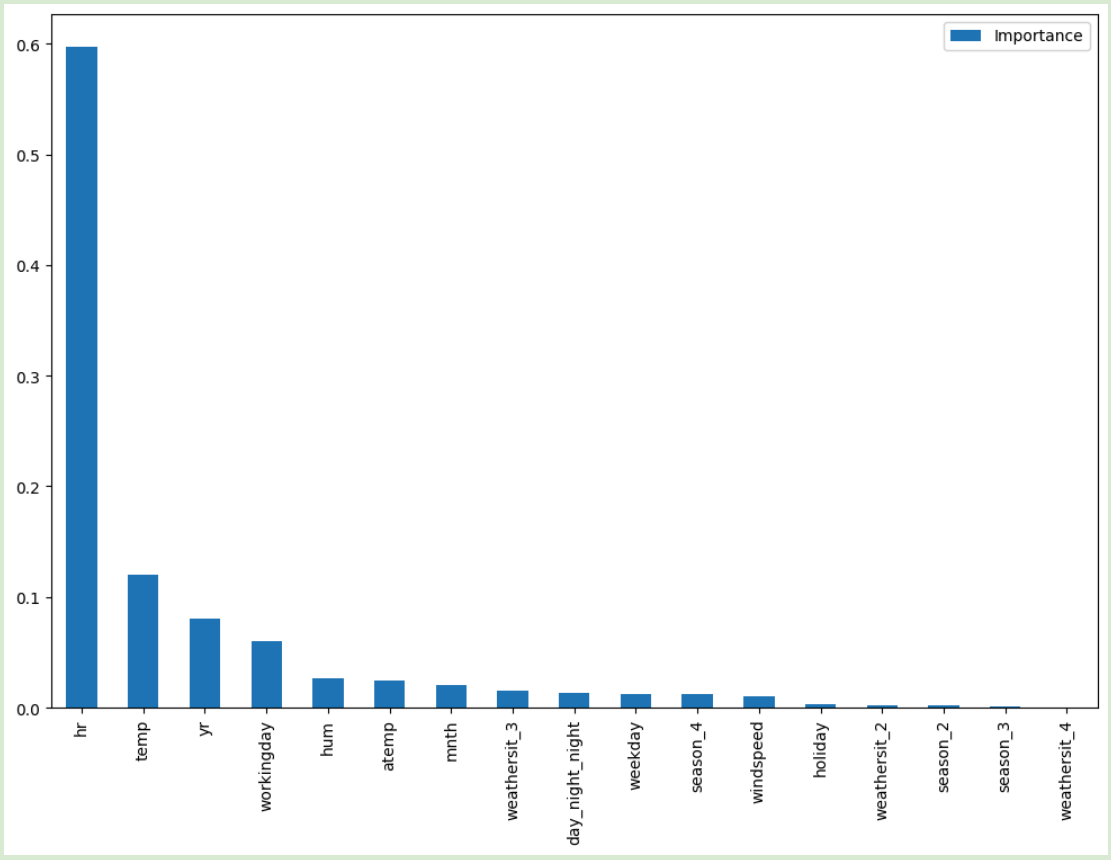
Saman Aijaz Siddiqui (B21BB027)

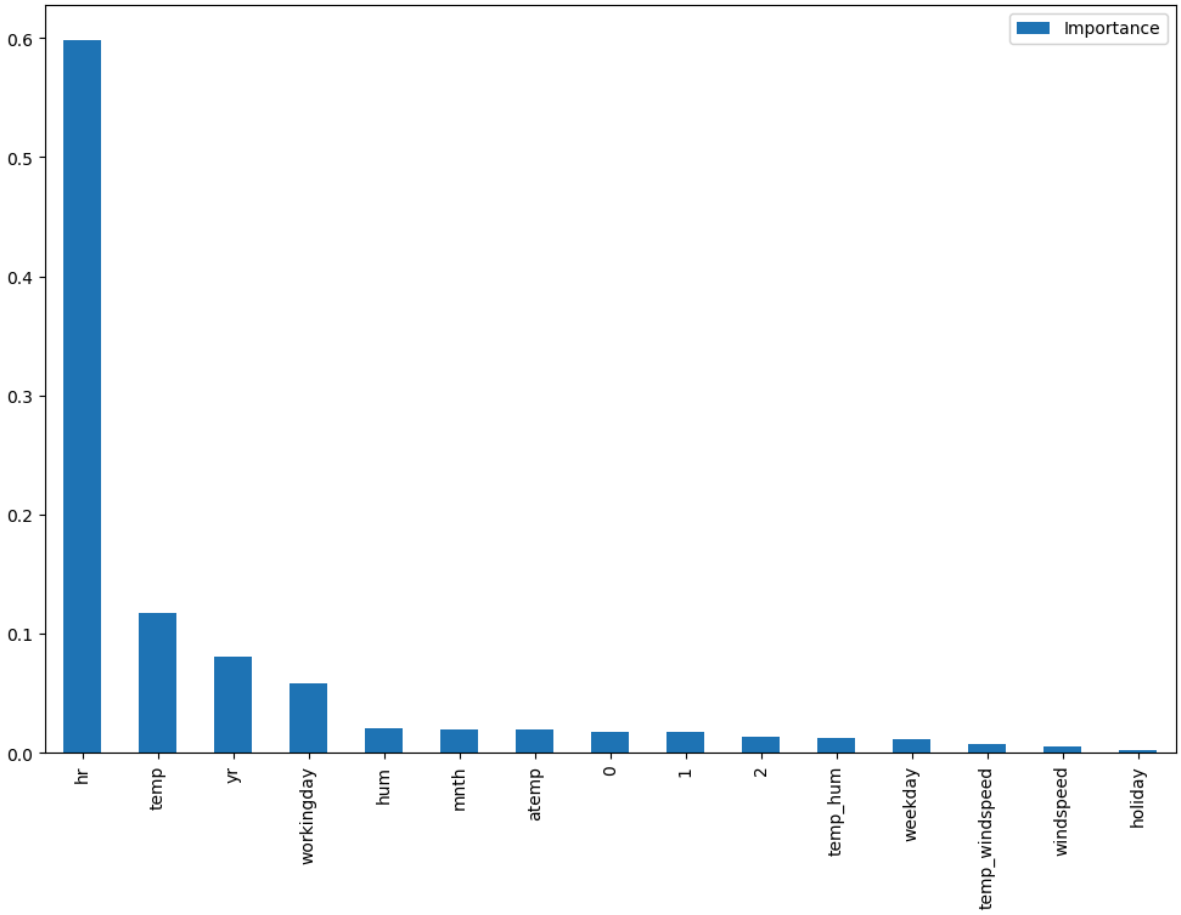
Assignment 02:

1. **Create Interaction Features:**

   Importance of Features:



I have created interaction features between numerical variables. This can help the model capture relationships that might not be evident from individual features alone. For example, the product of `temp` (temperature) and `hum` (humidity) could capture the effect of humid heat on bike rentals.



Justification:

- **`temp * hum`**: High temperature combined with high humidity might deter bike rentals more significantly than each factor alone.
- **`temp * windspeed`**: On windy days, higher temperatures might not feel as hot, potentially affecting bike rental behavior differently compared to still air.

2. Replaced `OneHotEncoder` with `TargetEncoder`:

`TargetEncoder` can be used to encode categorical variables based on the mean of the target variable (in this case, `cnt`). This approach can capture some information about the relationship between the categories and the target variable, potentially improving model performance.

3. Impact Evaluation:

After implementing `TargetEncoder`, the model compares performance (MSE and R^2) to the previous model that used `OneHotEncoder`.

Train `LinearRegressor`:

Training a `LinearRegressor` using both Scikit-Learn and from scratch:

a. Using Scikit-Learn:
Number of features after encoding: 15
Mean Squared Error: 1776.7196810182302
R-squared: 0.9438908384249323

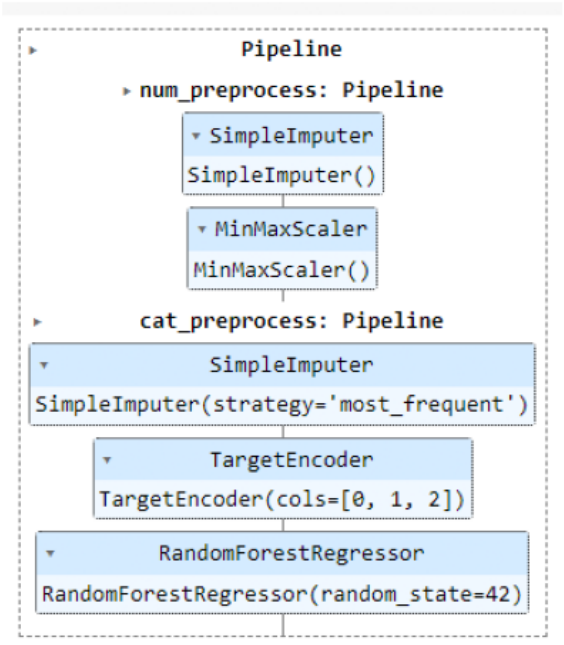
b. Training Linear Regression from Scratch:
Linear Regression MSE (From Scratch): 14974.133860641261
Linear Regression R-squared (From Scratch): 0.5271138687719688

Comparison:

- **Option (a) using Scikit-Learn** is better, because:

1. **Mean Squared Error (MSE)**: This metric measures the average of the squares of the errors—that is, the average squared difference between the estimated values and the actual value. A lower MSE indicates a better fit of the model to the data.
    - Option (a) MSE: 1776.72
    - Option (b) MSE: 14974.13
2. The MSE of Option (a) is significantly lower than that of Option (b), indicating that the Scikit-Learn model has a much smaller average error.
3. **R-squared**: This metric represents the proportion of the variance for the dependent variable that's explained by the independent variables in the model. An R-squared closer to 1 indicates a better fit of the model.
    - Option (a) R-squared: 0.944
    - Option (b) R-squared: 0.527
4. The R-squared value of Option (a) is much higher than that of Option (b), indicating that the Scikit-Learn model explains a much higher proportion of the variance in the dependent variable.

4. Integrate MLflow for Experiment Tracking:



The pipeline is as follows: