Lead Scoring Case Study

Presented By:

Saurabh Mudgal

Shashi Ranjan Kumar

Introduction

This case study aims to give you an idea of applying models to help identifying and converting hot leads in a real business scenario.

In this case study, apart from applying the techniques, we have also developed a basic understanding of changing focus to other areas of improvement while targets are met.

Business Understanding

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 website and browse for courses listed there.

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.

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%.

X Education gets a lot of leads, its lead conversion rate is very poor means leads who show interest don't get a deal.

Business Objective

- This case study aims to 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, and it is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

- Also, This case study helps identifying the problems presented by the company by building model which should be able to adjust to if the company's requirement changes in the future so it should handle these as well.

Steps Involved

Data understanding and preparation

- Data Sourcing
- Data Inspection
- Inspecting the null values

Data Cleaning and Manipulation

- Check data types
- Conversions from negative to positive
- Outliers

Data Analysis

- Data imbalance
- Correlation matrix analysis
- Univariate Analysis and Bivariate Analysis for numerical variable
- Load the previous data

Steps Involved

Feature Scaling and Dummy Variables

- StandardScaler for scaling the numeric data columns
- Dummy variables for categorical data columns

Model Building

- Logistic Regression
- RFE to identify the variables
- ROC Curve
- Making prediction

Model Evaluation

- Confusion matrix
- Accuracy, Sensitivity, Specificity

Model Presentation

Conclusions and recommendations

Handling Missing Data

- In the process of data cleaning and manipulation we dropped the columns having null values greater than 40%.

- There were few columns like: 'Do Not Call', 'Search', 'Newspaper Article', 'Magazine', 'X Education Forums',

- 'Digital Advertisement', 'Through Recommendations', 'Newspaper', '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' which had either 1 unique value or second values is almost negligible compare to first.

Some of the columns like: 'Specialization', 'How did you hear about X Education', 'Lead Profile', 'City' had Select character , hence this was replaced with NaN value and later removed based on the analysis.

- Since some of the columns having null value percentage below 40 and seemed to have null values those were imputed using median values.

- some categorical variables (Lead Source, Specialization, tags, Last Activity, Last Notable Activity) were having data which was quite low in frequency hence clubbed such data together with in variable with different names (Others, Management_Specialization, Not Specified, Others, Other_Notable_Activity) respectively.

Change in Datatype and Negative values

- Some of the variables were changed into numeric Data type:

- 'TARGET','CNT_CHILDREN', 'AMT_INCOME_TOTAL', 'AMT_CREDIT','AMT_ANNUITY', 'REGION_POPULATION_RELATIVE', 'DAYS_BIRTH','DAYS_EMPLOYED', 'DAYS_REGISTRATION', 'DAYS_ID_PUBLISH', 'HOUR_APPR_PROCESS_START', 'LIVE_REGION_NOT_WORK_REGION', 'REG_CITY_NOT_LIVE_CITY', 'REG_CITY_NOT_WORK_CITY', 'LIVE_CITY_NOT_WORK_CITY'

- Below columns had negative values which were converted into positive values using abs() function:

- 'DAYS_BIRTH', 'DAYS_EMPLOYED', 'DAYS_REGISTRATION', 'DAYS_ID_PUBLISH'

Exploratory Data Analysis

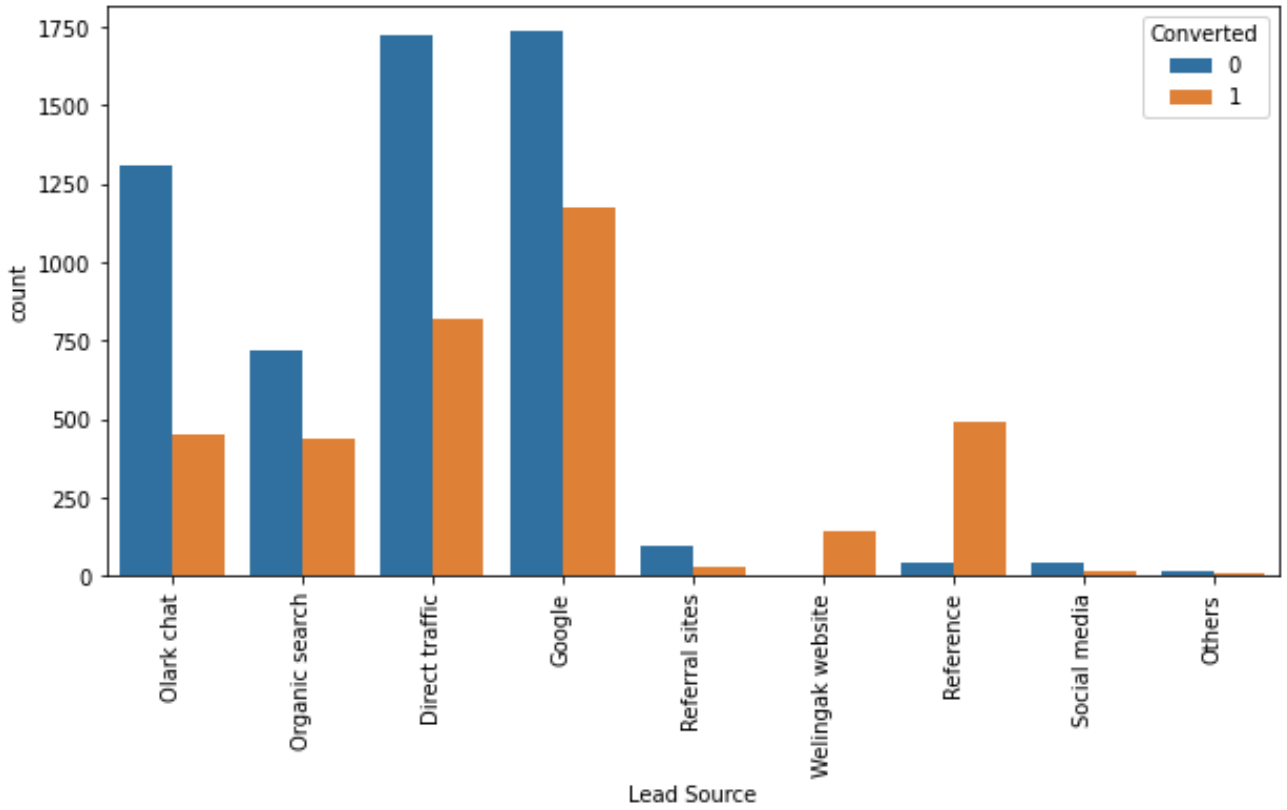
Catagorical Data Analysis

Conclusion for Lead Source on based on Converted 0 and 1::

1.Maximum leads are coming through Google and Direct Traffic.

2.Conversion rate for Reference and Welingak website is maximum.

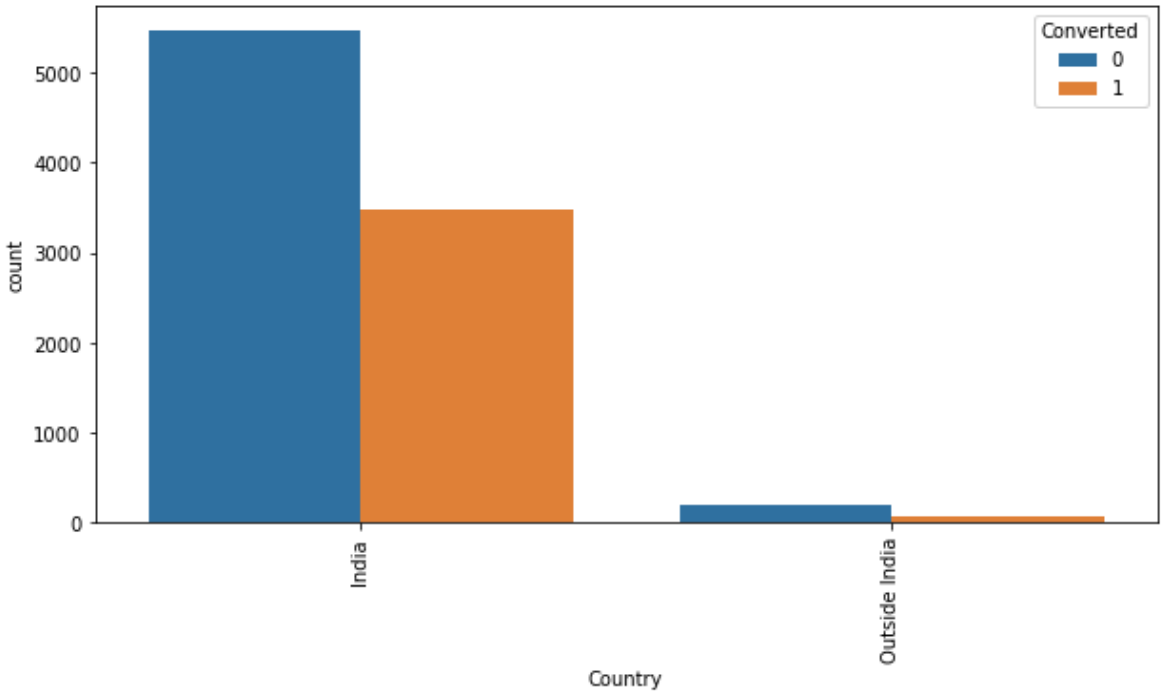
3.To improve the conversion rate , team should more focus on Olark Search , Organic Search, Direct Traffic and Google. Also, focus should be on generating more keads from reference and welingak website.



Conclusion for Country on based on Converted 0 and 1:

1.Except India other countries have very less values so we categorise it as India and outside India.

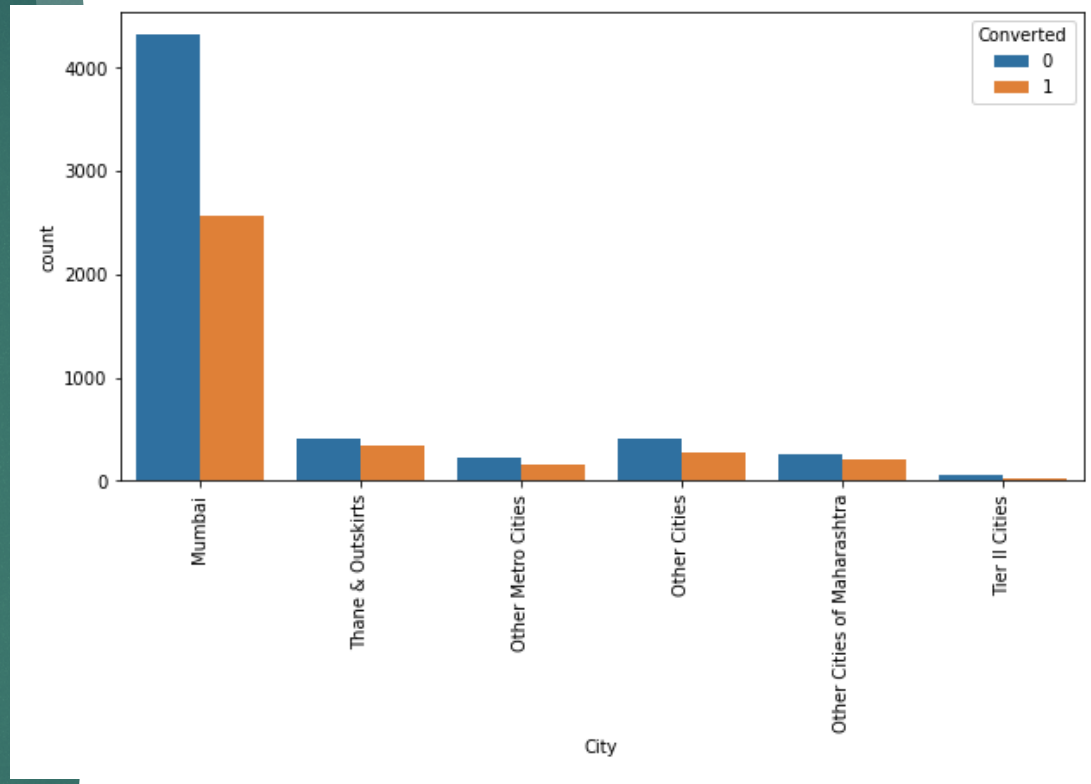
2.India has the maximum conversion rate and for outside india doesn't have much significance.



Conclusion for City based on converted target 0 and 1:

1.Mumbai is having highest number of occurences, hence we dropped this column. Also, all other major values belong to one state only.

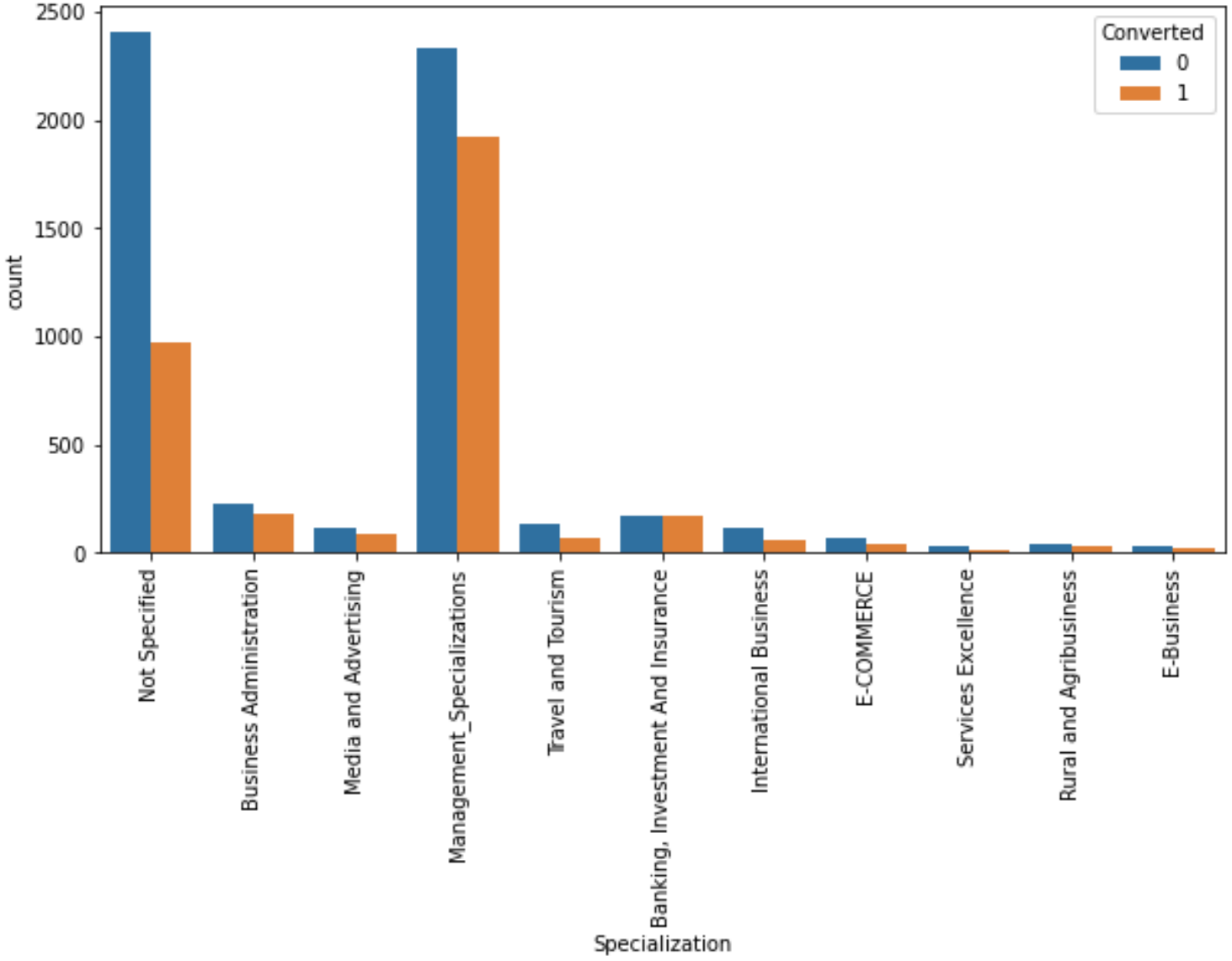
2.Nan Values were imputed with Mumbai since Mumbai has the highest frequency.



Conclusion for Specialization from Target 0 and 1 based on Converted:

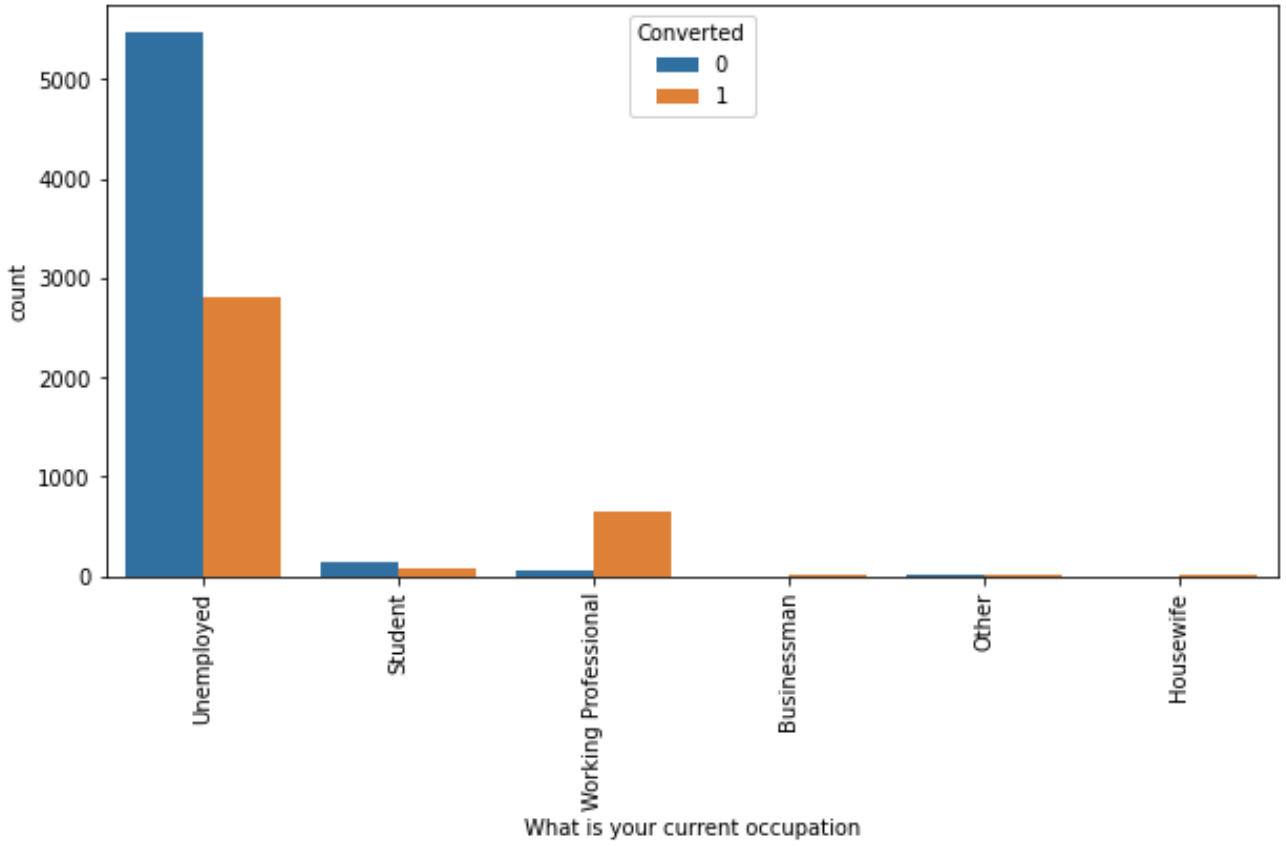
1.Since most of the specialisation from management and had very less frequency, so we combined it together and categorised it as Management_Specilization.

2.After doing analysis we see leads which are having maximum numbers are basically from management also having the maximum leads converted and shows similar trends. It's better we combine these categories.



Conclusion for What is your current occupation from Target 0 and 1 based on Converted:

1.Unemployed category has the maximum number of leads converted which means most of the leads are either passed colleges or not working at all but workingprofessional has the highest conversion rate which means they are looking for upskilling themselves.

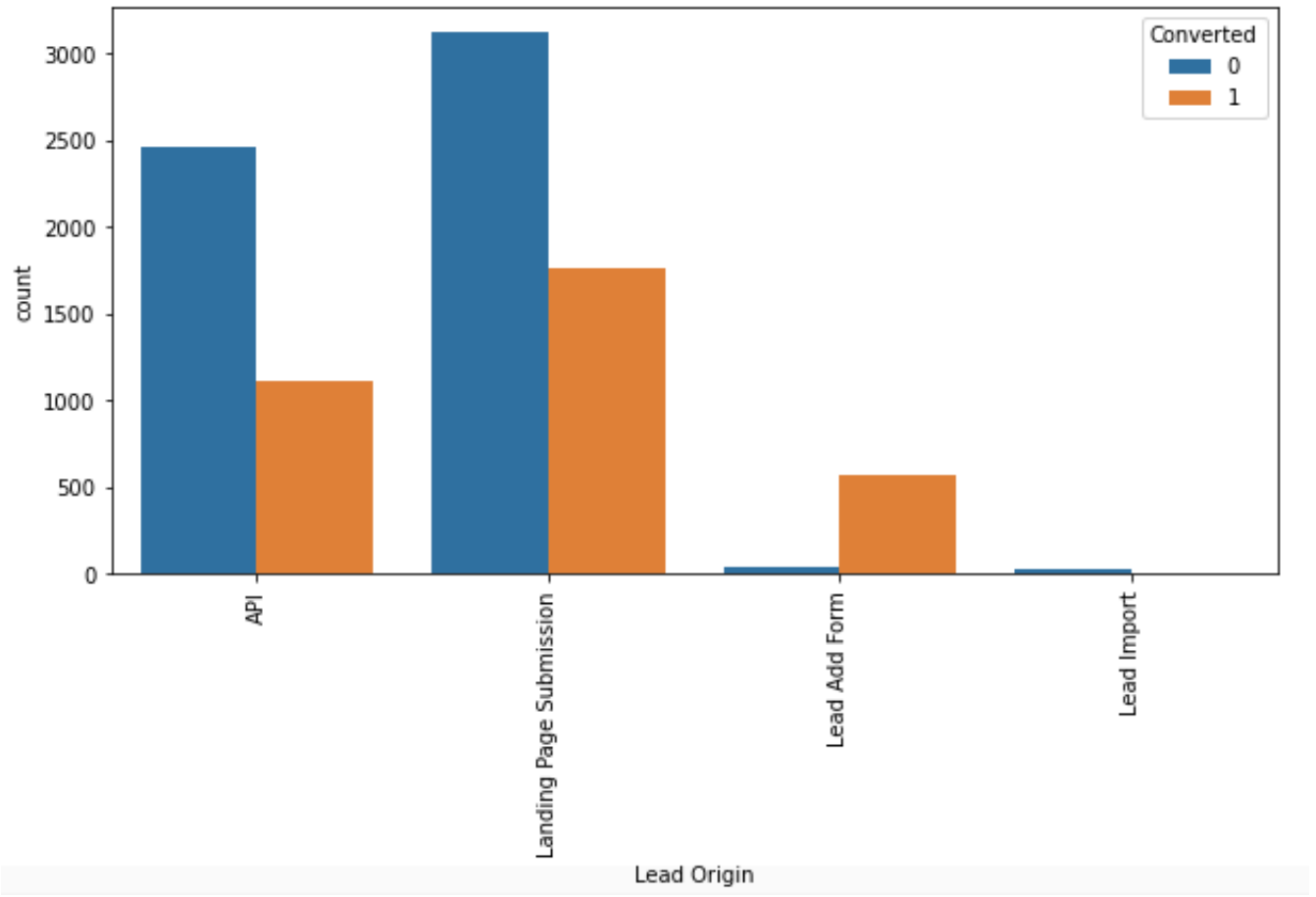


Conclusion for Lead Origin from Target 0 and 1 based on Converted:

1.API and Landing Page Submission both have higher number of leads and conversion too.

2.Lead Add Form has a very high conversion rate but number of leads is not so high.

3.Lead Import has very less leads.

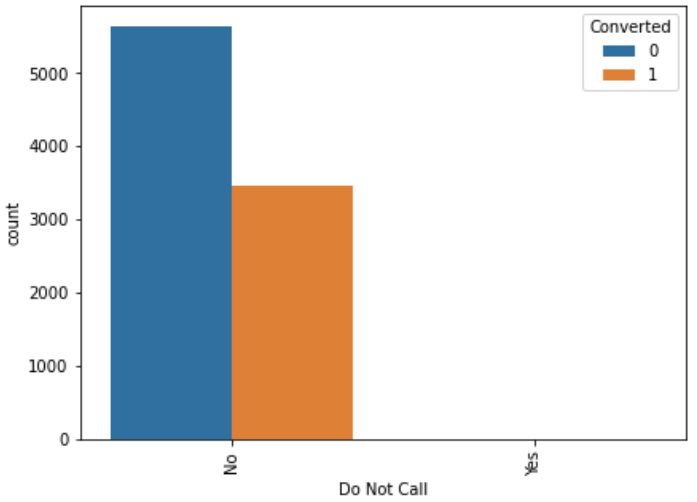


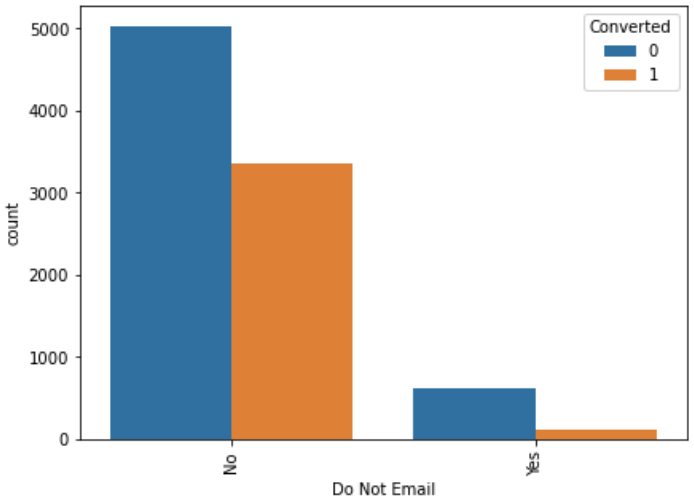
Conclusion for Do not Call and Do Not Email from Target 0 and 1 based on Converted:

1.Leads which said No to Do Not Call and Do Not Email have higher conversion rate.

2.There seems no leads for DO Not Call category who chose Yes .

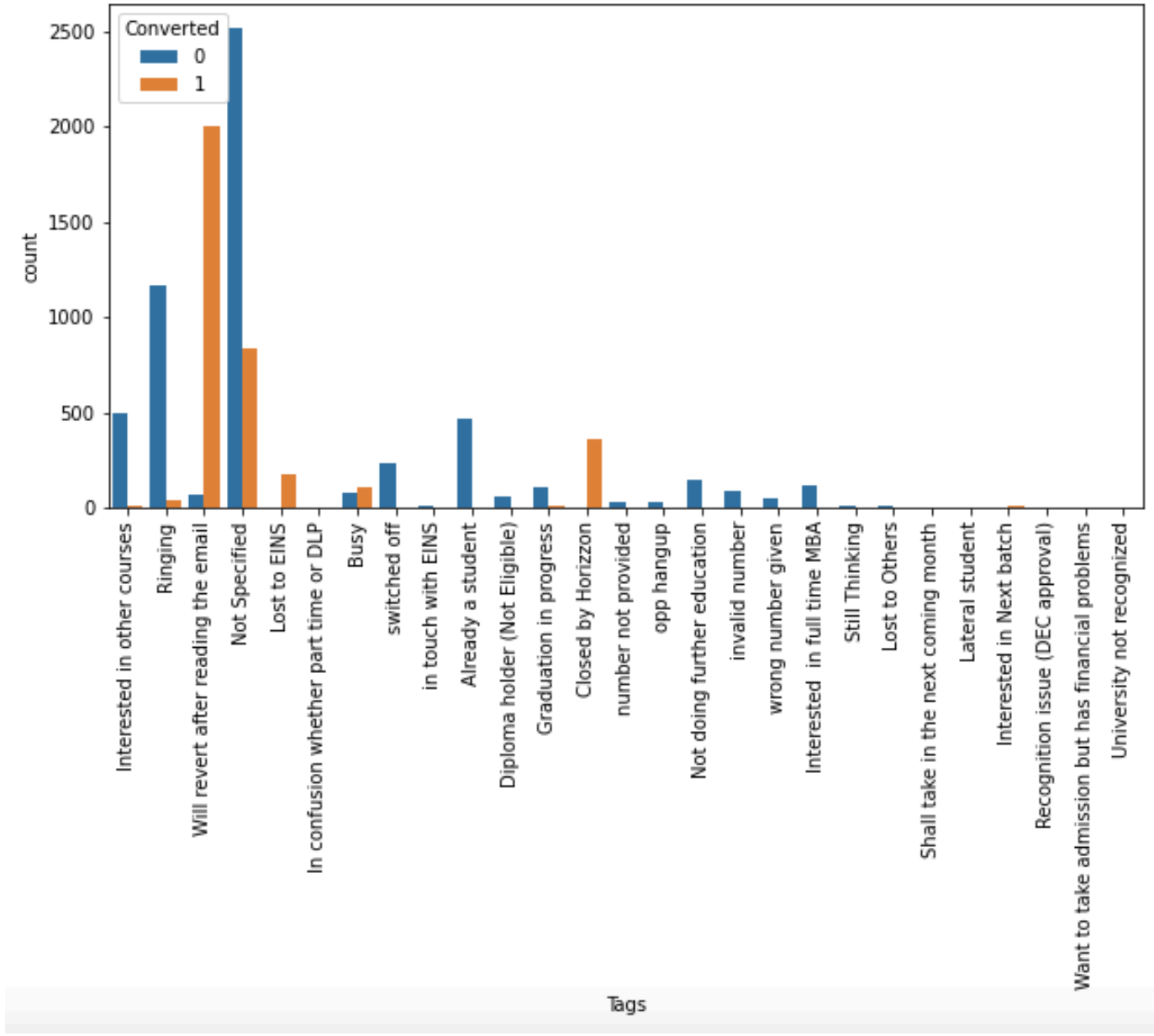
3.But still we have some converted leads for Do Not Email who chose yes.





Conclusion for Tags from Target 0 and 1 based on Converted:

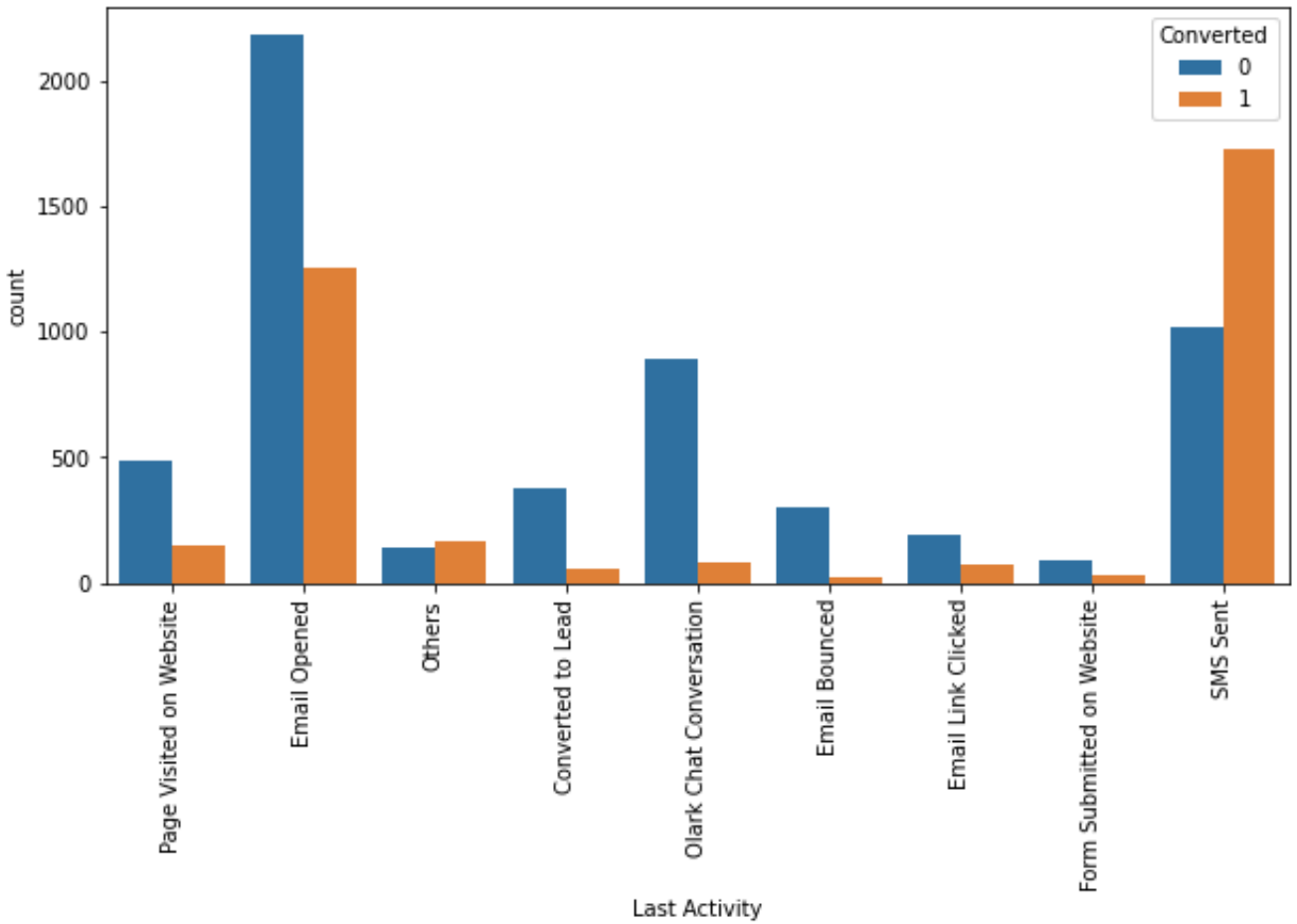
1.Not Specified has maximum leads but category ‘will revert after reading the email has the highest conversion rate. We can dropped Tags column.



Conclusion for Last Activity from Target 0 and 1 based on Converted:

1.SMS Sent has the highest conversion rate among all the categories .

2.Email Opened and Olark Chat Conversation have maximum number of leads.

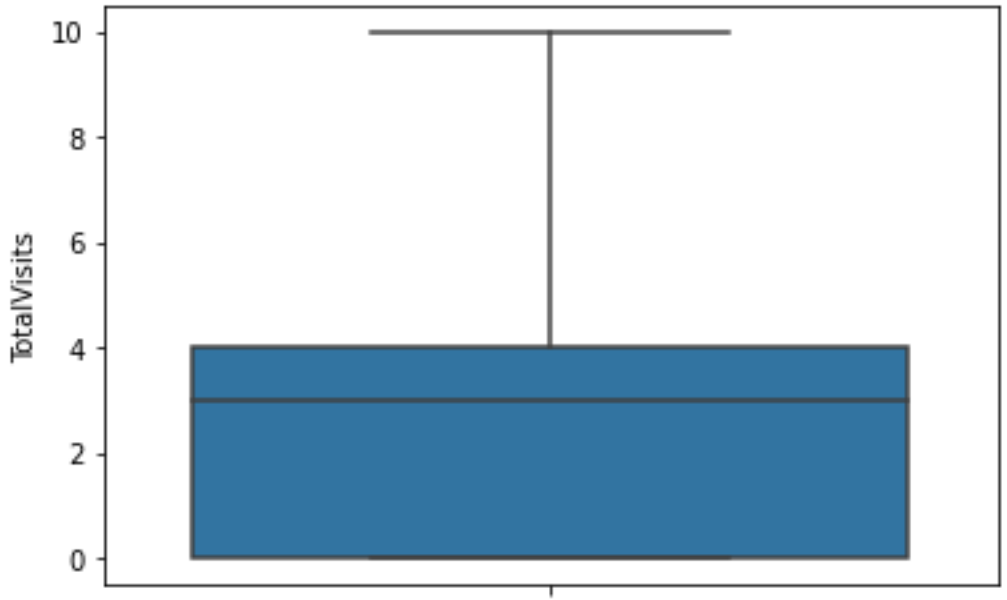


Numerical Data Analysis

Conclusion for Total Visits:

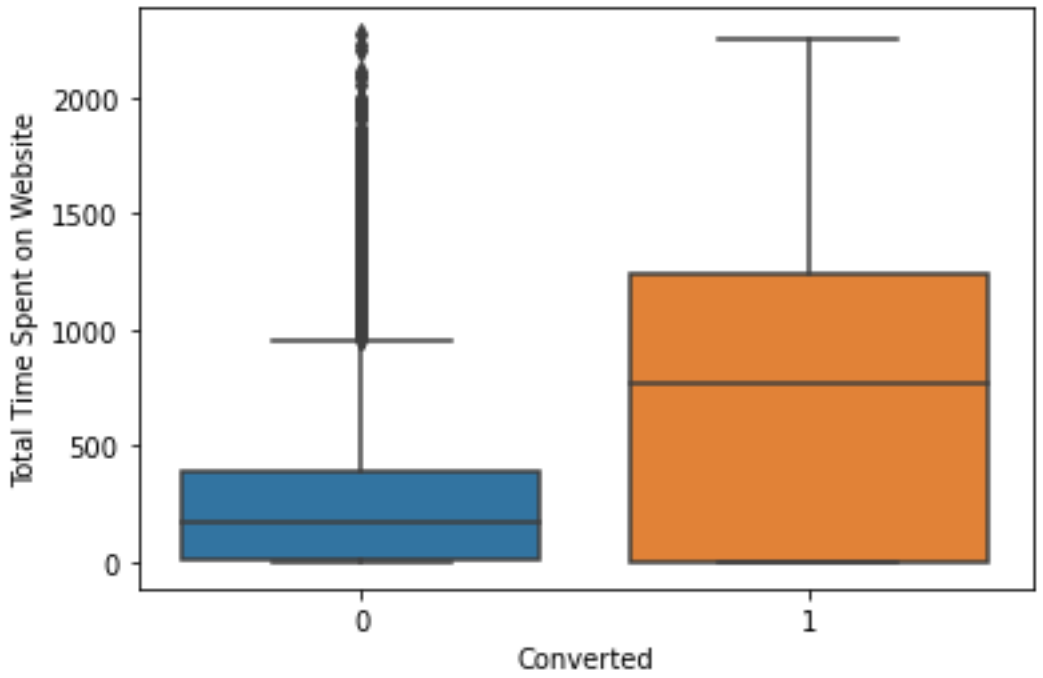
1.Total Visits had some outliers , hence from top and bottom removed 1% outliers.

2.Median is above the half .



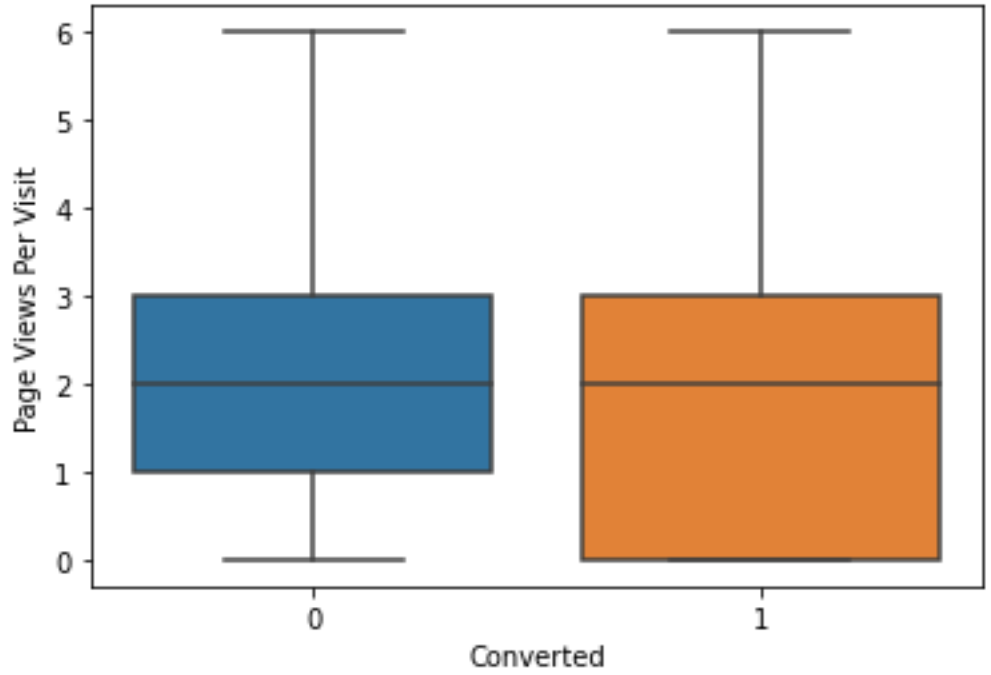
Conclusion for Total Time Spent On Website from Target 0 and 1 based on Converted:

1.Seeing the box plots we can conclude people who have visited the website more are the hot leads.



Conclusion for Page Views Per Visit from Target 0 and 1 based on Converted:

1.People even who have visited less pages are hot leads , may be they just visited the pages for the course they were interested in.



Model Building

Model Building:

1.Splitting the data into train and test data sets.

2.First step is to break data sets into 70:30 ratio for train and test respectively.

3.Use RFE for feature selection.

4.Ran RFE for 15 features.

5. Building model by dropping the variables whose p-value is greater than 0.05 and VIF value is greater than 5.

6.Predictions on test data set.

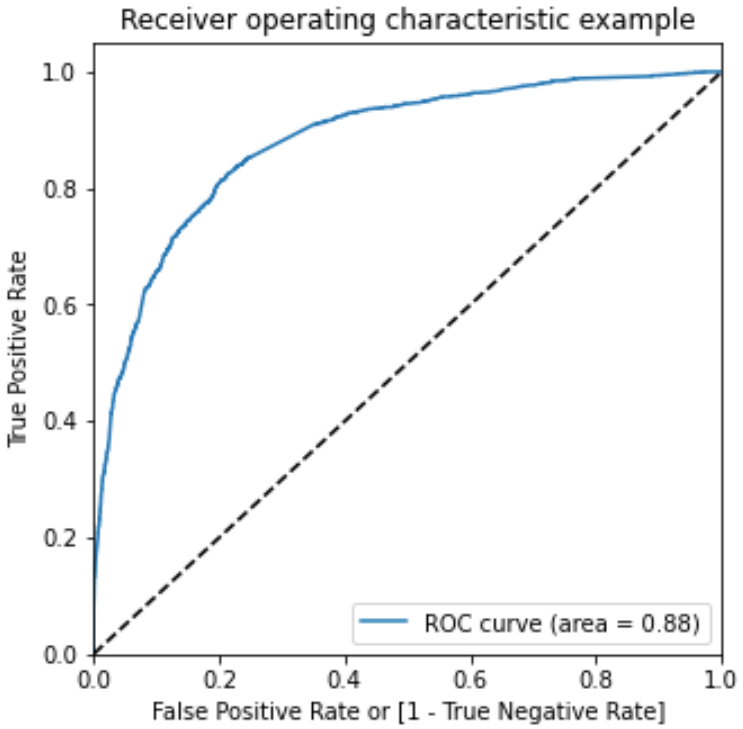
7.Overall accuracy is 81.32%.

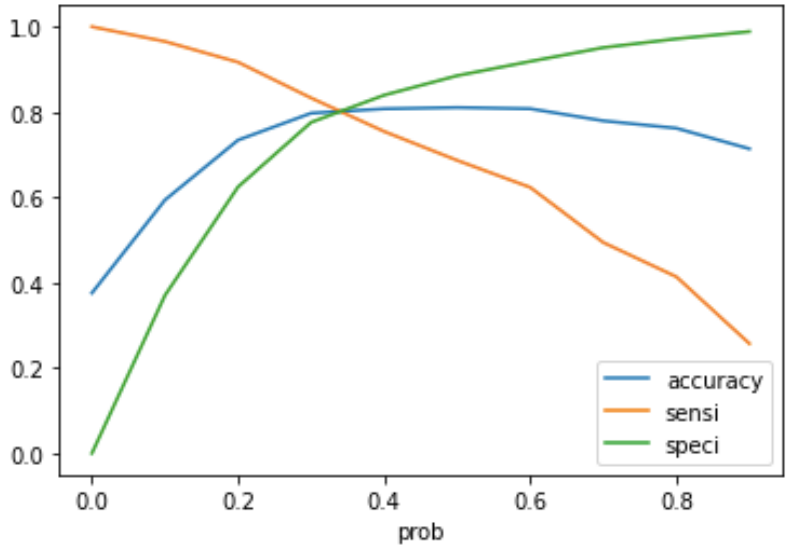
ROC Curve:

1.ROC Curve value we are getting is 0.88, which is a really good value and shows a good predictive model.

2.Finding Optimal Cut Off point is that where we get balanced sensitivity and specificity.

3.From the curve above, 0.37 is the optimum point to take it as a cutoff probability.





Conclusion:

1. 1. The Accuracy, Precision and Recall score we got from test set in acceptable range.
2. 2. We have high recall score than precision score which we were exactly looking for.
3. 3. In business terms, this model has an ability to adjust with the company's requirements in coming future.
4. 4. This concludes that the model is in stable state.
5. 5. Important features responsible for good conversion rate or the ones' which contributes more towards the probability of a lead getting converted are :
6. - Total Time Spent on Website
7. - Lead Origin_Landing Page Submission
8. - What is your current occupation_Working ProfessionalLoan purpose 'Repair' has higher number of unsuccessful payments on time.
9.

Thanks!