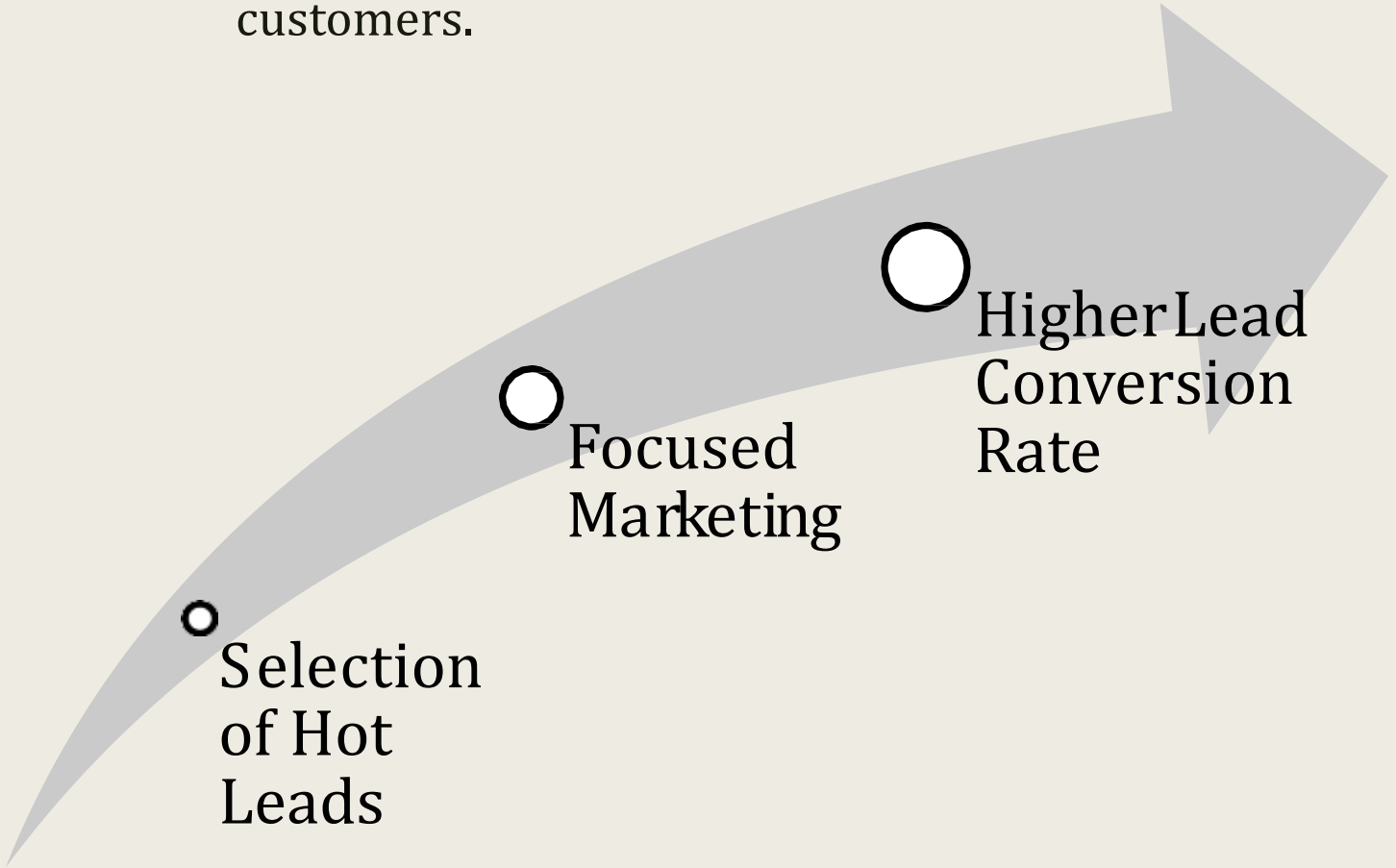
LEAD SCORING CASE STUDY

Focused business approach using logistic regression technique

Chaitanya Devarshi

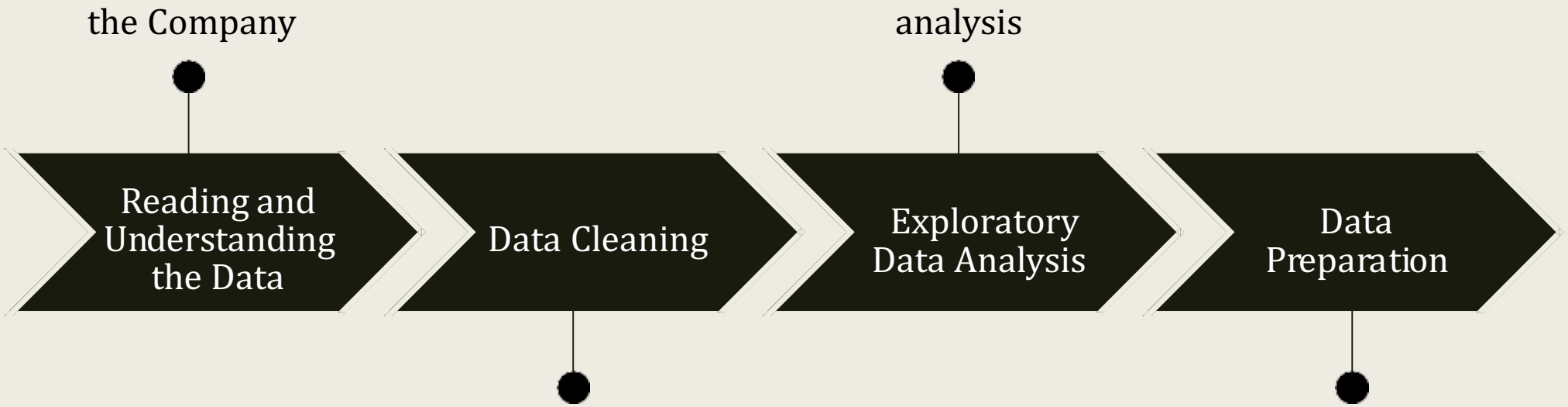
Business Objective

To help X Education select most promising leads ***(Hot Leads)***, i.e. the leads that are most Nlikelyto convert into paying customers.

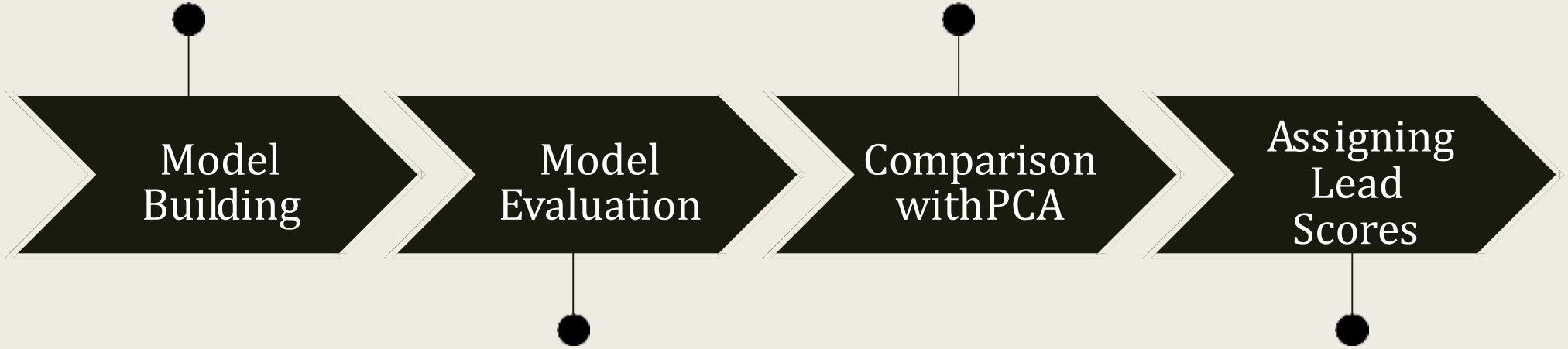


Importing and Observing the past data provided by

Univariate and Bivariate



- Missing value imputation
- Removing duplicate data
- Outlier treatment
- Dropping unnecessary columns and other redundancies
- Feature standardization
- Feature selection using RFE • Manual feature elimination based on p-values and VIFs
- Dummy variable creation
- Building another model using PCA
- Comparing the two models

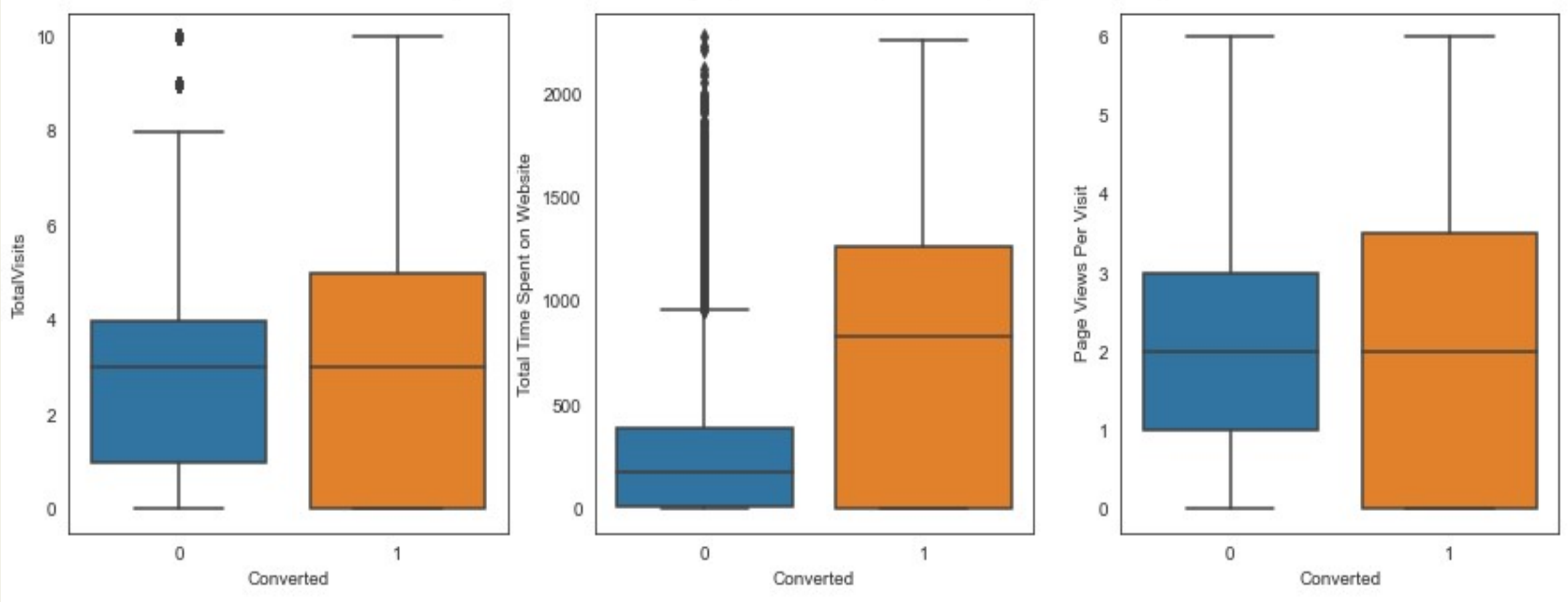


- Evaluating model based on various evaluation metrics

- Finalizing the first model
- Using predicted probabilities to calculate Lead Scores:

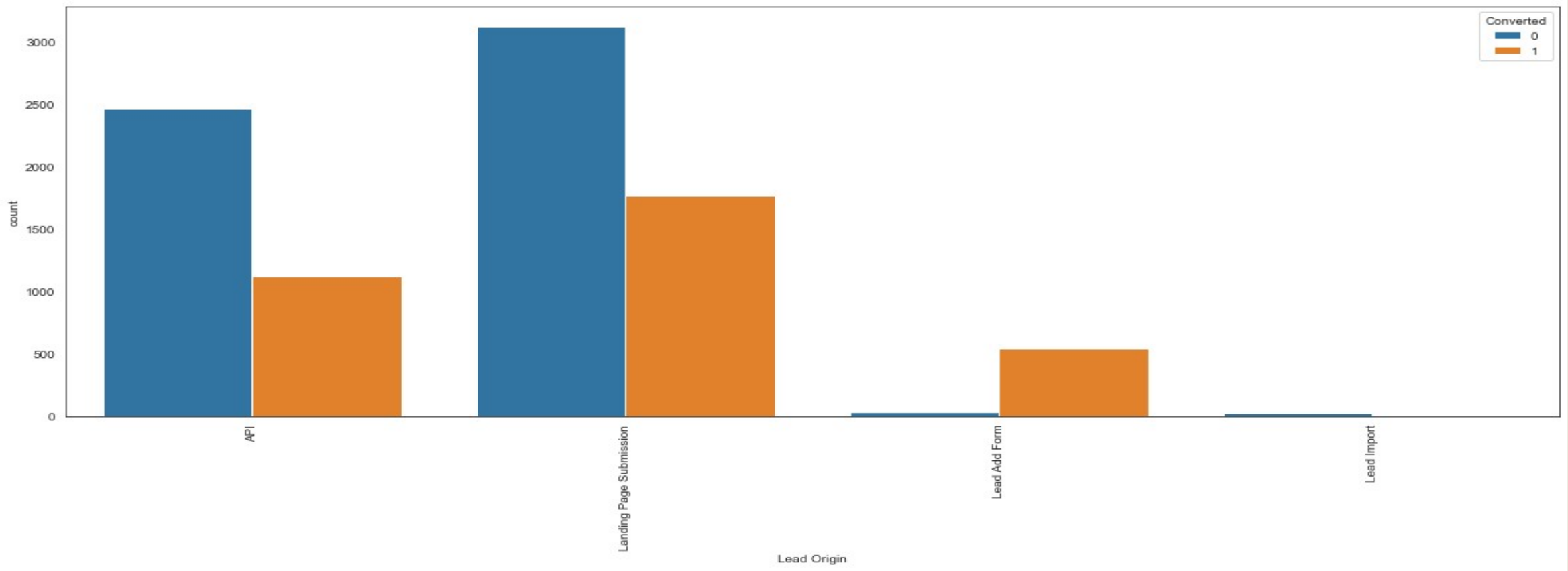
- Finding the optimal **LeadScore= Probability* 100** probability threshold

Numerical Variables



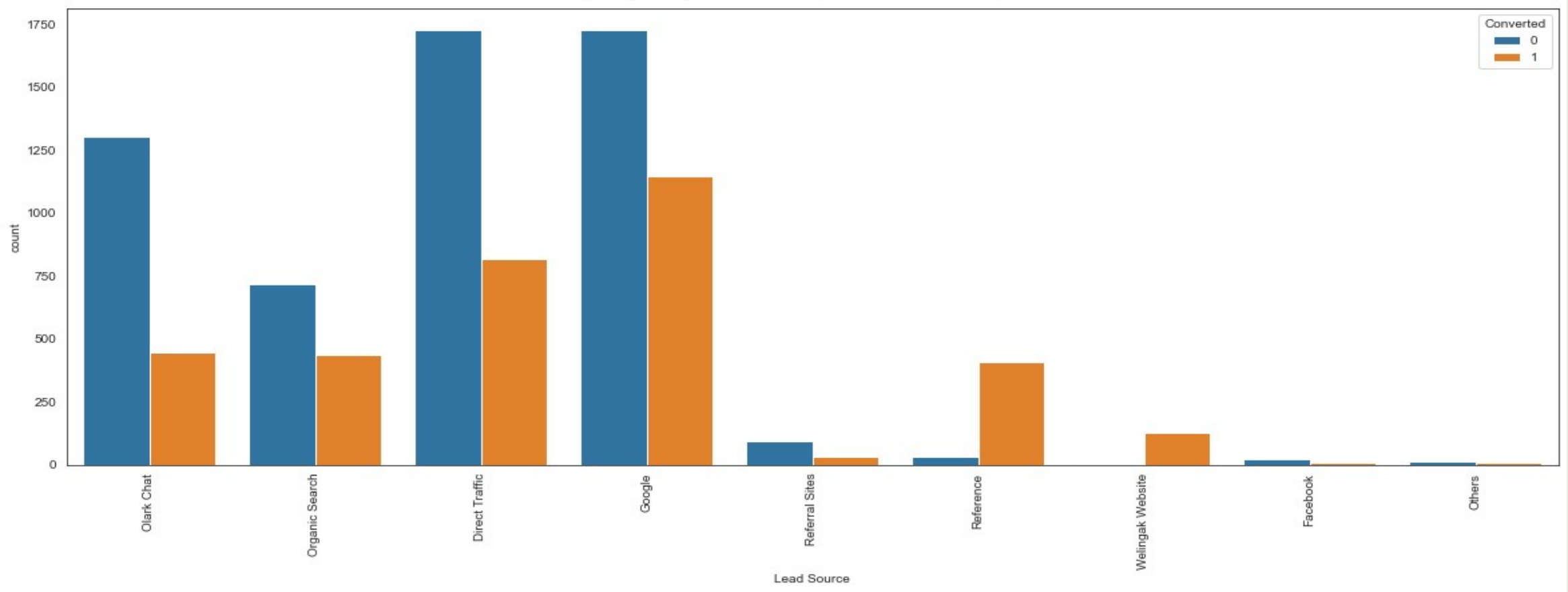
Peoplespendingmoretimeonwebsitearemorelikelytogetconverted.

Lead Origin



- ***'API*** and ***'LandingPageSubmission*** generate the most leads but have less conversion rates, whereas ***'LeadAddForm*** generates less leads but conversion rate is great.
- **Try to increase conversion rate for 'API' and 'LandingPageSubmission', and increase leads generation using 'LeadAddForm'.**

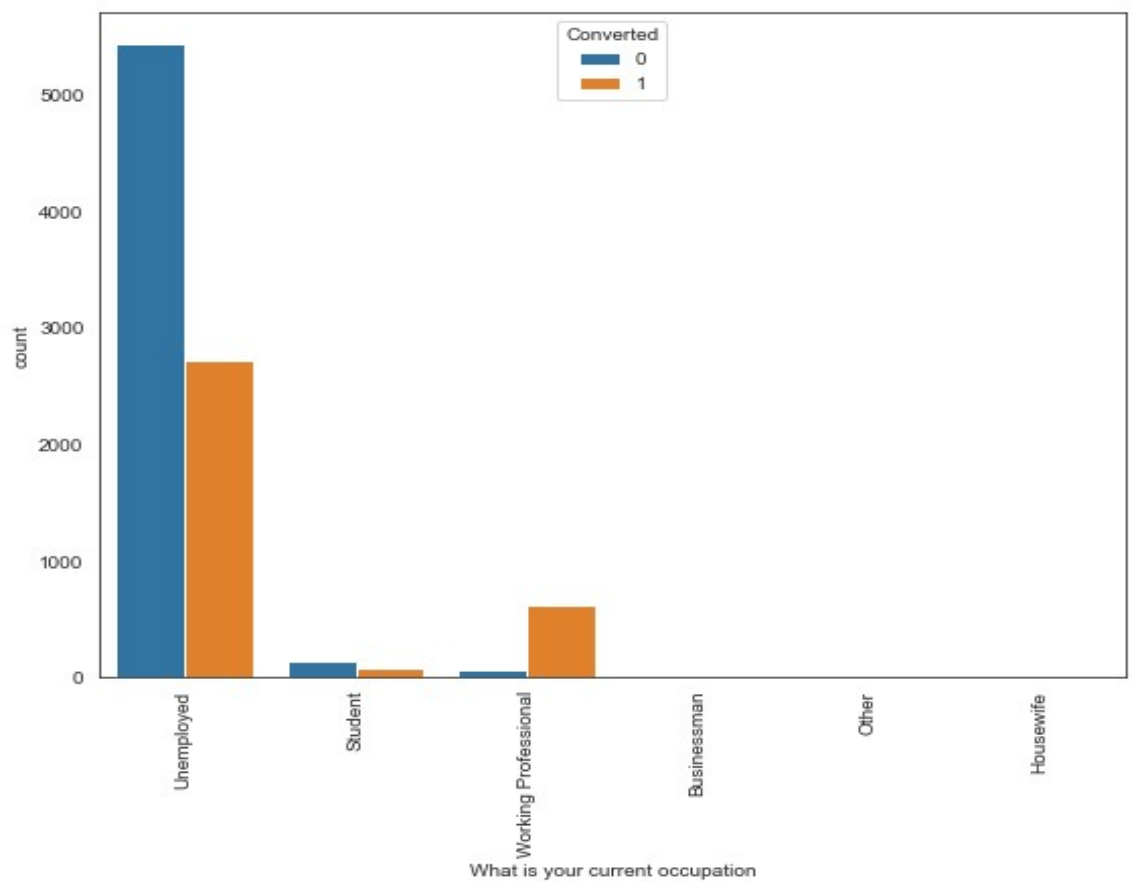
Lead Source

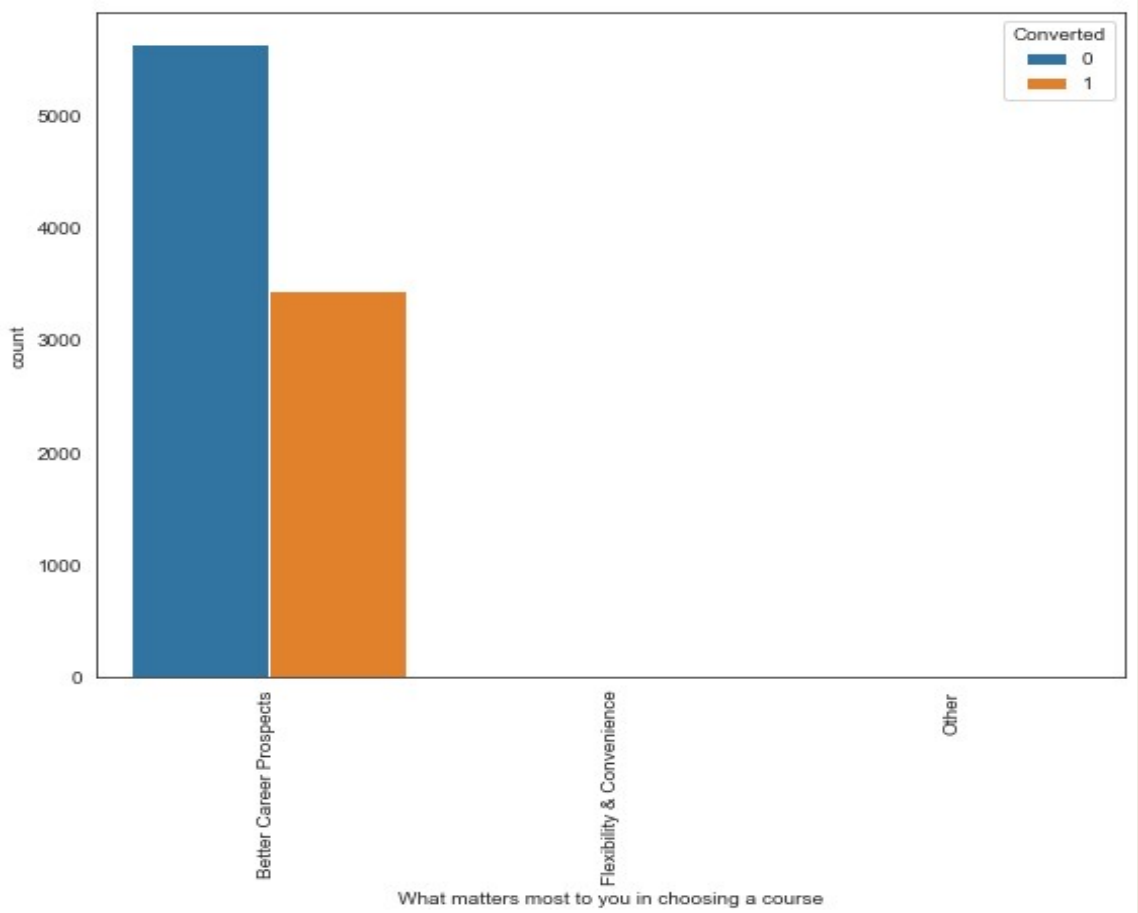


Very high conversion rates for lead sources '**Reference**' and '**Welingak Website'.**

Most leads are generated through **'DirectTraffic**'and '**Google**'.

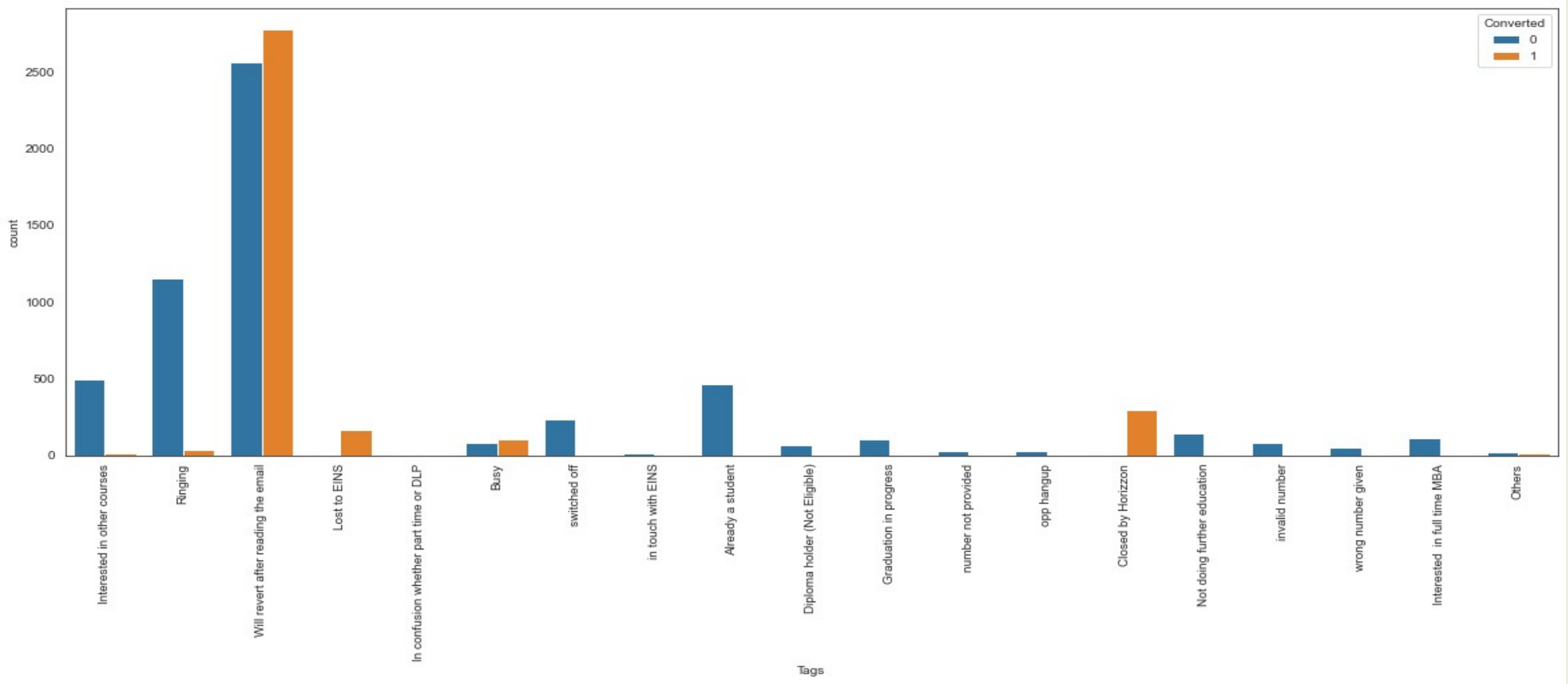
Current Occupation



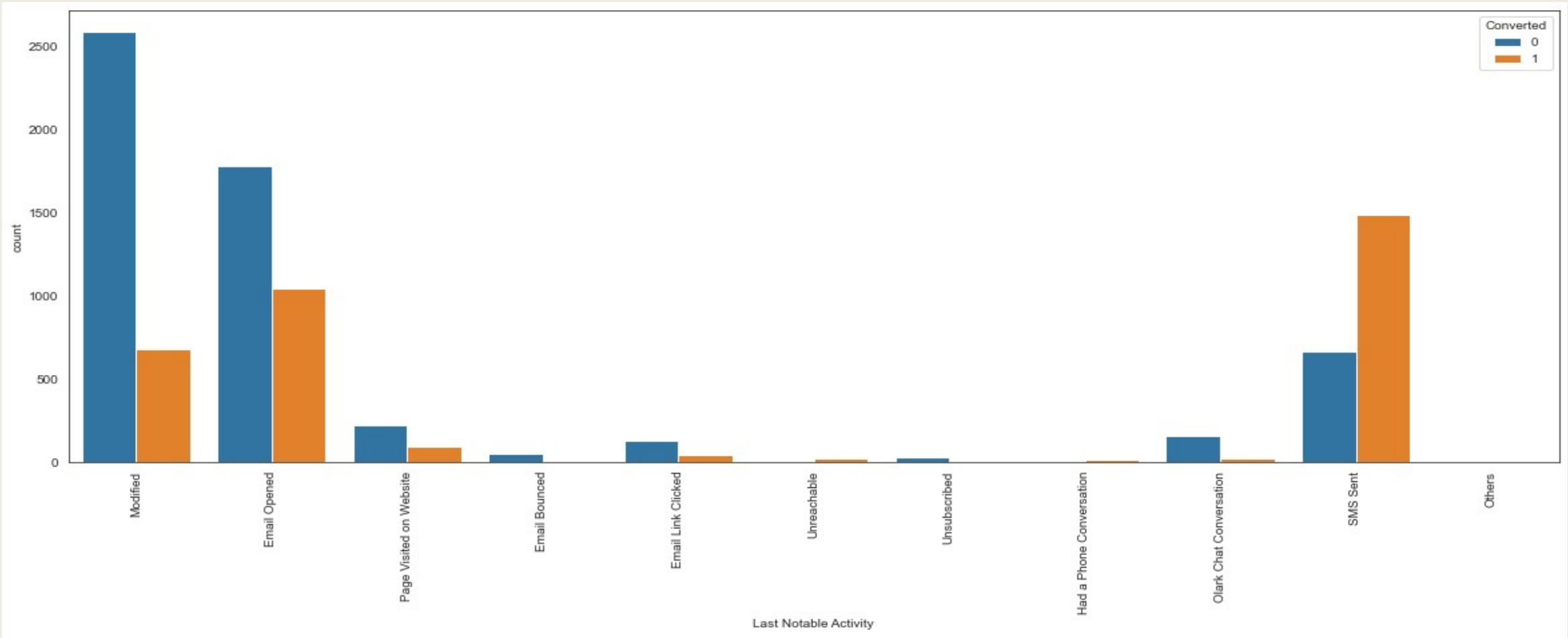


Working Professionals are most likelyto get converted.

Tags


Converted
0
1
count
2500
2000
1500
1000
500
0
Interested in other courses
Ringing
Will revert after reading the email
Lost to EINS
In confusion whether part time or DLP
Busy
switched off
in touch with EINS
Already a student
Diploma holder (Not Eligible)
Graduation in progress
number not provided
opp hangup
Closed by Horizzon
Not doing further education
invalid number
wrong number given
Interested in full time MBA
Others
Tags

High conversion rates for tags '**Willrevertafterreading theemail**', '**Closedby Horizon**', '**LosttoEINS**', and '**Busy**'.



Last Notable Activity

Highest conversion rate is for the last notable activity '**SMS Sent**'.

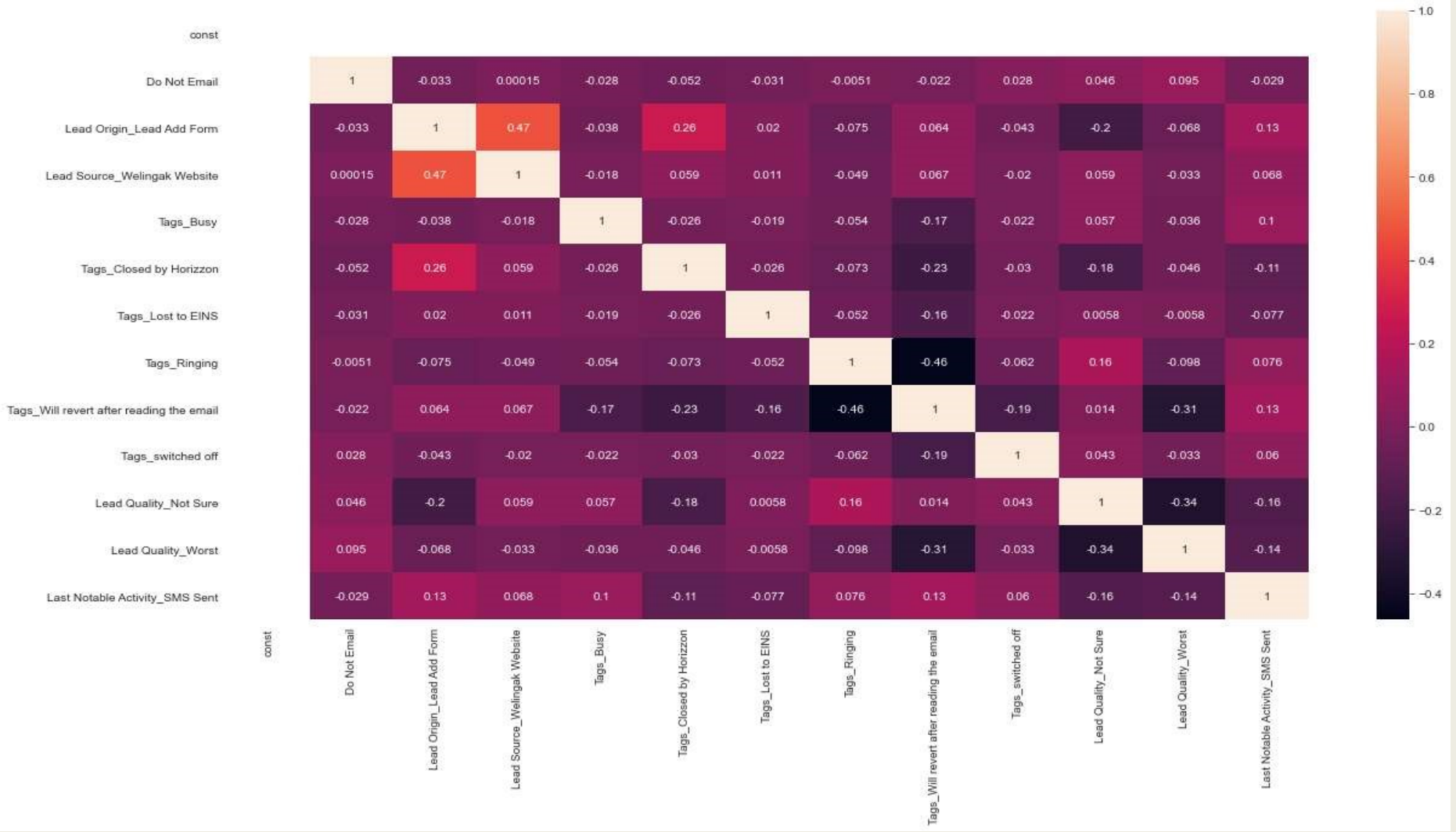
Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6351
Model:	GLM	Df Residuals:	6338
Model Family:	Binomial	Df Model:	12
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1601.0
Date:	Mon, 18 May 2020	Deviance:	3202.0
Time:	02:23:54	Pearson chi2:	3.48e+04
No. Iterations:	8		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.9192	0.211	-9.080	0.000	-2.333	-1.505
Do Not Email	-1.2835	0.212	-6.062	0.000	-1.698	-0.868
Lead Origin_Lead Add Form	1.2035	0.368	3.267	0.001	0.482	1.925
Lead Source_Welingak Website	3.2825	0.820	4.002	0.000	1.675	4.890
Tags_Busy	3.8043	0.330	11.525	0.000	3.157	4.451
Tags_Closed by Horizzon	7.9789	0.762	10.467	0.000	6.485	9.473
Tags_Lost to EINS	9.1948	0.753	12.209	0.000	7.719	10.671
Tags_Ringing	-1.8121	0.336	-5.401	0.000	-2.470	-1.154
Tags_Will revert after reading the email	3.9906	0.228	17.508	0.000	3.544	4.437
Tags_switched off	-2.4456	0.586	-4.171	0.000	-3.595	-1.297
Lead Quality_Not Sure	-3.5218	0.126	-28.036	0.000	-3.768	-3.276
Lead Quality_Worst	-3.9106	0.856	-4.567	0.000	-5.589	-2.232
Last Notable Activity_SMS Sent	2.7395	0.120	22.907	0.000	2.505	2.974

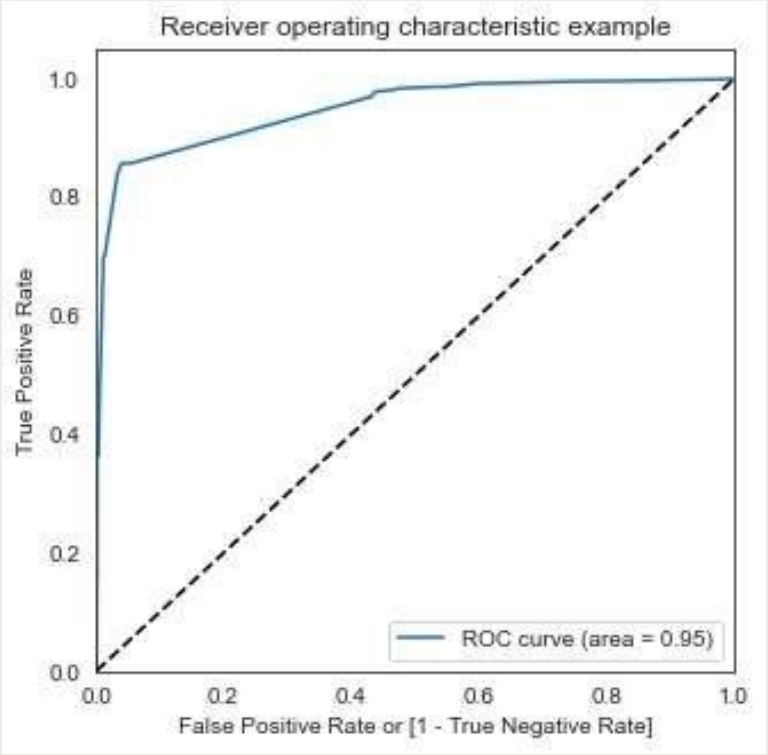
Final Model Summary: All p values are zero.

Heatmap


const
Do Not Email
Lead Origin_Lead Add Form
Lead Source_Welingak Website
Tags_Busy
Tags_Closed by Horizzon
Tags_Lost to EINS
Tags_Ringing
Tags_Will revert after reading the email
Tags_switched off
Lead Quality_Not Sure
Lead Quality_Worst
Last Notable Activity_SMS Sent
1 -0.033 0.00015 -0.028 -0.052 -0.031 -0.0051 -0.022 0.028 0.046 0.095 -0.029
-0.033 1 0.47 -0.038 0.26 0.02 -0.075 0.064 -0.043 -0.2 -0.068 0.13
0.00015 0.47 1 -0.018 0.059 0.011 -0.049 0.067 -0.02 0.059 -0.033 0.068
-0.028 -0.038 -0.018 1 -0.026 -0.019 -0.054 -0.17 -0.022 0.057 -0.036 0.1
-0.052 0.26 0.059 -0.026 1 -0.026 -0.073 -0.23 -0.03 -0.18 -0.046 -0.11
-0.031 0.02 0.011 -0.019 -0.026 1 -0.052 -0.16 -0.022 0.0058 -0.0058 -0.077
-0.0051 -0.075 -0.049 -0.054 -0.073 -0.052 1 -0.46 -0.062 0.16 -0.098 0.076
-0.022 0.064 0.067 -0.17 -0.23 -0.16 -0.46 1 -0.19 0.014 -0.31 0.13
0.028 -0.043 -0.02 -0.022 -0.03 -0.022 -0.062 -0.19 1 0.043 -0.033 0.06
0.046 -0.2 0.059 0.057 -0.18 0.0058 0.16 0.014 0.043 1 -0.34 -0.16
0.095 -0.068 -0.033 -0.036 -0.046 -0.0058 -0.098 -0.31 -0.033 -0.34 1 -0.14
-0.029 0.13 0.068 0.1 -0.11 -0.077 0.076 0.13 0.06 -0.16 -0.14 1
const
Do Not Email
Lead Origin_Lead Add Form
Lead Source_Welingak Website
Tags_Busy
Tags_Closed by Horizzon
Tags_Lost to EINS
Tags_Ringing
Tags_Will revert after reading the email
Tags_switched off
Lead Quality_Not Sure
Lead Quality_Worst
Last Notable Activity_SMS Sent
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

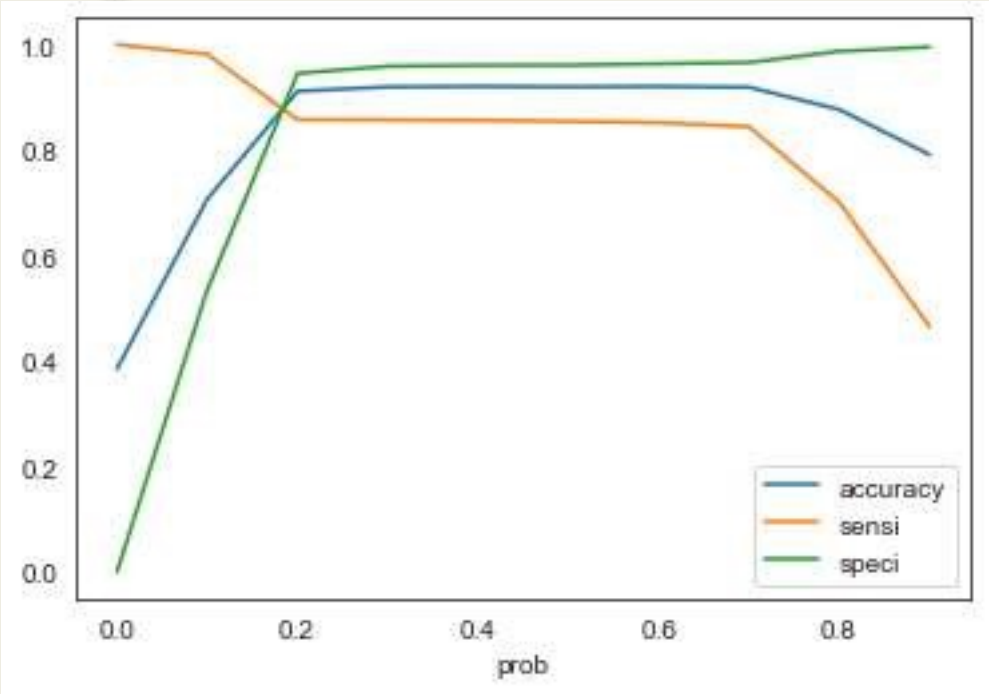
ROC curve

Correlationsbetween features in the final model are **negligible.**



Area undercurve= 0.95

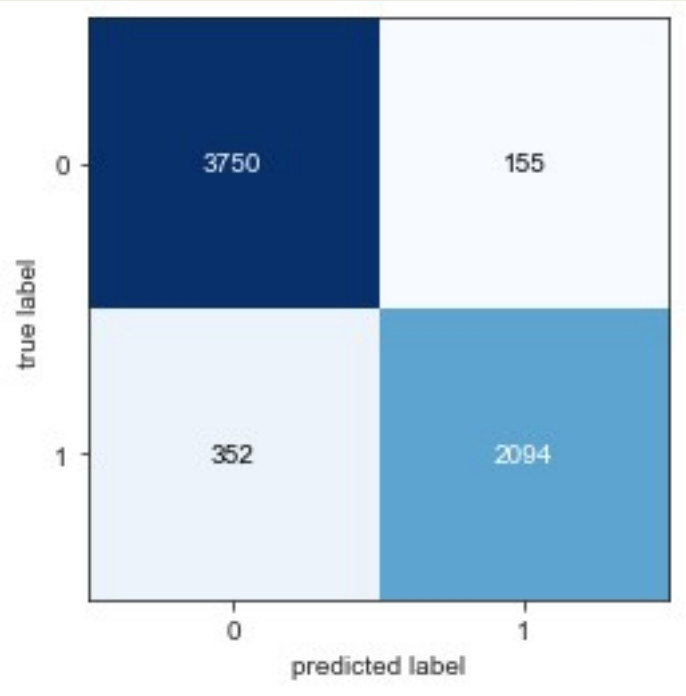
Finding Optimal Threshold



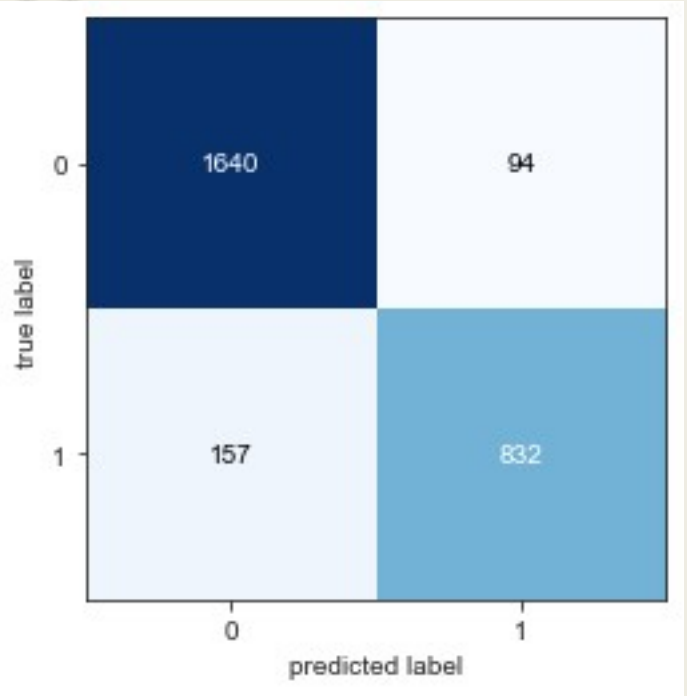
Graph showing changes in Sensitivity, Specificity and Accuracy with changes in the probability threshold values

Optimalcutoff= 0.20

Confusion Matrix Final Results



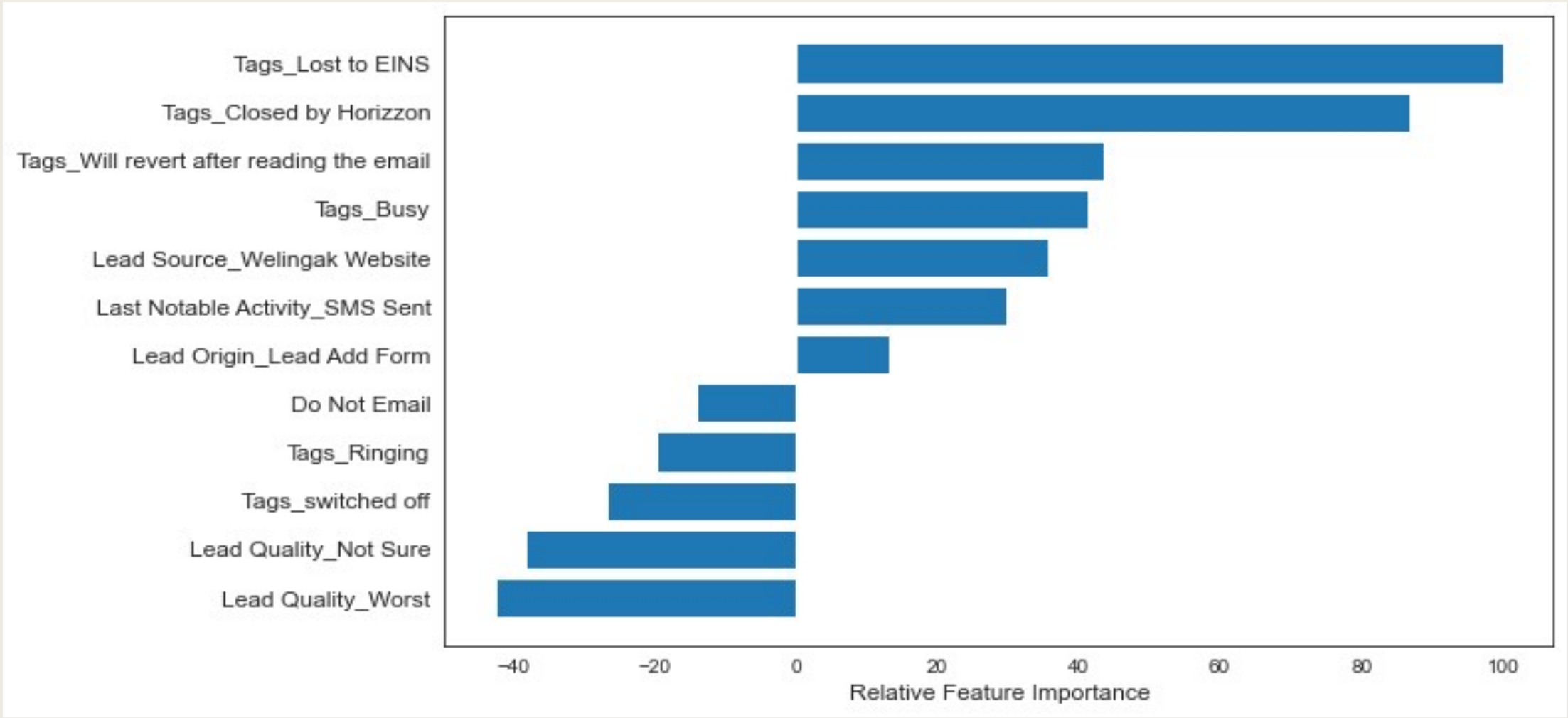
For train set



For test set

Data	Train set	Test set
Accuracy	**0.9111**	**0.9078**
Sensitivity	0.8573	0.8412
Specificity	0.9449	0.9457
FalsePositiveRate	0.0550	0.0542
PositivePredictiveValue	0.9070	0.8984
Negative PredictiveValue	0.9135	0.9126
AUC	0.9488	0.9388

Relative Importance Of Features


Tags_Lost to EINS
Tags_Closed by Horizzon
Tags_Will revert after reading the email
Tags_Busy
Lead Source_Welingak Website
Last Notable Activity_SMS Sent
Lead Origin_Lead Add Form
Do Not Email
Tags_Ringing
Tags_switched off
Lead Quality_Not Sure
Lead Quality_Worst
−40
−20
0
20
40
60
80
100
Relative Feature Importance

Feature Importance

- Three variables which contribute most towards the probability of a lead conversion in decreasing order of impact are:
    1. ***Tags_Lost toEINS***
    2. **Tags_ClosedbyHorizzon**
    3. **Tags_Willrevert afterreadingthe email**
- These are dummy features created from the categorical variable Tags.
- All three **contribute positively** towards the probability of a lead conversion.
- These results indicate that the company should **focusmoreonthe leads withthesethree tags**.

Situation1: Company has interns for 2 months. They wish to make lead conversion more aggressive. They want almost all of the potential leads to be converted and hence, want to make phone calls to as much of such people as possible.

Solution:

- $\boldsymbol{Sensitivity = TruePositives / (TruePositives + FalseNegatives)}$
- Sensitivity can be defined as the number of actual conversions predicted correctly out of total number of actual conversions. As we saw earlier, sensitivity decreases as the threshold increases.
- High sensitivity implies that our model will correctly predict almost all leads who are likely to convert. At the same time, it may overestimate and misclassify some of the non-conversions as conversions.

- As the company has extra man-power for two months and wants to make the lead conversion more aggressive, it is a good strategy to go for **highsensitivity**.
  To achieve high sensitivity, we need to **choosea lowthresholdvalue**.

Situation2: At times, the company reaches its target for a quarter before the deadline. It wants the sales team to focus on some new work. So during this time, the company's aim is to not make phone calls unless it's extremely necessary.

Solution:

- $\boldsymbol{Specificity = TrueNegatives/\ (TrueNegatives + FalsePositives)}$
- Specificity can be defined as the number of actual non-conversions predicted correctly out of total number of actual non-conversions. It increases as the threshold increases.
- High specificity implies that our model will correctly predict almost all leads who are not likely to convert. At the same time, it may misclassify some of the conversions as non-conversions.

- As the company has already reached its target for a quarter and doesn't want to make unnecessary phone calls, it is a good strategy to go for **highspecificity**.
- It will ensure that the phone calls are only made to customers who have a very high probability of conversion. To achieve high specificity, we need to **choose a highthresholdvalue**.

Recommendations

- By referring to the data visualizations, focus on
  - *Increasing the conversion rates for the categories generating more leads and*
  - *Generating more leads for categories having high conversion rates.*

- Pay attention to the relative importance of the features in the model and their positive or negative impact on the probability of conversion.
- Based on varying business needs, modify the probability threshold value for identifying potential leads.