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? (3 marks)

   Ans:

   1. If we take season categories, in the "fall" session there are maximum number of bikes has been used by people and "spring" has the lowest count.

   2. If we take Year categories, we can say that bike sharing company name as "boombike" gain significant amount of popularity from 2018 to 2019 because in highest count of 2018 is the median or 50th quartile of 2019 which is good.

   3. Month variables give similar result as season like in Aug, Sep and Oct get highest count since in these months rainfall happens.

   4. If we take the weekdays and working day variables, they have very balance impact on there each category.

   5. In the case of holidays, when there is no holidays, count is vary between 3k to 6k in most of the cases but when there is holiday count vary between 2k to 6k in most cases.

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

   Ans:

   1. By using drop_first we can drop additional variables which is created by pandas get_dummies function.
   2. E.g., there are 3 categories A,B and C ,so there dummy variables will be like this A => (1,0,0) , B =>(0,1,0) and C => (0,0,1). Means A will denote by A=1 , B=0, C=0 and B will denote by A=0 , B=1, C=0 and C will denote by A=0 , B=0, C=1.

    3. So, if we drop A using drop_first then A can be denote by B=0 and C=0. In short 1st variable can be denoted by other variables that is why we 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? (1 mark)

   Ans: Variables name as “temp” and “atemp” both have highest correlation i.e., 0.65 with target variable.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

   Ans:

   1. We focus on statical significance of the variable and VIF (variance inflation factor)

   2. That means we mainly see all P-values of all variables and their coefficient.

   3. We also look at the R^2, F-state and P(F-State) like P(F-state) should be very low or close to zero that indicate that model fit is not by chance, and R^2 should be high or closest to one.

   4. Regarding P-value it should be zero or close to zero which say variable is significant.

   5. In the case of checking VIF lower is better because VIF mainly checking multicollinearity which means predictors should not be corelated among them self or very low corelated.

   6. If we got hight P-value and High VIF straight away, we drop that variable if we got High P-value but low VIF then we drop that columns and observe the effect on other and same with low P-value and High VIF.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 mark)

   Ans: As per my final model Year(yr), temp and 3rd category of weather (weather_3 => Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds).

General Subjective Questions

Questions:

1. Explain the linear regression algorithm in detail. (4 marks)

   Ans:

   1. Linear regression is a statistical regression method used for predictive analysis and shows the relationship between the continuous variables.
   2. Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), consequently called linear regression.
   3. If there is a single input variable (x), such linear regression is called simple linear regression.
   4. And if there is more than one input variable, such linear regression is called multiple linear regression.
   5. The linear regression model gives a sloped straight line describing the relationship within the variables.
   6. To predict the y variable or dependent variable its used slop intersection formula.

      $$Y = mX + C \Rightarrow Y = a_0 + a_1X$$

      Where Y = Dependent Variable, X = Independent variable,

      a_1 = slope of line,

      a_0 = Y – axis intersection.

2. Explain the Anscombe’s quartet in detail. (3 marks)

   Ans:
   Anscombe’s Quartet can be defined as a group of four data sets which are nearly identical when we describe them in statistics, but there is some minute difference in the dataset that make regression model fools. If we plot them on a scatter plot, then we can find that they have very different distributions and appear differently.

3. What is Pearson’s R? (3 marks)

   Ans:
   1. It is the covariance of two variables, divided by the product of their standard deviations.
   2. The Pearson's correlation coefficient varies between -1 and +1
   3. It’s also known as Pearson correlation coefficient.
   4. Formula for Pearson’s R is:

   R = sum ((X_i - X_mean)* (Y_i - Y_mean))/sqrt (sum ((X_i - X_mean)^2)* sum ((Y_i - Y_mean)^2)) where X_i And Y_i are the variables values from data.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

   Ans:
   1. Scaling is a kind off pre-processing which we applied on data so that all variables should be in same range which helps algorithm to speeding in calculation.
   2. Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done, then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we must do scaling to bring all the variables to the same level of magnitude.
   3. scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

4. Normalized Scaling is also called min-max scaling. In this It shrink all of the data in the range of 0 and 1.
   X = X – Min(x) / Max(x) – Min(x)
5. Standardization Scaling brings all of the data into a standard normal distribution which has mean zero and standard deviation one.
   X = X – Min(x) / SD(X)

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

   Ans:

   1. If we see the formula for VIF is 1/1-R^2.
   2. Here R^2 denotes that how much variable is co-related to other variables.
   3. when R^2 = 1 then VIF = Infinity.
   4. That means when there is a perfect co-relation then VIF will be infinity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

   Ans:

   1. Quantile-Quantile (Q-Q) plot is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.
   2. This helps in a scenario of linear regression when we have train and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.
   3. We can use Q-Q plot to determine many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.