Answers of 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?

Answer :

We have 2 categorical variables: **'season'** and **'weathersit**.' The effect of these categorical variables on the dependent variable, 'cnt' is as follows:

A. **Season**:

- The 'season' variable is converted to dummy variables, with 'season_1' (spring) as the reference category.
- The coefficients of the 'season' dummy variables indicate how each season affects bike demand compared to spring.
- Positive coefficients for 'season_2' (summer), 'season_3' (fall), and 'season_4' (winter) would suggest that these seasons tend to increase bike rental demand compared to spring.
- Inferences:

a. Summer and fall are likely to have a positive effect on bike rental demand compared to spring.

b. Winter may also have a positive effect, possibly due to holiday or winter sport-related demand.

B. **Weathersit**:

- The 'weathersit' variable is converted into dummy variables, with 'weathersit_1' (clear, few clouds) as the reference category.
- The coefficients of the 'weathersit' dummy variables indicate how different weather conditions affect bike demand compared to clear, few clouds.
- Negative coefficients for 'weathersit_2' (misty/cloudy), 'weathersit_3' (light rain/snow), and 'weathersit_4' (heavy rain/snow) would suggest that these weather conditions tend to decrease bike rental demand compared to clear, few clouds.
- Inferences:

a. Misty/cloudy weather is likely to have a negative effect on bike rental demand compared to clear conditions.

b. Light rain/snow and heavy rain/snow are also likely to have a negative impact on demand.

These coefficients are important when making inferences. Statistically significant coefficients are more reliable indicators of the variables' effect on the dependent variable.

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

Answer :

Using drop_first=True when creating dummy variables is a standard practice in regression analysis, as it helps address the issue of multicollinearity, enhances model interpretability, simplifies the model, and makes the estimation process more efficient. It ensures that your regression model is more reliable and valid for making inferences about the relationships between categorical variables and the target variable.

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

Answer :

The variable **'registered'** has the highest positive correlation with the target variable **'cnt'** with a correlation value of **0.95**. This high **positive** correlation suggests that the number of registered users has a **strong positive influence** on the total bike rental count.

Additionally, the **R-squared score** on the test set is **0.83**, which means that our linear regression model explains approximately **83%** of the variance in the target variable **'cnt**.' This is a relatively **good R-squared score**, indicating that the model provides a reasonable fit to the data.

A high correlation between 'registered' and 'cnt' is expected since 'cnt' is the sum of 'casual' and 'registered' users, and 'registered' users represent a significant portion of the total rentals.

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

Answer :

As we can see, the variable 'registered' has the highest positive correlation with the target variable 'cnt' with a correlation value of 0.95. This high positive correlation suggests that the number of registered users has a strong positive influence on the total bike rental count.

Also, the R-squared score on the test set is 0.83, which means that your linear regression model explains approximately 83% of the variance in the target variable 'cnt.' This is a relatively good R-squared score, indicating that the model provides a reasonable fit to the data.

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

Answer :

As we can see here, yr, season_4 and weathersit_3 are the top 3 features contributing significantly towards explaining the demand of the shared bikes.

These 3 are the features with the largest absolute coefficient values and hence are the most significant contributors to the model.

Answers of General Subjective Questions :

1. **Explain the linear regression algorithm in detail.**

   Answer :

   Linear regression is a supervised machine learning algorithm used for modeling the relationship between a dependent variable and one or more independent variables. It is one of the simplest and most popular machine learning algorithms, and has a wide range of applications in various fields, such as forecasting, trend analysis, and correlation exploration.

   The linear regression algorithm works by fitting a line to the data points in such a way that the sum of the squared residuals is minimized. The residuals are the differences between the actual values of the dependent variable and the predicted values based on the fitted line.

   Below is the detailed explanation of the linear regression algorithm:

   a. Collect data: The first step is to collect a dataset that contains the independent and dependent variables.

   b. Prepare the data: Once the data has been collected, it is important to prepare it for the linear regression algorithm. This may involve cleaning the data, handling missing values, and scaling the data.

   c. Choose a linear regression algorithm: There are several different linear regression algorithms available. The best algorithm to choose will depend on the specific dataset and the desired results.

   d. Train the model: Once a linear regression algorithm has been chosen, it is necessary to train the model on the training data.

e. Evaluate the model: Once the model has been trained, it is important to evaluate its performance on the held-out test data. This will help to assess how well the model generalizes to new data.

f. Make predictions: Once the model has been evaluated and found to be performing well, it can be used to make predictions on new data.

2. Explain the Anscombe's quartet in detail.

Answer :

Anscombe's quartet is a famous dataset in statistics that consists of four small datasets that have nearly identical simple descriptive statistics (e.g., means, variances, and correlation coefficients), but they exhibit markedly different patterns when graphically analyzed. It was created by the British statistician Francis Anscombe in 1973 to emphasize the importance of data visualization and exploratory data analysis in statistical modeling.

The key lesson from Anscombe's quartet is that it is not sufficient to rely solely on summary statistics to understand a dataset. Data visualization and exploratory data analysis are crucial for uncovering the underlying patterns and relationships in the data. Different datasets with the same summary statistics may require entirely different modeling approaches, and this highlights the importance of context and domain knowledge in statistical analysis.

Anscombe's quartet is a reminder that it is important to plot data before analyzing it, and to be aware of the potential impact of outliers and other influential observations. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

Anscombe's quartet is often used in statistics and data science education to teach students about the importance of data visualization and the limitations of summary statistics. It is also used in research to study the robustness of statistical methods to different types of data distributions.

3. What is Pearson's R?

Answer :

Pearson's correlation coefficient, often denoted as "Pearson's R" or simply "r," is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It is widely used to assess the degree to which two variables are related to each other.

Pearson's R, also known as the Pearson correlation coefficient, is a measure of the linear correlation between two quantitative variables. It is a number between -1 and 1, where -1

indicates a perfect negative correlation, 1 indicates a perfect positive correlation, and 0 indicates no correlation.

Pearson's R can be used to measure the strength of the relationship between two variables, such as the relationship between height and weight, or the relationship between test scores and years of study. It can also be used to test whether the relationship between two variables is statistically significant.

To calculate Pearson's R, the following formula is used:

Pearson's R = (covariance of x and y) / (standard deviation of x * standard deviation of y)

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

Answer :

Scaling in machine learning is the process of transforming the values of features in a dataset to a similar scale. It is performed to ensure that all features contribute equally to the model and to avoid the domination of features with larger values.

Normalized scaling is a type of scaling that transforms the values of features to a range of 0 to 1. This is done by subtracting the minimum value from each feature and then dividing by the maximum value minus the minimum value.

Standardized scaling is another type of scaling that transforms the values of features to a mean of 0 and a standard deviation of 1. This is done by subtracting the mean from each feature and then dividing by the standard deviation.

Scaling is performed for the following reasons:

1. **Avoiding Numerical Instability**: Some machine learning algorithms are sensitive to the scale of input features. If features have significantly different scales, the algorithm may become numerically unstable and may not converge to an optimal solution.
2. **Improving Model Performance**: Scaling can help improve the performance of some machine learning algorithms, such as gradient descent-based optimization methods. It allows the algorithm to converge faster and find better model parameters.
3. **Interpretable Coefficients**: In linear models, such as linear regression, scaling the features ensures that the coefficients of the model are more interpretable, as they represent the change in the dependent variable associated with a one-unit change in the corresponding feature.
4. **Comparison of Variables**: Scaling makes it easier to compare the importance or contribution of different features to the model, as all features are on a similar scale.

Scaling in machine learning is the process of transforming the values of features in a dataset

to a similar scale. It is performed to ensure that all features contribute equally to the model and to avoid the domination of features with larger values.

Normalized scaling is a type of scaling that transforms the values of features to a range of 0 to 1. This is done by subtracting the minimum value from each feature and then dividing by the maximum value minus the minimum value.

Standardized scaling is another type of scaling that transforms the values of features to a mean of 0 and a standard deviation of 1. This is done by subtracting the mean from each feature and then dividing by the standard deviation.

Why is scaling performed?

Scaling is performed for a number of reasons, including:

- To improve the performance of machine learning algorithms: Many machine learning algorithms are sensitive to the scale of the data. Scaling the data can help to ensure that all features are treated equally and that the algorithms can converge more quickly.
- To prevent certain features from dominating the model: If features have different scales, then features with larger scales may dominate the model and have an excessive impact on the results. Scaling the data can help to ensure that all features contribute equally to the model.
- To make the data more interpretable: Scaling the data can make it easier to interpret the results of machine learning models. For example, if the data is scaled to a range of 0 to 1, then it is easy to see which features have the greatest impact on the model.

Difference between normalized scaling and standardized scaling:

Normalized scaling and standardized scaling are two of the most common scaling techniques used in machine learning. The main difference between these two techniques is that normalized scaling transforms the values of features to a range of 0 to 1, while standardized scaling transforms the values of features to a mean of 0 and a standard deviation of 1.

Normalized scaling is useful when the features have different units of measurement. For example, if one feature is measured in meters and another feature is measured in kilograms, then normalized scaling can be used to transform the values of these features to the same range.

Standardized scaling is useful when the features have different distributions. For example, if one feature follows a normal distribution and another feature follows a skewed distribution, then standardized scaling can be used to transform the values of these features to the same distribution.

Normalized scaling scales the features to a specified range, preserving the relationships between data points and features, while standardized scaling transforms the features to

have a mean of 0 and a standard deviation of 1, making them directly comparable and robust to outliers. The choice between the two scaling techniques depends on the specific requirements of the machine learning algorithm and the nature of the data.

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

Answer :

The Variance Inflation Factor (VIF) is a statistical measure used to assess multicollinearity in a regression model, particularly in multiple linear regression. Multicollinearity occurs when two or more independent variables in a regression model are highly correlated, which can lead to issues in model interpretation and estimation of coefficients. A high VIF value indicates that the variance of the coefficient estimates is inflated due to multicollinearity.

A VIF of infinity (or, more precisely, extremely large values) occurs when there is perfect multicollinearity in the model. Perfect multicollinearity happens when two or more independent variables in the model are perfectly correlated, meaning that one of them can be exactly predicted from the others. This leads to a situation where the coefficient estimates become unstable and indeterminate.

If VIF is infinite, then it is necessary to take steps to address the multicollinearity in the model. This may involve removing redundant variables, combining highly correlated variables, or transforming the data.

Below can be followed to avoid perfect multicollinearity:

- Carefully select your independent variables: Make sure that your independent variables are not redundant or highly correlated with each other.
- Use domain knowledge: Use your knowledge of the problem domain to identify and remove redundant or highly correlated variables.
- Transform the data: You may be able to reduce multicollinearity by transforming the data. For example, you could take the logarithm of the variables or standardize the variables.

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

Answer :

A Q-Q plot, or quantile-quantile plot, is a graphical method for comparing the distributions of two datasets. It is constructed by plotting the quantiles of one dataset against the quantiles of the other dataset. If the two datasets have the same distribution, then the points on the Q-Q plot will fall along a straight line. If the two datasets have different distributions, then the points on the Q-Q plot will deviate from the straight line.

Q-Q plots are often used in linear regression to assess whether the residuals of the model are normally distributed. Normality of the residuals is one of the assumptions of linear regression. If the residuals are not normally distributed, then the results of the linear regression model may be unreliable.

To construct a Q-Q plot for the residuals of a linear regression model, the following steps are taken:

1. Calculate the quantiles of the residuals.
2. Calculate the quantiles of a standard normal distribution.
3. Plot the quantiles of the residuals against the quantiles of the standard normal distribution.

If the residuals are normally distributed, then the points on the Q-Q plot will fall along a straight line. If the residuals are not normally distributed, then the points on the Q-Q plot will deviate from the straight line.

Q-Q plots are an important tool for assessing the normality of the residuals of a linear regression model. If the residuals are not normally distributed, then it may be necessary to transform the data or use a different statistical model.

Here are some examples of how Q-Q plots can be used in linear regression:

- To assess whether the residuals of a linear regression model are normally distributed.
- To compare the distributions of the residuals of two different linear regression models.
- To identify outliers in the data.
- To detect non-linear relationships in the data.

Q-Q plots are a powerful tool for understanding the distributions of data and for diagnosing problems with linear regression models.

tuneshare
more_vert