Lead Scoring Case Study

The company requires a model wherein a lead score is assigned 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 dataset is read and inspected for basic sanity checks. Unimportant variables with no variance are dropped. The value select is imputed to null as it is almost equal to null. There missing values, outliers in the dataset which are imputed correctly with significant methods. EDA on the dataset for outlier check and univariate analysis and bivariate analysis is done. Binary variables are mapped to 0/1 for yes/no and dummy variables are created for categorical variables with multiple levels. The obtained dataset is scaled for fitting the model on a logistic Regressor. The model is built with train-test split of 70,30 and the model has gone through several eliminations based on the high p-value and high vif values, the final model obtained is obtained with significant p-values i.e., less than 0.05 and vif values less than 5. The model is fit on the test data and following above steps a final model is obtained. The model is evaluated using the evaluation metrics i.e., confusion matrix. The confusion matrix is generated for both train and test sets, accuracy, specificity and sensitivity are calculated.

We obtained the following metrics for the test and train sets:

Train set:Accuracy:81.6, Sensitivity: 69.9, Specificity:88.9

Test Set: accuracy:81.7, sensitivity:73.8, specificity: 86.4

ROC curve is plotted to check the tradeoffs between specificity and sensitivity, an optimal cut off of 0.3 is obtained from the model.

Precision and Recall metrics are calculated for the train dataset and the tradeoffs are mapped.

Precision Score:79.54

recall score:69.87

Precision and Recall metrics are calculated for the train dataset and the tradeoffs are mapped.

Precision Score:78.08

recall score:73.88

It was found that the variables that mattered the most in the potential buyers are (In descending order)

- Lead Origin_Lead Add Form
- What is your current occupation_Working Professional
- Last Notable Activity_Had a Phone Conversation
- Lead Source
    - Welingak Website
    - Olark Chat
- Total Time Spent on Website

The Model has achieved 80% accuracy required by the company.

The company should concentrate more on people who spent more time on the website.

The company can use alternative methods of contacting like email and sms after the goal is reached.

The company has to concentrate more on people who are working professionals with interest for career growth.

The people who had an chat or a phone conversation previously can be called again as there is achance for lead conversion.
The company can ignore calls to students as they are already studying and have no income.
The company can ignore calls to housewifes.
The company should concentrate more who often visit the website and receive sms or emails.