

**DP-100.46q**

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://twitter.com/VCE_Plus)

<https://vceplus.com/>

### **Designing and Implementing a Data Science Solution on Azure**

#### **Question Set 1**

#### **QUESTION 1**

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.



<https://vceplus.com/> 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 2

You plan to build a team data science environment. Data for training models in machine learning pipelines will be over 20 GB in size.

You have the following requirements:

- Models must be built using Caffe2 or Chainer frameworks.
- Data scientists must be able to use a data science environment to build the machine learning pipelines and train models on their personal devices in both connected and disconnected network environments.

Personal devices must support updating machine learning pipelines when connected to a network.

You need to select a data science environment.

Which environment should you use?

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



**Correct Answer:** A

**Section:** [none]

**Explanation**

**Explanation/Reference:**

Explanation:

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

DSVM integrates with Azure Machine Learning.

Incorrect Answers:

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

References:

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

### QUESTION 3

You are implementing a machine learning model to predict stock prices.

The model uses a PostgreSQL database and requires GPU processing.

You need to create a virtual machine that is pre-configured with the required tools.

What should you do?

- A. Create a Data Science Virtual Machine (DSVM) Windows edition.
- B. Create a Geo AI Data Science Virtual Machine (Geo-DSVM) Windows edition.
- C. Create a Deep Learning Virtual Machine (DLVM) Linux edition.
- D. Create a Deep Learning Virtual Machine (DLVM) Windows edition.

**Correct Answer:** A

**Section:** [none]

**Explanation**

**Explanation/Reference:**

Explanation:

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

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

Incorrect Answers:

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

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

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

### QUESTION 4

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 5**

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 6**

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

You need to format the data for the Weka environment.

Which module should you use?

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

**Correct Answer:** C

**Section:** [none]

**Explanation**

**Explanation/Reference:**

Explanation:

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

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

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

### QUESTION 7

You plan to create a speech recognition deep learning model.

The model must support the latest version of Python.

You need to recommend a deep learning framework for speech recognition to include in the Data Science Virtual Machine (DSVM).

What should you recommend?

- A. Rattle
- B. TensorFlow
- C. Weka
- D. Scikit-learn



**Correct Answer:** B

**Section:** [none]

**Explanation**

**Explanation/Reference:**

Explanation:

TensorFlow is an open source library for numerical computation and large-scale machine learning. It uses Python to provide a convenient front-end API for building applications with the framework

TensorFlow can train and run deep neural networks for handwritten digit classification, image recognition, word embeddings, recurrent neural networks, sequenceto-sequence models for machine translation, natural language processing, and PDE (partial differential equation) based simulations.

Incorrect Answers:

A: Rattle is the R analytical tool that gets you started with data analytics and machine learning.

C: Weka is used for visual data mining and machine learning software in Java.

D: Scikit-learn is one of the most useful library for machine learning in Python. It is on NumPy, SciPy and matplotlib, this library contains a lot of effiecient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction.

Reference:

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

### QUESTION 8

You plan to use a Deep Learning Virtual Machine (DLVM) to train deep learning models using Compute Unified Device Architecture (CUDA) computations.

You need to configure the DLVM to support CUDA.

What should you implement?

- A. Solid State Drives (SSD)
- B. Computer Processing Unit (CPU) speed increase by using overclocking
- C. Graphic Processing Unit (GPU)
- D. High Random Access Memory (RAM) configuration
- E. Intel Software Guard Extensions (Intel SGX) technology

**Correct Answer: C**

**Section: [none]**

**Explanation**

**Explanation/Reference:**

Explanation:

A Deep Learning Virtual Machine is a pre-configured environment for deep learning using GPU instances.

References: <https://azuremarketplace.microsoft.com/en-au/marketplace/apps/microsoft-ads.dsvm-deep-learning>

### QUESTION 9

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

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

What should you create?

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





**Correct Answer:** E

**Section:** [none]

**Explanation**

**Explanation/Reference:**

Explanation:

Caffe2 and PyTorch is supported by Data Science Virtual Machine for Linux.

Microsoft offers Linux editions of the DSVM on Ubuntu 16.04 LTS and CentOS 7.4.

Only the DSVM on Ubuntu is preconfigured for Caffe2 and PyTorch.

Incorrect Answers:

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

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



## 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 scaling strategy for the local penalty detection data.

Which normalization type should you use?

- A. Streaming
- B. Weight

C. Batch



<https://vceplus.com/>

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>

**Testlet 2**

**Case study**

This is a case study. Case studies are not timed separately. You can use as much exam time as you would like to complete each case. However, there may be additional case studies and sections on this exam. You must manage your time to ensure that you are able to complete all questions included on this exam in the time provided.

To answer the questions included in a case study, you will need to reference information that is provided in the case study. Case studies might contain exhibits and other resources that provide more information about the scenario that is described in the case study. Each question is independent of the other questions in this case study.

At the end of this case study, a review screen will appear. This screen allows you to review your answers and to make changes before you move to the next section of the exam. After you begin a new section, you cannot return to this section.

### **To start the case study**

To display the first question in this case study, click the Next button. Use the buttons in the left pane to explore the content of the case study before you answer the questions. Clicking these buttons displays information such as business requirements, existing environment, and problem statements. If the case study has an All Information tab, note that the information displayed is identical to the information displayed on the subsequent tabs. When you are ready to answer a question, click the Question button to return to the question.

### **Overview**

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

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

### **Datasets**

There are two datasets in CSV format that contain property details for two cities, London and Paris. You add both files to Azure Machine Learning Studio as separate datasets to the starting point for an experiment. Both datasets contain the following columns:

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

An initial investigation shows that the datasets are identical in structure apart from the MedianValue column. The smaller Paris dataset contains the MedianValue in text format, whereas the larger London dataset contains the MedianValue in numerical format.

### Data issues

#### Missing values

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

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

#### Model fit

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

## **Experiment requirements**

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

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

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

## **Model training**

### **Permutation Feature Importance**

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

### **Hyperparameters**

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

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

## **Testing**

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

## **Cross-validation**

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

## **Linear regression module**

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

## **Data visualization**



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

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

### **QUESTION 1**

#### **DRAG DROP**

You need to 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****Answer Area**

Score Matchbox Recommender

Apply Transformation

Evaluate Recommender

Evaluate Model



Train Model



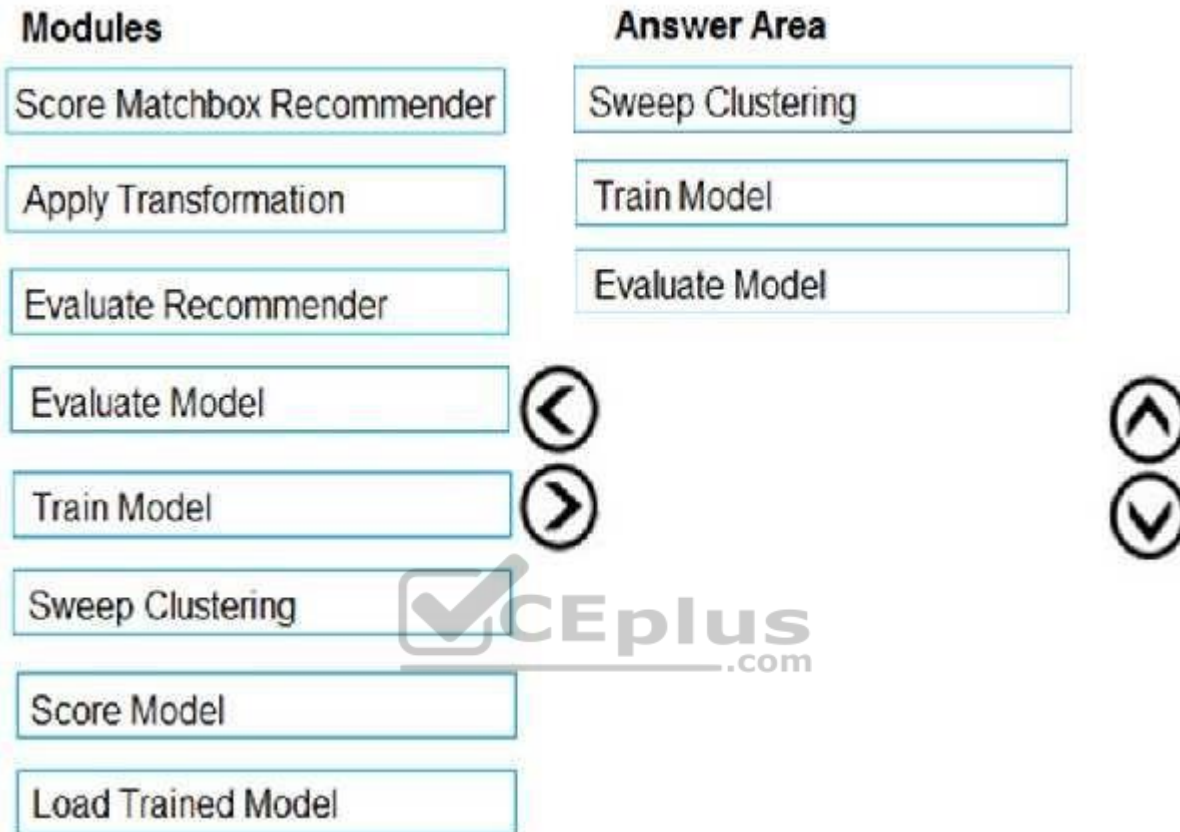
Sweep Clustering

Score Model

Load Trained Model



**Correct Answer:**



**Section:** [none]

**Explanation**

**Explanation/Reference:**

Explanation:

Step 1: Sweep Clustering

Start by using the "Tune Model Hyperparameters" module to select the best sets of parameters for each of the models we're considering.

One of the interesting things about the "Tune Model Hyperparameters" module is that it not only outputs the results from the Tuning, it also outputs the Trained Model.

Step 2: Train Model

Step 3: Evaluate Model

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

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

References: <http://breaking-bi.blogspot.com/2017/01/azure-machine-learning-model-evaluation.html>

## QUESTION 2

You need to visually identify whether outliers exist in the Age column and quantify the outliers before the outliers are removed.

Which three Azure Machine Learning Studio modules should you use? Each correct answer presents part of the solution.

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

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



**Correct Answer:** ABC

**Section:** [none]

**Explanation**

**Explanation/Reference:**

Explanation:

B: To have a global view, the summarize data module can be used. Add the module and connect it to the data set that needs to be visualized.

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

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

You can use the Clip Values module in Azure Machine Learning Studio, to identify and optionally replace data values that are above or below a specified threshold. This is useful when you want to remove outliers or replace them with a mean, a constant, or other substitute value.

References: <https://blogs.msdn.microsoft.com/azuredev/2017/05/27/data-cleansing-tools-in-azure-machine-learning/> <https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/clip-values>



### Question Set 3

#### QUESTION 1

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

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

You are analyzing a numerical dataset which contains missing values in several columns.

You must clean the missing values using an appropriate operation without affecting the dimensionality of the feature set.

You need to analyze a full dataset to include all values.

Solution: 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 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: 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 3

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

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

You are a data scientist using Azure Machine Learning Studio.

You need to normalize values to produce an output column into bins to predict a target column.

Solution: Apply 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 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 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 5**

You are solving a classification task.



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

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

Which value should you use?

- A.  $k=0.5$
- B.  $k=0.01$
- C.  $k=5$
- D.  $k=1$

**Correct Answer:** C

**Section:** [none]

**Explanation**

**Explanation/Reference:**

Explanation:

Leave One Out (LOO) cross-validation

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

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

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

#### **QUESTION 6**

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

You need to divide data into two distinct datasets.

Which module should you use?

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

**Correct Answer:** C

**Section:** [none]

**Explanation**

**Explanation/Reference:**

Explanation:

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

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

**QUESTION 7**

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

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

Which parameter should you use?

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



**Correct Answer: C**

**Section: [none]**

**Explanation**

**Explanation/Reference:**

Explanation:

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

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

**QUESTION 8**

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
Threshold Filter	Remove potential duplicates from a dataset.	Remove Duplicate Rows

**Section:** [none]

**Explanation**

**Explanation/Reference:**

Explanation:

Box 1: Clean Missing Data

Box 2: SMOTE

Use the SMOTE module in Azure Machine Learning Studio to increase the number of underrepresented cases in a dataset used for machine learning. SMOTE is a better way of increasing the number of rare cases than simply duplicating existing cases.

Box 3: Convert to Indicator Values

Use the Convert to Indicator Values module in Azure Machine Learning Studio. The purpose of this module is to convert columns that contain categorical values into a series of binary indicator columns that can more easily be used as features in a machine learning model. Box 4: Remove Duplicate Rows

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

**QUESTION 9**

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

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

You are 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 10

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

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

You are a data scientist using Azure Machine Learning Studio.

You need to normalize values to produce an output column into bins to predict a target column.

Solution: Apply a Quantiles normalization with a QuantileIndex normalization.

Does the solution meet the goal?

- A. Yes
- B. No

**Correct Answer:** B

**Section:** [none]

**Explanation**

**Explanation/Reference:**

Explanation:

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

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

#### QUESTION 11

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

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

You are creating a new experiment in Azure Machine Learning Studio.

One class has a much smaller number of observations than the other classes in the training set.

You need to select an appropriate data sampling strategy to compensate for the class imbalance.

Solution: You use the Scale and Reduce sampling mode.

Does the solution meet the goal?

- A. Yes
- B. No

**Correct Answer:** B

**Section:** [none]

**Explanation**

**Explanation/Reference:**

Explanation:

Instead use the Synthetic Minority Oversampling Technique (SMOTE) sampling mode.

Note: SMOTE is used to increase the number of underrepresented cases in a dataset used for machine learning. SMOTE is a better way of increasing the number of rare cases than simply duplicating existing cases.

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

## QUESTION 12

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

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

You are creating a new experiment in Azure Machine Learning Studio.

One class has a much smaller number of observations than the other classes in the training set.

You need to select an appropriate data sampling strategy to compensate for the class imbalance.

Solution: You use the Synthetic Minority Oversampling Technique (SMOTE) sampling mode.

Does the solution meet the goal?

- A. Yes
- B. No

**Correct Answer:** A

**Section:** [none]

**Explanation**

**Explanation/Reference:**

Explanation:

SMOTE is used to increase the number of underrepresented cases in a dataset used for machine learning. SMOTE is a better way of increasing the number of rare cases than simply duplicating existing cases.

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

**QUESTION 13**

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

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

You are creating a new experiment in Azure Machine Learning Studio.

One class has a much smaller number of observations than the other classes in the training set.

You need to select an appropriate data sampling strategy to compensate for the class imbalance.

Solution: You use the Stratified split for the sampling mode.

Does the solution meet the goal?

- A. Yes
- B. No

**Correct Answer:** B

**Section:** [none]

**Explanation**



**Explanation/Reference:**

Explanation:

Instead use the Synthetic Minority Oversampling Technique (SMOTE) sampling mode.

Note: SMOTE is used to increase the number of underrepresented cases in a dataset used for machine learning. SMOTE is a better way of increasing the number of rare cases than simply duplicating existing cases.

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

**QUESTION 14**

You are creating a machine learning model.

You need to identify outliers in the data.

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



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

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

**Correct Answer:** BE

**Section:** [none]

**Explanation**

**Explanation/Reference:**

Explanation:

The box-plot algorithm can be used to display outliers.

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

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

#### **QUESTION 15**

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

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

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

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

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

**Correct Answer:** BE

**Section: [none]**

**Explanation**

**Explanation/Reference:**

Explanation:

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

E: Summarize Data statistics are useful when you want to understand the characteristics of the complete dataset. For example, you might need to know:

How many missing values are there in each column?

How many unique values are there in a feature column?

What is the mean and standard deviation for each column?

The module calculates the important scores for each column, and returns a row of summary statistics for each variable (data column) provided as input.

Incorrect Answers:

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

C: With Python, you can perform tasks that aren't currently supported by existing Studio modules such as:

Visualizing data using matplotlib

Using Python libraries to enumerate datasets and models in your workspace

Reading, loading, and manipulating data from sources not supported by the Import Data module

D: The purpose of the Convert to Indicator Values module is to convert columns that contain categorical values into a series of binary indicator columns that can more easily be used as features in a machine learning model.

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

## QUESTION 16

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

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

You are analyzing a numerical dataset which contains missing values in several columns.

You must clean the missing values using an appropriate operation without affecting the dimensionality of the feature set.

You need to analyze a full dataset to include all values.

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

Does the solution meet the goal?

- A. Yes
- B. No

**Correct Answer:** B

**Section:** [none]

**Explanation**

**Explanation/Reference:**

Explanation:

Instead use the Multiple Imputation by Chained Equations (MICE) method.

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

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

References: <https://methods.sagepub.com/reference/encyc-of-research-design/n211.xml> <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3074241/>

## Question Set 1

### QUESTION 1

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

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

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

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

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

**Correct Answer:** BD

**Section:** [none]

**Explanation**

**Explanation/Reference:**

Explanation:

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

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

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

## QUESTION 2

You are performing feature engineering on a dataset.

You must add a feature named CityName and populate the column value with the text **London**.

You need to add the new feature to the dataset.

Which Azure Machine Learning Studio module should you use?

- A. Edit Metadata
- B. Filter Based Feature Selection
- C. Execute Python Script
- D. Latent Dirichlet Allocation

**Correct Answer:** A

**Section:** [none]

**Explanation**

**Explanation/Reference:**

Explanation:

Typical metadata changes might include marking columns as features.

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

**QUESTION 3**

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

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

You need to produce the distribution.



<https://vcceplus.com/>

Which type of distribution should you produce?

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

**Correct Answer:** D

**Section:** [none]

**Explanation**

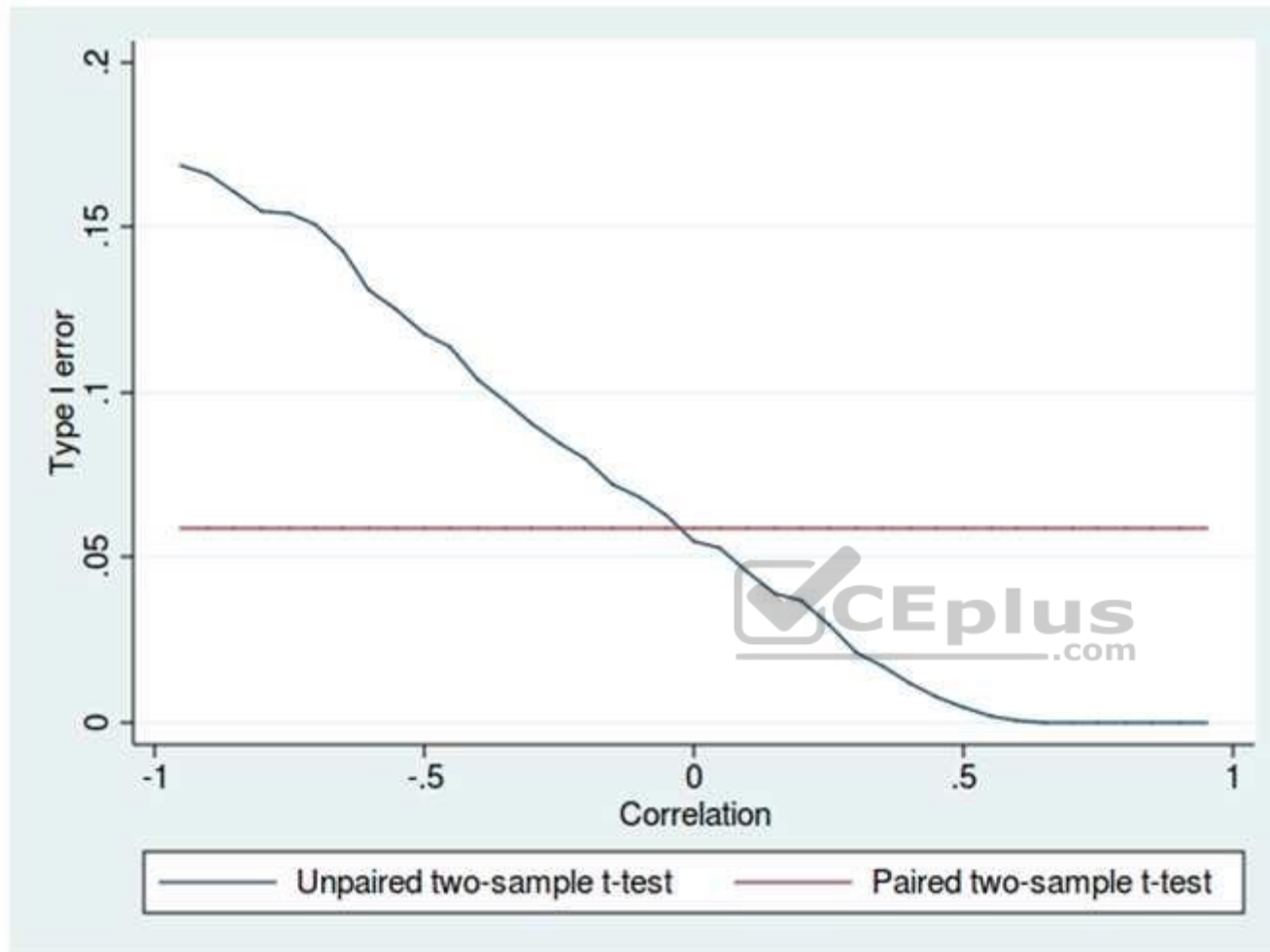
**Explanation/Reference:**

Explanation:

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

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





Reference:

<https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/test-hypothesis-using-t-test> [https://en.wikipedia.org/wiki/Student%27s\\_t-test](https://en.wikipedia.org/wiki/Student%27s_t-test)

Testlet 1

## Case study

### Overview

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

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

### Current environment

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

### Penalty detection and sentiment

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

### Advertisements

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

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



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

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

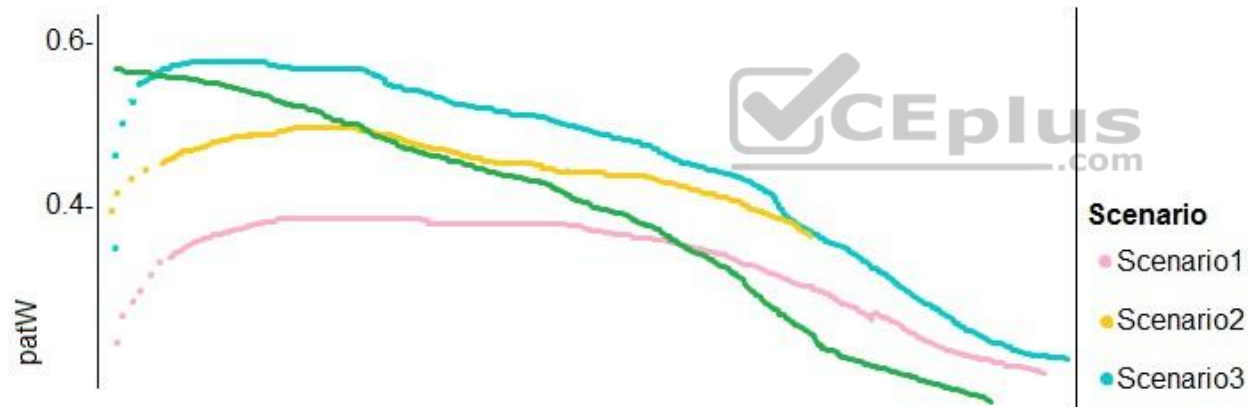
The ad propensity model uses cost factors shown in the following diagram:

		Actual	
		1	0
Predicted	0	1	2
	1	2	1

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

		Actual	
		1	0
Predicted	0	1	5
	1	5	1

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



### QUESTION 1

DRAG DROP

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

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

Select and Place:

### Actions

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



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]

Explanation

Explanation/Reference:

Explanation:

Step 1: Define a cross-entropy function activation

When using a neural network to perform classification and prediction, it is usually better to use cross-entropy error than classification error, and somewhat better to use cross-entropy error than mean squared error to evaluate the quality of the neural network.

Step 2: Add cost functions for each target state.

Step 3: Evaluated the distance error metric.

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

## QUESTION 2

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

Which technique should you use?

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

**Correct Answer:** A

**Section:** [none]

**Explanation**

### Explanation/Reference:

Explanation:

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

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

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

Scenario:

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

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

References:

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

### QUESTION 3

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

Which technique should you use?

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

**Correct Answer:** A

**Section:** [none]

**Explanation**

**Explanation/Reference:**

Explanation:

Scenario:

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



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

**Question Set 2**

### 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: Accuracy, Precision, Recall, F1 score, and AUC.

Does the solution meet the goal?

- A. Yes
- B. No

**Correct Answer:** B

**Section:** [none]

**Explanation**



**Explanation/Reference:**

Explanation:

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

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

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

Does the solution meet the goal?

- A. Yes
- B. No

**Correct Answer:** B

**Section:** [none]

**Explanation**

**Explanation/Reference:**

Explanation:

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

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

### QUESTION 3

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>

**QUESTION 4**

You are creating a binary classification by using a two-class logistic regression model.

You need to evaluate the model results for imbalance.

Which evaluation metric should you use?

- A. Relative Absolute Error
- B. AUC Curve
- C. Mean Absolute Error
- D. Relative Squared Error
- E. Accuracy
- F. Root Mean Square Error

**Correct Answer:** B

**Section:** [none]

**Explanation**

**Explanation/Reference:**

Explanation:



One can inspect the true positive rate vs. the false positive rate in the Receiver Operating Characteristic (ROC) curve and the corresponding Area Under the Curve (AUC) value. The closer this curve is to the upper left corner, the better the classifier's performance is (that is maximizing the true positive rate while minimizing the false positive rate). Curves that are close to the diagonal of the plot, result from classifiers that tend to make predictions that are close to random guessing.

References: <https://docs.microsoft.com/en-us/azure/machine-learning/studio/evaluate-model-performance#evaluating-a-binary-classification-model>

### QUESTION 5

#### 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>High</div> <div>Low</div> <div>Identical</div>	<div>▼</div> <div>High</div> <div>Low</div> <div>Identical</div>
15	<div>▼</div> <div>High</div> <div>Low</div> <div>Identical</div>	<div>▼</div> <div>High</div> <div>Low</div> <div>Identical</div>

**Correct Answer:**

## Answer Area

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

Section: [none]

Explanation

Explanation/Reference:

Explanation:

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

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

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

### QUESTION 6

You are building a machine learning model for translating English language textual content into French language textual content.

You need to build and train the machine learning model to learn the sequence of the textual content.

Which type of neural network should you use?

- A. Multilayer Perceptions (MLPs)
- B. Convolutional Neural Networks (CNNs)
- C. Recurrent Neural Networks (RNNs)
- D. Generative Adversarial Networks (GANs)

**Correct Answer:** C

**Section:** [none]

**Explanation**

**Explanation/Reference:**

Explanation:

To translate a corpus of English text to French, we need to build a recurrent neural network (RNN).

Note: RNNs are designed to take sequences of text as inputs or return sequences of text as outputs, or both. They're called recurrent because the network's hidden layers have a loop in which the output and cell state from each time step become inputs at the next time step. This recurrence serves as a form of memory. It allows contextual information to flow through the network so that relevant outputs from previous time steps can be applied to network operations at the current time step.

References: <https://towardsdatascience.com/language-translation-with-rnns-d84d43b40571>

### QUESTION 7

You create a binary classification model.

You need to evaluate the model performance.

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

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

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

**Correct Answer:** BC

**Section: [none]**

**Explanation**

**Explanation/Reference:**

Explanation:

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

Note: A very natural question is: 'Out of the individuals whom the model, how many were classified correctly (TP)?'

This question can be answered by looking at the Precision of the model, which is the proportion of positives that are classified correctly.

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

### QUESTION 8

You use the Two-Class Neural Network module in Azure Machine Learning Studio to build a binary classification model. You use the Tune Model Hyperparameters module to tune accuracy for the model.

You need to configure the Tune Model Hyperparameters module.

Which two values should you use? Each correct answer presents part of the solution.

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

- A. Number of hidden nodes
- B. Learning Rate
- C. The type of the normalizer
- D. Number of learning iterations
- E. Hidden layer specification

**Correct Answer:** DE

**Section: [none]**

**Explanation**

**Explanation/Reference:**

Explanation:

D: For Number of learning iterations, specify the maximum number of times the algorithm should process the training cases.

E: For Hidden layer specification, select the type of network architecture to create.

Between the input and output layers you can insert multiple hidden layers. Most predictive tasks can be accomplished easily with only one or a few hidden layers.

References:

<https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/two-class-neural-network>

### QUESTION 9

You create a binary classification model by using Azure Machine Learning Studio.

You must tune hyperparameters by performing a parameter sweep of the model. The parameter sweep must meet the following requirements:

- iterate all possible combinations of hyperparameters
- minimize computing resources required to perform the sweep

You need to perform a parameter sweep of the model.

Which parameter sweep mode should you use?

- A. Random sweep
- B. Sweep clustering
- C. Entire grid
- D. Random grid

**Correct Answer:** D

**Section:** [none]

**Explanation**

#### Explanation/Reference:

Explanation:

Maximum number of runs on random grid: This option also controls the number of iterations over a random sampling of parameter values, but the values are not generated randomly from the specified range; instead, a matrix is created of all possible combinations of parameter values and a random sampling is taken over the matrix. This method is more efficient and less prone to regional oversampling or undersampling.

If you are training a model that supports an integrated parameter sweep, you can also set a range of seed values to use and iterate over the random seeds as well. This is optional, but can be useful for avoiding bias introduced by seed selection.

Incorrect Answers:

B: If you are building a clustering model, use Sweep Clustering to automatically determine the optimum number of clusters and other parameters.

C: Entire grid: When you select this option, the module loops over a grid predefined by the system, to try different combinations and identify the best learner. This option is useful for cases where you don't know what the best parameter settings might be and want to try all possible combination of values.

Reference:



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

#### QUESTION 10

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]

**Explanation**

**Explanation/Reference:**

Explanation:

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

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

#### QUESTION 11

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

You need to define the possible termination conditions.

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

NOTE: Each correct selection is worth one point.

- A. Centroids do not change between iterations.
- B. The residual sum of squares (RSS) rises above a threshold.

- C. The residual sum of squares (RSS) falls below a threshold.
- D. A fixed number of iterations is executed.
- E. The sum of distances between centroids reaches a maximum.

**Correct Answer:** ACD

**Section:** [none]

**Explanation**

**Explanation/Reference:**

Explanation:

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

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

References: <https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/k-means-clustering> <https://nlp.stanford.edu/IR-book/html/htmledition/k-means-1.html>

## QUESTION 12

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

The CNN model you build shows signs of overfitting.

You need to reduce overfitting and converge the model to an optimal fit.

Which two actions should you perform? Each correct answer presents a complete solution.

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

- A. Add an additional dense layer with 512 input units.
- B. Add L1/L2 regularization.
- C. Use training data augmentation.
- D. Reduce the amount of training data.
- E. Add an additional dense layer with 64 input units.

**Correct Answer:** BD

**Section:** [none]

**Explanation**

**Explanation/Reference:**

Explanation:

B: Weight regularization provides an approach to reduce the overfitting of a deep learning neural network model on the training data and improve the performance of the model on new data, such as the holdout test set.

Keras provides a weight regularization API that allows you to add a penalty for weight size to the loss function.

Three different regularizer instances are provided; they are:

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

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

By avoiding training all nodes on all training data, dropout decreases overfitting.

References:

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

[https://en.wikipedia.org/wiki/Convolutional\\_neural\\_network](https://en.wikipedia.org/wiki/Convolutional_neural_network)



<https://vcceplus.com/>