MLflow Model Tracking Report

1. Introduction

This report outlines the experiments conducted to compare the performance of Linear Regression and Random Forest models on the California housing dataset. The models were trained, logged, and evaluated using MLflow, a popular platform for managing the machine learning lifecycle.

2. Dataset

The dataset used for this experiment is the **California housing dataset**, which includes various features related to housing prices. The target variable to predict is the median house value.

3. Model Comparison

Mean Squared Error (MSE) Values

The following table summarizes the Mean Squared Error (MSE) values obtained for each model:

Model	MSE
Linear Regression	0.5559
Random Forest	0.2554

Analysis

- **Linear Regression**: The MSE of the Linear Regression model is **0.5559**. This indicates a relatively higher error in predictions, suggesting that this model does not capture the underlying patterns in the data effectively.
- **Random Forest**: The MSE of the Random Forest model is **0.2554**, significantly lower than that of Linear Regression. This indicates that the Random Forest model provides more accurate predictions by better capturing the complexities in the data.

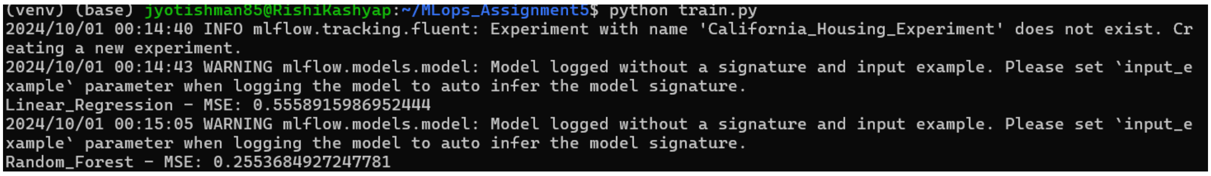
Conclusion

Based on the MSE values, the **Random Forest** model is identified as the better-performing model for this task, demonstrating superior accuracy in predicting housing prices compared to the Linear Regression model.

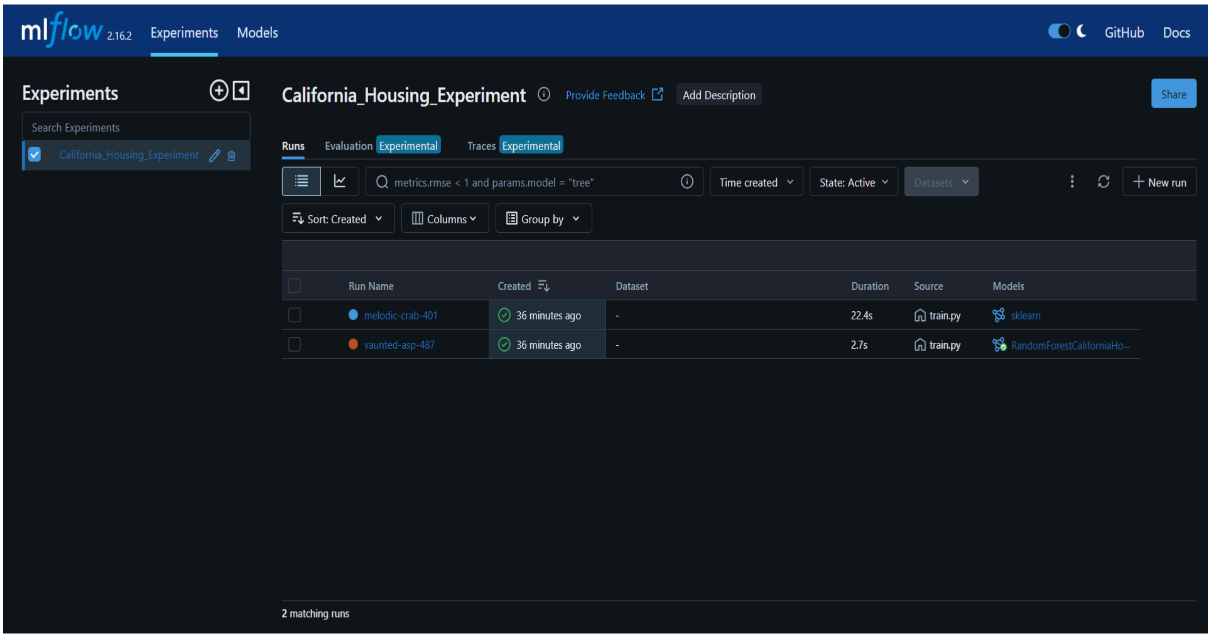
4. MLflow UI Screenshots

The MLflow UI was used to track the experiments, log metrics, and visualize model performance. Below are the relevant screenshots from the MLflow UI:

Experiment Overview:



```
(venv) (base) jyotishman85@RishiKashyap:~/MLops_Assignment5$ python train.py
2024/10/01 00:14:40 INFO mlflow.tracking.fluent: Experiment with name 'California_Housing_Experiment' does not exist. Cr
eating a new experiment.
2024/10/01 00:14:43 WARNING mlflow.models.model: Model logged without a signature and input example. Please set `input_e
xample` parameter when logging the model to auto infer the model signature.
Linear_Regression - MSE: 0.5558915986952444
2024/10/01 00:15:05 WARNING mlflow.models.model: Model logged without a signature and input example. Please set `input_e
xample` parameter when logging the model to auto infer the model signature.
Random_Forest - MSE: 0.2553684927247781
```

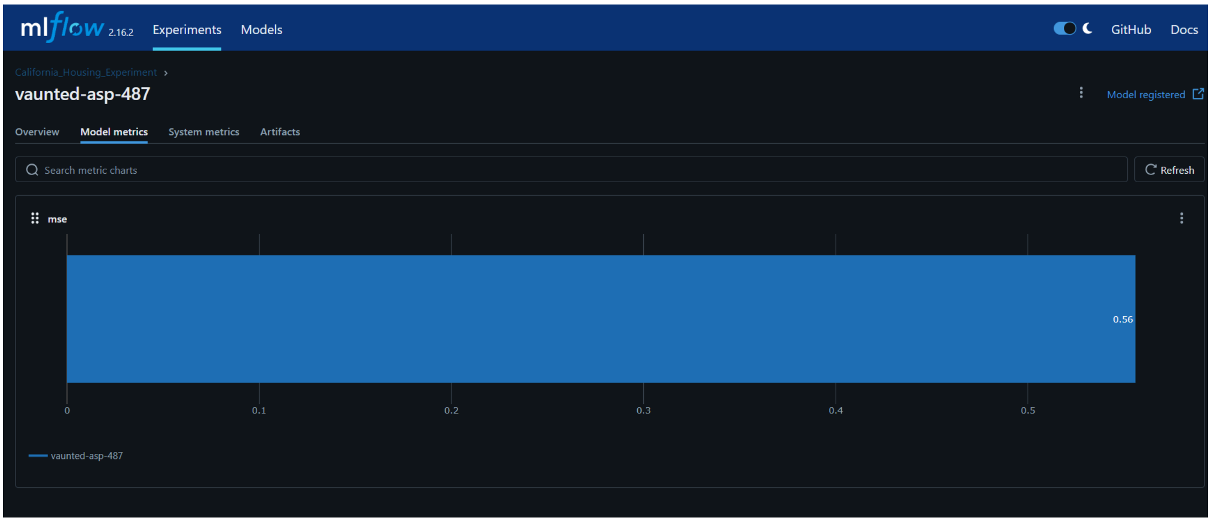


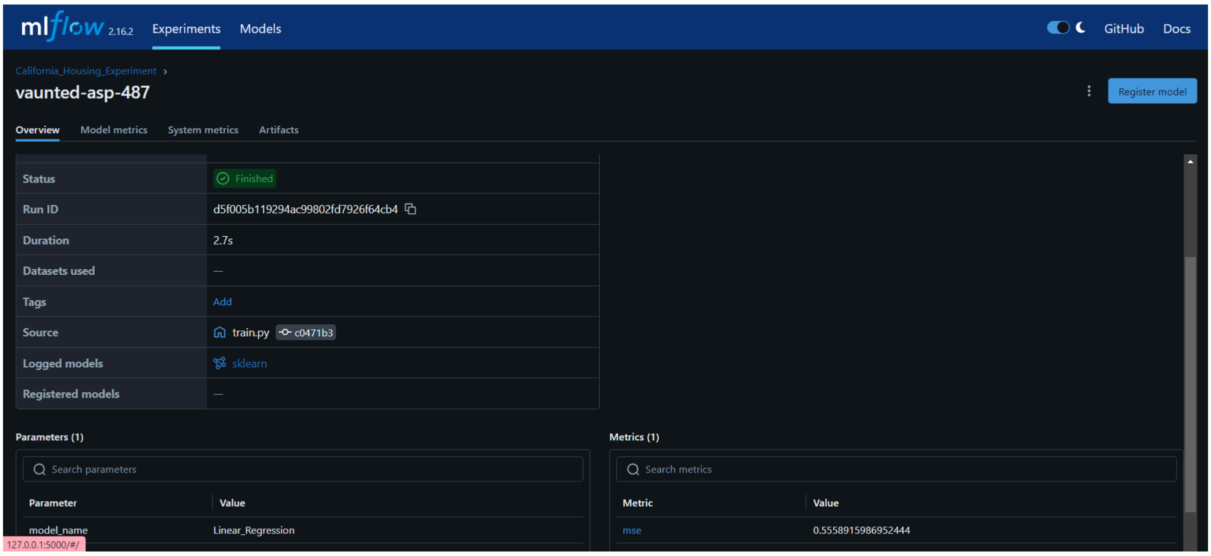
California_Housing_Experiment

☐	Run Name	Created	Dataset	Duration	Source	Models
☐	● melodic-crab-401	✓ 36 minutes ago	-	22.4s	train.py	sklearn
☐	● vaunted-asp-487	✓ 36 minutes ago	-	2.7s	train.py	RandomForestCaliforniaHo...

2 matching runs

Linear Regression Details:





vaunted-asp-487

Status	Finished
Run ID	d5f005b119294ac99802fd7926f64cb4
Duration	2.7s
Datasets used	—
Tags	Add
Source	train.py c0471b3
Logged models	sklearn
Registered models	—

Parameters (1)

Search parameters

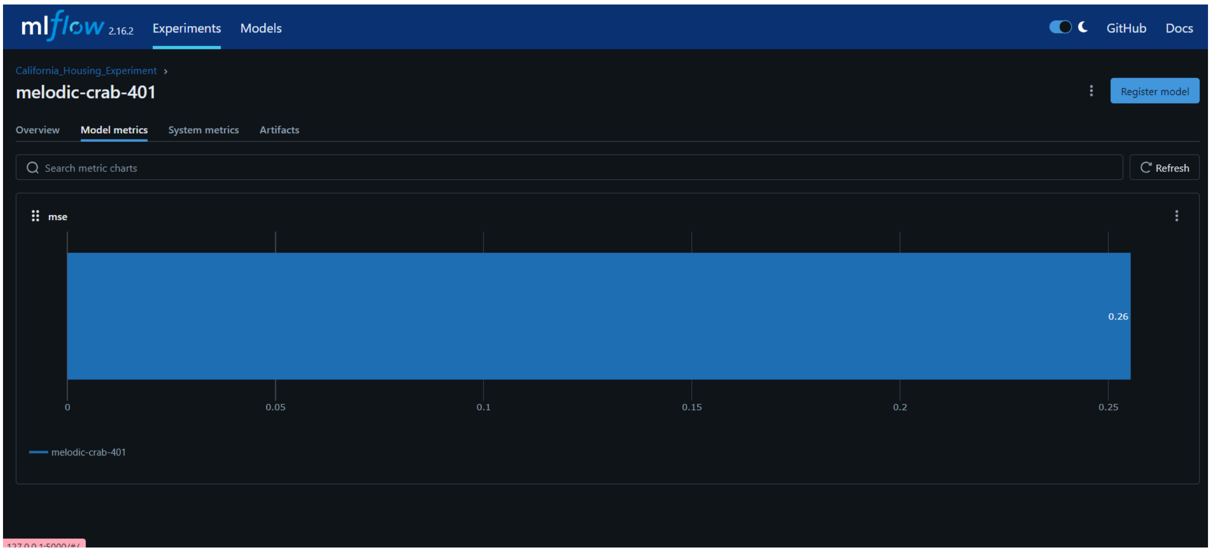
Parameter	Value
model_name	Linear_Regression

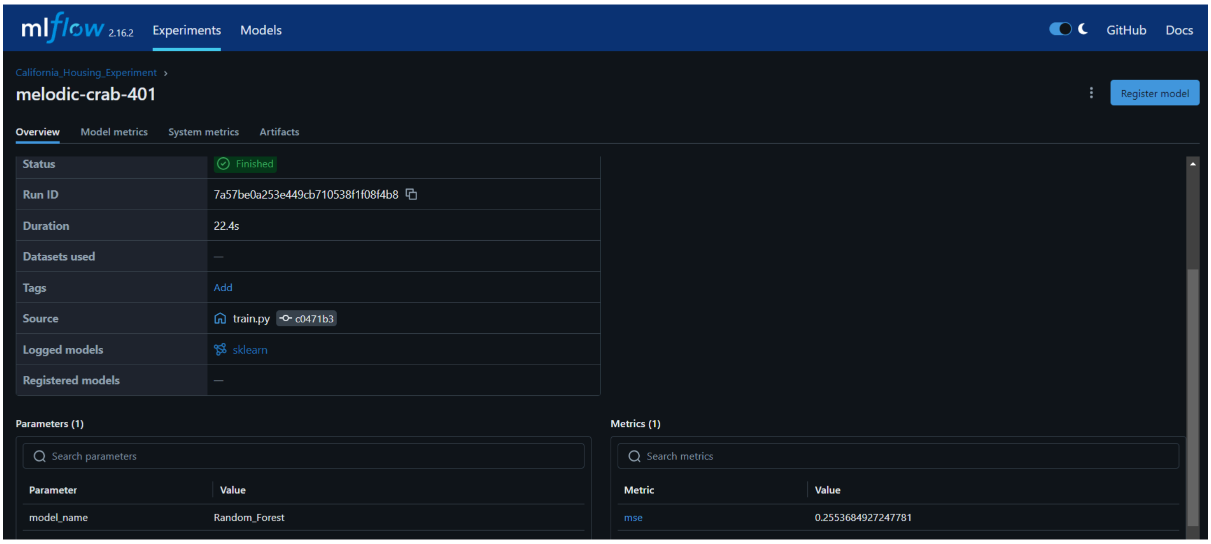
Metrics (1)

Search metrics

Metric	Value
mse	0.5558915986952444

Random Forest Details:





Status	✓ Finished
Run ID	7a57be0a253e449cb710538f1f08f4b8
Duration	22.4s
Datasets used	—
Tags	Add
Source	train.py c0471b3
Logged models	sklearn
Registered models	—

Parameters (1)

Parameter	Value
model_name	Random_Forest

Metrics (1)

Metric	Value
mse	0.2553684927247781

5. Model Registration

The Random Forest model, identified as the best performer, was registered in MLflow's Model Registry for future use. This allows for versioning and management of the model in production environments.

6. Final Remarks

The experiments conducted illustrate the power of using MLflow for managing machine learning workflows. The comparison of the two models highlights the effectiveness of the Random Forest algorithm for this particular dataset and task.