Lead Scoring Case Study

Smita Baruah & Ayan Chattaraj

Problem Statement

An education company named X Education sells online courses to industry professionals. 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.

Our Objective

Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

In order to achieve the above objective our model sensitivity should be high(>70%)

1. Approach

- ➔ Inspecting the Dataset
- ➔ Checking for Data Imbalance
- ➔ Exploratory Data Analysis
- ➔ Variable Transformation for Modelling
- ➔ Train and test split and feature scaling
- ➔ Model Building
- ➔ Checking performance of the Model
- ➔ Recommendation and conclusion

Inspecting the Dataset

- Size: 9240 rows and 37 columns
- Data Types: object, int64 and float64
- 7 numerical columns and 30 categorical columns
- Missing values Present: Yes

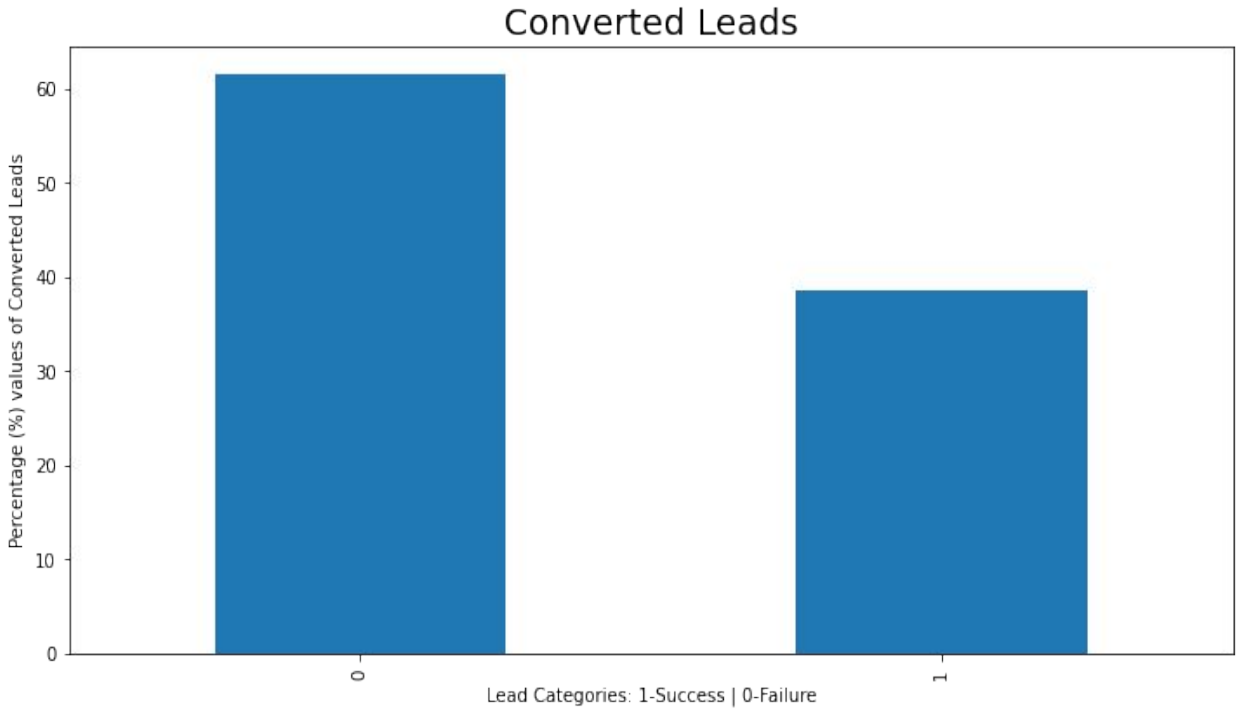
Missing Value Handling

Based on the quantum of missing values, columns with values greater than 40% missing were dropped.

```
Prospect ID                                    0.00
Lead Number                                    0.00
Lead Origin                                    0.00
Lead Source                                    0.39
Do Not Email                                   0.00
Do Not Call                                    0.00
Converted                                      0.00
TotalVisits                                    1.48
Total Time Spent on Website                    0.00
Page Views Per Visit                           1.48
Last Activity                                  1.11
Country                                       26.63
Specialization                                36.58
How did you hear about X Education            78.46
What is your current occupation               29.11
What matters most to you in choosing a course 29.32
Search                                         0.00
Magazine                                       0.00
Newspaper Article                              0.00
X Education Forums                             0.00
Newspaper                                      0.00
Digital Advertisement                          0.00
Through Recommendations                        0.00
Receive More Updates About Our Courses         0.00
Tags                                          36.29
Lead Quality                                  51.59
Update me on Supply Chain Content              0.00
Get updates on DM Content                      0.00
Lead Profile                                  74.19
City                                          39.71
Asymmetrique Activity Index                   45.65
Asymmetrique Profile Index                    45.65
Asymmetrique Activity Score                   45.65
Asymmetrique Profile Score                    45.65
I agree to pay the amount through cheque       0.00
A free copy of Mastering The Interview         0.00
Last Notable Activity                          0.00
dtype: float64
```

Checking for Data Imbalance

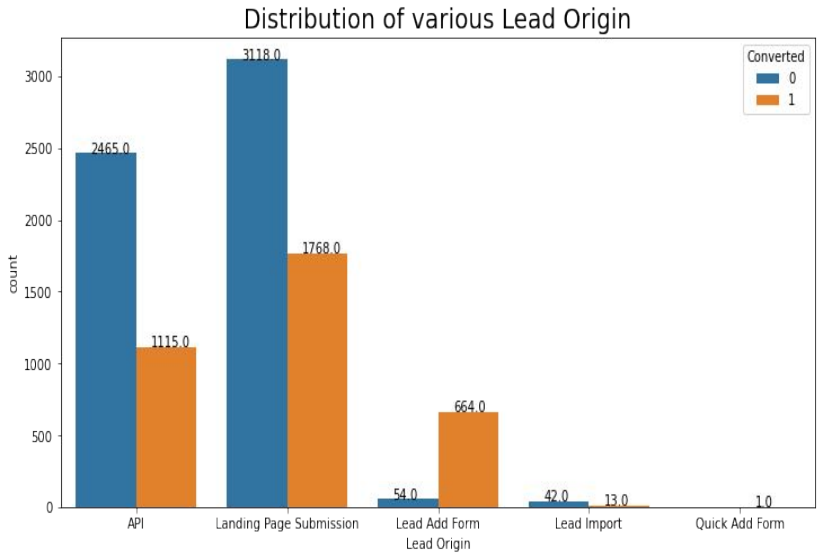
Our dataset constitutes of 38% successfully converted leads data, rest i.e. 62% are not converted data.

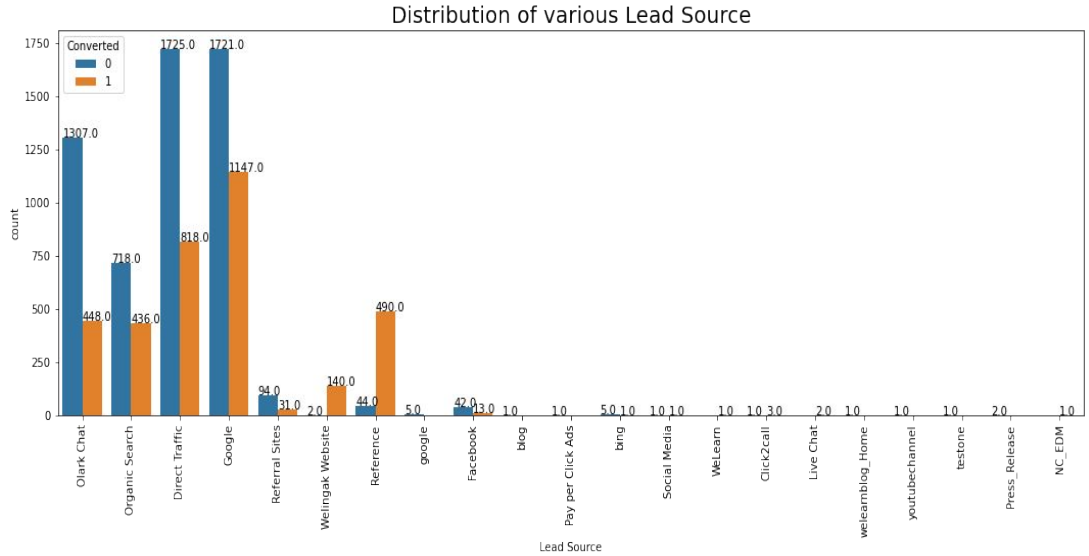


Data Wrangling

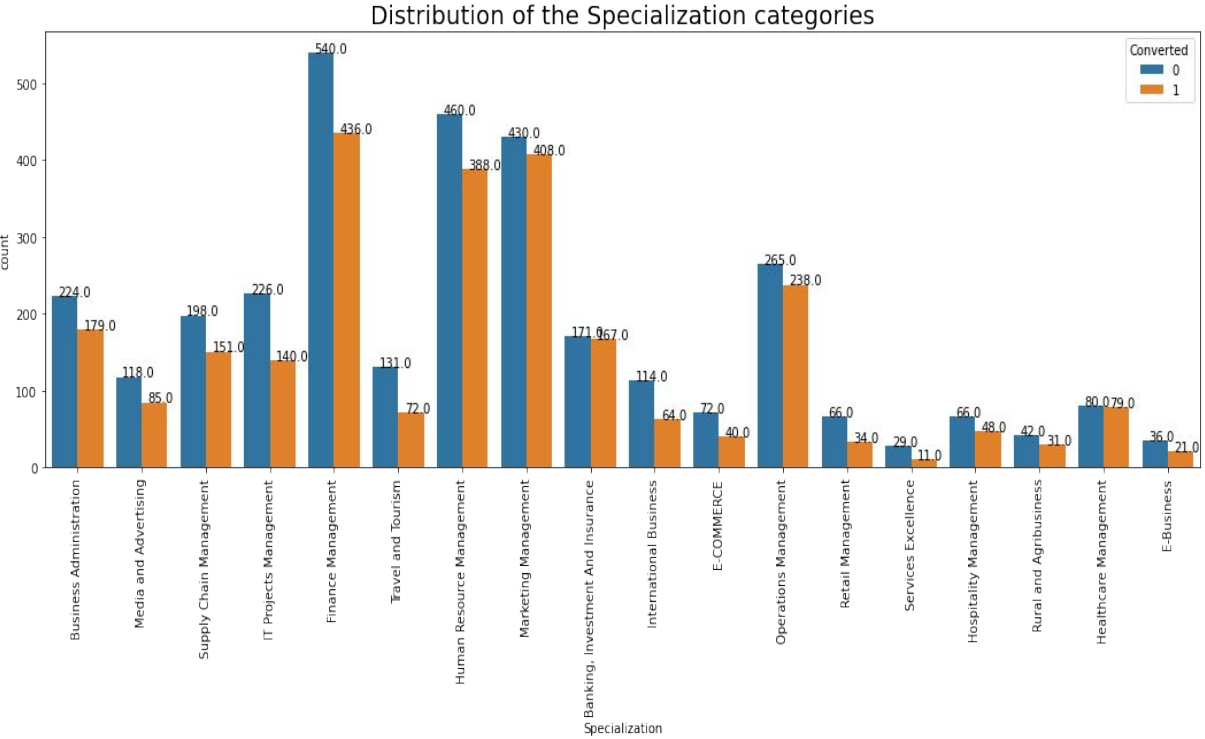
- EDA on Categorical and Numerical columns
- Outlier Handling
- Missing value Imputation

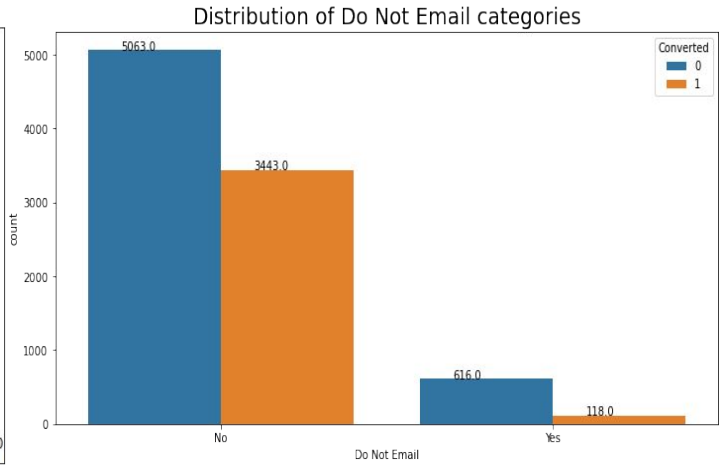
EDA on categorical variables



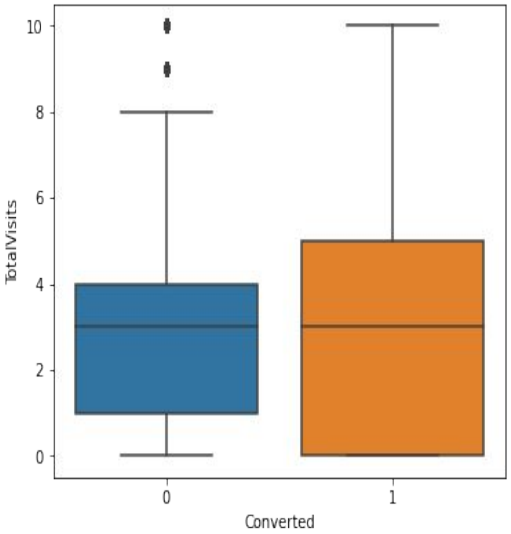


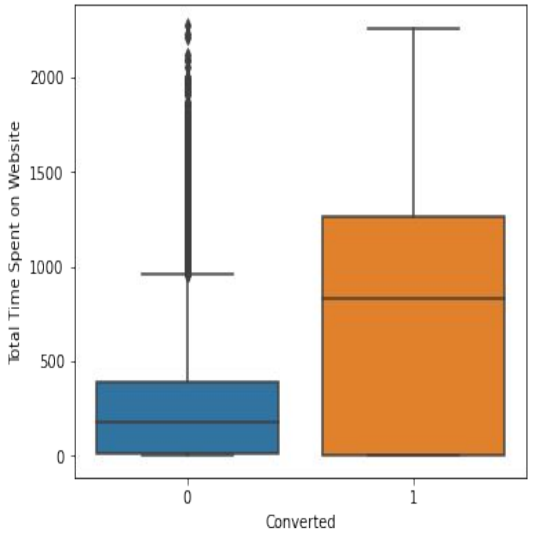
EDA on categorical variables

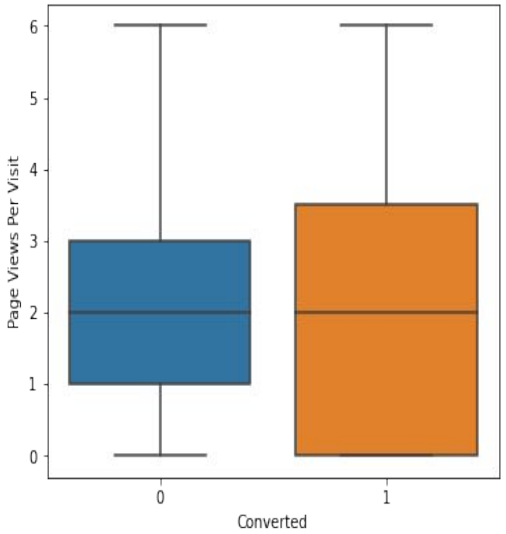




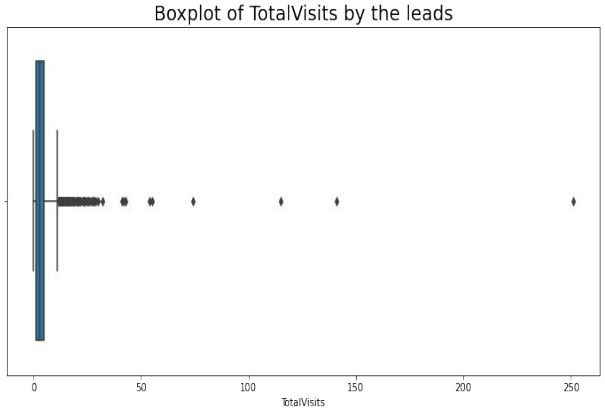
EDA on numeric variables

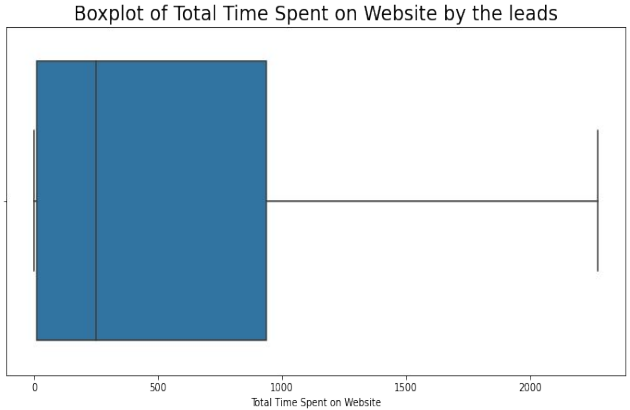


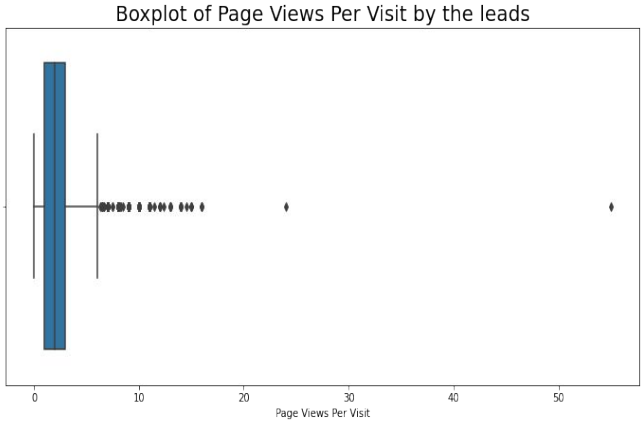




Outliers







The outliers in features 'Total Visits', 'Page Views Per Visit' have outliers and were handled by capping at 0.95 th quantiles

Final Model with p-values and VIF

Generalized Linear Model Regression Results

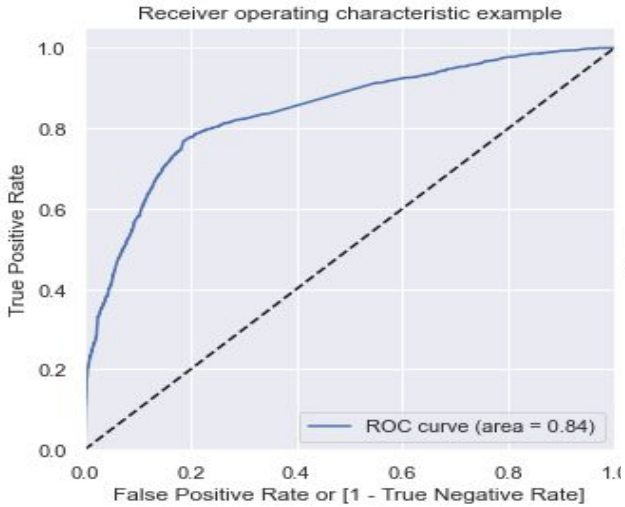
Dep. Variable:	Converted	**No. Observations:**	6468
Model:	GLM	**Df Residuals:**	6456
Model Family:	Binomial	**Df Model:**	11
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-3019.3
Date:	Tue, 15 Mar 2022	**Deviance:**	6038.5
Time:	09:51:38	**Pearson chi2:**	7.89e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

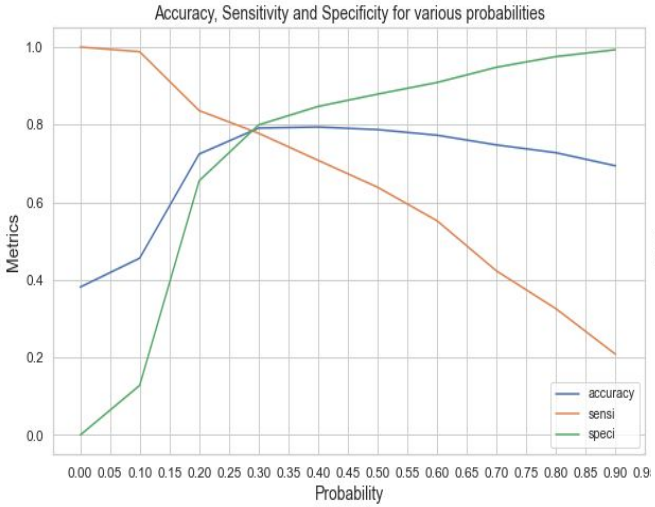
	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.3622	0.075	-4.839	0.000	-0.509	-0.215
Do Not Email	-1.3757	0.158	-8.724	0.000	-1.685	-1.067
Total Time Spent on Website	1.0961	0.037	29.578	0.000	1.023	1.169
Lead Origin_Lead Add Form	3.0726	0.188	16.305	0.000	2.703	3.442
Lead Source_Direct Traffic	-0.9243	0.105	-8.809	0.000	-1.130	-0.719
Lead Source_Google	-0.5815	0.098	-5.936	0.000	-0.773	-0.389
Lead Source_Organic Search	-0.6770	0.121	-5.581	0.000	-0.915	-0.439
Lead Source_Referral Sites	-1.1884	0.302	-3.935	0.000	-1.780	-0.596
Lead Source_Welingak Website	2.1284	0.741	2.872	0.004	0.676	3.581
Specialization_IT Projects Management	0.3929	0.169	2.324	0.020	0.062	0.724
Specialization_Rural and Agribusiness	0.7562	0.367	2.060	0.039	0.037	1.476
What is your current occupation_Working Professional	2.8691	0.180	15.924	0.000	2.516	3.222

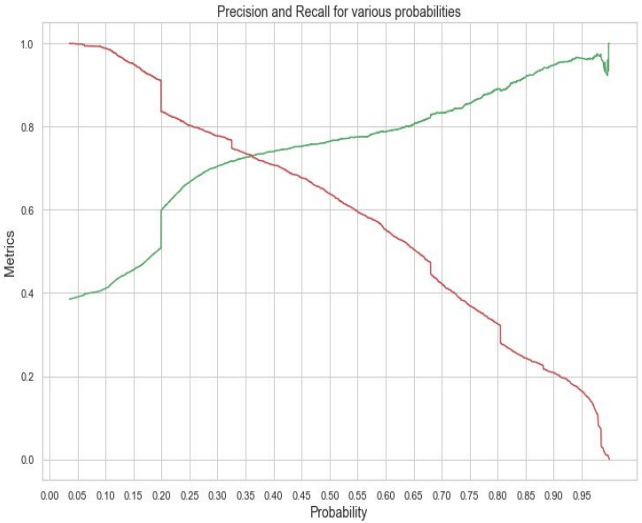
	Features	VIF
2	Lead Origin_Lead Add Form	1.36
7	Lead Source_Welingak Website	1.24
10	What is your current occupation_Working Profes...	1.16
3	Lead Source_Direct Traffic	1.14
0	Do Not Email	1.11
1	Total Time Spent on Website	1.09
4	Lead Source_Google	1.08
8	Specialization_IT Projects Management	1.07
5	Lead Source_Organic Search	1.04
9	Specialization_Rural and Agribusiness	1.01
6	Lead Source_Referral Sites	1.00

In final Model all p-values were < 0.05 and VIF were <5

Model Evaluation

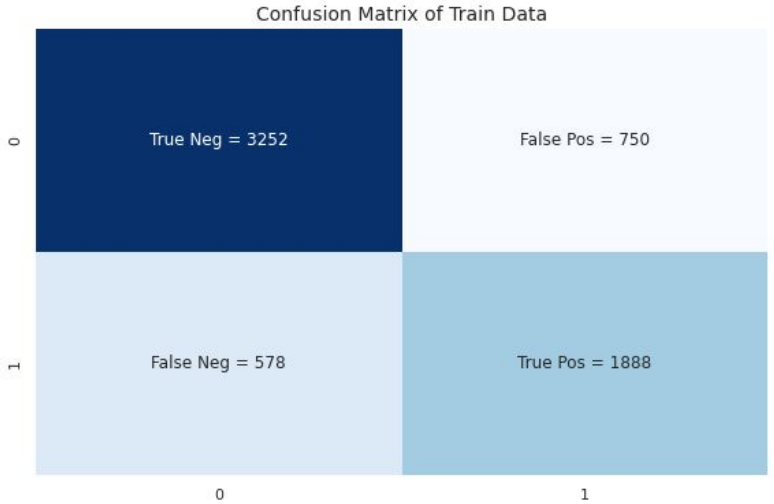


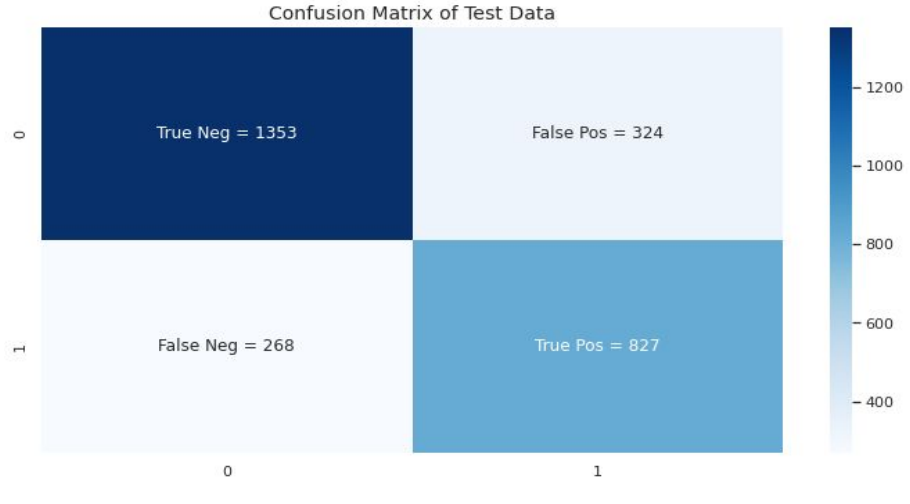




Based on these information threshold of 0.29 was chosen for model prediction

Confusion matrix and metrics





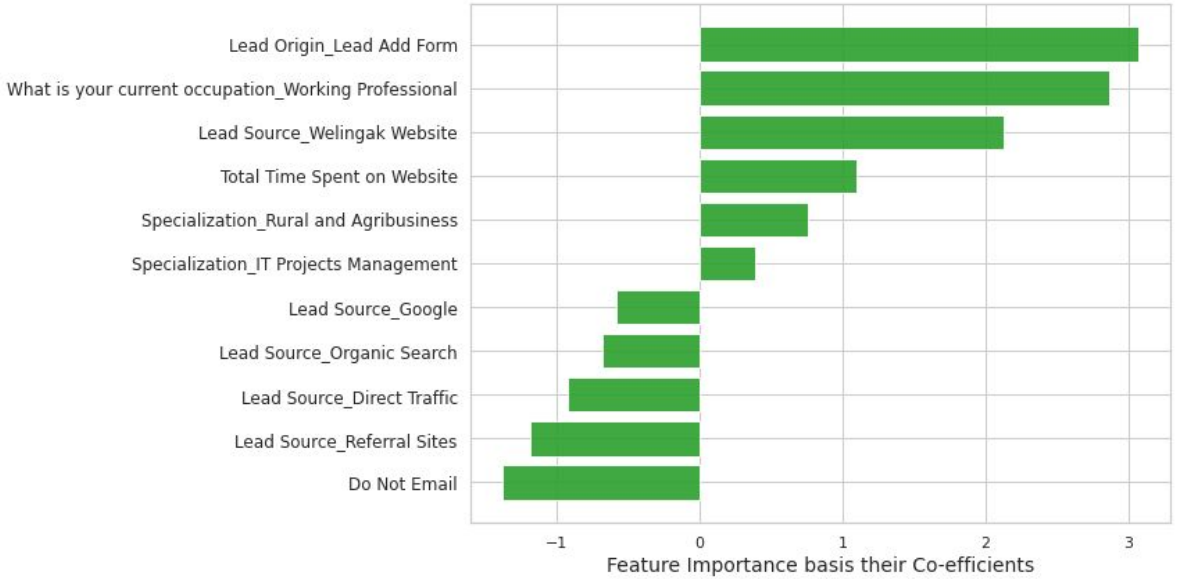
The Evaluation Metrics for the train Dataset:

- Accuracy is : 0.79
- Sensitivity is : 0.76
- Specificity is : 0.81
- Precision is : 0.71
- f1 score is : 0.73

The Evaluation Metrics for the test Dataset:

- Accuracy is : 0.78
- Sensitivity is : 0.75
- Specificity is : 0.80
- Precision is : 0.71
- f1 score is : 0.73

Recommendation and Conclusion



In order for the X-Education to achieve its target , focus should be on the leads coming from the below sources:-:

- The Lead origin as Lead Add form
- The Current Occupation as Working Professional
- The Lead Source coming from Welingak Website
- Spending lot of time on the website
- With specialization either Rural and Agri Business or IT Projects Management

Thank You