Summary

The Lead Score Case Study is a good experience to understand how a Logistic Regression model can help to solve a big business problem. The X Education can identify the promising customer without much effort. The company can cut the unnecessary cost spending on making phone calls to everyone and utilise the sales team effectively during on and off the phone calls.

I have started the analysis by data Inspection to understand the number of entries, columns and data type. Data cleaning is done to check the null values and drop the columns if the null values are greater than 40%. Assigned the binary values to binary variables as well as dummy values to categorical variables. To train the data in order to understand the hidden pattern of the data, the data set has been divided into train and test. Scaling is employed in order to standardise the data to avoid the impact of extreme values during the analysis. Selection feature such as Recursive feature elimination (RFE) is used to select the most appropriate feature/variables for the analysis. The selected features have been fed into the GLM () function or runned the Logistic Regression on selected features. The model is able to provide the coefficient and P-values of each variable and dropped such columns where the P-values are higher than 0.05. Also, Variance Inflation Factor (VIF) has been employed to avoid the highly correlated variables by obtaining the respective VIF values. I have ensured that VIF of all the selected variables are less than 5. Prediction has done upon the actual values in terms of their probability.

To assess or evaluate the model, the required probability must be obtained with the help of assumed probability. Initially 0.5 probability was assumed to obtain the scores of Accuracy, Sensitivity and Specificity later the actual probability was obtained with the help of cut-off values of all the three measure respectively. It is found that the actual probability was 0.4. Model evaluation is done with the help of the confusion matrix.

The following measures are used to Evaluate the performance of the model:

- Accuracy
- Recall
- Precision
- Sensitivity
- Specificity

The model would able to provide around 80% of True Positive result. Evaluation is done for both train and test data. The Recall rate of both train and test data is about 80%. The lead scores for each of the customers is generated.

The total data set is divided into two parts such as ‘Hot Leads’ and ‘Cold Leads’. The ‘Hot Leads’ are the leads which has Lead Scores more than 85. It is found that 336 entries were identified as ‘Hot Leads’. I can recommend the company that the 336 customers can be converted easily to business. Also, I could able to identify the important features contributing towards high lead score. I have learnt that most of the conversion had happened by those customers who visited the website very frequently. I believe that the customer more depends upon the content of the website to make their decisions.