Lead Scoring Case Study

Presented By:
Rahul Yadav
Sheetal Rudrawar
Rohith Nagula

Problem Statement

- X education data and they sell online courses to industry professionals. People interested in thecourse land on the website and browse courses.

- The company markets its courses on several websites and search engines like like Google. Once these people land on the website, they might browse the courses or fill up a form for the courseor watch some videos. When these these people fill up a form providing their email address or phone number, they are classified to be a lead.

- X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with higherlead 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%.

Steps Followed

- ❑ First will take the dataset, understand it and checks the columns description to understand all the parameters involved in the problem.

- ❑ After importing the dataset and the libraries will find out the shape, info, description, which will give us the no of rows, columns, null values, mean, median, standard deviation.

- ❑ The next step is Data Cleaning, which is the important step, to check the % of null values and how we can impute those values, dropping the columns which are not relevant to the problem at the time.

- ❑ So, after checking the missing values, dropped the columns which were having missingvalues more than 40% for further positive analysis.

❑ At the start, dataset had 37 columns, after dropping 40%, was left with 30 columns.

❑ Dropped Lead Number and Prospect ID as well, because these are just generated for thereference of the lead.

❑ Then, ‘Select’ was present in most of the columns so we have replaced it with NaN.

❑ Dropped 4 more columns which didn’t seem relevant for the analysis, Now left with 26 columns.

❑ Some of the columns(‘’Country, Specialization, How did you hear about X Education ,Whatis your current occupation, What matters most to you you in choosing a course”), still have null values present, we looked into them individually to see what can be done.

- ❑ We did value counts and imputed the columns with the mode as there are categorical columns.

- ❑ Dropped Country column because in country column- India and Nan combine willgive you 97% of the distribution, so its safe to drop it.

- ❑ Dropped How did you hear about X Education as well because of the null valueswhich were 78%.

- ❑ I have imputed categorical columns with mode and combined the columns for easyand better analysis.

- ❑ Now, we are left with 23 columns and all the columns have 0 null values.

- ❑ We checked the conversion rate at the start, which is 45.5%.

Exploratory Data Analysis

- cat_cols=['Lead Origin’, 'Lead Source’, 'Do Not Email’, 'Do Not Call’, 'LastActivity’, 'Specialization’, ' What is your current occupation’, 'What matters most to you in choosing a course’, 'Search’, 'X Education Forums’, 'Newspaper’, 'Digital Advertisement’, 'Through Recommendations’, 'Receive More Updates About Our Courses’, 'City’, 'Afree copy of Mastering The Interview’, 'Last Notable Activity’]

- con_cols=['Converted','TotalVisits','Total Time Spent on Website’, 'PageViews Per Visit']

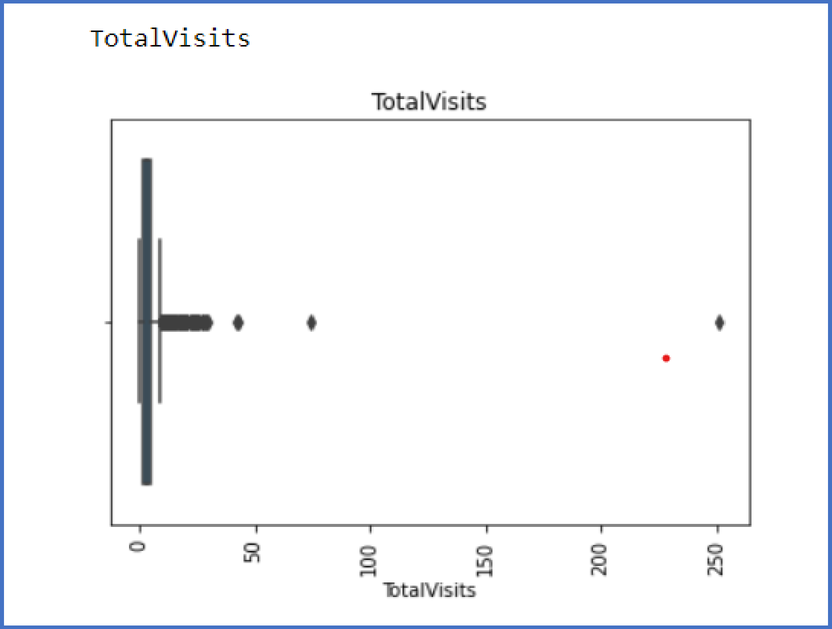
Univariate Analysis of Catergotical Columns :-

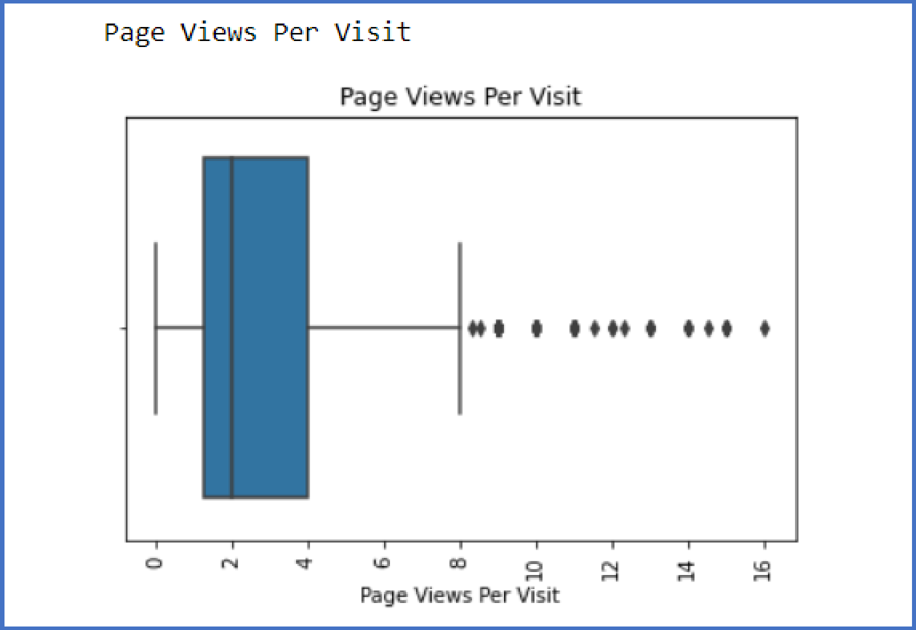
- Lead origin is mostly **landed on the page submission** than API.
- Majority of the leads are from **GOOGLE,** then direct traffic , Olark Chat and organic engines.
- Most people said **NO** to **email and calls.**
- **Email opened** was the **highest activity performed** by a customer then SMS sent comes after it.
- **Management courses are more in demand** and after that customer haven't specified the **specialization**.
- **Unemployed** tops in occupation working professional comes afterwards.
- **Better Career Prospects matters** most to the customers who wants to choose a course.

- Big **NO** to **search , Education forums, Newspaper, Digital advertisement.**
- **No one** has come through **recommendations.**
- **No one** wants to receive updates about the courses**.**
- Most people have **not mentioned their City** after that Mumbai comes in the list.
- **No** to A f**ree copy** of Mastering The Interview.
- **Modified, email opened and SMS sent** was the **Last Notable Activity** for most of the customers

Checking outliers of continuous columns :-

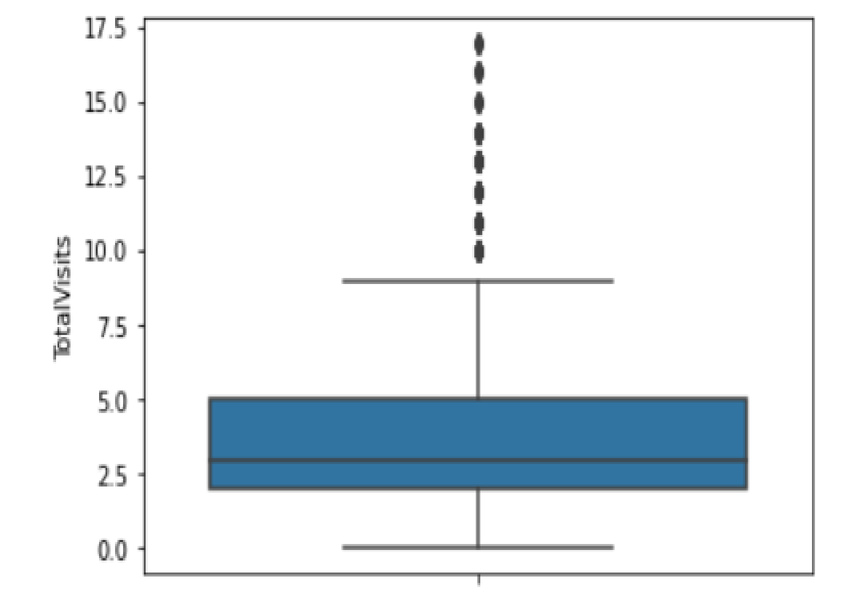
As we can see from box plot of continuous columns there are no such outliers present in the data. Though 'total visits' and 'Page Views Per Visit' has some outliers.

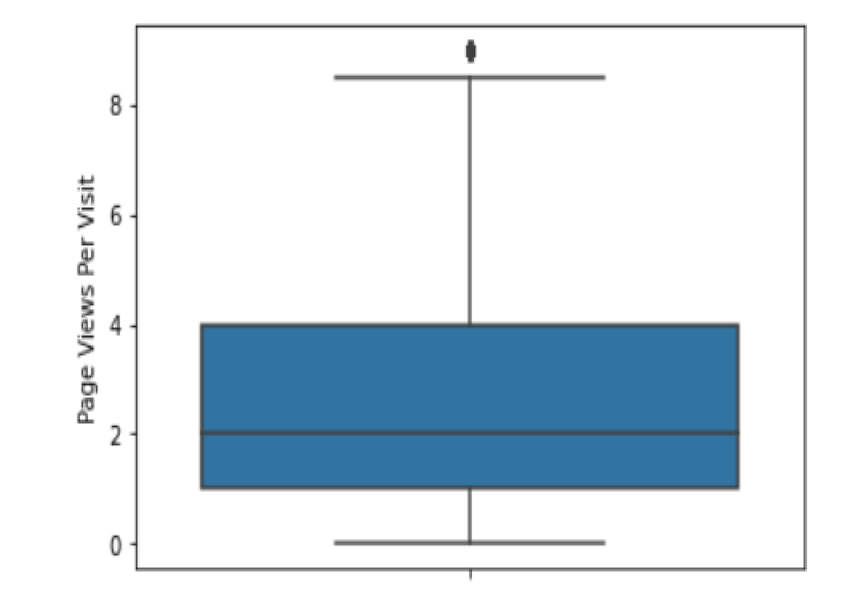

```
TotalVisits
```


```
Page Views Per Visit
```

Outliers Treatment :-

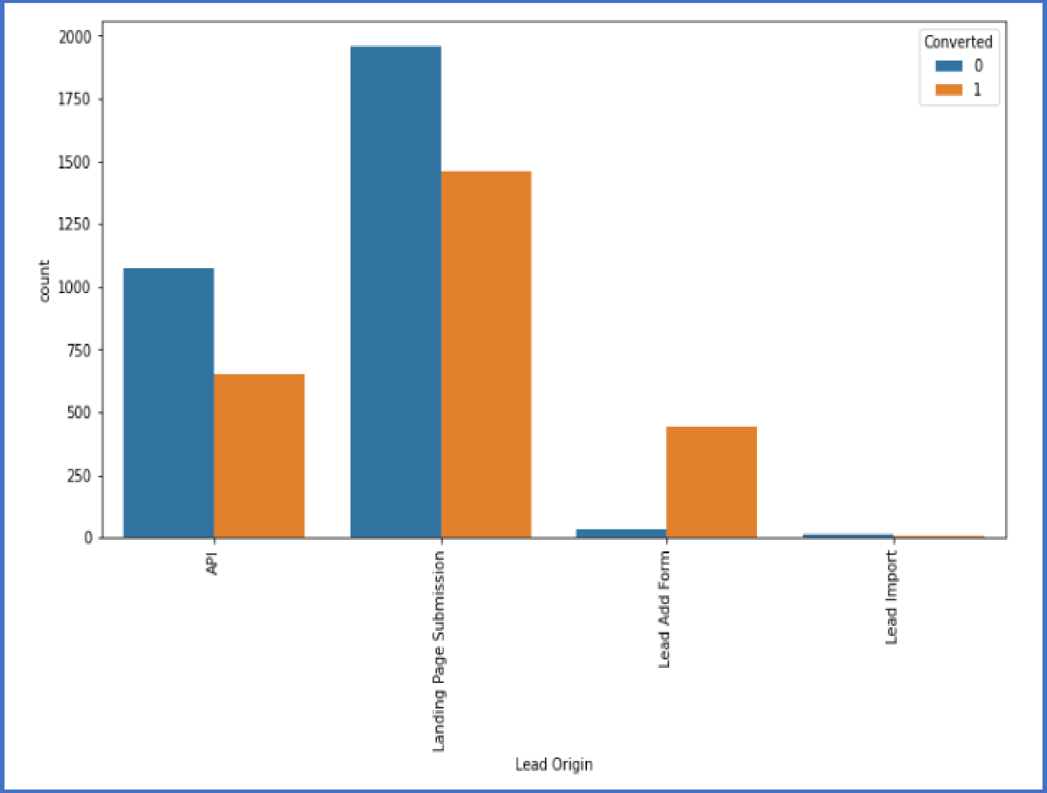
Treatment of outliers by removing top and bottom 1% of the outlier values.





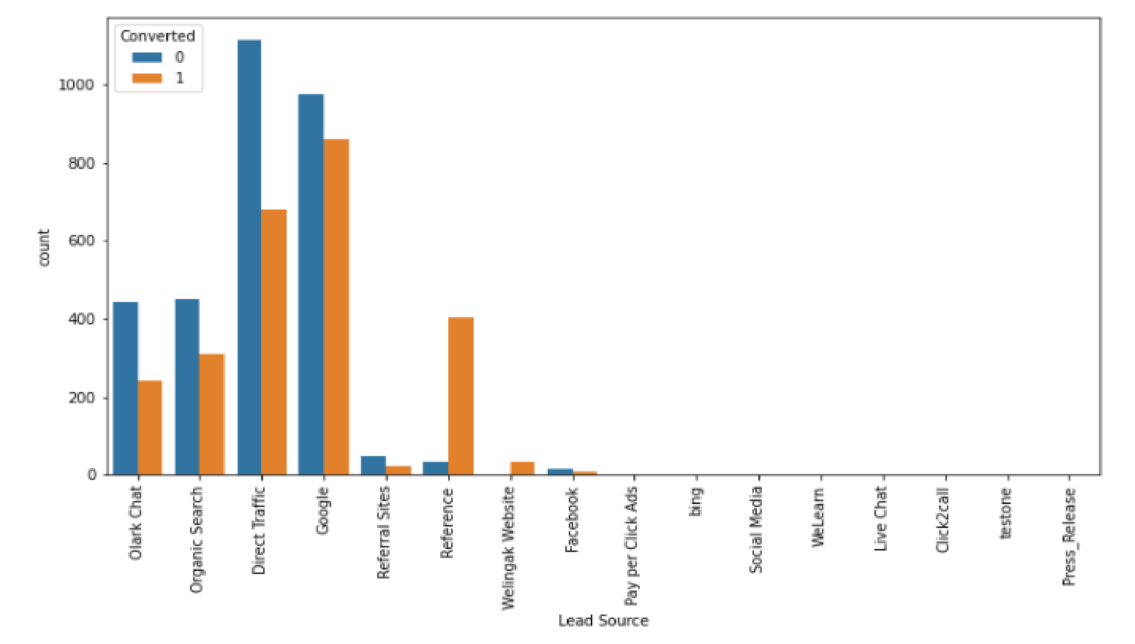
Bivariate Analysis

Bivariate analysis with target column converted



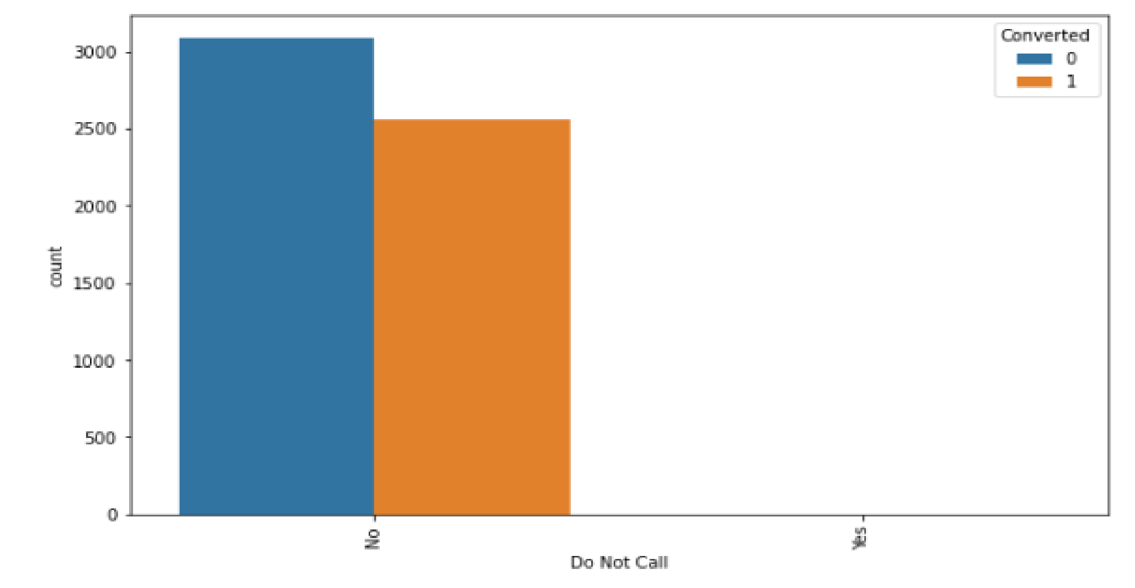
Lead Origin

Landing page submission has had High lead conversations.



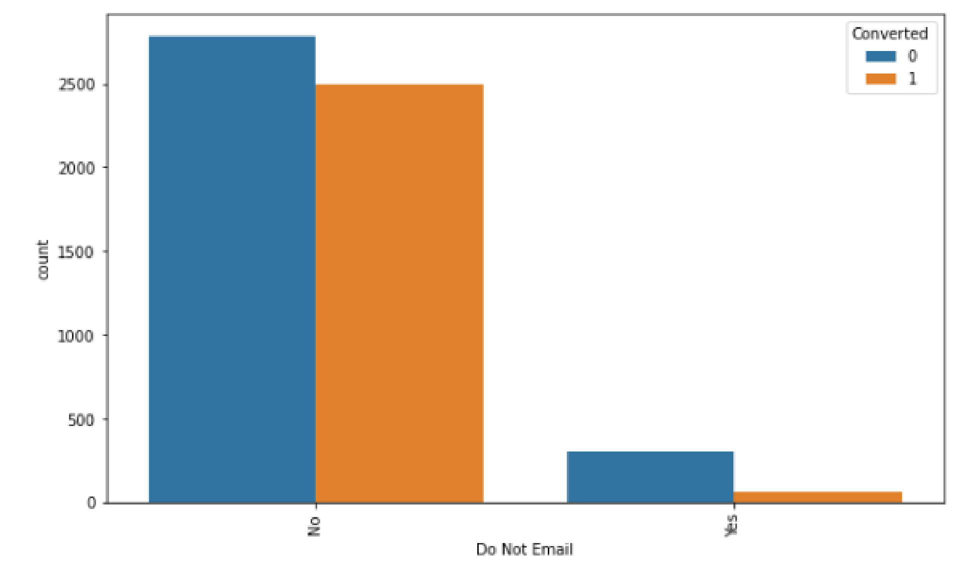
Lead Source

Majority of the leads are from **GOOGLE,** then direct traffic , Olark Chat and organic engines.



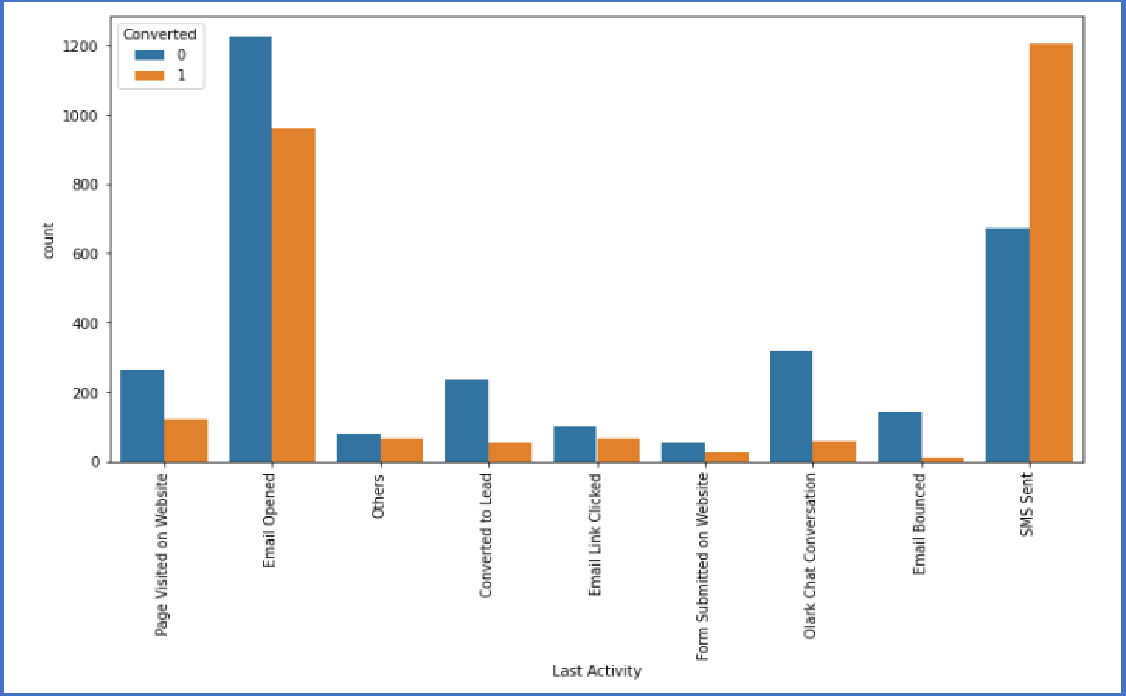
Do Not Call

Most people said **NO** to **calls**



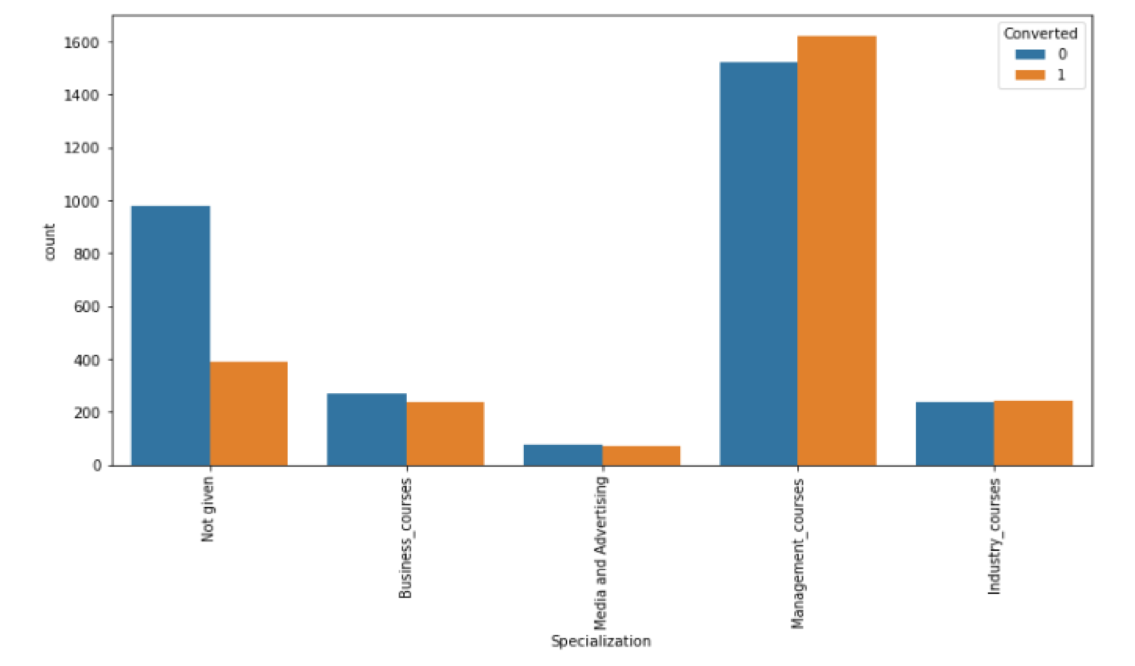
Do Not Email

Most people said **NO** to **Email.**



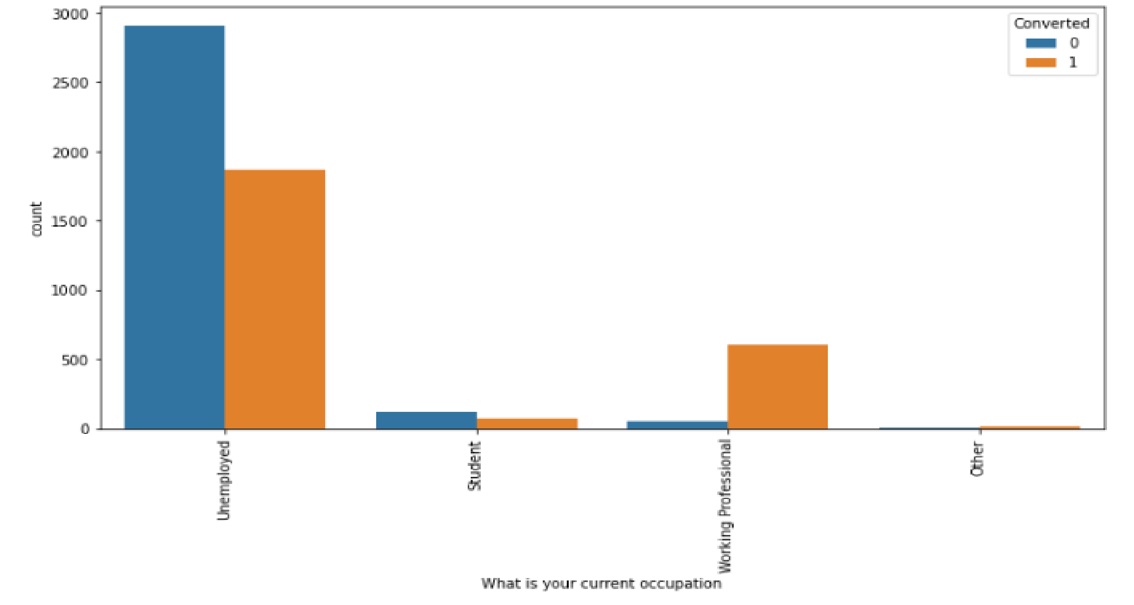
Last Activity

Email opened was the **highest activity performed** by a customer then SMS sent comes after it.



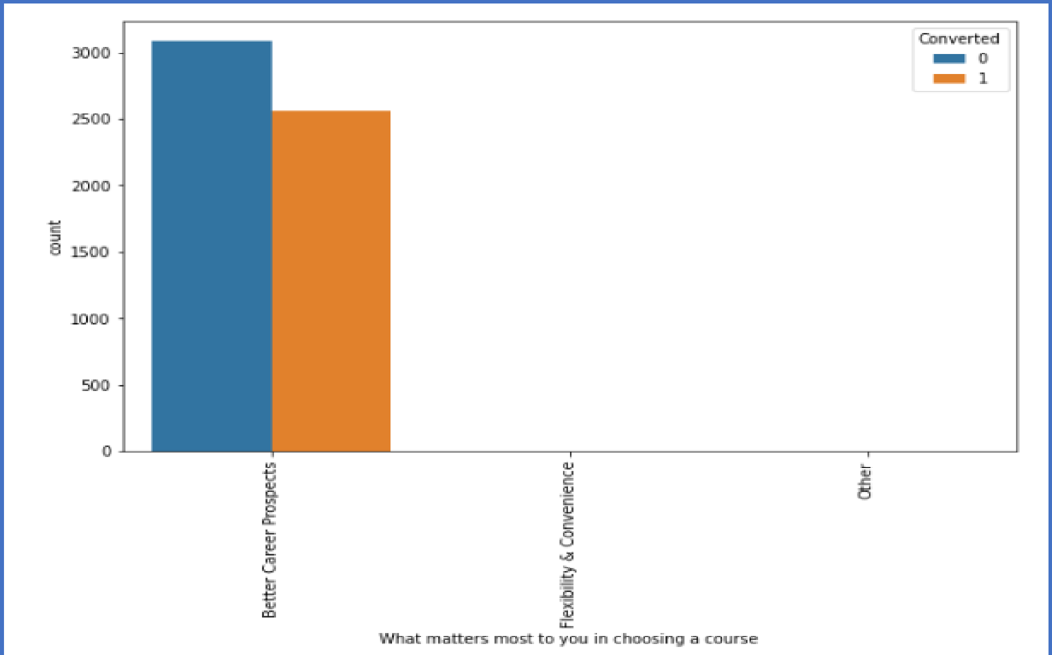
Specialization

Management courses are more in demand and after that customer haven't specified the **specialization.**



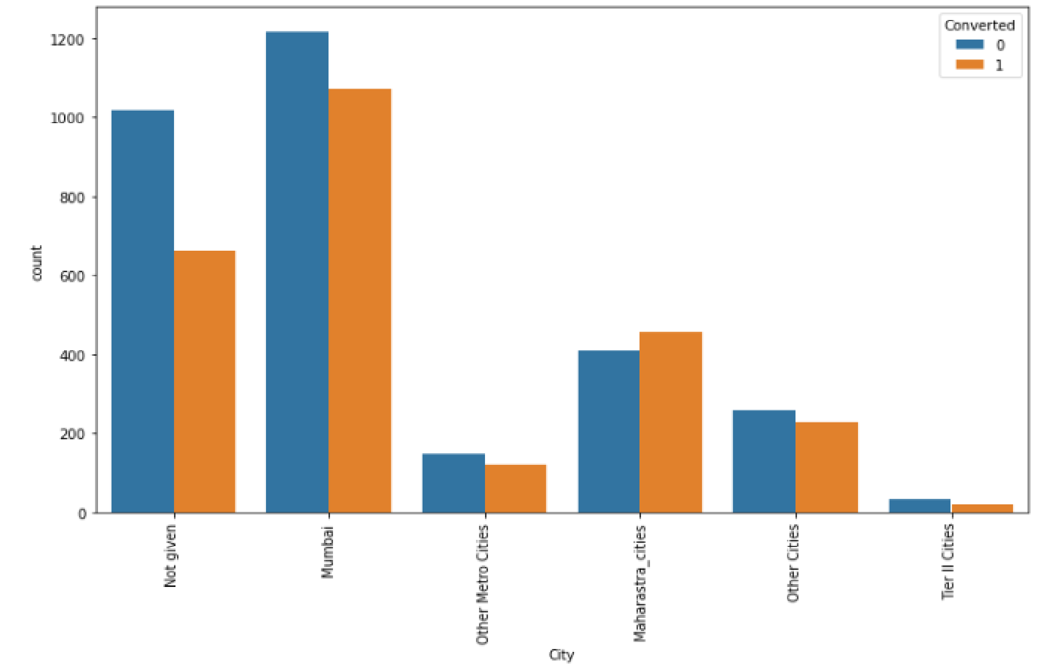
What is your Current Occupation

Unemployed tops in occupation working professional comes afterwards.



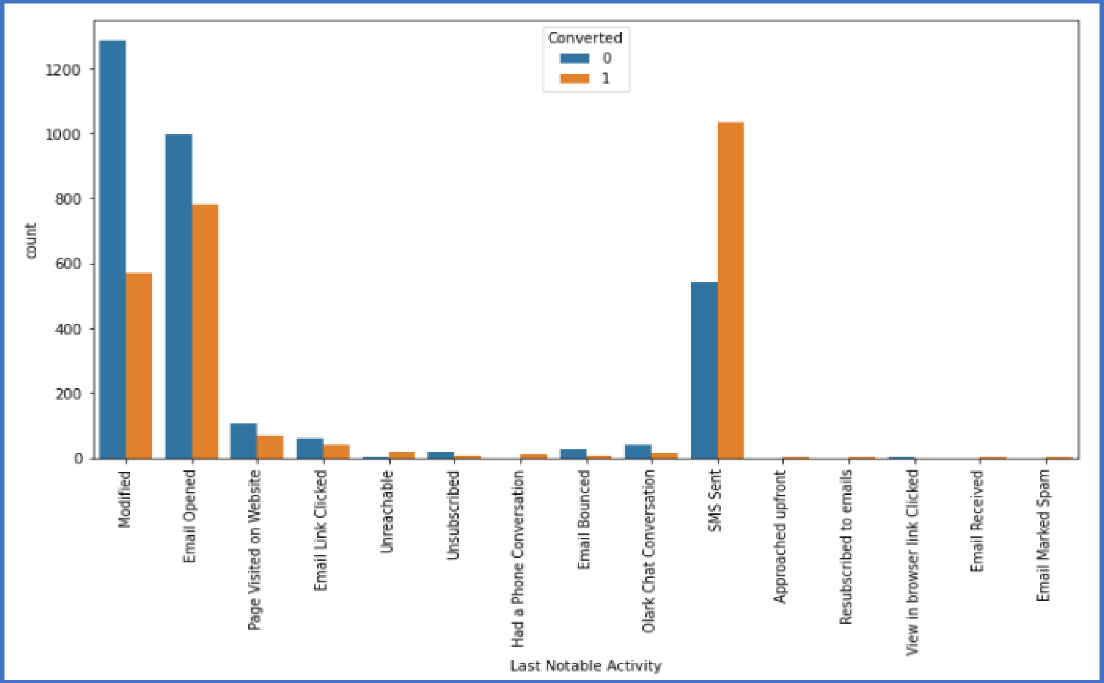
What matters most to you in chossing a course

Better Career Prospects matters most to the customers who wants to choose a course.



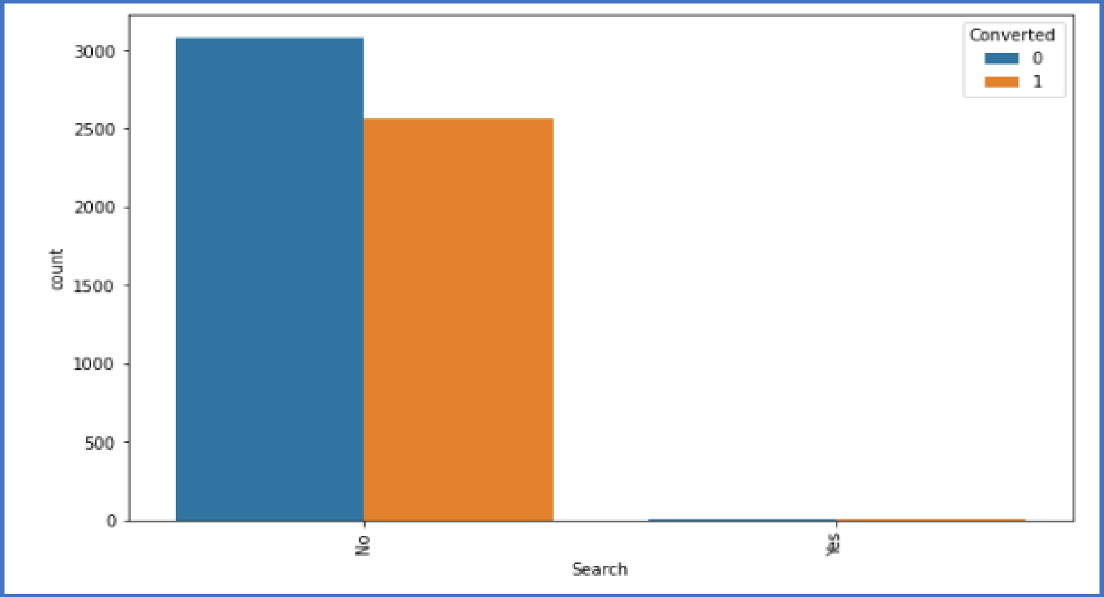
City

Most people have **not mentioned their City** after that Mumbai comes in the list.



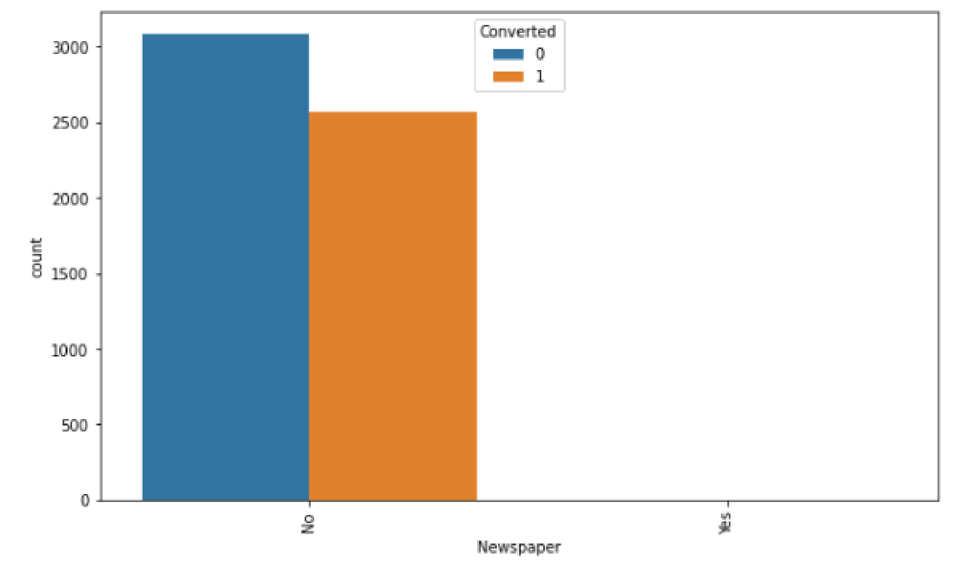
Last Notable Activity

Modified, email opened and SMS sent was the **Last Notable Activity** for most of the customers.



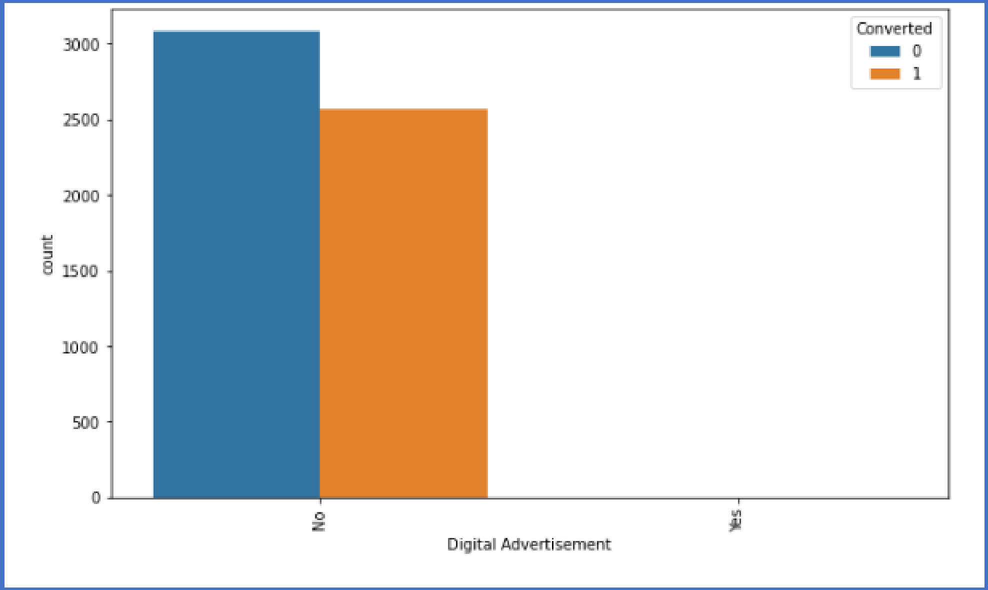
Search

The graph shows that Searches are Not good source of leads.



Newspaper

The graph shows that Newspaper is Not a good source of leads.



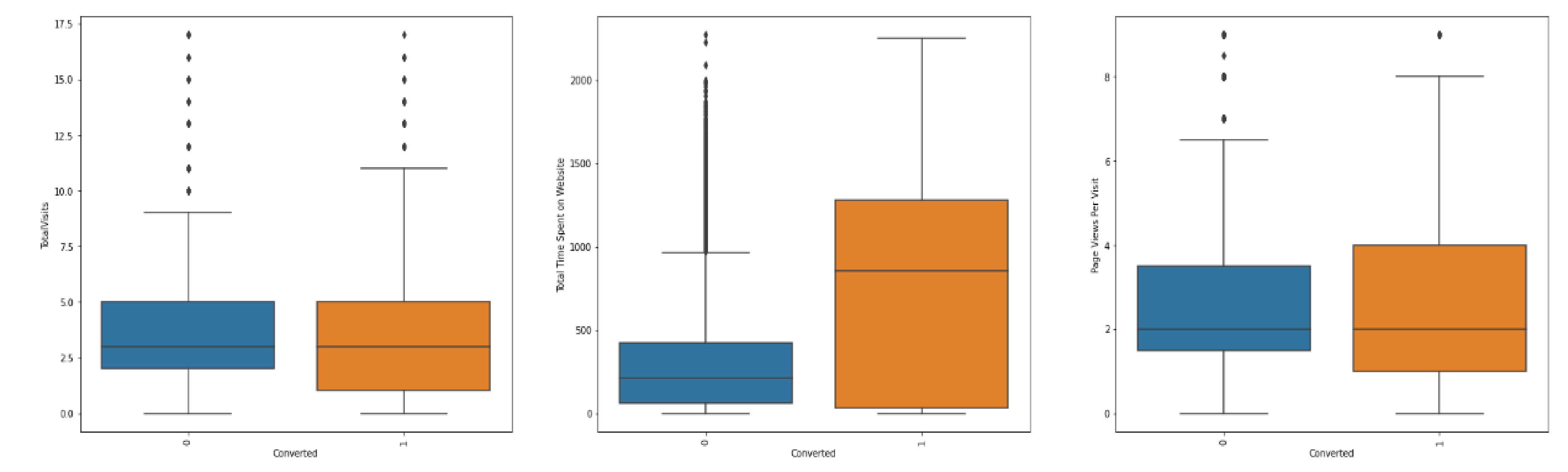
Digital Advertisement

The graph shows that Digital Advertisement is Not a good source of leads.

Inference :-

- Landed on the page submission and API will mostly get converted.
- leads are from GOOGLE, direct traffic ,Olark Chat, organic engines and through reference gets converted.
- People who have Opened email and sent SMS mostly gets converted.
- Management courses specialization and not mentioned anything mostly gets converted.
- Unemployed & working professional will get converted.
- People who wants a Better Career Prospects will choose a course.
- Most people have not mentioned their City, Mumbai and other Maharashtra cities will take a course.
- People who has done Last Notable Activity Modified, email opened and SMS sent will get converted
- who said no to Search, newspaper, digital advertisement will get converted.
- who said no to Do not email or calls will get converted.

Numerical columns with Target column



- Median of TotalVisits - converted and non converted leads is almostsame. Can't say anything on the basis of median.

- Leads **who spends more time on website will likely to** get **converted** so, company should engage their customers on website by making it more interesting.

- Median of Pages views per visit - converted and non converted leadsis almost same. Can't say anything on the basis of median

Data Preparation

- Converted into binary columns.
- Created dummy columns.
- Then, train-test split the data into 70-30% split.
- Then, Recursive feature elimination and went with 15 variables.

Model Building

- Started with model building, checked the p -values and VIF.
- Then if the p values look higher than the range 0.05 dropped that column and did this again and again until we have got p-values and VIF under control.
- So, we build 7 models in total and then we freeze the process and went ahead with the analysis on the model 7.
- Then we predicted the values and made confusion matrix, checked theaccuracy , sensitivity, specificity, positive and negative predictive values.
- Plotted ROC curve and we have got the area of 0.87.
- Find the optimal cut-offpoint - created columns with different probability cutoffs, calculated the accuracy, sensitivity, specificity on various cut-offs.

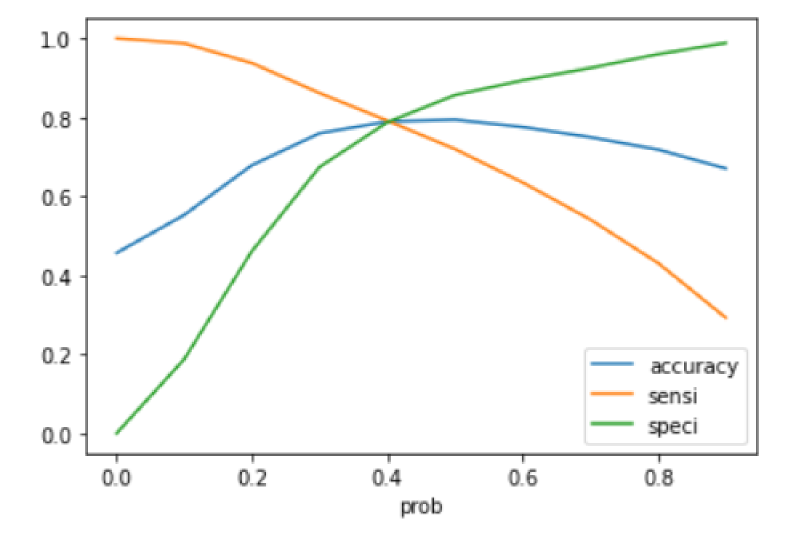
- Plotted the graph of various cut-offs.
- And, choose 0.4 the optimal cut-off point.
- Then, taking 0.4 the prob, created a lead score column.
- Again calculated the accuracy, sensitivity, specificity.
- Then we plotted precision recall curve and again checked the precision, recall values.

Predictions on Test-Set

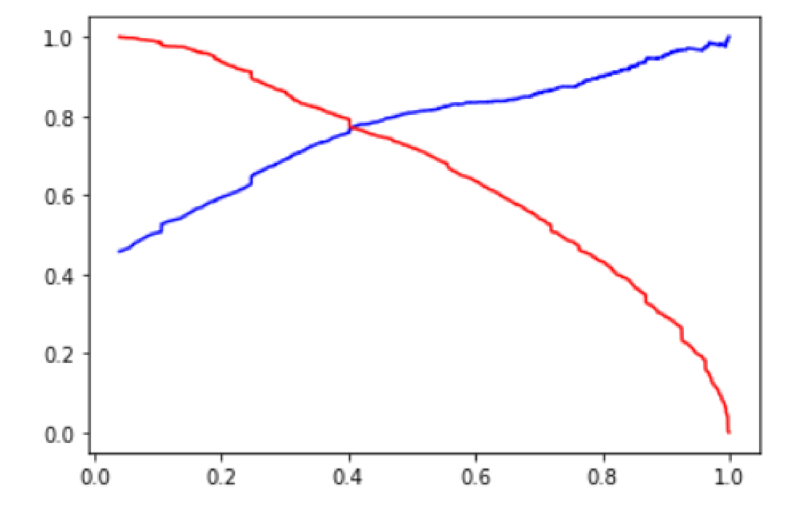
- Transform the test set, predicted the values on the basis of optimal prob, created lead-score column.

- Calculated the accuracy, sensitivity, specificity, precision and recall.

Model Evaluation – Train Data Set



- Accuracy = 78.9 %
- Sensitivity = 71.9 %
- Specificity = 85.6 %



- Precision = 80.8 %
- Recall = 71.9 %

Model Evaluation – Test Data Set

- Accuracy = 80 %
- Sensitivity = 81.5 %
- Specificity = 79.1 %

- Precision = 76 %
- Recall = 81.5 %

So, we have got a good model of accuracy around 80% and we can present this the CEO.

SUMMARY

There are a lot of leads generated in the initial stage but only a few of them come out as paying customers. In the process to convert them into hot leads, you need to polish the potential leads by educating the leads about the course,what are their requirements, what is their aim and constantly communicating them. First, churn out the best prospects from the leads you have received. 'Total Visits’, 'Total Time Spent on Website' , 'Page Views Per Visit’, ‘Last Activity’,‘Last Notable activity’ which contribute most towards the probability of a lead getting converted. Then, you should keep a list of leads in reach to inform them about new courses, services, job offers and future higher studies. Check each lead carefully so that you can tailor the information you send to them. A fine plan to check the requirements of each and every lead will go to convert them into as prospects. Focus should be converted leads. Have a video session with them, let them ask questions as much as possible to clear out all their doubts. Give them time they need to compare and check .Then at last, make further inquiries with the leads to determine their intention and mentality to join online courses.

“Thank You.. this is all from our end! ”