Subjective 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? (3 marks)

Ans: By analysing the P-values and VIF, I can see that 'yr', 'holiday' and 'weekday' have the most impact on the dependent variable. Other variables like Season and MNTH seemed to also have high impact, however, they had very high VIF and P-Value respectively

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

Ans: Generally, for a categorical variable with K levels, we need to have K-1 variables to represent it via dummies. drop_first=True accomplishes this

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

Ans: atemp and temp were almost the same in terms of positive correlation with the target.

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

Ans: I analysed the residuals and visualized them on a distribution plot. I wanted to see if the residuals followed a normal distribution (zero mean). Visually I saw that the residuals had normal distribution

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

Ans: temp, yr and windspeed are the top 3 features explaining the demand of the shared bikes. This is based on their coefficient values

General Subjective Questions

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

Ans: It is an ML algorithm that tries to fit a line on the provided inputs. The best fit line is given by the one with the least residual sum of squares. Given this objective, the coefficients for each variable along with the constant (bias value) is adjusted incrementally using gradient descent

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

Ans: It is a specialized dataset consisting of 4 data. Each of these have the same descriptive statistical values, such as mean, variance, correlation, etc. However, they have very different characteristics when we graph each of the datasets. This is used to show the significance of visualization and was first formulated by Francis Anscombe in 1973

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

Ans: Also known as Pearson correlation coefficient, it is a measure of linear correlation between two variables. It is given by the ratio of the covariance of the two variables and the product of their stdeviations. The value can range between -1 and 1. Values greater than zero indicate positive correlation and values less than zero indicate negative correlation

4. What is scaling? Why is scaling performed?

Ans: Scaling is performed mainly for interpretability. By scaling the values of variables to a certain range, we stabilize the values of the coefficients, which leads to better interpretability. Scaling also helps with faster convergence of the gradient descent optimization

5. What is the difference between normalized scaling and standardized scaling? (3 marks)

Ans: normalized scaling squishes the values to be between 0 and 1. Standardized scaling transforms to values such that they have zero mean and unit variance. Both are used to scale values and can be used as part of data preparation. Normalized scaling is generally preferred as it takes care of outliers, however, standardized scaling may be preferred for some datasets

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

Ans: This can happen when R-Squared for that variable is exactly 1. Which means, that variable is 100% explained by all the other variables. Such a variable certainly needs to be dropped

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

Ans: Stands for Quantile-Quantile plot. T is used to ascertain whether 2 datasets come from the same distribution. If they do, then in the plot that will approximately fall on the line Y=X. In linear regression, we can use Q-Q pots to ascertain whether the residuals are normally distributed. We can also use Q-Q plots to check if the actual target values and the predicted values belong to the same distribution. This can be a visual measure of the goodness of fit