Bike Sharing
Case Study

Assignment-based Subjective Questions

Topic: Linear Regression
By: Ramesh Velivela

Batch: ML61 EPGP in AI & ML

Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
2. Why is it important to use drop_first=True during dummy variable creation?
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
4. How did you validate the assumptions of Linear Regression after building the model on the training set?
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

- The categorical variables available in the assignment are "**season**", "**workingday**", "**weathersit**", "**weekday**", "**yr**", "**holiday**", and "**mnth**".
- "**season**" – Based on the data available, the most favorable seasons for biking are summer and fall. Higher targets can be planned in summer and fall with strategic advertising. Spring has significant low consumption ratio.
- "**workingday**" – Working day represents weekday and weekend/holiday information. The registered users are renting bikes on working days whereas casual users prefer the bikes on non-working days. This effect is nullified when we look at the total count because of the contradictory behavior of registered and casual users. Registered and casual users' identity and relevant strategy for working and not working days shall help to increase the numbers.
- "**weathersit**" – Most favorable weather condition is the clean / few clouds days. Registered users count is comparatively high even on the light rainy days, so the assumption can be drawn that the bikes are being used for daily commute to the workplace. There is no data available for heavy rain/snow days.
- "**weekday**" – If we consider "cnt" column we do not find any significant pattern with the weekday. However if the relation is plotted with "registered" users, we observe that bike usage is higher on working days. And with "casual" users it opposite.
- "**yr**" – 2 years data is available and the increase in the bikes has increased from 2018 to 2019.
- "**holiday**" – Holiday consumption of bikes if compared within "registered" and "casual" users then the observation is "casual" users are using bikes more on holiday.
- "**mnth**" – The bike rental ratio is higher for June, July, August, September and October months.

2. Why is it important to use drop_first=True during dummy variable creation?

- The dummy variables are created by using one-hot encoding, to cover the range of values of categorical variable.
- Each dummy variable have 1 and 0 values. 1 is used to depict the presence and 0 for absence of the respective category.
- If the category variable has 3 categories, there will be 3 dummy variables.
- The drop_first = True to be used while creating dummy variables to drop the base/reference category.
- The reason for this is to avoid the multi-collinearity getting added into the model if all dummy variables are included.
- The reference category can be easily deduced where 0 is present in a single row for all the other dummy variables of a particular category.

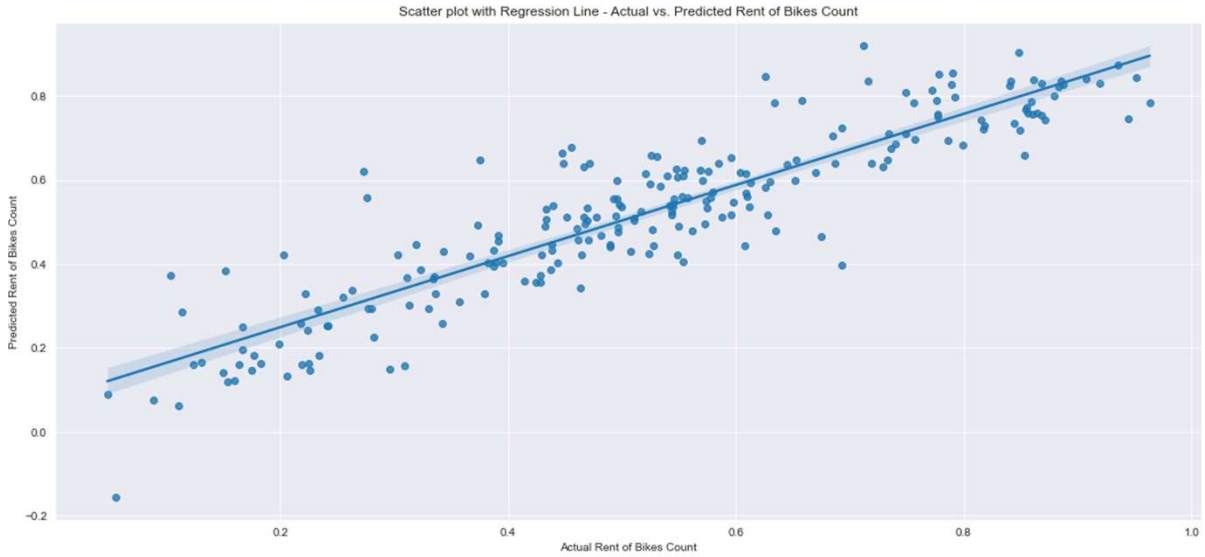
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

- “temp” is the variable which has the highest correlation with target variable i.e. 0.63.
- The casual and registered variables are actually part of the target variable as values of these columns sum up to get the target variable, hence ignoring the correlation of these 2 variables.
- “atemp” is the derived parameter from temp, humidity and windspeed, hence not considering it as it is eliminated in the model preparation.

4. How did you validate the assumptions of Linear Regression after building the model on the training set?

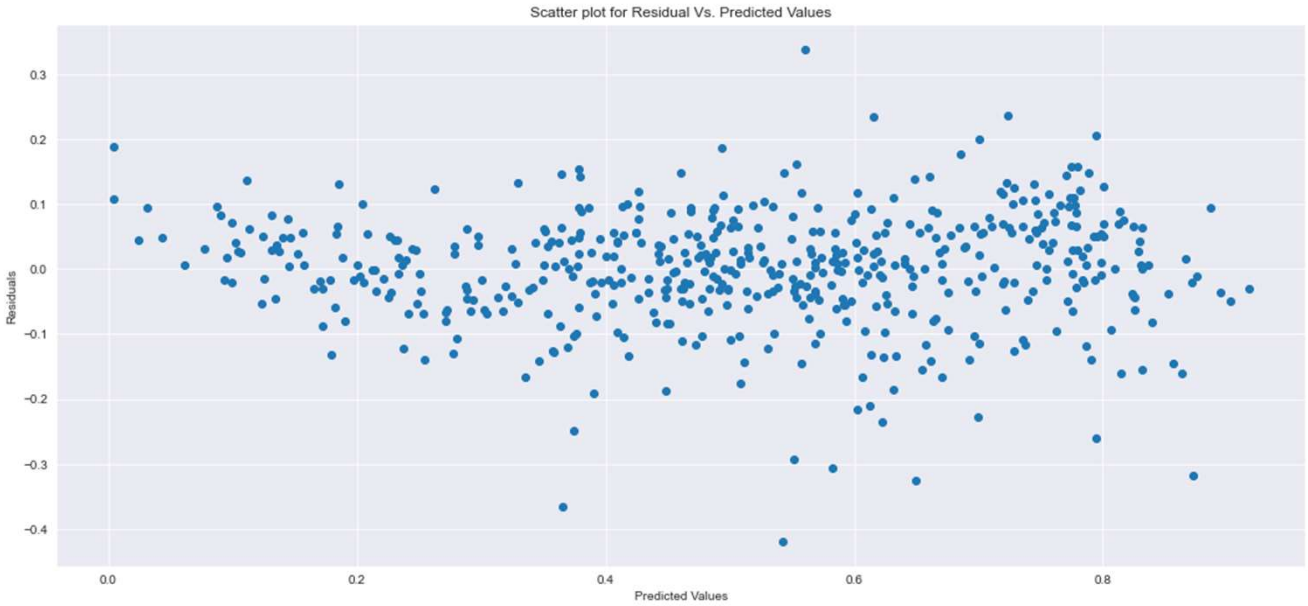
1. **Linear relationship between independent and dependent variables:**

The linearity is validated by looking at the points distributed symmetrically around the diagonal line of the actual vs predicted plot as shown in the below figure.



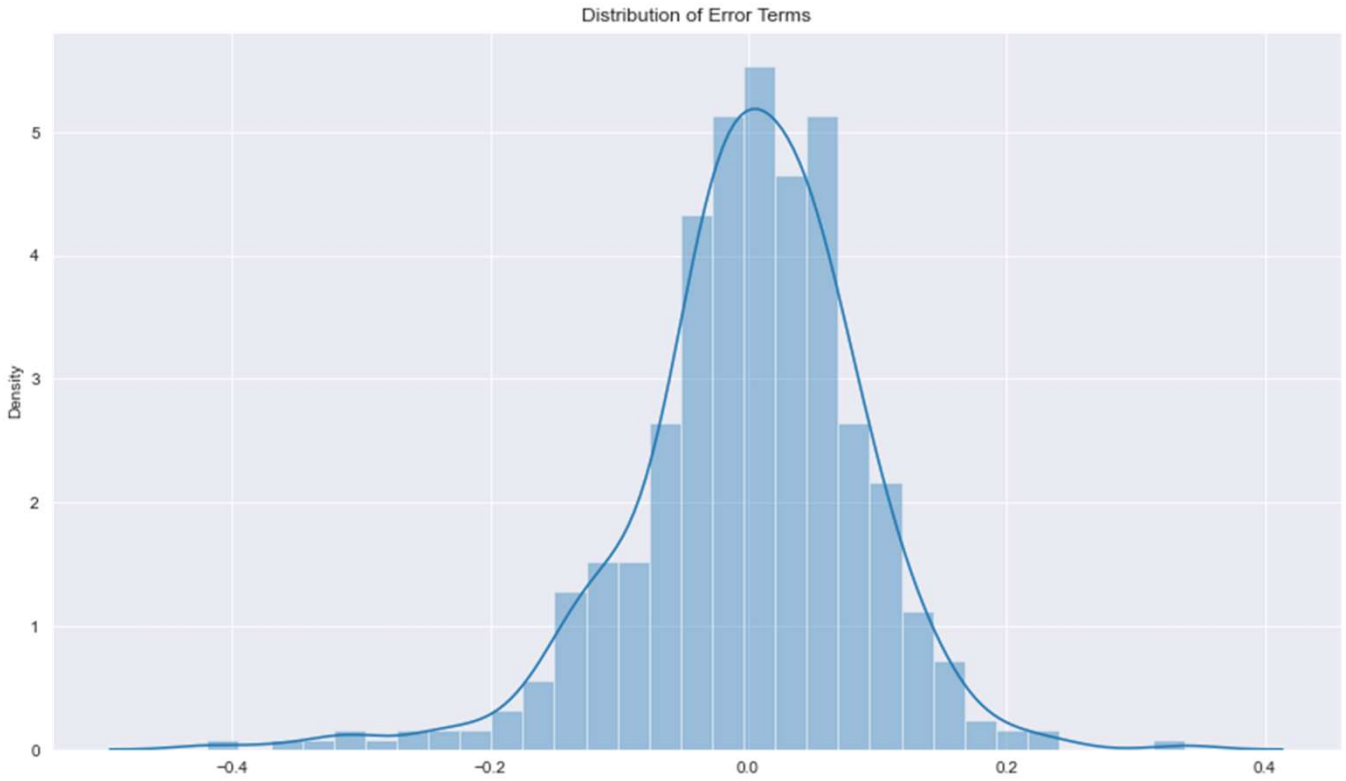
2. Error terms are independent of each other:

We can see there is no specific Pattern observed in the Error Terms with respect to Prediction, hence we can say Error terms are independent of each other



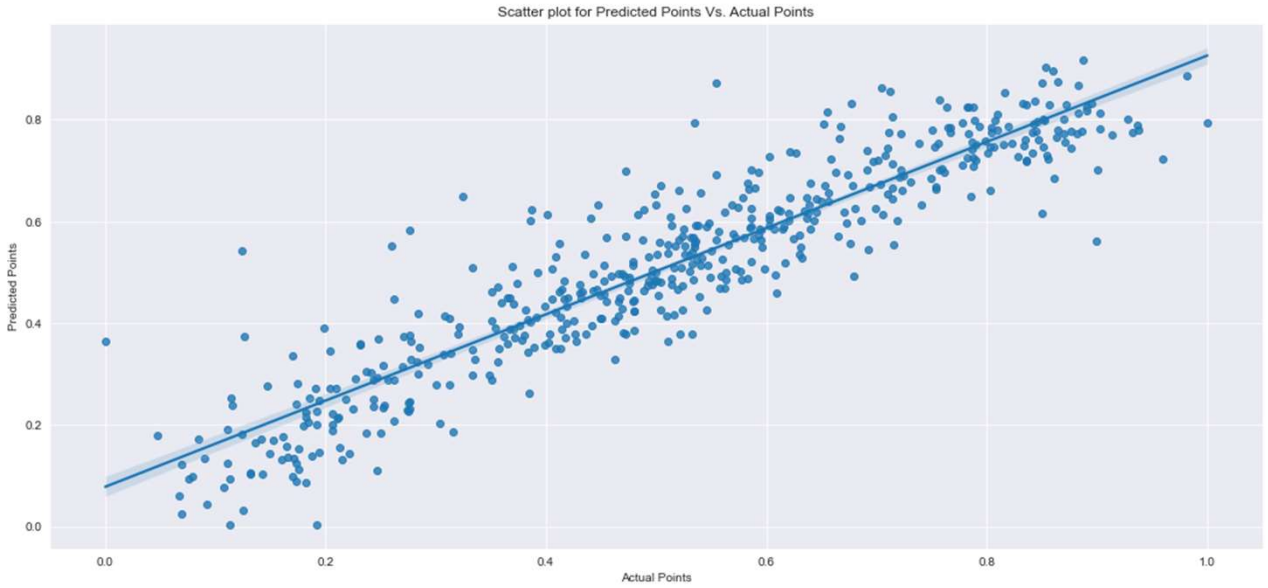
3. Error terms are normally distributed:

Histogram and distribution plot helps to understand the normal distribution of error terms along with the mean of 0. The figure below clearly depicts the same.



4. Error terms have constant variance (homoscedasticity):

We can see Error Terms have approximately a Constant Variance, hence it follows the Assumption of Homoscedasticity.



5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

- The top 3 variables are:

1. **Weathersit:** Temperature is the Most Significant Feature which affects the Business positively, Whereas the other Environmental condition such as Raining, Humidity, Windspeed and Cloudy affects the Business negatively**.**
2. **'Yr':** The growth year on year seems organic given the geological attributes.
3. **'season':** Winter season is playing the crucial role in the demand of shared bikes.

Thanks