Lead Scoring Case Study

Group Members:

1. Abhiram S Bharadwaj
2. Amaljith AP
3. Gaurav Sharma

Problem Statement

An X Education need help to select the most promising leads, i.e. the leads that are most likely to convert into paying customers.

The company requires us to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with 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%.

Goals and Objective

There are quite a few goals for this case study.

- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.
- There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

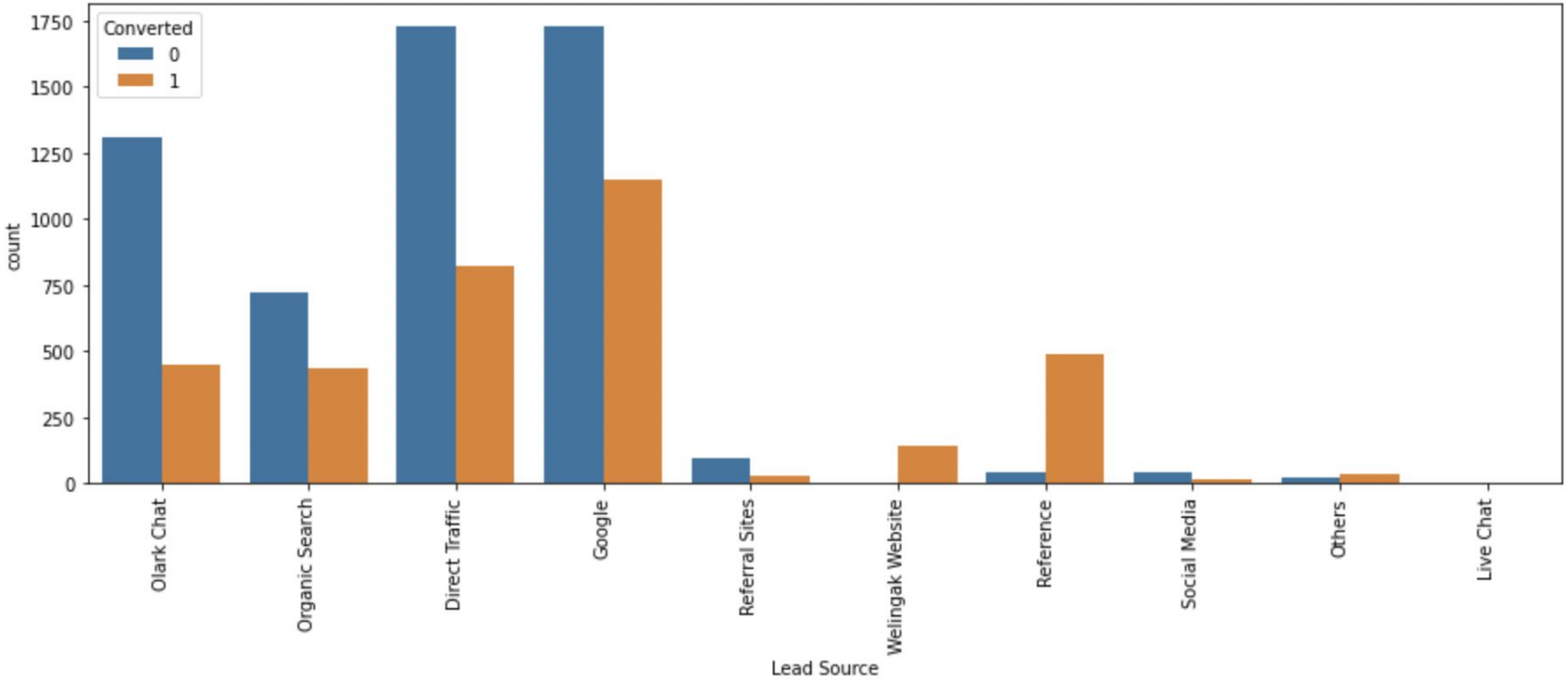
Solution Method

- Data Cleaning and Manipulation
  - Checked and handled duplicate null value and missing values.
  - Dropped all columns with more than 40% NULL data.
  - Dropped all columns which are not useful for analysis. (e.g. Prospect ID and Lead Number are just identifying features and are unique to each row. So we dropped).
  - Check and handled outliers in the data.
- EDA
  - Univariate data analysis.
  - Bivariate data analysis.
- Scaling & Dummy variables and encoding of the data.
- Model building and prediction using Logistic regression.
- Model Validation.
- Model presentation.
- Observation.

Date Manipulation

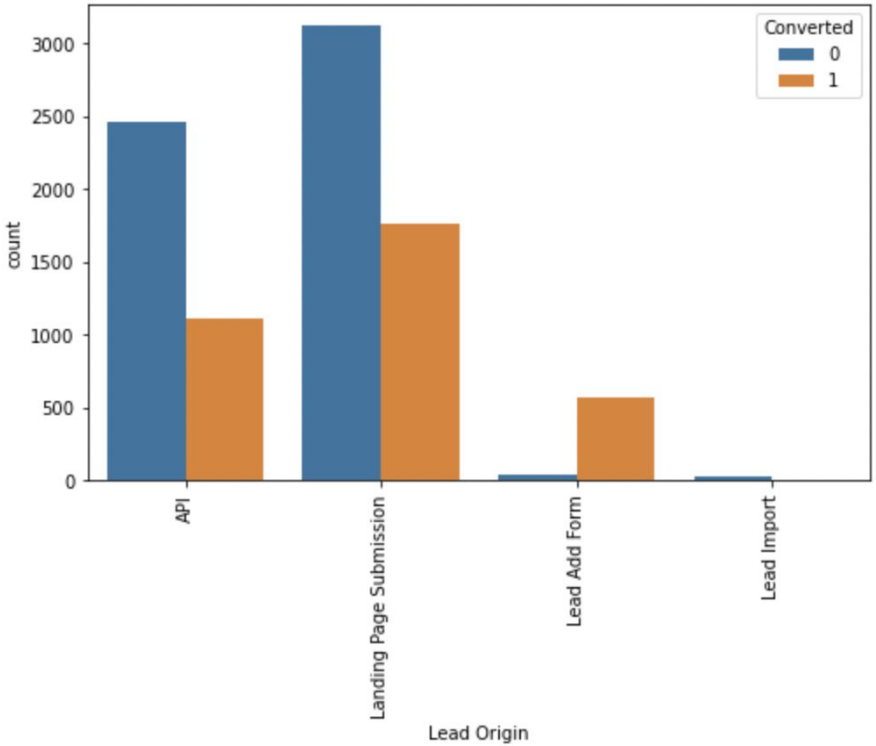
- Total number of rows are 37 and Total number of columns are 9240.
- Removing the “Prospect ID” and “Lead Number” which is not necessary for the analysis.
- Single value features like “Magazine”, “Receive More Updates About Our Courses”, “Update me on Supply”.
- 97% of the data in Country column belongs to India. This is abnormally skewed and this column can be dropped.
- Some of the columns have almost half of the data as null. Let us drop all columns with more than 40% NULL data : 'Lead Quality', 'Asymmetrique Activity Index', 'Asymmetrique Activity Score', 'Asymmetrique Profile Index', 'Asymmetrique Profile Score'.
- We have some data in-accuracies in Lead sources. We have 2868 entries for Google. We also have 36 Null values. Let us fix the inaccuracies and also group values with low number of occurance to one bucket.

EDA



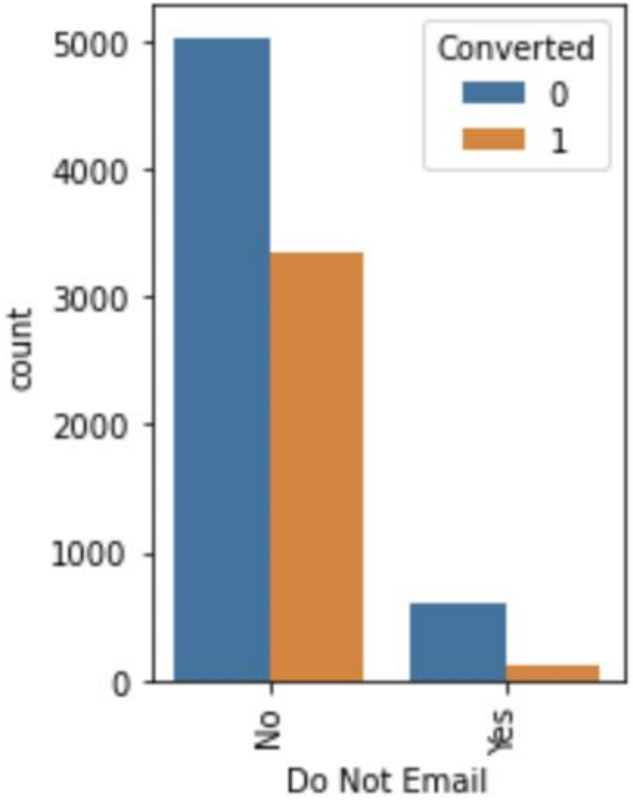
➢ Google and Direct Traffic respectively generate the most leads as well as conversions.

EDA



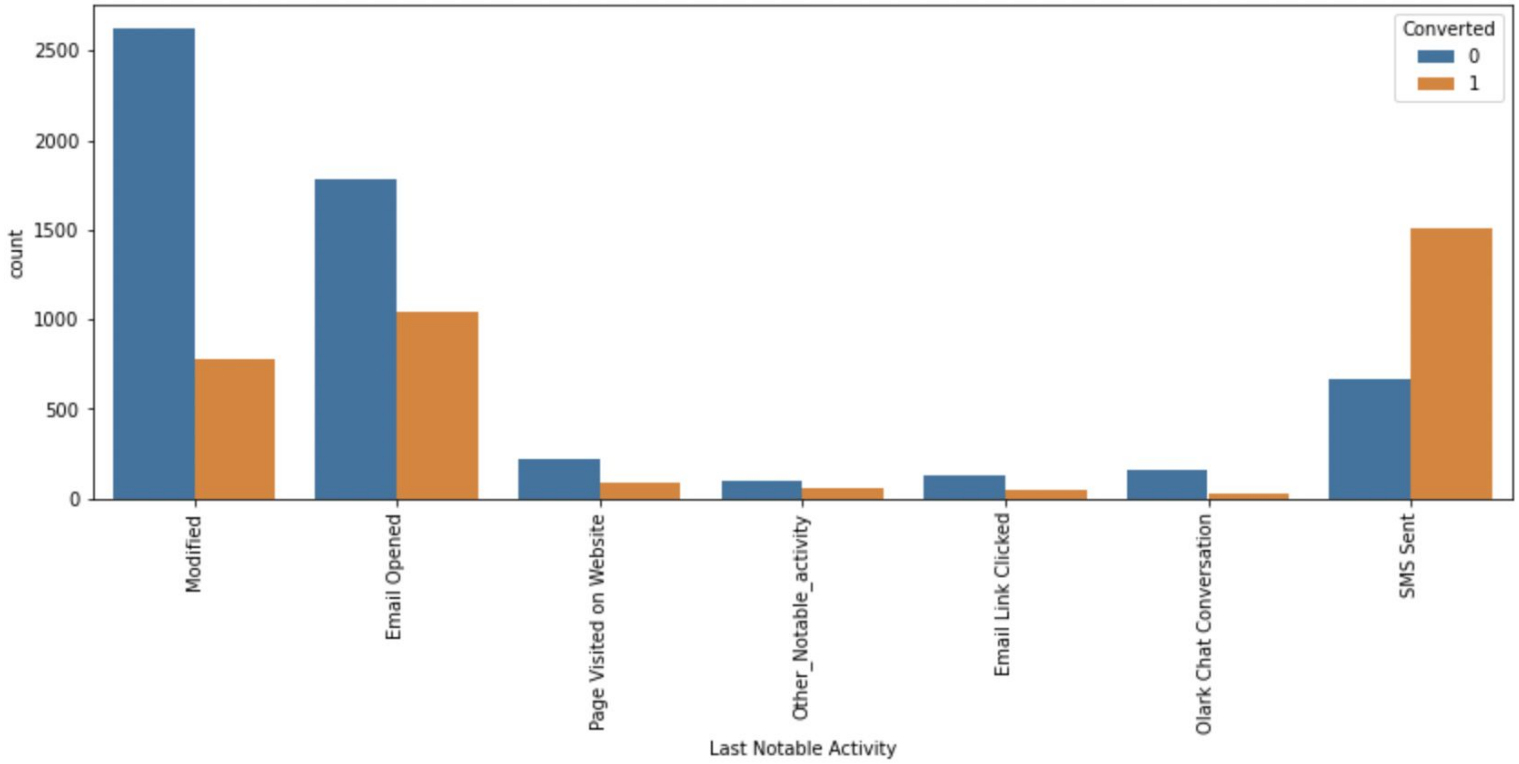
- Maximum number of leads seem to be coming from API and Landing page submission.
- %Conversion of Lead Add Form is extremely high.

EDA



- For the Do not Email feature more people answered with ‘No’ as compared with ‘Yes’
- There is higher conversion rate saying ‘No’ in comparison with people saying ‘Yes’ for Do not Email

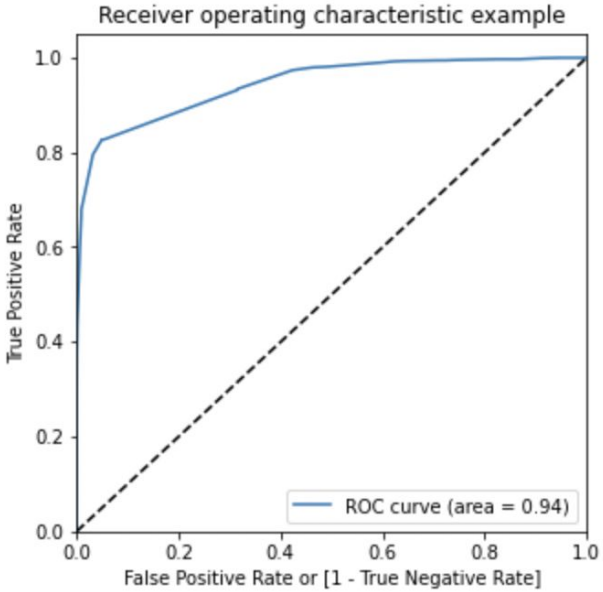
EDA

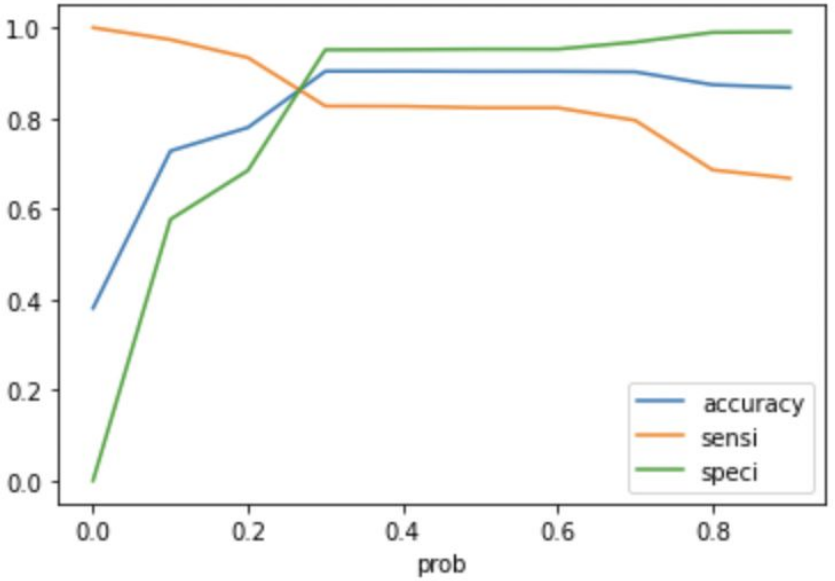

Converted
0
1
count
2500
2000
1500
1000
500
0
Modified
Email Opened
Page Visited on Website
Other_Notable_activity
Email Link Clicked
Olark Chat Conversation
SMS Sent
Last Notable Activity

Model Building

- Splitting the Data into Training and Testing sets.
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Using RFE for Feature Selection.
- Building model by iteratively removing the variable whose p- value is greater than 0.05 and vif value is greater than 5.
- Predictions on test data set.
- Overall accuracy over 90%.

ROC Curve





- The ROC Curve should be a value close to 1. We are getting a good value of 0.94 indicating a good predictive model.
- From the second graph it is visible that the optimal cut off is closed at 0.3.

Conclusion

Features/Variables that has the highest significance in predicting whether a lead is hot or not is listed below in descending order:

1. Lead Origin
2. Specialization
3. Lead Source
4. Last Activity
5. Last Notable Activity
6. Tags

A few other key points to keep in mind:

1. People who have a tag of “Will revert after reading email” have a very high chance of being converted and these people will need to be treated as priority 1 hot leads.
2. People whose phones are switched off or leave their phone ringing are extremely unlikely to enrol to any of the courses. So we can reduce the energy and effort spent to try and convert these leads.
3. Even though direct traffic and google sources have extremely high number of leads, the conversion rate is not satisfactory. Efforts should be made to improve the conversion rate for these sources which are extremely effective in producing leads.
4. People who have a tag of “Already a student” also have a high chance of not enrolling for another course. Actions need to be taken to better understand why this is so that the underlying issues can be addressed.

Thank You