Assignment-based Subjective Questions

Author: Swapnil Raut

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

    - The demand bike increased in the year 2019 when compared with year 2018.
    - Month Jun to Sep is the period when bike demand is high. The Month Jan is the lowest demand month.
    - Bike demand is high in working days as compared to holiday.
    - The bike demand is high when weather is clear and Few clouds however demand is less in case of Lightsnow and light rainfall.
    - Demand is high when temperatures is between 20 to 35 Degree Celsius.

2. **Why is it important to use drop_first=True during dummy variable creation?**
   drop_first=True option helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

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

   Correlation between “temp and atemp” is highest followed by Correlation between “registered and count” variables.

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

   The simple way to determine if this assumption is valid or not is by creating a scatter plot x vs y. If the data points fall on a straight line in the graph, there is a linear relationship between the dependent and the independent variables, and the assumption is valid.

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

   Below 3 features are contributing significantly for demand of the shared bikes.
    a. Temperature
    b. Season
    c. whether

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

   Linear regression is basic forms of machine learning where we train a model to predict the behaviour of your data based on some variables. In the case of linear regression as the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

   Mathematically, linear regression equation can be written as:

   $y = a + bx$

   where:-

   b = Slope of the line

   a = y-intercept of the line

   x = Independent variable from dataset

   y = Dependent variable from dataset.

2. **Explain the Anscombe's quartet in detail.**

   Anscombe's Quartet can be defined as a group of four data sets which are almost identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

3. **What is Pearson's R?**

   Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive.

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

   Scaling means transformation of data so that it will fits within a specific scale, like 0-100 or 0-1.

   Sometime collected data set contains features highly varying in magnitudes, units and range. Scaling bring all the variables to the same level of magnitude.

In normalized scaling minimum and maximum value of features are used for scaling whereas in standardized scaling mean and standard deviation is used for scaling.

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

   In case of perfect correlation, the VIF is infinite. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity.

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

   Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. The purpose of Q Q plots is to find out if two sets of data come from the same distribution.