MLOPS ASSIGNMENT-2 REPORT

Name: Ratnesh Dubey

Roll Number: M24CSA027

AIM 1

Create at least two new interaction features between numerical variables (e.g.,

temp * hum). Justify your choice of features and explain how they might

improve the model's predictive performance.

OUTCOME 1

2 new features (temp_hum, temp_windspeed) created

High temperature and high humidity may decrease the number of bike rentals on that day and moderate temperature and moderate humidity may increase the number of bike rentals per day as the weather is more pleasant, more people might come out from homes, thus the feature temp_hum can affect the model performance.

A mild day might feel colder with strong winds, potentially decreasing bike rentals. On the other hand, moderate wind could make hot days feel cooler and more pleasant for biking. This interaction can capture the compounded effect of wind and temperature on the decision to rent bikes.

These interaction terms allow the model to account for non-linear relationships between the predictors and the target variable. Instead of independently assuming a linear effect of temperature or windspeed, the model can now learn how these variables combine to influence bike rentals, thus improving the generalized performance of the model.

AIM 2

Replace the OneHotEncoder with TargetEncoder for categorical variables. Evaluate how this change impacts the model's performance compared to one-hot encoding.

OUTCOME 2

One Hot Encoder Performance:

Mean Squared Error: 1859.7503947386544, R-squared: 0.9412687119400371

Target Encoding Performance:

Mean Squared Error: 1793.641308661706, R-squared: 0.9433564500519636

We can see with Targe Encoding, we are getting a less Mean Squared Error compared to One Hot Encoding, but the R-square in Target Encoding is a little bit higher than One Hot Encoding, overall Target Encoding seems to perform better here.

AIM 3

Train LinearRegressor: a. Using the package, b. Write/Train it by scratch following the steps of a linear regressor. Compare their performance using metrics like Mean Squared Error (MSE) and R-squared.

OUTCOME 3

Linear Regression from Package Performance:

Mean Squared Error: 14973.691511022034, R-squared: 0.5271278382610913

Linear Regression from Scratch Performance:

Mean Squared Error: 15467.611031944612, R-squared: 0.5115297613665739

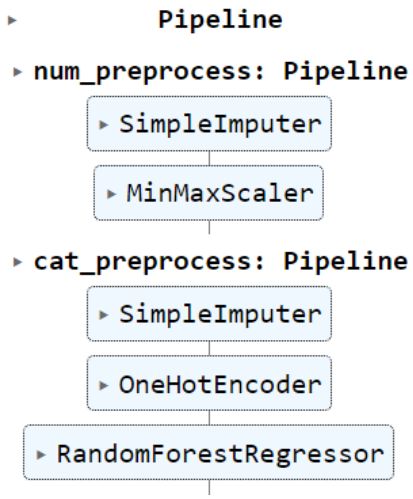
The comparison between the Linear Regression model from the package and the one implemented from scratch shows that both models perform similarly, but the package implementation has a slight edge. The package model achieved a lower Mean Squared Error (MSE) of 14,973.69 compared to 15,467.61 for the scratch implementation, indicating better accuracy. Additionally, the R-squared value for the package model is higher at 0.527, compared to 0.512 for the scratch model, suggesting that the package model explains slightly more variance in the data. Overall, while both models are fairly close in performance, the package implementation is slightly more effective.

AIM 4

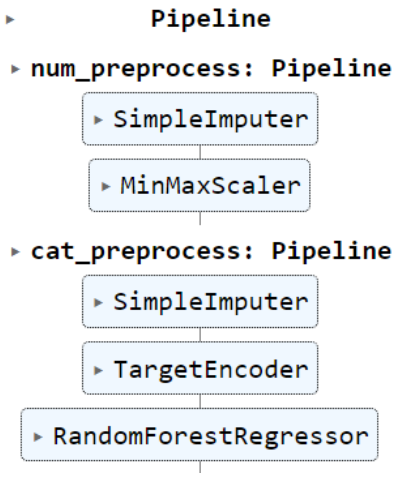
Save the screenshot of MLpipelines

OUTCOME 4

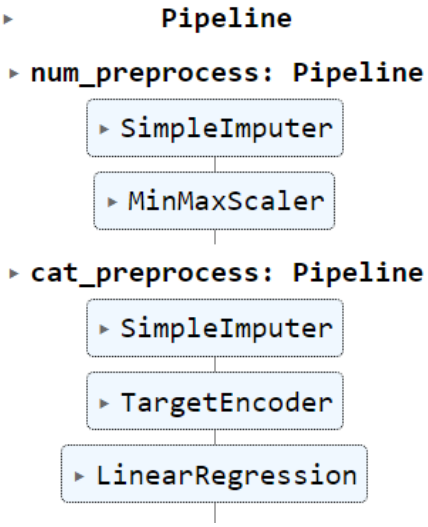
With One Hot Encoding and Random Forest



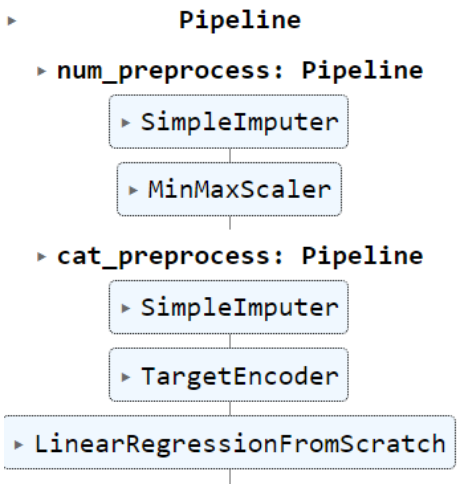
With Target Encoding and Random Forest



With the Linear Regression Package version



With the Linear Regression From Scratch version


Pipeline
num_preprocess: Pipeline
SimpleImputer
MinMaxScaler
cat_preprocess: Pipeline
SimpleImputer
TargetEncoder
LinearRegressionFromScratch