70-774

Number: 70-774
Passing Score: 800
Time Limit: 120 min
File Version: 1.0

70-774

Perform Cloud Data Science with Azure Machine Learning

Exam A

QUESTION 1
DRAG DROP

Note: This question is part of a series of questions that use the same scenario. For your convenience, the scenario is repeated in each question. Each question presents a different goal and answer choices, but the text of the scenario is exactly the same in each question in this series.

A travel agency named Margie’s Travel sells airline tickets to customers in the United States.

Margie’s Travel wants you to provide insights and predictions on flight delays. The agency is considering implementing a system that will communicate to its customers as the flight departure nears about possible delays due to weather conditions. The flight data contains the following attributes: ▪ DepartureDate: The departure date aggregated at a per hour granularity

- Carrier: The code assigned by the IATA and commonly used to identify a carrier
- OriginAitportID: An identification number assigned by the USDOT to identify a unique airport (the flight’s origin)
- DestAirportID: An identification number assigned by the USDOT to identify a unique airport (the flight’s destination)
- DepDel: The departure delay in minutes
- DepDel30: A Boolean value indicating whether the departure was delayed by 30 minutes or more (a value of 1 indicates that the departure was delayed by 30 minutes or more)

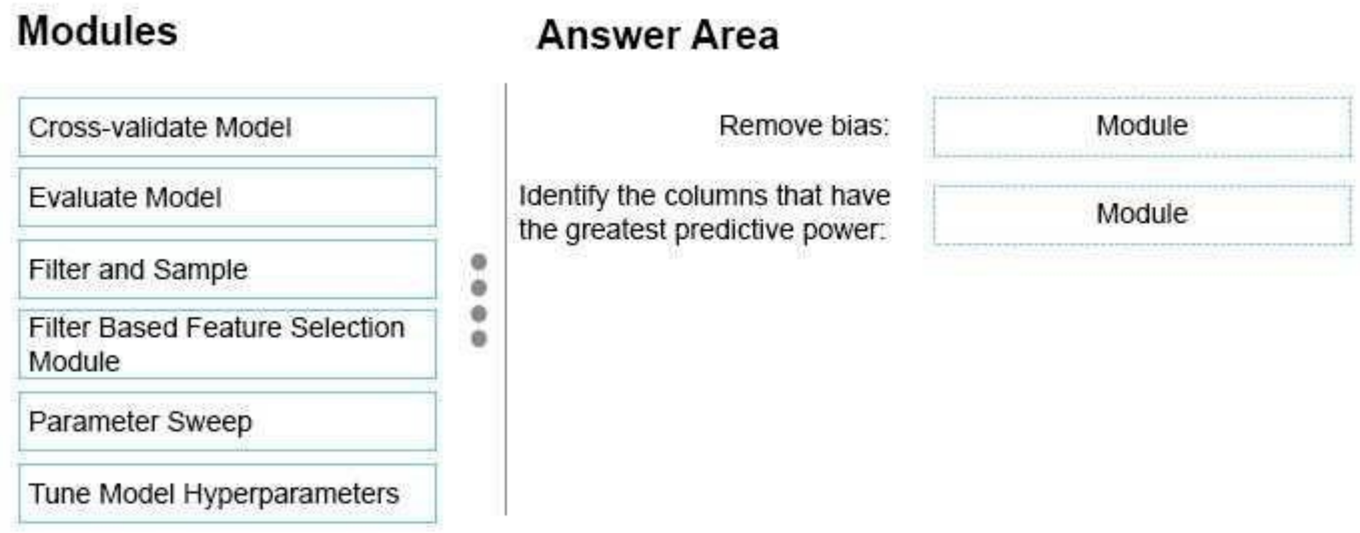
The weather data contains the following attributes: AirportID, ReadingDate (YYYY/MM/DD HH), SkyConditionVisibility, WeatherType, WindSpeed, StationPressure, PressureChange, and HourlyPrecip.

You need to remove the bias and to identify the columns in the input dataset that have the greatest predictive power.

Which module should you use for each requirement? To answer, drag the appropriate modules to the correct requirements. 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:



Modules

Cross-validate Model

Evaluate Model

Filter and Sample

Filter Based Feature Selection Module

Parameter Sweep

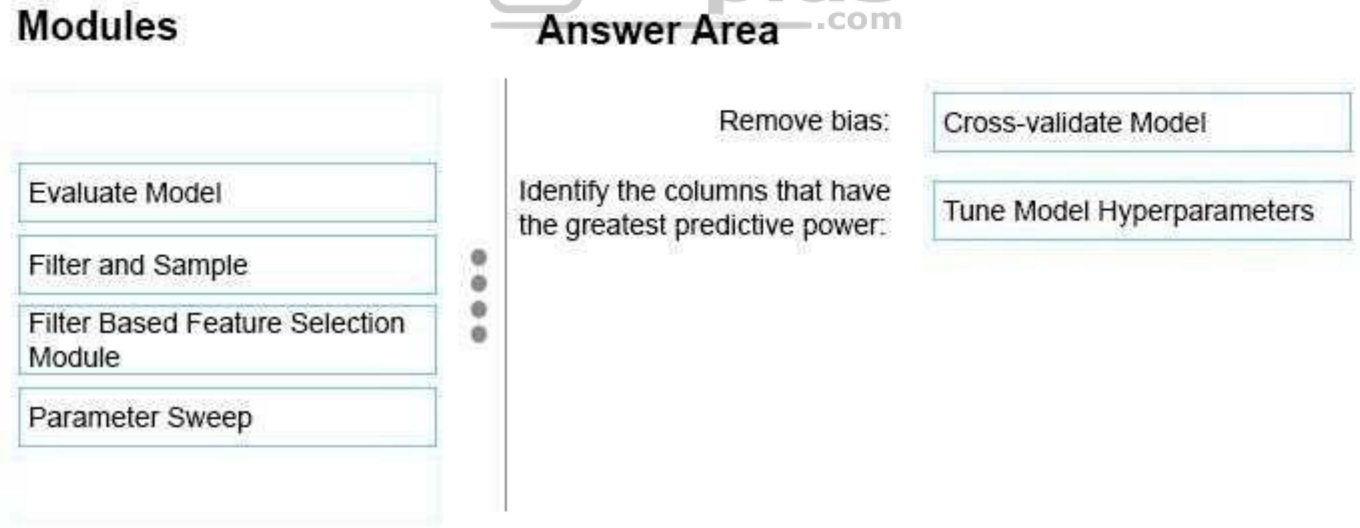
Tune Model Hyperparameters

Answer Area

Remove bias: Module

Identify the columns that have the greatest predictive power: Module

Correct Answer:



Modules

Evaluate Model

Filter and Sample

Filter Based Feature Selection Module

Parameter Sweep

Answer Area

Remove bias: Cross-validate Model

Identify the columns that have the greatest predictive power: Tune Model Hyperparameters

Section: (none)
Explanation

Explanation/Reference:
References:
https://gallery.cortanaintelligence.com/Experiment/Binary-Classification-Flight-delay-prediction-3 https://msdn.microsoft.com/library/azure/038d91b6-c2f2-42a1-9215-1f2c20ed1b40

QUESTION 2
Note: This question is part of a series of questions that use the same scenario. For your convenience, the scenario is repeated in each question. Each question presents a different goal and answer choices, but the text of the scenario is exactly the same in each question in this series.

A travel agency named Margie’s Travel sells airline tickets to customers in the United States.

Margie’s Travel wants you to provide insights and predictions on flight delays. The agency is considering implementing a system that will communicate to its customers as the flight departure nears about possible delays due to weather conditions. The flight data contains the following attributes: ▪ DepartureDate: The departure date aggregated at a per hour granularity

- Carrier: The code assigned by the IATA and commonly used to identify a carrier
- OriginAitportID: An identification number assigned by the USDOT to identify a unique airport (the flight’s origin)
- DestAirportID: An identification number assigned by the USDOT to identify a unique airport (the flight’s destination)
- DepDel: The departure delay in minutes
- DepDel30: A Boolean value indicating whether the departure was delayed by 30 minutes or more (a value of 1 indicates that the departure was delayed by 30 minutes or more)

The weather data contains the following attributes: AirportID, ReadingDate (YYYY/MM/DD HH), SkyConditionVisibility, WeatherType, WindSpeed, StationPressure, PressureChange, and HourlyPrecip.

You have an untrained Azure Machine Learning model that you plan to train to predict flight delays.

You need to assess the variability of the dataset and the reliability of the predictions from the model.

Which module should you use?

A. Cross-Validate Model
B. Evaluate Model

C. Tune Model Hyperparameters
D. Train Model
E. Score Model

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:
Explanation:

References: https://msdn.microsoft.com/en-us/library/azure/dn905852.aspx

QUESTION 3
Note: This question is part of a series of questions that use the same scenario. For your convenience, the scenario is repeated in each question. Each question presents a different goal and answer choices, but the text of the scenario is exactly the same in each question in this series.

A travel agency named Margie's Travel sells airline tickets to customers in the United States.

Margie's Travel wants you to provide insights and predictions on flight delays. The agency is considering implementing a system that will communicate to its customers as the flight departure nears about possible delays due to weather conditions. The flight data contains the following attributes: ▪ DepartureDate: The departure date aggregated at a per hour granularity
▪ Carrier: The code assigned by the IATA and commonly used to identify a carrier
▪ OriginAitportID: An identification number assigned by the USDOT to identify a unique airport (the flight's origin)

- DestAirportID: An identification number assigned by the USDOT to identify a unique airport (the flight's destination)
- DepDel: The departure delay in minutes
- DepDel30: A Boolean value indicating whether the departure was delayed by 30 minutes or more (a value of 1 indicates that the departure was delayed by 30 minutes or more)

The weather data contains the following attributes: AirportID, ReadingDate (YYYY/MM/DD HH), SkyConditionVisibility, WeatherType, WindSpeed, StationPressure, PressureChange, and HourlyPrecip.

You plan to predict flight delays that are 30 minutes or more.

You need to build a training model that accurately fits the data. The solution must minimize over fitting and minimize data leakage.

Which attribute should you remove?

A. OriginAirportID
B. DepDel

C. DepDel30
D. Carrier
E. DestAirportID

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:
Explanation:

QUESTION 4
Note: This question is part of a series of questions that use the same or similar answer choices. An answer choice may be correct for more than one question in the series. Each question is independent of the other questions in this series. Information and details provided in a question apply only to that question.

You need to remove rows that have an empty value in a specific column. The solution must use a native module.

Which module should you use?

A. Execute Python Script
B. Tune Model Hyperparameters
C. Normalize Data

D. Select Columns in Dataset
E. Import Data
F. Edit Metadata
G. Clip Values
H. Clean Missing Data

Correct Answer: H
Section: (none)
Explanation

Explanation/Reference:
Explanation:

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

QUESTION 5
Note: This question is part of a series of questions that use the same or similar answer choices. An answer choice may be correct for more than one question in the series. Each question is independent of the other questions in this series. Information and details provided in a question apply only to that question.

You have a non-tabular file that is saved in Azure Blob storage.

You need to download the file locally, access the data in the file, and then format the data as a dataset.

Which module should you use?

A. Execute Python Script
B. Tune Model Hyperparameters
C. Normalize Data
D. Select Columns in Dataset
E. Import Data
F. Edit Metadata
G. Clip Values
H. Clean Missing Data

Correct Answer: E
Section: (none)
Explanation

Explanation/Reference:
Explanation:

References: https://msdn.microsoft.com/en-us/library/azure/mt674698.aspx

QUESTION 6
Note: This question is part of a series of questions that use the same or similar answer choices. An answer choice may be correct for more than one question in the series. Each question is independent of the other questions in this series. Information and details provided in a question apply only to that question.

You have a dataset that contains a column named Column1. Column1 is empty.

You need to omit Column1 from the dataset. The solution must use a native module.

Which module should you use?

A. Execute Python Script
B. Tune Model Hyperparameters
C. Normalize Data
D. Select Columns in Dataset
E. Import Data
F. Edit Metadata
G. Clip Values
H. Clean Missing Data

Correct Answer: D
Section: (none)
Explanation

Explanation/Reference:
Explanation:

References: https://msdn.microsoft.com/en-us/library/azure/dn905883.aspx

QUESTION 7
Note: This question is part of a series of questions that use the same or similar answer choices. An answer choice may be correct for more than one question in the series. Each question is independent of the other questions in this series. Information and details provided in a question apply only to that question.

You need to use only one percent of an Apache Hive data table by conducting random sampling by groups.

Which module should you use?

A. Execute Python Script
B. Tune Model Hyperparameters
C. Normalize Data
D. Select Columns in Dataset

E. Import Data
F. Edit Metadata
G. Clip Values
H. Clean Missing Data

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:
Explanation:

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

QUESTION 8
From the Cortana Intelligence Gallery, you deploy a solution.

You need to modify the solution.

What should you use?

A. Azure Stream Analytics
B. Microsoft Power BI Desktop
C. Azure Machine Learning Studio
D. R Tools for Visual Studio

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:
Explanation:

References: https://docs.microsoft.com/en-us/azure/machine-learning/studio/gallery-experiments

QUESTION 9
You are building an Azure Machine Learning workflow by using Azure Machine Learning Studio.

You create an Azure notebook that supports the Microsoft Cognitive Toolkit.

You need to ensure that the stochastic gradient descent (SGD) configuration maximizes the samples per second and supports parallel modeling that is managed by a parameter server.

Which SGD algorithm should you use?

A. DataParallelASGD
B. DataParallelSGD
C. ModelAveragingSGD
D. BlockMomentumSGD

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:

QUESTION 10

You are analyzing taxi trips in New York City. You leverage the Azure Data Factory to create data pipelines and to orchestrate data movement.

You plan to develop a predictive model for 170 million rows (37 GB) of raw data in Apache Hive by using Microsoft R Server to identify which factors contribute to the passenger tipping behavior.

All of the platforms that are used for the analysis are the same. Each worker node has eight processor cores and 26 GB of memory.

Which type of Azure HDInsight cluster should you use to produce results as quickly as possible?

A. Hadoop
B. HBase
C. Interactive Hive
D. Spark

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:
Explanation:

References: https://azure.microsoft.com/en-gb/blog/general-availability-of-hdinsight-interactive-query-blazing-fast-data-warehouse-style-queries-on-hyper-scale-data-2/

QUESTION 11

You plan to use the Data Science Virtual Machine for development, but you are unfamiliar with R scripts.

You need to generate R code for an experiment.

Which IDE should you use?

A. XgBoost
B. Rattle
C. Vowpal Wabbit
D. R Tools for Visual Studio

Correct Answer: B
Section: (none)

Explanation

Explanation/Reference:
Explanation:

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

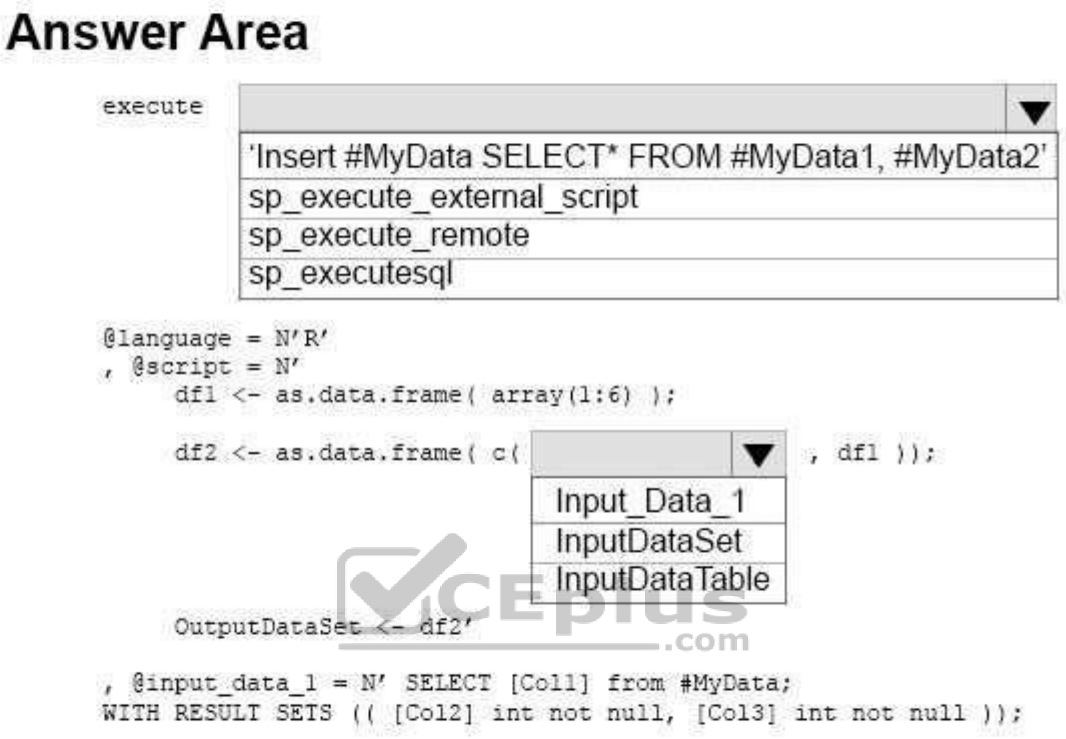
QUESTION 12
HOTSPOT

You need to use R code in a Transact-SQL statement to merge the repeating values 1 through 6 with Col1 in a table.

Which 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

```
execute
```

- 'Insert #MyData SELECT* FROM #MyData1, #MyData2'
- sp_execute_external_script
- sp_execute_remote
- sp_executesql

```
@language = N’R’
, @script = N’
     df1 <- as.data.frame( array(1:6) );

     df2 <- as.data.frame( c(              , df1 ));
```

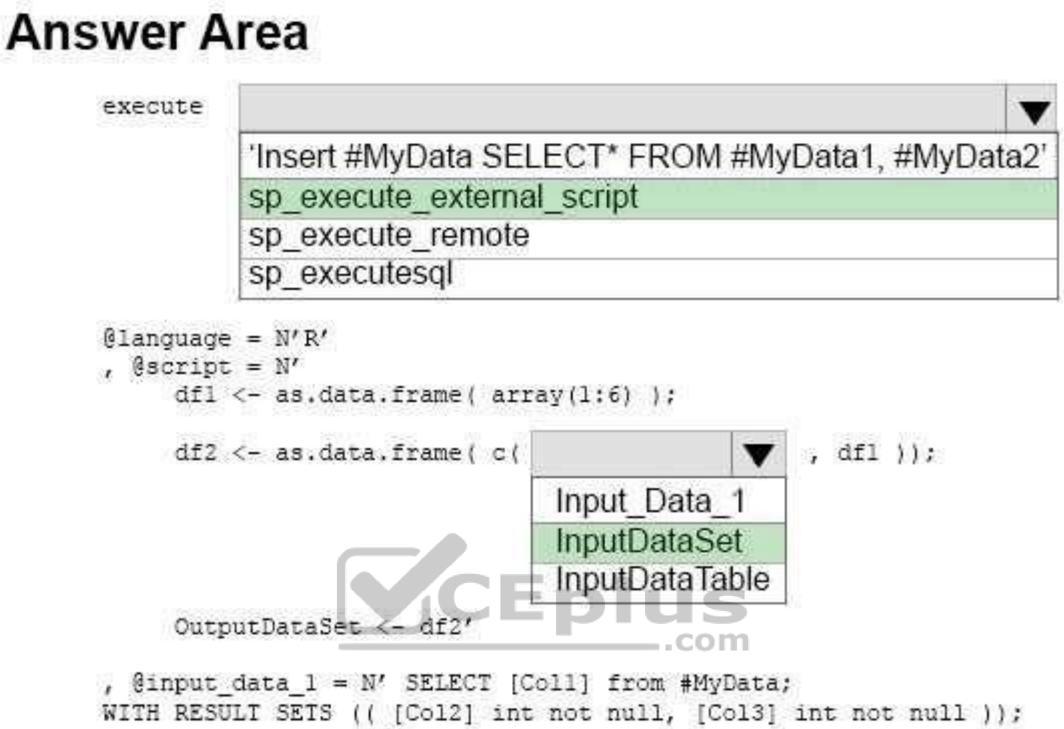
- Input_Data_1
- InputDataSet
- InputDataTable

```
     OutputDataSet <- df2’

, @input_data_1 = N’ SELECT [Col1] from #MyData;
WITH RESULT SETS (( [Col2] int not null, [Col3] int not null ));
```

Correct Answer:

Answer Area



```
execute
```

- 'Insert #MyData SELECT* FROM #MyData1, #MyData2'
- [✓ sp_execute_external_script]
- sp_execute_remote
- sp_executesql

```
@language = N'R'
, @script = N'
     df1 <- as.data.frame( array(1:6) );

     df2 <- as.data.frame( c(                , df1 ));
```

- Input_Data_1
- [✓ InputDataSet]
- InputDataTable

```
     OutputDataSet <- df2'

, @input_data_1 = N' SELECT [Col1] from #MyData;
WITH RESULT SETS (( [Col2] int not null, [Col3] int not null ));
```

Section: (none)
Explanation

Explanation/Reference:
References: https://docs.microsoft.com/en-us/sql/advanced-analytics/tutorials/rtsql-r-and-sql-data-types-and-data-objects

QUESTION 13
You are performing exploratory analysis of files that are encoded in a complex proprietary format. The format requires disk intensive access to several dependent files in HDFS.

You need to build an Azure Machine Learning model by using a canopy clustering algorithm. You must ensure that changes to proprietary file formats can be maintained by using the least amount of effort.

Which Machine Learning library should you use?

A. MicrosoftML
B. scikit-learn
C. SparkRD. Mahout

Correct Answer: D
Section: (none)
Explanation

Explanation/Reference:
Explanation:

QUESTION 14
Note: This question is part of a series of questions that use the same scenario. For your convenience, the scenario is repeated in each question. Each question presents a different goal and answer choices, but the text of the scenario is exactly the same in each question in this series.

You plan to create a predictive analytics solution for credit risk assessment and fraud prediction in Azure Machine Learning. The Machine Learning workspace for the solution will be shared with other users in your organization. You will add assets to projects and conduct experiments in the workspace.

The experiments will be used for training models that will be published to provide scoring from web services.

The experiment for fraud prediction will use Machine Learning modules and APIs to train the models and will predict probabilities in an Apache Hadoop ecosystem.

You plan to configure the resources for part of a workflow that will be used to preprocess data from files stored in Azure Blob storage. You plan to use Python to preprocess and store the data in Hadoop.

You need to get the data into Hadoop as quickly as possible.

Which three actions should you perform? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

A. Create an Azure virtual machine (VM), and then configure MapReduce on the VM.
B. Create an Azure HDInsight Hadoop cluster.
C. Create an Azure virtual machine (VM), and then install an IPython Notebook server.
D. Process the files by using Python to store the data to a Hadoop instance.
E. Create the Machine learning experiment, and then add an Execute Python Script module.

Correct Answer: BDE
Section: (none)
Explanation

Explanation/Reference:
Explanation:

QUESTION 15
DRAG DROP

Note: This question is part of a series of questions that use the same scenario. For your convenience, the scenario is repeated in each question. Each question presents a different goal and answer choices, but the text of the scenario is exactly the same in each question in this series.

You plan to create a predictive analytics solution for credit risk assessment and fraud prediction in Azure Machine Learning. The Machine Learning workspace for the solution will be shared with other users in your organization. You will add assets to projects and conduct experiments in the workspace.

The experiments will be used for training models that will be published to provide scoring from web services.

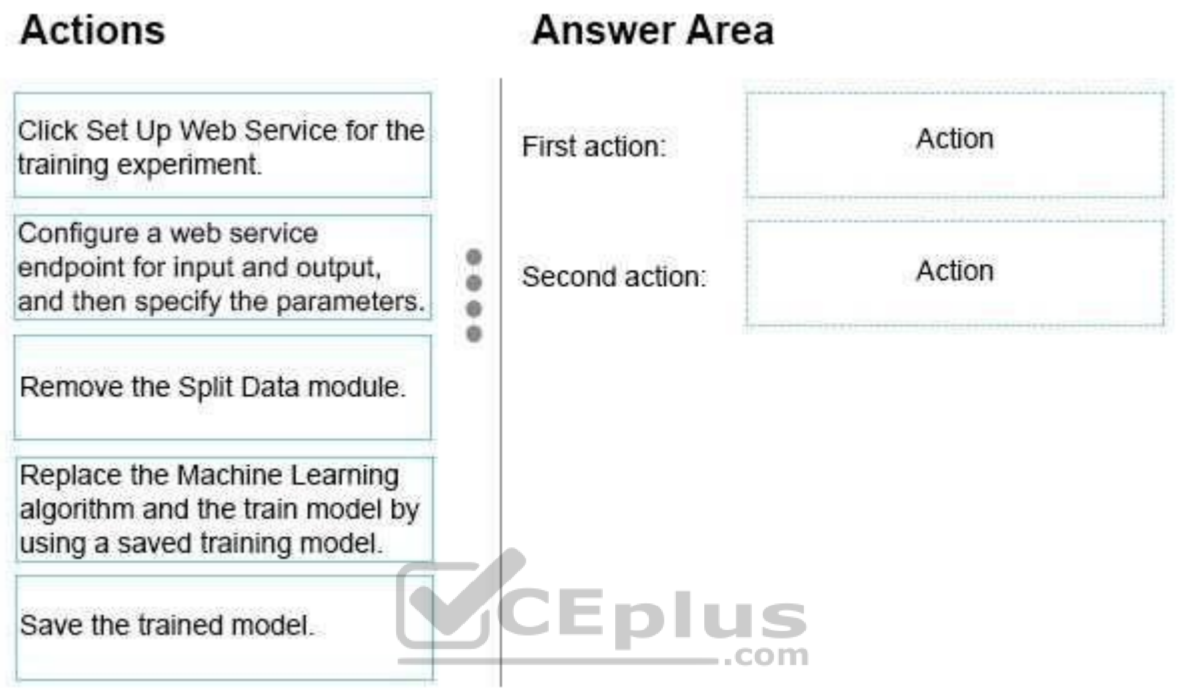
The experiment for fraud prediction will use Machine Learning modules and APIs to train the models and will predict probabilities in an Apache Hadoop ecosystem.

You finish training the model and are ready to publish a predictive web service that will provide the users with the ability to specify the data source and the save location of the results. The model includes a Split Data module.

Which two actions should you perform to convert the Machine Learning experiment to a predictive web service? To answer, drag the appropriate actions to the correct targets. Each action 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:



Actions

Click Set Up Web Service for the training experiment.

Configure a web service endpoint for input and output, and then specify the parameters.

Remove the Split Data module.

Replace the Machine Learning algorithm and the train model by using a saved training model.

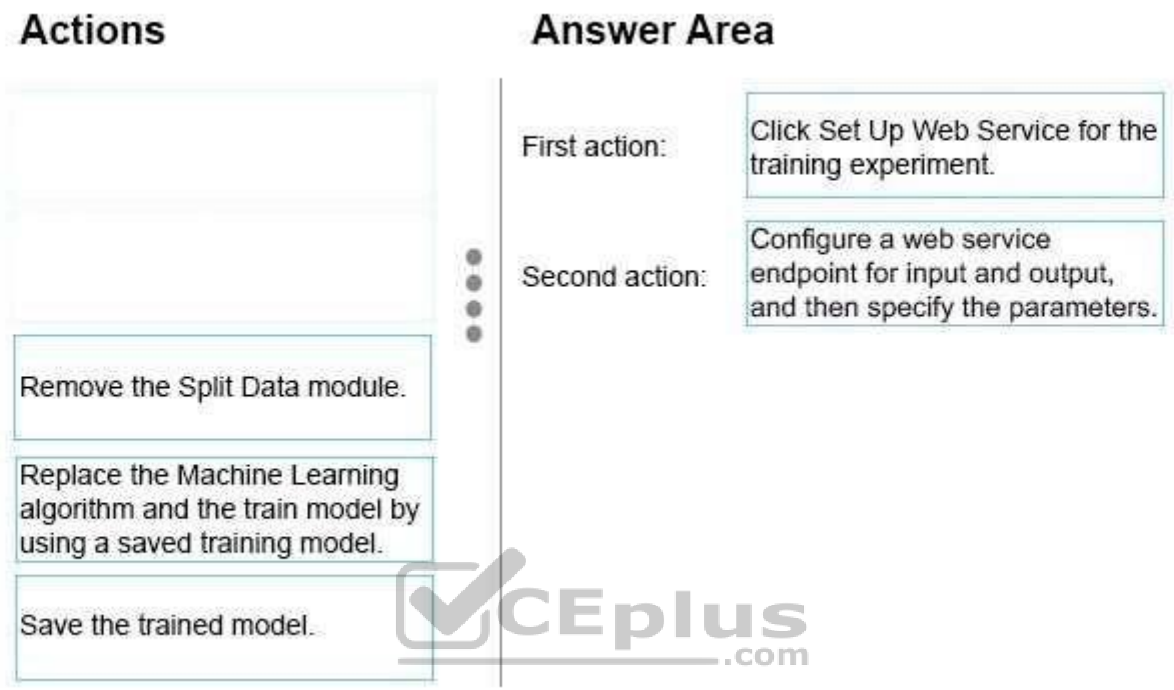
Save the trained model.

Answer Area

First action: Action

Second action: Action

Correct Answer:



Actions

Remove the Split Data module.

Replace the Machine Learning algorithm and the train model by using a saved training model.

Save the trained model.

Answer Area

First action: Click Set Up Web Service for the training experiment.

Second action: Configure a web service endpoint for input and output, and then specify the parameters.

Section: (none)
Explanation

Explanation/Reference:
References: https://docs.microsoft.com/en-us/azure/machine-learning/studio/convert-training-experiment-to-scoring-experiment

QUESTION 16
Note: This question is part of a series of questions that use the same scenario. For your convenience, the scenario is repeated in each question. Each question presents a different goal and answer choices, but the text of the scenario is exactly the same in each question in this series.

You plan to create a predictive analytics solution for credit risk assessment and fraud prediction in Azure Machine Learning. The Machine Learning workspace for the solution will be shared with other users in your organization. You will add assets to projects and conduct experiments in the workspace.

The experiments will be used for training models that will be published to provide scoring from web services.

The experiment for fraud prediction will use Machine Learning modules and APIs to train the models and will predict probabilities in an Apache Hadoop ecosystem.

You need to alter the list of columns that will be used for predicting fraud for an input web service endpoint. The columns from the original data source must be retained while running the Machine Learning experiment.

Which module should you add after the web service input module and before the prediction module?

A. Edit Metadata
B. Import Data
C. SMOTE
D. Select Columns in Dataset

Correct Answer: D
Section: (none)
Explanation

Explanation/Reference:
Explanation:

QUESTION 17
You have an Azure Machine Learning experiment.

You discover that a model causes many errors in a production dataset. The model causes only few errors in the training data.

What is the cause of the errors?

A. overfitting
B. generalization
C. underfitting
D. a simple predictor

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:
Explanation: