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: D

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 correct. 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 correct. 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: B

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.

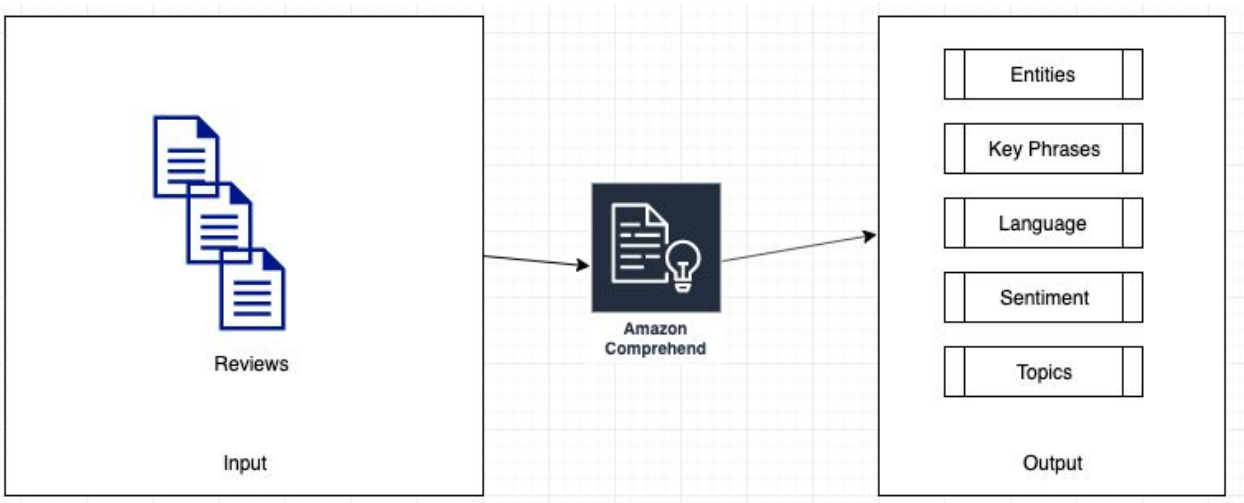
Option B is incorrect. The Neural Topic Model algorithm is used to group documents into topics using the statistical distribution of words within the documents. This algorithm would not be the most efficient choice for detecting offensive or unsafe language.

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 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 any 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 then 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