DP-100.examcollection.premium.exam.153q

Number: DP-100
Passing Score: 800
Time Limit: 120 min
File Version: 5.0

DP-100

Designing and Implementing a Data Science Solution on Azure

Version 5.0

Define and prepare the development environment

Question Set 1

QUESTION 1
You are developing a hands-on workshop to introduce Docker for Windows to attendees.

You need to ensure that workshop attendees can install Docker on their devices.

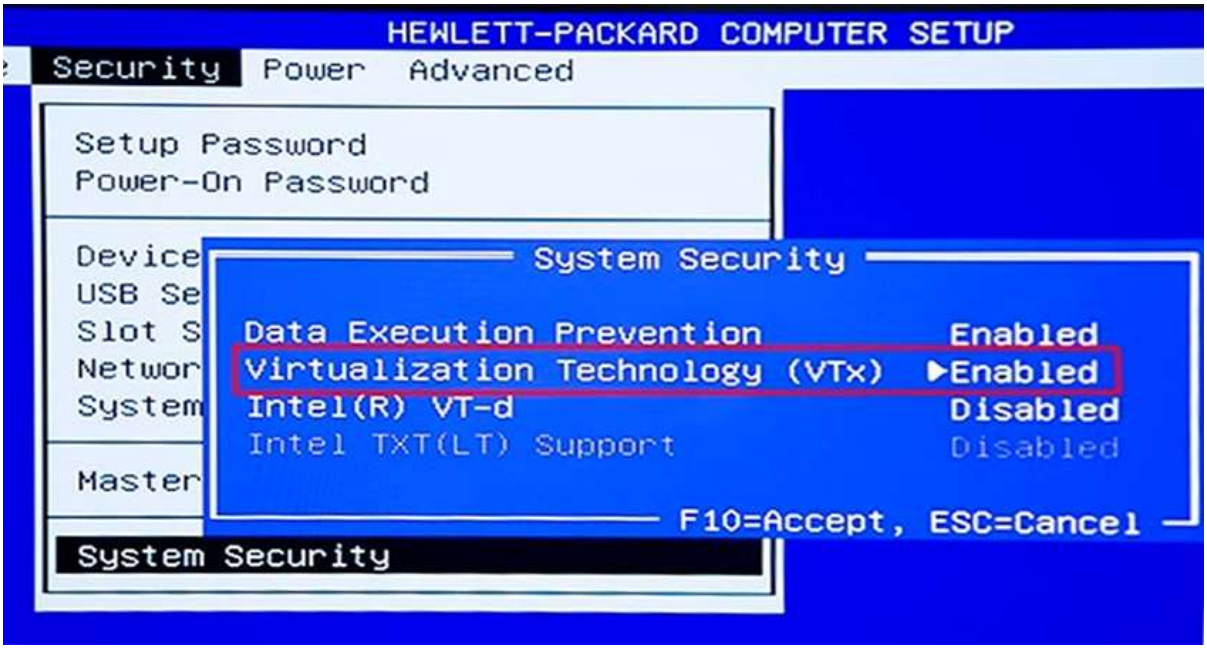
Which two prerequisite components should attendees install on the devices? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

A. Microsoft Hardware-Assisted Virtualization Detection Tool
B. Kitematic
C. BIOS-enabled virtualization
D. VirtualBox
E. Windows 10 64-bit Professional

Correct Answer: CE
Section: (none)
Explanation

Explanation/Reference:
Explanation:
C: Make sure your Windows system supports Hardware Virtualization Technology and that virtualization is enabled.
Ensure that hardware virtualization support is turned on in the BIOS settings. For example:



E: To run Docker, your machine must have a 64-bit operating system running Windows 7 or higher.

References:
https://docs.docker.com/toolbox/toolbox_install_windows/

https://blogs.technet.microsoft.com/canitpro/2015/09/08/step-by-step-enabling-hyper-v-for-use-on-windows-10/

QUESTION 2
Your team is building a data engineering and data science development environment.

The environment must support the following requirements:

- support Python and Scala
- compose data storage, movement, and processing services into automated data pipelines
- the same tool should be used for the orchestration of both data engineering and data science
- support workload isolation and interactive workloads
- enable scaling across a cluster of machines

You need to create the environment.

What should you do?

A. Build the environment in Apache Hive for HDInsight and use Azure Data Factory for orchestration.
B. Build the environment in Azure Databricks and use Azure Data Factory for orchestration.
C. Build the environment in Apache Spark for HDInsight and use Azure Container Instances for orchestration.
D. Build the environment in Azure Databricks and use Azure Container Instances for orchestration.

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:
In Azure Databricks, we can create two different types of clusters.
- Standard, these are the default clusters and can be used with Python, R, Scala and SQL
- High-concurrency

Azure Databricks is fully integrated with Azure Data Factory.

Incorrect Answers:
D: Azure Container Instances is good for development or testing. Not suitable for production workloads.

References:
https://docs.microsoft.com/en-us/azure/architecture/data-guide/technology-choices/data-science-and-machine-learning

QUESTION 3
DRAG DROP

You are building an intelligent solution using machine learning models.

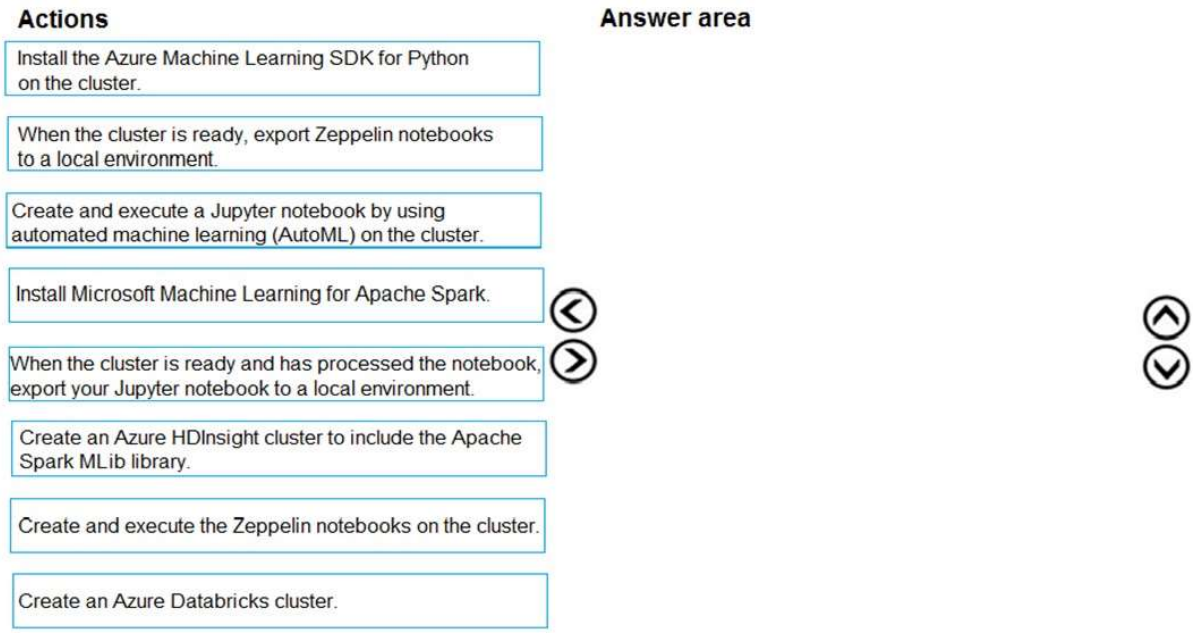
The environment must support the following requirements:

- Data scientists must build notebooks in a cloud environment
- Data scientists must use automatic feature engineering and model building in machine learning pipelines.
- Notebooks must be deployed to retrain using Spark instances with dynamic worker allocation.
- Notebooks must be exportable to be version controlled locally.

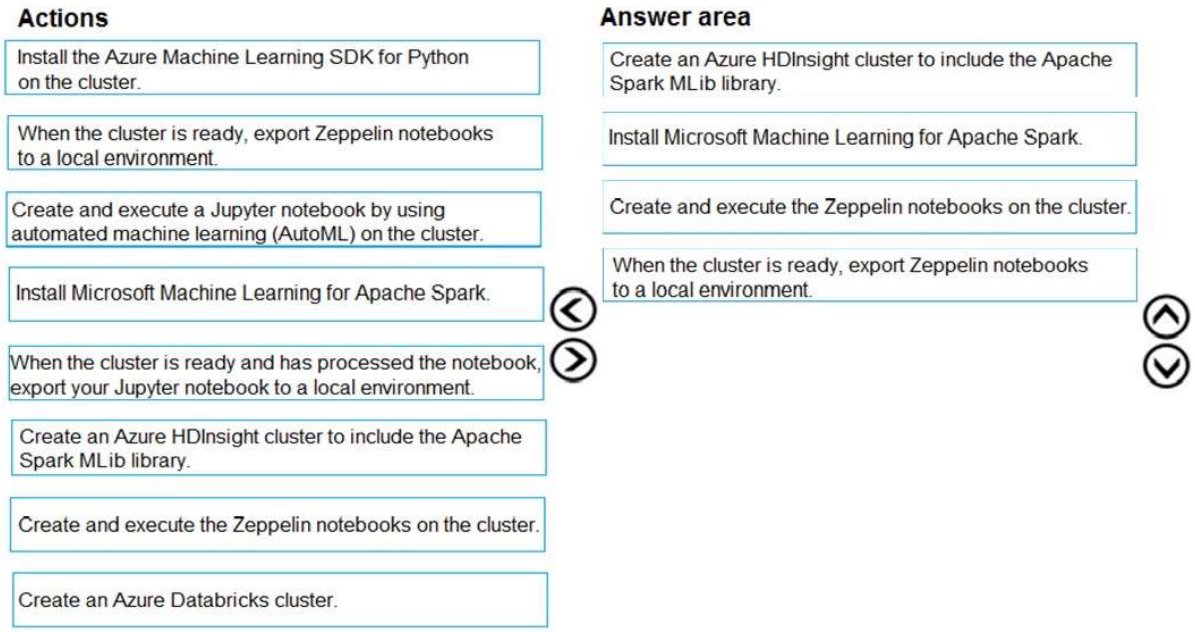
You need to create the environment.

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.

Select and Place:



Correct Answer:



Section: (none)
Explanation

Explanation/Reference:
Explanation:

Step 1: Create an Azure HDInsight cluster to include the Apache Spark Mlib library

Step 2: Install Microsot Machine Learning for Apache Spark
You install AzureML on your Azure HDInsight cluster.
Microsoft Machine Learning for Apache Spark (MMLSpark) provides a number of deep learning and data science tools for Apache Spark, including seamless integration of Spark Machine Learning pipelines with Microsoft Cognitive Toolkit (CNTK) and OpenCV, enabling you to quickly create powerful, highly-scalable predictive and analytical models for large image and text datasets.

Step 3: Create and execute the Zeppelin notebooks on the cluster

Step 4: When the cluster is ready, export Zeppelin notebooks to a local environment.
Notebooks must be exportable to be version controlled locally.

References:
https://docs.microsoft.com/en-us/azure/hdinsight/spark/apache-spark-zeppelin-notebook

https://azuremlbuild.blob.core.windows.net/pysparkapi/intro.html

QUESTION 4
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)

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:
Explanation:
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.

References:
https://docs.microsoft.com/en-us/azure/machine-learning/data-science-virtual-machine/overview

QUESTION 5
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.

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:
Explanation:
In the DSVM, your training models can use deep learning algorithms on hardware that's based on graphics processing units (GPUs).

PostgreSQL is available for the following operating systems: Linux (all recent distributions), 64-bit installers available for macOS (OS X) version 10.6 and newer – Windows (with installers available for 64-bit version; tested on latest versions and back to Windows 2012 R2.

Incorrect Answers:

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.

C, 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.

Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/data-science-virtual-machine/overview

QUESTION 6
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

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:
Explanation:
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

QUESTION 7
You must store data in Azure Blob Storage to support Azure Machine Learning.

You need to transfer the data into Azure Blob Storage.

What are three possible ways to achieve the goal? Each correct answer presents a complete solution.

NOTE: Each correct selection is worth one point.

A. Bulk Insert SQL Query
B. AzCopy
C. Python script
D. Azure Storage Explorer
E. Bulk Copy Program (BCP)

Correct Answer: BCD
Section: (none)
Explanation

Explanation/Reference:
Explanation:
You can move data to and from Azure Blob storage using different technologies:

- Azure Storage-Explorer
- AzCopy
- Python
- SSIS

References:
https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/move-azure-blob

QUESTION 8
You are moving a large dataset from Azure Machine Learning Studio to a Weka environment.

You need to format the data for the Weka environment.

Which module should you use?

A. Convert to CSV
B. Convert to Dataset
C. Convert to ARFF
D. Convert to SVMLight

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Use the Convert to ARFF module in Azure Machine Learning Studio, to convert datasets and results in Azure

Machine Learning to the attribute-relation file format used by the Weka toolset. This format is known as ARFF.

The ARFF data specification for Weka supports multiple machine learning tasks, including data preprocessing, classification, and feature selection. In this format, data is organized by entites and their attributes, and is contained in a single text file.

References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/convert-to-arff

QUESTION 9
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

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
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

QUESTION 10
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. Solid State Drives (SSD)
B. Computer Processing Unit (CPU) speed increase by using overclocking
C. Graphic Processing Unit (GPU)
D. High Random Access Memory (RAM) configuration

E. Intel Software Guard Extensions (Intel SGX) technology

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:
Explanation:
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

QUESTION 11
You plan to use a Data Science Virtual Machine (DSVM) with the open source deep learning frameworks Caffe2 and PyTorch.

You need to select a pre-configured DSVM to support the frameworks.

What should you create?

A. Data Science Virtual Machine for Windows 2012
B. Data Science Virtual Machine for Linux (CentOS)
C. Geo AI Data Science Virtual Machine with ArcGIS
D. Data Science Virtual Machine for Windows 2016
E. Data Science Virtual Machine for Linux (Ubuntu)

Correct Answer: E
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Caffe2 and PyTorch is supported by Data Science Virtual Machine for Linux.
Microsoft offers Linux editions of the DSVM on Ubuntu 16.04 LTS and CentOS 7.4.
Only the DSVM on Ubuntu is preconfigured for Caffe2 and PyTorch.

Incorrect Answers:
D: Caffe2 and PytOCH are only supported in the Data Science Virtual Machine for Linux.

References:
https://docs.microsoft.com/en-us/azure/machine-learning/data-science-virtual-machine/overview

QUESTION 12
HOTSPOT

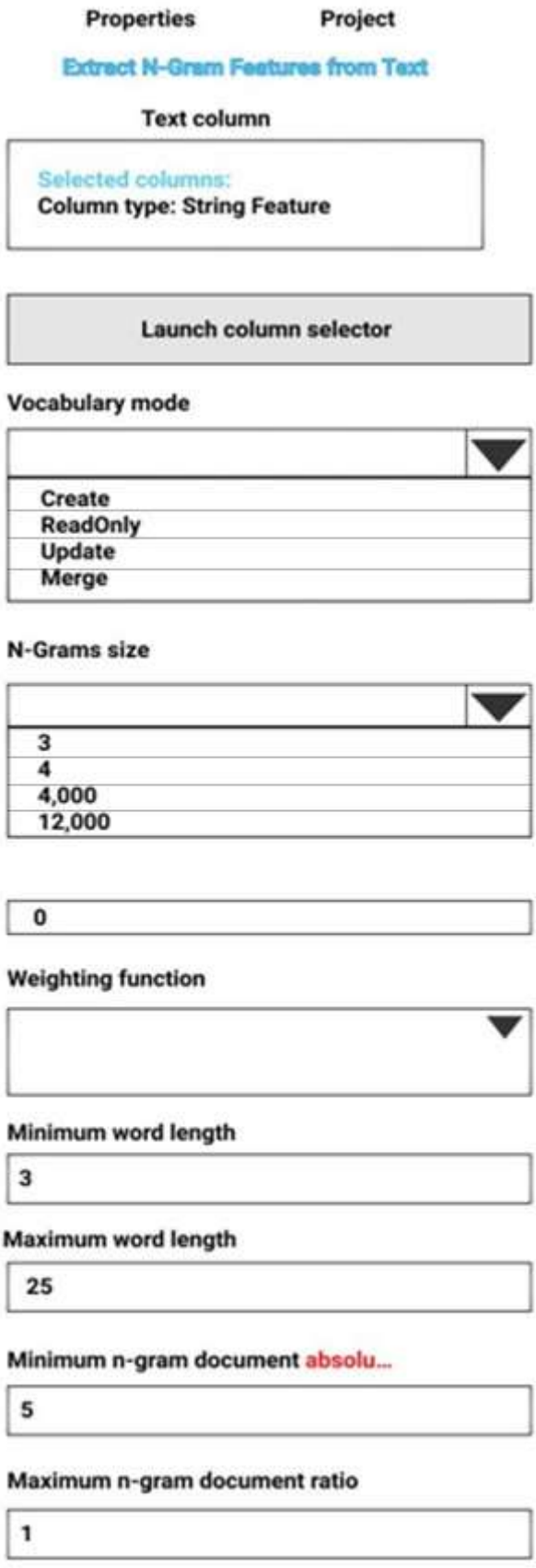
You are performing sentiment analysis using a CSV file that includes 12,000 customer reviews written in a short sentence format. You add the CSV file to Azure Machine Learning Studio and configure it as the starting point dataset of an experiment. You add the Extract N-Gram Features from Text module 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.

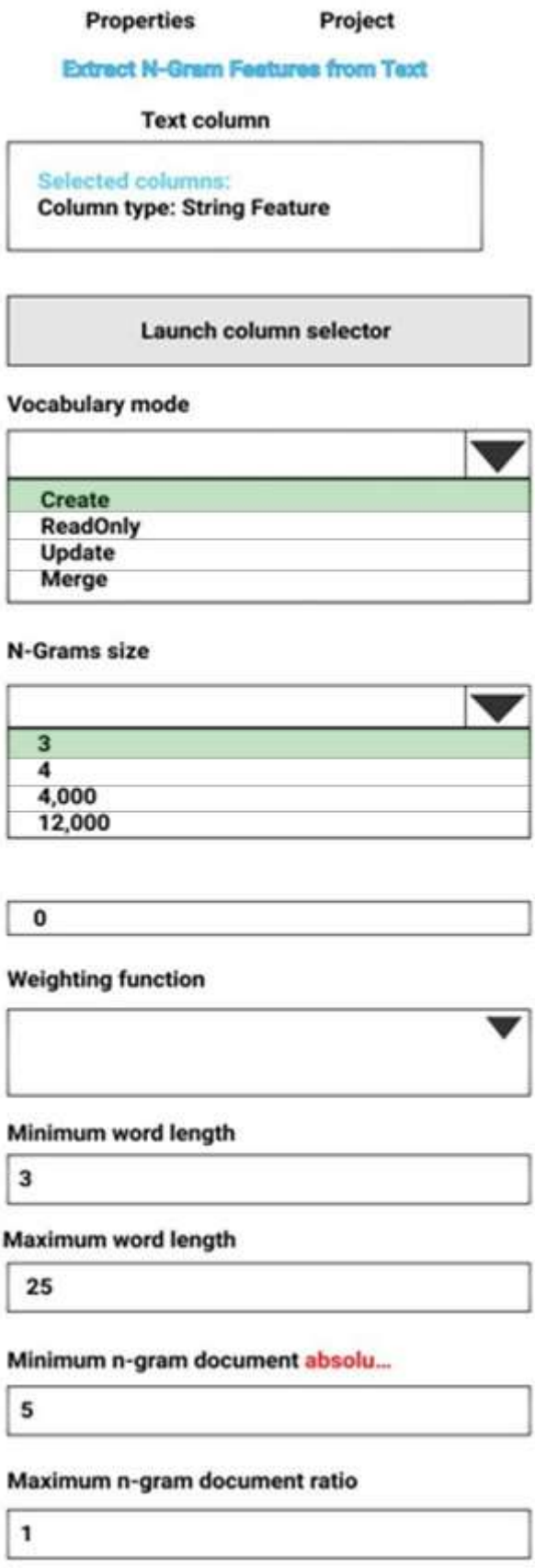
What should you select? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:


Properties
Project
Extract N-Gram Features from Text
Text column
Selected columns:
Column type: String Feature
Launch column selector
Vocabulary mode
Create
ReadOnly
Update
Merge
N-Grams size
3
4
4,000
12,000
0
Weighting function
Minimum word length
3
Maximum word length
25
Minimum n-gram document absolu...
5
Maximum n-gram document ratio
1

Correct Answer:



Section: (none)
Explanation

Explanation/Reference:
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-text

QUESTION 13
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

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:
Explanation:
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

QUESTION 14
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)

Correct Answer: D
Section: (none)
Explanation

Explanation/Reference:
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

QUESTION 15
DRAG DROP

You configure a Deep Learning Virtual Machine for Windows.

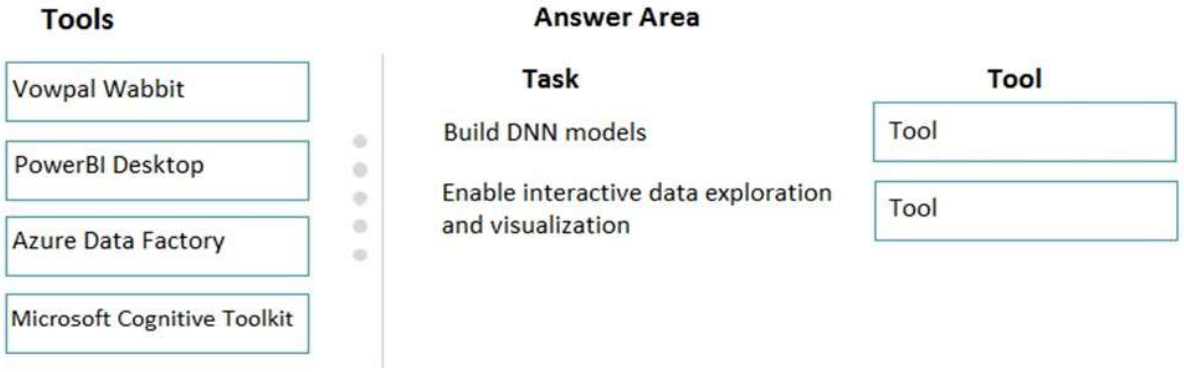
You need to recommend tools and frameworks to perform the following:

- Build deep neural network (DNN) models
- Perform interactive data exploration and visualization

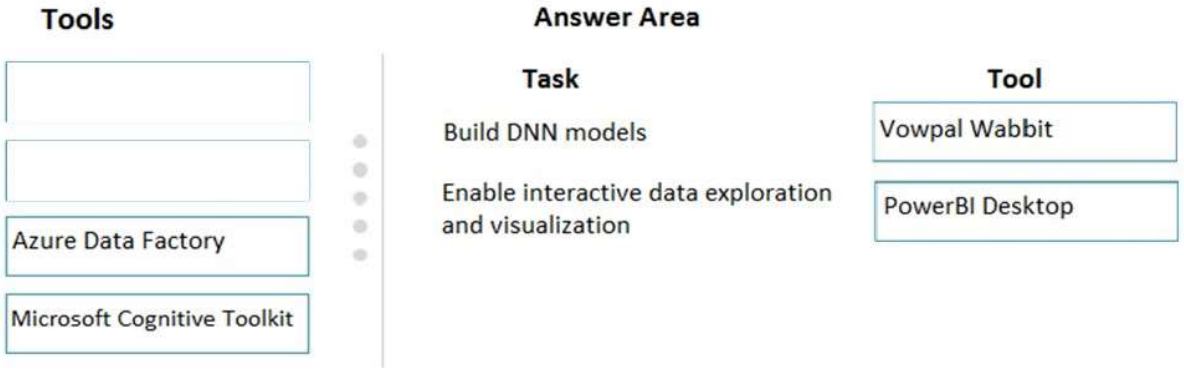
Which tools and frameworks should you recommend? To answer, drag the appropriate tools to the correct tasks. Each tool may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.

NOTE: Each correct selection is worth one point.

Select and Place:



Correct Answer:



Section: (none)
Explanation

Explanation/Reference:
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-model

https://docs.microsoft.com/en-us/azure/architecture/data-guide/scenarios/interactive-data-exploration

QUESTION 16
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

Correct Answer: AD
Section: (none)
Explanation

Explanation/Reference:
Explanation:
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

QUESTION 17
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:

```
hyperdrive = HyperDriveConfig(estimator=your_estimator,
 hyperparameter_sampling=your_params,
 policy=policy,
 primary_metric_name='AUC',
 primary_metric_goal=PrimaryMetricGoal.MAXIMIZE,
 max_total_runs=6,
 max_concurrent_runs=4)
```

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:

```
from sklearn.metrics import roc_auc_score
import logging
# code to train model omitted
auc = roc_auc_score(y_test, y_predicted)
logging.info("AUC: " + str(auc))
```

Does the solution meet the goal?

A. Yes
B. No

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:
Explanation:
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

QUESTION 18
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:

```
hyperdrive = HyperDriveConfig(estimator=your_estimator,
 hyperparameter_sampling=your_params,
 policy=policy,
 primary_metric_name='AUC',
 primary_metric_goal=PrimaryMetricGoal.MAXIMIZE,
 max_total_runs=6,
 max_concurrent_runs=4)
```

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:

```
import json, os
from sklearn.metrics import roc_auc_score
# code to train model omitted
auc = roc_auc_score(y_test, y_predicted)
os.makedirs("outputs", exist_ok = True)
with open("outputs/AUC.txt", "w") as file_cur:
    file_cur.write(auc)
```

Does the solution meet the goal?

A. Yes
B. No

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation
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

QUESTION 19
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:

```
hyperdrive = HyperDriveConfig(estimator=your_estimator,
 hyperparameter_sampling=your_params,
 policy=policy,
 primary_metric_name='AUC',
 primary_metric_goal=PrimaryMetricGoal.MAXIMIZE,
 max_total_runs=6,
 max_concurrent_runs=4)
```

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:

```
import numpy as np
from sklearn.metrics import roc_auc_score
# code to train model omitted
auc = roc_auc_score(y_test, y_predicted)
print(np.float(auc))
```

Does the solution meet the goal?

A. Yes
B. No

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation
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

QUESTION 20
HOTSPOT

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

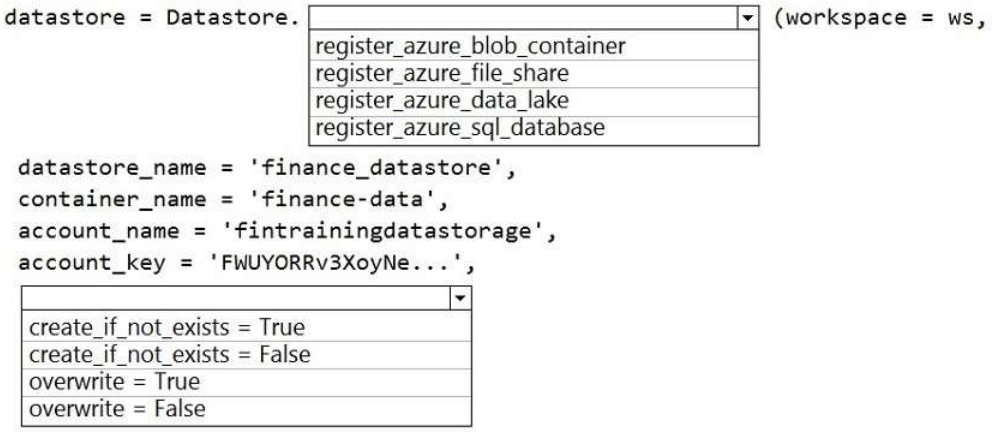
You need to register the container as a datastore in an Azure Machine Learning workspace and ensure that an error will be raised if the container does not exist.

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

NOTE: Each correct selection is worth one point.

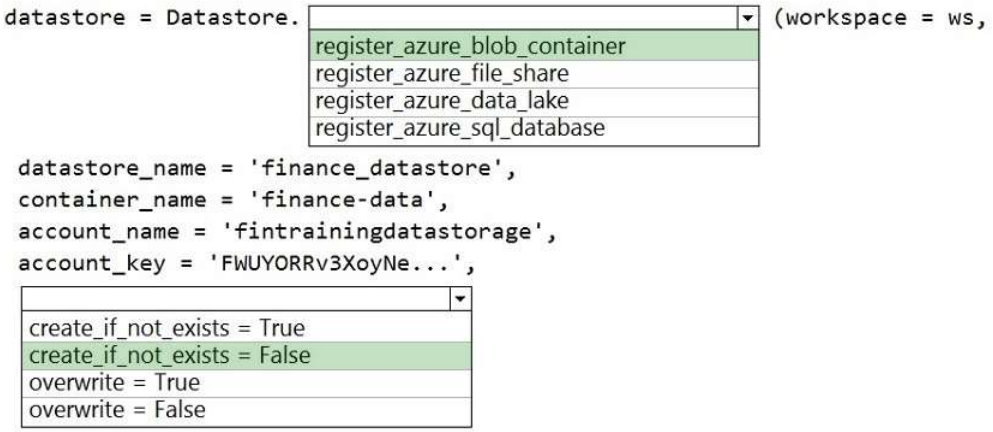
Hot Area:

Answer Area



Correct Answer:

Answer Area



Section: (none)
Explanation

Explanation/Reference:
Explanation:

Box 1: register_azure_blob_container
Register an Azure Blob Container to the datastore.

Box 2: create_if_not_exists = False
Create the file share if it does not exists, defaults to False.

Reference:
https://docs.microsoft.com/en-us/python/api/azureml-core/azureml.core.datastore.datastore

QUESTION 21
You plan to provision an Azure Machine Learning Basic edition workspace for a data science project.

You need to identify the tasks you will be able to perform in the workspace.

Which three tasks will you be able to perform? Each correct answer presents a complete solution.

NOTE: Each correct selection is worth one point.

A. Create a Compute Instance and use it to run code in Jupyter notebooks.
B. Create an Azure Kubernetes Service (AKS) inference cluster.
C. Use the designer to train a model by dragging and dropping pre-defined modules.
D. Create a tabular dataset that supports versioning.
E. Use the Automated Machine Learning user interface to train a model.

Correct Answer: ABD
Section: (none)
Explanation

Explanation/Reference:
Incorrect Answers:
C, E: The UI is included the Enterprise edition only.

Reference:
https://azure.microsoft.com/en-us/pricing/details/machine-learning/

QUESTION 22
HOTSPOT

A coworker registers a datastore in a Machine Learning services workspace by using the following code:

```
Datastore.register_azure_blob_container(workspace=ws,
 datastore_name='demo_datastore',
 container_name='demo_datacontainer',
 account_name='demo_account',
 account_key='0A0A0A-0A0A00A-0A00A0A0A0A0A',
 create_if_not_exists=True)
```

You need to write code to access the datastore from a notebook.

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.

Hot Area:

Answer Area

```
import azureml.core
from azureml.core import Workspace, Datastore
ws = Workspace.from_config()
datastore = [          ].get([     ], ' [          ] ')
```

Workspace	ws	demo_datastore
Datastore	run	demo_datacontainer
Experiment	experiment	demo_account
Run	log	Datastore

Correct Answer:

Answer Area

```
import azureml.core
from azureml.core import Workspace, Datastore
ws = Workspace.from_config()
datastore = [Workspace / Datastore / Experiment / Run].get([ws / run / experiment / log], ' [demo_datastore / demo_datacontainer / demo_account / Datastore] ')
```

Section: (none)
Explanation

Explanation/Reference:
Explanation:

Box 1: DataStore
To get a specific datastore registered in the current workspace, use the get() static method on the Datastore class:
# Get a named datastore from the current workspace
datastore = Datastore.get(ws, datastore_name='your datastore name')

Box 2: ws

Box 3: demo_datastore

Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-access-data

QUESTION 23
A set of CSV files contains sales records. All the CSV files have the same data schema.

Each CSV file contains the sales record for a particular month and has the filename sales.csv. Each file in stored in a folder that indicates the month and year when the data was recorded. The folders are in an Azure blob container for which a datastore has been defined in an Azure Machine Learning workspace. The folders are organized in a parent folder named sales to create the following hierarchical structure:

At the end of each month, a new folder with that month's sales file is added to the **sales** folder.

You plan to use the sales data to train a machine learning model based on the following requirements:

- You must define a dataset that loads all of the sales data to date into a structure that can be easily converted to a dataframe.
- You must be able to create experiments that use only data that was created before a specific previous

month, ignoring any data that was added after that month.
- You must register the minimum number of datasets possible.

You need to register the sales data as a dataset in Azure Machine Learning service workspace.

What should you do?

A. Create a tabular dataset that references the datastore and explicitly specifies each 'sales/*mm-yyyy*/sales.csv' file every month. Register the dataset with the name **sales_dataset** each month, replacing the existing dataset and specifying a tag named **month** indicating the month and year it was registered. Use this dataset for all experiments.
B. Create a tabular dataset that references the datastore and specifies the path 'sales/*/sales.csv', register the dataset with the name **sales_dataset** and a tag named **month** indicating the month and year it was registered, and use this dataset for all experiments.
C. Create a new tabular dataset that references the datastore and explicitly specifies each 'sales/*mm-yyyy*/sales.csv' file every month. Register the dataset with the name **sales_dataset_MM-YYYY** each month with appropriate MM and YYYY values for the month and year. Use the appropriate month-specific dataset for experiments.
D. Create a tabular dataset that references the datastore and explicitly specifies each 'sales/*mm-yyyy*/sales.csv' file. Register the dataset with the name **sales_dataset** each month as a new version and with a tag named **month** indicating the month and year it was registered. Use this dataset for all experiments, identifying the version to be used based on the **month** tag as necessary.

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Specify the path.

Example:
The following code gets the workspace existing workspace and the desired datastore by name. And then passes the datastore and file locations to the path parameter to create a new TabularDataset, weather_ds.

```
from azureml.core import Workspace, Datastore, Dataset

datastore_name = 'your datastore name'

# get existing workspace
workspace = Workspace.from_config()

# retrieve an existing datastore in the workspace by name
datastore = Datastore.get(workspace, datastore_name)

# create a TabularDataset from 3 file paths in datastore
datastore_paths = [(datastore, 'weather/2018/11.csv'),
                   (datastore, 'weather/2018/12.csv'),
                   (datastore, 'weather/2019/*.csv')]

weather_ds = Dataset.Tabular.from_delimited_files(path=datastore_paths)
```

QUESTION 24
DRAG DROP

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.

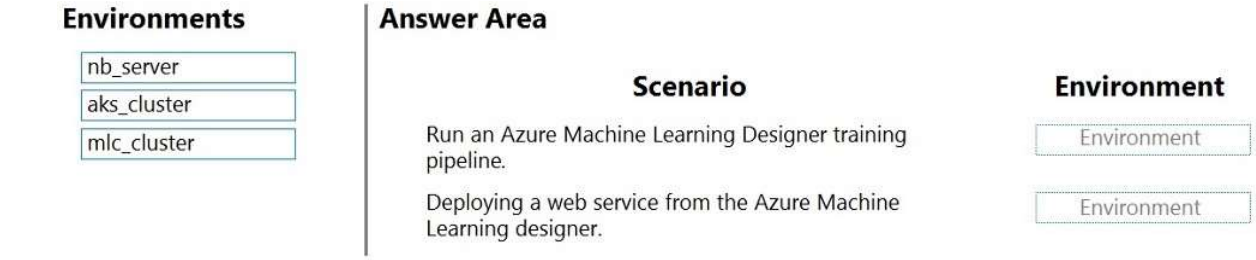
Environment name	Compute type
nb_server	Compute Instance
aks_cluster	Azure Kubernetes Service
mlc_cluster	Machine Learning Compute

You need to determine which compute environment to use for the following scenarios.

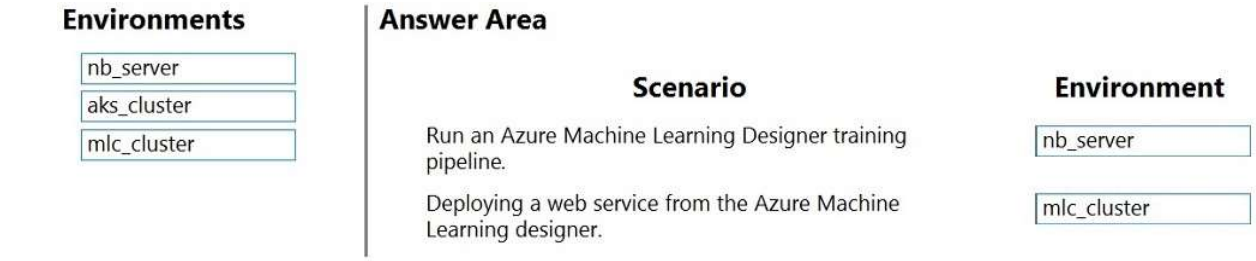
Which compute types should you use? To answer, drag the appropriate compute environments to the correct scenarios. Each compute environment may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.

NOTE: Each correct selection is worth one point.

Select and Place:



Correct Answer:



Section: (none)
Explanation

Explanation/Reference:
Explanation:

Box 1: nb_server

Training targets	Automated ML	ML pipelines	Azure Machine Learning designer
Local computer	yes		
Azure Machine Learning compute cluster	yes & hyperparameter tuning	yes	yes
Azure Machine Learning compute instance	yes & hyperparameter tuning	yes	yes
Remote VM	yes & hyperparameter tuning	yes	
Azure Databricks	yes (SDK local mode only)	yes	
Azure Data Lake Analytics		yes	
Azure HDInsight		yes	
Azure Batch		yes	

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

QUESTION 25
HOTSPOT

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:

```
from azureml.core.compute import ComputeTarget, AmlCompute
from azureml.core.compute_target import ComputeTargetException
the_cluster_name = "ComputeOne"
try:
  the_cluster = ComputeTarget(workspace=ws, name=the_cluster_name)
  print('Step1')
except ComputeTargetException:
  config = AmlCompute.provisioning_configuration(vm_size='STANDARD_DS12_v2', max_nodes=4)
  the_cluster = ComputeTarget.create(ws, the_cluster_name, config)
  print('Step2')
```

For each of the following statements, select Yes if the statement is true. Otherwise, select No.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

	Yes	No
A new machine learning compute resource is created with a virtual machine size of STANDARD_DS12_v2 and a maximum of four nodes.	○	○
Any experiments configured to use `the_cluster` will run on ComputeOne.	○	○
The text **Step1** will be printed to the screen.	○	○

Correct Answer:

Answer Area

	Yes	No
A new machine learning compute resource is created with a virtual machine size of STANDARD_DS12_v2 and a maximum of four nodes.	○	○
Any experiments configured to use `the_cluster` will run on ComputeOne.	○	○
The text **Step1** will be printed to the screen.	○	○

Section: (none)
Explanation

Explanation/Reference:
Explanation:

Box 1:Yes
ComputeTargetException class: An exception related to failures when creating, interacting with, or configuring a compute target. This exception is commonly raised for failures attaching a compute target, missing headers, and unsupported configuration values.

Create(workspace, name, provisioning_configuration)
Provision a Compute object by specifying a compute type and related configuration.

This method creates a new compute target rather than attaching an existing one.

Box 2: Yes

Box 3: No
The line before print('Step1') will fail.

Reference:

https://docs.microsoft.com/en-us/python/api/azureml-core/azureml.core.compute.computetarget

QUESTION 26
HOTSPOT

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.

Hot Area:

Answer Area

Virtual machine size

Search by name...

Name ↑	vCPUs	GPUs	RAM	Resource disk
BASIC_A0	1		0.75 GB	20 GB
STANDARD_D3_V2	4		14 GB	200 GB
STANDARD_E64_V3	64		432 GB	1,600 GB
STANDARD_M64LS	64		512 GB	2,000 GB
STANDARD_NC12	12	2	112 GB	680 GB
STANDARD_NC24	24	4	224 GB	1,440 GB

Correct Answer:

Answer Area

Virtual machine size

Search by name...

Name ↑	vCPUs	GPUs	RAM	Resource disk
BASIC_A0	1		0.75 GB	20 GB
STANDARD_D3_V2	4		14 GB	200 GB
STANDARD_E64_V3	64		432 GB	1,600 GB
STANDARD_M64LS	64		512 GB	2,000 GB
STANDARD_NC12	12	2	112 GB	680 GB
STANDARD_NC24	24	4	224 GB	1,440 GB

Section: (none)
Explanation

Explanation/Reference:
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

QUESTION 27
You use the following code to run a script as an experiment in Azure Machine Learning:

```
from azureml.core import Workspace, Experiment, Run
from azureml.core import RunConfig, ScriptRunConfig
ws = Workspace.from_config()
run_config = RunConfiguration()
run_config.target='local'
script_config = ScriptRunConfig(source_directory='./script', script='experiment.py', run_config=run_config)
experiment = Experiment(workspace=ws, name='script experiment')
run = experiment.submit(config=script_config)
run.wait_for_completion()
```

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()`

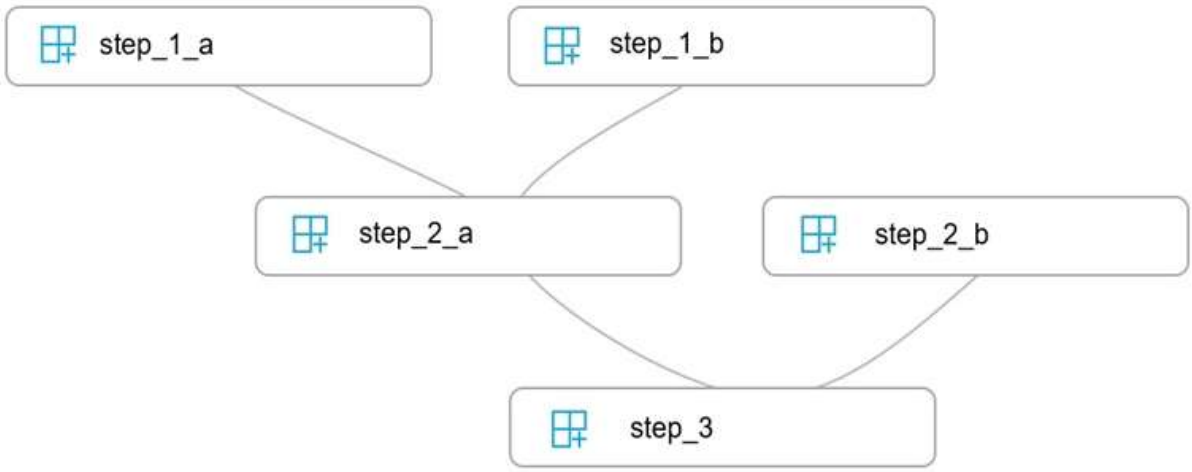
Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:
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

QUESTION 28
You write five Python scripts that must be processed in the order specified in Exhibit A – which allows the same modules to run in parallel, but will wait for modules with dependencies.

You must create an Azure Machine Learning pipeline using the Python SDK, because you want to script to create the pipeline to be tracked in your version control system. You have created five PythonScriptSteps and have named the variables to match the module names.



You need to create the pipeline shown. Assume all relevant imports have been done.

Which Python code segment should you use?

A.
```
p = Pipeline(ws, steps=[[[[step_1_a, step_1_b], step_2_a],  step_2_b], step_3])
```

B.
```
pipeline_steps = {
    "Pipeline": {
        "run": step_3,
        "run_after": {[
            {"run": step_2_a,
                "run_after":
                    [{"run": step_1_a},
                     {"run": step_1_b}]
                },
            {"run": step_2_b}]
        }
    }
}
p = Pipeline(ws, steps=pipeline_steps)
```

C.
```
step_2_a.run_after(step_1_b)
step_2_a.run_after(step_1_a)
step_3.run_after(step_2_b)
step_3.run_after(step_2_a)
p = Pipeline(ws, steps=[step_3])
```

D.
```
p = Pipeline(ws, steps=[step_1_a, step_1_b, step_2_a, step_2_b, step_3])
```

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:
Explanation:
The steps parameter is an array of steps. To build pipelines that have multiple steps, place the steps in order in this array.

Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-use-parallel-run-step

QUESTION 29
HOTSPOT

You are preparing to use the Azure ML SDK to run an experiment and need to create compute. You run the following code:

```
from azureml.core.compute import ComputeTarget, AmlCompute
from azureml.core.compute_target import ComputeTargetException
ws = Workspace.from_config()
cluster_name = 'aml-cluster'
try:
  training_compute = ComputeTarget(workspace=ws, name=cluster_name)
except ComputeTargetException:
  compute_config = AmlCompute.provisioning_configuration(vm_size='STANDARD_D2_V2', vm_priority='lowpriority',
max_nodes=4)
  training_compute = ComputeTarget.create(ws, cluster_name, compute_config)
  training_compute.wait_for_completion(show_output=True)
```

For each of the following statements, select Yes if the statement is true. Otherwise, select No.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

	Yes	No
If a training cluster named aml-cluster already exists in the workspace, it will be deleted and replaced.	○	○
The `wait_for_completion()` method will not return until the aml-cluster compute has four active nodes.	○	○
If the code creates a new aml-cluster compute target, it may be preempted due to capacity constraints.	○	○
The aml-cluster compute target is deleted from the workspace after the training experiment completes.	○	○

Correct Answer:

Answer Area

	Yes	No
If a training cluster named aml-cluster already exists in the workspace, it will be deleted and replaced.	○	○
The `wait_for_completion()` method will not return until the aml-cluster compute has four active nodes.	○	○
If the code creates a new aml-cluster compute target, it may be preempted due to capacity constraints.	○	○
The aml-cluster compute target is deleted from the workspace after the training experiment completes.	○	○

Section: (none)
Explanation

Explanation/Reference:
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

QUESTION 30
You create a datastore named **training_data** that references a blob container in an Azure Storage account. The blob container contains a folder named **csv_files** in which multiple comma-separated values (CSV) files are stored.

You have a script named train.py in a local folder named ./script that you plan to run as an experiment using an estimator. The script includes the following code to read data from the csv_files folder:

```
import os
import argparse
import pandas as pd

from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from azureml.core import Run

run = Run.get_context()
parser = argparse.ArgumentParser()
parser.add_argument('--data-folder', type=str, dest='data_folder', help='data reference')
args = parser.parse_args()

data_folder = args.data_folder
csv_files = os.listdir(data_folder)
training_data = pd.concat((pd.read_csv(os.path.join(data_folder,csv_file)) for csv_file in csv_files))

# Code goes on to split the training data and train a logistic regression model
```

You have the following script.

```
from azureml.core import Workspace, Datastore, Experiment
from azureml.train.sklearn import SKLearn

ws = Workspace.from_config()
exp = Experiment(workspace=ws, name='csv_training')
ds = Datastore.get(ws, datastore_name='training_data')
data_ref = ds.path('csv_files')

# Code to define estimator goes here

run = exp.submit(config=estimator)
run.wait_for_completion(show_output=True)
```

You need to configure the estimator for the experiment so that the script can read the data from a data reference named data_ref that references the csv_files folder in the training_data datastore.

Which code should you use to configure the estimator?

A.
```
estimator = SKLearn(source_directory='./script',
 inputs=[data_ref.as_named_input('data-folder').to_pandas_dataframe()],
 compute_target='local',
 entry_script='train.py')
```

B.
```
script_params = {
    '--data-folder': data_ref.as_mount()
}
estimator = SKLearn(source_directory='./script',
 script_params=script_params,
 compute_target='local',
 entry_script='train.py'
```

C.
```
estimator = SKLearn(source_directory='./script',
 inputs=[data_ref.as_named_input('data-folder').as_mount()],
 compute_target='local',
 entry_script='train.py')
```

D.
```
script_params = {
    '--data-folder': data_ref.as_download(path_on_compute='csv_files')
}
estimator = SKLearn(source_directory='./script',
 script_params=script_params,
 compute_target='local',
 entry_script='train.py'
```

E.
```
estimator = SKLearn(source_directory='./script',
 inputs=[data_ref.as_named_input('data-folder').as_download(path_on_compute='cs
 compute_target='local',
 entry_script='train.py')
```

Correct Answer: B
Section: (none)

Explanation

Explanation/Reference:
Explanation:
Besides passing the dataset through the inputs parameter in the estimator, you can also pass the dataset through script_params and get the data path (mounting point) in your training script via arguments. This way, you can keep your training script independent of azureml-sdk. In other words, you will be able use the same training script for local debugging and remote training on any cloud platform.

Example:

```
from azureml.train.sklearn import SKLearn

script_params = {
    # mount the dataset on the remote compute and pass the mounted path as an argument to the training script
    '--data-folder': mnist_ds.as_named_input('mnist').as_mount(),
    '--regularization': 0.5
}

est = SKLearn(source_directory=script_folder,
              script_params=script_params,
              compute_target=compute_target,
              environment_definition=env,
              entry_script='train_mnist.py')

# Run the experiment
run = experiment.submit(est)
run.wait_for_completion(show_output=True)
```

Incorrect Answers:
A: Pandas DataFrame not used.

Reference:
https://docs.microsoft.com/es-es/azure/machine-learning/how-to-train-with-datasets

QUESTION 31
DRAG DROP

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.

Select and Place:

Actions	Answer Area
Configure a DataTransferStep() to fetch new image data from public web portal, running on the cpu-compute compute target.	
Configure an EstimatorStep() to run an estimator that runs the bird_classifier_train.py model training script on the gpu_compute compute target.	
Configure a PythonScriptStep() to run both image_fetcher.py and image_resize.py on the cpu-compute compute target.	
Configure an EstimatorStep() to run an estimator that runs the bird_classifier_train.py model training script on the cpu_compute compute target.	
Configure a PythonScriptStep() to run image_fetcher.py on the cpu-compute compute target.	
Configure a PythonScriptStep() to run image_resize.py on the cpu-compute compute target.	
Configure a PythonScriptStep() to run bird_classifier_train.py on the cpu-compute compute target.	
Configure a PythonScriptStep() to run bird_classifier_train.py on the gpu-compute compute target.	

Correct Answer:

Actions	Answer Area
Configure a DataTransferStep() to fetch new image data from public web portal, running on the cpu-compute compute target.	Configure a DataTransferStep() to fetch new image data from public web portal, running on the cpu-compute compute target.
Configure an EstimatorStep() to run an estimator that runs the bird_classifier_train.py model training script on the gpu_compute compute target.	Configure a PythonScriptStep() to run image_resize.py on the cpu-compute compute target.
Configure a PythonScriptStep() to run both image_fetcher.py and image_resize.py on the cpu-compute compute target.	Configure an EstimatorStep() to run an estimator that runs the bird_classifier_train.py model training script on the gpu_compute compute target.
Configure an EstimatorStep() to run an estimator that runs the bird_classifier_train.py model training script on the cpu_compute compute target.	
Configure a PythonScriptStep() to run image_fetcher.py on the cpu-compute compute target.	
Configure a PythonScriptStep() to run image_resize.py on the cpu-compute compute target.	
Configure a PythonScriptStep() to run bird_classifier_train.py on the cpu-compute compute target.	
Configure a PythonScriptStep() to run bird_classifier_train.py on the gpu-compute compute target.	

Section: (none)
Explanation

Explanation/Reference:
Explanation:

Step 1: Configure a DataTransferStep() to fetch new image data…

Step 2: Configure a PythonScriptStep() to run image_resize.y on the cpu-compute compute target.

Step 3: Configure the EstimatorStep() to run training script on the gpu_compute computer target.

The PyTorch estimator provides a simple way of launching a PyTorch training job on a compute target.

Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-train-pytorch

QUESTION 32
HOTSPOT

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.

Hot Area:

Answer Area

Context	Resource
Test	ds-workstation notebook VM aks-compute cluster cpu-compute cluster gpu-compute cluster
Production	ds-workstation notebook VM aks-compute cluster cpu-compute cluster gpu-compute cluster

Correct Answer:

Answer Area

Context	Resource
Test	**ds-workstation notebook VM** aks-compute cluster cpu-compute cluster gpu-compute cluster
Production	ds-workstation notebook VM aks-compute cluster cpu-compute cluster **gpu-compute cluster**

Section: (none)
Explanation

Explanation/Reference:
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

QUESTION 33
You create a deep learning model for image recognition on Azure Machine Learning service using GPU-based training.

You must deploy the model to a context that allows for real-time GPU-based inferencing.

You need to configure compute resources for model inferencing.

Which compute type should you use?

A. Azure Container Instance
B. Azure Kubernetes Service
C. Field Programmable Gate Array
D. Machine Learning Compute

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:
You can use Azure Machine Learning to deploy a GPU-enabled model as a web service. Deploying a model on Azure Kubernetes Service (AKS) is one option. The AKS cluster provides a GPU resource that is used by the model for inference.

Inference, or model scoring, is the phase where the deployed model is used to make predictions. Using GPUs instead of CPUs offers performance advantages on highly parallelizable computation.

Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-deploy-inferencing-gpus

QUESTION 34
You create a batch inference pipeline by using the Azure ML SDK. You run the pipeline by using the following code:

```
from azureml.pipeline.core import Pipeline
from azureml.core.experiment import Experiment

pipeline = Pipeline(workspace=ws, steps=[parallelrun_step])
pipeline_run = Experiment(ws, 'batch_pipeline').submit(pipeline)
```

You need to monitor the progress of the pipeline execution.

What are two possible ways to achieve this goal? Each correct answer presents a complete solution.

NOTE: Each correct selection is worth one point.

A. Run the following code in a notebook:

```
from azureml.contrib.interpret.explanation.explanation_client import ExplanationClient
client = ExplanationClient.from_run(pipeline_run)
explanation = client.download_model_explanation()
explanation = client.download_model_explanation(top_k=4)
global_importance_values = explanation.get_ranked_global_values()
global_importance_names = explanation.get_ranked_global_names()
print('global importance values: {}'.format(global_importance_values))
print('global importance names: {}'.format(global_importance_names))
```

B. Use the Inference Clusters tab in Machine Learning Studio.

C. Use the Activity log in the Azure portal for the Machine Learning workspace.

D. Run the following code in a notebook:

```
from azureml.widgets import RunDetails
RunDetails(pipeline_run).show()
```

E. Run the following code and monitor the console output from the PipelineRun object:

```
pipeline_run.wait_for_completion(show_output=True)
```

Correct Answer: DE
Section: (none)
Explanation

Explanation/Reference:
Explanation:
A batch inference job can take a long time to finish. This example monitors progress by using a Jupyter widget. You can also manage the job's progress by using:

- Azure Machine Learning Studio.
- Console output from the PipelineRun object.

```
from azureml.widgets import RunDetails
RunDetails(pipeline_run).show()

pipeline_run.wait_for_completion(show_output=True)
```

Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-use-parallel-run-step#monitor-the-parallel-run-job

QUESTION 35
You train and register a model in your Azure Machine Learning workspace.

You must publish a pipeline that enables client applications to use the model for batch inferencing. You must use a pipeline with a single ParallelRunStep step that runs a Python inferencing script to get predictions from the input data.

You need to create the inferencing script for the ParallelRunStep pipeline step.

Which two functions should you include? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

A. `run(mini_batch)`

B. `main()`

C. `batch()`

D. `init()`
E. `score(mini_batch)`

Correct Answer: AD
Section: (none)
Explanation

Explanation/Reference:
Reference:
https://github.com/Azure/MachineLearningNotebooks/tree/master/how-to-use-azureml/machine-learning-pipelines/parallel-run

QUESTION 36
You deploy a model as an Azure Machine Learning real-time web service using the following code.

```
# ws, model, interference_config, and deployment_config defined previously
service = Model.deploy(ws, 'classification-service', [model], inference_config, deployment_config)
service.wait_for_deployment(True)
```

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()`

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:
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

QUESTION 37
HOTSPOT

You deploy a model in Azure Container Instance.

You must use the Azure Machine Learning SDK to call the model API.

You need to invoke the deployed model using native SDK classes and methods.

How should you complete the command? To answer, select the appropriate options in the answer areas.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

```
from azureml.core import Workspace
```

[▼]
- from azureml.core.webservice import requests
- from azureml.core.webservice import Webservice
- from azureml.core.webservice import LocalWebservice

```
import json
ws = Workspace.from_config()
service_name = "mlmodel1-service"
service = Webservice(name=service_name, workspace=ws)
x_new = [[2,101.5,1,24,21], [1,89.7,4,41,21]]
input_json = json.dumps({"data": x_new})
```

[▼]
- predictions = service.run(input_json)
- predictions = requests.post(service.scoring_uri, input_json)
- predictions = service.deserialize(ws, input_json)

Correct Answer:

Answer Area

```
from azureml.core import Workspace
```

[▼]
- from azureml.core.webservice import requests
- from azureml.core.webservice import Webservice
- from azureml.core.webservice import LocalWebservice

```
import json
ws = Workspace.from_config()
service_name = "mlmodel1-service"
service = Webservice(name=service_name, workspace=ws)
x_new = [[2,101.5,1,24,21], [1,89.7,4,41,21]]
input_json = json.dumps({"data": x_new})
```

[▼]
- predictions = service.run(input_json)
- predictions = requests.post(service.scoring_uri, input_json)
- predictions = service.deserialize(ws, input_json)

Section: (none)
Explanation

Explanation/Reference:
Explanation:

Box 1: from azureml.core.webservice import Webservice
The following code shows how to use the SDK to update the model, environment, and entry script for a web service to Azure Container Instances:

```
from azureml.core import Environment
from azureml.core.webservice import Webservice
from azureml.core.model import Model, InferenceConfig
```

Box 2: predictions = service.run(input_json)

Example: The following code demonstrates sending data to the service:

```
import json

test_sample = json.dumps({'data': [
    [1, 2, 3, 4, 5, 6, 7, 8, 9, 10],
    [10, 9, 8, 7, 6, 5, 4, 3, 2, 1]
]})

test_sample = bytes(test_sample, encoding='utf8')

prediction = service.run(input_data=test_sample)
print(prediction)
```

Reference:
https://docs.microsoft.com/bs-latn-ba/azure/machine-learning/how-to-deploy-azure-container-instance

https://docs.microsoft.com/en-us/azure/machine-learning/how-to-troubleshoot-deployment

QUESTION 38
You create a multi-class image classification deep learning model.

You train the model by using PyTorch version 1.2.

You need to ensure that the correct version of PyTorch can be identified for the inferencing environment when the model is deployed.

What should you do?

A. Save the model locally as a**.pt** file, and deploy the model as a local web service.
B. Deploy the model on computer that is configured to use the default Azure Machine Learning conda environment.
C. Register the model with a **.pt** file extension and the default **version** property.
D. Register the model, specifying the **model_framework** and **model_framework_version** properties.

Correct Answer: D
Section: (none)
Explanation

Explanation/Reference:
Explanation:
framework_version: The PyTorch version to be used for executing training code.

Reference:
https://docs.microsoft.com/en-us/python/api/azureml-train-core/azureml.train.dnn.pytorch?view=azure-ml-py

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:

		Actual	
		1	**0**
Predicted	**0**	1	2
	1	2	1

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

		Actual	
		1	**0**
Predicted	**0**	1	5
	1	5	1

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

QUESTION 1
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

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:
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

QUESTION 2
HOTSPOT

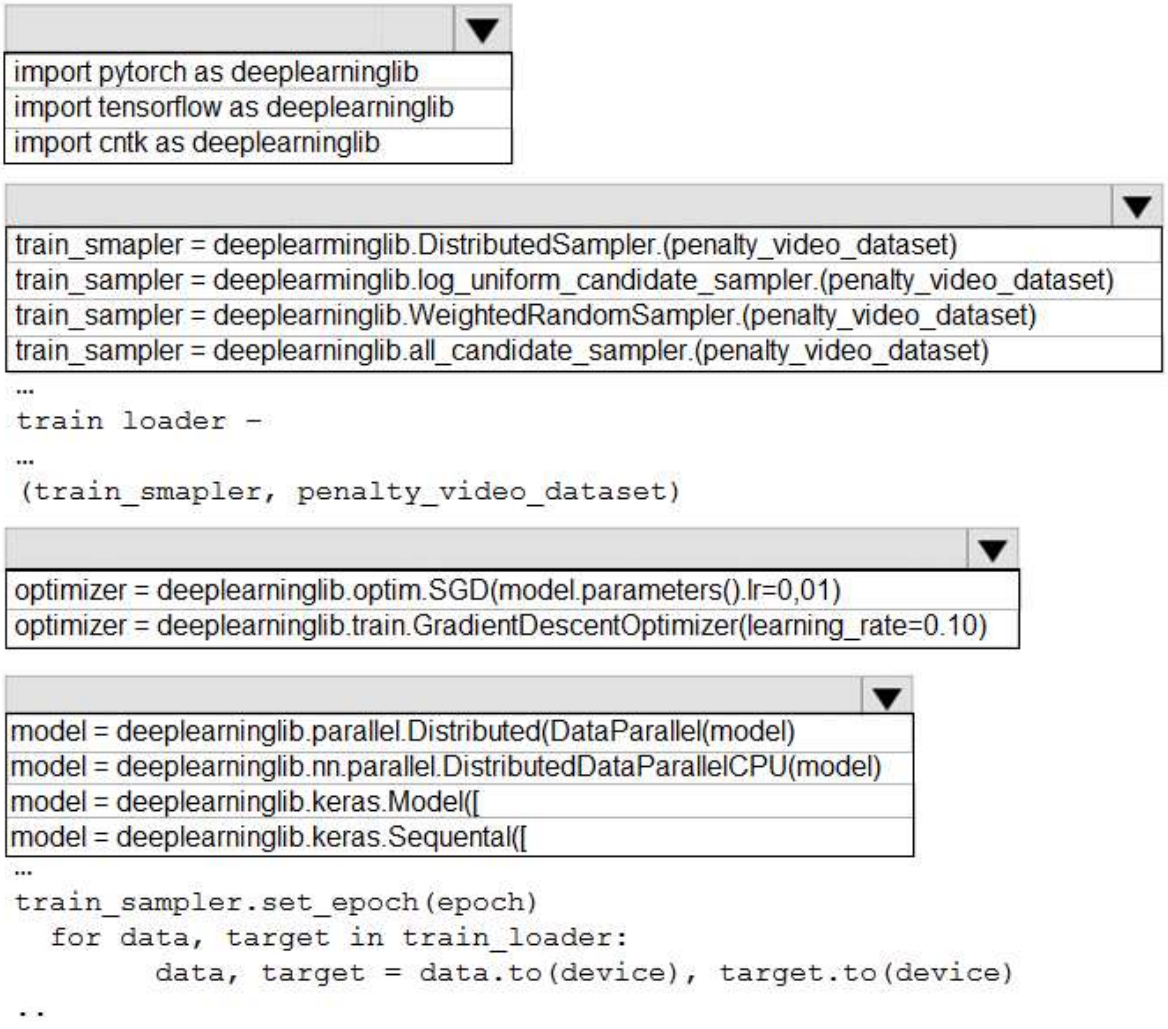
You need to use the Python language to build a sampling strategy for the global penalty detection models.

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.

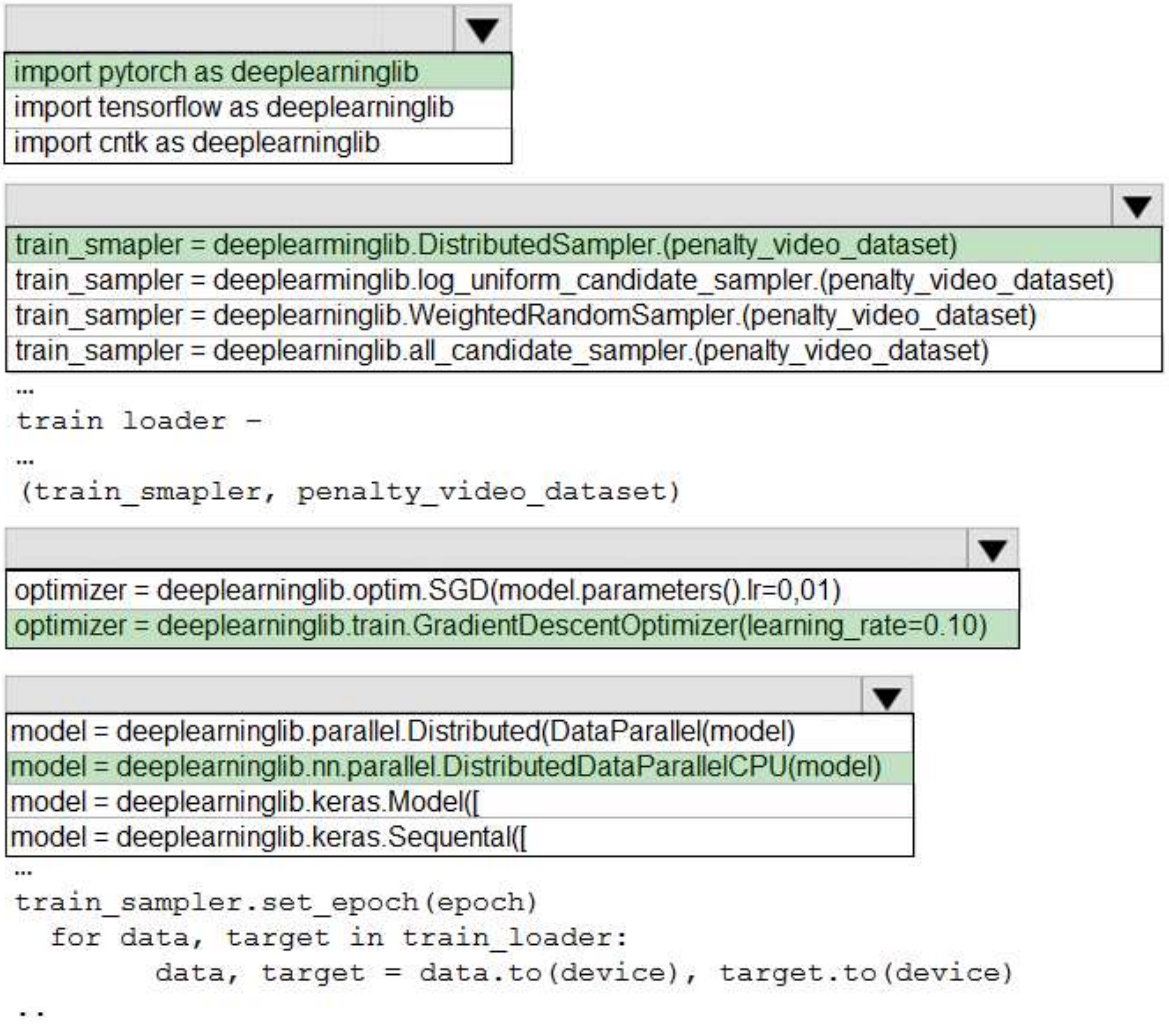
Hot Area:

Answer Area



Correct Answer:

Answer Area



Section: (none)
Explanation

Explanation/Reference:
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)

Incorrect Answers: ..SGD..

Scenario: All penalty detection models show inference phases using a Stochastic Gradient Descent (SGD) are running too slow.

Box 4: .. nn.parallel.DistributedDataParallel..
DistributedSampler(Sampler): The sampler that restricts data loading to a subset of the dataset.
    It is especially useful in conjunction with :class:`torch.nn.parallel.DistributedDataParallel`.

References:
https://github.com/pytorch/pytorch/blob/master/torch/utils/data/distributed.py

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 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. 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:

Column heading	Description
CapitaCrimeRate	per capita crime rate by town
Zoned	proportion of residential land zoned for lots over 25.000 square feet
NonRetailAcres	proportion of retail business acres per town
NextToRiver	proximity of a property to the river
NitrogenOxideConcentration	nitric oxides concentration (parts per 10 million)
AvgRoomsPerHouse	average number of rooms per dwelling
Age	proportion of owner-occupied units built prior to 1940
DistanceToEmploymentCenter	weighted distances to employment centers
AccessibilityToHighway	index of accessibility to radial highways to a value of two decimal places
Tax	full value property tax rate per $10,000
PupilTeacherRatio	pupil to teacher ratio by town
ProfessionalClass	professional class percentage
LowerStatus	percentage lower status of the population
MedianValue	median value of owner-occupied homes in $1000s

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 use a feature selection algorithm to analyze the relationship between the MedianValue 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 Machine Learning Studio to compare the Two-Class Decision Forest and the Two-Class Decision Jungle modules with one another.

QUESTION 1
HOTSPOT

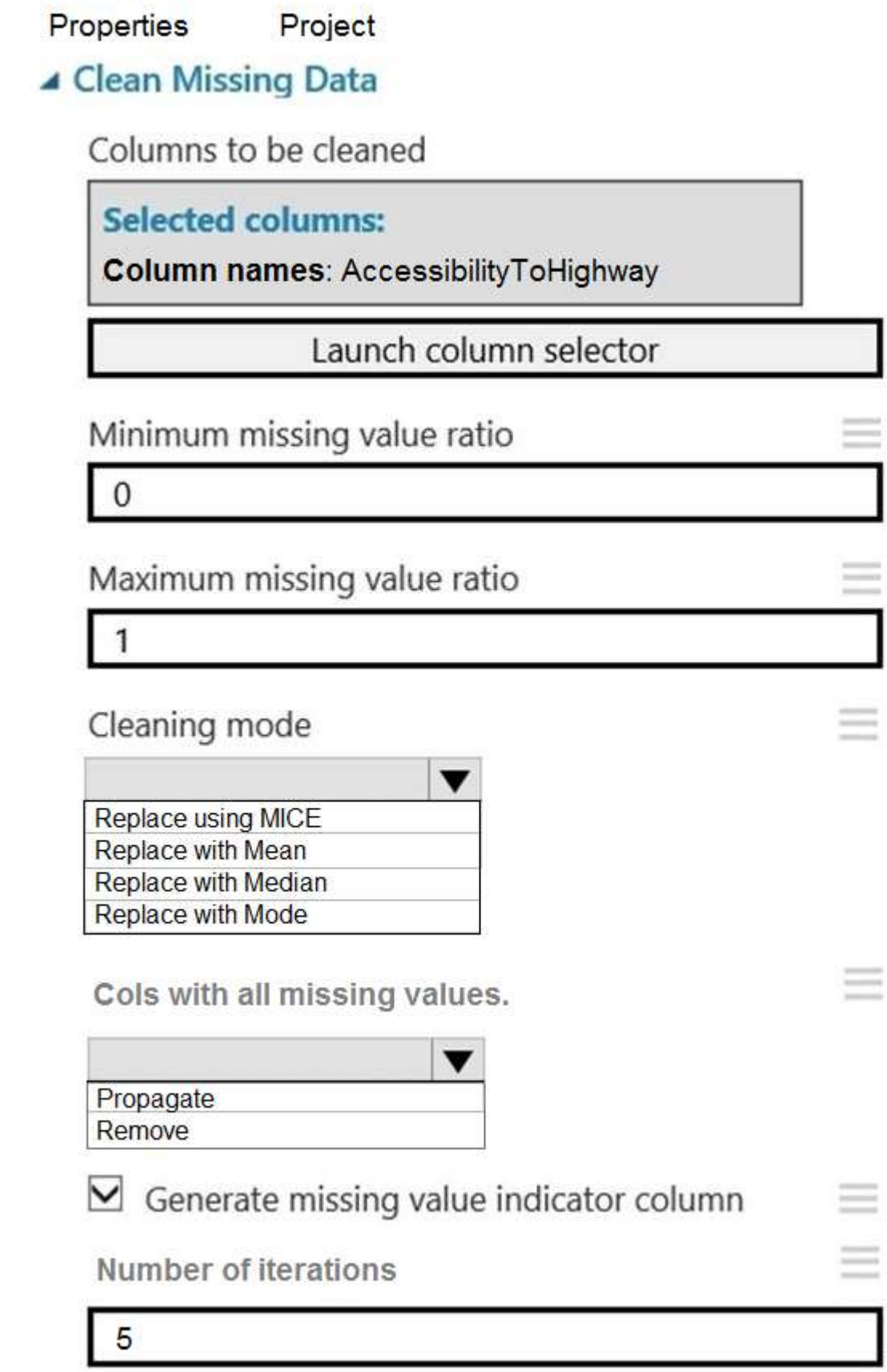
You need to replace the missing data in the AccessibilityToHighway columns.

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.

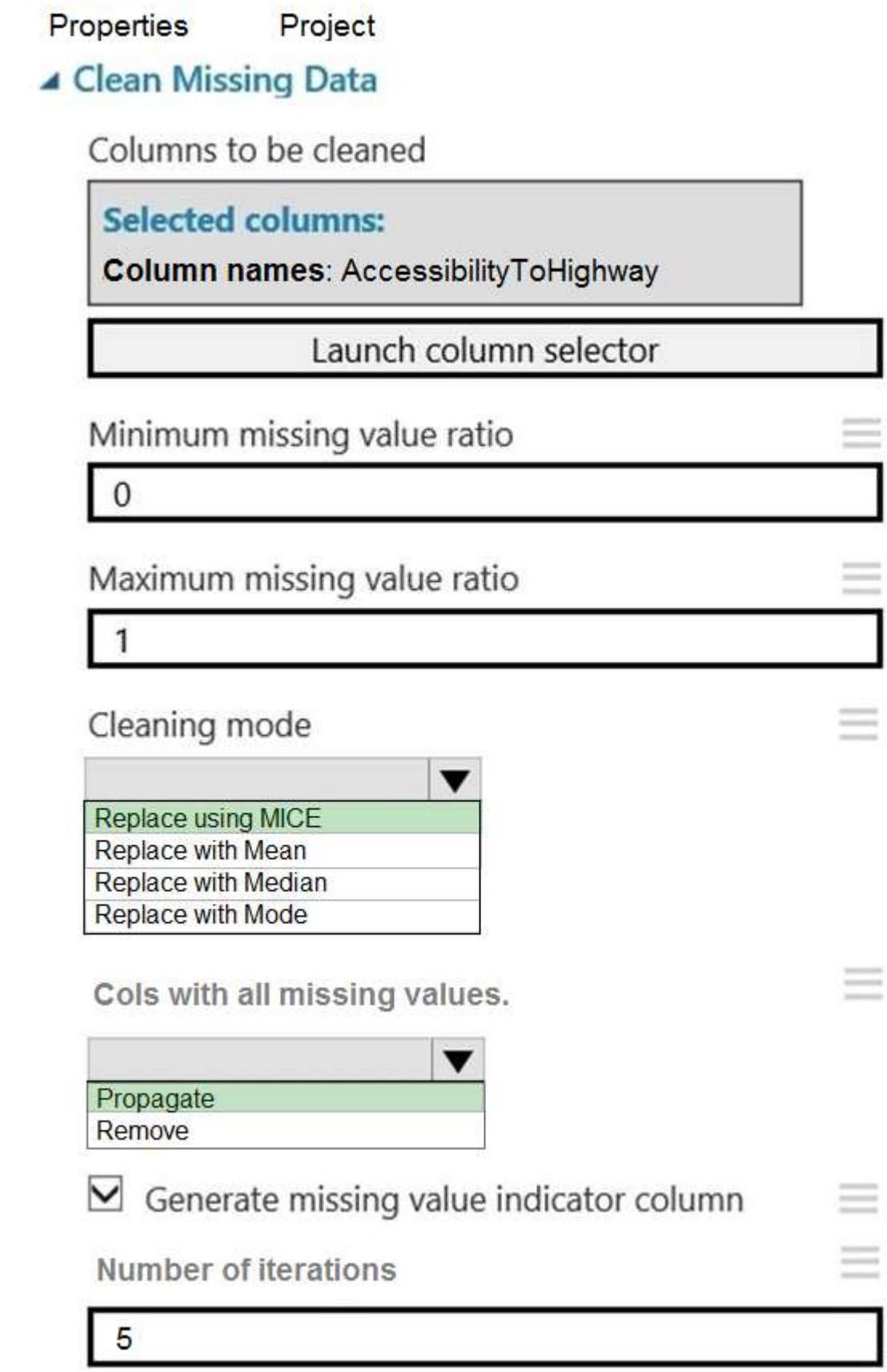
Hot Area:

Answer Area


Properties
Project
Clean Missing Data
Columns to be cleaned
Selected columns:
Column names: AccessibilityToHighway
Launch column selector
Minimum missing value ratio
0
Maximum missing value ratio
1
Cleaning mode
Replace using MICE
Replace with Mean
Replace with Median
Replace with Mode
Cols with all missing values.
Propagate
Remove
Generate missing value indicator column
Number of iterations
5

Correct Answer:

Answer Area


Properties
Project
Clean Missing Data
Columns to be cleaned
Selected columns:
Column names: AccessibilityToHighway
Launch column selector
Minimum missing value ratio
0
Maximum missing value ratio
1
Cleaning mode
Replace using MICE
Replace with Mean
Replace with Median
Replace with Mode
Cols with all missing values.
Propagate
Remove
Generate missing value indicator column
Number of iterations
5

Section: (none)
Explanation

Explanation/Reference:
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.

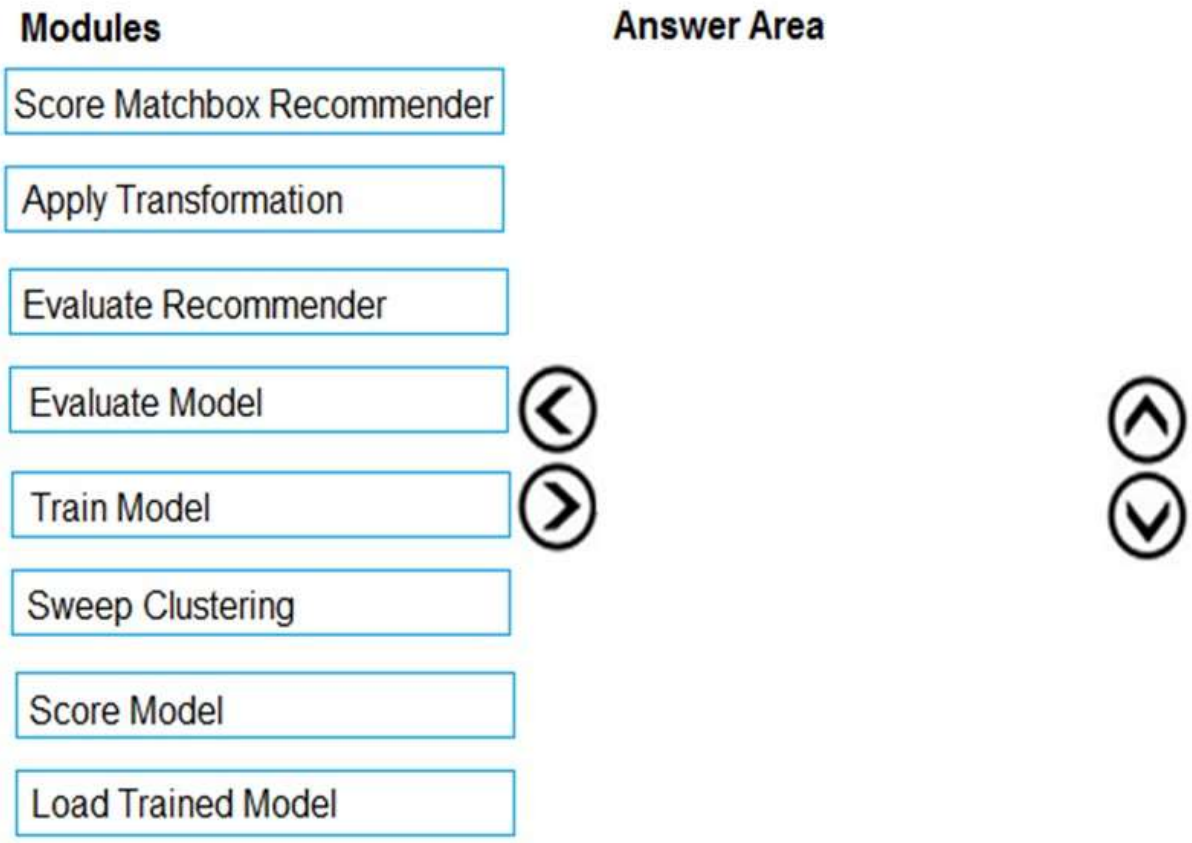
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/clean-missing-data

QUESTION 2
DRAG DROP

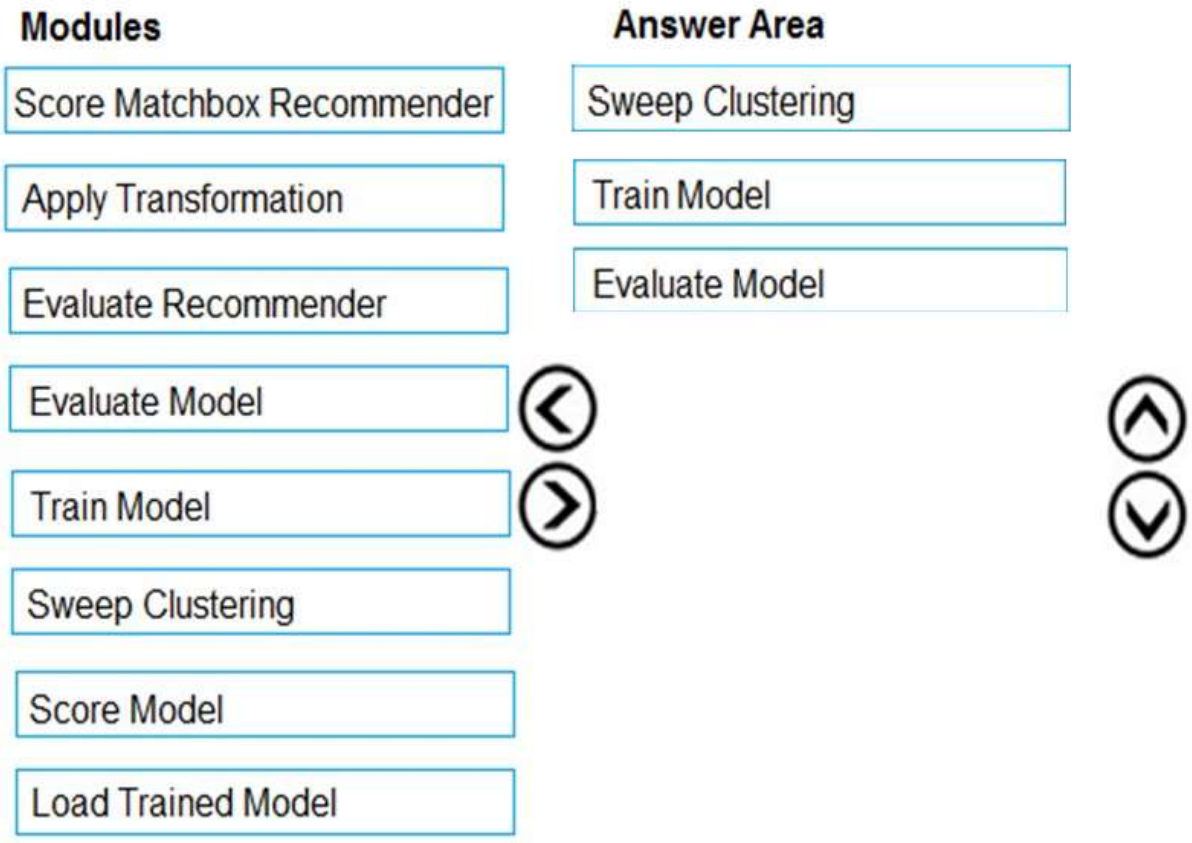
You need to produce a visualization for the diagnostic test evaluation according to the data visualization requirements.

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.

Select and Place:



Correct Answer:



Section: (none)
Explanation

Explanation/Reference:
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

QUESTION 3
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? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

A. Create Scatterplot
B. Summarize Data
C. Clip Values
D. Replace Discrete Values
E. Build Counting Transform

Correct Answer: ABC
Section: (none)
Explanation

Explanation/Reference:
Explanation:
B: To have a global view, the summarize data module can be used. Add the module and connect it to the data set that needs to be visualized.
A: One way to quickly identify Outliers visually is to create scatter plots.

C: The easiest way to treat the outliers in Azure ML is to use the Clip Values module. It can identify and optionally replace data values that are above or below a specified threshold.

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.

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

QUESTION 4
HOTSPOT

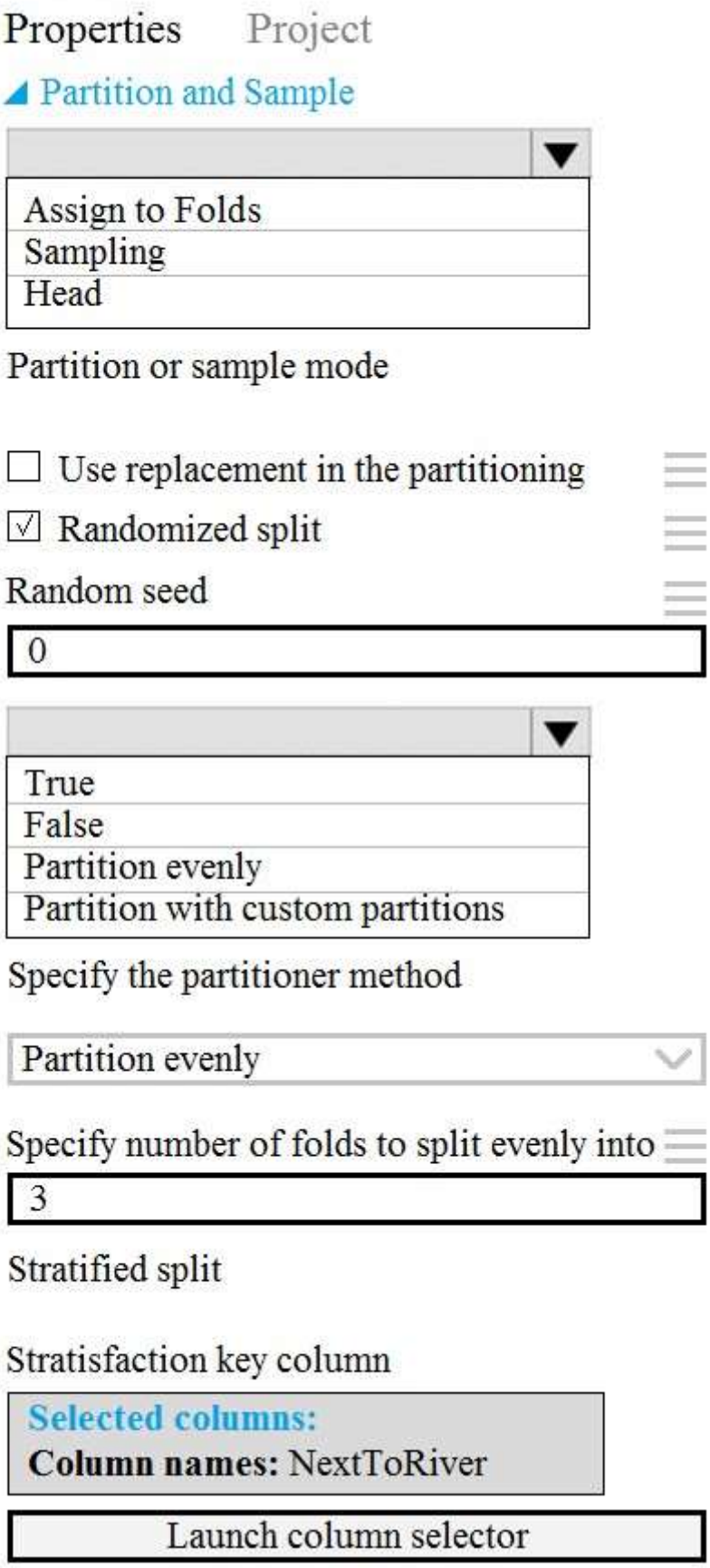
You need to identify the methods for dividing the data according to the testing requirements.

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

NOTE: Each correct selection is worth one point.

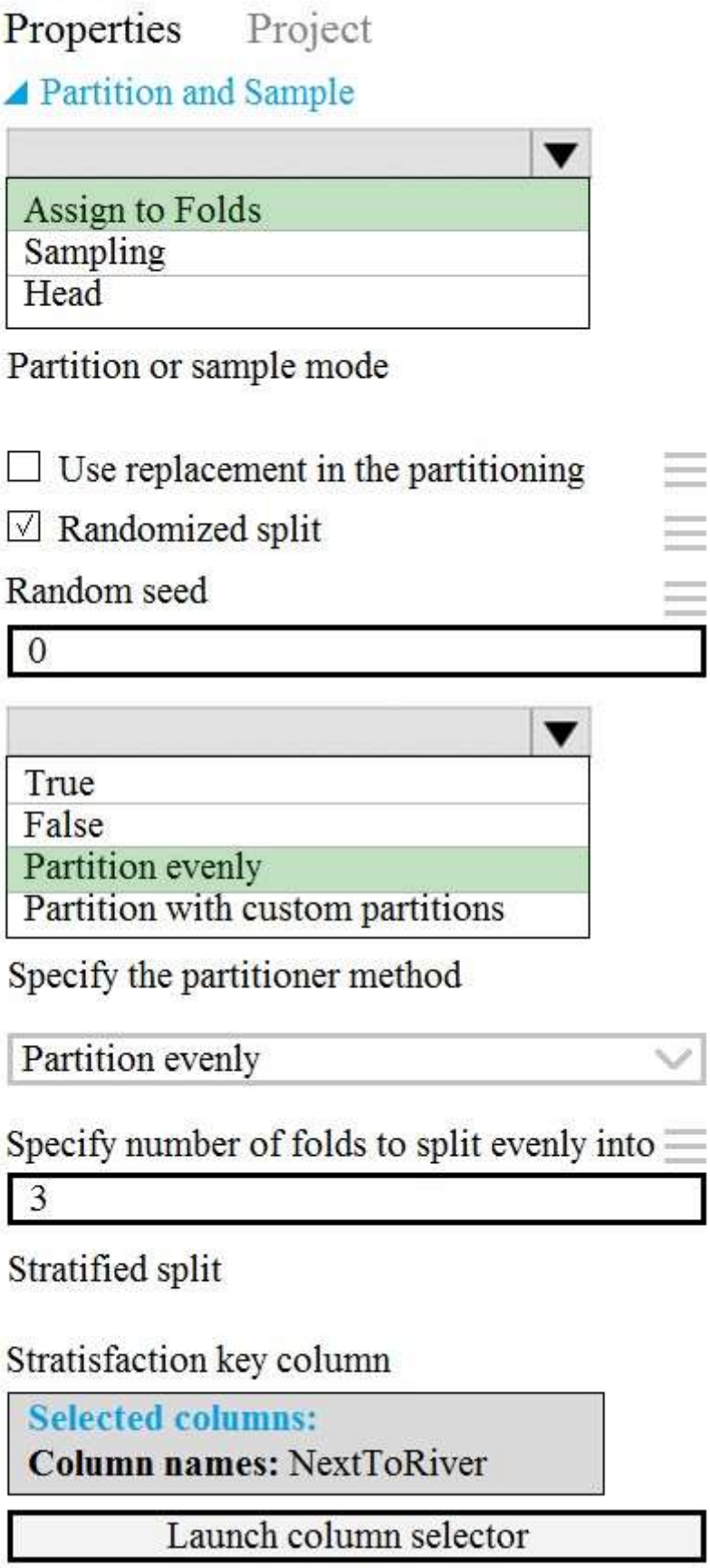
Hot Area:

Answer Area


Properties
Project
Partition and Sample
Assign to Folds
Sampling
Head
Partition or sample mode
Use replacement in the partitioning
Randomized split
Random seed
0
True
False
Partition evenly
Partition with custom partitions
Specify the partitioner method
Partition evenly
Specify number of folds to split evenly into
3
Stratified split
Stratisfaction key column
Selected columns:
Column names: NextToRiver
Launch column selector

Correct Answer:

Answer Area


Properties
Project
Partition and Sample
Assign to Folds
Sampling
Head
Partition or sample mode
Use replacement in the partitioning
Randomized split
Random seed
0
True
False
Partition evenly
Partition with custom partitions
Specify the partitioner method
Partition evenly
Specify number of folds to split evenly into
3
Stratified split
Stratisfaction key column
Selected columns:
Column names: NextToRiver
Launch column selector

Section: (none)
Explanation

Explanation/Reference:
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

QUESTION 5
HOTSPOT

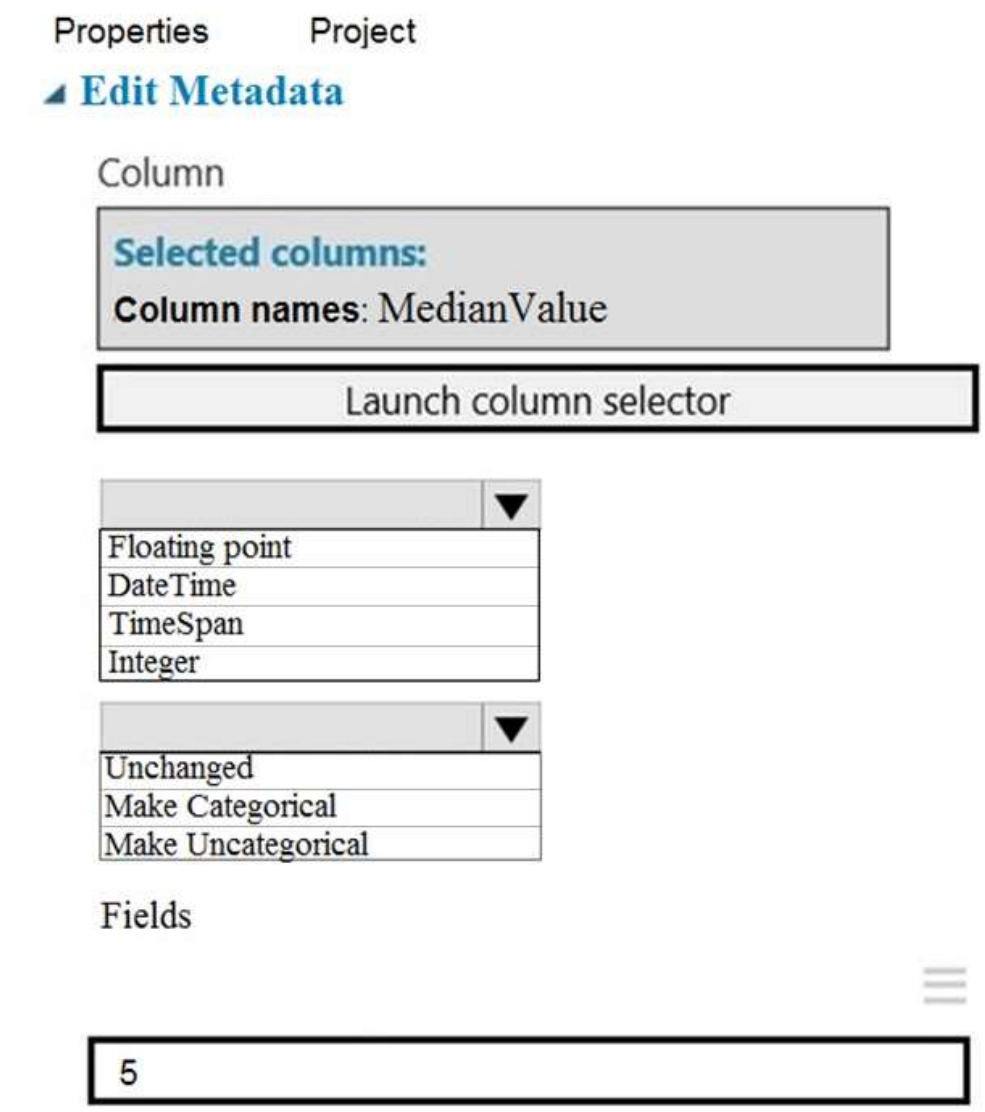
You need to configure the Edit Metadata module so that the structure of the datasets match.

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

NOTE: Each correct selection is worth one point.

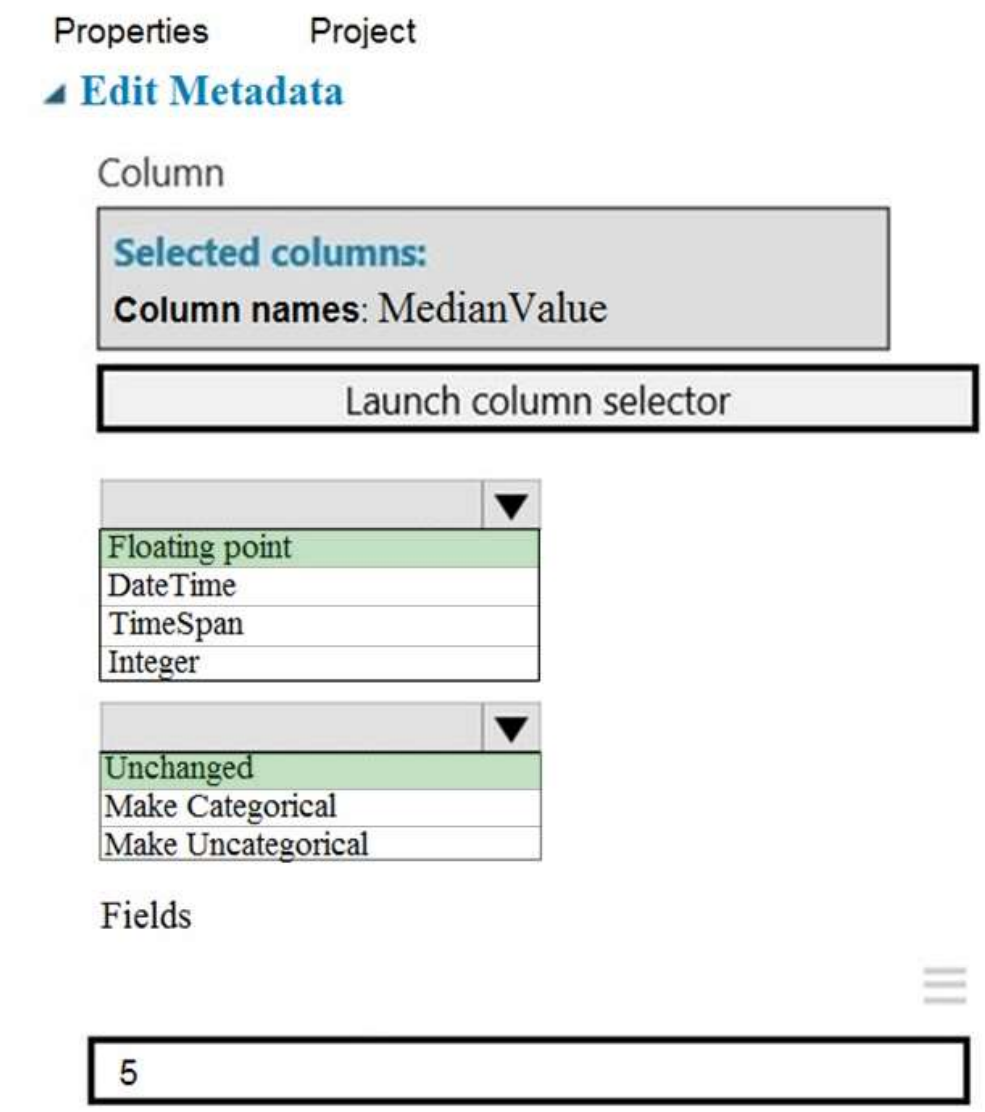
Hot Area:

Answer Area



Correct Answer:

Answer Area



Section: (none)
Explanation

Explanation/Reference:
Explanation:

Box 1: Floating point
Need floating point for Median values.

Scenario: 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.

Box 2: Unchanged

Note: Select the Categorical option to specify that the values in the selected columns should be treated as categories.

For example, you might have a column that contains the numbers 0,1 and 2, but know that the numbers actually mean "Smoker", "Non smoker" and "Unknown". In that case, by flagging the column as categorical you can ensure that the values are not used in numeric calculations, only to group data.

Prepare data for modeling

Question Set 3

QUESTION 1
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

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Use the Multiple Imputation by Chained Equations (MICE) method.

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

QUESTION 2
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

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
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

QUESTION 3
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

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
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

QUESTION 4
HOTSPOT

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).

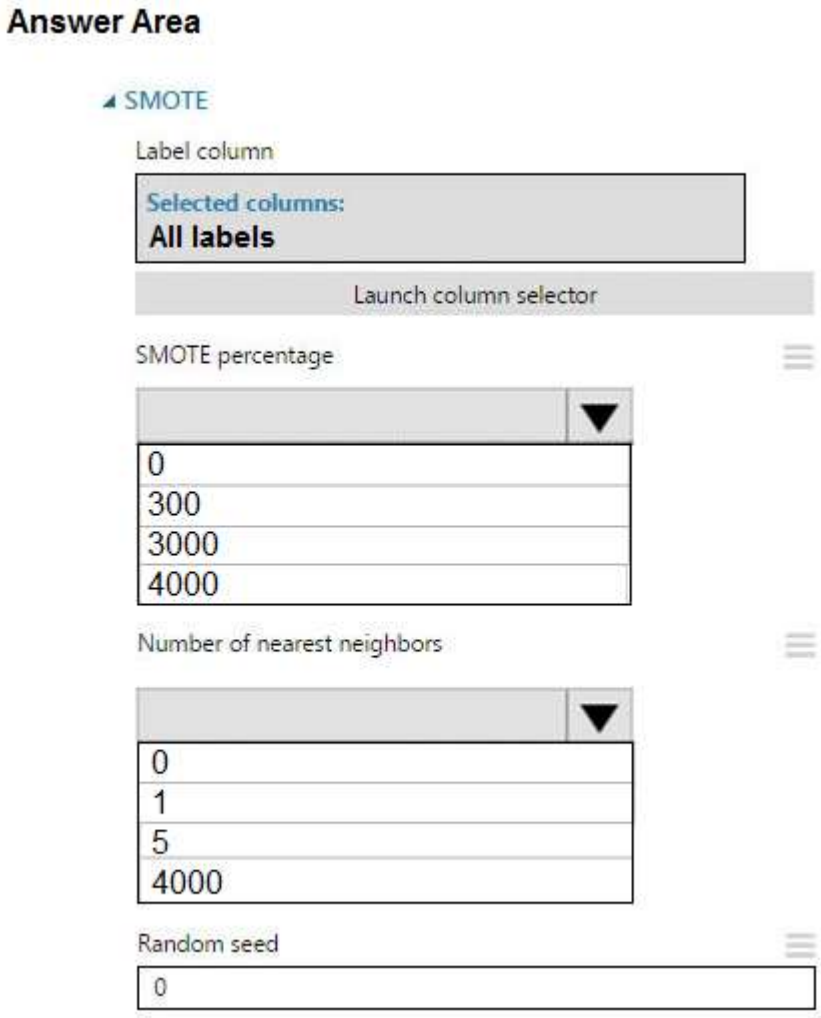
The training set is imbalances between two classes. You must increase the number of training examples for class 1 to 4,000 by using 5 data rows. You add the Synthetic Minority Oversampling Technique (SMOTE) module to the experiment.

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.

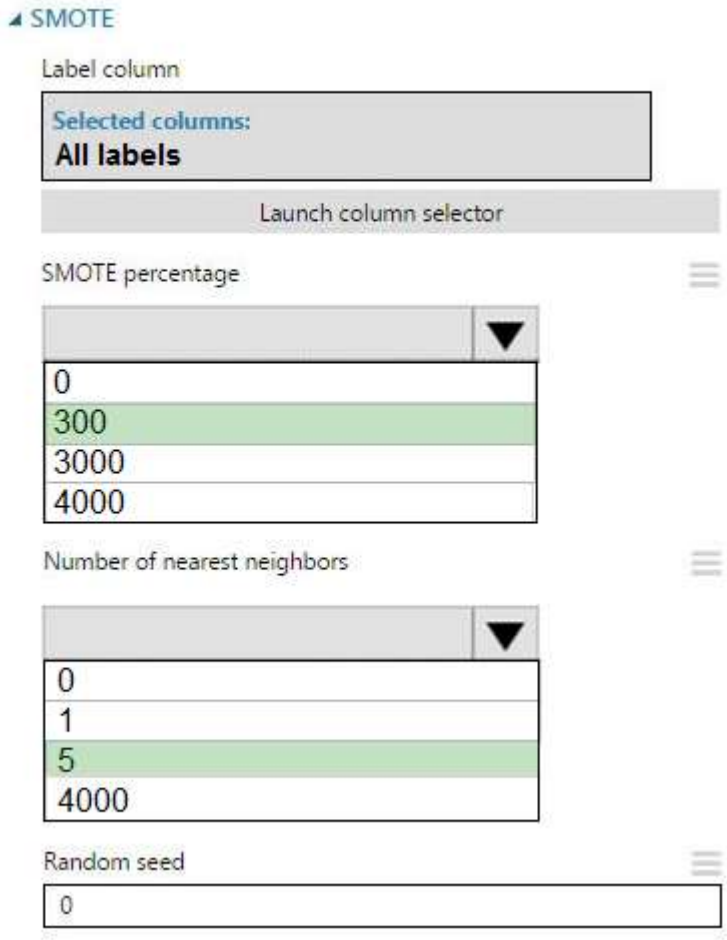
NOTE: Each correct selection is worth one point.

Hot Area:



Correct Answer:

Answer Area



Section: (none)
Explanation

Explanation/Reference:
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

QUESTION 5
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

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:
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.

QUESTION 6
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. Load Trained Model
C. Partition and Sample
D. Tune Model-Hyperparameters

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Partition and Sample with the Stratified split option outputs multiple datasets, partitioned using the rules you specified.

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

QUESTION 7
DRAG DROP

You are creating an experiment by using Azure Machine Learning Studio.

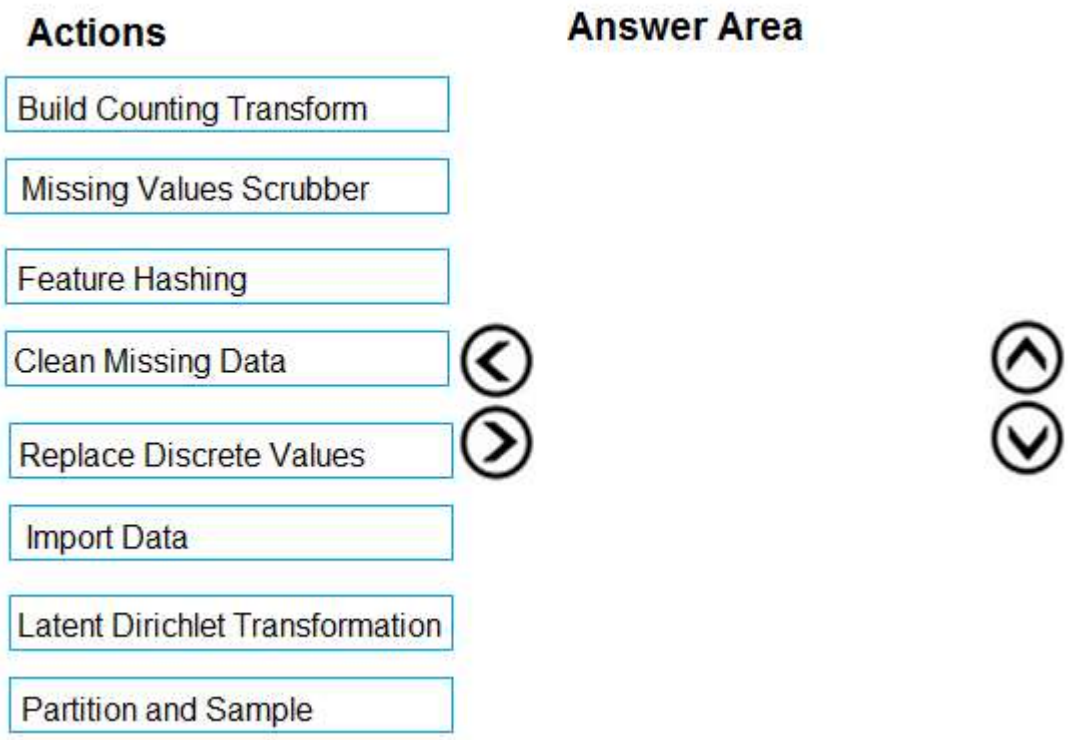
You must divide the data into four subsets for evaluation. There is a high degree of missing values in the data. You must prepare the data for analysis.

You need to select appropriate methods for producing the experiment.

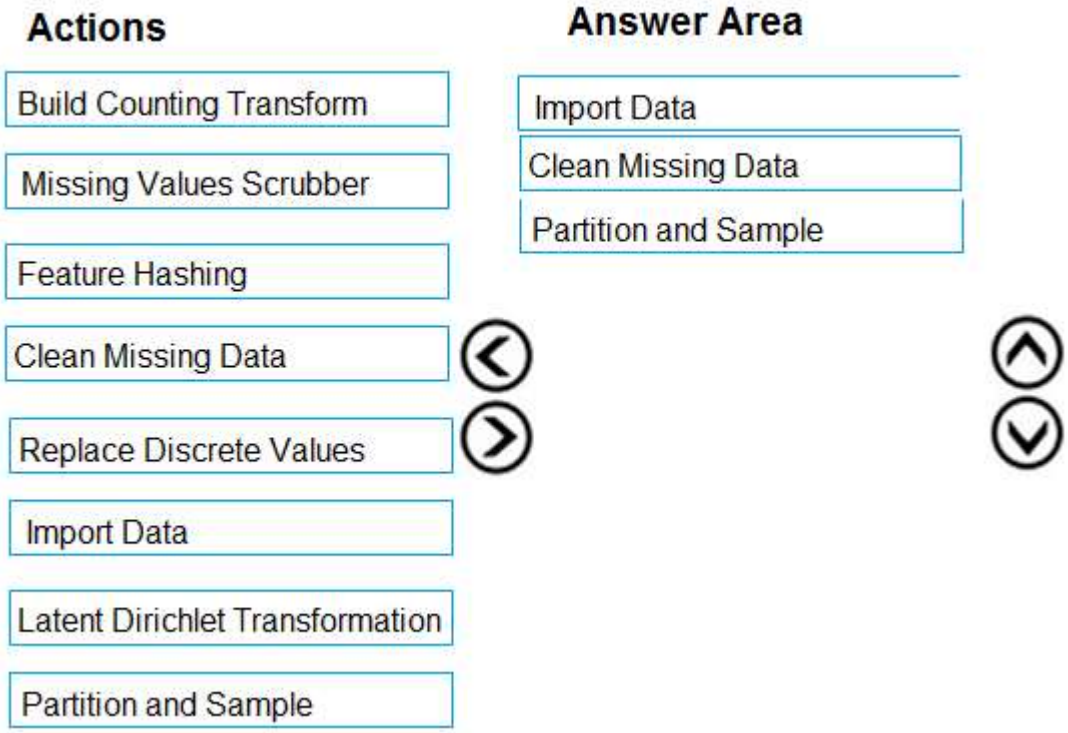
Which three modules should you run in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.

NOTE: More than one order of answer choices is correct. You will receive credit for any of the correct orders you select.

Select and Place:



Correct Answer:



Section: (none)
Explanation

Explanation/Reference:
Explanation:

The Clean Missing Data module in Azure Machine Learning Studio, to remove, replace, or infer missing values.

Incorrect Answers:

- Latent Direchlet Transformation: Latent Dirichlet Allocation module in Azure Machine Learning Studio, to group otherwise unclassified text into a number of categories. Latent Dirichlet Allocation (LDA) is often used in natural language processing (NLP) to find texts that are similar. Another common term is topic modeling.
- Build Counting Transform: Build Counting Transform module in Azure Machine Learning Studio, to analyze training data. From this data, the module builds a count table as well as a set of count-based features that can be used in a predictive model.
- Missing Value Scrubber: The Missing Values Scrubber module is deprecated.
- Feature hashing: Feature hashing is used for linguistics, and works by converting unique tokens into integers.
- Replace discrete values: the Replace Discrete Values module in Azure Machine Learning Studio is used to generate a probability score that can be used to represent a discrete value. This score can be useful for understanding the information value of the discrete values.

References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/clean-missing-data

QUESTION 8
HOTSPOT

You are retrieving data from a large datastore by using Azure Machine Learning Studio.

You must create a subset of the data for testing purposes using a random sampling seed based on the system clock.

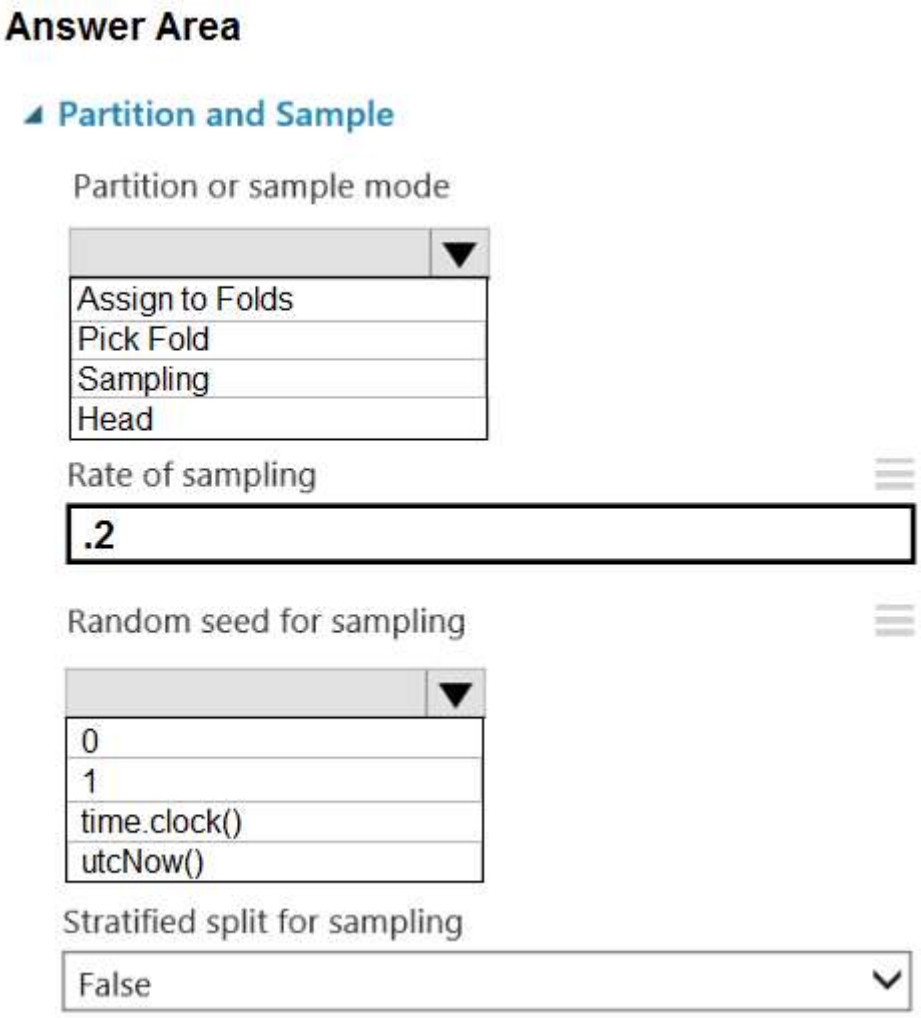
You add the Partition and Sample module to your experiment.

You need to select the properties for the module.

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

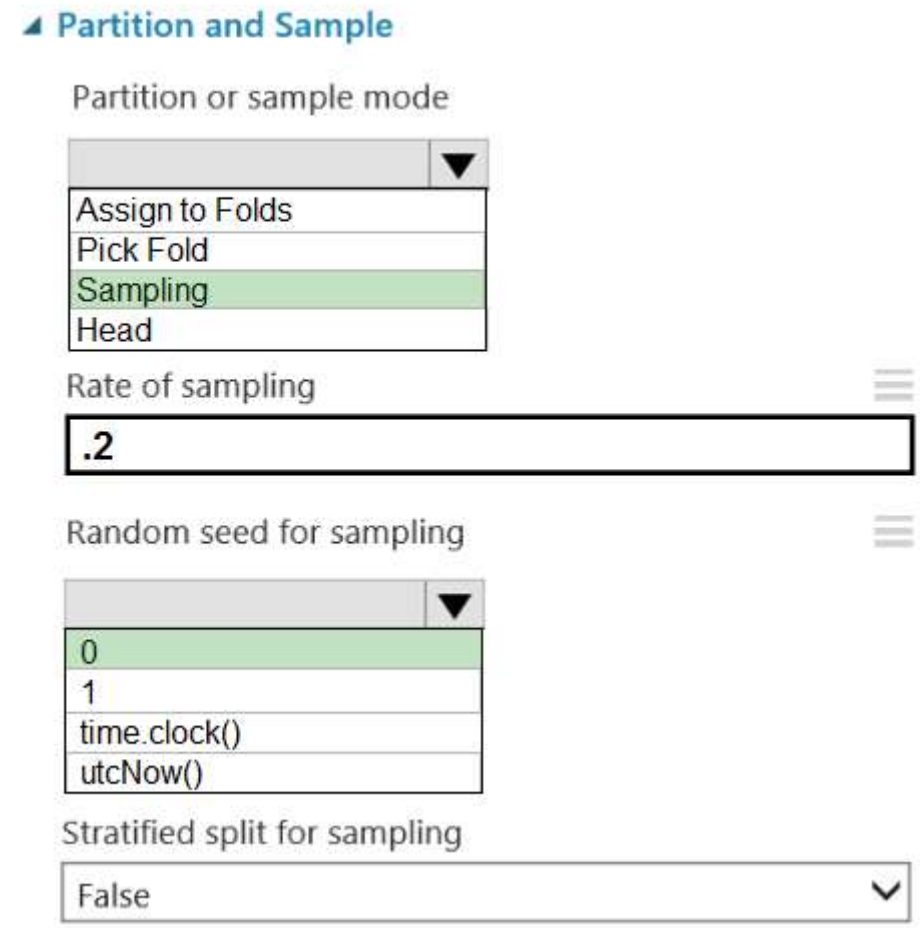
NOTE: Each correct selection is worth one point.

Hot Area:



Correct Answer:

Answer Area



Section: (none)
Explanation

Explanation/Reference:
Explanation:

Box 1: Sampling
Create a sample of data
This option supports simple random sampling or stratified random sampling. This is useful if you want to create a smaller representative sample dataset for testing.

1. Add the Partition and Sample module to your experiment in Studio, and connect the dataset.
2. Partition or sample mode: Set this to Sampling.
3. Rate of sampling. See box 2 below.

Box 2: 0
3. Rate of sampling. Random seed for sampling: Optionally, type an integer to use as a seed value.

This option is important if you want the rows to be divided the same way every time. The default value is 0, meaning that a starting seed is generated based on the system clock. This can lead to slightly different results each time you run the experiment.

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

QUESTION 9
You are creating a machine learning model. You have a dataset that contains null rows.

You need to use the Clean Missing Data module in Azure Machine Learning Studio to identify and resolve the null and missing data in the dataset.

Which parameter should you use?

A. Replace with mean
B. Remove entire column
C. Remove entire row
D. Hot Deck
E. Custom substitution value
F. Replace with mode

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Remove entire row: Completely removes any row in the dataset that has one or more missing values. This is useful if the missing value can be considered randomly missing.

References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/clean-missing-data

QUESTION 10
DRAG DROP

You are analyzing a raw dataset that requires cleaning.

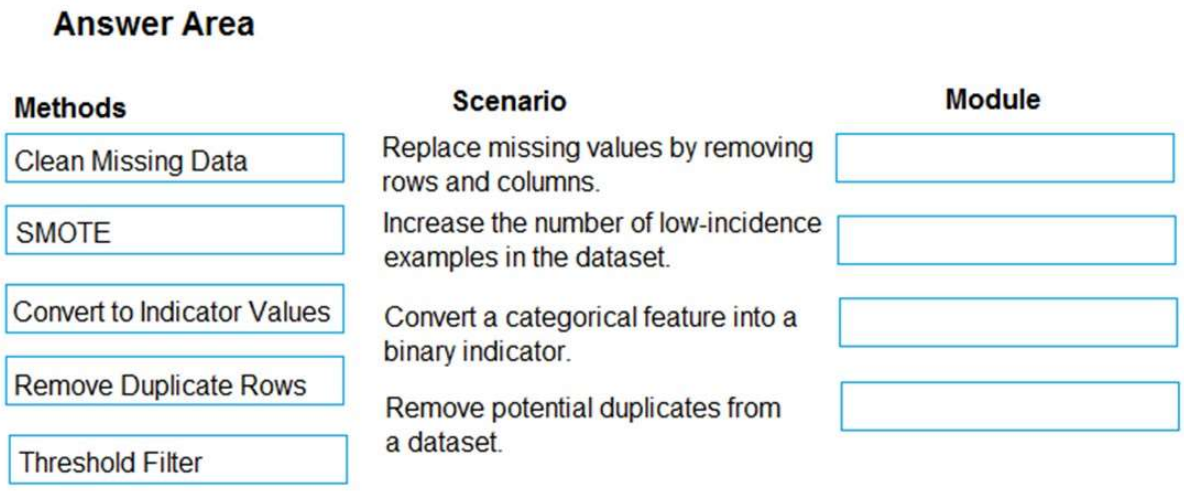
You must perform transformations and manipulations by using Azure Machine Learning Studio.

You need to identify the correct modules to perform the transformations.

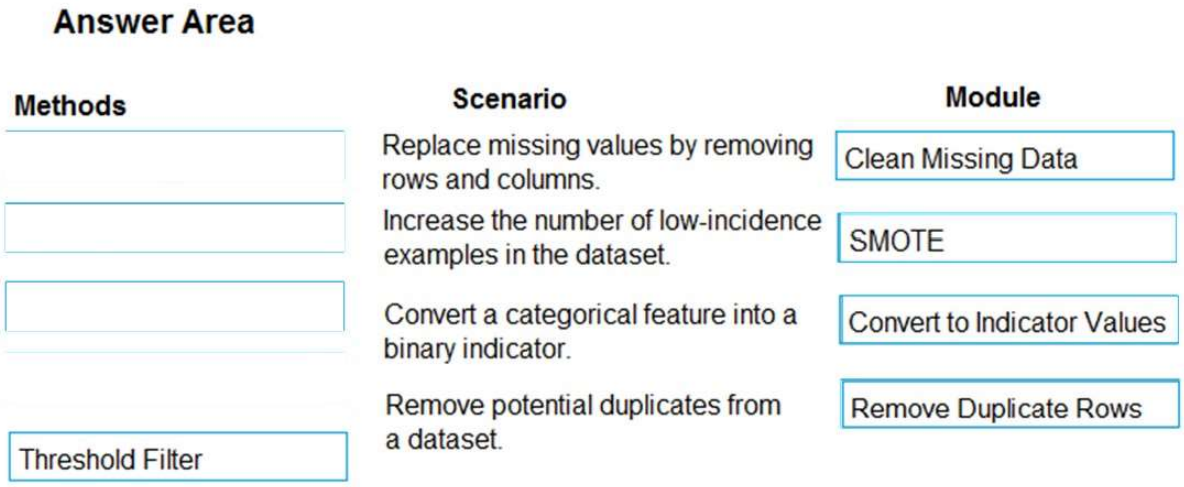
Which modules should you choose? To answer, drag the appropriate modules to the correct scenarios. Each module may be used once, more than once, or not at all.
You may need to drag the split bar between panes or scroll to view content.

NOTE: Each correct selection is worth one point.

Select and Place:



Correct Answer:



Section: (none)
Explanation

Explanation/Reference:
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

QUESTION 11
HOTSPOT

You have a Python data frame named **salesData** in the following format:

	shop	2017	2018
0	Shop X	34	25
1	Shop Y	65	76
2	Shop Z	48	55

The data frame must be unpivoted to a long data format as follows:

	shop	year	value
0	Shop X	2017	34
1	Shop Y	2017	65
2	Shop Z	2017	48
3	Shop X	2018	25
4	Shop Y	2018	76
5	Shop Z	2018	55

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.

Hot Area:

Answer Area

```
import pandas as pd
salesData = pd.melt(  [ ▼ ]  ,id_vars='  [ ▼ ]  ',value_vars=  [ ▼ ]  )
```

dataFrame	shop	'shop'
pandas	year	'year'
salesData	value	['year']
year	Shop X, Shop Y, Shop Z	['2017', '2018]

Correct Answer:

Answer Area

```
import pandas as pd
salesData = pd.melt(  [ ▼ ]  ,id_vars='  [ ▼ ]  ',value_vars=  [ ▼ ]  )
```

dataFrame	shop	'shop'
pandas	year	'year'
salesData	value	['year']
year	Shop X, Shop Y, Shop Z	['2017', '2018]

Section: (none)
Explanation

Explanation/Reference:
Explanation:

Box 1: dataFrame
Syntax: pandas.melt(frame, id_vars=None, value_vars=None, var_name=None, value_name='value', col_level=None)[source]

Where frame is a DataFrame

Box 2: shop
Paramter id_vars id_vars : tuple, list, or ndarray, optional
Column(s) to use as identifier variables.

Box 3: ['2017','2018']
value_vars : tuple, list, or ndarray, optional
Column(s) to unpivot. If not specified, uses all columns that are not set as id_vars.

Example:
```
df = pd.DataFrame({'A': {0: 'a', 1: 'b', 2: 'c'},
...                    'B': {0: 1, 1: 3, 2: 5},
...                    'C': {0: 2, 1: 4, 2: 6}})

pd.melt(df, id_vars=['A'], value_vars=['B', 'C'])
   A variable  value
0  a        B      1
1  b        B      3
2  c        B      5
3  a        C      2
4  b        C      4
5  c        C      6
```

References:
https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.melt.html

QUESTION 12
HOTSPOT

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:

Name	Description
IsPlaySoccer	Values can be 1 and 0.
Gender	Values can be M or F.
PrevExamMarks	Stores values from 0 to 100
Height	Stores values in centimeters
Weight	Stores values in kilograms

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.

Hot Area:

Answer Area

Category	Variables
Categorical variables	Gender, IsPlaySoccer Gender, PrevExamMarks, Height, Weight PrevExamMarks, Height, Weight IsPlaySoccer
Continuous variables	Gender, IsPlaySoccer Gender, PrevExamMarks, Height, Weight PrevExamMarks, Height, Weight IsPlaySoccer

Correct Answer:

Answer Area

Category	Variables
Categorical variables	Gender, IsPlaySoccer Gender, PrevExamMarks, Height, Weight PrevExamMarks, Height, Weight IsPlaySoccer
Continuous variables	Gender, IsPlaySoccer Gender, PrevExamMarks, Height, Weight PrevExamMarks, Height, Weight IsPlaySoccer

Section: (none)
Explanation

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

QUESTION 13
HOTSPOT

You plan to preprocess text from CSV files. You load the Azure Machine Learning Studio default stop words list.

You need to configure the Preprocess Text module to meet the following requirements:

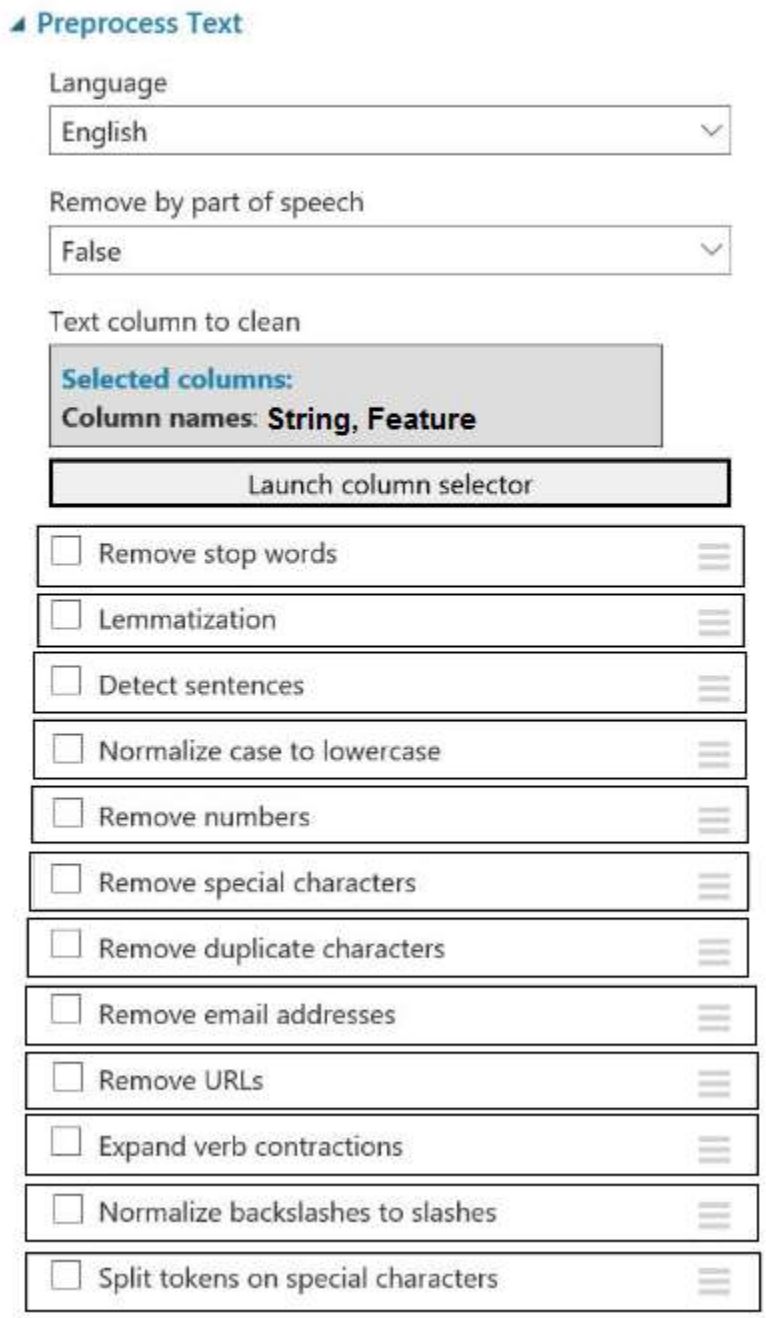
- Ensure that multiple related words from a single canonical form.
- Remove pipe characters from text.
- Remove words to optimize information retrieval.

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

NOTE: Each correct selection is worth one point.

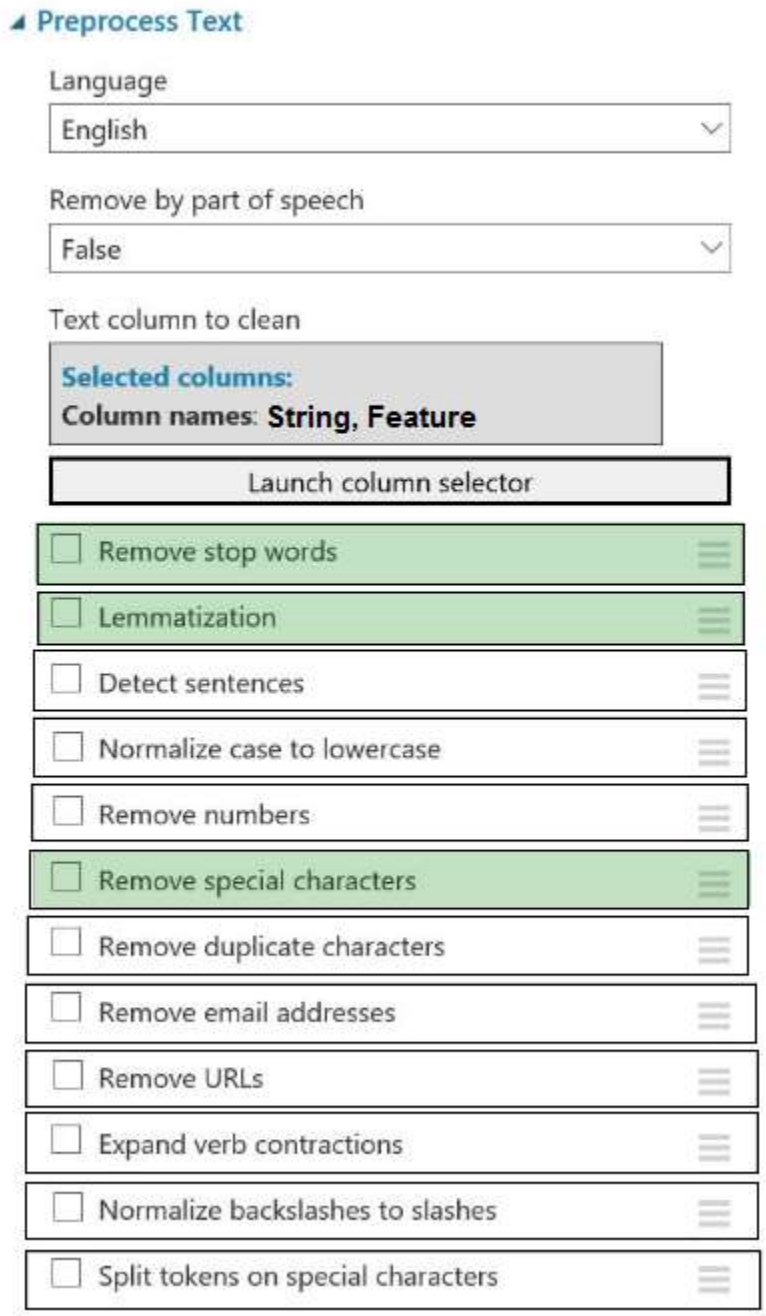
Hot Area:

Answer Area



Correct Answer:

Answer Area



Section: (none)
Explanation

Explanation/Reference:
Explanation:

Box 1: Remove stop words
Remove words to optimize information retrieval.
Remove stop words: Select this option if you want to apply a predefined stopword list to the text column. Stop word removal is performed before any other processes.

Box 2: Lemmatization
Ensure that multiple related words from a single canonical form.
Lemmatization converts multiple related words to a single canonical form

Box 3: Remove special characters
Remove special characters: Use this option to replace any non-alphanumeric special characters with the pipe | character.

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

QUESTION 14
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 Studio to perform feature engineering on a dataset.

You need to normalize values to produce a feature column grouped into bins.

Solution: Apply an Entropy Minimum Description Length (MDL) binning mode.

Does the solution meet the goal?

A. Yes
B. No

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Entropy MDL binning mode: This method requires that you select the column you want to predict and the column or columns that you want to group into bins. It then makes a pass over the data and attempts to determine the number of bins that minimizes the entropy. In other words, it chooses a number of bins that allows the data column to best predict the target column. It then returns the bin number associated with each row of your data in a column named <colname>quantized.

References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/group-data-into-bins

QUESTION 15
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

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
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

QUESTION 16
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

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
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

QUESTION 17
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

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:
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

QUESTION 18
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 Stratified split for the sampling mode.

Does the solution meet the goal?

A. Yes
B. No

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
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

QUESTION 19
You are creating a machine learning model.

You need to identify outliers in the data.

Which two visualizations can you use? Each correct answer presents a complete solution.

NOTE: Each correct selection is worth one point.

A. Venn diagram
B. Box plot
C. ROC curve
D. Random forest diagram
E. Scatter plot

Correct Answer: BE
Section: (none)
Explanation

Explanation/Reference:
Explanation:
The box-plot algorithm can be used to display outliers.

One other way to quickly identify Outliers visually is to create scatter plots.

References:
https://blogs.msdn.microsoft.com/azuredev/2017/05/27/data-cleansing-tools-in-azure-machine-learning/

QUESTION 20
You are analyzing a dataset by using Azure Machine Learning Studio.

You need to generate a statistical summary that contains the p-value and the unique count for each feature column.

Which two modules can you use? Each correct answer presents a complete solution.

NOTE: Each correct selection is worth one point.

A. Computer Linear Correlation
B. Export Count Table
C. Execute Python Script
D. Convert to Indicator Values
E. Summarize Data

Correct Answer: BE
Section: (none)
Explanation

Explanation/Reference:
Explanation:
The Export Count Table module is provided for backward compatibility with experiments that use the Build Count Table (deprecated) and Count Featurizer (deprecated) modules.

E: Summarize Data statistics are useful when you want to understand the characteristics of the complete

dataset. For example, you might need to know:

How many missing values are there in each column?
How many unique values are there in a feature column?
What is the mean and standard deviation for each column?
The module calculates the important scores for each column, and returns a row of summary statistics for each variable (data column) provided as input.

Incorrect Answers:
A: The Compute Linear Correlation module in Azure Machine Learning Studio is used to compute a set of Pearson correlation coefficients for each possible pair of variables in the input dataset.

C: With Python, you can perform tasks that aren't currently supported by existing Studio modules such as:
Visualizing data using matplotlib
Using Python libraries to enumerate datasets and models in your workspace
Reading, loading, and manipulating data from sources not supported by the Import Data module

D: The purpose of the Convert to Indicator Values 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.

References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/export-count-table

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

QUESTION 21
You are evaluating a completed binary classification machine learning model.

You need to use the precision as the evaluation metric.

Which visualization should you use?

A. Violin plot
B. Gradient descent
C. Box plot
D. Binary classification confusion matrix

Correct Answer: D
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Incorrect Answers:
A: A violin plot is a visual that traditionally combines a box plot and a kernel density plot.

B: Gradient descent is a first-order iterative optimization algorithm for finding the minimum of a function. To find a local minimum of a function using gradient descent, one takes steps proportional to the negative of the gradient (or approximate gradient) of the function at the current point.

C: A box plot lets you see basic distribution information about your data, such as median, mean, range and quartiles but doesn't show you how your data looks throughout its range.

References:
https://machinelearningknowledge.ai/confusion-matrix-and-performance-metrics-machine-learning/

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 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: Use the Last Observation Carried Forward (LOCF) method to impute the missing data points.

Does the solution meet the goal?

A. Yes
B. No

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Instead use the Multiple Imputation by Chained Equations (MICE) method.

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: Last observation carried forward (LOCF) is a method of imputing missing data in longitudinal studies. If a person drops out of a study before it ends, then his or her last observed score on the dependent variable is used for all subsequent (i.e., missing) observation points. LOCF is used to maintain the sample size and to reduce the bias caused by the attrition of participants in a study.

References:
https://methods.sagepub.com/reference/encyc-of-research-design/n211.xml

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3074241/

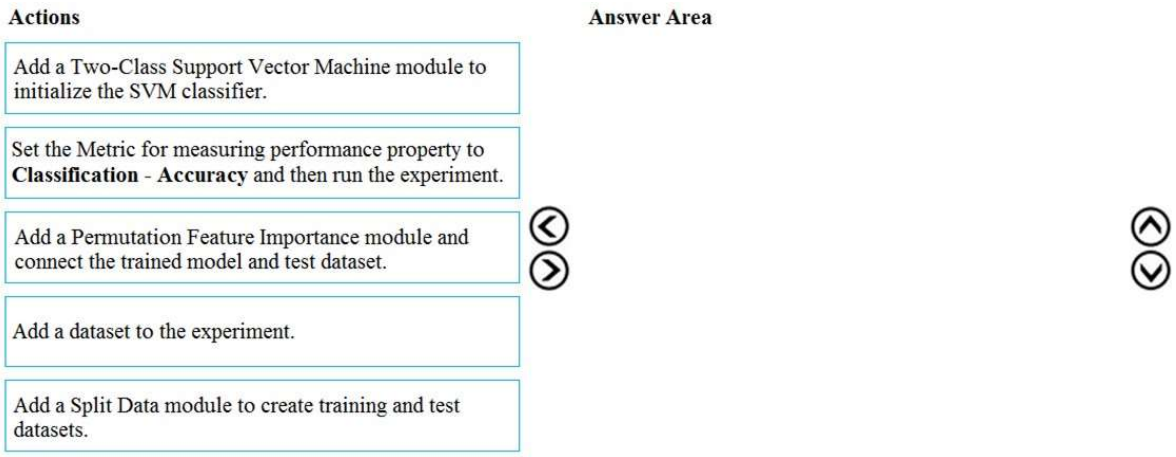
QUESTION 23
DRAG DROP

You have a dataset that contains over 150 features. You use the dataset to train a Support Vector Machine (SVM) binary classifier.

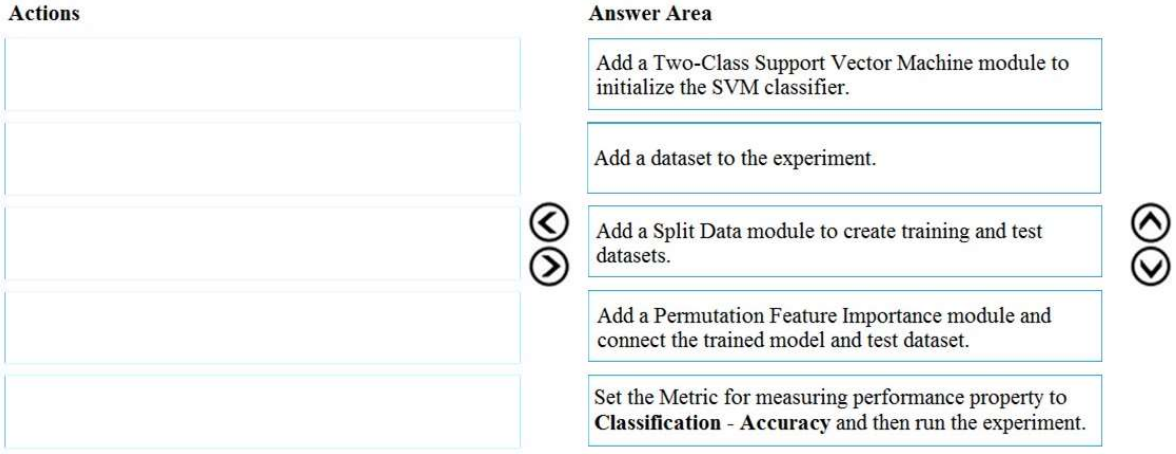
You need to use the Permutation Feature Importance module in Azure Machine Learning Studio to compute a set of feature importance scores for the dataset.

In which order should you perform the actions? To answer, move all actions from the list of actions to the answer area and arrange them in the correct order.

Select and Place:



Correct Answer:



Section: (none)
Explanation

Explanation/Reference:
Explanation:

Step 1: Add a Two-Class Support Vector Machine module to initialize the SVM classifier.

Step 2: Add a dataset to the experiment

Step 3: Add a Split Data module to create training and test dataset.
To generate a set of feature scores requires that you have an already trained model, as well as a test dataset.

Step 4: Add a Permutation Feature Importance module and connect to the trained model and test dataset.

Step 5: Set the Metric for measuring performance property to Classification - Accuracy and then run the experiment.

Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/two-class-support-vector-machine

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

QUESTION 24

HOTSPOT

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.

Hot Area:

Answer Area

```
from sklearn import linear_model
import [ pandas as df | numpy as df | scipy as df ]
dataset = df.read_csv("data\\ProductSales.csv")
ProductCategoryMapping = {"Bikes":1, "Cars":2, "Boats": 3,
"Vans": 4}
dataset['ProductCategoryMapping'] =
dataset['ProductCategory'].[ map[ProductCategoryMapping] | reduce[ProductCategoryMapping] | transpose[ProductCategoryMapping] ]
regr = linear_model.LinearRegression()
X_train = dataset[['ProductCategoryMapping', 'ProductSize',
'ProductCost']]
y_train = dataset[['Sales']]
regr.fit(X_train, y_train)
```

Correct Answer:

Answer Area

```
from sklearn import linear_model
import [pandas as df | numpy as df | scipy as df]
dataset = df.read_csv("data\\ProductSales.csv")
ProductCategoryMapping = {"Bikes":1, "Cars":2, "Boats": 3,
"Vans": 4}
dataset['ProductCategoryMapping'] =
dataset['ProductCategory'].[map[ProductCategoryMapping] | reduce[ProductCategoryMapping] | transpose[ProductCategoryMapping]]
regr = linear_model.LinearRegression()
X_train = dataset[['ProductCategoryMapping', 'ProductSize',
'ProductCost']]
y_train = dataset[['Sales']]
regr.fit(X_train, y_train)
```

Section: (none)
Explanation

Explanation/Reference:
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/

QUESTION 25
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

Correct Answer: D
Section: (none)
Explanation

Explanation/Reference:
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.

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/

QUESTION 26
HOTSPOT

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.

Hot Area:

Answer Area

Description	Error type
A person has a disease. The model classifies the case as having a disease.	True Positives True Negatives False Positives False Negatives
A person does not have a disease. The model classifies the case as having no disease.	True Positives True Negatives False Positives False Negatives
A person does not have a disease. The model classifies the case as having a disease.	True Positives True Negatives False Positives False Negatives
A person has a disease. The model classifies the case as having no disease.	True Positives True Negatives False Positives False Negatives

Correct Answer:

Answer Area

Description	Error type
A person has a disease. The model classifies the case as having a disease.	True Positives
A person does not have a disease. The model classifies the case as having no disease.	True Negatives
A person does not have a disease. The model classifies the case as having a disease.	False Positives
A person has a disease. The model classifies the case as having no disease.	False Negatives

Section: (none)
Explanation

Explanation/Reference:
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

QUESTION 27
HOTSPOT

You are using the Azure Machine Learning Service to automate hyperparameter exploration of your neural network classification model.

You must define the hyperparameter space to automatically tune hyperparameters using random sampling according to following requirements:

- The learning rate must be selected from a normal distribution with a mean value of 10 and a standard deviation of 3.
- Batch size must be 16, 32 and 64.
- Keep probability must be a value selected from a uniform distribution between the range of 0.05 and 0.1.

You need to use the param_sampling method of the Python API for the Azure Machine Learning Service.

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.

Hot Area:

Answer Area

```
from azureml.train.hyperdrive import RandomParameterSampling
param_sampling = RandomParameterSampling( {
"learning_rate" :  [                ▼] ,
                   uniform(10,3)
                   normal(10,3)
                   choice(10,3)
                   Loguniform(10,3)
"batch_size":      [                ▼] ,
                   choice(16,32,64)
                   choice(range(16,64))
                   normal(16,32,64)
                   normal(range(16,64))
"keep_probability" : [              ▼]
                   choice(range(0.05, 0.1))
                   uniform(0.05, 0.1)
                   normal(0.05, 0.1)
                   lognormal(0.05, 0.1)
    }
)
```

Correct Answer:

Answer Area

```
from azureml.train.hyperdrive import RandomParameterSampling
param_sampling = RandomParameterSampling( {
"learning_rate" :  [uniform(10,3) | normal(10,3) | choice(10,3) | Loguniform(10,3)] ,
"batch_size":  [choice(16,32,64) | choice(range(16,64)) | normal(16,32,64) | normal(range(16,64))] ,
"keep_probability" :  [choice(range(0.05, 0.1)) | uniform(0.05, 0.1) | normal(0.05, 0.1) | lognormal(0.05, 0.1)]
    }
)
```

Section: (none)
Explanation

Explanation/Reference:
Explanation:

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.

Example:

```
from azureml.train.hyperdrive import RandomParameterSampling
param_sampling = RandomParameterSampling( {
    "learning_rate": normal(10, 3),
    "keep_probability": uniform(0.05, 0.1),
    "batch_size": choice(16, 32, 64)
  }
```

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

QUESTION 28
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

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:
References:
https://notebooks.azure.com/

QUESTION 29
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

Correct Answer: A
Section: (none)
Explanation

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

QUESTION 30
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: Remove the entire column that contains the missing data point.

Does the solution meet the goal?

A. Yes
B. No

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Use the Multiple Imputation by Chained Equations (MICE) method.

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

QUESTION 31
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 Principal Components Analysis (PCA) sampling mode.

Does the solution meet the goal?

A. Yes
B. No

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
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.

Incorrect Answers:
The Principal Component Analysis module in Azure Machine Learning Studio (classic) is used to reduce the dimensionality of your training data. The module analyzes your data and creates a reduced feature set that captures all the information contained in the dataset, but in a smaller number of features.

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/principal-component-analysis

QUESTION 32
You are creating a new experiment in Azure Machine Learning Studio. You have a small dataset that has missing values in many columns. The data does not require the application of predictors for each column. You plan to use the Clean Missing Data.

You need to select a data cleaning method.

Which method should you use?

A. Replace using Probabilistic PCA
B. Normalization
C. Synthetic Minority Oversampling Technique (SMOTE)
D. Replace using MICE

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Replace using Probabilistic PCA: Compared to other options, such as Multiple Imputation using Chained Equations (MICE), this option has the advantage of not requiring the application of predictors for each column. Instead, it approximates the covariance for the full dataset. Therefore, it might offer better performance for datasets that have missing values in many columns.

References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/clean-missing-data

QUESTION 33
You are evaluating a completed binary classification machine learning model.

You need to use the precision as the evaluation metric.

Which visualization should you use?

A. violin plot

B. Gradient descent
C. Scatter plot
D. Receiver Operating Characteristic (ROC) curve

Correct Answer: D
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Receiver operating characteristic (or ROC) is a plot of the correctly classified labels vs. the incorrectly classified labels for a particular model.

Incorrect Answers:
A: A violin plot is a visual that traditionally combines a box plot and a kernel density plot.

B: Gradient descent is a first-order iterative optimization algorithm for finding the minimum of a function. To find a local minimum of a function using gradient descent, one takes steps proportional to the negative of the gradient (or approximate gradient) of the function at the current point.

C: A scatter plot graphs the actual values in your data against the values predicted by the model. The scatter plot displays the actual values along the X-axis, and displays the predicted values along the Y-axis. It also displays a line that illustrates the perfect prediction, where the predicted value exactly matches the actual value.

References:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-understand-automated-ml#confusion-matrix

QUESTION 34
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

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
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.

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. Split Data
B. Load Trained Model
C. Assign Data to Clusters
D. Group Data into Bins

Correct Answer: D
Section: (none)
Explanation

Explanation/Reference:
Explanation:
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.

References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/group-data-into-bins

QUESTION 36
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.

Correct Answer: D
Section: (none)
Explanation

Explanation/Reference:
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

QUESTION 37

DRAG DROP

You plan to explore demographic data for home ownership in various cities. The data is in a CSV file with the following format:

```
age,city,income,home_owner
21,Chicago,50000,0
35,Seattle,120000,1
23,Seattle,65000,0
45,Seattle,130000,1
18,Chicago,48000,0
```

You need to run an experiment in your Azure Machine Learning workspace to explore the data and log the results. The experiment must log the following information:

- the number of observations in the dataset
- a box plot of income by home_owner
- a dictionary containing the city names and the average income for each city

You need to use the appropriate logging methods of the experiment's run object to log the required information.

How should you complete the code? To answer, drag the appropriate code segments to the correct locations. Each code segment 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.

Select and Place:

Code segments

log
log_list
log_row
log_table
log_image

Answer Area

```
from azureml.core import Experiment, Run
import pandas as pd
import matplotlib.pyplot as plt
# Create an Azure ML experiment in workspace
experiment = Experiment(workspace = ws, name = "demo-experiment")
# Start logging data from the experiment
run = experiment.start_logging()
# load the dataset
data = pd.read_csv('research/demographics.csv')
# Log the number of observations
row_count = (len(data))
run. Segment ("observations", row_count)
# Log box plot for income by home_owner
fig = plt.figure(figsize=(9, 6))
ax = fig.gca()
data.boxplot(column = 'income', by = "home_owner", ax = ax)
ax.set_title('income by home_owner')
ax.set_ylabel('income')
run. Segment (name = 'income_by_home_owner', plot = fig)
# Create a dataframe of mean income per city
mean_inc_df = data.groupby('city')['income'].agg(np.mean).to_frame().reset_index()
# Convert to a dictionary
mean_inc_dict = mean_inc_df.to_dict('dict')
# Log city names and average income dictionary
run. Segment (name="mean_income_by_city", value= mean_inc_dict)
# Complete tracking and get link to details
run.complete()
```

Correct Answer:

Code segments

log_list

log_row

Answer Area

```
from azureml.core import Experiment, Run
import pandas as pd
import matplotlib.pyplot as plt
# Create an Azure ML experiment in workspace
experiment = Experiment(workspace = ws, name = "demo-experiment")
# Start logging data from the experiment
run = experiment.start_logging()
# load the dataset
data = pd.read_csv('research/demographics.csv')
# Log the number of observations
row_count = (len(data))
run. log          ("observations", row_count)
# Log box plot for income by home_owner
fig = plt.figure(figsize=(9, 6))
ax = fig.gca()
data.boxplot(column = 'income', by = "home_owner", ax = ax)
ax.set_title('income by home_owner')
ax.set_ylabel('income')
run. log_image    (name = 'income_by_home_owner', plot = fig)
# Create a dataframe of mean income per city
mean_inc_df = data.groupby('city')['income'].agg(np.mean).to_frame().reset_index()
# Convert to a dictionary
mean_inc_dict = mean_inc_df.to_dict('dict')
# Log city names and average income dictionary
run. log_table    (name="mean_income_by_city", value= mean_inc_dict)
# Complete tracking and get link to details
run.complete()
```

Section: (none)
Explanation

Explanation/Reference:
Explanation:

Box 1: log
The number of observations in the dataset.

run.log(name, value, description='')
Scalar values: Log a numerical or string value to the run with the given name. Logging a metric to a run causes that metric to be stored in the run record in the experiment. You can log the same metric multiple times within a run, the result being considered a vector of that metric.

Example: run.log("accuracy", 0.95)

Box 2: log_image
A box plot of income by home_owner.

log_image Log an image to the run record. Use log_image to log a .PNG image file or a matplotlib plot to the run. These images will be visible and comparable in the run record.

Example: run.log_image("ROC", plot=plt)

Box 3: log_table
A dictionary containing the city names and the average income for each city.

log_table: Log a dictionary object to the run with the given name.

QUESTION 38
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. `environment_definition = {"training_data":training_ds}`
B. `inputs = [training_ds.as_named_input('training_ds')]`
C. `script_params = {"--training_ds":training_ds}`
D. `source_directory = training_ds`

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
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%20Training.ipynb

QUESTION 39
You register a file dataset named csv_folder that references a folder. The folder includes multiple comma-separated values (CSV) files in an Azure storage blob container.

You plan to use the following code to run a script that loads data from the file dataset. You create and instantiate the following variables:

Variable	Description
remote_cluster	References the Azure Machine Learning compute cluster
ws	References the Azure Machine Learning workspace

You have the following code:

```
from azureml.train.estimator import Estimator
file_dataset = ws.datasets.get('csv_folder')
estimator = Estimator(source_directory=script_folder,

compute_target = remote_cluster,
entry_script ='script.py')
run = experiment.submit(config=estimator)
run.wait_for_completion(show_output=True)
```

You need to pass the dataset to ensure that the script can read the files it references.

Which code segment should you insert to replace the code comment?

A. `inputs=[file_dataset.as_named_input('training_files')],`
B. `inputs=[file_dataset.as_named_input('training_files').as_mount()],`
C. `inputs=[file_dataset.as_named_input('training_files').to_pandas_dataframe()],`
D. `script_params={'--training_files': file_dataset},`

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Example:
from azureml.train.estimator import Estimator

```
script_params = {
    # to mount files referenced by mnist dataset
    '--data-folder': mnist_file_dataset.as_named_input('mnist_opendataset').as_mount(),
    '--regularization': 0.5
}

est = Estimator(source_directory=script_folder,
                script_params=script_params,
                compute_target=compute_target,
                environment_definition=env,
                entry_script='train.py')
```

Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/tutorial-train-models-with-aml

QUESTION 40
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

Correct Answer: D
Section: (none)
Explanation

Explanation/Reference:
Explanation:
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

QUESTION 41
You define a datastore named **ml-data** for an Azure Storage blob container. In the container, you have a folder named train that contains a file named data.csv. You plan to use the file to train a model by using the Azure Machine Learning SDK.

You plan to train the model by using the Azure Machine Learning SDK to run an experiment on local compute.

You define a DataReference object by running the following code:

```
from azureml.core import Workspace, Datastore, Environment
from azureml.train.estimator import Estimator
ws = Workspace.from_config()
ml_data = = Datastore.get(ws, datastore_name='ml-data')
data_ref = ml_data.path('train').as_download(path_on_compute='train_data')
estimator = Estimator(source_directory='experiment_folder',
  script_params={'--data-folder': data_ref},
  compute_target = 'local',
  entry_script='training.py')
run = experiment.submit(config=estimator)
run.wait_for_completion(show_output=True)
```

You need to load the training data.

Which code segment should you use?

A.
```
import os
import argparse
import pandas as pd

parser = argparse.ArgumentParser()
parser.add_argument('--data-folder', type=str, dest='data_folder')
data_folder = args.data_folder
data = pd.read_csv(os.path.join(data_folder,'ml-data','train_data','data.csv'))
```

B.
```
import os
import argparse
import pandas as pd

parser = argparse.ArgumentParser()
parser.add_argument('--data-folder', type=str, dest='data_folder')
data_folder = args.data_folder
data = pd.read_csv(os.path.join(data_folder,'train','data.csv'))
```

C.
```
import pandas as pd

data = pd.read_csv('./data.csv')
```

D.
```
import os
import argparse
import pandas as pd

parser = argparse.ArgumentParser()
parser.add_argument('--data-folder', type=str, dest='data_folder')
data_folder = args.data_folder
data = pd.read_csv(os.path.join('ml_data', data_folder,'data.csv'))
```

E.
```
import os
import argparse
import pandas as pd

parser = argparse.ArgumentParser()
parser.add_argument('--data-folder', type=str, dest='data_folder')
data_folder = args.data_folder
data = pd.read_csv(os.path.join(data_folder,'data.csv'))
```

Correct Answer: E
Section: (none)
Explanation

Explanation/Reference:
Explanation:

Example:
data_folder = args.data_folder
# Load Train and Test data
train_data = pd.read_csv(os.path.join(data_folder, 'data.csv'))

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

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:

		Actual	
		1	**0**
Predicted	**0**	1	2
	1	2	1

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

		Actual	
		1	**0**
Predicted	**0**	1	5
	1	5	1

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

QUESTION 1
DRAG DROP

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.

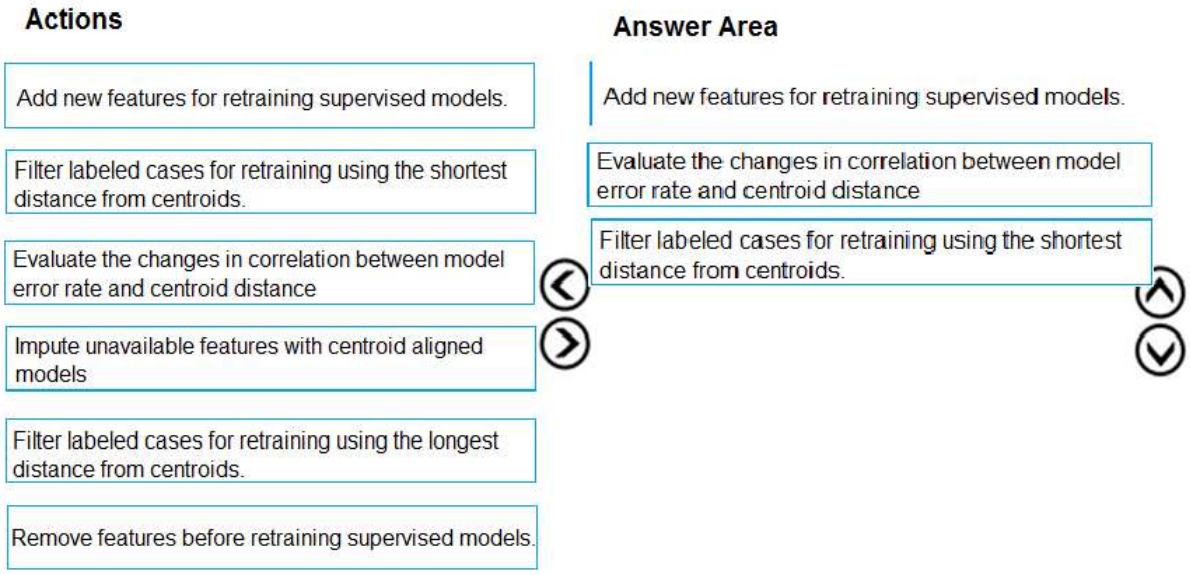
Select and Place:

Actions

- Add new features for retraining supervised models.
- Filter labeled cases for retraining using the shortest distance from centroids.
- Evaluate the changes in correlation between model error rate and centroid distance
- Impute unavailable features with centroid aligned models
- Filter labeled cases for retraining using the longest distance from centroids.
- Remove features before retraining supervised models.

Answer Area

Correct Answer:



Section: (none)
Explanation

Explanation/Reference:
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

QUESTION 2
You need to implement a feature engineering strategy for the crowd sentiment local models.

What should you do?

A. Apply an analysis of variance (ANOVA).
B. Apply a Pearson correlation coefficient.
C. Apply a Spearman correlation coefficient.
D. Apply a linear discriminant analysis.

Correct Answer: D
Section: (none)
Explanation

Explanation/Reference:
Explanation:
The linear discriminant analysis method works only on continuous variables, not categorical or ordinal variables.

Linear discriminant analysis is similar to analysis of variance (ANOVA) in that it works by comparing the means of the variables.

Scenario:
Data scientists must build notebooks in a local environment using automatic feature engineering and model building in machine learning pipelines.
Experiments for local crowd sentiment models must combine local penalty detection data.
All shared features for local models are continuous variables.

Incorrect Answers:
B: The Pearson correlation coefficient, sometimes called Pearson's R test, is a statistical value that measures the linear relationship between two variables. By examining the coefficient values, you can infer something about the strength of the relationship between the two variables, and whether they are positively correlated or negatively correlated.

C: Spearman's correlation coefficient is designed for use with non-parametric and non-normally distributed data. Spearman's coefficient is a nonparametric measure of statistical dependence between two variables, and is sometimes denoted by the Greek letter rho. The Spearman's coefficient expresses the degree to which two variables are monotonically related. It is also called Spearman rank correlation, because it can be used with ordinal variables.

References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/fisher-linear-discriminant-analysis

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

Perform Feature Engineering

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 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. 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:

Column heading	Description
CapitaCrimeRate	per capita crime rate by town
Zoned	proportion of residential land zoned for lots over 25.000 square feet
NonRetailAcres	proportion of retail business acres per town
NextToRiver	proximity of a property to the river
NitrogenOxideConcentration	nitric oxides concentration (parts per 10 million)
AvgRoomsPerHouse	average number of rooms per dwelling
Age	proportion of owner-occupied units built prior to 1940
DistanceToEmploymentCenter	weighted distances to employment centers
AccessibilityToHighway	index of accessibility to radial highways to a value of two decimal places
Tax	full value property tax rate per $10,000
PupilTeacherRatio	pupil to teacher ratio by town
ProfessionalClass	professional class percentage
LowerStatus	percentage lower status of the population
MedianValue	median value of owner-occupied homes in $1000s

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 use a feature selection algorithm to analyze the relationship between the MedianValue 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 Machine Learning Studio to compare the Two-Class Decision Forest and the Two-Class Decision Jungle modules with one another.

QUESTION 1
HOTSPOT

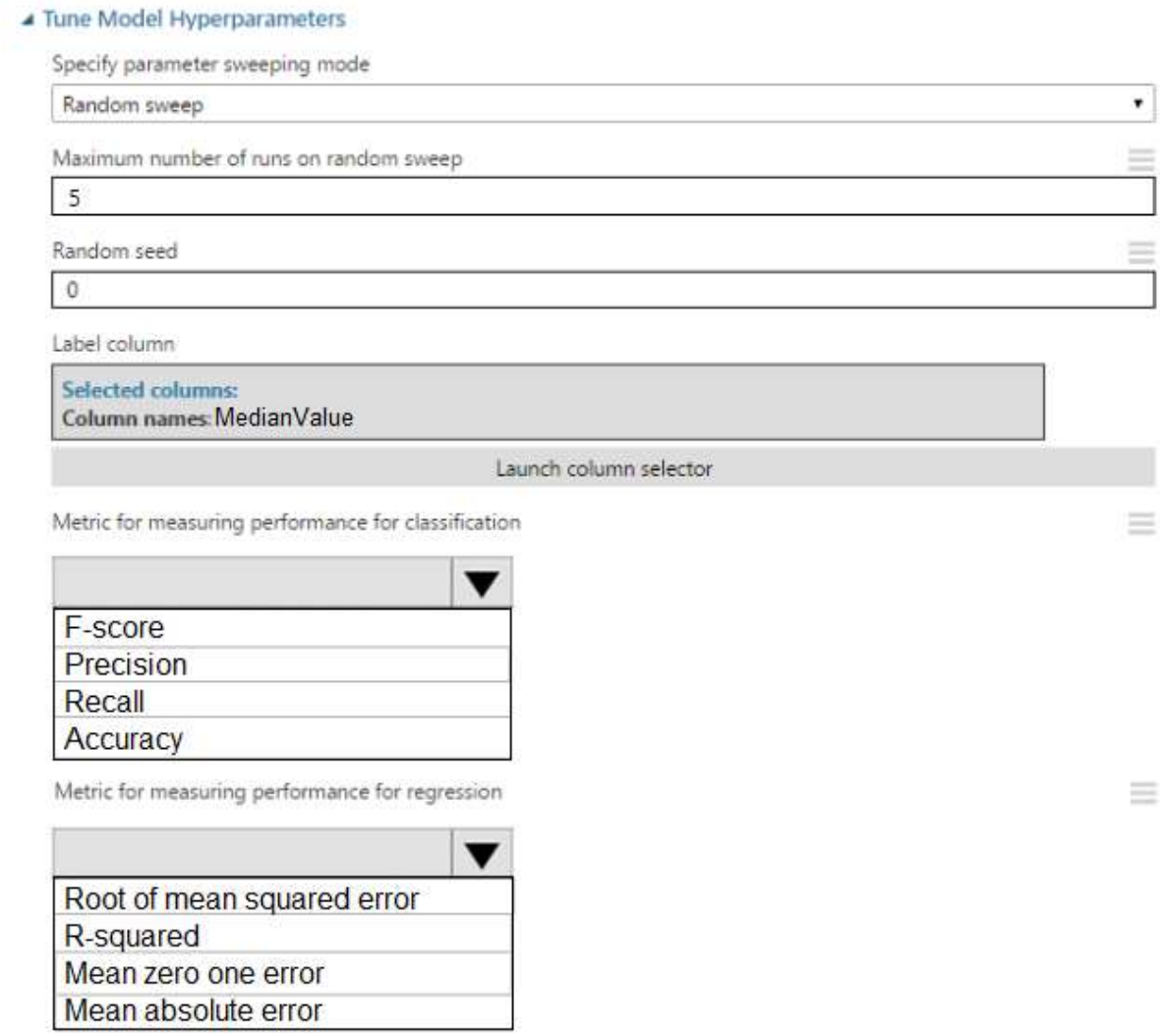
You need to set up the Permutation Feature Importance module according to the model training requirements.

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

NOTE: Each correct selection is worth one point.

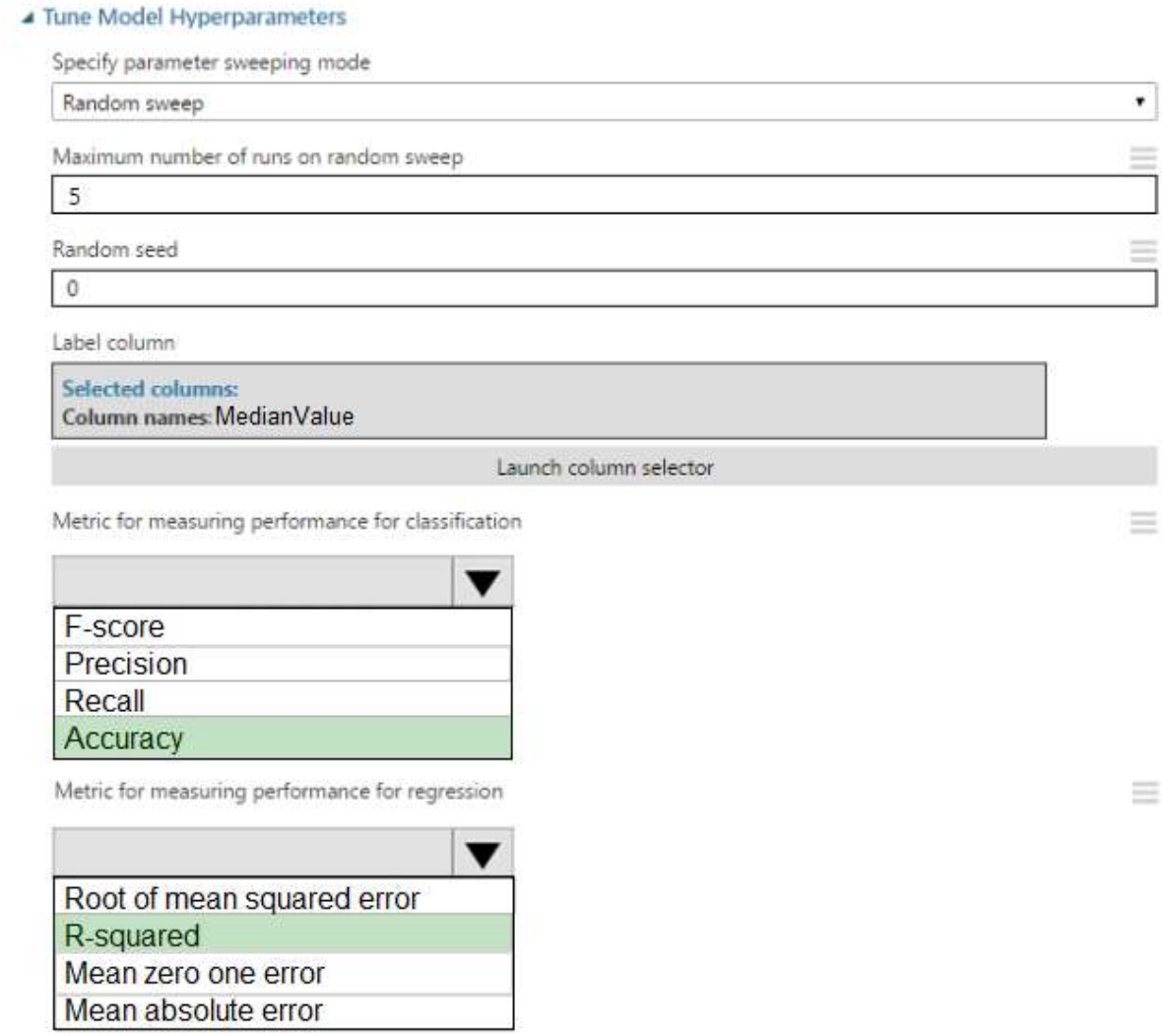
Hot Area:

Answer Area



Correct Answer:

Answer Area



Section: (none)
Explanation

Explanation/Reference:
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

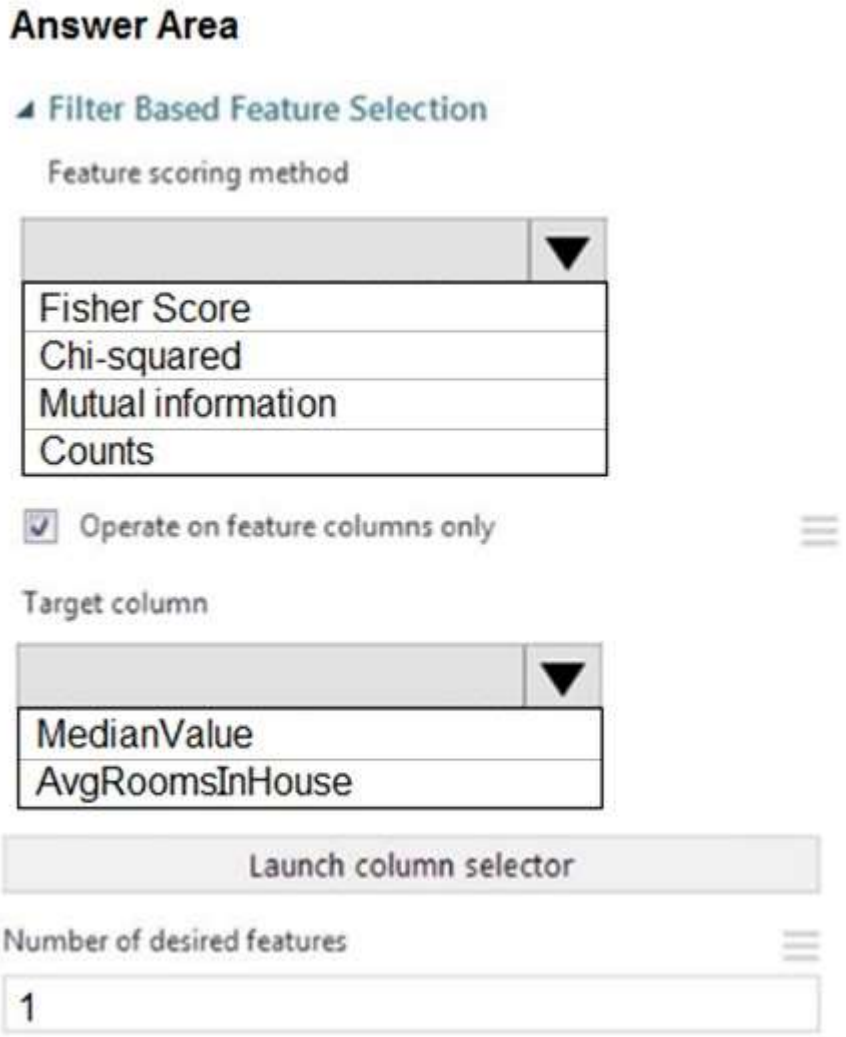
QUESTION 2
HOTSPOT

You need to configure the Feature Based Feature Selection module based on the experiment requirements and datasets.

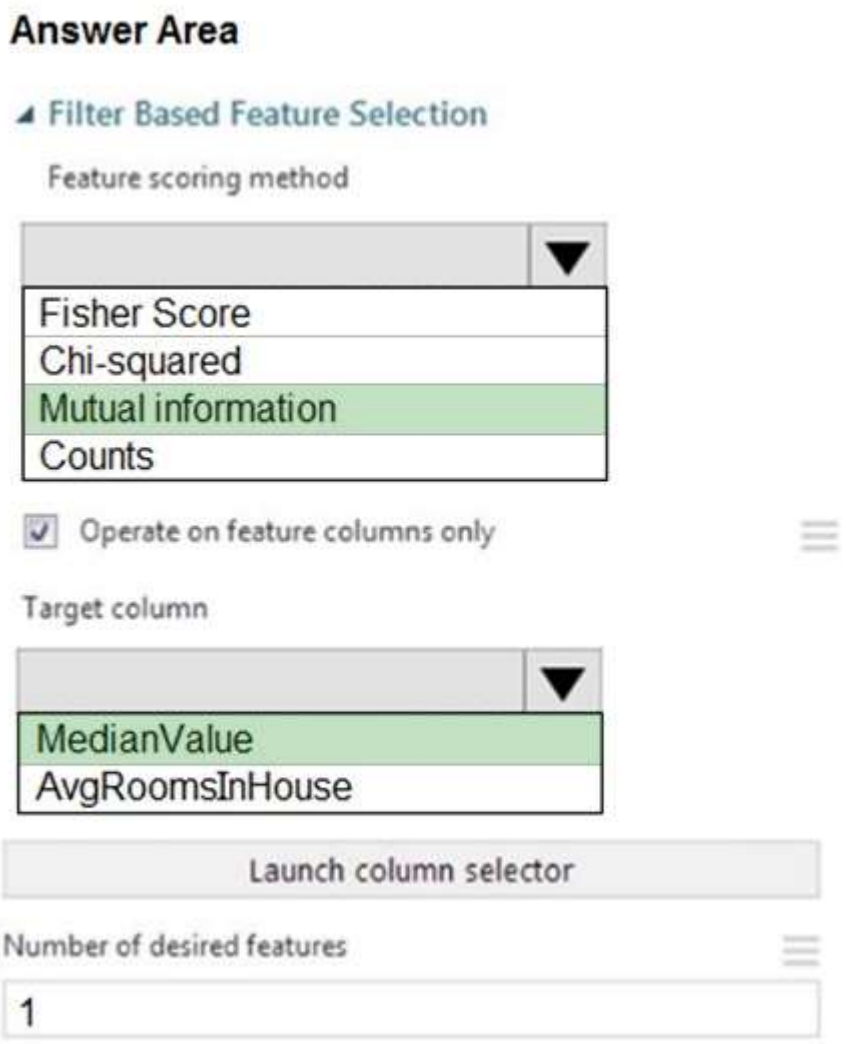
How should you configure the module properties? To answer, select the appropriate options in the dialog box in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:



Correct Answer:



Section: (none)
Explanation

Explanation/Reference:
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

QUESTION 3
You need to select a feature extraction method.

Which method should you use?

A. Mutual information
B. Mood's median test
C. Kendall correlation
D. Permutation Feature Importance

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:
Explanation:
In statistics, the Kendall rank correlation coefficient, commonly referred to as Kendall's tau coefficient (after the Greek letter τ), is a statistic used to measure the ordinal association between two measured quantities.
It is a supported method of the Azure Machine Learning Feature selection.

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.

References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/feature-selection-modules

QUESTION 4
HOTSPOT

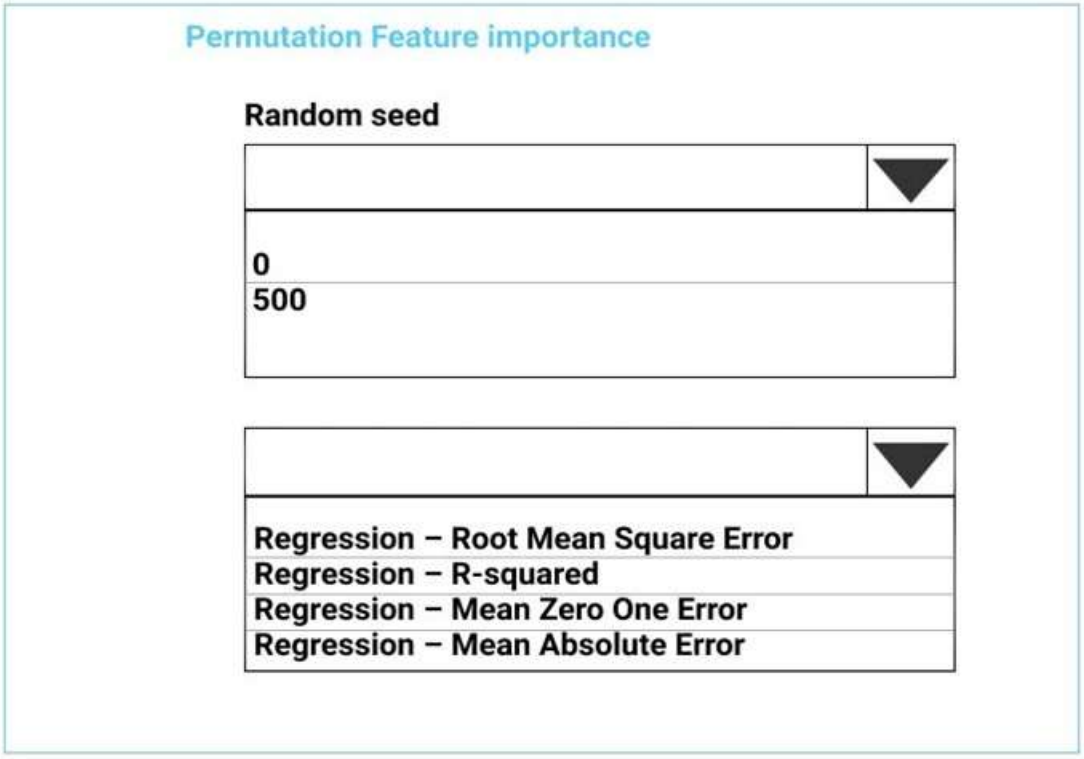
You need to configure the Permutation Feature Importance module for the model training requirements.

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.

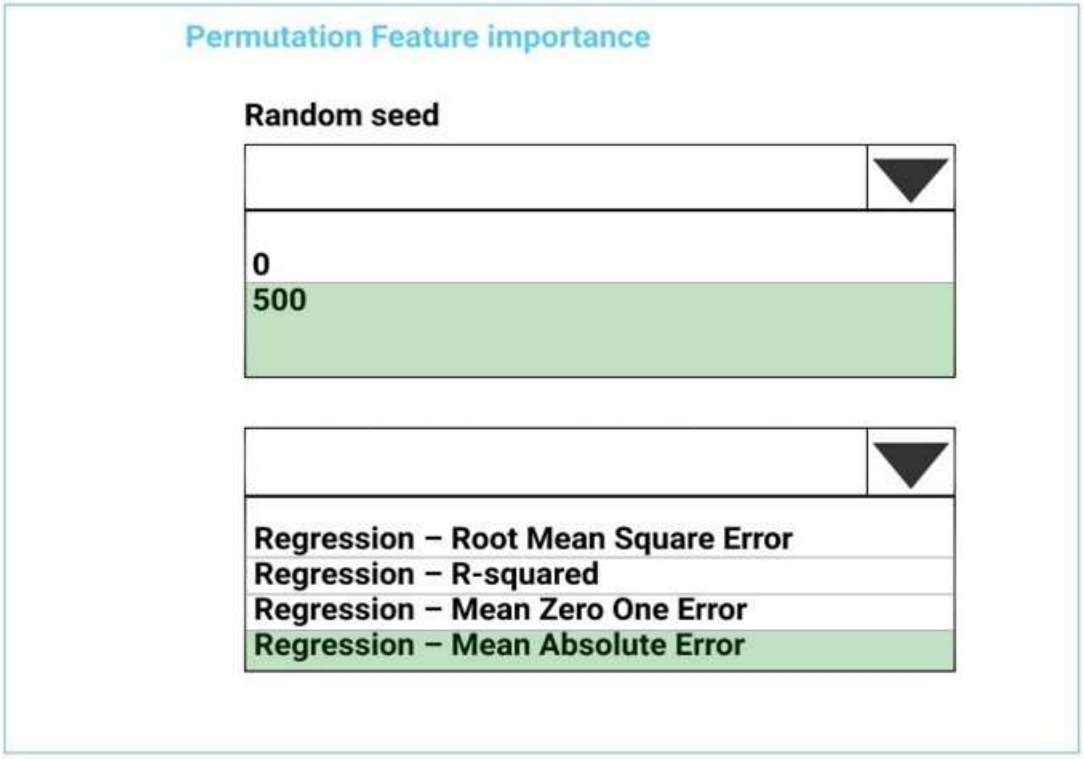
Hot Area:

Answer Area



Correct Answer:

Answer Area



Section: (none)
Explanation

Explanation/Reference:
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

QUESTION 5
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

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:
Explanation:
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

QUESTION 1
You are building a regression model for 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? Each 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 must be whole numbers.
C. The label data must be non-discrete.
D. The label data must be a positive value.
E. The label data can be positive or negative.

Correct Answer: BD
Section: (none)
Explanation

Explanation/Reference:
Explanation:
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.

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

QUESTION 2
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

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Typical metadata changes might include marking columns as features.

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

QUESTION 3
HOTSPOT

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:

Name	Description
X_train	training feature set
Y_train	training class labels
x_train	testing feature set
y_train	testing class labels

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.

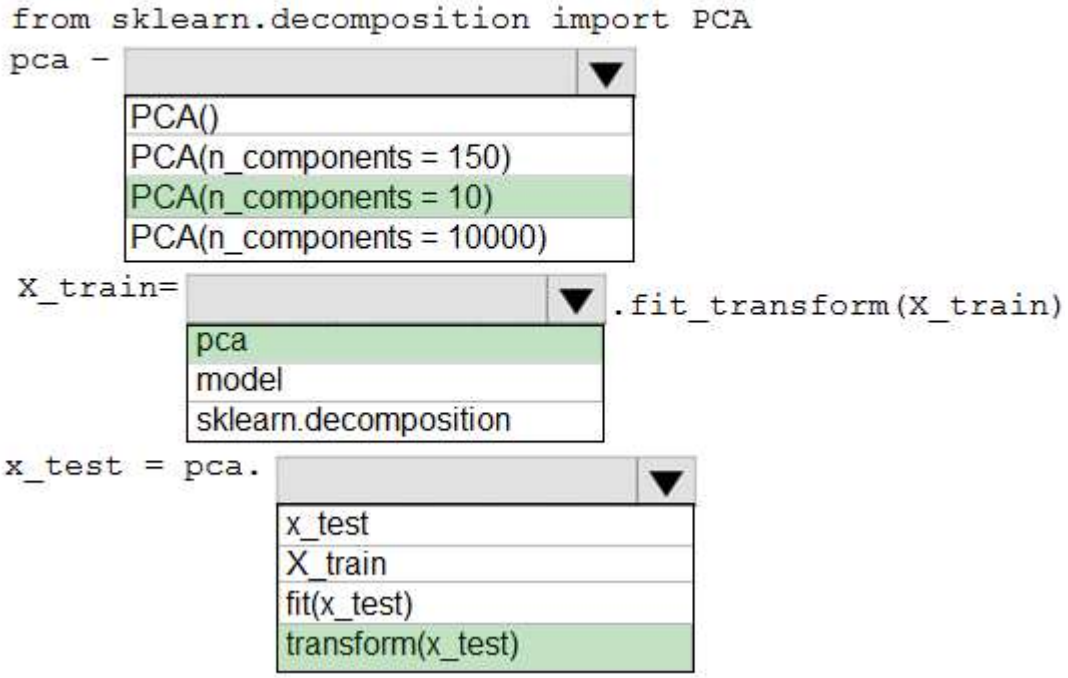
Hot Area:

Answer Area

```
from sklearn.decomposition import PCA
pca = [ PCA() | PCA(n_components = 150) | PCA(n_components = 10) | PCA(n_components = 10000) ]
X_train= [ pca | model | sklearn.decomposition ].fit_transform(X_train)
x_test = pca.[ x_test | X_train | fit(x_test) | transform(x_test) ]
```

Correct Answer:

Answer Area



Section: (none)
Explanation

Explanation/Reference:
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.

Reference:
https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

QUESTION 4
HOTSPOT

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:

	X	Y	Z
X	1	0.149676	-0.106276
Y	0.149676	1	0.859122
Z	-0.106276	0.859122	1

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.

Hot Area:

Answer Area

What is the r-value for the correlation of Y to Z?

-0.106276
0.149676
0.859122
1

Which type of relationship exists between Z and Y in the feature set?

a positive linear relationship
a negative linear relationship
no linear relationship

Correct Answer:

Answer Area

What is the r-value for the correlation of Y to Z?

-0.106276
0.149676
0.859122
1

Which type of relationship exists between Z and Y in the feature set?

a positive linear relationship
a negative linear relationship
no linear relationship

Section: (none)
Explanation

Explanation/Reference:
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

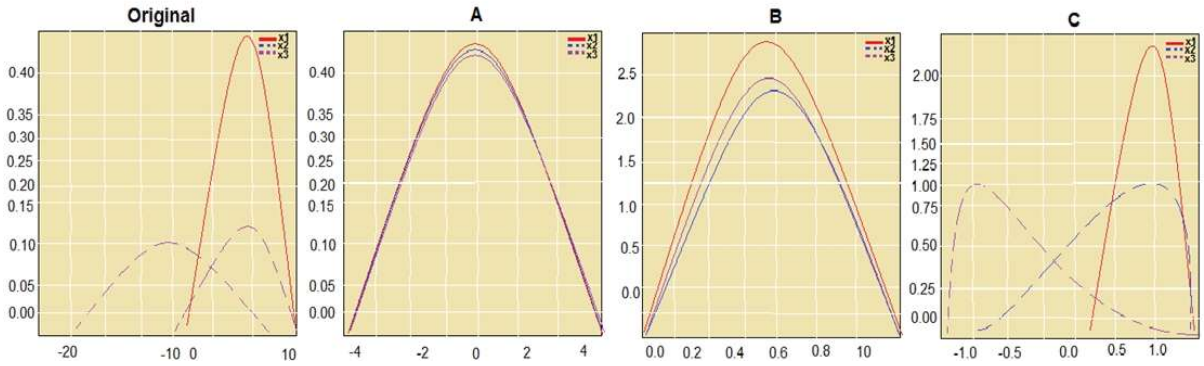
0 denotes no linear relationship between the two variables.

References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/compute-linear-correlation

QUESTION 5
HOTSPOT

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.

Hot Area:

Answer Area

Question	Answer choice
Which scaler is used in graph A?	Standard Scaler Min Max Scale Normalizer
Which scaler is used in graph B?	Standard Scaler Min Max Scale Normalizer
Which scaler is used in graph C?	Standard Scaler Min Max Scale Normalizer

Correct Answer:

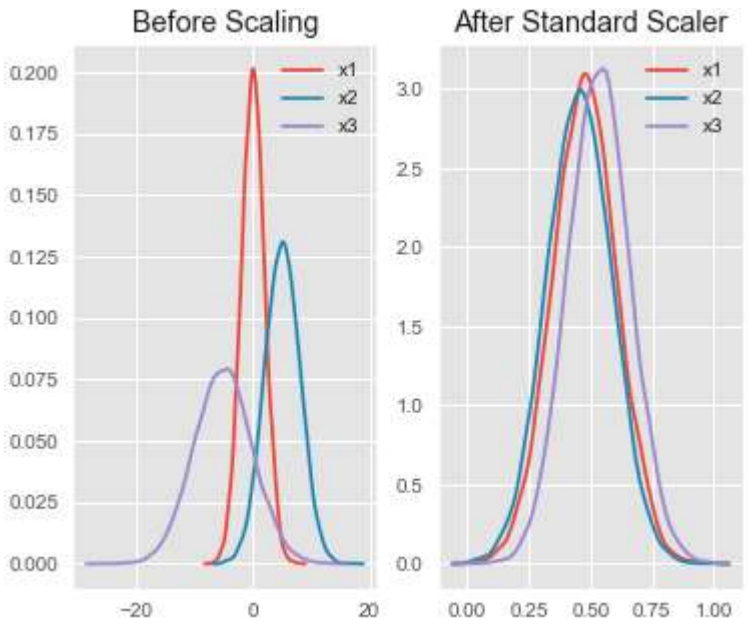
Answer Area

Question	Answer choice
Which scaler is used in graph A?	Standard Scaler Min Max Scale Normalizer
Which scaler is used in graph B?	Standard Scaler Min Max Scale Normalizer
Which scaler is used in graph C?	Standard Scaler Min Max Scale Normalizer

Section: (none)
Explanation

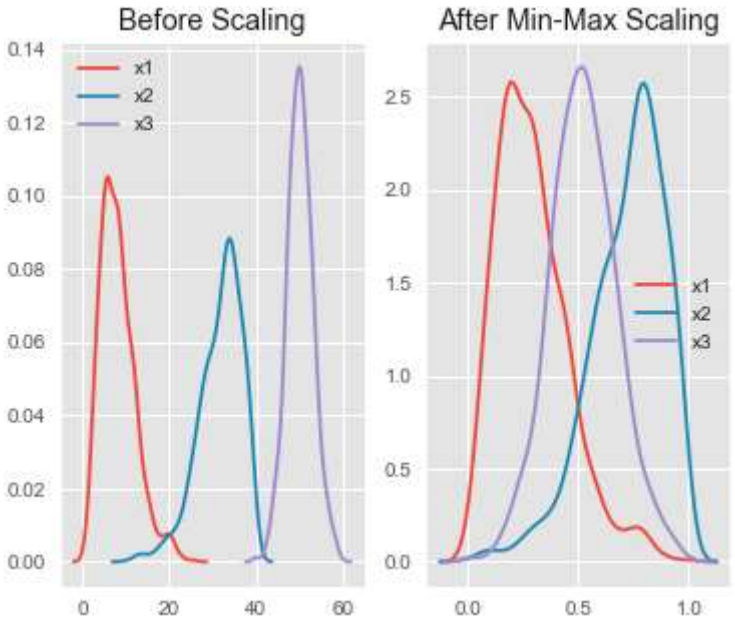
Explanation/Reference:
Explanation:

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/

QUESTION 6
You are determining if two sets of data are significantly different from one another by using Azure Machine Learning Studio.

Estimated values in one set of data may be more than or less than reference values in the other set of data. You must produce a distribution that has a constant Type I error as a function of the correlation.

You need to produce the distribution.

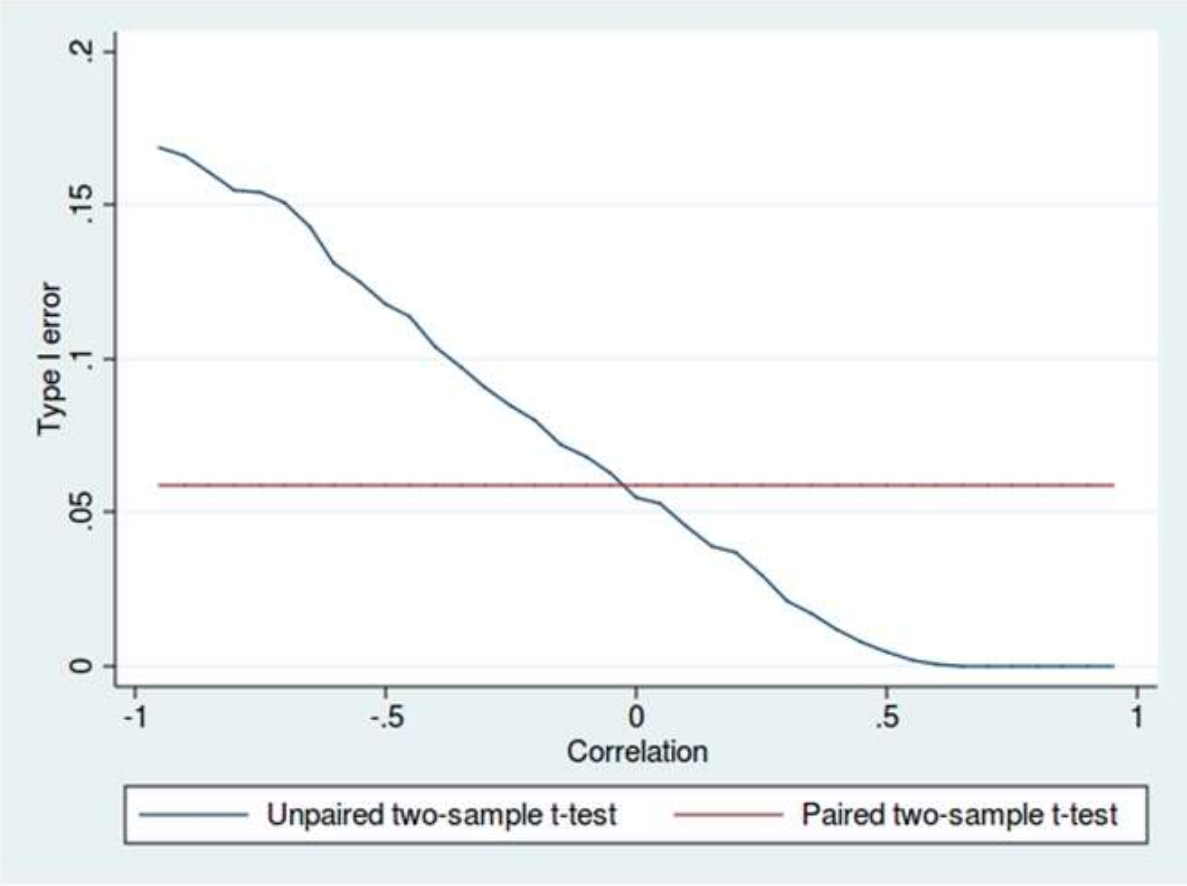
Which type of distribution should you produce?

A. Unpaired t-test with a two-tail option
B. Unpaired t-test with a one-tail option
C. Paired t-test with a one-tail option
D. Paired t-test with a two-tail option

Correct Answer: D
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Choose a one-tail or two-tail test. The default is a two-tailed test. This is the most common type of test, in which the expected distribution is symmetric around zero.

Example: Type I error of unpaired and paired two-sample t-tests as a function of the correlation. The simulated random numbers originate from a bivariate normal distribution with a variance of 1.



Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/test-hypothesis-using-t-test

https://en.wikipedia.org/wiki/Student%27s_t-test

QUESTION 7
DRAG DROP

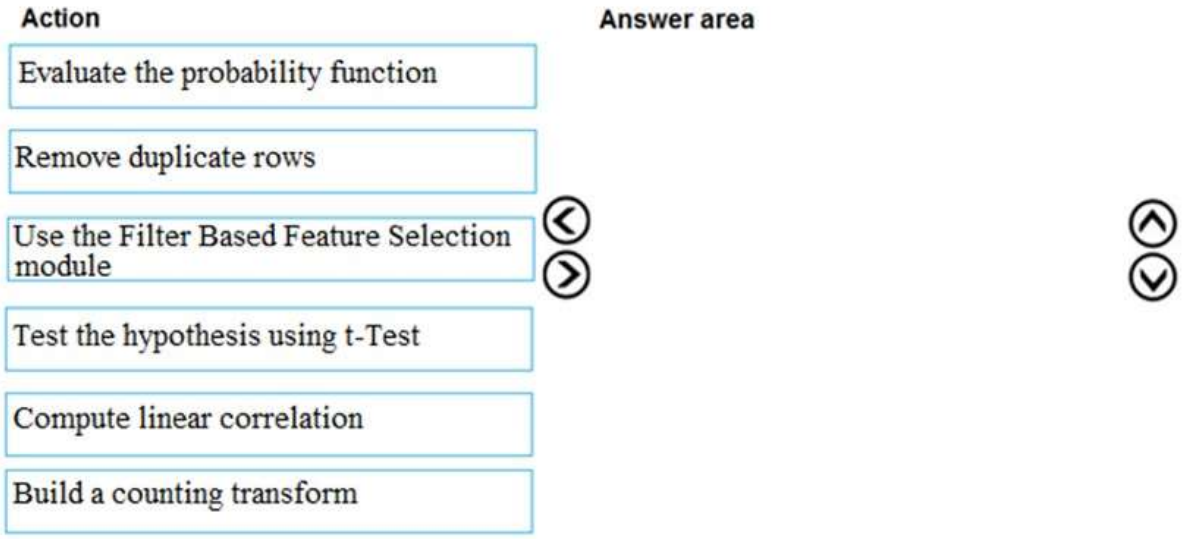
You are producing a multiple linear regression model in Azure Machine Learning Studio.

Several independent variables are highly correlated.

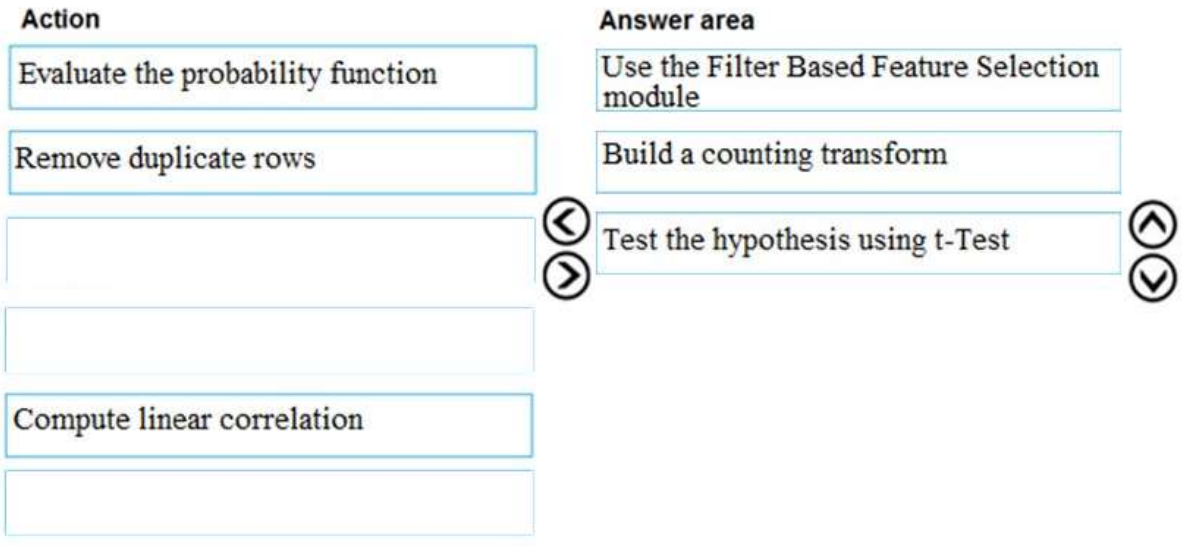
You need to select appropriate methods for conducting effective feature engineering on all the data.

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.

Select and Place:



Correct Answer:



Section: (none)
Explanation

Explanation/Reference:
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

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

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

QUESTION 8
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. Extract N-Gram Features from Text
B. Edit Metadata
C. Preprocess Text
D. Apply SQL Transformation

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Typical metadata changes might include marking columns as features.

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

QUESTION 9
HOTSPOT

You have a multi-class image classification deep learning model that uses a set of labeled photographs. You create the following code to select hyperparameter values when training the model.

```
from azureml.train.hyperdrive import BayesianParameterSampling
param_sampling = BayesianParametersSampling ({
 "learning_rate": uniform(0.01, 0.1),
 "batch_size": choice(16, 32, 64, 128)}
)
```

For each of the following statements, select Yes if the statement is true. Otherwise, select No.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

	Yes	No
Hyperparameter combinations for the runs are selected based on how previous samples performed in the previous experiment run.	○	○
The learning rate value 0.09 might be used during model training.	○	○
You can define an early termination policy for this hyperparameter tuning run.	○	○

Correct Answer:

Answer Area

	Yes	No
Hyperparameter combinations for the runs are selected based on how previous samples performed in the previous experiment run.	○	○
The learning rate value 0.09 might be used during model training.	○	○
You can define an early termination policy for this hyperparameter tuning run.	○	○

Section: (none)
Explanation

Explanation/Reference:
Explanation:

Box 1: Yes
Hyperparameters are adjustable parameters you choose to train a model that govern the training process itself. Azure Machine Learning allows you to automate hyperparameter exploration in an efficient manner, saving you significant time and resources. You specify the range of hyperparameter values and a maximum number of training runs. The system then automatically launches multiple simultaneous runs with different parameter configurations and finds the configuration that results in the best performance, measured by the metric you choose. Poorly performing training runs are automatically early terminated, reducing wastage of compute resources. These resources are instead used to explore other hyperparameter configurations.

Box 2: Yes
uniform(low, high) - Returns a value uniformly distributed between low and high

Box 3: No
Bayesian sampling does not currently support any early termination policy.

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

QUESTION 10
You run an automated machine learning experiment in an Azure Machine Learning workspace. Information about the run is listed in the table below:

Experiment	Run ID	Status	Created on	Duration
auto_ml_clasification	AutoML_1234567890-123	Completed	11/11/2019 11:00:00 AM	00:27:11

You need to write a script that uses the Azure Machine Learning SDK to retrieve the best iteration of the experiment run.

Which Python code segment should you use?

A.
```
from azureml.core import Workspace
from azureml.train.automl.run import AutoMLRun
ws = Workspace.from_config()
automl_ex = ws.experiments.get('auto_ml_classification')
best_iter = automl_ex.archived_time.find('11/11/2019 11:00:00 AM')
```

B.
```
from azureml.core import Workspace
from azureml.train.automl.run import AutoMLRun
automl_ex = ws.experiments.get('auto_ml_classification')
automl_run = AutoMLRun(automl_ex, 'AutoML_1234567890-123')
best_iter = automl_run.current_run
```

C.
```
from azureml.core import Workspace
from azureml.train.automl.run import AutoMLRun
ws = Workspace.from_config()
automl_ex = ws.experiments.get('auto_ml_classification')
best_iter = list(automl_ex.get_runs())[0]
```

D.
```
from azureml.core import Workspace
from azureml.train.automl.run import AutoMLRun
ws = Workspace.from_config()
automl_ex = ws.experiments.get('auto_ml_classification')
automl_run = AutoMLRun(automl_ex, 'AutoML_1234567890-123')
best_iter = automl_run.get_output()[0]
```

E.
```
from azureml.core import Workspace
from azureml.train.automl.run import AutoMLRun
ws = Workspace.from_config()
automl_ex = ws.experiments.get('auto_ml_classification')
best_iter = automl_ex.get_runs('AutoML_1234567890-123')
```

Correct Answer: D
Section: (none)
Explanation

Explanation/Reference:
Explanation:
The get_output method on automl_classifier returns the best run and the fitted model for the last invocation. Overloads on get_output allow you to retrieve the best run and fitted model for any logged metric or for a particular iteration.

In [ ]:

best_run, fitted_model = local_run.get_output()

Reference:
https://notebooks.azure.com/azureml/projects/azureml-getting-started/html/how-to-use-azureml/automated-machine-learning/classification-with-deployment/auto-ml-classification-with-deployment.ipynb

QUESTION 11
You have a comma-separated values (CSV) file containing data from which you want to train a classification model.

You are using the Automated Machine Learning interface in Azure Machine Learning studio to train the classification model. You set the task type to Classification.

You need to ensure that the Automated Machine Learning process evaluates only linear models.

What should you do?

A. Add all algorithms other than linear ones to the blocked algorithms list.
B. Set the Exit criterion option to a metric score threshold.
C. Clear the option to perform automatic featurization.
D. Clear the option to enable deep learning.
E. Set the task type to **Regression**.

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Automatic featurization can fit non-linear models.

Reference:
https://econml.azurewebsites.net/spec/estimation/dml.html

https://docs.microsoft.com/en-us/azure/machine-learning/how-to-use-automated-ml-for-ml-models

QUESTION 12
You are a data scientist working for a bank and have used Azure ML to train and register a machine learning model that predicts whether a customer is likely to repay a loan.

You want to understand how your model is making selections and must be sure that the model does not violate government regulations such as denying loans based on where an applicant lives.

You need to determine the extent to which each feature in the customer data is influencing predictions.

What should you do?

A. Enable data drift monitoring for the model and its training dataset.
B. Score the model against some test data with known label values and use the results to calculate a confusion matrix.
C. Use the Hyperdrive library to test the model with multiple hyperparameter values.
D. Use the interpretability package to generate an explainer for the model.
E. Add tags to the model registration indicating the names of the features in the training dataset.

Correct Answer: D
Section: (none)
Explanation

Explanation/Reference:
Explanation:
When you compute model explanations and visualize them, you're not limited to an existing model explanation for an automated ML model. You can also get an explanation for your model with different test data. The steps in this section show you how to compute and visualize engineered feature importance based on your test data.

Incorrect Answers:
A: In the context of machine learning, data drift is the change in model input data that leads to model performance degradation. It is one of the top reasons where model accuracy degrades over time, thus monitoring data drift helps detect model performance issues.

B: A confusion matrix is used to describe the performance of a classification model. Each row displays the instances of the true, or actual class in your dataset, and each column represents the instances of the class that was predicted by the model.

C: Hyperparameters are adjustable parameters you choose for model training that guide the training process. The HyperDrive package helps you automate choosing these parameters.

Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-machine-learning-interpretability-automl

QUESTION 13
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**.

Correct Answer: ADF
Section: (none)
Explanation

Explanation/Reference:
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.

QUESTION 14
HOTSPOT

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.

```
Feature          Importance
0                1.5627435610083558
2                0.6077689312583112
4                0.5574002432900718
3                0.42858759955671777
1                0.3501361539771977
```

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

```
# Assume required modules are imported

ws = Workspace.from_config()
feature_importances = explanation.[ ▼ ]( workspace = ws,
                                        experiment_name='train_and_explain',
                                        run_id='train_and_explain_12345')
explanation = client.[ ▼ ]()
feature_importances = explanation.[ ▼ ]()

for key, value in feature_importances.items():
            print(key, "\t", value)
```

feature_importances = explanation.	from_run list_model_explanations from_run_id download_model_explanation
explanation = client.	upload_model_explanation list_model_explanations run download_model_explanation
feature_importances = explanation.	explanation explanation_client get_feature_important_dict download_model_explanation

Correct Answer:

Answer Area

```
# Assume required modules are imported

ws = Workspace.from_config()
feature_importances = explanation.[ ▼ ]( workspace = ws,
                                        experiment_name='train_and_explain',
                                        run_id='train_and_explain_12345')
explanation = client.[ ▼ ]()
feature_importances = explanation.[ ▼ ]()

for key, value in feature_importances.items():
            print(key, "\t", value)
```

feature_importances = explanation.	(workspace = ws,
from_run	experiment_name='train_and_explain',
list_model_explanations	run_id='train_and_explain_12345')
from_run_id	
download_model_explanation	

explanation = client.	()
upload_model_explanation	
list_model_explanations	
run	
download_model_explanation	

feature_importances = explanation.	()
explanation	
explanation_client	
get_feature_important_dict	
download_model_explanation	

Section: (none)
Explanation

Explanation/Reference:
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

QUESTION 15
HOTSPOT

You are using the Hyperdrive feature in Azure Machine Learning to train a model.

You configure the Hyperdrive experiment by running the following code:

```
from azureml.train.hyperdrive import RandomParameterSampling
param_sampling = RandomParameterSampling( {
      "learning_rate": normal(10, 3),
      "keep_probability": uniform(0.05, 0.1),
      "batch_size": choice(16, 32, 64, 128)
      "number_of_hidden_layers": choice(range(3,5))
   }
)
```

For each of the following statements, select Yes if the statement is true. Otherwise, select No.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

	Yes	No
By defining sampling in this manner, every possible combination of the parameters will be tested.	○	○
Random values of the learning_rate parameter will be selected from a normal distribution with a mean of 10 and a standard deviation of 3.	○	○
The keep_probability parameter value will always be either **0.05** or **0.1**.	○	○
Random values for the number_of_hidden_layers parameter will be selected from a normal distribution with a mean of 3 and a standard deviation of 5.	○	○

Correct Answer:

Answer Area

	Yes	No
By defining sampling in this manner, every possible combination of the parameters will be tested.	○	○
Random values of the learning_rate parameter will be selected from a normal distribution with a mean of 10 and a standard deviation of 3.	○	○
The keep_probability parameter value will always be either **0.05** or **0.1**.	○	○
Random values for the number_of_hidden_layers parameter will be selected from a normal distribution with a mean of 3 and a standard deviation of 5.	○	○

Section: (none)
Explanation

Explanation/Reference:
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

QUESTION 16
You plan to use automated machine learning to train a regression model. You have data that has features which have missing values, and categorical features with few distinct values.

You need to configure automated machine learning to automatically impute missing values and encode categorical features as part of the training task.

Which parameter and value pair should you use in the AutoMLConfig class?

A. `featurization = 'auto'`
B. `enable_voting_ensemble = True`
C. `task = 'classification'`
D. `exclude_nan_labels = True`
E. `enable_tf = True`

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Featurization str or FeaturizationConfig
Values: 'auto' / 'off' / FeaturizationConfig
Indicator for whether featurization step should be done automatically or not, or whether customized featurization should be used.

Column type is automatically detected. Based on the detected column type preprocessing/featurization is done as follows:

Categorical: Target encoding, one hot encoding, drop high cardinality categories, impute missing values.
Numeric: Impute missing values, cluster distance, weight of evidence.
DateTime: Several features such as day, seconds, minutes, hours etc.

Text: Bag of words, pre-trained Word embedding, text target encoding.

Reference:
https://docs.microsoft.com/en-us/python/api/azureml-train-automl-client/azureml.train.automl.automlconfig.automlconfig

Develop models

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:

		Actual	
		1	**0**
Predicted	**0**	1	2
	1	2	1

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

		Actual	
		1	**0**
Predicted	**0**	1	5
	1	5	1

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

QUESTION 1
DRAG DROP

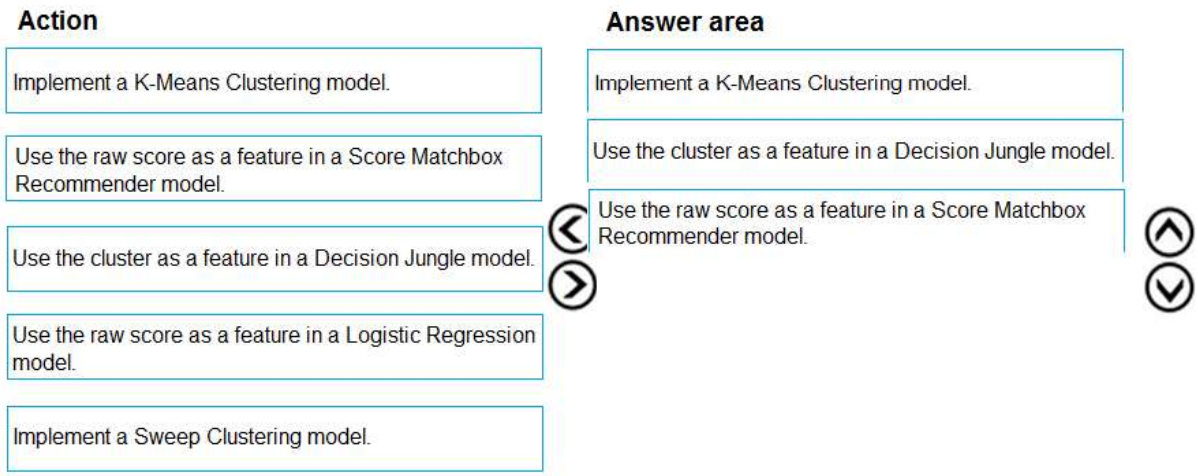
You need to define a modeling strategy for ad response.

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.

Select and Place:

Action	Answer area
Implement a K-Means Clustering model.	
Use the raw score as a feature in a Score Matchbox Recommender model.	
Use the cluster as a feature in a Decision Jungle model.	
Use the raw score as a feature in a Logistic Regression model.	
Implement a Sweep Clustering model.	

Correct Answer:



Section: (none)
Explanation

Explanation/Reference:
Explanation:

Step 1: Implement a K-Means Clustering model

Step 2: Use the cluster as a feature in a Decision jungle model.
Decision jungles are non-parametric models, which can represent non-linear decision boundaries.

Step 3: Use the raw score as a feature in a Score Matchbox Recommender model
The goal of creating a recommendation system is to recommend one or more "items" to "users" of the system. Examples of an item could be a movie, restaurant, book, or song. A user could be a person, group of persons, or other entity with item preferences.

Scenario:
Ad response rated declined.
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.
Ad response models must support non-linear boundaries of features.

References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/multiclass-decision-jungle

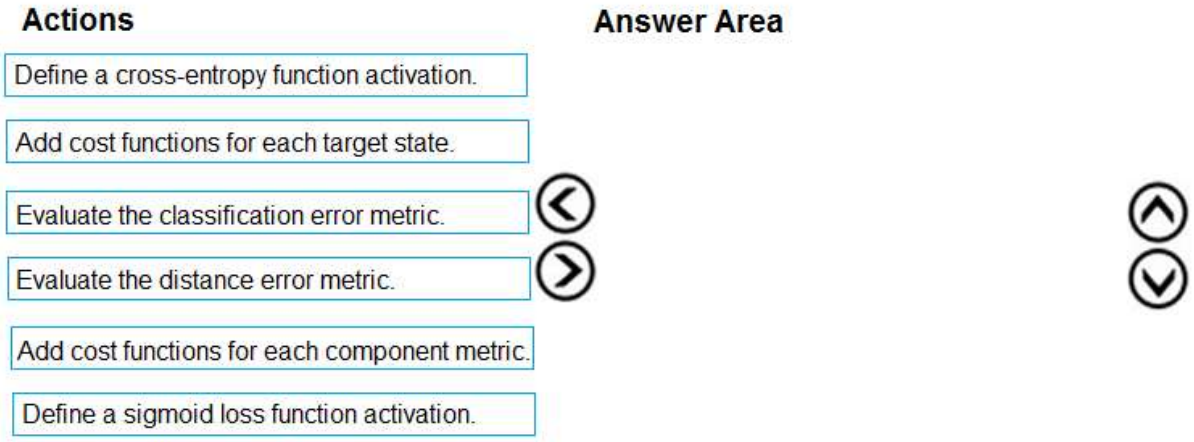
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/score-matchbox-recommender

QUESTION 2
DRAG DROP

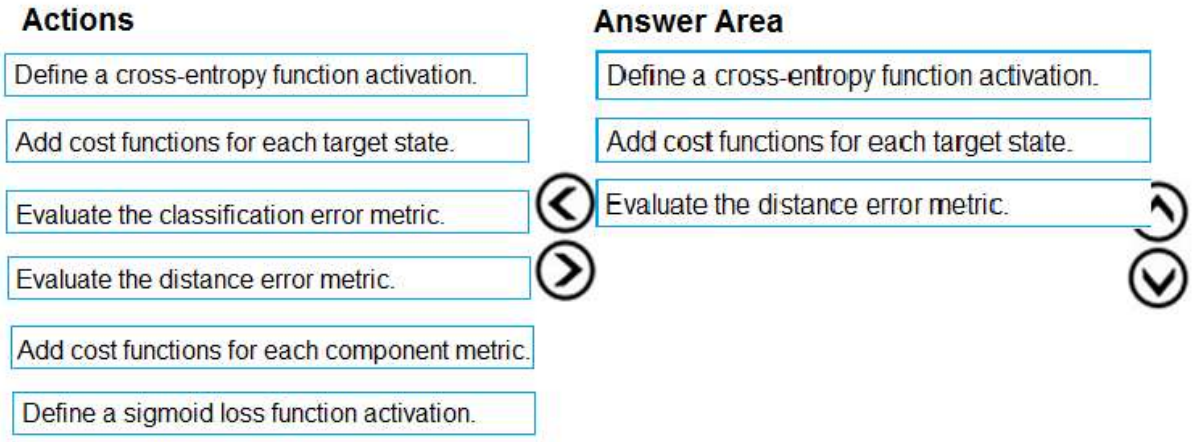
You need to define an evaluation strategy for the crowd sentiment models.

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.

Select and Place:



Correct Answer:



Section: (none)
Explanation

Explanation/Reference:
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/

QUESTION 3
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.

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:
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

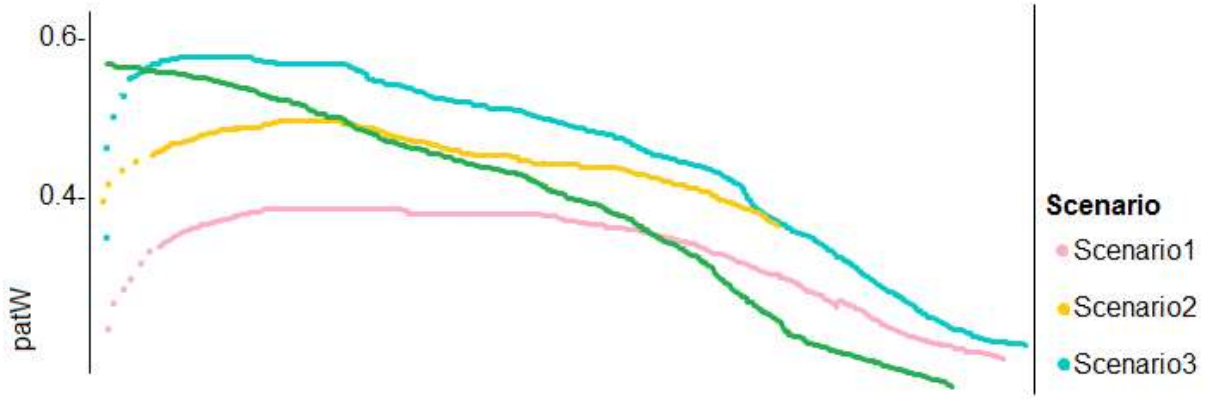
QUESTION 4
You need to implement a new cost factor scenario for the ad response models as illustrated in the performance curve exhibit.

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.

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:
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%.

Develop models

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 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. 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:

Column heading	Description
CapitaCrimeRate	per capita crime rate by town
Zoned	proportion of residential land zoned for lots over 25.000 square feet
NonRetailAcres	proportion of retail business acres per town
NextToRiver	proximity of a property to the river
NitrogenOxideConcentration	nitric oxides concentration (parts per 10 million)
AvgRoomsPerHouse	average number of rooms per dwelling
Age	proportion of owner-occupied units built prior to 1940
DistanceToEmploymentCenter	weighted distances to employment centers
AccessibilityToHighway	index of accessibility to radial highways to a value of two decimal places
Tax	full value property tax rate per $10,000
PupilTeacherRatio	pupil to teacher ratio by town
ProfessionalClass	professional class percentage
LowerStatus	percentage lower status of the population
MedianValue	median value of owner-occupied homes in $1000s

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 use a feature selection algorithm to analyze the relationship between the MedianValue 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 Machine Learning Studio to compare the Two-Class Decision Forest and the Two-Class Decision Jungle modules with one another.

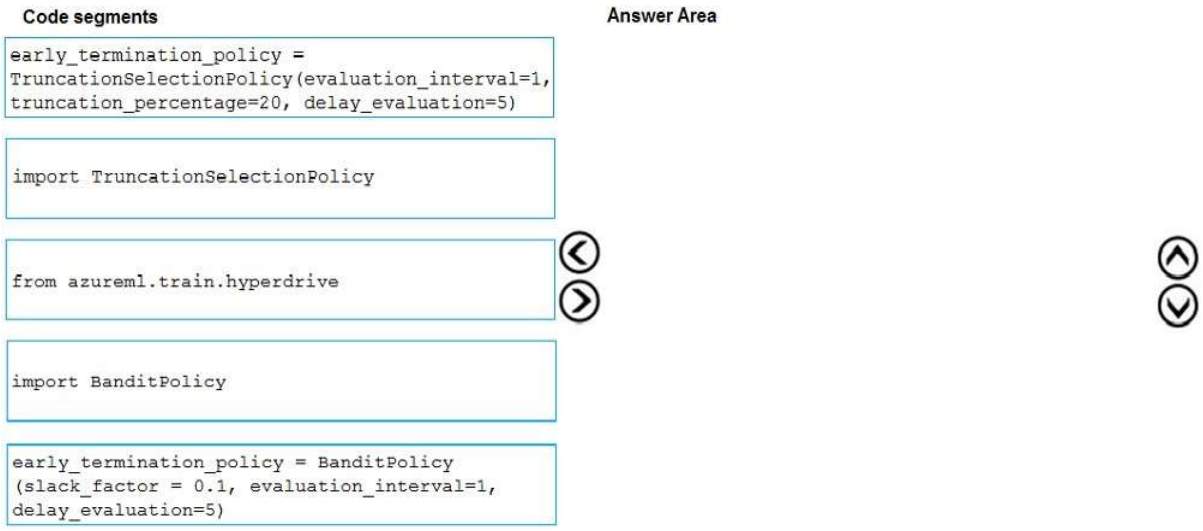
QUESTION 1
DRAG DROP

You need to implement an early stopping criteria policy for model training.

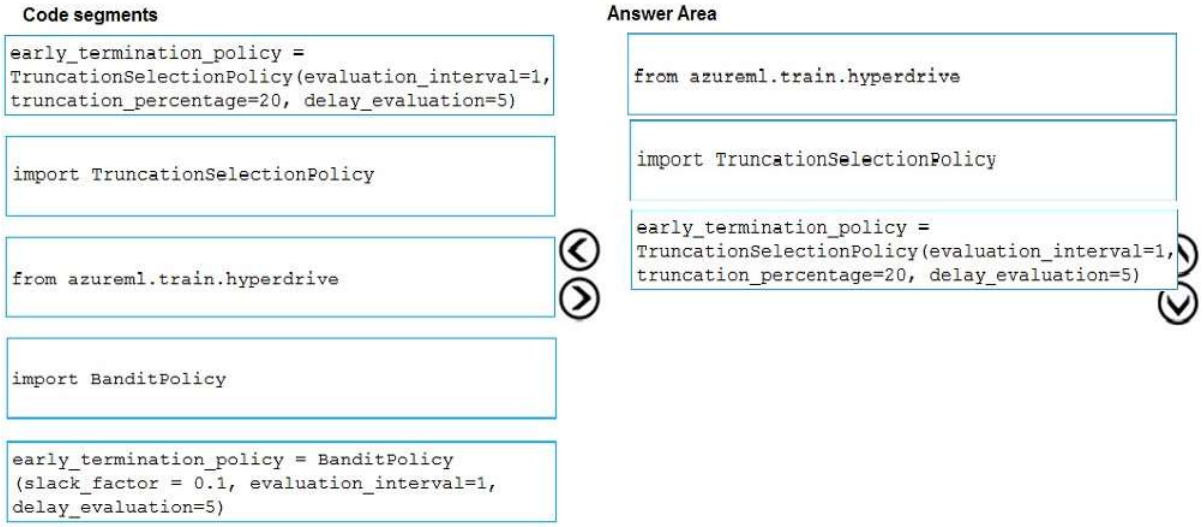
Which three code segments should you use to develop the solution? To answer, move the appropriate code segments from the list of code segments to the answer area and arrange them in the correct order.

NOTE: More than one order of answer choices is correct. You will receive credit for any of the correct orders you select.

Select and Place:



Correct Answer:



Section: (none)
Explanation

Explanation/Reference:
Explanation:

You need to implement an early stopping criterion on models that provides savings without terminating promising jobs.

Truncation selection cancels a given percentage of lowest performing runs at each evaluation interval. Runs are compared based on their performance on the primary metric and the lowest X% are terminated.

Example:
from azureml.train.hyperdrive import TruncationSelectionPolicy
early_termination_policy = TruncationSelectionPolicy(evaluation_interval=1, truncation_percentage=20, delay_evaluation=5)

Incorrect Answers:
Bandit is a termination policy based on slack factor/slack amount and evaluation interval. The policy early terminates any runs where the primary metric is not within the specified slack factor / slack amount with respect to the best performing training run.

Example:
from azureml.train.hyperdrive import BanditPolicy
early_termination_policy = BanditPolicy(slack_factor = 0.1, evaluation_interval=1, delay_evaluation=5

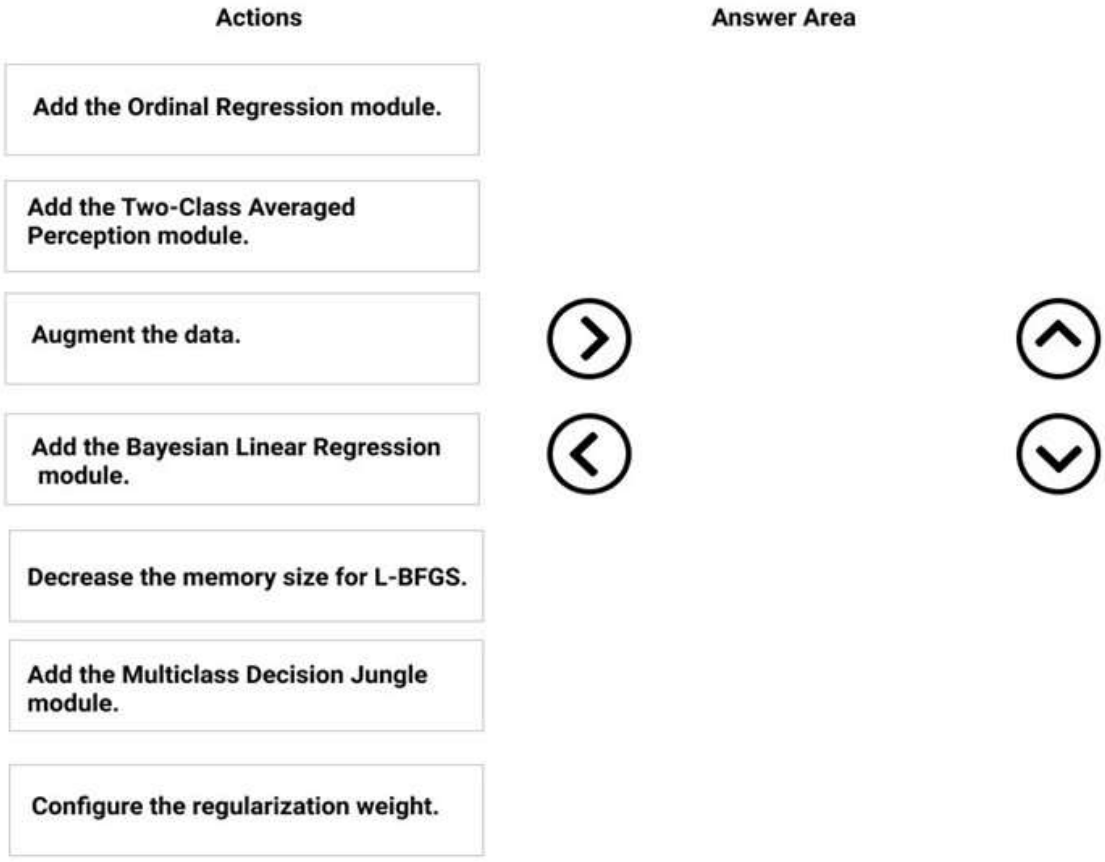
References:
https://docs.microsoft.com/en-us/azure/machine-learning/service/how-to-tune-hyperparameters

QUESTION 2
DRAG DROP

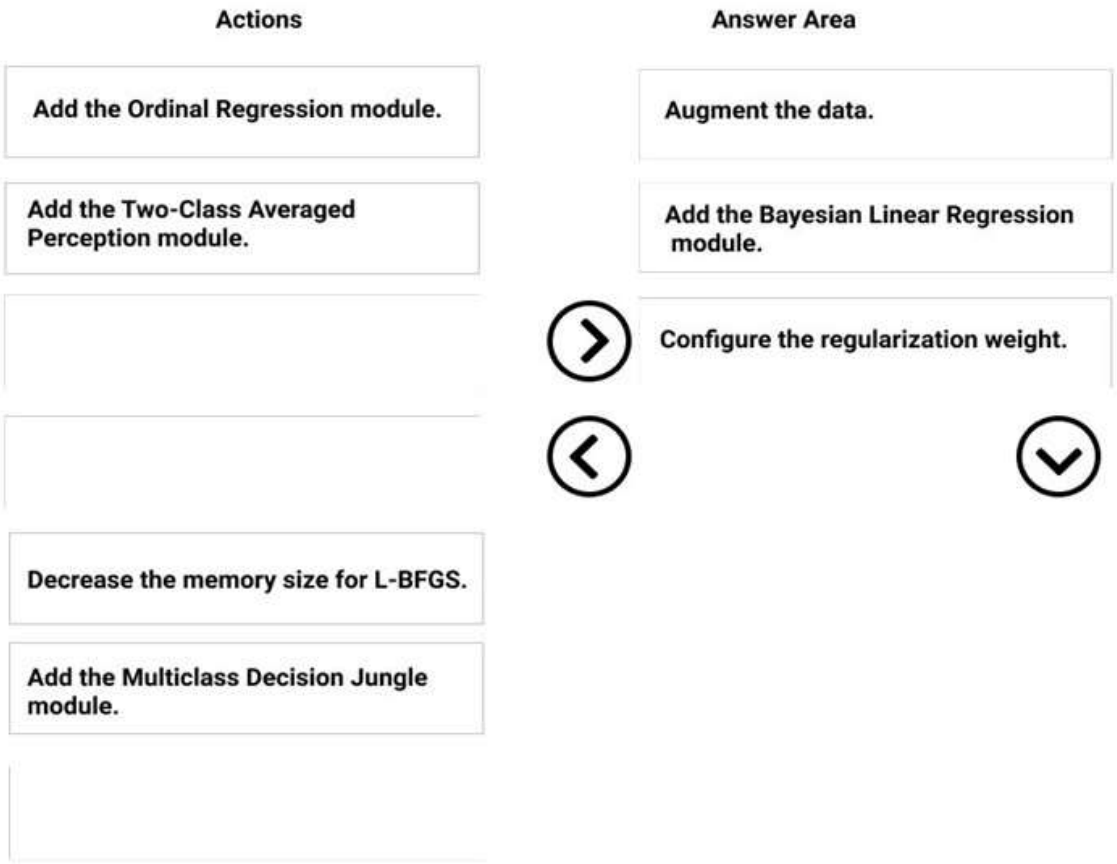
You need to correct the model fit issue.

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.

Select and Place:



Correct Answer:



Section: (none)
Explanation

Explanation/Reference:
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

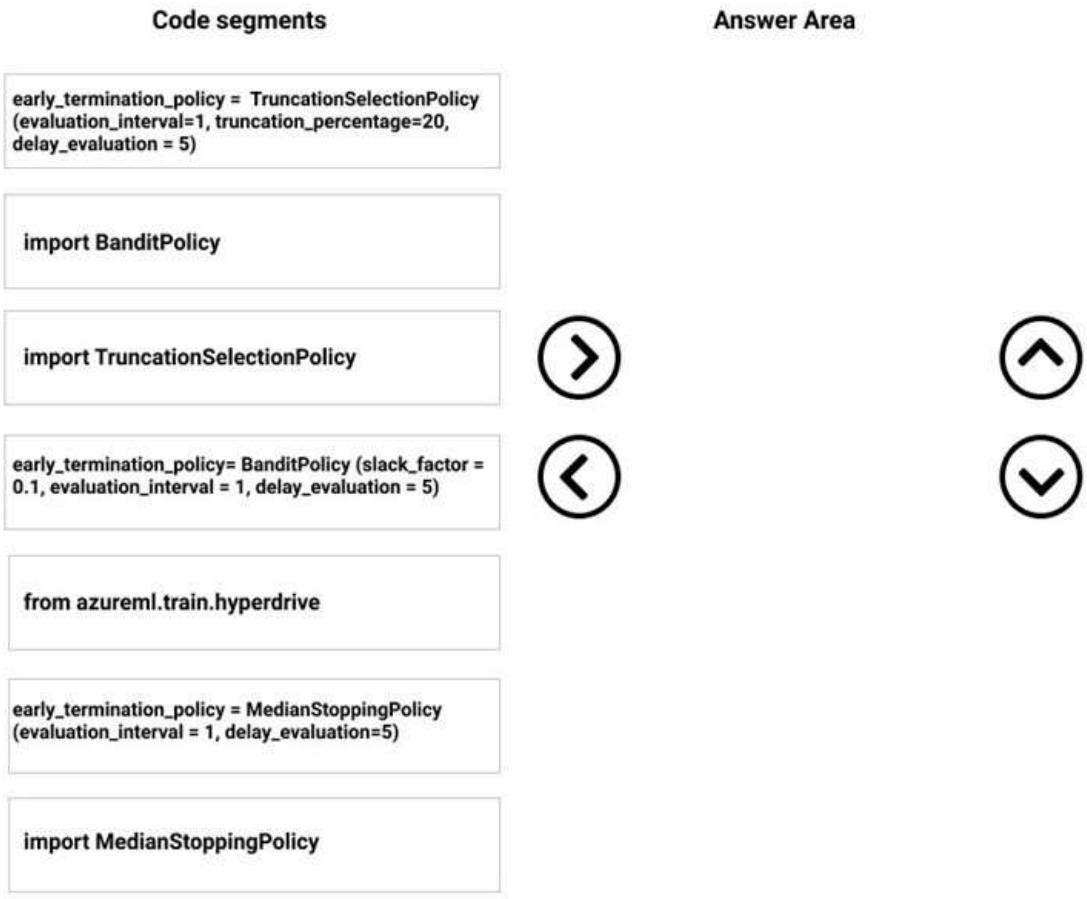
QUESTION 3
DRAG DROP

You need to implement early stopping criteria as stated in the model training requirements.

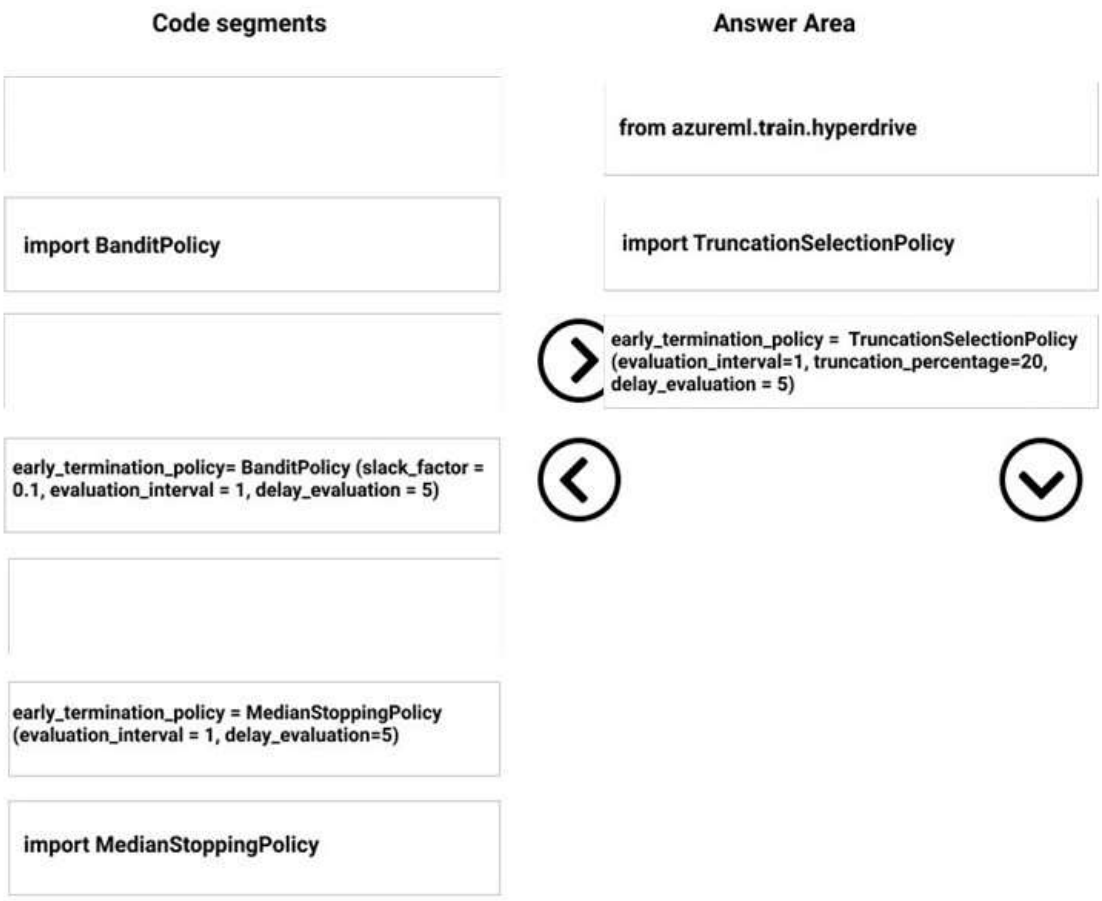
Which three code segments should you use to develop the solution? To answer, move the appropriate code segments from the list of code segments to the answer area and arrange them in the correct order.

NOTE: More than one order of answer choices is correct. You will receive the credit for any of the correct orders you select.

Select and Place:



Correct Answer:



Section: (none)
Explanation

Explanation/Reference:
Explanation:

Step 1: from azureml.train.hyperdrive

Step 2: Import TruncationCelectionPolicy
Truncation selection cancels a given percentage of lowest performing runs at each evaluation interval. Runs are compared based on their performance on the primary metric and the lowest X% are terminated.

Scenario: 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.

Step 3: early_terminiation_policy = TruncationSelectionPolicy..

Example:
from azureml.train.hyperdrive import TruncationSelectionPolicy
early_termination_policy = TruncationSelectionPolicy(evaluation_interval=1, truncation_percentage=20, delay_evaluation=5)
In this example, the early termination policy is applied at every interval starting at evaluation interval 5. A run will be terminated at interval 5 if its performance at interval 5 is in the lowest 20% of performance of all runs at interval 5.

Incorrect Answers:

Median:
Median stopping is an early termination policy based on running averages of primary metrics reported by the runs. This policy computes running averages across all training runs and terminates runs whose performance is worse than the median of the running averages.

Slack:
Bandit is a termination policy based on slack factor/slack amount and evaluation interval. The policy early terminates any runs where the primary metric is not within the specified slack factor / slack amount with respect to the best performing training run.

References:
https://docs.microsoft.com/en-us/azure/machine-learning/service/how-to-tune-hyperparameters

Develop models

Question Set 3

QUESTION 1
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

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:
Explanation:
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

QUESTION 2
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: Accuracy, Precision, Recall, F1 score, and AUC.

Does the solution meet the goal?

A. Yes
B. No

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Those are metrics for evaluating classification models, instead use: Mean Absolute Error, Root Mean Absolute Error, Relative Absolute Error, Relative Squared Error, and the Coefficient of Determination.

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

QUESTION 3
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: Relative Squared Error, Coefficient of Determination, Accuracy, Precision, Recall, F1 score, and AUC.

Does the solution meet the goal?

A. Yes
B. No

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Relative Squared Error, Coefficient of Determination are good metrics to evaluate the linear regression model, but the others are metrics for classification models.

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

QUESTION 4
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. Root Mean Square Error
B. Coefficient of determination
C. Recall
D. Precision
E. Mean absolute error

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:
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:
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.

C: Recall is the fraction of all correct results returned by the model.

D: Precision is the proportion of true results over all positive results.

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

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

QUESTION 5
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
E. Accuracy
F. Root Mean Square Error

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
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-binary-classification-model

QUESTION 6
HOTSPOT

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:

Algorithm	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Relative Squared Error
Bayesian Liner	3.276025	4.655442	0.511436	0.282138
Neural Network	2.676538	3.621476	0.417847	0.17073
Boosted Decision Tree	2.168847	2.878077	0.338589	0.107831
Linear	6.350005	8.720718	0.99133	0.99002
Decision Forest	2.390206	3.315 164	0.373146	0.14307

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.

Hot Area:

Answer Area

Which algorithm minimizes differences between actual and predicted values?

- Bayesian Linear Regression
- Neutral Network Regression
- Boosted Decision Tree Regression
- Linear Regression
- Decision Forest Regression

Which approach should you use to find the best parameters for a Linear Regression model for the Online Gradient Descent method?

- Set the Decrease learning rate option to True.
- Set the Decrease learning rate option to False.
- Set the Create trainer mode option to Parameter Range.
- Increase the number of epochs.
- Decrease the number of epochs.

Correct Answer:

Answer Area

Which algorithm minimizes differences between actual and predicted values?

- Bayesian Linear Regression
- Neutral Network Regression
- **Boosted Decision Tree Regression**
- Linear Regression
- Decision Forest Regression

Which approach should you use to find the best parameters for a Linear Regression model for the Online Gradient Descent method?

- Set the Decrease learning rate option to True.
- Set the Decrease learning rate option to False.
- **Set the Create trainer mode option to Parameter Range.**
- Increase the number of epochs.
- Decrease the number of epochs.

Section: (none)
Explanation

Explanation/Reference:
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

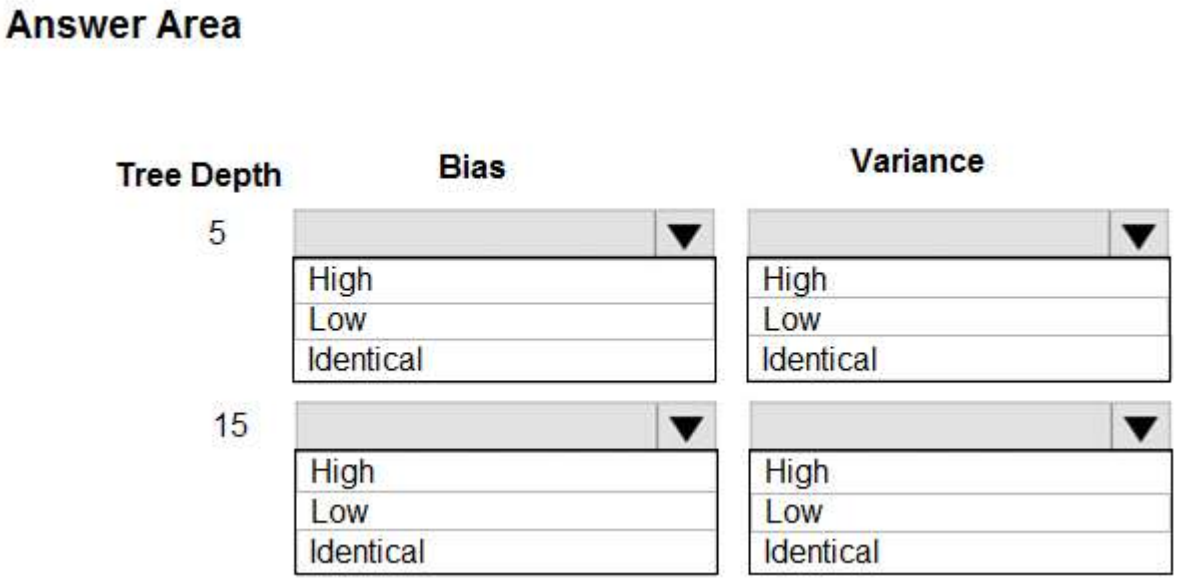
QUESTION 7
HOTSPOT

You are using a decision tree algorithm. You have trained a model that generalizes well at a tree depth equal to 10.

You need to select the bias and variance properties of the model with varying tree depth values.

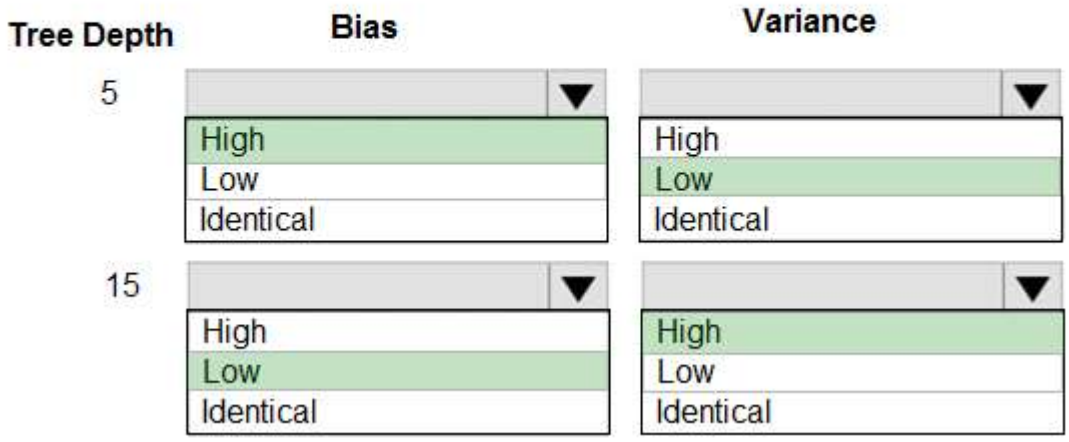
Which properties should you select for each tree depth? To answer, select the appropriate options in the answer area.

Hot Area:



Correct Answer:

Answer Area



Section: (none)
Explanation

Explanation/Reference:
Explanation:

In decision trees, the depth of the tree determines the variance. A complicated decision tree (e.g. deep) has low bias and high variance.

Note: In statistics and machine learning, the bias–variance tradeoff is the property of a set of predictive models whereby models with a lower bias in parameter estimation have a higher variance of the parameter estimates across samples, and vice versa. Increasing the bias will decrease the variance. Increasing the variance will decrease the bias.

References:
https://machinelearningmastery.com/gentle-introduction-to-the-bias-variance-trade-off-in-machine-learning/

QUESTION 8
DRAG DROP

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.

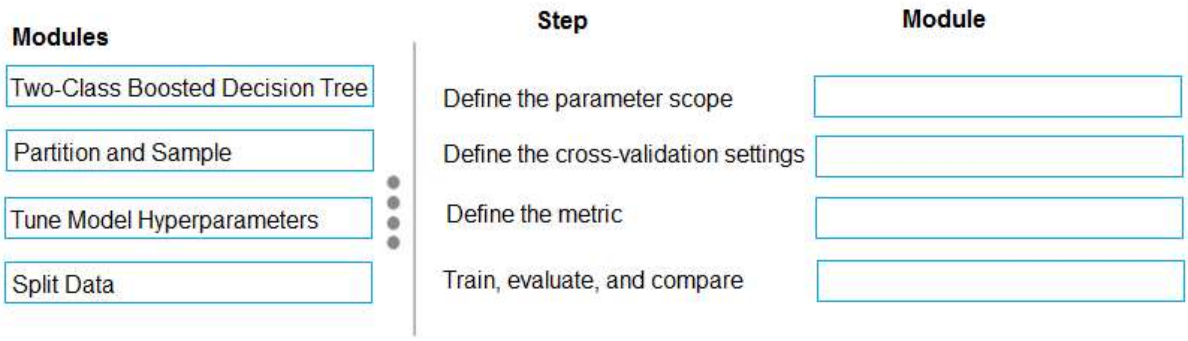
You need to identify a parameter set for the new model using Azure Machine Learning Studio.

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.

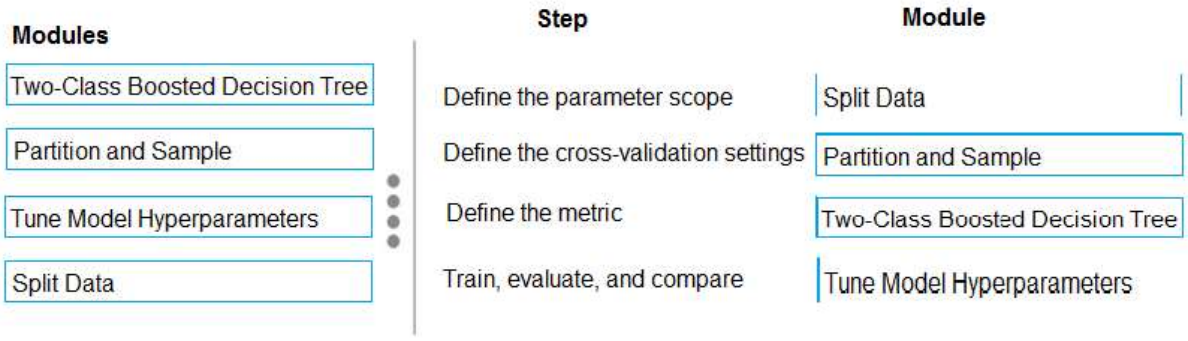
Select and Place:

Answer Area



Correct Answer:

Answer Area



Section: (none)
Explanation

Explanation/Reference:
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

QUESTION 9
HOTSPOT

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:

```
from sklearn.svm import svc
import numpy as np
svc = SVC(kernel= 'linear', class_weight= 'balanced', C-1.0, random_state-0)
model1 = svc.fit(X_train, y)
```

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.

Hot Area:

Answer Area

Code Segment	Evaluation Statement
class_weight=balanced	Automatically select the performance metrics for the classification. Automatically adjust weights directly proportional to class frequencies in the input data. Automatically adjust weights inversely proportional to class frequences in the input data.
C parameter	Penalty parameter Degreee of polynomial kernel function Size of the kernel cache

Correct Answer:

Answer Area

Code Segment	Evaluation Statement
class_weight=balanced	Automatically select the performance metrics for the classification. Automatically adjust weights directly proportional to class frequencies in the input data. Automatically adjust weights inversely proportional to class frequences in the input data.
C parameter	Penalty parameter Degreee of polynomial kernel function Size of the kernel cache

Section: (none)
Explanation

Explanation/Reference:
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

QUESTION 10
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)

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:
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

QUESTION 11
You create a binary classification model.

You need to evaluate the model performance.

Which two metrics can you use? Each correct answer presents a complete solution.

NOTE: Each correct selection is worth one point.

A. relative absolute error
B. precision
C. accuracy
D. mean absolute error
E. coefficient of determination

Correct Answer: BC
Section: (none)

Explanation

Explanation/Reference:
Explanation:
The evaluation metrics available for binary classification models are: Accuracy, Precision, Recall, F1 Score, and AUC.

Note: A very natural question is: 'Out of the individuals whom the model, how many were classified correctly (TP)?'
This question can be answered by looking at the Precision of the model, which is the proportion of positives that are classified correctly.

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

QUESTION 12
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 configure the Tune Model Hyperparameters module.

Which two values 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

Correct Answer: DE
Section: (none)
Explanation

Explanation/Reference:
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

QUESTION 13
HOTSPOT

You are evaluating a Python NumPy array that contains six data points defined as follows:

```
data = [10, 20, 30, 40, 50, 60]
```

You must generate the following output by using the k-fold algorithm implantation in the Python Scikit-learn machine learning library:

```
train: [10 40 50 60], test: [20 30]
train: [20 30 40 60], test: [10 50]
```

```
train: [10 20 30 50], test: [40 60]
```

You need to implement a cross-validation to generate the output.

How should you complete the code segment? To answer, select the appropriate code segment in the dialog box in the answer area.

NOTE: Each correct selection is worth one point.

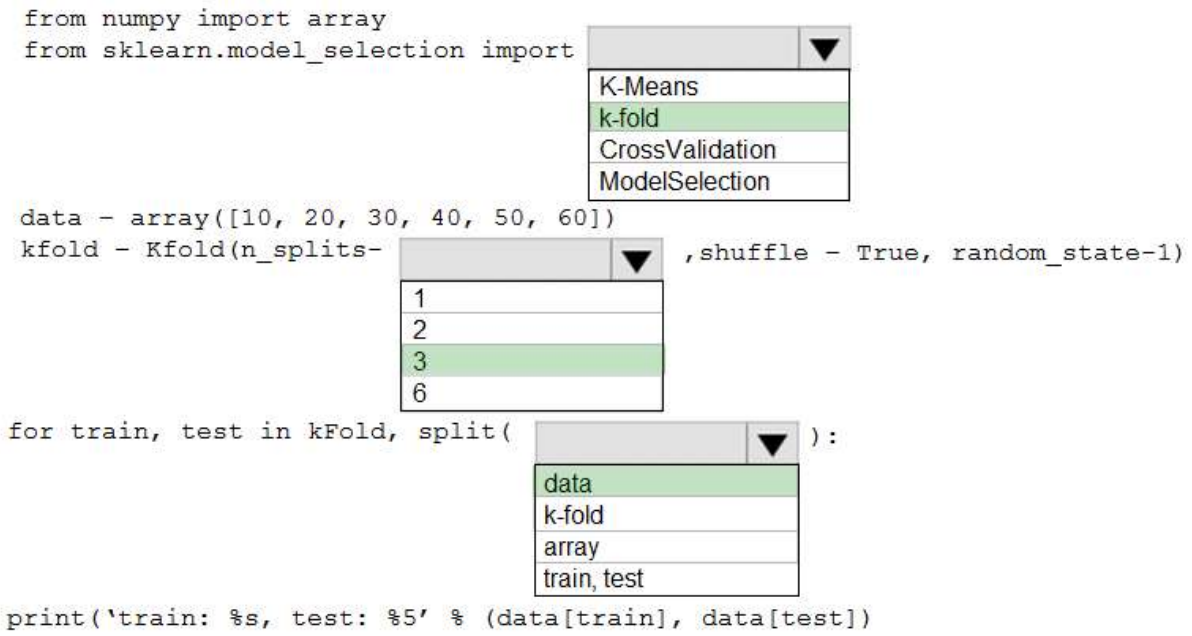
Hot Area:

Answer Area

```
from numpy import array
from sklearn.model_selection import [ K-Means | k-fold | CrossValidation | ModelSelection ]
data = array([10, 20, 30, 40, 50, 60])
kfold = Kfold(n_splits= [ 1 | 2 | 3 | 6 ] ,shuffle = True, random_state=1)
for train, test in kFold, split( [ data | k-fold | array | train, test ] ):
print('train: %s, test: %5' % (data[train], data[test])
```

Correct Answer:

Answer Area



Section: (none)
Explanation

Explanation/Reference:
Explanation:

Box 1: k-fold

Box 2: 3
K-Folds cross-validator provides train/test indices to split data in train/test sets. Split dataset into k consecutive folds (without shuffling by default).
The parameter n_splits (int, default=3) is the number of folds. Must be at least 2.

Box 3: data

Example: Example:

```
>>>
>>> from sklearn.model_selection import KFold
>>> X = np.array([[1, 2], [3, 4], [1, 2], [3, 4]])
>>> y = np.array([1, 2, 3, 4])
>>> kf = KFold(n_splits=2)
>>> kf.get_n_splits(X)
2
>>> print(kf)
KFold(n_splits=2, random_state=None, shuffle=False)
>>> for train_index, test_index in kf.split(X):
...    print("TRAIN:", train_index, "TEST:", test_index)
...    X_train, X_test = X[train_index], X[test_index]
...    y_train, y_test = y[train_index], y[test_index]
TRAIN: [2 3] TEST: [0 1]
TRAIN: [0 1] TEST: [2 3]
```

References:

https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html

QUESTION 14
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

Correct Answer: D
Section: (none)
Explanation

Explanation/Reference:
Explanation:
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.

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

QUESTION 15
You are building a recurrent neural network to perform a binary classification.

You review the training loss, validation loss, training accuracy, and validation accuracy for each training epoch.

You need to analyze model performance.

You need to identify whether the classification model is overfitted.

Which of the following is correct?

A. 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.

B. The training loss decreases while the validation loss increases when training the model.
C. The training loss stays constant and the validation loss decreases when training the model.
D. The training loss increases while the validation loss decreases when training the model.

Correct Answer: B
Section: (none)
Explanation

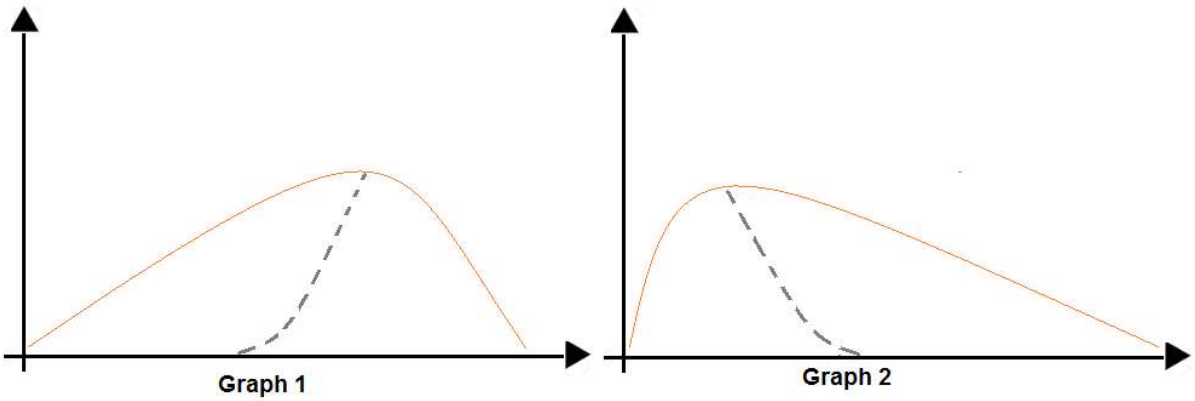
Explanation/Reference:
Explanation:
An overfit model is one where performance on the train set is good and continues to improve, whereas performance on the validation set improves to a point and then begins to degrade.

References:
https://machinelearningmastery.com/diagnose-overfitting-underfitting-lstm-models/

QUESTION 16
HOTSPOT

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.

Hot Area:

Answer Area

Question	Answer choice
Which type of distribution is shown for the dataset density curve of Graph 1?	Negative skew Positive skew Normal distribution Bimodal distribution
Which type of distribution is shown for the dataset density curve of Graph 2?	Negative skew Positive skew Normal distribution Bimodal distribution

Correct Answer:

Answer Area

Question	Answer choice
Which type of distribution is shown for the dataset density curve of Graph 1?	Negative skew **Positive skew** Normal distribution Bimodal distribution
Which type of distribution is shown for the dataset density curve of Graph 2?	**Negative skew** Positive skew Normal distribution Bimodal distribution

Section: (none)
Explanation

Explanation/Reference:
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

QUESTION 17
You are performing clustering by using the K-means algorithm.

You need to define the possible termination conditions.

Which three conditions can you use? Each correct answer presents a complete solution.

NOTE: Each correct selection is worth one point.

A. Centroids do not change between iterations.
B. The residual sum of squares (RSS) rises above a threshold.
C. The residual sum of squares (RSS) falls below a threshold.
D. A fixed number of iterations is executed.
E. The sum of distances between centroids reaches a maximum.

Correct Answer: ACD
Section: (none)
Explanation

Explanation/Reference:
Explanation:
AD: The algorithm terminates when the centroids stabilize or when a specified number of iterations are completed.

C: A measure of how well the centroids represent the members of their clusters is the residual sum of squares or RSS, the squared distance of each vector from its centroid summed over all vectors. RSS is the objective function and our goal is to minimize it.

References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/k-means-clustering

https://nlp.stanford.edu/IR-book/html/htmledition/k-means-1.html

QUESTION 18
You are a data scientist building a deep convolutional neural network (CNN) for image classification.

The CNN model you build 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. Add an additional dense layer with 512 input units.
B. Add L1/L2 regularization.
C. Use training data augmentation.
D. Reduce the amount of training data.
E. Add an additional dense layer with 64 input units.

Correct Answer: BD
Section: (none)
Explanation

Explanation/Reference:
Explanation:
B: Weight regularization provides an approach to reduce the overfitting of a deep learning neural network model on the training data and improve the performance of the model on new data, such as the holdout test set.
Keras provides a weight regularization API that allows you to add a penalty for weight size to the loss function.

Three different regularizer instances are provided; they are:

- L1: Sum of the absolute weights.
- L2: Sum of the squared weights.
- L1L2: Sum of the absolute and the squared weights.

D: Because a fully connected layer occupies most of the parameters, it is prone to overfitting. One method to reduce overfitting is dropout. At each training stage, individual nodes are either "dropped out" of the net with probability 1-p or kept with probability p, so that a reduced network is left; incoming and outgoing edges to a dropped-out node are also removed.
By avoiding training all nodes on all training data, dropout decreases overfitting.

References:
https://machinelearningmastery.com/how-to-reduce-overfitting-in-deep-learning-with-weight-regularization/

https://en.wikipedia.org/wiki/Convolutional_neural_network

QUESTION 19
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. Recommender Split
B. Regular Expression Split
C. Relative Expression Split
D. Split Rows with the Randomized split parameter set to true

Correct Answer: D
Section: (none)
Explanation

Explanation/Reference:
Explanation:
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.

Incorrect Answers:
B: Regular Expression Split: Choose this option when you want to divide your dataset by testing a single column for a value.
C: Relative Expression Split: Use this option whenever you want to apply a condition to a number column.

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

QUESTION 20
HOTSPOT

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.

Hot Area:

Answer Area

Parameter	Value
Splitting mode	Split rows Recommender Split Regular Expression Split Relative Expression Split
Fraction of rows in the first output dataset	0.75 0.25 0.5 1
Randomized split	True False
Stratified split	True False

Correct Answer:

Answer Area

Parameter	Value
Splitting mode	Split rows Recommender Split Regular Expression Split Relative Expression Split
Fraction of rows in the first output dataset	0.75 0.25 0.5 1
Randomized split	True False
Stratified split	True False

Section: (none)
Explanation

Explanation/Reference:
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

QUESTION 21
HOTSPOT

You are tuning a hyperparameter for an algorithm. The following table shows a data set with different hyperparameter, training error, and validation errors.

Hyperparameter (H)	Training error (TE)	Validation error (VE)
1	105	95
2	200	85
3	250	100
4	105	100
5	400	50

Use the drop-down menus to select the answer choice that answers each question based on the information presented in the graphic.

Hot Area:

Answer Area

Question	Answer Choise
Which H value should you select based on the data?	1 2 3 4 5
What H value displays the poorest training result?	1 2 3 4 5

Correct Answer:

Answer Area

Question	Answer Choise
Which H value should you select based on the data?	1 2 3 **4** 5
What H value displays the poorest training result?	1 2 3 4 **5**

Section: (none)
Explanation

Explanation/Reference:
Explanation:

Box 1: 4
Choose the one which has lower training and validation error and also the closest match.
Minimize variance (difference between validation error and train error).

Box 2: 5
Minimize variance (difference between validation error and train error).

Reference:
https://medium.com/comet-ml/organizing-machine-learning-projects-project-management-guidelines-2d2b85651bbd

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 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

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:
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

QUESTION 23
HOTSPOT

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.

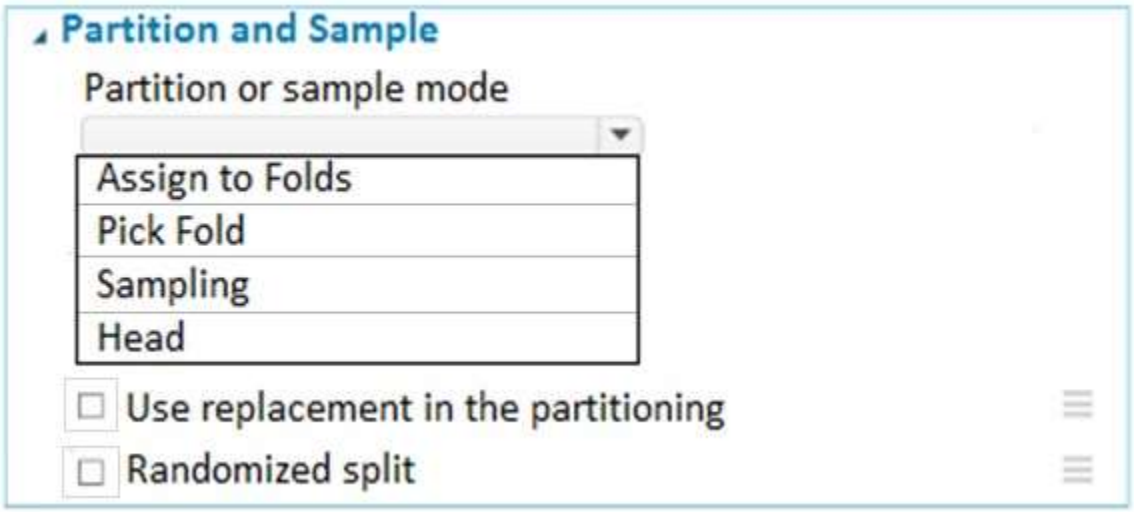
You need to configure the module. You must meet the following requirements:

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

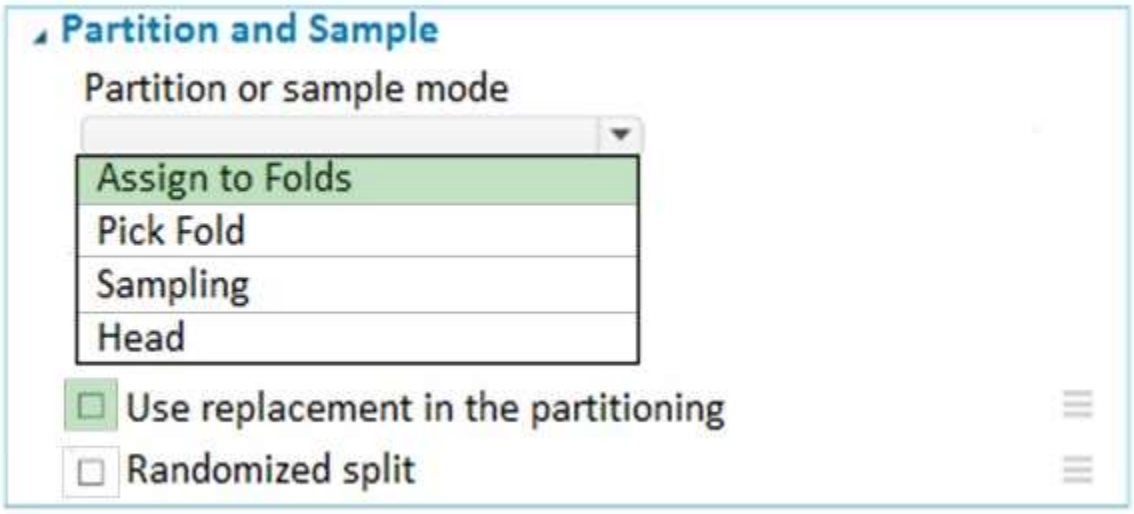
How should you configure the module? To answer, select the appropriate options in the dialog box in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:



Correct Answer:



Section: (none)
Explanation

Explanation/Reference:
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.

1. Add the Partition and Sample module to your experiment in Studio (classic), and connect the dataset.
2. For Partition or sample mode, select Assign to Folds.
3. 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.
4. 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.
5. 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.

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

QUESTION 24
You are building a binary classification model by using a supplied training set.

The training set is imbalanced between two classes.

You need to resolve the data imbalance.

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. Penalize the classification
B. Resample the dataset using undersampling or oversampling
C. Normalize the training feature set
D. Generate synthetic samples in the minority class
E. Use accuracy as the evaluation metric of the model

Correct Answer: ABD
Section: (none)
Explanation

Explanation/Reference:
Explanation:
A: Try Penalized Models
You can use the same algorithms but give them a different perspective on the problem.
Penalized classification imposes an additional cost on the model for making classification mistakes on the minority class during training. These penalties can bias the model to pay more attention to the minority class.

B: You can change the dataset that you use to build your predictive model to have more balanced data.
This change is called sampling your dataset and there are two main methods that you can use to even-up the classes:

- Consider testing under-sampling when you have an a lot data (tens- or hundreds of thousands of instances or more)
- Consider testing over-sampling when you don't have a lot of data (tens of thousands of records or less)

D: Try Generate Synthetic Samples
A simple way to generate synthetic samples is to randomly sample the attributes from instances in the minority class.

Reference:
https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/

QUESTION 25
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:

```
datastore = ws.get_default_datastore()
data_output = pd.read_csv("traindata.csv")
process_step = PythonScriptStep(script_name="process.py",
 arguments=["--data_for_train", data_output],
 outputs=[data_output],compute_target=aml_compute,
 source_directory=process_directory)
train_step = PythonScriptStep(script_name="train.py",
 arguments=["--data_for_train", data_output],
 inputs=[data_output],compute_target=aml_compute,
 source_directory=train_directory)
pipeline = Pipeline(workspace=ws, steps=[process_step, train_step])
```

Does the solution meet the goal?

A. Yes
B. No

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:
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=azure-ml-py

QUESTION 26
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:

```
datastore = ws.get_default_datastore()
data_output = PipelineData("processed_data", datastore=datastore)
process_step = PythonScriptStep(script_name="process.py",
 arguments=["--data_for_train", data_output],
 outputs=[data_output], compute_target=aml_compute,
 source_directory=process_directory)
pipeline = Pipeline(workspace=ws, steps=[process_step])
```

Does the solution meet the goal?

A. Yes
B. No

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:
train_step is missing.

Reference:
https://docs.microsoft.com/en-us/python/api/azureml-pipeline-core/azureml.pipeline.core.pipelinedata?view=azure-ml-py

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 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:

```
datastore = ws.get_default_datastore()
data_input = PipelineData("raw_data", datastore=rawdatastore)
data_output = PipelineData("processed_data", datastore=datastore)
process_step = PythonScriptStep(script_name="process.py",
 arguments=["--data_for_train", data_input],
 outputs=[data_output], compute_target=aml_compute,
 source_directory=process_directory)
train_step = PythonScriptStep(script_name="train.py",
 arguments=["--data_for_train", data_input], inputs=[data_output],
 compute_target=aml_compute, source_directory=train_directory)
pipeline = Pipeline(workspace=ws, steps=[process_step, train_step])
```

Does the solution meet the goal?

A. Yes
B. No

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
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=azure-ml-py

QUESTION 28
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:

```
from azureml.train.sklearn import SKLearn
sk_est = SKLearn(source_directory='./scripts',
 compute_target=aml-compute,
 entry_script='train.py')
```

Does the solution meet the goal?

A. Yes
B. No

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:
Explanation:
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

QUESTION 29
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:

```
from azureml.train.dnn import TensorFlow
sk_est = TensorFlow(source_directory='./scripts',
 compute_target=aml-compute,
 entry_script='train.py')
```

Does the solution meet the goal?

A. Yes
B. No

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:
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

QUESTION 30
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:

```
from azureml.train.estimator import Estimator
sk_est = Estimator(source_directory='./scripts',
 compute_target=aml-compute,
 entry_script='train.py',
 conda_packages=['scikit-learn'])
```

Does the solution meet the goal?

A. Yes
B. No

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:
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

QUESTION 31
DRAG DROP

You create machine learning models by using Azure Machine Learning.

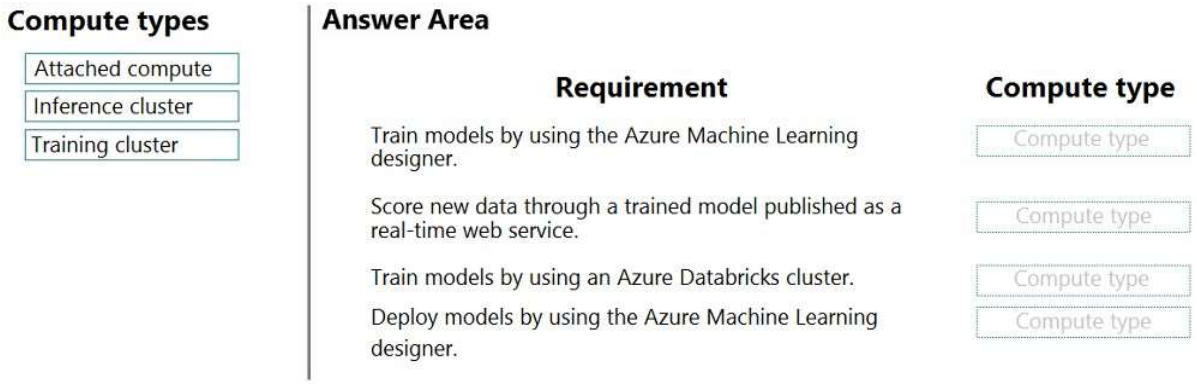
You plan to train and score models by using a variety of compute contexts. You also plan to create a new compute resource in Azure Machine Learning studio.

You need to select the appropriate compute types.

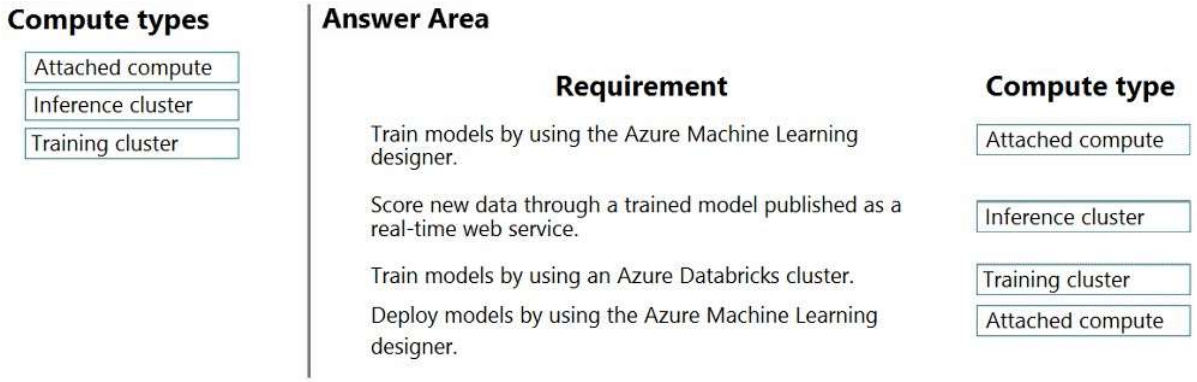
Which compute types should you select? To answer, drag the appropriate compute types to the correct requirements. Each compute type may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.

NOTE: Each correct selection is worth one point.

Select and Place:



Correct Answer:



Section: (none)
Explanation

Explanation/Reference:
Explanation:

Box 1: Attached compute

Training targets	Automated ML	ML pipelines	Azure Machine Learning designer
Local computer	yes		
Azure Machine Learning compute cluster	yes & hyperparameter tuning	yes	yes
Azure Machine Learning compute instance	yes & hyperparameter tuning	yes	yes

Box 2: Inference cluster

Box 3: Training cluster

Box 4: Attached compute

QUESTION 32
DRAG DROP

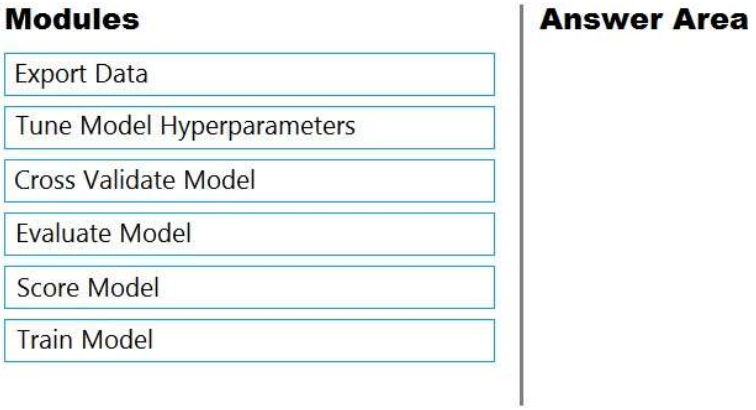
You are building an experiment using the Azure Machine Learning designer.

You split a dataset into training and testing sets. You select the Two-Class Boosted Decision Tree as the algorithm.

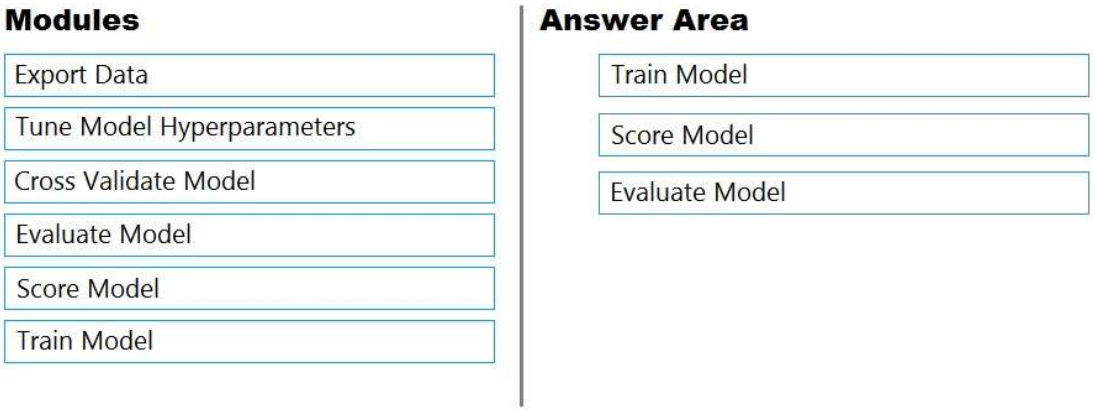
You need to determine the Area Under the Curve (AUC) of the model.

Which three 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.

Select and Place:



Correct Answer:



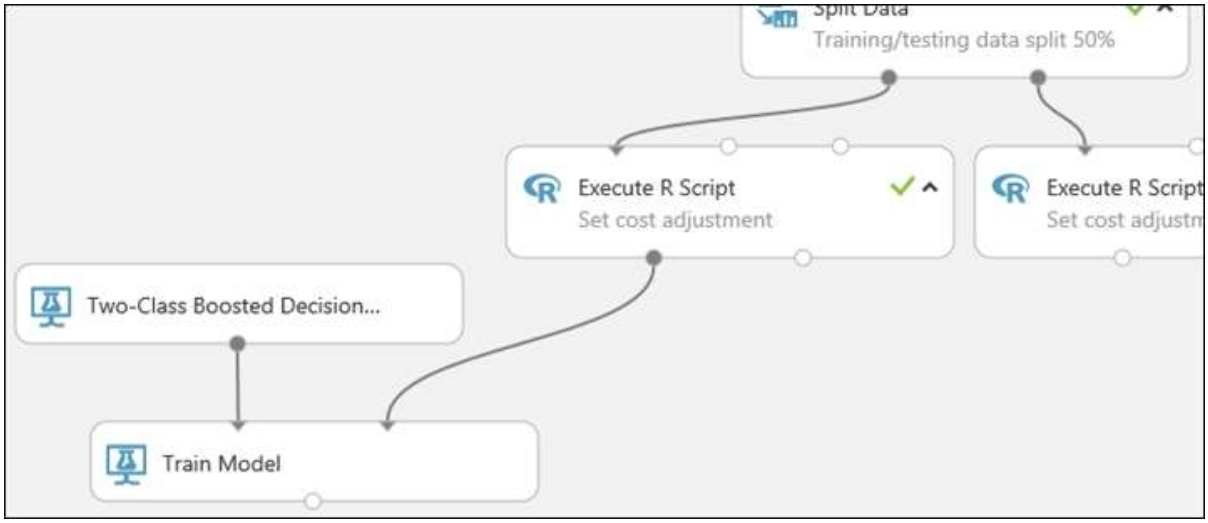
Section: (none)
Explanation

Explanation/Reference:
Explanation:

Step 1: Train Model
Two-Class Boosted Decision Tree
First, set up the boosted decision tree model.

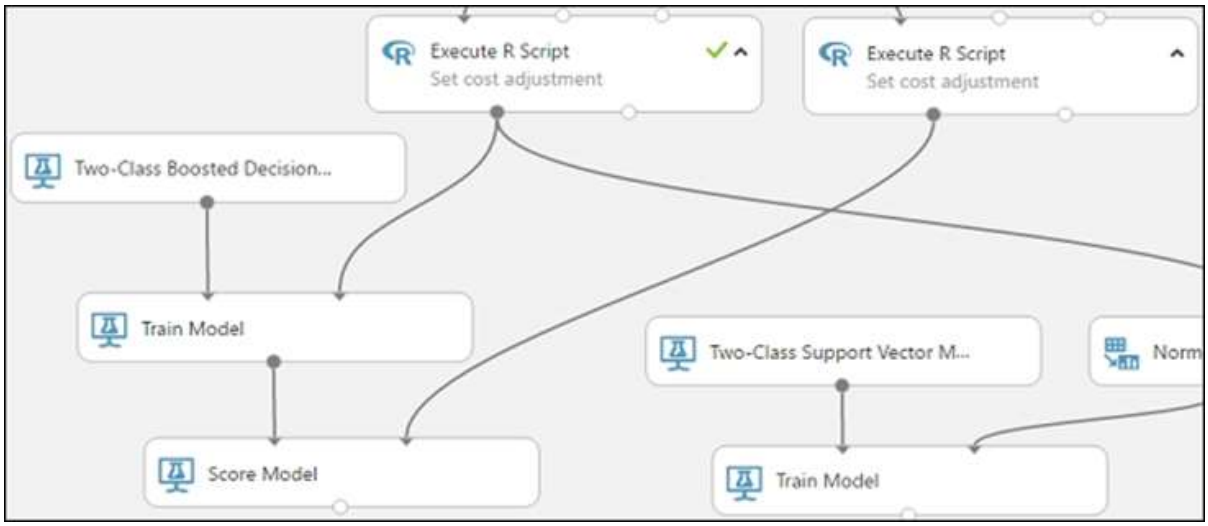
1. Find the Two-Class Boosted Decision Tree module in the module palette and drag it onto the canvas.
2. Find the Train Model module, drag it onto the canvas, and then connect the output of the Two-Class Boosted Decision Tree module to the left input port of the Train Model module.
   The Two-Class Boosted Decision Tree module initializes the generic model, and Train Model uses training data to train the model.
3. Connect the left output of the left Execute R Script module to the right input port of the Train Model module (in this tutorial you used the data coming from the left side of the Split Data module for training).

This portion of the experiment now looks something like this:



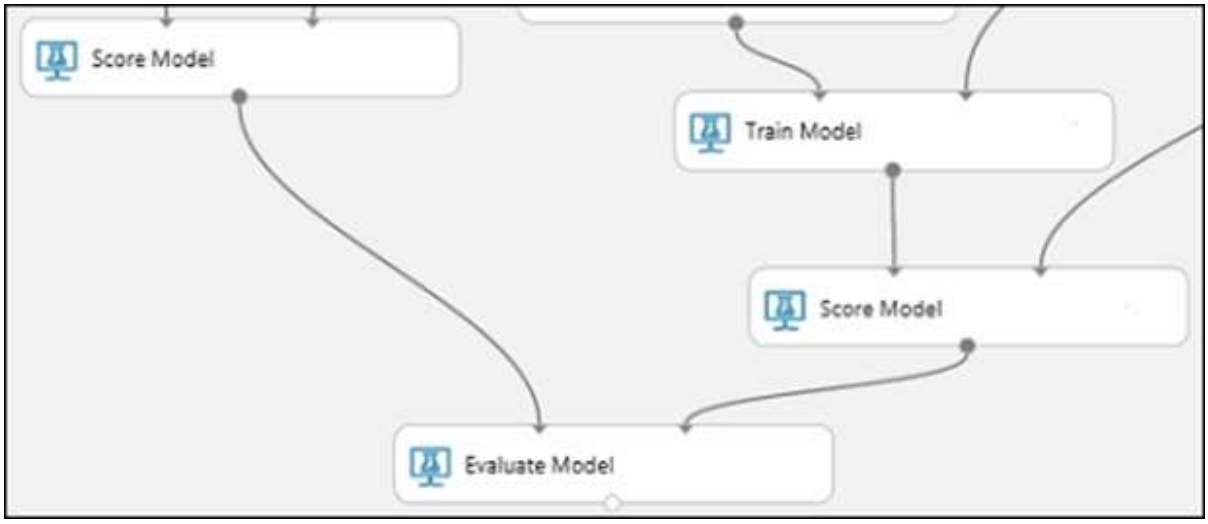
Step 2: Score Model
Score and evaluate the models
You use the testing data that was separated out by the Split Data module to score our trained models. You can then compare the results of the two models to see which generated better results.

Add the Score Model modules
1. Find the Score Model module and drag it onto the canvas.
2. Connect the Train Model module that's connected to the Two-Class Boosted Decision Tree module to the left input port of the Score Model module.
3. Connect the right Execute R Script module (our testing data) to the right input port of the Score Model module.



Step 3: Evaluate Model
To evaluate the two scoring results and compare them, you use an Evaluate Model module.

1. Find the Evaluate Model module and drag it onto the canvas.
2. Connect the output port of the Score Model module associated with the boosted decision tree model to the left input port of the Evaluate Model module.
3. Connect the other Score Model module to the right input port.



QUESTION 33
You create a multi-class image classification deep learning model that uses a set of labeled images. You create a script file named **train.py** that uses the PyTorch 1.3 framework to train the model.

You must run the script by using an estimator. The code must not require any additional Python libraries to be installed in the environment for the estimator. The time required for model training must be minimized.

You need to define the estimator that will be used to run the script.

Which estimator type should you use?

A. TensorFlow
B. PyTorch
C. SKLearn
D. Estimator

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:
For PyTorch, TensorFlow and Chainer tasks, Azure Machine Learning provides respective PyTorch, TensorFlow, and Chainer estimators to simplify using these frameworks.

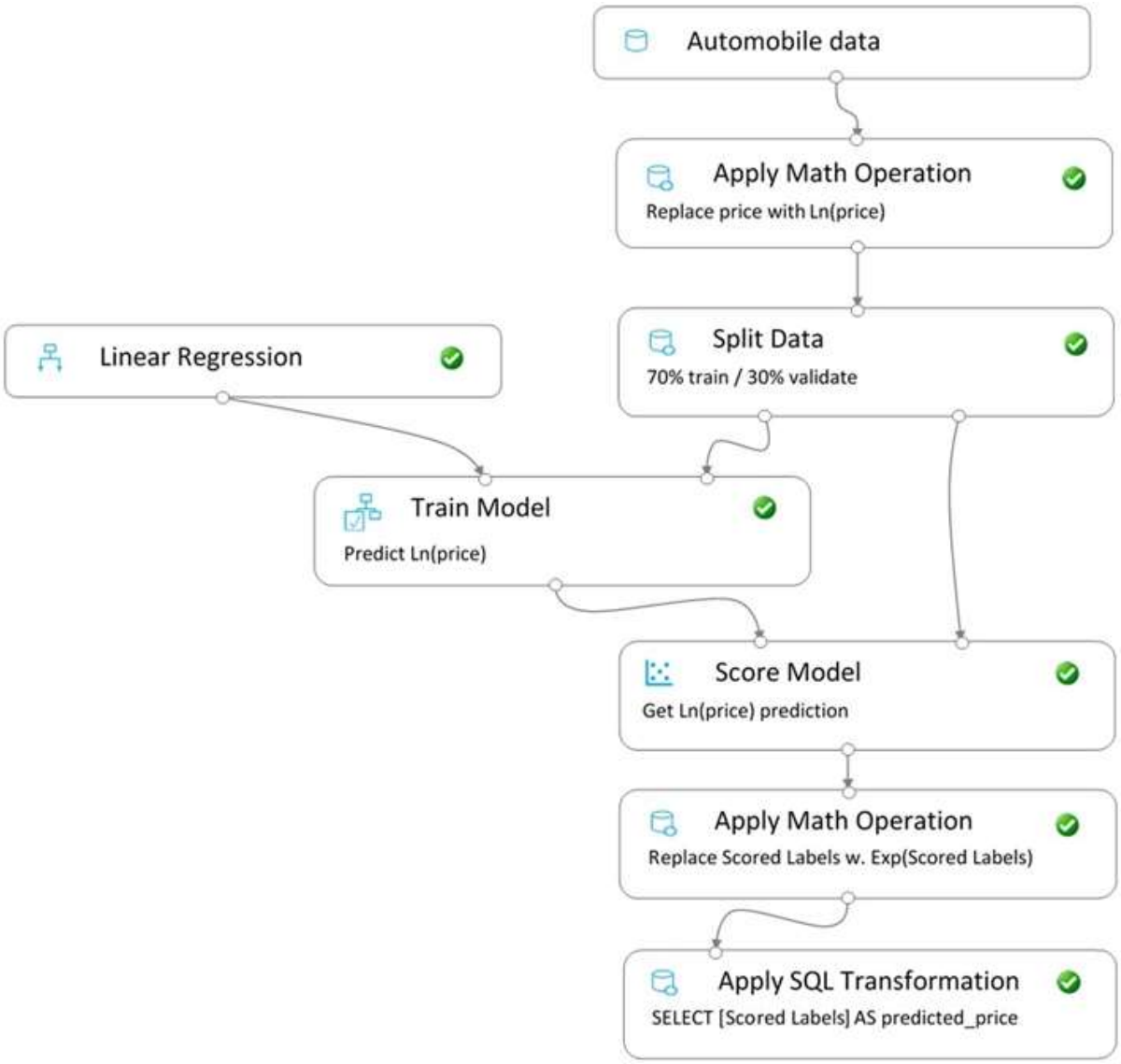
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-train-ml-models

QUESTION 34
You create a pipeline in designer to train a model that predicts automobile prices.

Because of non-linear relationships in the data, the pipeline calculates the natural log (Ln) of the prices in the training data, trains a model to predict this natural log of price value, and then calculates the exponential of the scored label to get the predicted price.

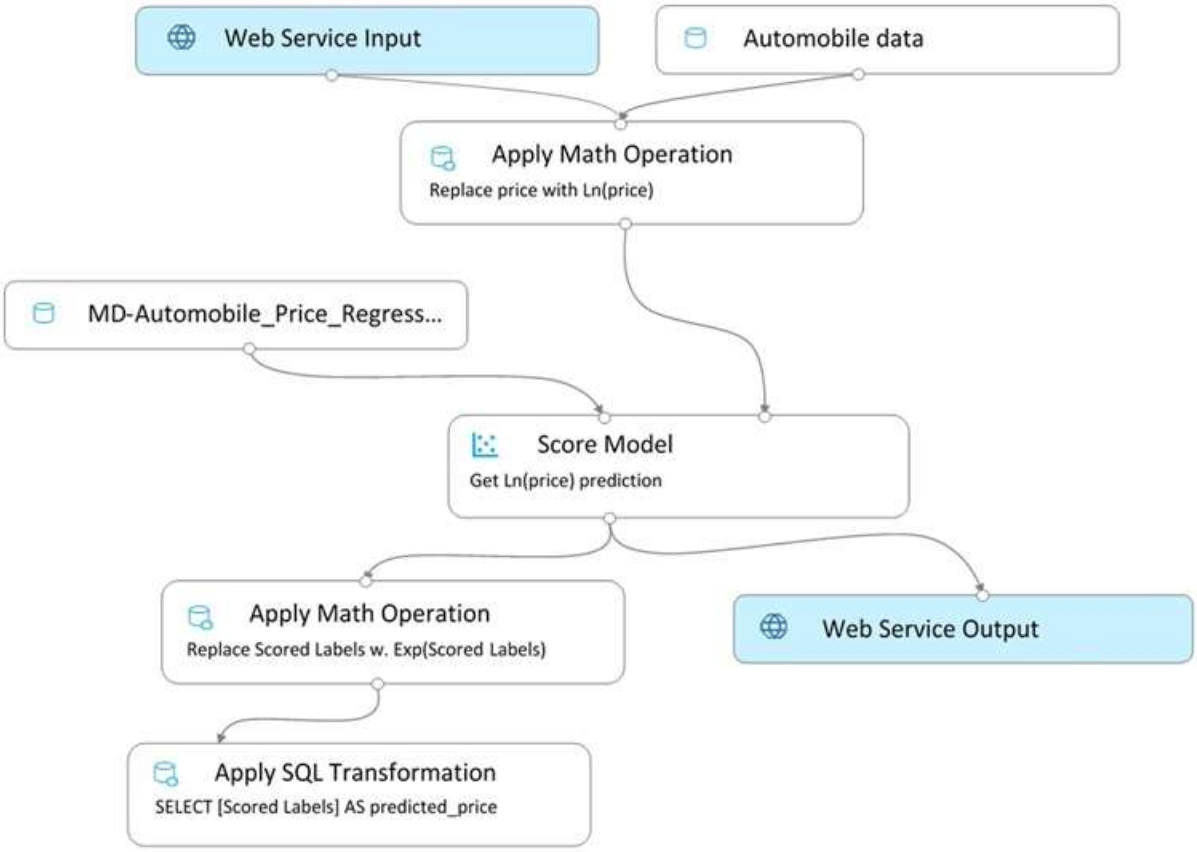
The training pipeline is shown in the exhibit. (Click the **Training pipeline** tab.)

Training pipeline



You create a real-time inference pipeline from the training pipeline, as shown in the exhibit. (Click the **Real-time pipeline** tab.)

Real-time pipeline



You need to modify the inference pipeline to ensure that the web service returns the exponential of the scored label as the predicted automobile price and that client applications are not required to include a price value in the input values.

Which three modifications must you make to the inference pipeline? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

A. Connect the output of the Apply SQL Transformation to the Web Service Output module.
B. Replace the Web Service Input module with a data input that does not include the price column.
C. Add a Select Columns module before the Score Model module to select all columns other than price.
D. Replace the training dataset module with a data input that does not include the price column.
E. Remove the Apply Math Operation module that replaces price with its natural log from the data flow.
F. Remove the Apply SQL Transformation module from the data flow.

Correct Answer: ACE
Section: (none)
Explanation

Explanation/Reference:

QUESTION 35
You train a model and register it in your Azure Machine Learning workspace. You are ready to deploy the model as a real-time web service.

You deploy the model to an Azure Kubernetes Service (AKS) inference cluster, but the deployment fails

because an error occurs when the service runs the entry script that is associated with the model deployment.

You need to debug the error by iteratively modifying the code and reloading the service, without requiring a re-deployment of the service for each code update.

What should you do?

A. Modify the AKS service deployment configuration to enable application insights and re-deploy to AKS.
B. Create an Azure Container Instances (ACI) web service deployment configuration and deploy the model on ACI.
C. Add a breakpoint to the first line of the entry script and redeploy the service to AKS.
D. Create a local web service deployment configuration and deploy the model to a local Docker container.
E. Register a new version of the model and update the entry script to load the new version of the model from its registered path.

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:
How to work around or solve common Docker deployment errors with Azure Container Instances (ACI) and Azure Kubernetes Service (AKS) using Azure Machine Learning.

The recommended and the most up to date approach for model deployment is via the Model.deploy() API using an Environment object as an input parameter. In this case our service will create a base docker image for you during deployment stage and mount the required models all in one call. The basic deployment tasks are:
1. Register the model in the workspace model registry.
2. Define Inference Configuration:
   a. Create an Environment object based on the dependencies you specify in the environment yaml file or use one of our procured environments.
   b. Create an inference configuration (InferenceConfig object) based on the environment and the scoring script.
3. Deploy the model to Azure Container Instance (ACI) service or to Azure Kubernetes Service (AKS).

QUESTION 36
HOTSPOT

You register the following versions of a model.

Model name	Model version	Tags	Properties
healthcare_model	3	'Training context':'CPU Compute'	value:87.43
healthcare_model	2	'Training context':'CPU Compute'	value:54.98
healthcare_model	1	'Training context':'CPU Compute'	value:23.56

You use the Azure ML Python SDK to run a training experiment. You use a variable named **run** to reference the experiment run.

After the run has been submitted and completed, you run the following code:

```
run.register_model(model_path='outputs/model.pkl',
 model_name='healthcare_model',
 tags={'Training context':'CPU Compute'} )
```

For each of the following statements, select Yes if the statement is true. Otherwise, select No.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

	Yes	No
The code will cause a previous version of the saved model to be overwritten.	○	○
The version number will now be 4.	○	○
The latest version of the stored model will have a property of value: 87.43.	○	○

Correct Answer:

Answer Area

	Yes	No
The code will cause a previous version of the saved model to be overwritten.	○	○
The version number will now be 4.	○	○
The latest version of the stored model will have a property of value: 87.43.	○	○

Section: (none)
Explanation

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

QUESTION 37
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. AUC_weighted
C. accuracy
D. normalized_root_mean_squared_error
E. spearman_correlation

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:
AUC_weighted is a Classification metric.

Note: AUC is the Area under the Receiver Operating Characteristic Curve. Weighted is the arithmetic mean of the score for each class, weighted by the number of true instances in each class.

Incorrect Answers:
A: normalized_mean_absolute_error is a regression metric, not a classification metric.

C: When comparing approaches to imbalanced classification problems, consider using metrics beyond accuracy such as recall, precision, and AUROC. It may be that switching the metric you optimize for during parameter selection or model selection is enough to provide desirable performance detecting the minority class.

D: normalized_root_mean_squared_error is a regression metric, not a classification metric.

Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-understand-automated-ml