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 : **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: 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 : **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: 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 the Data Recipes for feature transformation in their machine learning process. When using the JSON-like syntax to create your data recipes, what is the order of the sections in the specification?

A) Assignments; Outputs; Groups
B) Outputs; Assignments; Groups
C) Assignments; Groups; Outputs
D) Groups; Assignments; Outputs

Answer: D

Explanation:

Option A is incorrect. The proper order is Groups; Assignments; Outputs

Option B is incorrect. The proper order is Groups; Assignments; Outputs

Option C is incorrect. The proper order is Groups; Assignments; Outputs

Option D is correct. From the Amazon Machine Learning developer guide titled Recipe Format Reference "Recipes have the following sections, which must appear in the order shown here: Groups, Assignments, Outputs"

Reference:
Please see the Amazon Machine Learning developer guide titled Feature Transformations with Data Recipes

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 and Amazon ML. Specifically, you and your team have decided to use the built-in Data Recipes to transform your data.

Which of the sections of the JSON-like format of the Data Recipes specification defines which variables you'll use for learning and the transformations you want to apply to them?

A) The Assignments section
B) The Outputs section
C) The Groups section
D) The Transformations section

Answer: B

Explanation:

Option A is incorrect. The Assignments section is used to create named intermediate variables that you can reuse in processing.

Option B is correct. The Outputs section is used to define which variables you'll use for learning and the transformations you want to apply to them.

Option C is incorrect. The Groups section allows you to group multiple variables so you can easily apply transformations to them.

Option D is incorrect. There is no Transformations section in the Data Recipes specification.

Reference:
Please see the Amazon Machine Learning developer guide titled Feature Transformations with Data Recipes, and the Amazon Machine Learning developer guide titled Recipe Format Reference