Report on Model Training and Performance Comparison Using MLFlow

This report analyses and compares two machine learning models—Linear Regression and Random Forest Regression—applied to a bike-sharing dataset. The goal is to predict the hourly count of bike rentals (`cnt`) based on various features such as weather conditions, time of day, and more. The models' performance is compared using metrics such as Mean Squared Error (MSE) and R-squared score (R^2), which are logged using MLflow.

1. Data Preparation Overview

The dataset `hour.csv` was preprocessed to include new features such as:

- **`avg_temp`**: Average temperature derived from `atemp` and `temp`.
- **`heat_index`**: Calculated as a combination of average temperature and humidity.
- **`day_night`**: Categorical feature representing whether the hour falls in the daytime or nighttime.

Several columns, such as `instant`, `casual`, `registered`, `atemp`, `temp`, and `hum`, were removed to eliminate redundancy and avoid data leakage. The final dataset was split into training (80%) and testing (20%) sets.

2. Preprocessing and Pipeline Setup

The preprocessing and model training were handled using a Scikit-learn pipeline. The features were processed as follows:

- **Numerical Features**: Scaled using `MinMaxScaler`.
- **Categorical Features**: Encoded using `TargetEncoder`.

A `ColumnTransformer` applied these transformations, followed by model training using the pipeline.

3. Model Training and Metrics Logging

Two models were trained and their performance metrics were logged using MLflow:

Linear Regression Model:

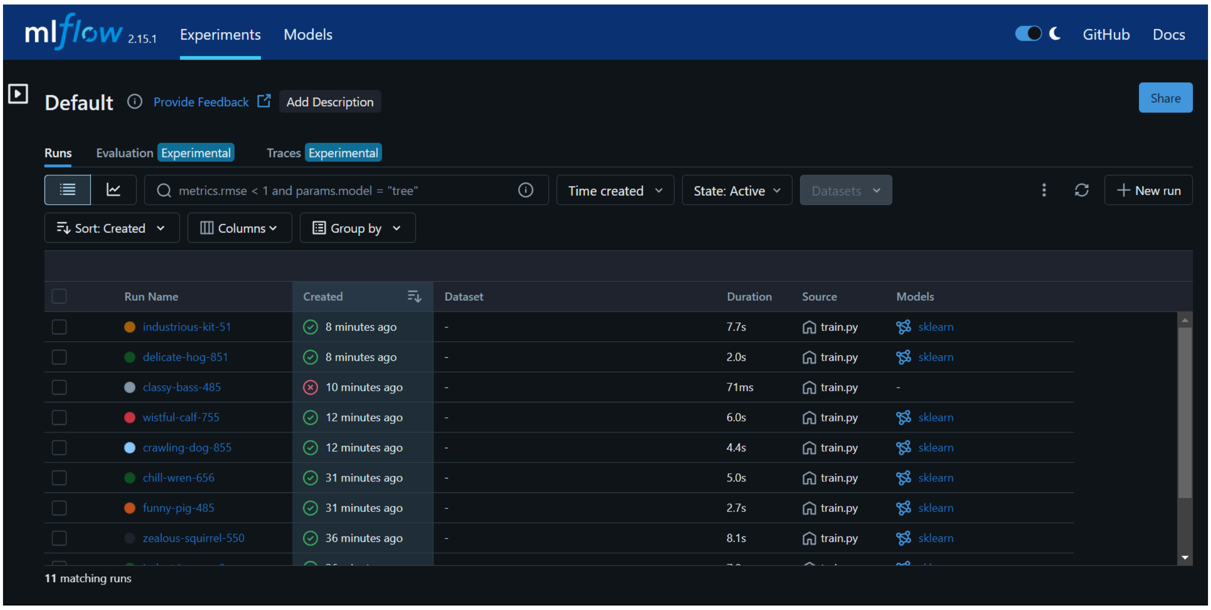
- **Mean Squared Error (MSE):** 15,151.6
- **R-Squared (R^2) Score:** 0.522

Interpretation: The MSE of 15,151.6 indicates the average squared difference between the actual and predicted values. This error value suggests that, on average, the predictions deviate moderately from the actual values. The R^2 score of 0.522 means that the model explains approximately 52.2% of the variance in the bike rental count. This is a moderate level of predictive accuracy, indicating that the model captures just over half of the variability in the data.

Random Forest Regression Model:

- **Mean Squared Error (MSE):** 1,814.7
- **R-Squared (R²) Score:** 0.943

Interpretation: The MSE of 1,814.7 for the Random Forest model is significantly lower compared to the Linear Regression model, indicating that the predictions are much closer to the actual values, with reduced error. The R² score of 0.943 suggests that the model explains around 94.3% of the variance in the bike rental count. This demonstrates a high level of predictive accuracy and indicates that the Random Forest model is far superior to the Linear Regression model in capturing the underlying patterns in the data.

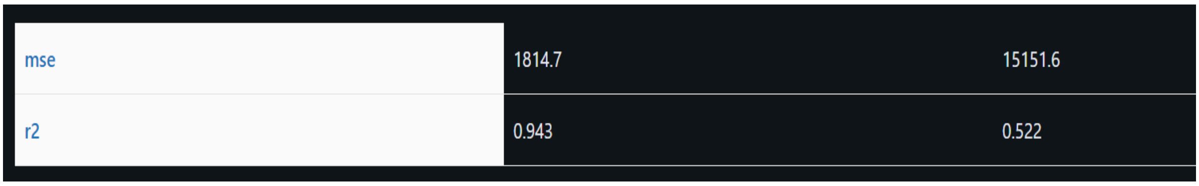


metrics.rmse < 1 and params.model = "tree"

Run Name	Created	Dataset	Duration	Source	Models
industrious-kit-51	8 minutes ago	-	7.7s	train.py	sklearn
delicate-hog-851	8 minutes ago	-	2.0s	train.py	sklearn
classy-bass-485	10 minutes ago	-	71ms	train.py	-
wistful-calf-755	12 minutes ago	-	6.0s	train.py	sklearn
crawling-dog-855	12 minutes ago	-	4.4s	train.py	sklearn
chill-wren-656	31 minutes ago	-	5.0s	train.py	sklearn
funny-pig-485	31 minutes ago	-	2.7s	train.py	sklearn
zealous-squirrel-550	36 minutes ago	-	8.1s	train.py	sklearn

11 matching runs

4. Performance Comparison



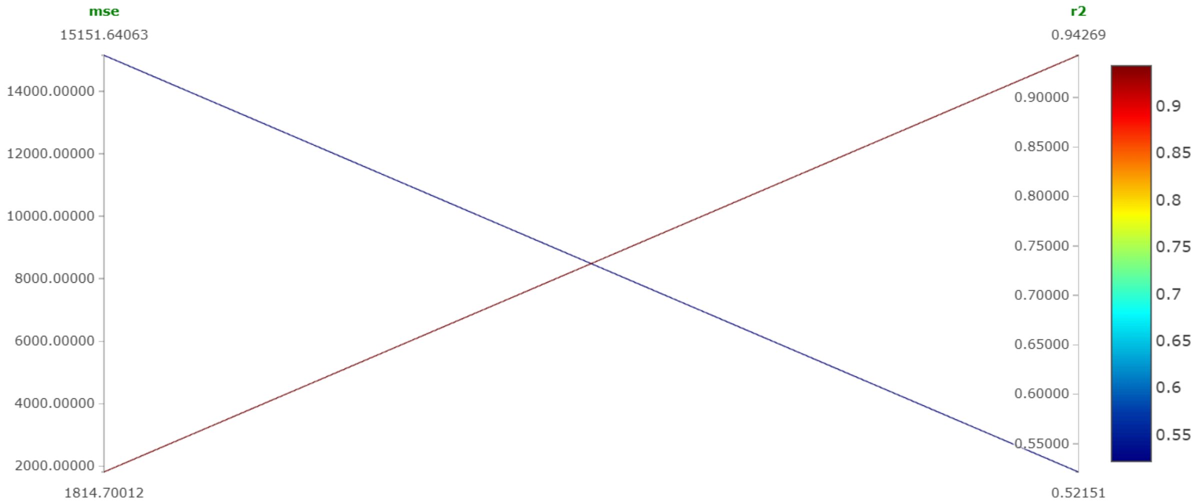
Metric	Random Forest Regression	Linear Regression
mse	1814.7	15151.6
r2	0.943	0.522

MSE Comparison:

The Random Forest model has a significantly lower MSE (1,814.7) compared to the Linear Regression model (15,151.6). This indicates that the Random Forest model fits the data much better, minimizing the squared prediction errors more effectively.

R² Score Comparison:

The R² score for the Random Forest model (0.943) is much higher than that of the Linear Regression model (0.522). This demonstrates that the Random Forest model explains a significantly larger portion of the variability in the bike rental counts. The Linear Regression model, while still moderate in performance, captures just over half of the variability, suggesting that it may not be the most suitable model for this dataset.



5. Conclusion

Model Performance: The Random Forest model outperforms the Linear Regression model significantly, as indicated by both the lower MSE and the higher R² score. The Random Forest model shows strong predictive power, whereas the Linear Regression model captures less than 55% of the variability in the data