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 - The categorical variables gives an informative understanding of how likely people are to use the shared bike , as can see per analysis we can see that if the temperature increase people favour to use the shared bike and as humidity increase people favour less to use shared bike and which season and month also we can see there is a huge demand.

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

Ans - Using drop first = True we use it to avoid multicollinearity , it improves model efficiency by reducing variables to n-1 category ,as one category becomes the reference baseline

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

Ans - The highest correlation with target variable is Registered

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

Ans – I have used the r2_score module to validate the testing performance on the test data after fitting the model on training data and also checked the root mean squared error and mean absolute percentage error to validate the assumptions of linear regression

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

Ans – Based on my model there 9 features contributing significantly towards explaining the demand of the shared bikes and they are temp, hum, windspeed, season_2, season_3, season_4, yr_1, mnth_9, weathersit_3

General Subjective Questions

1) Explain the linear regression algorithm in detail.

Ans - The linear regression method is used to create a representation of the connection, between a dependent variable (y) and one or more independent variables (x) using a straight line or hyperplane equation. By calculating coefficients that minimize the squared differences between predicted and actual values this approach helps in predicting and analyzing data assuming a linear relationship, between variables and specific statistical assumptions.

2) Explain the Anscombe’s quartet in detail

Ans - Anscombe's quartet consists of four small datasets that possess almost identical simple descriptive statistics, such as the mean, variance, and correlation. However, when these datasets are graphed, they exhibit distinctly different relationships. This underscores the significance of data visualization as a means to comprehend data. The quartet was devised by statistician Francis Anscombe to illustrate the limitations of relying solely on summary statistics for data understanding and making inferences. It serves as a compelling reminder that exploring and visualizing data is always necessary to unveil concealed patterns and relationships.

3)What is Pearson’s R?

Ans - The Pearson correlation coefficient, also known as Pearson's R, is a statistical measure that assesses the linear correlation or relationship between two continuous variables. It assigns a value between -1 and 1, where -1 signifies a strong negative linear relationship, 1 signifies a strong positive linear relationship, and 0 indicates no linear relationship. Pearson's R is extensively utilized in statistics to quantify the magnitude and direction of the connection between variables, making it an essential tool for comprehending data relationships and making predictions.

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

Ans - Scaling is the process of transforming data values to a common scale or range. This is done to allow for comparison or use of variables with different units and ranges in data analysis and machine learning models. There are two common methods of scaling: normalized scaling and standardized scaling.Normalized scaling, also known as Min-Max scaling, maps data to a range between 0 and 1. This preserves the relative relationships between data points while bringing them onto a common scale. It is useful when the exact values of the data points are less important than their relative positions.Standardized scaling, also known as Z-score scaling, transforms data to have a mean of 0 and a standard deviation of 1. This centers the data around zero and makes it more robust to outliers. Standardized scaling is often used when the absolute values and distributions of the data are important. The choice between normalized scaling and standardized scaling depends on the specific requirements and characteristics of the data and the modeling technique being used.

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

Ans - When there is perfect multicollinearity in a dataset, an infinite Variance Inflation Factor (VIF) occurs. Perfect multicollinearity happens when one or more independent variables can be accurately predicted from the others. This leads to issues in linear regression modeling because the VIF, which is used to measure the correlation between variables, becomes undefined. The calculation of VIF involves matrix inversion, which is not possible when perfect multicollinearity exists. To address this problem, highly correlated variables should be removed or combined before conducting a regression analysis.

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

Ans - A Q-Q plot, short for Quantile-Quantile plot, is a graphical tool used in statistics to compare the distribution of a dataset with a theoretical distribution. It is commonly used to evaluate the normality assumption in linear regression. The plot compares the quantiles of the residuals of a regression model to the quantiles of a normal distribution. By visually examining the plot, we can determine if the residuals follow a normal distribution. Ideally, the points on the plot will fall along a straight line, indicating that the residuals are normally distributed. Any deviations from this straight line suggest departures from normality, which can affect the reliability of the regression model's predictions. Therefore, the Q-Q plot is an important diagnostic tool in ensuring that the assumptions of linear regression are met.

Jacob Asir