Assignment-based Subjective Questions

Question -: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Answer -:

- Season: Most of the bike booking were happening in fall season with a median of over 5000 booking (for the period of 2 years). This was followed by summer & winter. This indicates, season can be a good predictor for the dependent variable
- Month: Based on mean 10% of the bike booking were happening in the months May, Jun, July, August, September & October with a median of over 4000 booking per month. This indicates, months has some trend for bookings and can be a good predictor for the dependent variable.
- weathersit: Almost 46% of the bike booking were happening during 'clear weather with a median of close to 5000 booking (for the period of 2 years). This was followed by Misty with 30% of total booking. This indicates, weathersit does show some trend towards the bike bookings can be a good predictor for the dependent variable.
- holiday: Almost 98% of the bike booking were happening when it is not a holiday which means this data is clearly biased. This indicates, holiday CANNOT be a good predictor for the dependent variable.
- weekday: weekday variable shows very close trend having their independent medians between 4000 to 5000 bookings. This variable can have some or no influence towards the predictor. I will let the model decide if this needs to be added or not.
- workingday: Almost all of the bike booking were same for working and non-working days. This indicates, workingday will not be a good predictor for the dependent variable

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

Answer -: It helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

Question -: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Answer -: Temperature has the highest correlation with target variable

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

Answer -:

The simple way to determine if this assumption is met 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 holds. As we can see in the note book y –test and y_predicted are on almost straight line, so our assumptions hold good.

Question -: Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Answer -: Below are the top three feature

- Temperature (temp) - A coefficient value of '0.4486' indicated that a unit increase in temp variable increases the bike hire numbers by 0.4486 units.
- Weather Situation (Light_snowrain) - A coefficient value of '-0.2907' indicated that, w.r.t Light snow rain, a unit increase in Light_snowrain variable decreases the bike hire numbers by 0.2907 units.
- Year (yr) - A coefficient value of '0.2341' indicated that a unit increase in yr variable increases the bike hire numbers by 0.2342 units.

General Subjective Questions

Question -: Explain the linear regression algorithm in detail. (4 marks)

Answer -: Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between a dependent variable and one or more independent features. When the number of the independent feature, is 1 then it is known as Univariate Linear regression, and in the case of more than one feature, it is known as multivariate linear regression. The goal of the algorithm is to find the best linear equation that can predict the value of the dependent variable based on the independent variables. The equation provides a straight line that represents the relationship between the dependent and independent variables. The slope of the line indicates how much the dependent variable changes for a unit change in the independent variable(s)

Question -: Explain the Anscombe's quartet in detail. (3 marks)

Answer -:

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x, y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough".

Question -: What is Pearson's R? (3 marks)

Answer -:

In statistics, the Pearson correlation coefficient (PCC) is a correlation coefficient that measures linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1. As with covariance

itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationships or correlations. As a simple example, one would expect the age and height of a sample of teenagers from a high school to have a Pearson correlation coefficient significantly greater than 0, but less than 1 (as 1 would represent an unrealistically perfect correlation).

$$R= n(\sum xy) – (\sum x)(\sum y) / \sqrt{[n\sum x^2-(\sum x)^2][n\sum y^2-(\sum y)^2}$$

Question -: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Answer -:

When you have a lot of independent variables in a model, a lot of them might be on very different scales which will lead amodel with very weird coefficients that might be difficult to interpret. So we need to scale features because of two reasons:

1. Ease of interpretation

2. Faster convergence for gradient descent methods

You can scale the features using two very popular method:

1. Standardizing: The variables are scaled in such a way that their mean is zero and standard deviation is one.

2. MinMax Scaling: The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F statistic, p-values, R-square, etc.

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

Answer -: When R-square is 1, VIF is infinite. This indicates that the regression predictions perfectly fit the data.

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

Answer -:

The purpose of the quantile-quantile (QQ) plot is to show if two data sets come from the same distribution. Plotting the first data set's quantiles along the x-axis and plotting the second data set's quantiles along the y-axis is how the plot is constructed.

Q-Q plots allow data scientists and other statisticians to graphically compare two probability distributions to determine normal distribution.