DP-100 – cases and exam questions

Module 1: Design a machine learning solution

Multiple choice questions

. Which of the following tasks is associated with designing a data ingestion solution for machine learning projects in Azure?

A. Implementing model training algorithms.
B. Choosing appropriate compute for machine learning.
C. Selecting features for model training.
D. Identifying the data source and desired data format.

.

. What could be a reason to train a model with GPU compute instead of CPU compute?

A. The model is trained in a Jupyter notebook.
B. The model is trained using Azure Databricks.
C. The model is trained using PyTorch
D. The model is trained using PySpark.

.

. When aiming to design a model deployment solution in Azure for machine learning projects, which aspect should you primarily consider first?

A. Selecting the appropriate machine learning algorithms.
B. Converting notebooks to scripts for production.
C. Choosing the appropriate compute resources.

.

. What is the main purpose of designing a machine learning operations (MLOps) solution?

A. Building and training machine learning models.
B. Making your machine learning workloads robust and reproducible.
C. Deploying your machine learning models to production.
D. Evaluating your machine learning models with appropriate metrics.

After training a machine learning model to predict customer churn, you need to evaluate its performance to ensure it meets the business requirements. Which of the following metrics is most appropriate for evaluating the performance of a classification model?

A. Mean Absolute Error (MAE)
B. Root Mean Squared Error (RMSE)
C. Precision, Recall, and F1 Score
D. R-squared (R^2)

Which of the following is a key benefit of using Azure Machine Learning environments?

A. Automatic scaling of compute resources
B. Version control for dependencies and packages
C. Built-in data visualization tools

D. Integrated development environment (IDE) for coding

In Azure Machine Learning, what is the primary purpose of an environment?

A. To manage user access and permissions
B. To define the compute resources for training models
C. To specify the software dependencies and configurations for experiments
D. To store and manage datasets

Open-ended questions

1. What is the benefit of storing data in a data storage solution on Azure, separated from compute?
2. What is the benefit of Spark compute?
3. What could be a reason for a team to choose Azure AI services to train a machine learning model?

Use case 1 - **Predicting Customer Churn for a Telecom Company**

Scenario:

You are a data scientist at a telecom company. The company is facing a high rate of customer churn and wants to implement a machine learning solution to predict which customers are likely to leave. This will help the company take proactive measures to retain these customers.

Objectives:

1. **Understand the Problem**: Discuss the business problem and the importance of predicting customer churn.
2. **Data Collection and Preparation**: Identify the types of data needed (e.g., customer demographics, usage patterns, customer service interactions) and discuss data cleaning and preprocessing steps.
3. **Feature Engineering**: Brainstorm potential features that could be useful for the model (e.g., average call duration, number of complaints, monthly charges).
4. **Model Selection**: Discuss different types of models that could be used (e.g., logistic regression, decision trees, random forests) and the criteria for selecting a model.
5. **Model Training and Evaluation**: Outline the steps for training the model, including splitting the data into training and test sets, and discuss evaluation metrics (e.g., accuracy, precision, recall, F1 score).
6. **Model Deployment**: Discuss how to deploy the model into a production environment and monitor its performance over time.
7. **Ethical Considerations**: Address any ethical concerns related to the use of customer data and the potential impact of the model's predictions.

Discussion Points:

- **Data Quality**: How to handle missing or inconsistent data.
- **Feature Importance**: How to determine which features are most important for predicting churn.
- **Model Interpretability**: The importance of having a model that is interpretable and explainable to stakeholders.
- **Bias and Fairness**: How to ensure the model does not unfairly target certain groups of customers.
- **Continuous Improvement**: Strategies for continuously improving the model as more data becomes available.

Part 2: Explore and configure the Azure Machine Learning workspace

Multiple choice questions

Which developer tool should be used when interacting with the Azure Machine Learning workspace with the purpose of automation?

A. The Azure portal.
B. The Azure Machine Learning Studio.

C. The Azure Machine Learning Python SDK.
D. The Azure Machine Learning extension for the Azure CLI.

Which URI should be used to connect to data stored in an Azure Data Lake (Gen2)?

A. http(s)
B. abfs(s)
C. azureml
D. blob

Which compute can be used for experimentation in notebooks?

A. Serverless Spark
B. Compute cluster
C. Containers
D. Kubernetes clusters

Which of the following best describes an MLTable in Azure Machine Learning?

A. A single file containing raw data
B. A directory containing multiple files
C. A YAML-based file that defines how data files should be loaded and transformed
D. A Python script used for data preprocessing

What is the primary advantage of using an MLTable over directly referencing files or folders in Azure Machine Learning?

A. MLTable allows for the execution of Python scripts
B. MLTable provides a detailed blueprint for data loading, including transformations and column type definitions
C. MLTable automatically trains machine learning models
D. MLTable is used to store model artifacts

Which of the following is a valid type of data asset in Azure Machine Learning SDK v2?

A. `uri_file`
B. `uri_folder`
C. `mltable`
D. **All of the above**

How do you create a data asset from a local folder using Azure ML SDK v2?

A. `DataAsset.create(path='path/to/folder')`
B. `DataAsset.from_path(path='path/to/folder')`
C. `Data(path='path/to/folder', type=AssetTypes.URI_FOLDER)`
D. `DataAsset.load(path='path/to/folder')`

Open-ended questions

1. Explain one benefit of using data assets in the Azure Machine Learning workspace.

2. Explain one approach to save compute costs.
3. Explain the difference between curated and custom environments in Azure Machine Learning.
4. What is a benefit of using components?

Module 2. Explore data, and run experiments

Part 1: Experiment with Azure ML

Multiple choice questions

What is the primary purpose of using Automated Machine Learning? Automate model deployment.

A. Automate data exploration.
B. Automate model selection and tuning.
C. Automate data visualization.

When using Automated Machine Learning to find the best classification model, what does the system optimize?

A. Training time
B. Model complexity
C. Model accuracy metrics
D. Number of features

How does MLflow contribute to tracking model training in notebooks? MLflow optimizes model hyperparameters.

A. MLflow facilitates data preprocessing.
B. MLflow logs and monitors model metrics.
C. MLflow automates model deployment.

Which of the following methods is commonly used for hyperparameter tuning in machine learning models?

A) Gradient Descent
B) Grid Search
C) Principal Component Analysis (PCA)
D) K-Means Clustering

Which of the following techniques can be used to reduce the computational cost of hyperparameter tuning?

A) Random Search
B) Grid Search
C) Gradient Descent
D) Principal Component Analysis (PCA)

In the context of hyperparameter tuning, what is the purpose of using cross-validation?

A) To increase the size of the training dataset
B) To ensure that the model is not overfitting to the training data
C) To reduce the number of hyperparameters
D) To convert categorical data into numerical data

Open-ended questions

A. What is a potential benefit of using Automated Machine Learning?
B. What is MLflow and why would it be used?
C. What are the different tasks that can be chosen in Automated Machine Learning

Use case 2 – Automl

Customer Use Case: Customer Churn Prediction

Scenario: A telecommunications company wants to predict customer churn to proactively retain customers. They have historical data on customer behavior, including usage patterns, service complaints, and demographic information. The company decides to use Azure Machine Learning's AutoML capabilities to build and deploy a churn prediction model.

Tasks:

1. **Data Preparation:**
   - **Question:** What steps would you take to prepare the data for AutoML?
2. **Configuring AutoML:**
   - **Question:** How would you configure an AutoML experiment in Azure Machine Learning?
3. **Running AutoML Experiment:**
   - **Question:** What steps are involved in running the AutoML experiment and selecting the best model?
4. **Model Deployment:**
   - **Question:** How can you deploy the best model to production using Azure Machine Learning?
5. **Model Monitoring and Management:**
   - **Question:** What strategies would you use to monitor and manage the deployed model?

Module 3: Train and deploy models

Multiple choice questions

When tracking a machine learning model with MLflow, what method should be used to log the value of the regularization parameter?
a. mlflow.log_param()
b. mlflow.log_metric()
c. mlflow.log_artifact()

When executing a sequence of multiple scripts in Azure Machine Learning, what type of job is run?
a. Single
b. Command
c. Sweep

d. Pipeline

What is an important aspect of ensuring code is production-ready?
a. Refactoring code to functions.
b. Adding comments to code for readability.
c. Converting scripts to notebooks.

What feature in the Responsible AI dashboard should be used to determine whether your findings related to fairness, error analysis, and causality are a result of your dataset's distribution?

A. Error analysis
B. Feature importance
C. Data analysis

Where in the MLmodel file will you find which framework was used to train the model?

A. Artifact path
B. Flavor
C. Signature

What is shown in the fairness assessment when different cohorts of the data perform differently when comparing selected performance metrics?

A. Disparity in model performance
B. Disparity in selection rate
C. Disparity in data cohorts

When deploying a model for real-time predictions, what is the easiest way to accomplish this?

A. Deploy an MLflow model to a batch endpoint.
B. Deploy a custom model to a batch endpoint.
C. Deploy an MLflow model to a managed online endpoint.

How can a managed online endpoint quickly be tested?

A. Through the Azure portal.
B. Through the Azure Machine Learning Studio.
C. Through a notebook in the Azure Machine Learning workspace.

Open-ended questions

A. What is the difference between a managed online endpoint and a batch endpoint?
B. How do you troubleshoot a batch scoring job?
C. What is the difference between logging a model as an artifact and logging it as a model?
D. What is the purpose of the MLmodel file?
E. What is the difference between aggregate and individual feature importance?

Use case 3 – MLFlow

Customer Use Case: Predictive Maintenance for Manufacturing Equipment

Scenario: A manufacturing company wants to implement a predictive maintenance system to reduce downtime and maintenance costs. They have historical data on equipment failures, maintenance logs, and sensor readings from various machines. The company wants to build a machine learning model to predict equipment failures before they occur. They have chosen to use MLflow to manage the machine learning lifecycle.

Tasks:

1. **Model Training:**
   - **Question:** How would you use MLflow to track experiments and model training?
2. **Model Deployment:**
   - **Question:** How can MLflow help in deploying the best model to production?
3. **Model Monitoring:**
   - **Question:** What strategies would you use to monitor the model's performance in production?
4. **Collaboration:**
   - **Question:** How can MLflow facilitate collaboration among data scientists and engineers in this project?

Use case 4 – Best practice for deploying models with Azure ML

A customer asks you about deploying models in Azure ML. Discuss why to use cloud solutions like Azure ML for deployment.

Optimize language models for AI applications

Prepare for model optimization

Question 1:What is the primary goal of fine-tuning a pre-trained language model?

- A) To train the model from scratch
- B) To adapt the model to a specific task or domain
- C) To reduce the size of the model
- D) To increase the number of parameters in the model

Question 2: Which technique is commonly used to prevent overfitting when fine-tuning language models?

- A) Increasing the learning rate
- B) Using dropout regularization
- C) Reducing the size of the training dataset
- D) Adding more layers to the model

Question 3: What is the purpose of using a validation dataset during model optimization?

- A) To train the model
- B) To test the model's performance after training
- C) To guide hyperparameters and evaluate performance
- D) To increase the size of the training dataset

Question 4: What is the role of learning rate scheduling in optimizing language models?

- A) To keep the learning rate constant throughout training
- B) To increase the learning rate as training progresses
- C) To adjust the learning rate dynamically to improve convergence
- D) To reduce the size of the model

Question 5: What is the primary purpose of testing a deployed model in the playground?

- A) To retrain the model with new data
- B) To evaluate the model's performance on real-world inputs
- C) To increase the model's accuracy
- D) To optimize the model's architecture

Optimize through prompt engineering and prompt flow

Question 1: What is the primary goal of prompt engineering in optimizing language models?

- A) To modify the model's architecture
- B) To design effective prompts that guide the model's responses
- C) To increase the size of the training dataset
- D) To reduce the number of parameters in the model

Question 2: Which technique can improve the effectiveness of prompts in language models?

- A) Using ambiguous phrasing
- B) Providing clear instructions and context
- C) Reducing the length of the prompt
- D) Avoiding examples in the prompt

Question 3: What is the purpose of prompt flow in optimizing language models?

- A) To test multiple prompts and select the best-performing one
- B) To reduce the computational cost of the model
- C) To increase the model's training speed
- D) To simplify the model's architecture

Question 4: Which feature of prompt flow allows users to evaluate the model's responses systematically?

- A) Randomized testing
- B) Automated scoring and comparison
- C) Manual evaluation
- D) Model retraining

Question 5: How can examples in a prompt improve the model's output?

- A) By reducing the model's accuracy
- B) By providing a clear pattern for the model to follow
- C) By increasing the complexity of the prompt
- D) By limiting the model's ability to generalize

Question 6: What is the primary purpose of defining chaining logic in the Prompt Flow SDK?

- A) To train the model on new data
- B) To create workflows that combine multiple prompts and model responses
- C) To reduce the size of the model
- D) To increase the number of model parameters

Question 7: Which feature in the Prompt Flow SDK allows chaining logic to handle conditional workflows?

- A) Static prompts
- B) Decision nodes
- C) Model retraining
- D) Batch processing

Question 8: How does chaining logic improve the performance of language models in complex tasks?

- A) By simplifying the model architecture
- B) By enabling sequential processing of prompts and responses
- C) By reducing the computational cost of the model
- D) By increasing the size of the training dataset

Question 9: What is the role of input validation in chaining logic workflows?

- A) To ensure the model generates diverse outputs
- B) To verify the correctness of inputs before processing
- C) To increase the model's training speed
- D) To reduce the number of prompts in the workflow

Optimize through Retrieval Augmented Generation (RAG)

Question 1: What is the main benefit of using Retrieval-Augmented Generation (RAG) in a data science solution?

- A) It replaces the need for embedding models
- B) It augments model responses with up-to-date external knowledge
- C) It reduces latency by eliminating retrieval steps
- D) It guarantees 100% factual accuracy

Question 2: In a typical RAG pipeline on Azure, which service is used to store and index vectors for fast similarity search?

- A) Azure Blob Storage
- B) Azure SQL Database
- C) Azure AI Search
- D) Azure Data Lake

Question 3: During model evaluation, you notice hallucinations in RAG outputs. Which approach is most effective to reduce them?

- A) Removing the retrieval step entirely
- B) Expanding the retrieval corpus indiscriminately

- C) Adding a filtering layer for relevance
- D) Training the model exclusively on generative tasks

Question 4: Which metric combination provides the best insight into the effectiveness of your RAG implementation?

- A) Retrieval recall and end-to-end answer accuracy
- B) Model parameter count and embedding vector size
- C) Latency and storage cost only
- D) Number of documents retrieved and CPU utilization

Optimize through fine-tuning

Question 1: What is the primary goal of fine-tuning a pre-trained model in a machine learning solution?

- A) Train a model from scratch on a large dataset
- B) Adapt the model to a specific domain or task with minimal data
- C) Compress model parameters for deployment
- D) Generate synthetic data for training

Question 2: Which hyperparameter is most critical to tune in order to avoid unstable updates and ensure convergence during fine-tuning?

- A) Batch size
- B) Learning rate
- C) Number of epochs
- D) Embedding dimension

Question 4: During fine-tuning, which dataset split is used to validate the model and tune hyperparameters while avoiding overfitting?

- A) Training set
- B) Validation set
- C) Test set
- D) Unseen set

Use Case: Legal Contract Review Assistant

Instructions: Read the use case and answer the questions below. Reflect on how to balance precision, latency, and cost when using only existing models in Azure AI Foundry. Use the following discussion point for discussion:

- *Design & Implementation*
- *Retrieval-Augmented Prompt Flow*
- *Prompt Engineering*
- *Performance Optimization*
- *Cost Management*
- *Monitoring & Evaluation*
- *Trade-offs between smaller vs. larger models in Foundry*

The use case

A corporate legal team needs an AI assistant/Agent that:

- Ingests large volumes of contracts (PDFs, Word docs)

- Extracts and highlights key clauses (e.g., termination, indemnification)

- Answers ad-hoc, contract-specific questions ("What's the notice period?")

- Flags high-risk language based on custom policy rules