Lead Scoring Assignment

Siddharth Dalal

Problem Statement

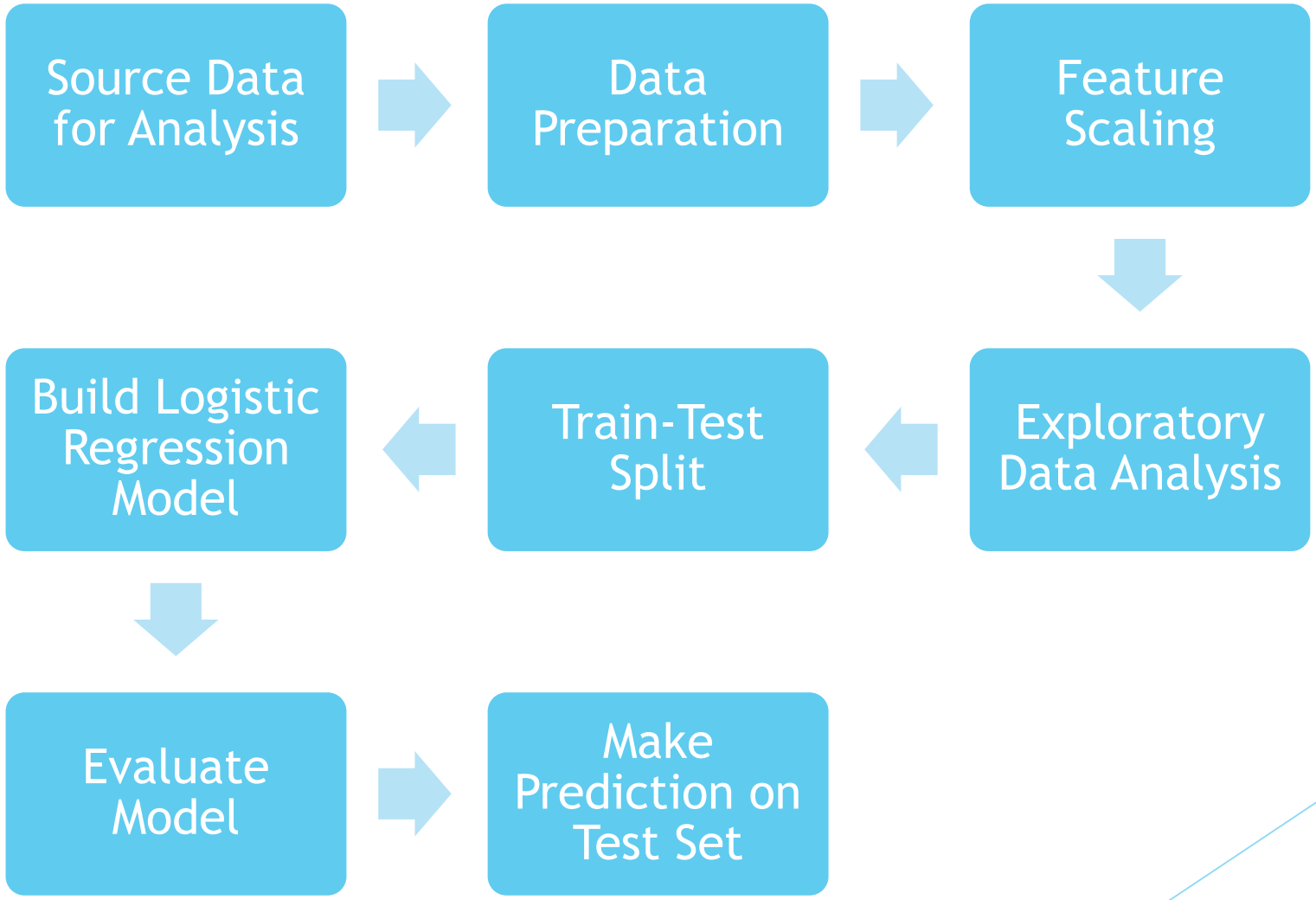
- An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.
- The company markets its courses on several websites and search engines like Google. Once these 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. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. 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.
- X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you 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

There are quite a few goals for this case study:

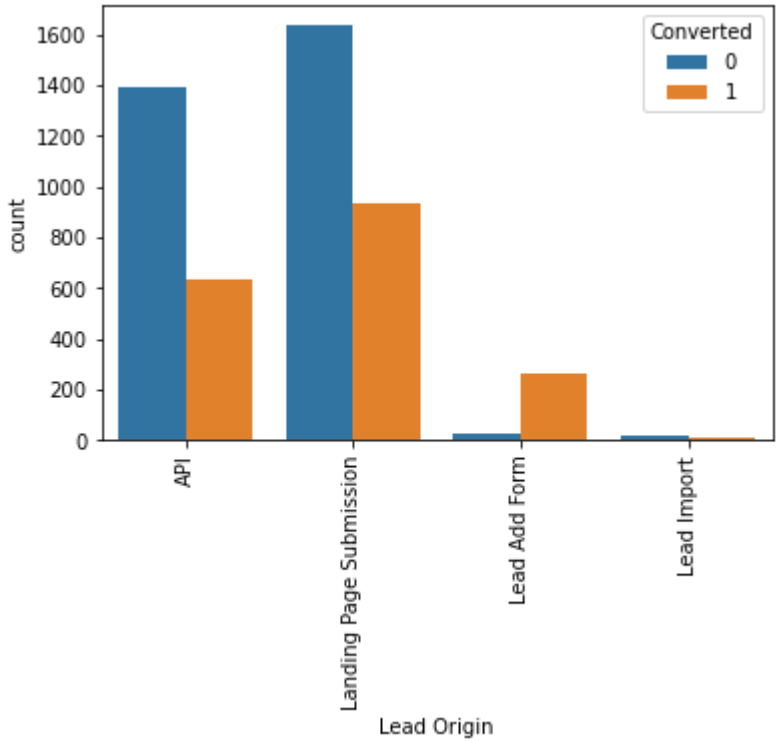
- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will 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.

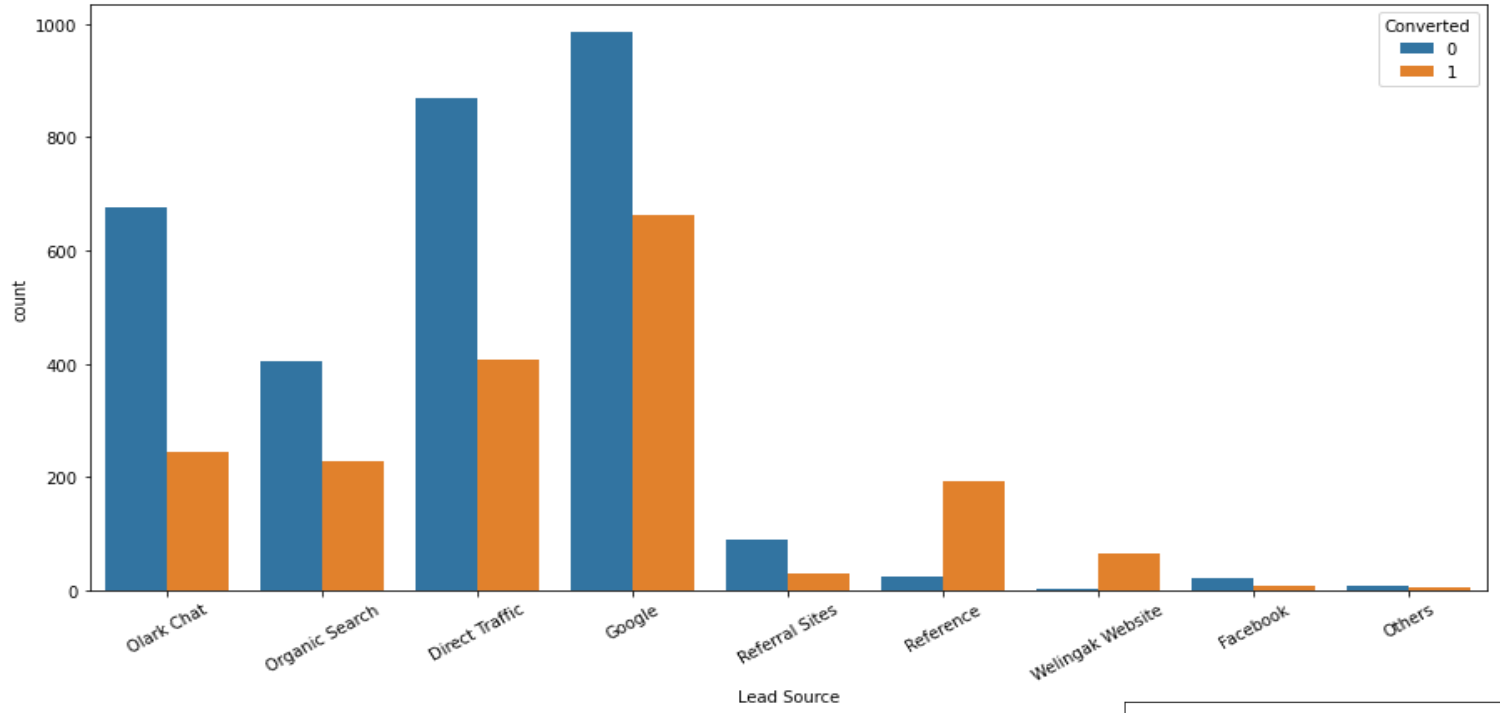
Approach


Source Data for Analysis
Data Preparation
Feature Scaling
Exploratory Data Analysis
Train-Test Split
Build Logistic Regression Model
Evaluate Model
Make Prediction on Test Set

EDA Inferences:

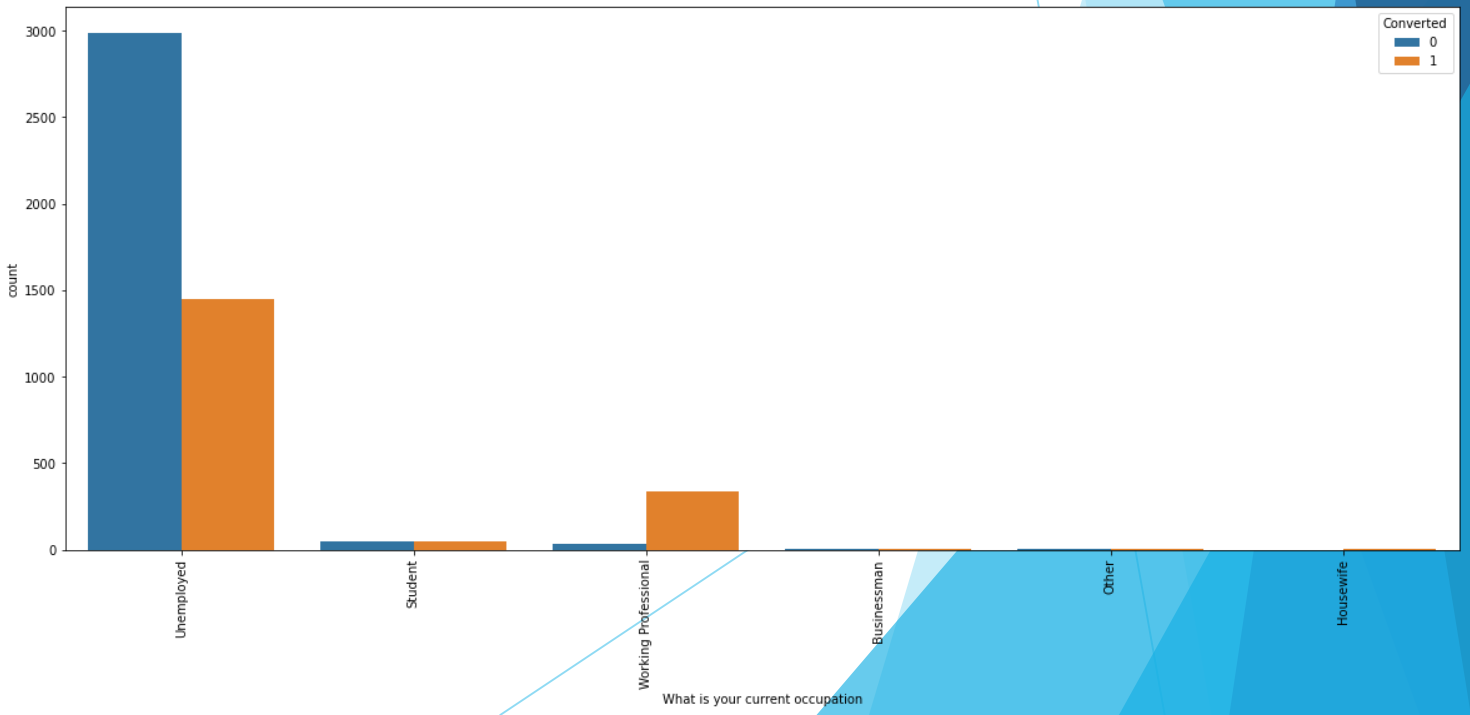
- We currently have a conversion rate of around 38%.
- Most leads are generated from API and Landing Page. They have conversion rate of 30-35%. Lead Add Form has more than 90% conversion rate but count of lead are not very high.



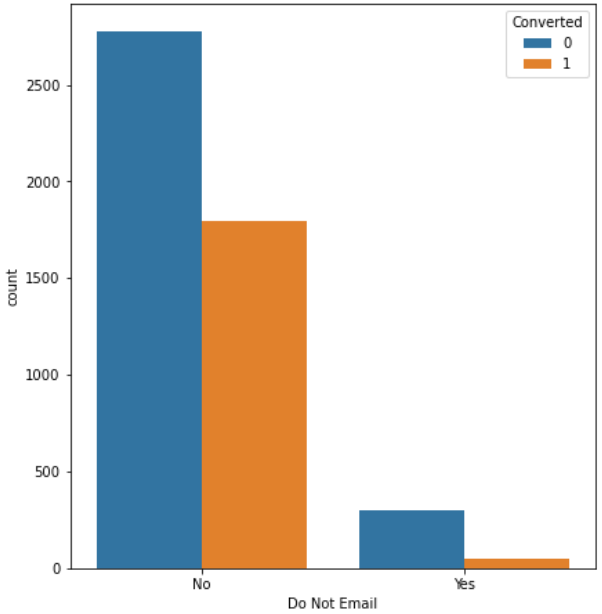


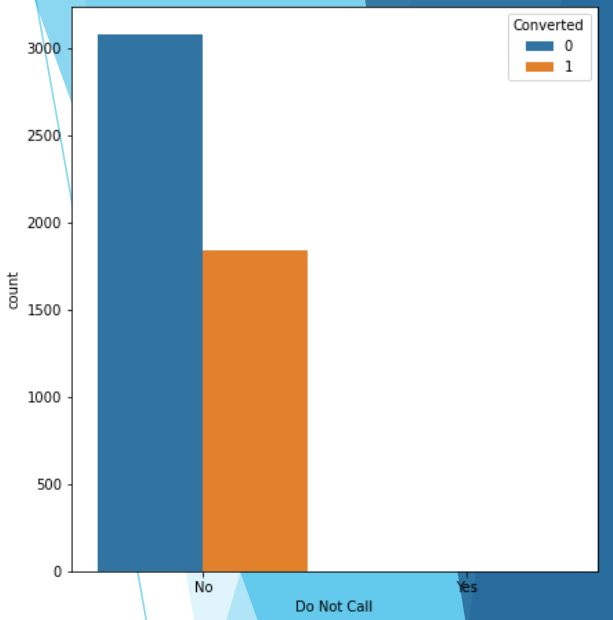
Google and Direct traffic generates maximum number of leads.

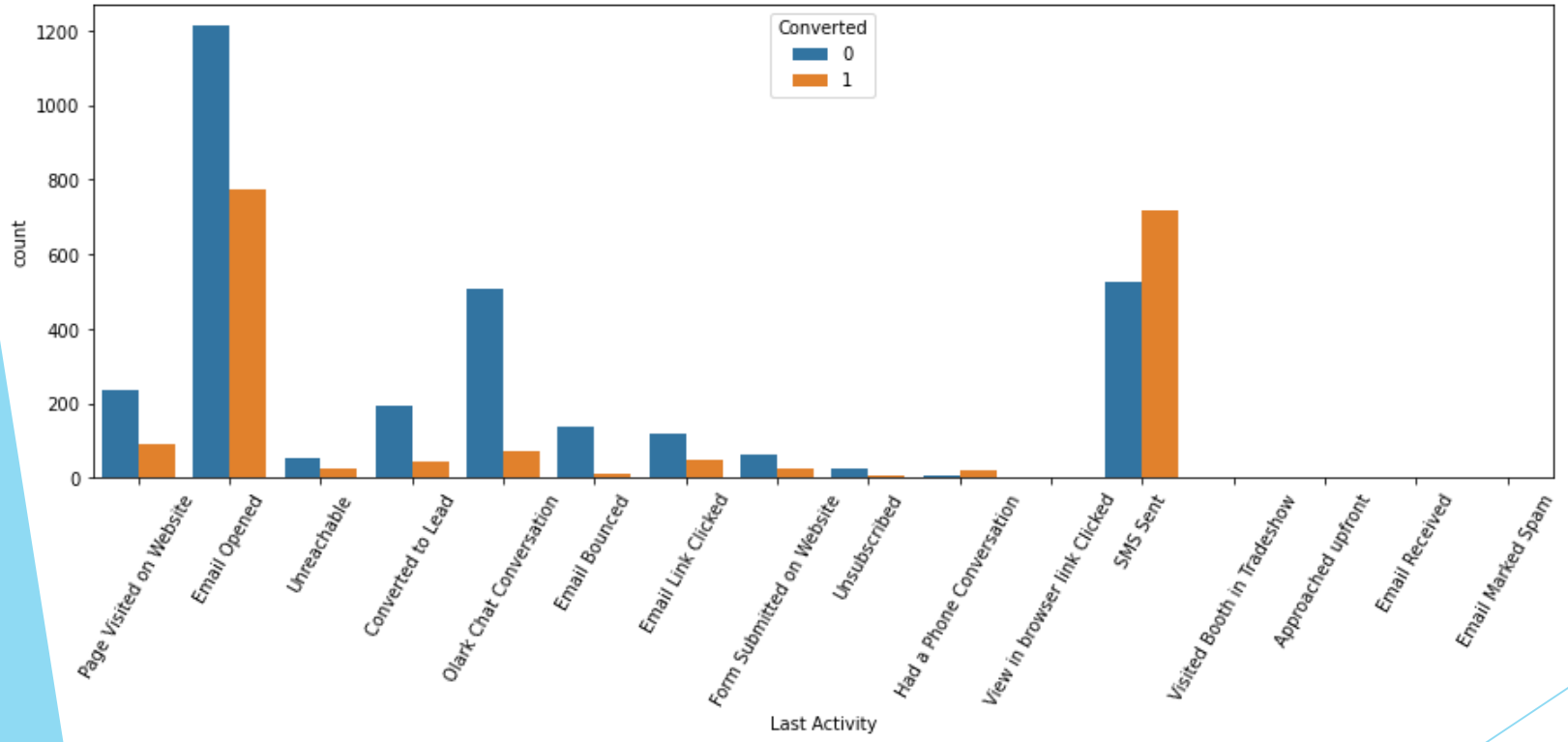
Unemployed leads are the most in numbers but has around 30-35% conversion rate. Working Professionals going for the course have high chances of joining it.



Major conversion has happened from Emails sent and Calls made

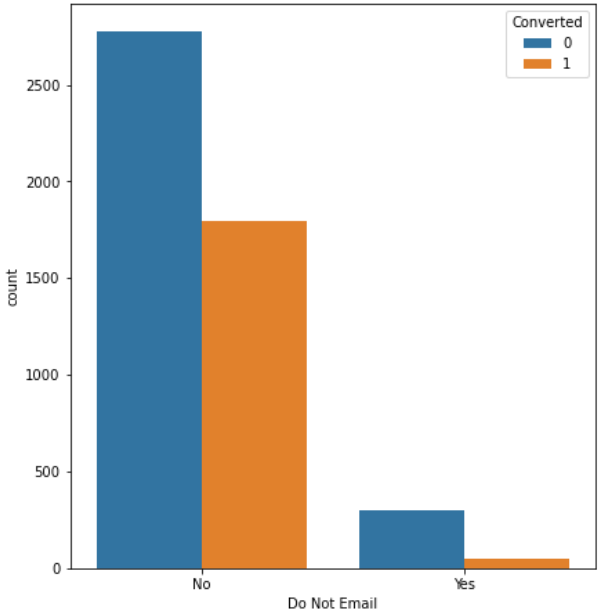


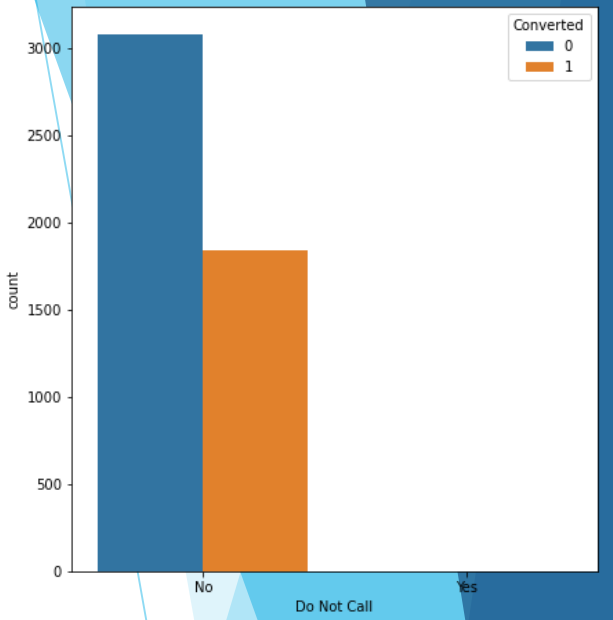


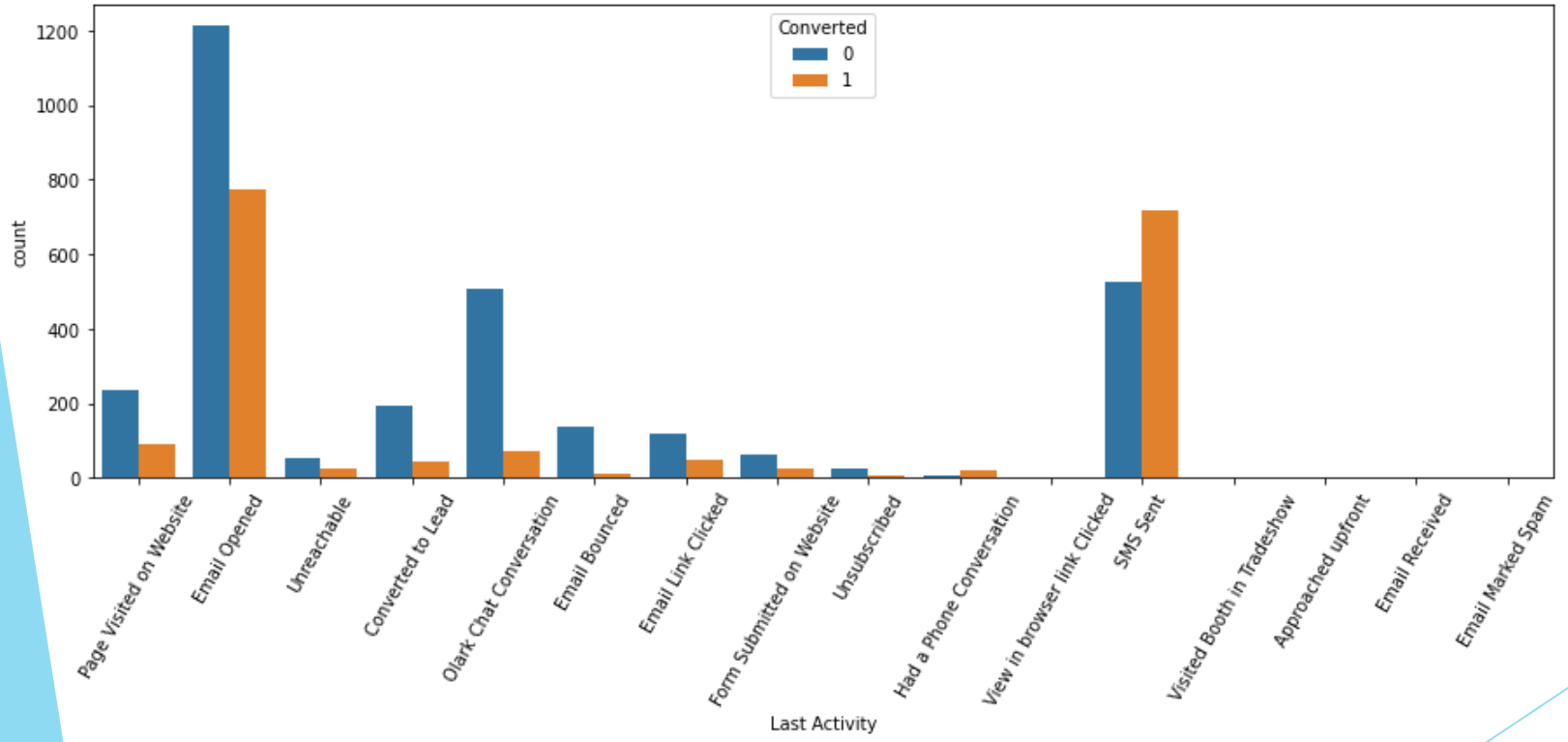


Last Activity value of SMS Sent' had more conversion.

Major conversion has happened from Emails sent and Calls made







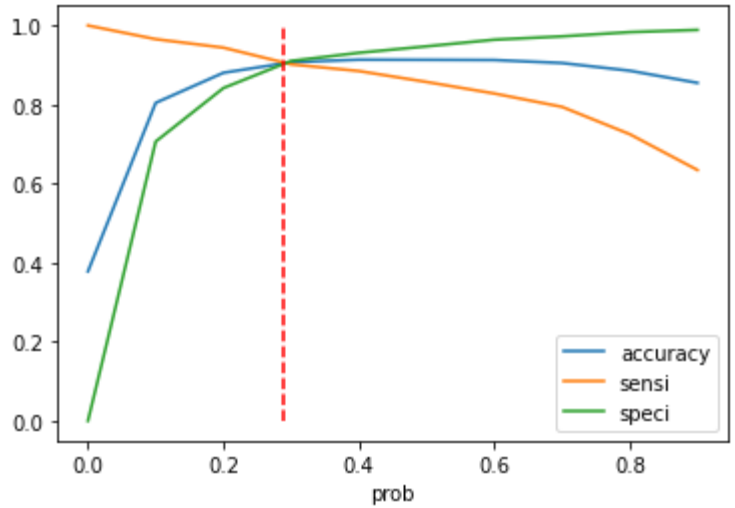
Last Activity value of SMS Sent' had more conversion.

Data Preparation and Feature Scaling

- Columns which have more than two levels were converted to dummies using pd.get_dummies function.
- Columns which have only two levels “Yes” and “No” were converted to numerical using binary mapping.
- RFE was used for feature selection to attain top 15 variables
- Applied min-max scaling on the variables ['TotalVisits', 'Page Views Per Visit', 'Total Time Spent on Website']

Model Building and Evaluation

- Data was split into train and test sets in the ratio of 70:30
- A logistic regression model was built by removing variables with high p values.
- Finally, an optimal model was obtained.
- 91% accuracy was achieved overall.
- Plotted ROC Curve
- Found Optimal cut off point to be 0.29



Conclusion

After running the model on the Train Dataset these are the figures we obtain:

- Accuracy: 91.3%
- Sensitivity: 85.70%
- Specificity: 94.7%

After running the model on the Test Dataset these are the figures we obtain:

- Accuracy: 88.1%
- Sensitivity: 92.76%
- Specificity: 85.4%

Conclusion

Top variables contributing to conversion:

Lead Source:

- Total Visits
- Total Time Spent on Website

Lead Origin:

- Lead Add Form

Lead source:

- Direct traffic
- Google
- Welingak website
- Organic search
- Referral Site

Last Activity:

- Do Not Email_Yes
- Last Activity_Email Bounced
- Olark chat conversation