Problem Description

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.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Now, although X Education gets a lot of leads, its lead conversion rate is very poor. 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. A typical lead conversion process can be represented using the following funnel:

The company requires to build a model wherein we need to 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%.

Approach

1. The I/P data has been thoroughly analyzed and cleaned for cases of
    -   Null values
    -   Unwanted Columns
    -   Outliers
2. EDA visualization has been performed on
    -   Categorical univariant
    -   Numerical univariant
    -   Bi-Variant – Numerical – Numerical
    -   Bi-Variant – Categorical to converted analysis.
3. Performing model building pre processing steps
    -   Converting Binary Categorical variable to 0 and 1
    -   Split multivalue categorical variables into dummy variables
    -   Drop unwanted columns including original multivalue categorical variables
4. Training Model
    -   Split Data into Train and Test
    -   Feature Scaling
    -   Check for Conversion ratio in train model
    -   Look for Corelation
    -   Feature Selection using RFE
    -   Building model using Logistic Regression
    -   Checking for P-Value and VIF to select appropriate features

- Predicting probability on train set and checking accuracy, simplicity and specificity ratio
- Plotting ROC curve
- Finding optimal Cutoff point
- Checking precision and recall tradeoff
- Making final prediction on test set
- Comparing accuracy, simplicity and specificity of test results with train results

Conclusion

As per model, below 3 features plays important role

1. Lead Origin - lead add form
2. What is your current occupation - working professional
3. Lead Profile - potential lead