MULTIPLE LINEAR REGRESSION ASSIGNMENT

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?

According to the Linear Regression Model obtained from the dataset using day.csv file, we can infer the following:

- The User Demand for the Bike Sharing Service is positively affected by workingday, weekday, summer, winter and temp variables.
- The User Demand for the Bike Sharing Service is negatively affected by, windspeed, Spring, holiday and Mist variables.

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

drop_first=True is necessary in creation of Dummy Variables, as it reduces the extra number of columns created. Therefore, it reduces the correlations created among dummy variables.

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

The Target Variable ‘Cnt’ is highly correlated with ‘temp’ and ‘atemp’ variable.

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

- Plot the histogram for Residual values obtained after model building and check whether it is is Normal Distributed.
- Plot Scatter of model predicted y values against Residuals and check if the error terms are independent throughout.

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

The variables such as ‘winter’, ‘year’ and ‘temp’ contribute significantly towards explaining the demand of the shared bikes.

MULTIPLE LINEAR REGRESSION ASSIGNMENT

General Subjective Question

1. Explain the Linear Regression Algorithm in detail?

Linear regression is a simple statistical regression method used in predictive analysis which shows the relationship between the continuous variables. Linear regression depicts the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), called linear regression. If there is a single input variable (x), such linear regression is called simple linear regression. And if there is more than one input variable, such linear regression is called multiple linear regression. The linear regression model gives a sloped straight line describing the relationship within the variables. The best fit line is obtained on basis of minimizing the cost function which is of least square method.

The algorithm divides the input data into train and test data. The model fit is iteratively calculated unless optimal p-values, R Squared and VIF values are obtained. Then, the predictions are made on test data.

2. Explain the Anscombe's Quartet in detail

Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analysing and model building, and the effect of other observations on statistical properties.

There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.

This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc.

3. What is Pearson's R?

In statistics, the Pearson correlation coefficient (PCC) also called as Pearson's r, or as simply as the correlation coefficient is a measure of linear correlation between two sets of data.

It is the ratio between the co-variance of two variables and the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.

The value R can be represented as follows:

- r = 1 implies the data is perfect linear with a positive slope.

- r = -1 implies the data is perfect linear with a negative slope.
- r = 0 implies there is no linear association.

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

It is a technique of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains variables which vary in high magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units, hence results in incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

It is important to note that scaling affects the coefficients only and not the other statistical parameters like t-statistic, F-statistic, p-values, R-squared, etc.

- Normalization/ Min-Max Scaling brings all of the data in the range of 0 and 1.
- sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.
- Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean zero and standard deviation one.
- sklearn.preprocessing.scale helps to implement standardization in python.

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

If VIF is infinity , then there is perfect correlation. This shows a perfect correlation between two independent variables. In the case, we get R2 =1, which leads to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be represented 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 analyse if a set of data came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets show a common distribution.

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

MULTIPLE LINEAR REGRESSION ASSIGNMENT

QQ Plots are used in Linear Regression to validate whether the Residuals are Normally Distributed

Vijay Prasad

13th April 2022