MLops Assignment-2

[Report]

Introduction

This report summarizes the enhancements and optimizations made to the MLOps pipeline for predicting bike rentals. The original pipeline, which utilized one-hot encoding for categorical features and a Random Forest Regressor, has been refined to explore alternative encoding techniques and evaluate the performance of a Linear Regression model.

Enhancements and Optimizations

1. **Feature Engineering:**

   - **New Interaction Features:** Created two new interaction features: temp_humidity and temp_windspeed to capture potential non-linear relationships between temperature, humidity, and windspeed.
   - **Justification:** These features might provide additional insights into how environmental factors influence bike rentals.

2. **Categorical Feature Encoding:**

   - **Target Encoding:** Replaced one-hot encoding with target encoding for the categorical features season, weathersit, and day_night.
   - **Evaluation:** Compared the performance of the model with both encoding techniques to assess the impact on prediction accuracy.

3. **Model Comparison:**

   - **Linear Regression:** Trained a Linear Regression model in addition to the Random Forest Regressor.
   - **Evaluation:** Compared the performance of both models using Mean Squared Error (MSE) and R-squared.

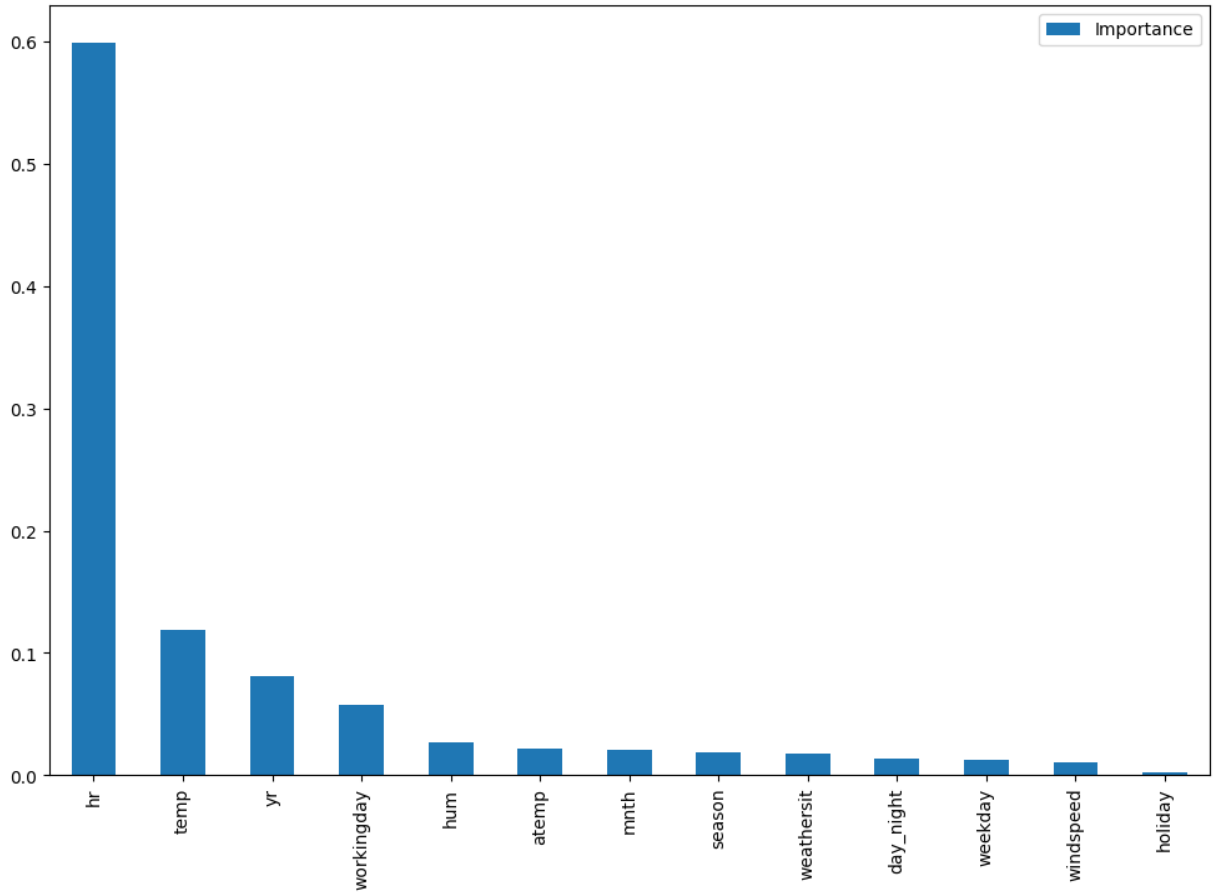
Results

Model Performance:

Model	MSE	R-squared
Random Forest (One-Hot Encoding)	1808.41	0.9429
Random Forest (Target Encoding)	1733.51	0.9453
Linear Regression	15112.95	0.5227

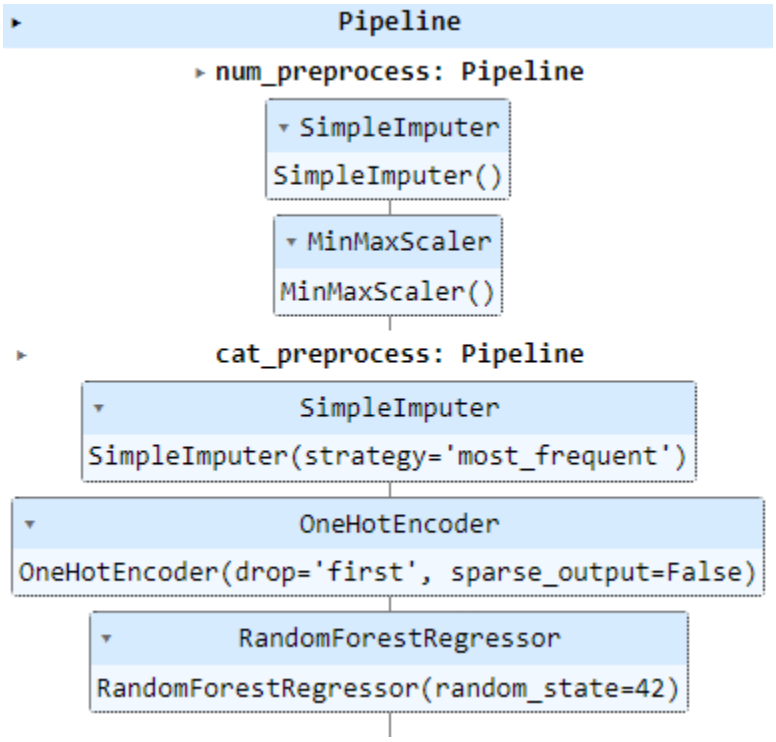
Graph:

This graph shows the importance of factors for bike rentals.

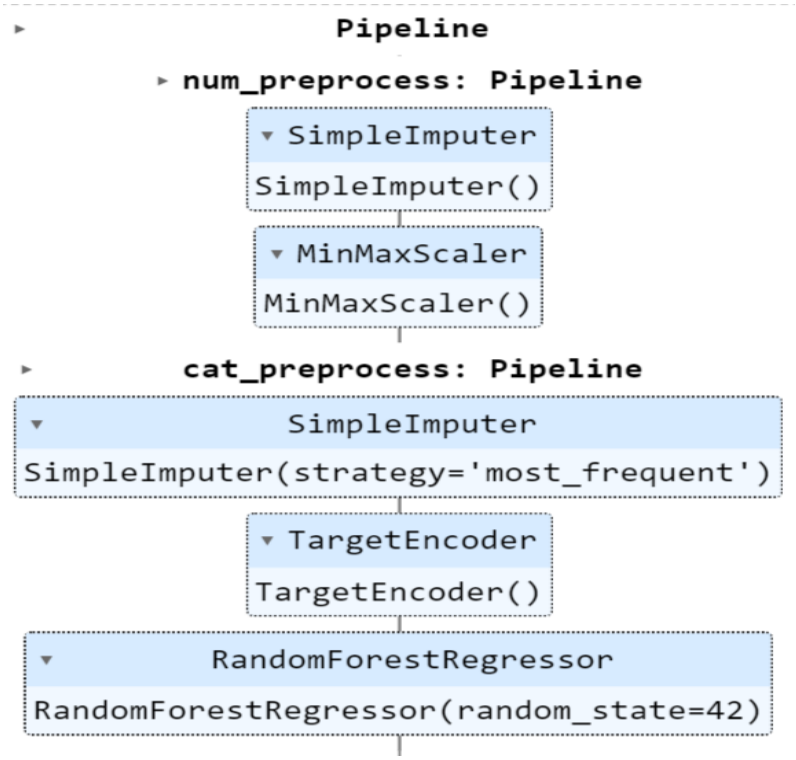


ML-Pipeline:

1) OneHotEncoder Pipeline



2) TargetEncoderr Pipeline



Analysis

- **Target Encoding Improvement:** The Random Forest model using target encoding achieved a slightly lower MSE and a higher R-squared compared to one-hot encoding, suggesting that target encoding was beneficial in this case.
- **Linear Regression Underperformance**: The Linear Regression model, despite being a simpler model, significantly underperformed compared to the Random Forest models, indicating that the non-linear relationships in the data are better captured by the Random Forest.

Conclusion

The MLOps pipeline has been enhanced by incorporating target encoding for categorical features. This change resulted in a slight improvement in the model's performance.