Free Test

Completed on 28-January-2021

Attempt	Marks Obtained	Your score	Time Taken	Result
01	0 / 15	0.0%	00 H 00 M 12 S	Failed

Domains wise Quiz Performance Report

No	Domain	Total Question	Correct	Incorrect	Unattempted	Marked as Review
1	Exploratory Data Analysis	5	0	1	4	0
2	ML Implementation and Operations	3	0	0	3	0
3	Modeling	4	0	0	4	0
4	Data Engineering	3	0	0	3	0
Total	All Domain	15	0	1	14	0

Review the Answers

Sorting by All

Question 1 **Unattempted**

Domain :Exploratory Data Analysis

You work for a financial services firm that wishes to further enhance their fraud detection capabilities. The firm has implemented fine grained transaction logging for all transactions their customers make using their credit cards. The fraud prevention department would like to use this data to produce dashboards to give them insight into their customer's transaction activity and to provide real-time fraud prediction.
You plan to build a fraud detection model using the transaction observation data with Amazon

SageMaker. Each transaction observation has a date-time stamp. In its raw form, the date-time stamp is not very useful in your prediction model since it is unique. Can you make use of the date-time stamp in your fraud prediction model, and if so how?

A. **No, you cannot use the date-time stamp since this data point will never occur again. Unique features like this will not help identify patterns in your data.**

B. **Yes, you can use the date-time stamp data point. You can just use feature selection to deselect the date-time stamp data point, thus dropping it from the learning process.**

C. **Yes, you can use the date-time stamp data point. You can transform the date-time stamp into features for the hour of the day, the day of the week, and the month.** ✓

D. **No you cannot use the date-time feature since there is no way to transform it into a unique data point.**

Explanation:

Answer: C

Option A is incorrect since you can use the date-time stamp if you use feature engineering to transform the data point into useful form.

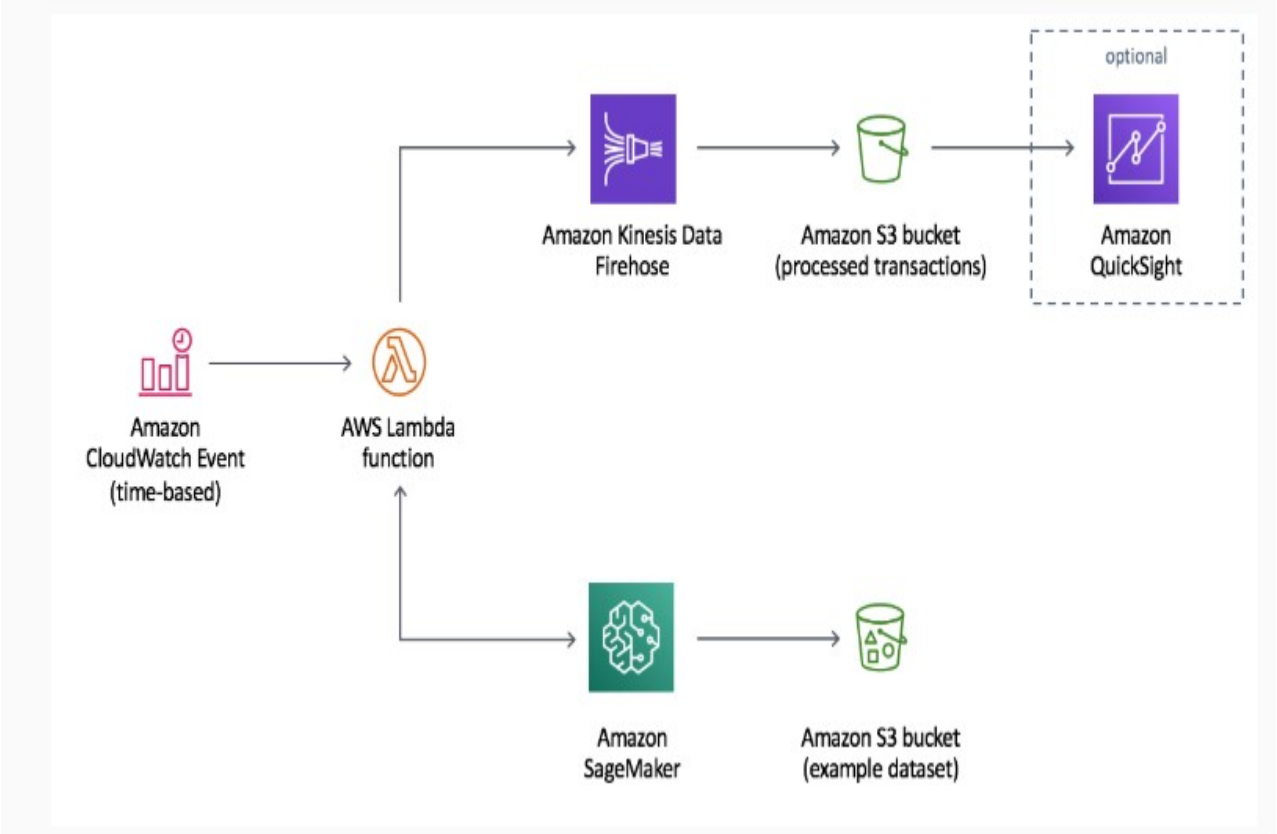
Option B is incorrect since this option is really just another way of ignoring, thus not using, the date-time stamp data point.

Option C is correct. You can transform the data point using feature engineering and thus gain value from it for the learning process of your model. (See the AWS Machine Learning blog post: **Simplify machine learning with XGBoost and Amazon SageMaker:** https://aws.amazon.com/blogs/machine-learning/simplify-machine-learning-with-xgboost-and-amazon-sagemaker/**)**

Option D is incorrect since we can transform the data point into unique features that represent the hour of the day, the day of the week, and the month, these variables could be useful to learn if the fraudulent activity tends to happen at a particular hour, day of the week, or month.

Diagram:

Here is a screen shot from the AWS Machine Learning documentation depicting a typical fraud detection machine learning solution:



Reference:

Please see the Amazon Machine Learning developer documentation:

https://docs.aws.amazon.com/machine-learning/latest/dg/feature-processing.html.

Ask our Experts Rate this Question?

View Queries open

Question 2 **Unattempted**

Domain :ML Implementation and Operations

You are deploying your data streaming pipeline for your machine learning environment. Your cloud formation stack has a Kinesis Data Firehose using the Data Transformation feature where you have configured Firehose to write to your S3 data lake. When you stream data through your Kinesis Firehose you notice that no data is arriving on your S3 bucket. What might be the problem that is causing the failure?

A. **Your lambda memory setting is set to the maximum value allowed**

B. **Your S3 bucket is in the same region as your Kinesis Data Firehose**

C. Your Kinesis Data Firehose buffer setting is set to the default value

D. Your lambda timeout value is set to the default value

Explanation:

Answer: D

Option A is incorrect. The maximum memory setting for lambda is 3 MB. Using the maximum memory would not cause Firehose to fail to write to S3. It will increase the cost of your solution however, since per the AWS documentation "Lambda allocates CPU power linearly in proportion to the amount of memory configured."

Option B is incorrect. Your S3 bucket used by Kinesis Data Firehose to output your data must be in the same region as your Firehose. Since they are in the same region, this would not cause a failure to write to the S3 bucket.

Option C is incorrect. The Kinesis Data Firehose documentation states that "Kinesis Data Firehose buffers incoming data before delivering it to Amazon S3. You can choose a buffer size (1–128 MBs) or buffer interval (60–900 seconds). The condition that is satisfied first triggers data delivery to Amazon S3." Using the default setting would not prevent Firehose from writing to S3.

Option D is correct. The lambda timeout value default is 3 seconds. For many Kinesis Data Firehose implementations, 3 seconds is not enough time to execute the transformation function.

Reference:

Please see the Amazon Kinesis Data Firehose developer guide documentation titled Configure Settings, the Amazon Kinesis Data Firehose developer guide documentation titled Amazon Kinesis Data Firehose Data Transformation, and the AWS Lambda developer guide documentation titled AWS Lambda Function Configuration.

Ask our Experts Rate this Question?

View Queries open

Question 3 **Unattempted**

Domain :Exploratory Data Analysis

You work as a machine learning specialist for a consulting firm where you are analyzing data about the consultants who work there in preparation for using the data in you machine

learning models. The features you have in your data are things like employee id, specialty, practice, job description, billing hours, and principle. The principle attribute is represented as 'yes' or 'no', whether the consultant has made principle level or not. For your initial analysis you need to identify the distribution of consultants and their billing hours for the given period. What visualization best describes this relationship?

A. **Scatter plot**

B. **Histogram** 

C. **Line chart**

D. **Box plot**

E. **Bubble chart**

Explanation:

Answer: B

Options A is incorrect. You are looking for a distribution on a single dimension: the consultants billing hours. From the Amazon QuickSite User Guide titled Working with Visual Types in Amazon QuickSight "A scatter chart shows a multiple distribution, i.e. two or three measures for a dimension."

Option B is correct. You are looking for a distribution of a single dimension: the consultants billing hours. From the wikipedia article titled Histogram "A histogram is an accurate representation of the distribution of numerical data. It is an estimate of the probability distribution of a continuous variable." The continuous variable in this question: the billing hours, binned into ranges (x axis), at a frequency: the number of consultants at a billing hour range (y axis).

Option C is incorrect. From the Amazon QuickSite User Guide titled Working with Visual Types in Amazon QuickSight "Use line charts to compare changes in measured values over a period of time." You are looking for a distribution not a comparison of changes over a period of time.

Option D is incorrect. From the Statistics How To article titled Types of Graphs Used in Math and Statistics "A boxplot, also called a box and whisker plot, is a way to show the spread and centers of a data set. Measures of spread include the interquartile range and the mean of the data set. Measures of center include the mean or average and median (the middle of a data set)." A Box Plot shows the distribution of multiple dimensions of the data. Once again, you are looking for a distribution of a single dimension, not a distribution on multiple dimensions.

Option E is incorrect. From the wikipedia article titled Bubble Chart "A bubble chart is a type of chart that displays three dimensions of data. Each entity with its triplet (*v1*, *v2*, *v3*) of associated data is plotted as a disk that expresses two of the *vi* values through the disk's *xy*

location and the third through its size." Once again, you are looking for a distribution of a single dimension, not a distribution on three dimensions.

Reference:

Please see the Amazon QuickSight user guide titled Working with Amazon QuickSight Visuals, and the Statistics How To article titled Types of Graphs Used in Math and Statistics

Ask our Experts Rate this Question?

View Queries open

Question 4 **Unattempted**

Domain :Modeling

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 (miles per gallon), 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** ✓

Explanation:

Answer: D

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)

Ask our Experts Rate this Question?

View Queries open

Question 5 **Unattempted**

Domain :Modeling

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

Explanation:

Answer: C

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

Ask our Experts Rate this Question?

View Queries open

Question 6 **Unattempted**

Domain :Data Engineering

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

Explanation:

Answer: D

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

Ask our Experts

Rate this Question?

View Queries open

Question 7 **Unattempted**

Domain :Exploratory Data Analysis

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 and performed feature engineering on 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. **Shuffle your data using a shuffling technique** ✓

Explanation:

Answer: D

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 and doing feature engineering of your data, which you have already done.

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

Option D is correct. For a linear regression model, once you have cleaned and engineered your data you need to shuffle 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

Ask our Experts Rate this Question? ☺ ☹

View Queries open

Question 8 **Unattempted**

Domain :Exploratory Data Analysis

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**

Explanation:

Answer: C

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

Ask our Experts Rate this Question?

View Queries open

Question 9 **Unattempted**

Domain :ML Implementation and Operations

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 Amazon 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

Explanation:

Answer: D

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.

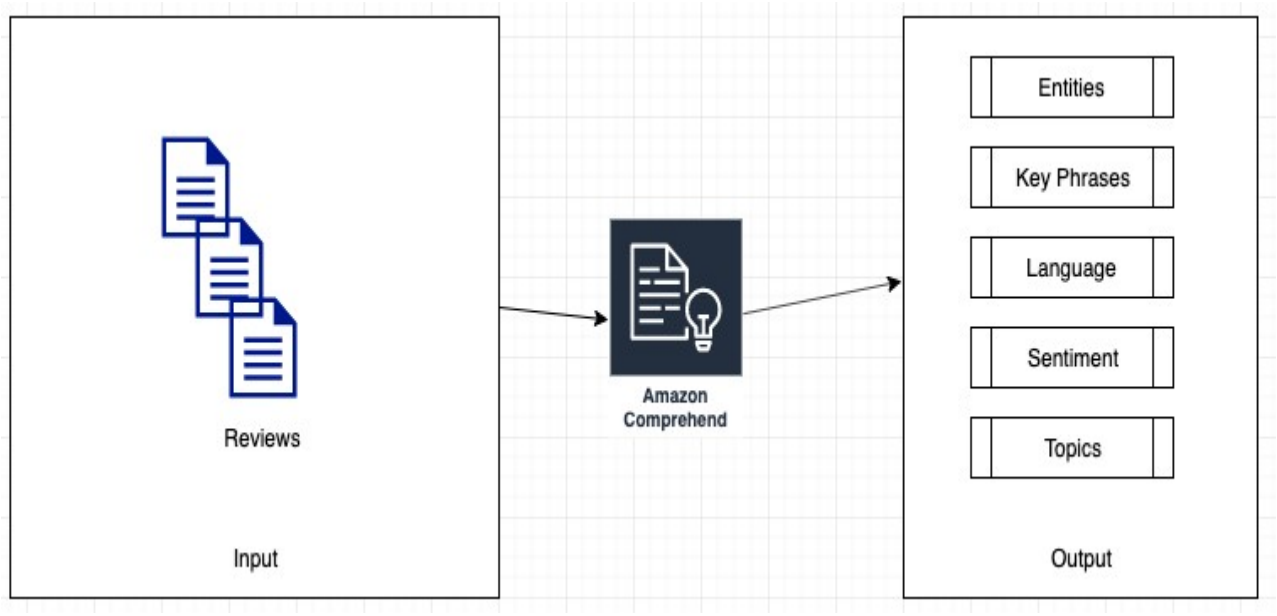
Option D is correct. The Comprehend service scans your unstructured review text and

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:



Ask our Experts

Rate this Question?

View Queries

open

Question 10

Unattempted

Domain :Modeling

You work for a software company that produces an online sports betting app. You are on the machine learning team responsible for building a model that predicts the likelihood of registered users to wager on a given event based on several features of sports events offered in the app. You and your team have selected the Linear Learner algorithm and have trained your model. You now wish to find the best set of hyperparameters for your model. You have chosen to use SageMaker's automatic model tuning and you have set your objective to validation:precision in your hyperparameter tuning job.
How do pass your tuning job settings into your hyperparameter tuning job? (Select THREE)

A. Define a JSON object and pass it as the value of the HyperParameterConfig to the HyperParameterTuningJob

B. Define a JSON object and pass it as the value of the HyperParameterTuningJobConfig to the CreateHyperParameterTuningJob ✓

C. In the JSON object specify the ranges of the hyperparameters you want to tune ✓

D. In the JSON object specify the limits of the hyperparameters you want to tune

E. In the JSON object specify the objective metric for the hyperparameter tuning job ✓

F. In the JSON object specify the MaxSequentialTrainingJobs parameter in the ResourceLimits section

Explanation:

Answers: B, C, E

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

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

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

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

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

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

Reference:

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

Ask our Experts Rate this Question?

View Queries open

Question 11 **Unattempted**

Domain :Data Engineering

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

A. Kinesis Data Analytics and its Hotspots function

B. Kinesis Data Analytics and its Random Cut Forest function ✓

C. Kinesis Data Firehose and its Hotspots function

D. Kinesis Data Streams and its Random Cut Forest function

E. Kinesis Data Streams and its Hotspots function

F. Kinesis Data Firehose and its Random Cut Forest function

Explanation:

Answer: B

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

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

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

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

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

Random_Cut_Forest.

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

Reference:

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

Ask our Experts Rate this Question?

View Queries open

Question 12 **Unattempted**

Domain :Data Engineering

You work for a startup ecommerce site that sells various consumer products. Your company has just launched their ecommerce website. The site provides the capability for your users to rate their purchases and the products they have purchased from your ecommerce site. You would like to use the review data to build a recommender machine learning model.
Since your ecommerce site is very new, you don't yet have a very large review dataset to use for your recommendation model. You have decided to use the Amazon Customer Reviews dataset available from the AWS website as a first data source for your machine learning model. Since your website sells similar products to the products sold on Amazon, you will use the Amazon Customer Reviews dataset as the basis for your initial training runs of your model. Once you have enough data from your own ecommerce site you'll use that data.
Your goal is to perform sentiment analysis on the review dataset to create your own dataset that will be the source used for your recommender machine learning model. Which set of AWS services would you use to build your data pipeline to produce your sentiment dataset for use by your SageMaker model?

A. S3 -> AWS Glue ETL -> Comprehend -> S3 -> SageMaker

B. S3 -> AWS Glue ETL -> Comprehend -> S3 -> Athena -> QuickSite -> SageMaker

C. S3 -> Kinesis Data Firehose -> Comprehend -> S3 -> SageMaker

D. S3 -> Kinesis Data Firehose -> Lambda -> S3 -> SageMaker

Explanation:

Answer: A

Option A is correct. The Amazon Customer Reviews dataset is stored on S3. You can use an AWS Glue ETL job to read the reviews from the Amazon dataset. The ETL job calls Comprehend for each review to get the sentiment for that review. The ETL job stores the sentiment enriched review data onto another S3 bucket in your account. Your SageMaker model uses the S3 bucket in your account as its dataset source for training your recommender model.

Option B is incorrect. This option has unnecessary steps. Specifically, you don't need Athena and QuickSite to produce your sentiment enriched dataset for your machine learning model.

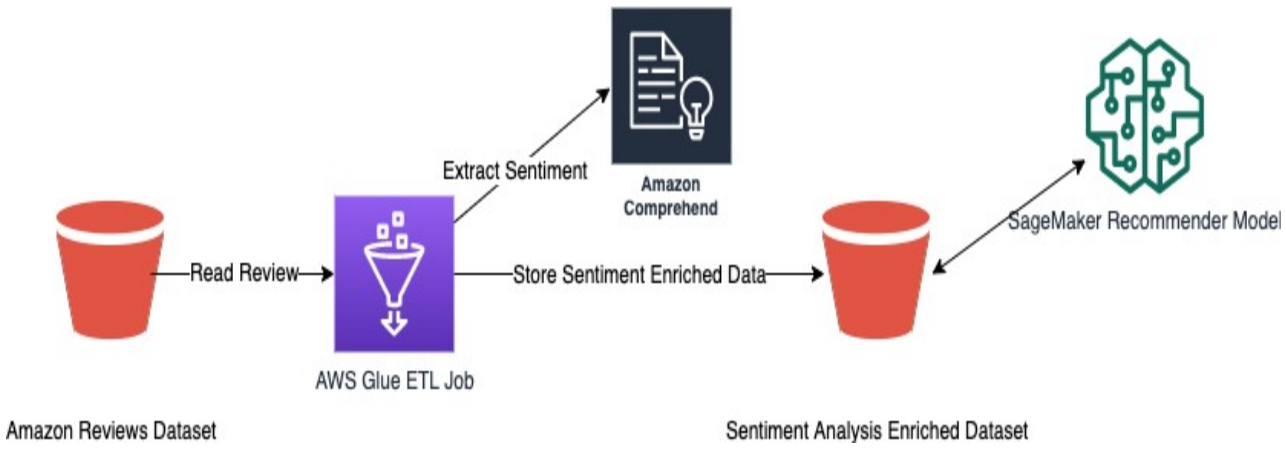
Option C is incorrect. The option uses Kinesis Data Firehose unnecessarily. The Amazon Customer Reviews dataset is stored on S3, there is no need to stream the data when you can simply read it using an ETL job. Also, if you used Kinesis Data Firehose to stream the data you would have to write a lambda function to call Comprehend for each streamed review data row.

Option D is incorrect. The option uses Kinesis Data Firehose unnecessarily. The Amazon Customer Reviews dataset is stored on S3, there is no need to stream the data when you can simply read it using an ETL job. That being said, this option does correctly combine Kinesis Data Firehose and lambda. However it lacks the Comprehend service. You would have to write your own sentiment analysis in your lambda function.

Reference:

Please see the data repository titled Registry of Open Data on AWS, the AWS Machine Learning blog titled How to scale sentiment analysis using Amazon Comprehend, AWS Glue and Amazon Athena, and the data set titled Amazon Customer Reviews Dataset

Here is a diagram of the proposed solution:



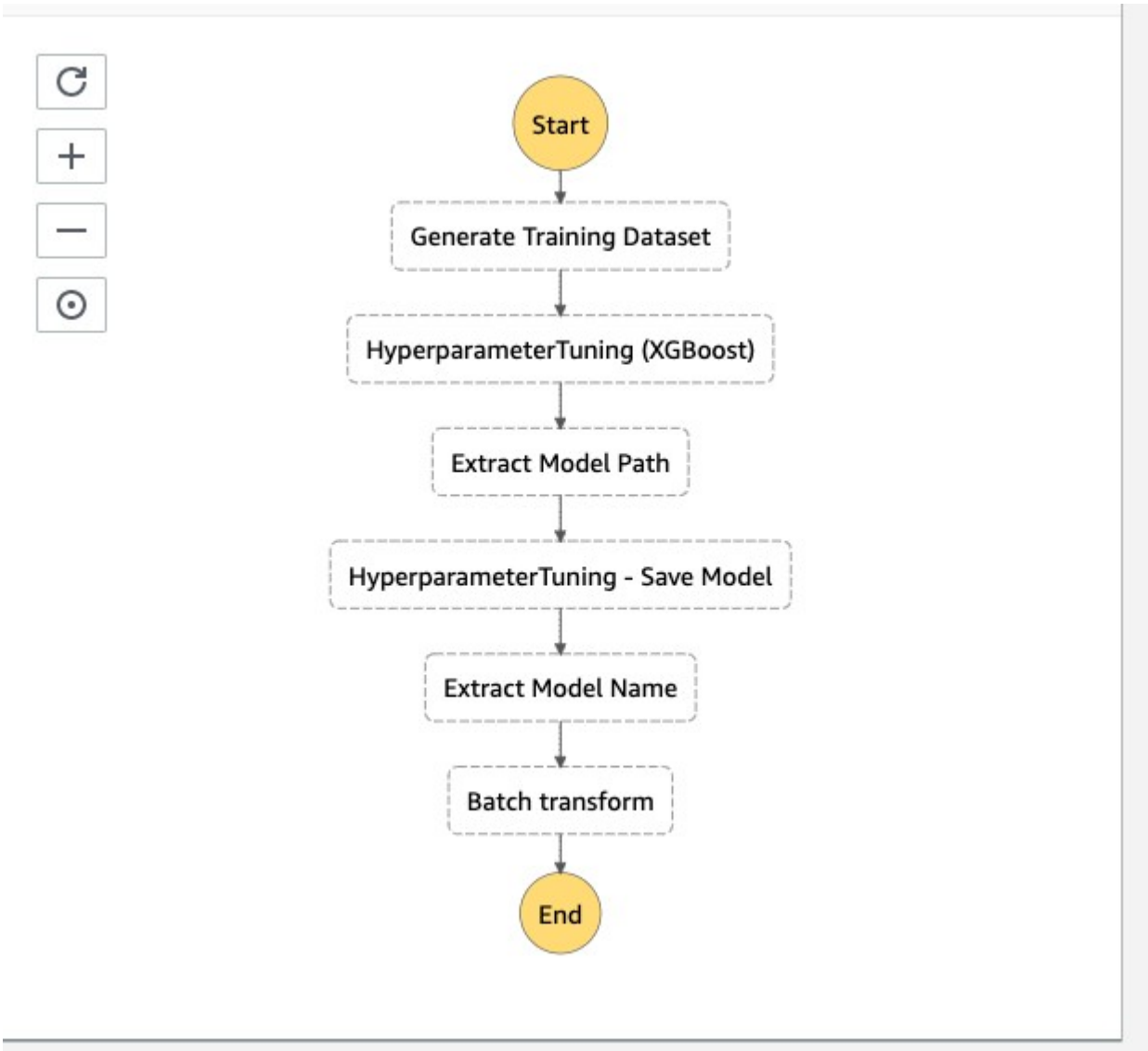
Ask our Experts Rate this Question?

View Queries open

Question 13 **Unattempted**

Domain :ML Implementation and Operations

You work for a mining company in their machine learning department. You and your team are working on a model to predict the minimum depth at which to drill in order to find various mineral deposits. You are building a model based on the XGBoost algorithm and your team is at the stage where you are running various models based on different hyperparameters in order to find the best hyperparameter settings. Because of the complexity of the problem you may have to run hundreds, or even thousands of hyperparameter tuning jobs to get the best result.
Your machine learning pipeline also includes a batch transform step to be executed after every hyperparameter tuning job. Your team lead has suggested that you use the Amazon Step Functions SageMaker integration capability to automate the execution of your many hyperparameter tuning jobs. You have setup your Step Functions environment and you have configured it as such:



You have written the following JSON-based Amazon States Language (ASL) for your State Machine (partial listing):

```
{
   "StartAt": "Generate Training Dataset",
   "States": {
      "Generate Training Dataset": {
         "Resource": "<GENERATE_LAMBDA_FUNCTION_ARN>",
         "Type": "Task",
         "Next": "HyperparameterTuning (XGBoost)"
      },
      "HyperparameterTuning (XGBoost)": {
         "Resource": "arn:<PARTITION>:states:::sagemaker:createHyperParameterTuningJob.sync",
         "Parameters": {
            "HyperParameterTuningJobName.$": "<JOB_NAME_FROM_LAMBDA>",
            "HyperParameterTuningJobConfig": {
               "Strategy": "Bayesian",
               "HyperParameterTuningJobObjective": {
                  "Type": "Minimize",
                  "MetricName": "validation:rmse"
               },
               "ResourceLimits": {
```

```
        "MaxNumberOfTrainingJobs": 2,
        "MaxParallelTrainingJobs": 2
    },
    "ParameterRanges": {
        "ContinuousParameterRanges": [{
                "Name": "alpha",
                "MinValue": "0",
                "MaxValue": "1000",
                "ScalingType": "Auto"
            },
            {
                "Name": "gamma",
                "MinValue": "0",
                "MaxValue": "5",
                "ScalingType": "Auto"
            }
        ],
        ...
```

Based on your Step Functions code, what is the type of metric you are using for your regression evaluation? Additionally, in the HyperparameterTuning (XGBoost) step, what happens when the alpha parameter increases through its range of 0 to 1,000? (Select TWO)

A. **Relative Mean Square Error**

B. **Gamma**

C. **Alpha**

D. **Root Mean Square Error** ✓

E. **Mean Square Error**

F. **As alpha increases, the model becomes more conservative** ✓

G. **As alpha increases, the model becomes less conservative**

H. **As alpha increases the model gains precision but sacrifices accuracy**

Explanation:

Answers: D, F

Option A is incorrect. The rmse metric acronym stands for Root Mean Square Error, not Relative Mean Square Error.

Option B is incorrect. The Gamma parameter defines the minimum loss reduction used to partition leaf nodes of the tree within the algorithm. This parameter is not used as a

regression evaluation objective.

Option C is incorrect. The Alpha parameter defines regularization terms on weights within the algorithm. This parameter is not used as a regression evaluation objective.

Option D is correct. Your code specifies the rmse metric as the objective on which to evaluate the tuning model run. The rmse acronym stands for Root Mean Square Error.

Option E is incorrect. Your code specifies the rmse metric as the objective on which to evaluate the tuning model run. The rmse acronym stands for Root Mean Square Error, not Mean Square Error.

Option F is correct. As the value of the alpha parameter increases, makes the model more conservative.

Option G is incorrect. As the value of the alpha parameter increases, makes the model more conservative, not less conservative.

Option H is incorrect. As the value of the alpha parameter increases, makes the model more conservative, it does not make the model gain precision while sacrificing accuracy.

Reference:

Please see the Amazon announcement titled Amazon SageMaker Announces New Machine Learning capabilities for Orchestration, Experimentation and Collaboration, the AWS Step Functions developer guide titled Manage Amazon SageMaker with Step Functions, the Amazon SageMaker developer guide titled Tune an XGBoost Model, and the XGBoost docs page titled XGBoost Parameters

Ask our Experts Rate this Question?

View Queries open

Question 14 **Unattempted**

Domain :Modeling

You work for a firm that produces cameras that can be used for research studies of animals in the wild. When placed in the wild, these cameras are used to identify individual animals and groups of animals as they pass in front of the camera. Researchers use your company's cameras to catalog animal traffic and specific animal counts in geographic areas where these animals are suspected to live. An example is the identification and counting of wolves in Canada and the far reaches of North America.
Using your company's cameras, you and your team of machine learning specialists have been contracted by the Wolf Conservation Center of North America to build a machine learning model to identify and count a specific species of wolf in remote areas of the Arctic Circle.
What type of machine learning problem are you trying to solve?

A. Linear regression

B. Binary classification

C. Multidimensional regression

D. Multiclass classification

Explanation:

Answer: B

Option A is incorrect. A linear regression is used to model the relationship between a dependent variable and one or more independent variables. For example: what will the sales in the North American region be when the GDP (Gross Domestic Product) is trending up and interest rates are trending down. You are trying to solve a classification problem with images as your inference data.

Option B is correct. A binary classification is used to classify an observation into one of two categories.
For example: based on the image data, is the animal in the image the wolf species we are looking for or not. You are trying to solve a binary classification problem: is the animal in the image the species I'm looking for or not? You are looking for a specific species of wolf.

Option C is incorrect. A multidimensional regression is used to find more than one real number values. For example: what is the height and width of the animal in the image? You are trying to solve a multiclass classification problem: what type of animal is in the image? You are looking for a specific species of wolf.

Option D is incorrect. A multiclass classification solves a classification problem where you have more than one class for your answer. For example: of all the animals identified in a given region, what type of animal is in the image? Of all the types of wolves identified to live in the Arctic Circle, what specific species of wolf is in the image? The problem we're trying to solve is whether this is the specific wolf species we're looking for or not? We are looking for one species class, therefore we should use a binary classification algorithm.

Reference:

Please see the Amazon Machine Learning developer guide titled Formulating the Problem, and the article titled Frame a problem as a machine learning problem or otherwise

Ask our Experts Rate this Question?

View Queries open

Question 15 **Incorrect**

Domain :Exploratory Data Analysis

You work for a retail athletic footwear company. Your company has just completed production of a new running shoe that contains IoT sensors in the shoe. These sensors are used to enhance the runner's running experience by giving detailed data about foot plant, distance, acceleration, gait, and other data points for use in personal running performance analysis.
You are on the machine learning team assigned the task of building a machine learning model to use the shoe IoT sensor data to make predictions of shoe life expectancy based on user wear and tear of the shoes. Instead of just using raw running miles as the predictor of shoe life, your model will use all of the IoT sensor data to produce a much more accurate prediction of remaining life of the shoes.
You are in the process of building your dataset for training your model and running inferences from your model. You need to clean the IoT sensor data before you use it for training or use it to provide inferences from your inference endpoint. You have decided to use Spark ML jobs within AWS Glue to build your feature transformation code. Which machine learning packages are the best choices for building your IoT sensor data transformer tasks in the simplest way possible? (Select THREE)

A. MLeap
B. MLib
✓ C. SparkML Serving Container
D. SparkML Batch Transform
E. MLTransform
F. SparkML MapReduce

Explanation:

Answers: A, B, C

Option A is correct. AWS Glue serializes Spark ML jobs into MLeap containers. You add these MLeap containers to your inference pipeline.

Option B is correct. Apache Spark MLib is a machine learning library that lets you build machine learning pipeline components where you can transform your data using the full

suite of standard transformers such as tokenizers, OneHotEncoders, normalizers, etc.

Option C is correct. The SparkML Serving Container allows you to deploy an Apache Spark ML pipeline in SageMaker.

Option D is incorrect. Batch Transformer is a feature of SageMaker that allows you to get inferences for an entire dataset. Batch Transform is not an Apache SparkML feature.

Option E is incorrect. There is no Apache SparkML feature called MLTransform.

Option F is incorrect. There is no Apache SparkML feature called MapReduce.

Reference:

Please see the Amazon SageMaker developer guide titled Feature Processing with Spark ML and Scikit-learn, the MLeap documentation, the SageMaker SparkML Serving Container GitHub repo, the Apache Spark MLib overview page, the Apache Spark MLib docs page titled Extracting, transforming, and selecting features, the Amazon SageMaker developer guide titled Deploy a Model on Amazon SageMaker Hosting Services, and the Amazon SageMaker developer guide titled Get Inferences for an Entire Dataset with Batch Transform

Ask our Experts

Rate this Question?

View Queries open

Finish Review