Multiple Linear Regression Assignment

Assignment-based Subjective Questions

Answers

1) Categorical variables in the dataset are season, Year, month, holiday, workingday, weekday and weathersit.

- The demand of bikes is highest in season 'fall' and least in 'spring' as compared to all the seasons.
- The demand of bikes increased in year '2019' as compared to year '2018'.
- The demand of bikes is highest in the months June till September in comparison to other months of the year which is in sync with the season 'fall'. The Month January is the lowest demand month.
- The demand of bikes in holidays is less as compared to the non-holidays.
- There is no significant change in bike demand with working day and non-working day.
- There is no significant change in bike demand with any day of the week.
- The bike demand is high when weather is clear and Few clouds however demand is less in case of Light snow and light rainfall. We do not have any data for Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog, so we cannot derive any conclusion.

2) "drop_first=True" is important because it helps reducing the extra column created during creation of dummy variables. In turn it reduces the correlation among the variables.

3) Highest correlated numeric variables to the target variable 'cnt' are 'temp' and 'atemp'.

4) After building the model on the train dataset, I validated the assumptions by checking if the R-squared value is more than 0.70, similarly checked for the value of adjusted R-squared o be more than 0.60. Values of AIC and BIC are very less. Considered checking the multicollinearity of all the variables considered by checking VIF values of each variable to be less than 5. Values of test R-squared and train R-squared are almost same. Also the mean squared error when plotted is around zero and its value is very less.

5) Top three features which are contributing significantly towards explaining the demand of the shared bikes are listed as follows,

- weathersit_Light rain_Light snow_Thunderstorm
- Year
- season_spring

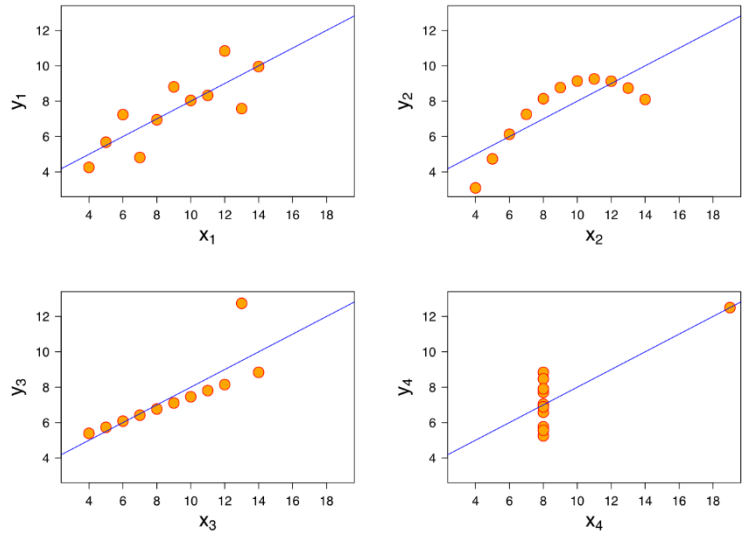
General Subjective Questions

Answers

1) It is a machine learning algorithm based on the output variable to be predicted is continuous and further is sub-divided into supervised learning where the past data with label is used for building the model. It achieves the prediction by performing regression tasks. It models a target variable based on independent variables. It is mostly used for finding out the relationship between the variables and forecasting. Different regression models differ based on, the kind of relationship between dependent and independent variables that they are considering and the number of independent variables being used.
Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.
We start the procedure by preparing the data as required by visually exploring the data. Then the data is split into train and test. Then we rescale the train data using one of the scalers and we fit it on the train data. After performing all the above-mentioned steps, the most significant independent variables are chosen if multiple. The constant is added, if necessary, based on the python package being used. Then we try to find the “best-fit” regression line using “Gradient descent” and is done so by ‘reducing’ the “cost function”.
Then the same variables are chosen from the test dataset and the numeric variables are transformed according to the model. Then the residual analysis is performed on the model and then the model is validated. There are multiple steps of validation. Some of the validation methods are listed as follows,

- Checking multicollinearity of independent variables using VIF.
- Checking the R-squared and adjusted R-squared values.
- Checking AIC and BIC values.
- Checking Homoscedasticity

2) Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x, y) points. They were constructed by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analysing it and the effect of outliers on statistical properties. It can be seen in the graphs below, that: -



- In the first one (top left), the scatter plot seems to be a linear relationship between x and y.
- In the second one (top right), there is a non-linear relationship between x and y.
- In the third one (bottom left), there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
- The fourth one (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

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) Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative. The Pearson coefficient correlation has a high statistical significance. It looks at the relationship between two variables. It seeks to draw a line through the data of two variables to show their relationship. The relationship of the variables is measured with the help Pearson correlation coefficient calculator. This linear relationship can be positive or negative. The Pearson's correlation coefficient varies between -1 and +1.

The formula of Pearson's correlation is as depicted below

$$r = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

4) Scaling is a technique where the numeric values of the continuous variables are assigned new values based on the scaling technique being employed. It changes the size of the values of the variables unaltering the effect in any manner on the dependent variable of those scaled variables. It is important because, if there is a vast difference in the range of few independent variables, say few ranging in thousands and few ranging in the tens, and it makes the underlying assumption that higher ranging numbers have superiority of some sort. So, these more significant number starts playing a more decisive role while training the model. The machine learning algorithm works on numbers and does not know what that number represents. So, for nullifying any sort of issues stated above, scaling is necessary.

Normalized scaling is typically rescaling the values between [0,1] so that the max value is 1 and minimum value is 0 in the entire set of values.

Standardized scaling is rescaling the values to have the mean as zero and standard deviation as 1 of the entire set of values.

5) The variance inflation factor (VIF) takes the value as infinity if the there exists a perfect correlation between any two variables or more considered in other words due to the presence high multicollinearity. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

6) Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. The purpose of Q-Q plots is to find out if two sets of data come from the same distribution. A 45-degree angle is plotted on the Q-Q plot, if the two data sets come from a common distribution, the points will fall on that reference line. Q-Q plot, is a graphical tool to help assess if a set of data credibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.