Assignment Submission

Aniket Shambharkar

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?

Ans. Observations from above boxplots for categorical variables:

1. The season box plots indicates that more bikes are rent during fall season.
2. The working day and holiday box plots indicate that more bikes are rent during normal working days than on weekends or holidays.
3. The month box plots indicates that more bikes are rent during september month.
4. The weekday box plots indicates that more bikes are rent during saturday.
5. The weathersit box plots indicates that more bikes are rent during Clear, Few clouds, Partly cloudy weather.

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

Ans. Not using drop_first=True would make the dummy variables correlated to each other and hence, redundant, which is not expected of our analysis.

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

Ans. By looking at the pair plot temp variable has the highest (0.63) correlation with target variable 'cnt'.

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

Ans. One of the fundamental assumptions of a linear regression model is that the error terms should correspond to a normal curve, when plotted on histogram. On Out[68], we spotted the same. Hence our assumption for Linear Regression is valid.

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

Ans. The Top 3 features contributing significantly towards the demands of share bikes are:
1) weathersit_Light_Snow(negative correlation).
2) yr_2019(Positive correlation).
3) temp(Positive correlation).

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans. An interpolation technique used to predict correlation between variables and how an independent variable is influenced by the dependent variable(s), is linear regression.

After looking into the data and cleaning it with exploratory data analysis, we split the dataset into training set (which would be used to train a model) and the testing set (which would be used to check how close is our model to the actual output). After checking the collinearity of variables and using the requisite variables to train the model and checking the R-value of the model and the p-values of dependent variables, after dealing/dropping the necessary columns and reiterating the steps (feature elimination), we come to a final model.

According to the conditions of linear regression which states that the error curve must be a normal one, we proceed to testing the model with the test dataset. The conclusion hence drawn on the model would be used to provide valuable insights/predictions on datapoints in the range of the model.

2. Explain the Anscombe’s quartet in detail.

Ans. A regression model is not always necessarily an exact one, it can also be fooled by some (smart) data! In certain cases, there are multiple datasets which are completely different but after training, the regression model looks the same. A group of four such datasets having identical descriptive statistics but with some peculiarities, is the Anscombe's quartet.

3. What is Pearson’s R?

Ans. Pearson's correlation coefficient, also known as Pearson's R, is a measure of the strength of correlation between two variables. It is commonly used in linear regression. The value of Pearson's R always lie between -1 and +1, the latter indicating a perfectly positive and linear correlation and the former indicating a perfectly linear negative regression. The values in between denotes the relative collinearity of two variables.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans. Scaling is necessary for a model to be functional with the appropriate range of coefficients. For e.g., if there were two independent variables named price and months on which the sale of car depended, the price range would be far too high because there are only 12 months in a year. In that case, scaling the variable price appropriately won't allow decimal errors to happen in the model. There are two types of scaling:

- Normalized scaling: This scaling is done to make the distribution of data into a Gaussian one. It doesn't have a preset range. Typically used in Neural networks broadly.
- Standardized scaling: The example given above is of standardized scaling. Here, the values of variable(s) is/are compressed into a specific range to suit the model.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans. If there is a perfect correlation between the dependent variable and independent variable(s), the R-squared value comes out to be 1. Hence VIF, which is (1/(1-R^2)) turns out to approach infinity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans. Q-Q plot is a graphical tool to assess if sets of data come from the same statistical distribution. It is particularly helpful in linear regression when we are given testing and training datasets differently. In this scenario, it becomes important to check whether both the data comes from the same background, in order to maintain the sanity of the model.