Lead Scoring Case Study

Logistic Regression

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.

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.

The Initial conversion rate is 38.5%, and the CEO has given a ballpark of the target lead conversion rate to be around 80%.

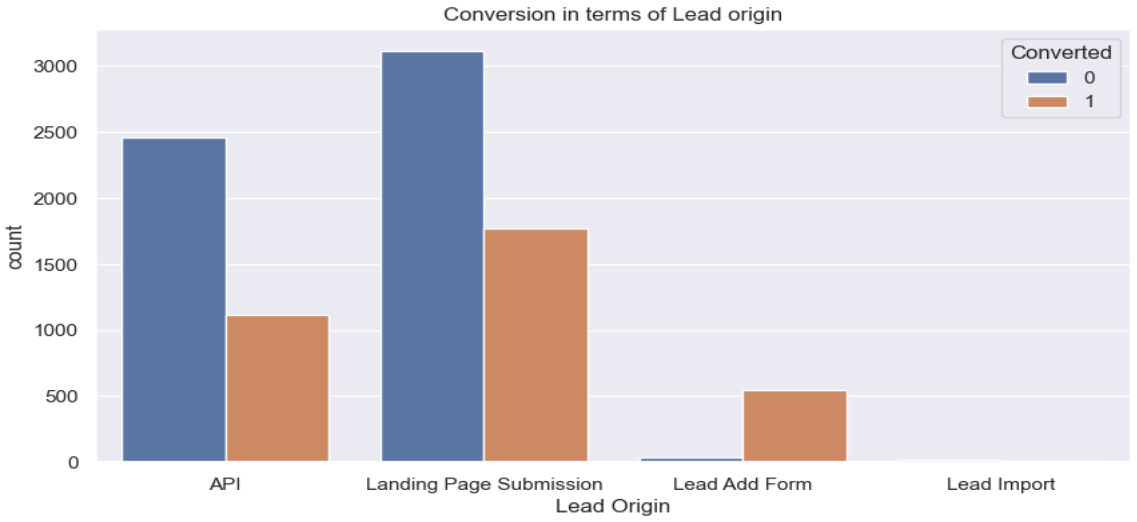
Objective

1. 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.
2. 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.

Strategy

- Import data
- Clean and prepare the acquired data for further analysis
- Exploratory data analysis for figuring out most helpful attributes for conversion
- Scaling features
- Prepare the data for model building
- Build a logistic regression model
- Assign a lead score for each leads
- Test the model on train set
- Evaluate model by different measures and metrics
- Test the model on test set
- Measure the accuracy of the model and other metrics for evaluation

EDA



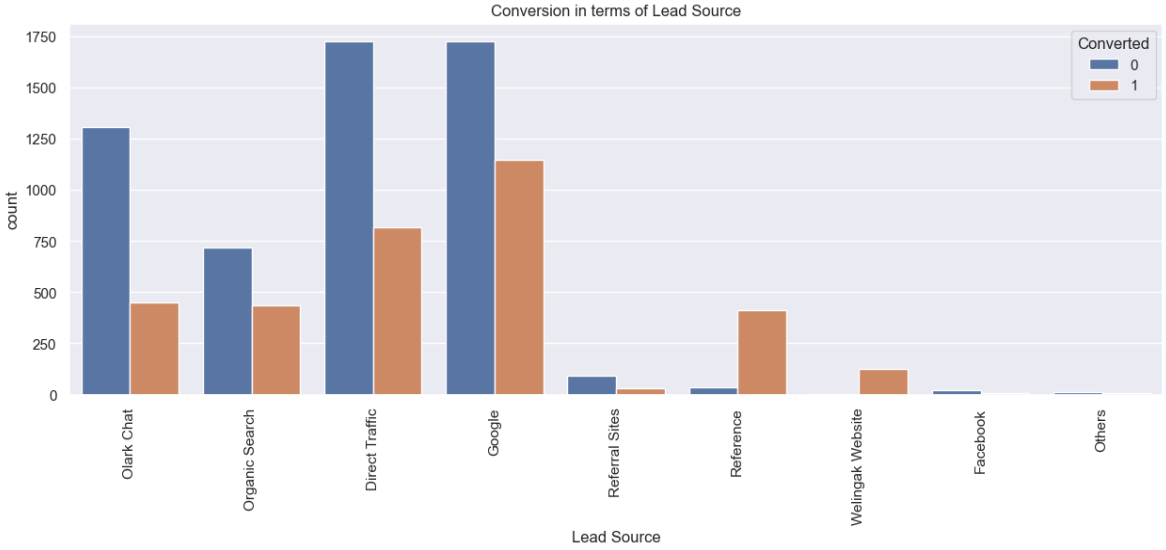
LEAD ORIGIN

Though less leads from Lead Add Form, it has the highest conversion rates

Lead Add Form & Landing Page Submission though has avg of 33% conversion, it generates the most no. of leads.

Lead Import has the least number of conversions and leads count.

To improve overall lead conversion rate, focus should be on improving lead conversion rate of API and Landing Page Submission. Also, generate more leads from Lead Add form since they have a very good conversion rate



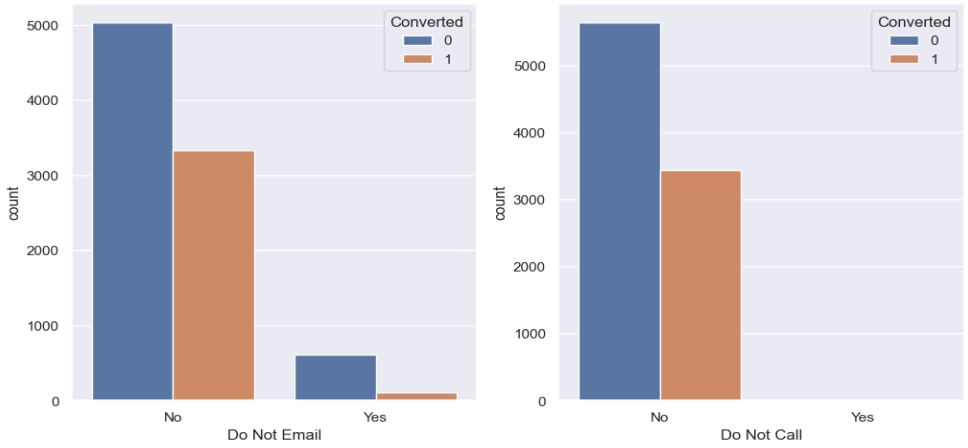
LEAD SOURCE

Google and direct traffic generates maximum number of leads but has conversion rate of 40% and 32% respectively.

Welingak website and References has highest conversion rates around 98% and 93% but generates a smaller number of leads.

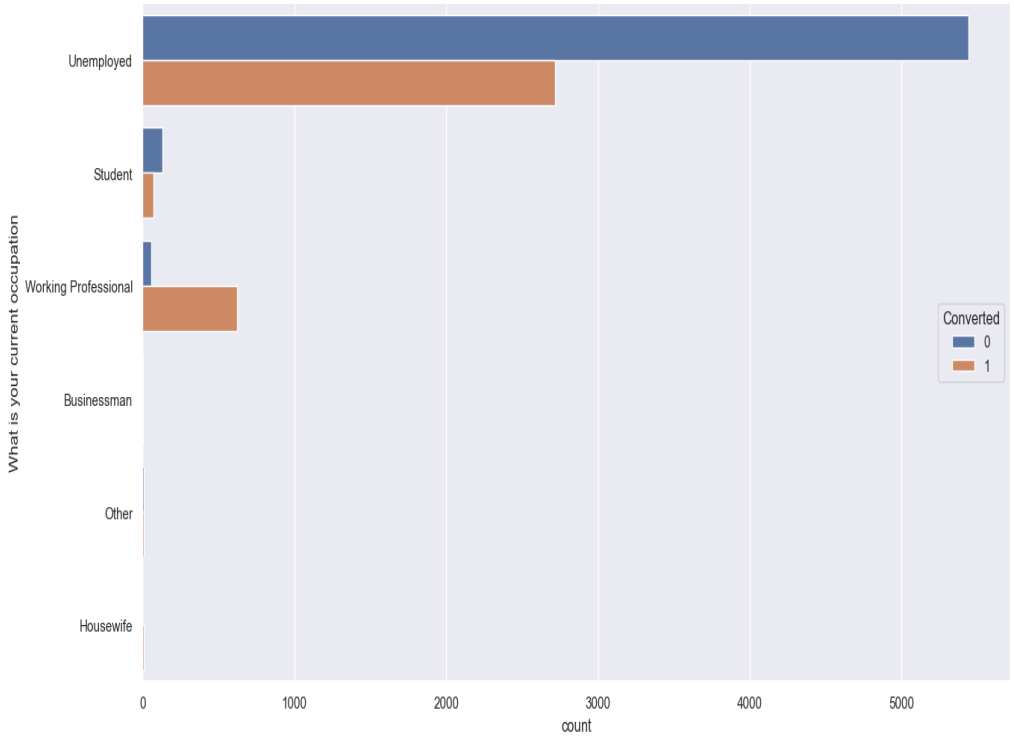
Olark chat and organic search has significant number of leads, but their conversion rate is low, around 26% and 38%.

To improve overall lead conversion rate, focus should be on improving lead conversion of olark chat, organic search, direct traffic and google lead source. Also, generate more leads from reference and welingak website since they have a very good conversion rate.



DO NOT EMAIL AND CALL

More than 95% of the leads do not prefer to be called or emailed.

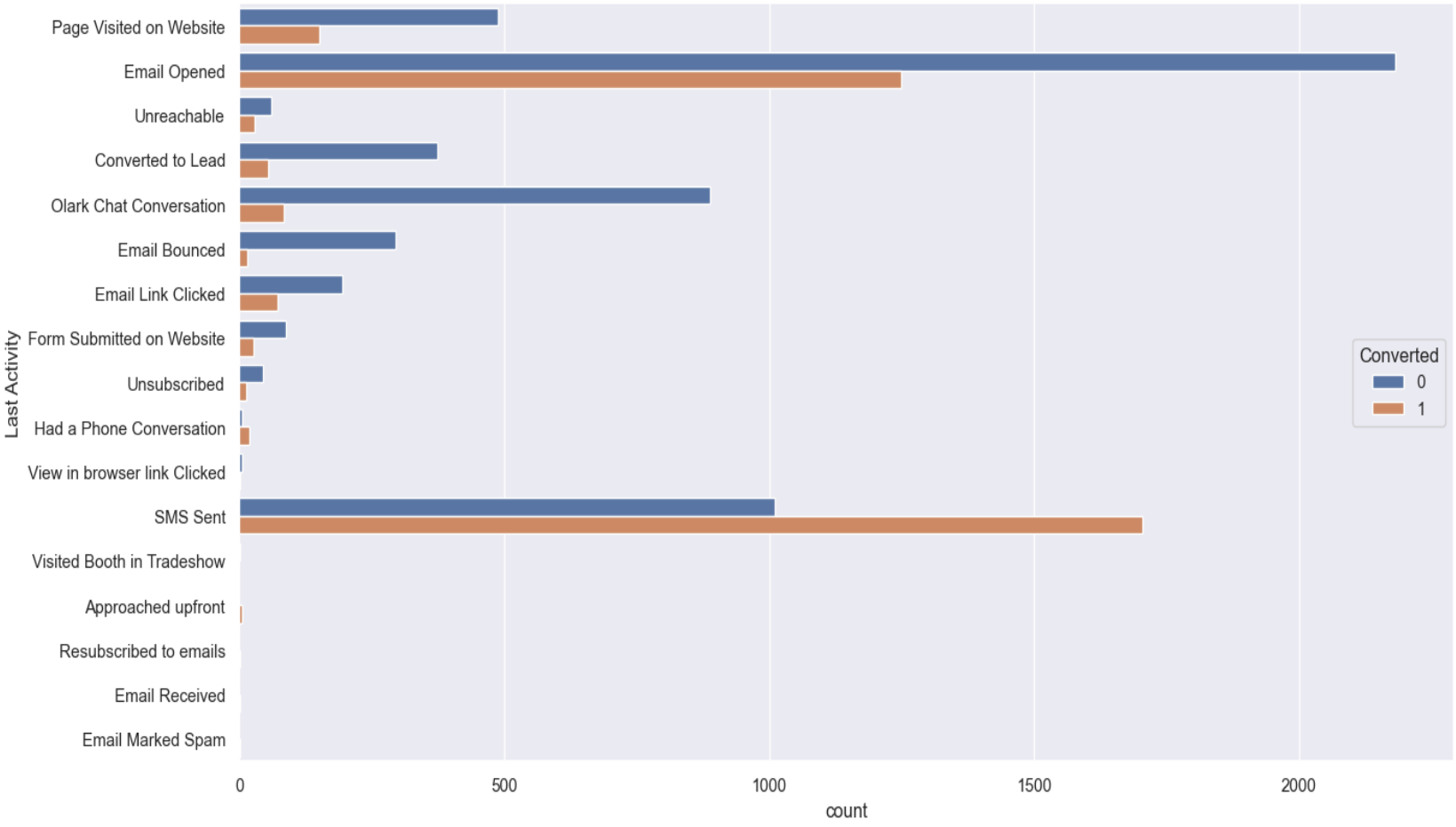


OCCUPATION

Working Professionals and Unemployed people generates maximum leads with conversion rates of 92% and 33%.

To improve overall lead conversion rate, focus should be on improving lead conversion of unemployed. Also, generate more leads from Working Professionals and Housewives.

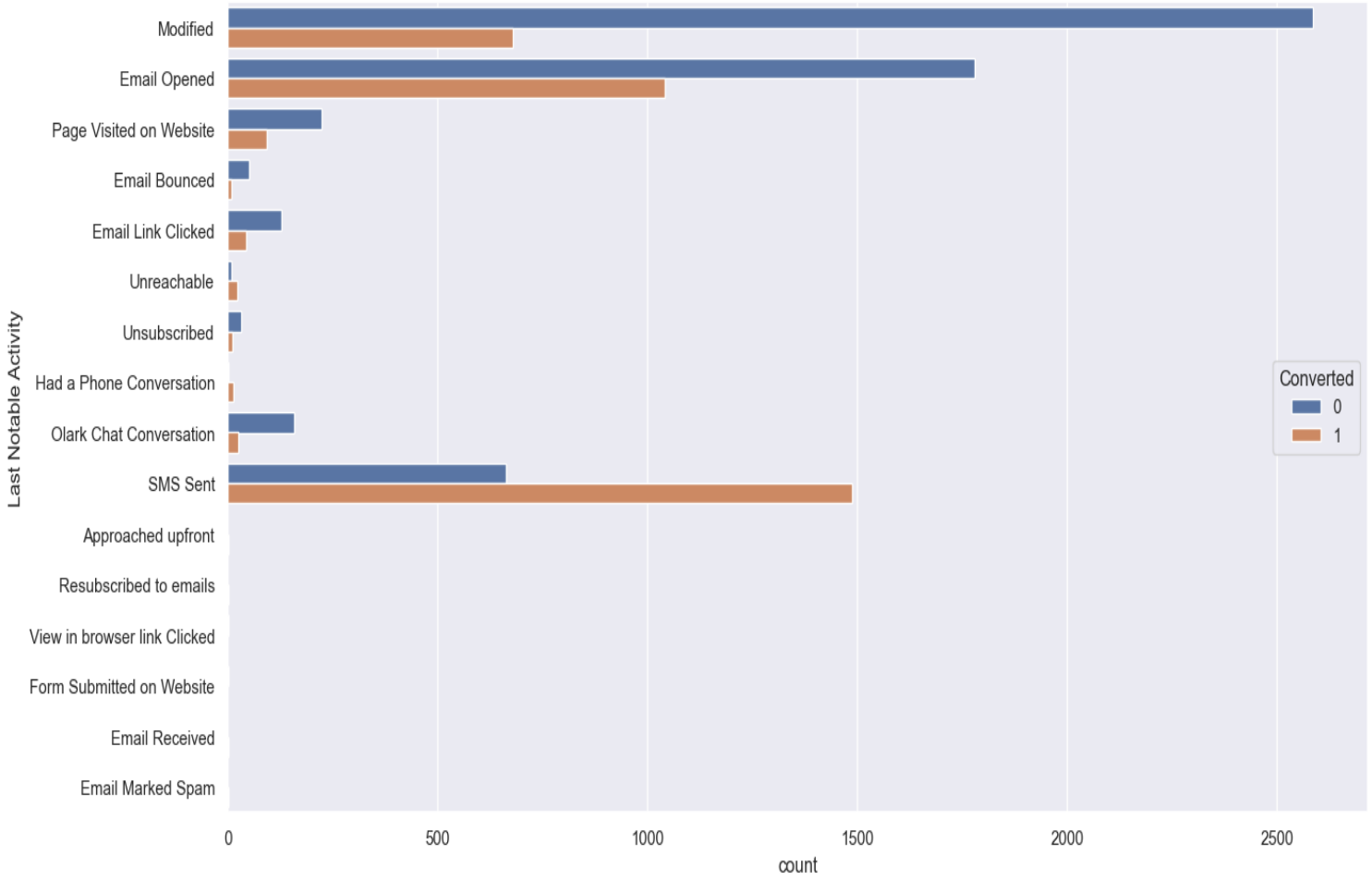
Last Activity



Maximum leads are generated from people with last activity: Email opened, and SMS sent, conversion rate is around 63% and 36%.

To improve overall lead conversion rate, focus should be on improving lead conversion of people with last activity: olark chat conversation, SMS sent and Page Visited on Website.

Last Notable Activity



SMS sent has a very high conversion rate.

General Insights from EDA.

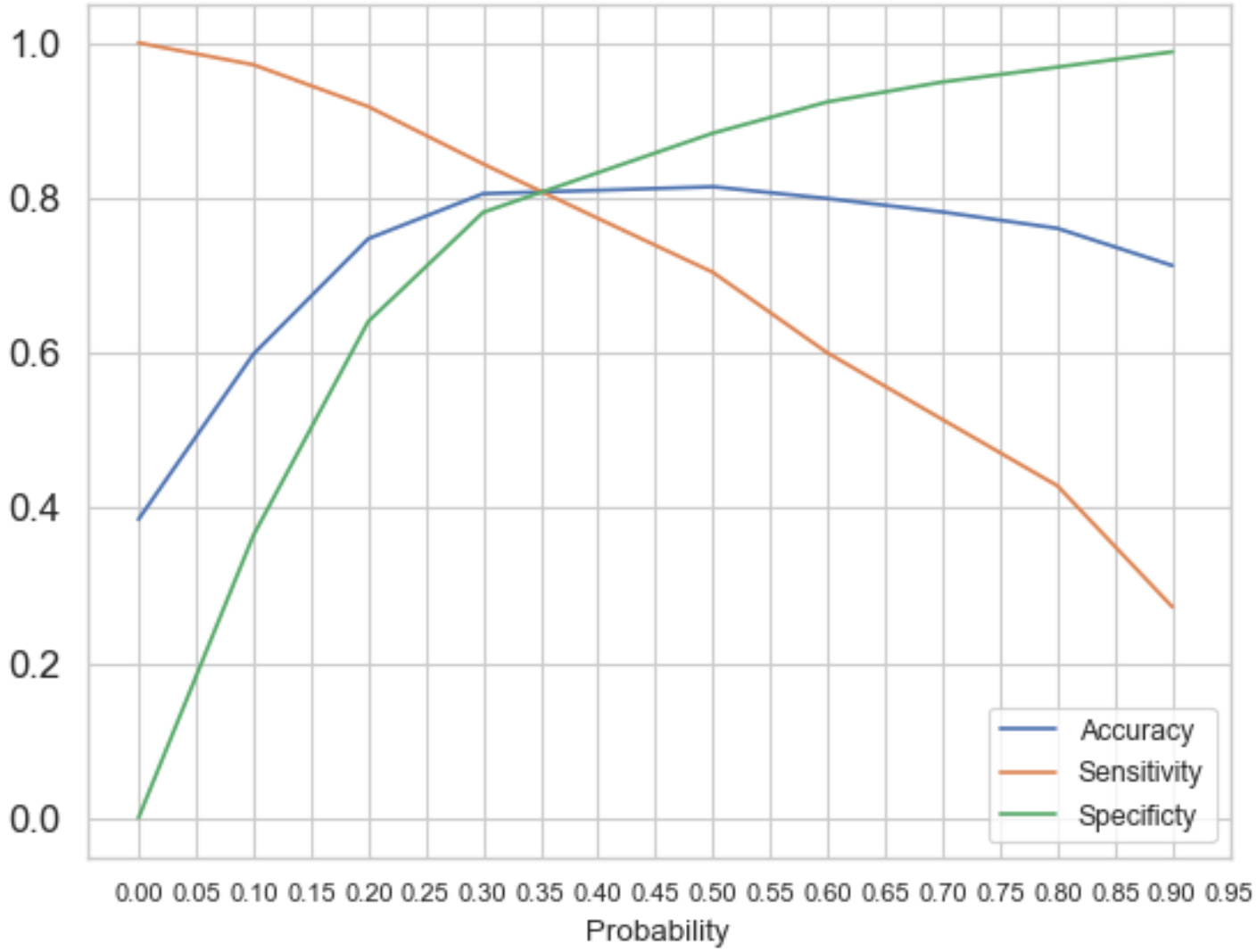
- To improve overall lead conversion rate, focus should be on improving lead conversion rate of API and Landing Page Submission. Also, generate more leads from Lead Add form since they have a very good conversion rate
- To improve overall lead conversion rate, focus should be on improving lead conversion of olark chat, organic search, direct traffic and google lead source. Also, generate more leads from reference and welingak website since they have a very good conversion rate.
- To improve overall lead conversion rate, focus should be on improving lead conversion of people with last activity: olark chat conversation, SMS sent and Page Visited on Website.
- Most of the Specialization has a more than 40 % conversion rate, with Finance Management and Human Resource Management having higher leads and conversion rates.
- To improve overall lead conversion rate, focus should be on improving lead conversion of unemployed. Also, generate more leads from Working Professionals and Housewives.
- Since maximum leads are from Mumbai City, focus should be on converting them more.
- Almost no one saw the advertisement through the above platforms.

Model Building

- Data Preparation – creating Dummy variable, converting binary variable and dropping variables with no such importance.
- Splitting into train and test – 70% and 30% respectively.
- Since too many variable, using RFE and select top 20 variable.
- Building the 1st model and eliminating variable based on high p-value(>0.05) and high VIF (>3).
- Re-building models until both p-values and VIR are below the threshold values.
- Model evaluation, with using the cut-off as 0.5.
- Plotting ROC Curve for optimal cut-off → 0.35
- Precision and Recall
- Making Predictions on Test test.
- Feature Importance

ROC CURVE

- ACCURACY – 80.7%
- SENSITIVITY – 81%
- SPECIFICITY – 80.5%



Final Insights for Sales team:

- The sales team of the X-Education should focus on the leads having lead origin: lead add form, occupation: Working Professional, Lead source: Wellingak website.
- Hot Leads are identified as Customers having lead score above 35. Sales Team of the company should first focus on the Hot Leads.
- There are many important variables like city, specialization, occupation which can potentially explain Conversion better. It is important for the management to make few of these information mandatory to fill, so that we can use in our model and suggest more important decisions for the business.
- High Sensitivity will ensure that almost all leads who are likely to Convert are correctly predicted whereas high Specificity will ensure that leads that are on the brink of the probability of getting Converted or not are not selected.
- If the Last Notable Activity is Modified, he/she may not be the potential lead.