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)

Answer: Categorical variables are playing an important role in impacting the target variable. Holiday, weather type 2 and 3 are having a negative impact whereas year 2019, winter and months 8 to 10 are having the positive impact.

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

Answer: It is important because it is the base variable which gets explained by the other dummy variables.

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

Answer: The atemp variable has the highest correlation with the target variable.

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

Answer:

a) The residuals follow normal distribution i.e. are independent.

b) Ensuring the model is not over fitted by keeping R2 below 1.

c) Confirmed that the multicollinearity is taken care for by keeping the VIF below or around 5.

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

Answer:

a) temp has the highest impact

b) yr 2019 has big positive impact

c) Weather type 3 has the highest negative impact

General Subjective Questions

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

Answer: The linear regression algorithm assumes that there is linear relationship between a target variable and one or more independent variables. Based on the historical data it establishes a linear relationship between target and independent variable, resulting in one intercept or slope constant and other variables where their impact on the target variable is defined by Bs of the variables. And there is an error term which explains the variation between predicted value and the actual value.

The linear regression algorithm ensures that the variables remain independent and the model is not overfitted so as to create bias in the model. It calculates the significance of the variable through its B value which should be not equal to 0. The linear regression model as is defined as

Y = c+ m Xi + e where i can be any number

Y is target variable, c is intercept or constant and Xi are the independent variable and e is the error term.

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

Anscombe's quartet is a group of four data sets which would give similar descriptive statistics but when you look at the distribution of the datasets, all datasets shows different behavior. This method is important to validate the suitability of regression model to the dataset where numerical statistic may give wrong interpretation.

3. What is Pearson's R? (3 marks)

Persons's R coefficient is used to measure the correlation between to values. It's values varies from -1 to +1. Where –ve sign signifies negative correlation and +ve sign signifies positive correlation. It is a relative measure which is used to identify where with change in the value of one variable there is relative change in the value of other variable.as

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

Scaling is bringing the values of the variable in comparative scale so that the Bs of the variables can be compared to identify the significance of the variables. Normalized scaling also known as min-max method bring the values of the variables within a maximum and minimum values and is calculated by the formula (x – min(x))/(max(x) – min(x)) where as the standardized scaling method brings the values centered around 0. Standardized scaling formula is (x-mean(x))/sd(x). The key difference between methods is the under normalized scaling method the outliers are merged with the maximum values.

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

VIC is calculated using the formula 1/(1-R^2). If the R is 1 the denominator would become 0 and VIF would be become infinite.

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

Q-Q plot is used to check the relationship between the categorical variables with the target variable. If the one value of the categorical variable is moving clearly in a particular direction which means having higher or lower Q1 and Q3 or Q2 in comparison to the other values. It defines that the variable is significant for the target variable.