Lead Scoring Case Study

Problem Statement

An education company named X Education sells online courses to industry professionals. The company markets its course on numerous websites and search engines like Google.

The company needs a model where you as a lead score is assigned to each of the leads such that customers with higher lead score should have a higher conversion chance and the customer with lower lead score have a lower conversion chance.

The CEO of the company, has given a ballpark of the target lead conversion rate around 30%.

Summary

There were many steps used for building this model. Let us see following are the steps that involved during building model is as follows:

Step1: Data Cleaning and Preparing Dataset:

The data was partially clean. There were many Null values; Duplicates were present in the data. We have checked for the first few rows also that it shows how data looks and spread.We have replaced all the Null values with appropriate methods, does the correct imputation for categorical Columns. We have also dropped the columns which were not necessary for the model evaluation or which were highly imbalanced.

Step2: Exploratory Data Analysis:

In this step Exploratory Data Analysis was done to check the condition of the data.we performed bivariate analysis with converted coulmns to see the relation how leads are related to columns. The numeric values seems good and no outliers were found. Also plotted maps to identify the columns which are related.

Step3:Feature Scaling:

At this stage we converted all the categorical column into a numerical value.

Step4: Splitting the Data into Train and Test Dataset:

In this step the dataset in the Test and Train sections with a proportion of 75%(75.2%) around 3% short of the prediction made on traing data set.

Step5: Precision Recall:

This method was used to recheck and a cutoff of 0.45 was found with precision around 77.9 % and recall around 75.2 % on the test data frame.

CONCLUSION

While we checked all the parameters as well as Precision and Recall metrics well we have considered the optimal cutoff based on sensitivity and specificity for calculating the final prediction. Accuracy, sensitivity and specificity values of test set are around 77%, 75% and 79% which are approximately closer to the respective values calculated using trained set. Also the lead score calculated in the trained set of data shows the conversion rate on the final prediction model is around 79%. Hence overall this model seems to be good.