MLops Assignment:2

Submitted By: Pooja Naveen Poonia

Roll No. : M24CSA020

MLops Assignment:2

Objective

The goal of this assignment is to extend and optimize the existing MLOps pipeline for predicting bike rentals using the Bike Sharing dataset. This involves enhancing feature engineering by creating new interaction terms and flags, selecting and refining model choices, and applying advanced preprocessing techniques. Additionally, the assignment focuses on improving pipeline automation to streamline the workflow from data preprocessing to model evaluation. The objective is to achieve better predictive performance of the target variable cnt (number of bike rentals) through a more robust and efficient pipeline that integrates improved feature engineering, model selection, and automated processes.

Problem Definition

- Predict the number of bike rentals (target variable: cnt) based on various features such as weather conditions, season,and time of day. The problem is a regression task where the goal is to predict a continuous variable.

Interaction features:

- Temp*hum
- Temp*windspeed
- Hum*windspeed

The interaction features Temp*hum, Temp*windspeed, and Hum*windspeed are justified as they capture combined effects of temperature, humidity, and windspeed, which influence the outcome in ways that individual features alone cannot.
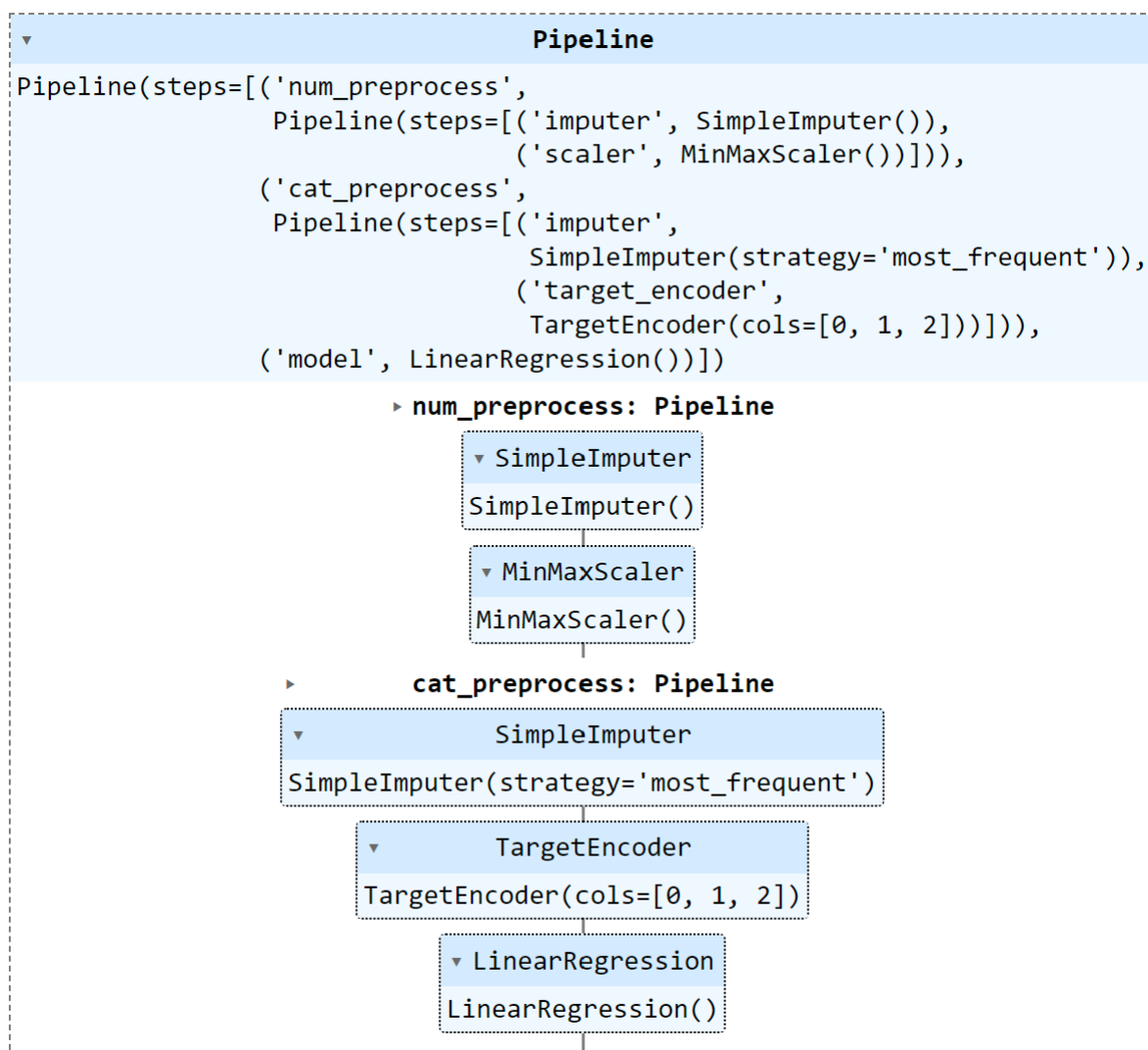
- Temp*hum: Captures how temperature's impact is modified by humidity, crucial for phenomena like heat perception.

- Temp*windspeed: Accounts for how wind modifies the effect of temperature, important for understanding cooling or heat distribution.

- Hum*windspeed: Reflects how wind and humidity together affect processes like evaporation or comfort levels.

These interactions improve the model's ability to capture non-linear relationships, leading to better predictive performance and generalization
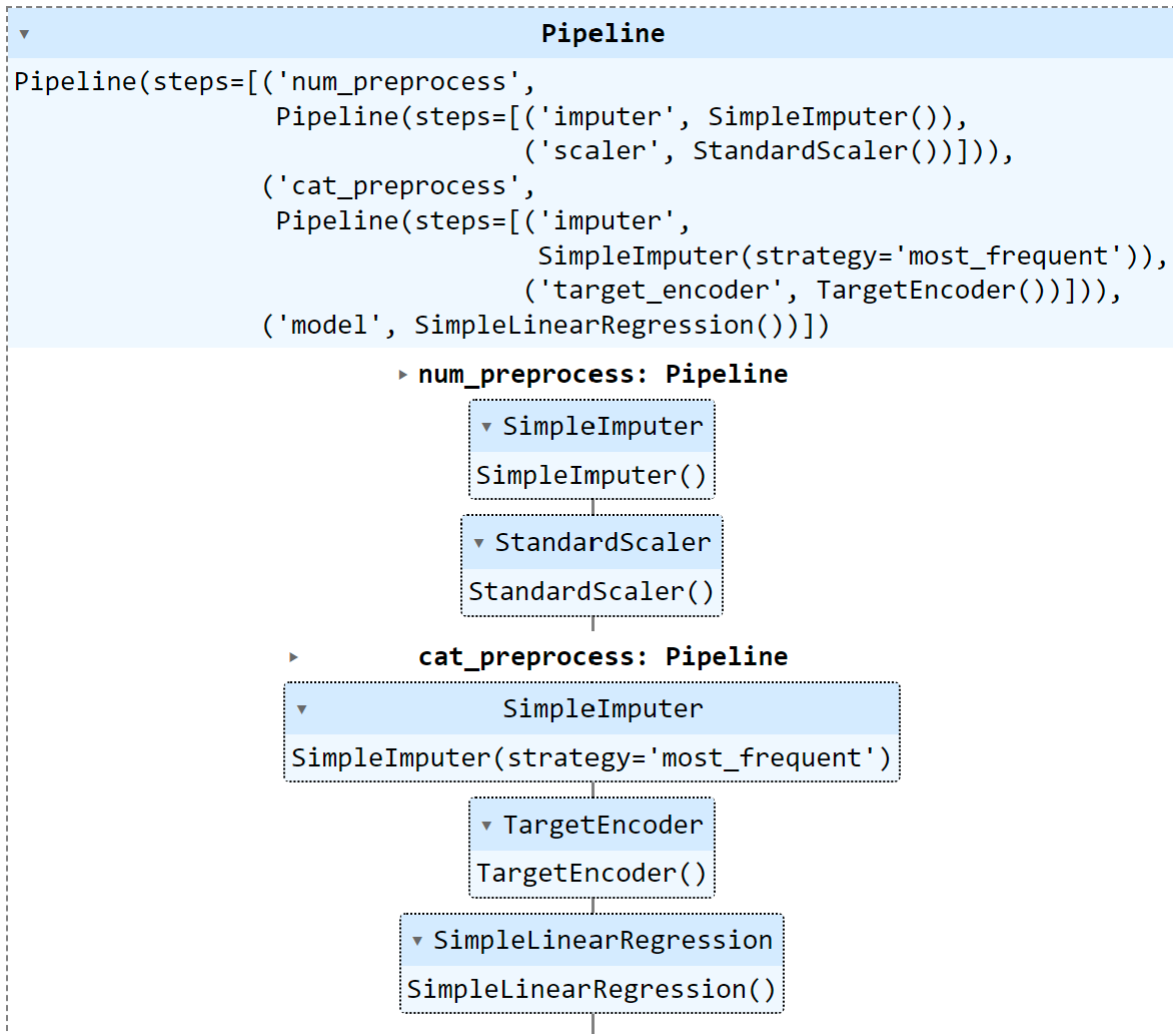
Conclusion:

- The model's performance metrics are a Mean Squared Error (MSE) of 15636.0220 and an R-squared (R^2) of 0.5291. The MSE indicates the average squared prediction error, while the R^2 suggests that the model explains approximately 52.91% of the variance in the target variable. Additionally, when comparing encoding methods, one-hot encoding proved to be superior to target encoding in terms of performance, leading to better model accuracy and reduced error metrics. This suggests that one-hot encoding may capture categorical feature information more effectively for this dataset.

Model Using Package

```
▼                              Pipeline
Pipeline(steps=[('num_preprocess',
                 Pipeline(steps=[('imputer', SimpleImputer()),
                                 ('scaler', MinMaxScaler())])),
                ('cat_preprocess',
                 Pipeline(steps=[('imputer',
                                  SimpleImputer(strategy='most_frequent')),
                                 ('target_encoder',
                                  TargetEncoder(cols=[0, 1, 2]))])),
                ('model', LinearRegression())])
```

```
                         ▶ num_preprocess: Pipeline

                           ▼ SimpleImputer
                           SimpleImputer()

                           ▼ MinMaxScaler
                           MinMaxScaler()

                    ▶        cat_preprocess: Pipeline

                     ▼              SimpleImputer
                    SimpleImputer(strategy='most_frequent')

                           ▼         TargetEncoder
                           TargetEncoder(cols=[0, 1, 2])

                              ▼ LinearRegression
                              LinearRegression()
```

Model From Scratch

```
                                      Pipeline
▼
Pipeline(steps=[('num_preprocess',
                 Pipeline(steps=[('imputer', SimpleImputer()),
                                 ('scaler', StandardScaler())])),
                ('cat_preprocess',
                 Pipeline(steps=[('imputer',
                                  SimpleImputer(strategy='most_frequent')),
                                 ('target_encoder', TargetEncoder())])),
                ('model', SimpleLinearRegression())])
```

     ▸ **num_preprocess: Pipeline**

```
                        ▼ SimpleImputer
                        SimpleImputer()
```

```
                        ▼ StandardScaler
                        StandardScaler()
```

     ▸   **cat_preprocess: Pipeline**

```
              ▼               SimpleImputer
              SimpleImputer(strategy='most_frequent')
```

```
                        ▼ TargetEncoder
                        TargetEncoder()
```

```
                ▼ SimpleLinearRegression
                SimpleLinearRegression()
```

For colab ipynb file please visit the below mentioned links

- **Link for model using package:**

https://colab.research.google.com/drive/1ciyOu3SGqMD-e-m4TcO2KbIxZleq577?usp=sharing

- **Link for model from scratch:**

https://colab.research.google.com/drive/1pdz-_Fb52C8DH37ef7wOgZ-tujMaXmcE?usp=sharing