Bike Sharing Assignment

Auditya Bhattaram

Assignment-based Subjective Questions

1. Effect of Categorical Variables on the Dependent Variable

The dataset includes several categorical variables such as season, yr, mnth, holiday, weekday, workingday, and weathersit. These categorical variables can significantly impact the dependent variable cnt - count of bike rentals

- Season can affect bike rentals due to weather conditions.

- Bike rentals might be higher or lower on holidays.

- Weekdays vs weekends can show different rental patterns.

- Workingday differentiates between working days and weekends/holidays.

- Weather conditions (clear, cloudy, rainy) directly impact bike usage. Weathersit variable

- Year (yr) & Month (mnth) can indicate changes in bike-sharing popularity over time & can infer seasons from months

2. Importance of drop_first=True in Dummy Variable Creation

When creating dummy variables for categorical features, using drop_first=True is important to avoid the dummy variable trap. By dropping the first category, we avoid multicollinearity and ensure the model remains interpretable and stable. Otherwise one feature closely defines the other feature as well thereby increasing multi coleniarity

For m levels of a feature we should be creating m-1 dummy variables.

3. Highest Correlation with the Target Variable

To determine which numerical variable has the highest correlation with the target variable (cnt), we would typically look at a pair-plot or correlation matrix. Based on common trends in bike-sharing datasets:

temp is showing a high positive correlation with bike rentals, as more people tend to rent bikes in favorable weather conditions.

4. Validating Assumptions of Linear Regression

Checked scatter plots of residuals vs. predicted values to ensure no patterns.

Ensured residuals are independent (Durbin-Watson test).

Plot residuals vs. predicted values to check for constant variance to check that there is no Homoscedasticity

Used Q-Q plots or histograms of residuals to check for normal distribution.

5. Top 3 Features Contributing to Bike Demand

From the final equation cnt = 0.183881 - 0.059810season_spring + 0.049720season_summer + 0.071669season_winter - 0.055068mnth_jul + 0.067145mnth_sept + 0.018564weekday_sat - 0.066934weathersit_moderate + 0.239401yr - 0.080819holiday + 0.504532temp - 0.179093*windspeed
temp, year & season are contributing to Bike demand

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a predictive modeling technique that identifies the relationship between the dependent variable (target) and independent variables (predictors). It demonstrates a linear relationship, indicating how the dependent variable's value changes with the independent variable's value. When there is a single input variable (x), it is termed simple linear regression. When there are multiple input variables, it is known as multiple linear regression. The linear regression model produces a sloped straight line that describes the relationship between the variables.

A regression line can exhibit either a Positive Linear Relationship or a Negative Linear Relationship. The objective of the linear regression algorithm is to determine the optimal values for a0 and a1 to find the best-fit line, which should have the least error. In Linear Regression, techniques like Recursive Feature Elimination (RFE) or Mean Squared Error (MSE) or cost function are used to identify the best possible values for a0 and a1, ensuring the best fit line for the data points

2. Explain the Anscombe's quartet in detail.

Anscombe's Quartet consists of four datasets that are nearly identical in simple descriptive statistics, but each has unique characteristics that can mislead a regression model. Despite having similar statistical properties, these datasets have very different distributions and look distinct when plotted on scatter plots. The quartet was created to emphasize the importance of graphing data before analysis and model building, and to show how other observations can affect statistical properties. Each of the four datasets has nearly the same statistical observations, including the variance and mean of all x and y points.

The first dataset fits a linear regression model, indicating a linear relationship between X and Y.

The second dataset does not show a linear relationship between X and Y, making it unsuitable for a linear regression model.

The third dataset contains outliers that cannot be handled by a linear regression model.

The fourth dataset has a high leverage point, resulting in a high correlation coefficient.

The conclusion is that regression algorithms can be deceived, so it is crucial to visualize data before building a machine learning model.

3. What is Pearson's R?

In statistics, Pearson's Correlation Coefficient, also known as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation, is a measure that quantifies the linear relationship between two variables.

Pearson's r quantifies the strength and direction of the linear relationship between two continuous variables. It ranges from -1 to 1:

+1 indicates a perfect positive linear relationship.

-1 indicates a perfect negative linear relationship.

0 indicates no linear relationship.

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

and standardized scaling?

Scaling involves transforming your data to fit within a specific range. It is a data pre-processing step that helps speed up calculations in an algorithm. Collected data often contains features with varying magnitudes, units, and ranges. Without scaling, algorithms may give undue weight to features with larger magnitudes, leading to incorrect modeling.

Differences between Normalizing Scaling and Standardizing Scaling:

1. Normalized Scaling: Uses the minimum and maximum values of features.

   Standardized Scaling: Uses the mean and standard deviation for scaling.

2. Normalized Scaling: Applied when features are on different scales.

   Standardized Scaling: Ensures zero mean and unit standard deviation.

3. Normalized Scaling: Scales values between (0,1) or (-1,1).

   Standardized Scaling: Values are not bounded within a specific range.

4. Normalized Scaling: Affected by outliers.

   Standardized Scaling: Less affected by outliers.

5. Normalized Scaling: Used when the distribution of data is unknown.

   Standardized Scaling: Used when the data distribution is normal.

6. Normalized Scaling: Also known as scaling normalization.

   Standardized Scaling: Also known as Z-score normalization.

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

An infinite value of the Variance Inflation Factor (VIF) indicates perfect multicollinearity among the independent variables in a regression model. This occurs when one independent variable can be perfectly predicted by a linear combination of the other independent variables.

1. It tends to Perfect Multicollinearity. Infinite VIF indicates that there is perfect multicollinearity, meaning that the independent variables are not providing unique information. This makes it impossible to estimate the regression coefficients uniquely.

2. It causes Model Instability. High multicollinearity can make the model coefficients unstable and unreliable, as small changes in the data can lead to large changes in the model.

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

A Q-Q plot, or quantile-quantile plot, is a graphical tool used to compare the distribution of a dataset to a theoretical distribution, typically the normal distribution. It helps to assess whether the data follows a specific distribution by plotting the quantiles of the data against the quantiles of the theoretical distribution.

1. Quantiles are points taken at regular intervals from the cumulative distribution function (CDF) of a random variable. For example, the median is the 0.5 quantile.
2. In a Q-Q plot, the quantiles of the sample data are plotted against the quantiles of the theoretical distribution. If the data follows the theoretical distribution, the points should lie approximately on a straight line.

Thank you