



Microsoft

Exam Questions DP-100

Designing and Implementing a Data Science Solution on Azure

NEW QUESTION 1

- (Exam Topic 3)

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

Answer: C

Explanation:

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

NEW QUESTION 2

- (Exam Topic 3)

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

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

You are creating a model to predict the price of a student's artwork depending on the following variables: the student's length of education, degree type, and art form.

You start by creating a linear regression model. You need to evaluate the linear regression model.

Solution: Use the following metrics: Mean Absolute Error, Root Mean Absolute Error, Relative Absolute Error, Relative Squared Error, and the Coefficient of Determination.

Does the solution meet the goal?

- A. Yes
- B. No

Answer: A

Explanation:

The following metrics are reported for evaluating regression models. When you compare models, they are ranked by the metric you select for evaluation.

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

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

Relative absolute error (RAE) is the relative absolute difference between expected and actual values; relative because the mean difference is divided by the arithmetic mean.

Relative squared error (RSE) similarly normalizes the total squared error of the predicted values by dividing by the total squared error of the actual values.

Mean Zero One Error (MZOE) indicates whether the prediction was correct or not. In other words: $\text{ZeroOneLoss}(x,y) = 1$ when $x \neq y$; otherwise 0.

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.

AUC.

References:

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

NEW QUESTION 3

- (Exam Topic 3)

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

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

The learning rate must be selected from a normal distribution with a mean value of 10 and a standard deviation of 3.

Batch size must be 16, 32 and 64.

Keep probability must be a value selected from a uniform distribution between the range of 0.05 and 0.1.

You need to use the `param_sampling` method of the Python API for the Azure Machine Learning Service. How should you complete the code segment? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

In random sampling, hyperparameter values are randomly selected from the defined search space. Random sampling allows the search space to include both discrete and continuous hyperparameters.

Example:

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

Reference:

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

NEW QUESTION 4

- (Exam Topic 3)

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

You need to select a data cleaning method. Which method should you use?

- A. Synthetic Minority
- B. Replace using Probabilistic PAC
- C. Replace using MICE
- D. Normalization

Answer: B

NEW QUESTION 5

- (Exam Topic 3)

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 value 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. Execute Python Script
- B. Export Count Table
- C. Convert to Indicator Values
- D. Summarize Data
- E. Compute linear Correlation

Answer: BE

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.

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>

NEW QUESTION 6

- (Exam Topic 3)

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

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

You need to select appropriate methods for producing the experiment.

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

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

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

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

NEW QUESTION 7

- (Exam Topic 3)

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

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

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

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Step 1: Add a Two-Class Support Vector Machine module to initialize the SVM classifier. Step 2: Add a dataset to the experiment

Step 3: Add a Split Data module to create training and test dataset.

To generate a set of feature scores requires that you have an already trained model, as well as a test dataset. Step 4: Add a Permutation Feature Importance module and connect to the trained model and test dataset. Step 5: Set the Metric for measuring performance property to Classification - Accuracy and then run the experiment.

Reference:

<https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/two-class-support-vector-mac> <https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/permutation-feature-importan>

NEW QUESTION 8

- (Exam Topic 3)

You are a data scientist building a deep convolutional neural network (CNN) for image classification. The CNN model you built 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. Reduce the amount of training data.
- B. Add an additional dense layer with 64 input units
- C. Add L1/L2 regularization.
- D. Use training data augmentation
- E. Add an additional dense layer with 512 input units.

Answer: AC

Explanation:

References:

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

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

NEW QUESTION 9

- (Exam Topic 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

Answer: A

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>

NEW QUESTION 10

- (Exam Topic 3)

You are building a regression model tot 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 can be positive or negative,
- C. The label data must be a positive value
- D. The label data must be non discrete.
- E. The data must be whole numbers.

Answer: CE

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>

NEW QUESTION 10

- (Exam Topic 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 in sequence? To answer, move the appropriate modules from the list of modules to the answer area and arrange them in the correct order.

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Create Scatterplot Summarize Data Clip Values

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>

NEW QUESTION 14

- (Exam Topic 2)

You need to identify the methods for dividing the data according to the testing requirements. Which properties should you select? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Scenario: Testing

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

Box 1: Assign to folds

Use Assign to folds option when you want to divide the dataset into subsets of the data. This option is also useful when you want to create a custom number of folds for cross-validation, or to split rows into several groups.

Not Head: Use Head mode to get only the first n rows. This option is useful if you want to test a pipeline on a small number of rows, and don't need the data to be balanced or sampled in any way.

Not Sampling: The Sampling option supports simple random sampling or stratified random sampling. This is useful if you want to create a smaller representative sample dataset for testing.

Box 2: Partition evenly

Specify the partitioner method: Indicate how you want data to be apportioned to each partition, using these options:

Partition evenly: Use this option to place an equal number of rows in each partition. To specify the number of output partitions, type a whole number in the Specify number of folds to split evenly into text box.

Reference:

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

NEW QUESTION 15

- (Exam Topic 3)

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. Partition and Sample
- B. Assign Data to Clusters
- C. Group Data into Bins
- D. Test Hypothesis Using t-Test

Answer: A

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>

NEW QUESTION 17

- (Exam Topic 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

Answer: B

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>

NEW QUESTION 19

- (Exam Topic 2)

You need to configure the Permutation Feature Importance module for the model training requirements. What should you do? To answer, select the appropriate options in the dialog box in the answer area. NOTE: Each correct selection is worth one point.

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Box 1: 500

For Random seed, type a value to use as seed for randomization. If you specify 0 (the default), a number is generated based on the system clock.

A seed value is optional, but you should provide a value if you want reproducibility across runs of the same experiment.

Here we must replicate the findings. Box 2: Mean Absolute Error

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

Regression. Choose one of the following: Precision, Recall, Mean Absolute Error, Root Mean Squared Error, Relative Absolute Error, Relative Squared Error, Coefficient of Determination

References:

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

NEW QUESTION 21

- (Exam Topic 2)

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

Which properties should you select? To answer, select the appropriate option-, m the answer area. NOTE: Each correct selection is worth one point.

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

NEW QUESTION 26

- (Exam Topic 2)

You need to correct the model fit issue.

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

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Step 1: Augment the data

Scenario: Columns in each dataset contain missing and null values. The datasets also contain many outliers.

Step 2: Add the Bayesian Linear Regression module.

Scenario: You produce a regression model to predict property prices by using the Linear Regression and Bayesian Linear Regression modules.

Step 3: Configure the regularization weight.

Regularization typically is used to avoid overfitting. For example, in L2 regularization weight, type the value to use as the weight for L2 regularization. We recommend that you use a non-zero value to avoid overfitting.

Scenario:

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

NEW QUESTION 28

- (Exam Topic 2)

You need to configure the Edit Metadata module so that the structure of the datasets match. Which configuration options should you select? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

NEW QUESTION 30

- (Exam Topic 2)

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

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

NOTE: Each correct selection is worth one point.

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Box 1: Mutual Information.

The mutual information score is particularly useful in feature selection because it maximizes the mutual information between the joint distribution and target variables in datasets with many dimensions.

Box 2: MedianValue

MedianValue is the feature column, it is the predictor of the dataset.

Scenario: The MedianValue and AvgRoomsInHouse columns both hold data in numeric format. You need to select a feature selection algorithm to analyze the relationship between the two columns in more detail.

References:

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

NEW QUESTION 35

- (Exam Topic 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

Answer: C

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>

NEW QUESTION 37

- (Exam Topic 1)

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.

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Scenario:

Experiments for local crowd sentiment models must combine local penalty detection data.

Crowd sentiment models must identify known sounds such as cheers and known catch phrases. Individual crowd sentiment models will detect similar sounds.

Note: Evaluate the changed in correlation between model error rate and centroid distance

In machine learning, a nearest centroid classifier or nearest prototype classifier is a classification model that assigns to observations the label of the class of training samples whose mean (centroid) is closest to the observation.

References: https://en.wikipedia.org/wiki/Nearest_centroid_classifier

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

NEW QUESTION 42

- (Exam Topic 1)

You need to implement a feature engineering strategy for the crowd sentiment local models. What should you do?

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

Answer: D

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.

NEW QUESTION 43

- (Exam Topic 1)

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.

Answer: A

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%.

NEW QUESTION 46

- (Exam Topic 1)

You need to define a process for penalty event detection.

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.

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

NEW QUESTION 50

- (Exam Topic 3)

You have a Python data frame named salesData in the following format: The data frame must be unpivoted to a long data format as follows:

You need to use the pandas.melt() function in Python to perform the transformation.

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

NOTE: Each correct selection is worth one point.

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Box 1: dataframe

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

Where frame is a DataFrame

Box 2: shop

Parameter id_vars : tuple, list, or ndarray, optional Column(s) to use as identifier variables.

Box 3: ['2017', '2018']

value_vars : tuple, list, or ndarray, optional

Column(s) to unpivot. If not specified, uses all columns that are not set as id_vars. Example:

```
df = pd.DataFrame({'A': {0: 'a', 1: 'b', 2: 'c'},
```

```
'B': {0: 1, 1: 3, 2: 5},
```

```
'C': {0: 2, 1: 4, 2: 6}})
```

```
pd.melt(df, id_vars=['A'], value_vars=['B', 'C'])
```

 A variable value

```
0 a B 1
```

```
1 b B 3
```

```
2 c B 5
```

```
3 a C 2
```

```
4 b C 4
```

```
5 c C 6
```

References:

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

NEW QUESTION 51

- (Exam Topic 3)

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)

Answer: BCD

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>

NEW QUESTION 56

- (Exam Topic 3)

You use Data Science Virtual Machines (DSVMs) for Windows and Linux in Azure. You need to access the DSVMs.

Which utilities should you use? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

NEW QUESTION 57

- (Exam Topic 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.

Which environment should you use?

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

Answer: A

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.

NEW QUESTION 61

- (Exam Topic 3)

You are using C-Support Vector classification to do a multi-class classification with an unbalanced training dataset. The C-Support Vector classification using Python code shown below:

You need to evaluate the C-Support Vector classification code.

Which evaluation statement should you use? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Box 1: Automatically adjust weights inversely proportional to class frequencies in the input data

The "balanced" mode uses the values of y to automatically adjust weights inversely proportional to class frequencies in the input data as $n_samples / (n_classes * np.bincount(y))$.

Box 2: Penalty parameter

Parameter: C : float, optional (default=1.0)

Penalty parameter C of the error term. References:

<https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html>

NEW QUESTION 63

- (Exam Topic 3)

You are developing a data science workspace that uses an Azure Machine Learning service. You need to select a compute target to deploy the workspace. What should you use?

- A. Azure Data Lake Analytics
- B. Azure Databrick .
- C. Apache Spark for HDInsight.
- D. Azure Container Service

Answer: D

Explanation:

Azure Container Instances can be used as compute target for testing or development. Use for low-scale CPU-based workloads that require less than 48 GB of RAM.

Reference:

<https://docs.microsoft.com/en-us/azure/machine-learning/service/how-to-deploy-and-where>

NEW QUESTION 64

- (Exam Topic 3)

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. NOTE: Each correct selection is worth one point.

- A. box plot
- B. scatter
- C. random forest diagram
- D. Venn diagram
- E. ROC curve

Answer: AB

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/>

NEW QUESTION 69

- (Exam Topic 3)

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

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

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

One class has a much smaller number of observations than tin- other classes in the training set. You need to select an appropriate data sampling strategy to compensate for the class imbalance. Solution: You use the Principal Components Analysis (PCA) sampling mode.

Does the solution meet the goal?

- A. Yes
- B. No

Answer: B

NEW QUESTION 70

- (Exam Topic 3)

You are performing a filter based feature selection for a dataset 10 build a multi class classifies by using Azure Machine Learning Studio.

The dataset contains categorical features that are highly correlated to the output label column.

You need to select the appropriate feature scoring statistical method to identify the key predictors. Which method should you use?

- A. Chi-squared
- B. Spearman correlation
- C. Kendall correlation
- D. Person correlation

Answer: D

Explanation:

Pearson's correlation statistic, or Pearson's correlation coefficient, is also known in statistical models as the r value. For any two variables, it returns a value that indicates the strength of the correlation
Pearson's correlation coefficient is the test statistics that measures the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship.

Reference:

<https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/filter-based-feature-selection> <https://www.statisticssolutions.com/pearsons-correlation-coefficient/>

NEW QUESTION 72

- (Exam Topic 3)

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

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

You are creating a model to predict the price of a student's artwork depending on the following variables: the student's length of education, degree type, and art form.

You start by creating a linear regression model.

You need to evaluate the linear regression model.

Solution: Use the following metrics: Relative Squared Error, Coefficient of Determination, Accuracy, Precision, Recall, F1 score, and AUC.

Does the solution meet the goal?

- A. Yes
- B. No

Answer: B

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>

NEW QUESTION 76

- (Exam Topic 3)

You have a dataset contains 2,000 rows. You are building a machine learning classification model by using Azure Machine Learning Studio. You add a Partition and Sample module to the experiment.

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

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

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

NOTE: Each correct selection is worth one point.

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

NEW QUESTION 79

- (Exam Topic 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: Replace each missing value using the Multiple Imputation by Chained Equations (MICE) method. Does the solution meet the goal?

A. Yes

B. NO

Answer: A

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>

NEW QUESTION 80

- (Exam Topic 3)

You create an experiment in Azure Machine Learning Studio. You add a training dataset that contains 10,000 rows. The first 9,000 rows represent class 0 (90 percent).

The remaining 1,000 rows represent class 1 (10 percent).

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

You need to configure the module.

Which values should you use? To answer, select the appropriate options in the dialog box in the answer area. NOTE: Each correct selection is worth one point.

A. Mastered

B. Not Mastered

Answer: A

Explanation:

Box 1: 300

You type 300 (%), the module triples the percentage of minority cases (3000) compared to the original dataset (1000).

Box 2: 5

We should use 5 data rows.

Use the Number of nearest neighbors option to determine the size of the feature space that the SMOTE algorithm uses when in building new cases. A nearest neighbor is a row of data (a case) that is very similar to some target case. The distance between any two cases is measured by combining the weighted vectors of all features.

By increasing the number of nearest neighbors, you get features from more cases.

By keeping the number of nearest neighbors low, you use features that are more like those in the original sample.

References:

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

NEW QUESTION 83

- (Exam Topic 3)

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

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

You are a data scientist using Azure Machine Learning Studio.

You need to normalize values to produce an output column into bins to predict a target column. Solution: Apply a Quantiles binning mode with a PQuantile normalization.

Does the solution meet the goal?

- A. Yes
- B. No

Answer: B

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>

NEW QUESTION 88

- (Exam Topic 3)

You are training a deep learning model to identify cats and dogs. You have 25,000 color images.

You must meet the following requirements:

- Reduce the number of training epochs.
- Reduce the size of the neural network.
- Reduce over-fitting of the neural network.

You need to select the image modification values.

Which value should you use? To answer, select the appropriate Options in the answer area. NOTE: Each correct selection is worth one point.

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

NEW QUESTION 92

- (Exam Topic 3)

You are evaluating a Python NumPy array that contains six data points defined as follows: `data = [10, 20, 30, 40, 50, 60]`

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

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

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

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

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

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

NOTE: Each correct selection is worth one point.

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Box 1: k-fold

Box 2: 3

K-F olds cross-validator provides train/test indices to split data in train/test sets. Split dataset into k consecutive folds (without shuffling by default).

The parameter n_splits (int, default=3) is the number of folds. Must be at least 2. Box 3: data

Example: Example:

>>>

```
>>> from sklearn.model_selection import KFold
```

```
>>> X = np.array([[1, 2], [3, 4], [1, 2], [3, 4]])
```

```
>>> y = np.array([1, 2, 3, 4])
```

```
>>> kf = KFold(n_splits=2)
```

```
>>> kf.get_n_splits(X) 2
```

```
>>> print(kf)
```

```
KFold(n_splits=2, random_state=None, shuffle=False)
```

```
>>> for train_index, test_index in kf.split(X): print("TRAIN:", train_index, "TEST:", test_index) X_train, X_test = X[train_index], X[test_index] y_train, y_test =
```

```
y[train_index], y[test_index] TRAIN: [2 3] TEST: [0 1]
```

```
TRAIN: [0 1] TEST: [2 3]
```

References:

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

NEW QUESTION 94

- (Exam Topic 3)

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

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

Ensure that multiple related words from a single canonical form.

Remove pipe characters from text.

Remove words to optimize information retrieval.

Which three options should you select? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Box 1: Remove stop words

Remove words to optimize information retrieval.

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

Box 2: Lemmatization

Ensure that multiple related words from a single canonical form. Lemmatization converts multiple related words to a single canonical form Box 3: Remove special characters

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

References:

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

NEW QUESTION 97

- (Exam Topic 3)

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

Answer: ADE

Explanation:

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>

NEW QUESTION 100

- (Exam Topic 3)

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

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

iterate all possible combinations of hyperparameters
minimize computing resources required to perform the sweep
You need to perform a parameter sweep of the model.
Which parameter sweep mode should you use?

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

Answer: D

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.

NEW QUESTION 103

- (Exam Topic 3)

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)

Answer: C

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>

NEW QUESTION 107

- (Exam Topic 3)

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

Answer: C

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>

NEW QUESTION 111

- (Exam Topic 3)

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

You need to select a data cleaning method. Which method should you use?

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

Answer: A

NEW QUESTION 114

- (Exam Topic 3)

You create an experiment in Azure Machine Learning Studio- You add a training dataset that contains 10,000 rows. The first 9,000 rows represent class 0 (90 percent). The first 1,000 rows represent class 1 (10 percent).

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

You need to configure the module.

Which values should you use? To answer, select the appropriate options in the dialog box in the answer area. NOTE: Each correct selection is worth one point.

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

NEW QUESTION 116

- (Exam Topic 3)

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

Answer: B

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-bina>

NEW QUESTION 117

- (Exam Topic 3)

You are creating a machine learning model in Python. The provided dataset contains several numerical columns and one text column. The text column represents a product's category. The product category will always be one of the following:

- Bikes
- Cars
- Vans
- Boats

You are building a regression model using the scikit-learn Python package.

You need to transform the text data to be compatible with the scikit-learn Python package.

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

NOTE: Each correct selection is worth one point.

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Box 1: pandas as df

Pandas takes data (like a CSV or TSV file, or a SQL database) and creates a Python object with rows and columns called data frame that looks very similar to table in a statistical software (think Excel or SPSS for example).

Box 2: transpose[ProductCategoryMapping] Reshape the data from the pandas Series to columns. Reference:
<https://datascienceplus.com/linear-regression-in-python/>

NEW QUESTION 121

.....

Thank You for Trying Our Product

We offer two products:

1st - We have Practice Tests Software with Actual Exam Questions

2nd - Questions and Answers in PDF Format

DP-100 Practice Exam Features:

- * DP-100 Questions and Answers Updated Frequently
- * DP-100 Practice Questions Verified by Expert Senior Certified Staff
- * DP-100 Most Realistic Questions that Guarantee you a Pass on Your FirstTry
- * DP-100 Practice Test Questions in Multiple Choice Formats and Updatesfor 1 Year

100% Actual & Verified — Instant Download, Please Click
[Order The DP-100 Practice Test Here](#)