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?

Ans. Year has a positive effect on the number of bike hires. With an increase in year, the number of bikes increases by 0.235519 units. 'Holiday' variable has a negative effect on number of bike hires. People tend to hire less bikes on holidays. In spring season, people hire less bikes and in winter season, people purchase more bikes. In September, people hire more bikes. In cloudy weather with light rain, people tend to hire less bikes.

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

Ans. drop_first=True is important to use, as it helps in reducing the extra columns that gets created during dummy variable creation. Hence it reduces the correlations created among dummy variables. Dropping the first columns as (p-1) dummies can explain p categories. In weathersit, first column was not dropped so as not to lose the info about severe weather situation.

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

Ans. 'Temp' and 'Atemp' has the highest correlation (0.65) with 'cnt' variable.

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

Ans. Residual Analysis: Errors are normally distributed with a mean of 0. Actual and predicted result follow the same pattern. The error terms are independent of each other.

R2 value for test predictions: R2 value for predictions on test data (0.80) is almost same as R2 value of train data(0.826). This is a good R-squared value, hence we can see our model is performing good even on unseen data (test data)

Homoscedacity: We can observe that variance of the residuals (error terms) is constant across predictions. i.e., error term does not vary much as the value of the predictor variable changes.

Plot Test vs Predicted value test: The prediction for test data is very close to actuals.

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

Ans. As per our final Model, the top 3 predictor variables that influences the bike booking are: temp, weathersit_Light Rain, yr. A unit increase in temp (Temperature) variable increase the bike hire numbers by 0.405976 units. A unit increase in weathersit_LightRain decrease the bike hire numbers by 0.287496 units. A unit increase in yr (Year) variable increase the bike hire numbers by 0.235519 units.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans. Linear regression is a supervised machine learning algorithm used for predicting a continuous numerical variable based on one or more independent variables. It establishes a linear relationship between the input features and the target variable. The goal of linear regression is to find the best-fitting plane that minimizes the difference between the predicted values and the actual data points. There are mainly two types of linear regression that are single and multiple linear regression.

In single linear regression, target variable is predicted based on only one dependent variable. But in case of multiple linear regression, target variable is being predicted using various dependent variables.

Key steps in linear regression model are data collection, data processing, model building, model testing and prediction using the final model. Linear regression may not perform well when the relationship is not truly linear or when there are non-linear interactions between variables, in which case more complex models like polynomial regression or other machine learning algorithms may be more appropriate.

2. Explain the Anscombe's quartet in detail.

Ans. Anscombe's quartet is a statistical demonstration that highlights the importance of data visualization and the limitations of relying solely on summary statistics. It consists of four datasets, each containing 11 data points, and when we calculate basic statistical properties (such as mean, variance, correlation, and regression parameters), they appear almost identical. However, when you plot the data, you can see that the datasets are significantly different. This paradoxical example was created by the British statistician Francis Anscombe in 1973 to emphasize the need for visual exploration of data and not just relying on statistical measures.

3. What is Pearson's R?

Ans. Pearson's correlation coefficient, often denoted as "r" or "Pearson's r," is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It was developed by Karl Pearson in the early 20th century and is widely used in statistics and data analysis. Pearson's correlation coefficient can take values between -1 and 1.

Pearson's correlation coefficient is a useful tool in statistics, as it helps assess the linear association between two variables. However, it assumes that the relationship is linear, and it may not capture non-linear relationships.

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

Ans. Scaling is the process of transforming the numerical values of features (variables) in a dataset to fit within a specific range or distribution. The primary reasons for performing scaling are to ensure that all features have a similar scale and to make the modeling process more effective. Scaling is particularly important for machine learning algorithms that are sensitive to the magnitude of features, such as gradient-based optimization algorithms (e.g., gradient descent) and distance-based algorithms (e.g., k-nearest neighbors or support vector machines).

Normalized scaling (min-max scaling) transforms data to a specified range, typically [0, 1], preserving the relative relationships between data points. It is sensitive to outliers as it uses the minimum and maximum values.

Standardized scaling (z-score scaling) rescales data to have a mean of 0 and a standard deviation of 1. It is less affected by outliers and makes it easier to compare the importance of different features, but it does not guarantee a specific range for the scaled data. The choice between the two depends on the specific requirements of data and analysis.

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

Ans. A VIF of infinity occurs when there is perfect multicollinearity among the predictor variables. It means that the two variables are so much correlated that it is very much needed to remove one variable from the model.

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

Ans. A Q-Q (Quantile-Quantile) plot is a graphical tool used in statistics to assess whether a dataset follows a specific theoretical distribution, typically the normal distribution. It is a visual representation of how the quantiles (percentiles) of the data compare to the quantiles of the theoretical distribution. The x-axis of the Q-Q plot represents the quantiles of the theoretical distribution, while the y-axis represents the quantiles of the observed data.

A Q-Q plot is a valuable diagnostic tool in linear regression for assessing normality, identifying outliers, and validating the distributional assumptions underlying the model, helping to ensure the reliability and validity of regression analysis.