Vendor: Microsoft

Exam Code:DP-100

Exam Name: Designing and Implementing a Data Science Solution on Azure

Version: Demo

QUESTION 1

You have the following code. The code prepares an experiment to run a script:

A. run = script_experiment.start_logging()

B. run = Run(experiment=script_experiment)

C. ws.get_run(run_id=experiment.id)

D. run = script_experiment.submit(config=script_config)

Correct Answer: D

The experiment class submit method submits an experiment and return the active created run.

Syntax: submit(config, tags=None, **kwargs)

Reference:

https://docs.microsoft.com/en-us/python/api/azureml-core/azureml.core.experiment.experiment

---

QUESTION 2

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while

others might not have a correct solution.

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

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

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

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

Solution: Replace each missing value using the Multiple Imputation by Chained Equations (MICE) method.

Does the solution meet the goal?

A. Yes

B. No

Correct Answer: A

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

the other variables in the data before filling in the missing values.

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

References: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3074241/

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

---

QUESTION 3

You develop and train a machine learning model to predict fraudulent transactions for a hotel booking website.

Traffic to the site varies considerably. The site experiences heavy traffic on Monday and Friday and much lower traffic on other days. Holidays are also high web traffic days.

You need to deploy the model as an Azure Machine Learning real-time web service endpoint on compute that can dynamically scale up and down to support demand.

Which deployment compute option should you use?

A. attached Azure Databricks cluster

B. Azure Container Instance (ACI)

C. Azure Kubernetes Service (AKS) inference cluster

D. Azure Machine Learning Compute Instance

E. attached virtual machine in a different region

Correct Answer: C

Azure Machine Learning compute cluster is a managed-compute infrastructure that allows you to easily create a single or multi-node compute. The compute is created within your workspace region as a resource that can be shared with other users in your workspace. The compute scales up automatically when a job is submitted, and can be put in an Azure Virtual Network.

Reference: https://docs.microsoft.com/en-us/azure/machine-learning/how-to-create-attach-compute-sdk

---

QUESTION 4

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

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 5

DRAG DROP

You build a binary classification model using the Azure Machine Learning Studio Two-Class Neural Network module.

You are preparing to configure the Tune Model Hyperparameters module for the purpose of tuning accuracy for the model.

Which of the following are valid parameters for the Two-Class Neural Network module? Answer by dragging the correct options from the list to the answer area.

Select and Place:

Options

| Depth of the tree |
| Random number seed |
| Optimization tolerance |
| The initial learning weights diameter |
| Lambda |
| Number of learning iterations |
| Project to the unit-sphere |

Answer

Correct Answer:

Options

Depth of the tree

Optimization tolerance

Lambda

Project to the unit-sphere

Answer

Random number seed

The initial learning weights diameter

Number of learning iterations

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

---

QUESTION 6

HOTSPOT

You create an Azure Databricks workspace and a linked Azure Machine Learning workspace.

You have the following Python code segment in the Azure Machine Learning workspace:

import mlflowimport mlflow.azuremlimport azureml.mlflowimport azureml.core

from azureml.core import Workspace

subscription_id = \\'subscription_id\\'resourse_group = \\'resource_group_name\\'workspace_name = \\'workspace_name\\'

ws = Workspace.get(name=workspace_name,subscription_id=subscription_id,resource_group=resource_group)

experimentName = "/Users/{user_name}/{experiment_folder}/{experiment_name}"mlflow.set_experiment(experimentName)

uri = ws.get_mlflow_tracking_uri()mlflow.set_tracking_uri(uri)

Instructions: For each of the following statements, select Yes if the statement is true. Otherwise, select No.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

	Yes	No
A resource group and Azure Machine Learning workspace will be created.	○	○
An Azure Databricks experiment will be tracked only in the Azure Machine Learning workspace.	○	○
The epoch loss metric is set to be tracked.	○	○

Correct Answer:

Answer Area

	Yes	No
A resource group and Azure Machine Learning workspace will be created.	○	⬤
An Azure Databricks experiment will be tracked only in the Azure Machine Learning workspace.	⬤	○
The epoch loss metric is set to be tracked.	⬤	○

Box 1: No

The Workspace.get method loads an existing workspace without using configuration files.

ws = Workspace.get(name="myworkspace", subscription_id=\\'\\', resource_group=\\'myresourcegroup\\')

Box 2: Yes

MLflow Tracking with Azure Machine Learning lets you store the logged metrics and artifacts from your local runs into your Azure Machine Learning workspace.

The get_mlflow_tracking_uri() method assigns a unique tracking URI address to the workspace, ws, and set_tracking_uri() points the MLflow tracking URI to that address.

Box 3: Yes

Note: In Deep Learning, epoch means the total dataset is passed forward and backward in a neural network once.

Reference: https://docs.microsoft.com/en-us/python/api/azureml-core/azureml.core.workspace.workspace
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-use-mlflow

---

QUESTION 7

You are solving a classification task.

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

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

Which value should you use?

A. k=1

B. k=10

C. k=0.5

D. k=0.9

Correct Answer: B

Leave One Out (LOO) cross-validation

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

LOO CV is sometimes useful but typically doesn\'t shake up the data enough. The estimates from each fold are highly correlated and hence their average can have high variance. This is why the usual choice is K=5 or 10. It provides a good

compromise for the bias-variance tradeoff.

QUESTION 8

HOTSPOT

You are developing a machine learning, experiment by using Azure. The following images show the input and output of a machine learning experiment:

Input / Output diagram
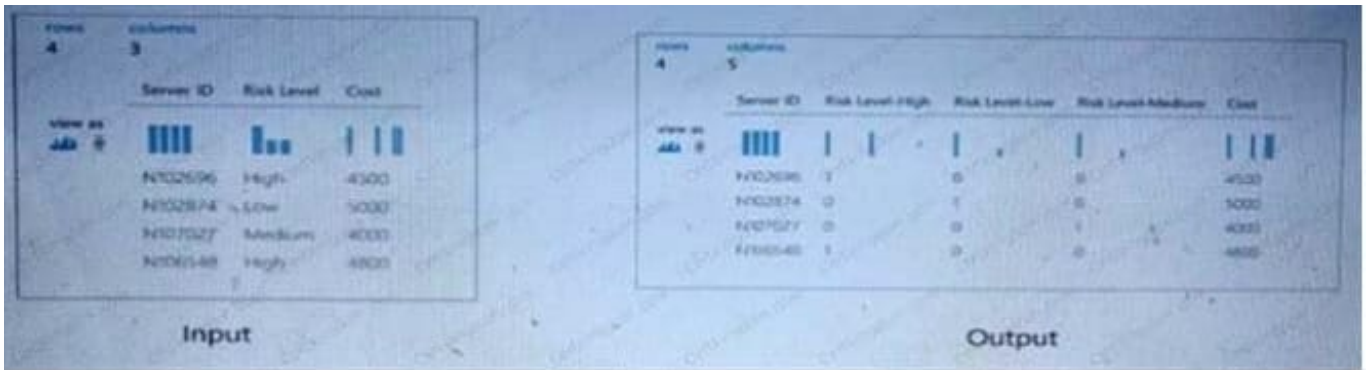
Use the drop-down menus to select the answer choice that answers each question based on the information presented in the graphic. NOTE: Each correct selection is worth one point.

Hot Area:

You need to perform the data transformation applied to the Risk Level column. Which module should you use?

▼
App*** Filter
Build Counting Transform
Convert to indicator Value

What is the expected input column type for this transformation?

▼
Categorical
Numerical
String

Correct Answer:

You need to perform the data transformation applied to the Risk Level column. Which module should you use?

▼
App*** Filter
Build Counting Transform
Convert to indicator Value

What is the expected input column type for this transformation?

▼
Categorical
Numerical
String

QUESTION 9

HOTSPOT

You are a lead data scientist for a project that tracks the health and migration of birds. You create a multi-image classification deep learning model that uses a set of labeled bird photos collected by experts. You plan to use the model to

develop a cross-platform mobile app that predicts the species of bird captured by app users.

You must test and deploy the trained model as a web service. The deployed model must meet the following requirements:

1.

An authenticated connection must not be required for testing.

2.

The deployed model must perform with low latency during inferencing.

3.

The REST endpoints must be scalable and should have a capacity to handle large number of requests when multiple end users are using the mobile application.

You need to verify that the web service returns predictions in the expected JSON format when a valid REST request is submitted.

Which compute resources should you use? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Context	Resource
Test	ds-workstation notebook VM ▼ ds-workstation notebook VM aks-compute cluster cpu-compute cluster gpu-compute cluster
Production	ds-workstation notebook VM ▼ ds-workstation notebook VM aks-compute cluster cpu-compute cluster gpu-compute cluster

Correct Answer:

Context

Test

Resource
ds-workstation notebook VM
aks-compute cluster
cpu-compute cluster
gpu-compute cluster

Production

Resource
ds-workstation notebook VM
aks-compute cluster
cpu-compute cluster
gpu-compute cluster

Box 1: ds-workstation notebook VM

An authenticated connection must not be required for testing.

On a Microsoft Azure virtual machine (VM), including a Data Science Virtual Machine (DSVM), you create local user accounts while provisioning the VM. Users then authenticate to the VM by using these credentials.

Box 2: gpu-compute cluster

Image classification is well suited for GPU compute clusters

Reference:

https://docs.microsoft.com/en-us/azure/machine-learning/data-science-virtual-machine/dsvm-common-identity

https://docs.microsoft.com/en-us/azure/architecture/reference-architectures/ai/training-deep-learning

---

QUESTION 10

You need to select a pre built development environment for a series of data science experiments. You must use the R language for the experiments.

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

NOTE: Each correct selection is worth one point.

A. MI.NET Library on a local environment

B. Azure Machine Learning Studio

C. Data Science Virtual Machine (OSVM)

D. Azure Data bricks

E. Azure Cognitive Services

Correct Answer: ABD

QUESTION 11

You plan to use the Hyperdrive feature of Azue Machine Learning to determine the optimal hyperparameter values when training a model. You must use Hyperdrive to try combinations of the following hyperparameter values. You must not apply an early termination policy.

1.

 learning_rate: any value between 0.001 and 0.1

2.

 batch_size: 16, 32, or 64

You need to configure the sampling method for the Hyperdrive experiment.

Which two sampling methods can you use? Each correct answer is a complete solution.

NOTE: Each correct selection is worth one point.

A. No sampling

B. Grid sampling

C. Bayesian sampling

D. Random sampling

Correct Answer: CD

C: Bayesian sampling is based on the Bayesian optimization algorithm and makes intelligent choices on the hyperparameter values to sample next. It picks the sample based on how the previous samples performed, such that the new sample

improves the reported primary metric. Bayesian sampling does not support any early termination policy

Example:

from azureml.train.hyperdrive import BayesianParameterSampling from azureml.train.hyperdrive import uniform, choice

param_sampling = BayesianParameterSampling({

"learning_rate": uniform(0.05, 0.1),

"batch_size": choice(16, 32, 64, 128)

}

)

D: 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. Incorrect Answers:

B: Grid sampling can be used if your hyperparameter space can be defined as a choice among discrete values and if you have sufficient budget to exhaustively search over all values in the defined search space. Additionally, one can use automated early termination of poorly performing runs, which reduces wastage of resources.

Example, the following space has a total of six samples:

from azureml.train.hyperdrive import GridParameterSampling from azureml.train.hyperdrive import choice

param_sampling = GridParameterSampling({

"num_hidden_layers": choice(1, 2, 3),

"batch_size": choice(16, 32)

}

)

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

---

QUESTION 12

You use Azure Machine Learning to train a model.

You must use a sampling method for tuning hyperparameters. The sampling method must pick samples based on how the model performed with previous samples.

You need to select a sampling method.

Which sampling method should you use?

A. Grid

B. Bayesian

C. Random

Correct Answer: B