**Foundations of Operationalizing AI**

**Individual Assignment 1**

**Talia Qaiser**

**Tqaiser**

**Step 1: Data Preparation**

**Data Cleaning**

Downloaded the data from the Zillow Dataset, it has columns like price, area, number of bedrooms and bathrooms etc. of houses.

I started with getting the information of the data and describing it. Then I got the different data types in the dataset.

There were no missing values in the dataset so it led to no further processing.

**Exploratory Data Analysis**

By using a simple pairplot I plotted the relationship between all the different variables. The relationship between price and area looked interesting so I dived deeper into it to see the distribution in both using histogram. The data was skewed a little to the right.

A correlation matrix was made between first the numeric variables only that showed a high correlation between price and area and price and number of bathrooms.

The categorical variables were then encoded and then a correlation matrix depicting all variables was made. This showed a highly negative correlation between a furnished house and an unfurnished house, rightfully so.

**Feature Engineering**

The numeric columns were standardized to be used in the machine learning models.

**Step 2: Model Building**

First, the MLFlow experiment was set by the name “MLFlow Housing Prediction” and auto log. was turned on.

Then the parameters for a Decision Tree model were set, the model trained and predicted and the Mean Absolute Error and Mean Squared error were calculated as the loss metrics.

The Decision Tree model was then run in MLFlow.

**Step 3: Model Evaluation**

I used mean absolute error as the evaluation metric.

The prices were then predicted and compared with the actual prices, the results were not that off from the actual prices.

MAE: 1035644.8623853211

MSE: 1819055990422.9358

**Step 4: Model Comparison**

For Model Comparison, I chose the following two models:

1. From the Kaggle competition: https://www.kaggle.com/code/krparekh24/zillow-home-price-prediction

2. From Github: https://github.com/Viveckh/LilHomie/blob/master/PricePredictionModel-RandomForest.ipynb

**Model 1: CatBoostRegressor**

Used the following parameters for the CatBoostRegressor model:

"iterations": 630, "learning\_rate": 0.03, "depth": 6, "l2\_leaf\_reg": 3, "random\_seed": seed

And used Mean Absolute Error as the evaluation metric:

MAE: 631205.8437346438

Ran the model using MLFlow.

**Model 2: Random Forest Regressor**

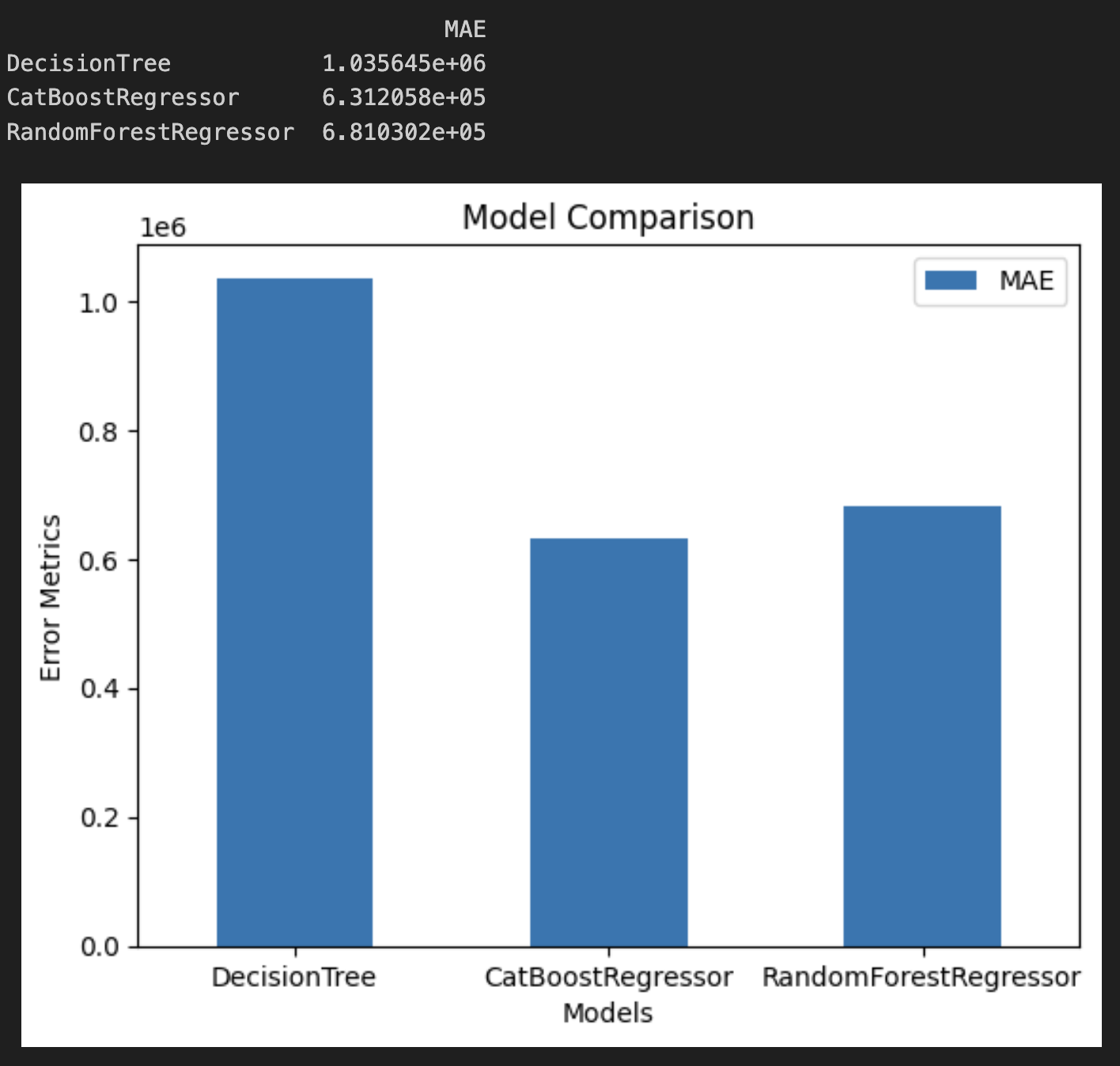
Used the following parameters for the Random Forest Regressor:

"n\_estimators" : 100, "n\_jobs" : -1

And used Mean Absolute Error as the evaluation metric:

MAE: 681030.2354740062

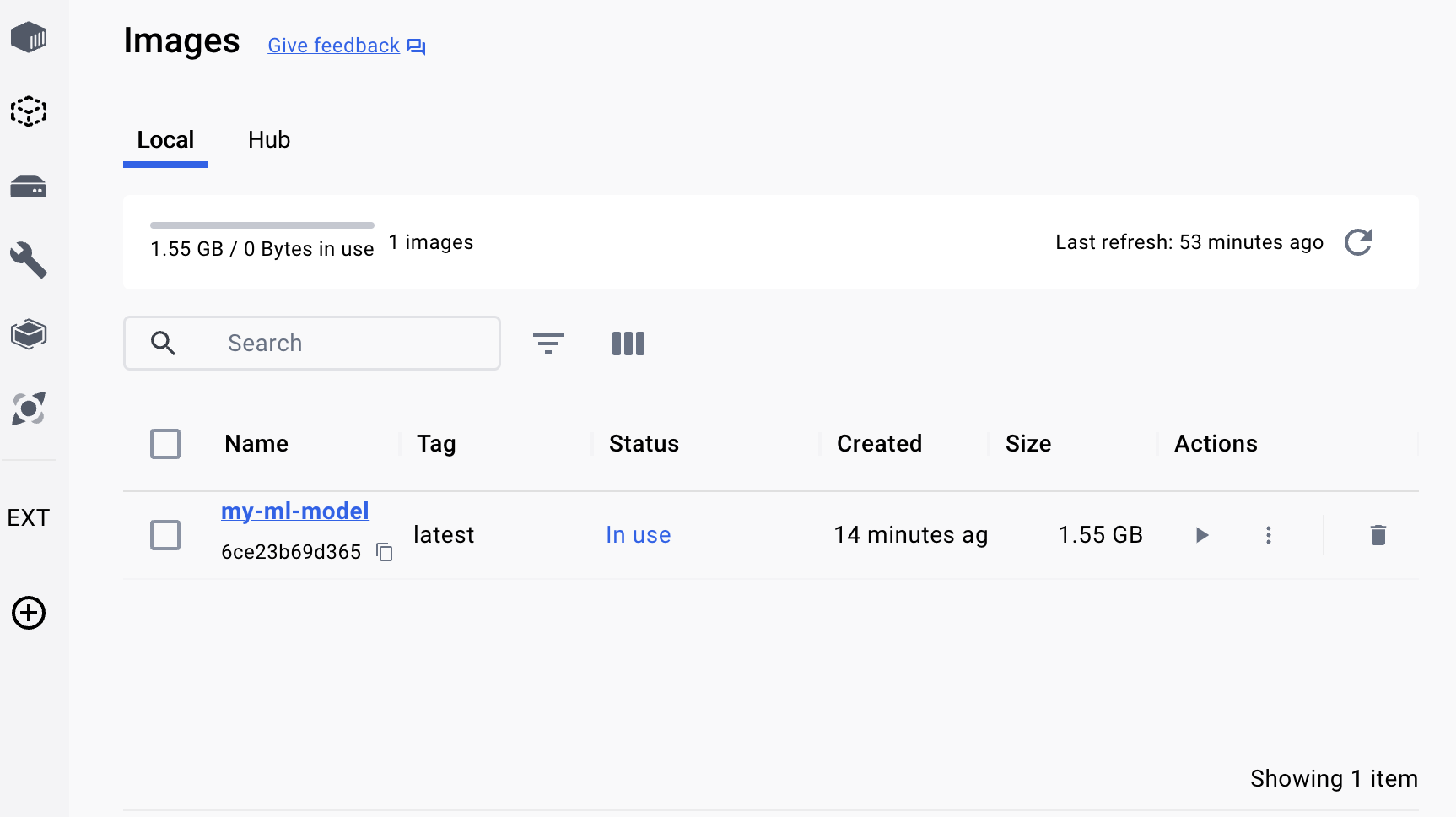
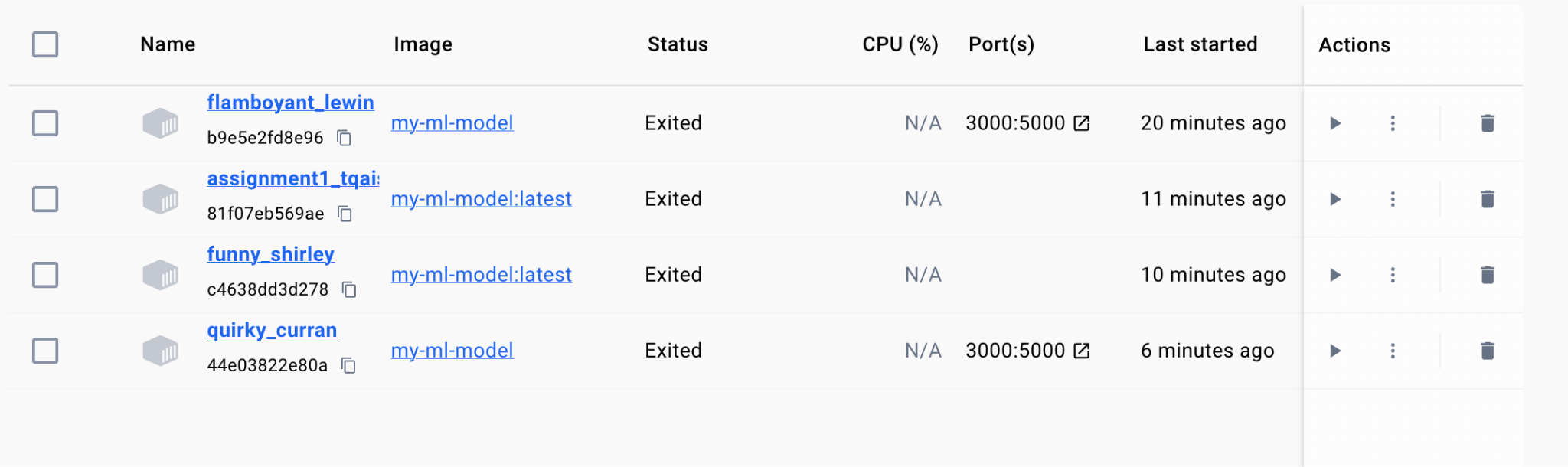
Ran the model using MLFlow.

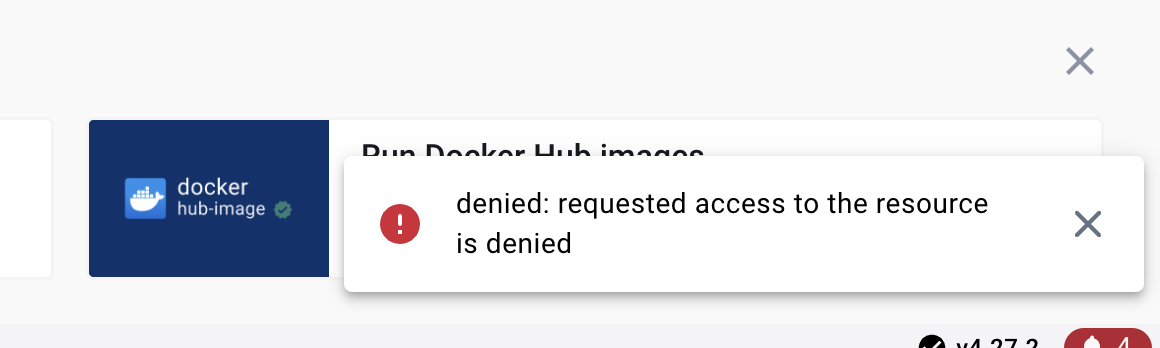


The CatBoostRegressor has the least mean absolute error. And that is the model that was in the Kaggle competition.

It has very specific parameters and runs really well for prediction models.

**Docker Hub:**



I tried running the docker. It worked locally but it did not get pushed to Docker hub due to some permission issues. I am unsure what the error is in that. 

**MLFlow Registry:**

<http://127.0.0.1:5000/#/experiments/128939817006926801?searchFilter=&orderByKey=attributes.start_time&orderByAsc=false&startTime=ALL&lifecycleFilter=Active&datasetsFilter=W10%3D&modelVersionFilter=All%20Runs&selectedColumns=attributes.%60Source%60,attributes.%60Models%60,attributes.%60Dataset%60&compareRunCharts=>

