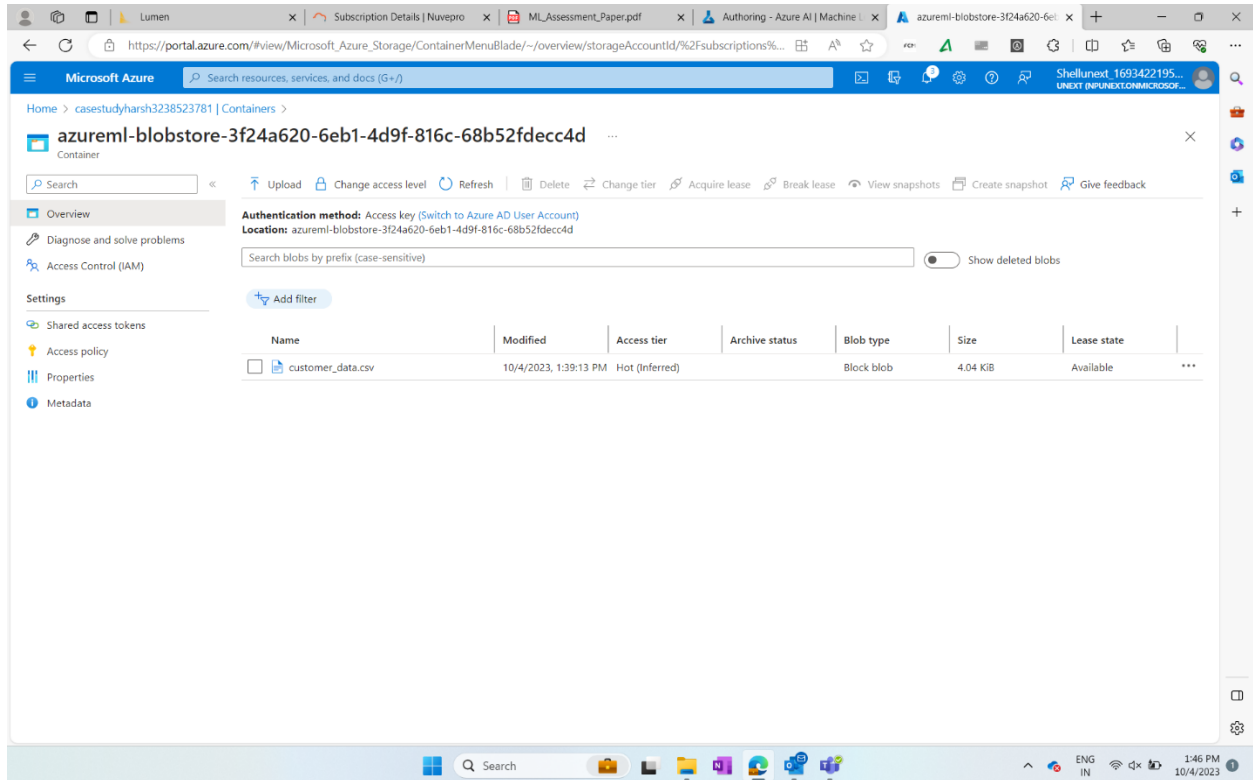


Case Study (ML)

Screenshots of Azure Machine Learning Studio

AzureML Blob Storage



The screenshot displays the Azure Machine Learning Studio interface, specifically the Blob Storage section for a container named `azureml-blobstore-3f24a620-6eb1-4d9f-816c-68b52fdecc4d`. The interface includes a search bar, a list of blobs, and a table of blob details.

Authentication method: Access key (Switch to Azure AD User Account)
Location: azureml-blobstore-3f24a620-6eb1-4d9f-816c-68b52fdecc4d

Search blobs by prefix (case-sensitive) ☐ Show deleted blobs

Add filter

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state	
<input type="checkbox"/> customer_data.csv	10/4/2023, 1:39:13 PM	Hot (Inferred)		Block blob	4.04 KiB	Available	...

Creation of data asset on AzureML Workspace

The screenshot displays the Azure ML Machine Learning Studio interface. The browser address bar shows the URL: https://ml.azure.com/dataset/CaseStudy_02/latest/details?wsid=/subscriptions/6e6fb922-d220-4e51-bf86-0cccc06be535/resourcegroup...

The interface is titled "Azure AI | Machine Learning Studio". The left sidebar contains navigation options: All workspaces, Home, Model catalog (PREVIEW), Authoring (Notebooks, Automated ML, Designer, Prompt flow (PREVIEW)), Assets (Data, Jobs, Components, Pipelines, Environments, Models, Endpoints), and Manage (Compute, Monitoring (PREVIEW)).

The main content area shows the details for the dataset "CaseStudy_02" (Version: 1 (latest)). The tabs include Details, Consume, Explore, Models, and Jobs. The "Details" tab is active, showing the following information:

- Attributes:**
 - Type: Folder (uri_folder)
 - Dataset type (from Azure ML v1 APIs): File
 - Created by: Shellunext unextIDA93
 - Files in dataset: 1
 - Total size of files in dataset: 4.044 KiB
 - Current version: 1
 - Latest version: 1
 - Created time: Oct 4, 2023 1:44 PM
 - Modified time: Oct 4, 2023 1:44 PM
- Tags:** No data
- Description:** Click edit icon to add a description
- Data sources:**
 - Datastore:** workspaceblobstore
 - Relative path:** customer_data.csv
 - Actions:** View in datastores browse, View in Azure Portal
 - Datastore URI:** azureml:/subscriptions/6e6fb922-d220-4e51-bf86-0cccc06be535/resourcegroup...
 - Storage URI:** https://casestudyharsh3238523781.blob.core.windows.net/azureml-blobstore-3...

The Windows taskbar at the bottom shows the time as 1:46 PM on 10/4/2023.

Configuring the workspace:

Train model:

The image displays two screenshots of the Azure AI Machine Learning Studio interface, illustrating the process of configuring a workspace and training a model.

Top Screenshot: Configuring the workspace

- The interface shows the "CaseStudyWorkspace_01" workspace.
- The "Clean Missing Data" component is selected, and its configuration panel is visible on the right. The "Columns to be cleaned" are set to "Age, AnnualIncome, SpendingScore". The "Minimum missing value ratio" is 0.0, and the "Maximum missing value ratio" is 1.0. The "Cleaning mode" is set to "Replace with mean".
- The "Split Data" component is also visible in the workspace, with its configuration panel on the right. The "Splitting mode" is set to "Split Rows", and the "Fraction of rows in the first output dataset" is 0.7.

Bottom Screenshot: Training the model

- The interface shows the "CaseStudyWorkspace_01" workspace.
- The "Train Model" component is selected, and its configuration panel is visible on the right. The "Splitting mode" is set to "Split Rows", and the "Fraction of rows in the first output dataset" is 0.7.
- The workspace diagram shows the flow of data from the "Clean Missing Data" component to the "Split Data" component, and then to the "Train Model" component. The "Train Model" component is connected to the "Split Data" component, and the output is labeled "Trained model".

Azure AI | Machine Learning Studio

Unnext > casestudyharsh > Designer > Authoring

split

Tags: All Add filter

Data Component

3 Split Data

Split Image Directory

Apply SQL Transformation

Boosted Decision Tree Regression

Untrained model

Train Model

Label column *

Column names: SpendingScore

Model explanations

False

Output settings

Input settings

Run settings

Node information

Component information

CaseStudyWorkspace_01

Save Pipeline interface

Configure & Submit

1:50 PM 10/4/2023

Azure AI | Machine Learning Studio

Unnext > casestudyharsh > Designer > Authoring

CaseStudyWorkspace_01

Save Pipeline interface

Configure & Submit

1:53 PM 10/4/2023

Flowchart illustrating the machine learning pipeline:

```
graph TD; A[CaseStudy_01] --> B[Data Input]; B --> C[Clean Missing Data clean_missing_data]; C --> D[Split Data split_data]; D --> E[Boosted Decision Tree Regression boosted_decision_tree_regression]; E --> F[Untrained model]; F --> G[Train Model train_model]; G --> H[Score Model score_model]; H --> I[Scored dataset scored_dataset]; I --> J[Evaluate Model evaluate_model]; J --> K[Evaluation results];
```

Hyperparameter training model:

The screenshot displays the Azure Machine Learning Studio interface. The left sidebar contains navigation options: All workspaces, Home, Model catalog, Authoring (Notebooks, Automated ML, Designer), Assets (Data, Jobs, Components, Pipelines, Environments, Models, Endpoints), and Manage (Compute, Monitoring). The main workspace is titled 'CaseStudyWorkspace_02' and shows a pipeline with a 'Boosted Decision Tree Regression' component. The 'Tune Model Hyperparameters' configuration panel is open on the right, showing the following settings:

- Specify parameter sweeping mode: Random sweep
- Maximum number of runs on random sweep: 5
- Random seed: 0
- Metric for measuring performance for classification: Accuracy
- Metric for measuring performance for regression: Mean absolute error
- Label column: SpendingScore
- Output settings: >
- Input settings: >
- Run settings: >

The bottom status bar indicates the system language is English (ENG IN) and the date is 10/4/2023.

Browser tabs: Lumen, Subscription Details | Nuvepro, ML_Assessment_Paper.pdf, Authoring - Azure AI | Machine L, azureml-blobstore-3f24a620-6e...

URL: https://ml.azure.com/visualinterface/authoring/Normal/04f014a6-a6f6-4983-9383-a08ca9bb8b33?wsid=/subscriptions/6e6fb922-d220-...

Azure AI | Machine Learning Studio

Unext > casestudyharsh > Designer > Authoring

Buttons: Undo, Redo, Validate, Show lineage, Clone, AutoSave, Configure & Submit

CaseStudyWorkspace_02

Save, Pipeline interface

Left sidebar:

- All workspaces
- Home
- Model catalog PREVIEW
- Authoring
 - Notebooks
 - Automated ML
 - Designer**
 - Prompt flow PREVIEW
- Assets
 - Data
 - Jobs
 - Components
 - Pipelines
 - Environments
 - Models
 - Endpoints
- Manage
 - Compute
 - Monitoring PREVIEW

Diagram:

```
graph TD; A[CaseStudy_02] --> B[Data Input]; B --> C[Clean Missing Data]; C --> D[Split Data]; D --> E[Train Model]; E --> F[Score Model]; F --> G[Evaluate Model];
```

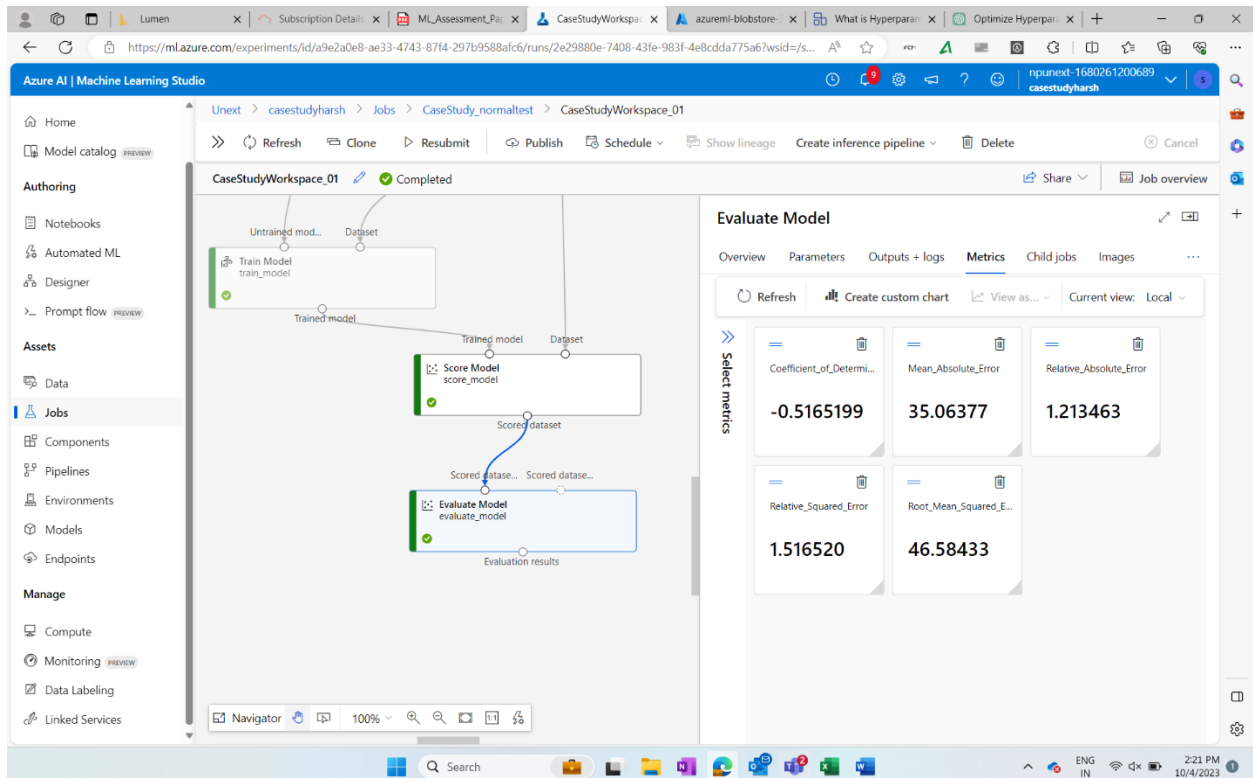
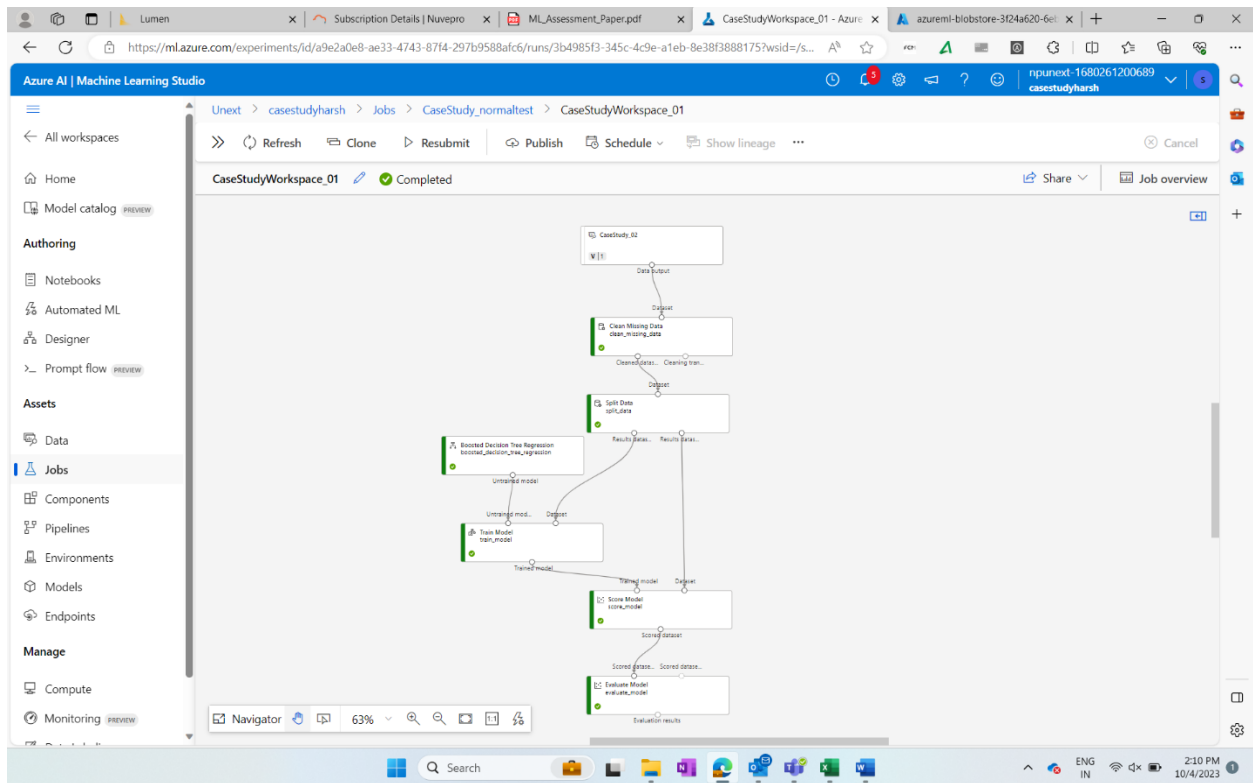
Diagram details:

- CaseStudy_02 (Data Input) connects to Clean Missing Data (Cleaning step).
- Clean Missing Data connects to Split Data (Splitting step).
- Split Data connects to Train Model (Training step).
- Train Model connects to Score Model (Scoring step).
- Score Model connects to Evaluate Model (Evaluation step).
- Train Model also connects to a sub-diagram: Untrained model connects to Tune Model Hyperparameters (Tuning step), which connects to Sweep results (Sweeping step), which connects to Trained Model (Training step), which connects to Train Model.

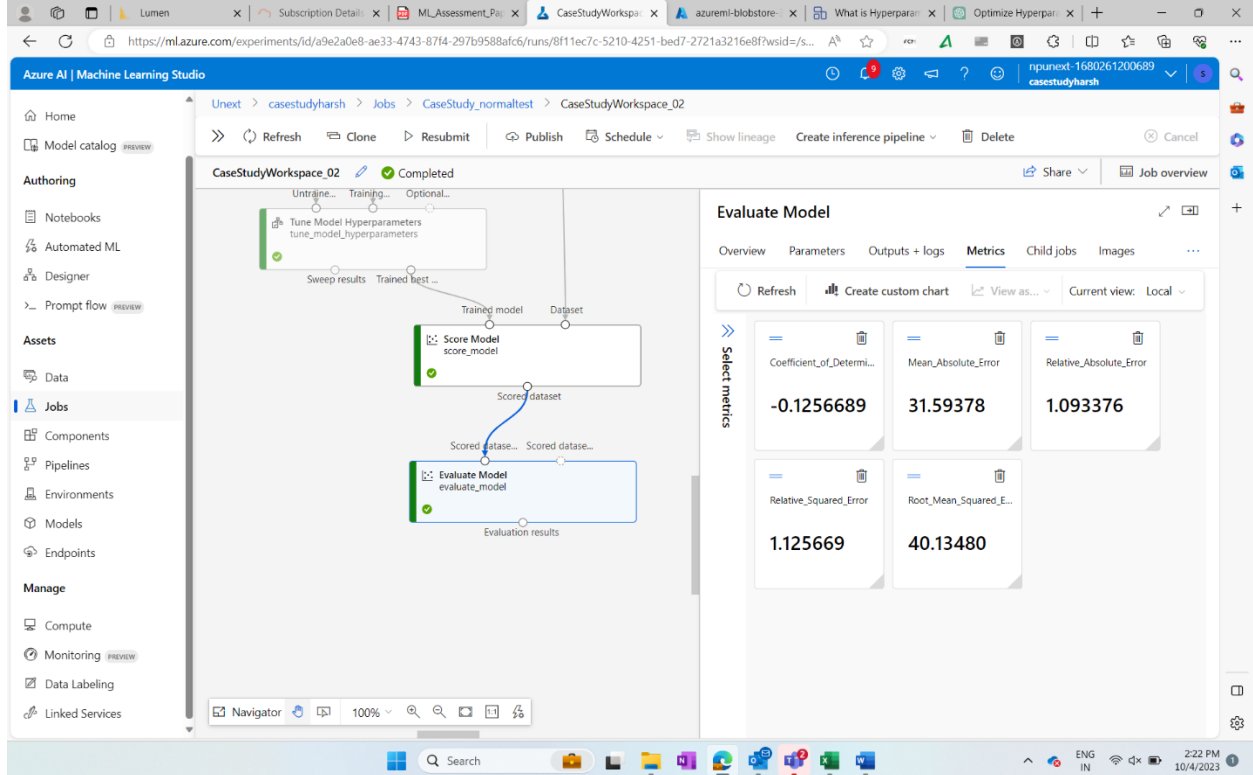
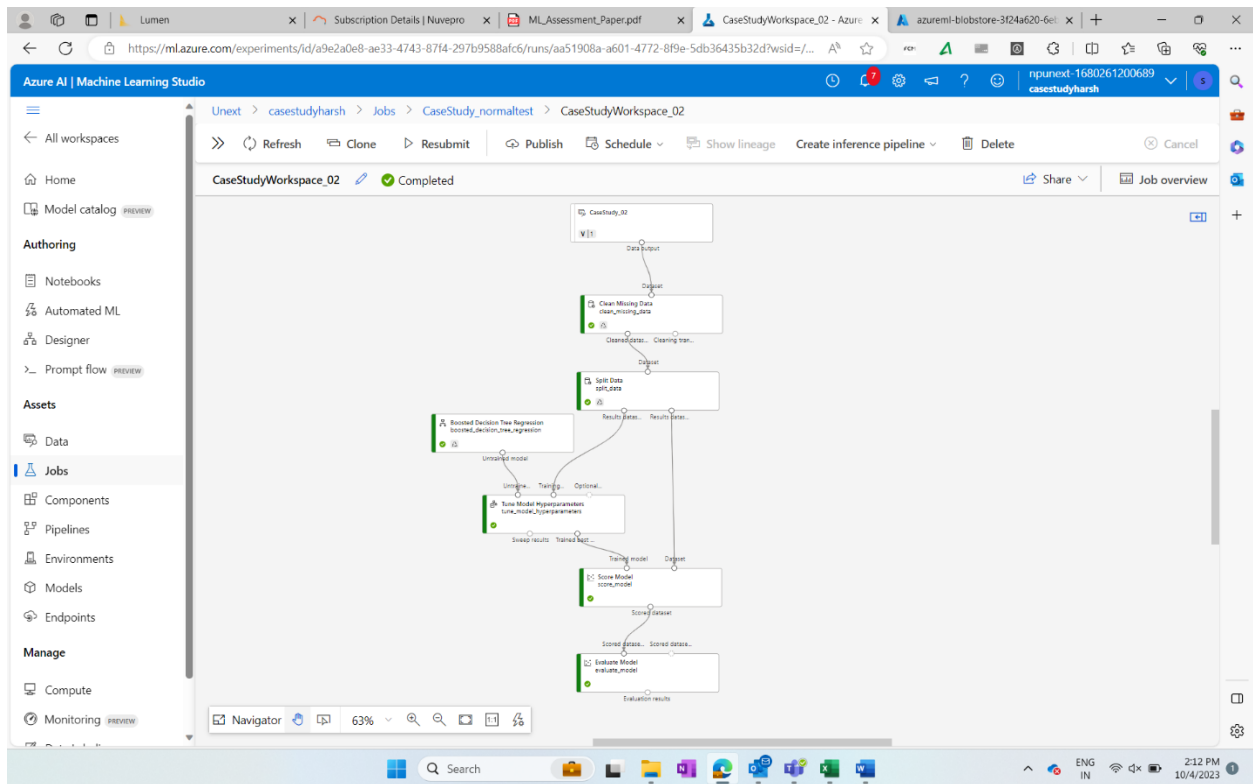
Bottom bar: Search, 60%, 10/4/2023, 1:59 PM

Outputs:

Train model:



Hyperparameter Training:



Question/Answers

1. Firstly, the data set is uploaded through blob storage account to pipeline. Then, `Clean_Missing_Data` is used to clear out any missing values, so that the garbage data is removed and only the data with relevant information is kept for training. Then, the data is split into training and testing data with 70% kept for training and 30% for testing. Finally, the split training data goes for training using the ML algorithm "Boosted Decision Tree Regression".
2. It is important to split the dataset into training and testing sets as the model shouldn't be tested with the same data that it has been trained with, as that won't test its capabilities and it would be impossible to check whether the model has been trained to recognize patterns in new, unfamiliar data. This helps in model evaluation as the model's performance is tested on separate data than what it's used to.
3. I have chosen Boosted decision tree algorithm for this problem as there was only one csv file, which had large amount of data. We can also use linear regression for the same, but I believe that a decision tree in this case would be better.
4. Hyperparameter tuning involves finding the best hyperparameter values for a machine learning algorithm to improve its performance. This optimization aims to minimize a predefined loss function, leading to better results with fewer errors.
Random search, as opposed to grid search, randomly explores hyperparameter combinations and returns the best-performing one after multiple iterations. It's efficient for handling many hyperparameters and wide search ranges, offering faster results without user-defined biases. However, it may not guarantee the absolute best hyperparameter combination due to its random approach.