Lead Score Case Study Summary

Problem Statement:

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses. Their company needs help in selecting most potential leads which are likely to convert into paying customers.

The company needs a model where a higher lead score is assigned to a customer who is more likely to convert and low lead score is assigned to a customer who is less likely to convert.

The CEO has given a ballpark of target lead conversion rate to be around 80%.

Solution summary:

Step1: Importing data and Inspecting dataframe

Read and analysed the data.

Step2: Data Cleaning

Handled duplicate rows and columns with no analytical significance by dropping them. Some columns are imputed, while other columns with high percentage of null values which are not useful for analysis are dropped.

Step3: Exploratory Data analysis

Performed Univariate and Bivariate analysis to understand data and found the variables which are significant and which are redundant. Also, we checked for outliers and imbalance, the converted column looked balanced.

Step4: Data Preparation

Outliers are properly treated and dummy variables are created. Splitting data into Train data set and Test data set with 70:30 proportions and then performing feature scaling using standard scaling.

Step5: Model Building

Feature selecting is done using RFE (Recursive Feature Elimination) and 20 top features have been selected. From the statistics generated, we recursively looked at p value and VIF value in order to select the most significant values that should be present in the final model and the insignificant values that are to be dropped. Finally we have 15 features, with p values less than or equal to 0.001 and VIF is less than 2.

Step6: Model Evaluation

Building ROC curve is plotted between True Positive rate and False positive rate. The area under ROC curve is 89%, which is good for the model. The best probability suggested using our code is 0.535 (53.5 in terms of lead score) as this has the best accuracy among the probabilities that have precision greater than 80%. A graph has been plotted to show the precision and recall optimal cut-off at 41.5, similarly another graph depicting the optimal probability cut-off based on the point of intersection of accuracy, precision and sensitivity. Based on the above model, we derived confusion matrix and specificity-89.8%, sensitivity-68%, accuracy-81.5% and precision-80.4% as well.

Step7: Conclusion

We used the derived model to predict on test data set, train data set and compared them to original values by creating a confusion matrix. The metrics for the test data set are not so far from the train data set. Hence we can say that the model is a good fit. As we achieved our specified result, this model is suggested to the X-Education company to be implemented for them to achieve more conversions from leads.