Linear regression_Jignesh Shinde

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? (3 marks)

I have done analysis on categorical columns using the boxplot and bar plot. Below are the few points we can infer from the visualization –
Most of the bookings has been done during the month of may, june, july, aug, sep and oct. Trend increased starting of the year till mid of the year and then it started decreasing as we approached the end of year. 2019 attracted more number of booking from the previous year, which shows good progress in terms of business. Clear weather attracted more booking which seems obvious.
Fall season seems to have attracted more booking. And, in each season the booking count has increased drastically from 2018 to 2019. When it's not holiday, booking seems to be less in number which seems reasonable as on holidays, people may want to spend time at home and enjoy with family. Booking seemed to be almost equal either on working day or non-working day. 2019 attracted more number of booking from the previous year, which shows good progress in terms of business. Thu, Fir, Sat and Sun have more number of bookings as compared to the start of the week.

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

Using `drop_first=True` during dummy variable creation in regression analysis is essential to prevent multicollinearity and the dummy variable trap. By dropping one dummy variable, we avoid perfect multicollinearity among the dummy variables and ensure independence, which leads to more stable coefficient estimates and meaningful interpretation of the model.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

'temp' variable has the highest correlation with the target variable.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

I have validated the assumption of Linear Regression Model based on below assumptions -
Independence of residuals

- Normality of error terms
- Multicollinearity check
- Homoscedasticity
- Linear relationship validation
- No auto-correlation

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

The top 3 features contributing significantly towards explaining the demand of the shared bikes are temp, winter and sep

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the independent and dependent variables. The objective is to find the best-fitting straight line (or hyperplane in higher dimensions) that describes this relationship.

Key assumptions include linearity, independence of observations, constant variance of residuals (homoscedasticity), and normality of residuals. The model equation typically takes the form (y = beta_0 + beta_1x + varepsilon), where (y) is the dependent variable, (x) is the independent variable, (beta_0) is the intercept, (beta_1) is the coefficient (slope), and (varepsilon) is the error term.

Parameter estimation is done using ordinary least squares (OLS), minimizing the sum of squared differences between observed and predicted values. Evaluation metrics include Mean Squared Error (MSE) and (R^2) (proportion of variance explained). Linear regression can be extended to handle multiple independent variables and complex relationships using techniques like polynomial regression, ridge regression, and lasso regression.

It's implemented in various programming languages and libraries such as Python (e.g., scikit-learn, statsmodels). Coefficients provide insights into the relationship between variables. Linear regression finds applications in economics, finance, biology, and social sciences for prediction, inference, and understanding data relationships.

2. Explain the Anscombe’s quartet in detail. (3 marks)

Anscombe's quartet is a set of four datasets that have nearly identical statistical properties but vastly different graphical representations. This collection was introduced by the statistician Francis Anscombe in 1973 to illustrate the importance of graphing data before analyzing it and to demonstrate the limitations of summary statistics in capturing the underlying structure of data. Here's a detailed explanation of each dataset in the quartet:

1. Dataset 1:

- Consists of 11 data points.

- Features a linear relationship between (x) and (y), with some random variation around the line.

- Summary statistics, such as mean, standard deviation, and correlation coefficient, accurately describe the data.

2. Dataset 2:

- Also consists of 11 data points.

- Exhibits a non-linear relationship between (x) and (y), forming a quadratic curve.

- Summary statistics remain similar to Dataset I despite the different relationship.

3. Dataset 3:

- Includes 11 data points as well.

- Contains an outlier that significantly influences the linear relationship between (x) and (y).

- The outlier has a disproportionate impact on summary statistics, such as the mean and correlation coefficient.

4. Dataset 4:

- Comprises 11 data points distributed along (x) values with one outlier.

- All data points have the same (x) value, except for the outlier.

- Summary statistics imply a perfect linear relationship, which is not evident from the data's graphical representation.

3. What is Pearson's R? (3 marks)

Pearson's R , or the Pearson correlation coefficient, measures the strength and direction of the linear relationship between two continuous variables. It ranges from -1 to 1, with 1 indicating a perfect positive linear relationship, -1 indicating a perfect negative linear relationship, and 0 indicating no linear relationship. It's calculated using the covariance of the variables divided by the product of their standard deviations. Pearson's R is widely used in statistics and data analysis to assess linear associations between variables, but it doesn't capture non-linear relationships.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Scaling is a preprocessing step used in data analysis and machine learning to adjust the range or scale of features. Its main goal is to make sure that all features contribute equally to the analysis, preventing some features from dominating due to their larger scales.

Why Scaling is Performed:

1. Equal Weighting: Scaling ensures that all features have similar ranges, preventing certain features from overshadowing others in the analysis or learning process.

2. Improved Performance: Many machine learning algorithms perform better when features are on similar scales. Scaling can enhance algorithm stability and convergence speed.

3. Interpretability: Scaling aids in interpreting coefficients or feature importance, especially in linear models or algorithms reliant on distance measures.

Types of Scaling:

1. Normalized Scaling (Min-Max Scaling): Scales data to a fixed range, typically between 0 and 1, preserving the original distribution.

2. Standardized Scaling (Z-score Scaling): Centers data around its mean and scales it based on its standard deviation, making it robust to outliers.

Difference between Normalized and Standardized Scaling:

- Normalized Scaling: Preserves the original distribution, scaling data to a fixed range.

- Standardized Scaling: Centers data around its mean, scaling it based on its standard deviation, making it robust to outliers and suitable for algorithms assuming normally distributed data.

The choice between these methods depends on the data characteristics and analysis requirements.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

The VIF (Variance Inflation Factor) measures the degree of multicollinearity between predictor variables in a regression model. A VIF value greater than 10 is often considered indicative of multicollinearity, but sometimes you may encounter cases where the VIF is reported as infinite. This occurs when there is perfect multicollinearity between predictor variables. Perfect multicollinearity happens when one predictor variable in a regression model can be perfectly predicted by a linear combination of other predictor variables. In other words, one of the predictor variables is a perfect linear function of one or more other predictor variables. When perfect multicollinearity exists, the determinant of the matrix of predictor variables becomes zero, leading to an infinite VIF value. This happens because the inverse of the matrix cannot be computed due to its singularity. Perfect multicollinearity can arise for various reasons, such as including a variable that is a constant multiple of another variable, or introducing dummy variables for categorical variables without dropping one of the categories. To address this issue, it's crucial to carefully examine the predictor variables and identify and remove any redundant variables that lead to perfect multicollinearity. This might involve rethinking the model specification, excluding certain variables, or combining correlated variables into composite variables.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

A Q-Q plot, or quantile-quantile plot, compares the quantiles of a given dataset to those of a theoretical distribution, typically the normal distribution. In linear regression, Q-Q plots are used to assess the normality assumption of residuals, crucial for valid inference. They visually reveal whether the residuals follow a normal distribution, aiding in model validation, inference validity, and potential

model improvement. If the plotted points deviate significantly from a straight line, it suggests departures from normality, prompting modelers to explore alternative specifications or transformations to enhance model performance and validity. Q-Q plots are essential tools for assessing and ensuring the validity of linear regression models.