Gokulnath_Linear Regression Bike Assignment Questions

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?
Based on the box plot ran on categorical variables, following are the inferences -

a. Based on seasons, mean goes higher from sprint to summer and summer to fall.
b. There is no major deviations in the dependent cnt output variables on holiday, workingday, weekday
c. On weathersit, when there is Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds, we see drop in count
d. On Month, between April to October there is spike in demand
e. As the year progress, demand is increasing

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

Based on VIF calculation we drop the higher VIF factor

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

Temperature has the highest correlation factor close to 0.63

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

a. We consider the R square to be above 0.7
b. There is no major difference between R square and Adjusted R Square
c. pvalues are less to indicate significance of the independent variables
d. Residual analysis indicate a normal distribution and there is no major skewness.

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

Month, temperature and year has more significance in affecting the output variable.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear Regression algorithm helps in prediction of regression problems based on number of features impacting the output. Since it is based on a linear model we arrive at the slope and intercept that has best fit using least square method to arrive at better R square and correlation coefficient of the model. The error terms based on residual analysis should not indicate any pattern and have a normal distribution.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is used to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties.

3. What is Pearson's R?

The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables. When one variable changes, the other variable changes in the same direction

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

Normalization scales in a range of [0,1] or [-1,1]. This is also called as min max scaling. Generally we prefer this as all the data is available in the similar scale.
Standardized scaling is the subtraction of the mean and then dividing by its standard deviation

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

VIF infinite indicate R square is close to 1 which indicates that a particular feature has very high correlation with all the other features. It is advised to drop those high VIF to get better coefficients.

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 a graphical plotting of the quantiles of two distributions with respect to each other