Lead Scoring Case Study

Binayak Biswas

Kiran Mekala

Pooja Kumari

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%.

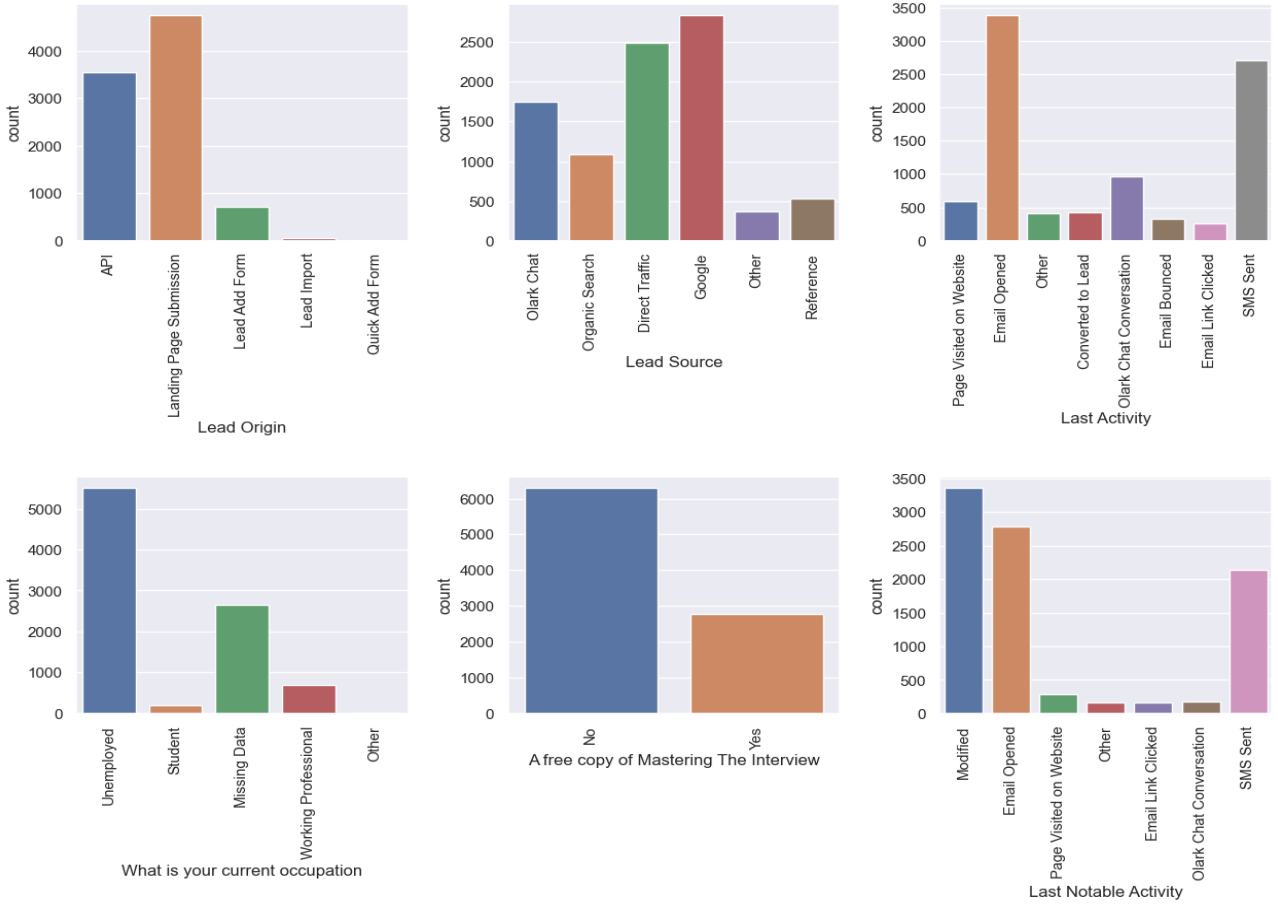
Business Goal

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.

Exploratory Data Analysis

Univariate Analysis (Categorical Columns)

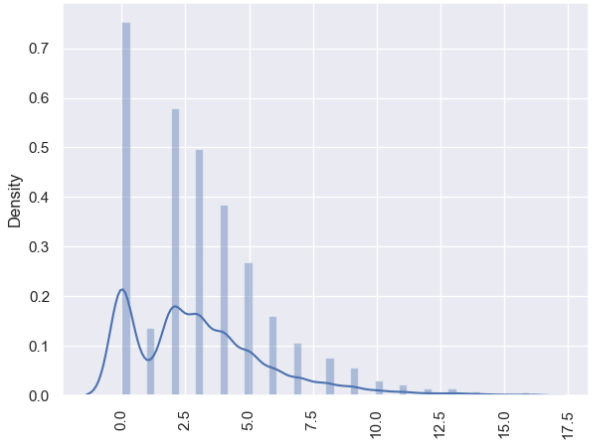
- API & Landing Page Submission are two major contributor of Lead Origin.
- Direct Traffic and Google are the two main source of Leads.
- Email Opened and SMS Sent are the major Last Activity.
- Most of the lead generated by Unemployed.
- Majority don't want a free copy of Mastering The Interview.

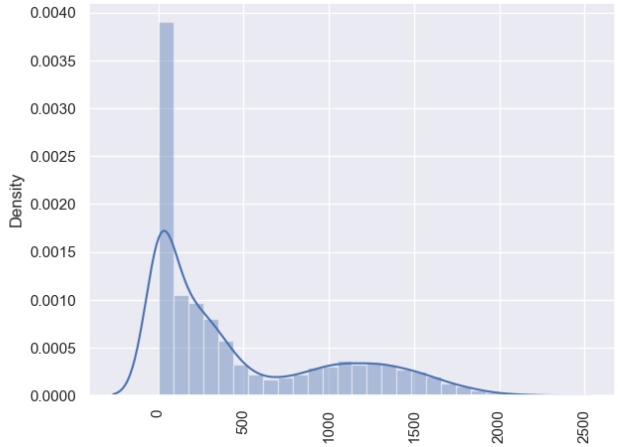


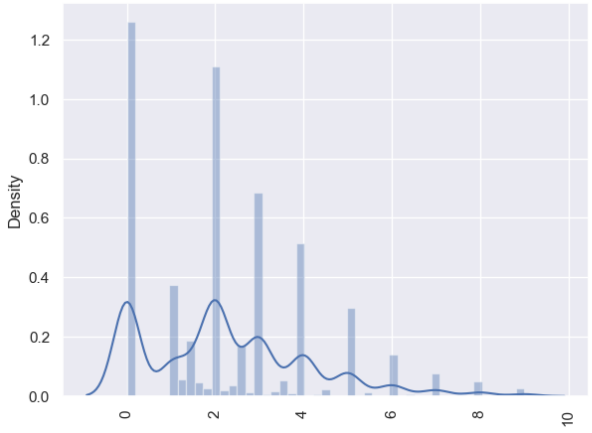
Exploratory Data Analysis

Univariate Analysis (Continuous Columns)

- None of the continuous variables are normally distributed.
- Outliers presence are not there.
- Totalvisits values are between 0-17, Total Time Spent on Website values are between 0-2500 and Page Views Per Visits values are between 0-10



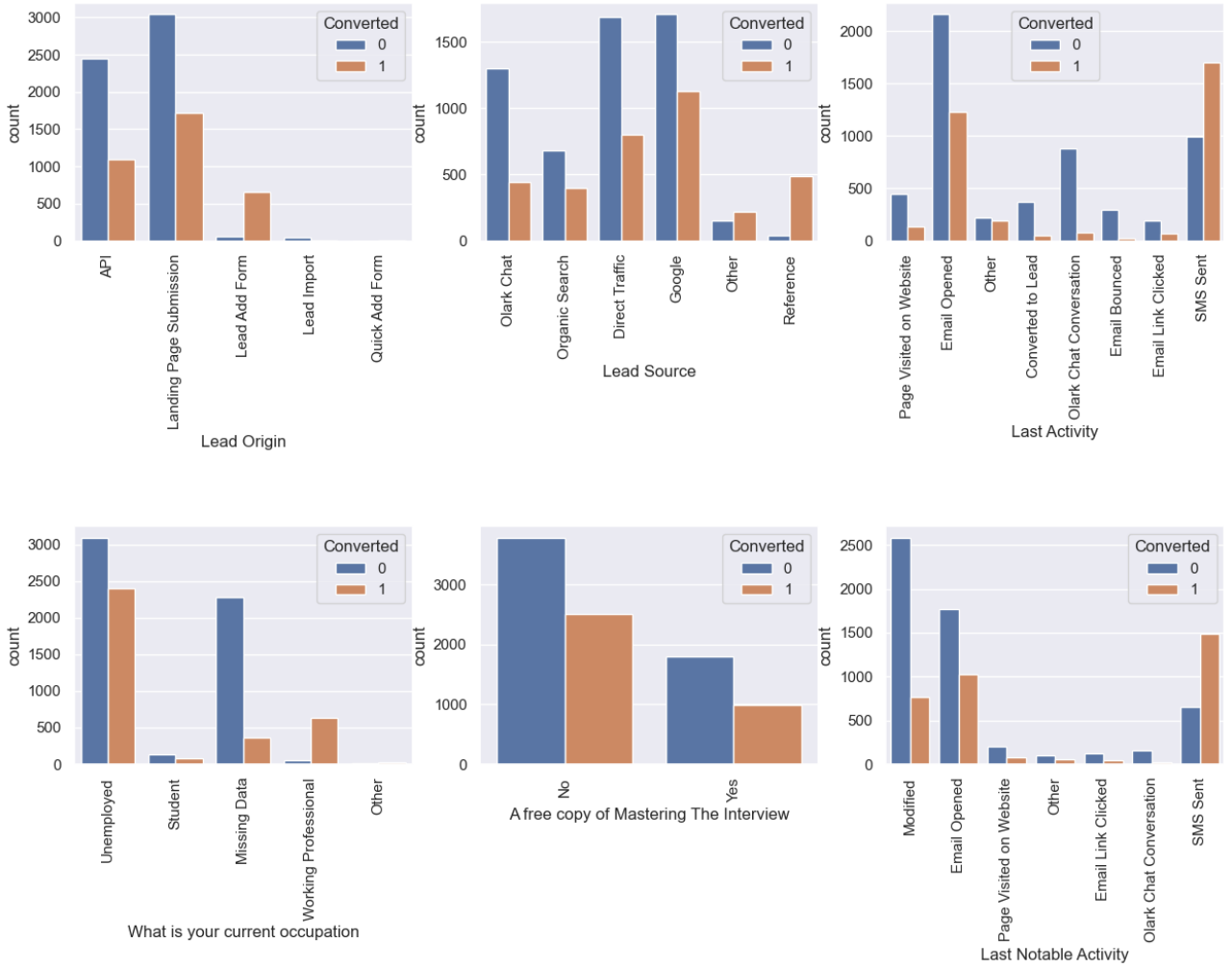




Exploratory Data Analysis

Bivariate Analysis (Continuous Columns)

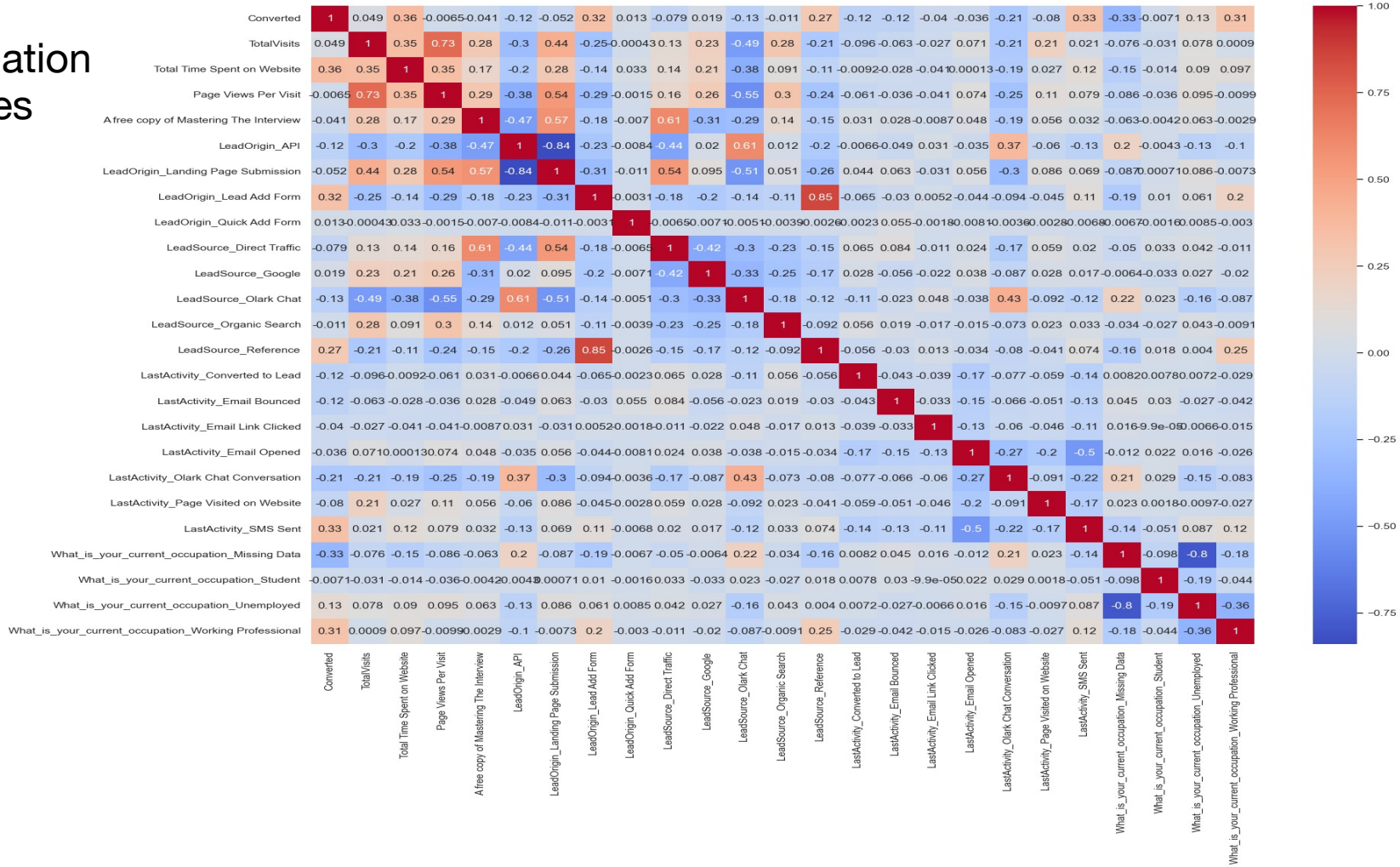
- Lead Origin : Hot leads are more in Landing Page Submission, API and Lead Add Form.
- Lead Source: Hot leads are higher in Direct Traffic and Google.
- Last Activity: Hot leads are higher in SMS Sent and EMAIL Opened.
- What is your current occupation: Hot leads are mostly generated by Unemployed and Working Professional.
- A free copy of Mastering The Interview: Hot leads are more with answer No.
- Last Notable Activity: Similar to Last Activity. - Lead Origin : Hot leads are more in Landing Page Submission, API and Lead Add Form.
- Lead Source: Hot leads are higher in Direct Traffic and Google.
- Last Activity: Hot leads are higher in SMS Sent and EMAIL Opened.
- What is your current occupation: Hot leads are mostly generated by Unemployed and Working Professional.
- A free copy of Mastering The Interview: Hot leads are more with answer No.
- Last Notable Activity: Similar to Last Activity.



Exploratory Data Analysis

Correlations

- Some variables have a high correlation
- Using RFE to select top 15 features

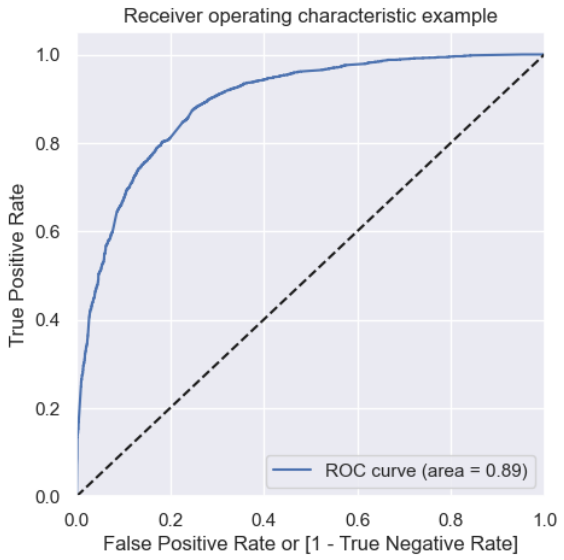


Model Evaluation

Plotting the ROC Curve

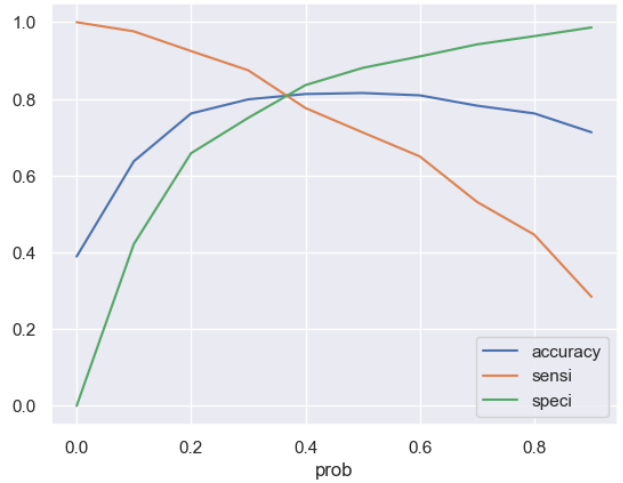
An ROC curve demonstrates several things:

- It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
- The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.



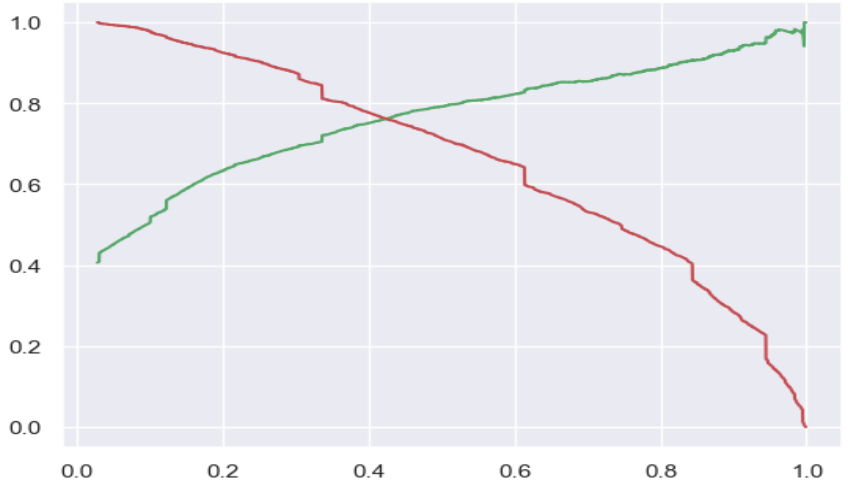
Finding Optimal Cutoff Point

- Optimal cutoff probability is that prob where we get balanced sensitivity and specificity



Precision and Recall Trade-off

- From the above curve we can see that precision & recall intersects at 0.41
- From the graph it is also clear that for having Recall >= 80% we have to keep cutoff <=0.32



Conclusion

Below are the variables that are most important (in order of priority) for being the potential buyers are:

- Total Time Spent on Website
- When LeadOrigin is 'Lead Add Form'
- When LastActivity was 'SMS Sent'
- When LeadSource is 'Olark Chat'
- TotalVisits
- When LastActivity was 'Email Opened'

Below are the variables that can lead a negative impact (in order of most impact) on the conversion:

- What_is_your_current_occupation has Missing Data
- What_is_your_current_occupation is 'Unemployed'
- What_is_your_current_occupation is 'Student'
- LastActivity was 'Email Bounced'
- LastActivity was 'Olark Chat Conversation'
- LeadSource was 'Direct Traffic'