# Microsoft.DP-100.v2022-03-16.q210

| | |
|---|---|
| **Exam Code:** | DP-100 |
| **Exam Name:** | Designing and Implementing a Data Science Solution on Azure |
| **Certification Provider:** | Microsoft |
| **Free Question Number:** | 210 |
| **Version:** | v2022-03-16 |
| **# of views:** | 4265 |
| **# of Questions views:** | 2100 |
| https://www.freepdfdumps.com/Microsoft.DP-100.v2022-03-16.q210.html | |

**NEW QUESTION: 1**
You need to define a process for penalty event detection.
Which three actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.
**Answer:**

**NEW QUESTION: 2**
You train a machine learning model.
You must deploy the model as a real-time inference service for testing. The service requires low CPU utilization and less than 48 MB of RAM. The compute target for the deployed service must initialize automatically while minimizing cost and administrative overhead.
Which compute target should you use?
**A.** Azure Kubernetes Service (AKS) inference cluster
**B.** Azure Machine Learning compute cluster
**C.** Azure Container Instance (ACI)
**D.** attached Azure Databricks cluster
**Answer: C (LEAVE A REPLY)**
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-deploy-and-where

**NEW QUESTION: 3**
You need to replace the missing data in the AccessibilityToHighway columns.
How should you configure the Clean Missing Data module? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
**Answer:**
Explanation
Box 1: Replace using MICE

Replace using MICE: For each missing value, this option assigns a new value, which is calculated by using a method described in the statistical literature as "Multivariate Imputation using Chained Equations" or "Multiple Imputation by Chained Equations". With a multiple imputation method, each variable with missing data is modeled conditionally using the other variables in the data before filling in the missing values.
Scenario: The AccessibilityToHighway column in both datasets contains missing values. The missing data must be replaced with new data so that it is modeled conditionally using the other variables in the data before filling in the missing values.
Box 2: Propagate
Cols with all missing values indicate if columns of all missing values should be preserved in the output.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/clean-missing-data

**NEW QUESTION: 4**
You need to produce a visualization for the diagnostic test evaluation according to the data visualization requirements.
Which three modules should you recommend be used in sequence? To answer, move the appropriate modules from the list of modules to the answer area and arrange them in the correct order.
**Answer:**
Explanation:
Step 1: Sweep Clustering
Start by using the "Tune Model Hyperparameters" module to select the best sets of parameters for each of the models we're considering.
One of the interesting things about the "Tune Model Hyperparameters" module is that it not only outputs the results from the Tuning, it also outputs the Trained Model.
Step 2: Train Model
Step 3: Evaluate Model
Scenario: You need to provide the test results to the Fabrikam Residences team. You create data visualizations to aid in presenting the results.
You must produce a Receiver Operating Characteristic (ROC) curve to conduct a diagnostic test evaluation of the model. You need to select appropriate methods for producing the ROC curve in Azure Machine Learning Studio to compare the Two-Class Decision Forest and the Two-Class Decision Jungle modules with one another.
References:
http://breaking-bi.blogspot.com/2017/01/azure-machine-learning-model-evaluation.html

**NEW QUESTION: 5**
For each of the following statements, select Yes if the statement is true. Otherwise, select No. NOTE: Each correct selection is worth one point.
**Answer:**

**NEW QUESTION: 6**

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You create a model to forecast weather conditions based on historical data.
You need to create a pipeline that runs a processing script to load data from a datastore and pass the processed data to a machine learning model training script.
Solution: Run the following code:
Does the solution meet the goal?
**A.** Yes
**B.** No
**Answer: (SHOW ANSWER)**
The two steps are present: process_step and train_step
The training data input is not setup correctly.
Note:
Data used in pipeline can be produced by one step and consumed in another step by providing a PipelineData object as an output of one step and an input of one or more subsequent steps.
PipelineData objects are also used when constructing Pipelines to describe step dependencies. To specify that a step requires the output of another step as input, use a PipelineData object in the constructor of both steps.
For example, the pipeline train step depends on the process_step_output output of the pipeline process step:
from azureml.pipeline.core import Pipeline, PipelineData
from azureml.pipeline.steps import PythonScriptStep
datastore = ws.get_default_datastore()
process_step_output = PipelineData("processed_data", datastore=datastore) process_step = PythonScriptStep(script_name="process.py", arguments=["--data_for_train", process_step_output], outputs=[process_step_output], compute_target=aml_compute, source_directory=process_directory)
train_step = PythonScriptStep(script_name="train.py", arguments=["--data_for_train", process_step_output], inputs=[process_step_output], compute_target=aml_compute, source_directory=train_directory) pipeline = Pipeline(workspace=ws, steps=[process_step, train_step]) Reference:
https://docs.microsoft.com/en-us/python/api/azureml-pipeline-core/azureml.pipeline.core.pipelinedata?view=azure-ml-py

**NEW QUESTION: 7**
You are solving a classification task.
You must evaluate your model on a limited data sample by using k-fold cross-validation. You start by configuring a k parameter as the number of splits.
You need to configure the k parameter for the cross-validation.
Which value should you use?
**A.** k=1

**B.** k=10
**C.** k=0.5
**D.** k=0.9
**Answer: B (LEAVE A REPLY)**
Explanation
Leave One Out (LOO) cross-validation
Setting K = n (the number of observations) yields n-fold and is called leave-one out cross-validation (LOO), a special case of the K-fold approach.
LOO CV is sometimes useful but typically doesn't shake up the data enough. The estimates from each fold are highly correlated and hence their average can have high variance.
This is why the usual choice is K=5 or 10. It provides a good compromise for the bias-variance tradeoff.

**NEW QUESTION: 8**
You are analyzing a raw dataset that requires cleaning.
You must perform transformations and manipulations by using Azure Machine Learning Studio.
You need to identify the correct modules to perform the transformations.
Which modules should you choose? To answer, drag the appropriate modules to the correct scenarios. Each module may be used once, more than once, or not at all.
You may need to drag the split bar between panes or scroll to view content.
NOTE: Each correct selection is worth one point.
**Answer:**
Explanation:
Box 1: Clean Missing Data
Box 2: SMOTE
Use the SMOTE module in Azure Machine Learning Studio to increase the number of underepresented cases in a dataset used for machine learning. SMOTE is a better way of increasing the number of rare cases than simply duplicating existing cases.
Box 3: Convert to Indicator Values
Use the Convert to Indicator Values module in Azure Machine Learning Studio. The purpose of this module is to convert columns that contain categorical values into a series of binary indicator columns that can more easily be used as features in a machine learning model.
Box 4: Remove Duplicate Rows
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/smote
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/convert-to-indicator-values

**NEW QUESTION: 9**
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.

After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You are a data scientist using Azure Machine Learning Studio.
You need to normalize values to produce an output column into bins to predict a target column.
Solution: Apply a Quantiles binning mode with a PQuantile normalization.
Does the solution meet the goal?
**A.** Yes
**B.** No
**Answer: B (LEAVE A REPLY)**
Explanation
Use the Entropy MDL binning mode which has a target column.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/group-data-into-bins

**NEW QUESTION: 10**
You use the following code to run a script as an experiment in Azure Machine Learning:
You must identify the output files that are generated by the experiment run.
You need to add code to retrieve the output file names.
Which code segment should you add to the script?
**A.** files = run.get_properties()
**B.** files= run.get_file_names()
**C.** files = run.get_details_with_logs()
**D.** files = run.get_metrics()
**E.** files = run.get_details()
**Answer: B (LEAVE A REPLY)**
You can list all of the files that are associated with this run record by called run.get_file_names() Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-track-experiments

**NEW QUESTION: 11**
You need to select an environment that will meet the business and data requirements.
Which environment should you use?
**A.** Azure HDInsight with Spark MLlib
**B.** Microsoft Machine Learning Server
**C.** Azure Cognitive Services
**D.** Azure Machine Learning Studio
**Answer: B (LEAVE A REPLY)**

**NEW QUESTION: 12**
You have a dataset that contains 2,000 rows. You are building a machine learning classification model by using Azure Learning Studio. You add a Partition and Sample module to the experiment.
You need to configure the module. You must meet the following requirements:

* Divide the data into subsets
* Assign the rows into folds using a round-robin method
* Allow rows in the dataset to be reused

How should you configure the module? To answer, select the appropriate options in the dialog box in the answer area.

NOTE: Each correct selection is worth one point.

**Answer:**

Explanation

Use the Split data into partitions option when you want to divide the dataset into subsets of the data. This option is also useful when you want to create a custom number of folds for cross-validation, or to split rows into several groups.

* Add the Partition and Sample module to your experiment in Studio (classic), and connect the dataset.
* For Partition or sample mode, select Assign to Folds.
* Use replacement in the partitioning: Select this option if you want the sampled row to be put back into the pool of rows for potential reuse. As a result, the same row might be assigned to several folds.
* If you do not use replacement (the default option), the sampled row is not put back into the pool of rows for potential reuse. As a result, each row can be assigned to only one fold.
* Randomized split: Select this option if you want rows to be randomly assigned to folds.

If you do not select this option, rows are assigned to folds using the round-robin method.

References:

https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/partition-and-sample

**NEW QUESTION: 13**

You are developing a deep learning model by using TensorFlow. You plan to run the model training workload on an Azure Machine Learning Compute Instance.

You must use CUDA-based model training.

You need to provision the Compute Instance.

Which two virtual machines sizes can you use? To answer, select the appropriate virtual machine sizes in the answer area.

NOTE: Each correct selection is worth one point.

**Answer:**

Explanation

CUDA is a parallel computing platform and programming model developed by Nvidia for general computing on its own GPUs (graphics processing units). CUDA enables developers to speed up compute-intensive applications by harnessing the power of GPUs for the parallelizable part of the computation.

Reference:

https://www.infoworld.com/article/3299703/what-is-cuda-parallel-programming-for-gpus.html

**NEW QUESTION: 14**

You are working on a classification task. You have a dataset indicating whether a student would like to play soccer and associated attributes. The dataset includes the following columns:
You need to classify variables by type.
Which variable should you add to each category? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
**Answer:**
References:
https://www.edureka.co/blog/classification-algorithms/

**NEW QUESTION: 15**
You create machine learning models by using Azure Machine Learning.
You plan to train and score models by using a variety of compute contexts. You also plan to create a new compute resource in Azure Machine Learning studio.
You need to select the appropriate compute types.
Which compute types should you select? To answer, drag the appropriate compute types to the correct requirements. Each compute type may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.
NOTE: Each correct selection is worth one point.
**Answer:**
Explanation
Box 1: Attached compute
Box 2: Inference cluster
Box 3: Training cluster
Box 4: Attached compute

**NEW QUESTION: 16**
You are building recurrent neural network to perform a binary classification.
The training loss, validation loss, training accuracy, and validation accuracy of each training epoch has been provided. You need to identify whether the classification model is over fitted.
Which of the following is correct?
**A.** The training loss decreases while the validation loss increases when training the model.
**B.** The training loss stays constant and the validation loss decreases when training the model.
**C.** The training loss increases while the validation loss decreases when training the model.
**D.** The training loss .stays constant and the validation loss stays on a constant value and close to the training loss value when training the model.
**Answer: C (LEAVE A REPLY)**

**NEW QUESTION: 17**
You have a dataset that includes home sales data for a city. The dataset includes the following columns.
Each row in the dataset corresponds to an individual home sales transaction.
You need to use automated machine learning to generate the best model for predicting the sales price based on the features of the house.
Which values should you use? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
**Answer:**
Reference:
https://docs.microsoft.com/en-us/learn/modules/create-regression-model-azure-machine-learning-designer

**NEW QUESTION: 18**
You are with a time series dataset in Azure Machine Learning Studio.
You need to split your dataset into training and testing subsets by using the Split Data module.
Which splitting mode should you use?
**A.** Regular Expression Split
**B.** Split Rows with the Randomized split parameter set to true
**C.** Relative Expression Split
**D.** Recommender Split
**Answer: B (LEAVE A REPLY)**
Split Rows: Use this option if you just want to divide the data into two parts. You can specify the percentage of data to put in each split, but by default, the data is divided 50-50.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/split-data

**NEW QUESTION: 19**
You need to implement a scaling strategy for the local penalty detection data.
Which normalization type should you use?
**A.** Streaming
**B.** Weight
**C.** Batch
**D.** Cosine
**Answer: C (LEAVE A REPLY)**

Post batch normalization statistics (PBN) is the Microsoft Cognitive Toolkit (CNTK) version of how to evaluate the population mean and variance of Batch Normalization which could be used in inference Original Paper.
In CNTK, custom networks are defined using the BrainScriptNetworkBuilder and described in the CNTK network description language "BrainScript." Scenario:
Local penalty detection models must be written by using BrainScript.
References:
https://docs.microsoft.com/en-us/cognitive-toolkit/post-batch-normalization-statistics Prepare data for modeling Testlet 2 Case study Overview You are a data scientist for Fabrikam Residences, a company specializing in quality private and commercial property in the United States. Fabrikam Residences is considering expanding into Europe and has asked you to investigate prices for private residences in major European cities. You use Azure Machine Learning Studio to measure the median value of properties. You produce a regression model to predict property prices by using the Linear Regression and Bayesian Linear Regression modules.
Datasets
There are two datasets in CSV format that contain property details for two cities, London and Paris, with the following columns:
The two datasets have been added to Azure Machine Learning Studio as separate datasets and included as the starting point of the experiment.
Dataset issues
The AccessibilityToHighway column in both datasets contains missing values. The missing data must be replaced with new data so that it is modeled conditionally using the other variables in the data before filling in the missing values.
Columns in each dataset contain missing and null values. The dataset also contains many outliers. The Age column has a high proportion of outliers. You need to remove the rows that have outliers in the Age column.
The MedianValue and AvgRoomsinHouse columns both hold data in numeric format. You need to select a feature selection algorithm to analyze the relationship between the two columns in more detail.
Model fit
The model shows signs of overfitting. You need to produce a more refined regression model that reduces the overfitting.
Experiment requirements
You must set up the experiment to cross-validate the Linear Regression and Bayesian Linear Regression modules to evaluate performance.
In each case, the predictor of the dataset is the column named MedianValue. An initial investigation showed that the datasets are identical in structure apart from the MedianValue column. The smaller Paris dataset contains the MedianValue in text format, whereas the larger London dataset contains the MedianValue in numerical format. You must ensure that the datatype of the MedianValue column of the Paris dataset matches the structure of the London dataset.
You must prioritize the columns of data for predicting the outcome. You must use non-parameters statistics to measure the relationships.
You must use a feature selection algorithm to analyze the relationship between the MedianValue and AvgRoomsinHouse columns.

Model training

Given a trained model and a test dataset, you need to compute the permutation feature importance scores of feature variables. You need to set up the Permutation Feature Importance module to select the correct metric to investigate the model's accuracy and replicate the findings.

You want to configure hyperparameters in the model learning process to speed the learning phase by using hyperparameters. In addition, this configuration should cancel the lowest performing runs at each evaluation interval, thereby directing effort and resources towards models that are more likely to be successful.

You are concerned that the model might not efficiently use compute resources in hyperparameter tuning. You also are concerned that the model might prevent an increase in the overall tuning time. Therefore, you need to implement an early stopping criterion on models that provides savings without terminating promising jobs.

Testing

You must produce multiple partitions of a dataset based on sampling using the Partition and Sample module in Azure Machine Learning Studio. You must create three equal partitions for cross-validation. You must also configure the cross-validation process so that the rows in the test and training datasets are divided evenly by properties that are near each city's main river. The data that identifies that a property is near a river is held in the column named NextToRiver. You want to complete this task before the data goes through the sampling process.

When you train a Linear Regression module using a property dataset that shows data for property prices for a large city, you need to determine the best features to use in a model. You can choose standard metrics provided to measure performance before and after the feature importance process completes. You must ensure that the distribution of the features across multiple training models is consistent.

Data visualization

You need to provide the test results to the Fabrikam Residences team. You create data visualizations to aid in presenting the results.

You must produce a Receiver Operating Characteristic (ROC) curve to conduct a diagnostic test evaluation of the model. You need to select appropriate methods for producing the ROC curve in Azure Machine Learning Studio to compare the Two-Class Decision Forest and the Two-Class Decision Jungle modules with one another.

Prepare data for modeling

Question Set 3

**NEW QUESTION: 20**

You need to implement a scaling strategy for the local penalty detection data.

Which normalization type should you use?

**A.** Streaming

**B.** Weight

**C.** Batch

**D.** Cosine

**Answer: (SHOW ANSWER)**

Post batch normalization statistics (PBN) is the Microsoft Cognitive Toolkit (CNTK) version of how to evaluate the population mean and variance of Batch Normalization which could be used in inference Original Paper.

In CNTK, custom networks are defined using the BrainScriptNetworkBuilder and described in the CNTK network description language "BrainScript." Scenario:
Local penalty detection models must be written by using BrainScript.
References:
https://docs.microsoft.com/en-us/cognitive-toolkit/post-batch-normalization-statistics Prepare data for modeling Testlet 2 Case study This is a case study. Case studies are not timed separately. You can use as much exam time as you would like to complete each case. However, there may be additional case studies and sections on this exam. You must manage your time to ensure that you are able to complete all questions included on this exam in the time provided.
To answer the questions included in a case study, you will need to reference information that is provided in the case study. Case studies might contain exhibits and other resources that provide more information about the scenario that is described in the case study. Each question is independent of the other questions in this case study.
At the end of this case study, a review screen will appear. This screen allows you to review your answers and to make changes before you move to the next section of the exam. After you begin a new section, you cannot return to this section.
To start the case study
To display the first question in this case study, click the Next button. Use the buttons in the left pane to explore the content of the case study before you answer the questions. Clicking these buttons displays information such as business requirements, existing environment, and problem statements. If the case study has an All Information tab, note that the information displayed is identical to the information displayed on the subsequent tabs. When you are ready to answer a question, click the Question button to return to the question.
Overview
You are a data scientist for Fabrikam Residences, a company specializing in quality private and commercial property in the Unites States. Fabrikam Residences is considering expanding into Europe and has asked you to investigate prices for private residences in major European cities.
You use Azure Machine Learning Studio to measure the median value of properties. You produce a regression model to predict property prices by using the Linear Regression and Bayesian Linear Regression modules.
Datasets
There are two datasets in CSV format that contain property details for two cities, London and Paris. You add both files to Azure Machine Learning Studio as separate datasets to the starting point for an experiment. Both datasets contain the following columns:
An initial investigation shows that the datasets are identical in structure apart from the MedianValue column. The smaller Paris dataset contains the MedianValue in text format, whereas the larger London dataset contains the MedianValue in numerical format.
Data issues
Missing values

The AccessibilityToHighway column in both datasets contains missing values. The missing data must be replaced with new data so that it is modeled conditionally using the other variables in the data before filling in the missing values.

Columns in each dataset contain missing and null values. The datasets also contain many outliers. The Age column has a high proportion of outliers. You need to remove the rows that have outliers in the Age column.

The MedianValue and AvgRoomsInHouse columns both hold data in numeric format. You need to select a feature selection algorithm to analyze the relationship between the two columns in more detail.

Model fit

The model shows signs of overfitting. You need to produce a more refined regression model that reduces the overfitting.

Experiment requirements

You must set up the experiment to cross-validate the Linear Regression and Bayesian Linear Regression modules to evaluate performance. In each case, the predictor of the dataset is the column named MedianValue. You must ensure that the datatype of the MedianValue column of the Paris dataset matches the structure of the London dataset.

You must prioritize the columns of data for predicting the outcome. You must use non-parametric statistics to measure relationships.

You must a feature selection algorithm to analyze the relationship between the MediaValue and AvgRoomsinHouse columns.

Model training

Permutation Feature Importance

Given a trained model and a test dataset, you must compute the Permutation Feature Importance scores of feature variables. You must be determined the absolute fit for the model.

Hyperparameters

You must configure hyperparameters in the model learning process to speed the learning phase. In addition, this configuration should cancel the lowest performing runs at each evaluation interval, thereby directing effort and resources towards models that are more likely to be successful.

You are concerned that the model might not efficiently use compute resources in hyperparameter tuning. You also are concerned that the model might prevent an increase in the overall tuning time. Therefore, must implement an early stopping criterion on models that provides savings without terminating promising jobs.

Testing

You must produce multiple partitions of a dataset based on sampling using the Partition and Sample module in Azure Machine Learning Studio.

Cross-validation

You must create three equal partitions for cross-validation. You must also configure the cross-validation process so that the rows in the test and training datasets are divided evenly by properties that are near each city's main river. You must complete this task before the data goes through the sampling process.

Linear regression module

When you train a Linear Regression module, you must determine the best features to use in a model. You can choose standard metrics provided to measure performance before and after the feature importance process completes. The distribution of features across multiple training models must be consistent.

Data visualization

You need to provide the test results to the Fabrikam Residences team. You create data visualizations to aid in presenting the results.

You must produce a Receiver Operating Characteristic (ROC) curve to conduct a diagnostic test evaluation of the model. You need to select appropriate methods for producing the ROC curve in Azure Learning Studio to compare the Two-Class Decision Forest and the Two-Class Decision Jungle modules with one another.

**NEW QUESTION: 21**

You need to implement a new cost factor scenario for the ad response models as illustrated in the performance curve exhibit.

Which technique should you use?

**A.** Set the threshold to 0.5 and retrain if weighted Kappa deviates +/- 5% from 0.45.

**B.** Set the threshold to 0.05 and retrain if weighted Kappa deviates +/- 5% from 0.5.

**C.** Set the threshold to 0.2 and retrain if weighted Kappa deviates +/- 5% from 0.6.

**D.** Set the threshold to 0.75 and retrain if weighted Kappa deviates +/- 5% from 0.15.

**Answer: A (LEAVE A REPLY)**

Scenario:

Performance curves of current and proposed cost factor scenarios are shown in the following diagram:

The ad propensity model uses a cut threshold is 0.45 and retrains occur if weighted Kappa deviated from 0.1 +/- 5%.

Develop models

Testlet 2

Case study

Overview

You are a data scientist for Fabrikam Residences, a company specializing in quality private and commercial property in the United States. Fabrikam Residences is considering expanding into Europe and has asked you to investigate prices for private residences in major European cities. You use Azure Machine Learning Studio to measure the median value of properties. You produce a regression model to predict property prices by using the Linear Regression and Bayesian Linear Regression modules.

Datasets

There are two datasets in CSV format that contain property details for two cities, London and Paris, with the following columns:

The two datasets have been added to Azure Machine Learning Studio as separate datasets and included as the starting point of the experiment.

Dataset issues

The AccessibilityToHighway column in both datasets contains missing values. The missing data must be replaced with new data so that it is modeled conditionally using the other variables in the data before filling in the missing values.

Columns in each dataset contain missing and null values. The dataset also contain many outliers. The Age column has a high proportion of outliers. You need to remove the rows that have outliers in the Age column.

The MedianValue and AvgRoomsinHouse columns both hold data in numeric format. You need to select a feature selection algorithm to analyze the relationship between the two columns in more detail.

Model fit

The model shows signs of overfitting. You need to produce a more refined regression model that reduces the overfitting.

Experiment requirements

You must set up the experiment to cross-validate the Linear Regression and Bayesian Linear Regression modules to evaluate performance.

In each case, the predictor of the dataset is the column named MedianValue. An initial investigation showed that the datasets are identical in structure apart from the MedianValue column. The smaller Paris dataset contains the MedianValue in text format, whereas the larger London dataset contains the MedianValue in numerical format. You must ensure that the datatype of the MedianValue column of the Paris dataset matches the structure of the London dataset.

You must prioritize the columns of data for predicting the outcome. You must use non-parameters statistics to measure the relationships.

You must use a feature selection algorithm to analyze the relationship between the MedianValue and AvgRoomsinHouse columns.

Model training

Given a trained model and a test dataset, you need to compute the permutation feature importance scores of feature variables. You need to set up the Permutation Feature Importance module to select the correct metric to investigate the model's accuracy and replicate the findings.

You want to configure hyperparameters in the model learning process to speed the learning phase by using hyperparameters. In addition, this configuration should cancel the lowest performing runs at each evaluation interval, thereby directing effort and resources towards models that are more likely to be successful.

You are concerned that the model might not efficiently use compute resources in hyperparameter tuning. You also are concerned that the model might prevent an increase in the overall tuning time. Therefore, you need to implement an early stopping criterion on models that provides savings without terminating promising jobs.

Testing

You must produce multiple partitions of a dataset based on sampling using the Partition and Sample module in Azure Machine Learning Studio. You must create three equal partitions for cross-validation. You must also configure the cross-validation process so that the rows in the test and training datasets are divided evenly by properties that are near each city's main river. The data that identifies that a property is near a river is held in the column named NextToRiver. You want to complete this task before the data goes through the sampling process.

When you train a Linear Regression module using a property dataset that shows data for property prices for a large city, you need to determine the best features to use in a model. You can choose standard metrics provided to measure performance before and after the feature importance process completes. You must ensure that the distribution of the features across multiple training models is consistent.

Data visualization

You need to provide the test results to the Fabrikam Residences team. You create data visualizations to aid in presenting the results.

You must produce a Receiver Operating Characteristic (ROC) curve to conduct a diagnostic test evaluation of the model. You need to select appropriate methods for producing the ROC curve in Azure Machine Learning Studio to compare the Two-Class Decision Forest and the Two-Class Decision Jungle modules with one another.
Develop models
Question Set 3

**NEW QUESTION: 22**
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You are a data scientist using Azure Machine Learning Studio.
You need to normalize values to produce an output column into bins to predict a target column.
Solution: Apply an Equal Width with Custom Start and Stop binning mode.
Does the solution meet the goal?
**A.** Yes
**B.** No
**Answer: B (LEAVE A REPLY)**
Explanation
Use the Entropy MDL binning mode which has a target column.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/group-data-into-bins

**NEW QUESTION: 23**
An organization uses Azure Machine Learning service and wants to expand their use of machine learning.
You have the following compute environments. The organization does not want to create another compute environment.
You need to determine which compute environment to use for the following scenarios.
Which compute types should you use? To answer, drag the appropriate compute environments to the correct scenarios. Each compute environment may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.
NOTE: Each correct selection is worth one point.
**Answer:**
Explanation:
Box 1: nb_server
Box 2: mlc_cluster
With Azure Machine Learning, you can train your model on a variety of resources or environments, collectively referred to as compute targets. A compute target can be a local machine or a cloud resource, such as an Azure Machine Learning Compute, Azure HDInsight or a remote virtual machine.

Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/concept-compute-target
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-set-up-training-targets

**NEW QUESTION: 24**
You are analyzing the asymmetry in a statistical distribution.
The following image contains two density curves that show the probability distribution of two datasets.
Use the drop-down menus to select the answer choice that answers each question based on the information presented in the graphic.
NOTE: Each correct selection is worth one point.
**Answer:**
Explanation
Box 1: Positive skew
Positive skew values means the distribution is skewed to the right.
Box 2: Negative skew
Negative skewness values mean the distribution is skewed to the left.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/compute-elementary-statistics

**NEW QUESTION: 25**
You need to build a feature extraction strategy for the local models.
How should you complete the code segment? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
**Answer:**
Explanation

**NEW QUESTION: 26**
You use an Azure Machine Learning workspace.
You have a trained model that must be deployed as a web service. Users must authenticate by using Azure Active Directory.
What should you do?
**A.** Deploy the model to Azure Kubernetes Service (AKS). During deployment, set the token_auth_enabled parameter of the target configuration object to true
**B.** Deploy the model to Azure Container Instances. During deployment, set the auch_enabled parameter of the target configuration object to true
**C.** Deploy the model to Azure Container Instances. During deployment, set the coken_auch_enabled parameter of the target configuration object to true
**D.** Deploy the model to Azure Kubernetes Service (AKS). During deployment, set the auch. enabled parameter of the target configuration object to true

**Answer: A (LEAVE A REPLY)**
To control token authentication, use the token_auth_enabled parameter when you create or update a deployment Token authentication is disabled by default when you deploy to Azure Kubernetes Service.
Note: The model deployments created by Azure Machine Learning can be configured to use one of two authentication methods:
key-based: A static key is used to authenticate to the web service.
token-based: A temporary token must be obtained from the Azure Machine Learning workspace (using Azure Active Directory) and used to authenticate to the web service.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-authenticate-web-service

**NEW QUESTION: 27**
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You train a classification model by using a logistic regression algorithm.
You must be able to explain the model's predictions by calculating the importance of each feature, both as an overall global relative importance value and as a measure of local importance for a specific set of predictions.
You need to create an explainer that you can use to retrieve the required global and local feature importance values.
Solution: Create a TabularExplainer.
Does the solution meet the goal?
**A.** Yes
**B.** No
Instead use Permutation Feature Importance Explainer (PFI).
Note 1:
Note 2: Permutation Feature Importance Explainer (PFI): Permutation Feature Importance is a technique used to explain classification and regression models. At a high level, the way it works is by randomly shuffling data one feature at a time for the entire dataset and calculating how much the performance metric of interest changes. The larger the change, the more important that feature is. PFI can explain the overall behavior of any underlying model but does not explain individual predictions.
**Answer: B (LEAVE A REPLY)**
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-machine-learning-interpretability

**NEW QUESTION: 28**
You use Azure Machine Learning Studio to build a machine learning experiment.
You need to divide data into two distinct datasets.
Which module should you use?

**A.** Split Data
**B.** Load Trained Model
**C.** Assign Data to Clusters
**D.** Group Data into Bins
**Answer: (SHOW ANSWER)**
The Group Data into Bins module supports multiple options for binning data. You can customize how the bin edges are set and how values are apportioned into the bins.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/group-data-into-bins

**NEW QUESTION: 29**
You are running a training experiment on remote compute in Azure Machine Learning.
The experiment is configured to use a conda environment that includes the mlflow and azureml-contrib-run packages.
You must use MLflow as the logging package for tracking metrics generated in the experiment.
You need to complete the script for the experiment.
How should you complete the code? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
**Answer:**
Explanation:
Box 1: import mlflow
Import the mlflow and Workspace classes to access MLflow's tracking URI and configure your workspace.
Box 2: mlflow.start_run()
Set the MLflow experiment name with set_experiment() and start your training run with start_run().
Box 3: mlflow.log_metric(' ..')
Use log_metric() to activate the MLflow logging API and begin logging your training run metrics.
Box 4: mlflow.end_run()
Close the run:
run.endRun()
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-use-mlflow

**NEW QUESTION: 30**
You are developing a machine learning, experiment by using Azure. The following images show the input and output of a machine learning experiment:
Use the drop-down menus to select the answer choice that answers each question based on the information presented in the graphic.
NOTE: Each correct selection is worth one point.
**Answer:**

**NEW QUESTION: 31**

You are performing sentiment analysis using a CSV file that includes 12.0O0 customer reviews written in a short sentence format. You add the CSV file to Azure Machine Learning Studio and Configure it as the starting point dataset of an experiment. You add the Extract N-Gram Features from Text module to the experiment to extract key phrases from the customer review column in the dataset.
You must create a new n-gram text dictionary from the customer review text and set the maximum n-gram size to trigrams.
You need to configure the Extract N Gram features from Text module.
What should you select? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
**Answer:**



**NEW QUESTION: 32**
You create an Azure Machine Learning compute resource to train models. The compute resource is configured as follows:
Minimum nodes: 2
Maximum nodes: 4
You must decrease the minimum number of nodes and increase the maximum number of nodes to the following values:
Minimum nodes: 0
Maximum nodes: 8
You need to reconfigure the compute resource.
What are three possible ways to achieve this goal? Each correct answer presents a complete solution.
NOTE: Each correct selection is worth one point.
**A.** Use the Azure Machine Learning studio.
**B.** Run the update method of the AmlCompute class in the Python SDK.
**C.** Use the Azure portal.
**D.** Use the Azure Machine Learning designer.
**E.** Run the refresh_state() method of the BatchCompute class in the Python SDK
**Answer: A,B,C (LEAVE A REPLY)**
Reference:
https://docs.microsoft.com/en-us/python/api/azureml-core/azureml.core.compute.amlcompute(class)

**NEW QUESTION: 33**

You use the Azure Machine Learning SDK in a notebook to run an experiment using a script file in an experiment folder.
The experiment fails.
You need to troubleshoot the failed experiment.
What are two possible ways to achieve this goal? Each correct answer presents a complete solution.
**A.** Use the get_metrics() method of the run object to retrieve the experiment run logs.
**B.** Use the get_details_with_logs() method of the run object to display the experiment run logs.
**C.** View the log files for the experiment run in the experiment folder.
**D.** View the logs for the experiment run in Azure Machine Learning studio.
**E.** Use the get_output() method of the run object to retrieve the experiment run logs.
**Answer: B,D (LEAVE A REPLY)**
Explanation
Use get_details_with_logs() to fetch the run details and logs created by the run.
You can monitor Azure Machine Learning runs and view their logs with the Azure Machine Learning studio.
Reference:
https://docs.microsoft.com/en-us/python/api/azureml-pipeline-core/azureml.pipeline.core.steprun
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-monitor-view-training-logs

**NEW QUESTION: 34**
You use the Azure Machine Learning SDK in a notebook to run an experiment using a script file in an experiment folder.
The experiment fails.
You need to troubleshoot the failed experiment.
What are two possible ways to achieve this goal? Each correct answer presents a complete solution.
**A.** View the logs for the experiment run in Azure Machine Learning studio.
**B.** Use the get_details_with_logsO method of the run object to display the experiment run logs.
**C.** Use the get.metricsO method of the run object to retrieve the experiment run logs.
**D.** View the log files for the experiment i un in the experiment folder.
**E.** Use the get_output() method of the run object to retrieve the experiment run logs.
**Answer: D,E (LEAVE A REPLY)**

**NEW QUESTION: 35**
You use Azure Machine Learning Studio to build a machine learning experiment.
You need to divide data into two distinct datasets.
Which module should you use?
**A.** Assign Data to Clusters
**B.** Partition and Sample
**C.** Test Hypothesis Using t-Test
**D.** Group Data into Bins
**Answer: B (LEAVE A REPLY)**

**NEW QUESTION: 36**
You plan to use the Hyperdrive feature of Azure Machine Learning to determine the optimal hyperparameter values when training a model.
You must use Hyperdrive to try combinations of the following hyperparameter values. You must not apply an early termination policy.
* learning_rate: any value between 0.001 and 0.1
* batch_size: 16, 32, or 64
You need to configure the sampling method for the Hyperdrive experiment.
Which two sampling methods can you use? Each correct answer is a complete solution.
NOTE: Each correct selection is worth one point.
**A.** No sampling
**B.** Grid sampling
**C.** Bayesian sampling
**D.** Random sampling
**Answer: (SHOW ANSWER)**
Explanation/Reference:
C: Bayesian sampling is based on the Bayesian optimization algorithm and makes intelligent choices on the hyperparameter values to sample next. It picks the sample based on how the previous samples performed, such that the new sample improves the reported primary metric.
Bayesian sampling does not support any early termination policy
Example:
from azureml.train.hyperdrive import BayesianParameterSampling
from azureml.train.hyperdrive import uniform, choice
param_sampling = BayesianParameterSampling( {
"learning_rate": uniform(0.05, 0.1),
"batch_size": choice(16, 32, 64, 128)
}
)
D: In random sampling, hyperparameter values are randomly selected from the defined search space. Random sampling allows the search space to include both discrete and continuous hyperparameters.
Incorrect Answers:
B: Grid sampling can be used if your hyperparameter space can be defined as a choice among discrete values and if you have sufficient budget to exhaustively search over all values in the defined search space. Additionally, one can use automated early termination of poorly performing runs, which reduces wastage of resources.
Example, the following space has a total of six samples:
from azureml.train.hyperdrive import GridParameterSampling
from azureml.train.hyperdrive import choice
param_sampling = GridParameterSampling( {
"num_hidden_layers": choice(1, 2, 3),
"batch_size": choice(16, 32)

}
)
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-tune-hyperparameters

**NEW QUESTION: 37**
You need to correct the model fit issue.
Which three actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.
**Answer:**
Explanation
Step 1: Augment the data
Scenario: Columns in each dataset contain missing and null values. The datasets also contain many outliers.
Step 2: Add the Bayesian Linear Regression module.
Scenario: You produce a regression model to predict property prices by using the Linear Regression and Bayesian Linear Regression modules.
Step 3: Configure the regularization weight.
Regularization typically is used to avoid overfitting. For example, in L2 regularization weight, type the value to use as the weight for L2 regularization. We recommend that you use a non-zero value to avoid overfitting.
Scenario:
Model fit: The model shows signs of overfitting. You need to produce a more refined regression model that reduces the overfitting.

**NEW QUESTION: 38**
You are developing deep learning models to analyze semi-structured, unstructured, and structured data types.
You have the following data available for model building:
Video recordings of sporting events
Transcripts of radio commentary about events
Logs from related social media feeds captured during sporting events
You need to select an environment for creating the model.
Which environment should you use?
**A.** Azure Cognitive Services
**B.** Azure Data Lake Analytics
**C.** Azure HDInsight with Spark MLib
**D.** Azure Machine Learning Studio
**Answer: A (LEAVE A REPLY)**
Azure Cognitive Services expand on Microsoft's evolving portfolio of machine learning APIs and enable developers to easily add cognitive features - such as emotion and video detection; facial, speech, and vision recognition; and speech and language understanding - into their applications. The goal of Azure Cognitive Services is to help developers create applications that can see, hear, speak, understand, and even begin to

reason. The catalog of services within Azure Cognitive Services can be categorized into five main pillars - Vision, Speech, Language, Search, and Knowledge.
References:
https://docs.microsoft.com/en-us/azure/cognitive-services/welcome

**NEW QUESTION: 39**
You need to define an evaluation strategy for the crowd sentiment models.
Which three actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.
**Answer:**
Explanation:
Step 1: Define a cross-entropy function activation
When using a neural network to perform classification and prediction, it is usually better to use cross-entropy error than classification error, and somewhat better to use cross-entropy error than mean squared error to evaluate the quality of the neural network.
Step 2: Add cost functions for each target state.
Step 3: Evaluated the distance error metric.
References:
https://www.analyticsvidhya.com/blog/2018/04/fundamentals-deep-learning-regularization-techniques/

**NEW QUESTION: 40**
You need to use the Python language to build a sampling strategy for the global penalty detection models.
How should you complete the code segment? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
**Answer:**
Explanation
Box 1: import pytorch as deeplearninglib
Box 2: ..DistributedSampler(Sampler)..
DistributedSampler(Sampler):
Sampler that restricts data loading to a subset of the dataset.
It is especially useful in conjunction with class:`torch.nn.parallel.DistributedDataParallel`. In such case, each process can pass a DistributedSampler instance as a DataLoader sampler, and load a subset of the original dataset that is exclusive to it.
Scenario: Sampling must guarantee mutual and collective exclusively between local and global segmentation models that share the same features.
Box 3: optimizer = deeplearninglib.train. GradientDescentOptimizer(learning_rate=0.10)

**NEW QUESTION: 41**
You need to define an evaluation strategy for the crowd sentiment models.
Which three actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.

**Answer:**

1 - Define a cross-entropy function activation

2 - Add cost functions for each target state.

3 - Evaluated the distance error metric.

Reference:

https://www.analyticsvidhya.com/blog/2018/04/fundamentals-deep-learning-regularization-techniques/

**NEW QUESTION: 42**

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.

After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

You are creating a model to predict the price of a student's artwork depending on the following variables: the student's length of education, degree type, and art form.

You start by creating a linear regression model.

You need to evaluate the linear regression model.

Solution: Use the following metrics: Mean Absolute Error, Root Mean Absolute Error, Relative Absolute Error, Relative Squared Error, and the Coefficient of Determination.

Does the solution meet the goal?

**A.** Yes

**B.** No

**Answer: A (LEAVE A REPLY)**

The following metrics are reported for evaluating regression models. When you compare models, they are ranked by the metric you select for evaluation.

Mean absolute error (MAE) measures how close the predictions are to the actual outcomes; thus, a lower score is better.

Root mean squared error (RMSE) creates a single value that summarizes the error in the model. By squaring the difference, the metric disregards the difference between over-prediction and under-prediction.

Relative absolute error (RAE) is the relative absolute difference between expected and actual values; relative because the mean difference is divided by the arithmetic mean.

Relative squared error (RSE) similarly normalizes the total squared error of the predicted values by dividing by the total squared error of the actual values.

Mean Zero One Error (MZOE) indicates whether the prediction was correct or not. In other words: ZeroOneLoss(x,y) = 1 when x!=y; otherwise 0.

Coefficient of determination, often referred to as R2, represents the predictive power of the model as a value between 0 and 1. Zero means the model is random (explains nothing); 1 means there is a perfect fit. However, caution should be used in interpreting R2 values, as low values can be entirely normal and high values can be suspect.

AUC.

Reference:

https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/evaluate-model

**NEW QUESTION: 43**
You need to configure the Feature Based Feature Selection module based on the experiment requirements and datasets.
How should you configure the module properties? To answer, select the appropriate options in the dialog box in the answer area.
NOTE: Each correct selection is worth one point.
**Answer:**
Explanation:
Box 1: Mutual Information.
The mutual information score is particularly useful in feature selection because it maximizes the mutual information between the joint distribution and target variables in datasets with many dimensions.
Box 2: MedianValue
MedianValue is the feature column, , it is the predictor of the dataset.
Scenario: The MedianValue and AvgRoomsinHouse columns both hold data in numeric format. You need to select a feature selection algorithm to analyze the relationship between the two columns in more detail.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/filter-based-feature-selection

**NEW QUESTION: 44**
You write code to retrieve an experiment that is run from your Azure Machine Learning workspace.
The run used the model interpretation support in Azure Machine Learning to generate and upload a model explanation.
Business managers in your organization want to see the importance of the features in the model.
You need to print out the model features and their relative importance in an output that looks similar to the following.
How should you complete the code? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
**Answer:**
Explanation:
Box 1: from_run_id
from_run_id(workspace, experiment_name, run_id)
Create the client with factory method given a run ID.
Returns an instance of the ExplanationClient.
Parameters
* Workspace Workspace An object that represents a workspace.
* experiment_name str The name of an experiment.
* run_id str A GUID that represents a run.
Box 2: list_model_explanations

list_model_explanations returns a dictionary of metadata for all model explanations available.
Returns
A dictionary of explanation metadata such as id, data type, explanation method, model type, and upload time, sorted by upload time Box 3: explanation Reference:
https://docs.microsoft.com/en-us/python/api/azureml-contrib-interpret/azureml.contrib.interpret.explanation.explanation_client.explanationclient?view=azure-ml-py

**NEW QUESTION: 45**
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You are a data scientist using Azure Machine Learning Studio.
You need to normalize values to produce an output column into bins to predict a target column.
Solution: Apply a Quantiles normalization with a QuantileIndex normalization.
Does the solution meet the GOAL?
**A.** Yes
**B.** No
**Answer: B (LEAVE A REPLY)**
Use the Entropy MDL binning mode which has a target column.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/group-data-into-bins

**NEW QUESTION: 46**
You have a Python data frame named salesData in the following format:
The data frame must be unpivoted to a long data format as follows:
You need to use the pandas.melt() function in Python to perform the transformation.
How should you complete the code segment? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
**Answer:**
Reference:
https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.melt.html

**NEW QUESTION: 47**
You use Azure Machine Learning to train and register a model.
You must deploy the model into production as a real-time web service to an inference cluster named service-compute that the IT department has created in the Azure Machine Learning workspace.
Client applications consuming the deployed web service must be authenticated based on their Azure Active Directory service principal.
You need to write a script that uses the Azure Machine Learning SDK to deploy the model. The necessary modules have been imported.
How should you complete the code? To answer, select the appropriate options in the answer area.
**Answer:**

**NEW QUESTION: 48**
You create a new Azure subscription. No resources are provisioned in the subscription.
You need to create an Azure Machine Learning workspace.
What are three possible ways to achieve this goal? Each correct answer presents a complete solution.
NOTE: Each correct selection is worth one point.
**A.** Run Python code that uses the Azure ML SDK library and calls the Workspace.create method with name, subscription_id, resource_group, and location parameters.
**B.** Use an Azure Resource Management template that includes a Microsoft.MachineLearningServices/ workspaces resource and its dependencies.
**C.** Use the Azure Command Line Interface (CLI) with the Azure Machine Learning extension to call the az group create function with --name and --location parameters, and then the az ml workspace create function, specifying -w and -g parameters for the workspace name and resource group.
**D.** Navigate to Azure Machine Learning studio and create a workspace.
**E.** Run Python code that uses the Azure ML SDK library and calls the Workspace.get method with name, subscription_id, and resource_group parameters.
**Answer: B,C,D (LEAVE A REPLY)**
Explanation
B: You can use an Azure Resource Manager template to create a workspace for Azure Machine Learning.
Example:
{"type": "Microsoft.MachineLearningServices/workspaces",
...
C: You can create a workspace for Azure Machine Learning with Azure CLI Install the machine learning extension.
Create a resource group: az group create --name <resource-group-name> --location <location> To create a new workspace where the services are automatically created, use the following command: az ml workspace create -w <workspace-name> -g <resource-group-name> D: You can create and manage Azure Machine Learning workspaces in the Azure portal.

* Sign in to the Azure portal by using the credentials for your Azure subscription.
* In the upper-left corner of Azure portal, select + Create a resource.
* Use the search bar to find Machine Learning.
* Select Machine Learning.
* In the Machine Learning pane, select Create to begin.

Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-create-workspace-template
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-manage-workspace-cli
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-manage-workspace

**NEW QUESTION: 49**
You use Data Science Virtual Machines (DSVMs) for Windows and Linux in Azure.
You need to access the DSVMs.
Which utilities should you use? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
**Answer:**

**NEW QUESTION: 50**
You are solving a classification task.
You must evaluate your model on a limited data sample by using k-fold cross validation. You start by configuring a k parameter as the number of splits.
You need to configure the k parameter for the cross-validation.
Which value should you use?
**A.** k=0.5
**B.** k=0
**C.** k=5
**D.** k=1
**Answer: (SHOW ANSWER)**
Explanation
Leave One Out (LOO) cross-validation
Setting K = n (the number of observations) yields n-fold and is called leave-one out cross-validation (LOO), a special case of the K-fold approach.
LOO CV is sometimes useful but typically doesn't shake up the data enough. The estimates from each fold are highly correlated and hence their average can have high variance.
This is why the usual choice is K=5 or 10. It provides a good compromise for the bias-variance tradeoff.

**NEW QUESTION: 51**
You are performing feature scaling by using the scikit-learn Python library for x.1 x2, and x3 features.
Original and scaled data is shown in the following image.
Use the drop-down menus to select the answer choice that answers each question based on the information presented in the graphic.

NOTE: Each correct selection is worth one point.

**Answer:**

Explanation:

Box 1: StandardScaler

The StandardScaler assumes your data is normally distributed within each feature and will scale them such that the distribution is now centred around 0, with a standard deviation of 1.

Example:

All features are now on the same scale relative to one another.

Box 2: Min Max Scaler

Notice that the skewness of the distribution is maintained but the 3 distributions are brought into the same scale so that they overlap.

Box 3: Normalizer

References:

http://benalexkeen.com/feature-scaling-with-scikit-learn/

**NEW QUESTION: 52**

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.

After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

You have a Python script named train.py in a local folder named scripts. The script trains a regression model by using scikit-learn. The script includes code to load a training data file which is also located in the scripts folder.

You must run the script as an Azure ML experiment on a compute cluster named aml-compute.

You need to configure the run to ensure that the environment includes the required packages for model training. You have instantiated a variable named aml-compute that references the target compute cluster.

Solution: Run the following code:

Does the solution meet the goal?

**A.** Yes

**B.** No

**Answer: (SHOW ANSWER)**

There is a missing line: conda_packages=['scikit-learn'], which is needed.

Correct example:

```
sk_est = Estimator(source_directory='./my-sklearn-proj',
script_params=script_params,
compute_target=compute_target,
entry_script='train.py',
conda_packages=['scikit-learn'])
```

Note:

The Estimator class represents a generic estimator to train data using any supplied framework.

This class is designed for use with machine learning frameworks that do not already have an Azure Machine Learning pre-configured estimator. Pre-configured estimators exist for Chainer, PyTorch, TensorFlow, and SKLearn.
Example:
from azureml.train.estimator import Estimator
script_params = {
# to mount files referenced by mnist dataset
'--data-folder': ds.as_named_input('mnist').as_mount(),
'--regularization': 0.8
}
Reference:
https://docs.microsoft.com/en-us/python/api/azureml-train-core/azureml.train.estimator.estimator

**NEW QUESTION: 53**
You plan to use a Deep Learning Virtual Machine (DLVM) to train deep learning models using Compute Unified Device Architecture (CUDA) computations.
You need to configure the DLVM to support CUDA.
What should you implement?
**A.** Intel Software Guard Extensions (Intel SGX) technology
**B.** Solid State Drives (SSD)
**C.** Graphic Processing Unit (GPU)
**D.** Computer Processing Unit (CPU) speed increase by using overcloking
**E.** High Random Access Memory (RAM) configuration
**Answer: (SHOW ANSWER)**
A Deep Learning Virtual Machine is a pre-configured environment for deep learning using GPU instances.
Reference:
https://azuremarketplace.microsoft.com/en-au/marketplace/apps/microsoft-ads.dsvm-deep-learning

**NEW QUESTION: 54**
You need to implement a new cost factor scenario for the ad response models as illustrated in the performance curve exhibit.
Which technique should you use?
**A.** Set the threshold to 0.5 and retrain if weighted Kappa deviates +/- 5% from 0.45.
**B.** Set the threshold to 0.05 and retrain if weighted Kappa deviates +/- 5% from 0.5.
**C.** Set the threshold to 0.2 and retrain if weighted Kappa deviates +/- 5% from 0.6.
**D.** Set the threshold to 0.75 and retrain if weighted Kappa deviates +/- 5% from 0.15.
**Answer: A (LEAVE A REPLY)**
Explanation
Scenario:
Performance curves of current and proposed cost factor scenarios are shown in the following diagram:
The ad propensity model uses a cut threshold is 0.45 and retrains occur if weighted Kappa deviated from 0.1

+/- 5%.

Topic 2, Case Study 2

Case study

Overview

You are a data scientist for Fabrikam Residences, a company specializing in quality private and commercial property in the United States. Fabrikam Residences is considering expanding into Europe and has asked you to investigate prices for private residences in major European cities. You use Azure Machine Learning Studio to measure the median value of properties. You produce a regression model to predict property prices by using the Linear Regression and Bayesian Linear Regression modules.

Datasets

There are two datasets in CSV format that contain property details for two cities, London and Paris, with the following columns:

The two datasets have been added to Azure Machine Learning Studio as separate datasets and included as the starting point of the experiment.

Dataset issues

The AccessibilityToHighway column in both datasets contains missing values. The missing data must be replaced with new data so that it is modeled conditionally using the other variables in the data before filling in the missing values.

Columns in each dataset contain missing and null values. The dataset also contains many outliers. The Age column has a high proportion of outliers. You need to remove the rows that have outliers in the Age column.

The MedianValue and AvgRoomsinHouse columns both hold data in numeric format. You need to select a feature selection algorithm to analyze the relationship between the two columns in more detail.

Model fit

The model shows signs of overfitting. You need to produce a more refined regression model that reduces the overfitting.

Experiment requirements

You must set up the experiment to cross-validate the Linear Regression and Bayesian Linear Regression modules to evaluate performance.

In each case, the predictor of the dataset is the column named MedianValue. An initial investigation showed that the datasets are identical in structure apart from the MedianValue column. The smaller Paris dataset contains the MedianValue in text format, whereas the larger London dataset contains the MedianValue in numerical format. You must ensure that the datatype of the MedianValue column of the Paris dataset matches the structure of the London dataset.

You must prioritize the columns of data for predicting the outcome. You must use non-parameters statistics to measure the relationships.

You must use a feature selection algorithm to analyze the relationship between the MedianValue and AvgRoomsinHouse columns.

Model training

Given a trained model and a test dataset, you need to compute the permutation feature importance scores of feature variables. You need to set up the Permutation Feature Importance module to select the correct metric to investigate the model's accuracy and replicate the findings.

You want to configure hyperparameters in the model learning process to speed the learning phase by using hyperparameters. In addition, this configuration should cancel the lowest performing runs at each evaluation interval, thereby directing effort and resources towards models that are more likely to be successful.
You are concerned that the model might not efficiently use compute resources in hyperparameter tuning. You also are concerned that the model might prevent an increase in the overall tuning time. Therefore, you need to implement an early stopping criterion on models that provides savings without terminating promising jobs.
Testing
You must produce multiple partitions of a dataset based on sampling using the Partition and Sample module in Azure Machine Learning Studio. You must create three equal partitions for cross-validation. You must also configure the cross-validation process so that the rows in the test and training datasets are divided evenly by properties that are near each city's main river. The data that identifies that a property is near a river is held in the column named NextToRiver. You want to complete this task before the data goes through the sampling process.
When you train a Linear Regression module using a property dataset that shows data for property prices for a large city, you need to determine the best features to use in a model. You can choose standard metrics provided to measure performance before and after the feature importance process completes. You must ensure that the distribution of the features across multiple training models is consistent.
Data visualization
You need to provide the test results to the Fabrikam Residences team. You create data visualizations to aid in presenting the results.
You must produce a Receiver Operating Characteristic (ROC) curve to conduct a diagnostic test evaluation of the model. You need to select appropriate methods for producing the ROC curve in Azure Machine Learning Studio to compare the Two-Class Decision Forest and the Two-Class Decision Jungle modules with one another.

**NEW QUESTION: 55**
You use Data Science Virtual Machines (DSVMs) for Windows and Linux in Azure.
You need to access the DSVMs.
Which utilities should you use? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
**Answer:**

**NEW QUESTION: 56**
You are a lead data scientist for a project that tracks the health and migration of birds. You create a multi-class image classification deep learning model that uses a set of labeled bird photographs collected by experts.
You have 100,000 photographs of birds. All photographs use the JPG format and are stored in an Azure blob container in an Azure subscription.
You need to access the bird photograph files in the Azure blob container from the Azure Machine Learning service workspace that will be used for deep learning model training. You must minimize data movement.
What should you do?

**A.** Create an Azure Data Lake store and move the bird photographs to the store.
**B.** Create an Azure Cosmos DB database and attach the Azure Blob containing bird photographs storage to the database.
**C.** Create and register a dataset by using TabularDataset class that references the Azure blob storage containing bird photographs.
**D.** Register the Azure blob storage containing the bird photographs as a datastore in Azure Machine Learning service.
**E.** Copy the bird photographs to the blob datastore that was created with your Azure Machine Learning service workspace.
**Answer: D (LEAVE A REPLY)**
Explanation
We recommend creating a datastore for an Azure Blob container. When you create a workspace, an Azure blob container and an Azure file share are automatically registered to the workspace.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-access-data

**NEW QUESTION: 57**
You plan to create a speech recognition deep learning model.
The model must support the latest version of Python.
You need to recommend a deep learning framework for speech recognition to include in the Data Science Virtual Machine (DSVM).
What should you recommend?
**A.** Rattle
**B.** TensorFlow
**C.** Weka
**D.** Scikit-learn
**Answer: B (LEAVE A REPLY)**
Explanation
Explanation:
TensorFlow is an open source library for numerical computation and large-scale machine learning. It uses Python to provide a convenient front-end API for building applications with the framework TensorFlow can train and run deep neural networks for handwritten digit classification, image recognition, word embeddings, recurrent neural networks, sequence-to-sequence models for machine translation, natural language processing, and PDE (partial differential equation) based simulations.
Incorrect Answers:
A: Rattle is the R analytical tool that gets you started with data analytics and machine learning.
C: Weka is used for visual data mining and machine learning software in Java.
D: Scikit-learn is one of the most useful library for machine learning in Python. It is on NumPy, SciPy and matplotlib, this library contains a lot of effiecient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction.
Reference:

https://www.infoworld.com/article/3278008/what-is-tensorflow-the-machine-learning-library-explained.html

**NEW QUESTION: 58**
You need to configure the Permutation Feature Importance module for the model framing requirements.
What should you do? To answer, select the appropriate options in the dialog box in the answer area-NOTE: Each correct selection is worth one point.
**Answer:**

**NEW QUESTION: 59**
You previously deployed a model that was trained using a tabular dataset named training-dataset, which is based on a folder of CSV files.
Over time, you have collected the features and predicted labels generated by the model in a folder containing a CSV file for each month. You have created two tabular datasets based on the folder containing the inference data: one named predictions-dataset with a schema that matches the training data exactly, including the predicted label; and another named features-dataset with a schema containing all of the feature columns and a timestamp column based on the filename, which includes the day, month, and year.
You need to create a data drift monitor to identify any changing trends in the feature data since the model was trained. To accomplish this, you must define the required datasets for the data drift monitor.
Which datasets should you use to configure the data drift monitor? To answer, drag the appropriate datasets to the correct data drift monitor options. Each source may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.
NOTE: Each correct selection is worth one point.
**Answer:**
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-monitor-datasets

**NEW QUESTION: 60**
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You are creating a model to predict the price of a student's artwork depending on the following variables: the student's length of education, degree type, and art form.
You start by creating a linear regression model.
You need to evaluate the linear regression model.
Solution: Use the following metrics: Mean Absolute Error, Root Mean Absolute Error, Relative Absolute Error, Accuracy, Precision, Recall, F1 score, and AUC.
Does the solution meet the goal?
**A.** Yes
**B.** No

**Answer: (SHOW ANSWER)**

Accuracy, Precision, Recall, F1 score, and AUC are metrics for evaluating classification models.

Note: Mean Absolute Error, Root Mean Absolute Error, Relative Absolute Error are OK for the linear regression model.

References:

https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/evaluate-model

**NEW QUESTION: 61**

You are producing a multiple linear regression model in Azure Machine Learning Studio.

Several independent variables are highly correlated.

You need to select appropriate methods for conducting effective feature engineering on all the data.

Which three actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.

**Answer:**

Explanation

Step 1: Use the Filter Based Feature Selection module

Filter Based Feature Selection identifies the features in a dataset with the greatest predictive power.

The module outputs a dataset that contains the best feature columns, as ranked by predictive power. It also outputs the names of the features and their scores from the selected metric.

Step 2: Build a counting transform

A counting transform creates a transformation that turns count tables into features, so that you can apply the transformation to multiple datasets.

Step 3: Test the hypothesis using t-Test

References:

https://docs.microsoft.com/bs-latn-ba/azure/machine-learning/studio-module-reference/filter-based-feature-selec

https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/build-counting-transform



**NEW QUESTION: 62**

You use the Azure Machine Learning Python SDK to define a pipeline to train a model.

The data used to train the model is read from a folder in a datastore.

You need to ensure the pipeline runs automatically whenever the data in the folder changes.

What should you do?

**A.** Set the regenerate_outputs property of the pipeline to True
**B.** Create a ScheduleRecurrance object with a Frequency of auto. Use the object to create a Schedule for the pipeline
**C.** Create a PipelineParameter with a default value that references the location where the training data is stored
**D.** Create a Schedule for the pipeline. Specify the datastore in the datastore property, and the folder containing the training data in the path_on_datascore property
**Answer: D (LEAVE A REPLY)**
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-trigger-published-pipeline

**NEW QUESTION: 63**
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You are creating a model to predict the price of a student's artwork depending on the following variables: the student's length of education, degree type, and art form.
You start by creating a linear regression model.
You need to evaluate the linear regression model.
Solution: Use the following metrics: Mean Absolute Error, Root Mean Absolute Error, Relative Absolute Error, Relative Squared Error, and the Coefficient of Determination.
Does the solution meet the goal?
**A.** Yes
**B.** No
**Answer: A (LEAVE A REPLY)**
The following metrics are reported for evaluating regression models. When you compare models, they are ranked by the metric you select for evaluation.
Mean absolute error (MAE) measures how close the predictions are to the actual outcomes; thus, a lower score is better.
Root mean squared error (RMSE) creates a single value that summarizes the error in the model. By squaring the difference, the metric disregards the difference between over-prediction and under-prediction.
Relative absolute error (RAE) is the relative absolute difference between expected and actual values; relative because the mean difference is divided by the arithmetic mean.
Relative squared error (RSE) similarly normalizes the total squared error of the predicted values by dividing by the total squared error of the actual values.
Mean Zero One Error (MZOE) indicates whether the prediction was correct or not. In other words: ZeroOneLoss(x,y) = 1 when x!=y; otherwise 0.
Coefficient of determination, often referred to as R2, represents the predictive power of the model as a value between 0 and 1. Zero means the model is random (explains nothing); 1 means there is a perfect fit.

However, caution should be used in interpreting R2 values, as low values can be entirely normal and high values can be suspect.
AUC.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/evaluate-model

**NEW QUESTION: 64**
You are analyzing a dataset containing historical data from a local taxi company. You are developing a regression model.
You must predict the fare of a taxi trip.
You need to select performance metrics to correctly evaluate the regression model.
Which two metrics can you use? Each correct answer presents a complete solution?
NOTE: Each correct selection is worth one point.
**A.** a Root Mean Square Error value that is low
**B.** an R-Squared value close to 0
**C.** an F1 score that is low
**D.** an R-Squared value close to 1
**E.** an F1 score that is high
**F.** a Root Mean Square Error value that is high
**Answer: A,D (LEAVE A REPLY)**
RMSE and R2 are both metrics for regression models.
A: Root mean squared error (RMSE) creates a single value that summarizes the error in the model. By squaring the difference, the metric disregards the difference between over-prediction and under-prediction.
D: Coefficient of determination, often referred to as R2, represents the predictive power of the model as a value between 0 and 1. Zero means the model is random (explains nothing); 1 means there is a perfect fit.
However, caution should be used in interpreting R2 values, as low values can be entirely normal and high values can be suspect.
Incorrect Answers:
C, E: F-score is used for classification models, not for regression models.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/evaluate-model

**NEW QUESTION: 65**
You are analyzing a raw dataset that requires cleaning.
You must perform transformations and manipulations by using Azure Machine Learning Studio.
You need to identify the correct modules to perform the transformations.
Which modules should you choose? To answer, drag the appropriate modules to the correct scenarios. Each module may be used once, more than once, or not at all.
You may need to drag the split bar between panes or scroll to view content.
NOTE: Each correct selection is worth one point.
**Answer:**

Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/smote
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/convert-to-indicator-values

**NEW QUESTION: 66**
You are evaluating a completed binary classification machine learning model.
You need to use the precision as the valuation metric.
Which visualization should you use?
**A.** Binary classification confusion matrix
**B.** box plot
**C.** coefficient of determination
**D.** Gradient descent
**Answer: (SHOW ANSWER)**

**NEW QUESTION: 67**
You are creating a machine learning model in Python. The provided dataset contains several numerical columns and one text column.
*Biker
*Cars
*Vans
*Boats
You are building a regression model using the scikit- learn Python package.
You need to transform the text data to be compatible with the scikit-learn Python package How should you complete the code segment? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
**Answer:**

**NEW QUESTION: 68**
You are developing a deep learning model by using TensorFlow. You plan to run the model training workload on an Azure Machine Learning Compute Instance.
You must use CUDA-based model training.
You need to provision the Compute Instance.
Which two virtual machines sizes can you use? To answer, select the appropriate virtual machine sizes in the answer area.
NOTE: Each correct selection is worth one point.
**Answer:**
Explanation:
CUDA is a parallel computing platform and programming model developed by Nvidia for general computing on its own GPUs (graphics processing units). CUDA enables developers to speed up compute-intensive applications by harnessing the power of GPUs for the parallelizable part of the computation.
Reference:

https://www.infoworld.com/article/3299703/what-is-cuda-parallel-programming-for-gpus.html

**NEW QUESTION: 69**
You plan to run a script as an experiment using a Script Run Configuration. The script uses modules from the scipy library as well as several Python packages that are not typically installed in a default conda environment.
You plan to run the experiment on your local workstation for small datasets and scale out the experiment by running it on more powerful remote compute clusters for larger datasets.
You need to ensure that the experiment runs successfully on local and remote compute with the least administrative effort.
What should you do?
**A.** Do not specify an environment in the run configuration for the experiment. Run the experiment by using the default environment.
**B.** Create a virtual machine (VM) with the required Python configuration and attach the VM as a compute target. Use this compute target for all experiment runs.
**C.** Create and register an Environment that includes the required packages. Use this Environment for all experiment runs.
**D.** Create a config.yaml file defining the conda packages that are required and save the file in the experiment folder.
**E.** Always run the experiment with an Estimator by using the default packages.
**Answer: C (LEAVE A REPLY)**
If you have an existing Conda environment on your local computer, then you can use the service to create an environment object. By using this strategy, you can reuse your local interactive environment on remote runs.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-use-environments

**NEW QUESTION: 70**
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You create a model to forecast weather conditions based on historical data.
You need to create a pipeline that runs a processing script to load data from a datastore and pass the processed data to a machine learning model training script.
Solution: Run the following code:
Does the solution meet the goal?
**A.** Yes
**B.** No
**Answer: B (LEAVE A REPLY)**
Explanation

Note: Data used in pipeline can be produced by one step and consumed in another step by providing a PipelineData object as an output of one step and an input of one or more subsequent steps.
Compare with this example, the pipeline train step depends on the process_step_output output of the pipeline process step:
from azureml.pipeline.core import Pipeline, PipelineData
from azureml.pipeline.steps import PythonScriptStep
datastore = ws.get_default_datastore()
process_step_output = PipelineData("processed_data", datastore=datastore) process_step = PythonScriptStep(script_name="process.py", arguments=["--data_for_train", process_step_output], outputs=[process_step_output], compute_target=aml_compute, source_directory=process_directory)
train_step = PythonScriptStep(script_name="train.py", arguments=["--data_for_train", process_step_output], inputs=[process_step_output], compute_target=aml_compute, source_directory=train_directory) pipeline = Pipeline(workspace=ws, steps=[process_step, train_step]) Reference:
https://docs.microsoft.com/en-us/python/api/azureml-pipeline-core/azureml.pipeline.core.pipelinedata?view=azu

**NEW QUESTION: 71**
You are creating an experiment by using Azure Machine Learning Studio.
You must divide the data into four subsets for evaluation. There is a high degree of missing values in the data.
You must prepare the data for analysis.
You need to select appropriate methods for producing the experiment.
Which three modules should you run in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.
NOTE: More than one order of answer choices is correct. You will receive credit for any of the correct orders you select.
**Answer:**
Explanation
The Clean Missing Data module in Azure Machine Learning Studio, to remove, replace, or infer missing values.

**NEW QUESTION: 72**
You use Azure Machine Learning to deploy a model as a real-time web service.
You need to create an entry script for the service that ensures that the model is loaded when the service starts and is used to score new data as it is received.
Which functions should you include in the script? To answer, drag the appropriate functions to the correct actions. Each function may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content NOTE: Each correct selection is worth one point.
**Answer:**
Explanation:
Box 1: init()

The entry script has only two required functions, init() and run(data). These functions are used to initialize the service at startup and run the model using request data passed in by a client. The rest of the script handles loading and running the model(s).
Box 2: run()
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-deploy-existing-model

**NEW QUESTION: 73**
You create a multi-class image classification deep learning experiment by using the PyTorch framework. You plan to run the experiment on an Azure Compute cluster that has nodes with GPU's.
You need to define an Azure Machine Learning service pipeline to perform the monthly retraining of the image classification model. The pipeline must run with minimal cost and minimize the time required to train the model.
Which three pipeline steps should you run in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.
**Answer:**
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-train-pytorch

**NEW QUESTION: 74**
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You are analyzing a numerical dataset which contains missing values in several columns.
You must clean the missing values using an appropriate operation without affecting the dimensionality of the feature set.
You need to analyze a full dataset to include all values.
Solution: Replace each missing value using the Multiple Imputation by Chained Equations (MICE) method.
Does the solution meet the goal?
**A.** Yes
**B.** No
**Answer: (SHOW ANSWER)**
Explanation/Reference:
Explanation:
Replace using MICE: For each missing value, this option assigns a new value, which is calculated by using a method described in the statistical literature as "Multivariate Imputation using Chained Equations" or "Multiple Imputation by Chained Equations". With a multiple imputation method, each variable with missing data is modeled conditionally using the other variables in the data before filling in the missing values.

Note: Multivariate imputation by chained equations (MICE), sometimes called "fully conditional specification" or "sequential regression multiple imputation" has emerged in the statistical literature as one principled method of addressing missing data. Creating multiple imputations, as opposed to single imputations, accounts for the statistical uncertainty in the imputations. In addition, the chained equations approach is very flexible and can handle variables of varying types (e.g., continuous or binary) as well as complexities such as bounds or survey skip patterns.
References:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3074241/
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/clean-missing-data

**NEW QUESTION: 75**
You need to configure the Feature Based Feature Selection module based on the experiment requirements and datasets.
How should you configure the module properties? To answer, select the appropriate options in the dialog box in the answer area.
NOTE: Each correct selection is worth one point.
**Answer:**
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/filter-based-feature-selection

**NEW QUESTION: 76**
You plan to use Hyperdrive to optimize the hyperparameters selected when training a model. You create the following code to define options for the hyperparameter experiment
For each of the following statements, select Yes if the statement is true. Otherwise, select No. NOTE: Each correct selection is worth one point.
**Answer:**
Reference:
https://docs.microsoft.com/en-us/python/api/azureml-train-core/azureml.train.hyperdrive.hyperdriveconfig
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-tune-hyperparameters



**NEW QUESTION: 77**
You deploy a model as an Azure Machine Learning real-time web service using the following code.

The deployment fails.
You need to troubleshoot the deployment failure by determining the actions that were performed during deployment and identifying the specific action that failed.
Which code segment should you run?
**A.** service.get_logs()
**B.** service.state
**C.** service.serialize()
**D.** service.update_deployment_state()
**Answer: (SHOW ANSWER)**
You can print out detailed Docker engine log messages from the service object. You can view the log for ACI, AKS, and Local deployments. The following example demonstrates how to print the logs.
# if you already have the service object handy
print(service.get_logs())
# if you only know the name of the service (note there might be multiple services with the same name but different version number) print(ws.webservices['mysvc'].get_logs()) Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-troubleshoot-deployment

**NEW QUESTION: 78**
You are creating a new Azure Machine Learning pipeline using the designer.
The pipeline must train a model using data in a comma-separated values (CSV) file that is published on a website. You have not created a dataset for this file.
You need to ingest the data from the CSV file into the designer pipeline using the minimal administrative effort.
Which module should you add to the pipeline in Designer?
**A.** Convert to CSV
**B.** Enter Data Manually
**C.** Import Data
**D.** Dataset
**Answer: (SHOW ANSWER)**
The preferred way to provide data to a pipeline is a Dataset object. The Dataset object points to data that lives in or is accessible from a datastore or at a Web URL. The Dataset class is abstract, so you will create an instance of either a FileDataset (referring to one or more files) or a TabularDataset that's created by from one or more files with delimited columns of data.
Example:
from azureml.core import Dataset
iris_tabular_dataset = Dataset.Tabular.from_delimited_files([(def_blob_store, 'train-dataset/iris.csv')])
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-create-your-first-pipeline

**NEW QUESTION: 79**
You are solving a classification task.

The dataset is imbalanced.
You need to select an Azure Machine Learning Studio module to improve the classification accuracy.
Which module should you use?
**A.** Fisher Linear Discriminant Analysis.
**B.** Filter Based Feature Selection
**C.** Synthetic Minority Oversampling Teachnique (SMOTE)
**D.** Permutation Feature Importance
**Answer: A (LEAVE A REPLY)**
Topic 1, Case Study 2
Overwiew
You are a data scientist for Fabrikam Residences, a company specializing in quality private and commercial property in the United States. Fabrikam Residences is considering expanding into Europe and has asked you to investigate prices for private residences m major European cities. You use Azure Machine Learning Studio to measure the- median value of properties. You produce a regression model to predict property prices by using the Linear Regression and Bayesian Linear Regression modules.
Datasets
There are two datasets in CSV format that contain property details for two cities. London and Paris, with the following columns:
The two datasets have been added to Azure Machine Learning Studio as separate datasets and included as the starting point of the experiment.
Datasets issues
The Accessibility ToHighway column in both datasets contains missing values. The missing data must be replaced with new data so that it is modeled conditionally using the other variables in the data before filling in the missing values.
Columns in each dataset contain missing and null values. The datasets also contain many outliers. The Age column has a high proportion of outliers. You need to remove the rows that have outliers in the Age column.
The MedianValue and AvgRoomsInHouse columms both hold data in numeric format. You need to select a feature selection algorithm to analyze the relationship between the two columns in more detail.
Model fit
The model shows signs of over fitting. You need to produce a more refined regression model that reduces the over fitting.
Experiment requirement
You must set up the experiment to cross-validate the Linear Regression And Bayesian Linear Regression modules to evaluate performance.
In cacti case, the predictor of the dataset is the column named MedianValue. An initial investigation showed that the datasets are Identical in structure apart from the MedianValue column. The smaller Paris dataset Paris dataset contains the MedianValue in text format, whereas the larger London dataset contains the Median Value in numerical format. You must ensure that the data type of the MedianValue column of the Paris dataset matches the Structure of the London dataset.
You must prioritize the columns of data for predicting the outcome. You must use non-parametric statistics to measure- the relationships.

You must use a feature selection algorithm to analyze the relationship between the MedianValue and AvgRoomsInHouse columns.

Model training

Given a trained model and a test dataset, you need to compute the permutation feature importance scores of feature variables. You need to set up the Permutation Feature Importance module to select the correct melt to investigate the model's accuracy and replicate the findings.

You want to configure hyper parameters the model learning process to speed the learning phase by using hyper parameters. In addition, this configuration should cancel the lowest performing runs at each evaluation interval, thereby directing effort and resources towards models that are more likely to be successful.

You are concerned that the model might not efficiently use computer resources in hyper parameter tuning. You also are concerned that the model might prevent an increase in the overall tuning time. Therefore, you need to implement an early stopping criterion on models that provides savings without terminating promising jobs.

Testing

You must produce multiple partitions of a dataset based on sampling using the Partition and Sample module Azure Machine Learning Studio You must create three partitions foe cross -validation. You must also configure validation process so that the rows in the test and training datasets are divided evenly by properties that are near each city's main river. The data that identifies that a property is near a river is held m the column named NextToRiver. You want to complete this task before the data goes through the sampling process.

When you train a Linear Regression module using a property dataset that shows data for property prices for large city, you need to determine the best features to use in a model. You can choose standard metrics provided to measure performance before and after the feature importance process completes. You must ensure that the distribution of the features across multiple training models is consistent.

Data visualization

You need to provide the test results to the Fabrikam Residences team. You create data visualizations to and in presenting the results.

You must produce a Receiver Operating Characteristic (ROC) curve to conduct a diagnostic test evaluation of the model. You need to select appropriate methods for producing the ROC curve in Azure Machine Learning Studio to compare the Two-Class Decision Jungle modules and the Two Class Decision Jungle modules with one another.

**NEW QUESTION: 80**

You need to identify the methods for dividing the data according to the testing requirements.

Which properties should you select? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

**Answer:**

Explanation:

Scenario: Testing

You must produce multiple partitions of a dataset based on sampling using the Partition and Sample module in Azure Machine Learning Studio.

Box 1: Assign to folds

Use Assign to folds option when you want to divide the dataset into subsets of the data. This option is also useful when you want to create a custom number of folds for cross-validation, or to split rows into several groups.
Not Head: Use Head mode to get only the first n rows. This option is useful if you want to test a pipeline on a small number of rows, and don't need the data to be balanced or sampled in any way.
Not Sampling: The Sampling option supports simple random sampling or stratified random sampling. This is useful if you want to create a smaller representative sample dataset for testing.
Box 2: Partition evenly
Specify the partitioner method: Indicate how you want data to be apportioned to each partition, using these options:
* Partition evenly: Use this option to place an equal number of rows in each partition. To specify the number of output partitions, type a whole number in the Specify number of folds to split evenly into text box.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/algorithm-module-reference/partition-and-sample

**NEW QUESTION: 81**
You need to replace the missing data in the AccessibilityToHighway columns.
How should you configure the Clean Missing Data module? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
**Answer:**

**NEW QUESTION: 82**
You have a dataset that contains 2,000 rows. You are building a machine learning classification model by using Azure Learning Studio. You add a Partition and Sample module to the experiment.
You need to configure the module. You must meet the following requirements:
* Divide the data into subsets
* Assign the rows into folds using a round-robin method
* Allow rows in the dataset to be reused
How should you configure the module? To answer, select the appropriate options in the dialog box in the answer area.
NOTE: Each correct selection is worth one point.
**Answer:**
Explanation
Use the Split data into partitions option when you want to divide the dataset into subsets of the data. This option is also useful when you want to create a custom number of folds for cross-validation, or to split rows into several groups.
* Add the Partition and Sample module to your experiment in Studio (classic), and connect the dataset.
* For Partition or sample mode, select Assign to Folds.
* Use replacement in the partitioning: Select this option if you want the sampled row to be put back into the pool of rows for potential reuse. As a result, the same row might be assigned to several folds.

* If you do not use replacement (the default option), the sampled row is not put back into the pool of rows for potential reuse. As a result, each row can be assigned to only one fold.
* Randomized split: Select this option if you want rows to be randomly assigned to folds.
If you do not select this option, rows are assigned to folds using the round-robin method.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/partition-and-sample

**NEW QUESTION: 83**
You need to modify the inputs for the global penalty event model to address the bias and variance issue.
Which three actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.
**Answer:**

**NEW QUESTION: 84**
You create a binary classification model by using Azure Machine Learning Studio.
You must tune hyperparameters by performing a parameter sweep of the model. The parameter sweep must meet the following requirements:
* iterate all possible combinations of hyperparameters
* minimize computing resources required to perform the sweep
* You need to perform a parameter sweep of the model.
Which parameter sweep mode should you use?
**A.** Random sweep
**B.** Sweep clustering
**C.** Entire grid
**D.** Random grid
**E.** Random seed
**Answer: D (LEAVE A REPLY)**
Maximum number of runs on random grid: This option also controls the number of iterations over a random sampling of parameter values, but the values are not generated randomly from the specified range; instead, a matrix is created of all possible combinations of parameter values and a random sampling is taken over the matrix. This method is more efficient and less prone to regional oversampling or undersampling.
If you are training a model that supports an integrated parameter sweep, you can also set a range of seed values to use and iterate over the random seeds as well. This is optional, but can be useful for avoiding bias introduced by seed selection.
Incorrect Answers:
B: If you are building a clustering model, use Sweep Clustering to automatically determine the optimum number of clusters and other parameters.
C: Entire grid: When you select this option, the module loops over a grid predefined by the system, to try different combinations and identify the best learner. This option is useful for cases where you don't know what the best parameter settings might be and want to try all possible combination of values.

E: If you choose a random sweep, you can specify how many times the model should be trained, using a random combination of parameter values.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/tune-model-hyperparameters

**NEW QUESTION: 85**
You are analyzing the asymmetry in a statistical distribution.
The following image contains two density curves that show the probability distribution of two datasets.
Use the drop-down menus to select the answer choice that answers each question based on the information presented in the graphic.
NOTE: Each correct selection is worth one point.
**Answer:**
Explanation
Box 1: Positive skew
Positive skew values means the distribution is skewed to the right.
Box 2: Negative skew
Negative skewness values mean the distribution is skewed to the left.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/compute-elementary-statistics

**NEW QUESTION: 86**
You have a Python script that executes a pipeline. The script includes the following code:
from azureml.core import Experiment
pipeline_run = Experiment(ws, 'pipeline_test').submit(pipeline)
You want to test the pipeline before deploying the script.
You need to display the pipeline run details written to the STDOUT output when the pipeline completes.
Which code segment should you add to the test script?
**A.** pipeline_run.get.metrics()
**B.** pipeline_run.wait_for_completion(show_output=True)
**C.** pipeline_param = PipelineParameter(name="stdout",
default_value="console")
**D.** pipeline_run.get_status()
**Answer: (SHOW ANSWER)**
wait_for_completion: Wait for the completion of this run. Returns the status object after the wait.
Syntax: wait_for_completion(show_output=False, wait_post_processing=False, raise_on_error=True)
Parameter: show_output Indicates whether to show the run output on sys.stdout.

**NEW QUESTION: 87**
You are performing a classification task in Azure Machine Learning Studio.

You must prepare balanced testing and training samples based on a provided data set.
You need to split the data with a 0.75:0.25 ratio.
Which value should you use for each parameter? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

**Answer:**

Explanation:
Box 1: Split rows
Use the Split Rows option if you just want to divide the data into two parts. You can specify the percentage of data to put in each split, but by default, the data is divided 50-50.
You can also randomize the selection of rows in each group, and use stratified sampling. In stratified sampling, you must select a single column of data for which you want values to be apportioned equally among the two result datasets.
Box 2: 0.75
If you specify a number as a percentage, or if you use a string that contains the "%" character, the value is interpreted as a percentage. All percentage values must be within the range (0, 100), not including the values 0 and 100.
Box 3: Yes
To ensure splits are balanced.
Box 4: No
If you use the option for a stratified split, the output datasets can be further divided by subgroups, by selecting a strata column.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/split-data

**NEW QUESTION: 88**

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You create an Azure Machine Learning service datastore in a workspace. The datastore contains the following files:
* /data/2018/Q1 .csv
* /data/2018/Q2.csv
* /data/2018/Q3.csv
* /data/2018/Q4.csv
* /data/2019/Q1.csv
All files store data in the following format:
id,M,f2,l
1,1,2,0

2,1,1,1
32,10
You run the following code:
You need to create a dataset named training_data and load the data from all files into a single data frame by using the following code:
Solution: Run the following code:
Does the solution meet the goal?
**A.** No
**B.** Yes
**Answer: B (LEAVE A REPLY)**

**NEW QUESTION: 89**
You are preparing to use the Azure ML SDK to run an experiment and need to create compute. You run the following code:
For each of the following statements, select Yes if the statement is true. Otherwise, select No.
NOTE: Each correct selection is worth one point.
**Answer:**
Reference:
https://notebooks.azure.com/azureml/projects/azureml-getting-started/html/how-to-use-azureml/training/train-on-amlcompute/train-on-amlcompute.ipynb
https://docs.microsoft.com/en-us/python/api/azureml-core/azureml.core.compute.computetarget

**NEW QUESTION: 90**
You are creating a machine learning model in Python. The provided dataset contains several numerical columns and one text column.
*Biker
*Cars
*Vans
*Boats
You are building a regression model using the scikit- learn Python package.
You need to transform the text data to be compatible with the scikit-learn Python package How should you complete the code segment? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
**Answer:**

**NEW QUESTION: 91**
You need to configure the Permutation Feature Importance module for the model training requirements.
What should you do? To answer, select the appropriate options in the dialog box in the answer area.
NOTE: Each correct selection is worth one point.
**Answer:**

Explanation
Box 1: 500
For Random seed, type a value to use as seed for randomization. If you specify 0 (the default), a number is generated based on the system clock.
A seed value is optional, but you should provide a value if you want reproducibility across runs of the same experiment.
Here we must replicate the findings.
Box 2: Mean Absolute Error
Scenario: Given a trained model and a test dataset, you must compute the Permutation Feature Importance scores of feature variables. You need to set up the Permutation Feature Importance module to select the correct metric to investigate the model's accuracy and replicate the findings.
Regression. Choose one of the following: Precision, Recall, Mean Absolute Error , Root Mean Squared Error, Relative Absolute Error, Relative Squared Error, Coefficient of Determination References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/permutation-feature-importan



**NEW QUESTION: 92**
You need to select a feature extraction method.
Which method should you use?
**A.** Mutual information
**B.** Pearson's correlation
**C.** Spearman correlation
**D.** Fisher Linear Discriminant Analysis
**Answer: C (LEAVE A REPLY)**
Spearman's rank correlation coefficient assesses how well the relationship between two variables can be described using a monotonic function.
Note: Both Spearman's and Kendall's can be formulated as special cases of a more general correlation coefficient, and they are both appropriate in this scenario.
Scenario: The MedianValue and AvgRoomsInHouse columns both hold data in numeric format. You need to select a feature selection algorithm to analyze the relationship between the two columns in more detail.
Incorrect Answers:

B: The Spearman correlation between two variables is equal to the Pearson correlation between the rank values of those two variables; while Pearson's correlation assesses linear relationships, Spearman's correlation assesses monotonic relationships (whether linear or not).
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/feature-selection-modules
Perform Feature Engineering Question Set 3

**NEW QUESTION: 93**
You have an Azure blob container that contains a set of TSV files. The Azure blob container is registered as a datastore for an Azure Machine Learning service workspace. Each TSV file uses the same data schema.
You plan to aggregate data for all of the TSV files together and then register the aggregated data as a dataset in an Azure Machine Learning workspace by using the Azure Machine Learning SDK for Python.
You run the following code.
For each of the following statements, select Yes if the statement is true. Otherwise, select No.
NOTE: Each correct selection is worth one point.
**Answer:**
Explanation
Box 1: No
FileDataset references single or multiple files in datastores or from public URLs. The TSV files need to be parsed.
Box 2: Yes
to_path() gets a list of file paths for each file stream defined by the dataset.
Box 3: Yes
TabularDataset.to_pandas_dataframe loads all records from the dataset into a pandas DataFrame.
TabularDataset represents data in a tabular format created by parsing the provided file or list of files.
Note: TSV is a file extension for a tab-delimited file used with spreadsheet software. TSV stands for Tab Separated Values. TSV files are used for raw data and can be imported into and exported from spreadsheet software. TSV files are essentially text files, and the raw data can be viewed by text editors, though they are often used when moving raw data between spreadsheets.
Reference:
https://docs.microsoft.com/en-us/python/api/azureml-core/azureml.data.tabulardataset

**NEW QUESTION: 94**
You are training machine learning models in Azure Machine Learning. You use Hyperdrive to tune the hyperparameters. In previous model training and tuning runs, many models showed similar performance. You need to select an early termination policy that meets the following requirements:
* accounts for the performance of all previous runs when evaluating the current run
* avoids comparing the current run with only the best performing run to date Which two early termination policies should you use? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.
**A.** Bandit

**B.** Median stopping
**C.** Default
**D.** Truncation selection
**Answer: B,C (LEAVE A REPLY)**
Explanation
The Median Stopping policy computes running averages across all runs and cancels runs whose best performance is worse than the median of the running averages.
If no policy is specified, the hyperparameter tuning service will let all training runs execute to completion.
Reference:
https://docs.microsoft.com/en-us/python/api/azureml-train-core/azureml.train.hyperdrive.medianstoppingpolicy
https://docs.microsoft.com/en-us/python/api/azureml-train-core/azureml.train.hyperdrive.truncationselectionpolic
https://docs.microsoft.com/en-us/python/api/azureml-train-core/azureml.train.hyperdrive.banditpolicy

**NEW QUESTION: 95**
You are using C-Support Vector classification to do a multi-class classification with an unbalanced training dataset. The C-Support Vector classification using Python code shown below:
You need to evaluate the C-Support Vector classification code.
Which evaluation statement should you use? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
**Answer:**
Explanation
Box 1: Automatically adjust weights inversely proportional to class frequencies in the input data The "balanced" mode uses the values of y to automatically adjust weights inversely proportional to class frequencies in the input data as n_samples / (n_classes * np.bincount(y)).
Box 2: Penalty parameter
Parameter: C : float, optional (default=1.0)
Penalty parameter C of the error term.
References:
https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

**NEW QUESTION: 96**
You need to define a process for penalty event detection.
Which three actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.
**Answer:**

**NEW QUESTION: 97**
The finance team asks you to train a model using data in an Azure Storage blob container named finance-data.

You need to register the container as a datastore in an Azure Machine Learning workspace and ensure that an error will be raised if the container does not exist.
How should you complete the code? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
**Answer:**
Reference:
https://docs.microsoft.com/en-us/python/api/azureml-core/azureml.core.datastore.datastore

**NEW QUESTION: 98**
You create a script for training a machine learning model in Azure Machine Learning service.
You create an estimator by running the following code:
For each of the following statements, select Yes if the statement is true. Otherwise, select No.
NOTE: Each correct selection is worth one point.
**Answer:**
Explanation:
Box 1: Yes
Parameter source_directory is a local directory containing experiment configuration and code files needed for a training job.
Box 2: Yes
script_params is a dictionary of command-line arguments to pass to the training script specified in entry_script.
Box 3: No
Box 4: Yes
The conda_packages parameter is a list of strings representing conda packages to be added to the Python environment for the experiment.

**NEW QUESTION: 99**
An organization creates and deploys a multi-class image classification deep learning model that uses a set of labeled photographs.
The software engineering team reports there is a heavy inferencing load for the prediction web services during the summer. The production web service for the model fails to meet demand despite having a fully-utilized compute cluster where the web service is deployed.
You need to improve performance of the image classification web service with minimal downtime and minimal administrative effort.
What should you advise the IT Operations team to do?
**A.** Create a new compute cluster by using larger VM sizes for the nodes, redeploy the web service to that cluster, and update the DNS registration for the service endpoint to point to the new cluster.
**B.** Increase the minimum node count of the compute cluster where the web service is deployed.
**C.** Increase the VM size of nodes in the compute cluster where the web service is deployed.
**D.** Increase the node count of the compute cluster where the web service is deployed.
**Answer: (SHOW ANSWER)**

**NEW QUESTION: 100**
You use the Two-Class Neural Network module in Azure Machine Learning Studio to build a binary classification model. You use the Tune Model Hyperparameters module to tune accuracy for the model.
You need to select the hyperparameters that should be tuned using the Tune Model Hyperparameters module.
Which two hyperparameters should you use? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.
**A.** Number of hidden nodes
**B.** Learning Rate
**C.** The type of the normalizer
**D.** Number of learning iterations
**E.** Hidden layer specification
**Answer: D,E (LEAVE A REPLY)**
Explanation
D: For Number of learning iterations, specify the maximum number of times the algorithm should process the training cases.
E: For Hidden layer specification, select the type of network architecture to create.
Between the input and output layers you can insert multiple hidden layers. Most predictive tasks can be accomplished easily with only one or a few hidden layers.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/two-class-neural-network

**NEW QUESTION: 101**
You are producing a multiple linear regression model in Azure Machine learning Studio.
Several independent variables are highly correlated.
You need to select appropriate methods for conducting elective feature engineering on all the data.
Which three actions should you perform in sequence? To answer, move the appropriate Actions from the list of actions to the answer area and arrange them in the correct order.
**Answer:**

**NEW QUESTION: 102**
You are tuning a hyperparameter for an algorithm. The following table shows a data set with different hyperparameter, training error, and validation errors.
**Answer:**

**NEW QUESTION: 103**
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.

After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You are creating a new experiment in Azure Machine Learning Studio.
One class has a much smaller number of observations than the other classes in the training set.
You need to select an appropriate data sampling strategy to compensate for the class imbalance.
Solution: You use the Scale and Reduce sampling mode.
Does the solution meet the goal?
**A.** Yes
**B.** No
**Answer: (SHOW ANSWER)**
Explanation
Explanation:
Instead use the Synthetic Minority Oversampling Technique (SMOTE) sampling mode.
Note: SMOTE is used to increase the number of underepresented cases in a dataset used for machine learning. SMOTE is a better way of increasing the number of rare cases than simply duplicating existing cases.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/smote

**NEW QUESTION: 104**
You need to configure the Permutation Feature Importance module for the model training requirements.
What should you do? To answer, select the appropriate options in the dialog box in the answer area.
NOTE: Each correct selection is worth one point.
**Answer:**
Explanation:
Box 1: 500
For Random seed, type a value to use as seed for randomization. If you specify 0 (the default), a number is generated based on the system clock.
A seed value is optional, but you should provide a value if you want reproducibility across runs of the same experiment.
Here we must replicate the findings.
Box 2: Mean Absolute Error
Scenario: Given a trained model and a test dataset, you must compute the Permutation Feature Importance scores of feature variables. You need to set up the Permutation Feature Importance module to select the correct metric to investigate the model's accuracy and replicate the findings.
Regression. Choose one of the following: Precision, Recall, Mean Absolute Error , Root Mean Squared Error, Relative Absolute Error, Relative Squared Error, Coefficient of Determination References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/permutation-feature-importance

**NEW QUESTION: 105**

You are a data scientist building a deep convolutional neural network (CNN) for image classification.
The CNN model you built shows signs of overfitting.
You need to reduce overfitting and converge the model to an optimal fit.
Which two actions should you perform? Each correct answer presents a complete solution.
NOTE: Each correct selection is worth one point.
**A.** Reduce the amount of training data.
**B.** Add an additional dense layer with 64 input units
**C.** Add L1/L2 regularization.
**D.** Use training data augmentation
**E.** Add an additional dense layer with 512 input units.
**Answer: A,C (LEAVE A REPLY)**
Reference:
https://machinelearningmastery.com/how-to-reduce-overfitting-in-deep-learning-with-weight-regularization/
https://en.wikipedia.org/wiki/Convolutional_neural_network

**NEW QUESTION: 106**
You are analyzing a raw dataset that requires cleaning.
You must perform transformations and manipulations by using Azure Machine Learning Studio.
You need to identify the correct modules to perform the transformations.
Which modules should you choose? To answer, drag the appropriate modules to the correct scenarios. Each module may be used once, more than once, or not at all.
You may need to drag the split bar between panes or scroll to view content.
NOTE: Each correct selection is worth one point.
**Answer:**
Explanation
Box 1: Clean Missing Data
Box 2: SMOTE
Use the SMOTE module in Azure Machine Learning Studio to increase the number of underepresented cases in a dataset used for machine learning. SMOTE is a better way of increasing the number of rare cases than simply duplicating existing cases.
Box 3: Convert to Indicator Values
Use the Convert to Indicator Values module in Azure Machine Learning Studio. The purpose of this module is to convert columns that contain categorical values into a series of binary indicator columns that can more easily be used as features in a machine learning model.
Box 4: Remove Duplicate Rows
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/smote
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/convert-to-indicator-values

**NEW QUESTION: 107**
You create machine learning models by using Azure Machine Learning.
You plan to train and score models by using a variety of compute contexts. You also plan to create a new compute resource in Azure Machine Learning studio.
You need to select the appropriate compute types.
Which compute types should you select? To answer, drag the appropriate compute types to the correct requirements. Each compute type may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.
NOTE: Each correct selection is worth one point.
**Answer:**
Explanation:
Box 1: Attached compute
Box 2: Inference cluster
Box 3: Training cluster
Box 4: Attached compute

**NEW QUESTION: 108**
You are building a regression model tot estimating the number of calls during an event.
You need to determine whether the feature values achieve the conditions to build a Poisson regression model.
Which two conditions must the feature set contain? I ach correct answer presents part of the solution. NOTE: Each correct selection is worth one point.
**A.** The label data must be a negative value.
**B.** The label data can be positive or negative,
**C.** The label data must be a positive value
**D.** The label data must be non discrete.
**E.** The data must be whole numbers.
**Answer: C,E (LEAVE A REPLY)**
Poisson regression is intended for use in regression models that are used to predict numeric values, typically counts. Therefore, you should use this module to create your regression model only if the values you are trying to predict fit the following conditions:
The response variable has a Poisson distribution.
Counts cannot be negative. The method will fail outright if you attempt to use it with negative labels.

A Poisson distribution is a discrete distribution; therefore, it is not meaningful to use this method with non-whole numbers.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/poisson-regression

**NEW QUESTION: 109**
You use Data Science Virtual Machines (DSVMs) for Windows and Linux in Azure.
You need to access the DSVMs.
Which utilities should you use? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
**Answer:**

**NEW QUESTION: 110**
You configure a Deep Learning Virtual Machine for Windows.
You need to recommend tools and frameworks to perform the following:
Build deep neural network (DNN) models
Perform interactive data exploration and visualization
Which tools and frameworks should you recommend? To answer, drag the appropriate tools to the correct tasks. Each tool may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.
NOTE: Each correct selection is worth one point.
**Answer:**
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/train-vowpal-wabbit-version-8-model
https://docs.microsoft.com/en-us/azure/architecture/data-guide/scenarios/interactive-data-exploration

**NEW QUESTION: 111**
You have a dataset created for multiclass classification tasks that contains a normalized numerical feature set with 10,000 data points and 150 features.
You use 75 percent of the data points for training and 25 percent for testing. You are using the scikit-learn machine learning library in Python. You use X to denote the feature set and Y to denote class labels.
You create the following Python data frames:
You need to apply the Principal Component Analysis (PCA) method to reduce the dimensionality of the feature set to 10 features in both training and testing sets.
How should you complete the code segment? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
**Answer:**
Explanation:
Box 1: PCA(n_components = 10)

Need to reduce the dimensionality of the feature set to 10 features in both training and testing sets.
Example:
from sklearn.decomposition import PCA
pca = PCA(n_components=2) ;2 dimensions
principalComponents = pca.fit_transform(x)
Box 2: pca
fit_transform(X[, y])fits the model with X and apply the dimensionality reduction on X.
Box 3: transform(x_test)
transform(X) applies dimensionality reduction to X.
References:
https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

**NEW QUESTION: 112**
You have a feature set containing the following numerical features: X, Y, and Z.
The Poisson correlation coefficient (r-value) of X, Y, and Z features is shown in the following image:
Use the drop-down menus to select the answer choice that answers each question based on the information presented in the graphic.
NOTE: Each correct selection is worth one point.
**Answer:**
Explanation
Box 1: 0.859122
Box 2: a positively linear relationship
+1 indicates a strong positive linear relationship
-1 indicates a strong negative linear correlation
0 denotes no linear relationship between the two variables.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/compute-linear-correlation

**NEW QUESTION: 113**
You are using Azure Machine Learning to train machine learning models. You need a compute target on which to remotely run the training script. You run the following Python code:
**Answer:**
Explanation:
Box 1: Yes
The compute is created within your workspace region as a resource that can be shared with other users.
Box 2: Yes
It is displayed as a compute cluster.
View compute targets
1. To see all compute targets for your workspace, use the following steps:
2. Navigate to Azure Machine Learning studio.

3. Under Manage, select Compute.
4. Select tabs at the top to show each type of compute target.
Box 3: Yes
min_nodes is not specified, so it defaults to 0.
Reference:
https://docs.microsoft.com/en-us/python/api/azureml-core/azureml.core.compute.amlcompute.amlcomputeprovisioningconfiguration
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-create-attach-compute-studio

**NEW QUESTION: 114**
You create an Azure Machine Learning workspace and set up a development environment. You plan to train a deep neural network (DNN) by using the Tensorflow framework and by using estimators to submit training scripts.
You must optimize computation speed for training runs.
You need to choose the appropriate estimator to use as well as the appropriate training compute target configuration.
Which values should you use? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
**Answer:**
Reference:
https://docs.microsoft.com/en-us/python/api/azureml-train-core/azureml.train.dnn

**NEW QUESTION: 115**
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You are using Azure Machine Learning to run an experiment that trains a classification model.
You want to use Hyperdrive to find parameters that optimize the AUC metric for the model. You configure a HyperDriveConfig for the experiment by running the following code:
You plan to use this configuration to run a script that trains a random forest model and then tests it with validation data. The label values for the validation data are stored in a variable named y_test variable, and the predicted probabilities from the model are stored in a variable named y_predicted.
You need to add logging to the script to allow Hyperdrive to optimize hyperparameters for the AUC metric.
Solution: Run the following code:
Does the solution meet the goal?
**A.** Yes
**B.** No
**Answer: (SHOW ANSWER)**
Use a solution with logging.info(message) instead.

Note: Python printing/logging example:
logging.info(message)
Destination: Driver logs, Azure Machine Learning designer
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-debug-pipelines

**NEW QUESTION: 116**
You configure a Deep Learning Virtual Machine for Windows.
You need to recommend tools and frameworks to perform the following:
Build deep neural network (DNN) models
Perform interactive data exploration and visualization
Which tools and frameworks should you recommend? To answer, drag the appropriate tools to the correct tasks. Each tool may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.
NOTE: Each correct selection is worth one point.
**Answer:**

**NEW QUESTION: 117**
You are creating a classification model for a banking company to identify possible instances of credit card fraud. You plan to create the model in Azure Machine Learning by using automated machine learning.
The training dataset that you are using is highly unbalanced.
You need to evaluate the classification model.
Which primary metric should you use?
**A.** normalized_mean_absolute_error
**B.** normalized_root_mean_squared_error
**C.** AUC.weighted
**D.** accuracy
**E.** [spearman_correlation
**Answer: A (LEAVE A REPLY)**

**NEW QUESTION: 118**
You have a dataset created for multiclass classification tasks that contains a normalized numerical feature set with 10,000 data points and 150 features.
You use 75 percent of the data points for training and 25 percent for testing. You are using the scikit-learn machine learning library in Python. You use X to denote the feature set and Y to denote class labels.
You create the following Python data frames:
You need to apply the Principal Component Analysis (PCA) method to reduce the dimensionality of the feature set to 10 features in both training and testing sets.
How should you complete the code segment? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
**Answer:**

Explanation
Box 1: PCA(n_components = 10)
Need to reduce the dimensionality of the feature set to 10 features in both training and testing sets.
Example:
from sklearn.decomposition import PCA
pca = PCA(n_components=2) ;2 dimensions
principalComponents = pca.fit_transform(x)
Box 2: pca
fit_transform(X[, y])fits the model with X and apply the dimensionality reduction on X.
Box 3: transform(x_test)
transform(X) applies dimensionality reduction to X.
References:
https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

**NEW QUESTION: 119**
You have a dataset that includes confidential dat
a. You use the dataset to train a model.
You must use a differential privacy parameter to keep the data of individuals safe and private.
You need to reduce the effect of user data on aggregated results.
What should you do?
**A.** Decrease the value of the epsilon parameter to reduce the amount of noise added to the data
**B.** Increase the value of the epsilon parameter to decrease privacy and increase accuracy
**C.** Decrease the value of the epsilon parameter to increase privacy and reduce accuracy
**D.** Set the value of the epsilon parameter to 1 to ensure maximum privacy
**Answer: C (LEAVE A REPLY)**
Differential privacy tries to protect against the possibility that a user can produce an indefinite number of reports to eventually reveal sensitive data. A value known as epsilon measures how noisy, or private, a report is. Epsilon has an inverse relationship to noise or privacy. The lower the epsilon, the more noisy (and private) the data is.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/concept-differential-privacy

**NEW QUESTION: 120**
You need to correct the model fit issue.
Which three actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.
**Answer:**

**NEW QUESTION: 121**
You need to select a feature extraction method.
Which method should you use?

**A.** Spearman correlation
**B.** Mutual information
**C.** Mann-Whitney test
**D.** Pearson's correlation
**Answer: A (LEAVE A REPLY)**
Spearman's rank correlation coefficient assesses how well the relationship between two variables can be described using a monotonic function.
Note: Both Spearman's and Kendall's can be formulated as special cases of a more general correlation coefficient, and they are both appropriate in this scenario.
Scenario: The MedianValue and AvgRoomsInHouse columns both hold data in numeric format. You need to select a feature selection algorithm to analyze the relationship between the two columns in more detail.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/feature-selection-modules



**NEW QUESTION: 122**
You need to implement a scaling strategy for the local penalty detection data.
Which normalization type should you use?
**A.** Streaming
**B.** Weight
**C.** Batch
**D.** Cosine
**Answer: C (LEAVE A REPLY)**
Explanation
Post batch normalization statistics (PBN) is the Microsoft Cognitive Toolkit (CNTK) version of how to evaluate the population mean and variance of Batch Normalization which could be used in inference Original Paper.
In CNTK, custom networks are defined using the BrainScriptNetworkBuilder and described in the CNTK network description language "BrainScript." Scenario:
Local penalty detection models must be written by using BrainScript.
References:
https://docs.microsoft.com/en-us/cognitive-toolkit/post-batch-normalization-statistics
Topic 1, Case Study 1
Overview

You are a data scientist in a company that provides data science for professional sporting events. Models will be global and local market data to meet the following business goals:
*Understand sentiment of mobile device users at sporting events based on audio from crowd reactions.
*Access a user's tendency to respond to an advertisement.
*Customize styles of ads served on mobile devices.
*Use video to detect penalty events.
Current environment
Requirements
* Media used for penalty event detection will be provided by consumer devices. Media may include images and videos captured during the sporting event and snared using social media. The images and videos will have varying sizes and formats.
* The data available for model building comprises of seven years of sporting event media. The sporting event media includes: recorded videos, transcripts of radio commentary, and logs from related social media feeds feeds captured during the sporting events.
*Crowd sentiment will include audio recordings submitted by event attendees in both mono and stereo Formats.
Advertisements
* Ad response models must be trained at the beginning of each event and applied during the sporting event.
* Market segmentation nxxlels must optimize for similar ad resporr.r history.
* Sampling must guarantee mutual and collective exclusivity local and global segmentation models that share the same features.
* Local market segmentation models will be applied before determining a user's propensity to respond to an advertisement.
* Data scientists must be able to detect model degradation and decay.
* Ad response models must support non linear boundaries features.
* The ad propensity model uses a cut threshold is 0.45 and retrains occur if weighted Kappa deviates from 0.1 +/-5%.
* The ad propensity model uses cost factors shown in the following diagram:
The ad propensity model uses proposed cost factors shown in the following diagram:
Performance curves of current and proposed cost factor scenarios are shown in the following diagram:
Penalty detection and sentiment
Findings
*Data scientists must build an intelligent solution by using multiple machine learning models for penalty event detection.
*Data scientists must build notebooks in a local environment using automatic feature engineering and model building in machine learning pipelines.
*Notebooks must be deployed to retrain by using Spark instances with dynamic worker allocation
*Notebooks must execute with the same code on new Spark instances to recode only the source of the data.
*Global penalty detection models must be trained by using dynamic runtime graph computation during training.
*Local penalty detection models must be written by using BrainScript.

* Experiments for local crowd sentiment models must combine local penalty detection data.
* Crowd sentiment models must identify known sounds such as cheers and known catch phrases. Individual crowd sentiment models will detect similar sounds.
* All shared features for local models are continuous variables.
* Shared features must use double precision. Subsequent layers must have aggregate running mean and standard deviation metrics Available.
segments
During the initial weeks in production, the following was observed:
*Ad response rates declined.
*Drops were not consistent across ad styles.
*The distribution of features across training and production data are not consistent.
Analysis shows that of the 100 numeric features on user location and behavior, the 47 features that come from location sources are being used as raw features. A suggested experiment to remedy the bias and variance issue is to engineer 10 linearly uncorrected features.
Penalty detection and sentiment
*Initial data discovery shows a wide range of densities of target states in training data used for crowd sentiment models.
*All penalty detection models show inference phases using a Stochastic Gradient Descent (SGD) are running too stow.
*Audio samples show that the length of a catch phrase varies between 25%-47%, depending on region.
*The performance of the global penalty detection models show lower variance but higher bias when comparing training and validation sets. Before implementing any feature changes, you must confirm the bias and variance using all training and validation cases.

**NEW QUESTION: 123**
You are performing feature engineering on a dataset.
You must add a feature named CityName and populate the column value with the text London.
You need to add the new feature to the dataset.
Which Azure Machine Learning Studio module should you use?
**A.** Edit Metadata
**B.** Preprocess Text
**C.** Execute Python Script
**D.** Latent Dirichlet Allocation
**Answer: A (LEAVE A REPLY)**
Explanation
Typical metadata changes might include marking columns as features.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/edit-metadata

**NEW QUESTION: 124**

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You are analyzing a numerical dataset which contains missing values in several columns.
You must clean the missing values using an appropriate operation without affecting the dimensionality of the feature set.
You need to analyze a full dataset to include all values.
Solution: Calculate the column median value and use the median value as the replacement for any missing value in the column.
Does the solution meet the goal?
**A.** Yes
**B.** No
**Answer: (SHOW ANSWER)**
Use the Multiple Imputation by Chained Equations (MICE) method.
Reference:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3074241/
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/clean-missing-data

**NEW QUESTION: 125**
You are developing a machine learning, experiment by using Azure. The following images show the input and output of a machine learning experiment:
Use the drop-down menus to select the answer choice that answers each question based on the information presented in the graphic.
NOTE: Each correct selection is worth one point.
**Answer:**

**NEW QUESTION: 126**
You have a model with a large difference between the training and validation error values.
You must create a new model and perform cross-validation.
You need to identify a parameter set for the new model using Azure Machine Learning Studio.
Which module you should use for each step? To answer, drag the appropriate modules to the correct steps. Each module may be used once or more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.
NOTE: Each correct selection is worth one point.
**Answer:**
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/partition-and-sample

**NEW QUESTION: 127**

You create a new Azure subscription. No resources are provisioned in the subscription.
You need to create an Azure Machine Learning workspace.
What are three possible ways to achieve this goal? Each correct answer presents a complete solution.
NOTE: Each correct selection is worth one point.
**A.** Run Python code that uses the Azure ML SDK library and calls the Workspacesreate method with name, subscriptionjd, resource_group, and location parameters.
**B.** Use the Azure Command Line Interface (CLI) with the Azure Machine Learning extension to call the az group create function with -name and -location parameters, and then the az ml workspace create function, specifying -w and -g parameters for the workspace name and resource group.
**C.** Use an Azure Resource Management template that includes a Microsoft.MachineLearningServices/workspaces resource and its dependencies.
**D.** Run Python code that uses the Azure ML SDK library and calls the Workspace.get method with name, subscriptionjd, and resource.group parameters.
**E.** Navigate to Azure Machine Learning studio and create a workspace.
**Answer: (SHOW ANSWER)**

**NEW QUESTION: 128**
You are developing a linear regression model in Azure Machine Learning Studio. You run an experiment to compare different algorithms.
The following image displays the results dataset output:
Use the drop-down menus to select the answer choice that answers each question based on the information presented in the image.
NOTE: Each correct selection is worth one point.
**Answer:**
Explanation:
Box 1: Boosted Decision Tree Regression
Mean absolute error (MAE) measures how close the predictions are to the actual outcomes; thus, a lower score is better.
Box 2:
Online Gradient Descent: If you want the algorithm to find the best parameters for you, set Create trainer mode option to Parameter Range. You can then specify multiple values for the algorithm to try.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/evaluate-model
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/linear-regression

**NEW QUESTION: 129**
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

You create an Azure Machine Learning service datastore in a workspace. The datastore contains the following files:
* /data/2018/Q1 .csv
* /data/2018/Q2.csv
* /data/2018/Q3.csv
* /data/2018/Q4.csv
* /data/2019/Q1.csv
All files store data in the following format:
id,M,f2,l
1,1,2,0
2,1,1,1
32,10
You run the following code:
You need to create a dataset named training_data and load the data from all files into a single data frame by using the following code:
Solution: Run the following code:
Does the solution meet the goal?
**A.** Yes
**B.** No
**Answer: A (LEAVE A REPLY)**

**NEW QUESTION: 130**
You are analyzing the asymmetry in a statistical distribution.
The following image contains two density curves that show the probability distribution of two datasets.
Use the drop-down menus to select the answer choice that answers each question based on the information presented in the graphic.
NOTE: Each correct selection is worth one point.
**Answer:**
Explanation:
Box 1: Positive skew
Positive skew values means the distribution is skewed to the right.
Box 2: Negative skew
Negative skewness values mean the distribution is skewed to the left.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/compute-elementary-statistics

**NEW QUESTION: 131**

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You plan to use a Python script to run an Azure Machine Learning experiment. The script creates a reference to the experiment run context, loads data from a file, identifies the set of unique values for the label column, and completes the experiment run:

```
from azureml.core import Run
import pandas as pd
run = Run.get_context()
data = pd.read_csv('data.csv')
label_vals = data['label'].unique()
# Add code to record metrics here
run.complete()
```

The experiment must record the unique labels in the data as metrics for the run that can be reviewed later.
You must add code to the script to record the unique label values as run metrics at the point indicated by the comment.
Solution: Replace the comment with the following code:

```
run.log_table('Label Values', label_vals)
```

Does the solution meet the goal?
**A.** Yes
**B.** No
**Answer: (SHOW ANSWER)**
Explanation
Instead use the run_log function to log the contents in label_vals:

```
for label_val in label_vals:
run.log('Label Values', label_val)
```

Reference:
https://www.element61.be/en/resource/azure-machine-learning-services-complete-toolbox-ai

**NEW QUESTION: 132**
You create an experiment in Azure Machine Learning Studio. You add a training dataset that contains 10,000 rows. The first 9,000 rows represent class 0 (90 percent).
The remaining 1,000 rows represent class 1 (10 percent).
The training set is imbalances between two classes. You must increase the number of training examples for class 1 to 4,000 by using 5 data rows. You add the Synthetic Minority Oversampling Technique (SMOTE) module to the experiment.
You need to configure the module.
Which values should you use? To answer, select the appropriate options in the dialog box in the answer area.
NOTE: Each correct selection is worth one point.

**Answer:**

Explanation:

Box 1: 300

You type 300 (%), the module triples the percentage of minority cases (3000) compared to the original dataset (1000).

Box 2: 5

We should use 5 data rows.

Use the Number of nearest neighbors option to determine the size of the feature space that the SMOTE algorithm uses when in building new cases. A nearest neighbor is a row of data (a case) that is very similar to some target case. The distance between any two cases is measured by combining the weighted vectors of all features.

By increasing the number of nearest neighbors, you get features from more cases.

By keeping the number of nearest neighbors low, you use features that are more like those in the original sample.

References:

https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/smote

**NEW QUESTION: 133**

You are running a training experiment on remote compute in Azure Machine Learning.

The experiment is configured to use a conda environment that includes the mlflow and azureml-contrib-run packages.

You must use MLflow as the logging package for tracking metrics generated in the experiment.

You need to complete the script for the experiment.

How should you complete the code? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

**Answer:**

Explanation:

Box 1: import mlflow

Import the mlflow and Workspace classes to access MLflow's tracking URI and configure your workspace.

Box 2: mlflow.start_run()

Set the MLflow experiment name with set_experiment() and start your training run with start_run().

Box 3: mlflow.log_metric(' ..')

Use log_metric() to activate the MLflow logging API and begin logging your training run metrics.

Box 4: mlflow.end_run()

Close the run:

run.endRun()

Reference:

https://docs.microsoft.com/en-us/azure/machine-learning/how-to-use-mlflow

**NEW QUESTION: 134**

For each of the following statements, select Yes if the statement is true. Otherwise, select No. NOTE: Each correct selection is worth one point.

**Answer:**

**NEW QUESTION: 135**

You are solving a classification task.

You must evaluate your model on a limited data sample by using k-fold cross-validation. You start by configuring a k parameter as the number of splits.

You need to configure the k parameter for the cross-validation.

Which value should you use?

**A.** k=0.5

**B.** k=0.01

**C.** k=5

**D.** k=1

**Answer: (SHOW ANSWER)**

Leave One Out (LOO) cross-validation

Setting K = n (the number of observations) yields n-fold and is called leave-one out cross-validation (LOO), a special case of the K-fold approach.

LOO CV is sometimes useful but typically doesn't shake up the data enough. The estimates from each fold are highly correlated and hence their average can have high variance.

This is why the usual choice is K=5 or 10. It provides a good compromise for the bias-variance tradeoff.

Perform Feature Engineering

Testlet 1

Case study

Overview

You are a data scientist in a company that provides data science for professional sporting events. Models will use global and local market data to meet the following business goals:

* Understand sentiment of mobile device users at sporting events based on audio from crowd reactions.
* Assess a user's tendency to respond to an advertisement.
* Customize styles of ads served on mobile devices.
* Use video to detect penalty events

Current environment

* Media used for penalty event detection will be provided by consumer devices. Media may include images and videos captured during the sporting event and shared using social media. The images and videos will have varying sizes and formats.
* The data available for model building comprises of seven years of sporting event media. The sporting event media includes; recorded video transcripts or radio commentary, and logs from related social media feeds captured during the sporting events.
* Crowd sentiment will include audio recordings submitted by event attendees in both mono and stereo formats.

Penalty detection and sentiment

* Data scientists must build an intelligent solution by using multiple machine learning models for penalty event detection.
* Data scientists must build notebooks in a local environment using automatic feature engineering and model building in machine learning pipelines.
* Notebooks must be deployed to retrain by using Spark instances with dynamic worker allocation.
* Notebooks must execute with the same code on new Spark instances to recode only the source of the data.
* Global penalty detection models must be trained by using dynamic runtime graph computation during training.
* Local penalty detection models must be written by using BrainScript.
* Experiments for local crowd sentiment models must combine local penalty detection data.
* Crowd sentiment models must identify known sounds such as cheers and known catch phrases. Individual crowd sentiment models will detect similar sounds.
* All shared features for local models are continuous variables.
* Shared features must use double precision. Subsequent layers must have aggregate running mean and standard deviation metrics available.

Advertisements

During the initial weeks in production, the following was observed:
* Ad response rated declined.
* Drops were not consistent across ad styles.
* The distribution of features across training and production data are not consistent Analysis shows that, of the 100 numeric features on user location and behavior, the 47 features that come from location sources are being used as raw features. A suggested experiment to remedy the bias and variance issue is to engineer 10 linearly uncorrelated features.
* Initial data discovery shows a wide range of densities of target states in training data used for crowd sentiment models.
* All penalty detection models show inference phases using a Stochastic Gradient Descent (SGD) are running too slow.
* Audio samples show that the length of a catch phrase varies between 25%-47% depending on region
* The performance of the global penalty detection models shows lower variance but higher bias when comparing training and validation sets. Before implementing any feature changes, you must confirm the bias and variance using all training and validation cases.
* Ad response models must be trained at the beginning of each event and applied during the sporting event.
* Market segmentation models must optimize for similar ad response history.
* Sampling must guarantee mutual and collective exclusively between local and global segmentation models that share the same features.
* Local market segmentation models will be applied before determining a user's propensity to respond to an advertisement.
* Ad response models must support non-linear boundaries of features.
* The ad propensity model uses a cut threshold is 0.45 and retrains occur if weighted Kappa deviated from 0.1 +/- 5%.
* The ad propensity model uses cost factors shown in the following diagram:

* The ad propensity model uses proposed cost factors shown in the following diagram:
* Performance curves of current and proposed cost factor scenarios are shown in the following diagram:

**NEW QUESTION: 136**
You are using the Hyperdrive feature in Azure Machine Learning to train a model.
You configure the Hyperdrive experiment by running the following code:
For each of the following statements, select Yes if the statement is true. Otherwise, select No.
NOTE: Each correct selection is worth one point.
**Answer:**
Explanation:
Box 1: Yes
In random sampling, hyperparameter values are randomly selected from the defined search space. Random sampling allows the search space to include both discrete and continuous hyperparameters.
Box 2: Yes
learning_rate has a normal distribution with mean value 10 and a standard deviation of 3.
Box 3: No
keep_probability has a uniform distribution with a minimum value of 0.05 and a maximum value of 0.1.
Box 4: No
number_of_hidden_layers takes on one of the values [3, 4, 5].
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-tune-hyperparameters



**NEW QUESTION: 137**
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You are creating a new experiment in Azure Machine Learning Studio.
One class has a much smaller number of observations than the other classes in the training set.
You need to select an appropriate data sampling strategy to compensate for the class imbalance.
Solution: You use the Synthetic Minority Oversampling Technique (SMOTE) sampling mode.
Does the solution meet the goal?

**A.** Yes
**B.** No
**Answer: A (LEAVE A REPLY)**
SMOTE is used to increase the number of underepresented cases in a dataset used for machine learning.
SMOTE is a better way of increasing the number of rare cases than simply duplicating existing cases.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/smote

**NEW QUESTION: 138**
You need to define an evaluation strategy for the crowd sentiment models.
Which three actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.
**Answer:**
1 - Define a cross-entropy function activation.
2 - Add cost functions for each target state.
3 - Evaluate the distance error metric.

**NEW QUESTION: 139**
You need to select a feature extraction method.
Which method should you use?
**A.** Mutual information
**B.** Pearson's correlation
**C.** Spearman correlation
**D.** Fisher Linear Discriminant Analysis
**Answer: (SHOW ANSWER)**
Spearman's rank correlation coefficient assesses how well the relationship between two variables can be described using a monotonic function.
Note: Both Spearman's and Kendall's can be formulated as special cases of a more general correlation coefficient, and they are both appropriate in this scenario.
Scenario: The MedianValue and AvgRoomsInHouse columns both hold data in numeric format. You need to select a feature selection algorithm to analyze the relationship between the two columns in more detail.
Incorrect Answers:
B: The Spearman correlation between two variables is equal to the Pearson correlation between the rank values of those two variables; while Pearson's correlation assesses linear relationships, Spearman's correlation assesses monotonic relationships (whether linear or not).
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/feature-selection-modules
Perform Feature Engineering Question Set 3

**NEW QUESTION: 140**

You are a data scientist working for a hotel booking website company. You use the Azure Machine Learning service to train a model that identifies fraudulent transactions.
You must deploy the model as an Azure Machine Learning real-time web service using the Model.deploy method in the Azure Machine Learning SDK. The deployed web service must return real-time predictions of fraud based on transaction data input.
You need to create the script that is specified as the entry_script parameter for the InferenceConfig class used to deploy the model.
What should the entry script do?
**A.** Start a node on the inference cluster where the web service is deployed.
**B.** Register the model with appropriate tags and properties.
**C.** Create a Conda environment for the web service compute and install the necessary Python packages.
**D.** Load the model and use it to predict labels from input data.
**E.** Specify the number of cores and the amount of memory required for the inference compute.
**Answer: (SHOW ANSWER)**
Explanation
The entry script receives data submitted to a deployed web service and passes it to the model. It then takes the response returned by the model and returns that to the client. The script is specific to your model. It must understand the data that the model expects and returns.
The two things you need to accomplish in your entry script are:
Loading your model (using a function called init())
Running your model on input data (using a function called run())
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-deploy-and-where

**NEW QUESTION: 141**
You create a training pipeline using the Azure Machine Learning designer. You upload a CSV file that contains the data from which you want to train your model.
You need to use the designer to create a pipeline that includes steps to perform the following tasks:
Select the training features using the pandas filter method.
Train a model based on the naive_bayes.GaussianNB algorithm.
Return only the Scored Labels column by using the query SELECT [Scored Labels] FROM t1; Which modules should you use? To answer, drag the appropriate modules to the appropriate locations. Each module name may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.
NOTE: Each correct selection is worth one point.
**Answer:**

**NEW QUESTION: 142**
You are implementing a machine learning model to predict stock prices.
The model uses a PostgreSQL database and requires GPU processing.
You need to create a virtual machine that is pre-configured with the required tools.

What should you do?
**A.** Create a Data Science Virtual Machine (DSVM) Windows edition.
**B.** Create a Geo AI Data Science Virtual Machine (Geo-DSVM) Windows edition.
**C.** Create a Deep Learning Virtual Machine (DLVM) Linux edition.
**D.** Create a Deep Learning Virtual Machine (DLVM) Windows edition.
**E.** Create a Data Science Virtual Machine (DSVM) Linux edition.
**Answer: E (LEAVE A REPLY)**
Incorrect Answers:
A, C: PostgreSQL (CentOS) is only available in the Linux Edition.
B: The Azure Geo AI Data Science VM (Geo-DSVM) delivers geospatial analytics capabilities from Microsoft's Data Science VM. Specifically, this VM extends the AI and data science toolkits in the Data Science VM by adding ESRI's market-leading ArcGIS Pro Geographic Information System.
D: DLVM is a template on top of DSVM image. In terms of the packages, GPU drivers etc are all there in the DSVM image. Mostly it is for convenience during creation where we only allow DLVM to be created on GPU VM instances on Azure.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/data-science-virtual-machine/overview

**NEW QUESTION: 143**
You have an Azure Machine Learning workspace that contains a CPU-based compute cluster and an Azure Kubernetes Services (AKS) inference cluster. You create a tabular dataset containing data that you plan to use to create a classification model.
You need to use the Azure Machine Learning designer to create a web service through which client applications can consume the classification model by submitting new data and getting an immediate prediction as a response.
Which three actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.
**Answer:**
1 - Create and start a Compute Instance
2 - Create and run a training pipeline..
3 - Create and run a real-time inference pipeline
Reference:
https://docs.microsoft.com/en-us/learn/modules/create-classification-model-azure-machine-learning-designer/

**NEW QUESTION: 144**
You need to define an evaluation strategy for the crowd sentiment models.
Which three actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.
**Answer:**
Explanation:
Scenario:

Experiments for local crowd sentiment models must combine local penalty detection data.
Crowd sentiment models must identify known sounds such as cheers and known catch phrases. Individual crowd sentiment models will detect similar sounds.
Note: Evaluate the changed in correlation between model error rate and centroid distance In machine learning, a nearest centroid classifier or nearest prototype classifier is a classification model that assigns to observations the label of the class of training samples whose mean (centroid) is closest to the observation.
References:
https://en.wikipedia.org/wiki/Nearest_centroid_classifier
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/sweep-clustering

**NEW QUESTION: 145**
You plan to build a team data science environment. Data for training models in machine learning pipelines will be over 20 GB in size.
You have the following requirements:
* Models must be built using Caffe2 or Chainer frameworks.
* Data scientists must be able to use a data science environment to build the machine learning pipelines and train models on their personal devices in both connected and disconnected network environments.
Personal devices must support updating machine learning pipelines when connected to a network.
You need to select a data science environment.
Which environment should you use?
**A.** Azure Machine Learning Service
**B.** Azure Machine Learning Studio
**C.** Azure Databricks
**D.** Azure Kubernetes Service (AKS)
**Answer: (SHOW ANSWER)**
The Data Science Virtual Machine (DSVM) is a customized VM image on Microsoft's Azure cloud built specifically for doing data science. Caffe2 and Chainer are supported by DSVM.
DSVM integrates with Azure Machine Learning.
Incorrect Answers:
B: Use Machine Learning Studio when you want to experiment with machine learning models quickly and easily, and the built-in machine learning algorithms are sufficient for your solutions.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/data-science-virtual-machine/overview

**NEW QUESTION: 146**
You are working on a classification task. You have a dataset indicating whether a student would like to play soccer and associated attributes. The dataset includes the following columns:
You need to classify variables by type.
Which variable should you add to each category? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

**Answer:**
Explanation
References:
https://www.edureka.co/blog/classification-algorithms/

**NEW QUESTION: 147**
You need to build a feature extraction strategy for the local models.
How should you complete the code segment? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
**Answer:**
Explanation

**NEW QUESTION: 148**
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You have a Python script named train.py in a local folder named scripts. The script trains a regression model by using scikit-learn. The script includes code to load a training data file which is also located in the scripts folder.
You must run the script as an Azure ML experiment on a compute cluster named aml-compute.
You need to configure the run to ensure that the environment includes the required packages for model training. You have instantiated a variable named aml-compute that references the target compute cluster.
Solution: Run the following code:
Does the solution meet the goal?
**A.** Yes
**B.** No
**Answer: B (LEAVE A REPLY)**
The scikit-learn estimator provides a simple way of launching a scikit-learn training job on a compute target. It is implemented through the SKLearn class, which can be used to support single-node CPU training.
Example:
from azureml.train.sklearn import SKLearn
}
estimator = SKLearn(source_directory=project_folder,
compute_target=compute_target,
entry_script='train_iris.py'
)
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-train-scikit-learn

**NEW QUESTION: 149**
You are working on a classification task. You have a dataset indicating whether a student would like to play soccer and associated attributes. The dataset includes the following columns:
You need to classify variables by type.
Which variable should you add to each category? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
**Answer:**
References:
https://www.edureka.co/blog/classification-algorithms/

**NEW QUESTION: 150**
A coworker registers a datastore in a Machine Learning services workspace by using the following code:
You need to write code to access the datastore from a notebook.
**Answer:**

**NEW QUESTION: 151**
You need to implement a new cost factor scenario for the ad response models as illustrated in the performance curve exhibit.
Which technique should you use?
**A.** Set the threshold to 0.5 and retrain if weighted Kappa deviates +/- 5% from 0.45.
**B.** Set the threshold to 0.05 and retrain if weighted Kappa deviates +/- 5% from 0.5.
**C.** Set the threshold to 0.2 and retrain if weighted Kappa deviates +/- 5% from 0.6.
**D.** Set the threshold to 0.75 and retrain if weighted Kappa deviates +/- 5% from 0.15.
**Answer: (SHOW ANSWER)**
Explanation
Scenario:
Performance curves of current and proposed cost factor scenarios are shown in the following diagram:
The ad propensity model uses a cut threshold is 0.45 and retrains occur if weighted Kappa deviated from 0.1 +/- 5%.



**NEW QUESTION: 152**

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You are a data scientist using Azure Machine Learning Studio.
You need to normalize values to produce an output column into bins to predict a target column.
Solution: Apply a Quantiles binning mode with a PQuantile normalization.
Does the solution meet the goal?
**A.** Yes
**B.** No
**Answer: B (LEAVE A REPLY)**
Use the Entropy MDL binning mode which has a target column.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/group-data-into-bins

**NEW QUESTION: 153**
You need to record the row count as a metric named row_count that can be returned using the get_metrics method of the Run object after the experiment run completes. Which code should you use?
**A.** run.log('row_count', rows)
**B.** run.log_table('row_count', rows)
**C.** run.tag('row_count', rows)
**D.** run.log_row('row_count', rows)
**E.** run.upload_file('row_count', './data.csv')
**Answer: A (LEAVE A REPLY)**

**NEW QUESTION: 154**
You plan to create a speech recognition deep learning model.
The model must support the latest version of Python.
You need to recommend a deep learning framework for speech recognition to include in the Data Science Virtual Machine (DSVM).
What should you recommend?
**A.** Rattle
**B.** TensorFlow
**C.** Weka
**D.** Scikit-learn
**Answer: B (LEAVE A REPLY)**
TensorFlow is an open source library for numerical computation and large-scale machine learning. It uses Python to provide a convenient front-end API for building applications with the framework TensorFlow can train and run deep neural networks for handwritten digit classification, image recognition, word embeddings,

recurrent neural networks, sequence-to-sequence models for machine translation, natural language processing, and PDE (partial differential equation) based simulations.
Incorrect Answers:
A: Rattle is the R analytical tool that gets you started with data analytics and machine learning.
C: Weka is used for visual data mining and machine learning software in Java.
D: Scikit-learn is one of the most useful library for machine learning in Python. It is on NumPy, SciPy and matplotlib, this library contains a lot of effiecient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction.
Reference:
https://www.infoworld.com/article/3278008/what-is-tensorflow-the-machine-learning-library-explained.html

**NEW QUESTION: 155**
You need to define a process for penalty event detection.
Which three actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.
**Answer:**
1 - Import the global model and build 1 the local model using PyTorch.
2 - Build the global model using PyTorch.
3 - Build the global model using TensorFlow
4 - Import the global model and build the local model using TensorFlow.

**NEW QUESTION: 156**
You are preparing to use the Azure ML SDK to run an experiment and need to create compute. You run the following code:
For each of the following statements, select Yes if the statement is true. Otherwise, select No.
NOTE: Each correct selection is worth one point.
**Answer:**
Explanation:
Box 1: No
If a training cluster already exists it will be used.
Box 2: Yes
The wait_for_completion method waits for the current provisioning operation to finish on the cluster.
Box 3: Yes
Low Priority VMs use Azure's excess capacity and are thus cheaper but risk your run being pre-empted.
Box 4: No
Need to use training_compute.delete() to deprovision and delete the AmlCompute target.
Reference:
https://notebooks.azure.com/azureml/projects/azureml-getting-started/html/how-to-use-azureml/training/train-on-amlcompute/train-on-amlcompute.ipynb
https://docs.microsoft.com/en-us/python/api/azureml-core/azureml.core.compute.computetarget

**NEW QUESTION: 157**
You create an Azure Databricks workspace and a linked Azure Machine Learning workspace.
You have the following Python code segment in the Azure Machine Learning workspace:
import mlflow
import mlflow.azureml
import azureml.mlflow
import azureml.core
from azureml.core import Workspace
subscription_id = 'subscription_id'
resourse_group = 'resource_group_name'
workspace_name = 'workspace_name'
ws = Workspace.get(name=workspace_name,
subscription_id=subscription_id,
resource_group=resource_group)
experimentName = "/Users/{user_name}/{experiment_folder}/{experiment_name}"
mlflow.set_experiment(experimentName) uri = ws.get_mlflow_tracking_uri() mlflow.set_tracking_uri(uri)
Instructions: For each of the following statements, select Yes if the statement is true. Otherwise, select No.
NOTE: Each correct selection is worth one point.
**Answer:**
Reference:
https://docs.microsoft.com/en-us/python/api/azureml-core/azureml.core.workspace.workspace
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-use-mlflow

**NEW QUESTION: 158**
You are preparing to build a deep learning convolutional neural network model for image classification. You create a script to train the model using CUDA devices. You must submit an experiment that runs this script in the Azure Machine Learning workspace. The following compute resources are available:
* a Microsoft Surface device on which Microsoft Office has been installed. Corporate IT policies prevent the installation of additional software
* a Compute Instance named ds-workstation in the workspace with 2 CPUs and 8 GB of memory
* an Azure Machine Learning compute target named cpu-cluster with eight CPU-based nodes
* an Azure Machine Learning compute target named gpu-cluster with four CPU and GPU-based nodes
**Answer:**

**NEW QUESTION: 159**
You are a lead data scientist for a project that tracks the health and migration of birds. You create a multi-class image classification deep learning model that uses a set of labeled bird photographs collected by experts.
You have 100,000 photographs of birds. All photographs use the JPG format and are stored in an Azure blob container in an Azure subscription.

You need to access the bird photograph files in the Azure blob container from the Azure Machine Learning service workspace that will be used for deep learning model training. You must minimize data movement.
What should you do?
**A.** Create an Azure Data Lake store and move the bird photographs to the store.
**B.** Create an Azure Cosmos DB database and attach the Azure Blob containing bird photographs storage to the database.
**C.** Create and register a dataset by using TabularDataset class that references the Azure blob storage containing bird photographs.
**D.** Register the Azure blob storage containing the bird photographs as a datastore in Azure Machine Learning service.
**E.** Copy the bird photographs to the blob datastore that was created with your Azure Machine Learning service workspace.
**Answer: D (LEAVE A REPLY)**
We recommend creating a datastore for an Azure Blob container. When you create a workspace, an Azure blob container and an Azure file share are automatically registered to the workspace.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-access-data

**NEW QUESTION: 160**
You are running a training experiment on remote compute in Azure Machine Learning.
The experiment is configured to use a conda environment that includes the mlflow and azureml-contrib-run packages.
You must use MLflow as the logging package for tracking metrics generated in the experiment.
You need to complete the script for the experiment.
How should you complete the code? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
**Answer:**
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-use-mlflow

**NEW QUESTION: 161**
You are a lead data scientist for a project that tracks the health and migration of birds. You create a multi-image classification deep learning model that uses a set of labeled bird photos collected by experts. You plan to use the model to develop a cross-platform mobile app that predicts the species of bird captured by app users.
You must test and deploy the trained model as a web service. The deployed model must meet the following requirements:
* An authenticated connection must not be required for testing.
* The deployed model must perform with low latency during inferencing.
* The REST endpoints must be scalable and should have a capacity to handle large number of requests when multiple end users are using the mobile application.

You need to verify that the web service returns predictions in the expected JSON format when a valid REST request is submitted.
Which compute resources should you use? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
**Answer:**
Explanation:
Box 1: ds-workstation notebook VM
An authenticated connection must not be required for testing.
On a Microsoft Azure virtual machine (VM), including a Data Science Virtual Machine (DSVM), you create local user accounts while provisioning the VM. Users then authenticate to the VM by using these credentials.
Box 2: gpu-compute cluster
Image classification is well suited for GPU compute clusters
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/data-science-virtual-machine/dsvm-common-identity
https://docs.microsoft.com/en-us/azure/architecture/reference-architectures/ai/training-deep-learning

**NEW QUESTION: 162**
You are creating a binary classification by using a two-class logistic regression model.
You need to evaluate the model results for imbalance.
Which evaluation metric should you use?
**A.** Relative Absolute Error
**B.** AUC Curve
**C.** Mean Absolute Error
**D.** Relative Squared Error
**Answer: B (LEAVE A REPLY)**
Explanation
One can inspect the true positive rate vs. the false positive rate in the Receiver Operating Characteristic (ROC) curve and the corresponding Area Under the Curve (AUC) value. The closer this curve is to the upper left corner, the better the classifier's performance is (that is maximizing the true positive rate while minimizing the false positive rate). Curves that are close to the diagonal of the plot, result from classifiers that tend to make predictions that are close to random guessing.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio/evaluate-model-performance#evaluating-a-binar

**NEW QUESTION: 163**
You are performing feature engineering on a dataset.
You must add a feature named CityName and populate the column value with the text London.
You need to add the new feature to the dataset.
Which Azure Machine Learning Studio module should you use?
**A.** Edit Metadata

**B.** Filter Based Feature Selection
**C.** Execute Python Script
**D.** Latent Dirichlet Allocation
**Answer: A (LEAVE A REPLY)**
Typical metadata changes might include marking columns as features.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/edit-metadata

**NEW QUESTION: 164**
You need to modify the inputs for the global penalty event model to address the bias and variance issue.
Which three actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.
**Answer:**
1 - Select the location data.
2 - Select the behavior data.
3 - Perform a Primary Component Analysis (PCA).
4 - Add a K-Means clustering module with 10 clusters.
5 - Bin the new data.
6 - Build rations.

**NEW QUESTION: 165**
You need to configure the Permutation Feature Importance module for the model training requirements.
What should you do? To answer, select the appropriate options in the dialog box in the answer area.
NOTE: Each correct selection is worth one point.
**Answer:**
Explanation:
Box 1: 500
For Random seed, type a value to use as seed for randomization. If you specify 0 (the default), a number is generated based on the system clock.
A seed value is optional, but you should provide a value if you want reproducibility across runs of the same experiment.
Here we must replicate the findings.
Box 2: Mean Absolute Error
Scenario: Given a trained model and a test dataset, you must compute the Permutation Feature Importance scores of feature variables. You need to set up the Permutation Feature Importance module to select the correct metric to investigate the model's accuracy and replicate the findings.
Regression. Choose one of the following: Precision, Recall, Mean Absolute Error , Root Mean Squared Error, Relative Absolute Error, Relative Squared Error, Coefficient of Determination References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/permutation-feature-importance

**NEW QUESTION: 166**
You are a lead data scientist for a project that tracks the health and migration of birds. You create a multi-image classification deep learning model that uses a set of labeled bird photos collected by experts. You plan to use the model to develop a cross-platform mobile app that predicts the species of bird captured by app users.
You must test and deploy the trained model as a web service. The deployed model must meet the following requirements:
* An authenticated connection must not be required for testing.
* The deployed model must perform with low latency during inferencing.
* The REST endpoints must be scalable and should have a capacity to handle large number of requests when multiple end users are using the mobile application.
You need to verify that the web service returns predictions in the expected JSON format when a valid REST request is submitted.
Which compute resources should you use? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
**Answer:**
Explanation
Box 1: ds-workstation notebook VM
An authenticated connection must not be required for testing.
On a Microsoft Azure virtual machine (VM), including a Data Science Virtual Machine (DSVM), you create local user accounts while provisioning the VM. Users then authenticate to the VM by using these credentials.
Box 2: gpu-compute cluster
Image classification is well suited for GPU compute clusters
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/data-science-virtual-machine/dsvm-common-identity
https://docs.microsoft.com/en-us/azure/architecture/reference-architectures/ai/training-deep-learning



**NEW QUESTION: 167**
You need to correct the model fit issue.
Which three actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.
**Answer:**
Explanation:

Step 1: Augment the data
Scenario: Columns in each dataset contain missing and null values. The datasets also contain many outliers.
Step 2: Add the Bayesian Linear Regression module.
Scenario: You produce a regression model to predict property prices by using the Linear Regression and Bayesian Linear Regression modules.
Step 3: Configure the regularization weight.
Regularization typically is used to avoid overfitting. For example, in L2 regularization weight, type the value to use as the weight for L2 regularization. We recommend that you use a non-zero value to avoid overfitting.
Scenario:
Model fit: The model shows signs of overfitting. You need to produce a more refined regression model that reduces the overfitting.
Incorrect Answers:
Multiclass Decision Jungle module:
Decision jungles are a recent extension to decision forests. A decision jungle consists of an ensemble of decision directed acyclic graphs (DAGs).
L-BFGS:
L-BFGS stands for "limited memory Broyden-Fletcher-Goldfarb-Shanno". It can be found in the wwo-Class Logistic Regression module, which is used to create a logistic regression model that can be used to predict two (and only two) outcomes.
References:
<https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/linear-regr ession>

**NEW QUESTION: 168**
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You are creating a new experiment in Azure Learning learning Studio.
One class has a much smaller number of observations than the other classes in the training You need to select an appropriate data sampling strategy to compensate for the class imbalance.
Solution: You use the Synthetic Minority Oversampling Technique (SMOTE) sampling mode.
Does the solution meet the goal?
**A.** Yes
**B.** No
**Answer: A (LEAVE A REPLY)**
Explanation
SMOTE is used to increase the number of underepresented cases in a dataset used for machine learning.
SMOTE is a better way of increasing the number of rare cases than simply duplicating existing cases.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/smote

**NEW QUESTION: 169**
You have a dataset created for multiclass classification tasks that contains a normalized numerical feature set with 10,000 data points and 150 features.
You use 75 percent of the data points for training and 25 percent for testing. You are using the scikit-learn machine learning library in Python. You use X to denote the feature set and Y to denote class labels.
You create the following Python data frames:
You need to apply the Principal Component Analysis (PCA) method to reduce the dimensionality of the feature set to 10 features in both training and testing sets.
How should you complete the code segment? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
**Answer:**
Explanation
Box 1: PCA(n_components = 10)
Need to reduce the dimensionality of the feature set to 10 features in both training and testing sets.
Example:
from sklearn.decomposition import PCA
pca = PCA(n_components=2) ;2 dimensions
principalComponents = pca.fit_transform(x)
Box 2: pca
fit_transform(X[, y])fits the model with X and apply the dimensionality reduction on X.
Box 3: transform(x_test)
transform(X) applies dimensionality reduction to X.
References:
https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

**NEW QUESTION: 170**
You have a model with a large difference between the training and validation error values.
You must create a new model and perform cross-validation.
You need to identify a parameter set for the new model using Azure Machine Learning Studio.
Which module you should use for each step? To answer, drag the appropriate modules to the correct steps.
Each module may be used once or more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.
NOTE: Each correct selection is worth one point.
**Answer:**
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/partition-and-sample

**NEW QUESTION: 171**
You have a feature set containing the following numerical features: X, Y, and Z.
The Poisson correlation coefficient (r-value) of X, Y, and Z features is shown in the following image:

Use the drop-down menus to select the answer choice that answers each question based on the information presented in the graphic.
NOTE: Each correct selection is worth one point.

**Answer:**

Explanation:
Box 1: 0.859122
Box 2: a positively linear relationship
+1 indicates a strong positive linear relationship
-1 indicates a strong negative linear correlation
0 denotes no linear relationship between the two variables.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/compute-linear-correlation

**NEW QUESTION: 172**

You need to identify the methods for dividing the data according to the testing requirements.
Which properties should you select? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

**Answer:**

Explanation
Scenario: Testing
You must produce multiple partitions of a dataset based on sampling using the Partition and Sample module in Azure Machine Learning Studio.
Box 1: Assign to folds
Use Assign to folds option when you want to divide the dataset into subsets of the data. This option is also useful when you want to create a custom number of folds for cross-validation, or to split rows into several groups.
Not Head: Use Head mode to get only the first n rows. This option is useful if you want to test a pipeline on a small number of rows, and don't need the data to be balanced or sampled in any way.
Not Sampling: The Sampling option supports simple random sampling or stratified random sampling. This is useful if you want to create a smaller representative sample dataset for testing.
Box 2: Partition evenly
Specify the partitioner method: Indicate how you want data to be apportioned to each partition, using these options:
Partition evenly: Use this option to place an equal number of rows in each partition. To specify the number of output partitions, type a whole number in the Specify number of folds to split evenly into text box.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/algorithm-module-reference/partition-and-sample

**NEW QUESTION: 173**

You are performing a filter-based feature selection for a dataset to build a multi-class classifier by using Azure Machine Learning Studio.
The dataset contains categorical features that are highly correlated to the output label column.
You need to select the appropriate feature scoring statistical method to identify the key predictors.
Which method should you use?
**A.** Kendall correlation
**B.** Spearman correlation
**C.** Chi-squared
**D.** Pearson correlation
**Answer: (SHOW ANSWER)**
Pearson's correlation statistic, or Pearson's correlation coefficient, is also known in statistical models as the r value. For any two variables, it returns a value that indicates the strength of the correlation Pearson's correlation coefficient is the test statistics that measures the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship.
Incorrect Answers:
C: The two-way chi-squared test is a statistical method that measures how close expected values are to actual results.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/filter-based-feature-selection
https://www.statisticssolutions.com/pearsons-correlation-coefficient/

**NEW QUESTION: 174**
You need to produce a visualization for the diagnostic test evaluation according to the data visualization requirements.
Which three modules should you recommend be used in sequence? To answer, move the appropriate modules from the list of modules to the answer area and arrange them in the correct order.
**Answer:**
Explanation
Step 1: Sweep Clustering
Start by using the "Tune Model Hyperparameters" module to select the best sets of parameters for each of the models we're considering.
One of the interesting things about the "Tune Model Hyperparameters" module is that it not only outputs the results from the Tuning, it also outputs the Trained Model.
Step 2: Train Model
Step 3: Evaluate Model
Scenario: You need to provide the test results to the Fabrikam Residences team. You create data visualizations to aid in presenting the results.

You must produce a Receiver Operating Characteristic (ROC) curve to conduct a diagnostic test evaluation of the model. You need to select appropriate methods for producing the ROC curve in Azure Machine Learning Studio to compare the Two-Class Decision Forest and the Two-Class Decision Jungle modules with one another.
References:
http://breaking-bi.blogspot.com/2017/01/azure-machine-learning-model-evaluation.html

**NEW QUESTION: 175**
You create an experiment in Azure Machine Learning Studio- You add a training dataset that contains 10.000 rows. The first 9.000 rows represent class 0 (90 percent). The first 1.000 rows represent class 1 (10 percent). The training set is unbalanced between two Classes. You must increase the number of training examples for class 1 to 4,000 by using data rows. You add the Synthetic Minority Oversampling Technique (SMOTE) module to the experiment.
You need to configure the module.
Which values should you use? To answer, select the appropriate options in the dialog box in the answer area.
NOTE: Each correct selection is worth one point.
**Answer:**

**NEW QUESTION: 176**
You are performing sentiment analysis using a CSV file that includes 12,000 customer reviews written in a short sentence format. You add the CSV file to Azure Machine Learning Studio and configure it as the starting point dataset of an experiment. You add the Extract N-Gram Features from Text module to the experiment to extract key phrases from the customer review column in the dataset.
You must create a new n-gram dictionary from the customer review text and set the maximum n-gram size to trigrams.
What should you select? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
**Answer:**
Explanation
Vocabulary mode: Create
For Vocabulary mode, select Create to indicate that you are creating a new list of n-gram features.
N-Grams size: 3
For N-Grams size, type a number that indicates the maximum size of the n-grams to extract and store. For example, if you type 3, unigrams, bigrams, and trigrams will be created.
Weighting function: Leave blank
The option, Weighting function, is required only if you merge or update vocabularies. It specifies how terms in the two vocabularies and their scores should be weighted against each other.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/extract-n-gram-features-from-

**NEW QUESTION: 177**
You need to correct the model fit issue.
Which three actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.
**Answer:**
1 - Augment the data
2 - Add the Bayesian Linear Regression module.
3 - Configure the regularization weight.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/linear-regr ession

**NEW QUESTION: 178**
You need to set up the Permutation Feature Importance module according to the model training requirements.
Which properties should you select? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
**Answer:**
Explanation
Box 1: Accuracy
Scenario: You want to configure hyperparameters in the model learning process to speed the learning phase by using hyperparameters. In addition, this configuration should cancel the lowest performing runs at each evaluation interval, thereby directing effort and resources towards models that are more likely to be successful.
Box 2: R-Squared

**NEW QUESTION: 179**
You are performing sentiment analysis using a CSV file that includes 12,000 customer reviews written in a short sentence format. You add the CSV file to Azure Machine Learning Studio and configure it as the starting point dataset of an experiment. You add the Extract N-Gram Features from Text module to the experiment to extract key phrases from the customer review column in the dataset.
You must create a new n-gram dictionary from the customer review text and set the maximum n-gram size to trigrams.
What should you select? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
**Answer:**
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/extract-n-gram-features-from-text

**NEW QUESTION: 180**
You create a multi-class image classification deep learning model that uses the PyTorch deep learning framework.
You must configure Azure Machine Learning Hyperdrive to optimize the hyperparameters for the classification model.
You need to define a primary metric to determine the hyperparameter values that result in the model with the best accuracy score.
Which three actions must you perform? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.
**A.** Set the primary_metric_goal of the estimator used to run the bird_classifier_train.py script to maximize.
**B.** Add code to the bird_classifier_train.py script to calculate the validation loss of the model and log it as a float value with the key loss.
**C.** Set the primary_metric_goal of the estimator used to run the bird_classifier_train.py script to minimize.
**D.** Set the primary_metric_name of the estimator used to run the bird_classifier_train.py script to accuracy.
**E.** Set the primary_metric_name of the estimator used to run the bird_classifier_train.py script to loss.
**F.** Add code to the bird_classifier_train.py script to calculate the validation accuracy of the model and log it as a float value with the key accuracy.
**Answer: (SHOW ANSWER)**
Explanation
AD:
primary_metric_name="accuracy",
primary_metric_goal=PrimaryMetricGoal.MAXIMIZE
Optimize the runs to maximize "accuracy". Make sure to log this value in your training script.
Note:
primary_metric_name: The name of the primary metric to optimize. The name of the primary metric needs to exactly match the name of the metric logged by the training script.
primary_metric_goal: It can be either PrimaryMetricGoal.MAXIMIZE or PrimaryMetricGoal.MINIMIZE and determines whether the primary metric will be maximized or minimized when evaluating the runs.
F: The training script calculates the val_accuracy and logs it as "accuracy", which is used as the primary metric.

**NEW QUESTION: 181**
You deploy a real-time inference service for a trained model.
The deployed model supports a business-critical application, and it is important to be able to monitor the data submitted to the web service and the predictions the data generates.
You need to implement a monitoring solution for the deployed model using minimal administrative effort.
What should you do?
**A.** View the explanations for the registered model in Azure ML studio.
**B.** Enable Azure Application Insights for the service endpoint and view logged data in the Azure portal.
**C.** View the log files generated by the experiment used to train the model.
**D.** Create an ML Flow tracking URI that references the endpoint, and view the data logged by ML Flow.

**Answer: A (LEAVE A REPLY)**



**NEW QUESTION: 182**
You are building a machine learning model for translating English language textual content into French language textual content.
You need to build and train the machine learning model to learn the sequence of the textual content.
Which type of neural network should you use?
**A.** Multilayer Perceptions (MLPs)
**B.** Convolutional Neural Networks (CNNs)
**C.** Recurrent Neural Networks (RNNs)
**D.** Generative Adversarial Networks (GANs)
**Answer: C (LEAVE A REPLY)**
Explanation
To translate a corpus of English text to French, we need to build a recurrent neural network (RNN).
Note: RNNs are designed to take sequences of text as inputs or return sequences of text as outputs, or both. They're called recurrent because the network's hidden layers have a loop in which the output and cell state from each time step become inputs at the next time step. This recurrence serves as a form of memory. It allows contextual information to flow through the network so that relevant outputs from previous time steps can be applied to network operations at the current time step.
References:
https://towardsdatascience.com/language-translation-with-rnns-d84d43b40571

**NEW QUESTION: 183**
You are developing a data science workspace that uses an Azure Machine Learning service.
You need to select a compute target to deploy the workspace.
What should you use?
**A.** Azure Data Lake Analytics
**B.** Azure Databricks
**C.** Azure Container Service
**D.** Apache Spark for HDInsight
**Answer: C (LEAVE A REPLY)**
Azure Container Instances can be used as compute target for testing or development. Use for low-scale CPU-based workloads that require less than 48 GB of RAM.

Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/service/how-to-deploy-and-where

**NEW QUESTION: 184**
You arc I mating a deep learning model to identify cats and dogs. You have 25,000 color images.
You must meet the following requirements:
* Reduce the number of training epochs.
* Reduce the size of the neural network.
* Reduce over-fitting of the neural network.
You need to select the image modification values.
Which value should you use? To answer, select the appropriate Options in the answer area.
NOTE: Each correct selection is worth one point.
**Answer:**

**NEW QUESTION: 185**
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You create a model to forecast weather conditions based on historical data.
You need to create a pipeline that runs a processing script to load data from a datastore and pass the processed data to a machine learning model training script.
Solution: Run the following code:
Does the solution meet the goal?
**A.** Yes
**B.** No
**Answer: (SHOW ANSWER)**
Explanation
The two steps are present: process_step and train_step
Note:
Data used in pipeline can be produced by one step and consumed in another step by providing a PipelineData object as an output of one step and an input of one or more subsequent steps.
PipelineData objects are also used when constructing Pipelines to describe step dependencies. To specify that a step requires the output of another step as input, use a PipelineData object in the constructor of both steps.
For example, the pipeline train step depends on the process_step_output output of the pipeline process step:
from azureml.pipeline.core import Pipeline, PipelineData
from azureml.pipeline.steps import PythonScriptStep
datastore = ws.get_default_datastore()

process_step_output = PipelineData("processed_data", datastore=datastore) process_step = PythonScriptStep(script_name="process.py", arguments=["--data_for_train", process_step_output], outputs=[process_step_output], compute_target=aml_compute, source_directory=process_directory) train_step = PythonScriptStep(script_name="train.py", arguments=["--data_for_train", process_step_output], inputs=[process_step_output], compute_target=aml_compute, source_directory=train_directory) pipeline = Pipeline(workspace=ws, steps=[process_step, train_step]) Reference: https://docs.microsoft.com/en-us/python/api/azureml-pipeline-core/azureml.pipeline.core.pipelinedata?view=azu

**NEW QUESTION: 186**

You are a data scientist creating a linear regression model.

You need to determine how closely the data fits the regression line.

Which metric should you review?

**A.** Coefficient of determination

**B.** Recall

**C.** Precision

**D.** Mean absolute error

**E.** Root Mean Square Error

**Answer: D (LEAVE A REPLY)**

Topic 2, Case Study 1

Overview

You are a data scientist in a company that provides data science for professional sporting events. Models will be global and local market data to meet the following business goals:

*Understand sentiment of mobile device users at sporting events based on audio from crowd reactions.

*Access a user's tendency to respond to an advertisement.

*Customize styles of ads served on mobile devices.

*Use video to detect penalty events.

Current environment

Requirements

* Media used for penalty event detection will be provided by consumer devices. Media may include images and videos captured during the sporting event and snared using social media. The images and videos will have varying sizes and formats.

* The data available for model building comprises of seven years of sporting event media. The sporting event media includes: recorded videos, transcripts of radio commentary, and logs from related social media feeds feeds captured during the sporting events.

*Crowd sentiment will include audio recordings submitted by event attendees in both mono and stereo Formats.

Advertisements

* Ad response models must be trained at the beginning of each event and applied during the sporting event.

* Market segmentation nxxlels must optimize for similar ad resporr.r history.

* Sampling must guarantee mutual and collective exclusivity local and global segmentation models that share the same features.
* Local market segmentation models will be applied before determining a user's propensity to respond to an advertisement.
* Data scientists must be able to detect model degradation and decay.
* Ad response models must support non linear boundaries features.
* The ad propensity model uses a cut threshold is 0.45 and retrains occur if weighted Kappa deviates from 0.1 +/-5%.
* The ad propensity model uses cost factors shown in the following diagram:
Penalty detection and sentiment
Findings
*Data scientists must build an intelligent solution by using multiple machine learning models for penalty event detection.
*Data scientists must build notebooks in a local environment using automatic feature engineering and model building in machine learning pipelines.
*Notebooks must be deployed to retrain by using Spark instances with dynamic worker allocation
*Notebooks must execute with the same code on new Spark instances to recode only the source of the data.
*Global penalty detection models must be trained by using dynamic runtime graph computation during training.
*Local penalty detection models must be written by using BrainScript.
* Experiments for local crowd sentiment models must combine local penalty detection data.
* Crowd sentiment models must identify known sounds such as cheers and known catch phrases. Individual crowd sentiment models will detect similar sounds.
* All shared features for local models are continuous variables.
* Shared features must use double precision. Subsequent layers must have aggregate running mean and standard deviation metrics Available.
segments
During the initial weeks in production, the following was observed:
*Ad response rates declined.
*Drops were not consistent across ad styles.
*The distribution of features across training and production data are not consistent.
Analysis shows that of the 100 numeric features on user location and behavior, the 47 features that come from location sources are being used as raw features. A suggested experiment to remedy the bias and variance issue is to engineer 10 linearly uncorrected features.
Penalty detection and sentiment
*Initial data discovery shows a wide range of densities of target states in training data used for crowd sentiment models.
*All penalty detection models show inference phases using a Stochastic Gradient Descent (SGD) are running too stow.
*Audio samples show that the length of a catch phrase varies between 25%-47%, depending on region.

*The performance of the global penalty detection models show lower variance but higher bias when comparing training and validation sets. Before implementing any feature changes, you must confirm the bias and variance using all training and validation cases.

**NEW QUESTION: 187**
You create a binary classification model to predict whether a person has a disease.
You need to detect possible classification errors.
Which error type should you choose for each description? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
**Answer:**
Explanation:
Box 1: True Positive
A true positive is an outcome where the model correctly predicts the positive class Box 2: True Negative A true negative is an outcome where the model correctly predicts the negative class.
Box 3: False Positive
A false positive is an outcome where the model incorrectly predicts the positive class.
Box 4: False Negative
A false negative is an outcome where the model incorrectly predicts the negative class.
Note: Let's make the following definitions:
"Wolf" is a positive class.
"No wolf" is a negative class.
We can summarize our "wolf-prediction" model using a 2x2 confusion matrix that depicts all four possible outcomes:
Reference:
https://developers.google.com/machine-learning/crash-course/classification/true-false-positive-negative

**NEW QUESTION: 188**
You need to identify the methods for dividing the data according to the testing requirements.
Which properties should you select? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
**Answer:**
Explanation:
Scenario: Testing
You must produce multiple partitions of a dataset based on sampling using the Partition and Sample module in Azure Machine Learning Studio.
Box 1: Assign to folds
Use Assign to folds option when you want to divide the dataset into subsets of the dat a. This option is also useful when you want to create a custom number of folds for cross-validation, or to split rows into several groups.

Not Head: Use Head mode to get only the first n rows. This option is useful if you want to test a pipeline on a small number of rows, and don't need the data to be balanced or sampled in any way.
Not Sampling: The Sampling option supports simple random sampling or stratified random sampling. This is useful if you want to create a smaller representative sample dataset for testing.
Box 2: Partition evenly
Specify the partitioner method: Indicate how you want data to be apportioned to each partition, using these options:
Partition evenly: Use this option to place an equal number of rows in each partition. To specify the number of output partitions, type a whole number in the Specify number of folds to split evenly into text box.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/algorithm-module-reference/partition-and-sample

**NEW QUESTION: 189**
You need to implement a model development strategy to determine a user's tendency to respond to an ad.
Which technique should you use?
**A.** Use a Relative Expression Split module to partition the data based on centroid distance.
**B.** Use a Relative Expression Split module to partition the data based on distance travelled to the event.
**C.** Use a Split Rows module to partition the data based on distance travelled to the event.
**D.** Use a Split Rows module to partition the data based on centroid distance.
**Answer: A (LEAVE A REPLY)**
Explanation
Split Data partitions the rows of a dataset into two distinct sets.
The Relative Expression Split option in the Split Data module of Azure Machine Learning Studio is helpful when you need to divide a dataset into training and testing datasets using a numerical expression.
Relative Expression Split: Use this option whenever you want to apply a condition to a number column. The number could be a date/time field, a column containing age or dollar amounts, or even a percentage. For example, you might want to divide your data set depending on the cost of the items, group people by age ranges, or separate data by a calendar date.
Scenario:
Local market segmentation models will be applied before determining a user's propensity to respond to an advertisement.
The distribution of features across training and production data are not consistent References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/split-data
Topic 1, Case Study 1
Overview
You are a data scientist in a company that provides data science for professional sporting events. Models will be global and local market data to meet the following business goals:
*Understand sentiment of mobile device users at sporting events based on audio from crowd reactions.
*Access a user's tendency to respond to an advertisement.
*Customize styles of ads served on mobile devices.
*Use video to detect penalty events.

Current environment

Requirements

* Media used for penalty event detection will be provided by consumer devices. Media may include images and videos captured during the sporting event and snared using social media. The images and videos will have varying sizes and formats.
* The data available for model building comprises of seven years of sporting event media. The sporting event media includes: recorded videos, transcripts of radio commentary, and logs from related social media feeds feeds captured during the sporting events.
*Crowd sentiment will include audio recordings submitted by event attendees in both mono and stereo Formats.

Advertisements

* Ad response models must be trained at the beginning of each event and applied during the sporting event.
* Market segmentation nxxlels must optimize for similar ad resporr.r history.
* Sampling must guarantee mutual and collective exclusivity local and global segmentation models that share the same features.
* Local market segmentation models will be applied before determining a user's propensity to respond to an advertisement.
* Data scientists must be able to detect model degradation and decay.
* Ad response models must support non linear boundaries features.
* The ad propensity model uses a cut threshold is 0.45 and retrains occur if weighted Kappa deviates from 0.1 +/-5%.
* The ad propensity model uses cost factors shown in the following diagram:

The ad propensity model uses proposed cost factors shown in the following diagram:

Performance curves of current and proposed cost factor scenarios are shown in the following diagram:

Penalty detection and sentiment

Findings

*Data scientists must build an intelligent solution by using multiple machine learning models for penalty event detection.
*Data scientists must build notebooks in a local environment using automatic feature engineering and model building in machine learning pipelines.
*Notebooks must be deployed to retrain by using Spark instances with dynamic worker allocation
*Notebooks must execute with the same code on new Spark instances to recode only the source of the data.
*Global penalty detection models must be trained by using dynamic runtime graph computation during training.
*Local penalty detection models must be written by using BrainScript.
* Experiments for local crowd sentiment models must combine local penalty detection data.
* Crowd sentiment models must identify known sounds such as cheers and known catch phrases. Individual crowd sentiment models will detect similar sounds.
* All shared features for local models are continuous variables.
* Shared features must use double precision. Subsequent layers must have aggregate running mean and standard deviation metrics Available.

segments
During the initial weeks in production, the following was observed:
*Ad response rates declined.
*Drops were not consistent across ad styles.
*The distribution of features across training and production data are not consistent.
Analysis shows that of the 100 numeric features on user location and behavior, the 47 features that come from location sources are being used as raw features. A suggested experiment to remedy the bias and variance issue is to engineer 10 linearly uncorrected features.
Penalty detection and sentiment
*Initial data discovery shows a wide range of densities of target states in training data used for crowd sentiment models.
*All penalty detection models show inference phases using a Stochastic Gradient Descent (SGD) are running too stow.
*Audio samples show that the length of a catch phrase varies between 25%-47%, depending on region.
*The performance of the global penalty detection models show lower variance but higher bias when comparing training and validation sets. Before implementing any feature changes, you must confirm the bias and variance using all training and validation cases.

**NEW QUESTION: 190**
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You train and register a machine learning model.
You plan to deploy the model as a real-time web service. Applications must use key-based authentication to use the model.
You need to deploy the web service.
Solution:
Create an AksWebservice instance.
Set the value of the auth_enabled property to True.
Deploy the model to the service.
Does the solution meet the goal?
**A.** Yes
**B.** No
**Answer: A (LEAVE A REPLY)**
Key-based authentication.
Web services deployed on AKS have key-based auth enabled by default. ACI-deployed services have key-based auth disabled by default, but you can enable it by setting auth_enabled = TRUE when creating the ACI web service. The following is an example of creating an ACI deployment configuration with key-based auth enabled.

deployment_config <- aci_webservice_deployment_config(cpu_cores = 1,
memory_gb = 1,
auth_enabled = TRUE)
Reference:
https://azure.github.io/azureml-sdk-for-r/articles/deploying-models.html

**NEW QUESTION: 191**
You are with a time series dataset in Azure Machine Learning Studio.
You need to split your dataset into training and testing subsets by using the Split Data module.
Which splitting mode should you use?
**A.** Split Rows with the Randomized split parameter set to true
**B.** Recommender Split
**C.** Relative Expression Split
**D.** Regular Expression Split
**Answer: (SHOW ANSWER)**
Topic 2, Case Study
Overview
You are a data scientist for Fabrikam Residences, a company specializing in quality private and commercial property in the United States. Fabrikam Residences is considering expanding into Europe and has asked you to investigate prices for private residences in major European cities. You use Azure Machine Learning Studio to measure the median value of properties. You produce a regression model to predict property prices by using the Linear Regression and Bayesian Linear Regression modules.
Datasets
There are two datasets in CSV format that contain property details for two cities, London and Paris, with the following columns:
The two datasets have been added to Azure Machine Learning Studio as separate datasets and included as the starting point of the experiment.
Dataset issues
The AccessibilityToHighway column in both datasets contains missing values. The missing data must be replaced with new data so that it is modeled conditionally using the other variables in the data before filling in the missing values.
Columns in each dataset contain missing and null values. The dataset also contains many outliers. The Age column has a high proportion of outliers. You need to remove the rows that have outliers in the Age column.
The MedianValue and AvgRoomsinHouse columns both hold data in numeric format. You need to select a feature selection algorithm to analyze the relationship between the two columns in more detail.
Model fit
The model shows signs of overfitting. You need to produce a more refined regression model that reduces the overfitting.
Experiment Requirements
You must set up the experiment to cross-validate the Linear Regression and Bayesian Linear Regression modules to evaluate performance.

In each case, the predictor of the dataset is the column named MedianValue. An initial investigation showed that the datasets are identical in structure apart from the MedianValue column. The smaller Paris dataset contains the MedianValue in text format, whereas the larger London dataset contains the MedianValue in numerical format. You must ensure that the datatype of the MedianValue column of the Paris dataset matches the structure of the London dataset.
You must prioritize the columns of data for predicting the outcome. You must use non-parameters statistics to measure the relationships.
You must use a feature selection algorithm to analyze the relationship between the MedianValue and AvgRoomsinHouse columns.
Model training
Given a trained model and a test dataset, you need to compute the permutation feature importance scores of feature variables. You need to set up the Permutation Feature Importance module to select the correct metric to investigate the model's accuracy and replicate the findings.
You want to configure hyperparameters in the model learning process to speed the learning phase by using hyperparameters. In addition, this configuration should cancel the lowest performing runs at each evaluation interval, thereby directing effort and resources towards models that are more likely to be successful.
You are concerned that the model might not efficiently use compute resources in hyperparameter tuning. You also are concerned that the model might prevent an increase in the overall tuning time. Therefore, you need to implement an early stopping criterion on models that provides savings without terminating promising jobs.
Testing
You must produce multiple partitions of a dataset based on sampling using the Partition and Sample module in Azure Machine Learning Studio. You must create three equal partitions for cross-validation. You must also configure the cross-validation process so that the rows in the test and training datasets are divided evenly by properties that are near each city's main river. The data that identifies that a property is near a river is held in the column named NextToRiver. You want to complete this task before the data goes through the sampling process.
When you train a Linear Regression module using a property dataset that shows data for property prices for a large city, you need to determine the best features to use in a model. You can choose standard metrics provided to measure performance before and after the feature importance process completes. You must ensure that the distribution of the features across multiple training models is consistent.
Data visualization
You need to provide the test results to the Fabrikam Residences team. You create data visualizations to aid in presenting the results.
You must produce a Receiver Operating Characteristic (ROC) curve to conduct a diagnostic test evaluation of the model. You need to select appropriate methods for producing the ROC curve in Azure Machine Learning Studio to compare the Two-Class Decision Forest and the Two-Class Decision Jungle modules with one another.

**NEW QUESTION: 192**
You are solving a classification task.
The dataset is imbalanced.

You need to select an Azure Machine Learning Studio module to improve the classification accuracy.
Which module should you use?
**A.** Permutation Feature Importance
**B.** Filter Based Feature Selection
**C.** Fisher Linear Discriminant Analysis
**D.** Synthetic Minority Oversampling Technique (SMOTE)
**Answer: (SHOW ANSWER)**
Explanation
Explanation:
Use the SMOTE module in Azure Machine Learning Studio (classic) to increase the number of underepresented cases in a dataset used for machine learning. SMOTE is a better way of increasing the number of rare cases than simply duplicating existing cases.
You connect the SMOTE module to a dataset that is imbalanced. There are many reasons why a dataset might be imbalanced: the category you are targeting might be very rare in the population, or the data might simply be difficult to collect. Typically, you use SMOTE when the class you want to analyze is under-represented.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/smote Prepare data for modeling Testlet 1 Case study Overview You are a data scientist in a company that provides data science for professional sporting events. Models will use global and local market data to meet the following business goals:
* Understand sentiment of mobile device users at sporting events based on audio from crowd reactions.
* Assess a user's tendency to respond to an advertisement.
* Customize styles of ads served on mobile devices.
* Use video to detect penalty events
Current environment
* Media used for penalty event detection will be provided by consumer devices. Media may include images and videos captured during the sporting event and shared using social media. The images and videos will have varying sizes and formats.
* The data available for model building comprises of seven years of sporting event media. The sporting event media includes; recorded video transcripts or radio commentary, and logs from related social media feeds captured during the sporting events.
* Crowd sentiment will include audio recordings submitted by event attendees in both mono and stereo formats.
Penalty detection and sentiment
* Data scientists must build an intelligent solution by using multiple machine learning models for penalty event detection.
* Data scientists must build notebooks in a local environment using automatic feature engineering and model building in machine learning pipelines.
* Notebooks must be deployed to retrain by using Spark instances with dynamic worker allocation.
* Notebooks must execute with the same code on new Spark instances to recode only the source of the data.

* Global penalty detection models must be trained by using dynamic runtime graph computation during training.
* Local penalty detection models must be written by using BrainScript.
* Experiments for local crowd sentiment models must combine local penalty detection data.
* Crowd sentiment models must identify known sounds such as cheers and known catch phrases. Individual crowd sentiment models will detect similar sounds.
* All shared features for local models are continuous variables.
* Shared features must use double precision. Subsequent layers must have aggregate running mean and standard deviation metrics available.
Advertisements
During the initial weeks in production, the following was observed:
* Ad response rated declined.
* Drops were not consistent across ad styles.
* The distribution of features across training and production data are not consistent Analysis shows that, of the 100 numeric features on user location and behavior, the 47 features that come from location sources are being used as raw features. A suggested experiment to remedy the bias and variance issue is to engineer 10 linearly uncorrelated features.
* Initial data discovery shows a wide range of densities of target states in training data used for crowd sentiment models.
* All penalty detection models show inference phases using a Stochastic Gradient Descent (SGD) are running too slow.
* Audio samples show that the length of a catch phrase varies between 25%-47% depending on region
* The performance of the global penalty detection models shows lower variance but higher bias when comparing training and validation sets. Before implementing any feature changes, you must confirm the bias and variance using all training and validation cases.
* Ad response models must be trained at the beginning of each event and applied during the sporting event.
* Market segmentation models must optimize for similar ad response history.
* Sampling must guarantee mutual and collective exclusively between local and global segmentation models that share the same features.
* Local market segmentation models will be applied before determining a user's propensity to respond to an advertisement.
* Ad response models must support non-linear boundaries of features.
* The ad propensity model uses a cut threshold is 0.45 and retrains occur if weighted Kappa deviated from 0.1 +/- 5%.
* The ad propensity model uses cost factors shown in the following diagram:
* The ad propensity model uses proposed cost factors shown in the following diagram:
* Performance curves of current and proposed cost factor scenarios are shown in the following diagram:

**NEW QUESTION: 193**
You use Azure Machine Learning to train and register a model.

You must deploy the model into production as a real-time web service to an inference cluster named service-compute that the IT department has created in the Azure Machine Learning workspace.
Client applications consuming the deployed web service must be authenticated based on their Azure Active Directory service principal.
You need to write a script that uses the Azure Machine Learning SDK to deploy the model. The necessary modules have been imported.
How should you complete the code? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
**Answer:**
Explanation
Box 1: AksCompute
Example:
aks_target = AksCompute(ws,"myaks")
# If deploying to a cluster configured for dev/test, ensure that it was created with enough
# cores and memory to handle this deployment configuration. Note that memory is also used by
# things such as dependencies and AML components.
deployment_config = AksWebservice.deploy_configuration(cpu_cores = 1, memory_gb = 1) service = Model.deploy(ws, "myservice", [model], inference_config, deployment_config, aks_target) Box 2: AksWebservice Box 3: token_auth_enabled=Yes Whether or not token auth is enabled for the Webservice.
Note: A Service principal defined in Azure Active Directory (Azure AD) can act as a principal on which authentication and authorization policies can be enforced in Azure Databricks.
The Azure Active Directory Authentication Library (ADAL) can be used to programmatically get an Azure AD access token for a user.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-deploy-azure-kubernetes-service
https://docs.microsoft.com/en-us/azure/databricks/dev-tools/api/latest/aad/service-prin-aad-token

**NEW QUESTION: 194**
You use the Azure Machine Learning service to create a tabular dataset named training.data. You plan to use this dataset in a training script.
You create a variable that references the dataset using the following code:
training_ds = workspace.datasets.get("training_data")
You define an estimator to run the script.
You need to set the correct property of the estimator to ensure that your script can access the training.data dataset Which property should you set?
A)
B)
C)
D)
**A.** Option A

**B.** Option B
**C.** Option C
**D.** Option D
**Answer: A (LEAVE A REPLY)**
Explanation
Example:
# Get the training dataset
diabetes_ds = ws.datasets.get("Diabetes Dataset")
# Create an estimator that uses the remote compute
hyper_estimator = SKLearn(source_directory=experiment_folder,
inputs=[diabetes_ds.as_named_input('diabetes')], # Pass the dataset as an input compute_target = cpu_cluster, conda_packages=['pandas','ipykernel','matplotlib'], pip_packages=['azureml-sdk','argparse','pyarrow'], entry_script='diabetes_training.py') Reference:
https://notebooks.azure.com/GraemeMalcolm/projects/azureml-primers/html/04%20-%20Optimizing%20Model

**NEW QUESTION: 195**
You plan to use a Deep Learning Virtual Machine (DLVM) to train deep learning models using Compute Unified Device Architecture (CUDA) computations.
You need to configure the DLVM to support CUDA.
What should you implement?
**A.** Intel Software Guard Extensions (Intel SGX) technology
**B.** Solid State Drives (SSD)
**C.** Graphic Processing Unit (GPU)
**D.** Computer Processing Unit (CPU) speed increase by using overcloking
**E.** High Random Access Memory (RAM) configuration
**Answer: (SHOW ANSWER)**
A Deep Learning Virtual Machine is a pre-configured environment for deep learning using GPU instances.
References:
https://azuremarketplace.microsoft.com/en-au/marketplace/apps/microsoft-ads.dsvm-deep-learning

**NEW QUESTION: 196**
You have a model with a large difference between the training and validation error values.
You must create a new model and perform cross-validation.
You need to identify a parameter set for the new model using Azure Machine Learning Studio.
Which module you should use for each step? To answer, drag the appropriate modules to the correct steps. Each module may be used once or more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.
NOTE: Each correct selection is worth one point.
**Answer:**
Explanation:

Box 1: Split data
Box 2: Partition and Sample
Box 3: Two-Class Boosted Decision Tree
Box 4: Tune Model Hyperparameters
Integrated train and tune: You configure a set of parameters to use, and then let the module iterate over multiple combinations, measuring accuracy until it finds a "best" model. With most learner modules, you can choose which parameters should be changed during the training process, and which should remain fixed.
We recommend that you use Cross-Validate Model to establish the goodness of the model given the specified parameters. Use Tune Model Hyperparameters to identify the optimal parameters.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/partition-and-sample



**NEW QUESTION: 197**
You create a multi-class image classification deep learning model.
The model must be retrained monthly with the new image data fetched from a public web portal. You create an Azure Machine Learning pipeline to fetch new data, standardize the size of images, and retrain the model.
You need to use the Azure Machine Learning SDK to configure the schedule for the pipeline.
Which four actions should you perform in sequence. To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.
**Answer:**
1 - Define a new Azure Machine Learning pipline StepRun object with the step ID of the first step in the pipline.
2 - Retrive the pipeline ID.
3 - Create a scheduleRecurrence(frequency*'Month', interval*1, start_time*'2019-01-01T00:00:00') object.
4 - Define a pipeline parameter named RunDate.

**NEW QUESTION: 198**
You are analyzing the asymmetry in a statistical distribution.
The following image contains two density curves that show the probability distribution of two datasets.
Use the drop-down menus to select the answer choice that answers each question based on the information presented in the graphic.
NOTE: Each correct selection is worth one point.
**Answer:**

Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/compute-elementary-statistics

**NEW QUESTION: 199**
You need to build a feature extraction strategy for the local models.
How should you complete the code segment? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
**Answer:**
Explanation

**NEW QUESTION: 200**
You use an Azure Machine Learning workspace.
You create the following Python code:
For each of the following statements, select Yes if the statement is true. Otherwise, select No.
NOTE: Each correct selection is worth one point.
**Answer:**
Reference:
https://docs.microsoft.com/en-us/python/api/azureml-core/azureml.core.scriptrunconfig
https://docs.microsoft.com/en-us/python/api/azureml-core/azureml.core.environment.environment

**NEW QUESTION: 201**
You need to visually identify whether outliers exist in the Age column and quantify the outliers before the outliers are removed.
Which three Azure Machine Learning Studio modules should you use in sequence? To answer, move the appropriate modules from the list of modules to the answer area and arrange them in the correct order.
Create Scatterplot
Summarize Data
Clip Values
You can use the Clip Values module in Azure Machine Learning Studio, to identify and optionally replace data values that are above or below a specified threshold. This is useful when you want to remove outliers or replace them with a mean, a constant, or other substitute value.
**Answer:**
References:
https://blogs.msdn.microsoft.com/azuredev/2017/05/27/data-cleansing-tools-in-azure-machine-learning/
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/clip-values

**NEW QUESTION: 202**

You are creating a machine learning model in Python. The provided dataset contains several numerical columns and one text column. The text column represents a product's category. The product category will always be one of the following:
Bikes
Cars
Vans
Boats
You are building a regression model using the scikit-learn Python package.
You need to transform the text data to be compatible with the scikit-learn Python package.
How should you complete the code segment? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
**Answer:**
Explanation:
Box 1: pandas as df
Pandas takes data (like a CSV or TSV file, or a SQL database) and creates a Python object with rows and columns called data frame that looks very similar to table in a statistical software (think Excel or SPSS for example.
Box 2: transpose[ProductCategoryMapping]
Reshape the data from the pandas Series to columns.
Reference:
https://datascienceplus.com/linear-regression-in-python/

**NEW QUESTION: 203**
You plan to deliver a hands-on workshop to several students. The workshop will focus on creating data visualizations using Python. Each student will use a device that has internet access.
Student devices are not configured for Python development. Students do not have administrator access to install software on their devices. Azure subscriptions are not available for students.
You need to ensure that students can run Python-based data visualization code.
Which Azure tool should you use?
**A.** Anaconda Data Science Platform
**B.** Azure BatchAI
**C.** Azure Notebooks
**D.** Azure Machine Learning Service
**Answer: C (LEAVE A REPLY)**
Reference:
https://notebooks.azure.com/

**NEW QUESTION: 204**
You create an Azure Machine Learning compute target named ComputeOne by using the STANDARD_D1 virtual machine image.

You define a Python variable named was that references the Azure Machine Learning workspace. You run the following Python code:
For each of the following statements, select Yes if the statement is true. Otherwise, select No.
NOTE: Each correct selection is worth one point.
**Answer:**
Reference:
https://docs.microsoft.com/en-us/python/api/azureml-core/azureml.core.compute.computetarget

**NEW QUESTION: 205**
You are performing feature engineering on a dataset.
You must add a feature named CityName and populate the column value with the text London.
You need to add the new feature to the dataset.
Which Azure Machine Learning Studio module should you use?
**A.** Edit Metadata
**B.** Preprocess Text
**C.** Execute Python Script
**D.** Latent Dirichlet Allocation
**Answer: A (LEAVE A REPLY)**
Explanation/Reference:
Explanation:
Typical metadata changes might include marking columns as features.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/edit-metadata Testlet 1
Case study Overview You are a data scientist in a company that provides data science for professional sporting events. Models will use global and local market data to meet the following business goals:
Understand sentiment of mobile device users at sporting events based on audio from crowd reactions.
Assess a user's tendency to respond to an advertisement.
Customize styles of ads served on mobile devices.
Use video to detect penalty events
Current environment
Media used for penalty event detection will be provided by consumer devices. Media may include images and videos captured during the sporting event and shared using social media. The images and videos will have varying sizes and formats.
The data available for model building comprises of seven years of sporting event media. The sporting event media includes; recorded video transcripts or radio commentary, and logs from related social media feeds captured during the sporting events.
Crowd sentiment will include audio recordings submitted by event attendees in both mono and stereo formats.
Penalty detection and sentiment
Data scientists must build an intelligent solution by using multiple machine learning models for penalty

event detection.
Data scientists must build notebooks in a local environment using automatic feature engineering and model building in machine learning pipelines.
Notebooks must be deployed to retrain by using Spark instances with dynamic worker allocation.
Notebooks must execute with the same code on new Spark instances to recode only the source of the data.
Global penalty detection models must be trained by using dynamic runtime graph computation during training.
Local penalty detection models must be written by using BrainScript.
Experiments for local crowd sentiment models must combine local penalty detection data.
Crowd sentiment models must identify known sounds such as cheers and known catch phrases.
Individual crowd sentiment models will detect similar sounds.
All shared features for local models are continuous variables.
Shared features must use double precision. Subsequent layers must have aggregate running mean and standard deviation metrics available.
Advertisements
During the initial weeks in production, the following was observed:
Ad response rated declined.
Drops were not consistent across ad styles.
The distribution of features across training and production data are not consistent
Analysis shows that, of the 100 numeric features on user location and behavior, the 47 features that come from location sources are being used as raw features. A suggested experiment to remedy the bias and variance issue is to engineer 10 linearly uncorrelated features.
Initial data discovery shows a wide range of densities of target states in training data used for crowd sentiment models.
All penalty detection models show inference phases using a Stochastic Gradient Descent (SGD) are running too slow.
Audio samples show that the length of a catch phrase varies between 25%-47% depending on region
The performance of the global penalty detection models shows lower variance but higher bias when comparing training and validation sets. Before implementing any feature changes, you must confirm the bias and variance using all training and validation cases.
Ad response models must be trained at the beginning of each event and applied during the sporting event.
Market segmentation models must optimize for similar ad response history.
Sampling must guarantee mutual and collective exclusively between local and global segmentation models that share the same features.
Local market segmentation models will be applied before determining a user's propensity to respond to an advertisement.
Ad response models must support non-linear boundaries of features.
The ad propensity model uses a cut threshold is 0.45 and retrains occur if weighted Kappa deviated

from 0.1 +/- 5%.
The ad propensity model uses cost factors shown in the following diagram:
The ad propensity model uses proposed cost factors shown in the following diagram:
Performance curves of current and proposed cost factor scenarios are shown in the following diagram:

**NEW QUESTION: 206**
You deploy a model as an Azure Machine Learning real-time web service using the following code.
The deployment fails.
You need to troubleshoot the deployment failure by determining the actions that were performed during deployment and identifying the specific action that failed.
Which code segment should you run?
**A.** service.get_logs()
**B.** service.state
**C.** service.serialize()
**D.** service.update_deployment_state()
**Answer: A (LEAVE A REPLY)**
Explanation
You can print out detailed Docker engine log messages from the service object. You can view the log for ACI, AKS, and Local deployments. The following example demonstrates how to print the logs.
# if you already have the service object handy
print(service.get_logs())
# if you only know the name of the service (note there might be multiple services with the same name but different version number) print(ws.webservices['mysvc'].get_logs()) Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-troubleshoot-deployment

**NEW QUESTION: 207**
You are performing a filter based feature selection for a dataset 10 build a multi class classifies by using Azure Machine Learning Studio.
The dataset contains categorical features that are highly correlated to the output label column.
You need to select the appropriate feature scoring statistical method to identify the key predictors. Which method should you use?
**A.** Chi-squared
**B.** Spearman correlation
**C.** Kendall correlation
**D.** Person correlation
**Answer: D (LEAVE A REPLY)**
Explanation
Pearson's correlation statistic, or Pearson's correlation coefficient, is also known in statistical models as the r value. For any two variables, it returns a value that indicates the strength of the correlation Pearson's correlation coefficient is the test statistics that measures the statistical relationship, or association, between

two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/filter-based-feature-selection
https://www.statisticssolutions.com/pearsons-correlation-coefficient/

**NEW QUESTION: 208**
You create an Azure Machine Learning workspace.
You must create a custom role named that meets the following requirements:
* Role members must not be able to delete the workspace.
* Role members must not be able to create, update, or delete compute resource in the workspace.
* Role members must not be able to add new users to the workspace.
You need to create a JSON file for the DataScientist role in the Azure Machine Learning workspace.
The custom role must enforce the restrictions specified by the IT Operations team.
Which JSON code segment should you use?
A)
B)
C)
D)
**A.** Option A
**B.** Option B
**C.** Option C
**D.** Option D
**Answer: A (LEAVE A REPLY)**
Explanation
The following custom role can do everything in the workspace except for the following actions:
* It can't create or update a compute resource.
* It can't delete a compute resource.
* It can't add, delete, or alter role assignments.
* It can't delete the workspace.
To create a custom role, first construct a role definition JSON file that specifies the permission and scope for the role. The following example defines a custom role named "Data Scientist Custom" scoped at a specific workspace level:
data_scientist_custom_role.json :
{
"Name": "Data Scientist Custom",
"IsCustom": true,
"Description": "Can run experiment but can't create or delete compute.",
"Actions": ["*"],

"NotActions": [
"Microsoft.MachineLearningServices/workspaces/*/delete",
"Microsoft.MachineLearningServices/workspaces/write",
"Microsoft.MachineLearningServices/workspaces/computes/*/write",
"Microsoft.MachineLearningServices/workspaces/computes/*/delete",
"Microsoft.Authorization/*/write"
],
"AssignableScopes": [
"/subscriptions/<subscription_id>/resourceGroups/<resource_group_name>/providers/Microsoft.MachineLearni
]
}
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-assign-roles

**NEW QUESTION: 209**
You need to identify the methods for dividing the data according to the testing requirements.
Which properties should you select? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
**Answer:**
Explanation
Scenario: Testing
You must produce multiple partitions of a dataset based on sampling using the Partition and Sample module in Azure Machine Learning Studio.
Box 1: Assign to folds
Use Assign to folds option when you want to divide the dataset into subsets of the data. This option is also useful when you want to create a custom number of folds for cross-validation, or to split rows into several groups.
Not Head: Use Head mode to get only the first n rows. This option is useful if you want to test a pipeline on a small number of rows, and don't need the data to be balanced or sampled in any way.
Not Sampling: The Sampling option supports simple random sampling or stratified random sampling. This is useful if you want to create a smaller representative sample dataset for testing.
Box 2: Partition evenly
Specify the partitioner method: Indicate how you want data to be apportioned to each partition, using these options:
* Partition evenly: Use this option to place an equal number of rows in each partition. To specify the number of output partitions, type a whole number in the Specify number of folds to split evenly into text box.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/algorithm-module-reference/partition-and-sample

**NEW QUESTION: 210**
You configure a Deep Learning Virtual Machine for Windows.

You need to recommend tools and frameworks to perform the following:
* Build deep neural network (DNN) models
* Perform interactive data exploration and visualization
Which tools and frameworks should you recommend? To answer, drag the appropriate tools to the correct tasks. Each tool may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.
NOTE: Each correct selection is worth one point.
**Answer:**
Explanation
Box 1: Vowpal Wabbit
Use the Train Vowpal Wabbit Version 8 module in Azure Machine Learning Studio (classic), to create a machine learning model by using Vowpal Wabbit.
Box 2: PowerBI Desktop
Power BI Desktop is a powerful visual data exploration and interactive reporting tool BI is a name given to a modern approach to business decision making in which users are empowered to find, explore, and share insights from data across the enterprise.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/train-vowpal-wabbit-version-8
https://docs.microsoft.com/en-us/azure/architecture/data-guide/scenarios/interactive-data-exploration