B is incorrect. Using a Pandas DataFrame to remove superfluous rows and features is part of cleaning you data, which you have already done.

Option C is incorrect. One-hot encoding is another way to clean your data in preparation for training. You have already completed the cleaning of your data.

Option D is correct. For a linear regression model, once you have cleaned your data you need to randomize the data to prevent overfitting and to reduce variance. (See Amazon Machine Learning developer guide titled The Amazon Machine Learning Process)

**Reference:**
Please see the Amazon Machine Learning developer guide titled Machine Learning Concepts, and the Amazon Machine Learning developer guide titled The Amazon Machine Learning Process

Question: 55
**Main Topic :** Machine Learning
**Sub Topic :** **Sanitize and prepare data for modeling**

**Domain:** Exploratory Data Analysis

**Question text**:
You work as a machine learning specialist at a retail shoe manufacturer. Your marketing department wants to do a promotion for a new running shoe they are about to release into their product pipeline. They need a model to predict sales of the new shoe using the purchase history of their registered customers based on past releases of new shoes.

You have decided to use a linear regression algorithm for your model. Your data has thousands of observations and 35 numeric features. While doing analysis to better understand your data you find 25 observations that have what looks like outlier data points. After speaking to your marketing department you learn that these values are valid. You also find several hundred observations that have some blank feature values.

How should you correct the outlier and blank feature problems?

A) Remove the observations with the outlier data points and replace the blank values with the null value
B) Remove the outlier and blank value observations
C) Remove the observations with the outlier data points and replace the blank values with the mean value
D) Remove the observations with the outlier data points and replace the blank values with the value 0

**Answer:** C

**Explanation:**
Option A is incorrect. Null values in an observation should be replaced since linear regression calculations will have a problem with null values. Therefore, you would not replace empty fields with null.

Option B is incorrect. Removing the observations with blank values will reduce the accuracy of your model's predictions since you have removed many features from the training dataset.

Option C is correct. You should remove the outlier observations. You should also replace the blank values with a meaningful value. The mean value is the best option of those listed.

Option D is incorrect. You should remove the outlier observations. You should also replace the blank values with a meaningful value. The 0 value is not the best option of those listed because the mean is invariably a better approximation than 0 for a continuous numeric value.

**Reference:**
Please see the Amazon Machine Learning developer guide titled Feature Processing

Question: 56
**Main Topic :** Machine Learning
**Sub Topic :** **Create data repositories for machine learning**

**Domain:** Data Engineering

**Question text**:
You work as a machine learning specialist at a hedge fund firm. Your firm is working on a new quant algorithm to predict when to enter and exit holdings in their portfolio. You are building a machine learning model to predict these entry and exit points in time. You have cleaned your data and you are now ready to split the data into training and test datasets.

Which splitting technique is best suited to your model's requirements?

A) Use k-fold cross validation to split the data
B) Sequentially splitting the data
C) Randomly splitting the data
D) Categorically splitting the data by holding

**Answer:** B

**Explanation:**
Option A is incorrect. Using k-fold cross validation will randomly split your data, but you need to consider the time-series nature of your data when splitting. So randomizing the data would eliminate the time element of your observations, making the datasets unusable for predicting price changes over time.

Option B is correct. By sequentially splitting the data you preserve the time element of your observations.

Option C is incorrect. Randomly splitting the data would eliminate the time element of your observations, making the datasets unusable for predicting price changes over time.

Option D is incorrect. If you split the data by a category such as the holding attribute you would create imbalanced training and test dataset since some holdings would only be in the training dataset and others would only be in the test dataset.

**Reference:**
Please see the Amazon Machine Learning developer guide titled Splitting Your Data

Question: 57
**Main Topic :** Machine Learning
**Sub Topic :** **Create data repositories for machine learning**

**Domain:** Data Engineering

**Question text**:
You work as a machine learning specialist for a software company that is developing a movie rating social media site where users can rate movies. You want to use your companies data to predict the ratings distribution of a movie based on the genre of the movie. Your training data contains a genre feature with a set of categories such as documentary, romance, etc. You have sorted your data by the genre feature and then used the Amazon ML sequential split option to split your data into training and test datasets.

When using your test dataset to verify your genre-prediction model you discover that the accuracy rate is very low. What could be the underlying problem?

A) You should have sorted by a different feature before you used the sequential split option
B) You should have split your data categorically by genre
C) You should have split your data sequentially by year
D) You should not have used the sequential split option

**Answer:** D

**Explanation:**
Option A is incorrect. Sorting the data by a different feature wouldn't solve the problem. You used the sequential option when splitting the data so you have not properly randomized your data.

Option B is incorrect. By categorically splitting the data by definition you will have some genre movies only in the training dataset and others only in the test dataset. This reduces the genre feature to a meaningless datapoint.

Option C is incorrect. Sequentially splitting the data by year wouldn't solve the problem. You used the sequential option when splitting the data so you have not properly randomized your data.

Option D is correct. You should not have used the sequential option when splitting your data. For this type of problem, in order to get proper generalization from your data, you need to randomize it.

**Reference:**
Please see the Amazon Machine Learning developer guide titled Splitting Your Data

Question: 58
**Main Topic :** Machine Learning
**Sub Topic :** **Analyze and visualize data for machine learning**

**Domain:** Exploratory Data Analysis

**Question text**:
You work as a machine learning specialist for a real estate company. You are using the kaggle housing prices data as your experimentation data to optimize your model before you use your model on the real estate data for your area of the country. You have a hypothesis that you can predict the price of a real estate property based on the foundation type. You have your data from kaggle but you want to make sure your model is not overly influenced by outliers.

What is the quickest way to identify outliers in your data?

A) Arrange your data points from lowest to highest; calculate the median of the data set; use a qualitative assessment to determine whether to remove outliers
B) Calculate the Z-Score for your data points
C) Visualize your data using scatter plots and/or box plots
D) Visualize your data using network and correlation matrices

**Answer:** C

**Explanation:**
Option A is incorrect. You can find your outliers using a quantitative assessment, but it will involve more effort and therefore more time than visualizing your data.

Option B is incorrect. The z-score of a data point shows how many standard deviations the data point is from the mean. This would help you find your outliers but it will involve more effort and therefore more time than visualizing your data.

Option C is correct. With large datasets, such as the real estate data you are using in this problem, the quickest way to find outliers is to visualize your data. The best plots for this task are the scatter plot and the box plot. (See the article titled How to Make your Machine Learning Models Robust to Outliers)

Option D is incorrect. Visualization is the quickest and easiest way to find outliers, but the network and/or correlation matrix charting choices will not show outliers. They are used to represent relations between data points as nodes. These relationships would not give you any information about the extremity of a data point.

**Reference:**
Please see the article titled How to Make your Machine Learning Models Robust to Outliers, and the article titled A Brief Overview of Outlier Detection Techniques

Question: 59
**Main Topic :** Machine Learning
**Sub Topic :** **Select the appropriate model(s) for a given machine learning problem**

**Domain:** Modeling

**Question text**:
You work as a machine learning specialist for a company that runs car rating website. Your company wants to build a price prediction model that is more accurate than their current model, which is a linear regression model using the age of the car as the single independent variable in the regression to predict the price. You have decided to add the horse power, fuel type, city mpg, drive wheels, and number of doors as independent variables in your model. You believe that adding these additional independent variables will give you a more accurate prediction of price.

Which type of algorithm will you now use for your prediction?

A) Logistic Regression
B) Decision Tree
C) Naive Bayes
D) Multivariate Regression

**Answer:** D

**Explanation:**
Option A is incorrect. Logistic regression is used for problems where you are trying to classify and estimate a discrete value (on or off, 1 or 0) based on a set of independent variables. In your problem you are trying to estimate a continuous numerical value: price, not a binary classification.

Option B is incorrect. A decision tree is a classification algorithm, so it is not a good fit for your continuous numerical value prediction problem.

Option C is incorrect. Naive Bayes is another classification algorithm, so it is not a good fit for your continuous numerical value prediction problem.

Option D is correct. You are trying to predict the price of a car (dependent variable) based on a number of independent variables (horse power, fuel type, city mpg, drive wheels, and number of doors, etc.) The Multivariate Regression algorithm is the best choice for this type of problem. (See the article Data Science Simplified Part 5: Multivariate Regression Models)

**Reference:**
Please see the Amazon Machine Learning developer guide titled Regression Model Insights, and the article titled Commonly Used Machine Learning Algorithms (with Python and R codes)

Question: 60
**Main Topic :** Machine Learning
**Sub Topic : Analyze and visualize data for machine learning**

**Domain:** Exploratory Data Analysis

**Question text**:
You work as a machine learning specialist for a company that produces a polling data and uses it for predictive modeling. Your company wants to build an election prediction model that uses multiple independent variables such as age of voter, religion, sex, registered affiliation, etc. to predict the candiate for which each observed voter will vote in the upcoming election.

Which type of algorithm is NOT a good choice to use for your prediction? (Select THREE)

A) Ordinary Least Squares Regression (OLSR)
B) Local Outlier Factor (LOF)
C) Naive Bayes
D) Least-Angle Regression (LARS)
E) K-Means

**Answers:** B, C, E

**Explanation:**
Option A is incorrect. Ordinary Least Squares Regression (OLSR) is a regression technique that predicts a dependent variable using one or more independent variables. This is exactly what you are trying to solve.

Option B is correct. The Local Outlier Factor (LOF) algorithm is used to discover outlier data points. So this would NOT be a good choice for your algorithm where you are trying to solve for a dependent variable based on multiple independent variables.

Option C is correct. The Naive Bayes algorithm is used as a classifier. So this would NOT be a good choice for your algorithm where you are trying to solve for a dependent variable based on multiple independent variables.

Option D is incorrect. Least-Angle Regression (LARS) is also a regression technique that predicts a dependent variable using one or more independent variables. This is exactly what you are trying to solve.

Option E is correct. The K-Means algorithm is used as a clustering algorithm, so it would NOT be a good choice for your algorithm where you are trying to solve for a dependent variable based on multiple independent variables.

**Reference:**
Please see the Amazon Machine Learning developer guide titled Regression Model Insights, and the article titled A Tour of the Most Popular Machine Learning Algorithms

Question: 61
**Main Topic :** Machine Learning
**Sub Topic :** **Identify and Implement a data-transformation solution**

**Domain:** Data Engineering

**Question text**:
You are a machine learning specialist for a research firm. Your team is using Amazon SageMaker and it's built-in scikit-learn library for feature transformation in your machine learning process. When using the SimpleImputer transformer to replace missing values in your observations, which strategy is the default strategy that your SageMaker scikit-learn code will use if you don't explicitly pass a strategy parameter?

A) constant
B) most_frequent
C) median
D) mean
E) mode

**Answer:** D

**Explanation:**
Option A is incorrect. The default strategy is mean. The constant strategy replaces the missing values with a constant you supply.

Option B is incorrect. The default strategy is mean. The most_frequent strategy replaces the missing values with the most frequent value along each column.

Option C is incorrect. The default strategy is mean. The median strategy replaces the missing values with the median along each column.

Option D is correct. The default strategy is mean. The mean strategy replaces the missing values with the mean along each column.

Option E is incorrect. There is no mode strategy in the SimpleImputer scikit-learn transformer.

**Reference:**
Please see the Amazon Machine Learning blog titled Preprocess input data before making predictions using Amazon SageMaker inference pipelines and Scikit-learn

Question: 62
**Main Topic :** Machine Learning
**Sub Topic : Identify and Implement a data-transformation**

**Domain:** Data Engineering

**Question text**:
You are a machine learning specialist for a gaming software startup. Your company is investigating ways to use machine learning to enhance their game software platform. The team has selected the Amazon SageMaker platform for their machine learning efforts. You are participating in the feature transformation process in preparation to creating your machine learning models. Instead of transforming your data before you use it in your SageMaker models, you and your team have decided to use the built-in transformations of SageMaker. Specifically, you and your team have decided to use the built-in OneHotEncoder transformer to transform your categorical data.

You have decided to drop one of the categories per feature because you suspect you may have perfectly collinear features. Which of the following is NOT a drop methodology used in the OnHotEncoder transformer?

A) None
B) Last
C) Array

D) First

**Answer:** B

**Explanation:**
Option A is incorrect. The OneHotEncoder transformer has the following methodologies you can use to drop one of the categories per feature: None, first, array. None is the default methodology.

Option B is correct. The OneHotEncoder transformer has the following methodologies you can use to drop one of the categories per feature: None, first, array. None is the default methodology. The OneHotEncoder transformer drop parameter does not offer a last methodology.

Option C is incorrect. The OneHotEncoder transformer has the following methodologies you can use to drop one of the categories per feature: None, first, array. None is the default methodology.

Option D is incorrect. The OneHotEncoder transformer has the following methodologies you can use to drop one of the categories per feature: None, first, array. None is the default methodology.

**Reference:**
Please see the Amazon Machine Learning blog titled Preprocess input data before making predictions using Amazon SageMaker inference pipelines and Scikit-learn, and the Scikit-learn api documentation OneHotEncoder

Question: 63
**Main Topic :** Machine Learning
**Sub Topic :** **Evaluate machine learning models**

**Domain:** Modeling

**Question text**:
You work as a machine learning specialist for a consulting firm that has the NFL as a client. You are working on the passer completion probability model using statistics from in-play metrics. You are running your linear learner model in Amazon SageMaker using a CSV file representation of your passer completion probability statistics. You are now running your inference.

Some of the features and their data types are listed below:

| Feature Name | Data Type |
|---|---|
| Passer age | Numeric |
| Length of pass | Numeric |
| Complete (yes/no) | Categorical |

| Feature Name | Data Type |
|---|---|
| Distance between receiver and nearest defender | Numeric |
| Play called (post, crossing, screen, etc.) | Categorical |

You are using the Complete feature as your prediction response feature. You are now making predictions on new data. When you interrogate the response of your model, which of the following do you expect to find?

A) score: the prediction produced by the model
B) score: the prediction produced by the model AND predicted_class which is an integer from 0 to num_classes-1
C) score: single floating point number measuring the strength of the prediction AND predicted_label which is 0 or 1
D) score: the prediction produced by the model OR predicted_label which is 0 or 1

**Answer:** C

**Explanation:**
Option A is incorrect. For a binary classification (complete yes or no) the model produces a score denoting the strength of the prediction AND a predicted_label denoting complete or not complete

Option B is incorrect. This option describes the response for a multiclass classification, but you are working with a binary classification.

Option C is correct. For a binary classification (complete yes or no) the model produces a score denoting the strength of the prediction AND a predicted_label denoting complete or not complete.

Option D is incorrect. For a binary classification (complete yes or no) the model produces a score denoting the strength of the prediction AND a predicted_label denoting complete or not complete.

**Reference:**
Please see the Amazon SageMaker developer guide titled Linear Learner Algorithm

Question: 64
**Main Topic :** Machine Learning
**Sub Topic :** **Perform hyperparameter optimization**

**Domain:** Modeling

**Question text**:
You work in the machine learning department of a major retail company. Your team is working on a model to predict the region that will have the highest sales for a given quarter. You have selected your observations from past sales cycles for all regions and split your data into training and evaluation datasets. You are now training your linear learner model in Amazon SageMaker and you are trying to select the model hyperparameters that give your team the best predictions.

You have set the predictor_type hyperparameter to binary_classifier. Which loss function hyperparameter setting is NOT one of your options?

A) auto
B) logistic
C) hinge_loss
D) softmax_loss

**Answer:** D

**Explanation:**
Option A is incorrect. The three hyperparameters values that you can set for the loss function are auto, logistic, and hinge_loss. The default for auto is logistic.

Option B is incorrect. The three hyperparameters values that you can set for the loss function are auto, logistic, and hinge_loss. The default for auto is logistic.

Option C is incorrect. The three hyperparameters values that you can set for the loss function are auto, logistic, and hinge_loss. The default for auto is logistic.

Option D is correct. The three hyperparameters values that you can set for the loss function are auto, logistic, and hinge_loss. The default for auto is logistic. The softmax_loss setting is an option if your predictor_type is set to multiclas_classifier.

**Reference:**
Please see the Amazon SageMaker developer guide titled Linear Learner Hyperparameters

Question: 65
**Main Topic :** Machine Learning
**Sub Topic :** **Perform hyperparameter optimization**

**Domain:** Modeling

**Question text**:
You work in the machine learning department of a major retail company. Your team is working on a model to classify customers by purchase history. Your marketing department wants to use the results of your model predictions to determine which customers should receive a new campaign offer. You have selected your observations and cleaned your data. You have also split your data into training and evaluation datasets. You are now training your k-means model in Amazon SageMaker and you are trying to select the model hyperparameters that give your marketing team the best predictions.

You have set the feature_dim hyperparameter to equal the number of features in your input data. You have set the k hyperparameter to 10, the number of clusters you estimate is appropriate for your model. You have set the epochs hyperparameter to 1 so that the model performs one pass over your data.

You need to report a score for your model. Which k-means hyperparameter allows you to select the metric types to report this scoring, and what are the available metric options?

A) extra_center_factor with msd, ssd, or [msd, ssd] as the available metric type values
B) score_metrics with mse, ssd, or [mse, ssd] as the available metric type values
C) eval_method with mse, ssd, or [mse, ssd] as the available metric type values
D) eval_metrics with msd, ssd, or [msd, ssd] as the available metric type values

**Answer:** D

**Explanation:**
Option A is incorrect. The hyperparameter you chose to report a score for your model is the eval_metrics hyperparameter. The eval_metrics hyperparameter has the allowed values of msd for Mean Square Error, ssd for Sum of Square Distance, and the option of both msd and ssd. The extra_center_factor is used to control the number of clusters.

Option B is incorrect. The hyperparameter you chose to report a score for your model is the eval_metrics hyperparameter. The eval_metrics hyperparameter has the allowed values of msd for Mean Square Error, ssd for Sum of Square Distance, and the option of both msd and ssd. The Amazon SageMaker k-means algorithm does not have a score_metrics hyperparameter.

Option C is incorrect. The hyperparameter you chose to report a score for your model is the eval_metrics hyperparameter. The eval_metrics hyperparameter has the allowed values of msd for Mean Square Error, ssd for Sum of Square Distance, and the option of both msd and ssd. The Amazon SageMaker k-means algorithm does not have a eval_method hyperparameter.

Option D is correct. The hyperparameter you chose to report a score for your model is the eval_metrics hyperparameter. The eval_metrics hyperparameter has the allowed values of msd for Mean Square Error, ssd for Sum of Square Distance, and the option of both msd and ssd.

**Reference:**
Please see the Amazon SageMaker developer guide titled K-Means Hyperparameters

START HERE WITH 4TH SET
Question: 66
**Main Topic :** Machine Learning
**Sub Topic : Select the appropriate model(s) for a given machine learning problem**

**Domain:** Modeling

**Question text**:
You are a machine learning specialist at a large online retailer. Your team is working on a recommender model for your online purchase workflow. The recommender will suggest similar items to the items the user has viewed or placed in their shopping cart. To find items that are similar to the item your customer is viewing, you want to compare other users who like each item. If these similar users like the same two items, then the probability the items are similar is higher.

Which Amazon SageMaker built-in algorithm is best suited to your use case?

A) Semantic Segmentation
B) K-Nearest Neighbor
C) Linear Learner
D) Random Cut Forest

**Answer:** B

**Explanation:**
Option A is incorrect. The semantic segmentation algorithm is used to develop computer vision applications. You are trying to find items that are similar to each other.

Option B is correct. The k-nearest neighbor algorithm is used to find items that are similar to each other. This is what you need to find similar items to recommend to a user in the online purchase workflow.

Option C is incorrect. The linear learner algorithm is used to show how a change in an independent variable affects a dependent variable. You are trying to find items that are similar to each other.

Option D is incorrect. The random cut forest algorithm is predominantly used to classify observations, such as whether a transaction is fraudulent or not. You are trying to find items that are similar to each other.

**Reference:**
Please see the Amazon SageMaker developer guide titled Use Amazon Sagemaker Built-in Algorithms

Question: 67
**Main Topic :** Machine Learning
**Sub Topic :** **Identify and implement a data-transformation solution**

**Domain:** Data Engineering

**Question text**:
You have just landed a position as a machine learning specialist at a large financial services firm. Your new team is working on a fraud detection model using the SageMaker built-in linear learner algorithm. You are gathering the data required for your machine learning model. The dataset you intend to produce will contain well over 5,000 objects that need to be labeled. Your team wants to control the costs of cleaning your data. Therefore, the team has decided to use SageMaker Ground Truth active learning to automate the labeling of your data.

The Ground Truth automated labeling job initially follows this set of steps:
- Selects a random sample of data
- sends the sample data to human workers
- uses the human-labeled data as validation data
- runs a SageMaker batch transform using the validation set which generates a quality metric used to estimate the potential quality of auto-labeling the rest of the unlabeled data
- runs a SageMaker batch transform on the unlabeled data
- data where the expected quality of automatically labeling the data is above the requested level of accuracy is labeled

After performing the above steps, what does Ground Truth do next to complete the labeling of ALL of your data?

A) Selects a new sample of unlabeled data and sends it to human workers; it uses the existing labeled data to verify the new human-labeled data; repeats this later set of steps until all the data in the dataset is labeled
B) Selects a new sample of unlabeled data and sends it to human workers; it uses the existing labeled data and the new human-labeled data to train a new model; repeats this later set of steps until all the data in the dataset is labeled

C) Selects a new sample of the most hard to identify unlabeled data and sends it to human workers; it uses the existing labeled data to verify the new human-labeled data; repeats this later set of steps until all the data in the dataset is labeled
D) Selects a new sample of the most hard to identify unlabeled data and sends it to human workers; it uses the existing labeled data and the new human-labeled data to train a new model; repeats this later set of steps until all the data in the dataset is labeled

**Answer:** D

**Explanation:**
Option A is incorrect. This option doesn't articulate that the selection of a new sample looks for the most hard to identify unlabeled data. It also doesn't state that the new human-labeled data is used with the existing labeled data to train a new model.

Option B is incorrect. This option doesn't articulate that the selection of a new sample looks for the most hard to identify unlabeled data.

Option C is incorrect. This option doesn't state that the new human-labeled data is used with the existing labeled data to train a new model.

Option D is correct. This is the set of steps Ground Truth uses to iterate over the unlabeled data using human labelers and model training to complete the labeling of your large dataset.

**Reference:**
Please see the Amazon SageMaker developer guide titled Amazon SageMaker Ground Truth, and the Amazon SageMaker developer guide titled Using Automated Data Labeling

Question: 68
**Main Topic :** Machine Learning
**Sub Topic :** **Perform feature engineering**

**Domain:** Exploratory Data Analysis

**Question text**:
You work for a major banking firm as a machine learning specialist. As part of the bank's fraud detection team, you are building a machine learning model to detect fraudulent transactions. Using your training dataset you have produced a Receiver Operating Characteristic (ROC) curve and it shows 99.99% accuracy. Your transaction dataset is very large, but 99.99% of the observations in your dataset represent non-fraudulent transactions. Therefore, the fraudulent observations are a minority class. Your dataset is very imbalanced.

Given you have the approval from your management team to produce the most accurate model possible, even if it means spending more time perfecting the model, what is the most effective technique to address the imbalance in your dataset?

A) Synthetic Minority Oversampling Technique (SMOTE) oversampling
B) Random oversampling
C) Generative Adversarial Networks (GANs) oversampling
D) Edited Nearest Neighbor undersampling

**Answer:** C

**Explanation:**
Option A is incorrect. The SMOTE technique creates new observations of the underrepresented class, in this case the fraudulent observations. These synthetic observations are almost identical to the original fraudulent observations. This technique is expeditious, but the types of synthetic observations it produces are not as useful as the unique observations created by other oversampling techniques.

Option B is incorrect. Random oversampling uses copies of some of the minority class observations (randomly selected) to augment the minority class observation set. These observations are exact replicas of existing minority class observations, making them less effective than observations created by other techniques that produce unique synthetic observations.

Option C is correct. The Generative Adversarial Networks (GANs) technique generates unique observations that more closely resemble the real minority observations without being so similar that they are almost identical. This results in more unique observations of your minority class that improve your model's accuracy by helping to correct the imbalance in your data.

Option D is incorrect. Using an undersampling technique would remove potentially useful majority class observations. Additionally, you would have to remove a very large number of your majority class observations to correct your imbalance that you would render your entire training dataset useless.

**Reference:**
Please see the wikipedia article titled Oversampling and undersampling in data analysis, and the article titled Imbalanced data and credit card fraud

Question: 69

**Main Topic :** Machine Learning
**Sub Topic :** **Evaluate machine learning models**

**Domain:** Modeling

**Question text**:
You work as a machine learning specialist for a car manufacturer that has developed driverless technology for their new line of cars. These cars require real-time machine learning models to perform all of the tasks of driving. You have trained multiple models, using different algorithms and/or different hyperparameters, as candidates to assist in lane line crossover detection using live data from sensors on the undercarriage of the car. You want to select one of these models as the model to go to production in the line of cars.

Using the various options available from SageMaker, which are the most effective method steps you should use to select the correct model? (Select TWO)

A) Use online testing with historical data
B) Deploy your trained models to beta endpoints, then using a jupyter notebook in your SageMaker instance, send inference requests to each model in turn using the AWS SDK for python or the SageMaker high-level python library and finally evaluate each model.
C) Use online testing with live data
D) Deploy your models to a SageMaker training instance, then train each model on a portion of the live data and finally evaluate each model
E) Deploy your models to a SageMaker endpoint, then send a portion of the live data to each model and finally evaluate each model

**Answers:** C, E

**Explanation:**
Option A is incorrect. For online testing you use live data. For offline testing you use historical data.

Option B is incorrect. When performing offline testing of your models, you deploy your trained models to alpha endpoints, not beta endpoints.

Option C is correct. For online testing you use live data. Testing with live data will allow you to perform the steps listed in option E.

Option D is incorrect. To use online testing, you deploy your models to a SageMaker endpoint, not a SageMaker training instance.

Option E is correct. To perform online testing of your models you deploy the models to a SageMaker endpoint and then send a portion of the data to each model (or production variant) allowing you to evaluate the models.

**Reference:**
Please see the SageMaker developer guide titled Validate a Machine Learning Model

Question: 70
**Main Topic :** Machine Learning
**Sub Topic : Select the appropriate model(s) for a given machine learning problem**

**Domain:** Modeling

**Question text**:
You work as a machine learning specialist for a large auto parts manufacturing company. You have been tasked with building a machine learning model to analyze images of car parts on your company's production lines to automatically classify the parts. The classified parts will then be placed in their appropriate warehouse containers by classification.

Some examples of the classifications are: electronics, trim, gasket, hose, etc. Since your company has many manufacturing plants across the globe, your classification model needs to be able to classify millions of high resolution images.

Which algorithm best fits your problem?

A) Object Detection
B) Convolutional Neural Network
C) Latent Dirichlet Allocation (LDA)
D) Factorization Machine

**Answer:** B

**Explanation:**
Option A is incorrect. The Object Detection algorithm is used to identify all instances of an object within an image. While this may be used in a naive approach to the image classification problem, it is not meant for image classification in the way and scale needed for your problem.

Option B is correct. The SageMaker built-in Image Classification algorithm uses a Convolutional Neural Network to classify images that supports multi-label classification. It scales to millions of images at high resolution. It solves this problem through convolution and multiple layers in the neural network. (See the article AWS SageMaker and CNN for Dog Breed Classification)

Option C is incorrect. The Latent Dirichlet Allocation algorithm is used for topic discovery within documents.

Option D is incorrect. The Factorization Machine algorithm can be used to classify observations, but it is used primarily to detect interactions between features. Examples include reaction to ads on a web page, or item recommendation.

**Reference:**
Please see the SageMaker developer guide titled Using Amazon SageMaker Built-in Algorithms

Question: 71
**Main Topic :** Machine Learning
**Sub Topic :** **Select the appropriate model(s) for a given machine learning problem**

**Domain:** Modeling

**Question text**:
You work as a machine learning specialist for medical research facility. Your research team is working on a brain tumor detection scanner to be used in hospitals across the country. The team has decided to use machine learning to detect tumors in the scans and to catalog the findings in a database that can be shared across medical facilities.

You have millions of brain scan data to use in your model. Also, you will have an incoming stream of new scans every day, so your volume is very high. Your research team requires that the model perform at scale and with very high accuracy due to the nature of the consequences of false negative predictions.

Which algorithm best fits your problem?

A) Object Detection
B) K-Means
C) Convolutional Neural Network
D) Random Cut Forest

**Answer:** C

**Explanation:**
Option A is incorrect. The Object Detection algorithm is used to identify all instances of an object within an image. You are trying to classify a high resolution image as either containing a tumor or not. You are not trying to identify, and surrounding with a bounding box, all elements in an image.

Option B is incorrect. The K-Means algorithm is used to find groups within data where the members of the group are similar. This would not work for our image classification problem.

Option C is correct. The SageMaker built-in Image Classification algorithm uses a Convolutional Neural Network to classify images. It breaks up each image into a series of tiles and then predicts what each tile contains. This is the optimal way to find a tumor within a larger brain scan image. (See the article Image Classification using Deep Neural Networks - A beginner friendly approach using TensorFlow)

Option D is incorrect. The Random Cut Forest algorithm is used to find abnormal data points with your dataset. It would not be the best choice for your image classification problem with large numbers of high resolution images in which you are trying to detect an anomaly.

**Reference:**
Please see the SageMaker developer guide titled Using Amazon SageMaker Built-in Algorithms, and the article titled How might companies use random forest models for predictions?

Question: 72
**Main Topic :** Machine Learning
**Sub Topic :** **Recommend and implement the appropriate machine learning services and features for a given problem**

**Domain:** ML Implementation and Operations

**Question text**:
You work as a machine learning specialist for an online retail company that sells health products. Your company allows users to enter reviews of the products they buy from the website. You want to make sure the reviews do not contain any offensive or unsafe content, such as obscenities or threatening language.

Which Amazon SageMaker algorithm or service will allow you to scan your user's review text in the simplest way?

A) BlazingText
B) Neural Topic Model (NTM)
C) Semantic Segmentation
D) Comprehend

**Answer:** D

**Explanation:**
Option A is incorrect. The BlazingText algorithm is used for natural language processing tasks like sentiment analysis, and named entity recognition. You should use all of these features when scanning your user's review text, however the BlazingText algorithm requires more developer effort and time than using the Comprehend service.

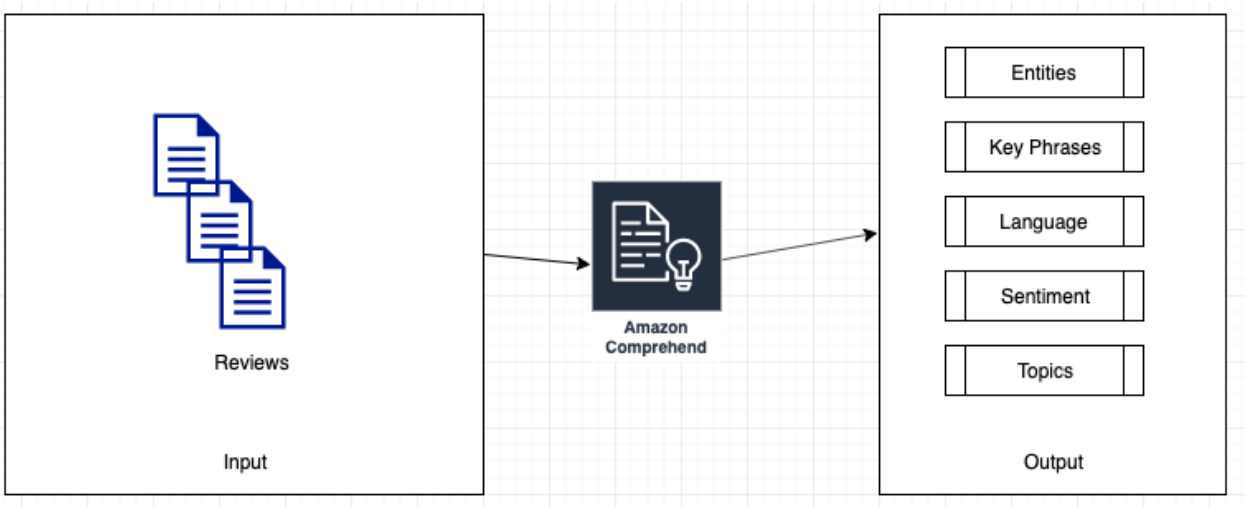
Option B is incorrect. The Neural Topic Model algorithm is used to group documents into topics using the statistical distribution of words within the documents. This algorithm would not be the most efficient choice for detecting offensive or unsafe language.

Option C is incorrect. The Semantic Segmentation algorithm is used for computer vision application, so it is not an algorithm you would use for text analysis.

Option D is correct. The Comprehend service scans your unstructured review text and analyzes it using SageMaker Natural Language Processing (NLP) algorithms to find key phrases, entities, and sentiments. This is the most expeditious and efficient option.

**Reference:**
Please see the Amazon SageMaker developer guide titled Using Amazon SageMaker Built-in Algorithms, and the Amazon Machine Learning blog titled Analyze content with Amazon Comprehend and Amazon SageMaker notebooks

Here is a diagram of the solution:




Question: 73
**Main Topic :** Machine Learning

**Sub Topic :** **Recommend and implement the appropriate machine learning services and features for a given problem**

**Domain:** ML Implementation and Operations

**Question text**:
You work as a machine learning specialist for a news organization that has a very active online community who contributes comments on your organization's news articles very frequently. Your news editors wish to use the comments from their users to gain insight into what interests them the most. Instead of just relying on the raw count of comments per article, the editors would like to use machine learning to find the underlying intent of the comments. This will allow them to understand their readers better so that they can provide more tailored articles for the most popular subjects.

You have decided to use Amazon Comprehend as your machine learning platform for this task. Which of the listed Comprehend APIs would give you the information your editors have requested? (Select THREE)

A) CreateDocumentClassifier
B) DetectSentiment
C) DetectSyntax
D) DetectEntities
E) DetectKeyPhrases
F) DetectDominantLanguage

**Answers:** B, D, E

**Explanation:**
Option A is incorrect. The CreateDocumentClassifier Comprehend API creates a document classifier that you use to categorize documents. Your editors want you to find the underlying intent of the comments.

Option B is correct. The DetectSentiment Comprehend API gives you the underlying sentiment (positive, neutral, mixed, or negative) of a string, such as a comment.

Option C is incorrect. The DetectSyntax Comprehend API gives you the part of speech of each word in a string. This would not help you understand the underlying intent of a comment.

Option D is correct. The DetectEntities Comprehend API finds named entities in text. This would help you find entities such as a news organization, politicians, celebrities, companies, etc. This information will help you identify the subject matter of the comments.

Option E is correct. The DetectKeyPhrases Comprehend API finds key noun phrases in text. This will also help you identify the subject matter of a comment.

Option F is incorrect. The DetectDominantLanguage Comprehend API finds the language (English, French, Spanish, etc.) used most frequently in the comments. This would not offer you much insight into the intent of a comment.

**Reference:**
Please see the Amazon Comprehend developer guide titled Amazon Comprehend

Question: 74
**Main Topic :** Machine Learning
**Sub Topic : Select the appropriate model(s) for a given machine learning problem**

**Domain:** Modeling

**Question text**:
You work as a machine learning specialist for a marketing consulting firm. Your firm has an online retailer as a client that wants to apply different marketing strategies per segment of their customer base. They have decided that the best way to segment their customers is by their purchase history. You have all of the online retailer purchase history from the last 5 years that you can use for your machine learning model.

Which type of machine learning algorithm would give you segmentation based on purchase history in the most expeditious manner?

A) K-Nearest Neighbors (KNN)
B) K-Means
C) Semantic Segmentation
D) Neural Topic Model (NTM)

**Answer:** B

**Explanation:**
Option A is incorrect. The k-nearest neighbor algorithm is used to find items that are similar to each other. This may find purchases that are similar to each other, but not customers that have similar purchase history. You would have to do additional modeling to use this algorithm.

Option B is correct. The K-Means algorithm is used to find groups within data where the members of the group are similar to each other but different from members of other groups. This is exactly what you are trying to solve: find groups of customers with similar purchase history.

Option C is incorrect. The semantic segmentation algorithm is used to develop computer vision applications. You are trying to solve a clustering problem, so this algorithm would not work for this problem.

Option D is incorrect. The Neural Topic Model algorithm is used to group documents into topics using the statistical distribution of words within the documents. You are trying to solve a clustering problem, so this algorithm would not work for this problem.

**Reference:**
Please see the Amazon SageMaker developer guide titled Using Amazon SageMaker Built-in Algorithms, and the article titled The 5 Clustering Algorithms Data Scientists Need to Know

Question: 75
**Main Topic :** Machine Learning
**Sub Topic :** **Perform feature engineering**

**Domain:** Exploratory Data Analysis

**Question text**:
You work for the security department of your firm. As part of securing your firm's email activity from phishing attacks you need to build a machine learning model that analyzes incoming email text to find word phrases like "you're a winner" or "click here now" to find potential phishing emails.

Which of the following text feature engineering techniques is the best solution for this task?

A) Orthogonal Sparse Bigram (OSB)
B) Term Frequency-Inverse Document Frequency (tf-idf)
C) Bag-of-Words
D) N-Gram

**Answer:** D

**Explanation:**
Option A is incorrect. The Orthogonal Sparse Bigram natural language processing algorithm creates groups of words and outputs the pairs of words that includes the first word. You are trying to classify an email as a phishing attack by having your model learn based on the presence of multi-word phrases in the email text, not pairs of words from the email text stream using the first word as the key.

Option B is incorrect. Term Frequency-Inverse Document Frequency determines how important a word is in a document by giving weights to words that are common and less common in the document. You are trying to classify an email as a phishing attack by having your model learn based on the presence of multi-word phrases in the email text, you are not trying to determine the importance of a word of phrase in the email text.

Option C is incorrect. The Bag-of-Words natural language processing algorithm creates tokens of the input document text and outputs a statistical depiction of the text. The statistical depiction, such as a histogram, shows the count of each word in the document. You are trying to classify an email as a phishing attack by having your model learn based on the presence of multi-word phrases in the email text, not individual words.

Option D is correct. The N-Gram natural language processing algorithm is used to find multi-word phrases in text, in this case an email. This suits your phishing detection task since you are trying to classify an email as a phishing attack by having your model learn based on the presence of multi-word phrases.

**Reference:**
Please see the article titled Introduction to Natural Language Processing for Text, and the article titled Document Classification Part 2: Text Processing (N-Gram Model & TF-IDF Model)

Question: 76
**Main Topic :** Machine Learning
**Sub Topic :** **Perform feature engineering**

**Domain:** Identify and implement a data-transformation solution

**Question text**:
You work for a car manufacturer as a machine learning specialist. Your marketing team wants to use a marketing strategy to market to different consumer segments based on how the features of each of their cars resonate with their customer base.

The dataset with which you have to work contains many features about each car, such as color, size, number of doors, number of speakers, type of roof, type of auto-assist, etc. Through your exploratory modeling you have found many of these features are redundant, meaning they don't offer anything further to your algorithm's performance.

Your dataset contains a large number of observations and a large number of features. How would you solve this redundant feature problem in the most efficient and expeditious manner?

A) Keep all the features and use the XGBoost algorithm to account for redundant features
B) Use Sparse Feature Graph to remove the redundant features

C) Use Principal Component Analysis to reduce the number of features
D) Keep all the features and use the Random Cut Forest algorithm to account for redundant features

**Answer:** C

**Explanation:**
Option A is incorrect. The XGBoost algorithm is used to predict a target variable in a very fast and efficient manner. However, the XGBoost will not automatically adjust for redundant features. The redundant features will act as a performance drag since you have a large number of features and a large number of observations.

Option B is incorrect. Removing the redundant features outright creates the risk of information loss. A better solution is to find composites of features that are uncorrelated, which is the technique used by Principal Component Analysis.

Option C is correct. Principal Component Analysis is a machine learning algorithm that reduces dimensionality within your data without sacrificing information. It does this by finding composites of features that are uncorrelated

Option D is incorrect. The Random Cut Forest algorithm is used to find atypical data points in a dataset, therefore it will not help find redundant features. The redundant features will act as a performance drag since you have a large number of features and a large number of observations.

**Reference:**
Please see the Amazon SageMaker developer guide titled Using Amazon SageMaker Built-in Algorithms, the Amazon SageMaker developer guide titled Principal Component Analysis (PCA) Algorithm, and the article titled Automatically Redundant Features Removal for Unsupervised Feature Selection via Sparse Feature Graph

Question: 77
**Main Topic :** Machine Learning
**Sub Topic :** **ML Implementation and Operations**

**Domain:** Deploy and operationalize machine learning solutions

**Question text**:
You work for an auto parts manufacturer as a machine learning specialist. You need to build a machine learning model that categorizes proprietary auto parts as they traverse your plant's production lines. You do not have any existing trained models from which to start your work.

You plan to use an image classification algorithm such as ResNet to classify the auto parts with one or more labels. The classified image data will then be used by your accounting department to dynamically keep the company's parts database updated with the newly produced units.

Since you are building a model to classify images of proprietary auto parts, which technique can you use within SageMaker to expedite the deployment and operation of your model?

A) Online learning
B) Incremental learning
C) Transfer learning
D) Out-of-core learning

**Answer:** C

**Explanation:**
Option A is incorrect. Online learning refers to the process of training your model incrementally by giving it data observations as individual observations or in mini-batches. This will train your model, but it won't expedite the process.

Option B is incorrect. Incremental learning would help expedite the training process if you are starting with an existing model and extending it with new data, specifically your proprietary auto parts images. However, you don't have any existing trained models from which to start your work.

Option C is correct. When you use transfer learning you start with an existing trained model, usually 'off the shelf' from a source such as ONNX Model Zoo. You take the existing trained model and apply it to your different but closely aligned observations. This saves you time in deploying and operationalizing your machine learning solution since you are starting from a pretrained model.

Option D is incorrect. Out-of-code learning is used to train huge datasets that you can't load into your server's memory. This algorithm loads some of the data, trains on that subset, loads another subset of observations, trains on that subset, and repeats this process until it has completed the training of all the observations. This process will not help you deploy and operationalize your model more expeditiously.

**Reference:**
Please see the Amazon SageMaker developer guide titled Using Amazon SageMaker Built-in Algorithms, the Amazon SageMaker machine learning blog titled Now easily perform incremental learning on Amazon SageMaker, and the article titled Transfer learning with MXNet Gluon

Question: 78
**Main Topic :** Machine Learning
**Sub Topic :** **Perform hyperparameter optimization**

**Domain:** Modeling

**Question text**:
You work as a machine learning specialist on a team tasked with designing an image recognition system that can adapt to new observations very quickly. You team is designing automated driving software for cars in a ride-share fleet. Your company wants to implement a service where when users hail a ride through your app on their mobile device, a nearby self-driving car arrives at the user's location. It has the desired route preloaded and is ready to take the user to their destination. Your team has decided to use the SageMaker Image Classification algorithm in your image recognition model.

The machine learning models powering this self-driving car fleet need to react very quickly to new observations, such as previously not encountered obstacles like different types and sized animals, etc. Which hyperparameter would you set, and to what value, to obtain the desired outcome?

A) early_stopping set to True
B) early_stopping set to False
C) learning_rate set to 0.1
D) learning_rate set to 0.8
E) use_pretrained_model set to 0
F) use_pretrained_model set to 1

**Answer:** D

**Explanation:**
Option A is incorrect. The early_stopping hyperparameter is used to decide whether to use early stopping during training. This hyperparameter allows you to terminate a training job early if it is observed that further training will not be necessary. Tuning this hyperparameter would not help your model react very quickly to new observations.

Option B is incorrect. The early_stopping hyperparameter is used to decide whether to use early stopping during training. This hyperparameter allows you to terminate a training job early if it is observed that further training will not be necessary. Tuning this hyperparameter would not help your model react very quickly to new observations.

Option C is incorrect. The learning_rate hyperparameter governs how quickly the model adapts to new or changing data. Valid values range from 0.0 to 1.0. Setting this hyperparameter to a

low value, such as 0.1, will make the model learn more slowly. This is not what you want, you want your model to learn very rapidly.

Option D is correct. The learning_rate hyperparameter governs how quickly the model adapts to new or changing data. Valid values range from 0.0 to 1.0. Setting this hyperparameter to a high value, such as 0.8, will make the model learn quickly. This is what you want, you want your model to learn very rapidly.

Option E is incorrect. The use_pretrained_model hyperparameter defines whether you want a pre-trained model to be loaded before training. This will not help you adapt quickly to new or changing observations.

Option F is incorrect. The use_pretrained_model hyperparameter defines whether you want a pre-trained model to be loaded before training. This will not help you adapt quickly to new or changing observations.

**Reference:**
Please see the Amazon SageMaker developer guide titled Image Classification Hyperparameters, and the Amazon Machine Learning blog titled Amazon SageMaker Automatic Model Tuning now supports early stopping of training jobs

Question: 79
**Main Topic :** Machine Learning
**Sub Topic :** **Deploy and operationalize machine learning solutions**

**Domain:** Machine Learning Implementation and Operations

**Question text**:
You work as a machine learning specialist for a gaming software company. You have trained and tested a machine learning model to predict gaming users likelihood of buying in-app purchases based on their player characteristics, such as playing time, levels achieved, etc. You are now ready to deploy your trained model onto the Amazon SageMaker Hosting service.

What are the three steps for deploying a model using Amazon SageMaker Hosting Services? (Select THREE)

A) Create a model in Amazon SageMaker including the S3 path where the model artifacts are stored and the Docker registry path for the inference image
B) Create a model in Amazon SageMaker including the S3 path where the model artifacts are stored and the Kubernetes registry path for the inference image
C) Create an endpoint configuration for a REST endpoint

D) Create an endpoint configuration for an HTTPS endpoint
E) Create an HTTPS endpoint
F) Create a REST endpoint

**Answers:** A, D, E

**Explanation:**
Option A is correct. From the Amazon SageMaker developer guide titled Deploy a Model on Amazon SageMaker Hosting Services "By creating a model, you tell Amazon SageMaker where it can find the model components. This includes the S3 path where the model artifacts are stored and the Docker registry path for the image that contains the inference code."

Option B is incorrect. The Amazon SageMaker Hosting Service expects to find the inference code in a Docker container, not in Kubernetes. (See the Amazon SageMaker developer guide titled Deploy a Model on Amazon SageMaker Hosting Services)

Option C is incorrect. The Amazon SageMaker Hosting Service uses an HTTPS endpoint (not a REST endpoint) to provide inferences from the model. (See the Amazon SageMaker developer guide titled Deploy a Model on Amazon SageMaker Hosting Services)

Option D is correct. The Amazon SageMaker Hosting Service uses an HTTPS endpoint to provide inferences from the model. This endpoint is configured to provide models to launch and instances on which to run them. (See the Amazon SageMaker developer guide titled Deploy a Model on Amazon SageMaker Hosting Services)

Option E is correct. The Amazon SageMaker Hosting Service uses an HTTPS endpoint to provide inferences from the model. Client applications send requests to the SageMaker runtime HTTPS endpoint to get inferences, in your case to get inferences on the probability that a gamer will buy in-app purchases.

Option F is incorrect. The Amazon SageMaker Hosting Service uses an HTTPS endpoint (not a REST endpoint) to provide inferences from the model.

**Reference:**
Please see the Amazon SageMaker developer guide titled Deploy a Model on Amazon SageMaker Hosting Services for an overview of the deployment of a SageMaker model.

Question: 80
**Main Topic :** Machine Learning
**Sub Topic :** **Create data repositories for machine learning**

**Domain:** Data Engineering

**Question text**:
You are building a data repository for your company's social media website that allows users to upload photos and videos to their personal stream. These photos and videos need to be labeled and classified so your company can use them to build direct marketing capabilities into your application based on machine learning. The direct marketing capability will be used to send targeted advertisements to users who have uploaded videos or photos of content that relates to a given products.

You are using Amazon SageMaker Ground Truth to label you user's photos and videos. Sometimes your Ground Truth human workers mislabel images and/or videos. Which SageMaker Ground Truth feature helps you continue to get high quality labeling in an automated way even when your workers occasionally mislabel?

A) Chaining labeling jobs
B) Label verification and adjustment
C) Batches for labeling tasks
D) Annotation consolidation

**Answer:** D

**Explanation:**
Option A is incorrect. Ground Truth chaining labeling jobs allows you to reuse datasets from previous labeling jobs. This feature would not help you address mislabeled images or videos.

Option B is incorrect. The Ground Truth label verification and adjustment feature allows you to have workers verify and correct labels that were mislabeled. This would help you correct mislabeled items, but it is not an automated process, it is manual.

Option C is incorrect. The Ground Truth batches for labeling tasks feature is used to send objects to your workers in batches. This would not help you correct mislabeled objects.

Option D is correct. The Ground Truth annotation consolidation feature allows you to combine the annotations of multiple workers to produce an automated probabilistic estimate of what the correct label should be.

**Reference:**
Please see the Amazon SageMaker developer guide titled Data Labeling, and the Amazon Machine Learning blog titled Use the wisdom of crowds with Amazon SageMaker Ground Truth to annotate data more accurately

Question: 81
**Main Topic :** Machine Learning

**Sub Topic :** **Create data repositories for machine learning**

**Domain:** Data Engineering

**Question text**:
You work as a machine learning specialist for a media sharing service. The media sharing service will be used by healthcare professionals to share images of x-rays, MRIs, and other medical imagery. The accuracy of labeling these images is of primary importance, since the labeling will be used in autodiagnostic software. As your team builds the data repository to be used by your machine learning algorithms, you need to use human manual labelers. You have decided to use Amazon Ground Truth for this purpose. Since accuracy is of prime importance, you have decided to use the annotation consolidation feature of Ground Truth to ensure proper labeling of the medical images.

Which of the Ground Truth annotation consolidation functions should you use for ensuring the accuracy of your labeling tasks? (Select TWO)

A) Bounding box
B) Semantic segmentation
C) Named entity
D) Output manifest
E) Mechanical turk

**Answers:** A, B

**Explanation:**
Option A is correct. The bounding box finds the most similar bounding boxes from workers and averages them, thus using the power of multiple workers to annotate your images more accurately.

Option B is correct. The semantic segmentation feature fuses the pixel annotations of multiple workers and applying a smoothing function to the image, thus using the power of multiple workers to annotate your images more accurately.

Option C is incorrect. The named entity feature is used with text annotation work, not image annotation.

Option D is incorrect. The Ground Truth output manifest allows the output of a labeling job to be used as the input to a machine learning model. This feature will not help ensure accuracy of worker annotations.

Option E is incorrect. The Ground Truth Mechanical Turk feature gives you access to a large pool of labeling workers. While increasing the number of workers at your disposal, this feature will not help ensure accuracy of worker annotations.

**Reference:**
Please see the Amazon SageMaker developer guide titled Annotation Consolidation, and the Amazon Machine Learning blog titled Use the wisdom of crowds with Amazon SageMaker Ground Truth to annotate data more accurately, and GitHub repository titled Amazon Sagemaker Examples Introduction to Ground Truth Labeling Jobs

Question: 82
**Main Topic :** Machine Learning
**Sub Topic :** **Recommend and implement the appropriate machine learning services and features for a given problem**

**Domain:** ML Implementation and Operations

**Question text**:
You work as a machine learning specialist for a large software company that has several huge data centers around the world. Your company has realized they could do a better job managing their data center power usage effectiveness (PUE) by implementing a machine learning system to automate the management of the many controls used to control their data center power usage. The machine learning model needs to take as inputs data from building management systems such as chillers, pumps, cooling units, the actual load from systems usage, etc. You have trained your model based on historical data of these inputs and the desired outcomes in these historical observations. Now you want to run your model to process real-time inferences while also continuing to learn from the new inferences.

Which combination of SageMaker algorithms and learning techniques should you use for your model to predict settings that optimize PUE on an ongoing basis?

A) Supervised learning using a Convolutional Neural Network algorithm
B) Unsupervised learning using a Multilayer Perceptron algorithm
C) Reinforcement learning using a Convolutional Neural Network algorithm
D) Unsupervised learning using a Sequence-to-Sequence Neural Network algorithm
E) Supervised learning using a Feedforward Neural Network algorithm

**Answer:** C

**Explanation:**
Option A is incorrect. In order to benefit from the trained model and then perform inferences while continuing to learn from the inferences, you cannot use supervised learning, you need to use reinforcement learning.

Option B is incorrect. The Multilayer Perceptron algorithm is used for speech recognition and translation.

Option C is correct. Reinforcement learning is used to continually update your model as new inference observations are encountered. Also, the Convolutional Neural Network algorithm is typically used in scenarios like this. (See the article titled A Practical Guide to Artificial Intelligence for the Data Center)

Option D is incorrect. The Sequence-to-Sequence Neural Network algorithm is used for machine translation and question answering systems.

Option E is incorrect. The Feedforward Neural Network algorithm is a simple neural network not capable of handling a complex problem like data center power usage effectiveness management. (See the article titled A Practical Guide to Artificial Intelligence for the Data Center)

**Reference:**
Please see the article titled A Practical Guide to Artificial Intelligence for the Data Center, the article titled Demystifying reinforcement learning and convolutional neural network, the wikipedia article titled Reinforcement learning, the wikipedia article titled Convolutional neural network, and the article titled A Comprehensive Guide to Types of Neural Networks

Question: 83
**Main Topic :** Machine Learning
**Sub Topic :** **Recommend and implement the appropriate machine learning services and features for a given problem**

**Domain:** ML Implementation and Operations

**Question text**:
You work as a machine learning specialist for a home maintenance automation company that produces robots to vacuum the floor, mow the lawn, and other automated worker tools. You have built and trained your model (starting from a pre-trained model from ImageNet) using the SageMaker built-in Object Detection algorithm. The Object Detection algorithm is used by the robots to detect objects that are obstacles or boundaries in their work area. You now need to have the robots run in real home settings using your model. You also want your robots to be able to communicate with each other if there is more than one robot in the operating area.

Which set of Amazon services will give you the most cost effective solution?

A) Amazon Elastic Inference and AWS IoT Greengrass

B) AWS RoboMaker and Amazon Sumerian
C) Amazon Rekognition and AWS IoT Greengrass
D) Amazon Rekognition and Amazon Sumerian

**Answer:** A

**Explanation:**
Option A is correct. Amazon Elastic Inference allows you to reduce the cost of your inference learning by up to 75% while giving you the inference processing (CPU, GPU, etc.) you need to process your obstacle and boundary observations. AWS IoT Greengrass gives you the capability to run inference on your robot devices and communicate with other IoT devices.

Option B is incorrect. Amazon Sumerian is used for augmented reality, which is not needed to solve your machine learning scenarios.

Option C is incorrect. Amazon Rekognition is used for image and video analysis. It would identify objects in your domain, but it wouldn't contribute to lowering the cost of your inference implementation.

Option D is incorrect. Amazon Rekognition is used for image and video analysis. It would identify objects in your domain, but it wouldn't contribute to lowering the cost of your inference implementation. Also, Amazon Sumerian is used for augmented reality, which is not needed to solve your machine learning scenarios.

**Reference:**
Please see the Amazon SageMaker Overview, particularly the Deploy and manage models in production section, the Amazon Elastic Inference Overview, the AWS News blog titled Amazon Elastic Inference – GPU-Powered Deep Learning Inference Acceleration, the Amazon SageMaker developer guide titled Object Detection Algorithm, the AWS IoT Greengrass Overview, the Amazon Sumerian Overview, and the Amazon Rekognition Overview

Question: 84
**Main Topic :** Machine Learning
**Sub Topic :** Analyze and visualize data for machine learning

**Domain:** Exploratory Data Analysis

**Question text**:
You work for a city government in their shared bike program as a machine learning specialist. You need to visualize the bike share location predictions you are producing on an hourly basis using your model inference you created using the SageMaker built-in K-Means algorithm. Your inference endpoint takes IoT data from your shared bikes as they are used throughout the city.

You also want to enrich your shared bike data with external data sources such as current weather and road conditions.

Which set of Amazon services would you use to create your visualization with the least amount of effort?

A) IoT Core -> IoT Analytics ->SageMaker -> QuickSight
B) IoT Core -> Kinesis Firehose -> SageMaker -> QuickSight
C) IoT Core -> Lambda -> SageMaker -> QuickSight
D) IoT Core -> IoT Greengrass -> QuickSight

**Answer:** A

**Explanation:**
Option A is correct. IoT Core collects data from each shared bike, IoT Analytics retrieves messages from the shared bikes as they stream data data, IoT Analytics also enriches the streaming data with your external data sources and sends the streaming data to your K-Means machine learning inference endpoint, QuickSight is then used to create your visualization. This approach requires the least amount of effort mainly because of the data enrichment feature of IoT Analytics.

Option B is incorrect. With this option you would have to create a lambda function to gather the data enrichment information (weather, road conditions) and enrich the data streams in your own code.

Option C is incorrect. Also, with this option you would have to add code to your lambda function to gather the data enrichment information (weather, road conditions) and enrich the data streams in your own code.

Option D is incorrect. IoT Greengrass is a service that you use to run local machine learning inference capabilities on connected devices. This approach would not easily integrate with your QuickSight visualization.

**Reference:**
Please see the AWS IoT Analytics overview, the Amazon SageMaker developer guide titled K-Means Algorithm, the AWS Big Data blog titled Build a Visualization and Monitoring Dashboard for IoT Data with Amazon Kinesis Analytics and Amazon QuickSight, the AWS IoT Analytics User Guide titled What IS AWS IoT Analytics?, and the AWS IoT Greengrass FAQs

Question: 85
**Main Topic :** Machine Learning
**Sub Topic :** Identify and implement a data-transformation solution

**Domain:** Data Engineering

**Question text**:
You work for a logistics company that specializes in the storage, movement, and control of massive amounts of packages. You are on the machine learning team assigned the task of building a machine learning model to assist in the control of your company's package logistics. Specifically, your model needs to predict the routes your package movers should take for optimal delivery and resource usage. The model requires various transformations to be performed on the data. You also want to get inferences on entire datasets once you have your model in production. Additionally, you won't need a persistent endpoint for applications to call to get inferences.

Which type of production deployment would you use to get predictions from your model in the most expeditious manner?

A) SageMaker Hosting Services
B) SageMaker Batch Transform
C) SageMaker Containers
D) SageMaker Elastic Inference

**Answer:** B

**Explanation:**
Option A is incorrect. SageMaker Hosting Services is used for applications to send requests to an HTTPS endpoint to get inferences. This type of deployment is used when you need a persistent endpoint for applications to call to get inferences.

Option B is correct. SageMaker Batch Transform is used to get inferences for an entire dataset and you don't need a persistent endpoint for applications to call to get inferences.

Option C is incorrect. SageMaker Containers is a service you can use to create your own Docker containers to deploy your models. This would not be the most expeditious option.

Option D is incorrect. SageMaker Elastic Interface is used to accelerate deep learning inference workloads. This service alone would not give you the batch transform capabilities you need.

**Reference:**
Please see the Amazon SageMaker developer guide titled Deploy a Model on Amazon SageMaker Hosting Services, the Amazon SageMaker developer guide titled Get Inferences for an Entire Dataset with Batch Transform, the Amazon Elastic Inference developer guide titled What Is Amazon Elastic Inference?, and the Amazon SageMaker developer guide titled Amazon SageMaker Containers: a Library to Create Docker Containers

Question: 86
**Main Topic :** Machine Learning
**Sub Topic :** Perform feature engineering

**Domain:** Exploratory Data Analysis

**Question text**:
You work for a flight diagnostics company that builds instrumentation for airline manufacturers. Your company's instrumentation hardware and software is used to detect flight pattern information such as flight path deviation, as well as airline component malfunction. Your team of machine learning specialists has created a model using the Random Cut Forest algorithm to be used to identify anomalies in the data. The streaming data that your instrumentation processes needs to be cleaned and transformed via feature engineering before passing it to your inference endpoint. You have created the pre-processing and post-processing steps (for cleaning and feature engineering) in your training process.

How can you implement the cleaning and feature engineering steps in your inference processing in the most efficient manner?

A) Execute the pre-processing in a client application before sending the data to your inference endpoint
B) Bundle and export the training pre-processing steps and deploy them to your inference container
C) Bundle and export the training pre-processing steps and deploy them as part of your Inference Pipeline
D) Bundle and export the training pre-processing steps and deploy them to IoT Core on the data emitting devices.

**Answer:** C

**Explanation:**
Option A is incorrect. Although you could execute your pre-processing steps in a client application before sending the data on to your inference end-point, this would require additional work on your part to build that client application and then incorporate your feature engineering scripts from your training process into it.

Option B is incorrect. You could also include your pre-processing steps in your inference container, however this requires more work on your part than using the SageMaker Inference Pipelines feature.

Option C is correct. SageMaker Inference Pipelines allows you to bundle and export your pre and post-processing steps from your training process and deploy them as part of your Inference Pipeline. Inference Pipelines are fully managed by AWS.

Option D is incorrect. Amazon IoT Core is used to facilitate device intercommunication. It is not a service you would use for pre-processing data streams for machine learning inference endpoints.

**Reference:**
Please see the Amazon announcement titled Announcing Enhancements for Data Processing and Feature Engineering, and Improved Framework Support with Amazon SageMaker, the Amazon SageMaker developer guide titled Deploy an Inference Pipeline, the AWS Machine Learning blog titled Use the built-in Amazon SageMaker Random Cut Forest algorithm for anomaly detection, and the AWS IoT Core Overview page

Question: 87
**Main Topic :** Machine Learning
**Sub Topic :** Frame business problems as machine learning

**Domain:** Modeling

**Question text**:
You work as a machine learning specialist for a farming corporation that wants to use in-ground soil sensors together with enrichment from geolocation, rainfall, and other weather information for the growing area to help identify crop growth stages. They want to use the crop growth information to increase yield and produce more product year over year. They also hope to increase the crop quality through this effort.

The machine learning models that you build for this solution will analyze various growing conditions, such as temperature and humidity so the farming corporation can schedule watering appropriately for the area.

What collection of AWS services would you use to implement a solution that first trains your model, then gathers the information from the in-ground sensors, then enriches the sensor data, and finally deploys the model to run inference on connected devices in the field?

A) SageMaker, IoT Core, IoT Analytics, IoT Greengrass
B) SageMaker, IoT Core, Kinesis Data Analytics, IoT Greengrass
C) SageMaker, IoT Code, Kinesis Data Streams, IoT Greengrass
D) SageMaker, IoT Core, IoT Analytics, Inference Pipeline

**Answer:** A

**Explanation:**
Option A is correct. SageMaker is used to create your model and train it initially. IoT Core sends the sensor data to IoT Analytics for enrichment and analysis. The pre-trained model is deployed into the field using IoT Greengrass so you can perform ML inference using the enriched data on the farm local devices in the field.

Option B is incorrect. You could use Kinesis Data Analytics to analyze your IoT device data streams, but IoT Analytics is built specifically for analyzing the highly unstructured IoT data, so it is a better choice.

Option C is incorrect. You could use Kinesis Data Streams to stream your IoT device data, but you would have to write lambda functions to perform the enrichment step. IoT Analytics is built specifically for analyzing and enriching the highly unstructured IoT data, so it is a better choice.

Option D is incorrect. Inference Pipeline is used to define and deploy pretrained SageMaker algorithms. Inference Pipeline does not have the IoT inference integration that IoY Greengrass has, so IoT Greengrass is a better choice for this problem.

**Reference:**
Please see the AWS IoT Greengrass ML Inference overview, the AWS IoT Analytics overview, the Amazon Kinesis Data Analytics overview, and Amazon SageMaker developer guide titled Deploy an Inference Pipeline

Question: 88
**Main Topic :** Machine Learning
**Sub Topic :** Perform hyperparameter optimization

**Domain:** Modeling

**Question text**:
You work for a transportation company as a machine learning specialist. You are currently working on a project to optimize container truck routes with the objective of minimizing empty container travel. For example, as a truck delivers its payload to a destination you want to have the container loaded for another route, you don't want the truck to move to another destination with an empty container. You have selected the SageMaker XGBoost algorithm for your model. You now need to tune your hyperparameters to get the optimum performance out of your model. You have chosen the Area Under the Curve (AUC) metric as your objective metric for your hyperparameter tuning job.

Which algorithm should you use as the SageMaker hyperparameter tuning algorithm to get your results in the minimal number of training jobs?

A) Random search
B) Bayesian Search
C) Linear Search
D) Depth First Search

**Answer:** B

**Explanation:**
Option A is incorrect. SageMaker uses two types of models to search for the optimum hyperparameters for your model: Random Search and Bayesian Search. For most models, Bayesian Search requires less training jobs to reach your optimal hyperparameter settings. (See the Amazon Machine Learning blog titled Amazon SageMaker automatic model tuning now supports random search and hyperparameter scaling)

Option B is correct. SageMaker uses two types of models to search for the optimum hyperparameters for your model: Random Search and Bayesian Search. For most models, Bayesian Search requires less training jobs to reach your optimal hyperparameter settings. (See the Amazon Machine Learning blog titled Amazon SageMaker automatic model tuning now supports random search and hyperparameter scaling)

Option C is incorrect. SageMaker hyperparameter tuning does not use Linear Search as a hyperparameter tuning model.

Option D is incorrect. SageMaker hyperparameter tuning does not use Depth First Search as a hyperparameter tuning model.

**Reference:**
Please see the Amazon SageMaker developer guide titled Configure and Launch a Hyperparameter Tuning Job, the Amazon SageMaker developer guide titled Automatic Model Tuning, and the Amazon SageMaker developer guide titled How Hyperparameter Tuning Works

Question: 89
**Main Topic :** Machine Learning
**Sub Topic :** Perform hyperparameter optimization

**Domain:** Modeling

**Question text**:
You work for a software company that produces an online sports betting app. You are on the machine learning team responsible for building a model that predicts the likelihood of registered

users to wager on a given event based on several features of sports events offered in the app. You and your team have selected the Linear Learner algorithm and have trained your model. You now wish to find the best set of hyperparameters for your model. You have chosen to use SageMaker's automatic model tuning and you have set your objective to validation:precision in your hyperparameter tuning job.

How do pass your tuning job settings into your hyperparameter tuning job? (Select THREE)

A) Define a JSON object and pass it as the value of the HyperParameterConfig to the HyperParameterTuningJob
B) Define a JSON object and pass it as the value of the HyperParameterTuningJobConfig to the CreateHyperParameterTuningJob
C) In the JSON object specify the ranges of the hyperparameters you want to tune
D) In the JSON object specify the limits of the hyperparameters you want to tune
E) In the JSON object specify the objective metric for the hyperparameter tuning job
F) In the JSON object specify the MaxSequentialTrainingJobs parameter in the ResourceLimits section

**Answers:** B, C, E

**Explanation:**
Option A is incorrect. The correct name of the value you use to pass your JSON object is HyperParameterTuningJobConfig and the name of the job is CreateHyperParameterTuningJob.

Option B is correct. To specify the hyperparameter settings for your hyperparameter tuning job you pass a JSON object as the HyperParameterTuningJobConfig parameter to the job named CreateHyperParameterTuningJob

Option C is correct. You specify the ranges of the hyperparameters you want to tune in the ParameterRanges section of the HyperParameterTuningJobConfig.

Option D is incorrect. You specify the ranges of the hyperparameters you want to tune in the ParameterRanges section of the HyperParameterTuningJobConfig, not the limits of the hyperparameters.

Option E is correct. In the HyperParameterTuningJobObjective section of the HyperParameterTuningJobConfig you set MetricName to the objective metric for the hyperparameter tuning job.

Option F is incorrect. There is no MaxSequentialTrainingJobs parameter in the ResourceLimits section of the HyperParameterTuningJobConfig.

**Reference:**

Please see the Amazon SageMaker developer guide titled Automatic Model Tuning, and the Amazon SageMaker developer guide titled Configure and Launch a Hyperparameter Tuning Job

Question: 90
**Main Topic :** Machine Learning
**Sub Topic :** Analyze and visualize data for machine learning

**Domain:** Exploratory Data Analysis

**Question text**:
You are working on a Linear Learner algorithm based model used to predict the quarterly sales for each region of your company's global sales force. The model needs to use data from your sales team's past sales performance, such as quantity of products sold, revenue generated, expenses incurred, sales force size, etc.

You and your team are in the process of training the model based on the SageMaker built-in Linear Learner algorithm. You want to track and monitor metrics, such as test objective loss and test precision as the model trains. Which AWS service(s) would you use to track and monitor these metrics? (Select THREE)

A) Specify the metrics you want to track using the AWS Management Dashboard for SageMaker
B) Specify the metrics you want to track using the AWS Management Console for SageMaker
C) Specify the metrics you want to track using the SageMaker Javascript SDK APIs
D) Specify the metrics you want to track using the SageMaker Python SDK APIs
E) Use the CloudWatch console for visualizing time-series curves of your metrics
F) Use the SageMaker Javascript SDK APIs to visualize your metrics programmatically

**Answers:** B, D, E,

**Explanation:**
Option A is incorrect. You can specify the metrics you want to track using the AWS Management Console for SageMaker, not the AWS Management Dashboard for SageMaker.

Option B is correct. To specify the metrics you want to track you use the AWS Management Console for SageMaker or the SageMaker Python SDK APIs.

Option C is incorrect. To specify the metrics you want to track you use the AWS Management Console for SageMaker or the SageMaker Python SDK APIs, not the SageMaker Javascript SDK APIs.

Option D is correct. To specify the metrics you want to track you use the AWS Management Console for SageMaker or the SageMaker Python SDK APIs.

Option E is correct. Once the model training starts, SageMaker streams the metrics you specified to CloudWatch where you can visualize time-series curves of your metrics.

Option F is incorrect. You can visualize your metrics either via the CloudWatch console, or the SageMaker Python SDK APIs, not the SageMaker Javascript SDK APIs.

**Reference:**
Please see the AWS Machine Learning Blog titled Easily monitor and visualize metrics while training models on Amazon SageMaker

Question: 91
**Main Topic :** Machine Learning
**Sub Topic :** Analyze and visualize data for machine learning

**Domain:** Exploratory Data Analysis

**Question text**:
You work for an oil refinery company where you are on one of their machine learning teams. Your team is responsible for building models that help the company decide where to place their exploratory drilling teams across the globe. Your team lead has decided to build your model based on the K-Means built-in SageMaker algorithm. The team lead has tasked you with providing metric visualization charts for the training runs of your team's model.

How would you go about visualizing the training metrics? (Select TWO)

A) In your SageMaker jupyter notebook, using the SageMaker python module called pandas.analytics, import TrainingAnalytics .
B) In your SageMaker jupyter notebook, using the SageMaker python module called sagemaker.analytics, import TrainingAnalytics.
C) In your SageMaker jupyter notebook, using the SageMaker python module called sagemaker.analytics, import TrainingJobAnalytics.
D) In your SageMaker jupyter notebook, using the SageMaker python module called pandas.analytics, import TrainingJobAnalytics.
E) Set one of the metric names to test:cross_entropy'
F) Set one of the metric names to test:msd'

**Answers:** C, F

**Explanation:**

Option A is incorrect. You use the SageMaker python module called sagemaker.analytics (not pandas.analytics) from which you import TrainingJobAnalytics (not TrainingAnalytics) to gain access to the python methods that allow you to visualize you metrics in charts.

Option B is incorrect. You use the SageMaker python module called sagemaker.analytics from which you import TrainingJobAnalytics (not TrainingAnalytics) to gain access to the python methods that allow you to visualize you metrics in charts.

Option C is correct. You use the SageMaker python module called sagemaker.analytics from which you import TrainingJobAnalytics to gain access to the python methods that allow you to visualize you metrics in charts.

Option D is incorrect. You use the SageMaker python module called sagemaker.analytics (not pandas.analytics) from which you import TrainingJobAnalytics to gain access to the python methods that allow you to visualize you metrics in charts.

Option E is incorrect. To set the metric name that you wish to visualize you need to give a valid metric for the algorithm you are training. The test:cross_entropy metric is not valid for a K-Means training run.

Option F is correct. To set the metric name that you wish to visualize you need to give a valid metric for the algorithm you are training. The test:msd metric is one of the two valid for a K-Means training run. The other valid metric for K-Means is test:ssd.

**Reference:**
Please see the AWS Machine Learning Blog titled Easily monitor and visualize metrics while training models on Amazon SageMaker, and the Amazon SageMaker developer guide titled Tune a K-Means model

Question: 92
**Main Topic :** Machine Learning
**Sub Topic :** Analyze and visualize data for machine learning

**Domain:** Exploratory Data Analysis

**Question text**:
You work for an online retailer as a machine learning specialist. Your team is building a deep learning model based on the Keras Sequential model to categorize the clothing your company's users post on their instagram feeds when they use one of the hashtags that refers to your company. You are the machine learning specialist assigned to building the training run visualization code to allow the team to monitor training metrics of the model as it trains.

How would you go about visualizing the training metrics? (Select TWO)

A) When creating your model training job in the SageMaker console, specify a regex pattern for the metrics that you want your model training script to write to your logs
B) When creating your model training job in the SageMaker console, specify the metrics that you want your model training script to write to your logs
C) Use the CloudWatch metrics dashboard to visualize the metrics that SageMaker automatically parsed from your logs and published for graphing and visualization.
D) Use the SageMaker metrics dashboard to visualize the metrics that SageMaker automatically parsed from your logs and published for graphing and visualization.
E) Write a python script in your SageMaker jupyter notebook to visualize the metrics that SageMaker automatically parsed from your logs and published for graphing and visualization.

**Answers:** A, C

**Explanation:**
Option A is correct. While creating your model training job in the SageMaker console, you specify a regex pattern that is used for the metrics that your model training script writes to your logs.

Option B is incorrect. While creating your model training job in the SageMaker console, you specify a regex pattern that is used for the metrics that your model training script writes to your logs. You can't specify the metrics directly, you must use a regex pattern.

Option C is correct. SageMaker parses from your logs the metrics which you wish to track and publishes them to CloudWatch. The CloudWatch metrics dashboard allows you to visualize your SageMaker training job metrics as graphs for visualization.

Option D is incorrect. The CloudWatch metrics dashboard allows you to visualize your SageMaker training job metrics as graphs for visualization, not the SageMaker metrics dashboard.

Option E is incorrect. You would not need to write a python script to visualize your metrics data since the CloudWatch metrics dashboard gives you this functionality.

**Reference:**
Please see the AWS Machine Learning Blog titled Easily monitor and visualize metrics while training models on Amazon SageMaker, the Amazon SageMaker developer guide titled Use TensorFlow with Amazon SageMaker, and the Tensorflow.org page titled Basic classification: Classify images of clothing

Question: 93
**Main Topic :** Machine Learning
**Sub Topic :** Identify and implement a data-ingestion solution

**Domain:** Data Engineering

**Question text**:
You work for a manufacturer of wifi connected radios. You company wants to use data captured when these radios are in use by their customers (such as how the hardware is performing, the applications that are running on the radio, and the content that's being streamed) to better serve their customers. You and your team of machine learning specialists have been asked to use the data captured when users play their radios to build a model that detects anomalies with the hardware performance.

What AWS service and function within that service will allow you to identify anomalies in the data stream?

A) Kinesis Data Analytics and its Hotspots function
B) Kinesis Data Analytics and its Random Cut Forest function
C) Kinesis Data Firehose and its Hotspots function
D) Kinesis Data Streams and its Random Cut Forest function
E) Kinesis Data Streams and its Hotspots function
F) Kinesis Data Firehose and its Random Cut Forest function

**Answer:** B

**Explanation:**
Option A is incorrect. The Kinesis Data Analytics Hotspot function is used to get information about dense regions in your data, not to identify outlier data, or anomalies, in your streaming data.

Option B is correct. The Kinesis Data Analytics Random_Cut_Forest function is used to identify outlier data, or anomalies, in your streaming data.

Option C is incorrect. Kinesis Data Firehose does not have functions like Hotspots or Random_Cut_Forest.

Option D is incorrect. Kinesis Data Streams does not have functions like Hotspots or Random_Cut_Forest.

Option E is incorrect. Kinesis Data Streams does not have functions like Hotspots or Random_Cut_Forest.

Option F is incorrect. Kinesis Data Firehose does not have functions like Hotspots or Random_Cut_Forest.

**Reference:**
Please see the Amazon Kinesis Data Analytics for SQL Applications Developer Guide titled Examples: Machine Learning, the Amazon Kinesis Data Analytics for SQL Applications Developer Guide titled Example: Detecting Data Anomalies on a Stream (RANDOM_CUT_FOREST Function), and the Amazon Kinesis Data Analytics for SQL Applications Developer Guide titled Example: Detecting Hotspots on a Stream (HOTSPOTS Function)

Question: 94
**Main Topic :** Machine Learning
**Sub Topic :** Perform feature engineering

**Domain:** Exploratory Data Analysis

**Question text**:
You work for a city electric scooter rental company. Your company supplies a fleet of electric scooters to different cities around the country. These scooters need to be managed as far as their location, their rental miles, their need for maintenance, etc. The company accumulates hundreds of data points on each scooter every day. You are on the machine learning team of your company where you have been assigned the job of building a machine learning model to track each scooter and decide when they are ready for maintenance. One would assume the decision for maintenance would be based predominantly on miles accumulated. Since you have so many features captured for a given scooter, you have decided you need to find the most predictive features in your model in order to avoid low model performance due to collinearity.

You have built your model in SageMaker using the built-in XGBoost algorithm. Using the XGBoost python API package, which type of booster and which API call would you use if you wanted to select the most predictive features based on the total gain across all splits in which the feature is used?

A) booster = gblinear using the get_fscore with importance_type parameter set to total_gain
B) booster = gblinear using the get_score with importance_type parameter set to gain
C) booster = gbtree using the get_score with importance_type parameter set to total_gain
D) booster = gbtree using the get_fscore with importance_type parameter set to gain
E) booster = dart using the get_fscore with importance_type parameter set to gain
F) booster = dart using the get_score with importance_type parameter set to total_gain

**Answer:** C

**Explanation:**
Option A is incorrect. To get the features based on the total gain across all splits in which the feature is used you need to use the gbtree booster and call get_score passing the parameter importance_type set to total_gain. Feature importance is defined only for base learner, or tree boosters. Feature importance is not defined for linear learners. The importance_type parameter is defined for the get_score API call, not the get_fscore API call.

Option B is incorrect. To get the features based on the total gain across all splits in which the feature is used you need to use the gbtree booster and call get_score passing the parameter importance_type set to total_gain. Feature importance is defined only for base learner, or tree boosters. Feature importance is not defined for linear learners. The importance_type parameter needs to be set to total_gain to get the total gain across all splits in which the feature is used. The importance_type parameter of gain gives you the average gain across all splits in which the feature is used.

Option C is correct. To get the features based on the total gain across all splits in which the feature is used you need to use the gbtree booster and call get_score passing the parameter importance_type set to total_gain.

Option D is incorrect. To get the features based on the total gain across all splits in which the feature is used you need to use the gbtree booster and call get_score passing the parameter importance_type set to total_gain. The importance_type parameter needs to be set to total_gain to get the total gain across all splits in which the feature is used. The importance_type parameter of gain gives you the average gain across all splits in which the feature is used.

Option E is incorrect. To get the features based on the total gain across all splits in which the feature is used you need to use the gbtree booster and call get_score passing the parameter importance_type set to total_gain. Feature importance is defined only for base learner, or tree boosters. Feature importance is not defined for dart boosters. The importance_type parameter needs to be set to total_gain to get the total gain across all splits in which the feature is used. The importance_type parameter of gain gives you the average gain across all splits in which the feature is used.

Option F is incorrect. To get the features based on the total gain across all splits in which the feature is used you need to use the gbtree booster and call get_score passing the parameter importance_type set to total_gain. Feature importance is defined only for base learner, or tree boosters. Feature importance is not defined for dart boosters.

**Reference:**
Please see the Amazon SageMaker developer guide titled XGBoost Algorithm, the Amazon SageMaker developer guide titled XGBoost Hyperparameters, and the XGBoost Python API Reference

Question: 95

**Main Topic :** Machine Learning
**Sub Topic :** Identify and implement a data-transformation solution

**Domain:** Data Engineering

**Question text**:
You work as a machine learning specialist for a retail chain that has recently purchased another retail chain and is in the process of merging the two chain's systems. Both retail chains have customer databases. Some of the firm's customers overlap, meaning that the same customer registered with both chains in the past. When merging the customer data stores of the two, now merged retail chains, you need to link duplicate customer data so that you can have one accurate customer data source.

You have been assigned the task of creating the new customer data source for the now merged retail chain. Instead of trying to find the duplicate customer data manually through traditional programming techniques, you have decided to use machine learning techniques to solve the problem.

You have determined that the AWS Glue Machine Learning FindMatches Transform is the best solution to this problem. Knowing that incorrectly linking what appear to be duplicate customers must be avoided at all costs, how should you configure the AWS Glue FindMatches ML Transform parameters to achieve the most efficient and accurate duplicate customer detection process?

A) Set the FindMatches precision-recall parameter to 'precision' and the accuracy-cost parameter to 'accuracy'
B) Set the FindMatches precision-recall parameter to 'precision' and the accuracy-cost parameter to 'lower cost'
C) Set the FindMatches precision-recall parameter to 'recall' and the accuracy-cost parameter to 'accuracy'
D) Set the FindMatches precision-recall parameter to 'recall' and the accuracy-cost parameter to 'lower cost'

**Answer:** A

**Explanation:**
Option A is correct. Setting the FindMatches precision-recall parameter to 'precision' minimizes false positives (when you don't have a match of a duplicate customer but mark it as a match mistakenly). This is what you want. Setting the FindMatches accuracy-cost parameter to 'accuracy' maximizes the transform accuracy of finding matching records as duplicate. This is also what you want.

Option B is incorrect. Setting the FindMatches precision-recall parameter to 'precision' minimizes false positives (when you don't have a match of a duplicate customer but mark it as a

match mistakenly). This is what you want. But, setting the accuracy-cost parameter to 'lower cost' favors cost or the speed of running the transform at the expense of the transform's accuracy. This may make your transform more performant, but your primary concern is avoiding linking customers incorrectly so you should set the accuracy-cost parameter to 'accuracy'.

Option C is incorrect. Setting the FindMatches precision-recall parameter to 'recall' minimizes false negatives (when you have a match of a duplicate customer but fail to detect it). This may cause customer frustration, but your primary concern is avoiding linking customers incorrectly.

Option D is incorrect. Setting the FindMatches precision-recall parameter to 'recall' minimizes false negatives (when you have a match of a duplicate customer but fail to detect it). This may cause customer frustration, but your primary concern is avoiding linking customers incorrectly.

**Reference:**
Please see the AWS Glue developer guide titled Machine Learning Transforms in AWS Glue, and the AWS Glue developer guide titled Tuning Machine Learning Transforms in AWS Glue

START LAST SET OF QUESTIONS HERE

Question: 96
**Main Topic :** Machine Learning
**Sub Topic :** Identify and implement a data-transformation solution

**Domain:** Data Engineering

**Question text**:
You work for a car rental firm in their car tracking department. Your team is responsible for building the machine learning solutions to track the company's fleet of cars. Each car is equipped with a GPS vehicle tracking device that emits IoT data. You are building a data transformation solution to take the GPS IoT data and transform it before storing it in S3 for use in your machine learning models.

You have decided to use Kinesis Data Firehose data transformation to pre-process your IoT data before storing it in S3. You have written your lambda function that pre-processes the data and you are now testing your data transformation process flow. When running your tests you see that Kinesis Data Firehose is rejecting every record as a data transformation failure. What could be the reason for the failure?

A) In your lambda function you have set the result to OK or Dropped for each record processed.

B) The transformed records from your lambda function contain the recordId and result parameters.
C) When creating your lambda function you used a lambda blueprint for data transformation from the AWS Serverless Application Repository.
D) When creating your lambda function you used a lambda blueprint for data transformation from the AWS Lambda console.

**Answer:** B

**Explanation:**
Option A is incorrect. The status of your transformed record produced by your lambda function can be Ok (the record was transformed successfully), Dropped (the record was dropped intentionally by your transformation logic), or ProcessingFailed (the record could not be transformed). A status of Ok or Dropped indicates to Kinesis Data Firehose that the record was successfully processed. A status of ProcessingFailed indicates a failed transformation. Your lambda function has set each record's status to either Ok or Dropped, so this option is incorrect.

Option B is correct. Transformed records received by Kinesis Data Firehose from lambda must contain the recordId, result, and data parameters. Your transformed records only contain the recordId and result parameters.

Option C is incorrect. You can use lambda blueprints from either the AWS Serverless Application Repository or the AWS Lambda console to create you transformation lambda function.

Option D is incorrect. You can use lambda blueprints from either the AWS Serverless Application Repository or the AWS Lambda console to create you transformation lambda function.

**Reference:**
Please see the Amazon Kinesis Data Firehose developer guide titled Amazon Kinesis Data Firehose Data Transformation

Question: 97
**Main Topic :** Machine Learning
**Sub Topic :** Apply basic AWS security practices to machine learning solutions

**Domain:** ML Implementation and Operations

**Question text**:
You work for a healthcare data provider company that gathers real-time streaming data from healthcare plan participants who have agreed to allow their insurance company use their health

data gathered by their wearable technology, such as internet connected watches and step counters. The plan participants receive discounts on their healthcare plan fees when participating in the data streaming effort. You are on the machine learning team that will use this data to better predict healthcare issues based on the gathered wearable data. Due to the secure nature of this personal information, you need to build encryption into your data pipeline for this effort.

How would you construct your data pipeline in the most secure way to ensure your data is encrypted as it moves from the IoT wearable devices to your machine learning data source?

A) Use IoT Analytics to gather the streaming data from the IoT devices, encrypt the data, and send it to your machine learning data source.
B) Use Kinesis Data Streams to gather the streaming data from the IoT devices. Have Kinesis Data Streams be the source of a Kinesis Data Firehose delivery stream which encrypts your data using an AWS Key Management Service (AWS KMS) key before storing the data at rest and then delivers the data to your S3 bucket used for your machine learning models.
C) Use Kinesis Data Streams to gather the streaming data from the IoT devices and encrypt your data using an AWS Key Management Service (AWS KMS) key before storing the data at rest. Then have Kinesis Data Streams be the source of a Kinesis Data Firehose delivery stream which delivers the data to your S3 bucket used for your machine learning models.
D) Use Kinesis Data Analytics to gather the streaming data from the IoT devices, encrypt the data, and send it to your machine learning data source.

**Answer:** C

**Explanation:**
Option A is incorrect. IoT Analytics is used to filter, transform, and enrich IoT data before storing the data in a time-series data store for analysis. IoT Analytics doesn't encrypt your data.

Option B is incorrect. Using Kinesis Data Streams to gather your IoT data and be the source for a Kinesis Data Firehose delivery stream is the correct choice. However, you would leverage Kinesis Data Streams to encrypt your data using an AWS Key Management Service (AWS KMS) key before storing the data at rest, not Kinesis Data Firehose. When you use a Kinesis data stream as the source of a Kinesis Data Firehose delivery stream, Kinesis Data Firehose does not store the data at rest. The data is stored at rest in the Kinesis Data Stream.

Option C is correct. You use Kinesis Data Streams to gather your IoT data and be the source for a Kinesis Data Firehose delivery stream. You also leverage Kinesis Data Streams to encrypt your data using an AWS Key Management Service (AWS KMS) key before storing the data at rest. Then Kinesis Data Streams is used as the source of your to Kinesis Data Firehose delivery stream, which delivers the data to your S3 bucket used for your machine learning models.

Option D is incorrect. You would have to use Kinesis Data Streams together with Kinesis Data Analytics to get the encryption needed for your solution.

**Reference:**
Please see the Amazon Kinesis Data Firehose developer guide titled Data Protection in Amazon Kinesis Data Firehose, the Amazon Kinesis Data Analytics overview page, the AWS IoT Analytics overview page, the AWS IoT Analytics user guide titled What Is AWS IoT Analytics, and the Amazon Kinesis Data Analytics for SQL Applications developers guide titled Data Protection in Amazon Kinesis Data Analytics for SQL Applications

Question: 98
**Main Topic :** Machine Learning
**Sub Topic :** Evaluate machine learning models

**Domain:** Modeling

**Question text**:
You work for a fantasy sports wagering software company as a machine learning specialist. You are the leader of a team of machine learning specialists who have been given the assignment of building a model to predict the over/under line for every professional football game each week of the NFL season. Due to the complex nature of the problem and its many feature combinations, you have your team experimenting with different datasets, algorithms, and hyperparameters to find the best combination for your machine learning problem. You don't want to limit the number of experiments your team can perform. Since you have a relatively large team of talented machine learning specialists, they will generate several hundred to over a thousand experiments over the course of your modeling effort.

Which Amazon machine learning service(s)/feature(s) should you use to help manage your team's experiments at scale?

A) Use Amazon SageMaker Inference Pipeline
B) Use Amazon SageMaker model tracking capability
C) Use Amazon SageMaker model experiments capability
D) Use Amazon SageMaker model containers capability

**Answer:** B

**Explanation:**
Option A is incorrect. The Amazon Inference Pipeline is used to deploy pretrined SageMaker algorithms packaged in Docker containers. You would not use Amazon Inference Pipeline to manage experiments at scale.

Option B is correct. You can use the Amazon SageMaker model tracking capability to search key model attributes such as hyperparameter values, the algorithm used, and tags associated with your team's models. This SageMaker capability allows you to manage your team's experiments at the scale of up to thousands of model experiments.

Option C is incorrect. There is no Amazon SageMaker feature called 'model experiments capability'

Option D is incorrect. There is no Amazon SageMaker feature called 'model containers capability'

**Reference:**
Please see the AWS announcement titled New Model Tracking Capabilities for Amazon SageMaker Are Now Generally Available, the Amazon SageMaker developer guide titled Manage Machine Learning Experiments, the AWS Machine Learning blog titled Using model attributes to track your training runs on Amazon SageMaker, the Amazon SageMaker developer guide titled Monitor and Analyze Training Jobs Using Metrics, and the Amazon SageMaker developer guide titled Deploy an Inference Pipeline

Question: 99
**Main Topic :** Machine Learning
**Sub Topic :** Evaluate machine learning models

**Domain:** Modeling

**Question text**:
You work on an application development team for a new start-up social media site. Your team is made up of data scientists and machine learning specialists, of which you are the lead machine learning specialist. Your team has built a model in SageMaker using the built-in linear learner algorithm. The team has performed several training runs in an effort to find the best datasets and hyperparameters. You have decided to use the SageMaker model tracking capability to manage the many training runs your team has produced.

You have asked your team to show you the results of their efforts to help you lead them in making the decision on which hyperparameters and test datasets to use. They have used the AWS SDK API for SageMaker to produce the data for your decision. The following is a section of code from their use of the SageMaker model tracking capability. What does the code do?

```
search_params = {
    "MaxResults": 10,
    "Resource": "TrainingJob",
    "SearchExpression": {
```

```
        “Filters”: [{
            “Name”: “Tags.Model”,
            “Operator”: “Equals”,
            “Value”: ”Model_Social_Media_Classifier”,
        }]},
    “SortBy”: “”Metrics.train:precision”,
    “SortOrder”: “Descending”
}
smclient = boto3.client(service_name=’sagemaker’)
results = smclient.search(**search_params)
```

A) It uses the SageMaker API to run at most 10 training jobs for a model called Model_Social_Media_Classifier and sorts the results by the model precision in descending order
B) It uses the SageMaker API to find the the 10 best hyperparameters (based on the precision metric) of a model that has been tagged as Model: Model_Social_Media_Classifier
C) It uses the SageMaker API to find the the 10 best training runs (based on their precision metric) of a model that has been tagged as Model: Model_Social_Media_Classifier
D) It uses the SageMaker API to run a training job called Model_Social_Media_Classifier and sorts the results by the precision metric in descending order for the 10 best results.

**Answer:** C

**Explanation:**
Option A is incorrect. The code uses the SageMaker python client API to search your team’s SageMaker resources (such as training run results) for a specific model’s training run results. It does not run any training jobs.

Option B is incorrect. The code uses the SageMaker python client API to search your team’s SageMaker resources (such as training run results) for a specific model’s training run results. It does not search for the best hyperparameters.

Option C is correct. The code uses the SageMaker python client API to search your team’s SageMaker resources (such as training run results) for a specific model’s training run results. It then sorts the results by the precision metric in descending order. This will allow you to see which training model run performed the best from a precision perspective. The results give that model’s algorithm, data sources, hyperparameter values, and metrics results.

Option D is incorrect. The code uses the SageMaker python client API to search your team’s SageMaker resources (such as training run results) for a specific model’s training run results. It does not run any training jobs.

**Reference:**

Please see the AWS announcement titled New Model Tracking Capabilities for Amazon SageMaker Are Now Generally Available, the AWS Machine Learning blog titled Using model attributes to track your training runs on Amazon SageMaker, the Amazon SageMaker developer guide titled Search, the Amazon SageMaker developer guide titled Manage Machine Learning Experiments, the AWS SageMaker Client boto3 docs, and the Amazon SageMaker developer guide titled Tune a Linear Learner Model

Question: 100
**Main Topic :** Machine Learning
**Sub Topic :** Apply basic AWS security practices to machine learning solutions

**Domain:** ML Implementation and Operations

**Question text**:
You work on an application development team for a financial services firm. You and your team are working on a mission critical project with a very aggressive timeline for implementation. For this project you are building a machine learning model to predict customer retention where you are using customer PII (Personal Identifiable Information) data. This data is very sensitive and is also controlled by SEC (Securities Exchange Commision) compliance regulations. Therefore, your data ingestion process and data storage must be highly secure. For this reason, you have a mandate to use encryption for all data storage.

How do you use SageMaker features to make sure all of your model artifacts are highly secure with the least amount of effort on your team's part?

A) Use SSL to encrypt your data on your S3 bucket (where you store your model artifacts and data) and your SageMaker jupyter notebooks. Then run your SageMaker training jobs, hyperparameter tuning jobs, batch transform jobs, and your inference endpoint using the default SageMaker IAM roles and policies.
B) Use SageMaker Neo, which encrypts your data at rest in your S3 bucket where you store your model artifacts and data. Then pass an AWS Key Management Service key to your SageMaker jupyter notebooks, training jobs, hyperparameter tuning jobs, batch transform jobs, and your inference endpoint to encrypt the S3 bucket.
C) Use encrypted S3 buckets for your model artifacts and data. Then pass an AWS Key Management Service key to your SageMaker jupyter notebooks, training jobs, hyperparameter tuning jobs, batch transform jobs, and your inference endpoint to encrypt the attached machine learning storage volume.
D) Use your customer owned AWS Key Management Service key to store your data on the ML EBS volume or in your S3 buckets, which you encrypt using your customer owned Key Management Service key. Pass your customer owned Key Management Service key to your SageMaker jupyter notebooks, training jobs, hyperparameter tuning jobs,

batch transform jobs, and your inference endpoint to encrypt the attached machine learning storage volume.

**Answer:** D

**Explanation:**
Option A is incorrect. To ensure your data is secure you use an AWS Key Management Service key to store your data and to pass to your SageMaker resources. You don't use SSL for this purpose.

Option B is incorrect. SageMaker Neo is a SageMaker service that allows you to train your model once and run it anywhere in the cloud and at the edge. SageMaker Neo does not provide encryption services.

Option C is incorrect. You should use AWS Key Management Service keys for your data and SageMaker resource encryption, but since your project requires encryption for regulatory compliance reasons, you need to use a customer owned KMS key.

Option D is correct. Since your project requires encryption for regulatory compliance reasons, you need to use a customer owned KMS key. You should use your customer owned AWS KMS key to store your data on the ML EBS volume or in your S3 buckets, which you encrypt using your customer managed KMS keys. You also should pass your customer owned KMS key to your SageMaker jupyter notebooks, training jobs, hyperparameter tuning jobs, batch transform jobs, and your inference endpoint to encrypt the attached machine learning storage volume.

**Reference:**
Please see the Amazon SageMaker developer guide titled Protecting Data at Rest Using Encryption, and the Amazon SageMaker Neo overview page

Question: 101
**Main Topic :** Machine Learning
**Sub Topic :** Identify and implement a data-ingestion solution

**Domain:** Data Engineering

**Question text**:
You work for a telecommunications service and internet provider company that has been in business for decades. Over the decades the company has built various types of application systems and database technologies on the evolving platforms of the time. Therefore, you have massive amounts of customer and company operational data on legacy mainframe systems and their associated data stores, such as aging relational databases.

Your team is attempting to build a machine learning model to use streaming data from the company's in-home routers, functioning as IoT (Internet of Things) devices, and use that data to help the company sell additional services to its customer base. The IoT data is unstructured, so you need to transform it to CSV format before you ingest it into your S3 buckets that you use to house your datasets for your SageMaker model. You also need to enrich the IoT data with real-time data from your legacy mainframe systems as the data streams into your AWS cloud environment.

Which set of Amazon services would you use to setup this data transformation and ingestion pipeline?

A) Use Kinesis Data Firehose to receive the streaming data from the IoT devices. Use the Kinesis Data Firehose lambda integration capability to enrich the IoT data with your legacy mainframe systems data and transform it to CSV before writing it to the S3 bucket used by your SageMaker model.
B) Have your legacy mainframe systems write to S3 and use AWS Storage Gateway to enrich the IoT data with your legacy system data and transform it to CSV before writing it to the S3 bucket used by your SageMaker model.
C) Have your legacy mainframe systems write to AWS Storage Gateway using the File Gateway configuration via an NFS (Network File System) connection. Use Kinesis Data Firehose to receive the streaming data from the IoT devices. Use the Kinesis Data Firehose lambda integration capability to enrich the IoT data with your legacy mainframe systems data and convert it to CSV before writing it to the S3 bucket used by your SageMaker model.
D) Use AWS Snowball to migrate your legacy mainframe data to your AWS account. Use Kinesis Data Firehose to receive the streaming data from the IoT devices. Use the Kinesis Data Firehose lambda integration capability to enrich the IoT data with your legacy mainframe systems data and convert it to CSV before writing it to the S3 bucket used by your SageMaker model.

**Answer:** C

**Explanation:**
Option A is incorrect. You can't enrich your IoT data with your mainframe data without first getting your mainframe data into your AWS cloud environment.

Option B is incorrect. You can't write directly from your mainframe systems to S3. You could use AWS Storage Gateway to get your mainframe data into your AWS cloud environment, but AWS Storage Gateway doesn't have the capability to enrich your IoT data.

Option C is correct. You can use AWS Storage Gateway using the File Gateway configuration via an NFS (Network File System) connection to move your data from your legacy mainframe systems into your AWS cloud environment. You can then use Kinesis Data Firehose lambda integration to to enrich the IoT data with your legacy mainframe systems data and convert it to

CSV. Finally, you can have your lambda function write your transformed data to your S3 bucket used by your SageMaker model.

Option D is correct. AWS Snowball moves data from your on premises environment to your AWS cloud environment in a one time batch. This wouldn't work since you need real-time integration of your legacy data with your IoT data.

**Reference:**
Please see the AWS whitepaper titled Building Big Data Storage Solutions (Data Lakes) for Maximum Flexibility

Question: 102
**Main Topic :** Machine Learning
**Sub Topic :** Train machine learning models

**Domain:** Modeling

**Question text**:
You work for a ride share software company as a machine learning specialist. You are working on a model to predict driver capacity based on a number of factors, such as location, time of day, weather, population density, age of car, etc. You have several million observations stretching back over 5 years across several geographic locations worldwide. You have performed feature engineering on your data and you have transformed it into 5 CSV files (one for each year) which you have uploaded to your S3 bucket training prefix.

Due to the large number of observations your management team anticipates that training this model could get costly, so they have asked you to keep the costs of your project as low as possible.

You have written the following python code using the SageMaker Python SDK in your SageMaker jupyter notebook:

```
s3_train =
        sagemaker.s3_input(s3_data='s3://{}/{}'.format(bucket, path_train),
                content_type='csv',distribution='ShardedByS3Key')

my_container = get_image_uri(boto3.Session().region_name, 'xgboost')
my_session = sagemaker.Session()
role = get_execution_role()

xgb = sagemaker.estimator.Estimator(my_container,
                            role,
                            train_instance_count=5,
```

```
                    train_instance_type='ml.m4.xlarge',
                    output_path=output_path,
                    sagemaker_session=my_session)

xgb.set_hyperparameters( max_depth=10,
                  eta=0.2,
                   gamma=4,
                  min_child_weight=40,
                  subsample=0.8,
                  silent=0,
                  objective='reg:linear',
                  early_stopping_rounds=10,
                  num_round=200 )

xgb.fit({'train': s3_train,
      'validation': s3_input_validation})
```

Using this code, how does SageMaker replicate your dataset to your Machine Learning instances for training?

A) SageMaker replicates the entire dataset on each of the 10 ML instances that are launched for training
B) SageMaker replicates the entire dataset on each of the 5 ML instances that are launched for training
C) SageMaker replicates a subset of your dataset on each of the 10 ML instances that are launched for training
D) SageMaker replicates a subset of your dataset on each of the 5 ML instances that are launched for training

**Answer:** D

**Explanation:**
Option A is incorrect. In the SageMaker API, when you set the distribution type parameter to ShardedByS3Key, SageMaker replicates a subset of your dataset on each of the ML instances you've defined.

Option B is incorrect. In the SageMaker API, when you set the distribution type parameter to ShardedByS3Key, SageMaker replicates a subset of your dataset on each of the ML instances you've defined. You define the quantity of the ML instances (in this case 5) in the train_instance_count parameter of the Estimator API call.

Option C is incorrect. It is correct that in the SageMaker API, when you set the distribution type parameter to ShardedByS3Key, SageMaker replicates a subset of your dataset on each of the

ML instances you've defined. You define the quantity of the ML instances (in this case 5) in the train_instance_count parameter of the Estimator API call.

Option D is correct. In the SageMaker API, when you set the distribution type parameter to ShardedByS3Key, SageMaker replicates a subset of your dataset on each of the ML instances you've defined. You define the quantity of the ML instances (in this case 5) in the train_instance_count parameter of the Estimator API call. Distributing your dataset across several instances, making your training much faster and therefore less expensive.

**Reference:**
Please see the Amazon SageMaker developer guide titled Train a Model with Amazon SageMaker, the Amazon SageMaker developer guide titled S3DataSource, and the AWS Machine Learning blog titled Amazon SageMaker Automatic Model Tuning becomes more efficient with warm start of hyperparameter tuning jobs )specifically the 'create a training estimator' section of the blog)

Question: 103
**Main Topic :** Machine Learning
**Sub Topic :** Train machine learning models

**Domain:** Modeling

**Question text**:
You work for a large healthcare diagnostics company. You are on the machine learning team responsible for predicting various anomalies in blood samples. You have data samples from all of the corporation's many testing facilities across the country. You have performed feature engineering and data cleaning on your dataset. You have also written the python code to split your dataset into training and test datasets. You are now ready to train your model for the first time.

You have written the following python code in your SageMaker jupyter notebook:

```
import sagemaker
from sagemaker.amazon.amazon_estimator import get_image_uri
from sagemaker import get_execution_role
container = get_image_uri(boto3.Session().region_name, 'xgboost')
role = get_execution_role()

s3_train = 's3://{}/{}/{}'.format(bucket, prefix, 'train')
s3_validation = 's3://{}/{}/{}'.format(bucket, prefix, 'validation')
s3_output = 's3://{}/{}/{}'.format(bucket, prefix, xgb_output)
```

```
xgb_model = sagemaker.estimator.Estimator(container,
                                role,
                                train_instance_count=1,
                                train_instance_type='ml.m4.xlarge',
                                train_volume_size = 5,
                                output_path=s3_output,
                                sagemaker_session=sagemaker.Session())

xgb_model.set_hyperparameters(max_depth = 2,
                        eta = 2,
                        gamma = 2,
                        min_child_weight = 2,
                        silent = 0,
                        objective = "multi:softmax",
                        num_class = 10,
                        num_round = 10)

train_channel = sagemaker.session.s3_input(s3_train, content_type='text/csv')
valid_channel = sagemaker.session.s3_input(s3_validation, content_type='text/csv')
data_channels = {'train': train_channel, 'validation': valid_channel}
xgb_model.fit(inputs=data_channels,  logs=True)
```

When you attempt to run this code in your SageMaker jupyter notebook it fails. You check the CloudWatch logs and find this error message:

```
AlgorithmError: u'2' is not valid under any of the given
schemas\n\nFailed validating u'oneOf' in
schema[u'properties'][u'feature_dim']:\n    {u'oneOf':
[{u'pattern': u'^([0]\.[0-9])$', u'type': u'string'},\n
{u'minimum': 0, u'type': u'integer'}]}\
```

What is the cause of your error?

A) You have used an invalid hyperparameter
B) You have used an invalid hyperparameter value
C) You have used an invalid train content_type
D) You have used an invalid objective

**Answer:** B

**Explanation:**

Option A is incorrect. If you had specified an invalid hyperparameter you would get an error such as:

```
ERROR 139623806805824 train.py:48]
Additional properties are not allowed (u'min_child_weigh' was
unexpected)
```

Option B is correct. You specified the value of 2 for the eta hyperparameter, but the valid range for this hyperparameter for the XGBoost algorithm is float range: [0,1]

Option C is incorrect. The valid content types for the XGBoost algorithm are text/libsvm (default) or text/csv. You have used text/csv, so your content type is valid.

Option D is incorrect. The objective multi:softmax is a valid setting for the XGBoost algorithm.

**Reference:**
Please see the Amazon SageMaker developer guide titled Logs for Built-in Algorithms, the Amazon SageMaker developer guide titled XGBoost Hyperparameters, and the XGBoost Parameters GitHub page (especially the Learning Task Parameters section)

Question: 104
**Main Topic :** Machine Learning
**Sub Topic :** Sanitize and prepare data for modeling

**Domain:** Exploratory Data Analysis

**Question text**:
You work for a retail clothing manufacturer that has a very active online web store. You have been assigned the task of building a model to contact customers for a direct marketing campaign based on their predicted receptiveness to the campaign. Some of your customers have been contacted in the past for other marketing campaigns. You don't want to contact these customers who have been contacted in the past for this latest campaign.

Before training this model, you need to clean your data and prepare it for the XGBoost algorithm you are going to use. You have written your cleaning/preparation code in your SageMaker notebook. Based on the following code, what happens on lines 19, 21, 22? (Select THREE)

```
1 import sagemaker
2 import boto3
3 from sagemaker.predictor import csv_serializer

4 import numpy as np
5 import pandas as pd
6 from time import gmtime, strftime
7 import os
```

```
8 region = boto3.Session().region_name
9 smclient = boto3.Session().client('sagemaker')

10 from sagemaker import get_execution_role
11 role = get_execution_role()

12 bucket = 'sagemakerS3Bucket'
13 prefix = 'sagemaker/xgboost'

14 !wget -N https://.../bank.zip
15 !unzip -o bank.zip
16 data = pd.read_csv('./bank/bank-full.csv', sep=';')
17 pd.set_option('display.max_columns', 500)
18 pd.set_option('display.max_rows', 5)

19 data['no_previous_campaign'] = np.where(data['contacted'] == 999, 1, 0)
20 data['not_employed'] = np.where(np.in1d(data['job'], ['student', 'retired', 'unempl']), 1, 0)
21 model_data = pd.get_dummies(data)
22 model_data = model_data.drop(['duration', 'employee.rate', 'construction.price.idex',
       'construction.confidence.idx','lifetime.rate', 'region'], axis=1)

23 train_data, validation_data, test_data = np.split(model_data.sample(frac=1,
       random_state=1729), [int(0.7 * len(model_data)), int(0.9*len(model_data))])

24 pd.concat([train_data['y_yes'], train_data.drop(['y_no', 'y_yes'], axis=1)],
       axis=1).to_csv('train.csv', index=False, header=False)
25 pd.concat([validation_data['y_yes'], validation_data.drop(['y_no', 'y_yes'], axis=1)],
       axis=1).to_csv('validation.csv', index=False, header=False)
26 pd.concat([test_data['y_yes'], test_data.drop(['y_no', 'y_yes'], axis=1)],
       axis=1).to_csv('test.csv', index=False, header=False)

27 boto3.Session().resource('s3').Bucket(bucket).Object(os.path.join(prefix,
       'train/train.csv')).upload_file('train.csv')
28 boto3.Session().resource('s3').Bucket(bucket).Object(os.path.join(prefix,
       'validation/validation.csv')).upload_file('validation.csv')
```

A) Splits bank dataset into train, validation, and test datasets
B) Sets the attribute no_previous_campaign to 999, 0, or 1 depending if the customer in the observation has been contacted via a previous campaign
C) Sets the attribute no_previous_campaign to 1 if the customer in the observation has not been contacted via a previous campaign or 0 if they have been contacted via a previous campaign

D) Converts categorical data to a set of indicator variables
E) Converts empty attributes to dummy variables
F) Removes features deemed inconsequential
G) Removes observations deemed inconsequential

**Answers:** C, D, F

**Explanation:**
Option A is incorrect. This option describes what happens on line 23, not what happens on lines 20, 21, or 22.

Option B is incorrect. Line 19 does not set the attribute no_previous_campaign to 999. It sets the attribute no_previous_campaign to 1 or 0 depending on whether the customer in the observation has been contacted via a previous campaign, as indicated by the value 999.

Option C is correct. Line 19 sets the attribute no_previous_campaign to 1 or 0 depending if the customer in the observation has been contacted via a previous campaign, as indicated by the value 999.

Option D is correct. Line 21 uses the pandas library get_dummies method to convert the categorical attributes in the dataframe to dummy (or indicator) variables.

Option E is incorrect. Line 21 does not convert empty attributes to dummy variables, it uses the pandas library get_dummies method to convert the categorical attributes in the dataframe to dummy (or indicator) variables.

Option F is correct. Line 22 removes (or drops) several features presumably because you have deemed the features inconsequential to the training of your model.

Option G is incorrect. Line 22, in this usage, calls the pandas drop method to remove features, not observations.

**Reference:**
Please see the SciPy numpy.where documentation (for line 19), the pandas get_dummies documentation (for line 21), and the pandas DataFrame.drop documentation (for line 22)

Question: 105
**Main Topic :** Machine Learning
**Sub Topic :** Perform feature engineering

**Domain:** Exploratory Data Analysis

**Question text**:

You work for the credit card division of a large financial services firm. You are a machine learning specialist working on a credit card transaction classification model. Your model will be used to classify your firm's customer transactions for use in direct marketing campaigns by your firm's marketing department. You have built your model based on the SageMaker pre-built Linear Learner algorithm. You have also deployed your model to an inference endpoint using an inference pipeline. You are performing your feature engineering via the SageMaker built-in feature transformers so you don't need to write your own feature engineering logic.

You have defined the containers for your pipeline using the CreateModel SageMaker API and you have created an inference endpoint using the SageMaker CreateEndpointConfig and CreateEndpoint APIs. You have decided to change your pipeline to use a different SageMaker feature transformer strategy (change the strategy from the default None to SingleRecord). How do you make this change to your inference pipeline?

A) Your pipeline model is mutable, meaning you can change it while it is running.
B) Your pipeline is immutable, but you can update your inference pipeline by deleting the old one and redeploying the new one using the SageMaker CreateEndpointConfig and CreateEndpoint APIs.
C) Your pipeline is immutable, but you can change your inference pipeline by deploying a new one using the ReplaceEndpoint API.
D) Your pipeline is immutable, but you can change your inference pipeline by deploying a new one using the UpdateEndpoint API.

**Answer:** D

**Explanation:**
Option A is incorrect. SageMaker inference pipelines are immutable, so you cannot change them while they are running.

Option B is incorrect. It is true that your inference pipeline is immutable, but you change it via the UpdateEndpoint API. You do not have to delete your pipeline and recreate it.

Option C is incorrect. Your inference pipeline is immutable, but you change it via the UpdateEndpoint API not a ReplaceEndpoint API. The is no ReplaceEndpoint API.

Option D is correct. Your inference pipeline is immutable. You change it by deploying a new one via the UpdateEndpoint API. SageMaker deploys the new inference pipeline, then switches incoming requests to the new one. SageMaker then deletes the resources associated with the old pipeline.

**Reference:**
Please see the Amazon SageMaker developer guide titled Deploy an Inference Pipeline, the Amazon SageMaker developer guide titled CreateModel, the Amazon SageMaker developer

guide titled UpdateModel, the Amazon SageMaker developer guide titled Use Amazon SageMaker Built-in Algorithms, and the SageMaker docs page titled Transformer

Question: 106
**Main Topic :** Machine Learning
**Sub Topic :** Identify and implement a data-transformation solution

**Domain:** Data Engineering

**Question text**:
You work for an online retailer as a machine learning specialist. Your team has been tasked with creating a machine learning model to identify similar products for a product comparison chart on many of the product pages on your website. Your website designers want to show a grid of a product compared to similar products. The grid will show price, review summary (stars), and key features of each product. You are at the stage in your development where you are gathering, cleaning, and transforming your data and training your model.

Using machine learning techniques, how can you determine similar product data for use in this grid in the most efficient manner?

A) Use the Linear Learner built-in SageMaker algorithm and set its predictor_type hyperparameter to binary_classifier
B) Use the XGBoost built-in SageMaker algorithm and set its objective hyperparameter to reg:logistic
C) Use the Linear Learner built-in SageMaker algorithm and set its predictor_type hyperparameter to regressor
D) Use the AWS Glue FindMatches ML Transform and set its precision-recall parameter to precision
E) Use the XGBoost built-in SageMaker algorithm and set its objective hyperparameter to reg:linear
F) Use the AWS Glue FindMatches ML Transform and set its precision-recall parameter to recall

**Answer:** D

**Explanation:**
Option A is incorrect. Using a Linear Learner algorithm based model with the binary_classifier predictor_type may help you find similar products, but it is not the most efficient technique listed in the options.

Option B is incorrect. Using a XGBoost algorithm based model with the reg:logistic objective may help you find similar products, but it is not the most efficient technique listed in the options.

Option C is incorrect. Using the Linear Learner algorithm with the regressor predictor_type would not be a good choice for a discrete categorization problem such as matching similar products.

Option D is correct. The AWS Glue FindMatches ML Transform uses machine learning capabilities to find matching records in your database, even when the records don't have exactly matching fields. This type of matching is perfect for finding similar products in a products table. Setting the FindMatches ML Transform precision_recall parameter to precision is the correct parameter setting. You use this setting when you want to minimize false positives. Meaning, you don't want to show two items as similar when they are not similar.

Option E is incorrect. Using the XGBoost algorithm with the reg:linear objective would not be a good choice for a discrete categorization problem such as matching similar products.

Option F is incorrect. The AWS Glue FindMatches ML Transform uses machine learning capabilities to find matching records in your database, even when the records don't have exactly matching fields. Setting the FindMatches ML Transform precision_recall parameter to recall is incorrect since this setting is used when you want to minimize false negatives. Meaning, the ML transform failed to find a match when a match actually existed. This is not an optimal result, but it is a better outcome than incorrectly identifying two items as similar when they really aren't (false positive).

**Reference:**
Please see the AWS Glue developer guide titled Machine Learning Transforms in AWS Glue, the AWS Glue developer guide titled Tuning Machine Learning Transforms in AWS Glue, and the Amazon SageMaker developer guide titled Use Amazon SageMaker Built-in Algorithms

Question: 107
**Main Topic :** Machine Learning
**Sub Topic :** Create data repositories for machine learning

**Domain:** Data Engineering

**Question text**:
You work for an online fashion retailer as a machine learning specialist. You are on a team of machine learning specialists and data scientists who have been given the responsibility of centralizing your company's product, customer, supplier, and materials data in one source. This new data source will be used for analytics and for making informed business decisions using KPIs (Key Performance Indicators). Your company has many different data sources where their product, customer, supplier, and materials data is stored. These data repositories are also housed on several different database technologies.

When you load the various data sources into your new centralized data source, you need to clean and classify the data as well. What is the most expeditious and efficient way to create this new centralized data source?

A) Use Amazon EMR and its built-in machine learning tool Apache Spark MLlib to extract the data from your disparate data sources, transform (clean and classify) the data, and load it into an S3 data lake.
B) Use AWS Glue crawlers to crawl your disparate data sources and create a metastore for your S3 data lake. Use AWS Glue to then extract, transform (clean and classify), and load the source data into your S3 data lake.
C) Use Amazon Kinesis Data Firehose to send the data from your disparate data sources to you S3 data lake. Use lambda integration with Kinesis Data Firehose to transform (clean and classify) your data as it loads into your S3 data lake.
D) Use AWS Lake Formation to collect and catalog the data from your disparate data sources, transform (clean and classify) your data, and load the data into your S3 data lake.

**Answer:** D

**Explanation:**
Option A is incorrect. Using Amazon EMR and its built-in machine learning tools will work to extract, transform, and load your disparate data sources into your S3 data lake, but it is not the quickest or simplest option given.

Option B is incorrect. Using AWS Glue and its crawlers will work to extract, transform, and load your disparate data sources into your S3 data lake, but it is not the quickest or simplest option given.

Option C is incorrect. Using Amazon Kinesis Data Firehoe and its lambda integration will work to extract, transform, and load your disparate data sources into your S3 data lake, but it is not the quickest or simplest option given.

Option D is correct. AWS Lake Formation builds on the capabilities of AWS Glue to simplify the creation of an S3 data lake. Once you define your disparate data sources to AWS Lake Formation, it crawls your data sources and moves the data into your S3 data lake. It uses machine learning algorithms to clean and classify your data. This is the simplest and most efficient option listed.

**Reference:**
Please see the AWS Lake Formation overview page, the Amazon EMR overview page, the AWS Big Data blog titled Build a Data Lake Foundation with AWS Glue and Amazon S3, and the Amazon Kinesis overview page

Question: 108
**Main Topic :** Machine Learning
**Sub Topic :** Identify and implement a data-transformation solution

**Domain:** Data Engineering

**Question text**:
You work for a major banking and financial services firm as a machine learning specialist. Your firm has decided to improve their fraud detection for specialized cases where fraudulent actors attempt to open accounts through your firm's banking and trading services. These services have websites where potential customers can open accounts by completing online forms. These services make use of your firm's highly secure customer and account data stores.

You have been assigned the task of determining when a known fraudulent actor attempts to open a new account. You have decided to build a machine learning solution to solve this problem. Since your firm has a very large customer base, several million customer accounts, you need to consider the performance as well as the precision of your fraud detection process.

You have decided to use the AWS Glue FindMatches ML Transform to process your online form data to find matching known fraudulent accounts in your firm's data stores. Knowing that detecting a fraudulent actor is of primary importance, how should you configure the AWS Glue FindMatches ML Transform parameters to achieve the most performant and accurate fraud detection process?

A) Set the FindMatches precision-recall parameter to 'precision' and the accuracy-cost parameter to 'accuracy'
B) Set the FindMatches precision-recall parameter to 'precision' and the accuracy-cost parameter to 'lower cost'
C) Set the FindMatches precision-recall parameter to 'recall' and the accuracy-cost parameter to 'accuracy'
D) Set the FindMatches precision-recall parameter to 'recall' and the accuracy-cost parameter to 'lower cost'

**Answer:** C

**Explanation:**
Option A is incorrect. Setting the FindMatches precision-recall parameter to 'precision' minimizes false positives (when you don't have a match of a fraudulent account but mark it as a match mistakenly). But you are more concerned about minimizing false negatives (when you have a match of a fraudulent account but fail to detect it).

Option B is incorrect. Setting the FindMatches precision-recall parameter to 'precision' minimizes false positives (when you don't have a match of a fraudulent account but mark it as a

match mistakenly). But you are more concerned about minimizing false negatives (when you have a match of a fraudulent account but fail to detect it).

Option C is correct. Setting the FindMatches precision-recall parameter to 'recall' minimizes false negatives (when you have a match of a fraudulent account but fail to detect it). This is what you want. Also, setting the FindMatches accuracy-cost parameter to 'accuracy' maximizes the transform accuracy of finding matching records as fraudulent.

Option D is incorrect. Setting the FindMatches precision-recall parameter to 'recall' minimizes false negatives (when you have a match of a fraudulent account but fail to detect it). This is what you want. But, setting the accuracy-cost parameter to 'lower cost' favors cost or the speed of running the transform at the expense of the transform's accuracy. This may make your transform more performant, but your primary concern is detecting a fraudulent actor so you should set the accuracy-cost parameter to 'accuracy'.

**Reference:**
Please see the AWS Glue developer guide titled Machine Learning Transforms in AWS Glue, and the AWS Glue developer guide titled Tuning Machine Learning Transforms in AWS Glue

Question: 109
**Main Topic :** Machine Learning
**Sub Topic :** Deploy and operationalize machine learning solutions

**Domain:** ML Implementation and Operations

**Question text**:
You work for a government census bureau in their machine learning group. Your team is working on a model that will be used to predict population movement based on many attributes of the population and the geographic regions in which they live and move to and from. Some of the dataset features are id, age, height, weight, family size, country of origin, etc. You have built your model using the SageMaker built-in linear learner algorithm. You have trained your model and deployed it using SageMaker Hosting Services. You are now ready to send inference requests to your inference endpoint. You have chosen to use CSV file data stored on one of your S3 buckets as your inference request data. Since you are processing large census data files you don't need sub-second latency.

Here is an example of the CSV file data:

| id | age | height (in.) | weight (lb) | family size | country of origin | ... |
|---|---|---|---|---|---|---|
| 6185 | 23 | 75 | 145 | 3 | USA | ... |
| 5437 | 54 | 80 | 187 | 7 | Canada | ... |

...

You know that the id attribute in your dataset is not relevant to your model's prediction results and you didn't use it when training your model. What is the simplest way you exclude this

attribute when you send prediction requests to your inference endpoint, but have the id attribute associated with the prediction results that your model outputs so you can easily analyze the prediction results?

A) Use SageMaker Batch Transform to run the predictions from your CSV file on your S3 bucket and have it exclude the id from the prediction request. Also have Batch Transform join the id attribute to the prediction results.
B) Use Kinesis Data Analytics to stream your prediction requests from your CSV file on your S3 bucket to your inference endpoint. Transform the prediction requests by removing the id attribute. Use Kinesis Data Analytics to join the id attribute to the prediction results.
C) Use Kinesis Data Analytics to stream your prediction requests from your CSV file on your S3 bucket to your inference endpoint. Transform the prediction requests by removing the id attribute. Use Kinesis Data Streams to join the id attribute to the prediction results.
D) Use Kinesis Data Firehose to run the predictions from your CSV file on your S3 bucket and have it exclude the id from the prediction request. Use Kinesis Data Streams to join the id attribute to the prediction results.

**Answer:** A

**Explanation:**
Option A is correct. The simplest way to first exclude the id attribute from the inference prediction requests and then join the id attribute to the prediction results is to use Amazon SageMaker Batch Transform.

Option B is incorrect. While you could use Kinesis DataAnalytics to exclude the id attribute from your prediction request and then to join the attribute with the prediction results, this would not be as simple a solution as just using Batch Transform pre and post processing.

Option C is incorrect. While you could use Kinesis Data Analytics to exclude the id attribute from your prediction requests and then use Kinesis Data Streams to join the id attribute to the prediction results possibly using a lambda function you would have to write, this approach would not be as simple as just using Batch Transform pre and post processing.

Option D is incorrect. You could use Kinesis Data Firehose Data Transformation to exclude your id attribute from your prediction requests and then use Kinesis Data Streams to join the id attribute to the prediction results possibly using a lambda function you would have to write, this approach would not be as simple as just using Batch Transform pre and post processing.

**Reference:**
Please see the AWS announcement titled SageMaker Batch Transform now enables associating prediction results with input attributes, the Amazon SageMaker developer guide titled Associate Prediction Results with Input Records, the Amazon SageMaker developer guide titled Deploy a Model on Amazon SageMaker Hosting Services, the AWS Lambda developer

guide titled Using AWS Lambda with Amazon Kinesis, and the Amazon Kinesis Data Firehose developer guide titled Amazon Kinesis Data Firehose Data Transformation

Question: 110
**Main Topic :** Machine Learning
**Sub Topic :** Train machine learning models

**Domain:** Modeling

**Question text**:
You work as a machine learning specialist for an auto manufacturer who produces several car models in several product lines. Example models include an LX model, an EX model, a Sport model, etc. These models have many similarities, but of course they also have defining differences. Each model has its own parts list entries in your company's parts database. When ordering commodity parts for these car models from auto parts manufacturers you want to produce the most efficient orders for each parts manufacturer by combining orders for similar parts lists. This will save your company money. You have decided to use the AWS Glue FindMatches Machine Learning Transform to find your matching parts lists.

You have created your data source file as a CSV, and you have also created your labeling file used to train your FindMatches transform. When you run your AWS Glue transform job it fails. Which of the following could be the root of the problem?

A) The labeling file is in the CSV format
B) The labeling file has labeling_set_id and label as its first two columns with the remaining columns matching the schema of the parts list data to be processed
C) Records in the labeling file that don't have any matches have unique labels
D) The labeling file is not encoded in UTF-8 without BOM (byte order mark)

**Answer:** D

**Explanation:**
Option A is incorrect. When using the AWS Glue FindMatches ML Transform, the labeling file must be in CSV format.

Option B is incorrect. When using the AWS Glue FindMatches ML Transform, the first two columns of the labeling file are required to be labeling_set_id and label. Also, the remaining columns must match the schema of the data to be processed.

Option C is correct. When using the AWS Glue FindMatches ML Transform, if a record doesn't have a match, it is assigned a unique label.

Option D is correct. When using the AWS Glue FindMatches ML Transform, the labeling file must be encoded as UTF-8 without BOM.

**Reference:**
Please see the AWS Glue developer guide titled Machine Learning Transforms in AWS Glue

Question: 111
**Main Topic :** Machine Learning
**Sub Topic :** Sanitize and prepare data for modeling

**Domain:** Exploratory Data Analysis

**Question text**:
You work as a machine learning specialist for an electric bicycle company. The electric bicycles your company produces have IoT sensors on them that transmit usage and maintenance information to your company data lake. You are using Kinesis Data Streams to gather the bicycle IoT data and store it into an S3 data store that you can use for your machine learning models. You are on the team that has the assignment of using the IoT data to predict when your customer's electric bicycles need maintenance.

The IoT data that the electric bicycles produce is unstructured, and sometimes, depending on the manufacturer of the IoT part, the data has a different schema structure. You need to clean and classify the IoT data before using it in your machine learning model. How can you build an ETL script to perform the necessary cleaning and classification knowing you have message data with differing schema structures?

A) Use AWS Glue to build a series of transforms that use Apache Spark SparkSQL DataRecord to pass the data from transform to transform. Each transform performing a different cleaning and/or transforming task.
B) Use AWS Glue to build a series of transforms that use Apache Spark SparkSQL DataFrames to pass the data from transform to transform. Each transform performing a different cleaning and/or transforming task.
C) Use AWS Glue to build a series of transforms that uses DynamicFrames to pass the data from transform to transform. Each transform performing a different cleaning and/or transforming task.
D) Use AWS Glue to build a series of transforms that uses DynamicRecord to pass the data from transform to transform. Each transform performing a different cleaning and/or transforming task.

**Answer:** C

**Explanation:**
Option A is incorrect. There is no DataRecord construct in Apache Spark SparkSQL.

Option B is incorrect. The Apache Spark SparkSQL DataFrame does not efficiently handle data with unknown schema structure. This option would produce suboptimal results.

Option C is correct. The AWS Glue DynamicFrame allows for each record to be self-describing so it can handle unknown or changing schemas.

Option D is incorrect. DynamicRecord represents a logical record within a DynamicFrame. It is a row in a DynamicFrame. So you wouldn't pass individual DynamicRecords from transform to transform, you pass a DynamicFrame.

**Reference:**
Please see the AWS Glue developer guide titled Machine Learning Transforms in AWS Glue, and the AWS Glue developer guide titled DynamicFrame Class

Question: 112
**Main Topic :** Machine Learning
**Sub Topic :** Identify and implement a data-ingestion solution

**Domain:** Data Engineering

**Question text**:
You work as a machine learning specialist for a book publishing company. Your company has several publishing data stores housed in relational databases across its infrastructure. Your company recently purchased another publishing company and are in the process of merging the two company's systems infrastructure. A part of this merger activity is joining the two publisher book databases. Your team has been given the assignment to build a data lake sourced from the two company's relational data stores.

How would you construct an ETL pipeline to achieve this goal? (Select FOUR)

A) Use AWS DataSync to ingest the relational data from your book data stores and store it in S3
B) Use an AWS Glue crawler to build your AWS Glue catalog
C) Have a lambda function triggered by an S3 trigger to start your AWS Glue crawler
D) Use an AWS Glue trigger to start your AWS Glue ETL job that processes/transforms your data and places it into your S3 data lake
E) Use a lambda function triggered by a CloudWatch event trigger to start your AWS Glue ETL job that processes/transforms your data and places it into your S3 data lake
F) Use AWS Database Migration Service to ingest the relational data from your book data stores and store it in S3

**Answers:** B, C, E, F

**Explanation:**
Option A is incorrect. AWS DataSync is used to ingest data from a Network File System (NFS), not relational databases.

Option B is correct. Once your data has been ingested from your databases, you need to catalog the data using an AWS Glue crawler.

Option C is correct. The AWS Glue crawler can be started by a lambda function that is triggered by an S3 object create event.

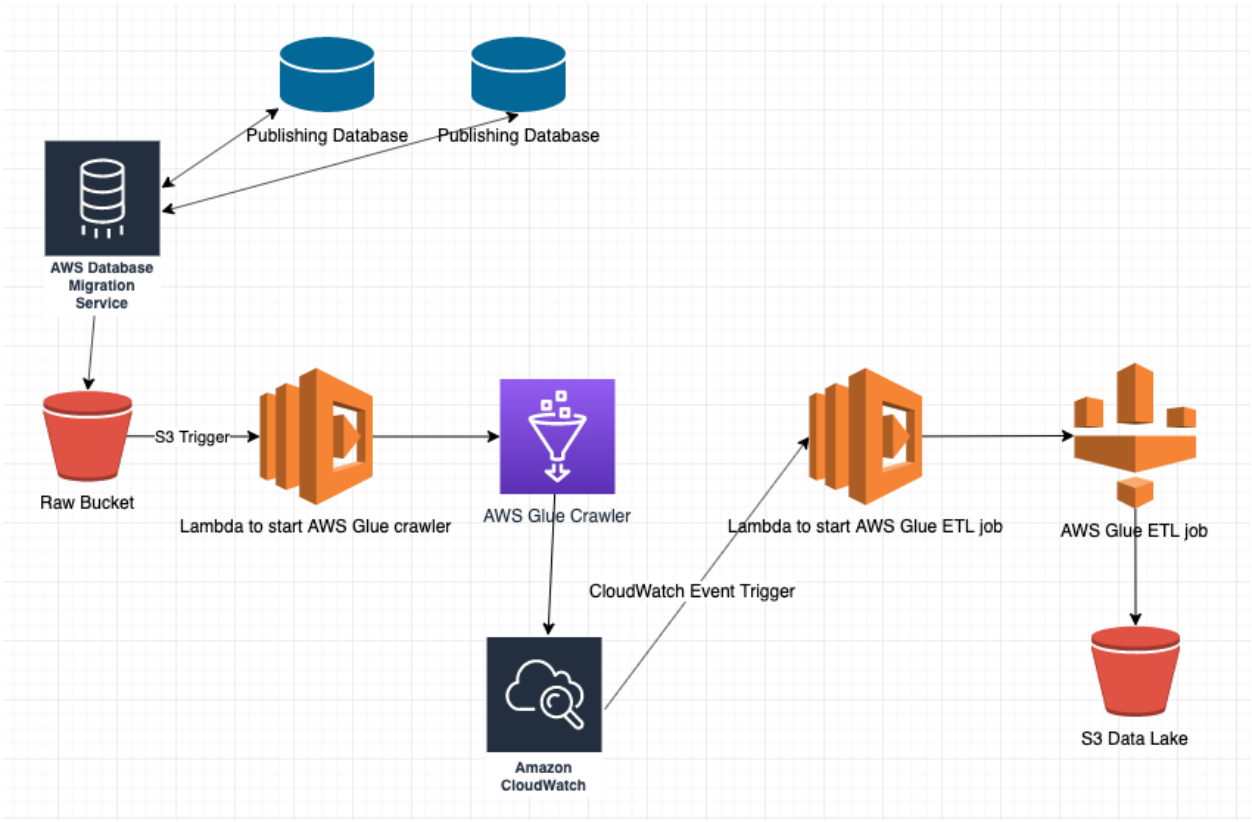
Option D is incorrect. It is not possible to start an AWS Glue ETL job from an AWS Glue trigger.

Option E is correct. You can have your AWS Glue ETL job started by a lambda function that is triggered by a CloudWatch event trigger.

Option F is correct. You can use the AWS Database Migration Service to ingest your data from your relational databases and then store the data in an S3 bucket.

**Reference:**
Please see the AWS Big Data blog titled Build and automate a serverless data lake using an AWS Glue trigger for the Data Catalog and ETL jobs, and the AWS article titled How can I automatically start an AWS Glue job when a crawler run completes?

Here is a diagram representing the proposed solution:




Question: 113
**Main Topic :** Machine Learning
**Sub Topic :** Deploy and operationalize machine learning solutions

**Domain:** ML Implementation and Operations

**Question text**:
You work for a sports wagering company as a machine learning specialist. Your team is responsible for building the machine learning models that produce the sports wager line for the NFL (National Football League) games each week. You are working on the line versus the spread model. For this model you have chosen the XGBoost algorithm. You have trained your model and deployed it to Amazon SageMaker Hosting Services where you are now ready to send inference requests to your model.

You are sending requests to your inference endpoint, but you are seeing that your inferences are failing. Which of these would NOT be the source of the problem? (Select TWO)

A) You have serialized your inference request in the text/csv format
B) You have serialized your inference request in the application/x-recordio-protobuf format
C) You have serialized your inference request in the text/libsvm format
D) You have serialized your inference request in the application/json format

**Answers:** A, C

**Explanation:**
Option A is correct. Inference endpoints built using the XGBoost algorithm only support the text/csv and text/libsvm request formats.

Option B is incorrect. Inference endpoints built using the XGBoost algorithm only support the text/csv and text.libsvm request formats. Your inference request will fail if you serialize your inference request using the application/x-recordio-protobuf format.

Option C is correct. Inference endpoints built using the XGBoost algorithm only support the text/csv and text/libsvm request formats.

Option D is incorrect. Inference endpoints built using the XGBoost algorithm only support the text/csv and text.libsvm request formats. Your inference request will fail if you serialize your inference request using the application/json format.

**Reference:**
Please see the Amazon SageMaker developer guide titled Deploy a Model on Amazon SageMaker Hosting Services, the Amazon SageMaker developer guide titled CreateEndpoint, and the Amazon SageMaker developer guide titled Common Data Formats for Inference

Question: 114
**Main Topic :** Machine Learning
**Sub Topic :** Train machine learning models

**Domain:** Modeling

**Question text**:
You work for a computer peripheral manufacturer that builds printers, external hard drives, etc. You are on the machine learning team where you are currently building a machine learning model to be used to find anomalies in the functional behavior of your company's line of printers. The printers generate IoT device messages that are streamed to your madel S3 bucket using Amazon Kinesis Data Streams. You have performed your data cleansing and data engineering of you IoT printer data. You are now ready to start training your model. You have chosen the Random Cut Forest SageMaker built-in algorithm for your model. You hope to find anomalies in your customer's printer activity by looking for outlier observations using your Random Cut Forest based model. Finding these anomalies will help your company provide better customer service.

You have started your first training job, but you see that your training job is failing. What may be the cause of this failure?

A) You have selected compute resources of the GPU compute instance class
B) You have selected compute resources of the CPU compute instance class
C) You have built your training data files using the CSV file type
D) You have built your training data files using the recordio-protobuf file type

**Answer:** A

**Explanation:**
Option A is correct. SageMaker only supports the CPU instance class for the Random Cut Forest algorithm.

Option B is incorrect. SageMaker only supports the CPU instance class for the Random Cut Forest algorithm. So selecting the instance class of CPU would not cause your training job to fail.

Option C is incorrect. SageMaker supports both the CSV and recordio-protobuf file types for training data files. So using the CSV file type for your training data would not cause your training job to fail.

Option D is incorrect. SageMaker supports both the CSV and recordio-protobuf file types for training data files. So using the recordio-protobuf file type for your training data would not cause your training job to fail.

**Reference:**
Please see the Amazon SageMaker developer guide titled Train a Model with Amazon SageMaker, the Amazon SageMaker developer guide titled Common Parameters for Built-In Algorithms

Question: 115
**Main Topic :** Machine Learning
**Sub Topic :** Evaluate machine learning models

**Domain:** Modeling

**Question text**:
You work for a power tool manufacturer as a machine learning specialist. You work in the battery powered power tool division where your team of machine learning specialists and data scientists have been tasked with building a model that predicts the lifespan of particular models

of power tools. You have selected the Linear Learner algorithm on which to build your model. You have cleaned and engineered your features for your training and test data. Your feature engineering transformations convert all feature attributes to integers or real numbers. You have also trained your model and have deployed it to Amazon SageMaker Hosting Services.

Your training dataset has this structure:

| model | power | battery Ah | use pattern | region | country |

For your client application inference requests, how would you structure the body argument for your invoke_endpoint call?

A) A string with this value: “547,3.5,1.5,23.4,2,43,1”
B) A string with this value: “547,3.5,1.5,23.4,2,43”
C) A string with this value: “Quite strike,battery,1.5,frequent,North America,US”
D) An array set to these values: [547,3.5,1.5,23.4,2,43]
E) A list set to these values: [547,3.5,1.5,23.4,2,43]

**Answer:** B

**Explanation:**
Option A is incorrect. The Linear Learner algorithm expects either CSV or recordio-protobuf as the inference request content type. For text/csv the value of the body argument for the invoke_endpoint API call should be a string with with comma separated values for each feature. This option has a comma separated string, but it has 7 values, when you only have 6 features in your data used to train your model.

Option B is correct. The Linear Learner algorithm expects either CSV or recordio-protobuf as the inference request content type. For text/csv the value of the body argument for the invoke_endpoint API call should be a string with with comma separated values for each feature. This option has a comma separated string and it has 6 values. You also have 6 features in your data used to train your model, so this inference request is structured correctly.

Option C is incorrect. The Linear Learner algorithm expects either CSV or recordio-protobuf as the inference request content type. Also, any transforms performed on the data for training also must be performed on inference request data before attempting to obtain an inference. The body argument in this option has not been transformed in the way your training data was transformed.

Option D is incorrect. The Linear Learner algorithm expects either CSV or recordio-protobuf as the inference request content type. Also, the body argument to the invoke_endpoint should be a string, not an array.

Option E is incorrect. The Linear Learner algorithm expects either CSV or recordio-protobuf as the inference request content type. Also, the body argument to the invoke_endpoint should be a string, not a list.

**Reference:**
Please see the Amazon SageMaker developer guide titled Train a Model with Amazon SageMaker, the Amazon SageMaker developer guide titled Common Parameters for Built-In Algorithms, and the Amazon SageMaker developer guide titled Common Data Formats for Inference

Question: 116
**Main Topic :** Machine Learning
**Sub Topic :** Evaluate machine learning models

**Domain:** Modeling

**Question text**:
You work for an Internet of Things (IoT) component manufacturer which builds servos, engines, sensors, etc. The IoT devices transmit usage and environment information back to AWS IoT Core via the MQTT protocol. You want to use a machine learning model to show how/where the use of your products is clustered in various regions around the world. This information will help your data scientists build KPI dashboards for use in improving your component engineering quality and performance. You have created, trained, and deployed to Amazon SageMaker Hosting Services your model based on the XGBoost algorithm. Your model is set up to receive inference requests from a lambda function that is triggered by the receipt of an IoT Core MQTT message via your Kinesis Data Streams instance.

What transform steps need to be done for each inference request. Also which steps are handled by your code versus by the inference algorithm? (Select TWO)

A) Inference request serialization (handled by the algorithm)
B) Inference request serialization (handled by your lambda code)
C) Inference request deserialization (handled by your lambda code)
D) Inference request deserialization (handled by the algorithm)
E) Inference request post serialization (handled by the algorithm)

**Answers:** B, D

**Explanation:**
Option A is incorrect. The inference request serialization must be completed by your lambda code. The algorithm needs to receive the inference request in serialized form.

Option B is correct. The inference request serialization must be completed by your lambda code.

Option C is incorrect. The inference request is deserialized by the algorithm in the response to the inference request. Your lambda code is responsible for serializing the inference request.

Option D is correct. The inference request is deserialized by the algorithm in the response to the inference request.

Option E is incorrect. There is no inference request post serialization step in the SageMaker inference request/response process.

**Reference:**
Please see the Amazon SageMaker developer guide titled Common Data Formats for Inference, the AWS IoT Core overview page, the AWS IoT developer guide titled Creating an AWS Lambda Rule

Question: 117
**Main Topic :** Machine Learning
**Sub Topic :** Recommend and implement the appropriate machine learning services and features for a given problem

**Domain:** ML Implementation and Operations

**Question text**:
You work for a farming equipment component manufacturer which builds farm product containers like corn silos, mike containers, etc. These containers have IoT sensors built into them that transmit information such as fill rate, capacity usage, etc. The IoT devices transmit their data back to your cloud environment via the MQTT protocol. You want to use a machine learning model to predict container usage buy region and by product stored. This information will help your management team use real-time dashboards to better understand their product marketing campaigns by region. You have created, trained, and deployed to Amazon SageMaker Hosting Services your model based on the Linear Learner algorithm.

What AWS services would you use to create your pipeline to feed your inference requests to you model? (Select THREE)

A) IoT Greengrass to receive the IoT device MQTT messages
B) IoT Core to receive the IoT device MQTT messages
C) Elastic Beanstalk to stream the IoT messages
D) Kinesis Data Streams to stream the IoT messages
E) A Lambda function to transform the IoT message data to the inference request serialization format

F) API Gateway to transform the IoT message data to the inference request serialization format
G) Route 53 to stream the IoT messages

**Answers:** B, D, E

**Explanation:**
Option A is incorrect. AWS IoT Greengrass is used to extend AWS to edge devices, such as your sensors in your farming containers. Greengrass is used to perform prediction directly on the devices themselves. IoT Core is a better option for receiving your MQTT IoT messages for processing via your machine learning inference running in Amazon SageMaker Hosting Services.

Option B is correct. IoT Core is designed to allow IoT devices interact with other AWS services, such as Kinesis Data Streams.

Option C is incorrect. Elastic Beanstalk is used to host web applications or worker nodes for web applications. You wouldn't use Elastic Beanstalk to stream IoT messages.

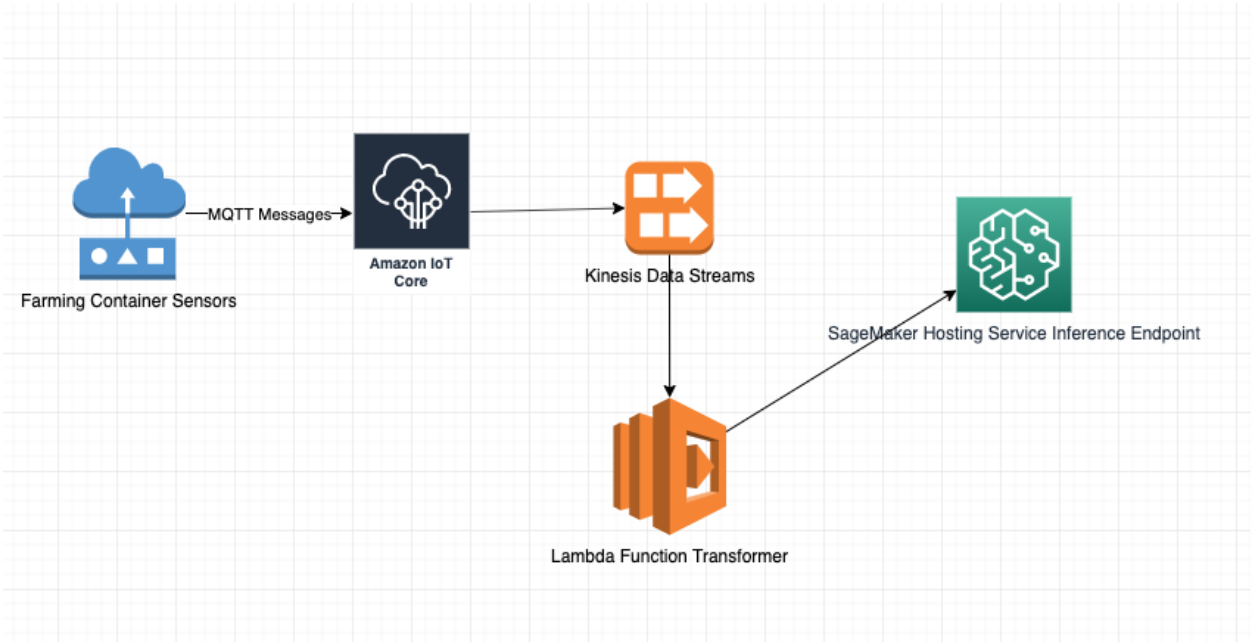
Option D is correct. Kinesis Data Streams can receive IoT messages from IoT Core and then stream to your SageMaker inference endpoint via a Lambda function.

Option E is correct. You can write a lambda function that is triggered by Kinesis Data Streams that transforms you IoT messages into the inference serialization format required by your inference endpoint.

Option F is incorrect. API Gateway is used to create an API request endpoint. You wouldn't ise API Gateway to transform IoT message data. You would have to have a lambda function behind your API Gateway to accomplish this.

Option G is incorrect. Route 53 is Amazon's DNS server implementation. You can't use Route 53 to transform messages.

**Reference:**
Please see the Amazon SageMaker developer guide titled Train a Model with Amazon SageMaker, the AWS IoT Core overview, the AWS IoT developer guide titled Creating an AWS Lambda Rule, and the AWS IoT Greengrass overview
Here is a diagram of the proposed solution:




Question: 118
**Main Topic :** Machine Learning
**Sub Topic :** Analyze and visualize data for machine learning

**Domain:** Exploratory Data Analysis

**Question text**:
You work for a management consulting firm as a machine learning specialist. You are on a team of data scientists and other machine learning specialists. Your team has been assigned the task of building a machine learning model to predict Return On Investment (ROI) for new potential engagements that your management consults may wish to take onto their book of business.

You have a dataset of past engagements that has many features that can help you define your problem as a machine learning problem. Before you decide on which machine learning algorithms to evaluate you wish to visualize the historical data to get an idea of the relationships between three of the key features of your dataset: ROI, investment time, and investment size.

Which type of visualization would best give you an idea of the relationship between these three features?

A) Pie chart
B) Tree map
C) Column histogram
D) Bar chart
E) Bubble chart
F) Line chart

**Answer:** E

**Explanation:**
Option A is incorrect. A pie chart is best used to show the portion of the total for each slice of the pie. This type of chart doesn't work well with three dimensions, such as ROI, investment time, and investment size.

Option B is incorrect. A tree map chart also shows the portion of the total. This type of chart is good for data with a long tail. But it also would not work well on three dimensions.

Option C is incorrect. Column histograms are distribution charts. They show how data is distributed of intervals. But you a looking for a visualization to show the relationship between three variables.

Option D is incorrect. A bar chart is a comparison chart. These types of charts are good for showing how feature values change over time or to show a static snapshot of how different variables compare with each other. But you are looking for the relationship between three variables, not change over time or a static snapshot comparison.

Option E is correct. A bubble chart is a relationship chart. For a relationship between two variables you could use a scatter chart. For the relationship between 3 variables, a bubble chart shows the relationships as such: x-axis for investment time, y-axis for ROI, and the bubble size for investment size.

Option F is incorrect. A line chart is used to show a comparison of variables changing over time. You are looking for a relationship between three variables, not how they change over time.

**Reference:**
Please see the AWS Data Visualization page, and the Amazon QuickSite overview page

Question: 119
**Main Topic :** Machine Learning
**Sub Topic :** Select the appropriate model(s) for a given machine learning problem

**Domain:** Modeling

**Question text**:
You work for a major retail chain in their web development area. You are on the machine learning team responsible for building a recommendation engine for the company's retail website where they sell many different items across many different categories. The recommendation engine will use customer data such as purchase history, credit rating, geographic location, household income, response to past marketing mailings, etc. Your marketing team has decided to send a marketing mailing to customers who have responded to

past mailings. They have two different content templates to use depending on the classification category of each customer. Your model needs to recommend which mailing template to use for each customer in the target customer dataset.

Which SageMaker built-in algorithm is best suited to this problem, and what value should you use for the predictor_type hyperparameter for the desired outcome? (Select TWO)

A) Linear Learner
B) classifier
C) regressor
D) Factorization Machine
E) multiclass_classifier
F) K-Means
G) binary_classifier
H) Neural Topic Model

**Answers:** D, G

**Explanation:**
Option A is incorrect. The Linear Learner is best suited for discrete classification problems. But you have already classified your customers, you are now trying to provide a discrete recommendation. The Factorization Machine algorithm is better suited for this type of problem.

Option B is incorrect. The classifier predictor_type hyperparameter value is not a valid choice for the Factorization Machine algorithm. The classifier predictor_type hyper parameter value is a valid choice for the K-Nearest-Neighbor algorithm.

Option C is incorrect. The regressor predictor_type hyperparameter value setting is used for regression type problems and therefore is not the correct choice for this type of problem. The regressor predictor_type hyperparameter setting is used when you are solving for a quantitative value. You are trying to solve for a discrete value.

Option D is correct. The Factorization Machine algorithm is a good choice for problems where you are trying to solve for a discrete recommendation.

Option E is incorrect. The multiclass_classifier predictor_type hyperparameter value is not a valid choice for the Factorization Machine algorithm. The multiclass_classifier predictor_type hyperparameter value is a valid choice for the Linear Learner algorithm.

Option F is incorrect. The K-Means algorithm is best used for grouping observations. You are trying to solve a discrete recommendation problem, you are not trying to group customers.

Option G is correct. The binary_classifier predictor_type hyperparameter value is the correct choice for this discrete recommendation problem where you are attempting to choose one of two possible outcomes (one of the two content templates).

Option H is incorrect. The Neural Topic Model algorithm is best suited to organizing documents into topics using groupings of words based on their statistical distribution within the documents. This algorithm is not a good choice for a discrete recommendation problem.

**Reference:**
Please see the Amazon SageMaker developer guide titled Use Amazon SageMaker Built-in Algorithms, and the AWS Machine Learning blog titled Build a movie recommender with factorization machines on Amazon SageMaker

Question: 120
**Main Topic :** Machine Learning
**Sub Topic :** Analyze and visualize data for machine learning

**Domain:** Exploratory Data Analysis

**Question text**:
You work for the city planning department of a major metropolitan city in the United States. You are on the city's machine learning team where you are responsible for creating a model that assists in the resource planning for police officers in the city. Each day the city has to assign police officers to each precinct according to varying parameters. You have data from the past several years for your city and other US cities of a similar makeup. You are in the process of deciding which algorithm to use for your police officer allocation model. Your goal is to predict the police officer allocation size for a given shift based on your dataset features.

Your city dataset has the following features:
1) Infrastructure average age
2) Square feet
3) Citizens
4) Precincts
5) Residences
6) Population density
7) Police officers

Before you select an algorithm you need to perform feature selection and dimensionality reduction of your features. You only want to select features that are relevant to your training dataset, i.e. dimensionality reduction. This process will help you prevent overfitting, and increase computation efficiency through simplification of the feature set.

You have chosen to use visualization techniques to decide which of your 7 features are the most important, or most relevant. In other words, which of your 7 features are needed to properly train your model.

Which visualization techniques are the best to use for this purpose? (Choose TWO)

A) Cat plot
B) Swarm plot
C) Pairs plot
D) Covariance matrix
E) Entropy matrix

**Answers:** C, D

**Explanation:**
Option A is incorrect. A catplot is used to show the relationship between a numerical value and one or more categorical variables using a visualization such as violinplot, boxenplot, etc. But you are trying to show relationships between pairs of data, such as police officers to population density, or police officers to precincts.

Option B is incorrect. A swarm plot is used to show categorical scatter plot data that shows the distribution of values for each feature. But you are trying to show relationships between pairs of data, such as police officers to population density, or police officers to precincts.

Option C is correct. A pairs plot is used to show the relationship between pairs of features as well as the distribution of one of the variables in relation to the other. This is what you need to analyze. You want to see which features correlate well with your police officers feature.

Option D is correct. A covariance matrix shows the degree of correlation between two features. This visualization gives you a numerical representation of the correlation, where the pairs plot gives you a visual representation as points plotted in two dimensional space.

Option E is incorrect. Entropy represents the measure of randomness in your features. This measure would not help you find the correlation between your target feature, police officers, and the potential training features.

**Reference:**
Please see the article titled Feature Selection and Dimensionality Reduction Using Covariance Matrix Plot, the article titled Visualizing Data with Pairs Plots in Python, and the article titled What is Entropy and why Information gain matter in Decision Trees?

Question: 121

**Main Topic :** Machine Learning
**Sub Topic :** Frame business problems as machine learning problems

**Domain:** Modeling

**Question text**:
You work for a firm that produces cameras that can be used for research studies of animals in the wild. When placed in the wild, these cameras are used to identify individual animals and groups of animals as they pass in front of the camera. Researchers use your company's cameras to catalog animal traffic and specific animal counts in geographic areas where these animals are suspected to live. An example is the identification and counting of wolves in Canada and the far reaches of North America.

Using your company's cameras, you and your team of machine learning specialists have been contracted by the Wolf Conservation Center of North America to build a machine learning model to identify and count a specific species of wolf in remote areas of the Arctic Circle.

What type of machine learning problem are you trying to solve?

A) Linear regression
B) Binary classification
C) Multidimensional regression
D) Multiclass classification

**Answer:** D

**Explanation:**
Option A is incorrect. A linear regression is used to model the relationship between a dependent variable and one or more independent variables. For example: what will the sales in the north american region be when the GDP (Gross Domestic Product) is trending up and interest rates are trending down. You are trying to solve a classification problem with images as your inference data.

Option B is incorrect. A binary classification is used to classify an observation into one of two categories. For example: based on the image data, is the animal in the image a wolf or not a wolf. You are trying to solve a multiclass classification problem, what type of animal is in the image? You are looking for a specific species of wolf.

Option C is incorrect. A multidimensional regression is used to find more than one real number values. For example: what is the height and width of the animal in the image? You are trying to solve a multiclass classification problem: what type of animal is in the image? You are looking for a specific species of wolf.

Option D is correct. A multiclass classification solves a classification problem where you have more than one class for your answer. For example: of all the animals identified in a given region, what type of animal is in the image? This is the type of problem you are trying to solve. Of all the types of wolves identified to live in the Arctic Circle, what specific species of wolf is in the image?

**Reference:**
Please see the Amazon Machine Learning developer guide titled Formulating the Problem, and the article titled Frame a problem as a machine learning problem or otherwise

Question: 122
**Main Topic :** Machine Learning
**Sub Topic :** Select the appropriate model(s) for a given machine learning problem

**Domain:** Modeling

**Question text**:
You work for a manufacturing firm that is attempting to build a rechargeable battery that has capacity multiple times greater than the current rechargeable batteries on the market. As a machine learning specialist on the team responsible for building a machine learning model that can predict the chemical component interaction that maximizes battery capacity, you have decided that none of the built-in algorithms available in SageMaker fit your problem as well as you would like. So you and your team have decided to create your own SageMaker algorithm resource. You'll use this custom algorithm to train and run inferences on your model.

Which of the following steps do you NOT need to complete to create your custom algorithm for use in SageMaker?

A) Create Docker containers for your training and inference code
B) Specify the hyperparameters that your algorithm supports
C) Specify the metrics that your algorithm sends to CloudWatch when training
D) The instance types your algorithm supports for training and inference
E) Whether your algorithm supports distributed inference across multiple instances

**Answer:** E

**Explanation:**
Option A is incorrect. SageMaker uses Docker containers for your custom algorithm training and hosting your algorithm.

Option B is incorrect. When you create a custom algorithm resource in SageMaker you need to specify the hyperparameters your algorithm will support.

Option C is incorrect. When you create a custom algorithm resource in SageMaker you need to specify the metrics that your algorithm will send to CloudWatch when running your training jobs.

Option D is incorrect. When you create a custom algorithm resource in SageMaker you need to specify the EC2 instance types your algorithm supports for training and inference.

Option E is correct. When you create a custom algorithm resource in SageMaker you need to specify whether it supports distributed training across multiple instances, not distributed inference.

**Reference:**
Please see the Amazon SageMaker developer guide titled Use Your Own Algorithms or Models with Amazon SageMaker, and the Amazon SageMaker developer guide titled Create an Algorithm Resource

Question: 123
**Main Topic :** Machine Learning
**Sub Topic :** Perform Feature Engineering

**Domain:** Exploratory Data Analysis

**Question text**:
You work for a company that manufactures cell phone peripherals such as bluetooth headphones and bluetooth selfie sticks. Your company has designed their products so that they act as IoT devices that send usage and diagnostic MQTT messages to your IoT Core service running in AWS. Your machine learning team wants to use this IoT message data to run inferences through your machine learning inference end-point. However, the IoT data is unstructured so you need to preprocess the data by performing feature engineering on the observations before they are fed into your inference endpoint.

You have decided to use a SageMaker Inference Pipeline to construct this machine learning solution. As you are defining the containers for your pipeline, one for feature engineering preprocessing, and one for inference predictions, which SageMaker CLI command and which parameter on that command do you need to run using the SageMaker CLI in order to build your inference pipeline?

A) CreateModel command with the EndpointArn request parameter
B) UpdateEndpoint command with the Containers parameter
C) CreateModel command with the PrimaryContainer request parameter
D) CreateModel command with the Containers request parameter
E) UpdateEndpoint command with the ModelArn parameter

**Answer:** D

**Explanation:**
Option A is incorrect. The SageMaker CLI CreateModel command is the correct command but EndpointArn is a response element of the UpdateEndpoint command.

Option B is incorrect. The SageMaker CLI UpdateEndpoint command is used to switch from an existing endpoint to a new endpoint. You would not use this command to create a new inference pipeline container.

Option C is incorrect. The SageMaker CLI CreateModel command is the correct command but you use thePrimaryContainer request parameter when you want to create a single container, not when you want to create an inference pipeline.

Option D is correct. The SageMaker CLI CreateModel command is the correct command and the Containers parameter is the correct parameter. The Containers request parameter is used to set the containers that make up your pipeline.

Option E is incorrect. The SageMaker CLI UpdateEndpoint command is used to switch from an existing endpoint to a new endpoint. You would not use this command to create a new inference pipeline container.

**Reference:**
Please see the Amazon SageMaker developer guide titled Deploy an Inference Pipeline, the Amazon SageMaker developer guide titled CreateModel, the Amazon SageMaker developer guide titled UpdateEndpoint, the AWS CLI Command Reference titled create-model, and the AWS IoT Core overview page

Question: 124
**Main Topic :** Machine Learning
**Sub Topic :** Perform Feature Engineering

**Domain:** Exploratory Data Analysis

**Question text**:
You work for a large manufacturer of consumer electronic devices. Your company wishes to build a machine learning model to predict which product has the most dedicated following among its consumer base. This product will receive funding for future investment in new models and/or enhancements to existing models. You and your machine learning team have a vast amount of observations of the use of the current product base. You know you and your team need to perform feature engineering on the large dataset before you use it to train your XGBoost algorithm based model for predictions.

What SageMaker feature can you use to perform the required feature engineering of your dataset in the most efficient way?

A) Automatic Model Tuning
B) Built-In Transforms
C) Batch Transform
D) Hosting Services

**Answer:** C

**Explanation:**
Option A is incorrect. The SageMaker Automatic Model Tuning feature is used to automatically adjusting thousands of different combinations of hyperparameters to give you the most accurate predictions for your model. But you are trying to perform feature engineering transformation prior to training, so this option is not correct.

Option B is incorrect. The Built-In Transforms feature is part of the AWS Glue service, not SageMaker.

Option C is correct. The SageMaker Batch Transform feature can be used to preprocess your data before using the data in your training runs.

Option D is incorrect. The SageMaker Hosting Services feature is used to allow your model to provide inferences once you've trained your model. But you are trying to perform feature engineering transformation prior to training, so this option is not correct.

**Reference:**
Please see the Amazon SageMaker developer guide titled Run Batch Transforms with Inference Pipelines, the Amazon SageMaker developer guide titled Get Inferences for an Entire Dataset with Batch Transform, the Amazon SageMaker Features overview page, the Amazon SageMaker developer guide titled Deploy a Model on Amazon SageMaker Hosting Services, and the AWS Glue developer guide titled Built-In Transforms

Question: 125
**Main Topic :** Machine Learning
**Sub Topic :** Evaluate Machine Learning Models

**Domain:** Modeling

**Question text**:

You work for a real estate ecommerce company. Your machine learning team is building a house price prediction model to be used on your company's site. This model will be used as a guide to users as an unbiased objective estimate of a given house's value. Your company has gathered an enormous dataset of house observations from across the United States. The observations in the dataset are categorized by region of the country. The housing data prices are mainly clustered by region across the dataset. However, each region has several outlier priced houses.

Since you have defined the housing price prediction work as a regression problem, you have selected the XGBoost SageMaker built-in algorithm on which to base your model. You are now ready to do your hyperparameter tuning so you need a good regression evaluation metric. Which of the following evaluation metrics best fit your problem?

A) MSE (Mean Squared Error)
B) AUC (Area Under the Curve)
C) ROC curve (Receiver Operating Characteristic) curve
D) MAE (Mean Absolute Error)

**Answer:** D

**Explanation:**
Option A is incorrect. The MSE metric is useful for measuring regression problems, however it does not handle outliers as well as the MAE metric. Your dataset has several outliers per region.

Option B is incorrect. The AUC metric is best used for classification type machine learning algorithms. You are using a regression algorithm.

Option C is incorrect. The AUC metric is best used for classification type machine learning algorithms. You are using a regression algorithm.

Option D is correct. The MAE is the correct regression metric to use when your dataset can be significantly influenced by outliers. Your dataset contains several outliers per region.

**Reference:**
Please see the article titled 20 Popular Machine Learning Metrics. Part 1: Classification & Regression Evaluation Metrics, the Amazon SageMaker developer guide titled XGBoost Algorithm, and the Amazon SageMaker developer guide titled Tune an XGBoost Model

Question: 126
**Main Topic :** Machine Learning
**Sub Topic :** Create data repositories for machine learning

**Domain:** Data Engineering

**Question text**:
You work for a startup ecommerce site that sells various consumer products. Your company has just launched their ecommerce website. The site provides the capability for your users to rate their purchases and the products they have purchased from your ecommerce site. You would like to use the review data to build a recommender machine learning model.

Since your ecommerce site is very new, you don't yet have a very large review dataset to use for your recommendation model. You have decided to use the Amazon Customer Reviews dataset available from the AWS website as a first data source for your machine learning model. Since your website sells similar products to the products sold on Amazon, you will use the Amazon Customer Reviews dataset as the basis for your initial training runs of your model. Once you have enough data from your own ecommerce site you'll use that data.

Your goal is to perform sentiment analysis on the review dataset to create your own dataset that will be the source used for your recommender machine learning model. Which set of AWS services would you use to build your data pipeline to produce your sentiment dataset for use by your SageMaker model?

A) S3 -> AWS Glue ETL -> Comprehend -> S3 -> SageMaker
B) S3 -> AWS Glue ETL -> Comprehend -> S3 -> Athena -> QuickSite -> SageMaker
C) S3 -> Kinesis Data Firehose -> Comprehend -> S3 -> SageMaker
D) S3 -> Kinesis Data Firehose -> Lambda -> S3 -> SageMaker

**Answer:** A

**Explanation:**
Option A is correct. The Amazon Customer Reviews dataset is stored on S3. You can use an AWS Glue ETL job to read the reviews from the Amazon dataset. The ETL job calls Comprehend for each review to get the sentiment for that review. The ETL job stores the sentiment enriched review data onto another S3 bucket in your account. Your SageMaker model uses the S3 bucket in your account as its dataset source for training your recommender model.

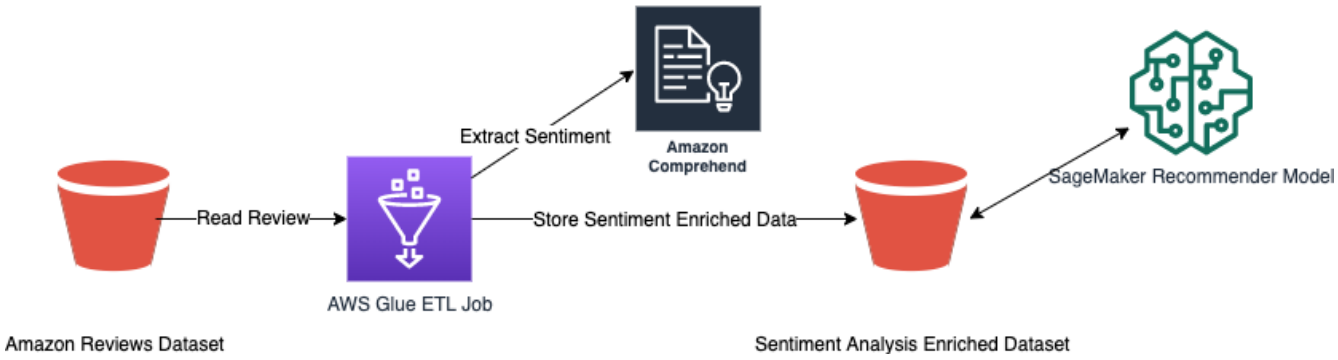
Option B is incorrect. This option has unnecessary steps. Specifically, you don't need Athena and QuickSite to produce your sentiment enriched dataset for your machine learning model.

Option C is incorrect. The option uses Kinesis Data Firehose unnecessarily. The Amazon Customer Reviews dataset is stored on S3, there is no need to stream the data when you can simply read it using an ETL job. Also, if you used Kinesis Data Firehose to stream the data you would have to write a lambda function to call Comprehend for each streamed review data row.

Option D is incorrect. The option uses Kinesis Data Firehose unnecessarily. The Amazon Customer Reviews dataset is stored on S3, there is no need to stream the data when you can simply read it using an ETL job. That being said, this option does correctly combine Kinesis Data Firehose and lambda. However it lacks the Comprehend service. You would have to write your own sentiment analysis in your lambda function.

**Reference:**
Please see the data repository titled Registry of Open Data on AWS, the AWS Machine Learning blog titled How to scale sentiment analysis using Amazon Comprehend, AWS Glue and Amazon Athena, and the data set titled Amazon Customer Reviews Dataset

Here is a diagram of the proposed solution:




Question: 127
**Main Topic :** Machine Learning
**Sub Topic :** Sanitize and prepare data for modeling

**Domain:** Exploratory Data Analysis

**Question text**:
You work for a retail athletic footwear company. Your company has just completed production of a new running shoe that contains IoT sensors in the shoe. These sensors are used to enhance the runner's running experience by giving detailed data about foot plant, distance, acceleration, gait, and other data points for use in personal running performance analysis.

You are on the machine learning team assigned the task of building a machine learning model to use the shoe IoT sensor data to make predictions of shoe life expectancy based on user wear and tear of the shoes. Instead of just using raw running miles as the predictor of shoe life, your model will use all of the IoT sensor data to produce a much more accurate prediction of remaining life of the shoes.

You are in the process of building your dataset for training your model and running inferences from your model. You need to clean the IoT sensor data before you use it for training or use it to provide inferences from your inference endpoint. You have decided to use Spark ML jobs within AWS Glue to build your feature transformation code. Which machine learning packages are the best choices for building your IoT sensor data transformer tasks in the simplest way possible? (Select THREE)

A) MLeap
B) MLib
C) SparkML Serving Container
D) SparkML Batch Transform
E) MLTransform
F) SparkML MapReduce

**Answers:** A, B, C

**Explanation:**
Option A is correct. AWS Glue serializes Spark ML jobs into MLeap containers. You add these MLeap containers to your inference pipeline.

Option B is correct. Apache Spark MLib is a machine learning library that lets you build machine learning pipeline components where you can transform your data using the full suite of standard transformers such as tokenizers, OneHotEncoders, normalizers, etc.

Option C is correct. The SparkML Serving Container allows you to deploy an Apache Spark ML pipeline in SageMaker.

Option D is incorrect. Batch Transformer is a feature of SageMaker that allows you to get inferences for an entire dataset. Batch Transform is not an Apache SparkML feature.

Option E is incorrect. There is no Apache SparkML feature called MLTransform.

Option F is incorrect. There is no Apache SparkML feature called MapReduce.

**Reference:**
Please see the Amazon SageMaker developer guide titled Feature Processing with Spark ML and Scikit-learn, the MLeap documentation, the SageMaker SparkML Serving Container GitHub repo, the Apache Spark MLib overview page, the Apache Spark MLib docs page titled Extracting, transforming, and selecting features, the Amazon SageMaker developer guide titled Deploy a Model on Amazon SageMaker Hosting Services, and the Amazon SageMaker developer guide titled Get Inferences for an Entire Dataset with Batch Transform

Question: 128
**Main Topic :** Machine Learning
**Sub Topic :** Frame business problems as machine learning problems

**Domain:** Modeling

**Question text**:
You work for a robotics company that is building a new product that allows commuters to ride electric skateboards to work. These skateboards are equipped with IoT sensors for safety measures. The sensors detect obstacles in the path of the skateboard and alert the rider with haptics and sound. The onboard software also uses the IoT sensor data to adjust the skateboard's performance based on its surroundings. This allows the rider who follows similar paths to work on their daily commute to have their skateboard become more adept at handling the surroundings commonly encountered on this path.

Which type of machine learning model would you use to build the onboard software for these commuter skateboards?

A) Unsupervised Learning model
B) Supervised Learning model
C) Reinforcement Learning model
D) Semi-Supervised Learning model

**Answer:** C

**Explanation:**
Option A is incorrect. Unsupervised learning is used to find patterns in your training dataset when you don't have preexisting labels. It is self-organizing. This type of model is not the best choice for learning an environment through exploration, which is what you are trying to do using your skateboard IoT sensor data. The better choice is Reinforcement Learning.

Option B is incorrect. Supervised learning is used when you have a training dataset that is labeled. In your skateboard learning example, you don't have any labels for your IoT sensor observations. Therefore, you could not use supervised learning for this type of problem.

Option C is correct. Reinforcement learning is used when you want to find the best way to achieve a goal or improve performance of a task. Your IoT sensor driven model is trying to improve the performance of the task of alerting for safety hazards as you ride the board through your daily commute environment.

Option D is incorrect. Semi-Supervised learning is used when you have a dataset with both labeled and unlabeled data from which to train your model. In the IoT sensor driven environment exploration use case, you will not have labeled data since you will discover new observations as

your IoT sensor equipped skateboard moves through its environment. Therefore, this type of machine learning model is not the best choice for this type of problem.

**Reference:**
Please see the Amazon SageMaker developer guide titled Reinforcement Learning with Amazon SageMaker RL, and the NVIDIA blog titled SuperVize Me: What's the Difference Between Supervised, Unsupervised, Semi-Supervised and Reinforcement Learning?