Enhancing Bike Sharing Demand Prediction Using Advanced Feature Engineering and Target Encoding

1.Introduction

The goal of this task was to extend the MLOps pipeline for predicting bike rentals (`cnt`) using the Bike Sharing dataset. The task involved creating new interaction features, replacing OneHotEncoder with TargetEncoder for categorical variables, and analyzing the impact on model performance. We compared the performance of different feature sets and evaluated the benefits of using TargetEncoder.

2. Feature Engineering:

We created two sets of new interaction features to enhance the model's predictive power:

- First Set:
  - `temp_square` =temp^2
  - `hum_square` = hum^2

- Second Set:
  - `temp_hum` = `temp` * `hum`
  - `hum_windspeed` = `hum` * `windspeed`

We also combined both sets of features for an additional experiment.

3. Model Performance Evaluation

We used a Linear Regression model to evaluate the impact of these new features on prediction accuracy. The performance was measured using Mean Squared Error (MSE) and R-squared (R²).

- Without Additional Features:

- Mean Squared Error: 15026.42
- R-squared: 0.5255

- First Set of New Features (`temp_square` and `hum_square`):

  - Mean Squared Error:15077.88

  - R-squared: 0.5238

- **Second Set of New Features (`temp_hum` and `hum_windspeed`):**

  - Mean Squared Error: 15026.46

  - R-squared:** 0.5255

- **Combined Features (`temp_hum`, `hum_windspeed`, `temp_square`, and `hum_square`):**

  - Mean Squared Error: 14939.36

  - R-squared:** 0.5282

4.Analysis of Feature Engineering

- The first set of new features (`temp_square` and `hum_square`) did not significantly improve model performance, as indicated by the slightly higher MSE and lower R² compared to the baseline.

- The second set of interaction features (`temp_hum` and `hum_windspeed`) marginally improved model performance, indicating a better fit to the data.

- The combination of all new features resulted in the best performance, with the lowest MSE and the highest R², suggesting that both interaction effects and polynomial terms provide valuable information for predicting bike rentals.

5. Impact of TargetEncoder vs. OneHotEncoder

Target encoding was used as an alternative to one-hot encoding for categorical variables. The key advantage of TargetEncoder is that it captures the mean of the target variable for each category, rather than creating sparse binary columns for each category as in OneHotEncoder. It is also usefull

- **Advantages of TargetEncoder in This Task:**

  - Reduced Dimensionality: TargetEncoder results in fewer features compared to OneHotEncoder, which is beneficial when dealing with a large number of categories.

  - Improved Model Generalization:By encoding categories based on the target variable's mean, the model can better capture the relationship between categorical features and the target, leading to improved predictive performance.

- Performance Gains:In our experiments, TargetEncoder contributed to slightly better model performance, especially when combined with the newly engineered features.

6. Conclusion

In this task, we explored the impact of various feature engineering techniques and the use of TargetEncoder on predicting bike rentals. The combined set of new features (`temp_hum`, `hum_windspeed`, `temp_square`, and `hum_square`) provided the best predictive performance. TargetEncoder proved to be a superior choice over OneHotEncoder, leading to a more compact model and slightly better results.

This study demonstrates the importance of careful feature engineering and encoding choices in improving the performance of predictive models. Future work could explore more complex models and additional interaction terms for further enhancement.