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)

Answer: There were 6 categorical variables in the dataset. All variables except `holiday` can be good predictor for the dependant variable as almost 97% of the bike booking were happening when it is not a holiday, which means that the data is clearly biased. While all variables, `season`, `mnth`, `weathersit`, `weekday`, `workingday` can be good predictors.

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

Answer: Using `drop_first=True` during dummy variable creation is important to avoid multicollinearity in linear regression.

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

Answer: `atemp` has the highest correlation.

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

Answer: **Residual Analysis**: Calculated the residuals by subtracting the predicted values from the actual target values in the training set. Then, I checked for linearity and homoscedasticity.
Multicollinearity: Calculated the Variance Inflation Factor (VIF) for each independent variable to assess multicollinearity. VIF values above a certain threshold (typically 5 or 10) indicate high multicollinearity.

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

Answer: `temp`, `yr` and `weathersit`.

General Subjective Questions

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

Answer: Linear regression is a supervised machine learning algorithm used for modeling the relationship between a dependent variable (target) and one or more independent variables (features or predictors).
Objective: The primary goal of linear regression is to find the best-fitting linear relationship between the independent variables (features) and the dependent variable (target). This relationship can be represented as a straight line equation:
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon$
y is the target variable.
$x_1, x_2, ..., x_k$ are the independent variables.
$\beta_0, \beta_1, \beta_2, ..., \beta_k$ are the coefficients that represent the intercept and the slopes of the linear equation.
ε represents the error term, accounting for the variability not explained by the model.
Model Fitting: To build the linear regression model, the algorithm determines the values of coefficients ($\beta_0, \beta_1, \beta_2$, etc.) that minimize the sum of squared differences between the actual target values and the values predicted by the model. This process is often referred to as "model training" or "model fitting."
Types of Linear Regression: Linear regression can take different forms, including:
Simple Linear Regression: Involves a single independent variable (e.g., predicting house prices based on square footage).

Multiple Linear Regression: Includes multiple independent variables (e.g., predicting a car's price based on its age, mileage, and other features).

Evaluation and Prediction: After fitting the model, it is essential to evaluate its performance. Common evaluation metrics include Mean Squared Error (MSE), R-squared (R^2), and Mean Absolute Error (MAE). These metrics measure how well the model fits the training data and generalizes to unseen data. Once the model is evaluated and deemed satisfactory, it can be used to make predictions on new data by plugging in the values of the independent variables.

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

Answer: Anscombe's quartet is a set of four small datasets that have nearly identical simple descriptive statistics but exhibit very different patterns and relationships when graphed. It was created by the statistician Francis Anscombe in 1973 to illustrate the importance of data visualization in understanding and analyzing data. The quartet is a classic example that highlights the limitations of relying solely on summary statistics without visualizing the data.

Creation and Composition: Anscombe's quartet consists of four separate datasets, each with 11 data points. Despite having different values, each dataset has nearly identical summary statistics, including means, variances, and correlation coefficients. This was a deliberate design to demonstrate that summary statistics alone can be misleading.

Data Characteristics: The four datasets within the quartet are distinct in their underlying patterns:

Dataset I: Exhibits a strong linear relationship between the two variables.

Dataset II: Also has a linear relationship but with an outlier that significantly influences the regression line.

Dataset III: Shows a non-linear relationship between the variables.

Dataset IV: Contains one data point that is an extreme outlier, which has a significant impact on the correlation and regression line.

Significance: Anscombe's quartet serves as a cautionary example in statistics and data analysis. It emphasizes that while summary statistics can provide an initial overview of a dataset, they may not reveal the full picture. The quartet demonstrates that the same summary statistics can correspond to completely different data distributions and relationships. This underlines the importance of data visualization and exploratory data analysis (EDA) to gain insights into the true nature of data, identify outliers, and validate statistical assumptions.

3. What is Pearson’s R? (3 marks)

Answer: Pearson's R, also known as the Pearson correlation coefficient or Pearson's correlation, is a statistic used to quantify the linear relationship or correlation between two continuous variables. It is a measure of the strength and direction of the linear association between two variables. Pearson's R is calculated as follows:

$R = (\Sigma((X_i - \bar{X})(Y_i - \bar{Y}))) / (n * \sigma_x * \sigma_y)$

Σ denotes summation over all data points (i from 1 to n).

X_i and Y_i represent individual data points.

$\bar{X}$ and $\bar{Y}$ are the means of X and Y, respectively.

σ_x and σ_y are the standard deviations of X and Y, respectively.

Interpretation: Pearson's R takes values between -1 and 1, where:

A positive value (closer to 1) indicates a positive linear correlation, meaning that as one variable increases, the other tends to increase as well.

A negative value (closer to -1) indicates a negative linear correlation, implying that as one variable increases, the other tends to decrease.

A value close to 0 suggests a weak or no linear correlation between the two variables.

Pearson's R is widely used in statistics and data analysis to assess the strength and direction of linear relationships between variables. It's a valuable tool for understanding how two variables are related and can help in making predictions or drawing insights from data.

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

Answer: Scaling is the process of transforming the values of variables (features) into a specific range or distribution. It is done to ensure that all features have a similar scale, which can be essential for various machine learning algorithms. Scaling is performed to make sure that no feature dominates others in the modeling process, particularly when using algorithms that rely on distance-based calculations or gradient descent.

Reasons for Scaling:

Equal Weightage: Scaling ensures that all features are on a similar scale, giving them equal weight in the model. This is crucial for algorithms like k-means clustering, support vector machines, and principal component analysis (PCA).

Faster Convergence: In gradient-based optimization algorithms (e.g., linear regression, neural networks), scaling helps in faster convergence and prevents the algorithm from taking longer routes to find the optimal solution.

Improved Model Performance: Scaling can lead to improved model performance, accuracy, and interpretability.

There are two common methods of scaling: normalized scaling and standardized scaling:

Difference Between Normalized and Standardized Scaling:

Range: Normalized scaling scales data to a specific range, typically between 0 and 1, preserving the relative relationships between data points. Standardized scaling centers the data around a mean of 0 and scales it based on the standard deviation.

Handling Outliers: Standardized scaling is more robust to outliers because it's based on the mean and standard deviation, while normalized scaling can be influenced by outliers since it depends on the minimum and maximum values.

Use Cases: Normalized scaling is suitable when you know the range of values is meaningful (e.g., percentages), and you want to maintain that range. Standardized scaling is often preferred when the distribution is not necessarily bounded and you want to ensure that all features have comparable means and variances.

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

Answer: The value of the Variance Inflation Factor (VIF) can become infinite when there is perfect multicollinearity among the independent variables. Perfect multicollinearity means that one or more independent variables in the model can be exactly predicted from a linear combination of the others.

For example, if you have a variable $X_1 = 2 * X_2$ in your model, there is perfect multicollinearity.

The formula to calculate the VIF for a particular independent variable is:

$VIF = 1 / (1 - R^2)$

Where R^2 is the coefficient of determination when the variable of interest is regressed on all the other independent variables. In the presence of perfect multicollinearity, R^2 becomes 1, and thus, VIF becomes 1 / (1 - 1), which equals infinity.

Interpretation: An infinite VIF indicates that the variance of the estimated coefficient for the variable with perfect multicollinearity is inflated to an extreme degree. In practical terms, it becomes impossible to provide meaningful and stable coefficient estimates for the variable with perfect multicollinearity. This can lead to issues in the model, such as unstable parameter estimates and challenges in making inferences.

To address the problem of infinite VIF, it's essential to identify and address multicollinearity in the model by removing one or more of the correlated variables, redefining the model, or using techniques like Ridge or Lasso regression, which can help mitigate multicollinearity issues.

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

(3 marks)

Answer: A Q-Q plot (Quantile-Quantile plot) is a graphical tool used in statistics and data analysis to assess whether a dataset follows a specific theoretical distribution, typically the normal distribution. It helps you visually compare the quantiles of your dataset to the quantiles of a theoretical distribution.

Use of Q-Q Plot: Q-Q plots are primarily used to check the assumption of normality in linear regression and other statistical analyses. The assumption of normality is essential because many statistical methods, including linear regression, assume that the errors (residuals) are normally distributed. Deviations from this assumption can impact the validity and reliability of the results.

Importance in Linear Regression: Q-Q plots are crucial in linear regression for the following reasons:

Normality Assumption: One of the fundamental assumptions of linear regression is that the residuals (the differences between observed values and predicted values) are normally distributed. A Q-Q plot helps verify whether this assumption holds by visually comparing the observed residuals to the expected quantiles of a normal distribution.

Detecting Departures from Normality: If the points in the Q-Q plot deviate from a straight line, it suggests departures from normality. These departures can include heavy tails, skewness, or outliers in the residuals.

Model Validity and Interpretability: Ensuring that the residuals are normally distributed is important for the validity of hypothesis tests, confidence intervals, and p-values in linear regression. It also enhances the interpretability of the model coefficients and the accuracy of predictions.

Date: 11-10-2023

(Vikas Sharma)