Lead Scoring Case Study

Darshan Patel
Satyajit Pattnaik

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 & Objectives

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.

Understanding the data

Understanding Target Variable: "Converted"

Findings

- Data is imbalanced, ratio is almost 62:38.
- Most of the leads aren't converted, the rate of conversion is 38% which seems to be poor for the company.
- To make this process more efficient, the company needs to identify the most potential leads, also known as 'Hot Leads'.
- If the company 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.
- This can be achieved by assigning 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.

Count of converted variable

1	3435
0	5639

Missing Data

Findings

- Many columns have a lot of missing data (35-80%), some have few missing data (20-35%) and many columns also have no missing data at all.
- For features with less missing values- can use regression to predict the missing values or fill with the mean of the values present, depending on the feature.
- For features with very high number of missing values- it is better to drop those columns as they give very less insight on analysis.
- As there's no thumb rule on what criteria do we delete the columns with high number of missing values, we have done a small analysis and have taken decisions, Analysis to be continued in next slides…

Initial Intuition from the data

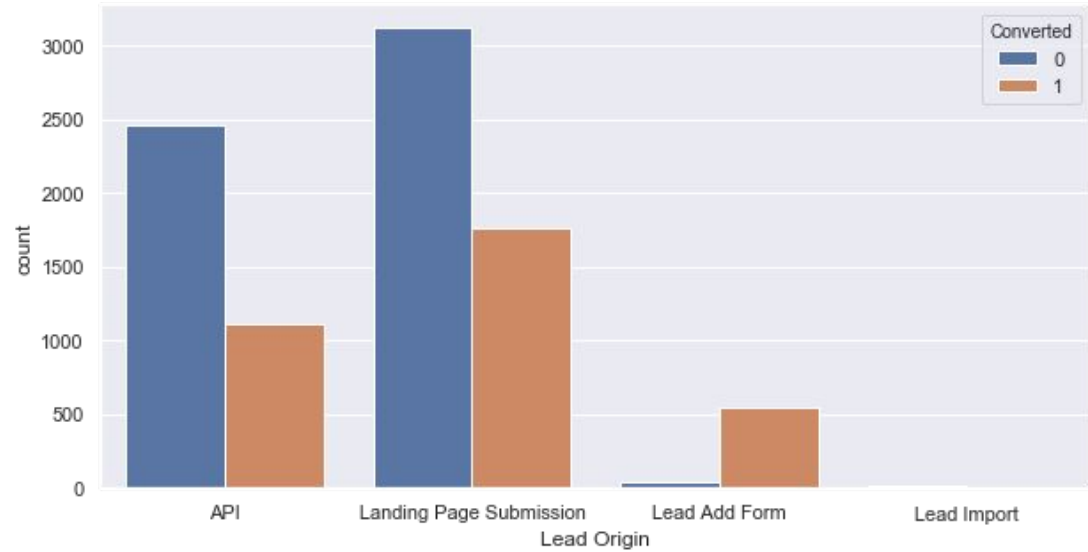Percentage of Missing values
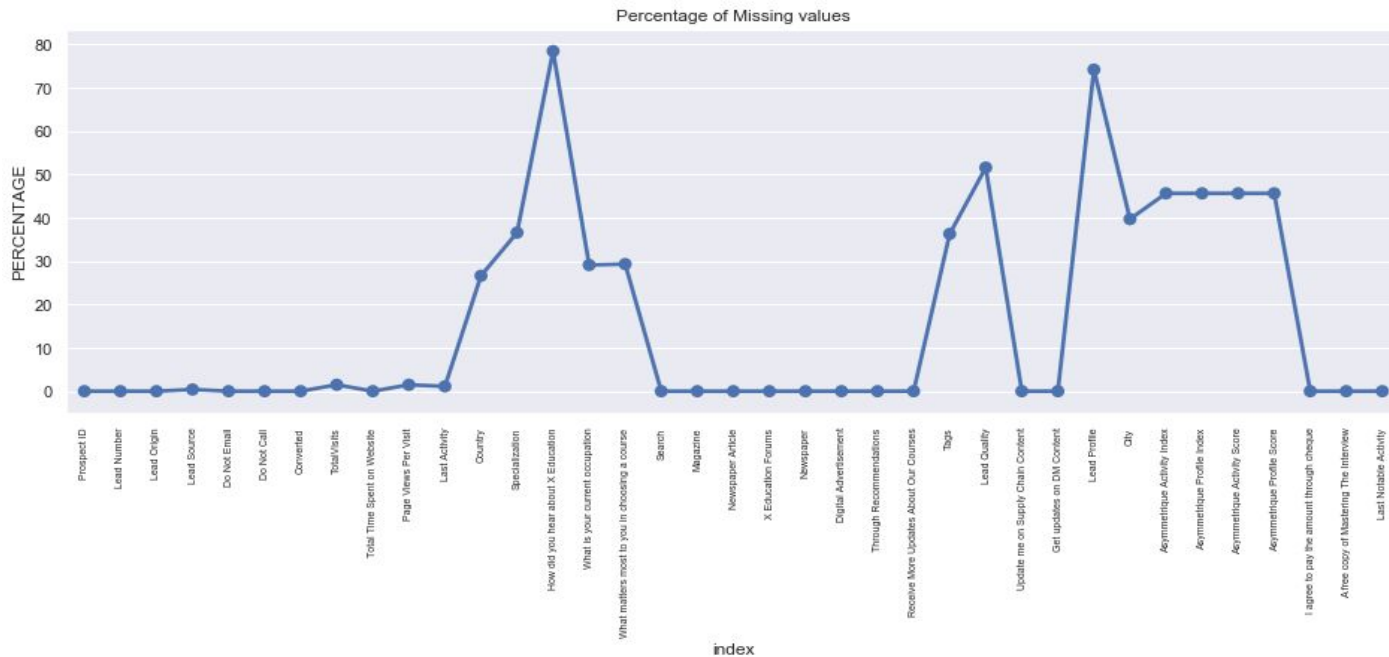
Initial Intuition

- Generally, there's no such thumb rule to delete any column having more than x% of null values, as we need to analyze each and every column to come to a conclusion.
- In this use case, we are considering a threshold of 45%, so any column beyond this value shall be discarded.

Categorical Analysis

Lead Origin Channel

Findings

❖ From above plot, API, Landing Page Submission and Lead Add Form are having impact on conversion rate.

❖ Even though, API and Landing Page Submission are having less 'Converted' rate, but the amount of lead is being generated can be considered.

❖ In case of Lead Add Form, we can clearly see the amount of 'Converted' is less but the converted rate is more. For 'Lead Import', the converted rate is very low.

❖ So, to improve overall lead conversion rate, the company needs to focus more on improving lead conversion of API and Landing Page Submission origin and generate more leads from Lead Add Form.

Lead Source
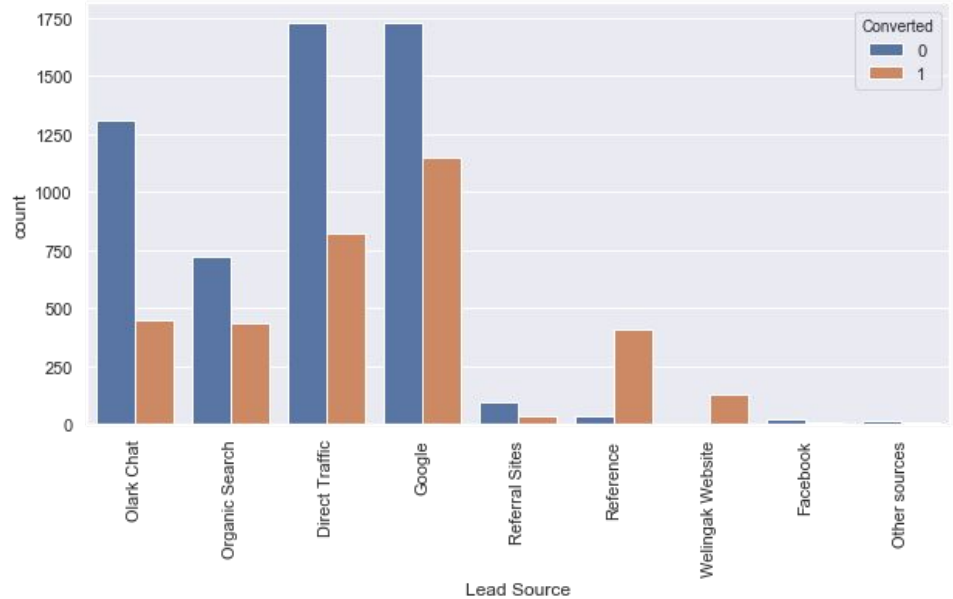
Findings

❖ Here, we can see Direct Traffic and Google can be helpful to increase conversion rate.
❖ The conversion rate in case of reference and welingak website is high.
❖ So, to improve overall lead conversion rate, the company needs to focus on increasing lead conversion rate of olark chat, organic search, direct traffic, and google. More Leads are generated from reference and welingak website.
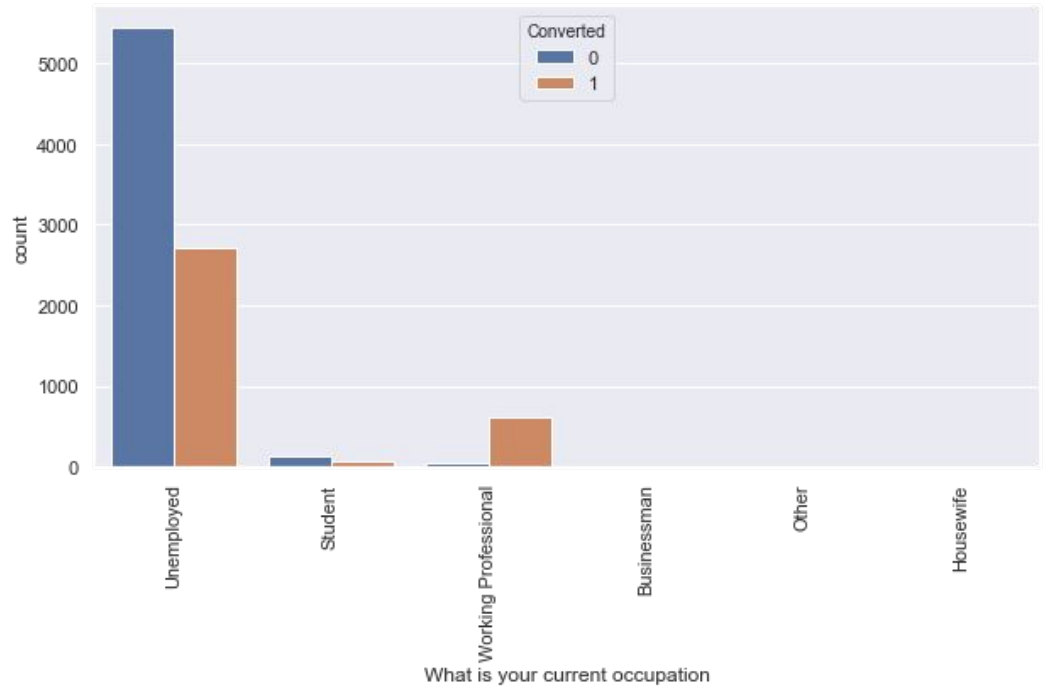
Last Activity
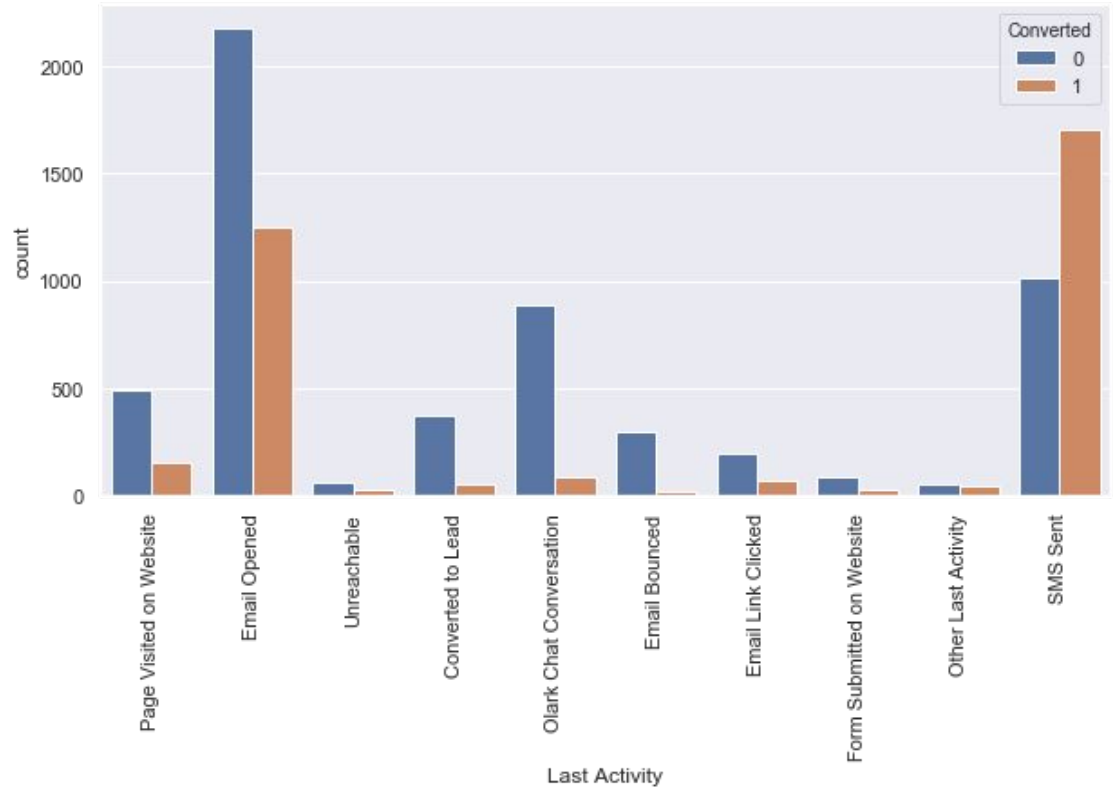
Findings

❖ Here, we can see most of the leads do the activities like Email and SMS sending.

❖ The conversion rate is more in case of SMS sent activity
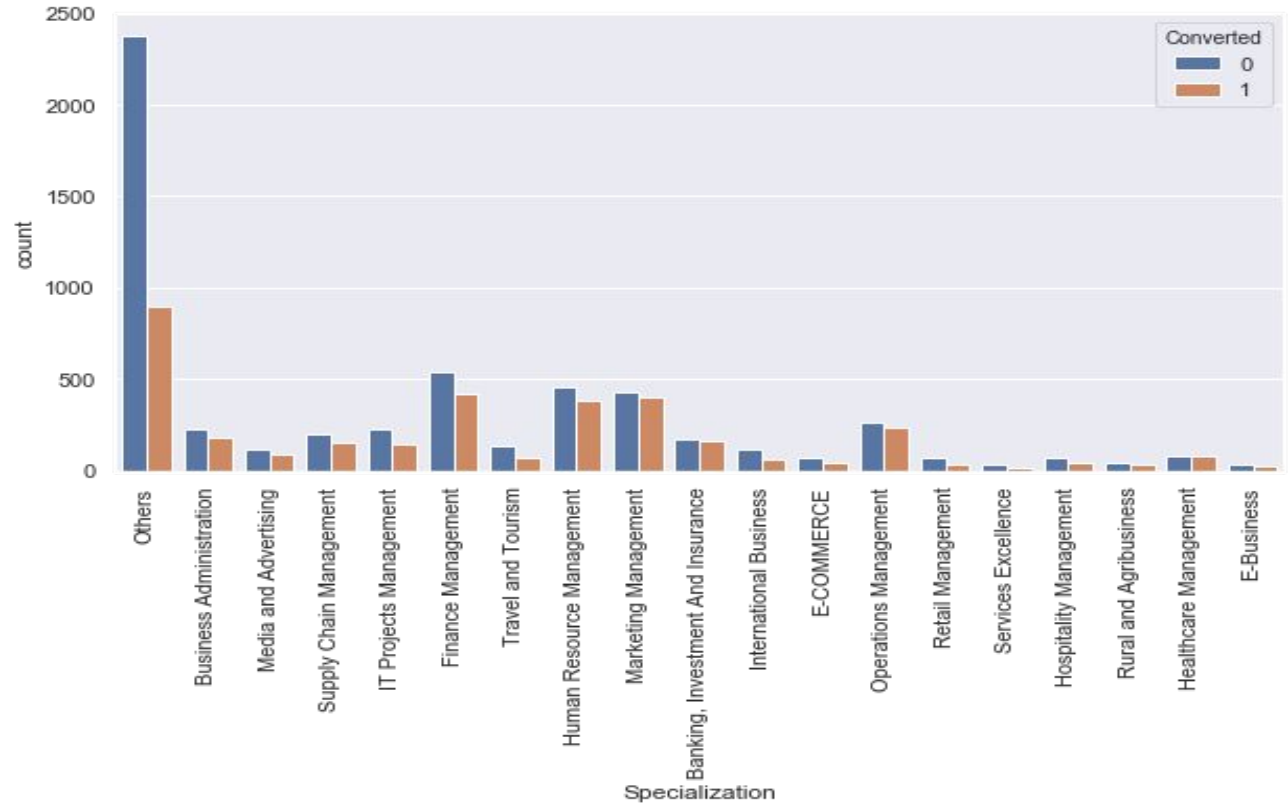
Occupation

Findings

❖ Working Professional are more likely to join the courses. So, our focus should be more on them.

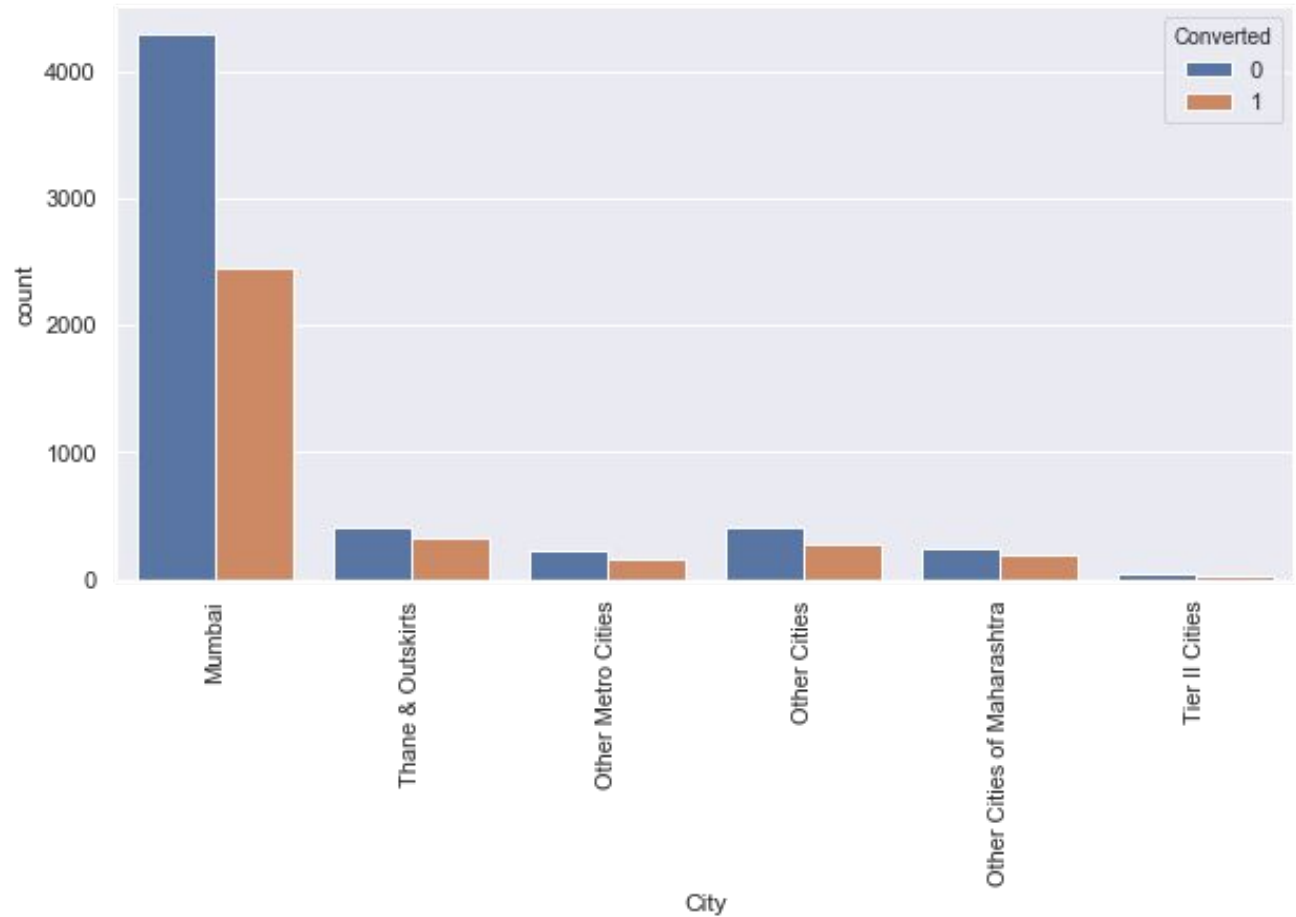❖ For Unemployed leads, the conversion rate is less but the number of joiners are more.

Specialization

❖ Most of the leads are having specialisation in other program.

City

Findings
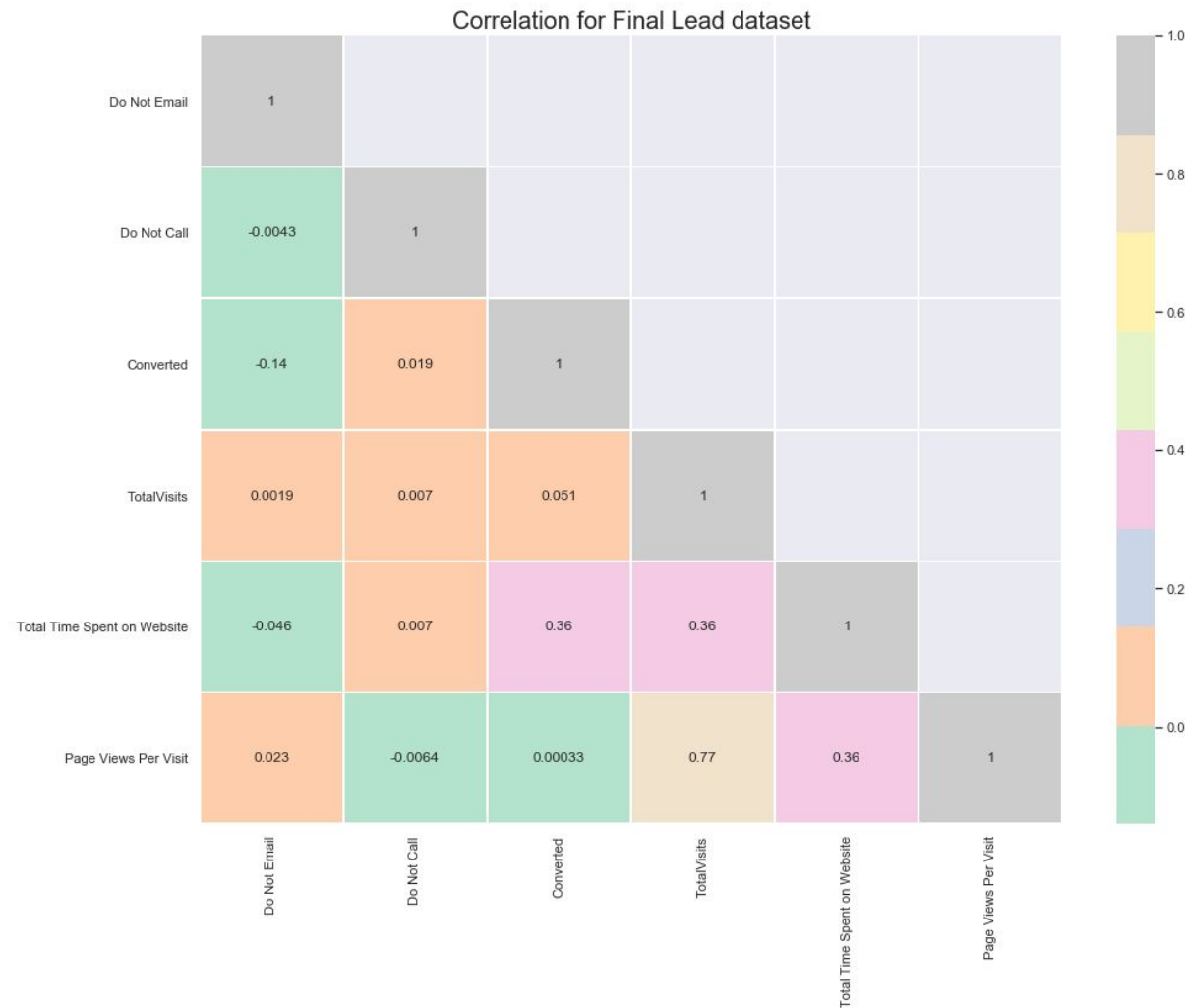
❖ Most of the leads are from Mumbai, so we should focus on Mumbai

Correlation Matrix

Correlation for Final Lead dataset

|	Do Not Email	Do Not Call	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit
Do Not Email	1					
Do Not Call	-0.0043	1				
Converted	-0.14	0.019	1			
TotalVisits	0.0019	0.007	0.051	1		
Total Time Spent on Website	-0.046	0.007	0.36	0.36	1	
Page Views Per Visit	0.023	-0.0064	0.00033	0.77	0.36	1

Findings

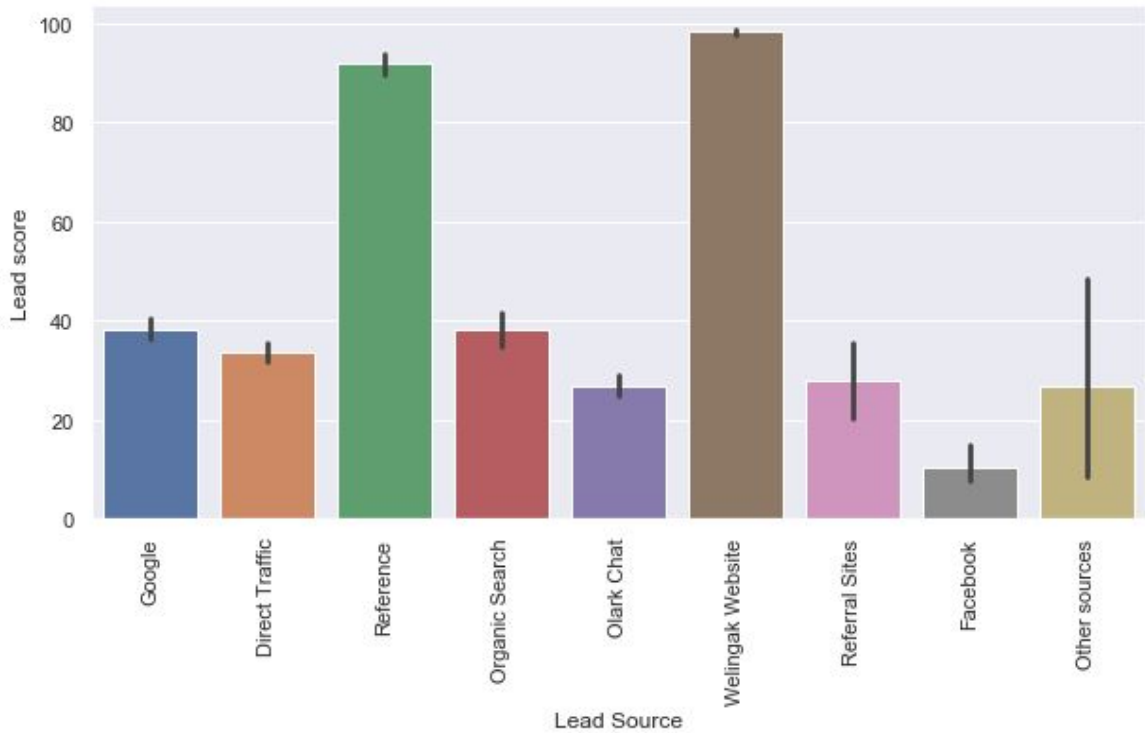Total time spent on website and Converted are highly correlated.

Lead Origin vs Lead Score
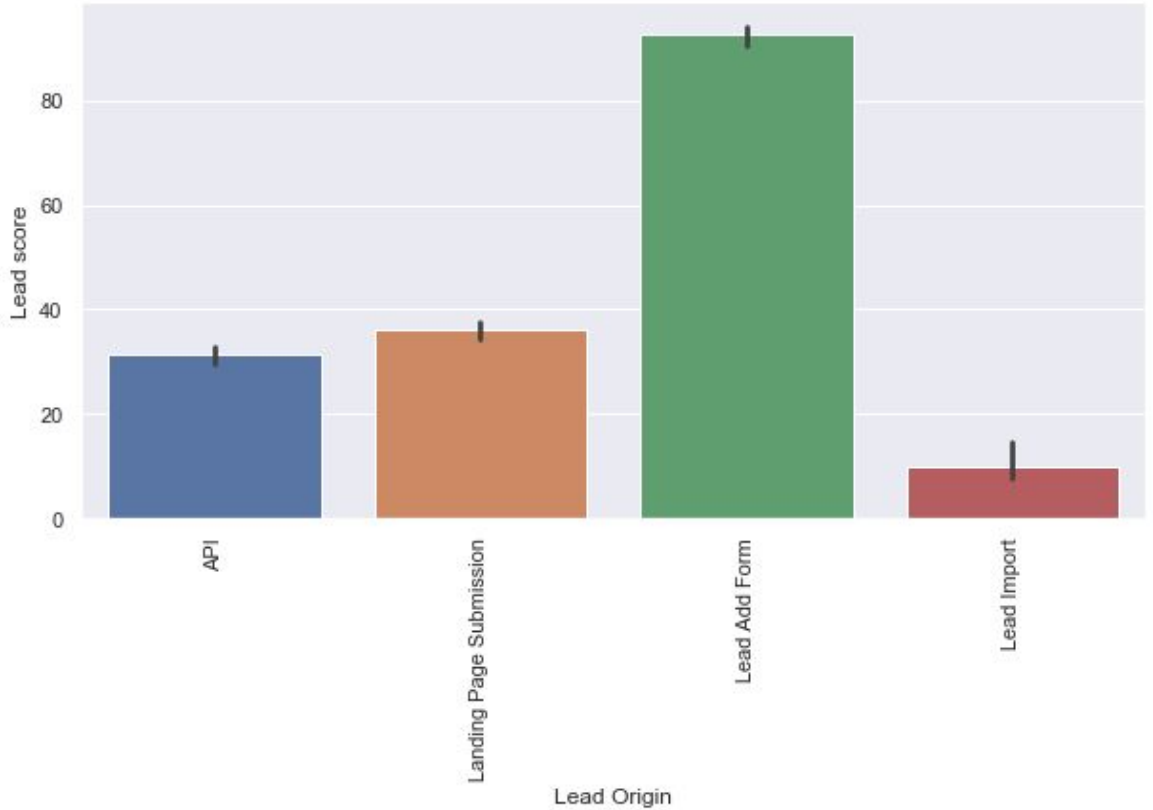
Findings

After calculating Lead score, we can see leads, who are added by lead add form.
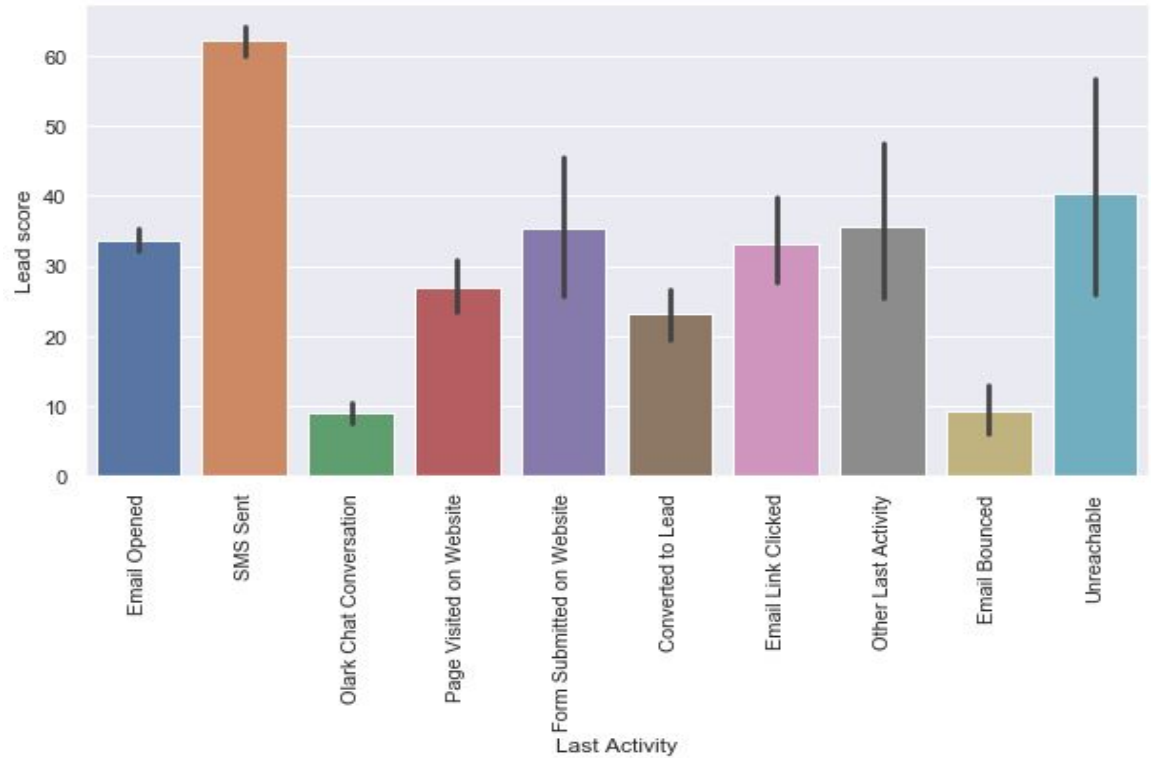
Lead Source vs Lead Score

Findings

Welingak and reference are the main source for leads.

Lead Score vs Last Activity

Findings

As per lead score, the last activity is done by leads are sending SMS, Email. So, These would be the notable way to communicate the leads.

Conclusion

- The lead scores are assigned to each lead and it will help X Education to identify the potential leads, called as 'Hot leads'. So that, the company can 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 potential leads can be decided on the basis of lead scores now.

- If the lead score is higher i.e. greater than 80, then the lead is hot and is most likely to convert.

- If the lead score is low i.e. let's say less than 50, then lead is low and will mostly not get converted.

Recommendations

X Education can focus on below criterias of the leads to identify the more potential leads.

- Lead Origin : API, Landing Page Submission and Lead Add
- Lead Source : Direct Traffic, Google, reference and welingak website
- Do Not Email
- Do Not Call
- TotalVisits
- Total Time Spent on Website
- Page Views Per Visit
- Last Activity : Email and SMS
- Specialization
- What is your current occupation : Unemployed & Working Professional
- Tags
- Lead Quality
- City : Mumbai

Thanks

The Team

Satyajit Pattnaik

Darshan Patel