REPORT
ASSIGNMENT -2
SUBJECT - ML OPs

Name: Suvigya Sharma
Roll No.: M24CSA033

- **TASK 1: New Interaction features**

1. temp_hum : Temperature * Humidity

Physical Relationship: Temperature and humidity are closely related in determining comfort levels and outdoor conditions. High temperatures combined with high humidity typically create uncomfortable conditions, which can significantly affect bike rentals. Conversely, high temperatures with low humidity may still be conducive to outdoor activities.

Impact on Rentals: People are more likely to rent bikes when the weather is comfortable. The interaction between temperature and humidity can capture more complex patterns in how these weather conditions jointly influence rental behavior, which a linear model might miss if these features are considered independently.

Possible Impact on Model Performance**:** By creating this interaction term, the model can capture non-linear effects where, for instance, bike rentals drop sharply beyond certain combinations of high temperature and high humidity, providing a more nuanced prediction of bike rentals.

Temp, hum have correlation of -0.07

2. temp_windspeed : Temperature * windspeed

Physical Relationship: Wind speed and temperature together influence the perceived temperature (wind chill effect). On colder days, high winds can make it feel much colder, possibly deterring bike rentals. On warmer days, a mild wind might make biking more pleasant.

Impact on Rentals: High temperatures with low wind speed might encourage bike rentals as the conditions are more pleasant. However, if the wind speed is high, even a comfortable temperature might deter some people from biking. This interaction can help capture these non-linear effects in the prediction model.

Possible Impact on Model Performance: By including this interaction term, the model can better understand how wind speed modifies the effect of temperature on bike rentals. This could lead to more accurate predictions, particularly on days with extreme weather conditions.

Temp, windspeed have correlation of -0.02

3. **hum_windspeed: Humidity * Windspeed** : Made no significant impact on Mean Squared Error and R squared error. Hence, removed.

- **TASK 2: Replacing One hot encoder with Target Encoder in the categorical features**

In the categorical pipeline, I applied Target Encoder to the following categorical features:

a. season : Season of the year (1:winter, 2:spring, 3:summer, 4:fall).
Encoded values after implementing pipeline = array([111.1145686 , 208.34406895, 236.01623665, 198.86885633])

b. weathersit : Weather situation (1: Clear, 2: Mist, 3: Light Snow/Rain, 4: Heavy Rain/Snow).

Encoded values after implementing pipeline = array([204.86927188, 175.16549296, 111.57928118, 171.67953962])

c. day_night :This column categorizes each value in the existing hr column as either 'day' or ‘night’ based on the hour of the day.

Encoded values after implementing pipeline = array([98.89413845, 265.22593258])

-> Comparison between Onehot Encoder and Target Encoder on Random Forest Regressor with added interaction features:

Onehot + Random Forest:
Mean Squared Error: 1808.4074990292243
R-squared: 0.9428901308176855

target encoder + random forest regressor:
Mean Squared Error: 1771.7479210545564
R-squared: 0.9440478475953121

Conclusion: The results suggest that target encoding provided a slight improvement in model performance over one-hot encoding. The lower MSE and higher R^2 indicate that target encoding is more effective in capturing the relationships between categorical features and the target variable (bike rentals), resulting in a more accurate and reliable model.

Practical Implication**:** In this context, target encoding is a preferable choice for handling categorical variables when using a Random Forest Regressor, especially if the goal is to optimize the predictive accuracy of the model.

- **TASK 3: Train Linear Regression Model**

a. Using Package
b. From scratch

Results (same):
MSE: 14974.133860641217
R^2: 0.5271138687719702

Linear Regression implementation from a package library (like sklearn) and a custom implementation written from scratch should yield the same results on the same dataset, assuming that both implementations use the Same optimization Method, handle Data in the same way, include regularization (if any) the same way.

Comparison between Onehot Encoder and Target Encoder on Linear Regressor with added interaction features:

Onehot + Linear Regression
Mean Squared Error: 14778.22
R-squared: 0.53

Target encoder + Linear Regression
Mean Squared Error: 14974.133860641217
R-squared: 0.5271138687719702

TASK 4: MLPipeline