Lead Score Case Study

By – Gantavya & Roshan

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.

Problem Statement

- 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

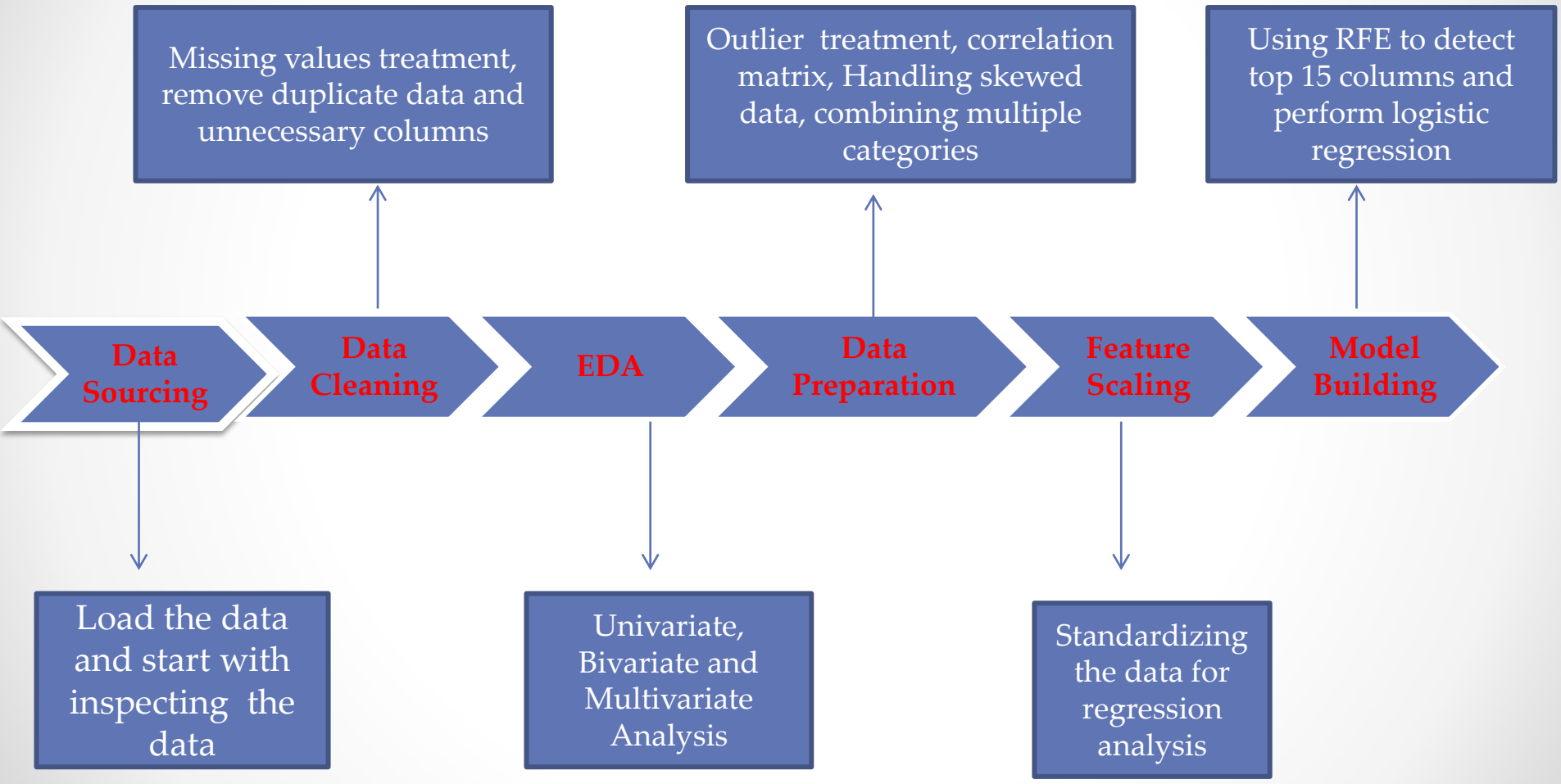
Proposed Solution

- Currently the conversion rate for the company is 30% which is very low. In order to get this number to **80%**, we will make of logistic regression method to accurately identify and highlight **HOT LEADS** by giving each lead a Lead score ranging from 0-100

- Once the data is filtered, the sales team then can focus on the leads with higher lead score for better conversion rate

- Finally, we have a smaller number of leads to focus on and customers that are potential buyers, the sales team can have effective communications with their leads for better conversion rate

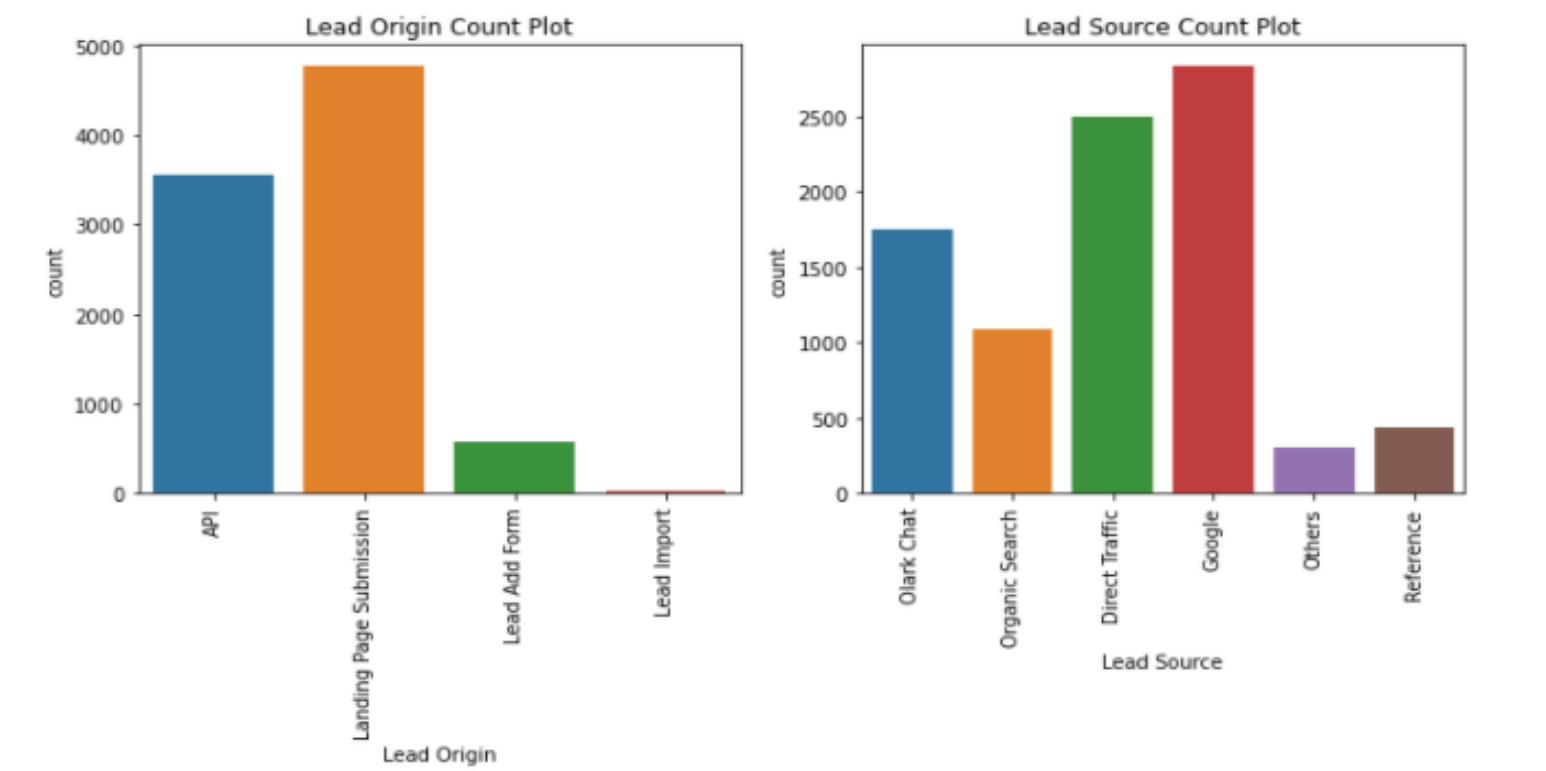
Methodology



Exploratory Data Analysis

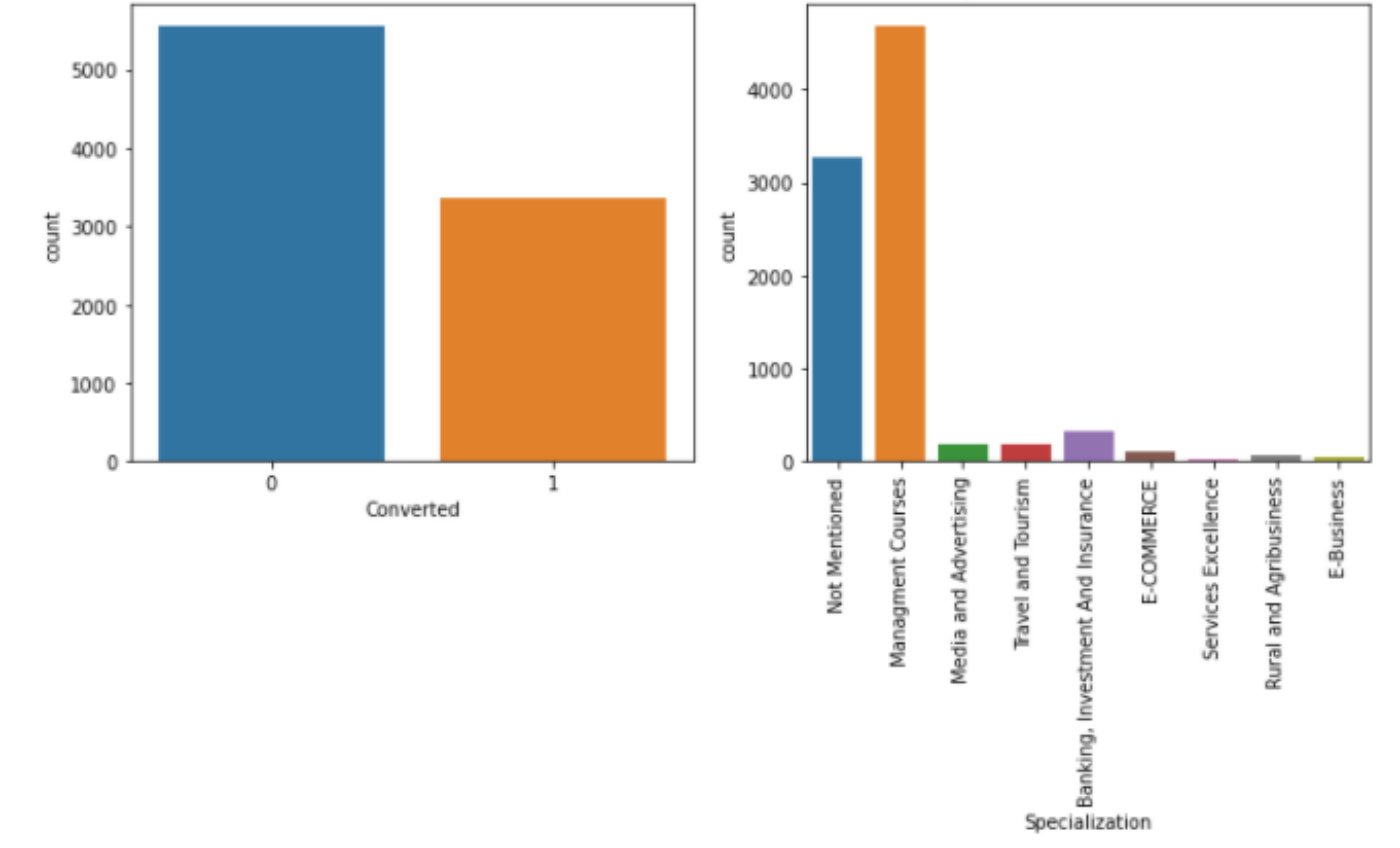
-Data Visualization

Univariate Data Analysis



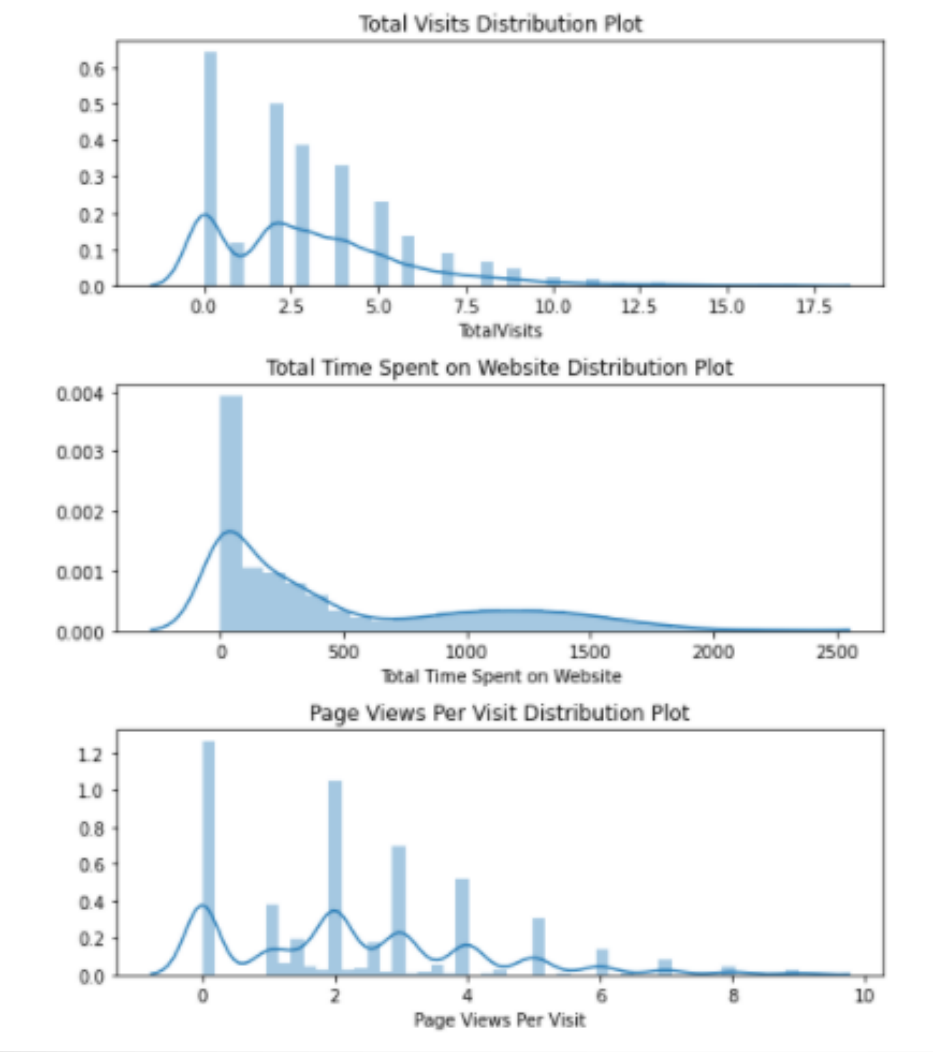
- In the Lead Origin Count Plot we can notice that there is a huge number of **landing page submissions** followed by API
- In the Lead Source Count Plot we can notice that the major source of traffic is from **Google** followed by Direct Traffic

Univariate Data Analysis



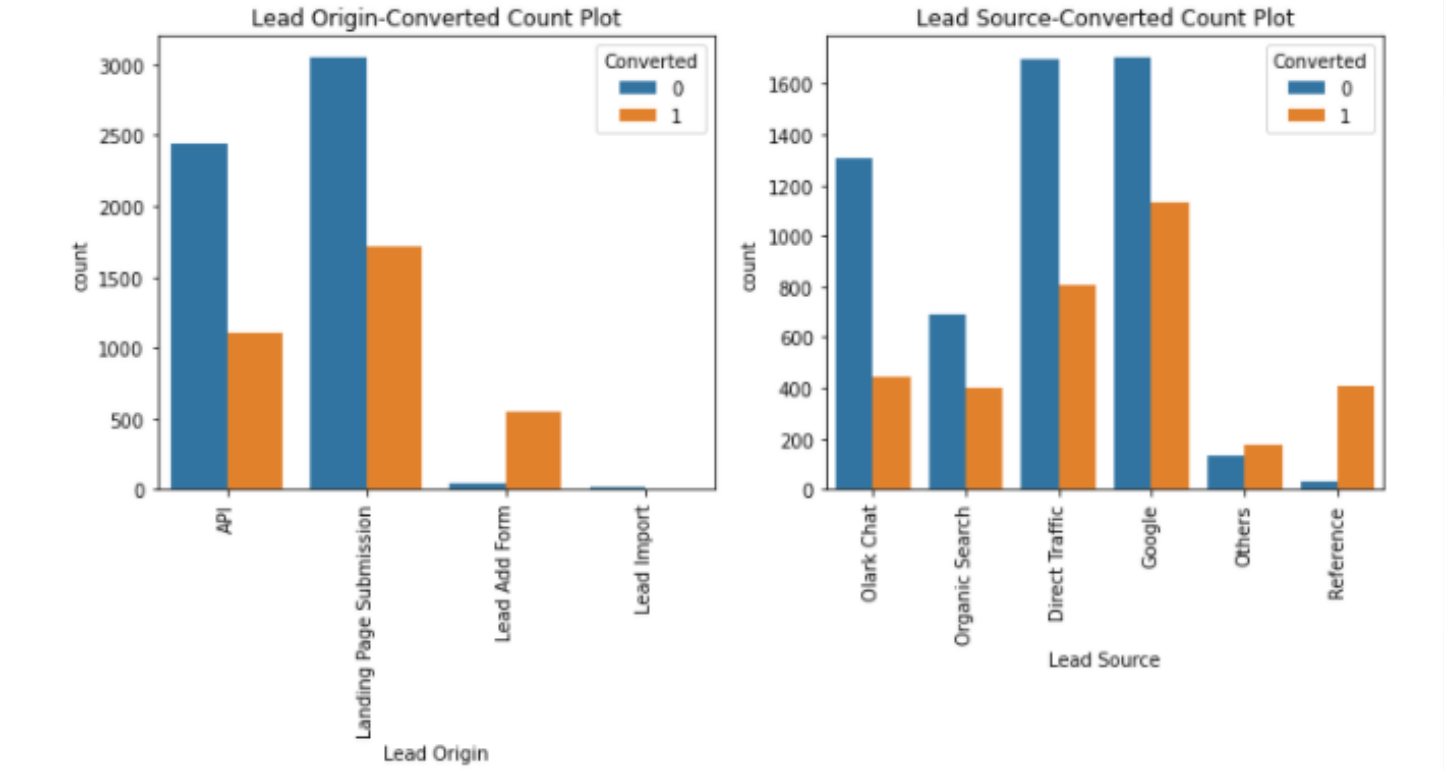
- In the Converted Count Plot we can notice that there almost twice the number of people who have not converted. There are a total of around **39% conversion rate**
- Most of the customers have a specialization in **Management Courses**

Numerical Univariate Analysis



- The maximum number of visits made by a customer to the website is of around **17 visits**.
- The maximum total time spent by a customer on the website is around **2272.**
- The maximum page views by customer is found to be around **9 views**.

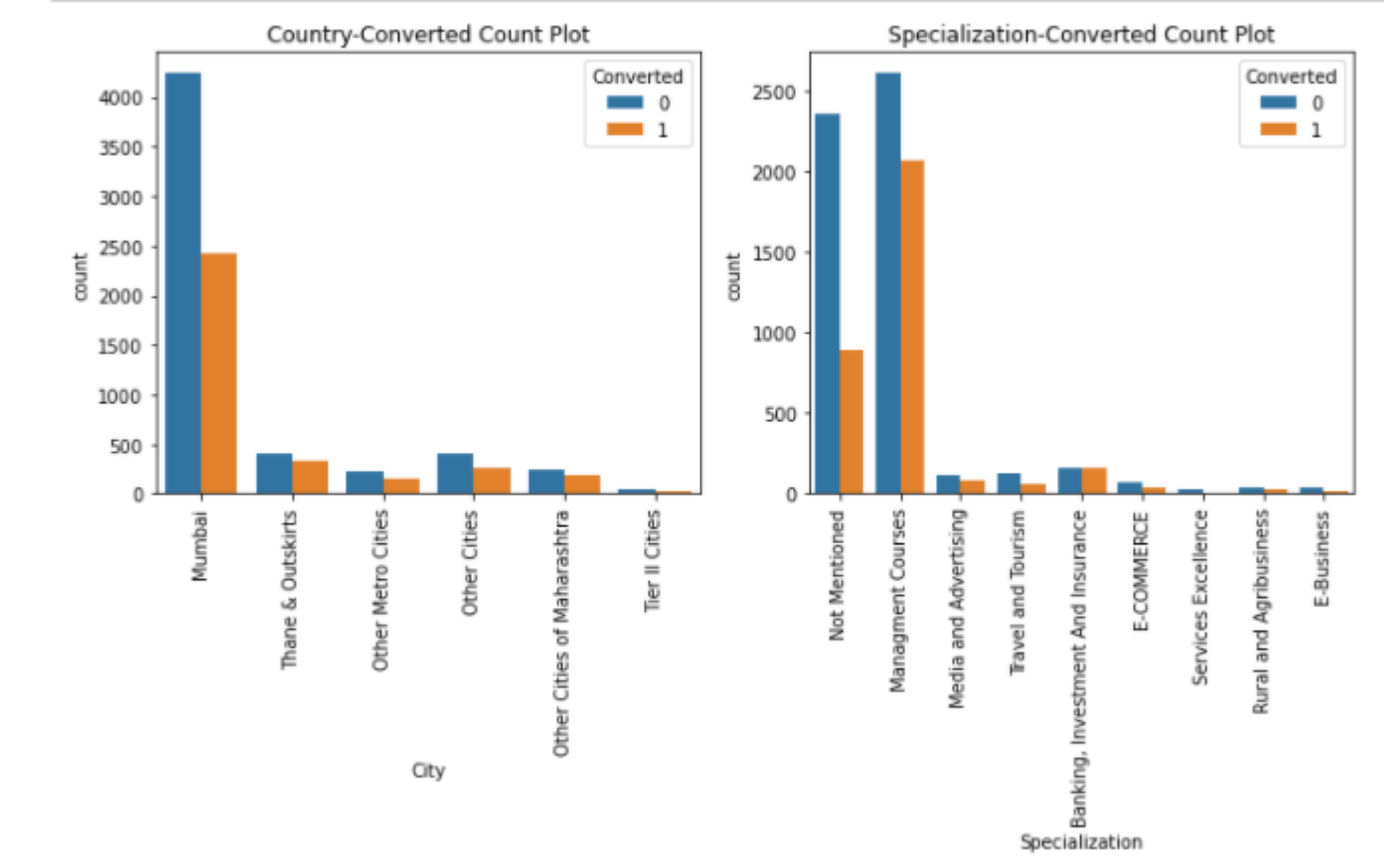
Multivariate Analysis



- In the Lead Origin Count Plot we can notice that most of the leads got converted at **landing page submission** category followed by API.

- In the Lead Source Count Plot we can notice that most of the leads got converted at by **Goggle** followed by Direct Traffic

Multivariate Analysis



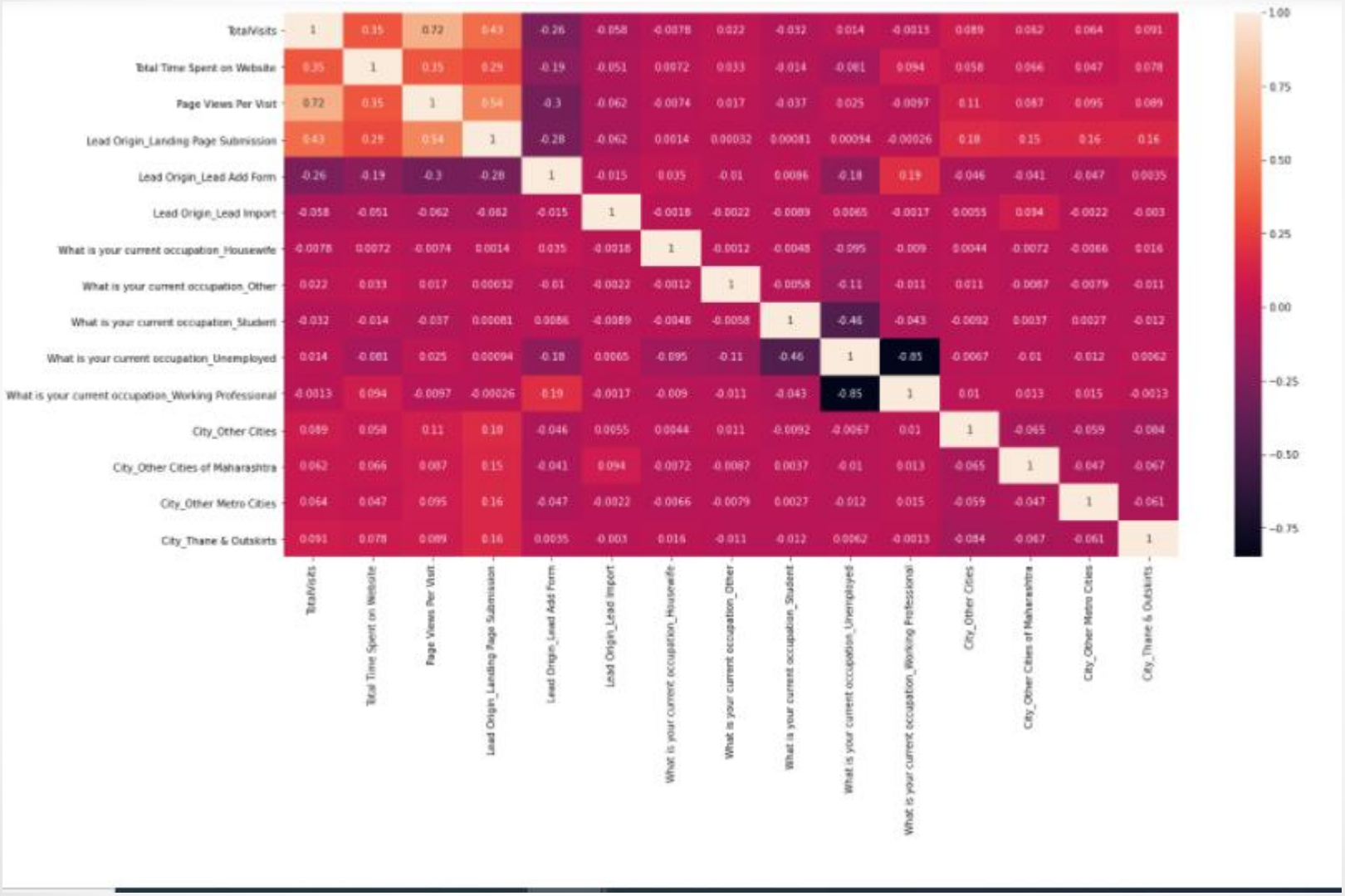
- In the City converted Count Plot we can notice that most of the leads got converted by customers from **Mumbai**.

- In the specialization Count Plot we can notice that most of the customers have a specialization in **Management Courses**

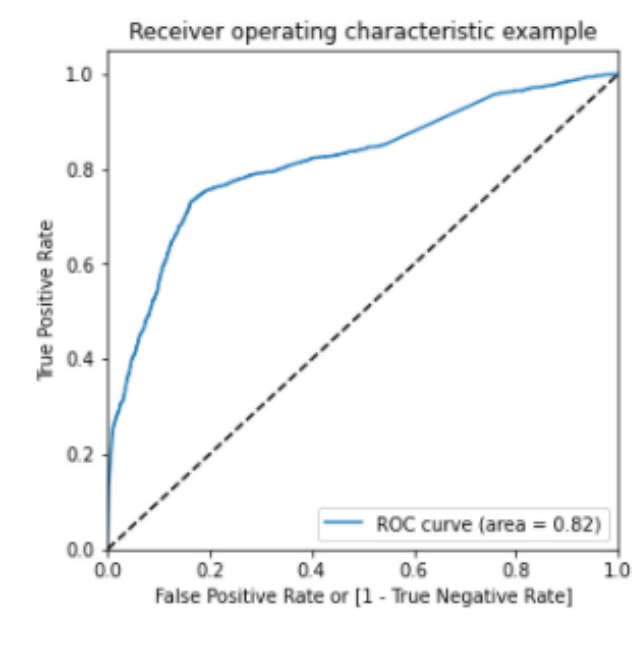
Top 15 Columns After **RFE**

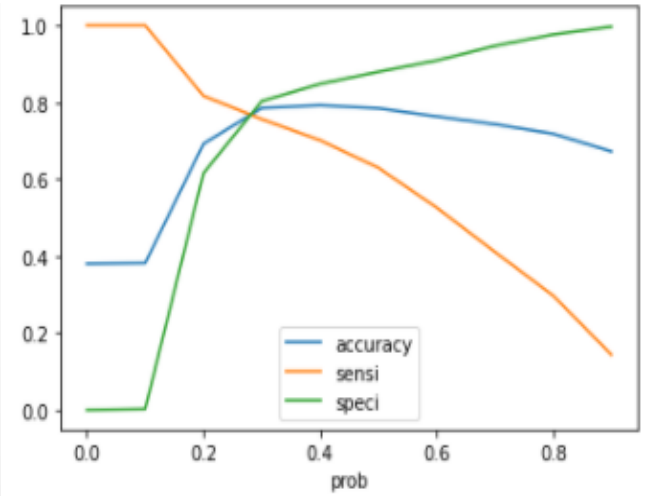
- TotalVisits
- Total Time Spent on Website
- Page Views Per Visit
- Lead Origin_Landing Page Submission
- Lead Origin_Lead Add Form
- Lead Origin_Lead Import
- What is your current occupation_Housewife
- What is your current occupation_Other
  What is your current occupation_Student
- What is your current occupation_Unemployed
- What is your current occupation_Working Professional
- City_Other Cities
- City_Other Cities of Maharashtra
- City_Other Metro Cities
- City_Thane & Outskirts

Correlation Matrix



Model Evaluation - Train Dataset





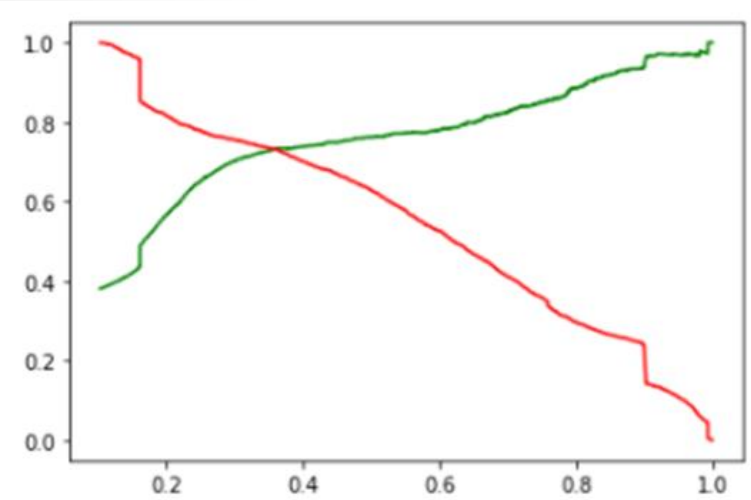
CONFUSION MATRIX

3167	704
592	1783

Accuracy = 79%
Sensitivity = 75%
Specificity = 81%
False Positive Rate = 18%
Positive predictive value = 71%
Negative predictive value = 89%

- **0.3** is the optimal cut-off based on Accuracy, Sensitivity and Specifity

Model Evaluation - Train Dataset



0.4 is the optimal cut-off based on Precision and Recall

CONFUSION MATRIX

3167	704
592	1783

Precision = 71.5%
Recall = 76%

Model Evaluation – Test Dataset

CONFUSION MATRIX

1361	323
242	752

Accuracy = 78%
Sensitivity = 75%
Specificity = 80%
Precision = 70%
Recall = 75%

Summary

- Model **Accuracy** on **Train** Data is **79%**
- Sensitivity of Train data is 75%
- Specificity On Train data is 81%
- Model **Accuracy** on **Test** Data is **78%**
- Sensitivity of Test data is 75%
- Specificity On Test data is 80%

Our Logistic Regression Model looks accurate and good enough when we compare the model on the train and test data set. We are getting the similar accuracy, Sensitivity and Specificity

Summary

Based on the efficiency of our model the client can now get an accurate idea of the probability of a hot-lead. This in return can help effectively boost the conversion rates of the sales team.

The company could employ different strategies to communicate with clients with different lead scores in order to ensure maximum conversion rates.

For example, customers with a high conversion rate can be directly contacted via call whereas customers with a low conversion rate could be provided with more content to boost the chances of conversion.