Bike Sharing Assignment

Subjective Questions Submission

Name: Mohammed Rizwan Shaik

Assignment-based Subjective Questions

Question 1: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer:

From the analysis of the categorical variables such as \`season\`, \`weathersit\`, \`yr\`, \`holiday\`, and \`workingday\`, we can infer the following effects on the dependent variable (\`cnt\`):

1. Season: The demand for shared bikes varies significantly across different seasons. For example, bike rentals are generally higher in the fall and summer compared to spring and winter, likely due to favorable weather conditions.
2. Weathersit: Weather conditions play a crucial role in bike rentals. Clear weather (weathersit=1) is associated with higher rentals, while adverse weather conditions like mist or snow (weathersit=2 or 3) lead to lower rentals.
3. Year (yr): The demand for bike rentals has increased from 2018 to 2019, indicating a growing trend in the use of shared bikes.
4. Holiday: The effect of holidays on bike rentals is less pronounced, but there tends to be a slight decrease in demand on holidays compared to non-holidays.
5. Workingday: There is a higher demand for bike rentals on working days compared to weekends and holidays.

Question 2: Why is it important to use \`drop_first=True\` during dummy variable creation?

Answer:

Using \`drop_first=True\` during dummy variable creation is important to avoid the dummy variable trap, which occurs when the dummy variables are highly correlated (perfect multicollinearity). By dropping the first category, we ensure that the regression model does not face issues due to this multicollinearity, making the model more stable and interpretable.

Question 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 analysis, `atemp` (feeling temperature) has the highest correlation with the target variable `cnt` (total bike rentals), indicating that as the perceived temperature increases, the number of bike rentals also tends to increase.

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

Answer:

To validate the assumptions of Linear Regression, the following steps were taken:

1. Linearity: Checked the scatter plots between the predictors and the target variable to ensure a linear relationship.
2. Independence: Ensured the independence of residuals by plotting residuals versus fitted values and checking for any patterns.
3. Homoscedasticity: Plotted residuals versus fitted values to ensure constant variance (homoscedasticity).
4. Normality: Used a Q-Q plot to check if the residuals followed a normal distribution.
5. Multicollinearity: Checked Variance Inflation Factor (VIF) values to detect multicollinearity among the predictors.

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

Answer:

Based on the final model, the top 3 features contributing significantly towards explaining the demand for shared bikes are:

1. `yr` (year)
2. `atemp` (feeling temperature)
3. `season_fall`

General Subjective Questions

Question 1: Explain the linear regression algorithm in detail.

Answer:

Linear regression is a fundamental statistical technique 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 the best-fitting straight line (or hyperplane in higher dimensions) that describes how the dependent variable changes as the independent variables change.

Types of Linear Regression:

1. Simple Linear Regression: Involves one dependent variable and one independent variable.
2. Multiple Linear Regression: Involves one dependent variable and multiple independent variables.

The equation of a simple linear regression is:

y = beta_0 + beta_1*x + error term

Where:

y is the dependent variable.

x is the independent variable.

beta_0 is the intercept.

beta_1 is the slope (coefficient).

In multiple linear regression, the model includes multiple predictors:

y = beta_0 + beta_1*x_1 + beta_2*x_2 + ... + beta_p*x_p + error term

The objective is to minimize the sum of the squared differences between the observed values and the values predicted by the linear model, using methods like Ordinary Least Squares (OLS).

Question 2: Explain the Anscombe's quartet in detail.

Answer:

Anscombe's quartet consists of four datasets that have nearly identical simple descriptive statistics (mean, variance, correlation, etc.), yet they are very different when graphed. This demonstrates the importance of visualizing data before analyzing it statistically. Each dataset in the quartet has:

1. The same mean for both x and y.
2. The same variance for both x and y.
3. The same correlation between x and y.
4. The same linear regression line.

Despite these similarities, the scatter plots of these datasets reveal distinct patterns (linear, non-linear, outliers, etc.), highlighting the potential pitfalls of relying solely on summary statistics.

Question 3: What is Pearson's R?

Answer:

Pearson's R, or Pearson's correlation coefficient, measures the linear relationship between two variables. It ranges from -1 to 1, where:

- 1 indicates a perfect positive linear relationship.
- -1 indicates a perfect negative linear relationship.
- 0 indicates no linear relationship.

Pearson's R is calculated as:

$$r = \frac{\sum \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum \left(x_i - \bar{x}\right)^2 \sum \left(y_i - \bar{y}\right)^2}}$$

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

Answer:

Scaling is a data preprocessing technique used to adjust the range of feature values in a dataset to a common scale. This is particularly important when the range of values across different features varies significantly. Scaling helps in standardizing the magnitude of the values of features, ensuring that each feature contributes proportionately to the distance calculations and the model's performance. Scaling is performed for several reasons:

1. Improves Algorithm Performance: Many machine learning algorithms, such as gradient descent-based algorithms, K-Nearest Neighbors (KNN), and Support Vector Machines (SVM), perform better and converge faster when the data is scaled. This is because these algorithms are sensitive to the magnitude of the input features.
2. Prevents Dominance of Features: When features have different scales, those with larger ranges may dominate the learning process, leading to biased models. Scaling ensures that all features contribute equally to the model.
3. Enhances Model Interpretability: Scaling can make the interpretation of coefficients in linear models easier and more meaningful, as it standardizes the units of measurement.
4. Improves Numerical Stability: Scaling can improve the numerical stability of algorithms, particularly those involving matrix operations, by reducing the chances of encountering extremely large or small values.

Aspect	Normalized Scaling (Min-Max Scaling)	Standardized Scaling (Z-score Scaling)
Definition	Adjusts data to a fixed range, typically [0, 1]	Adjusts data to have a mean of 0 and a standard deviation of 1
Parameters Used	Minimum and maximum values of the dataset	Mean (average) and standard deviation of the dataset
Range of Transformed Data	Typically [0, 1]	No fixed range (values can be positive or negative)
Use Case	Suitable for data that does not follow a normal distribution and when a specific range is needed	Suitable for data that follows a normal distribution or when normality is assumed by the algorithm
Algorithms Benefited	Beneficial for algorithms like neural networks and image processing	Beneficial for algorithms like linear regression, logistic regression, and PCA

Impact on Data	Rescales features to a specified range	Centers data around the mean and scales it according to the standard deviation
Effect of Outliers	Highly sensitive to outliers as they affect the min and max values	Less sensitive to outliers as it standardizes based on the mean and standard deviation

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

Answer:

The value of VIF (Variance Inflation Factor) becomes infinite when there is perfect multicollinearity among the predictors. This means that one predictor can be expressed as an exact linear combination of one or more other predictors, leading to a determinant of zero in the matrix inversion process used to calculate VIF.

Question 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 is a graphical tool to compare the distribution of a dataset to a theoretical distribution (typically the normal distribution). It plots the quantiles of the data against the quantiles of the theoretical distribution.

In linear regression, a Q-Q plot of the residuals is used to check the normality assumption. If the residuals are normally distributed, the points on the Q-Q plot will lie approximately along a straight line. Deviations from this line indicate departures from normality, which can affect the validity of hypothesis tests and confidence intervals.