QUESTION 50
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.

Answer: B
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 51
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

Answer: A
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 52
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.

Answer: BD
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 53
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

QUESTION 54
Hotspot Question
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.

Answer Area

▲ SMOTE

Label column

Selected columns: **All labels**

Launch column selector

SMOTE percentage ☰

▼
0
300
3000
4000

Number of nearest neighbors ☰

▼
0
1
5
4000

Random seed ☰

0

Answer:

Answer Area

▲ SMOTE

Label column

```
Selected columns:
All labels
```

```
            Launch column selector
```

SMOTE percentage   ≡

```
                                        ▼
0
300
3000
4000
```

Number of nearest neighbors   ≡

```
                                        ▼
0
1
5
4000
```

Random seed   ≡

```
0
```

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

QUESTION 55
Drag and Drop Question
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.

Actions

Build Counting Transform

Missing Values Scrubber

Feature Hashing

Clean Missing Data

Replace Discrete Values

Import Data

Latent Dirichlet Transformation

Partition and Sample

Answer Area

Answer:

Actions

Build Counting Transform

Missing Values Scrubber

Feature Hashing

Replace Discrete Values

Latent Dirichlet Transformation

Answer Area

Import Data

Clean Missing Data

Partition and Sample

Explanation:
The Clean Missing Data module in Azure Machine Learning Studio, to remove, replace, or infer missing values.
Incorrect Answers:
Latent Direchlet Transformation: Latent Dirichlet Allocation module in Azure Machine Learning Studio, to group otherwise unclassified text into a number of categories. Latent Dirichlet Allocation (LDA) is often used in natural language processing (NLP) to find texts that are similar. Another common term is topic modeling.
Build Counting Transform: Build Counting Transform module in Azure Machine Learning Studio, to analyze training

data. From this data, the module builds a count table as well as a set of count-based features that can be used in a predictive model.

Missing Value Scrubber: The Missing Values Scrubber module is deprecated.

Feature hashing: Feature hashing is used for linguistics, and works by converting unique tokens into integers.

Replace discrete values: the Replace Discrete Values module in Azure Machine Learning Studio is used to generate a probability score that can be used to represent a discrete value. This score can be useful for understanding the information value of the discrete values.

References:

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

QUESTION 56
Hotspot Question
You are retrieving data from a large datastore by using Azure Machine Learning Studio.
You must create a subset of the data for testing purposes using a random sampling seed based on the system clock.
You add the Partition and Sample module to your experiment.
You need to select the properties for the module.
Which values should you select? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

Answer Area

◢ **Partition and Sample**

Partition or sample mode

▼
Assign to Folds
Pick Fold
Sampling
Head

Rate of sampling ≡

.2

Random seed for sampling ≡

▼
0
1
time.clock()
utcNow()

Stratified split for sampling

False ⌄

Answer:

Answer Area

▲ **Partition and Sample**

Partition or sample mode

	▼
Assign to Folds	
Pick Fold	
Sampling	
Head	

Rate of sampling ☰

.2

Random seed for sampling ☰

	▼
0	
1	
time.clock()	
utcNow()	

Stratified split for sampling

False ✓

Explanation:
Box 1: Sampling
Create a sample of data
This option supports simple random sampling or stratified random sampling. This is useful if you want to create a smaller representative sample dataset for testing.
1. Add the Partition and Sample module to your experiment in Studio, and connect the dataset.
2. Partition or sample mode: Set this to Sampling.
3. Rate of sampling. See box 2 below.
Box 2: 0
3. Rate of sampling. Random seed for sampling: Optionally, type an integer to use as a seed value.
This option is important if you want the rows to be divided the same way every time. The default value is 0, meaning that a starting seed is generated based on the system clock. This can lead to slightly different results each time you run the experiment.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/partition-and-sample

QUESTION 57
Drag and Drop Question
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.

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		

Answer:

Answer Area

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

Explanation:
Box 1: Clean Missing Data
Box 2: SMOTE
Use the SMOTE module in Azure Machine Learning Studio to increase the number of underepresented 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 58
Hotspot Question
You have a Python data frame named salesData in the following format:

	shop	2017	2018
0	Shop X	34	25
1	Shop Y	65	76
2	Shop Z	48	55

The data frame must be unpivoted to a long data format as follows:

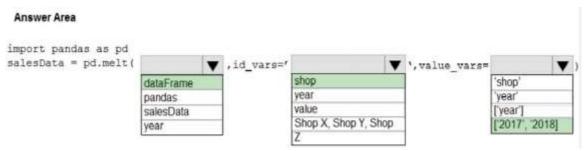|	shop	year	value
0	Shop X	2017	34
1	Shop Y	2017	65
2	Shop Z	2017	48
3	Shop X	2018	25
4	Shop Y	2018	76
5	Shop Z	2018	55

You need to use the pandas.melt() function in Python to perform the transformation.
How should you complete the code segment? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

Answer Area

```
import pandas as pd
salesData = pd.melt(      ▼     ,id_vars='      ▼      ',value_vars=      ▼      )
                    dataFrame          shop                        'shop'
                    pandas             year                        'year'
                    salesData          value                       ['year']
                    year               Shop X, Shop Y, Shop         ['2017', '2018']
                                       Z
```

Answer:

Answer Area

```
import pandas as pd
salesData = pd.melt(      ▼     ,id_vars='      ▼      ',value_vars=      ▼      )
                    dataFrame          shop                        'shop'
                    pandas             year                        'year'
                    salesData          value                       ['year']
                    year               Shop X, Shop Y, Shop         ['2017', '2018']
                                       Z
```

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
Paramter id_vars 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

QUESTION 59
Hotspot Question
You are working on a classification task. You have a dataset indicating whether a student would like to play soccer and associated attributes. The dataset includes the following columns:

Name	Description
IsPlaySoccer	Values can be 1 and 0.
Gender	Values can be M or F.
PrevExamMarks	Stores values from 0 to 100
Height	Stores values in centimeters
Weight	Stores values in kilograms

You need to classify variables by type.
Which variable should you add to each category? To answer, select the appropriate options in the answer area.
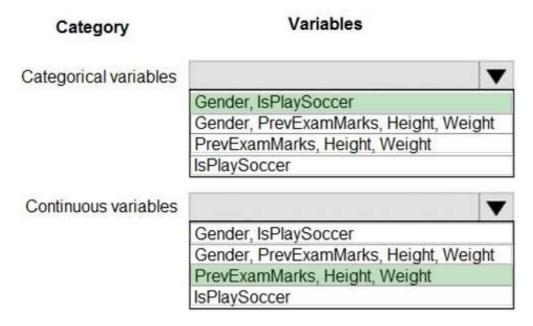NOTE: Each correct selection is worth one point.

Answer Area

Category	Variables
Categorical variables	▼

Gender, IsPlaySoccer
Gender, PrevExamMarks, Height, Weight
PrevExamMarks, Height, Weight
IsPlaySoccer

| Continuous variables | ▼ |

Gender, IsPlaySoccer
Gender, PrevExamMarks, Height, Weight
PrevExamMarks, Height, Weight
IsPlaySoccer

Answer:

Answer Area

Category	Variables
Categorical variables	▼

Gender, IsPlaySoccer
Gender, PrevExamMarks, Height, Weight
PrevExamMarks, Height, Weight
IsPlaySoccer

Continuous variables	▼

Gender, IsPlaySoccer
Gender, PrevExamMarks, Height, Weight
PrevExamMarks, Height, Weight
IsPlaySoccer

Explanation:
https://www.edureka.co/blog/classification-algorithms/

QUESTION 60
Hotspot Question
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.

Answer Area

▲ Preprocess Text

Language

English ⌄

Remove by part of speech

False ⌄

Text column to clean

Selected columns:
Column names: String, Feature

Launch column selector

☐ Remove stop words ≡

☐ Lemmatization ≡

☐ Detect sentences ≡

☐ Normalize case to lowercase ≡

☐ Remove numbers ≡

☐ Remove special characters ≡

☐ Remove duplicate characters ≡

☐ Remove email addresses ≡

☐ Remove URLs ≡

☐ Expand verb contractions ≡

☐ Normalize backslashes to slashes ≡

☐ Split tokens on special characters ≡

Answer:

Answer Area

▲ Preprocess Text

Language

| English | ∨ |

Remove by part of speech

| False | ∨ |

Text column to clean

Selected columns:
Column names: **String, Feature**

| Launch column selector |

☐ Remove stop words	≡
☐ Lemmatization	≡
☐ Detect sentences	≡
☐ Normalize case to lowercase	≡
☐ Remove numbers	≡
☐ Remove special characters	≡
☐ Remove duplicate characters	≡
☐ Remove email addresses	≡
☐ Remove URLs	≡
☐ Expand verb contractions	≡
☐ Normalize backslashes to slashes	≡
☐ Split tokens on special characters	≡

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