Vendor: Microsoft

Exam Code: DP-100

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

Version: DEMO

Case Study 1
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.

QUESTION 1
You need to implement a scaling strategy for the local penalty detection data.

Which normalization type should you use?

A. Streaming
B. Weight
C. Batch
D. Cosine

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

Case Study 2
Overview
You are a data scientist for Fabrikam Residences, a company specializing in quality private and commercial property in the United States. Fabrikam Residences is considering expanding into Europe and has asked you to investigate prices for private residences in major European cities. You use Azure Machine Learning Studio to measure the median value of properties. You produce a regression model to predict property prices by using the Linear Regression and Bayesian Linear Regression modules.

Datasets

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

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

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

QUESTION 2
You need to select a feature extraction method.

Which method should you use?

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

Answer: C
Explanation:
In statistics, the Kendall rank correlation coefficient, commonly referred to as Kendall's tau coefficient (after the Greek letter), is a statistic used to measure the ordinal association between two measured quantities.
It is a supported method of the Azure Machine Learning Feature selection.
Scenario: When you train a Linear Regression module using a property dataset that shows data for property prices for a large city, you need to determine the best features to use in a model. You can choose standard metrics provided to measure performance before and after the feature importance process completes. You must ensure that the distribution of the features across multiple training models is consistent.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/feature-selection-modules

QUESTION 3
QUESTION 198
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 train a classification model by using a logistic regression algorithm.

You must be able to explain the model's predictions by calculating the importance of each feature, both as an overall global relative importance value and as a measure of local importance for a specific set of predictions.

You need to create an explainer that you can use to retrieve the required global and local feature importance values.

Solution: Create a MimicExplainer.

Does the solution meet the goal?

A. Yes
B. No

Answer: B
Explanation:
Instead use Permutation Feature Importance Explainer (PFI).
Note 1: Mimic explainer is based on the idea of training global surrogate models to mimic blackbox models.
A global surrogate model is an intrinsically interpretable model that is trained to approximate the predictions of any black box model as accurately as possible. Data scientists can interpret the surrogate model to draw conclusions about the black box model.
Note 2: Permutation Feature Importance Explainer (PFI): Permutation Feature Importance is a technique used to explain classification and regression models. At a high level, the way it works is by randomly shuffling data one feature at a time for the entire dataset and calculating how much the performance metric of interest changes. The larger the change, the more important that feature is. PFI can explain the overall behavior of any underlying model but does not explain individual predictions.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-machine-learning-interpretability

QUESTION 4
You retrain an existing model.

You need to register the new version of a model while keeping the current version of the model in the registry.

What should you do?

A. Register a model with a different name from the existing model and a custom property named version with the value 2.
B. Register the model with the same name as the existing model.
C. Save the new model in the default datastore with the same name as the existing model. Do not register the new model.
D. Delete the existing model and register the new one with the same name.

Answer: B
Explanation:
Model version: A version of a registered model. When a new model is added to the Model Registry, it is added as Version 1. Each model registered to the same model name increments the version number.
Reference:
https://docs.microsoft.com/en-us/azure/databricks/applications/mlflow/model-registry

QUESTION 5
You run an experiment that uses an AutoMLConfig class to define an automated machine learning task with a maximum of ten model training iterations. The task will attempt to find the best performing model based on a metric named accuracy.

You submit the experiment with the following code:

```
from azureml.core.experiment import Experiment
automl_experiment = Experiment(ws, 'automl_experiment')
automl_run = automl_experiment.submit(automl_config, show_output=True)
```

You need to create Python code that returns the best model that is generated by the automated machine learning task.

Which code segment should you use?

A. best_model = automl_run.get_details()
B. best_model = automl_run.get_metrics()
C. best_model = automl_run.get_file_names()[1]
D. best_model = automl_run.get_output()[1]

Answer: D
Explanation:
The get_output method returns the best run and the fitted model.
Reference:
https://notebooks.azure.com/azureml/projects/azureml-getting-started/html/how-to-use-azureml/automated-machine-learning/classification/auto-ml-classification.ipynb

QUESTION 6
You use the Azure Machine Learning SDK to run a training experiment that trains a classification model and calculates its accuracy metric.

The model will be retrained each month as new data is available.

You must register the model for use in a batch inference pipeline.

You need to register the model and ensure that the models created by subsequent retraining experiments are registered only if their accuracy is higher than the currently registered model.

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

NOTE: Each correct selection is worth one point.

A. Specify a different name for the model each time you register it.
B. Register the model with the same name each time regardless of accuracy, and always use the latest version of the model in the batch inferencing pipeline.
C. Specify the model framework version when registering the model, and only register subsequent models if this value is higher.
D. Specify a property named accuracy with the accuracy metric as a value when registering the model, and only register subsequent models if their accuracy is higher than the accuracy property value of the currently registered model.
E. Specify a tag named accuracy with the accuracy metric as a value when registering the model, and only register subsequent models if their accuracy is higher than the accuracy tag value of the currently registered model.

Answer: CE
Explanation:
E: Using tags, you can track useful information such as the name and version of the machine learning library used to train the model. Note that tags must be alphanumeric.
Reference:
https://notebooks.azure.com/xavierheriat/projects/azureml-getting-started/html/how-to-use-azureml/deployment/register-model-create-image-deploy-service/register-model-create-image-deploy-service.ipynb

QUESTION 7
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.

```
- learning_rate: any value between 0.001 and 0.1
- 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

Answer: CD
Explanation:
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 8
You are training machine learning models in Azure Machine Learning. You use Hyperdrive to tune the hyperparameter.

In previous model training and tuning runs, many models showed similar performance.

You need to select an early termination policy that meets the following requirements:

```
- accounts for the performance of all previous runs when evaluating the current run
- avoids comparing the current run with only the best performing run to date
```

Which two early termination policies should you use? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

A. Median stopping
B. Bandit
C. Default
D. Truncation selection

Answer: AC
Explanation:
The Median Stopping policy computes running averages across all runs and cancels runs whose best performance is worse than the median of the running averages.
If no policy is specified, the hyperparameter tuning service will let all training runs execute to completion.
Incorrect Answers:
B: BanditPolicy defines an early termination policy based on slack criteria, and a frequency and

delay interval for evaluation.
The Bandit policy takes the following configuration parameters:
slack_factor: The amount of slack allowed with respect to the best performing training run. This factor specifies the slack as a ratio.
D: The Truncation selection policy periodically cancels the given percentage of runs that rank the lowest for their performance on the primary metric. The policy strives for fairness in ranking the runs by accounting for improving model performance with training time. When ranking a relatively young run, the policy uses the corresponding (and earlier) performance of older runs for comparison. Therefore, runs aren't terminated for having a lower performance because they have run for less time than other runs.
Reference:
https://docs.microsoft.com/en-us/python/api/azureml-train-core/azureml.train.hyperdrive.medianstoppingpolicy
https://docs.microsoft.com/en-us/python/api/azureml-train-core/azureml.train.hyperdrive.truncationselectionpolicy
https://docs.microsoft.com/en-us/python/api/azureml-train-core/azureml.train.hyperdrive.banditpolicy

QUESTION 9
You are a data scientist working for a hotel booking website company. You use the Azure Machine Learning service to train a model that identifies fraudulent transactions.

You must deploy the model as an Azure Machine Learning real-time web service using the Model.deploy method in the Azure Machine Learning SDK. The deployed web service must return real-time predictions of fraud based on transaction data input.

You need to create the script that is specified as the entry_script parameter for the InferenceConfig class used to deploy the model.

What should the entry script do?

A. Register the model with appropriate tags and properties.
B. Create a Conda environment for the web service compute and install the necessary Python packages.
C. Load the model and use it to predict labels from input data.
D. Start a node on the inference cluster where the web service is deployed.
E. Specify the number of cores and the amount of memory required for the inference compute.

Answer: C
Explanation:
The entry script receives data submitted to a deployed web service and passes it to the model. It then takes the response returned by the model and returns that to the client. The script is specific to your model.
It must understand the data that the model expects and returns.
The two things you need to accomplish in your entry script are:
Loading your model (using a function called init())
Running your model on input data (using a function called run()) Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-deploy-and-where

QUESTION 10
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

Answer: D
Explanation:
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 11
You use the Azure Machine Learning SDK in a notebook to run an experiment using a script file in an experiment folder.

The experiment fails.

You need to troubleshoot the failed experiment.

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

A. Use the get_metrics() method of the run object to retrieve the experiment run logs.
B. Use the get_details_with_logs() method of the run object to display the experiment run logs.
C. View the log files for the experiment run in the experiment folder.
D. View the logs for the experiment run in Azure Machine Learning studio.
E. Use the get_output() method of the run object to retrieve the experiment run logs.

Answer: BD
Explanation:
Use get_details_with_logs() to fetch the run details and logs created by the run.
You can monitor Azure Machine Learning runs and view their logs with the Azure Machine Learning studio.
Incorrect Answers:
A: You can view the metrics of a trained model using run.get_metrics().
E: get_output() gets the output of the step as PipelineData.
Reference:
https://docs.microsoft.com/en-us/python/api/azureml-pipeline-core/azureml.pipeline.core.steprun
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-monitor-view-training-logs

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 have a Python script named train.py in a local folder named scripts. The script trains a regression model by using scikit-learn. The script includes code to load a training data file which is also located in the scripts folder.

You must run the script as an Azure ML experiment on a compute cluster named aml-compute.

You need to configure the run to ensure that the environment includes the required packages for model training. You have instantiated a variable named aml-compute that references the target compute cluster.

Solution: Run the following code:

```
from azureml.train.estimator import Estimator
sk_est = Estimator(source_directory='./scripts',
 compute_target=aml-compute,
 entry_script='train.py')
```

Does the solution meet the goal?

A. Yes
B. No

Answer: B
Explanation:
There is a missing line: conda_packages=['scikit-learn'], which is needed.
Correct example:
sk_est = Estimator(source_directory='./my-sklearn-proj',
script_params=script_params,
compute_target=compute_target,
entry_script='train.py',
conda_packages=['scikit-learn'])
Note:
The Estimator class represents a generic estimator to train data using any supplied framework. This class is designed for use with machine learning frameworks that do not already have an Azure Machine Learning pre-configured estimator. Pre-configured estimators exist for Chainer, PyTorch, TensorFlow, and SKLearn.
Example:
from azureml.train.estimator import Estimator
script_params = {
# to mount files referenced by mnist dataset
'--data-folder': ds.as_named_input('mnist').as_mount(),
'--regularization': 0.8
}
Reference:
https://docs.microsoft.com/en-us/python/api/azureml-train-core/azureml.train.estimator.estimator

QUESTION 13
Hotspot Question

You plan to use Hyperdrive to optimize the hyperparameters selected when training a model. You create the following code to define options for the hyperparameter experiment:

```
import azureml.train.hyperdrive.parameter_expressions as pe
from azureml.train.hyperdrive import GridParameterSampling, HyperDriveConfig

param_sampling = GridParameterSampling({
  "max_depth" : pe.choice(6, 7, 8, 9),
  "learning_rate" : pe.choice(0.05, 0.1, 0.15)
  })
hyperdrive_run_config = HyperDriveConfig(
  estimator = estimator,
  hyperparameter_sampling = param_sampling,
  policy = None,
  primary_metric_name = "auc",
  primary_metruc_goal = PrimaryMetricGoal.MAXIMIZE,
  max_total_runs = 50,
  max_concurrent_runs = 4)
```

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

NOTE: Each correct selection is worth one point.

Answer Area

	Yes	No
There will be 50 runs for this hyperparameter tuning experiment.	○	○
You can use the policy parameter in the HyperDriveConfig class to specify a security policy.	○	○
The experiment will create a run for every possible value for the learning rate parameter between 0.05 and 0.15.	○	○

Answer:

Answer Area

	Yes	No
There will be 50 runs for this hyperparameter tuning experiment.	○	○
You can use the policy parameter in the HyperDriveConfig class to specify a security policy.	○	○
The experiment will create a run for every possible value for the learning rate parameter between 0.05 and 0.15.	○	○

Explanation:
Box 1: No
max_total_runs (50 here)
The maximum total number of runs to create. This is the upper bound; there may be fewer runs when the sample space is smaller than this value.

Box 2: Yes
Policy EarlyTerminationPolicy
The early termination policy to use. If None - the default, no early termination policy will be used.

Box 3: No
Discrete hyperparameters are specified as a choice among discrete values. choice can be:
one or more comma-separated values
a range object
any arbitrary list object

Reference:
https://docs.microsoft.com/en-us/python/api/azureml-train-core/azureml.train.hyperdrive.hyperdriveconfig
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-tune-hyperparameters

QUESTION 14
Hotspot Question

You are preparing to build a deep learning convolutional neural network model for image classification. You create a script to train the model using CUDA devices.

You must submit an experiment that runs this script in the Azure Machine Learning workspace.

The following compute resources are available:

```
- a Microsoft Surface device on which Microsoft Office has been
installed. Corporate IT policies prevent the installation of additional
software
- a Compute Instance named ds-workstation in the workspace with 2 CPUs
and 8 GB of memory
- an Azure Machine Learning compute target named cpu-cluster with eight
CPU-based nodes
- an Azure Machine Learning compute target named gpu-cluster with four
CPU and GPU-based nodes
```

You need to specify the compute resources to be used for running the code to submit the experiment, and for running the script in order to minimize model training time.

Which resources should the data scientist use? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

Resource type	Option
Run code to submit the experiment	the Microsoft Surface device the ds-workstation notebook VM the cpu-cluster compute target the gpu-cluster compute target
Run the training script	the ds-workstation notebook VM the cpu-compute target the gpu-compute target the Microsoft Surface device

Answer:

Answer Area

Resource type	Option
Run code to submit the experiment	the Microsoft Surface device **the ds-workstation notebook VM** the cpu-cluster compute target the gpu-cluster compute target
Run the training script	the ds-workstation notebook VM the cpu-compute target **the gpu-compute target** the Microsoft Surface device

Explanation:
Box 1: the ds-workstation notebook VM

Box 2: the gpu-compute target
Just as GPUs revolutionized deep learning through unprecedented training and inferencing performance, RAPIDS enables traditional machine learning practitioners to unlock game-changing performance with GPUs. With RAPIDS on Azure Machine Learning service, users can accelerate the entire machine learning pipeline, including data processing, training and inferencing, with GPUs from the NC_v3, NC_v2, ND or ND_v2 families. Users can unlock performance gains of more than 20X (with 4 GPUs), slashing training times from hours to minutes and dramatically reducing time-to-insight.

Reference:
https://azure.microsoft.com/sv-se/blog/azure-machine-learning-service-now-supports-nvidia-s-rapids/

QUESTION 15
Hotspot Question

A biomedical research company plans to enroll people in an experimental medical treatment trial.

You create and train a binary classification model to support selection and admission of patients to the trial. The model includes the following features: Age, Gender, and Ethnicity.

The model returns different performance metrics for people from different ethnic groups.

You need to use Fairlearn to mitigate and minimize disparities for each category in the Ethnicity feature.

Which technique and constraint should you use? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

Option	Value
Technique	Grid search Outlier detection Dimensionality reduction
Constraint	Demographic parity False-positive rate parity

Answer:

Answer Area

Option	Value
Technique	Grid search
Constraint	Demographic parity

Explanation:
Box 1: Grid Search
Fairlearn open-source package provides postprocessing and reduction unfairness mitigation algorithms: ExponentiatedGradient, GridSearch, and ThresholdOptimizer.
Note: The Fairlearn open-source package provides postprocessing and reduction unfairness mitigation algorithms types:
Reduction: These algorithms take a standard black-box machine learning estimator (e.g., a LightGBM model) and generate a set of retrained models using a sequence of re-weighted training datasets.
Post-processing: These algorithms take an existing classifier and the sensitive feature as input.
Box 2: Demographic parity
The Fairlearn open-source package supports the following types of parity constraints: Demographic parity, Equalized odds, Equal opportunity, and Bounded group loss.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/concept-fairness-ml

QUESTION 16
You train and register a machine learning model. You create a batch inference pipeline that uses the model to generate predictions from multiple data files.
You must publish the batch inference pipeline as a service that can be scheduled to run every night.
You need to select an appropriate compute target for the inference service.
Which compute target should you use?

A. Azure Machine Learning compute instance
B. Azure Machine Learning compute cluster
C. Azure Kubernetes Service (AKS)-based inference cluster
D. Azure Container Instance (ACI) compute target

Answer: B
Explanation:
Azure Machine Learning compute clusters is used for Batch inference. Run batch scoring on serverless compute. Supports normal and low-priority VMs. No support for real-time inference.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/concept-compute-target

Thank You for Trying Our Product

Braindump2go Certification Exam Features:

★ More than **99,900** Satisfied Customers Worldwide.

★ Average **99.9%** Success Rate.

★ **Free Update** to match latest and real exam scenarios.

★ **Instant Download** Access! No Setup required.

★ Questions & Answers are downloadable in **PDF** format and **VCE** test engine format.

★ Multi-Platform capabilities - **Windows, Laptop, Mac, Android, iPhone, iPod, iPad**.

★ **100%** Guaranteed Success or **100%** Money Back Guarantee.

★ **Fast**, helpful support **24x7**.

View list of all certification exams: http://www.braindump2go.com/all-products.html

10% Discount Coupon Code: **ASTR14**