LEAD SCORE CASE STUDY

- CHANDAN MOHANTY
- SONAL SHARMA
- SAHIL SAMAL

CONTENTS

- Problem Statement
- Objective
- Approach
- Data Insight
- Factors Responsible in Driving Leads
- Model Metrics
- Conclusion

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. 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.

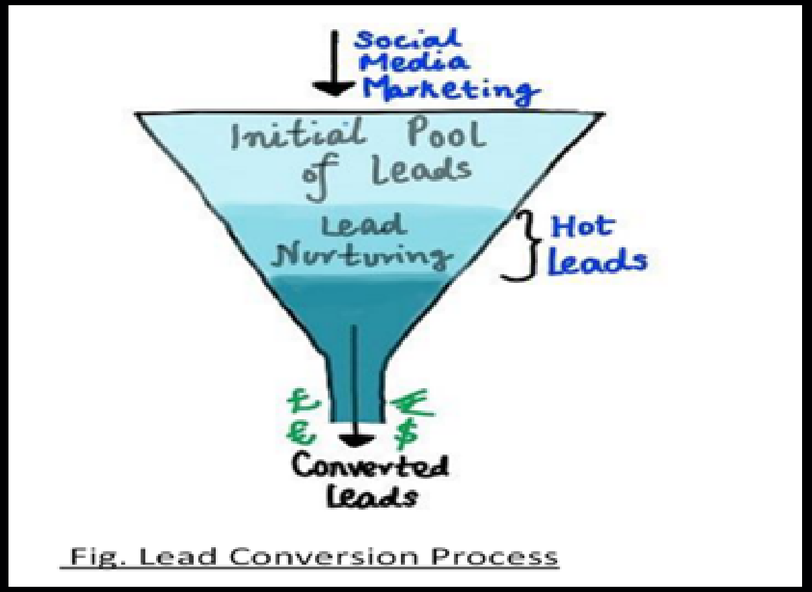


Fig. Lead Conversion Process

OBJECTIVE

- To help the company in selecting the most potential leads, also known as ‘Hot Leads’ whose lead conversion rate is around 80%.
- To build a model wherein a lead score is assigned 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.
- Help the sales team to divert their focus on potential leads & avoid them from making useless phone calls.

APPROACH

- Analysing Patterns
- Driving Factors
- Correlations
- Recommendations

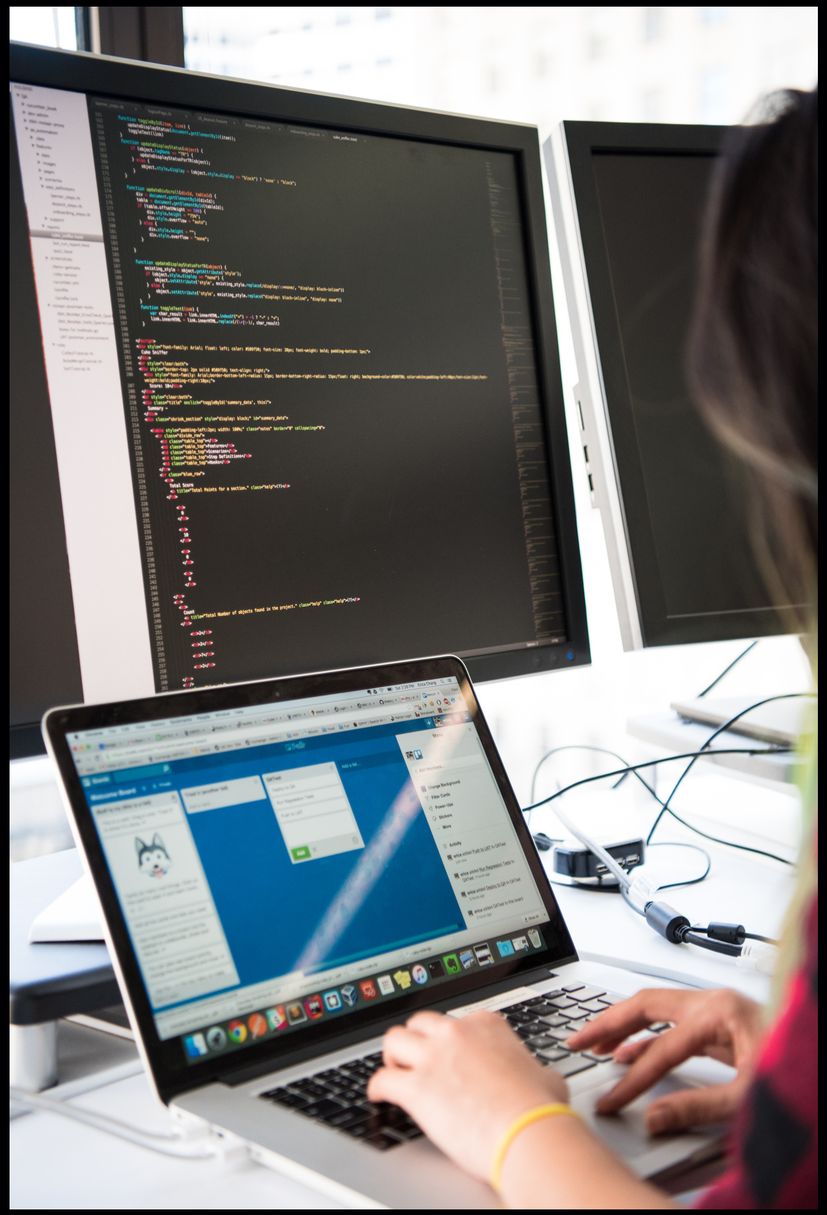
ANALYSING PATTERNS

Using Exploratory Data Analysis, we have analysed the patterns present in the Dataset which will provide us intuition that the which features will help in driving the lead conversion.

DRIVING FACTORS

Looking at the below data we get an intuition that how the variables are distributed.

	Lead Number	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Asymmetrique Activity Score	Asymmetrique Profile Score
count	9240.000000	9240.000000	9103.000000	9240.000000	9103.000000	5022.000000	5022.000000
mean	617188.435606	0.385390	3.445238	487.698268	2.362820	14.306252	16.344883
std	23405.995698	0.486714	4.854853	548.021466	2.161418	1.386694	1.811395
min	579533.000000	0.000000	0.000000	0.000000	0.000000	7.000000	11.000000
25%	596484.500000	0.000000	1.000000	12.000000	1.000000	14.000000	15.000000
50%	615479.000000	0.000000	3.000000	248.000000	2.000000	14.000000	16.000000
75%	637387.250000	1.000000	5.000000	936.000000	3.000000	15.000000	18.000000
max	660737.000000	1.000000	251.000000	2272.000000	55.000000	18.000000	20.000000



CORRELATIONS

Identifying correlations amongst variables to identify the variability in data and identify most important features that can help in driving the conversion of leads

RECOMMENDATIONS

Focus on features that can expedite the conversion of leads

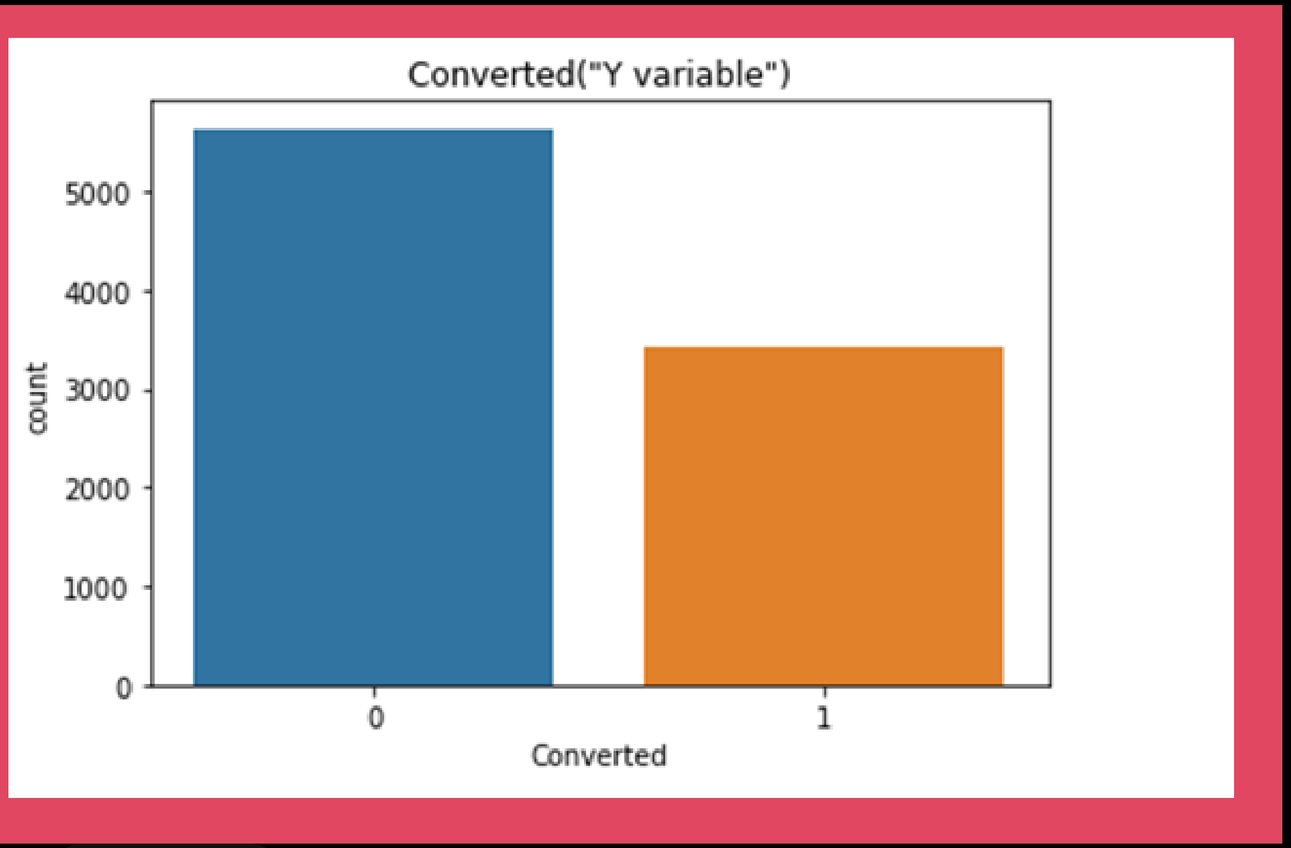
DATA INSIGHTS

We have total 9240 entries of unique customers and we needs to identify out of these which have the highest probability of getting converted.

DECISION CRITERIA

*Potential Leads can be bifurcated on the basis of Leads Score (which is probability of getting converted)

- ·Out of 9240 entries we see that around 37.8% of leads are converted and 73% of leads are not converted.



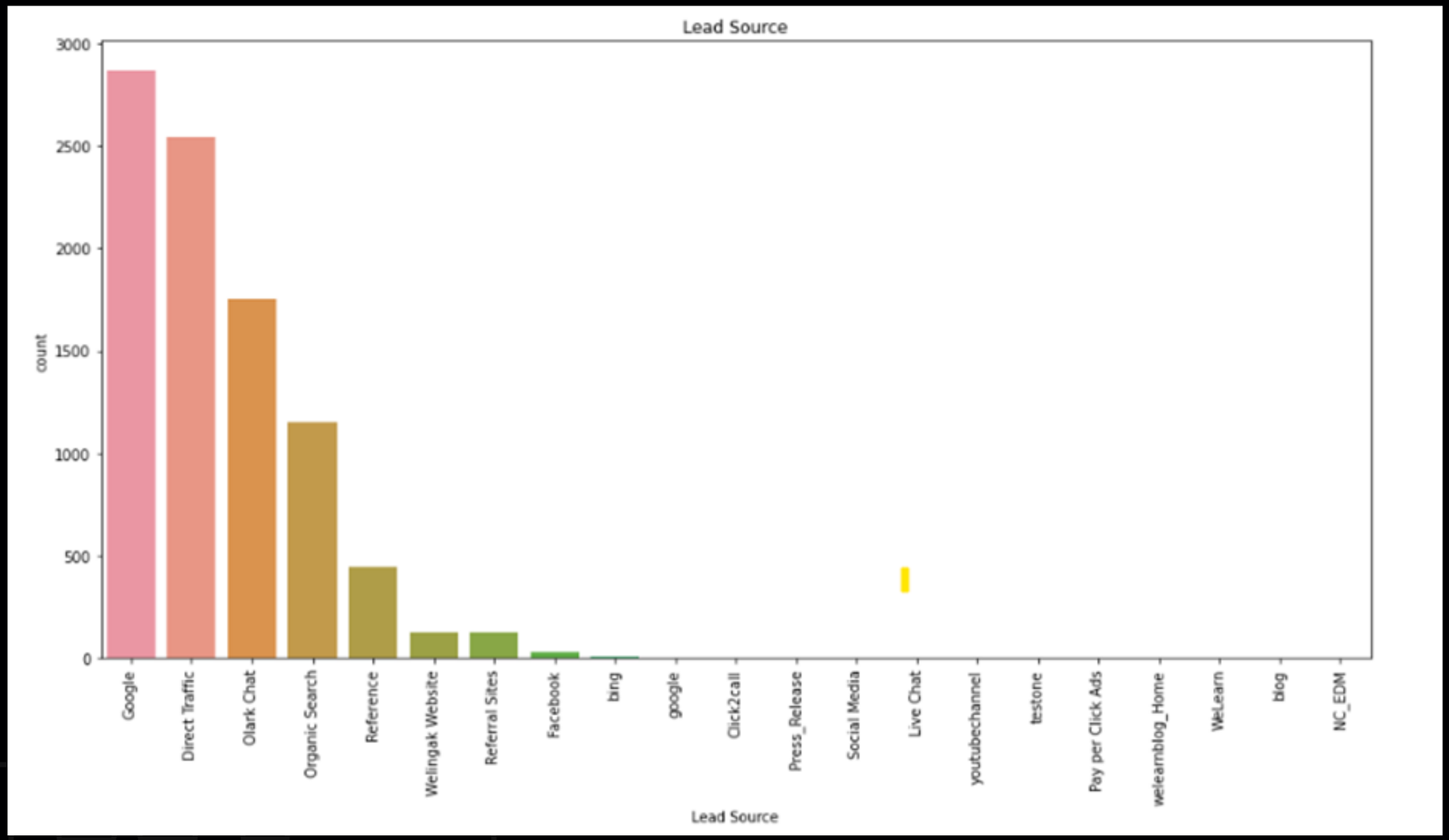
37.8% of the 'Converted' data is 1 ie. 37.8% of the leads are converted. This means we have enough data of converted leads for modelling.

Task: Identify solution so that the lead conversion rate could be increased.

Lets see the spread of Categorical Columns w.r.t Converted Columns.

LEAD SOURCE

- Majority source of the lead is Google & Direct Traffic.
- Lead source from Google has highest probability of conversion.
- Leads with source Reference has maximum probability of conversion.



- As it can be seen from the graph, number of leads generated by many of the sources are negligible. There are sufficient numbers till Facebook. We can convert all others in one single category of 'Others'.

- 'Direct Traffic' and 'Google' generate maximum number of leads while maximum conversion rate is achieved through 'Reference' and 'Welingak Website'.

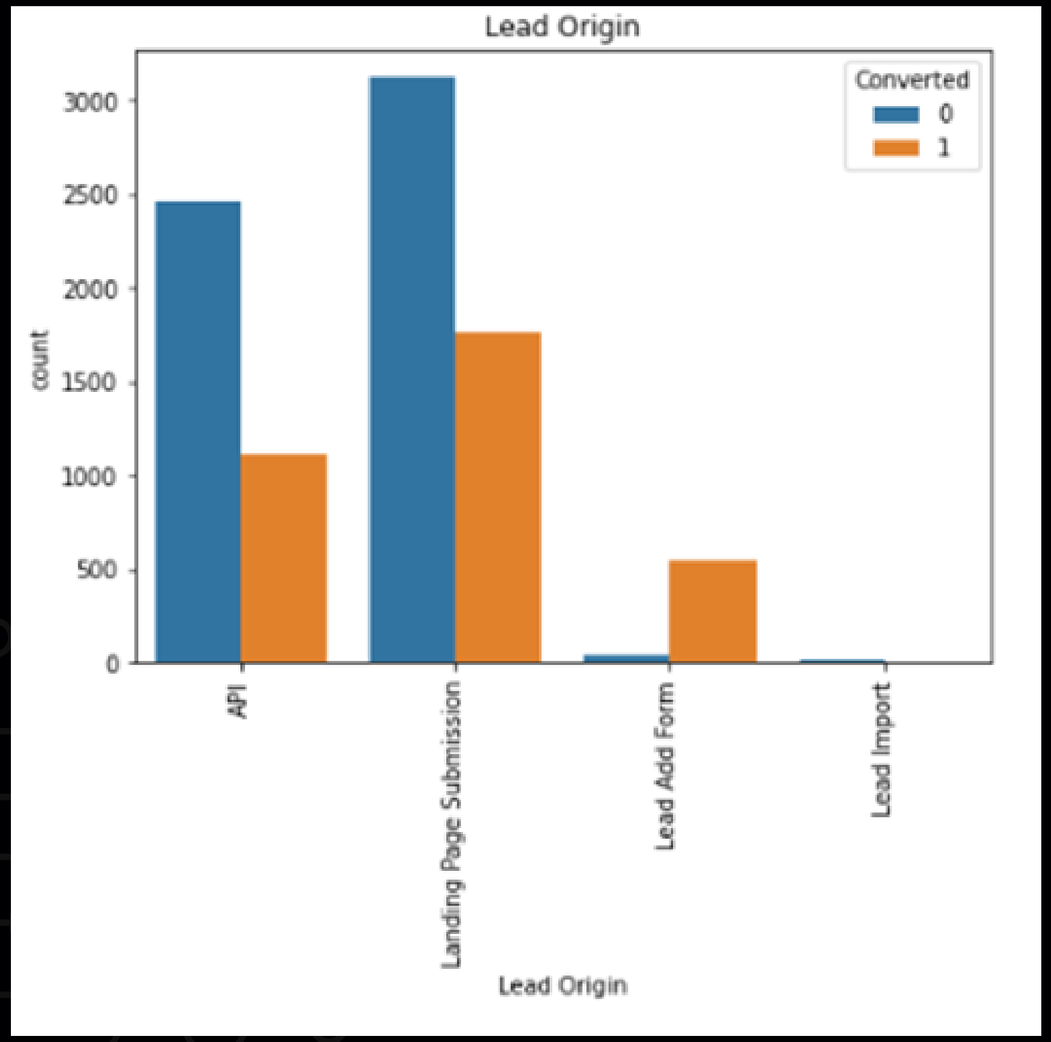
LEAD ORIGIN

Customers who were identified as Leads from Landing Page submission, constitute most of the leads.

Customers originating from Lead Add Form have high probability of conversion. These Customers are very few.

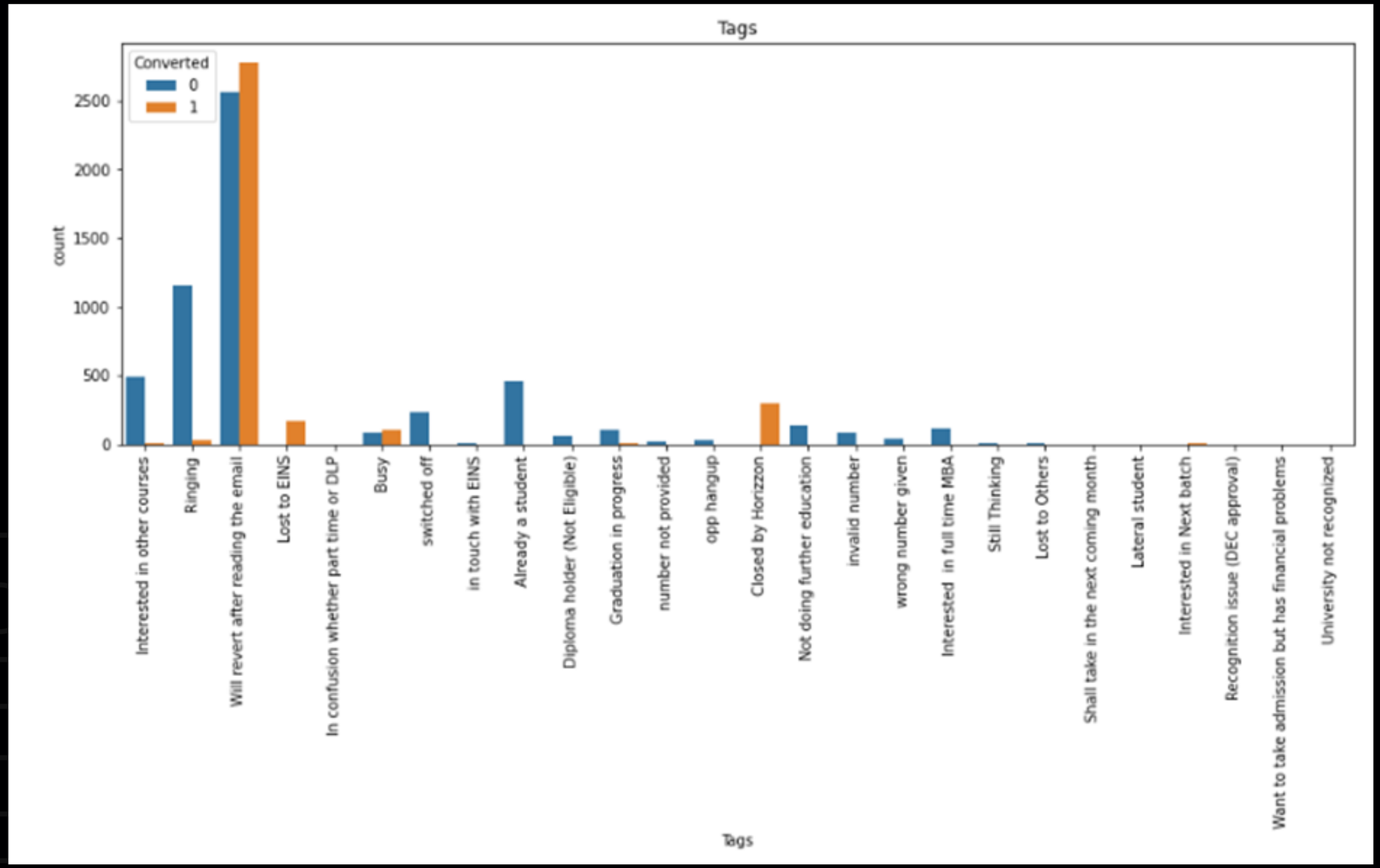
Lead origin-API & Lead Import have the least conversion rate. Customers from Lead Import are very few.

OBSERVATION FROM LEAD ORIGIN



'API' and 'Landing Page Submission' generate the most leads but have less conversion rates of around 30%. Whereas, 'Lead Add Form' generates less leads but conversion rate is great. We should try to increase conversion rate for 'API' and 'Landing Page Submission', and increase leads generation using 'Lead Add Form'. 'Lead Import' does not seem very significant.

TAGS

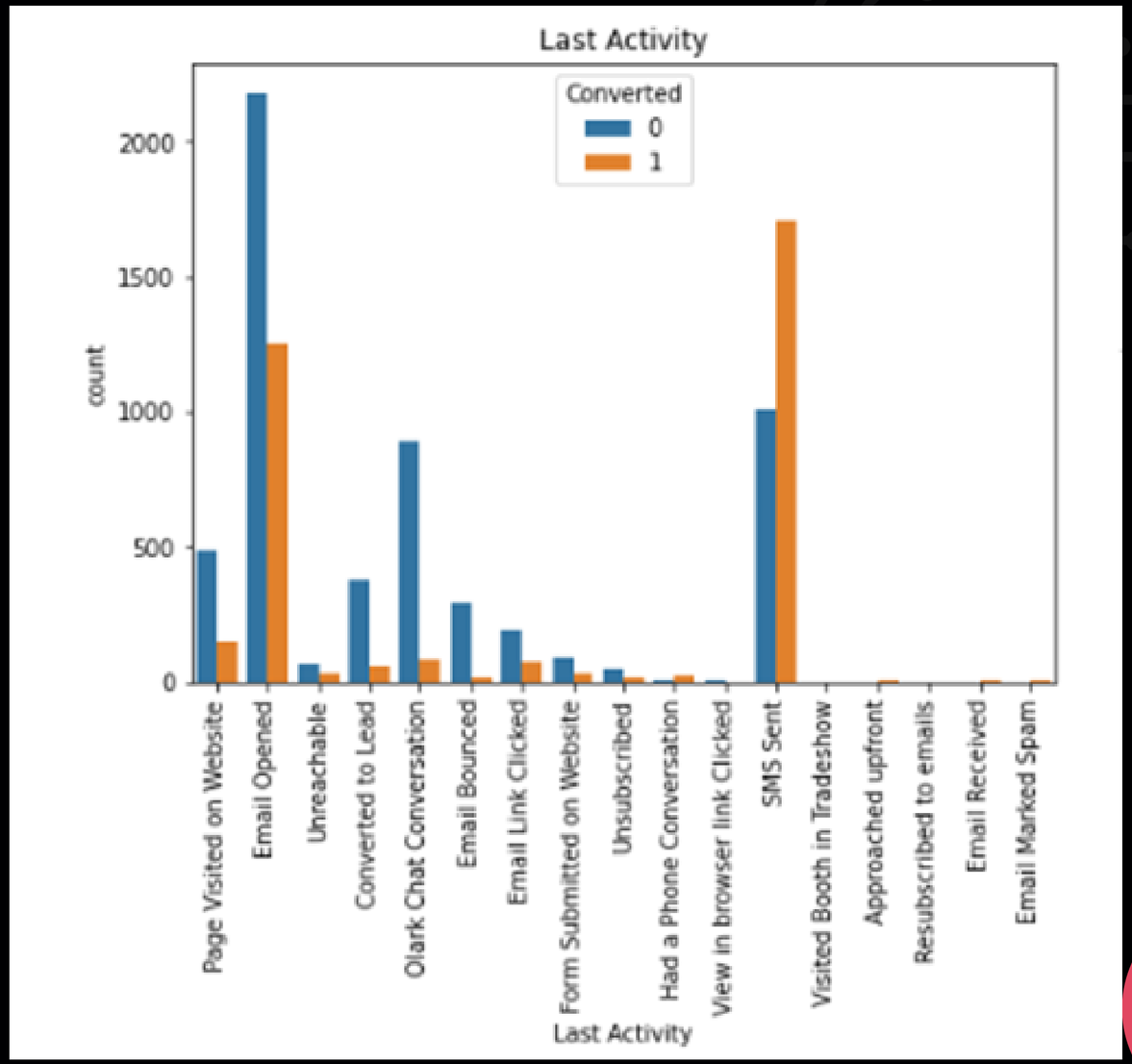


More focus shall be given on the leads as will revert after reading the mail they have higher rate of conversion.

LAST ACTIVITY

Customers whose last activity was SMS Sent have higher conversion rate which is around 63%.

Customers who last activity was Email Opened constitute majority of the customers. They have around 36% of conversion rate.

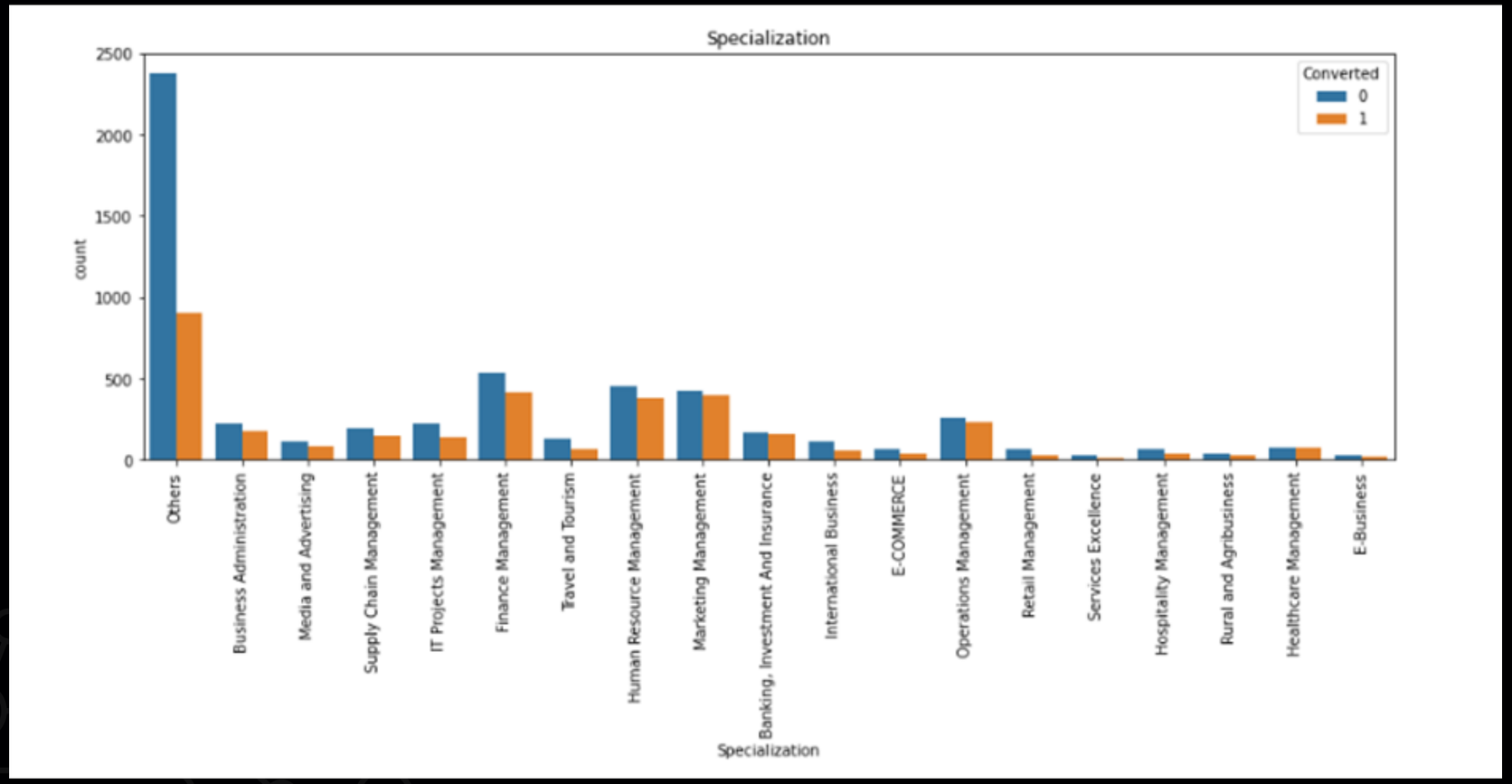


OBSERVATION FOR LAST ACTIVITY

- Highest number of lead are generated where the last activity is 'Email Opened' while maximum conversion rate is for the activity of 'SMS Sent'. Its conversion rate is significantly high.

- Categories after the 'SMS Sent' have almost negligible effect. We can aggregate them all in one single category.

SPECIALIZATION



Observations for Specialization :

Conversion rates are mostly similar across different specializations.

HEATMAP OF CORRELATION

Let us observe the correlation among the numerical columns.

We can observe that the variables are not highly correlated with each other. But still there is multicollinearity among some features

OBSERVATION

EDA

• An Exploratory Data Analysis was done to analyse the data using visual techniques and how they react to target field (Converted). It was used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations.

- It was found that a lot of attributes in the categorical variables were irrelevant.
- The numeric values had outliers which were handled.

After validating Bar plots for all the categorical values:

- lead Source: google and Google can be merged
- Country can be segregated as India, NA and US and Others, there are many similar fields
- Field with one Unique value can be dropped as no one is interested for these offers and this might make the model skewed

Observations:

- Most converted leads are originated from Lead add and landing page submission
- Most Conversion has happened from Google, Direct traffic and reference
- Leads don’t want calls or mails
- Most converted are unemployed and looking for better career prospects
- Most leads didn't see any adds in search, magazine, Education forum, newspaper or digital forum

FACTORS RESPONSIBLE IN DRIVING LEADS

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6468
Model:	GLM	**Df Residuals:**	6455
Model Family:	Binomial	**Df Model:**	12
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2172.9
Date:	Mon, 14 Nov 2022	**Deviance:**	4345.8
Time:	22:06:39	**Pearson chi2:**	6.99e+03
No. Iterations:	23		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Total Time Spent on Website	2.6140	0.145	18.090	0.000	2.331	2.897
Lead Source_Welingak Website	4.1041	1.013	4.050	0.000	2.118	6.090
Last Activity_SMS Sent	1.3714	0.080	17.179	0.000	1.215	1.528
Tags_Already a student	-5.8411	1.005	-5.811	0.000	-7.811	-3.871
Tags_Closed by Horizzon	6.3636	1.007	6.321	0.000	4.390	8.337
Tags_Interested in full time MBA	-3.5460	0.732	-4.843	0.000	-4.981	-2.111
Tags_Interested in other courses	-3.4159	0.370	-9.226	0.000	-4.142	-2.690
Tags_Lost to EINS	4.1605	0.522	7.976	0.000	3.138	5.183
Tags_Not doing further education	-24.1964	1.15e+04	-0.002	0.998	-2.26e+04	2.26e+04
Tags_Other	-3.3338	0.262	-12.740	0.000	-3.847	-2.821
Tags_Ringing	-4.5834	0.214	-21.457	0.000	-5.002	-4.165
Tags_switched off	-4.9952	0.517	-9.654	0.000	-6.009	-3.981
Last Notable Activity_Modified	-2.0326	0.079	-25.609	0.000	-2.188	-1.877

Below features are most important ones which are responsible for leads conversion:

- Total Time Spent on Website
- Lead Source_Welingak Website
- Last Activity_SMS Sent
- Tags_Already a student
- Tags_Closed by Horizzon
- Tags_Interested in full time MBA
- Tags_Lost to EINS
- Tags_Not doing further education
- Tags_Other
- Tags_Ringing
- Tags_switched off
- Last Notable Activity_Modified

MODEL METRICS

i) Train-Test split of Data:
The split was done at 70% and 30% for train and test data respectively.

ii) Model Building:
• Firstly, RFE was done to attain the top 15 relevant variables.
• Secondly VIF values and p-value were used to remove few more variable which were creating influence (The variables with VIF < 5 and p-value < 0.05 were kept).

iii) Model Evaluation:
• Confusion matrix and ROC was created for train Data, then calculation was made to find out the accuracy, sensitivity and specificity which came to be around

Train set:
Accuracy: 0.8630179344465059
Sensitivity: 0.7821939586645469
Specificity: 0.9144736842105263

Test set:
Accuracy: 0.8647186147186147
Sensitivity: 0.7923444976076555
Specificity: 0.9085118702953098

iv) Prediction:

Prediction was done on the test data frame and with an optimum cut off as 0.6 with accuracy ~86%.

CONCLUSION

From EDA the attributes which is of great interest

- Most converted leads are originated from Lead add and landing page submission
- Most Conversion has happened from Google, Direct traffic and reference
- Leads don’t want calls or mails
- Most converted are unemployed and looking for better career prospects
- Most leads didn't see any adds in search, magazine , Education forum, newspaper or digital forum

It was found that the attributes that mattered the most in prediction are as follows

- Total Time Spent on Website
- Last Notable Activity Modified
- Last Activity_SMS Sent
- Interested in other courses
- Closed by Horizzon
- Ringing
- Already a student
- Lost to EINS
- Welingak Website
- Not doing further education
- Interested in full time MBA

Keeping these in mind the X Education can flourish as they have a very high chance to get
almost all the potential buyers to change their mind and buy their courses.


THANK YOU