LEAD SCORING CASE STUDY

Focused business approach using logistic regression technique

Group Members

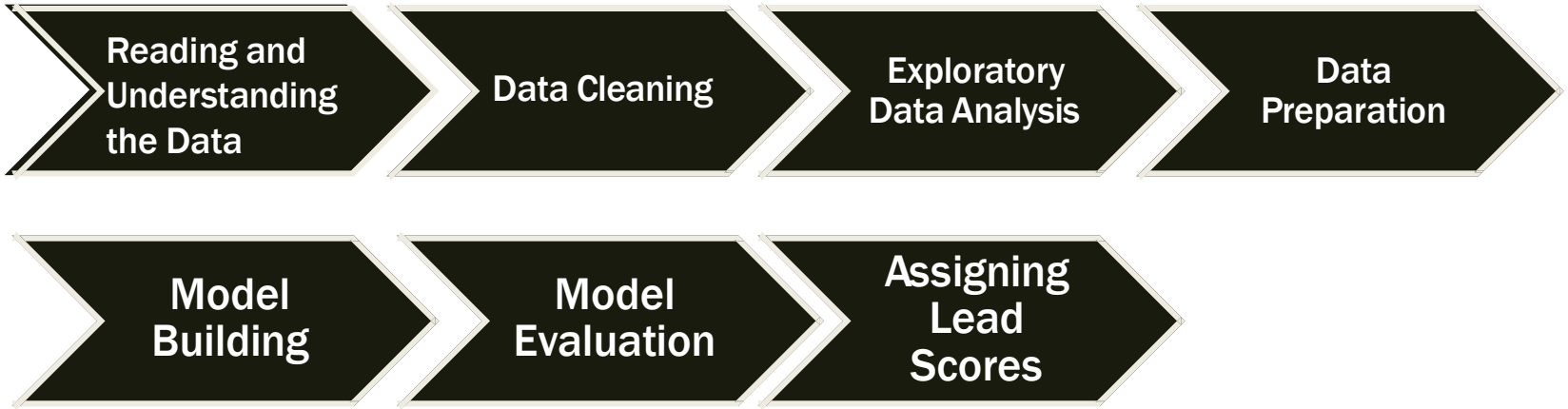
Nishanth Thulasiram & Girisankar

BUSINESS OBJECTIVE

- To help X Education select most promising leads, i.e. the leads that are most likely to convert into paying customers.
- 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.

METHODOLOGY

- To build a Logistic Regression model that assigns lead scores to all leads such that the customers with higher lead score have a higher conversion chance and vice versa.
- Target Lead Conversion Rate ≈ 80%



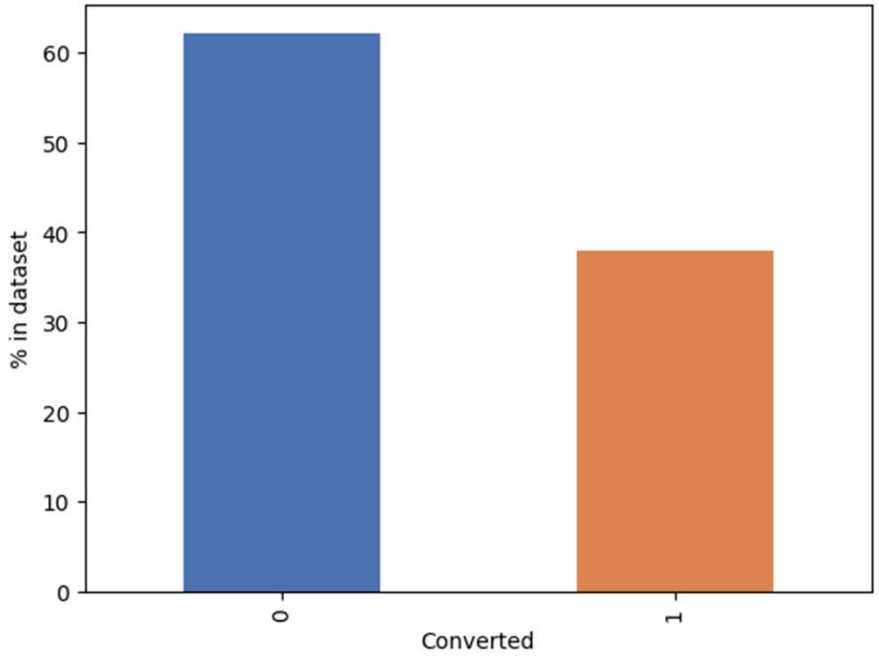
DATA MANIPULATION

- Total Number of Rows =37, Total Number of Columns =9240.
- Here we checked for discrepancies in the dataset
- Checking for any column names correction
- Checking for null values and imputing them with appropriate methods
- We used mode imputation for categorical columns.
- We used mean imputation for numerical columns, if there is no skewness in data.
- We used median imputation for numerical columns, if there is skewness in the data.

DATA VISUALIZATION & OUTLIERS TREATMENT

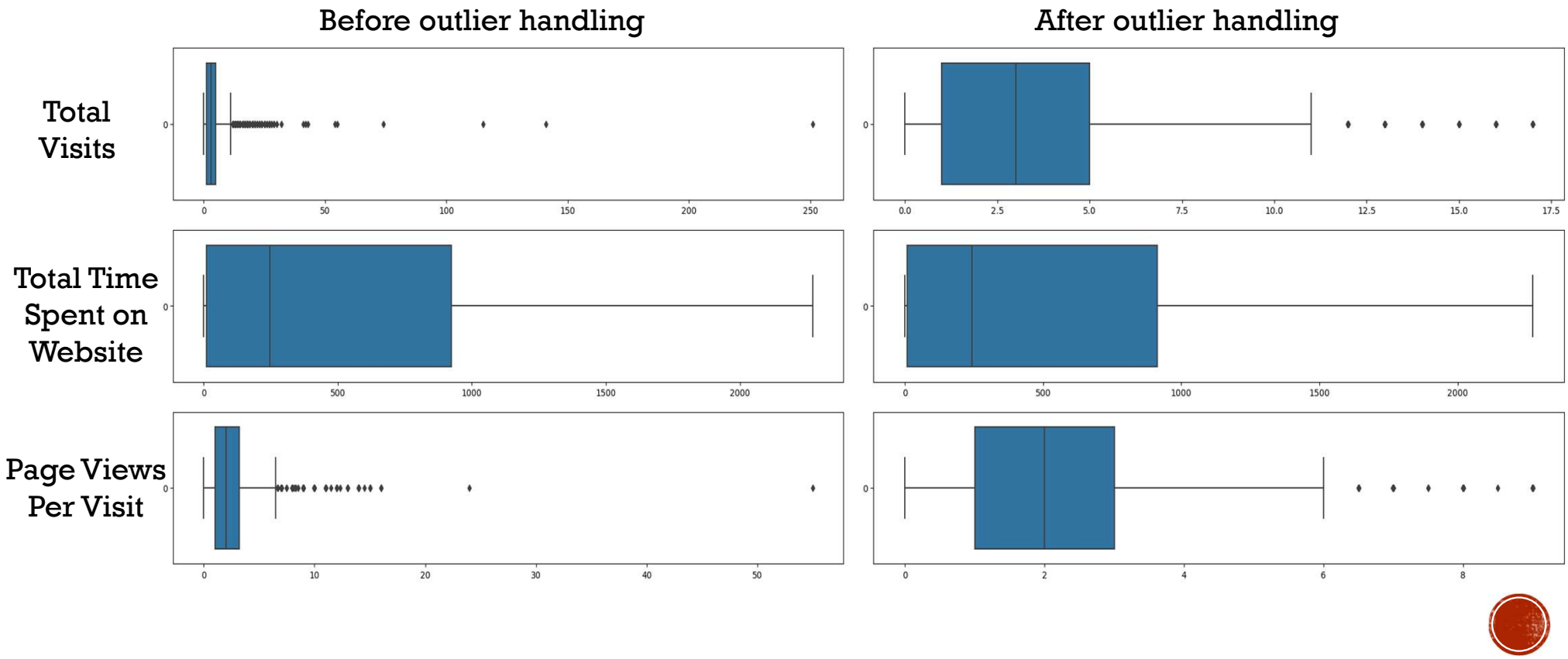
- We performed univariate analysis on categorical column to see which columns makes more sense and removed those columns whose variance is nearly zero.
- We performed bivariate analysis on categorical columns to see how they vary w.r.t Converted column.
- We performed univariate analysis on numerical columns by plotting box plots to see are there any outliers in the data or not.
- We performed bivariate analysis on numerical columns with Converted column to see how the leads are related to these columns.
- We have used IQR method to treat the outliers in the data set.
- In this step we also plotted the correlation matrix to identify the columns which are correlated.

DATA VISUALIZATION

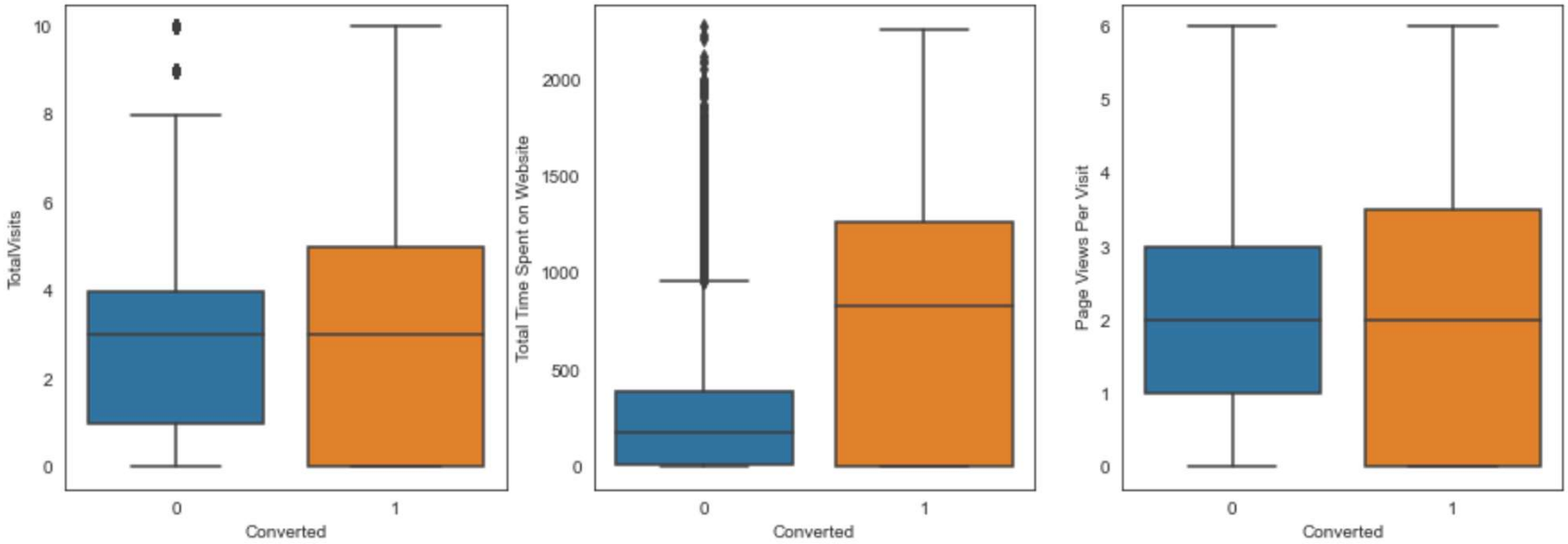


In the data frame 37.8% of the leads are converted. This means we have enough data of converted leads for modelling.

NUMERICAL DATA

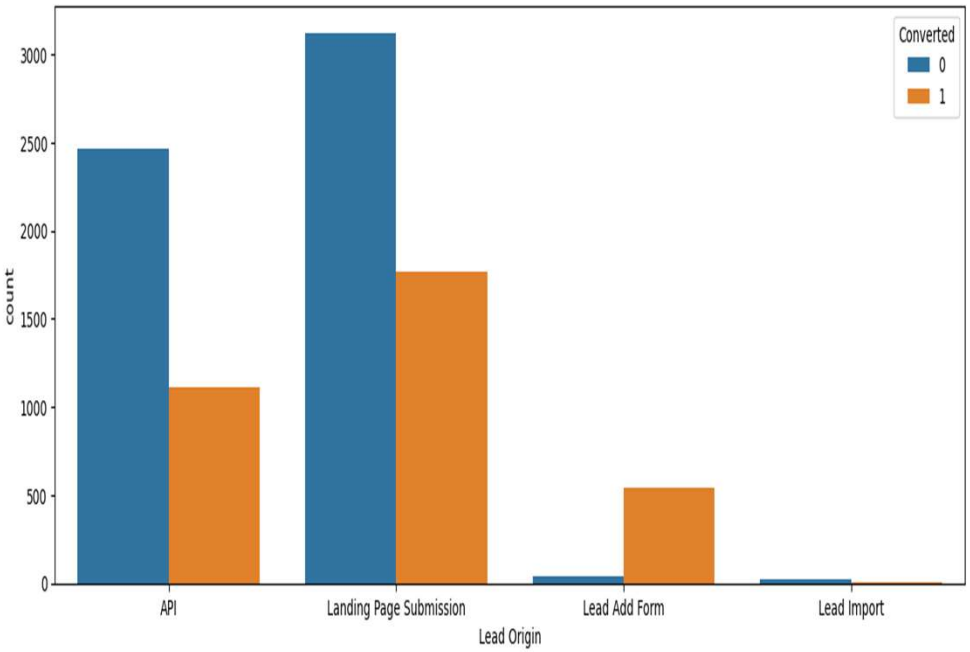


NUMERICAL DATA



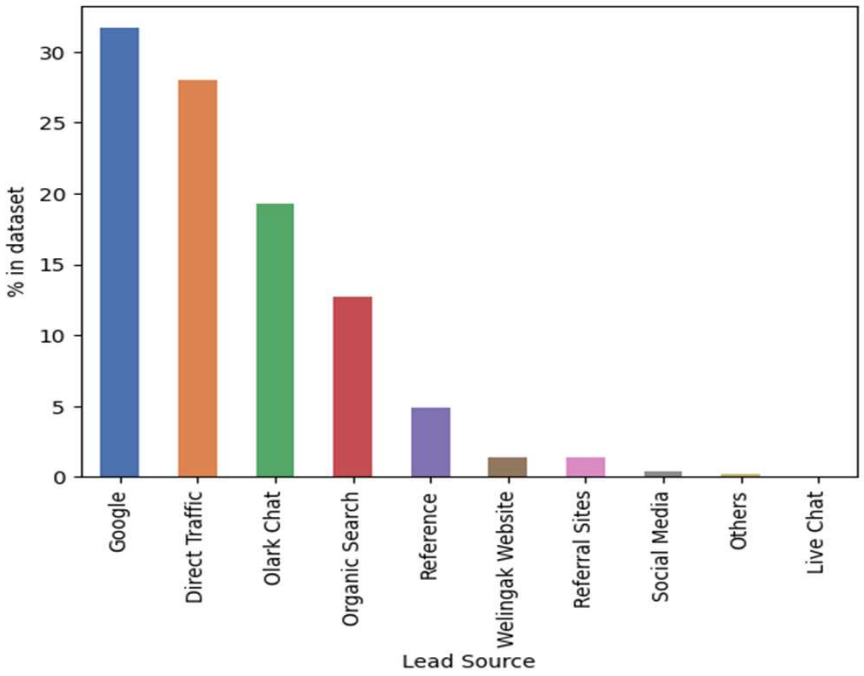
People spending more time on website are more likely to get converted

LEAD ORIGIN



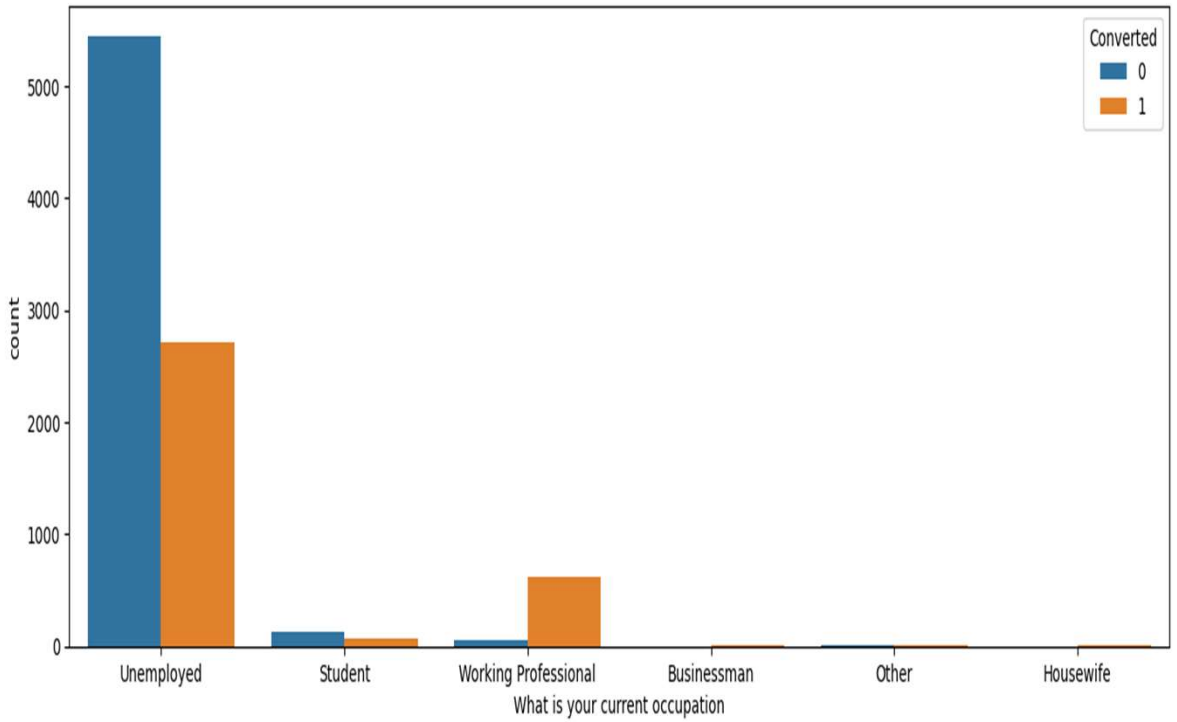
- API and Landing Page Submission bring higher number.
- Lead Add Form has a very high conversion rate but count of leads are not very high.
- Lead Import and Quick Add Form get very few leads.
- In order to improve overall lead conversion rate, we have to improve lead converion of API and Landing Page Submission origin and generate more leads from Lead Add Form.

LEAD SOURCE



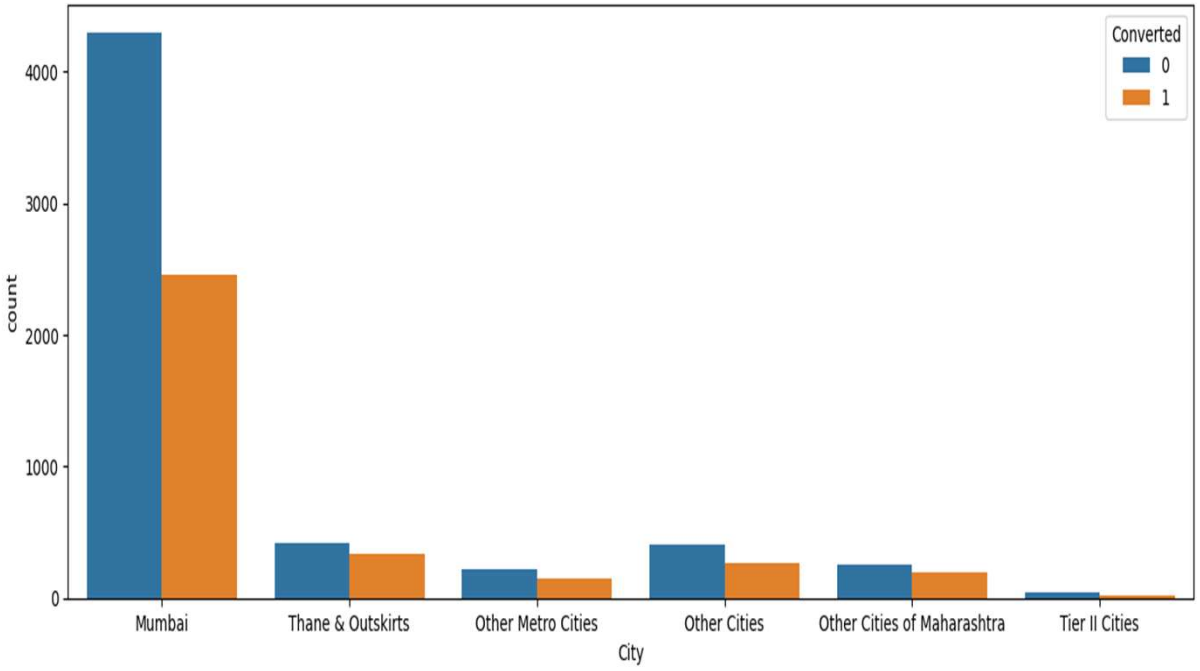
- Maximum number of leads are generated by Google and Direct traffic.
- Conversion Rate of reference leads and leads through welingak website is high.
- To improve overall lead conversion rate, focus should be on improving lead converion of olark chat, organic search, direct traffic, and google leads and generate more leads from reference and welingak website.

WHAT IS YOUR CURRENT OCCUPATION



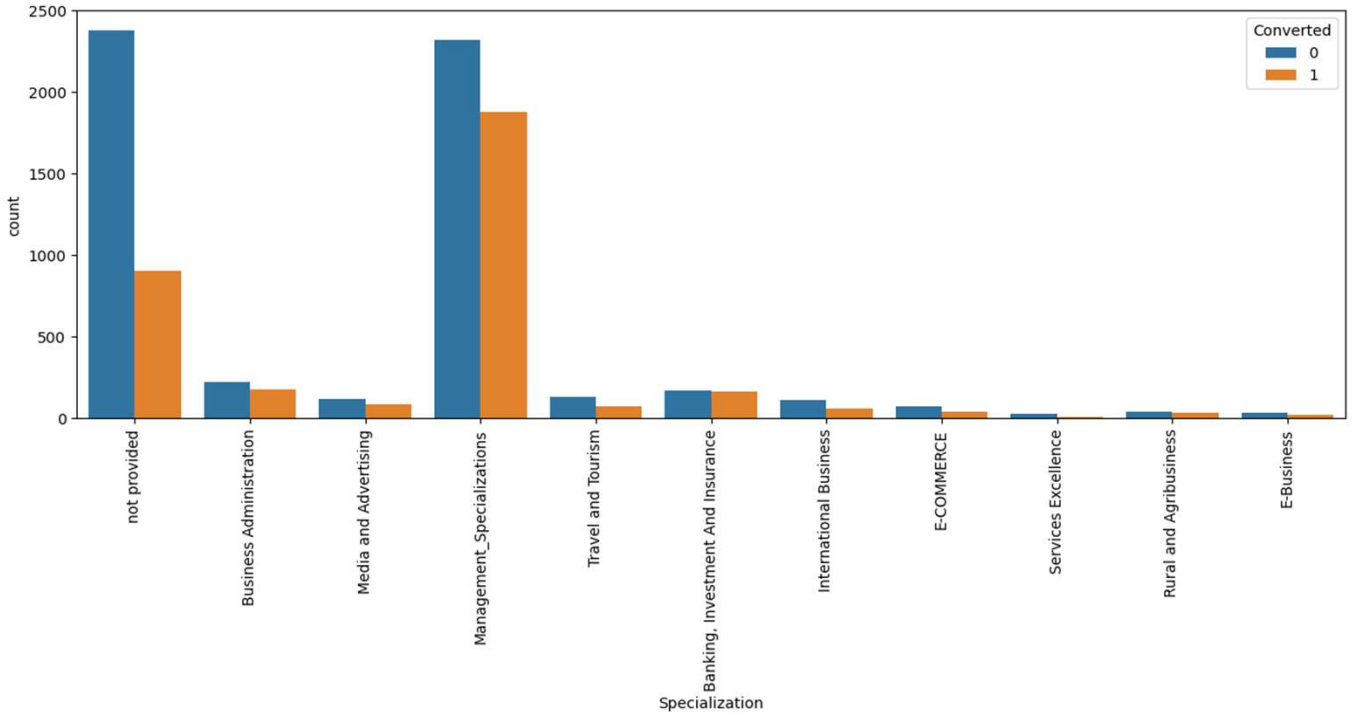
- Working Professionals going for the course have high chances of joining it.
- Unemployed leads are the most in terms of Absolute numbers.

CITY



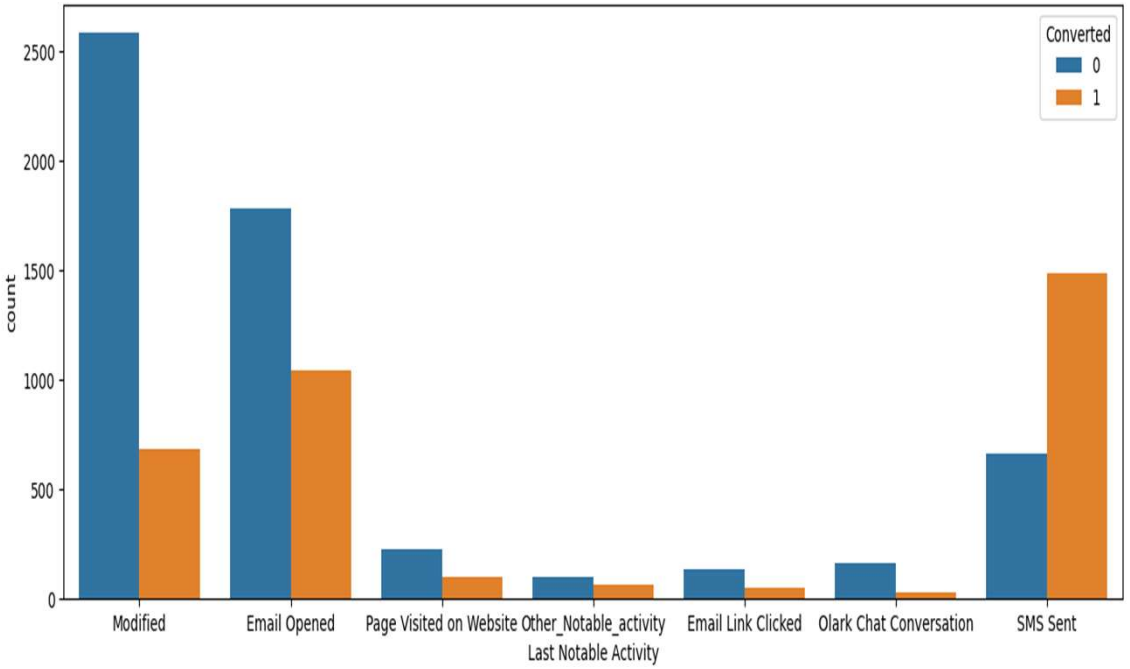
- In City, most of the leads are generated for 'Mumbai'.

SPECIALIZATION



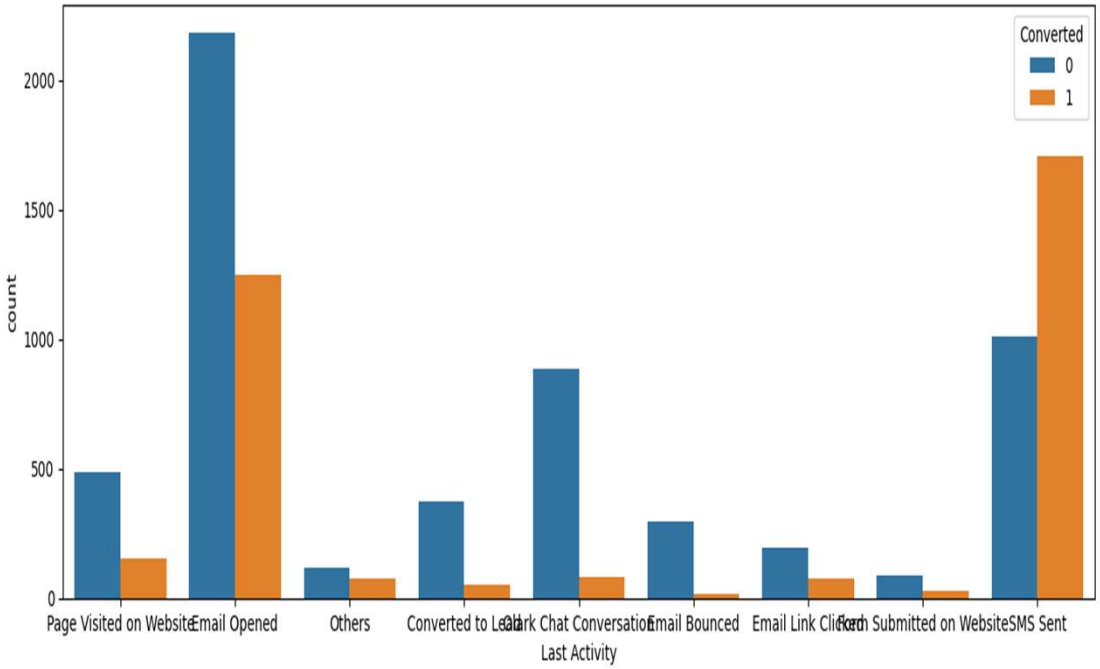
- Most of them are looking into the management specialization.

LAST NOTABLE ACTIVITY



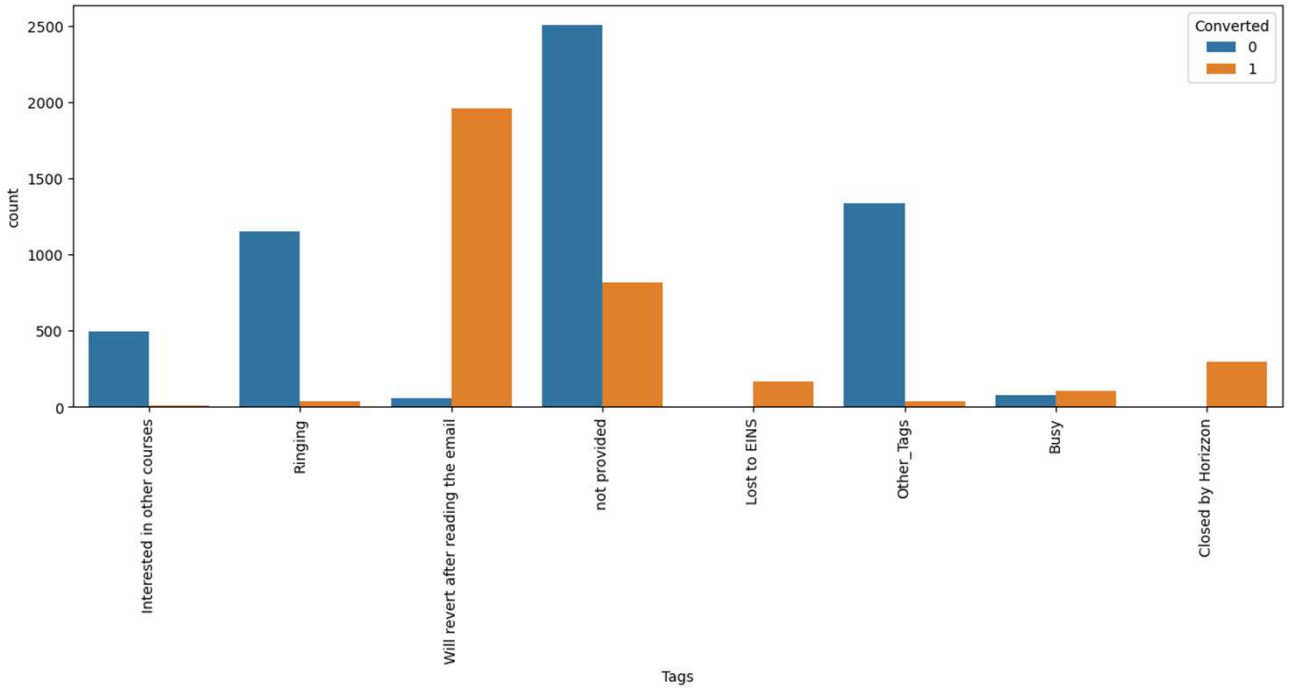
- Most generated leads for the category 'Modified' & most conversion rates in 'SMS Sent' and 'Email Opened' activities.

LAST ACTIVITY



- Leads are generated where the last activity is 'Email Opened' while maximum conversion rate is for the activity of 'SMS Sent'.

TAGS



- Leads are generated where the Tag is 'Will revert after reading the email' option.

MODEL EVALUATION - MODEL 1

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6246
Model:	GLM	Df Residuals:	6230
Model Family:	Binomial	Df Model:	15
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1240.1
Date:	Tue, 20 Jun 2023	Deviance:	2480.3
Time:	13:06:01	Pearson chi2:	8.90e+03
No. Iterations:	8	Pseudo R-squ. (CS):	0.6059
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.5859	0.096	-16.496	0.000	-1.774	-1.398
Total Time Spent on Website	1.0830	0.061	17.798	0.000	0.964	1.202
Lead Origin_Lead Add Form	1.7763	0.425	4.177	0.000	0.943	2.610
Lead Source_Olark Chat	1.2092	0.144	8.406	0.000	0.927	1.491
Lead Source_Welingak Website	3.4826	0.848	4.107	0.000	1.821	5.145
Last Activity_Email Bounced	-1.4939	0.541	-2.760	0.006	-2.555	-0.433
Last Activity_SMS Sent	1.3820	0.231	5.994	0.000	0.930	1.834
Tags_Closed by Horizzon	6.4254	0.738	8.705	0.000	4.979	7.872
Tags_Interested in other courses	-2.0275	0.394	-5.143	0.000	-2.800	-1.255
Tags_Lost to EINS	5.3589	0.533	10.063	0.000	4.315	6.403
Tags_Other_Tags	-2.6413	0.228	-11.605	0.000	-3.087	-2.195
Tags_Ringing	-3.6868	0.255	-14.470	0.000	-4.186	-3.187
Tags_Will revert after reading the email	4.4231	0.190	23.339	0.000	4.052	4.794
Last Notable Activity_Modified	-1.4460	0.155	-9.352	0.000	-1.749	-1.143
Last Notable Activity_Olark Chat Conversation	-1.7724	0.420	-4.216	0.000	-2.597	-0.948
Last Notable Activity_SMS Sent	0.7924	0.263	3.009	0.003	0.276	1.309

	Features	VIF
14	Last Notable Activity_SMS Sent	6.52
5	Last Activity_SMS Sent	6.36
12	Last Notable Activity_Modified	1.92
1	Lead Origin_Lead Add Form	1.77
11	Tags_Will revert after reading the email	1.60
2	Lead Source_Olark Chat	1.46
0	Total Time Spent on Website	1.44
3	Lead Source_Welingak Website	1.31
6	Tags_Closed by Horizzon	1.20
9	Tags_Other_Tags	1.18
7	Tags_Interested in other courses	1.13
10	Tags_Ringing	1.11
4	Last Activity_Email Bounced	1.10
8	Tags_Lost to EINS	1.06
13	Last Notable Activity_Olark Chat Conversation	1.06

MODEL EVALUATION - MODEL 2

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6246
Model:	GLM	Df Residuals:	6231
Model Family:	Binomial	Df Model:	14
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1244.7
Date:	Tue, 20 Jun 2023	Deviance:	2489.4
Time:	13:09:21	Pearson chi2:	8.84e+03
No. Iterations:	8	Pseudo R-squ. (CS):	0.6054
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.5349	0.094	-16.374	0.000	-1.719	-1.351
Total Time Spent on Website	1.0773	0.061	17.788	0.000	0.959	1.196
Lead Origin_Lead Add Form	1.7481	0.430	4.064	0.000	0.905	2.591
Lead Source_Olark Chat	1.2349	0.143	8.653	0.000	0.955	1.515
Lead Source_Welingak Website	3.4706	0.853	4.068	0.000	1.799	5.143
Last Activity_Email Bounced	-1.3974	0.545	-2.566	0.010	-2.465	-0.330
Last Activity_SMS Sent	1.9819	0.116	17.108	0.000	1.755	2.209
Tags_Closed by Horizzon	6.5893	0.739	8.917	0.000	5.141	8.038
Tags_Interested in other courses	-1.9646	0.392	-5.005	0.000	-2.734	-1.195
Tags_Lost to EINS	5.4909	0.535	10.265	0.000	4.443	6.539

	Features	VIF
1	Lead Origin_Lead Add Form	1.77
11	Tags_Will revert after reading the email	1.55
12	Last Notable Activity_Modified	1.55
5	Last Activity_SMS Sent	1.46
0	Total Time Spent on Website	1.44
2	Lead Source_Olark Chat	1.43
3	Lead Source_Welingak Website	1.31
6	Tags_Closed by Horizzon	1.20
9	Tags_Other_Tags	1.16
7	Tags_Interested in other courses	1.11
10	Tags_Ringing	1.10
4	Last Activity_Email Bounced	1.09
8	Tags_Lost to EINS	1.06
13	Last Notable Activity_Olark Chat Conversation	1.06

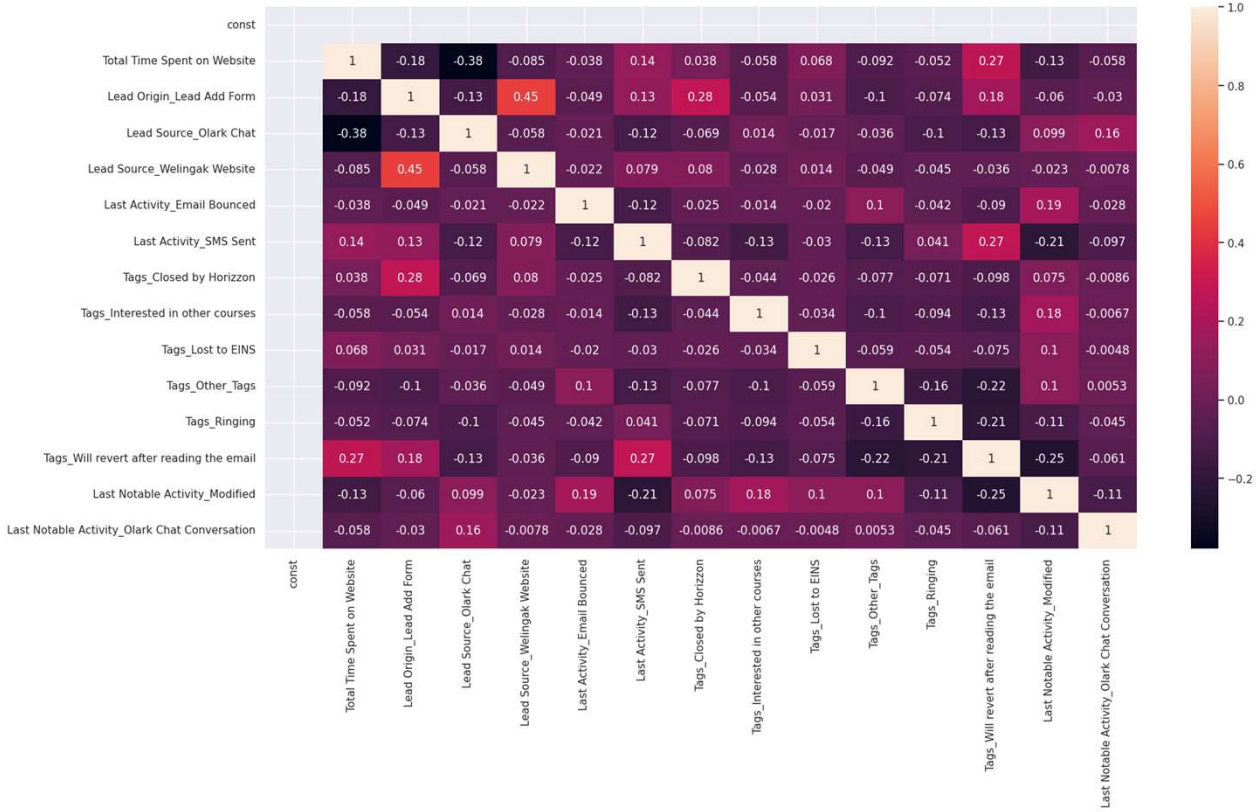
FINAL MODEL

- All p-values are very close to zero. VIFs for all features are very low.
- There is hardly any multicollinearity present.
- Training accuracy of 92.64% at a probability threshold of 0.05 is also very good.
- Confusion Matrix:

  [3701 170]

  [290 2085]
- Training Accuracy: 0.9263528658341339
- Sensitivity: 0.8778947368421053
- Specificity: 0.9560836993025058
- False postive rate - predicting the lead conversion when the lead does not convert: 0.043916300697494186
- Positive predictive value: 0.9246119733924612
- Negative predictive value: 0.9273365071410674

LOOKING AT CORRELATIONS



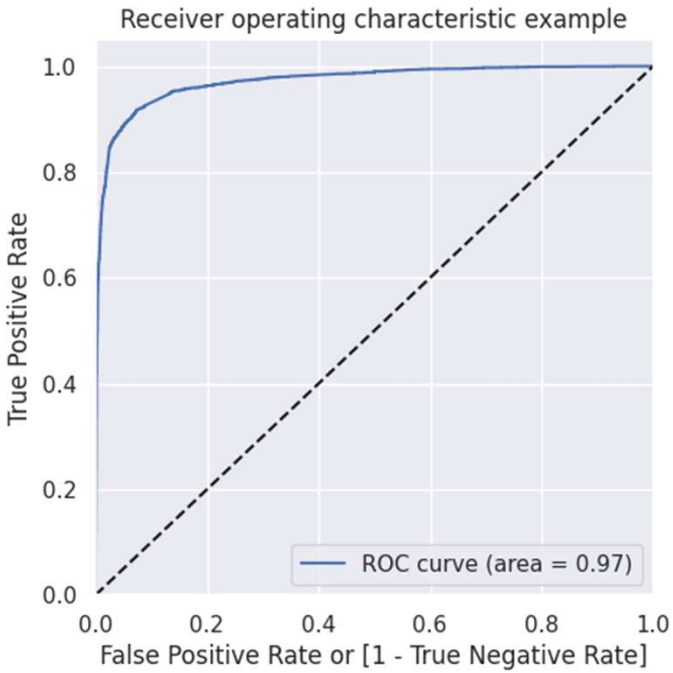
- From VIF values and heat maps, we can see that there is not much multicollinearity present.
- All variables have a good value of VIF.
- These features seem important from the business aspect as well. So we need not drop any more variables and we can proceed with making predictions using this model only

THE ROC CURVE

An ROC curve demonstrates several things:

- It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
- The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

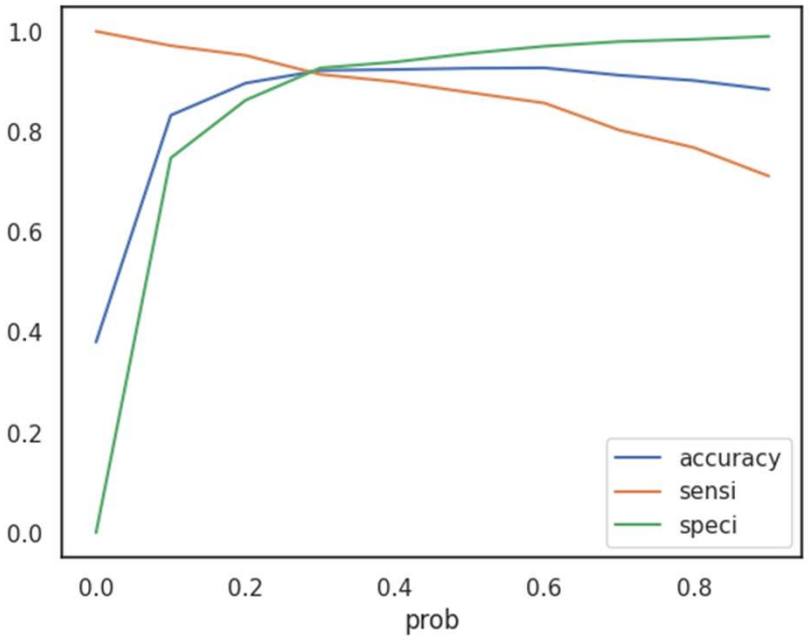
Area under curve (auc) is approximately 0.97 which is very close to ideal auc of 1.



FINDING OPTIMAL CUTOFF POINT

- Optimal cutoff probability is the prob where we get balanced sensitivity and specificity.

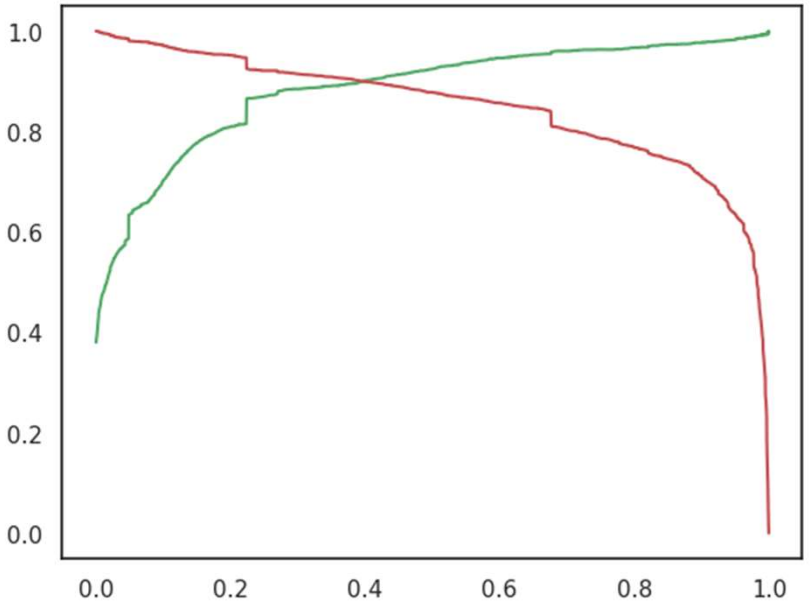
0.3 is the optimum point to take as a cutoff probability



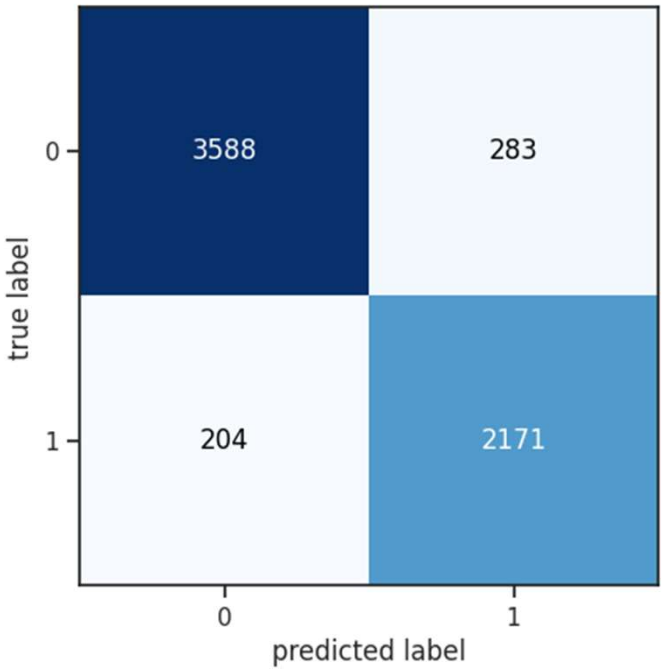
PRECISION AND RECALL TRADEOFF

To make predictions on the test dataset, optimum cutoff was considered as obtained from Precision recall graph of the train dataset as shown in the figure

- We can observe that **0.4** is the tradeoff between Precision and Recall. Thus we can safely choose to consider any Prospect Lead with Conversion Probability higher than 40% to be a hot Lead

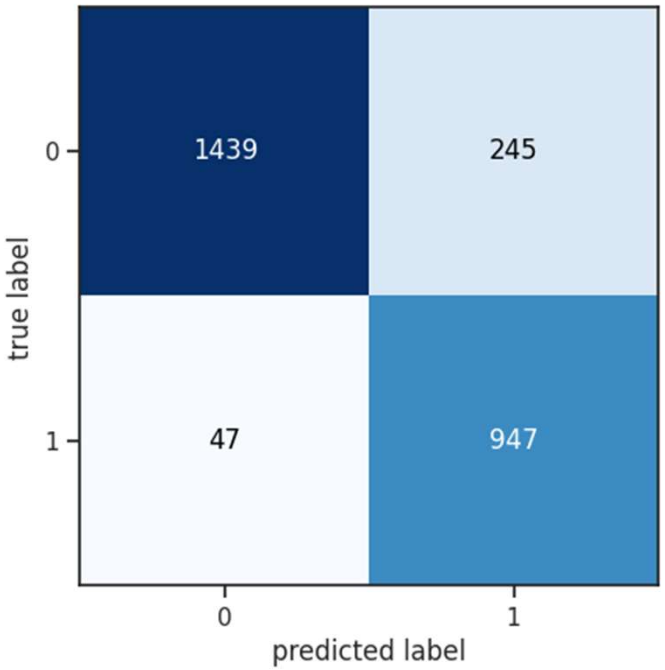


CONFUSION MATRIX ON TRAIN DATA



- Accuracy: 0.9263528658341339
- Sensitivity: 0.9141052631578948
- Specificity: 0.9268922758977008
- False positive rate - predicting the lead conversion when the lead does not convert: 0.07310772410229915
- Positive predictive value: 0.884678076609617
- Negative predictive value: 0.9462025316455697

CONFUSION MATRIX ON TEST DATA



- Accuracy: 0.8909634055265123
- Sensitivity: 0.9527162977867203
- Specificity: 0.8545130641330166
- False positive rate - predicting the lead conversion when the lead does not convert: 0.14548693586698339
- Positive predictive value: 0.7944630872483222
- Negative predictive value: 0.968371467025572

CLASSIFICATION REPORT

Train Data

	precision	recall	f1-score	support
Not Converted	0.95	0.93	0.94	3871
Converted	0.88	0.91	0.9	2375
accuracy			0.92	6246
macro avg	0.92	0.92	0.92	6246
weighted avg	0.92	0.92	0.92	6246

Test Data

	precision	recall	f1-score	support
Not Converted	0.97	0.85	0.91	1684
Converted	0.79	0.95	0.87	994
accuracy			0.89	2678
macro avg	0.88	0.9	0.89	2678
weighted avg	0.9	0.89	0.89	2678

DETERMINING FEATURE IMPORTANCE

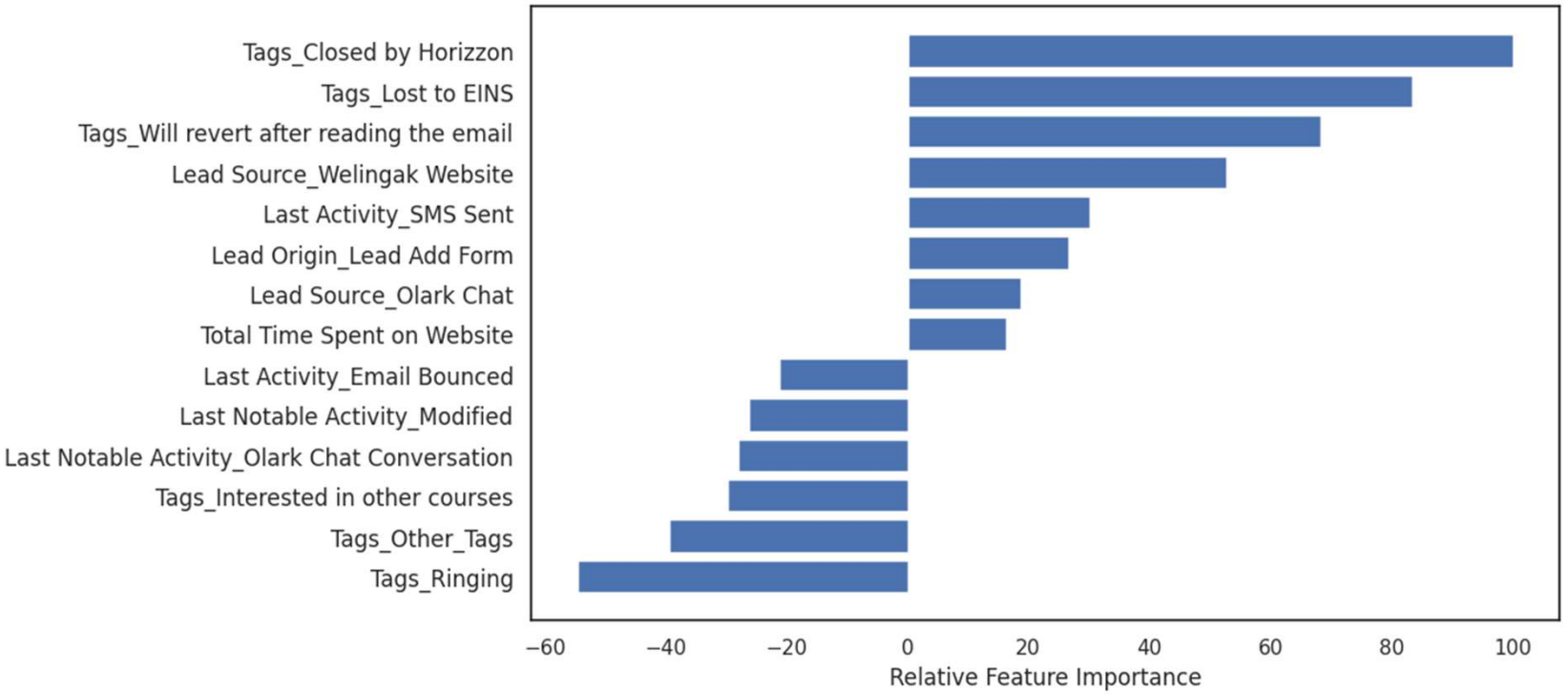
Features with corresponding coefficients in final model

Features	Coefficients
Tags_Closed by Horizzon	6.59
Tags_Lost to EINS	5.49
Tags_Will revert after reading the email	4.5
Lead Source_Welingak Website	3.47
Last Activity_SMS Sent	1.98
Lead Origin_Lead Add Form	1.75
Lead Source_Olark Chat	1.23
Total Time Spent on Website	1.08
Last Activity_Email Bounced	-1.4
Last Notable Activity_Modified	-1.73
Last Notable Activity_Olark Chat Conversation	-1.85
Tags_Interested in other courses	-1.96
Tags_Other_Tags	-2.6
Tags_Ringing	-3.59

Features with their relative importance

Features	Coefficients
Tags_Closed by Horizzon	100
Tags_Lost to EINS	83.33
Tags_Will revert after reading the email	68.25
Lead Source_Welingak Website	52.67
Last Activity_SMS Sent	30.08
Lead Origin_Lead Add Form	26.53
Lead Source_Olark Chat	18.74
Total Time Spent on Website	16.35
Last Activity_Email Bounced	-21.21
Last Notable Activity_Modified	-26.3
Last Notable Activity_Olark Chat Conversation	-28.11
Tags_Interested in other courses	-29.82
Tags_Other_Tags	-39.4
Tags_Ringing	-54.53

GRAPH OF RELATIVE IMPORTANCE



THANK YOU