

**DP-100.examcollection.premium.exam.78q**

## 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]

### 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]

### 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:**

**Actions**

Install the Azure Machine Learning SDK for Python on the cluster.

When the cluster is ready, export Zeppelin notebooks to a local environment.

Create and execute a Jupyter notebook by using automated machine learning (AutoML) on the cluster.

Install Microsoft Machine Learning for Apache Spark.

When the cluster is ready and has processed the notebook, export your Jupyter notebook to a local environment.

Create an Azure HDInsight cluster to include the Apache Spark Mlib library.

Create and execute the Zeppelin notebooks on the cluster.

Create an Azure Databricks cluster.

**Answer area**



**Correct Answer:**

**Actions**

Install the Azure Machine Learning SDK for Python on the cluster.

When the cluster is ready, export Zeppelin notebooks to a local environment.

Create and execute a Jupyter notebook by using automated machine learning (AutoML) on the cluster.

Install Microsoft Machine Learning for Apache Spark.

When the cluster is ready and has processed the notebook, export your Jupyter notebook to a local environment.

Create an Azure HDInsight cluster to include the Apache Spark Mlib library.

Create and execute the Zeppelin notebooks on the cluster.

Create an Azure Databricks cluster.

**Answer area**

Create an Azure HDInsight cluster to include the Apache Spark Mlib library.

Install Microsoft Machine Learning for Apache Spark.

Create and execute the Zeppelin notebooks on the cluster.

When the cluster is ready, export Zeppelin notebooks to a local environment.



**Section: [none]**

**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]

#### 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.
- E. Create a Data Science Virtual Machine (DSVM) Linux edition.

**Correct Answer:** E

**Section:** [none]

#### 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 MLlib
- D. Azure Machine Learning Studio

**Correct Answer:** A  
**Section:** [none]

#### 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]

#### 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]

#### 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. Deeplearning4j

**Correct Answer:** B  
**Section:** [none]

#### 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 Extensions (Intel SGX) technology

**Correct Answer:** C

**Section:** [none]

#### 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]

#### 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:

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

Section: [none]



## 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

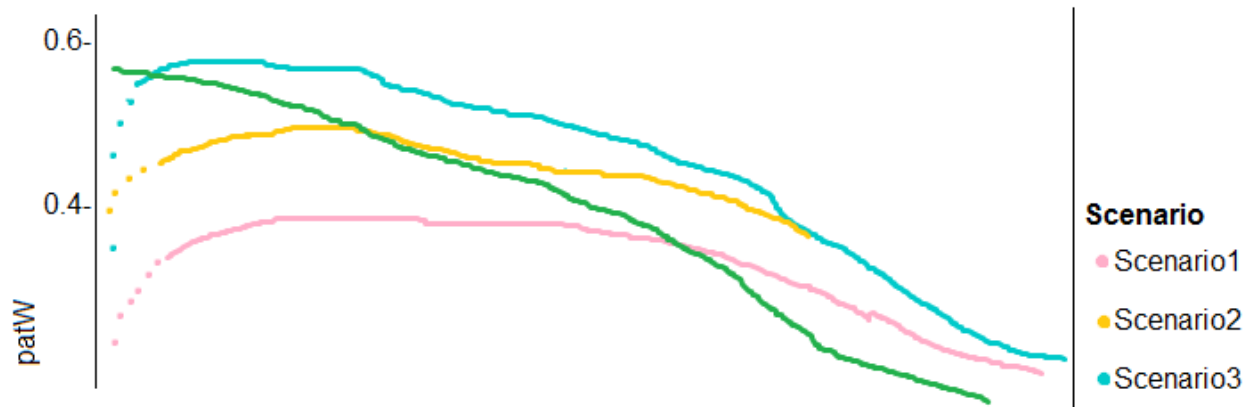
- The performance of the global penalty detection models shows lower variance but higher bias when comparing training and validation sets. Before implementing any feature changes, you must confirm the bias and variance using all training and validation cases.
- Ad response models must be trained at the 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 exclusivity 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 retrain occurs 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]

### 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

```
import torch as deeplearninglib
import tensorflow as deeplearninglib
import cntk as deeplearninglib
```

```
train_sampler = deeplearninglib.DistributedSampler(penalty_video_dataset)
train_sampler = deeplearninglib.log_uniform_candidate_sampler(penalty_video_dataset)
train_sampler = deeplearninglib.WeightedRandomSampler(penalty_video_dataset)
train_sampler = deeplearninglib.all_candidate_sampler(penalty_video_dataset)
```

```
...
train_loader =
...
(train_sampler, penalty_video_dataset)
```

```
optimizer = deeplearninglib.optim.SGD(model.parameters(), lr=0.01)
optimizer = deeplearninglib.train.GradientDescentOptimizer(learning_rate=0.10)
```

```
model = deeplearninglib.parallel.Distributed(DataParallel(model))
model = deeplearninglib.nn.parallel.DistributedDataParallel(CPU(model))
model = deeplearninglib.keras.Model([
model = deeplearninglib.keras.Sequential([
...
train_sampler.set_epoch(epoch)
for data, target in train_loader:
    data, target = data.to(device), target.to(device)
..
```

Correct Answer:

## Answer Area

```
import torch as deeplearninglib
import tensorflow as deeplearninglib
import cntk as deeplearninglib
```

```
train_sampler = deeplearninglib.DistributedSampler(penalty_video_dataset)
train_sampler = deeplearninglib.log_uniform_candidate_sampler(penalty_video_dataset)
train_sampler = deeplearninglib.WeightedRandomSampler(penalty_video_dataset)
train_sampler = deeplearninglib.all_candidate_sampler(penalty_video_dataset)
```

```
...
train_loader =
...
(train_sampler, penalty_video_dataset)
```

```
optimizer = deeplearninglib.optim.SGD(model.parameters(), lr=0.01)
optimizer = deeplearninglib.train.GradientDescentOptimizer(learning_rate=0.10)
```

```
model = deeplearninglib.parallel.Distributed(DataParallel(model))
model = deeplearninglib.nn.parallel.DistributedDataParallel(CPU(model))
model = deeplearninglib.keras.Model([
model = deeplearninglib.keras.Sequential([
...
train_sampler.set_epoch(epoch)
    for data, target in train_loader:
        data, target = data.to(device), target.to(device)
..
```

Section: [none]

## **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 the 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 a feature selection algorithm to analyze the relationship between the MediaValue and AvgRoomsInHouse columns.

## **Model training**

### **Permutation Feature Importance**

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

### **Hyperparameters**

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

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

### **Testing**

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

### **Cross-validation**

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

### **Linear regression module**

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

### **Data visualization**

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

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

## **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]

## 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:

### Modules

Score Matchbox Recommender

Apply Transformation

Evaluate Recommender

Evaluate Model

Train Model

Sweep Clustering

Score Model

Load Trained Model

### Answer Area



Correct Answer:

Modules		Answer Area
Score Matchbox Recommender		Sweep Clustering
Apply Transformation		Train Model
Evaluate Recommender		Evaluate Model
Evaluate Model	⬅	⬆
Train Model	➡	⬇
Sweep Clustering		
Score Model		
Load Trained Model		

**Section:** [none]

### 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]

### Question Set 3

#### QUESTION 1

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]

#### 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 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]

#### 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 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]**

#### QUESTION 4

**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]**

#### QUESTION 5

**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]

#### QUESTION 6

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

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

You 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]

#### QUESTION 7

##### 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 imbalanced 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:**



## Answer Area

### SMOTE

Label column

Selected columns:

**All labels**

Launch column selector

SMOTE percentage

	▼
0	
300	
3000	
4000	

Number of nearest neighbors

	▼
0	
1	
5	
4000	

Random seed

0
---

**Correct Answer:**

## Answer Area

### SMOTE

Label column

Selected columns:  
**All labels**

Launch column selector

SMOTE percentage

	▼
0	
300	
3000	
4000	

Number of nearest neighbors

	▼
0	
1	
5	
4000	

Random seed

0
---

Section: [none]

### QUESTION 8

You are solving a classification task.

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

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

Which value should you use?

- A.  $k=0.5$
- B.  $k=0$
- C.  $k=5$
- D.  $k=1$

**Correct Answer: C**

Section: [none]

### QUESTION 9

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]

### QUESTION 10

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:**

<b>Actions</b>		<b>Answer Area</b>
Build Counting Transform		
Missing Values Scrubber		
Feature Hashing		
Clean Missing Data	⬅	⬆
Replace Discrete Values	➡	⬇
Import Data		
Latent Dirichlet Transformation		
Partition and Sample		

**Correct Answer:**

Actions		Answer Area
Build Counting Transform		Import Data
Missing Values Scrubber		Clean Missing Data
Feature Hashing		Partition and Sample
Clean Missing Data	⬅	⬆
Replace Discrete Values	➡	⬇
Import Data		
Latent Dirichlet Transformation		
Partition and Sample		

Section: [none]

**QUESTION 11**  
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:**

## Answer Area

### ▴ Partition and Sample

Partition or sample mode

	▼
Assign to Folds	
Pick Fold	
Sampling	
Head	

Rate of sampling

.2
----

Random seed for sampling

	▼
0	
1	
time.clock()	
utcNow()	

Stratified split for sampling

False	▼
-------	---

Correct Answer:

## Answer Area

### ▲ Partition and Sample

Partition or sample mode

	▼
Assign to Folds	
Pick Fold	
Sampling	
Head	

Rate of sampling

.2
----

Random seed for sampling

	▼
0	
1	
time.clock()	
utcNow()	

Stratified split for sampling

False	▼
-------	---

Section: [none]

#### QUESTION 12

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]

#### QUESTION 13

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:**

### Answer Area

Methods	Scenario	Module
Clean Missing Data	Replace missing values by removing rows and columns.	
SMOTE	Increase the number of low-incidence examples in the dataset.	
Convert to Indicator Values	Convert a categorical feature into a binary indicator.	
Remove Duplicate Rows	Remove potential duplicates from a dataset.	
Threshold Filter		

**Correct Answer:**

### Answer Area

Methods	Scenario	Module
	Replace missing values by removing rows and columns.	Clean Missing Data
	Increase the number of low-incidence examples in the dataset.	SMOTE
	Convert a categorical feature into a binary indicator.	Convert to Indicator Values
	Remove potential duplicates from a dataset.	Remove Duplicate Rows
Threshold Filter		

**Section:** [none]

### QUESTION 14

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(
```

dataFrame
pandas
salesData
year

```
, id_vars=
```

shop
year
value
Shop X, Shop Y, Shop
/

```
, value_vars=
```

'shop'
'year'
['year']
['2017', '2018']

```
)
```

**Correct Answer:**

**Answer Area**

```
import pandas as pd
salesData = pd.melt(
```

dataFrame
pandas
salesData
year

```
, id_vars=
```

shop
year
value
Shop X, Shop Y, Shop
/

```
, value_vars=
```

'shop'
'year'
['year']
['2017', '2018']

```
)
```

**Section:** [none]

## QUESTION 15

### 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	<div><div></div><div>Gender, IsPlaySoccer</div><div>Gender, PrevExamMarks, Height, Weight</div><div>PrevExamMarks, Height, Weight</div><div>IsPlaySoccer</div></div>
Continuous variables	<div><div></div><div>Gender, IsPlaySoccer</div><div>Gender, PrevExamMarks, Height, Weight</div><div>PrevExamMarks, Height, Weight</div><div>IsPlaySoccer</div></div>

**Correct Answer:**

## Answer Area

Category	Variables
Categorical variables	<div><div>▼</div><div>Gender, IsPlaySoccer</div><div>Gender, PrevExamMarks, Height, Weight</div><div>PrevExamMarks, Height, Weight</div><div>IsPlaySoccer</div></div>
Continuous variables	<div><div>▼</div><div>Gender, IsPlaySoccer</div><div>Gender, PrevExamMarks, Height, Weight</div><div>PrevExamMarks, Height, Weight</div><div>IsPlaySoccer</div></div>

Section: [none]

### QUESTION 16

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

### Preprocess Text

Language

English



Remove by part of speech

False



Text column to clean

**Selected columns:**

**Column names: String, Feature**

Launch column selector

☐

Remove stop words

☐

Lemmatization

☐

Detect sentences

☐

Normalize case to lowercase

☐

Remove numbers

☐

Remove special characters

☐

Remove duplicate characters

☐

Remove email addresses

☐

Remove URLs

☐

Expand verb contractions

☐

Normalize backslashes to slashes

☐

Split tokens on special characters



Correct Answer:

## Answer Area

### Preprocess Text

Language

English

Remove by part of speech

False

Text column to clean

**Selected columns:**

**Column names: String, Feature**

Launch column selector

<input checked="" type="checkbox"/>	Remove stop words	≡
<input checked="" type="checkbox"/>	Lemmatization	≡
<input type="checkbox"/>	Detect sentences	≡
<input type="checkbox"/>	Normalize case to lowercase	≡
<input type="checkbox"/>	Remove numbers	≡
<input checked="" type="checkbox"/>	Remove special characters	≡
<input type="checkbox"/>	Remove duplicate characters	≡
<input type="checkbox"/>	Remove email addresses	≡
<input type="checkbox"/>	Remove URLs	≡
<input type="checkbox"/>	Expand verb contractions	≡
<input type="checkbox"/>	Normalize backslashes to slashes	≡
<input type="checkbox"/>	Split tokens on special characters	≡

Section: [none]

#### 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 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]**

#### 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 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]**

#### 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 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]**

### QUESTION 20

**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]**

### QUESTION 21

**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]**

### QUESTION 22

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]

**QUESTION 23**

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]

**QUESTION 24**

You are evaluating a completed binary classification machine learning model.

You need to use the precision as the valuation metric.

Which visualization should you use?

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

**Correct Answer:** D

**Section:** [none]

## 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



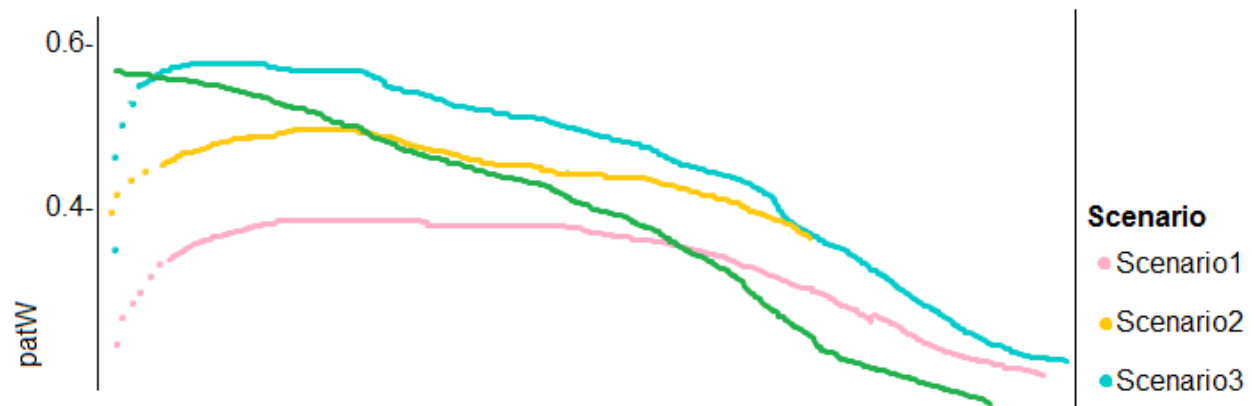
- The performance of the global penalty detection models shows lower variance but higher bias when comparing training and validation sets. Before implementing any feature changes, you must confirm the bias and variance using all training and validation cases.
- Ad response models must be trained at the 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 exclusivity 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 retrain occurs 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:

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

Section: [none]

## 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]

## **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 the 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 a feature selection algorithm to analyze the relationship between the MediaValue and AvgRoomsInHouse columns.

## **Model training**

### **Permutation Feature Importance**

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

### **Hyperparameters**

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

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

### **Testing**

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

### **Cross-validation**

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

### **Linear regression module**

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

### **Data visualization**

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

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

## **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

### ▲ Tune Model Hyperparameters

Specify parameter sweeping mode

Random sweep ▼

Maximum number of runs on random sweep

5

Random seed

0

Label column

Selected columns:

Column names: MedianValue

Launch column selector

Metric for measuring performance for classification

	▼
F-score	
Precision	
Recall	
Accuracy	

Metric for measuring performance for regression

	▼
Root of mean squared error	
R-squared	
Mean zero one error	
Mean absolute error	

Correct Answer:

## Answer Area

### ▲ Tune Model Hyperparameters

Specify parameter sweeping mode

Random sweep ▼

Maximum number of runs on random sweep

5

Random seed

0

Label column

Selected columns:

Column names: MedianValue

Launch column selector

Metric for measuring performance for classification

F-score

Precision

Recall

Accuracy

Metric for measuring performance for regression

Root of mean squared error

R-squared

Mean zero one error

Mean absolute error

Section: [none]

### 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:**



## Answer Area

### Filter Based Feature Selection

Feature scoring method

	▼
Fisher Score	
Chi-squared	
Mutual information	
Counts	

☒ Operate on feature columns only



Target column

	▼
MedianValue	
AvgRooms/nHouse	

Launch column selector

Number of desired features



1

Correct Answer:

## Answer Area

### Filter Based Feature Selection

Feature scoring method

	▼
Fisher Score	
Chi-squared	
Mutual information	
Counts	

☒ Operate on feature columns only

Target column

	▼
MedianValue	
AvgRooms/nHouse	

Launch column selector

Number of desired features

1
---

Section: [none]

### 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]

### 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

### Permutation Feature importance

Random seed

	▼
0	
500	

	▼
Regression – Root Mean Square Error	
Regression – R-squared	
Regression – Mean Zero One Error	
Regression – Mean Absolute Error	

Correct Answer:

## Answer Area

### Permutation Feature importance

Random seed

	▼
0	
500	

	▼
Regression – Root Mean Square Error	
Regression – R-squared	
Regression – Mean Zero One Error	
Regression – Mean Absolute Error	

Section: [none]

### 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]

#### 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. Preprocess Text
- C. Execute Python Script
- D. Latent Dirichlet Allocation

**Correct Answer:** A

**Section:** [none]

#### 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 = 
X_train = .fit_transform(X_train)
x_test = pca.
```

PCA()
PCA(n_components = 150)
PCA(n_components = 10)
PCA(n_components = 10000)

pca
model
sklearn.decomposition

x_test
X_train
fit(x_test)
transform(x_test)

**Correct Answer:**

## Answer Area

```
from sklearn.decomposition import PCA
pca = 
PCA()
PCA(n_components = 150)
PCA(n_components = 10)
PCA(n_components = 10000)

X_train = .fit_transform(X_train)
pca
model
sklearn.decomposition

x_test = pca.
x_test
X_train
fit(x_test)
transform(x_test)
```

Section: [none]

### 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]

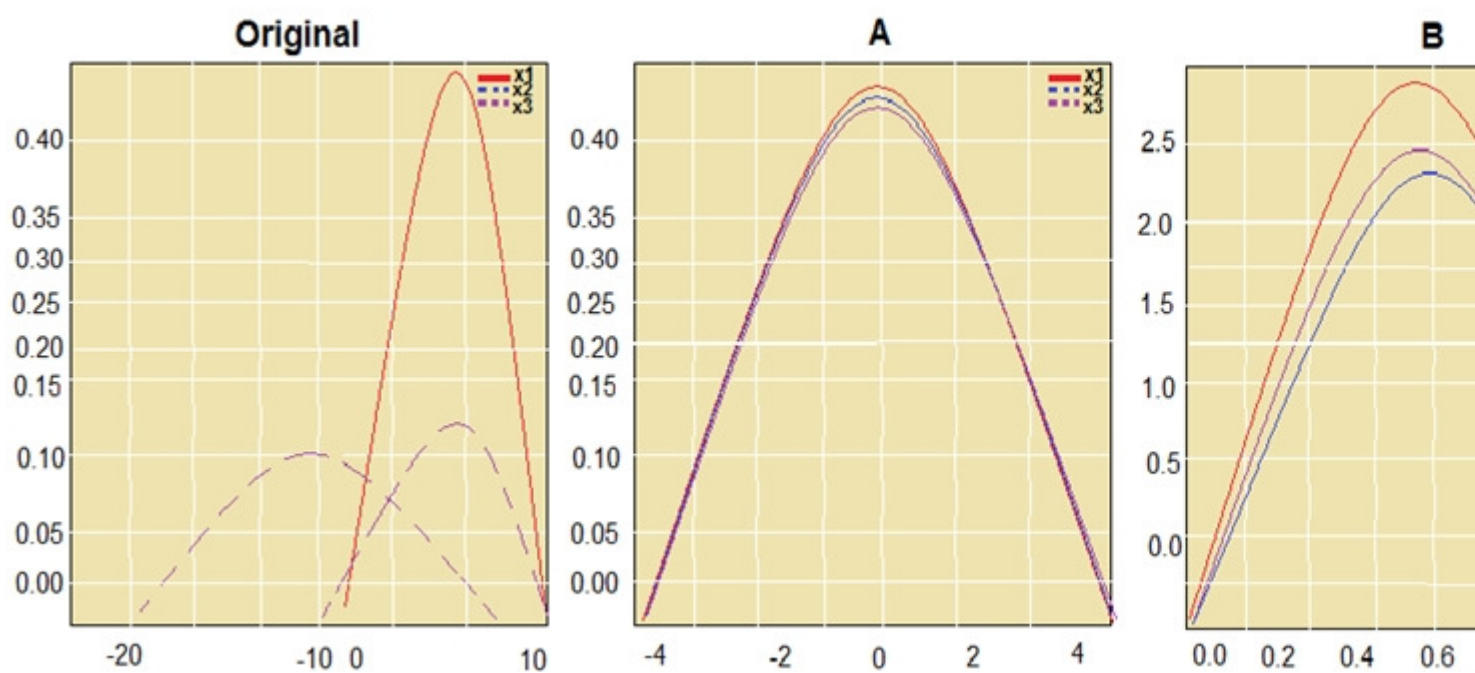
### QUESTION 5

HOTSPOT

You are performing feature scaling by using the scikit-learn Python library for x1, 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]

## 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

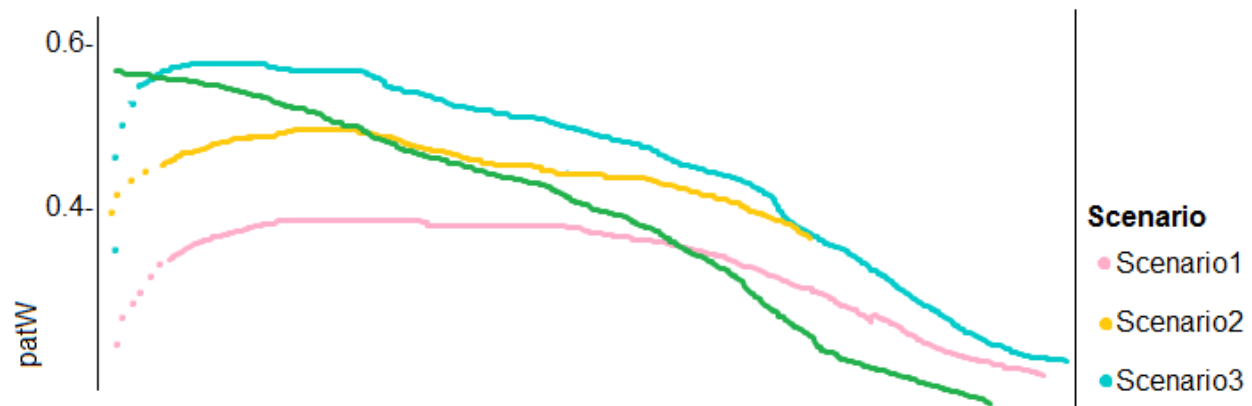
- The performance of the global penalty detection models shows lower variance but higher bias when comparing training and validation sets. Before implementing any feature changes, you must confirm the bias and variance using all training and validation cases.
- Ad response models must be trained at the 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 exclusivity 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 retrain occurs 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

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.

#### Answer area



Correct Answer:

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

Section: [none]

## 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:

Actions	Answer Area
Define a cross-entropy function activation.	
Add cost functions for each target state.	
Evaluate the classification error metric.	
Evaluate the distance error metric.	
Add cost functions for each component metric.	
Define a sigmoid loss function activation.	

Correct Answer:

### Actions

Define a cross-entropy function activation.

Add cost functions for each target state.

Evaluate the classification error metric.

Evaluate the distance error metric.

Add cost functions for each component metric.

Define a sigmoid loss function activation.

### Answer Area

Define a cross-entropy function activation.

Add cost functions for each target state.

Evaluate the distance error metric.



Section: [none]

#### 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]

#### 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]



## **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 the 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 a feature selection algorithm to analyze the relationship between the MediaValue and AvgRoomsInHouse columns.

## **Model training**

### **Permutation Feature Importance**

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

### **Hyperparameters**

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

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

### **Testing**

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

### **Cross-validation**

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

### **Linear regression module**

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

### **Data visualization**

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

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

### **QUESTION 1**

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:**

**Code segments**

```
early_termination_policy =  
TruncationSelectionPolicy(evaluation_interval=1,  
truncation_percentage=20, delay_evaluation=5)
```

```
import TruncationSelectionPolicy
```

```
from azureml.train.hyperdrive
```

```
import BanditPolicy
```

```
early_termination_policy = BanditPolicy  
(slack_factor = 0.1, evaluation_interval=1,  
delay_evaluation=5)
```

**Answer Area****Correct Answer:****Code segments**

```
early_termination_policy =  
TruncationSelectionPolicy(evaluation_interval=1,  
truncation_percentage=20, delay_evaluation=5)
```

```
import TruncationSelectionPolicy
```

```
from azureml.train.hyperdrive
```

```
import BanditPolicy
```

```
early_termination_policy = BanditPolicy  
(slack_factor = 0.1, evaluation_interval=1,  
delay_evaluation=5)
```

**Answer Area**

```
from azureml.train.hyperdrive
```

```
import TruncationSelectionPolicy
```





```
early_termination_policy =  
TruncationSelectionPolicy(evaluation_interval=1,  
truncation_percentage=20, delay_evaluation=5)
```

**Section: [none]****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:**

Actions		Answer Area
Add the Ordinal Regression module.		
Add the Two-Class Averaged Perception module.		
Augment the data.		
Add the Bayesian Linear Regression module.		
Decrease the memory size for L-BFGS.		
Add the Multiclass Decision Jungle module.		
Configure the regularization weight.		

Correct Answer:

Actions		Answer Area
Add the Ordinal Regression module.		Augment the data.
Add the Two-Class Averaged Perception module.		Add the Bayesian Linear Regression module.
	➤	Configure the regularization weight.
	⬅	⏏
Decrease the memory size for L-BFGS.		
Add the Multiclass Decision Jungle module.		

Section: [none]

**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:**

### Code segments

```
early_termination_policy = TruncationSelectionPolicy  
(evaluation_interval=1, truncation_percentage=20,  
delay_evaluation = 5)
```

```
import BanditPolicy
```

```
import TruncationSelectionPolicy
```

```
early_termination_policy= BanditPolicy (slack_factor =  
0.1, evaluation_interval = 1, delay_evaluation = 5)
```

```
from azureml.train.hyperdrive
```

```
early_termination_policy = MedianStoppingPolicy  
(evaluation_interval = 1, delay_evaluation=5)
```

```
import MedianStoppingPolicy
```

### Answer Area



Correct Answer:

### Code segments

import BanditPolicy

early\_termination\_policy= BanditPolicy (slack\_factor = 0.1, evaluation\_interval = 1, delay\_evaluation = 5)

early\_termination\_policy = MedianStoppingPolicy (evaluation\_interval = 1, delay\_evaluation=5)

import MedianStoppingPolicy

### Answer Area

from azureml.train.hyperdrive

import TruncationSelectionPolicy



early\_termination\_policy = TruncationSelectionPolicy (evaluation\_interval=1, truncation\_percentage=20, delay\_evaluation = 5)



Section: [none]



### 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]

#### 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]

#### 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]**

#### 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]**

#### 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]**

#### 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
Bayesian Linear	3.276025	4.655442	0.511436
Neural Network	2.676538	3.621476	0.417847
Boosted Decision Tree	2.168847	2.878077	0.338589
Linear	6.350005	8.720718	0.99133
Decision Forest	2.390206	3.315 164	0.373146


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	
Neural 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]

### 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:

## Answer Area

Tree Depth	Bias	Variance
5	<div><div></div><div>High</div><div>Low</div><div>Identical</div></div>	<div><div></div><div>High</div><div>Low</div><div>Identical</div></div>
15	<div><div></div><div>High</div><div>Low</div><div>Identical</div></div>	<div><div></div><div>High</div><div>Low</div><div>Identical</div></div>

Correct Answer:

## Answer Area

Tree Depth	Bias	Variance
5	<div><div></div><div>High</div><div>Low</div><div>Identical</div></div>	<div><div></div><div>High</div><div>Low</div><div>Identical</div></div>
15	<div><div></div><div>High</div><div>Low</div><div>Identical</div></div>	<div><div></div><div>High</div><div>Low</div><div>Identical</div></div>

Section: [none]

### 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**

Modules	Step	Module
Two-Class Boosted Decision Tree	Define the parameter scope	
Partition and Sample	Define the cross-validation settings	
Tune Model Hyperparameters	Define the metric	
Split Data	Train, evaluate, and compare	

**Correct Answer:**

**Answer Area**

Modules	Step	Module
Two-Class Boosted Decision Tree	Define the parameter scope	Split Data
Partition and Sample	Define the cross-validation settings	Partition and Sample
Tune Model Hyperparameters	Define the metric	Two-Class Boosted Decision Tree
Split Data	Train, evaluate, and compare	Tune Model Hyperparameters

**Section:** [none]

### 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	<div>▼</div> <div>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 frequencies in the input data.</div>
C parameter	<div>▼</div> <div>Penalty parameter Degree of polynomial kernel function Size of the kernel cache</div>

Correct Answer:

## Answer Area

Code Segment	Evaluation Statement
class_weight=balanced	<div>▼</div> <div>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 frequencies in the input data.</div>
C parameter	<div>▼</div> <div>Penalty parameter Degree of polynomial kernel function Size of the kernel cache</div>

Section: [none]

### 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]

### 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]

### 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]

### 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.

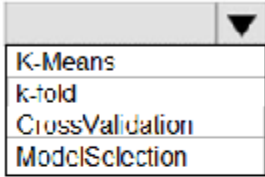
How should you complete the code segment? To answer, select the appropriate code segment in the dialog box in the answer area.

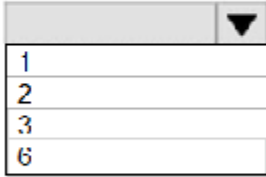
**NOTE:** Each correct selection is worth one point.

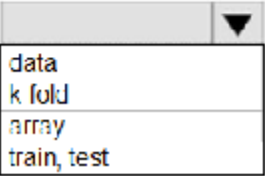
**Hot Area:**



## Answer Area

```
from numpy import array
from sklearn.model_selection import 

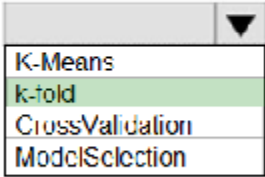
data = array([10, 20, 30, 40, 50, 60])
kfold = Kfold(n_splits=, shuffle = True, random state=1)

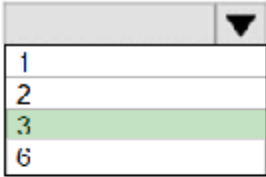
for train, test in kfold.split():

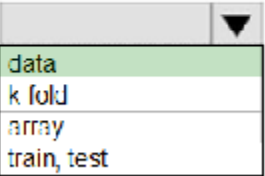
    print('train: %s, test: %s' % (data[train], data[test]))
```

Correct Answer:

## Answer Area

```
from numpy import array
from sklearn.model_selection import 

data = array([10, 20, 30, 40, 50, 60])
kfold = Kfold(n_splits=, shuffle = True, random state=1)

for train, test in kfold.split():

    print('train: %s, test: %s' % (data[train], data[test]))
```

Section: [none]

### 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
- E. Random seed

**Correct Answer:** D

**Section:** [none]

### QUESTION 15

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

The training loss, validation loss, training accuracy, and validation accuracy of each training epoch has been provided.

You need to identify whether the classification model is 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

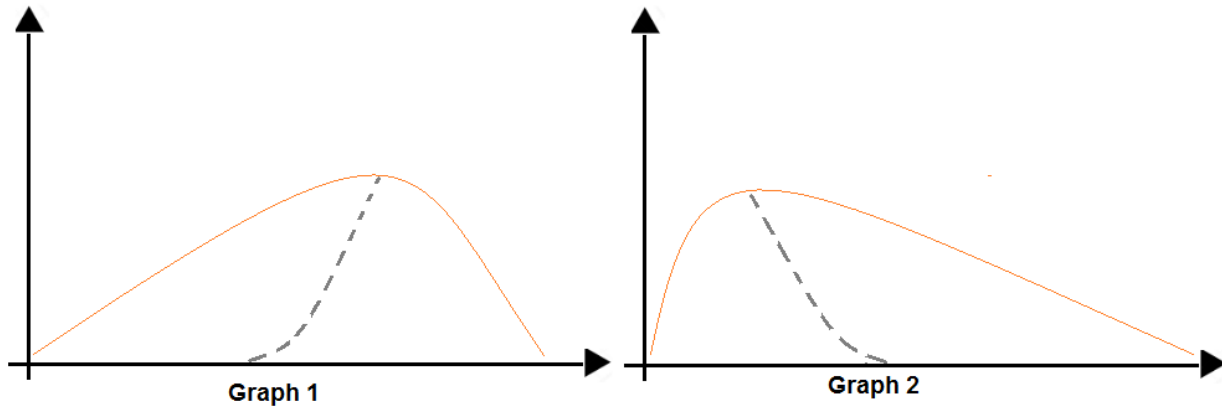
**Section:** [none]

### 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?	<input type="text"/> <ul style="list-style-type: none"> <li>Negative skew</li> <li>Positive skew</li> <li>Normal distribution</li> <li>Bimodal distribution</li> </ul>
Which type of distribution is shown for the dataset density curve of Graph 2?	<input type="text"/> <ul style="list-style-type: none"> <li>Negative skew</li> <li>Positive skew</li> <li>Normal distribution</li> <li>Bimodal distribution</li> </ul>

**Correct Answer:**

## Answer Area

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

Section: [none]

### 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]

### QUESTION 18

You are 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]**

**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]**