Lead Scoring Case Study

Anusha G

Adesh Sonar

Kishore K C U

Problem Statement

- X Education sells online courses to industry professionals.
- 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.

Business Objective:

- X education wants to know most promising leads.
- For that they want to build a Model which identifies the hot leads.
- Deployment of the model for the future use.

Data Manipulation

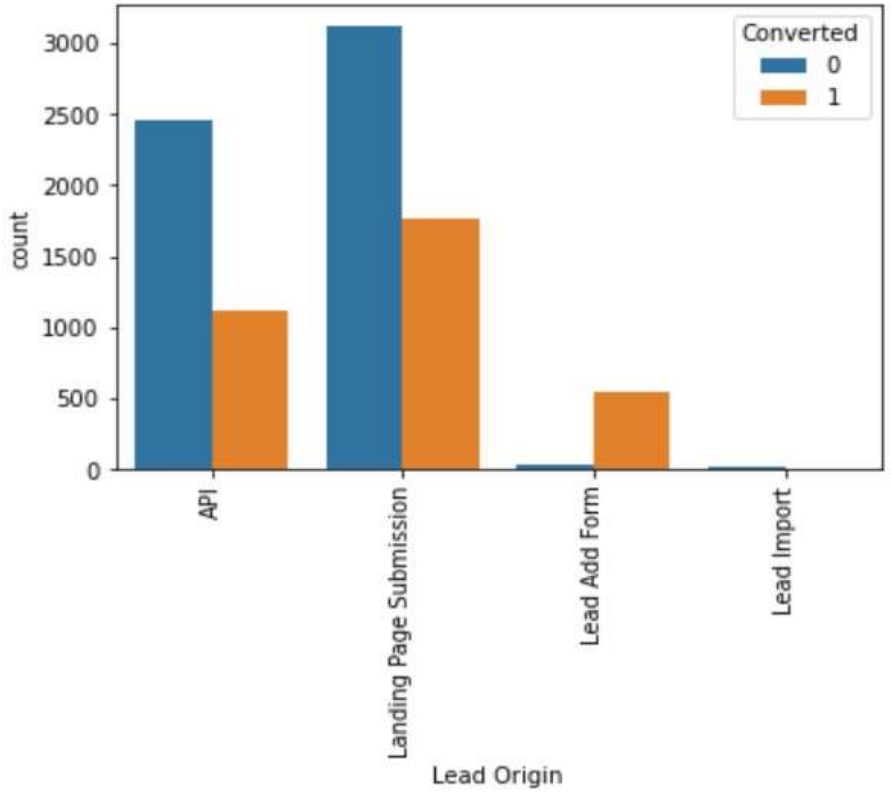
- Total Number of Rows =9240, Total Number of Columns =37.
- Single value features like “Magazine”, “Receive More Updates About Our Courses”, “Update me on Supply Chain Content”, “Get updates on DM Content”, “I agree to pay the amount through cheque” etc. have been dropped.
- Removing the “Prospect ID” and “Lead Number” which is not necessary for the analysis.
- After checking for the value counts for some of the object type variables, we find some of the features which has no enough variance, which we have dropped, the features are: “What matters most to you in choosing course”, “Search”, “Newspaper Article”, “X Education Forums”, “Newspaper”, “Digital Advertisement” etc.
- Dropping the columns having more than 70% as missing value such as ‘How did you hear about X Education’ and ‘Lead Profile’.

Exploratory Data Analysis (EDA)

Lead Origin

- Count of API and Landing Page Submission has considerable lead count and conversion rate is moderate.
- Although the count of lead in Lead Add Form is lower, the conversion rate is higher.
- Lead import has very less count.

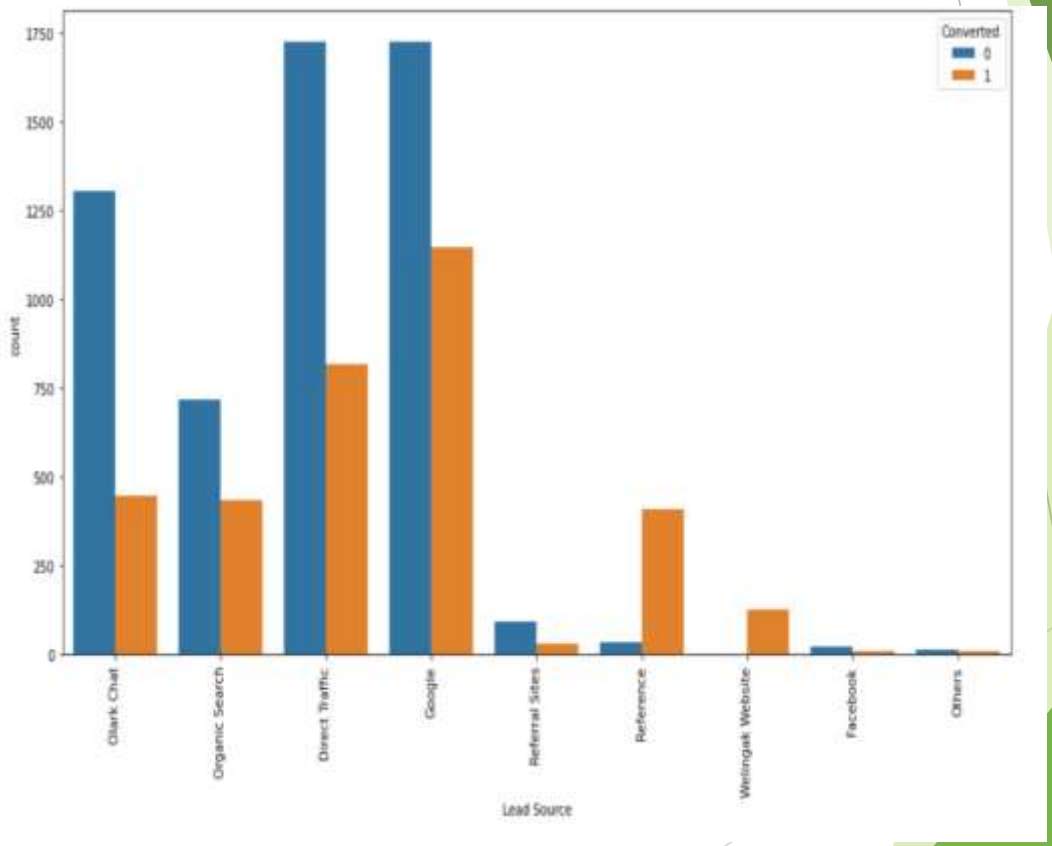
To improve overall lead conversion rate, we need to focus more on improving lead conversion of API and Landing Page Submission origin and generate more leads from Lead Add Form.



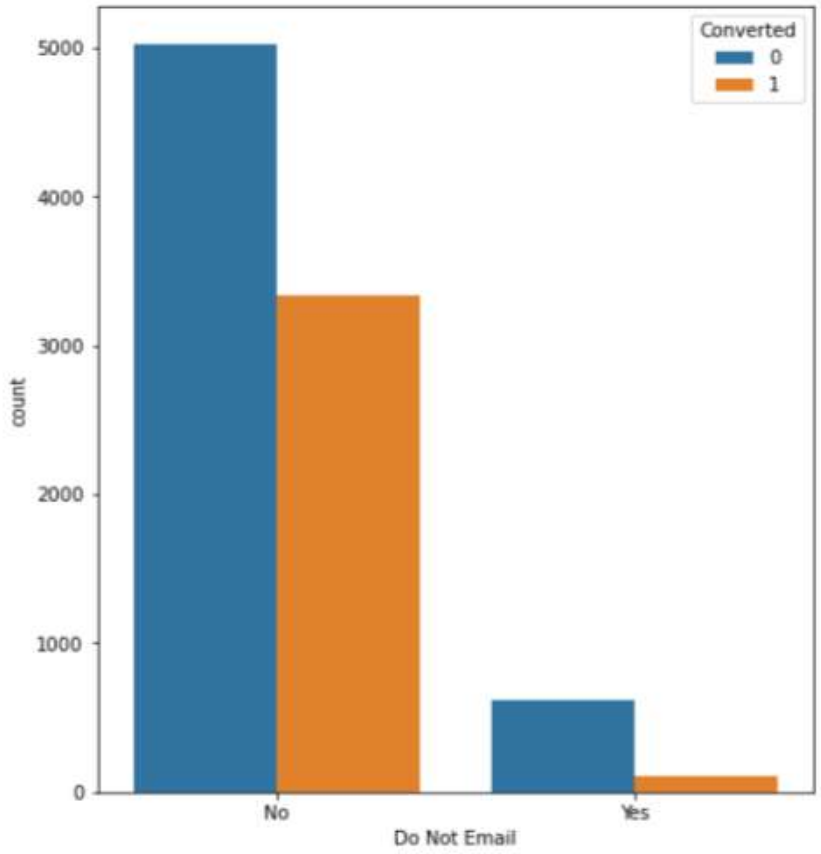
Lead Source

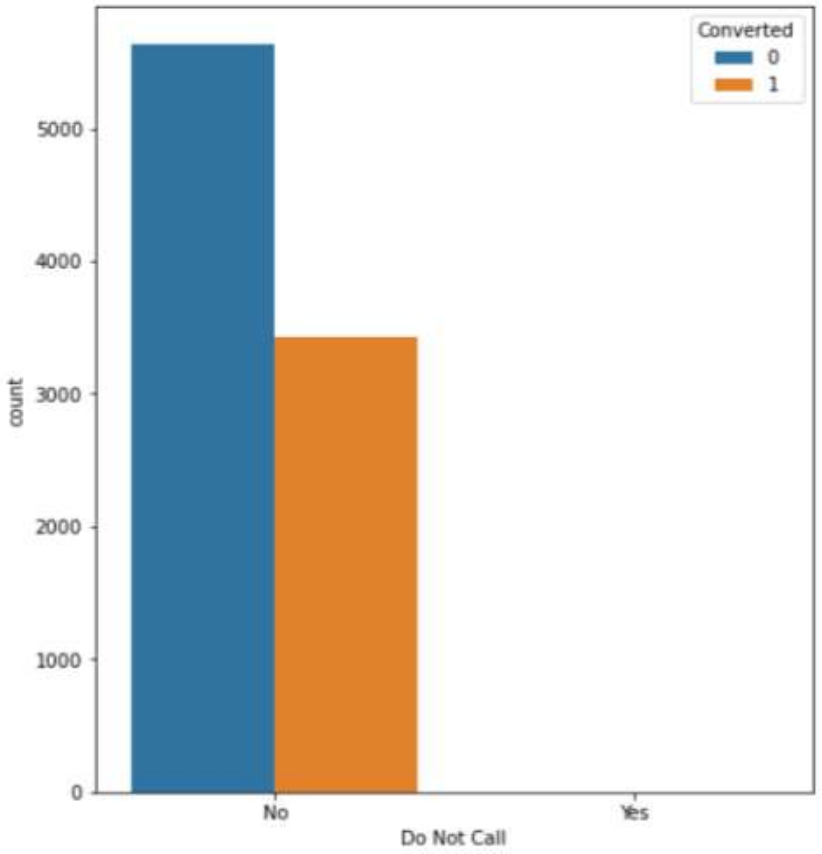
- The count of lead from Direct Traffic and Google have Maximum non-conversion rate.
- Although the count of lead is less in Reference and Welingak Website, the conversion rate is on higher side.
- Facebook and Others categories have less number of leads.

We should focus on improving lead conversion from the categories olark chat, organic search, direct traffic and google since they generate good amount of leads. This way conversion of leads can be improved.



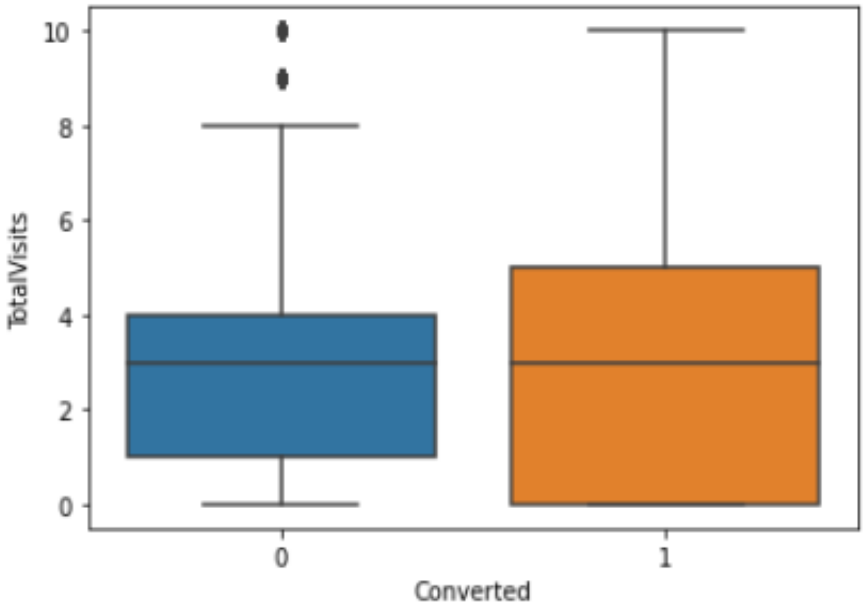
Do Not Email, Do not Call





Total Visits

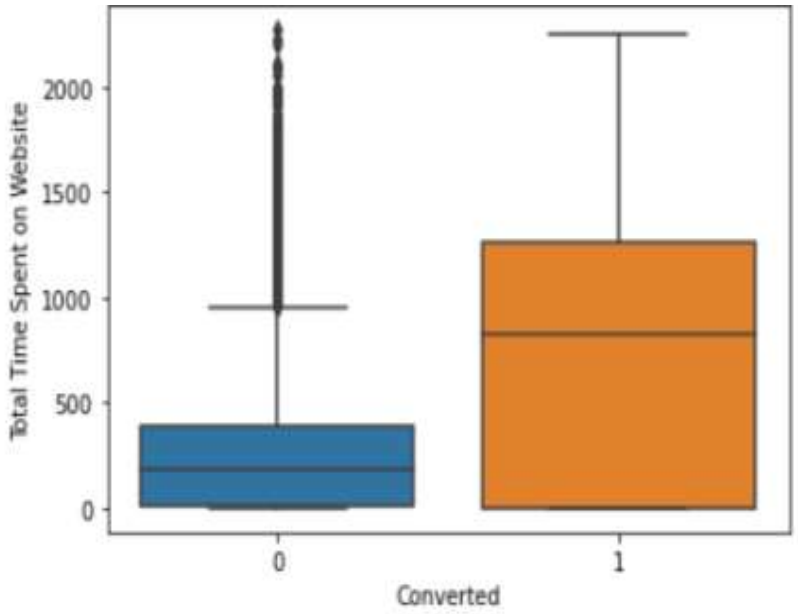
We cannot draw any inference based on total visits, although the median for both of them is same.



Total Time spent on Website

- The Leads who spend more time on website are more likely to be converted.

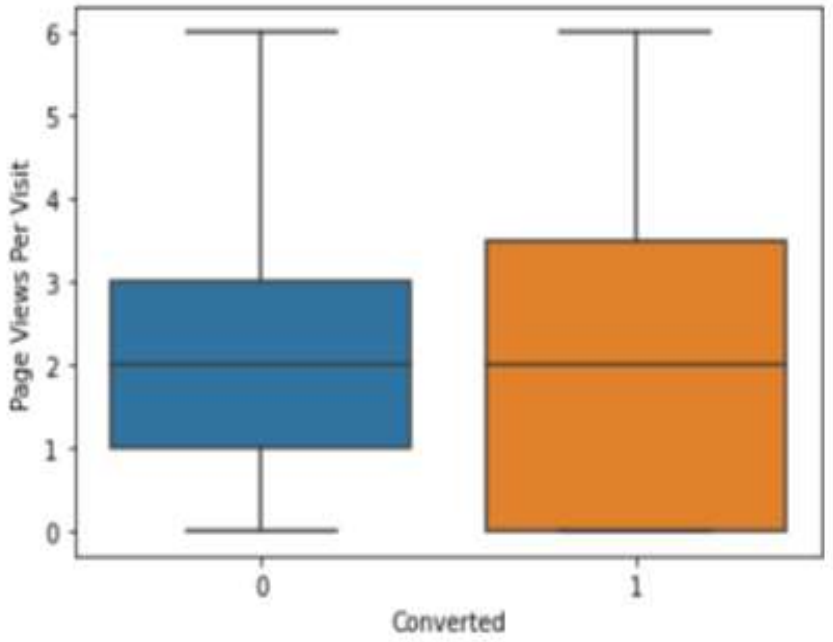
Website should be made more engaging to make leads spend more time.



Page Views Per Visit

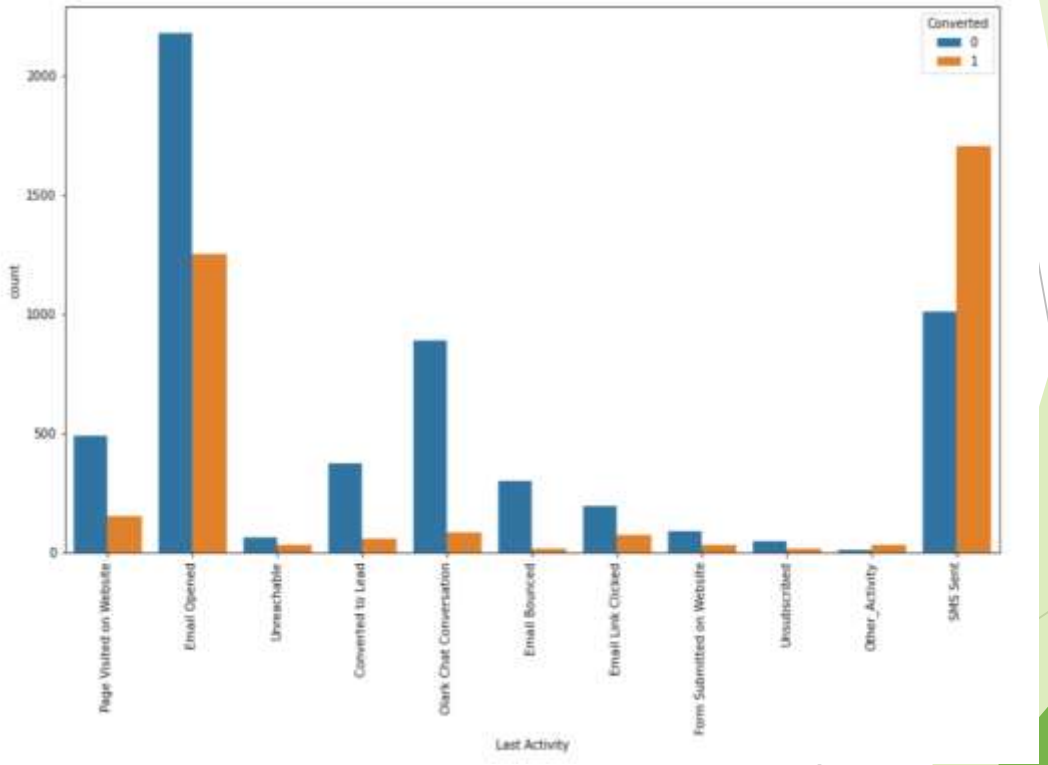
- The median for both converted and non-converted leads is same.

We cannot draw any inference based on Page Views Per Visit



Last Activity

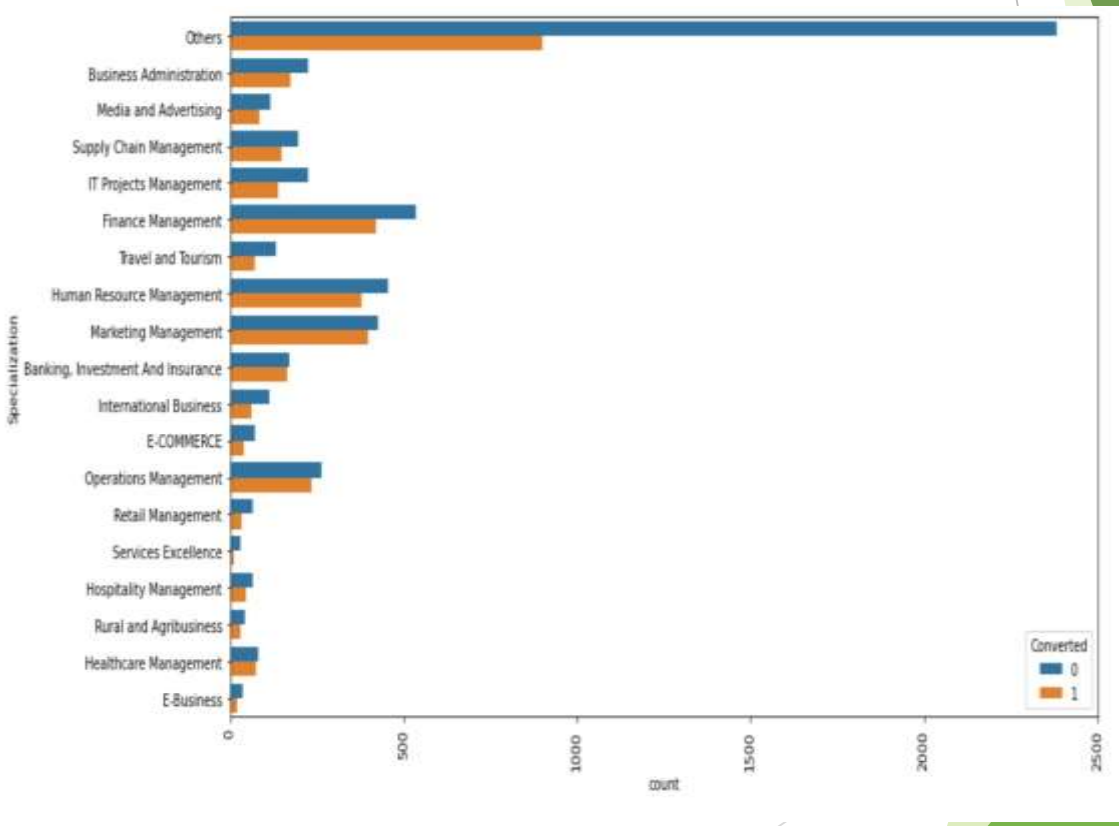
- Most of the lead have opened their Email as their last activity.
- Conversion rate for the leads is higher in SMS sent category when compared to all.



Specialization

- The Maximum incoming leads are more interested in other specializations.

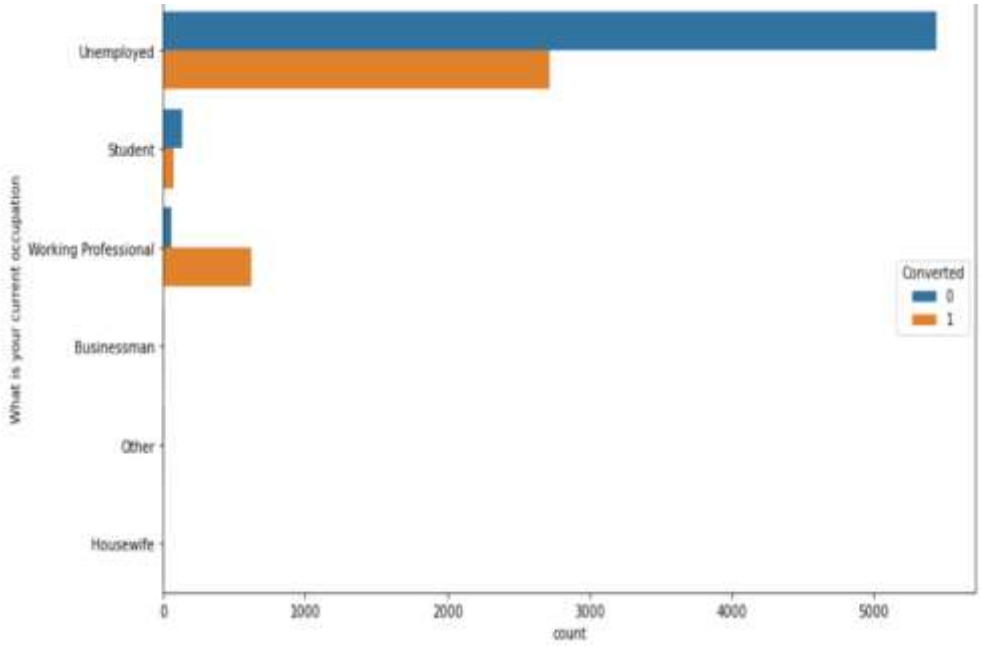
Since the maximum incoming leads are interested in other specializations we should focus on making feasible plans so that the conversion rate increases in that specialization.



What is your current occupation?

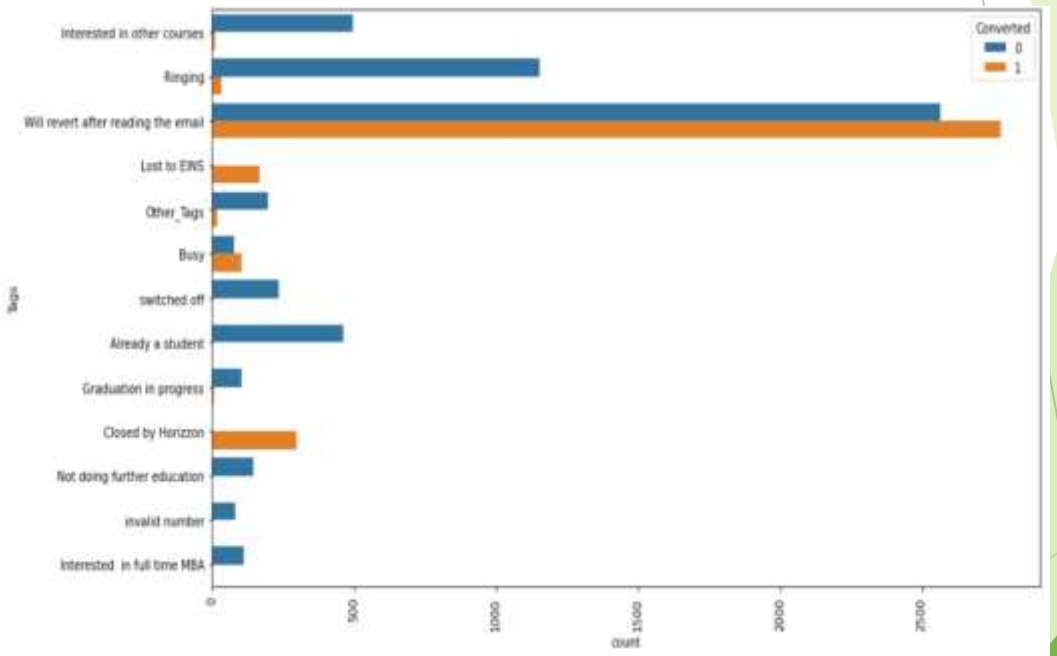
- The incoming leads for the 'Unemployed' category is higher, although the conversion rate is lower.
- The incoming leads for the 'Working Professional' category is lower, although the conversion rate is higher.

We should focus more on the conversion rate of unemployed category also we should concentrate on increasing the leads in working professional category.



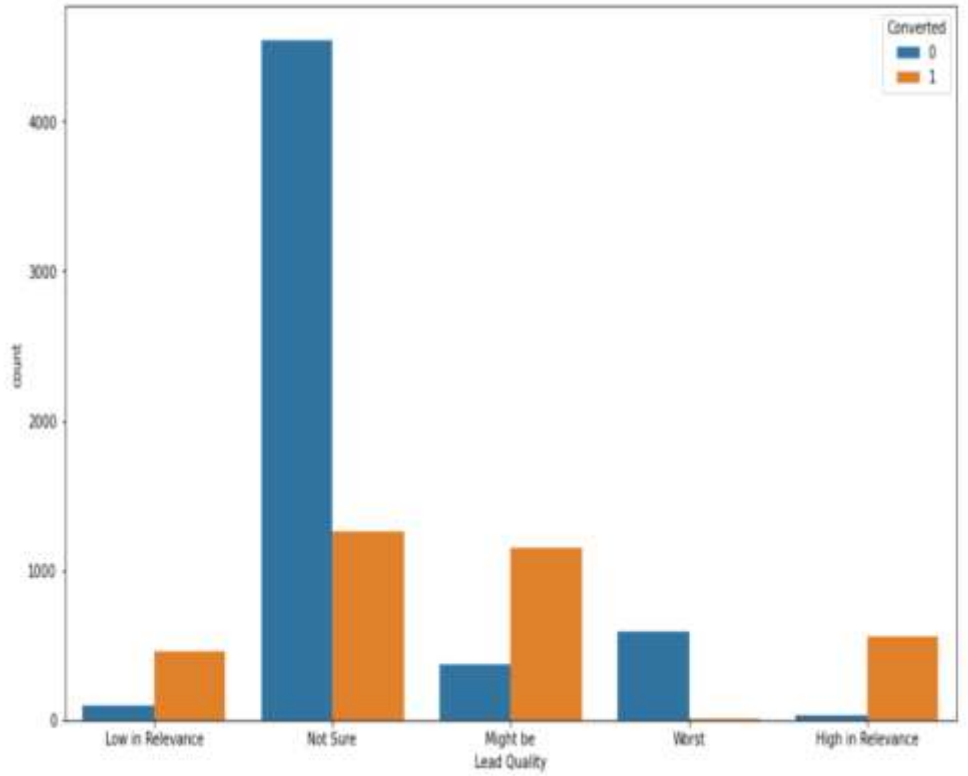
Tags

- There is a higher conversion rate in the "Will revert after reading the email" category.



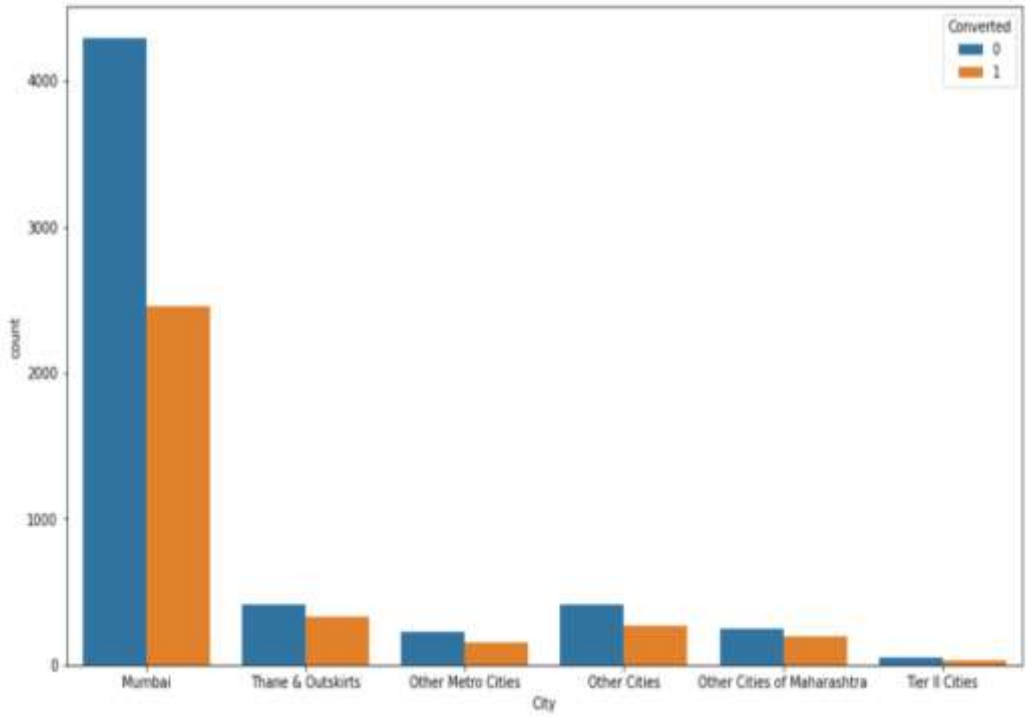
Lead Quality

- The incoming leads for the 'Not Sure' category is higher, although the conversion rate is lower.
- The incoming leads for the 'High in Relevance' category is lower, although the conversion rate is higher.



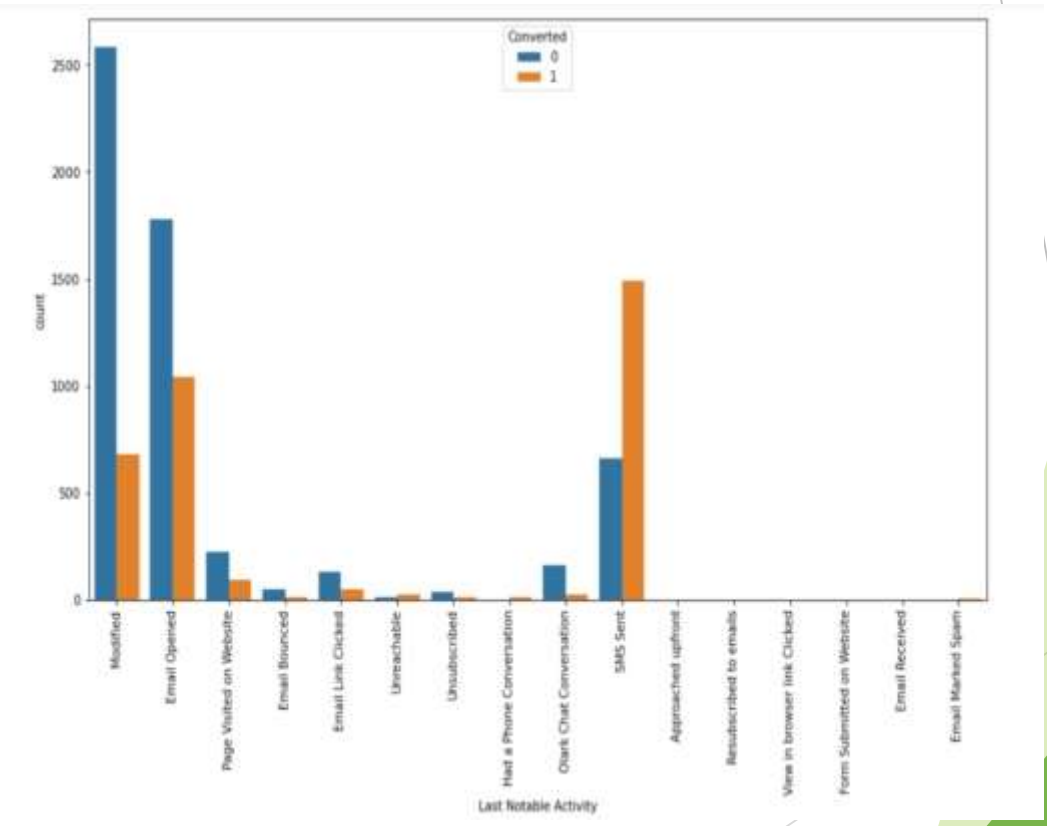
City

- We can observe that conversion rate for Mumbai is higher.
- The people from Tier II Cities are least interested.



Last Notable Activity

- Conversion rate for the leads is higher in SMS sent category when compared to all.



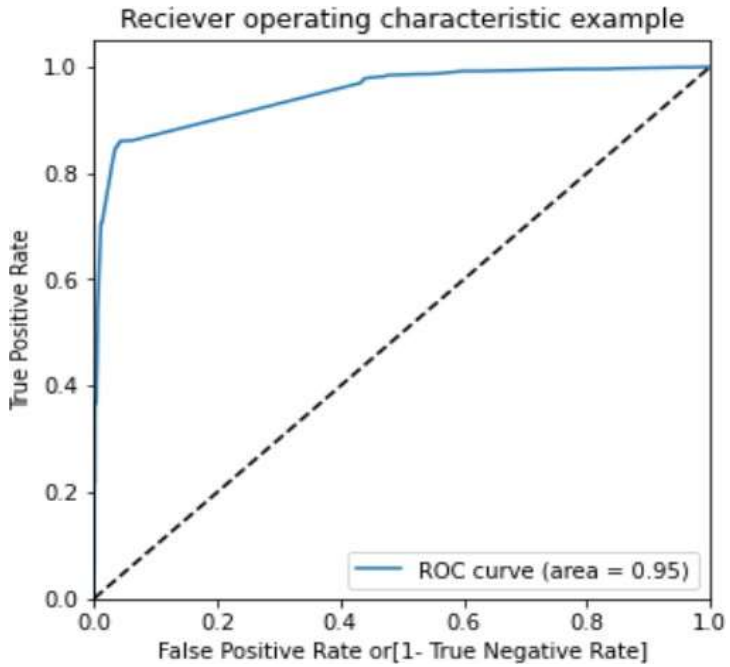
Data Preparation

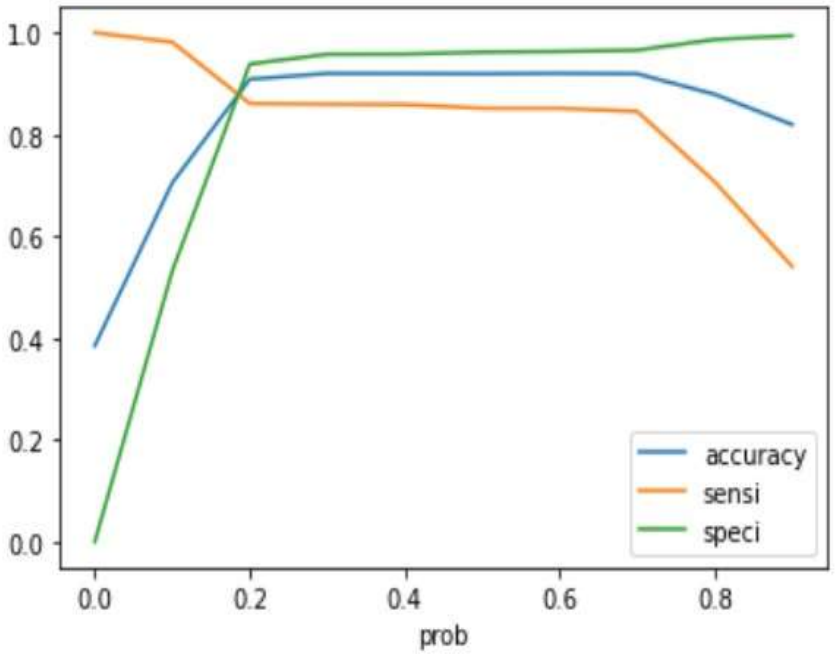
- Conversion of categorical variables into Binary variables. (Do not Email, Do not Call)
- Dummy Variables are created for object type variables
- Total rows for analysis – 9074
- Total columns for Analysis – 86

Model Building

- Splitting the Data into Training and Testing Sets
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Use RFE for Feature Selection
- Running RFE with 15 variables as output
- Building Model by removing the variable whose p- value is greater than 0.05 and VIF value is greater than 5
- Predictions on Train data set
- Overall accuracy is 90%

ROC Curve





- The ROC Curve should be a value close to 1. We are getting a good value of 0.99 indicating a good predictive model.
- From the curve above, 0.2 is the optimum point to take it as a cutoff probability.

Conclusion

- It was found out that the variables that mattered the most in the potential buyers are:
- Do not Email
- Last notable activity was SMS sent.
- When the lead source was:
  - Wellingak website
- When the tags were:
  - Busy
  - Closed by Horizon
  - Lost to EINS
  - Ringing
  - Will revert after reading email
  - Switched Off
- When the lead origin is lead add form and lead import.
- When the lead quality was:
  - Not sure
  - Worst
- When the current occupation was:
  - Unemployed