

DP-100.25q

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

DP-100



Website: <https://vceplus.com>

VCE to PDF Converter: <https://vceplus.com/vce-to-pdf/>

Facebook: <https://www.facebook.com/VCE.For.All.VN/>

Twitter : https://twitter.com/VCE_Plus

<https://vceplus.com/>

Designing and Implementing a Data Science Solution on Azure (beta)

Question Set 1

QUESTION 1

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

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

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

NOTE: Each correct selection is worth one point.

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

Correct Answer: CE

Section: [none]

Explanation

Explanation/Reference:

Explanation:

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





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

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

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

QUESTION 2

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

The environment must support the following requirements:

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

- enable scaling across a cluster of machines

You need to create the environment.

What should you do?

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

Correct Answer: B

Section: [none]

Explanation

Explanation/Reference:

Explanation:

In Azure Databricks, we can create two different types of clusters.

- Standard, these are the default clusters and can be used with Python, R, Scala and SQL
- High-concurrency

Azure Databricks is fully integrated with Azure Data Factory.

Incorrect Answers:

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

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

QUESTION 3

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.



<https://vceplus.com/>

Which environment should you use?

- A. Azure Machine Learning Service
- B. Azure Machine Learning Studio
- C. Azure Databricks
- D. Azure Kubernetes Service (AKS)

Correct Answer: A

Section: [none]

Explanation

Explanation/Reference:

Explanation:

The Data Science Virtual Machine (DSVM) is a customized VM image on Microsoft's Azure cloud built specifically for doing data science. Caffe2 and Chainer are supported by DSVM.

DSVM integrates with Azure Machine Learning.

Incorrect Answers:

B: Use Machine Learning Studio when you want to experiment with machine learning models quickly and easily, and the built-in machine learning algorithms are sufficient for your solutions.

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

QUESTION 4

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]

Explanation

Explanation/Reference:

Incorrect Answers:

A, C: PostgreSQL (CentOS) is only available in the Linux Edition.

B: The Azure Geo AI Data Science VM (Geo-DSVM) delivers geospatial analytics capabilities from Microsoft's Data Science VM. Specifically, this VM extends the AI and data science toolkits in the Data Science VM by adding ESRI's market-leading ArcGIS Pro Geographic Information System.

D: DLVM is a template on top of DSVM image. In terms of the packages, GPU drivers etc are all there in the DSVM image. Mostly it is for convenience during creation where we only allow DLVM to be created on GPU VM instances on Azure.

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

QUESTION 5

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]

Explanation

Explanation/Reference:

Explanation:

Azure Cognitive Services expand on Microsoft's evolving portfolio of machine learning APIs and enable developers to easily add cognitive features – such as emotion and video detection; facial, speech, and vision recognition; and speech and language understanding – into their applications. The goal of Azure Cognitive Services is to help developers create applications that can see, hear, speak, understand, and even begin to reason. The catalog of services within Azure Cognitive Services can be categorized into five main pillars - Vision, Speech, Language, Search, and Knowledge.

References: <https://docs.microsoft.com/en-us/azure/cognitive-services/welcome>



QUESTION 6

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

You need to transfer the data into Azure Blob Storage.

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

NOTE: Each correct selection is worth one point.

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

Correct Answer: BCD

Section: [none]

Explanation

Explanation/Reference:

Explanation:

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

Azure Storage-Explorer

AzCopy

Python

SSIS

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

QUESTION 7

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

You need to format the data for the Weka environment.

Which module should you use?

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



Correct Answer: C

Section: [none]

Explanation

Explanation/Reference:

Explanation:

Use the Convert to ARFF module in Azure Machine Learning Studio, to convert datasets and results in Azure Machine Learning to the attribute-relation file format used by the Weka toolset. This format is known as ARFF.

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

References:

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

Testlet 1

Case study

Overview

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

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

Current environment

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

Penalty detection and sentiment

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

Advertisements

During the initial weeks in production, the following was observed:

- Ad response rated declined.

- Drops were not consistent across ad styles.
- The distribution of features across training and production data are not consistent

Analysis shows that, of the 100 numeric features on user location and behavior, the 47 features that come from location sources are being used as raw features. A suggested experiment to remedy the bias and variance issue is to engineer 10 linearly uncorrelated features.

- Initial data discovery shows a wide range of densities of target states in training data used for crowd sentiment models.
 - All penalty detection models show inference phases using a Stochastic Gradient Descent (SGD) are running too slow.
 - Audio samples show that the length of a catch phrase varies between 25%-47% depending on region
 - The performance of the global penalty detection models shows lower variance but higher bias when comparing training and validation sets. Before implementing any feature changes, you must confirm the bias and variance using all training and validation cases.
 - Ad response models must be trained at the beginning of each event and applied during the sporting event.
 - Market segmentation models must optimize for similar ad response history.
 - Sampling must guarantee mutual and collective exclusively between local and global segmentation models that share the same features. ▪
- Local market segmentation models will be applied before determining a user's propensity to respond to an advertisement.
- Ad response models must support non-linear boundaries of features.
 - The ad propensity model uses a cut threshold is 0.45 and retrains occur if weighted Kappa deviated from 0.1 +/- 5%. ▪

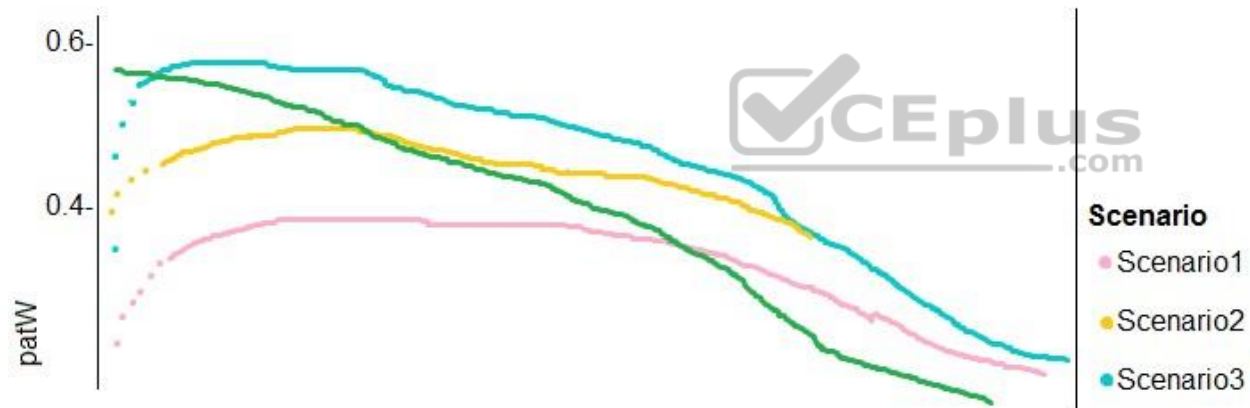
The ad propensity model uses cost factors shown in the following diagram:

		Actual	
		1	0
Predicted	0	1	2
	1	2	1

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

		Actual	
		1	0
Predicted	0	1	5
	1	5	1

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



QUESTION 1

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

Which normalization type should you use?

- A. Streaming
- B. Weight

C. Batch D. Cosine

Correct Answer: C

Section: [none]

Explanation

Explanation/Reference:

Explanation:

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

In CNTK, custom networks are defined using the BrainScriptNetworkBuilder and described in the CNTK network description language "BrainScript."

Scenario:

Local penalty detection models must be written by using BrainScript.

References:

<https://docs.microsoft.com/en-us/cognitive-toolkit/post-batch-normalization-statistics>



Question Set 2

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]

Explanation



Explanation/Reference:

References:

<https://notebooks.azure.com/>

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]

Explanation

Explanation/Reference:

Explanation:

Replace using MICE: For each missing value, this option assigns a new value, which is calculated by using a method described in the statistical literature as "Multivariate Imputation using Chained Equations" or "Multiple Imputation by Chained Equations". With a multiple imputation method, each variable with missing data is modeled conditionally using the other variables in the data before filling in the missing values.

Note: Multivariate imputation by chained equations (MICE), sometimes called “fully conditional specification” or “sequential regression multiple imputation” has emerged in the statistical literature as one principled method of addressing missing data. Creating multiple imputations, as opposed to single imputations, accounts for the statistical uncertainty in the imputations. In addition, the chained equations approach is very flexible and can handle variables of varying types (e.g., continuous or binary) as well as complexities such as bounds or survey skip patterns.

References: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3074241/> <https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/clean-missing-data>

QUESTION 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]

Explanation

Explanation/Reference:

Explanation:

Use the Multiple Imputation by Chained Equations (MICE) method.

References: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3074241/> <https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/clean-missing-data>

QUESTION 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]

Explanation

Explanation/Reference:

Explanation:

Use the Multiple Imputation by Chained Equations (MICE) method.

References: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3074241/> <https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/clean-missing-data>

QUESTION 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]

Explanation

Explanation/Reference:

Explanation:

Use the Entropy MDL binning mode which has a target column.

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

QUESTION 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]

Explanation

Explanation/Reference:

Explanation:

Use the Entropy MDL binning mode which has a target column.

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

QUESTION 7

You are solving a classification task.

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

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

Which value should you use?

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

Correct Answer: C

Section: [none]

Explanation

Explanation/Reference:

Explanation:

Leave One Out (LOO) cross-validation

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

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

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

QUESTION 8

You use Azure Machine Learning Studio to build a machine learning experiment.

You need to divide data into two distinct datasets.

Which module should you use?

- A. Assign Data to Clusters
- B. Load Trained Model
- C. Partition and Sample
- D. Tune Model-Hyperparameters

Correct Answer: C

Section: [none]

Explanation



Explanation/Reference:

Explanation:

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

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

QUESTION 9

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

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

Which parameter should you use?

- A. Replace with mean
- B. Remove entire column

- C. Remove entire row
- D. Hot Deck

Correct Answer: B

Section: [none]

Explanation

Explanation/Reference:

Explanation:

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

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



Testlet 1

Case study

Overview

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

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

Current environment

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

Penalty detection and sentiment

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

Advertisements

During the initial weeks in production, the following was observed:

- Ad response rated declined.

- Drops were not consistent across ad styles.
- The distribution of features across training and production data are not consistent

Analysis shows that, of the 100 numeric features on user location and behavior, the 47 features that come from location sources are being used as raw features. A suggested experiment to remedy the bias and variance issue is to engineer 10 linearly uncorrelated features.

- Initial data discovery shows a wide range of densities of target states in training data used for crowd sentiment models.
 - All penalty detection models show inference phases using a Stochastic Gradient Descent (SGD) are running too slow.
 - Audio samples show that the length of a catch phrase varies between 25%-47% depending on region
 - The performance of the global penalty detection models shows lower variance but higher bias when comparing training and validation sets. Before implementing any feature changes, you must confirm the bias and variance using all training and validation cases.
 - Ad response models must be trained at the beginning of each event and applied during the sporting event.
 - Market segmentation models must optimize for similar ad response history.
 - Sampling must guarantee mutual and collective exclusively between local and global segmentation models that share the same features. ▪
- Local market segmentation models will be applied before determining a user's propensity to respond to an advertisement.
- Ad response models must support non-linear boundaries of features.
 - The ad propensity model uses a cut threshold is 0.45 and retrain occurs if weighted Kappa deviated from 0.1 +/- 5%. ▪

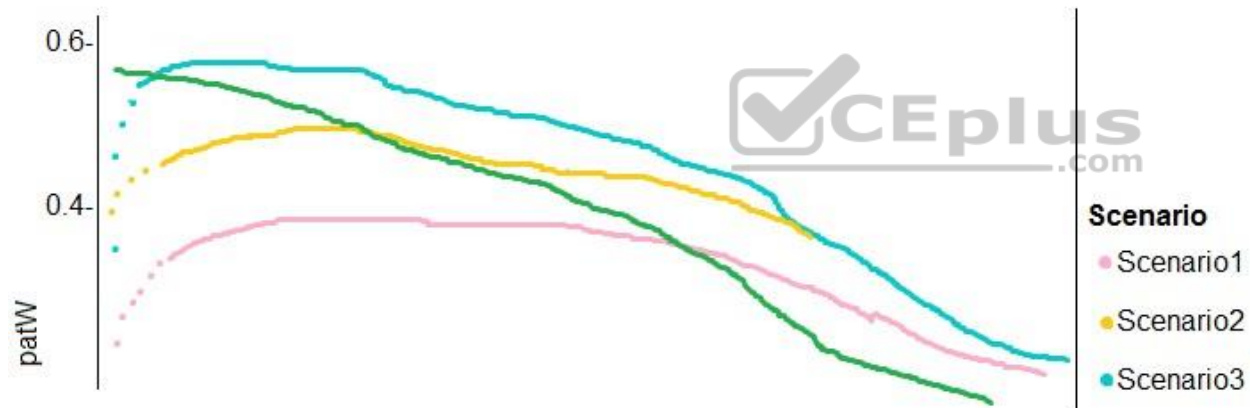
The ad propensity model uses cost factors shown in the following diagram:

		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 feature engineering strategy for the crowd sentiment local models.

What should you do?

- Apply an analysis of variance (ANOVA).
- Apply a Pearson correlation coefficient.

- C. Apply a Spearman correlation coefficient.
- D. Apply a linear discriminant analysis.

Correct Answer: D

Section: [none]

Explanation

Explanation/Reference:

Explanation:

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

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

Scenario:

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

Incorrect Answers:

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

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

References:

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

<https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/compute-linear-correlation> **Testlet 2**

Case study

Overview

You are a data scientist for Fabrikam Residences, a company specializing in quality private and commercial property in the United States. Fabrikam Residences is considering expanding into Europe and has asked you to investigate prices for private residences in major European cities. You use Azure Machine Learning

Studio to measure the median value of properties. You produce a regression model to predict property prices by using the Linear Regression and Bayesian Linear Regression modules. **Datasets**

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

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

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

Dataset issues

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

Columns in each dataset contain missing and null values. The dataset also contains many outliers. The Age column has a high proportion of outliers. You need to remove the rows that have outliers in the Age column. The MedianValue and AvgRoomsInHouse columns both hold data in numeric format. You need to select a feature selection algorithm to analyze the relationship between the two columns in more detail.

Model fit

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

Experiment requirements

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

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

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

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

Model training

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

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

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

Testing

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

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

Data visualization

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

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

QUESTION 1

You need to select a feature extraction method.

Which method should you use?

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

Correct Answer: C

Section: [none]

Explanation

Explanation/Reference:

Explanation:

In statistics, the Kendall rank correlation coefficient, commonly referred to as Kendall's tau coefficient (after the Greek letter τ), is a statistic used to measure the ordinal association between two measured quantities.

It is a supported method of the Azure Machine Learning Feature selection.

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

References:

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



<https://vceplus.com/>



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 using Azure Machine Learning Studio to perform feature engineering on a dataset.

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

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

Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: A

Section: [none]

Explanation



Explanation/Reference:

Explanation:

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

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

QUESTION 2

You are building a regression model for estimating the number of calls during an event.

You need to determine whether the feature values achieve the conditions to build a Poisson regression model.

Which two conditions must the feature set contain? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

- A. The label data must be a negative value.
- B. The label data must be whole numbers.
- C. The label data must be non-discrete.
- D. The label data must be a positive value.
- E. The label data can be positive or negative.

Correct Answer: BD

Section: [none]

Explanation

Explanation/Reference:

Explanation:

Poisson regression is intended for use in regression models that are used to predict numeric values, typically counts. Therefore, you should use this module to create your regression model only if the values you are trying to predict fit the following conditions:

- The response variable has a Poisson distribution.
- Counts cannot be negative. The method will fail outright if you attempt to use it with negative labels.
- A Poisson distribution is a discrete distribution; therefore, it is not meaningful to use this method with non-whole numbers.

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

QUESTION 3

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]

Explanation

Explanation/Reference:

Explanation:

Typical metadata changes might include marking columns as features.

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



Testlet 1

Case study

Overview

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

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

Current environment

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

Penalty detection and sentiment

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

Advertisements

During the initial weeks in production, the following was observed:

- Ad response rated declined.

- Drops were not consistent across ad styles.
- The distribution of features across training and production data are not consistent

Analysis shows that, of the 100 numeric features on user location and behavior, the 47 features that come from location sources are being used as raw features. A suggested experiment to remedy the bias and variance issue is to engineer 10 linearly uncorrelated features.

- Initial data discovery shows a wide range of densities of target states in training data used for crowd sentiment models.
 - All penalty detection models show inference phases using a Stochastic Gradient Descent (SGD) are running too slow.
 - Audio samples show that the length of a catch phrase varies between 25%-47% depending on region
 - The performance of the global penalty detection models shows lower variance but higher bias when comparing training and validation sets. Before implementing any feature changes, you must confirm the bias and variance using all training and validation cases.
 - Ad response models must be trained at the beginning of each event and applied during the sporting event.
 - Market segmentation models must optimize for similar ad response history.
 - Sampling must guarantee mutual and collective exclusively between local and global segmentation models that share the same features. ▪
- Local market segmentation models will be applied before determining a user's propensity to respond to an advertisement.
- Ad response models must support non-linear boundaries of features.
 - The ad propensity model uses a cut threshold is 0.45 and retrain occurs if weighted Kappa deviated from 0.1 +/- 5%. ▪

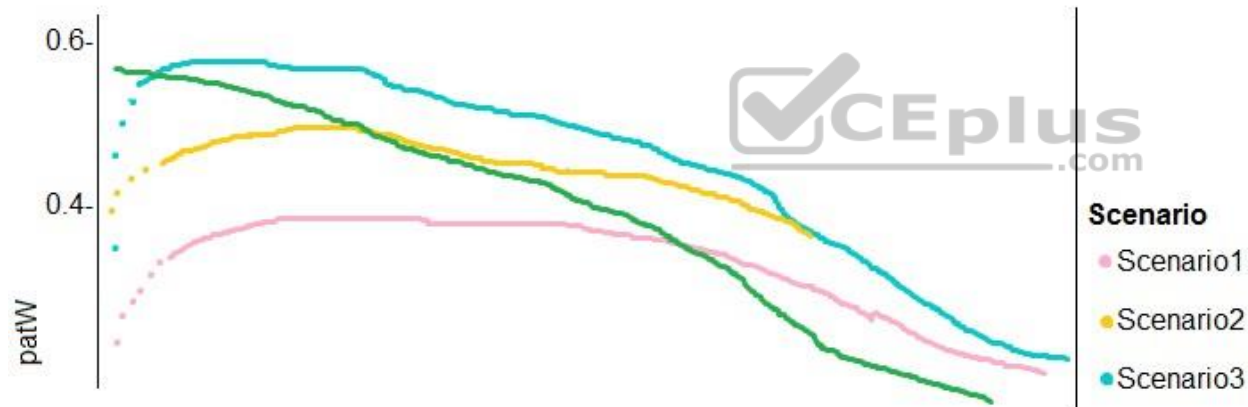
The ad propensity model uses cost factors shown in the following diagram:

		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 model development strategy to determine a user's tendency to respond to an ad.

Which technique should you use?

- Use a Relative Expression Split module to partition the data based on centroid distance.
- Use a Relative Expression Split module to partition the data based on distance travelled to the event.

- C. Use a Split Rows module to partition the data based on distance travelled to the event.
- D. Use a Split Rows module to partition the data based on centroid distance.

Correct Answer: A

Section: [none]

Explanation

Explanation/Reference:

Explanation:

Split Data partitions the rows of a dataset into two distinct sets.

The Relative Expression Split option in the Split Data module of Azure Machine Learning Studio is helpful when you need to divide a dataset into training and testing datasets using a numerical expression.

Relative Expression Split: Use this option whenever you want to apply a condition to a number column. The number could be a date/time field, a column containing age or dollar amounts, or even a percentage. For example, you might want to divide your data set depending on the cost of the items, group people by age ranges, or separate data by a calendar date.

Scenario:

Local market segmentation models will be applied before determining a user's propensity to respond to an advertisement.

The distribution of features across training and production data are not consistent

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

QUESTION 2

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

Which technique should you use?

- A. Set the threshold to **0.5** and retrain if weighted Kappa deviates +/- 5% from 0.45.
- B. Set the threshold to **0.05** and retrain if weighted Kappa deviates +/- 5% from 0.5.
- C. Set the threshold to **0.2** and retrain if weighted Kappa deviates +/- 5% from 0.6.
- D. Set the threshold to **0.75** and retrain if weighted Kappa deviates +/- 5% from 0.15.

Correct Answer: A

Section: [none]

Explanation

Explanation/Reference:

Explanation:

Scenario:

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



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

Question Set 2

QUESTION 1

You are a data scientist creating a linear regression model.

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

Which metric should you review?

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

Correct Answer: B

Section: [none]

Explanation

Explanation/Reference:

Explanation:

Coefficient of determination, often referred to as R^2 , represents the predictive power of the model as a value between 0 and 1. Zero means the model is random (explains nothing); 1 means there is a perfect fit. However, caution should be used in interpreting R^2 values, as low values can be entirely normal and high values can be suspect.

Incorrect Answers:

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

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

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

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

References:

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



<https://vceplus.com/>

