Lead Scoring Case Study

Problem Statement

An education company named X Education sells online courses to industry professionals. Through marketing and past referrals, X education acquires the information of leads.
The aim of this case study is to identify 'hot leads' - those leads that are most likely to convert into paying customers, so that X education can target these promising leads to ensure a high conversion rate.

Approach

The dataset provided contains the information of past leads like last activity, occupation, etc. and also whether or not that lead has converted. We began the data cleaning process by removing columns that had a very high null percentage and also those columns that do not have enough variance. Then we imputed the null values with either mode or median, depending on the column datatype.

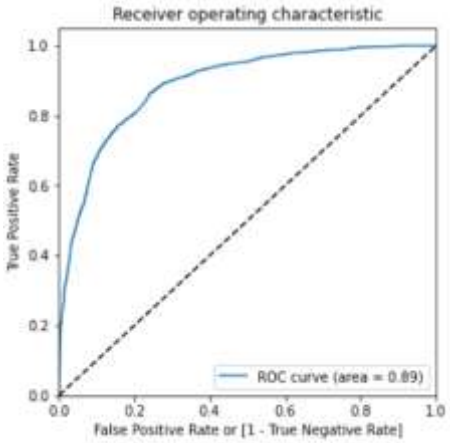
We then performed Exploratory Data Analysis to check for imbalance in the target and also to find some interesting patterns. We observed some patterns such as low conversion rate for people who have not provided details of their occupation and specialization.

The next step was to convert all features into numeric format by creating dummy variables. After doing so, we identified the highly correlated columns and dropped them. After which we had 62 columns.

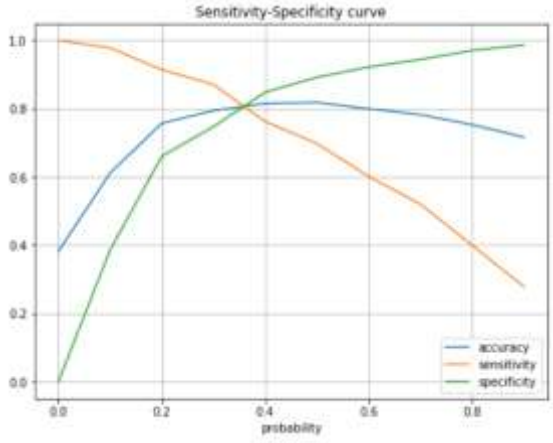
We then split the data into train and test datasets with a 70:30 ratio and performed feature scaling using MinMaxScaler. We proceeded to build the model by first reducing the number of columns from 62 to 25. This was done using recursive feature elimination(RFE).

We then built a model using features selected by RFE, identified the columns having p-value >0.05 and VIF >5 and dropped them. In this manner, we arrived at the final model with 17 columns having 80% accuracy when the cut-off probability is 0.5.

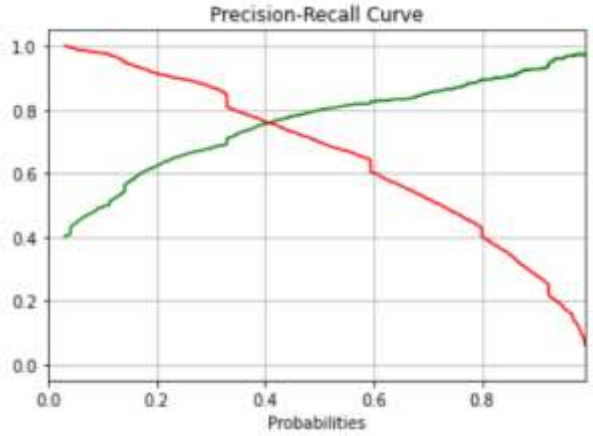
To check the goodness of the model, we plotted the ROC curve.



We see that area under the curve is high indicating that the model is performing well. To find the correct value of the probability cut-off, we plotted the sensitivity-specificity curve.



Looking at this plot, we have chosen the cut-off to be 0.35. At this point, the values of all three metrics are close to 0.8, which is a good indicator. Leads having conversion probability >0.35 will be classified as 'Hot leads'. We also looked at the Precision-Recall curve, which intersected at 0.4.



However, the business requirement was being met using 0.35 as the cut-off, hence, we didn’t change the cut-off value.

Conclusion and learnings:

We built a logistic regression model with sensitivity, specificity and accuracy close to 80%. The model assigns a score between 0 and 100 for each of the leads, where a higher score is assigned to leads who visit the website frequently, spend more time on the website, but view less pages in each visit, are working professionals, and originated via Lead Add Form. Also, it is observed that leads that are sourced through "Welingak Website" are more promising, so this website can be promoted to attract more 'Hot leads'.