Lead Scoring Case Study Summary

Problem Statement:

An education company named X Education sells online courses to industry professionals. The typical lead conversion rate at X education is around 30%.

To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

The Objective of the problem is to build a model that will assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Solution Summary

Reading and understanding the data:

- Identifying Categorical and Numerical columns.
- Number of null and other redundant values present in the data such as many of the categorical columns were containing "Select" as a value.

Data Cleaning

- During data analysis it was found that a lot of categorical columns containing Yes/no as values were highly imbalanced with high percentage of the values as "No". Such Columns having "No" as response for more than 90% rows were dropped.
- Columns having "NAN" percentage grater than 45% were dropped. Columns for which sum of "NAN" and "Select" was grater than 50% were dropped.
- For the remaining columns the "NAN" were replaced by mode.
- For Some of the Categorical columns the no of categories were very high, with some of the categories having only one count, categories in such columns with count less than 100 were clubbed together and named "Others".
- For Columns having a significant no of values as "Select", the distribution of target variable for rows having "Select" was different from the distribution of target variable for mode, So "Select" was replaced by "Not Mentioned".
- Outliers were replaced by "Q3+1.5(IQR)"

Data Analysis

- Exploratory data analysis of the data set was done to find visible trends and get the feel how the data is oriented.

Data Preparation

- Min Max scaler was used for scaling the numerical Columns.
- The dataframe was split into X and Y ,
- Columns having two categories were mapped to 0 and 1, and Columns having more than two categories were split into dummies, finally the two dataframes were Merged and Duplicate columns dropped
- Correlation plot was plotted and Highly Correlated Dummy features were dropped.
- The X as well as y were split in train and test set in 70:30 ratio.

Model Building

- Statsmodel.GLM was used to build the first model and it was noticed that a lot of features are having high significance level. RFE was used for feature elimination with no of features as 18.
- After RFE manual Feature elimination was done and features having significance level grater than 5%, and VIF greater than 5 were dropped.
- The final model is having 15 features with 81.5% accuracy at .49 cut-off probability.
- Finally ROC curve was plotted to optimize the cut-off probability.
- With optimal cut-off probability of .34, the Accuracy, Specificity and Sensitivity are all around 80%.

Metrics

Based on the Objective of the analysis Precession of the model was used to arrive at the best model

- The Precession for optimal cut-off value of .34 is 72% , which is much lower than the required 80%.
- Precession-Recall trade-off: Precession and Recall are both equal to 75% at .42 cut-off probability, precession is still less than 80% mark:
- At a cut-off probability of .49, the precession is 78.9 on train set, but recall /Sensitivity drops to around 70% .
- on test Set the model gives Accuracy of 81.6%, Specificity of 88.7%, Sensitivity of 68.7% at cut-off probability of .49 The Precession on test set is grater than 79.8% so the model is not overfitting on training set and as it is giving a Precession of around 80% on test set this model with cut-off of .49 was chosen
- Finally a new lead can be fed into the model and its “Lead Score” should be checked if lead Score is greater than 49 it’s a “hot lead”.