- Form a team 3-5 members

Each team has to perform the following

- Define a problem statement [10 marks]
- Build a tabular and a pdf dataset [10 marks]
- Use the following to work on your problem statement
    - Automated ML [10 marks]
- Use a LLM of your choice to perform search on private data or literature related to the dataset you have picked: [30 marks]

- Use Generative AI to generate test cases and comments for your notebook [10 marks]
-
- Design the architecture [30 marks], which should enforce fault tolerance and high availability

Once completed, create a github repository, upload relevant screenshots and notebooks in there.

Define a Problem Statement:

A real estate prediction system that uses available data and machine learning algorithms to predict housing prices based on features and location. Such systems can be incredibly valuable for property buyers to help them in the following use cases:

- Predict housing prices based on the selected features
- Correlation between Housing price and location attributes
- Get the overall picture of the USA housing prices w.r.t. locations
- How house attributes (bedroom, bathroom count) strongly correlate with the price? Are there any hidden patterns?

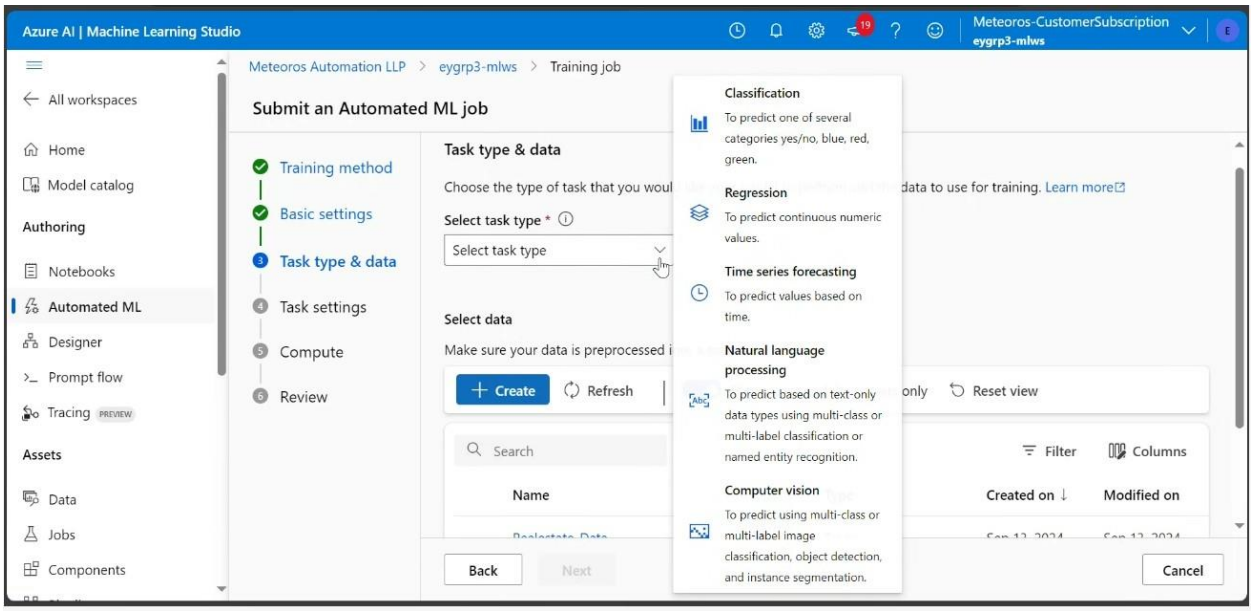
Build a tabular and a pdf dataset

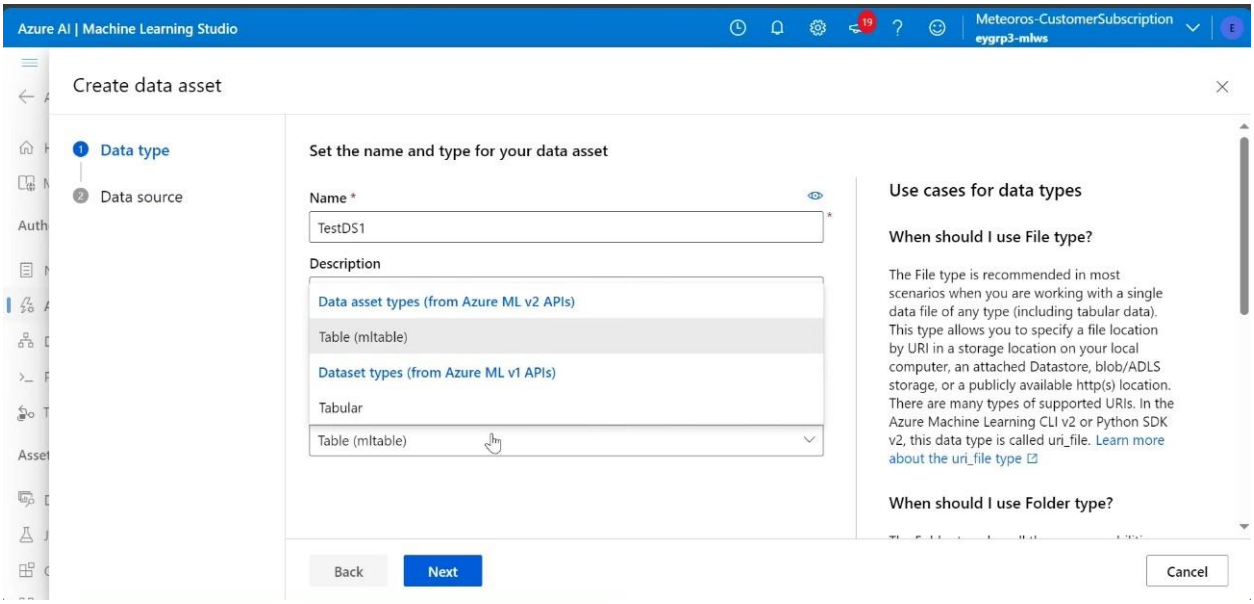
- Use the following to refer to the dataset: USA Real Estate Dataset (kaggle.com)

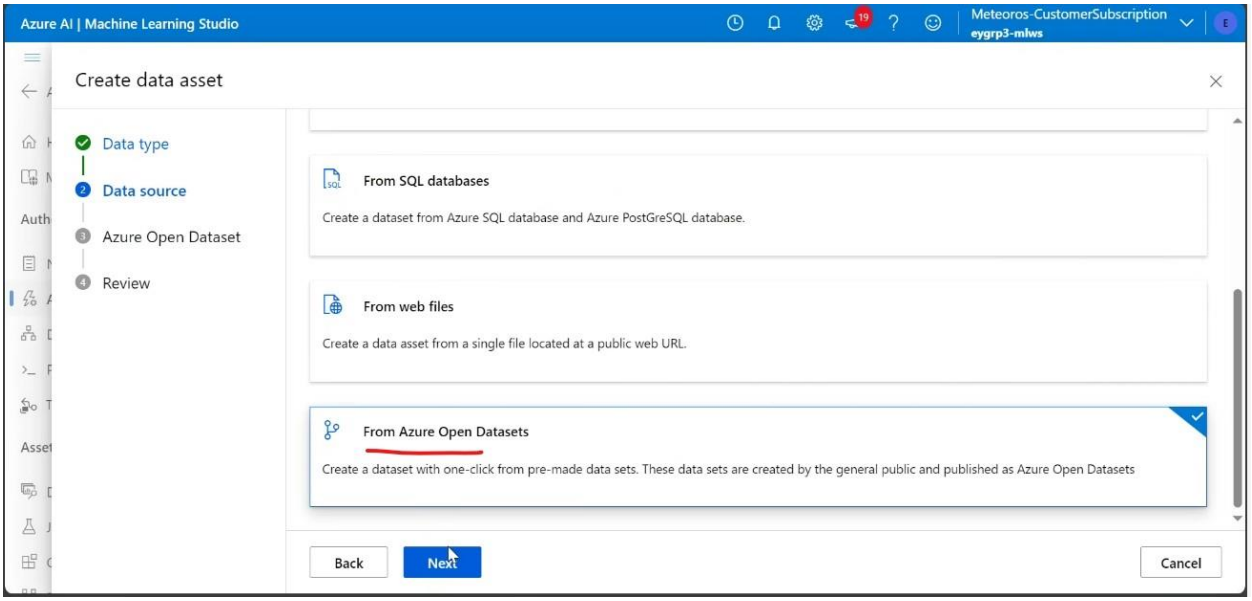
brokered_	status	price	bed	bath	acre_lot	street	city	state	zip_code	house_siz	prev_sold_date
103378	for_sale	105000	3	2	0.12	1962661	Adjuntas	Puerto Ric	00601	920	
52707	for_sale	80000	4	2	0.08	1902874	Adjuntas	Puerto Ric	00601	1527	
103379	for_sale	67000	2	1	0.15	1404990	Juana Dia	Puerto Ric	00795	748	
31239	for_sale	145000	4	2	0.1	1947675	Ponce	Puerto Ric	00731	1800	
34632	for_sale	65000	6	2	0.05	331151	Mayaguez	Puerto Ric	00680		
103378	for_sale	179000	4	3	0.46	1850806	San Sebas	Puerto Ric	00612	2520	
1205	for_sale	50000	3	1	0.2	1298094	Ciales	Puerto Ric	00639	2040	
50739	for_sale	71600	3	2	0.08	1048466	Ponce	Puerto Ric	00731	1050	
81909	for_sale	100000	2	1	0.09	734904	Ponce	Puerto Ric	00730	1092	
65672	for_sale	300000	5	3	7.46	1946226	Las Maria	Puerto Ric	00670	5403	
52707	for_sale	89000	3	2	13.39	1902814	Isabela	Puerto Ric	00662	1106	
52707	for_sale	150000	3	2	0.08	1773902	Juana Dia	Puerto Ric	00795	1045	
46019	for_sale	155000	3	2	0.1	1946165	Lares	Puerto Ric	00669	4161	
52707	for_sale	79000	5	2	0.12	1761024	Utuado	Puerto Ric	00641	1620	
88441	for_sale	649000	5	5	0.74	1879215	Ponce	Puerto Ric	00731	2677	
50739	for_sale	120000	3	2	0.08	17854	Yauco	Puerto Ric	00698	1100	
51202	for_sale	235000	4	4	0.22	13687	Mayaguez	Puerto Ric	00680	3450	
12876	for_sale	105000	3	2	0.08	1868721	Ponce	Puerto Ric	00728	1500	
109906	for_sale	575000	3	2	3.88	1312671	San Sebas	Puerto Ric	00685	4000	
46019	for_sale	140000	6	3	0.25	6710	Anasco	Puerto Ric	00610	1230	
52707	for_sale	50000	2	1	0.23	1902835	Yauco	Puerto Ric	00698	621	
52707	for_sale	165000	6	3	0.1	117231	Moca	Puerto Ric	00676	3000	
81909	for_sale	189000	3	1	2	1307740	Coamo	Puerto Ric	00769	1213	

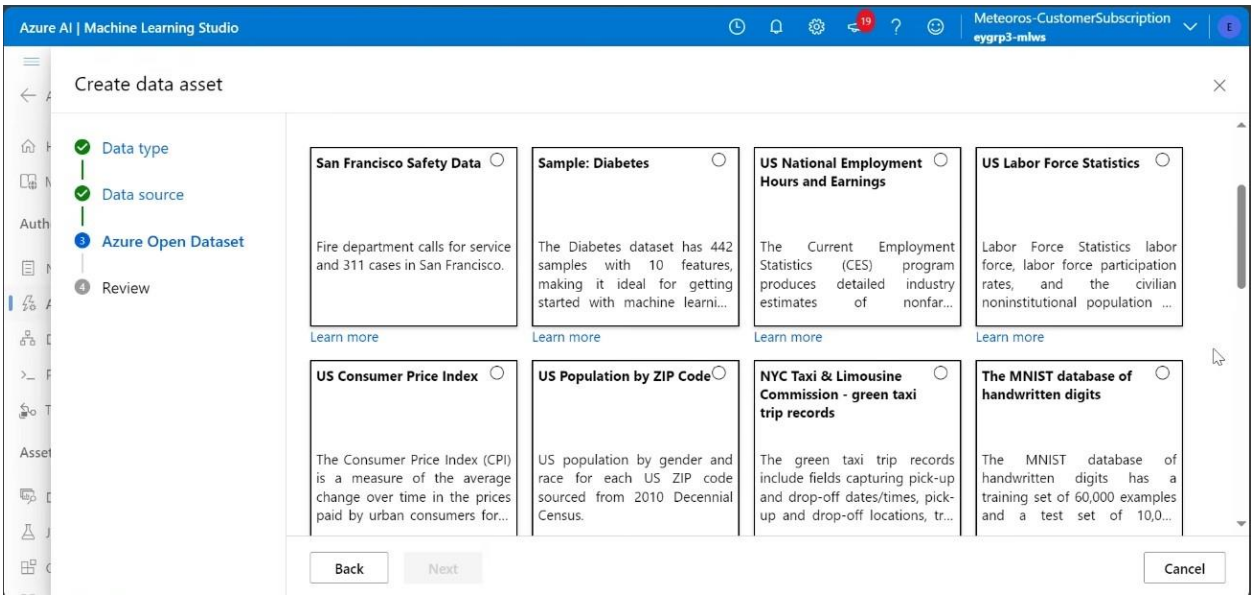
Use the following to work on your problem statement

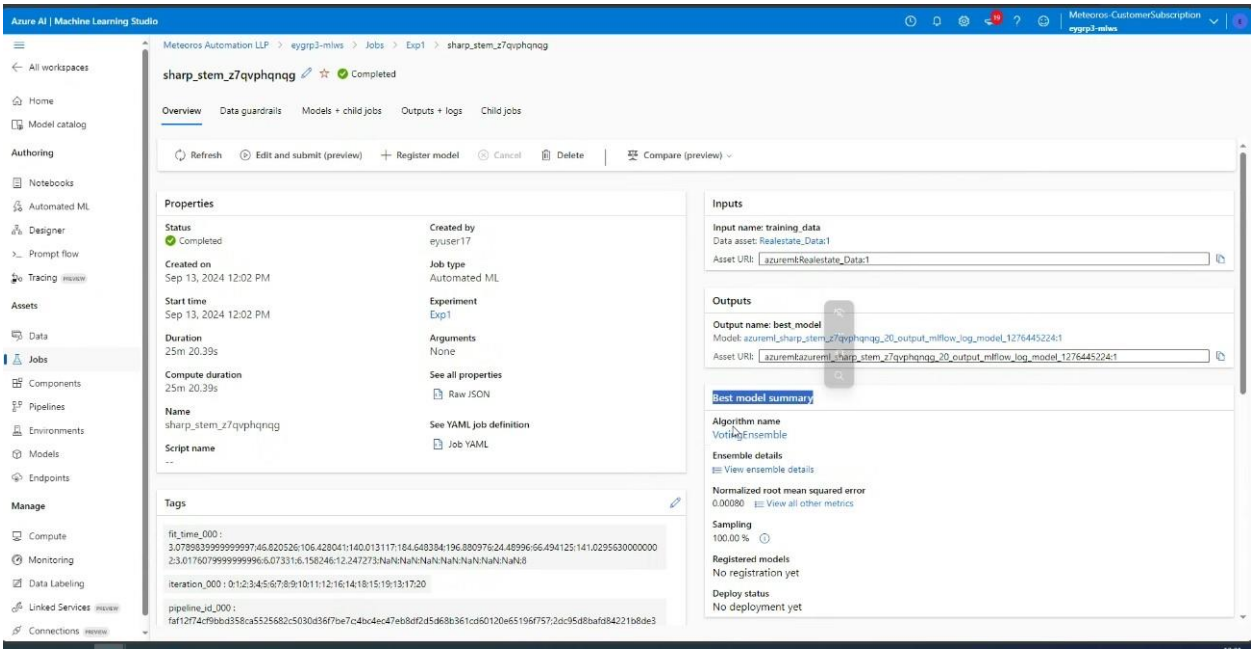
Below are the steps taken to build Automated ML Model: (Screenshots taken while creating the Automated ML Model)

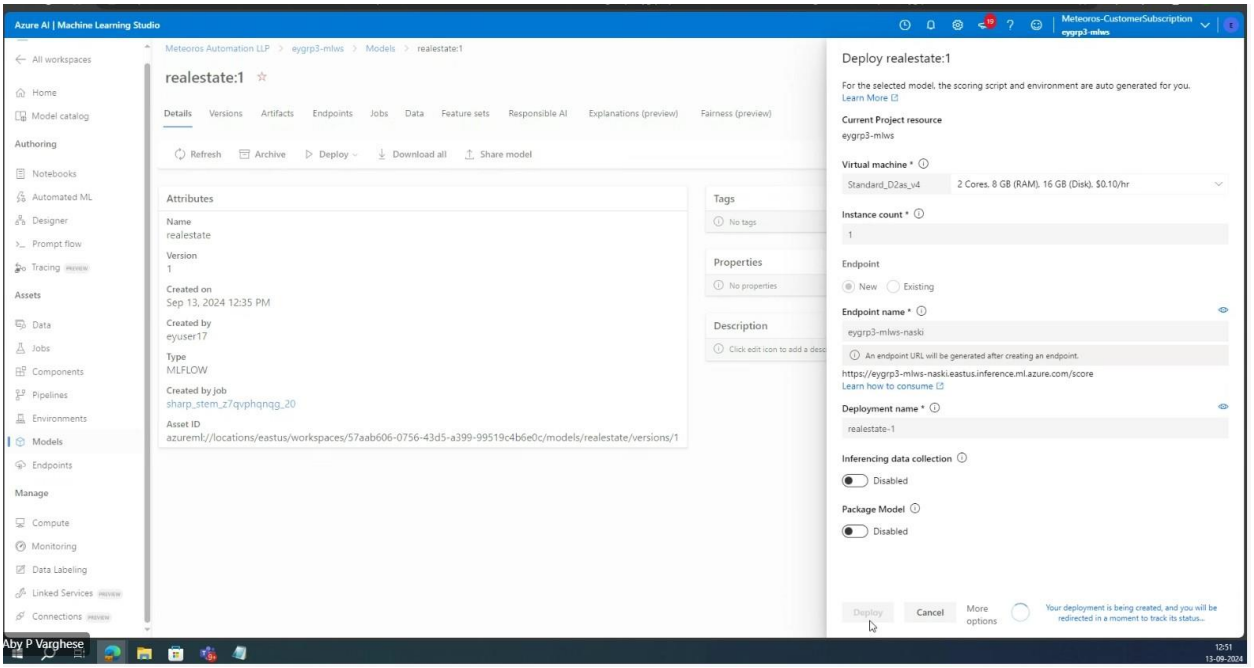


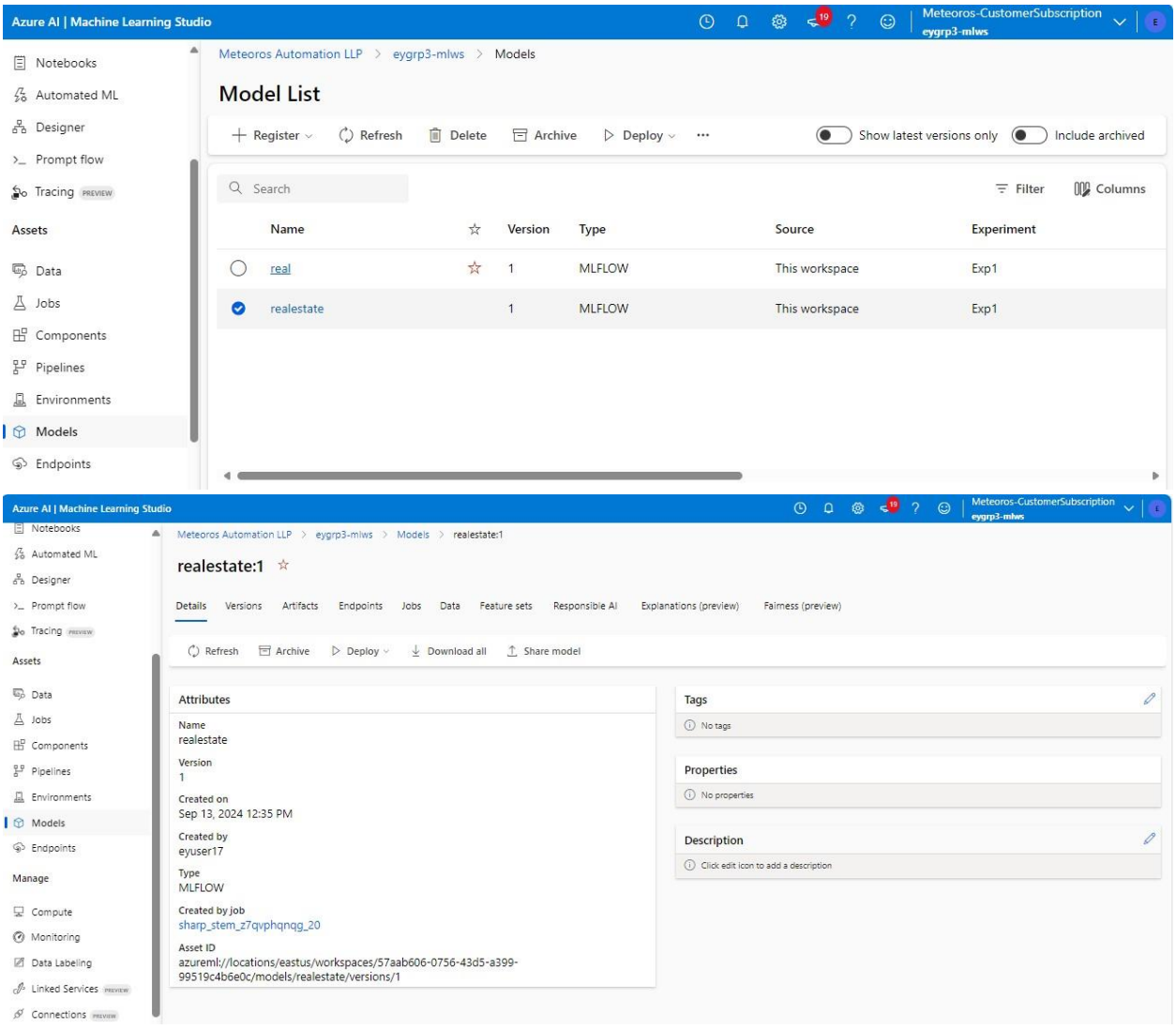



Azure AI | Machine Learning Studio
Meteoros-CustomerSubscription
eygrp3-mlws
Create data asset
Data type
Data source
Azure Open Dataset
Review
From SQL databases
Create a dataset from Azure SQL database and Azure PostGreSQL database.
From web files
Create a data asset from a single file located at a public web URL.
From Azure Open Datasets
Create a dataset with one-click from pre-made data sets. These data sets are created by the general public and published as Azure Open Datasets
Back
Next
Cancel

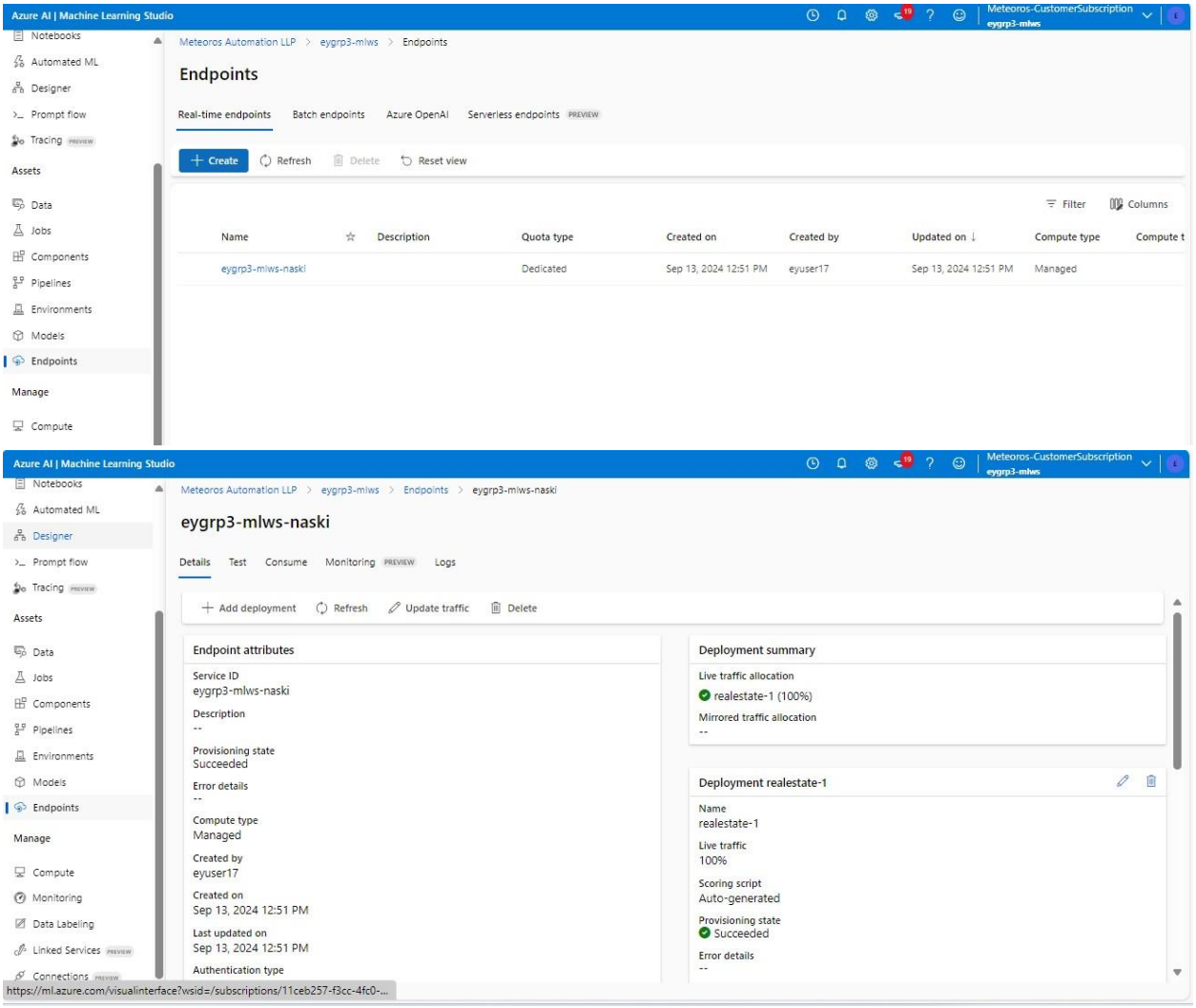

Azure AI | Machine Learning Studio
Meteoros-CustomerSubscription
eygrp3-mlws
Create data asset
Data type
Data source
Azure Open Dataset
Review
San Francisco Safety Data
Fire department calls for service and 311 cases in San Francisco.
Learn more
Sample: Diabetes
The Diabetes dataset has 442 samples with 10 features, making it ideal for getting started with machine learni...
Learn more
US National Employment Hours and Earnings
The Current Employment Statistics (CES) program produces detailed industry estimates of nonfar...
Learn more
US Labor Force Statistics
Labor Force Statistics labor force, labor force participation rates, and the civilian noninstitutional population ...
Learn more
US Consumer Price Index
The Consumer Price Index (CPI) is a measure of the average change over time in the prices paid by urban consumers for...
US Population by ZIP Code
US population by gender and race for each US ZIP code sourced from 2010 Decennial Census.
NYC Taxi & Limousine Commission - green taxi trip records
The green taxi trip records include fields capturing pick-up and drop-off dates/times, pick-up and drop-off locations, tr...
The MNIST database of handwritten digits
The MNIST database of handwritten digits has a training set of 60,000 examples and a test set of 10,0...
Back
Next
Cancel


Azure AI | Machine Learning Studio
Meteoros-CustomerSubscription
eygrp3-mlws
Meteoros Automation LLP > eygrp3-mlws > Jobs > Exp1 > sharp_stem_z7qvphqnqg
All workspaces
Home
Model catalog
Authoring
Notebooks
Automated ML
Designer
Prompt flow
Tracing
Assets
Data
Jobs
Components
Pipelines
Environments
Models
Endpoints
Manage
Compute
Monitoring
Data Labeling
Linked Services
Connections
sharp_stem_z7qvphqnqg
Completed
Overview
Data guardrails
Models + child jobs
Outputs + logs
Child jobs
Refresh
Edit and submit (preview)
Register model
Cancel
Delete
Compare (preview)
Properties
Status
Completed
Created by
eyuser17
Created on
Sep 13, 2024 12:02 PM
Job type
Automated ML
Start time
Sep 13, 2024 12:02 PM
Experiment
Exp1
Duration
25m 20.39s
Arguments
None
Compute duration
25m 20.39s
See all properties
Raw JSON
Name
sharp_stem_z7qvphqnqg
See YAML job definition
Job YAML
Script name
--
Tags
fit_time_000 :
3.07898399999999997;46.820526;106.428041;140.013117;184.648384;196.880976;24.48996;66.494125;141.0295630000002;3.0176079999999996;6.07331;6.158246;12.247273;NaN;NaN;NaN;NaN;NaN;NaN;NaN;8
iteration_000 : 0;1;2;3;4;5;6;7;8;9;10;11;12;16;14;18;15;19;13;17;20
pipeline_id_000 :
faf12f74cf9bbd358ca5525682c5030d36f7be7c4bc4ec47eb8df2d5d68b361cd60120e65196f757;2dc95d8bafd84221b8de3
Inputs
Input name: training_data
Data asset: Realestate_Data:1
Asset URI: azureml:Realestate_Data:1
Outputs
Output name: best_model
Model: azureml_sharp_stem_z7qvphqnqg_20_output_mlflow_log_model_1276445224:1
Asset URI: azureml:azureml_sharp_stem_z7qvphqnqg_20_output_mlflow_log_model_1276445224:1
Best model summary
Algorithm name
VotingEnsemble
Ensemble details
View ensemble details
Normalized root mean squared error
0.00080
View all other metrics
Sampling
100.00 %
Registered models
No registration yet
Deploy status
No deployment yet


Azure AI | Machine Learning Studio
Meteoros-CustomerSubscription
eygrp3-mlws
Meteoros Automation LLP > eygrp3-mlws > Models > realestate:1
realestate:1
Details
Versions
Artifacts
Endpoints
Jobs
Data
Feature sets
Responsible AI
Explanations (preview)
Fairness (preview)
Refresh
Archive
Deploy
Download all
Share model
Attributes
Name
realestate
Version
1
Created on
Sep 13, 2024 12:35 PM
Created by
eyuser17
Type
MLFLOW
Created by job
sharp_stem_z7qvphqnqg_20
Asset ID
azureml://locations/eastus/workspaces/57aab606-0756-43d5-a399-99519c4b6e0c/models/realestate/versions/1
Tags
No tags
Properties
No properties
Description
Deploy realestate:1
For the selected model, the scoring script and environment are auto generated for you.
Learn More
Current Project resource
eygrp3-mlws
Virtual machine *
Standard_D2as_v4
2 Cores, 8 GB (RAM), 16 GB (Disk), $0.10/hr
Instance count *
1
Endpoint
New
Existing
Endpoint name *
eygrp3-mlws-naski
An endpoint URL will be generated after creating an endpoint.
https://eygrp3-mlws-naski.eastus.inference.ml.azure.com/score
Learn how to consume
Deployment name *
realestate-1
Inferencing data collection
Disabled
Package Model
Disabled
Deploy
Cancel
More options
Your deployment is being created, and you will be redirected in a moment to track its status...
Aby P Varghese
12:51
13-09-2024

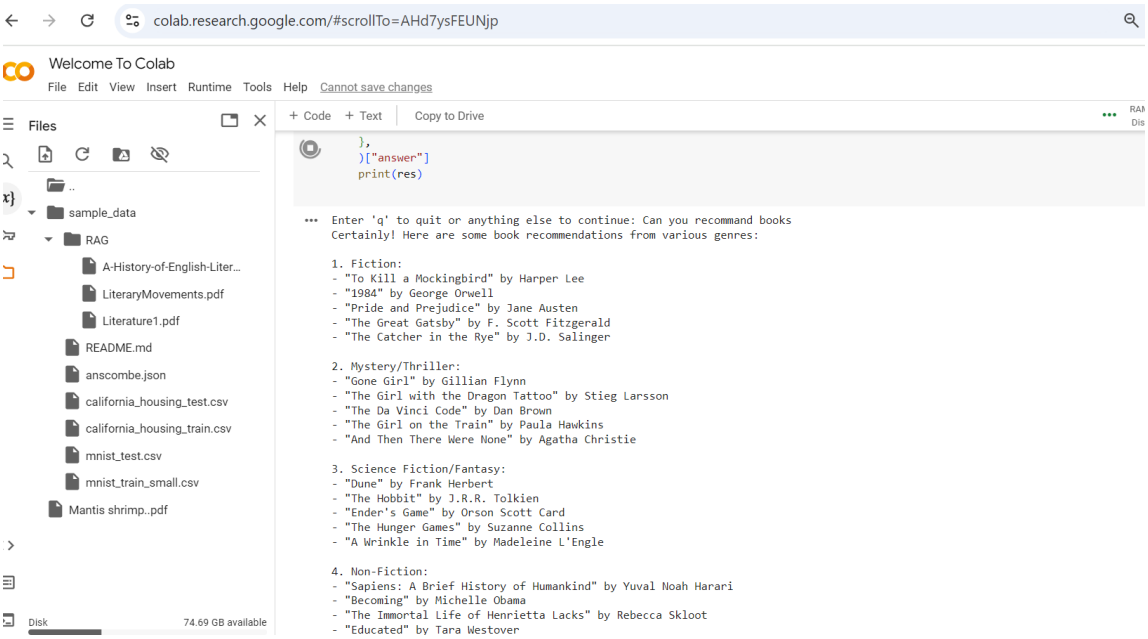

Azure AI | Machine Learning Studio
Meteoros-CustomerSubscription
eygrp3-mlws
Notebooks
Automated ML
Designer
Prompt flow
Tracing PREVIEW
Assets
Data
Jobs
Components
Pipelines
Environments
Models
Endpoints
Meteoros Automation LLP > eygrp3-mlws > Models
Model List
Register
Refresh
Delete
Archive
Deploy
Show latest versions only
Include archived
Search
Filter
Columns
Name
Version
Type
Source
Experiment
real
1
MLFLOW
This workspace
Exp1
realestate
1
MLFLOW
This workspace
Exp1

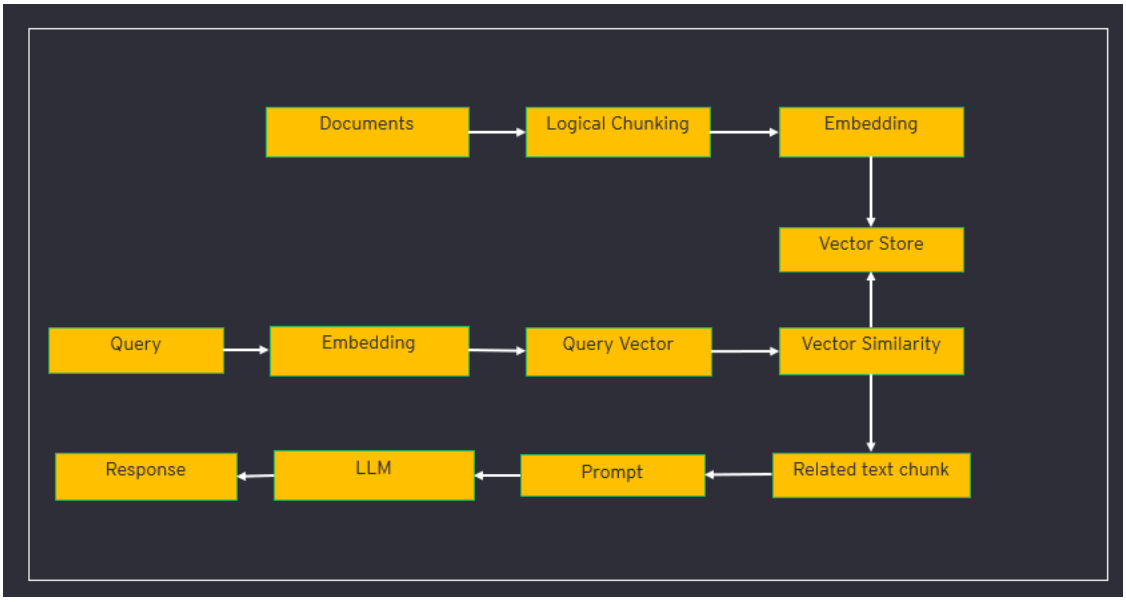
Azure AI | Machine Learning Studio
Meteoros-CustomerSubscription
eygrp3-mlws
Meteoros Automation LLP > eygrp3-mlws > Models > realestate:1
realestate:1
Details
Versions
Artifacts
Endpoints
Jobs
Data
Feature sets
Responsible AI
Explanations (preview)
Fairness (preview)
Refresh
Archive
Deploy
Download all
Share model
Attributes
Name
realestate
Version
1
Created on
Sep 13, 2024 12:35 PM
Created by
eyuser17
Type
MLFLOW
Created by job
sharp_stem_z7qvphqnqg_20
Asset ID
azureml://locations/eastus/workspaces/57aab606-0756-43d5-a399-99519c4b6e0c/models/realestate/versions/1
Tags
No tags
Properties
No properties
Description
Click edit icon to add a description



Use a LLM of your choice to perform search on private data or literature related to the dataset you have picked

We have created Literature related system using the private data on Literature related information. This system gives any questions and answers related to Literature related information. The code base LangChain pipeline to do the job.


colab.research.google.com/#scrollTo=AHd7ysFEUNjp
Welcome To Colab
File Edit View Insert Runtime Tools Help Cannot save changes
Files
+ Code + Text Copy to Drive
},
)["answer"]
print(res)
Enter 'q' to quit or anything else to continue: Can you recommand books
Certainly! Here are some book recommendations from various genres:
1. Fiction:
- "To Kill a Mockingbird" by Harper Lee
- "1984" by George Orwell
- "Pride and Prejudice" by Jane Austen
- "The Great Gatsby" by F. Scott Fitzgerald
- "The Catcher in the Rye" by J.D. Salinger
2. Mystery/Thriller:
- "Gone Girl" by Gillian Flynn
- "The Girl with the Dragon Tattoo" by Stieg Larsson
- "The Da Vinci Code" by Dan Brown
- "The Girl on the Train" by Paula Hawkins
- "And Then There Were None" by Agatha Christie
3. Science Fiction/Fantasy:
- "Dune" by Frank Herbert
- "The Hobbit" by J.R.R. Tolkien
- "Ender's Game" by Orson Scott Card
- "The Hunger Games" by Suzanne Collins
- "A Wrinkle in Time" by Madeleine L'Engle
4. Non-Fiction:
- "Sapiens: A Brief History of Humankind" by Yuval Noah Harari
- "Becoming" by Michelle Obama
- "The Immortal Life of Henrietta Lacks" by Rebecca Skloot
- "Educated" by Tara Westover
sample_data
RAG
A-History-of-English-Liter...
LiteraryMovements.pdf
Literature1.pdf
README.md
anscombe.json
california_housing_test.csv
california_housing_train.csv
mnist_test.csv
mnist_train_small.csv
Mantis shrimp..pdf
Disk 74.69 GB available


Documents
Logical Chunking
Embedding
Vector Store
Query
Embedding
Query Vector
Vector Similarity
Response
LLM
Prompt
Related text chunk

Design the architecture [30 marks], which should enforce fault tolerance and high availability

The following is the proposed Scalable, Highly Available, Fault Tolerance Infrastructure Architecture Diagram for the problem: