Linear Regression Subjective Questions-Answers

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?

Answer:

Categorical variables like season and weathersit have a significant impact on the bike demand (cnt). For example:

season influences bike demand as people tend to rent more bikes during favorable seasons like summer and fall compared to winter.

weathersit affects demand because adverse weather conditions such as heavy rain or snow reduce the likelihood of bike rentals.

These variables capture seasonal and weather-related variations in demand, making them important predictors.

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

Answer:
Using drop_first=True eliminates one category from the dummy variables, effectively reducing multicollinearity in the dataset. Multicollinearity occurs when one dummy variable can be perfectly predicted by the others (perfectly correlated). By dropping the first category, we prevent the "dummy variable trap" while retaining the necessary information.

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

Answer:
From the pair-plot and the correlation matrix:

registered has the highest positive correlation with cnt. This indicates that the number of registered users strongly drives total bike demand. (Note: While registered is dropped in the model as a subset of cnt, it is evident in exploratory analysis.)

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

Answer:
The following steps validated the assumptions:

Linearity: Checked the scatter plot of residuals to ensure no distinct pattern existed, indicating linearity.

Normality of Residuals: Plotted a residual histogram (or Q-Q plot) to confirm residuals followed a normal distribution.

Homoscedasticity: Verified constant variance of residuals using residual scatter plots.

Multicollinearity: Checked the Variance Inflation Factor (VIF) values to ensure predictors weren't highly collinear.

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

Answer:
From the model coefficients:

Temperature (temp): Higher temperatures increase bike demand significantly.

Year (yr): Bike-sharing systems gain popularity over time, with 2019 showing higher demand compared to 2018.

Humidity (hum): Lower humidity positively correlates with higher bike demand.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer:

Linear regression is a supervised learning algorithm used to model the relationship between a dependent variable (`y`) and one or more independent variables (`X`) by fitting a straight line:

- The line is defined by: $$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$
- **Key Steps**:
    - Identify predictors (independent variables) and the target variable (dependent).
    - Train the model using ordinary least squares (OLS) to minimize the sum of squared residuals.
    - Make predictions and evaluate model accuracy using metrics like R-squared. Linear regression works best when assumptions such as linearity, independence, normality, and homoscedasticity of residuals are met.

2. Explain the Anscombe's quartet in detail.

Answer:

Anscombe's quartet consists of four datasets that share similar statistical properties (mean, variance, correlation) but differ significantly when visualized. It demonstrates:

1. The importance of visualizing data before drawing conclusions.

2. How distinct patterns and relationships can be hidden behind identical summary statistics.

3. What is Pearson's R?

Answer:

Pearson's R (correlation coefficient) measures the linear relationship between two continuous variables:

- **Range:** -1 to +1
- A value of +1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 indicates no linear relationship. It's widely used to understand how one variable predicts the behavior of another.

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

Answer:

Scaling is the process of transforming data to bring all features onto a similar scale.

- **Why Perform Scaling?**
  - Features with larger magnitudes can dominate the model, leading to biased predictions.
  - Scaling ensures effective model convergence (especially for models like linear regression and SVM).
- **Differences**:
  - **Normalization:** Scales data to a range [0,1] or [-1,1].
  - **Standardization:** Centers the data around 0 with a standard deviation of 1.

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

Answer:

Infinite VIF values occur when perfect multicollinearity exists in the data, i.e., one predictor is a perfect linear combination of others. To fix this:

- Remove redundant variables.
- Use dimensionality reduction techniques.

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

Answer:

A Q-Q (Quantile-Quantile) plot compares the distribution of residuals to a normal distribution. It's used to:

1. Validate the assumption of normality for residuals.

2. Identify outliers or deviations from normality. The closer the points lie to the diagonal, the more normal the distribution of residuals.