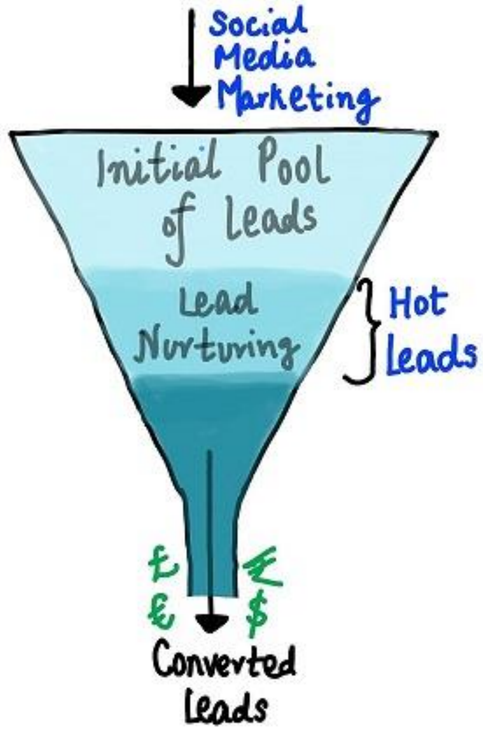
Lead Conversion Rate

X EDUCATION

BY- PARUL BANSAL

Problem

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. 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. 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 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%.**



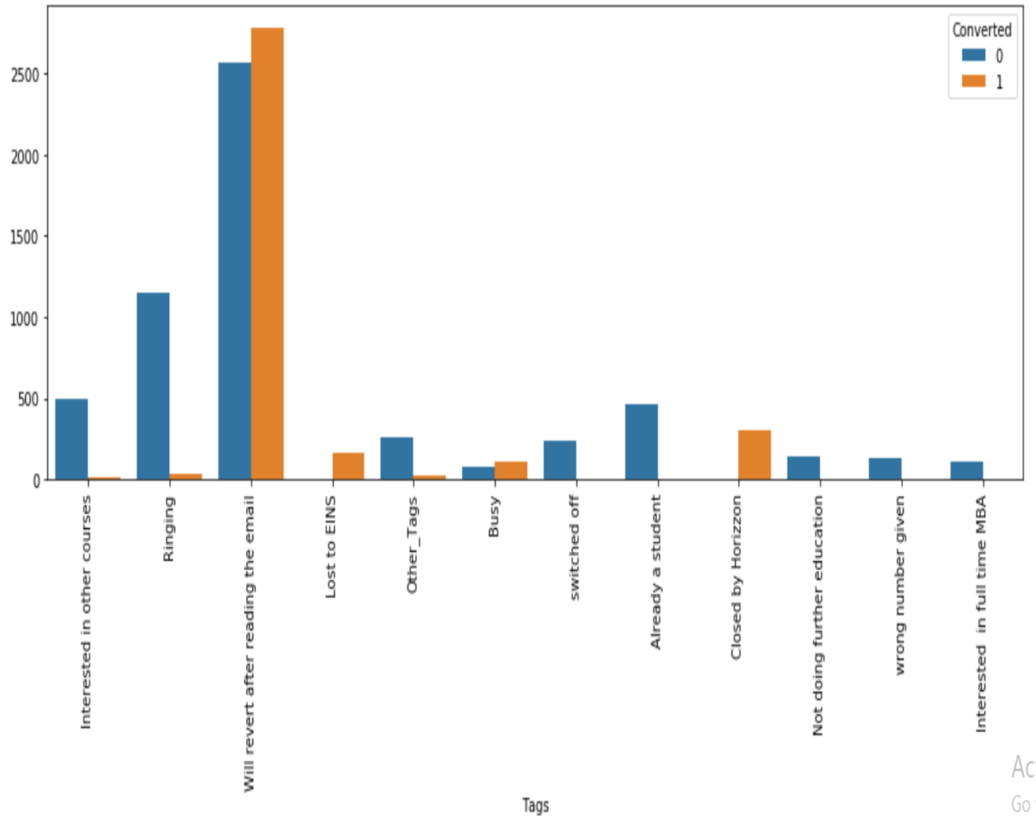
About Dataset

Leads dataset has been provided from the past with around **9000 data points.** This dataset consists of **total 37 attributes** such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity ,etc. The target variable, in this case, is the column ‘Converted’ which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn’t converted. Dataset has two primary keys – Prospect Id and the lead number. Most of the columns are categorical in nature whereas 5 of them are numerical.

Many of the categorical variables have a level called 'Select' which means that customer did not select any option from the list, so it is as good as a Null value.

Lead Conversion rate for entire dataset is around 37% hence we have a fair representation of the both the variables- converted and not converted.

Data Statistics



Inference 1. Tags that have high conversion ratio

a) Lost to EIMS and Chosen by horizon

b) Will revert after reading the mail is the most chosen tag, with conversion ratio of more than 50%.

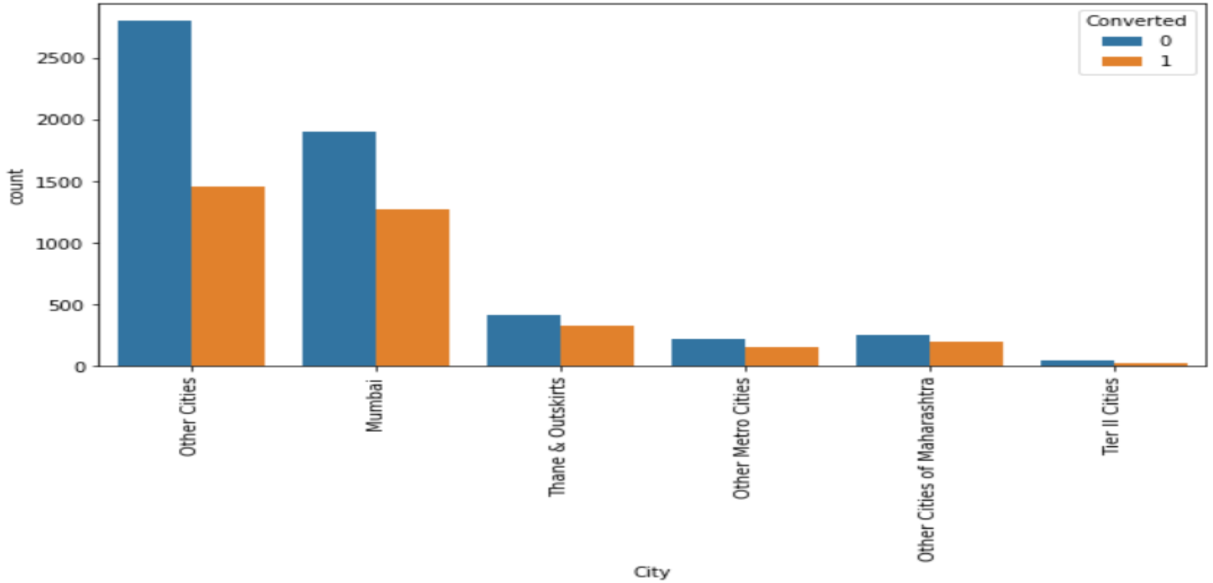
Demographics

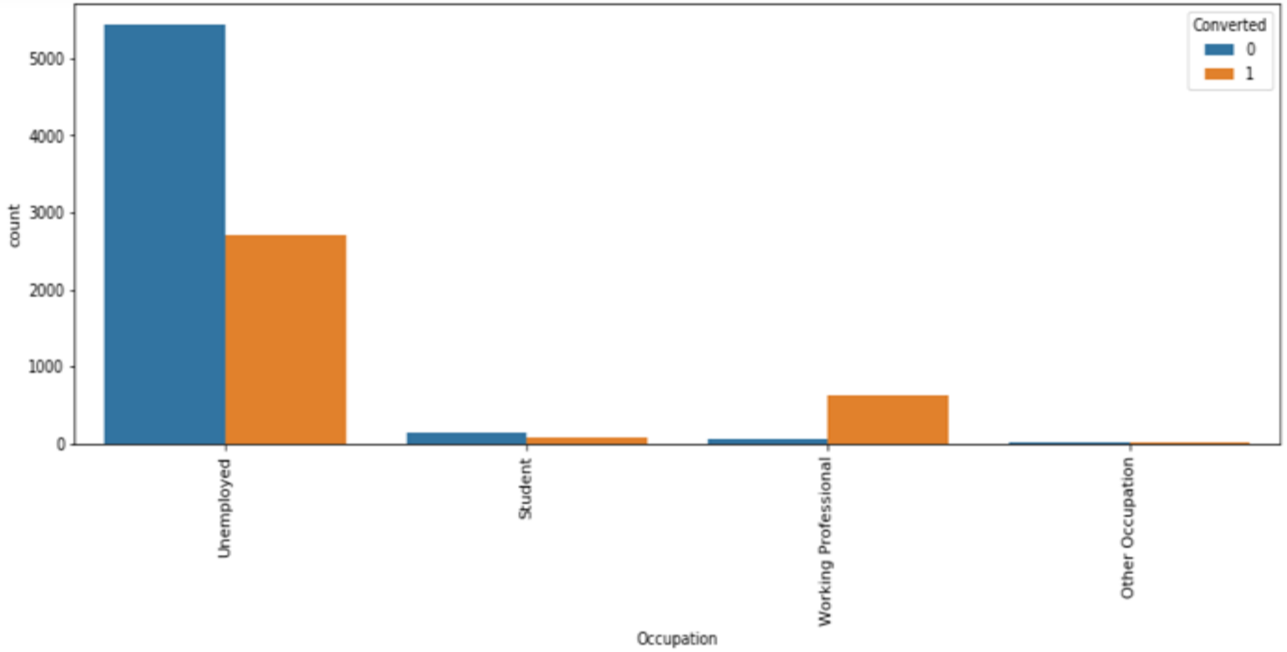
Most of the customers belong to India.

Mumbai has the highest number of customers with conversion ratio of 30%.

Working Professionals going for the course have high chances of joining it.

Unemployed leads are the most in numbers but has around 30-35% conversion rate.





Lead Origin and Source

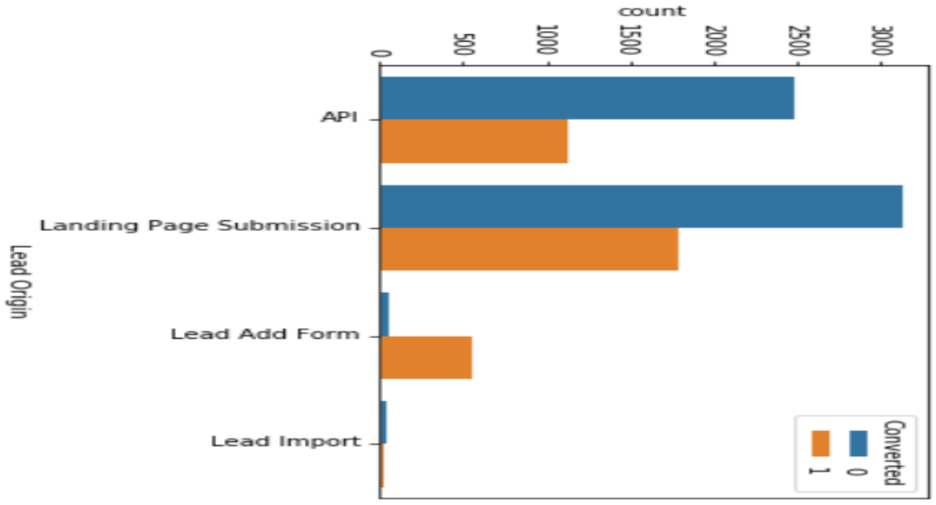
a) API and Landing Page Submission have 30-35% conversion rate but count of lead originated from them are considerable.

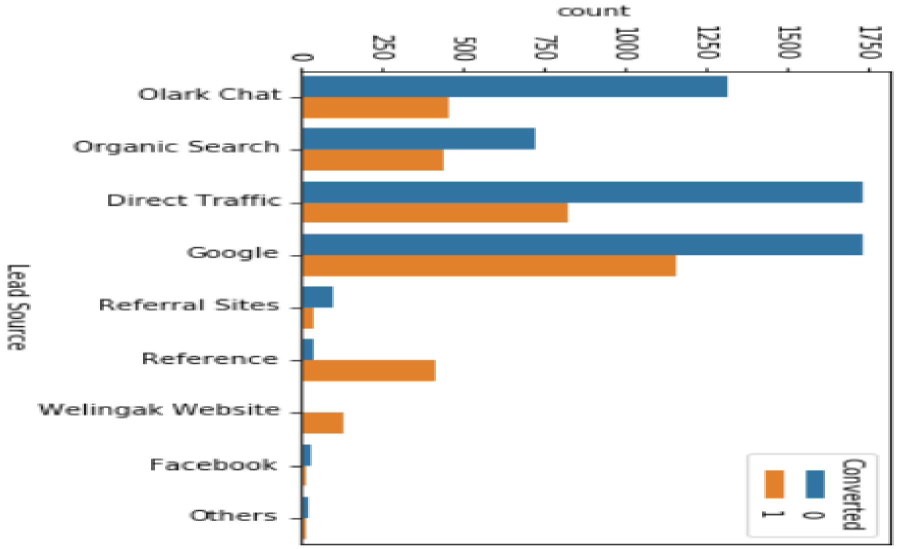
b) Lead Add Form has more than 90% conversion rate but count of lead are not very high.

c) Lead Import are very less in count.

d) Google and Direct traffic generates maximum number of leads.

e) Conversion Rate of reference leads and leads through Welingak website is high.





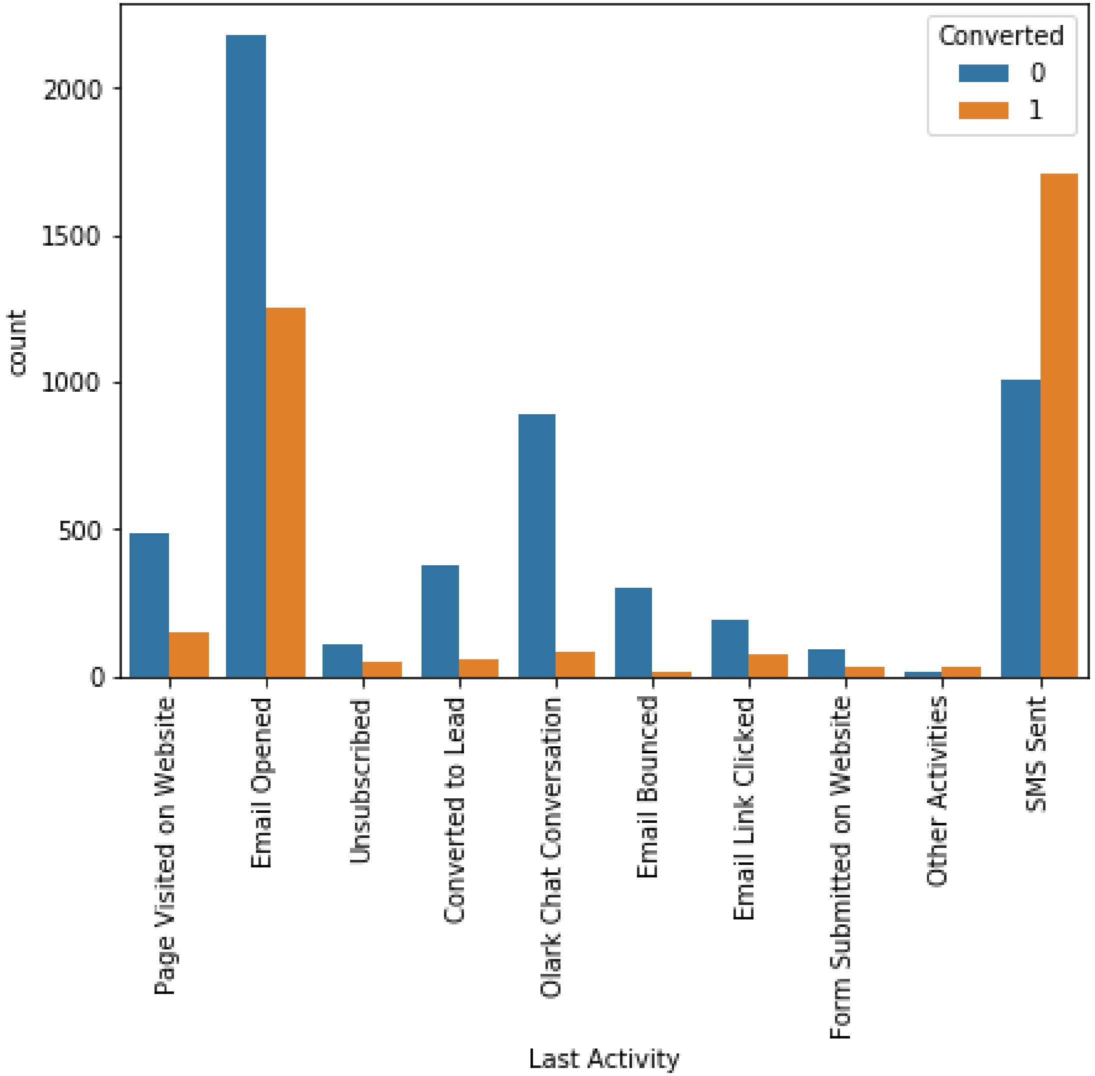
Last Activity

Last Activity and Last Notable activity are the same metircs

Most of the lead have their Email opened as their last activity.

Conversion rate for leads with last activity as SMS Sent is almost 60%

Conversion rate for other activities is very low.

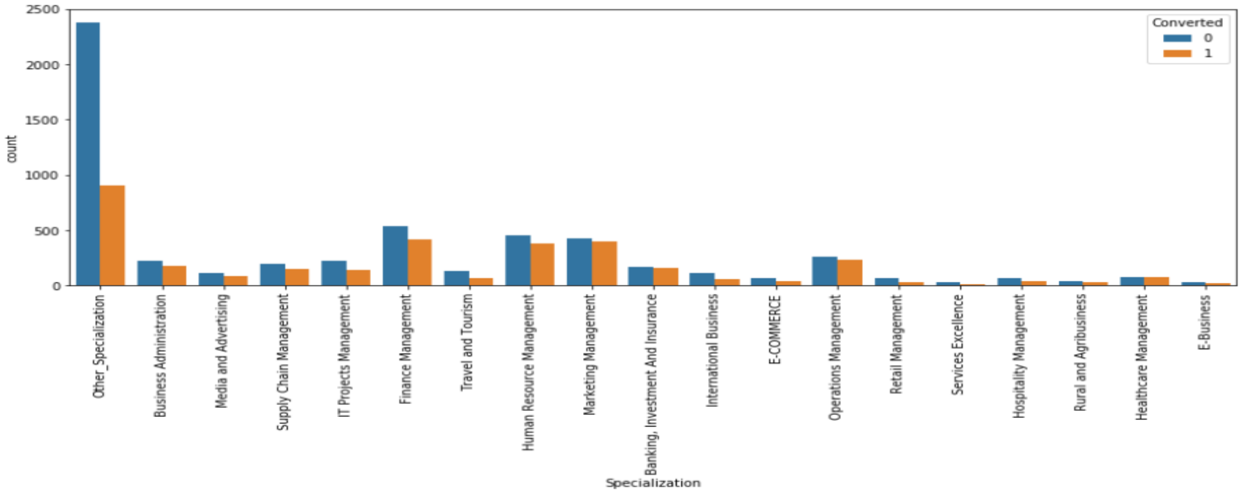


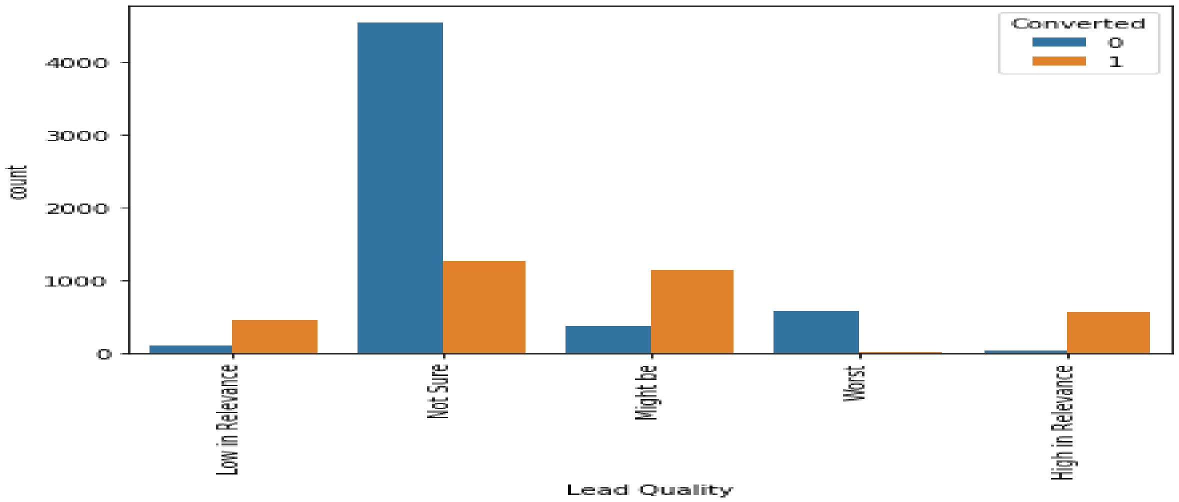
Specialization

Focus on specializations with higher conversion rate such as finance, operations and ecommerce.

- Low in Relevance, High in Relevance and Might be have good conversion rate but with less leads.

- Not Sure have very low conversion rate but having maximum leads.

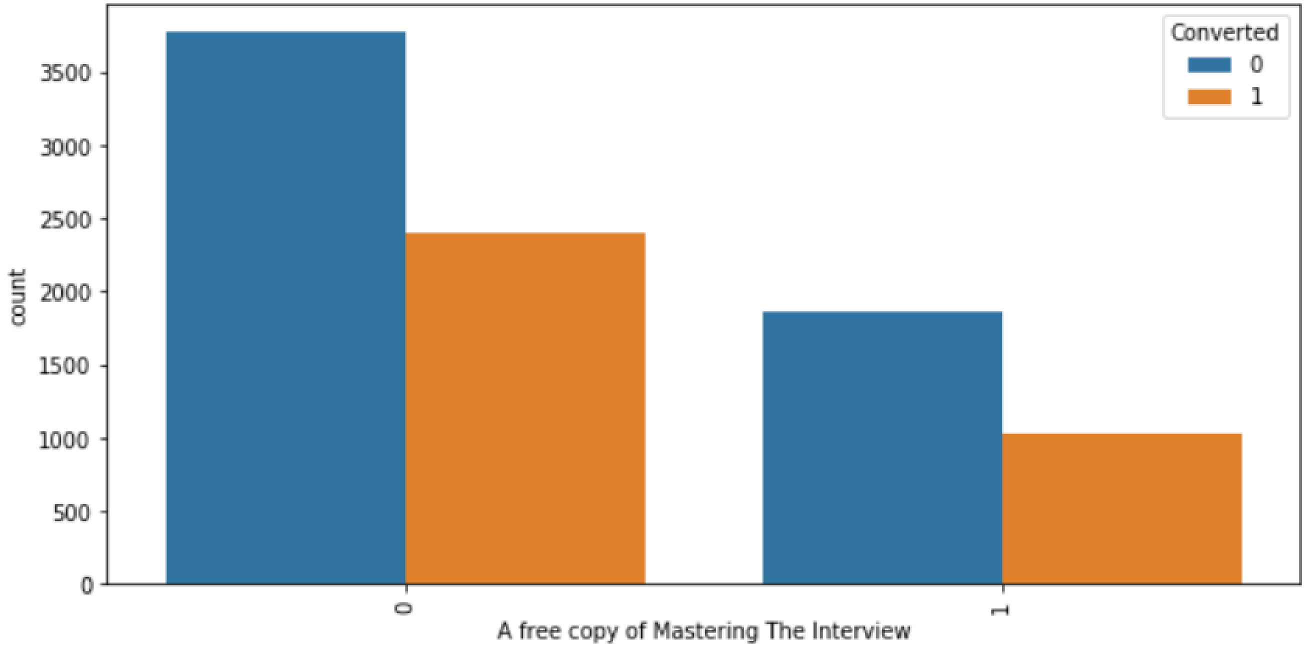




Digital Marketing

Newspaper Article, Newspaper, X Education forms, Search, Digital Marketting, Magazine, all of these have 99% columns same only 1 value.

A free copy of Mastering The Interview has good representation of both yes and no. Also the conversion rate for both the answers is almost the same.



Data Preparation

All the null values were handled accordingly.

The Columns which has only 1 value or the columns which had majority of the data as a single value were not considered for model preparation- Xeducation forms, Magazine, Search etc.

Asymmetric index data have very high percentage of missing data , hence those columns were not considered for the model preparation as well.

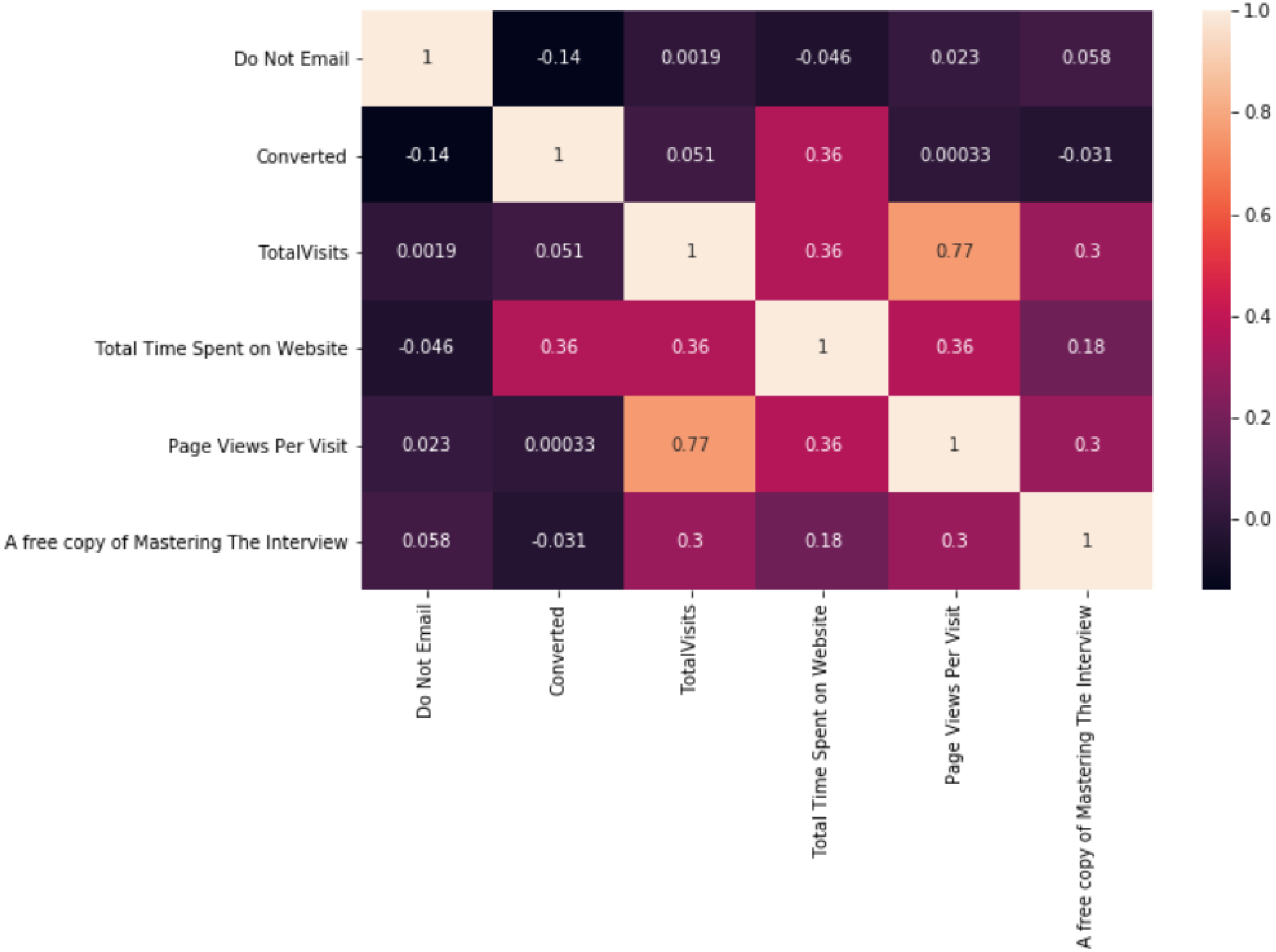
Test Train split – 80% test data, 20% training data.

After the spilt the numeric data was standardized between -1 and 1 for Total Views, Total Page Views and Total time spent on the website. This was done for the training dataset.

Yes and No was converted into the 1 and 0 respectively.

Using heat map page views was found to have high correlation with other metrics, hence was not considered for model building.

Other categorical variables were included by converting them to dummy variables.



	Do Not Email	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	A free copy of Mastering The Interview
Do Not Email	1	-0.14	0.0019	-0.046	0.023	0.058
Converted	-0.14	1	0.051	0.36	0.00033	-0.031
TotalVisits	0.0019	0.051	1	0.36	0.77	0.3
Total Time Spent on Website	-0.046	0.36	0.36	1	0.36	0.18
Page Views Per Visit	0.023	0.00033	0.77	0.36	1	0.3
A free copy of Mastering The Interview	0.058	-0.031	0.3	0.18	0.3	1

Model Building

Initially model was build using all the features. There were total 69 variables .

Using RFE top 20 features were selected. They are :- 'Do Not Email', 'Total Time Spent on Website', 'Lead Origin_Lead Add Form', 'Lead Source_Olark Chat', 'Lead Source_Welingak Website', 'Last Activity_Olark Chat Conversation', 'Last Activity_SMS Sent', 'Last Activity_Unsubscribed‘, 'Occupation_Student', 'Occupation_Unemployed', 'Tags_Busy','Tags_Closed by Horizzon', 'Tags_Lost to EINS','Tags_Not doing further education', 'Tags_Ringing', 'Tags_Will revert after reading the email', 'Tags_switched off', 'Tags_wrong number given', 'Lead Quality_Not Sure', 'Lead Quality_Worst‘.

Then the model was trained multiple times. At each step we check the p value, and continue till p value is less and 0.005 for all the variables.

Finally we check the vif and train the model again, till the vif is less than 3 for all the variables.

Final Model

Accuracy – 87.5%

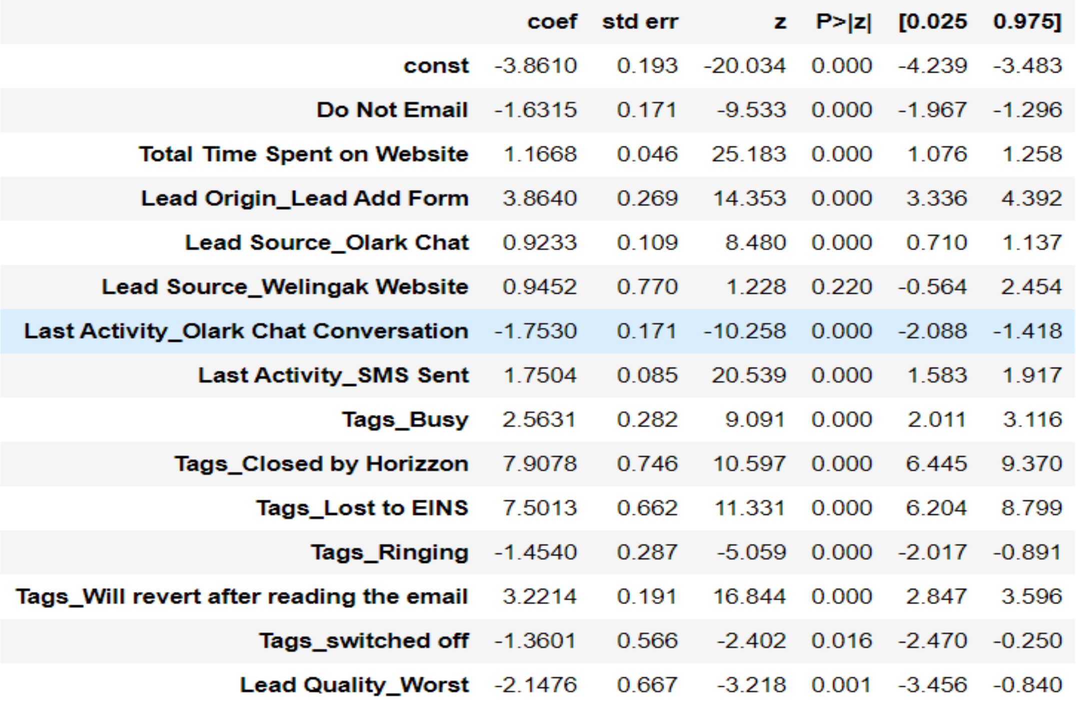
Sensitivity- 83%

Specificity-90%

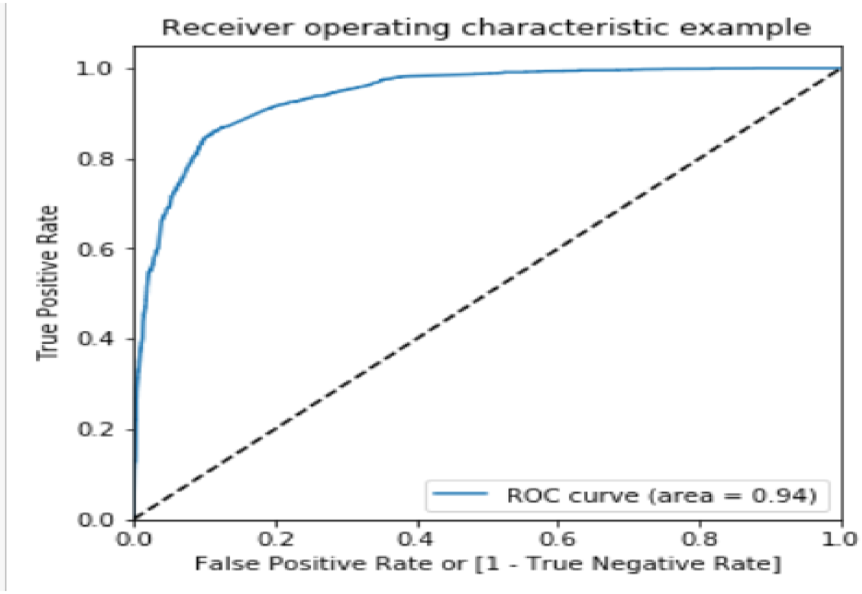
False Positive Rate – 9%

Positive prediction rate -84.3%

Negative prediction rate- 89.5%



	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.8610	0.193	-20.034	0.000	-4.239	-3.483
Do Not Email	-1.6315	0.171	-9.533	0.000	-1.967	-1.296
Total Time Spent on Website	1.1668	0.046	25.183	0.000	1.076	1.258
Lead Origin_Lead Add Form	3.8640	0.269	14.353	0.000	3.336	4.392
Lead Source_Olark Chat	0.9233	0.109	8.480	0.000	0.710	1.137
Lead Source_Welingak Website	0.9452	0.770	1.228	0.220	-0.564	2.454
Last Activity_Olark Chat Conversation	-1.7530	0.171	-10.258	0.000	-2.088	-1.418
Last Activity_SMS Sent	1.7504	0.085	20.539	0.000	1.583	1.917
Tags_Busy	2.5631	0.282	9.091	0.000	2.011	3.116
Tags_Closed by Horizzon	7.9078	0.746	10.597	0.000	6.445	9.370
Tags_Lost to EINS	7.5013	0.662	11.331	0.000	6.204	8.799
Tags_Ringing	-1.4540	0.287	-5.059	0.000	-2.017	-0.891
Tags_Will revert after reading the email	3.2214	0.191	16.844	0.000	2.847	3.596
Tags_switched off	-1.3601	0.566	-2.402	0.016	-2.470	-0.250
Lead Quality_Worst	-2.1476	0.667	-3.218	0.001	-3.456	-0.840



Final Model

The cutoff for conversion rate was found to be 0.35 . That is lead havig score greater than 35 is potential candidate for conversion.

Accuracy – 87.5%

Sensitivity- 86.4%

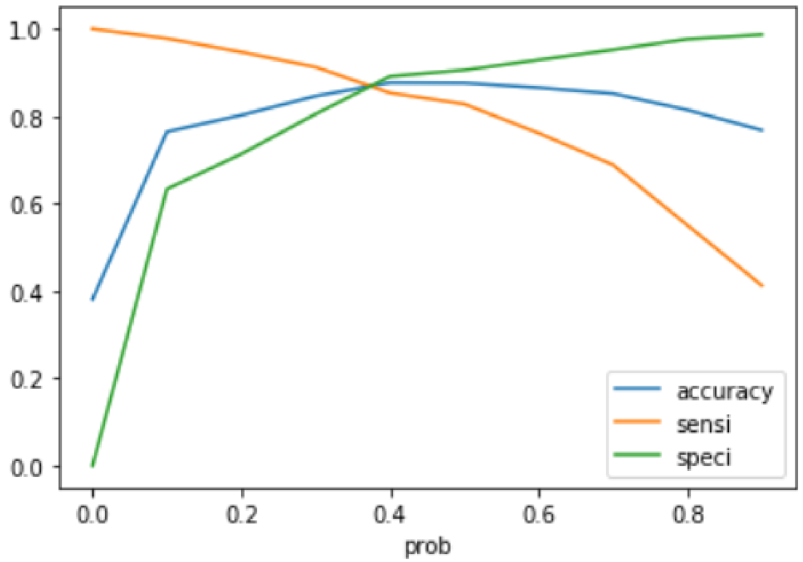
Specificity-88.2%

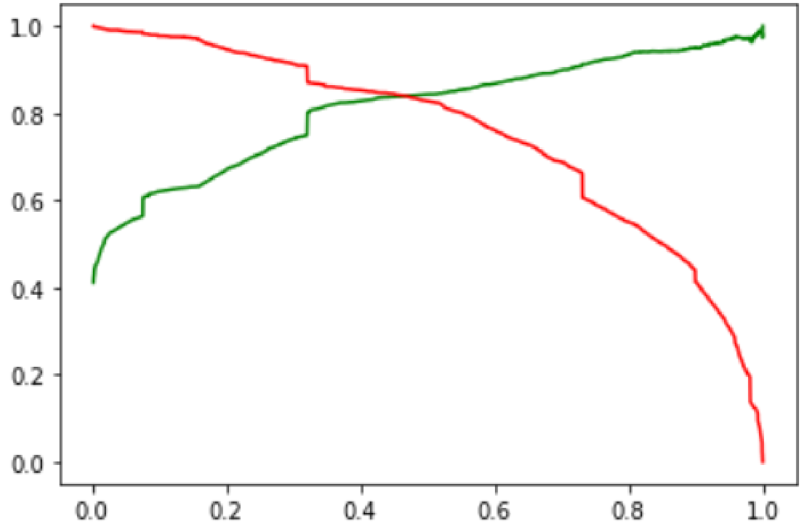
False Positive Rate – 11%

Positive prediction rate -81.3%

Negative prediction rate- 91.5%

Trade of between precision and recall – 45%





Results on test data set

Accuracy – 87.5%

Sensitivity- 86.7%

Specificity-87.5%

Results are similar for the test and training dataset . Hence our model works fine

By comparing test set and train set we can say that results are close and we are getting very good accuracy around 90% which shows that model is predicting good lead scores.

The CEO, in particular has given a ballpark of the target lead conversion rate to be around 80% and our precision rate is around 80% hence this condition is satisfied.

Initially conversion rate around 38% but with the help of the model built conversion rate is improved and it’s around 80% which can make a good and profitable difference between company's past records and future records.

Business Interpretation

Company can deploy the model built and can predict the values of ‘Hot Leads’ which will help the sales team to focus more on communicating with the potential leads rather than making calls to everyone which will help to optimize the cost and time for the company.

Lead Source: WElingkak WEbsite, Although the number of Users are less, but there is almost 100 percent conversion Thus if it is focused more then a very good Conversion rate can be achieved. The strategy should be to promote this source

Lead Origin: Add_Form : This origin has 93% conversion rate, this should not be neglected at all and in fact if the origin is Add_Form then more priority should be given to the user as it has higher chances to convert to HOT leads

Lead Notable Activity: Olark Chat : this reveals that there are a very large number of users who are using Olark chat, and the conversion rate here is not so high, so keeping in mind the number of users, and their interest in online conversation, focus should be given to look for more potential leads, so that we don’t miss a large number of enquiring users.