Bike Sharing Assignment

Assignment-based Subjective Questions

Q1.) From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

A1.) The observations made from analysis of categorical variables:

- Bike demand is observed to be lower in spring than other seasons, while it is higher from June to September and lowest in January.
- The year 2019 shows a higher demand for bikes compared to the year 2018.
- The demand for bikes is relatively lower during holidays than non-holidays.
- There is no significant difference in bike demand between weekdays and weekends or working and non-working days.

Q2.) Why is it important to use **drop_first=True** during dummy variable creation?

A2.) It is important to use **drop_first=True** during dummy variable creation to avoid the issue of multicollinearity in the dataset.

Q3.) Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

A3.) The plot shows that the temperature variable has the strongest correlation (0.63) with the target variable 'cnt'.

Q4.) How did you validate the assumptions of Linear Regression after building the model on the training set?

A4.) One common method of residual analysis is to plot the residuals against the predicted values to check for normal distribution with constant variance. This helps to ensure that the assumptions of linearity, homoscedasticity, and normality of residuals are met. Therefore, To validate the assumptions of linear regression, a graph of the residuals (Y - Y predicted values) was plotted to analyze the distribution of the error terms.

Q5.) Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

A5.) The final model's top three contributing features, ranked by their variable coefficients, are: weathersit_Light Snow (-1.3520), yr (1.0329), and season_spring (-0.6009).

Bike Sharing Assignment

General Subjective Questions

Q1.) Explain the linear regression algorithm in detail.

A1.) Linear regression is a statistical algorithm used to model the relationship between a dependent variable (target) and one or more independent variables (predictors). The goal of linear regression is to find a linear equation that describes the relationship between the variables. The equation takes the form:

$y = \beta0 + \beta1x1 + \beta2x2 + ... + \beta n*xn$

where y is the dependent variable, x1, x2, ..., xn are the independent variables, β0 is the intercept, and β1, β2, ..., βn are the coefficients that describe the relationship between the independent variables and the dependent variable.

Q2.) Explain the Anscombe's quartet in detail.

A2.) Anscombe's quartet is a set of four datasets, each consisting of eleven (x, y) points. These datasets have nearly identical statistical properties, but they look quite different when graphed. The four datasets have the same mean and variance for both the x and y variables, the same correlation coefficient, and the same linear regression line. However, when plotted, they show vastly different patterns, highlighting the importance of visualizing data.

The first dataset appears to have a linear relationship between x and y, and a strong correlation. The second dataset also appears to have a linear relationship, but with a single outlier that has a large effect on the correlation coefficient. The third dataset has a non-linear relationship, but still has the same statistical properties as the first two datasets. Finally, the fourth dataset has a clear non-linear relationship, but with no correlation.

Q3.) What is Pearson's R?

A3.) Pearson's R is a measure of the linear correlation between two continuous variables, ranging from -1 (perfect negative correlation) to +1 (perfect positive correlation), with 0 indicating no correlation. It is commonly used to measure the strength and direction of the relationship between two variables in statistics and data analysis.

Q4.) What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

A4.) Scaling is the process of transforming variables to a standard scale, usually between 0 and 1, or with a mean of 0 and a standard deviation of 1. Scaling is performed to standardize variables so that they have the same range of values, making it easier to compare and interpret their effects in a statistical model.

Normalized scaling scales the data so that it has a range between 0 and 1, based on the minimum and maximum values of the variable. This type of scaling is useful when the distribution of the variable is not normal, and we want to ensure that all values are between 0 and 1.

Standardized scaling scales the data so that it has a mean of 0 and a standard deviation of 1, based on the distribution of the variable. This type of scaling is useful when we want to compare the relative importance of different variables in a statistical model. Standardized scaling is also useful when the data is normally distributed.

Q5.) You might have observed that sometimes the value of VIF is infinite. Why does this happen?

A5.) The VIF (Variance Inflation Factor) is a measure of multicollinearity among the predictor variables in a regression model. The VIF is calculated by dividing the variance of the estimated regression coefficient for a particular predictor variable by the variance of that predictor variable.

Sometimes, the value of VIF can be infinite, which happens when the variance of a predictor variable is zero. This can occur when all of the values for a particular predictor variable are the same, such as when a categorical variable has only one level. In this case, the VIF cannot be calculated using the usual formula, and it is reported as infinite.

Q6.) What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

A6.) A Q-Q plot, short for quantile-quantile plot, is a graphical technique used to compare two probability distributions. It plots the quantiles of one distribution against the corresponding quantiles of another distribution. In linear regression, Q-Q plots are used to check if the residuals of the model follow a normal distribution. If the residuals follow a normal distribution, then it suggests that the model is well-fitted and accurate.

The importance of Q-Q plots in linear regression is that they help to validate the assumptions of the model. Specifically, a Q-Q plot is used to validate the assumption of normality of residuals. If the residuals are normally distributed, then the model is more likely to be accurate and reliable. On the other hand, if the residuals do not follow a normal distribution, it may indicate that there are issues with the model, such as non-linearity, heteroscedasticity, or outliers.