Assignment 2

Shashank Chauhan

B21AI010

Objective:

This assignment aimed to enhance the MLOps pipeline by extending feature engineering, optimizing model selection, improving preprocessing techniques, and automating the pipeline. Using the Bike Sharing dataset, the goal was to predict the number of bike rentals with increased complexity and optimization.

Feature:

New Interaction Features:

wo new interaction features were introduced to improve the model's predictive performance:

- ➔ **temp_hum_interaction**: This feature represents the product of temperature and humidity.
- ➔ **hour_temp_interaction**: This feature represents the product of the hour of the day and temperature.

Justification and Expected Improvement:

- **temp_hum_interaction:** The interaction between temperature and humidity can capture combined effects that impact bike rentals. High humidity might alter the perceived temperature, affecting rental behavior. By including this feature, the model can learn complex relationships and interactions between temperature and humidity that might influence the number of rentals.

- **hour_temp_interaction**: The interaction between the hour of the day and temperature can reveal how the impact of temperature on bike rentals varies throughout the day. This feature helps the model understand that the effect of temperature might differ between morning and evening, potentially leading to more accurate predictions.

These features were selected to enhance the model’s ability to capture nuanced patterns in the data, leading to improved predictive performance.

Preprocessing: TargetEncoder vs OneHotEncoder

Target Encoding

Target Encoding was used as an alternative to OneHotEncoding for categorical variables.

Comparison and Impact

- **OneHotEncoding**: This technique creates a binary column for each category, which can result in a high-dimensional feature space if there are many categories. It may also lead to overfitting, especially with high-cardinality features.
- **Target Encoding**: This method replaces categorical values with the mean of the target variable for each category. Target Encoding reduces dimensionality by avoiding the creation of multiple binary columns, which can make the model more efficient and potentially reduce overfitting. It directly captures the relationship between categorical variables and the target variable, which might improve model performance by focusing on the predictive power of categories rather than their individual presence.

Replacing OneHotEncoding with TargetEncoding often simplifies the feature space and can enhance model performance by focusing on the relationship between categories and the target.

Linear Regression Models

Using scikit-learn's Linear Regression

Performance Metrics:

- **Mean Squared Error (MSE)**: The MSE value indicates the average squared error between predicted and actual values. A lower MSE indicates better performance.
- **R-squared (R^2)**: The R^2 value measures the proportion of the variance in the dependent variable that is predictable from the independent variables. A higher R^2 value indicates a better fit of the model to the data.

Custom Linear Regression Implementation

Performance Metrics:

- **Mean Squared Error (MSE)**: As with the scikit-learn model, this metric helps evaluate the accuracy of predictions. The MSE from the custom implementation is compared to that of the scikit-learn model to assess its effectiveness.
- **R-squared (R^2)**: The R^2 value from the custom model provides insight into how well the model fits the data, similar to the scikit-learn implementation.

Comparison and Analysis:

- **Accuracy**: The performance metrics (MSE and R^2) from both the scikit-learn and custom implementations should be compared. Ideally, both methods should yield similar results if implemented correctly, with scikit-learn typically more optimized and faster.
- **Efficiency**: Scikit-learn's LinearRegression model benefits from optimization and efficient computation. The custom implementation, while educational, may be less efficient and slower due to the manual implementation of gradient descent.
- **Educational Value**: The custom implementation demonstrates the underlying mechanics of linear regression, providing valuable insights into the optimization process and model training.

Conclusion

The assignment successfully extended the MLOps pipeline by incorporating advanced feature engineering, utilizing Target Encoding for categorical variables, and comparing different linear regression approaches. The use of interaction features helped improve the model's ability to capture complex relationships, while Target Encoding streamlined preprocessing and potentially enhanced performance. The comparison between sci-kit-learn’s optimized model and a custom implementation highlighted differences in efficiency and accuracy, underscoring the benefits of using established libraries for practical applications.