LINEAR REGRESSION CASE STUDY

UPGRAD & IIITB, DATA SCIENCE DSC65 – FEB 2024

BY G B SHRUTHI

ASSIGNMENT BASED QUESTIONS

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

- 'season','yr','mnth','weekday','weathersit','workingday','holiday‘ are the categorical variables from the data.
- In the rainy season, the count is slightly less when compared with clear sky.
- Holidays effects the no of bike count.
- Year also plays an important role of effect on dependent variable.

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

Ans: Since if there are n levels in a column then there should be n-1 no of dummy variables. Without using drop_first=True would make the dummy variables correlated to each other and hence, redundant, which is not expected of our analysis.

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

Ans: temp and atemp variables have the highest correlation from the data.

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

Ans: One of the assumptions of Linear regression after building the model is that the error terms or residuals should be normally distributed. Also the Adjusted R-squared values of Trained dataset and test dataset must be similar or nearer.

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

Ans: The top 3 features directly influencing the count are the features with highest coefficients are temp, year → Positively correlated. season_spring, windspeed and weathersit_Light Snow & Rain (Negatively correlated).

SUBJECTIVE BASED QUESTIONS

1. EXPLAIN THE LINEAR REGRESSION ALGORITHM IN DETAIL

- Linear Regression is a type of Machine learning algorithm mainly a supervised machine learning algorithm that learns from the labelled datasets and maps the data points to the linear functions.
- Linear Regression algorithm computes the linear relationship between the dependent variable and one or more independent variables by fitting a linear equation.
- When there is only one independent variable, then it is known as Simple Linear Regression.
- When there are more than one independent variables, then it is known as Multiple Linear Regression.
- Equation of linear regression is $Y=mx+c$; m = slope c = intercept

Assumptions of Simple Linear Regression:

- Linear Equation: The dependent and independent variables have a linear relationship with one another. If the relationship is not linear, then linear regression will not be an accurate model.
- Normality: The residuals should be normally distributed which means it gives a bell curve shape. If not, then it is not an accurate model.
- Independence: The observations of the dataset should be independent.
- Homoscedasticity: The independent variables should have constant variance.

2. EXPLAIN THE ANSCOMBE'S QUARTET IN DETAIL

- Anscombe's Quartet comprises of 4 datasets with nearly summary statistics, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations and linear regression lines but having different representations when we scatter plots on graph.
- It is a group of datasets that have same mean, standard deviations and regression line but which are qualitatively different.
- The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

3. WHAT IS PEARSON'S R?

- The Pearson correlation coefficient is the most common way of measuring the linear correlation. It is a number from -1 to 1 that measures the strength and direction of the relationship between two variables.

Types of correlations:

- Positive correlation: Between 0 and 1. When one changes the other also changes in the same direction.
- Negative correlation: Between -1 and 0. When one changes the other also changes in the opposite direction.
- No correlation: At 0. There is no relationship between the variables.

4. WHAT IS SCALING? WHY IS SCALING PERFORMED? WHAT IS THE DIFFERENCE BETWEEN NORMALIZED SCALING AND STANDARDIZED SCALING?

Feature Scaling: It is a technique to standardize the independent features present in the data in a fixed range. It is performed during the pre-processing step of linear regression model creation to handle highly varying magnitudes or values etc.

It is used for various benefits:

- Algorithm performance algorithm
- Preventing numerical stability
- It ensures that each characteristic is considered
- It guarantees that all features are on a comparable scale or not.

TYPES OF SCALING

- Normalization: This method rescales the features the features to a fixed range, usually 0 to 1. from sklearn.preprocessing import MinMaxScaler is used. The formula for calculating the scaled value of a feature is:

Normalized Value = Value - Min/ Max – Min

- Standardization: This process involves rescaling the distribution of values so that the mean of observed values is aligned to 0 and the standard deviation to 1. from sklearn.preprocessing import StandardScaler is used. The formula for calculating the scaled value of a feature is:

Standardized value = (x-mean)/standard deviation

5. YOU MIGHT HAVE OBSERVED THAT SOMETIMES THE VALUE OF VIF IS INFINITE. WHY DOES THIS HAPPEN ?

The value of Variation Inflation factor (VIF) is INFINITE means that the value R^2 squared is equal to 1. That is if there is perfect correlation, then VIF=Infinity. This shows the perfect correlation between two variables of the dataset. i.e., VIF = 1/1-R^2 → 1/1-1 = Infinite.

To solve this we need to drop one of the variables from the dataset which is causing the multicollinearity. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

6. WHAT IS A Q-Q PLOT? EXPLAIN THE USE AND IMPORTANCE OF A Q-Q PLOT IN LINEAR REGRESSION.

Quantile- Quantile (Q-Q) plot, is a graphical tool to help us access if a set of data which came from some different types of distributions like normal, exponential, uniform etc. It also helps to determine if two datasets come from same population with a common distribution.

A QQ plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that the data sets are from populations with same distributions.

THANK YOU