Assignment-based Subjective Questions

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

Answer
One cannot tell how much each of the independent variables affects the dependent variable just by doing analysis of the categorical variables. What can be observed at that point is how the count of the total rentals is distributed around the categorical variables. After the model building,some of the categorical variables were dropped and the rest of the are selected are significant because they have a p-value of 0.00.

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

Answer
We drop the first one because we can deduce what the value is with the other dummy variables excluding the first column of the dummy variables. This reduces complexity and simplifies the model building process.

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

Answer
atemp has the highest correlation.

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

Answer
I plotted a histogram of the error terms derived from prediction on the train data set and can observe that the error terms have a normal distribution. Also, a scatter plot of the error terms and the dependent variables also does not show any clear pattern in how the error terms are distributed.

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

Answer

1. Temp
2. 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds(drizzling)
3. Yr

General Subjective Questions

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

Below are the steps:

1. Identify the dependent and independent variables.
2. Determine if there is a linear relationship between the dependent and independent variables. This means that if the equation of a straight line can be used to obtain/predict dependent variable y from independent variables X1, X2 .. Xn. This can be determinedusing pair plots to visualize the data and better understand how they are related with each other.
3. Split data into train and test data sets
4. Perform scaling and/or creations of dummy variables from categorical variables.
5. Derive a model by finding the line of best fit. The line of best fit is the one that minimizes the error obtained from a prediction to the lowest. The equation of line that produces the lowest mean squared error would be the best model.
6. Perform residual analysis and validate that there is a normal distribution of the error terms and that there is no relationship between the error tests.
7. Test the data on a test data different from those used to train the model set and compare the result of the predicted values with the actual values.

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

The concept of Anscombe's quartet expressed how four different data sets could have similar descriptive statistics such as mean and variance but vary significantly in the manner the data isdistributed. This is only seen when data is plotted in a scatter plot. This is important because regression models can easily misinterpret these kinds and data and produce a similar model for them despite the significant difference in how the data is distributed. Hence it is important to visualize the model first and after a model is obtained, confirm that the model can explain the data based on the visualization also. This way you can derive a better fit that better explains how the data is distributed.

What is Pearson's R? (3 marks)

This is a way to determine correlation coefficient between two variables. It is calculated by dividing the product of the variance between the two variables by the product of their standard deviations. There are certain assumptions that must be considered to obtain accurate value for the correlation. One of such would be to eliminate outliers. Another would be that eachdata point is independent of the other. Another is that the population correlation is zero.

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

Scaling refers to expressing your data within the range of a certain scale. For instance, between 1-100 or 0-1.

Scaling is important in multiple linear regression because it makes it easier to interpret the coefficients of the model. It allows you to be able to compare different variables on equal footing for example different currencies.

Normalization is a process that transforms your data so that it will fit into a normal distribution or better yet a transformation of data so that it forms a normal distribution curve.

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

You will obtain an infinite Variable inflation factor (VIF) if the values for which the VIF was computed are perfectly correlated. The higher the VIF the more indication of multicollinearity among the variables

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

It is also known as quantile-quantile plot, and it is used to determine if two data sets were obtained from populations with a common distribution. It is a plot of the quantiles of the first data set against that of the second data set. A 45-degree line is used as a baseline to determine likelihood that the data sets have come from populations with similar distribution or not. The farther away the data points from the line, the greater the tendency that they have been obtained from populations with different distributions.

The Q-Q plot is important when you need to know if two data set come from population with similar distribution so that you can location and scale estimates can use data from both datasets to determine a common locations and scale estimates