Linear Regression Subjective Questions

Assignment-based Subjective Questions

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

Ans 1.

Heatmap:

Some of the variables as yr, Weekday_0, weekday_6 are highly corelated to the dependent variable.

BoxPlot:

- Fall season has the highest bookings or the month of sept.
- Year 2019 has the higher number of bookings
- Higher booking on non-holiday, for both casual and registered users.
- Weekends has a higher number of bookings
- People prefer too book bikes more on a sunny day.
- We can say that temp and atemp shows a linear relationship with the cnt of bikes on that day.
- And humidity and windspeed is not showing any linear relationship

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

Ans 2. So that for N level of a feature there will N-1 Dummy Variables created. By setting drop_first=True during the creation of dummy variables, we instruct the function to drop the first category of each categorical variable before creating dummy variables. This ensures that we avoid the dummy variable trap and mitigate the issue of multicollinearity. It also reduces the dimensionality of the dataset, making it more efficient to work with.

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

Ans 3. Temp seems to have highest corelation with Target Variable for both Registered and Casual Users

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

Ans 4. Linearity: The relationship between the dependent variable and the independent variables is linear. This means that the change in the dependent variable should be proportional to the change in the independent variable.

Independence: The observations used in the regression is independent of each other. This means that the value of one observation is influenced by the value of another observation.

Homoscedasticity: The variance of the residuals is constant across all levels of the independent variable(s).

Normality: The residuals are normally distributed.

No multicollinearity: The independent variables are highly correlated with each other.

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

Ans 5. Top 3 Features for:

1. Registered Users:
    a. Temperature
    b. Year
    c. Clear Weather
2. Casual Users:
    a. Temprature
    b. Weekday 6
    c. Weekday 0

General Subjective Questions

Ques 1. Explain the linear regression algorithm in detail.

Ans 1.

Linear regression is a statistical method for modeling the relationship between a dependent variable and one or more independent variables. In simple linear regression, there is only one independent variable, whereas in multiple linear regression, there are multiple independent variables. The goal of linear regression is to find the line or hyperplane that best fits the data by minimizing the sum of the squared errors.

The mathematical formula for simple linear regression is: y = b0 + b1*x + e

y = b0 + b1x1 + b2x2 + ... + bn*xn + e

To find the values of b0, b1, b2, ..., bn, we use a method called Ordinary Least Squares (OLS). OLS minimizes the sum of the squared errors between the predicted and actual values of the dependent variable. Once we have the values of the coefficients, we can use the formula to predict the values of the dependent variable for new values of the independent variables.

Linear regression has many variants like Ridge regression, Lasso regression, Elastic Net, etc., which are used to handle multicollinearity, overfitting, and other issues that arise in linear regression.

Ques 2. Explain the Anscombe's quartet in detail.

Ans 2. Anscombe's quartet is a collection of four datasets that have identical statistical properties but different visual representations. Each dataset contains 11 (x, y) points that have been created to have almost the same summary statistics (mean, variance, correlation, etc.). These datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of graphical exploratory data analysis in identifying patterns and outliers.

The four datasets in Anscombe's quartet are:

1. Dataset I: This dataset has a linear relationship between x and y. The regression line fits the data well, with an R-squared value of 0.67.
2. Dataset II: This dataset has a non-linear relationship between x and y. The regression line does not fit the data well, with an R-squared value of 0.02.
3. Dataset III: This dataset has a linear relationship between x and y, but with an outlier. The regression line fits the data well, with an R-squared value of 0.67, but the outlier is clearly visible in the plot.
4. Dataset IV: This dataset has a non-linear relationship between x and y, with an outlier. The regression line does not fit the data well, with an R-squared value of 0.02, and the outlier is clearly visible in the plot.

The main point of Anscombe's quartet is that summary statistics such as mean, variance, and correlation can be misleading when trying to understand a dataset. It is important to visually explore the data and identify patterns and outliers that may not be apparent from summary statistics alone. Graphical exploratory data analysis allows us to better understand the relationship between variables and make more informed decisions when performing statistical analyses.

Ques 3. What is Pearson's R?

Ans 3. Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure that quantifies the linear relationship between two continuous variables. It is denoted by the symbol "r" and ranges between -1 and 1.

A value of +1 indicates a perfect positive correlation, where an increase in one variable is associated with an increase in the other variable at the same rate. A value of -1 indicates a perfect negative correlation, where an increase in one variable is associated with a decrease in the other variable at the same rate. A value of 0 indicates no linear correlation, meaning there is no relationship between the variables.

Pearson's R is calculated as the covariance of two variables divided by the product of their standard deviations. It is commonly used in regression analysis and other statistical analyses to understand the relationship between variables and to make predictions based on that relationship. However, it assumes that the relationship between the variables is linear, that the variables are normally distributed, and that there are no outliers in the data.

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

Ans 4. Scaling is the process of transforming numerical data into a standard range to facilitate analysis and model building. It involves converting the values of the features to a specific scale, which makes them comparable and removes the impact of their magnitudes.

Scaling is performed because some machine learning algorithms are sensitive to the magnitude of the features. If the range of the feature values is large, it can cause issues such as slow convergence and difficulty in finding the optimal solution. Scaling also ensures that all features are on the same scale, making it easier to compare them and identify patterns in the data.

Normalized scaling and standardized scaling are two common methods for scaling data.

Normalized scaling involves scaling the data so that the values fall within a specific range, typically between 0 and 1. This is done by subtracting the minimum value of the data and dividing by the range of the data (i.e., the maximum value minus the minimum value). The formula for normalized scaling is:

$x' = (x - \min(x)) / (\max(x) - \min(x))$

Standardized scaling, on the other hand, involves transforming the data so that it has a mean of 0 and a standard deviation of 1. This is done by subtracting the mean of the data and dividing by the standard deviation of the data. The formula for standardized scaling is:

$x' = (x - \text{mean}(x)) / \text{std}(x)$

The main difference between normalized scaling and standardized scaling is that normalized scaling only scales the data to a specific range, while standardized scaling also standardizes the data by transforming it to have a mean of 0 and a standard deviation of 1. In other words, normalized scaling preserves the original distribution of the data, while standardized scaling transforms the data to a standard normal distribution.

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

Ans 5. In statistics, the Variance Inflation Factor (VIF) is a measure of the extent to which the variance of the estimated regression coefficient is increased due to collinearity among the predictor variables. It quantifies the degree to which the variance of an estimated regression coefficient is increased because of the correlation of the predictor variable with the other predictor variables in the model.

In some cases, the value of VIF may be infinite. This occurs when the predictor variable can be perfectly predicted by a linear combination of the other predictor variables in the model. In other words, there is a perfect linear relationship between the predictor variable and the other predictor variables in the model, leading to a perfect multicollinearity. When this happens, the regression model cannot be estimated using the standard techniques, as the coefficient estimates cannot be obtained due to the perfect collinearity.In practice, it is important to detect and handle perfect multicollinearity in the data before running a regression analysis. One approach to handling perfect multicollinearity is to remove one of the correlated variables from the model. Another approach is to use regularization techniques, such as Ridge or Lasso regression, to handle the collinearity.

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

Ans 6. A Q-Q plot (quantile-quantile plot) is a graphical technique used to check the normality of a dataset. It compares the distribution of a given set of data to the normal distribution by plotting the quantiles of the sample against the corresponding quantiles of the normal distribution.

In the context of linear regression, Q-Q plots are important because linear regression models assume that the residuals (the differences between the predicted values and actual values) are normally distributed. If the residuals are not normally distributed, this may indicate that the model is not appropriate for the data, and the results of the regression analysis may be unreliable.A Q-Q plot can help us visually assess the normality of the residuals. If the residuals are normally distributed, the Q-Q plot should form a straight line. However, if the residuals deviate significantly from a straight line, this may indicate that the normality assumption is violated, and the linear regression model may not be appropriate for the data.