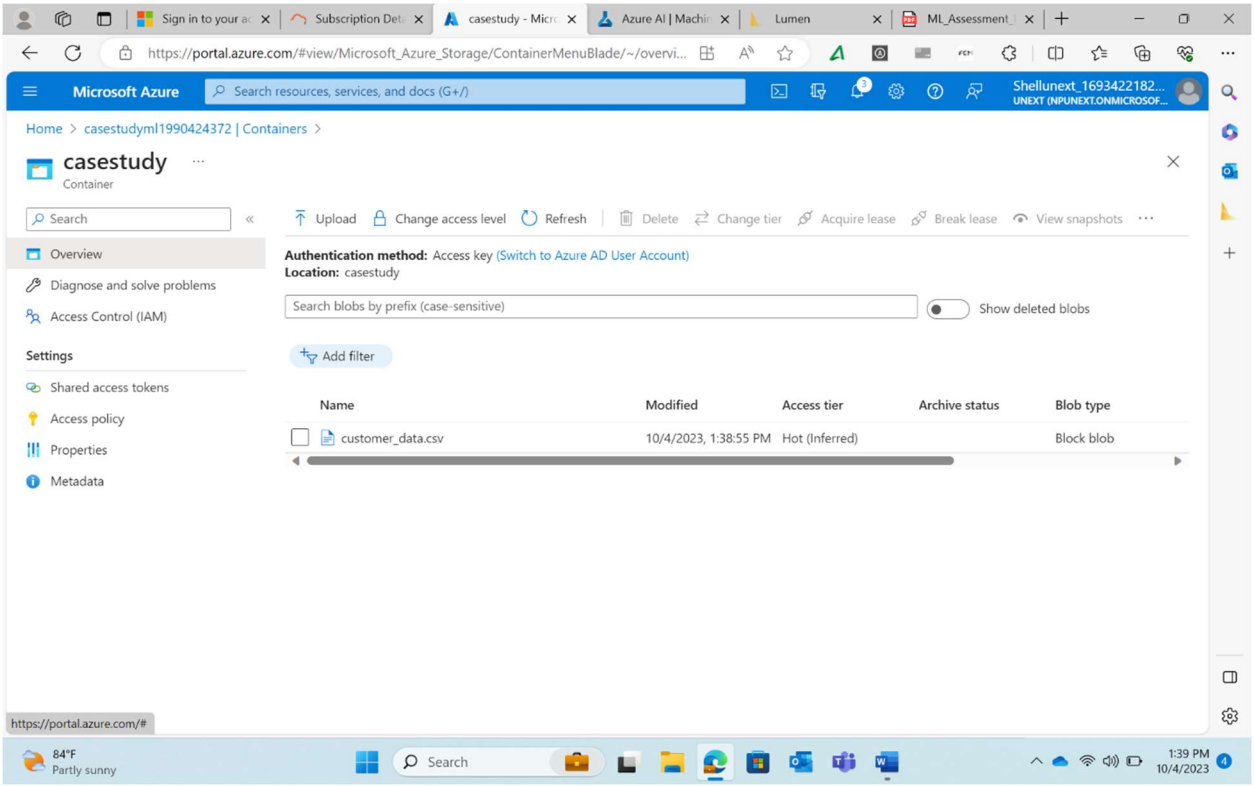
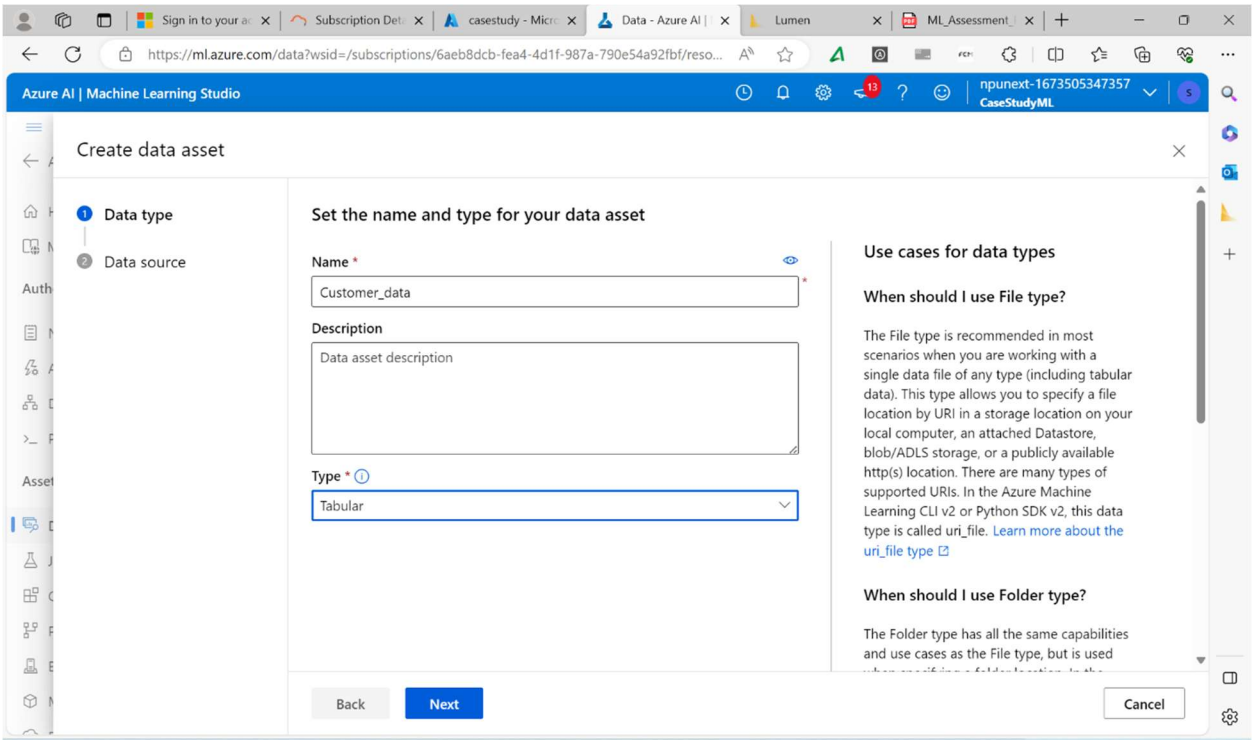
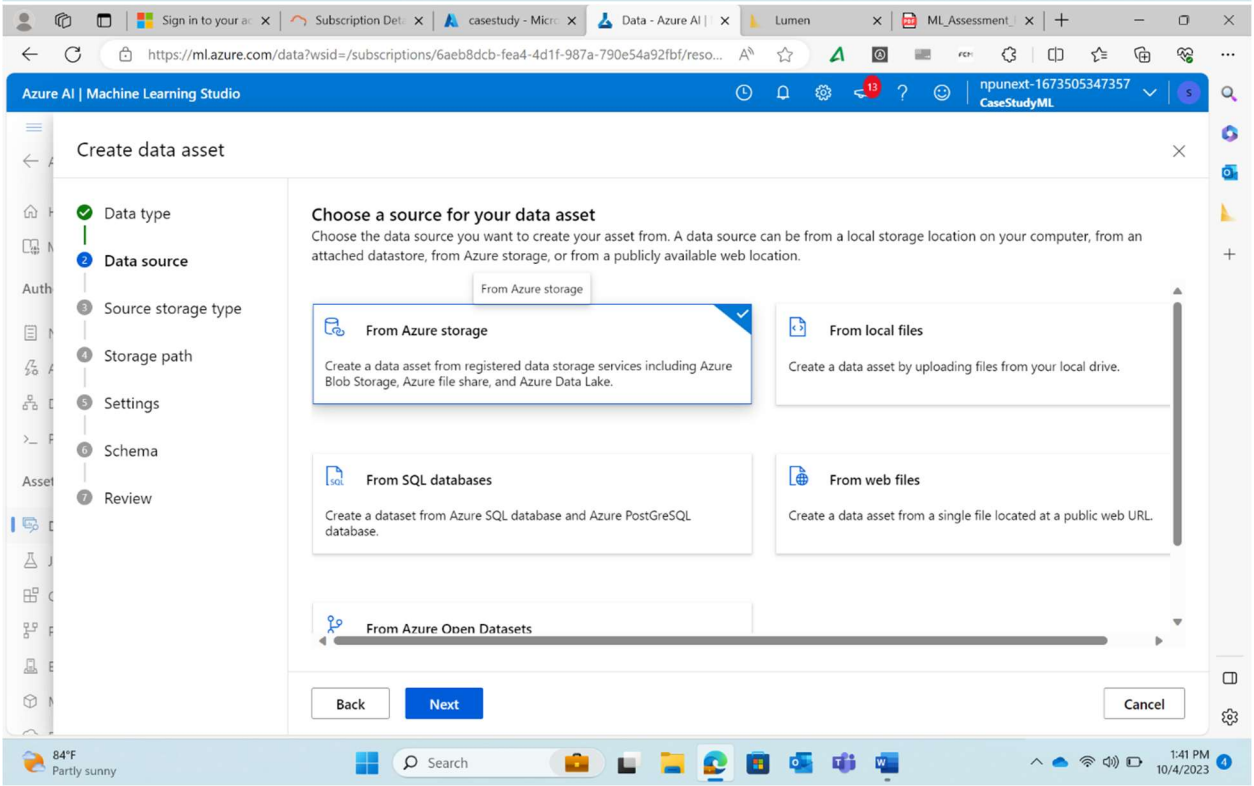
Case Study ML

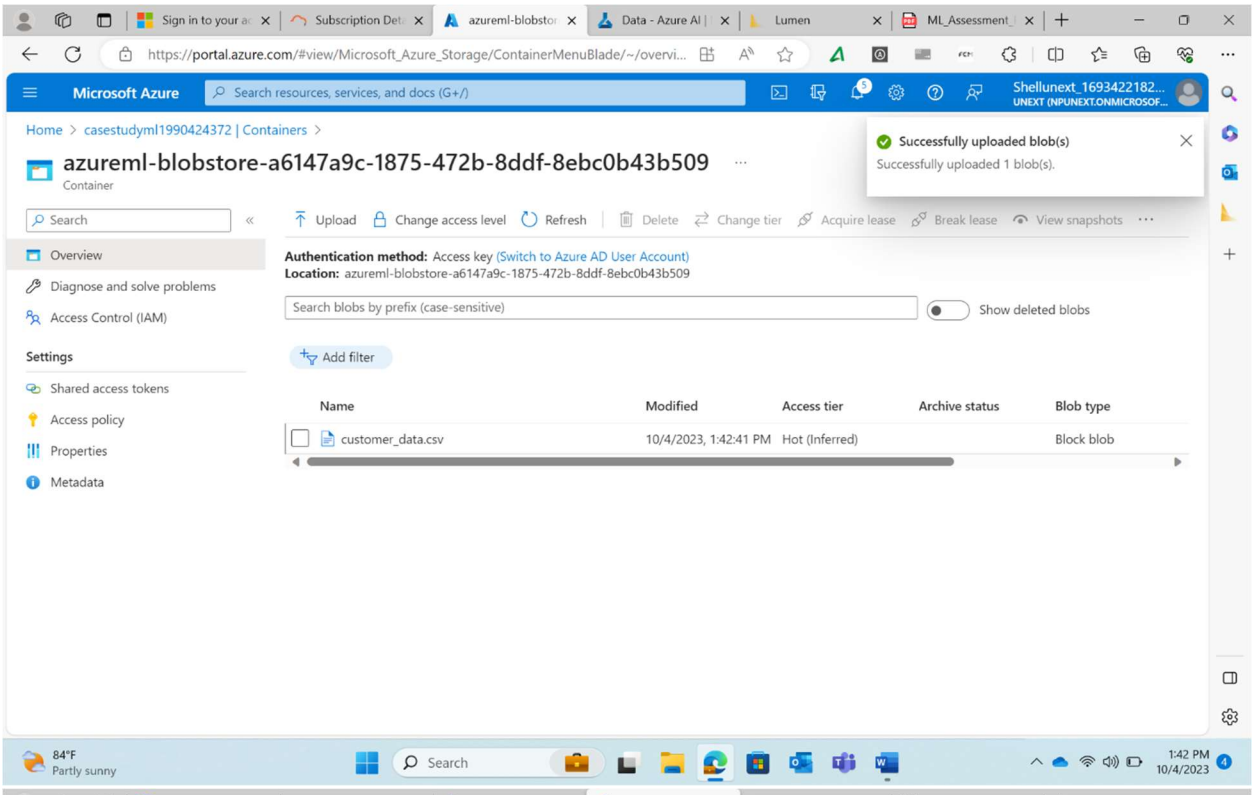
- By Thosic Tarun Bellana

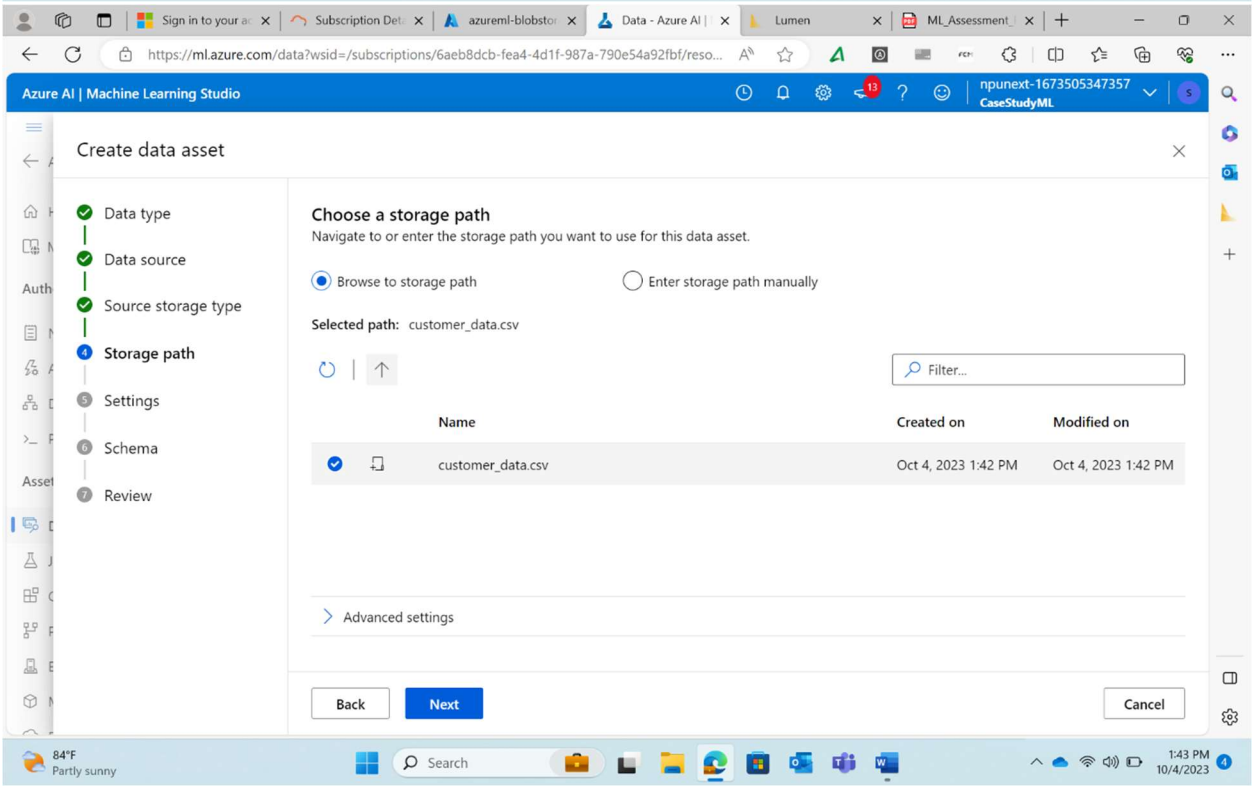
Data Preparation:

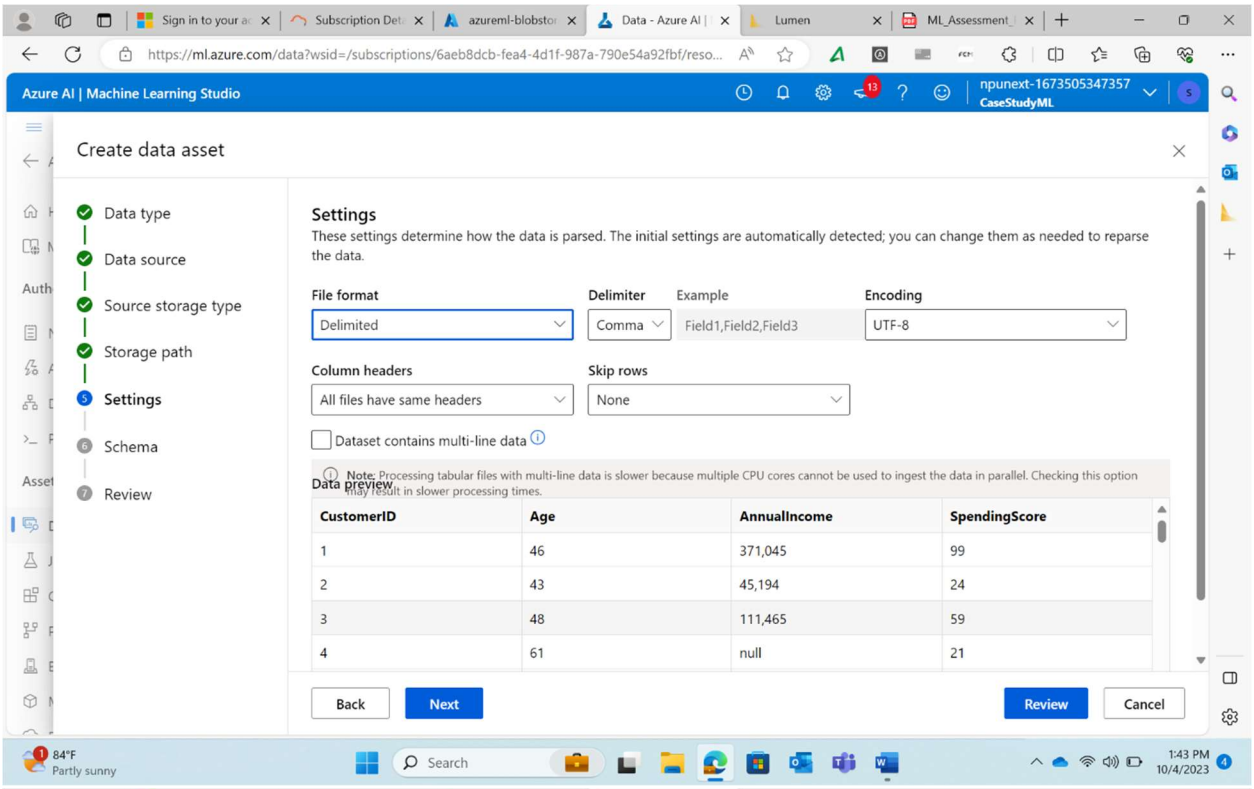


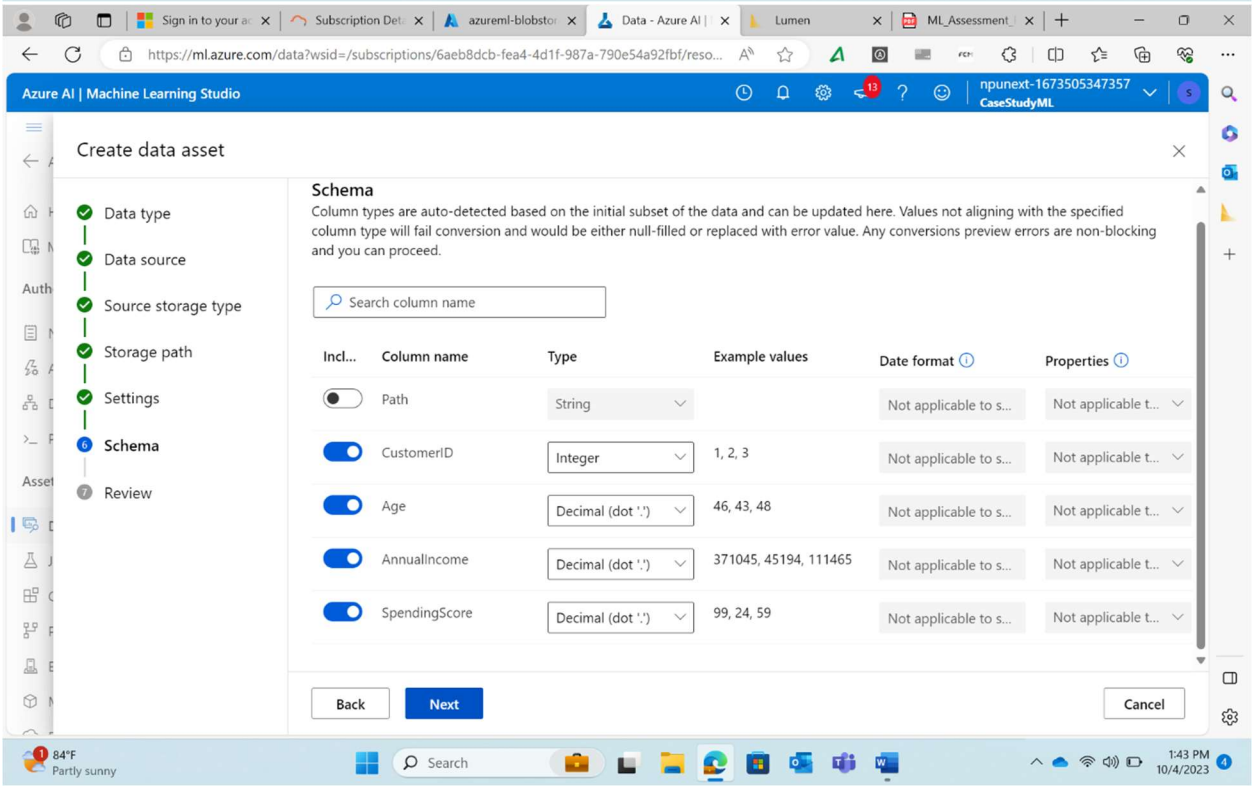

Azure AI | Machine Learning Studio
npunext-1673505347357
CaseStudyML
Create data asset
Data type
Data source
Set the name and type for your data asset
Name *
Customer_data
Description
Data asset description
Type *
Tabular
Use cases for data types
When should I use File type?
The File type is recommended in most scenarios when you are working with a single data file of any type (including tabular data). This type allows you to specify a file location by URI in a storage location on your local computer, an attached Datastore, blob/ADLS storage, or a publicly available http(s) location. There are many types of supported URIs. In the Azure Machine Learning CLI v2 or Python SDK v2, this data type is called uri_file. Learn more about the uri_file type
When should I use Folder type?
The Folder type has all the same capabilities and use cases as the File type, but is used
Back
Next
Cancel

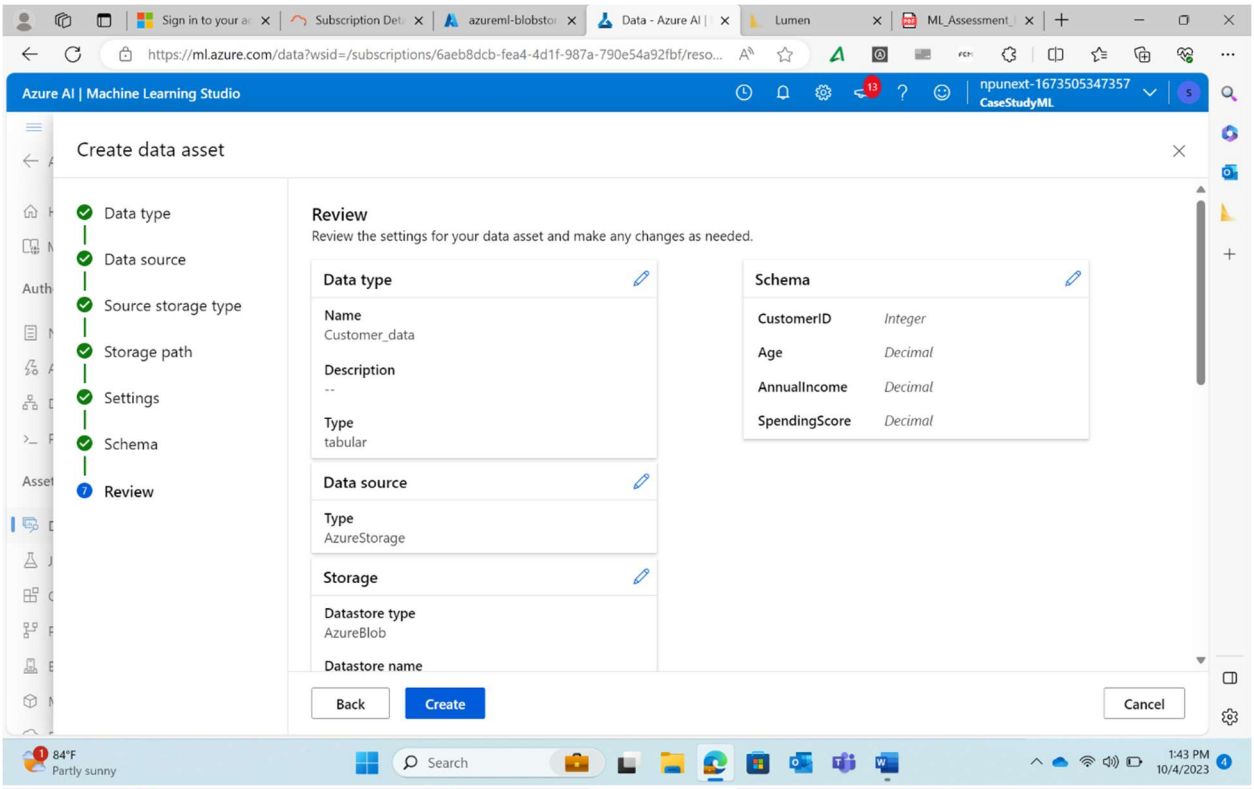

Azure AI | Machine Learning Studio
npunext-1673505347357
CaseStudyML
Create data asset
Data type
Data source
Source storage type
Storage path
Settings
Schema
Review
Choose a source for your data asset
Choose the data source you want to create your asset from. A data source can be from a local storage location on your computer, from an attached datastore, from Azure storage, or from a publicly available web location.
From Azure storage
From Azure storage
Create a data asset from registered data storage services including Azure Blob Storage, Azure file share, and Azure Data Lake.
From local files
Create a data asset by uploading files from your local drive.
From SQL databases
Create a dataset from Azure SQL database and Azure PostGreSQL database.
From web files
Create a data asset from a single file located at a public web URL.
From Azure Open Datasets
Back
Next
Cancel

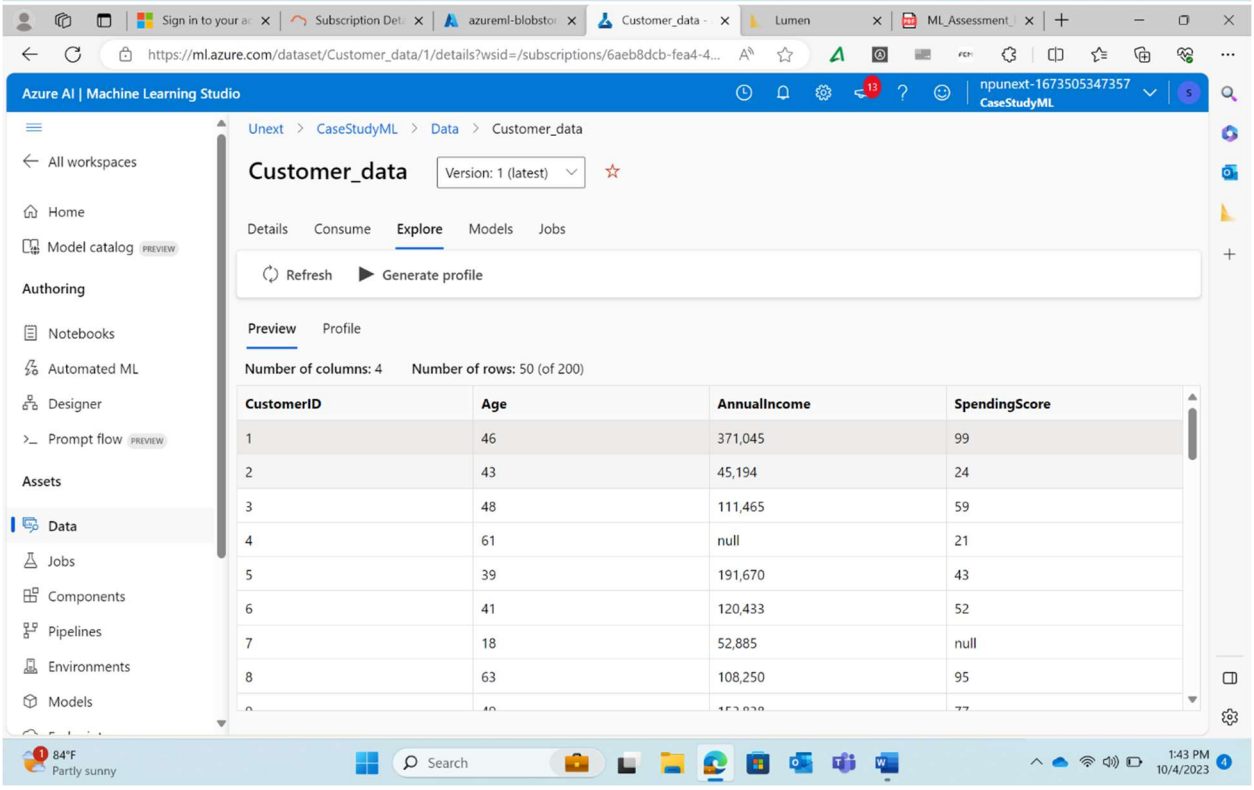

Microsoft Azure
Search resources, services, and docs (G+/)
Home > casestudyml1990424372 | Containers >
azureml-blobstore-a6147a9c-1875-472b-8ddf-8ebc0b43b509
Container
Successfully uploaded blob(s)
Successfully uploaded 1 blob(s).
Search
Overview
Diagnose and solve problems
Access Control (IAM)
Settings
Shared access tokens
Access policy
Properties
Metadata
Upload
Change access level
Refresh
Delete
Change tier
Acquire lease
Break lease
View snapshots
Authentication method: Access key (Switch to Azure AD User Account)
Location: azureml-blobstore-a6147a9c-1875-472b-8ddf-8ebc0b43b509
Search blobs by prefix (case-sensitive)
Show deleted blobs
Add filter
Name
Modified
Access tier
Archive status
Blob type
customer_data.csv
10/4/2023, 1:42:41 PM
Hot (Inferred)
Block blob


Azure AI | Machine Learning Studio
npunext-1673505347357
CaseStudyML
Create data asset
Data type
Data source
Source storage type
Storage path
Settings
Schema
Review
Choose a storage path
Navigate to or enter the storage path you want to use for this data asset.
Browse to storage path
Enter storage path manually
Selected path: customer_data.csv
Filter...
Name
Created on
Modified on
customer_data.csv
Oct 4, 2023 1:42 PM
Oct 4, 2023 1:42 PM
Advanced settings
Back
Next
Cancel


Azure AI | Machine Learning Studio
npunext-1673505347357
CaseStudyML
Create data asset
Data type
Data source
Source storage type
Storage path
Settings
Schema
Review
Settings
These settings determine how the data is parsed. The initial settings are automatically detected; you can change them as needed to reparse the data.
File format
Delimited
Delimiter
Comma
Example
Field1,Field2,Field3
Encoding
UTF-8
Column headers
All files have same headers
Skip rows
None
Dataset contains multi-line data
Note: Processing tabular files with multi-line data is slower because multiple CPU cores cannot be used to ingest the data in parallel. Checking this option may result in slower processing times.
Data preview
CustomerID
Age
AnnualIncome
SpendingScore
1
46
371,045
99
2
43
45,194
24
3
48
111,465
59
4
61
null
21
Back
Next
Review
Cancel


Azure AI | Machine Learning Studio
npunext-1673505347357
CaseStudyML
Create data asset
Data type
Data source
Source storage type
Storage path
Settings
Schema
Review
Schema
Column types are auto-detected based on the initial subset of the data and can be updated here. Values not aligning with the specified column type will fail conversion and would be either null-filled or replaced with error value. Any conversions preview errors are non-blocking and you can proceed.
Search column name
Incl...
Column name
Type
Example values
Date format
Properties
Path
String
CustomerID
Integer
1, 2, 3
Age
Decimal (dot '.')
46, 43, 48
AnnualIncome
Decimal (dot '.')
371045, 45194, 111465
SpendingScore
Decimal (dot '.')
99, 24, 59
Not applicable to s...
Not applicable t...
Back
Next
Cancel

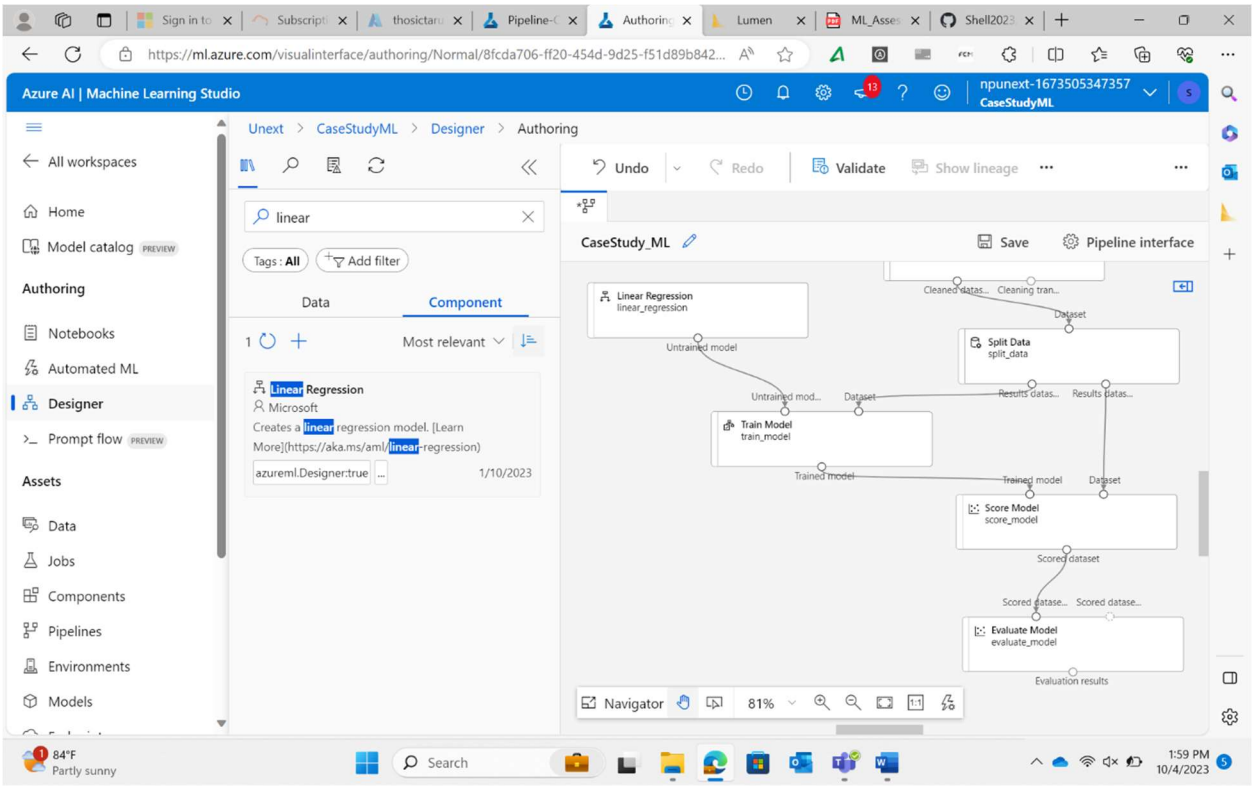




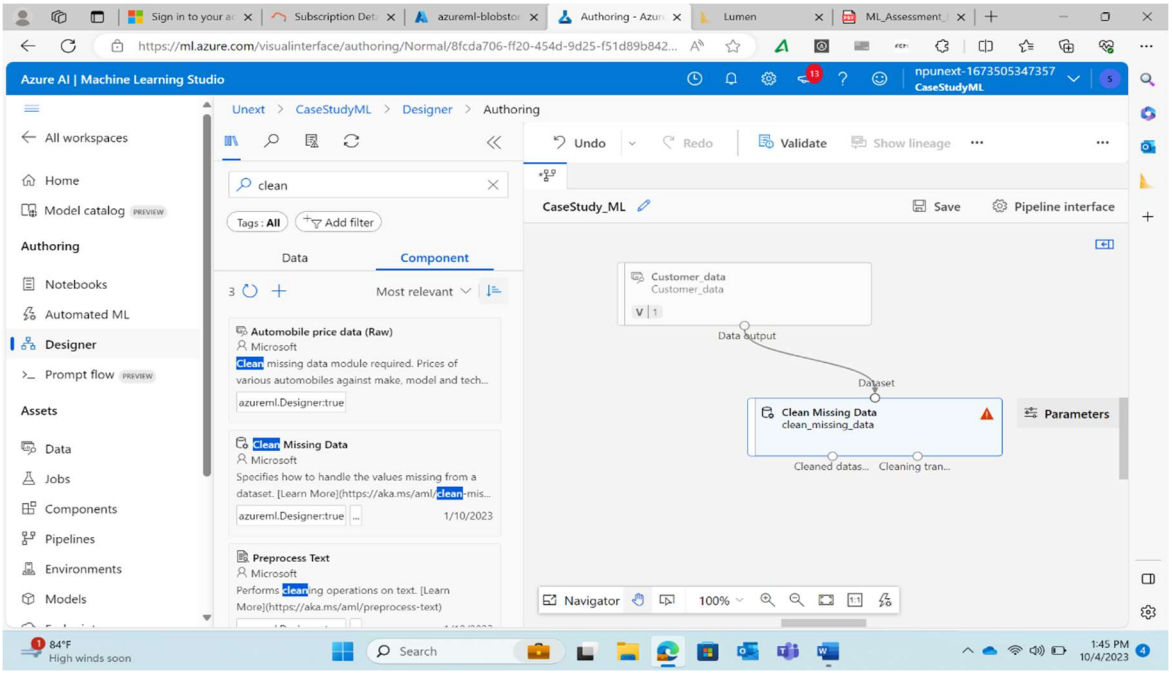
CustomerID	Age	AnnualIncome	SpendingScore
1	46	371,045	99
2	43	45,194	24
3	48	111,465	59
4	61	null	21
5	39	191,670	43
6	41	120,433	52
7	18	52,885	null
8	63	108,250	95

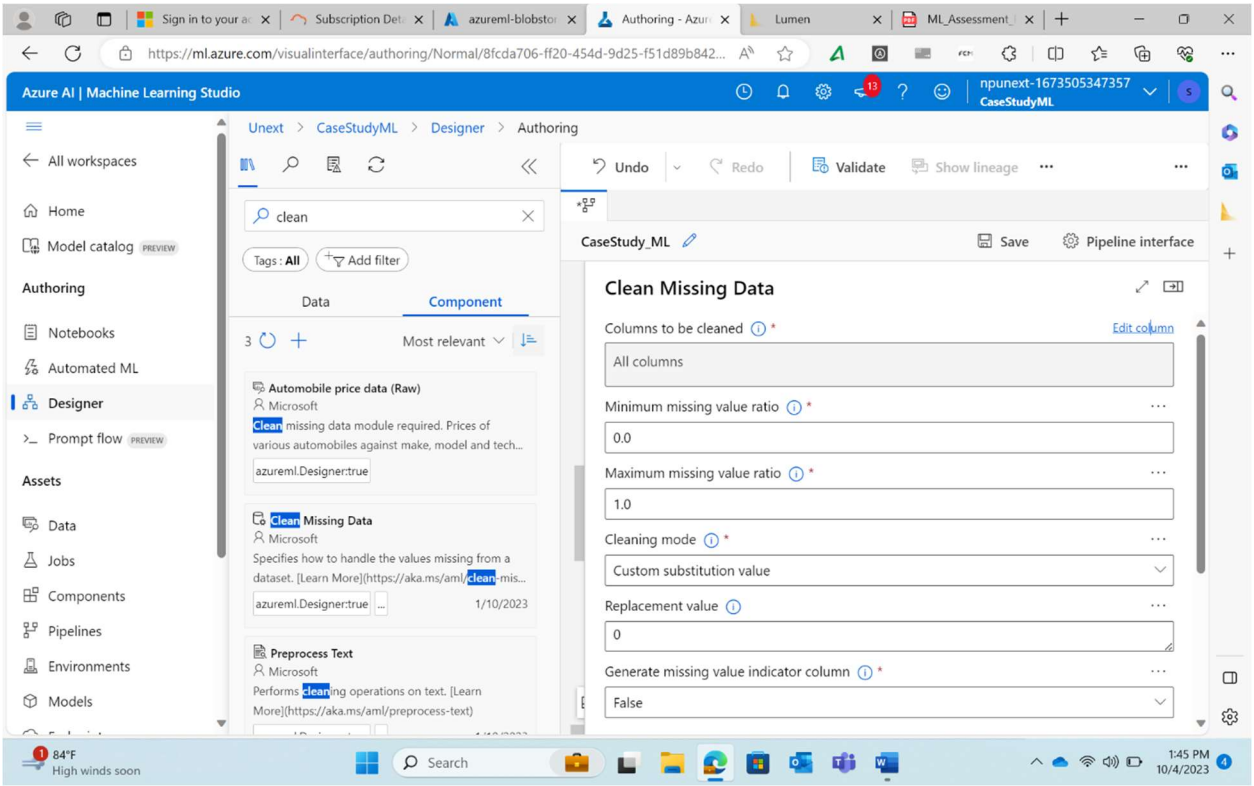
Model Development:

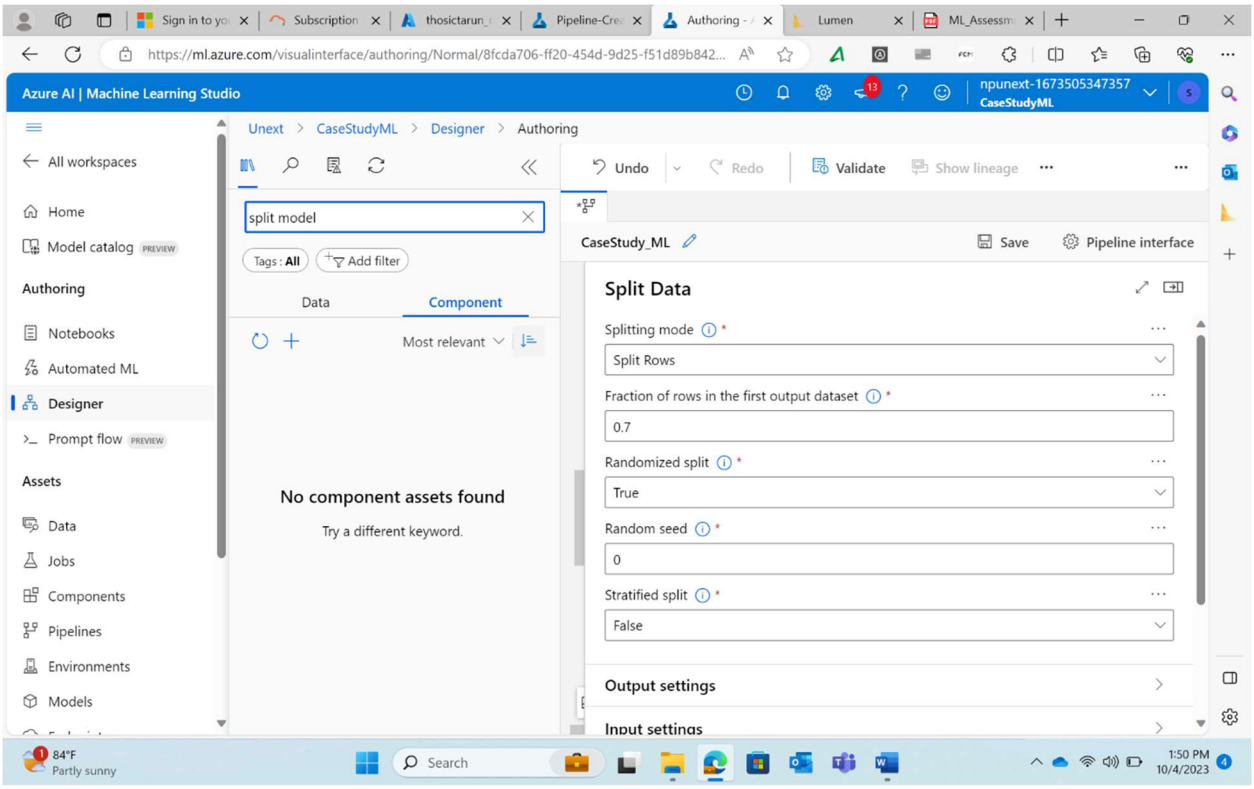
a) Choosing ML Algorithm(Linear Regression):



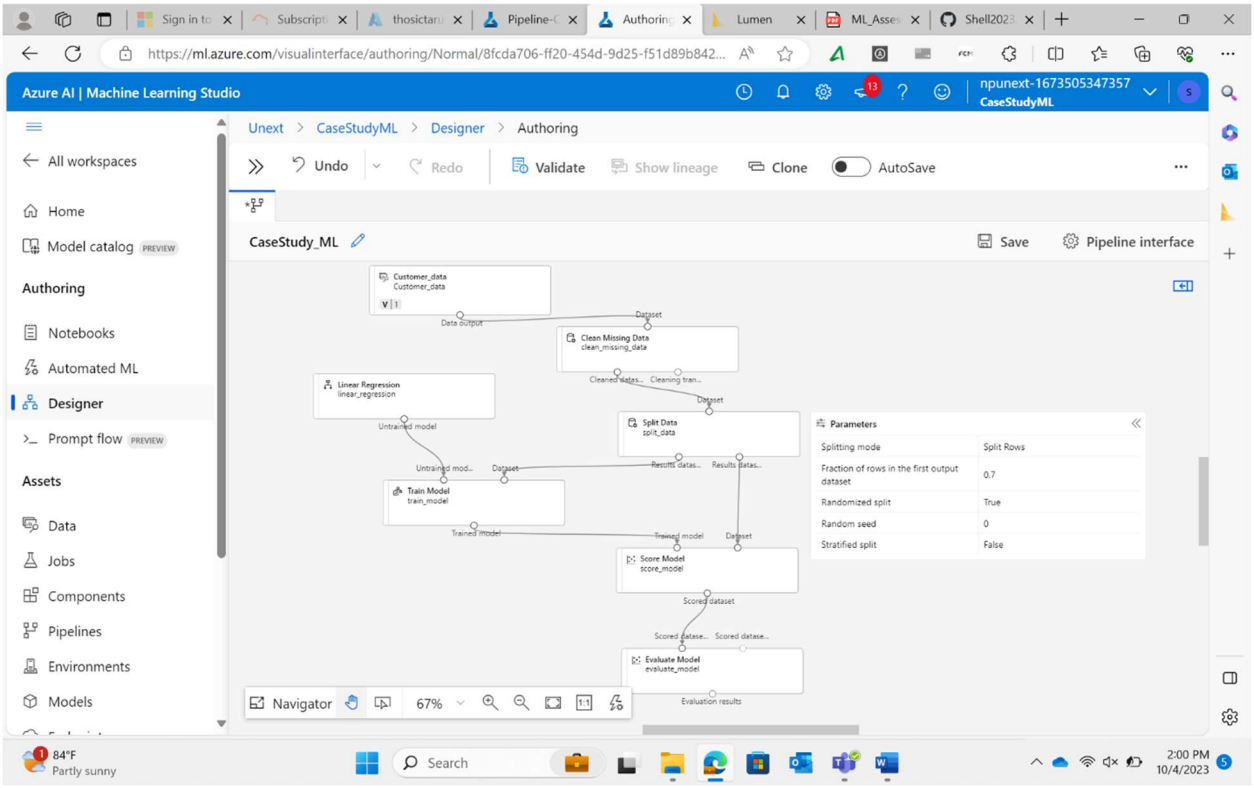
b) Splitting The dataset:



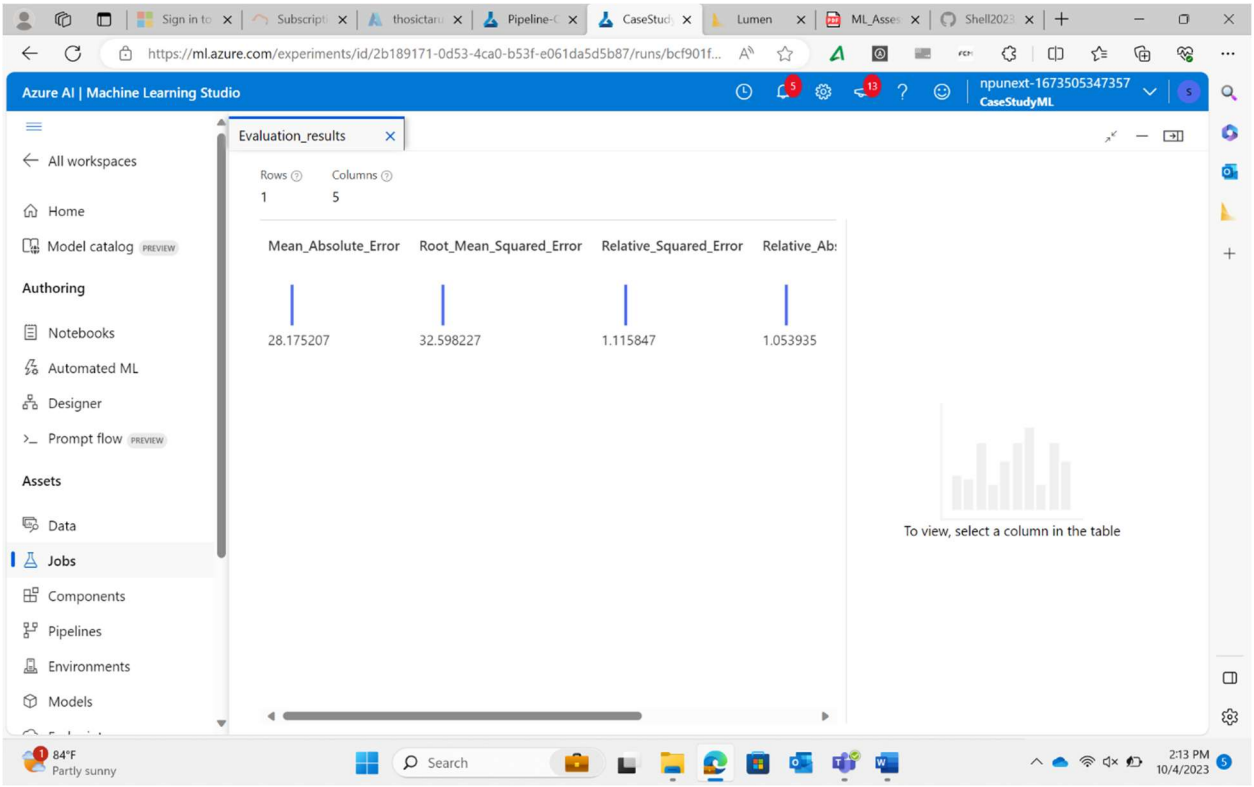


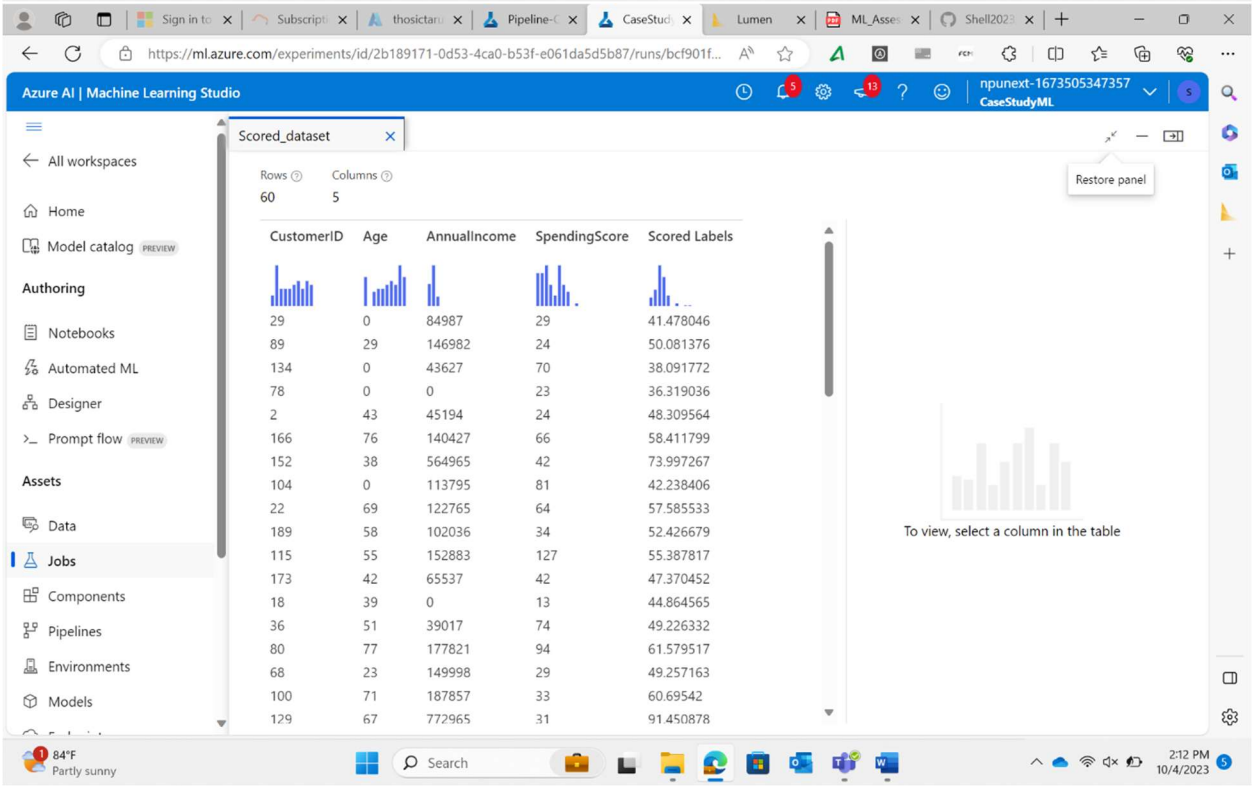


c) Training The model :

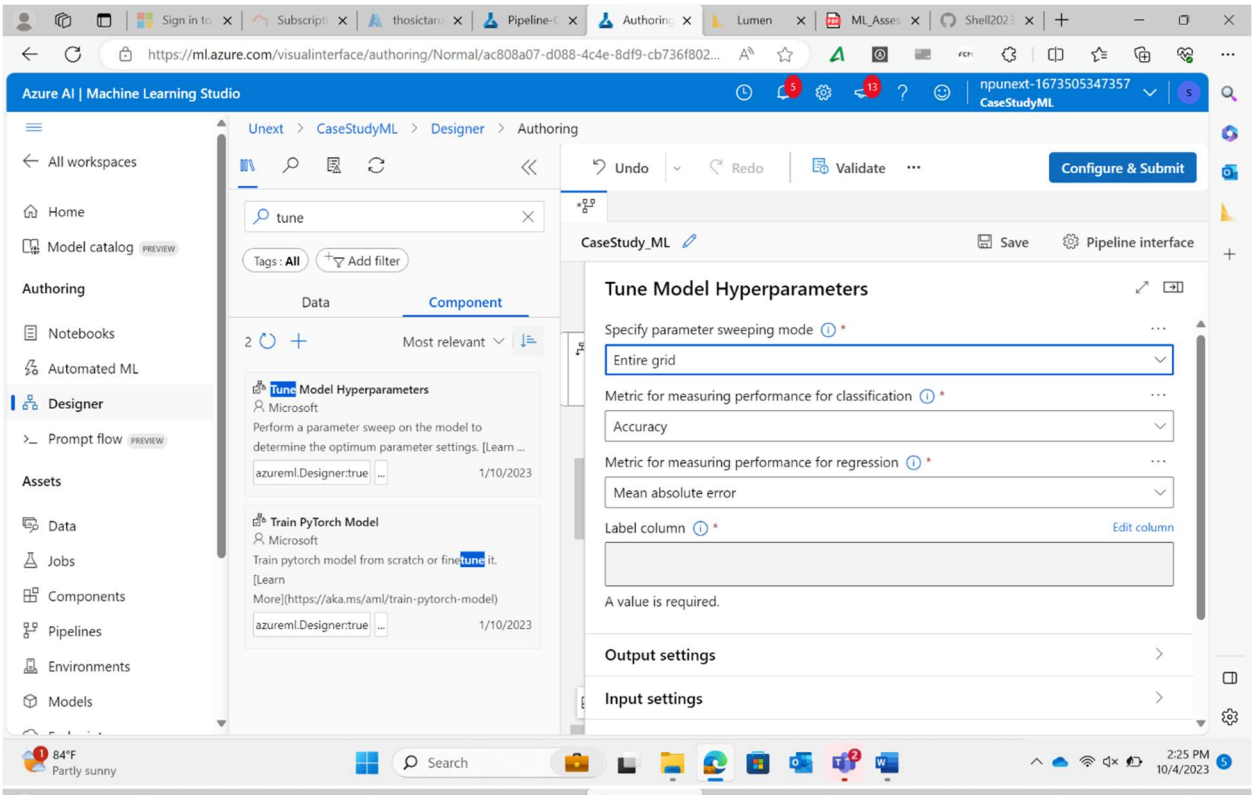


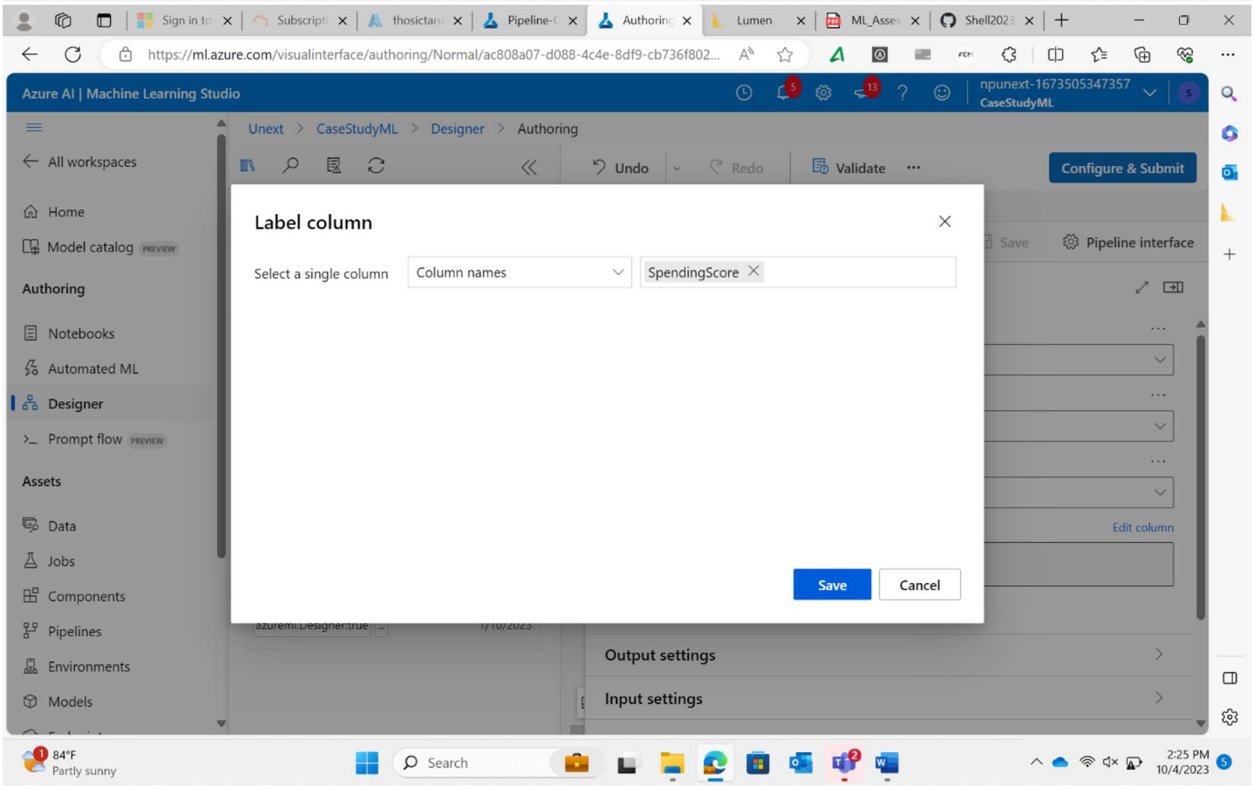
d)Evaluating Models Performance:

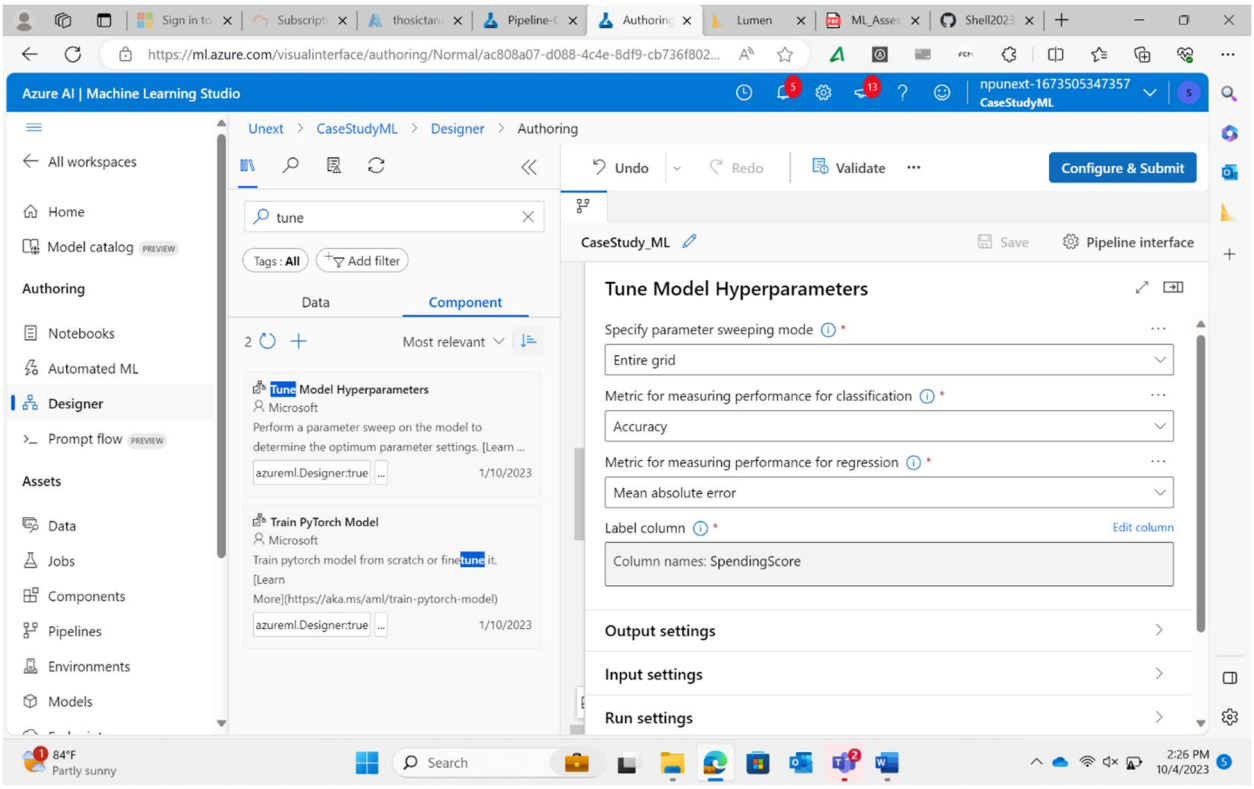


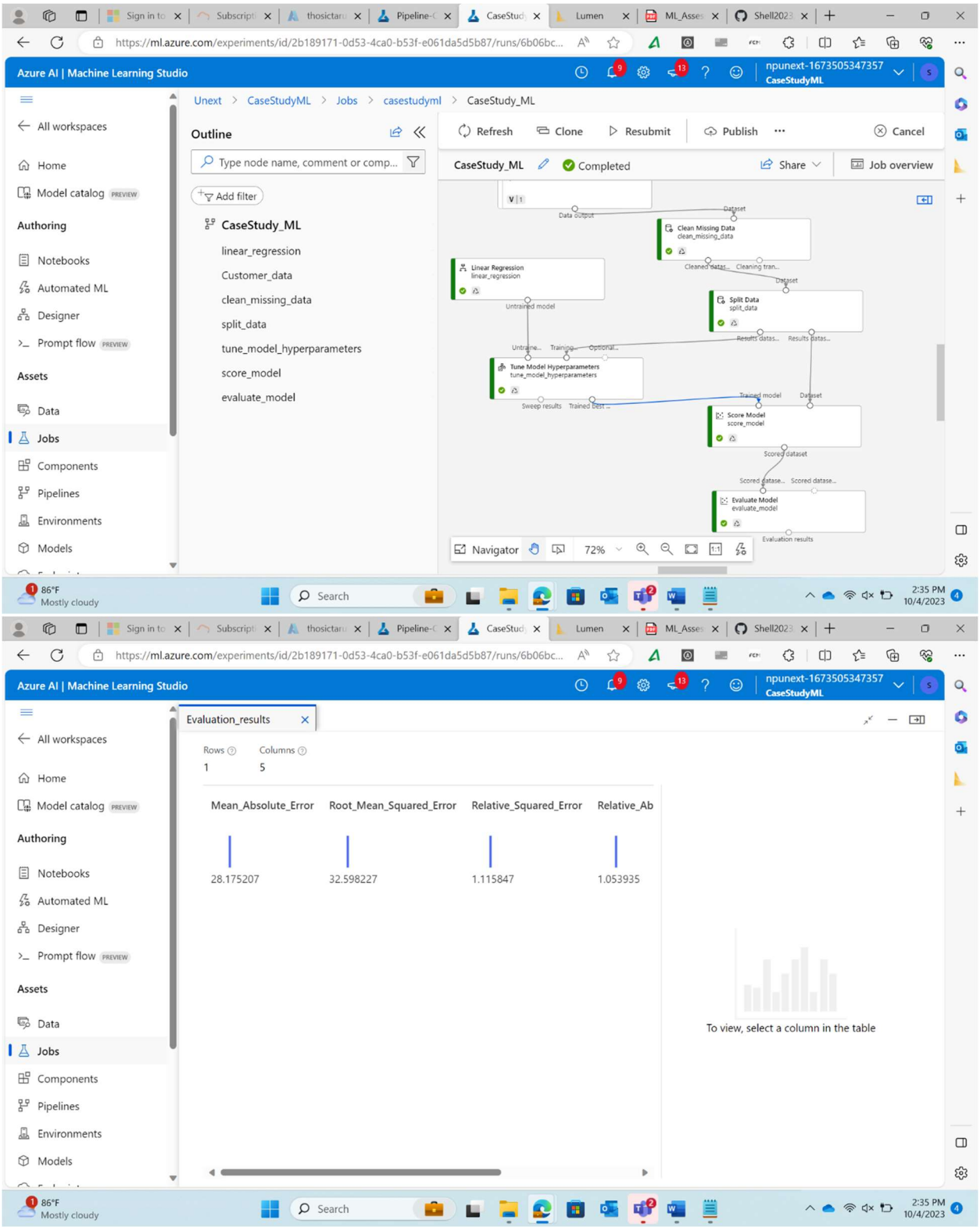


HyperParameter Tuning:


Azure AI | Machine Learning Studio
Unext > CaseStudyML > Designer > Authoring
All workspaces
Home
Model catalog
Authoring
Notebooks
Automated ML
Designer
Prompt flow
Assets
Data
Jobs
Components
Pipelines
Environments
Models
tune
Tags : All
Add filter
Data
Component
Most relevant
Tune Model Hyperparameters
Microsoft
Perform a parameter sweep on the model to determine the optimum parameter settings. [Learn ...
azureml.Designer:true
1/10/2023
Train PyTorch Model
Microsoft
Train pytorch model from scratch or fine tune it. [Learn More](https://aka.ms/aml/train-pytorch-model)
azureml.Designer:true
1/10/2023
Undo
Redo
Validate
Configure & Submit
CaseStudy_ML
Save
Pipeline interface
Tune Model Hyperparameters
Specify parameter sweeping mode *
Entire grid
Metric for measuring performance for classification *
Accuracy
Metric for measuring performance for regression *
Mean absolute error
Label column *
Edit column
A value is required.
Output settings
Input settings


Azure AI | Machine Learning Studio
Unext > CaseStudyML > Designer > Authoring
Label column
Select a single column
Column names
SpendingScore
Save
Cancel
Output settings
Input settings


Azure AI | Machine Learning Studio
Unext > CaseStudyML > Designer > Authoring
All workspaces
Home
Model catalog PREVIEW
Authoring
Notebooks
Automated ML
Designer
Prompt flow PREVIEW
Assets
Data
Jobs
Components
Pipelines
Environments
Models
tune
Tags : All
Add filter
Data
Component
Most relevant
Tune Model Hyperparameters
Microsoft
Perform a parameter sweep on the model to determine the optimum parameter settings. [Learn ...
azureml.Designer:true
1/10/2023
Train PyTorch Model
Microsoft
Train pytorch model from scratch or fine tune it. [Learn More](https://aka.ms/aml/train-pytorch-model)
azureml.Designer:true
1/10/2023
Undo
Redo
Validate
Configure & Submit
CaseStudy_ML
Save
Pipeline interface
Tune Model Hyperparameters
Specify parameter sweeping mode *
Entire grid
Metric for measuring performance for classification *
Accuracy
Metric for measuring performance for regression *
Mean absolute error
Label column *
Edit column
Column names: SpendingScore
Output settings
Input settings
Run settings
84°F
Partly sunny
Search
2:26 PM
10/4/2023



Answers for Assessment Questions:

1)Key steps involved in preparing the dataset for training are cleaning the dataset, by replacing null values, splitting the dataset into test and train data sets, and choosing the appropriate learning algorithm etc. and then score model and evaluate for evaluating the measures.

2) It is important to split the same dataset into test and train datasets

3) I choose Linear Regression for predicting the customer's purchasing behaviour.

4)Hyperparameter tuning is used while training for the optimized predictions , where the errors are minimal compared to train data component.