MLops Assignment Report

By- Tanishq Kankerwal (B21BB032)

Experiment Tracking with MLflow

Objective - My goal is to to use the mlflow to see the performance of different models and manage experiment metadata. The steps include model training, performance comparison, and logging the best-performing model in the MLflow Model Registry.

Dataset used

Housing Prices Dataset

It contains the following columns.

crim: Per capita crime rate by town.
zn: Proportion of residential land zoned for lots over 25,000 sq. ft.
indus: Proportion of non-retail business acres per town.
chas: Charles River dummy variable (1 if tract bounds river; 0 otherwise).
nox: Nitric oxides concentration (parts per 10 million).
rm: Average number of rooms per dwelling.
age: Proportion of owner-occupied units built prior to 1940.
dis: Weighted distances to five Boston employment centers.
rad: Index of accessibility to radial highways.
tax: Full-value property tax rate per $10,000.
ptratio: Pupil-teacher ratio by town.
b: b=1000(Bk−0.63)2b = 1000(Bk - 0.63)^2b=1000(Bk−0.63)2 where BkBkBk is the proportion of Black residents by town.
lstat: Percentage of lower status of the population.
medv: Median value of owner-occupied homes in $1000s (target variable).

Models used

Linear Regression

Random Forest

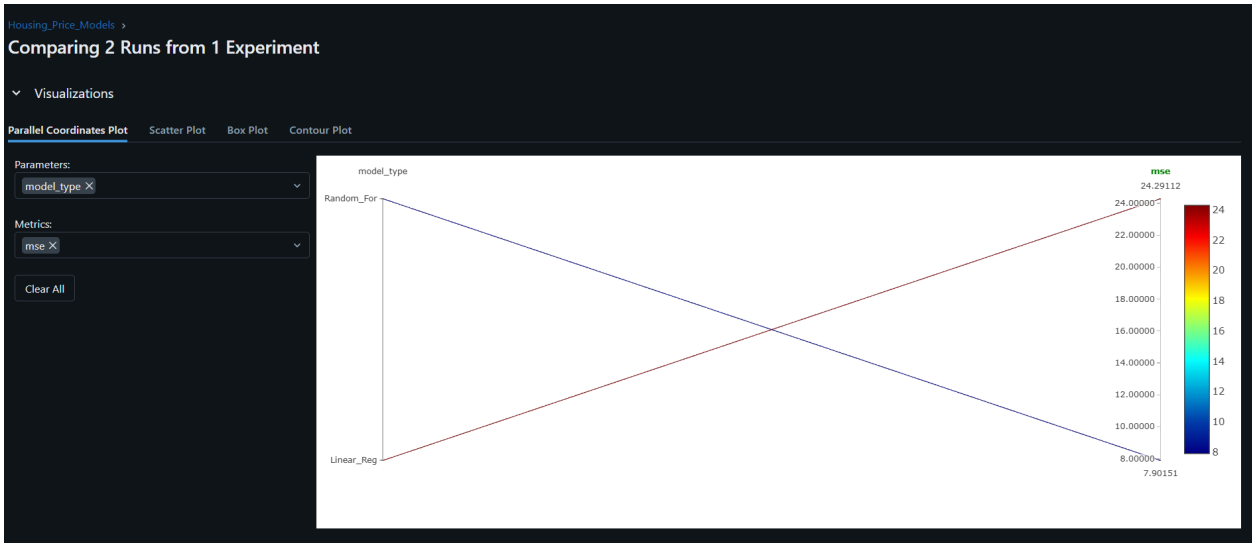
Both models were employed to predict the median value of homes based on various housing features. By comparing their Mean Squared Error (MSE) values in MLflow, you can evaluate which model performs better and under what circumstances. The combination of a simple model (Linear Regression) and a more complex model (Random Forest) provides a comprehensive understanding of the dataset and its predictive capabilities.

Model training

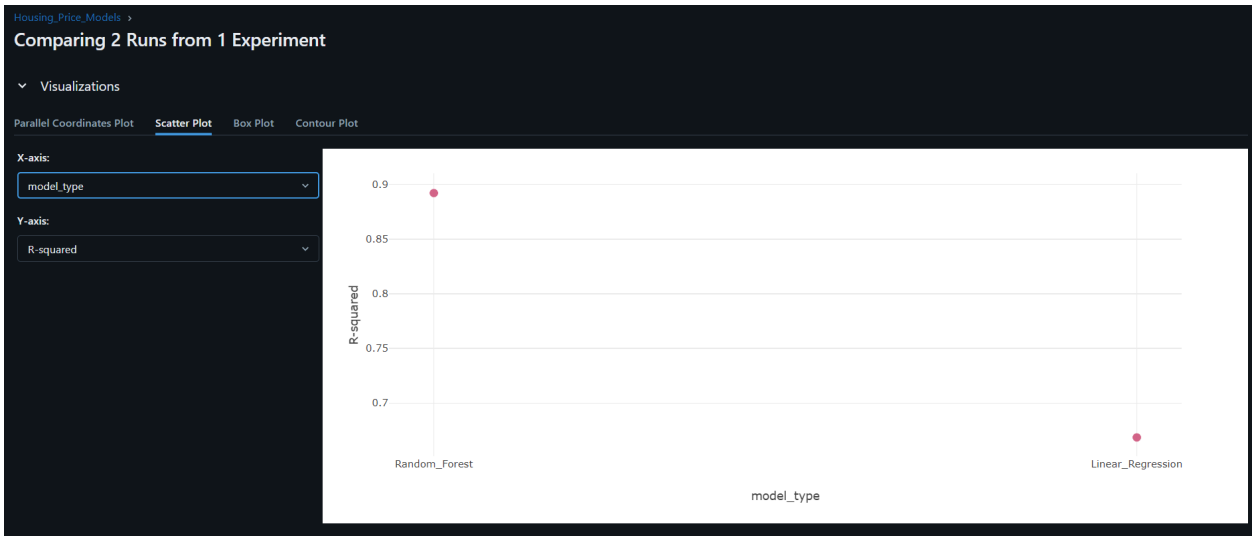
I wrote the code for training the model using 2 approaches, one is linear regression and another is random forest.

Metrics used for comparison

Mean squared Error



R- squared




```
Linear_Regression - MSE: 24.291119474973385
Linear_Regression - R-squared: 0.6687594935356338
2024/10/01 12:01:07 WARNING mlflow.models.model: Mo
Random_Forest - MSE: 7.901513892156864
Random_Forest - R-squared: 0.8922527442109116
```

For Linear Regression:

MSE: 24.291119474973385

R-squared: 0.6687594935356338

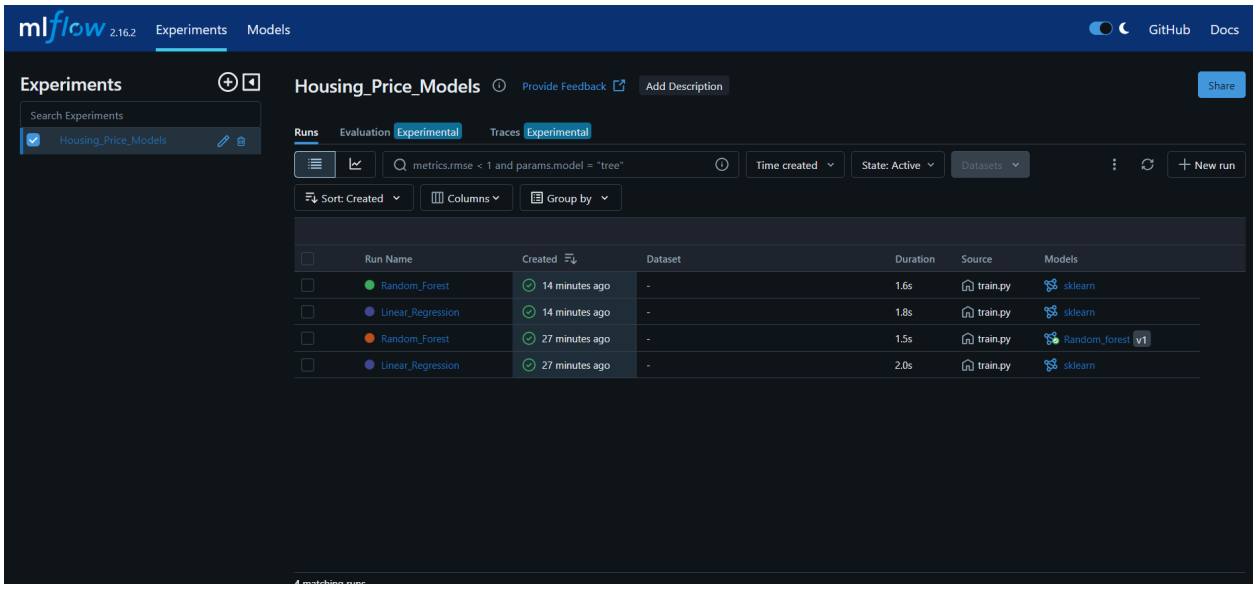
For Random Forest:

MSE: 7.901513892156864

R-squared: 0.8922527442109116

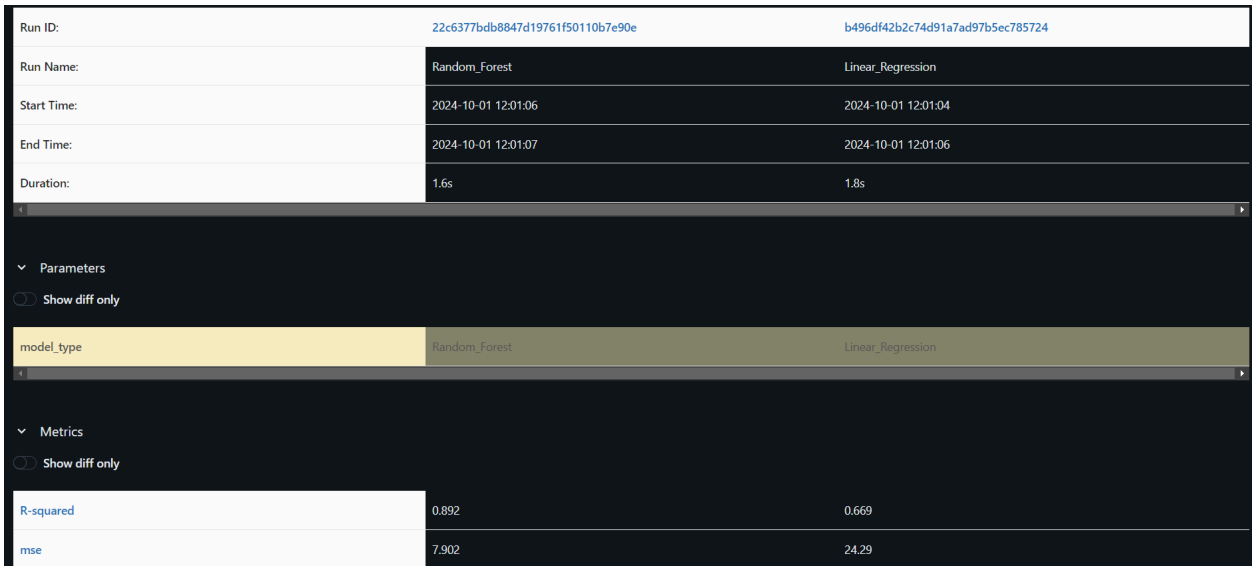
ML Flow for visualization

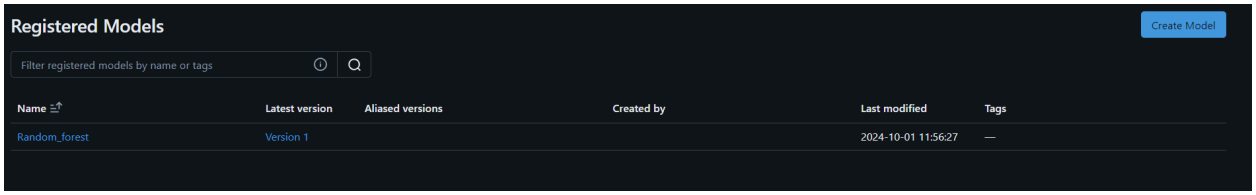
I used ml flow to see the visualization



Registered Models

Compare the model first





Registered the best model which was Random forest

Conclusion

- **Performance**: The **Random Forest** model significantly outperforms **Linear Regression** in terms of both MSE and R squared. The lower MSE and higher R squared value for Random Forest demonstrate that it provides more accurate predictions and captures more of the data's variance, making it a better model for this dataset.
- **Interpretability**: While Linear Regression is simpler and easier to interpret due to its linear assumptions, it falls short in performance compared to Random Forest, which handles complex and non-linear interactions among features effectively.
- **Recommendation**: Based on these results, **Random Forest** is the recommended model for predicting housing prices due to its superior predictive power and ability to generalize well to unseen data.