CASE STUDY - LEAD SCORING

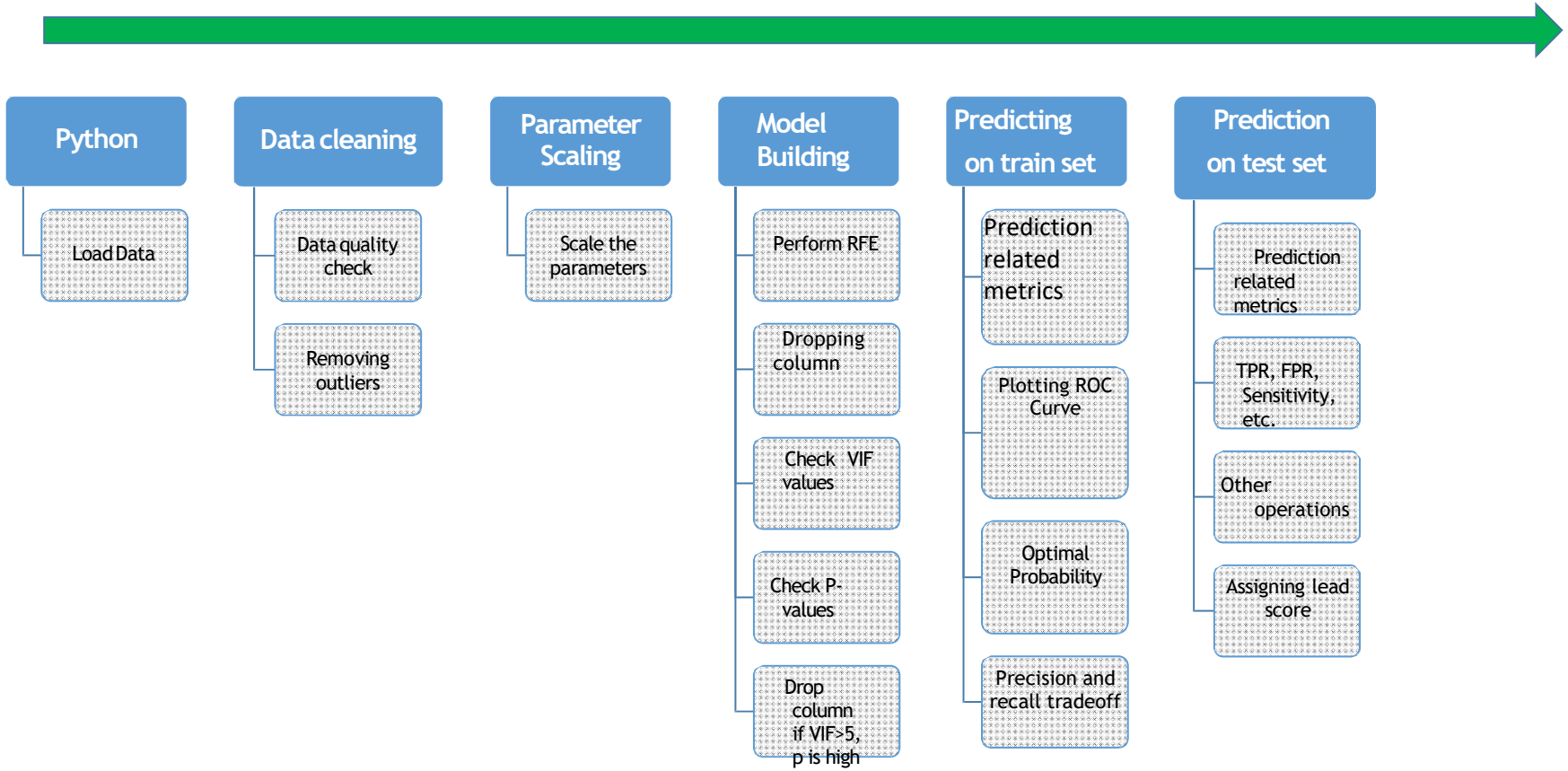
Member name: Vikul Aggarwal

Member name: Ankit Sahu

Problem Statement

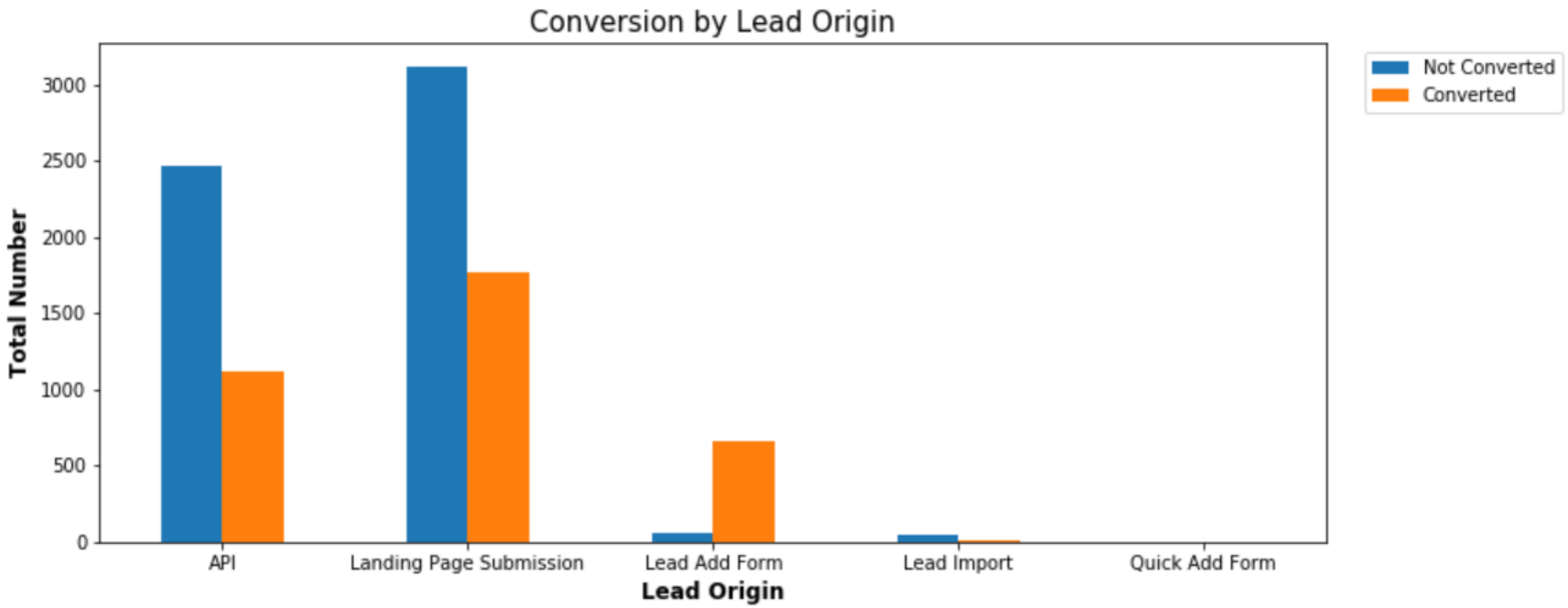
- **Problem**: Help the company ‘X Education’ to select the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- **Aim:** We need to build a model wherein we have to assign a lead score to each of the leads such that customers with high lead score have a higher conversion chance and the customers with a low lead score have a low chance of conversion.
- **Data:** Leads dataset from the past with around 9000 data points. This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc.
- **Target Variable** :‘Converted’ which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn’t converted.

Problem solving methodology



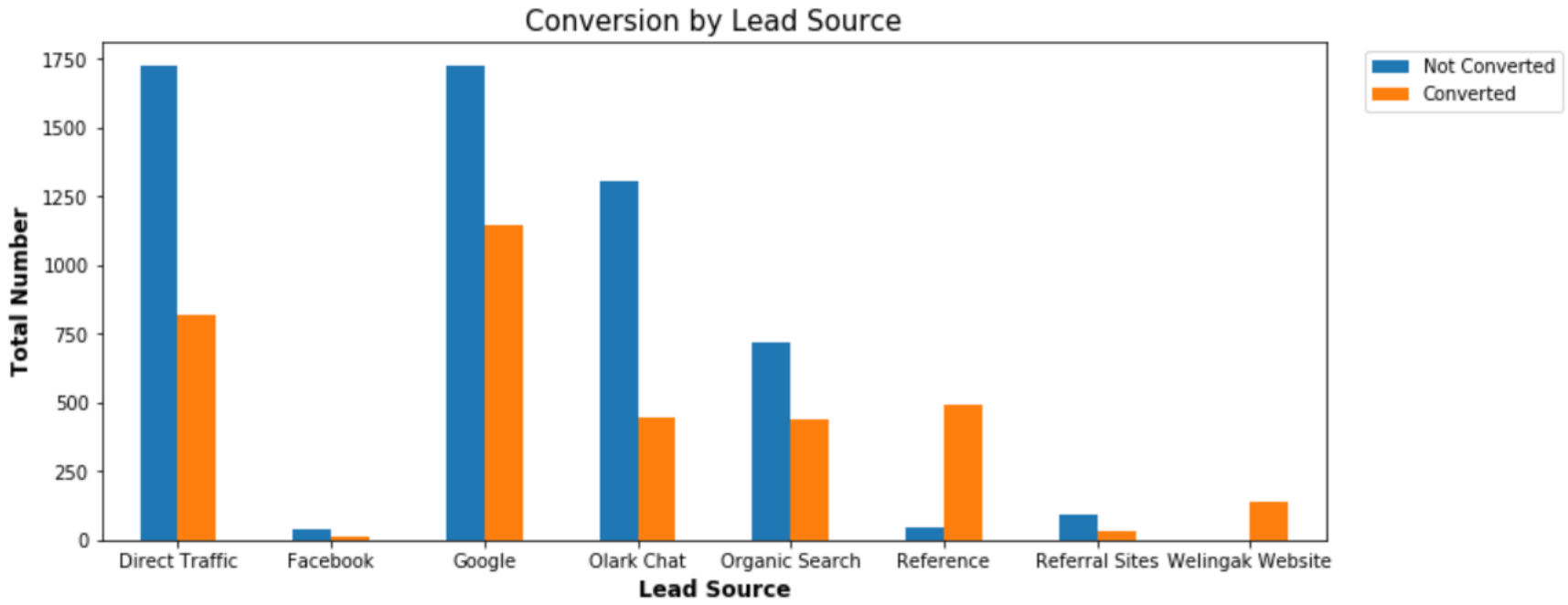
Conversion by Lead Origin

API and Landing Page Submission have low conversion ratio, but high volume. Lead Add form has very high conversion ratio, but low volume.



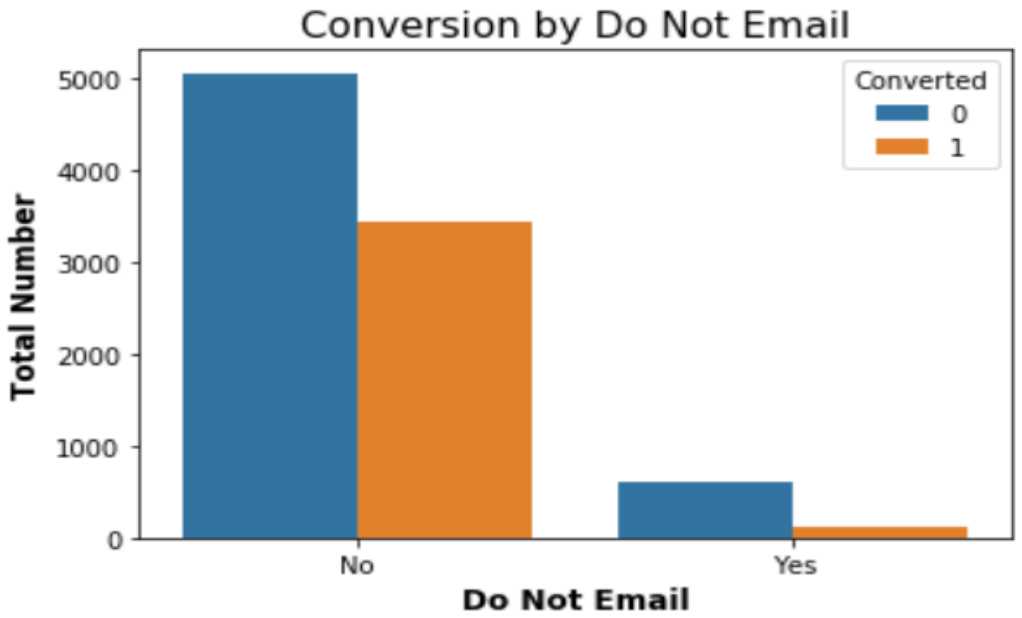
Conversion by Lead Source

Direct Traffic and Google have high volume of lead conversions. Conversion rates of leads through Reference and Welingak Website are high.



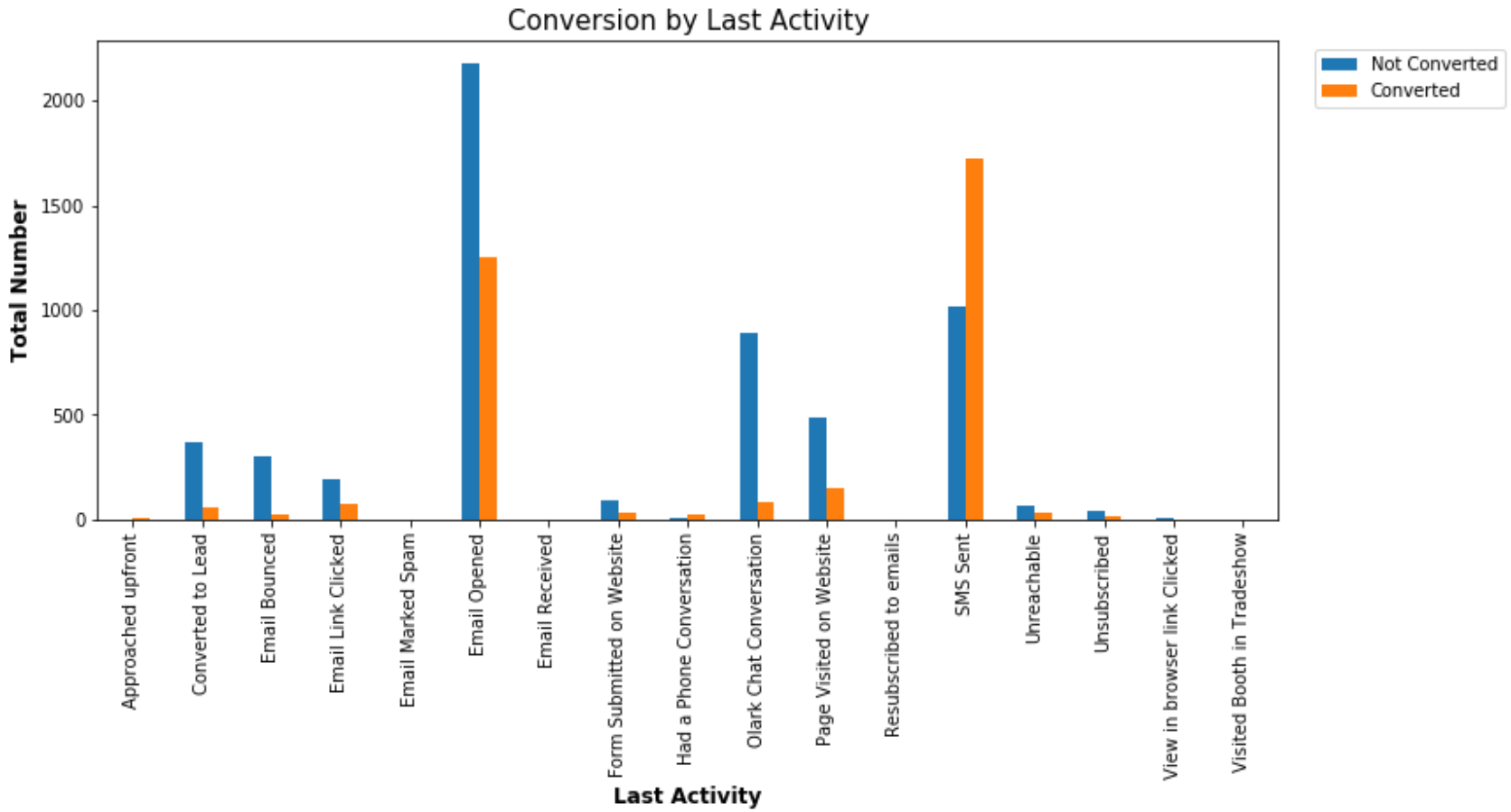
Conversion By Do not Email

Leads who select 'Do Not Email' have a high volume of conversion.



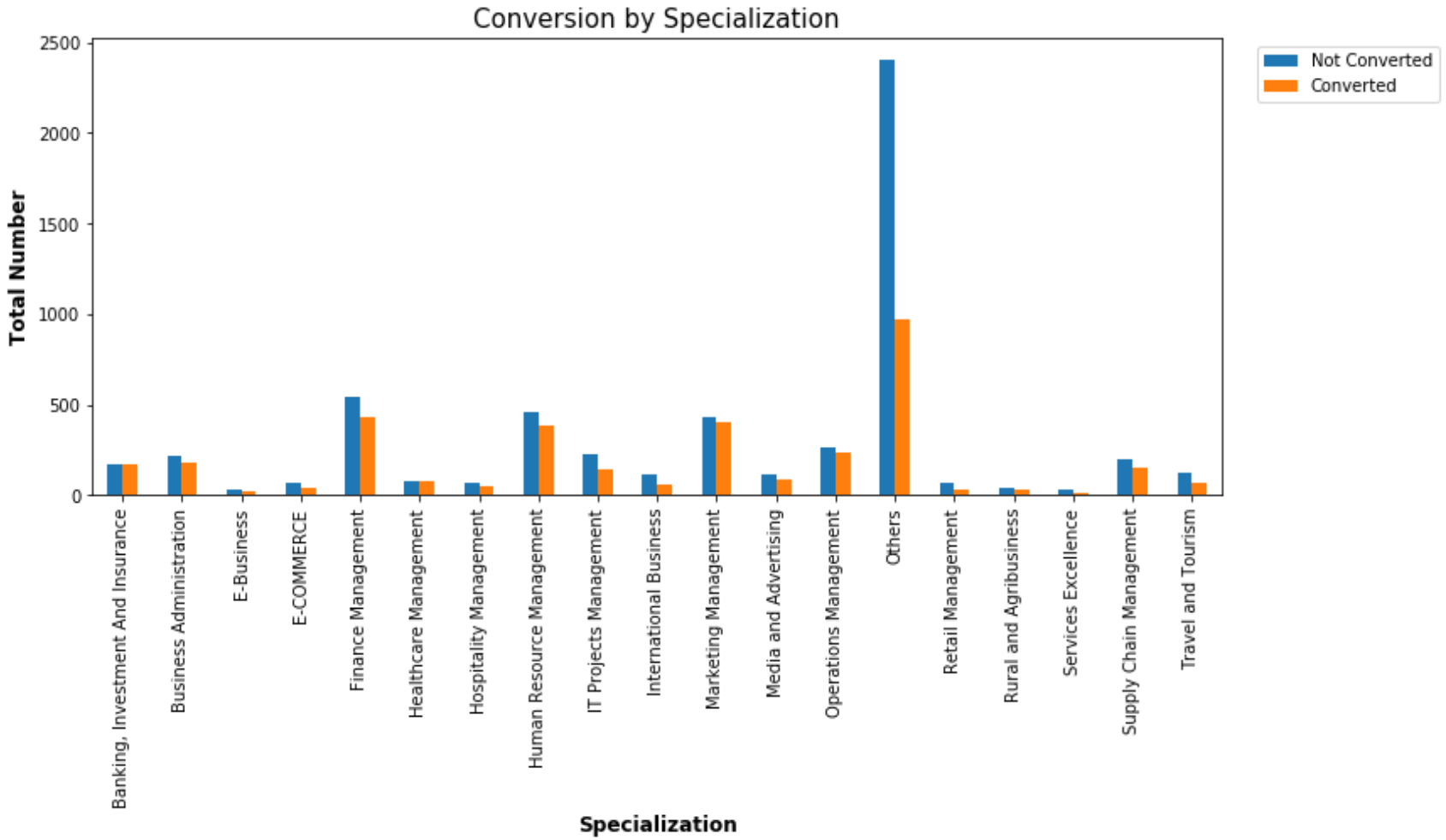
Conversion By Last Activity

Conversion rate for 'SMS Sent' is the highest. Email Opened last activity generates most number of leads.



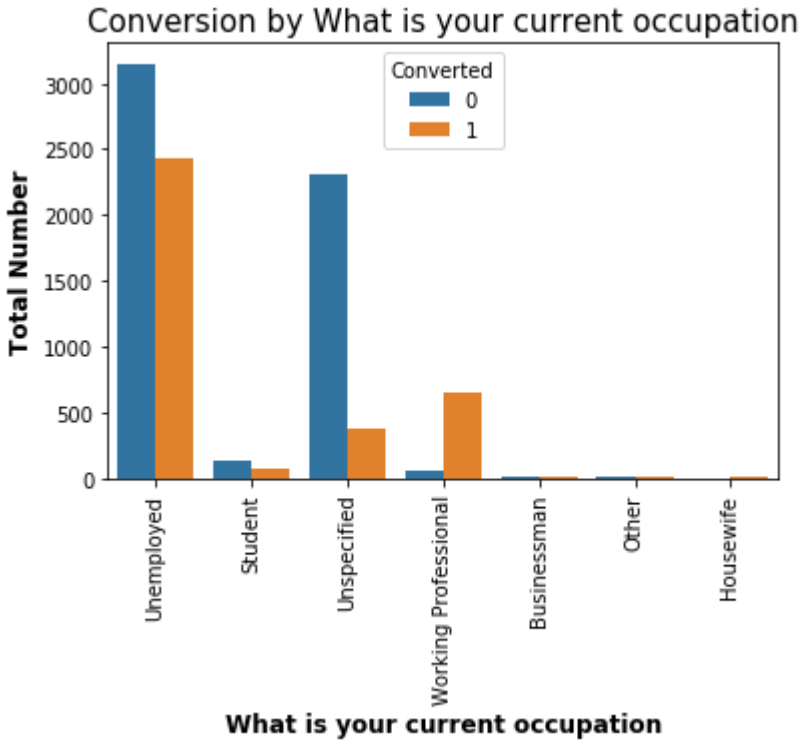
Conversion By Specialization

The conversion volume is the highest for which the Specialization imputed with ‘Others’.



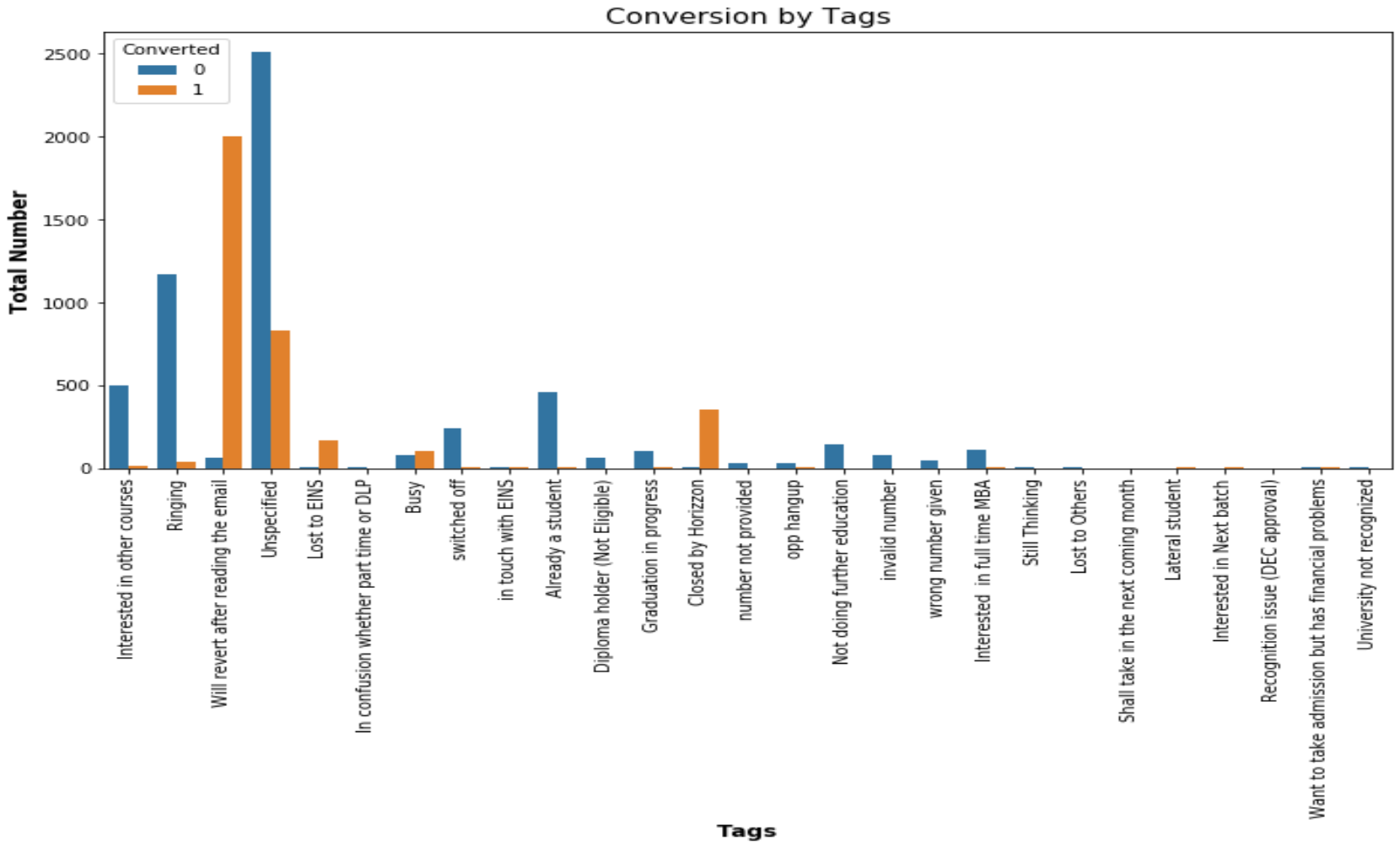
Conversion by Current Occupation

A very high volume of conversions are from Unemployed leads



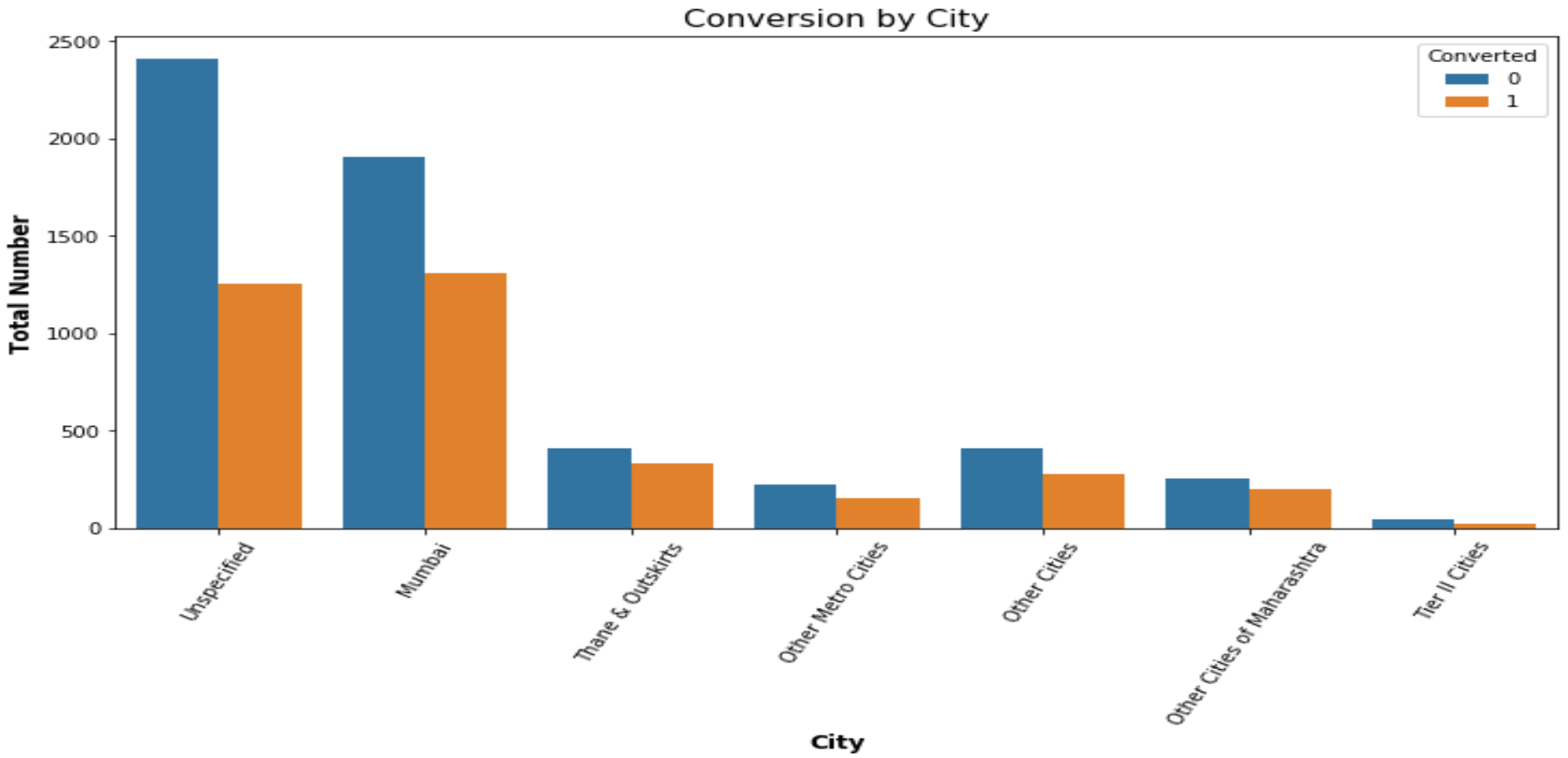
Conversion by Tags

There are some categories like University not recognized, Recognition issue (DEC approval), Lost to Others, wrong number given, number not provided, invalid number etc which almost never convert



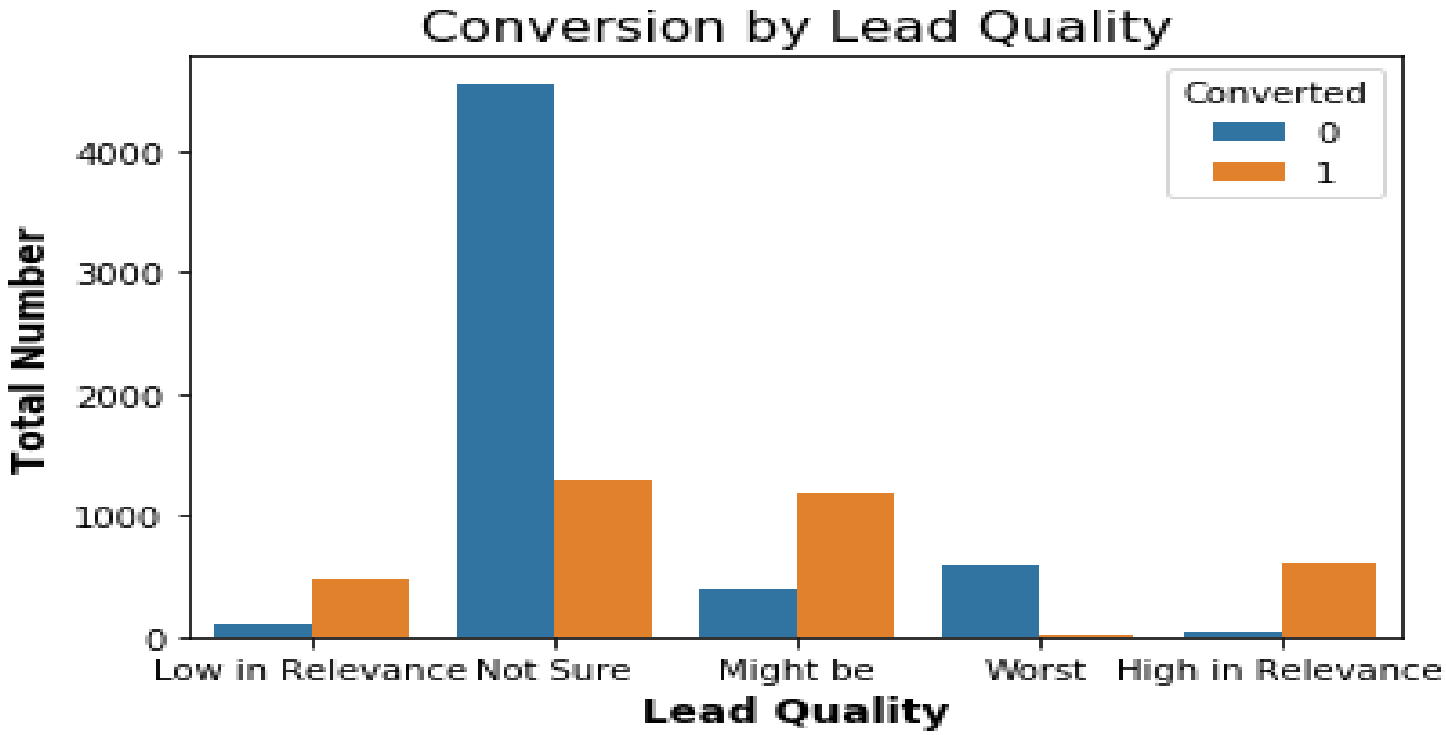
Conversion by City

There seems to be a high conversion rate from Metro Cities like Mumbai in comparison to Tier-II cities.



Conversion By Lead Quality

Might be has the highest Conversion ratio. Low in Relevance and High in Relevance have more Converted leads than not Converted. Rest types have low conversion ratio.



Generalized Linear Model Regression Results for the chosen model

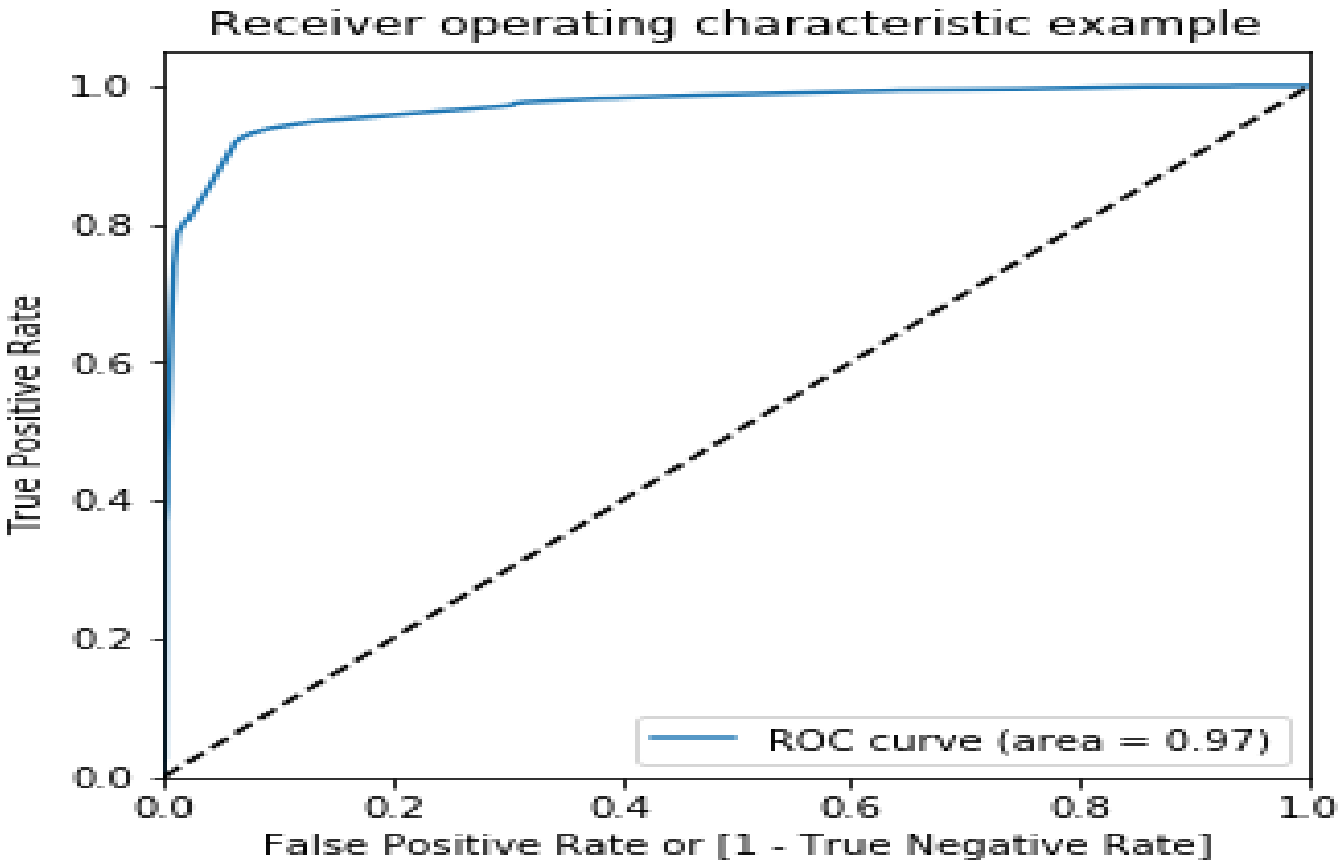
Dep. Variable:	Converted	**No. Observations:**	6335
Model:	GLM	**Df Residuals:**	6322
Model Family:	Binomial	**Df Model:**	12
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1257.5
Date:	Sun, 09 Jun 2019	**Deviance:**	2515.1
Time:	20:07:53	**Pearson chi2:**	1.12e+04
No. Iterations:	8	**Covariance Type:**	nonrobust

	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.1694	0.207	-15.331	0.000	-3.575	-2.764
Lead Source_Welingak Website	3.1827	1.024	3.109	0.002	1.176	5.189
Last Activity_SMS Sent	2.1615	0.118	18.323	0.000	1.930	2.393
What is your current occupation_Unspecified	-2.4769	0.141	-17.554	0.000	-2.753	-2.200
Tags_Busy	2.4606	0.290	8.479	0.000	1.892	3.029
Tags_Closed by Horizzon	8.7293	0.745	11.719	0.000	7.269	10.189
Tags_Lost to EINS	9.1703	0.652	14.062	0.000	7.892	10.448
Tags_Ringing	-1.5686	0.297	-5.290	0.000	-2.150	-0.987
Tags_Unspecified	3.5306	0.230	15.347	0.000	3.080	3.982
Tags_Will revert after reading the email	6.6086	0.265	24.964	0.000	6.090	7.127
Tags_switched off	-2.1103	0.623	-3.385	0.001	-3.332	-0.888
Lead Quality_Worst	-2.1231	0.727	-2.922	0.003	-3.547	-0.699
Last Notable Activity_Modified	-1.5345	0.126	-12.174	0.000	-1.782	-1.287

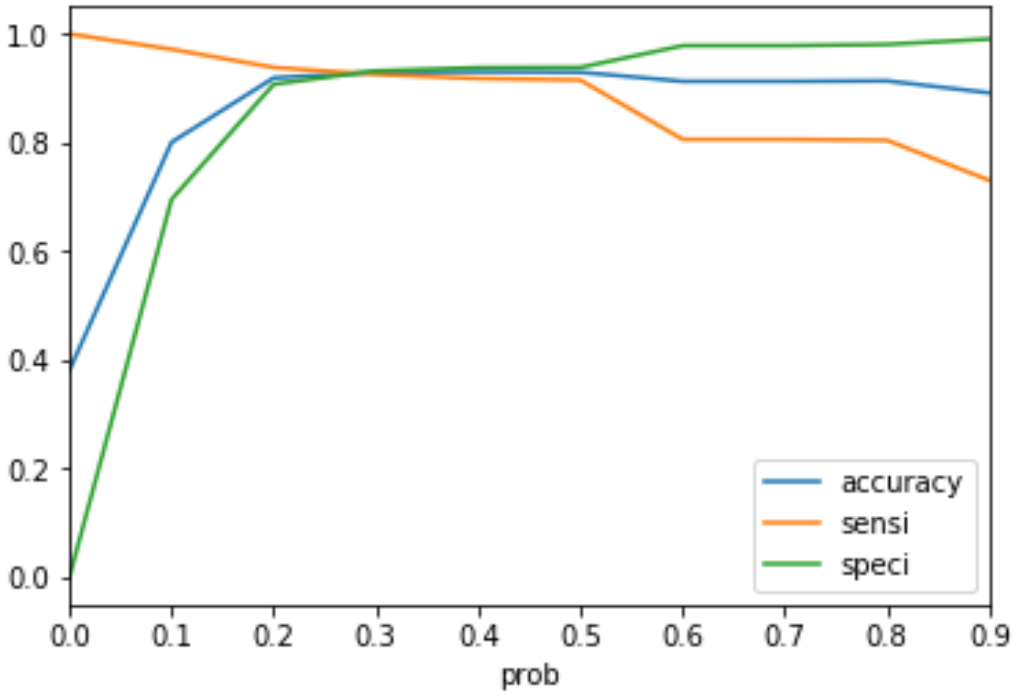
VIF scores for the features retained in the chosen model

Features	VIF
Lead Source_Welingak Website	1.12
Tags_Lost to EINS	1.06
Tags_Closed by Horizzon	1.05
Tags_Busy	1.04
Tags_switched off	1.03
Lead Quality_Worst	0.42
What is your current occupation_Unspecified	0.21
Last Notable Activity_Modified	0.15
Last Activity_SMS Sent	0.12
Tags_Will revert after reading the email	0.11
Tags_Unspecified	0.10
Tags_Ringing	0.07

ROC Curve


Receiver operating characteristic example
True Positive Rate
False Positive Rate or [1 - True Negative Rate]
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.97)

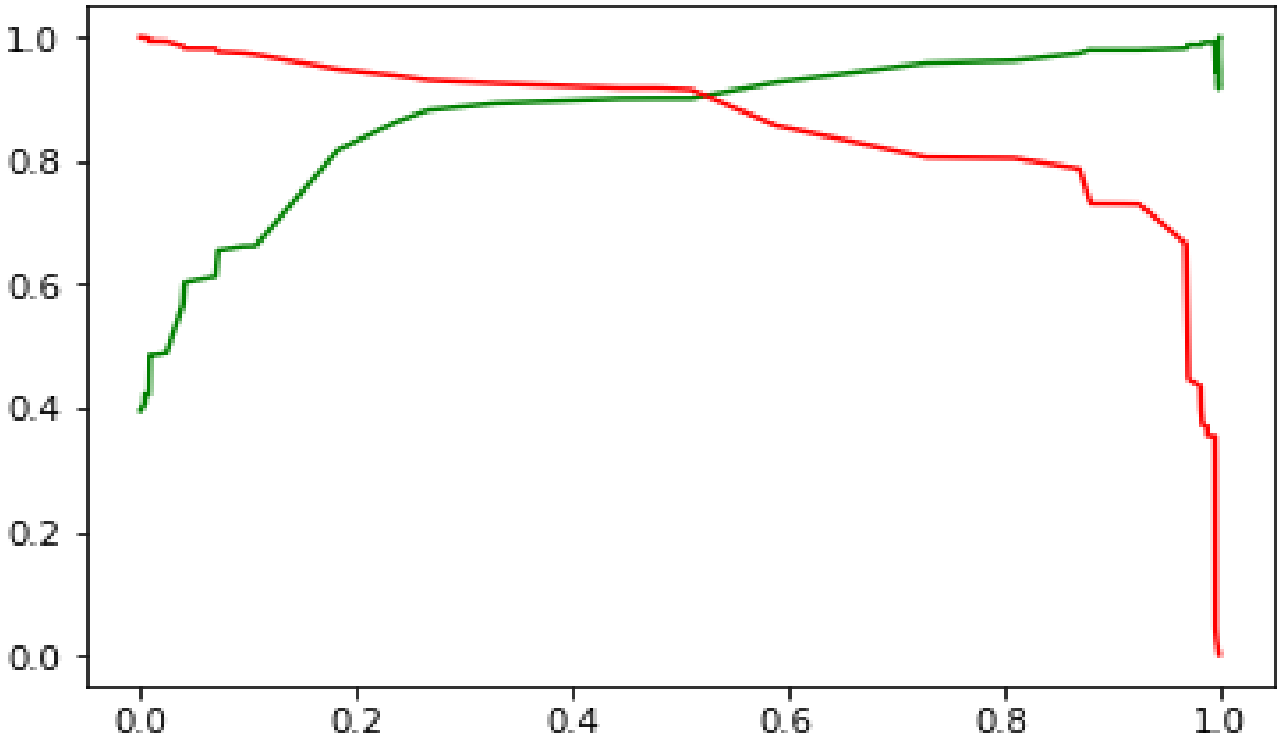
Sensitivity-Specificity Trade Off



From the curve above, 0.3 is the optimum point to take it as a cutoff probability.

Precision Recall Curve

Precision-Recall Curve gives an optimal probability of ~0.55



Results and Conclusions

```
Overall Accuracy: 0.9199684293606946
Confusion Matrix:
[[3760  163]
 [ 344 2068]]
Sensitivity: 0.857379767827529
Specificity: 0.9584501656895233
False Positive Rate: 0.04154983431047667
Positive Predictive Value: 0.9269385925593904
Negative Predictive Value: 0.9161793372319688
Precision: 0.9269385925593904
Recall: 0.857379767827529
```

Above are the metrics for the train data.

```
Overall Accuracy: 0.927098674521355
Confusion Matrix:
[[1636   66]
 [ 132  882]]
Sensitivity: 0.8698224852071006
Specificity: 0.9612220916568742
False Positive Rate: 0.038777908343125736
Positive Predictive Value: 0.930379746835443
Negative Predictive Value: 0.9253393665158371
Precision: 0.930379746835443
Recall: 0.8698224852071006
```

Above are the metrics on the test data.

Recommendations to increase lead conversion

- Lead Origin - 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 - Focus should be on improving lead conversion of Olark chat, organic search, direct traffic, and google leads and on generating more leads from reference and welingak website.
- Leads spending more time on the website are more likely to be converted. Website should be made more engaging to make leads spend more time.
- 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%. These activities can be tracked for better lead conversion.
- 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. These two categories should be pursued aggressively.
- Leads tagged as ‘Will revert after reading the email’ have a very high volume of conversion, and should be pursued aggressively.
- Most leads are from Mumbai with around 30% conversion rate. Leads from Mumbai can be pursued more aggressively after taking their Email/Call preferences.