hictior
onsectem
ue il

h / Statistic 2

lacius exerspit, nes excescient hictior
ute liquam voles aliscit inulpa consectem
fuga. Edia volum volupit incillique il
t Hitem quamoccu

7 survey

Graph / Statistic 3

Sequidel lacius exerspit, nes excescient hictior iatemquid ute liquam voles aliscit inulpa consectem aut a corem fuga. Edia volum volupit incillique il nihilicillo vent Hitem quamoccu

according to 2017 survey

NEXT PAGE
Analysis of Net Funds

Lead Scoring Case Study

Problem Statement

- An education company named X Education sells online courses to industry professionals. On any given day, a large number of professionals interested in the courses visit their website and look up courses. These individuals are referred to as leads.
- After obtaining these leads, sales team members begin calling, sending emails, etc. Some of the leads are converted through this method, but the majority are not. Now, although X Education gets a lot of leads, its lead conversion rate is very poor.

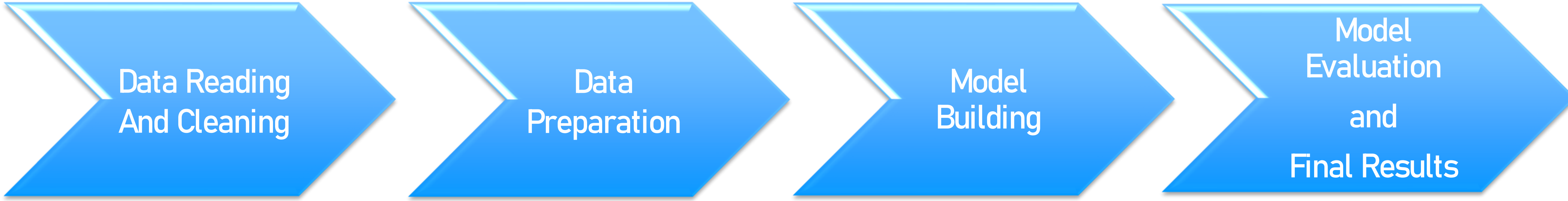
Business Objective

- X Education has asked us to assist them in identifying the most promising leads, or those who are most likely to convert into paying clients.

- The company expects us to create a model in which we give a lead score to each lead so that customers with higher lead scores have a higher conversion chance and customers with lower lead scores have a lower conversion chance.

Social Media Marketing
Initial Pool of Leads
Lead Nurturing
Hot Leads
£ € € $
Converted Leads

Overall Approach



Data Reading And Cleaning

- ✓ Read and Understand data
- ✓ Drop or impute outliers and missing values
- ✓ Exploratory data analysis

Data Preparation

- ✓ Creating a dummy variable
- ✓ Splitting the data into train and test dataset
- ✓ Feature scaling of numerical variables

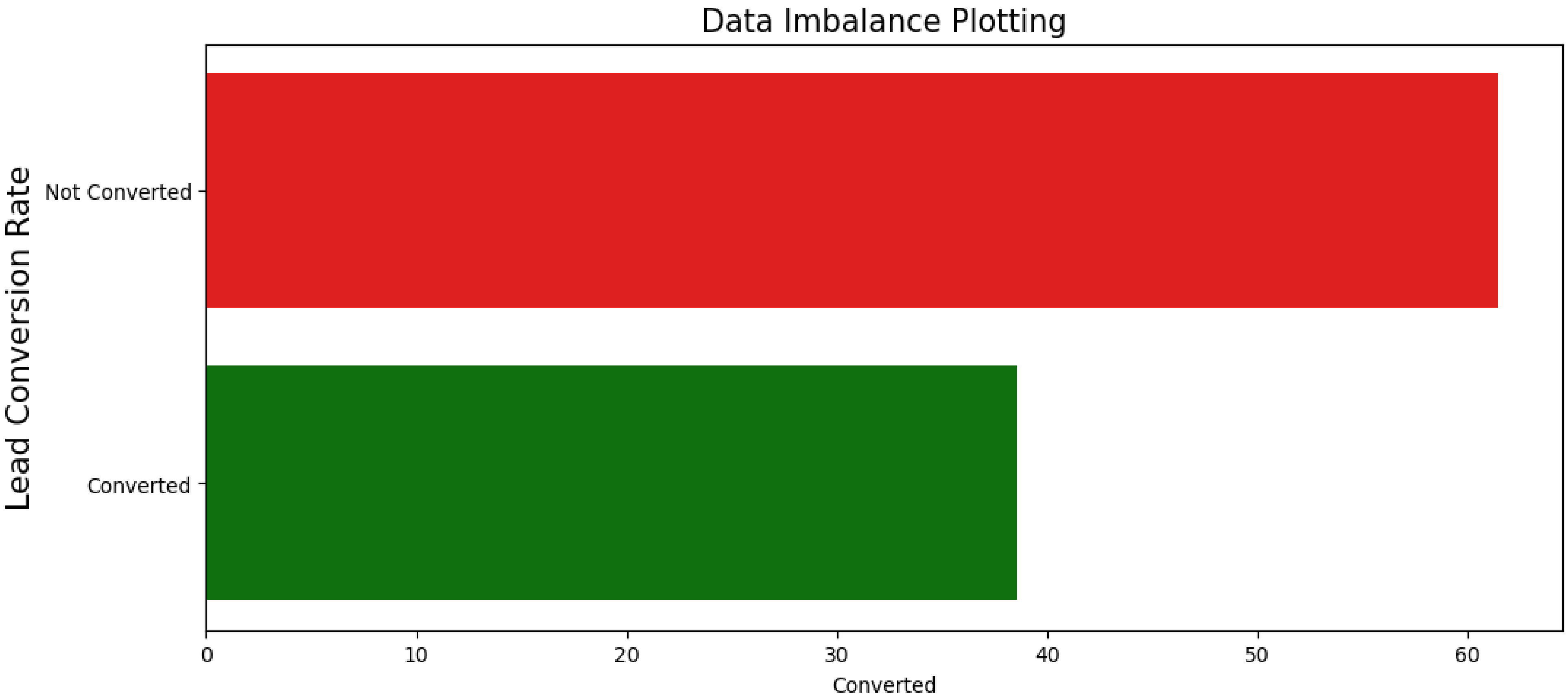
Model Building

- ✓ Selecting features using RFE, VIF, and P-Values
- ✓ Finding optimal cut-off for prediction
- ✓ Make predictions on the Train dataset

Model Evaluation and Final Results

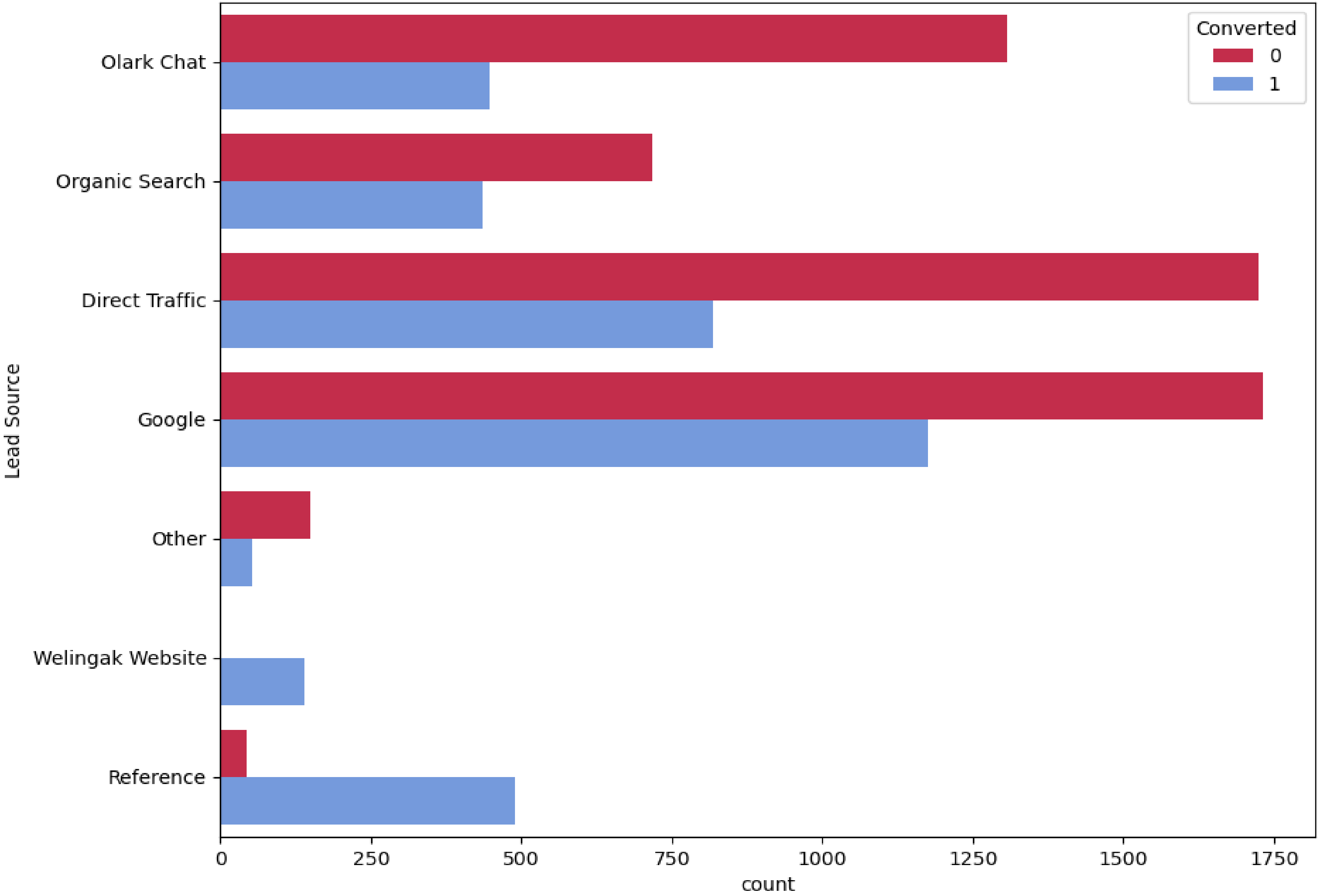
- ✓ Evaluate the final model with different evaluation metrics like Recall, Precision, and Specificity
- ✓ Make predictions on the Test dataset
- ✓ Assign a lead score to all variables in the dataset

Exploratory Data Analysis



A dataset with a class imbalance of 61.46% negative class and 38.54% positive class is considered moderately imbalanced.

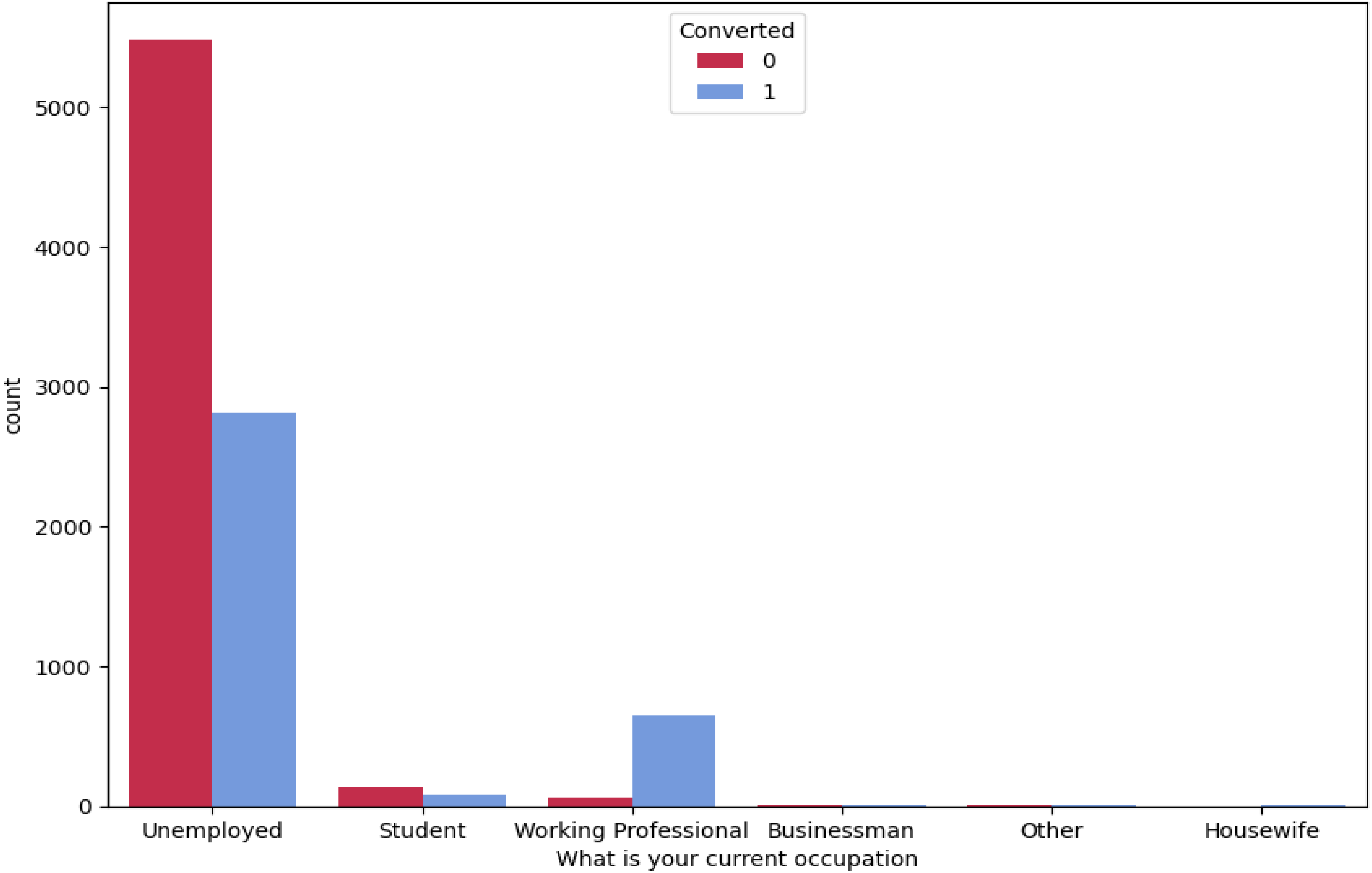
Lead Sources



- From the graph, we can infer that the major lead count is from direct traffic and google.

- Leads from Reference and Welingak Website have the highest conversion rate

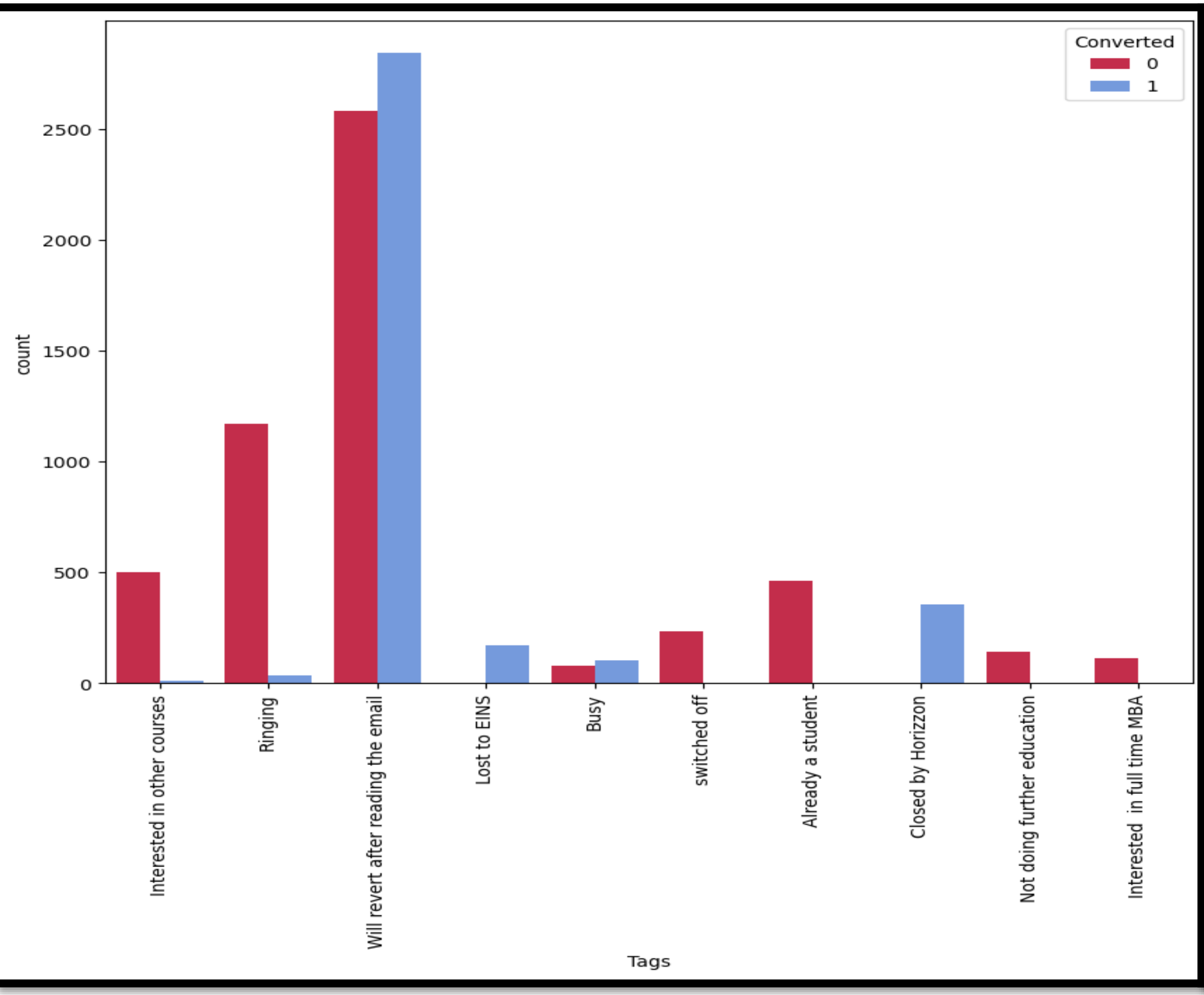
What is your current occupation



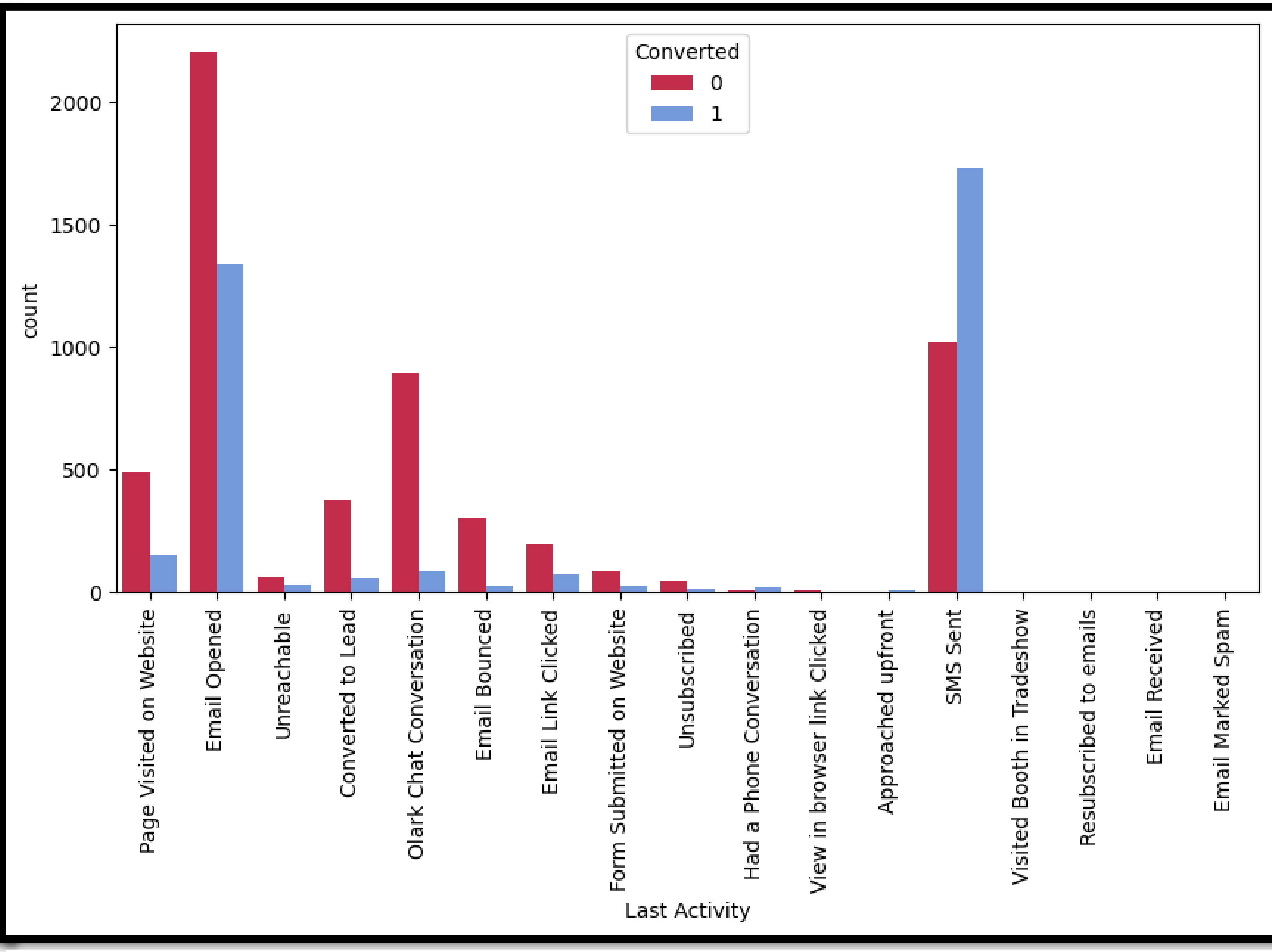
- In this graph we can clearly see that the majority of leads are Unemployed and around 50% of them are converted into a paying customer.

- The category of working professionals have the highest conversion rate

Tags



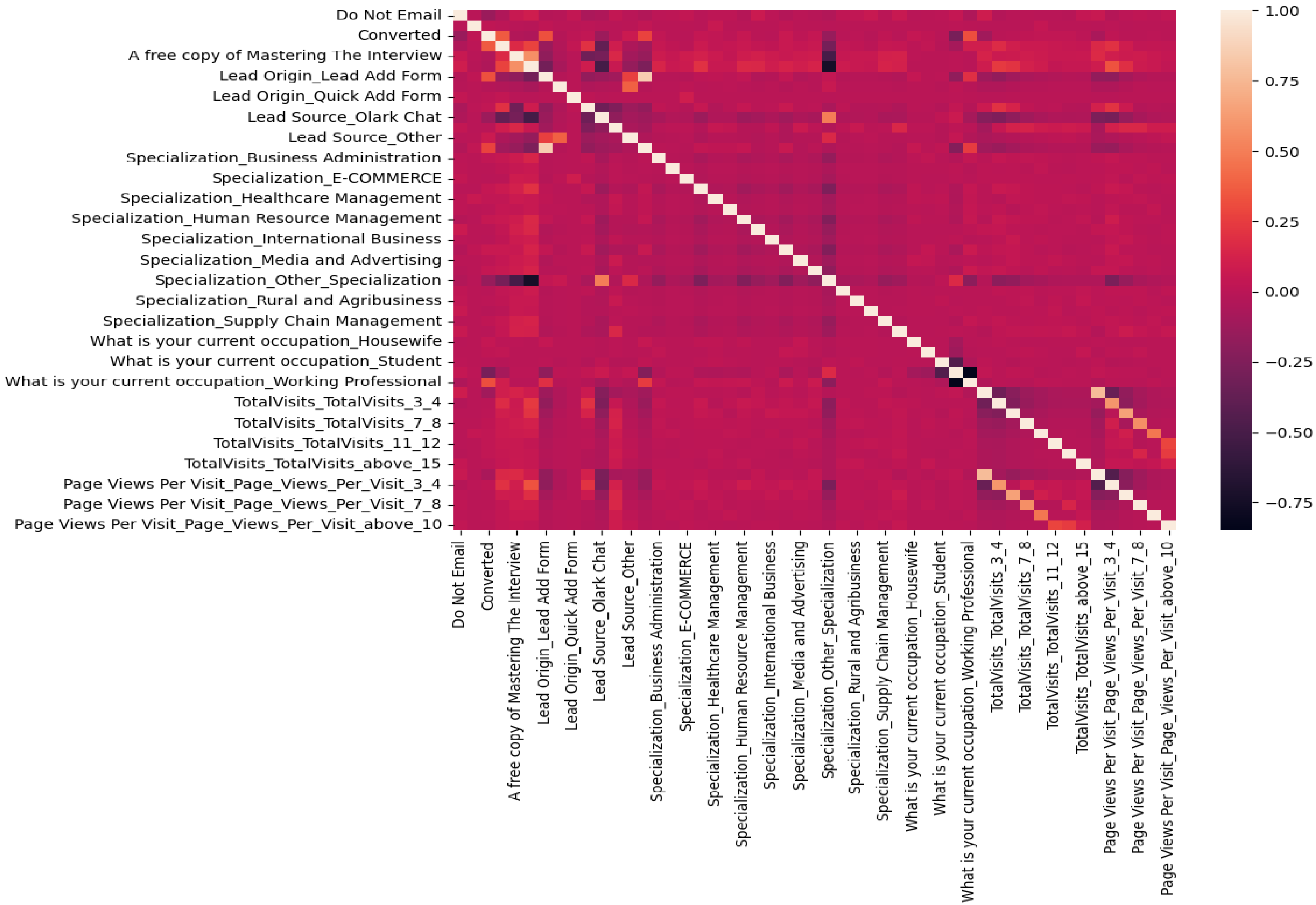
Last Activity



- Majority of leads tagged as ' Will revert after reading the email' are having the highest conversion rate

- We should focus on boosting the conversion rate of those leads with the most recent action as Email Opened, as well as trying to improve the count of those with the most recent activity as SMS Sent.¶

Heatmap To Check The Correlation Among Variables



In this heatmap since there are a lot of variables we couldn't find which features are highly correlated so we proceeded with building our model and based on the p-values and VIFs.

Model Building and Evaluating Process

- Created dummy variables of all the categorical columns
- Splatted the data in the ratio of 70:30 for train and test
- Choose 15 variables with the help of RFE
- Started building the model by eliminating variables whose P-Value is greater than 0.05 and whose VIF is greater than 5.
- Then we made predictions on Train and Test dataset and evaluated the model based on the factors mentioned below.

Train Data Set metrics

Specificity Score: 76.91

Precision Score: 68.14

Recall Score: 80.13

Test Data Set metrics:

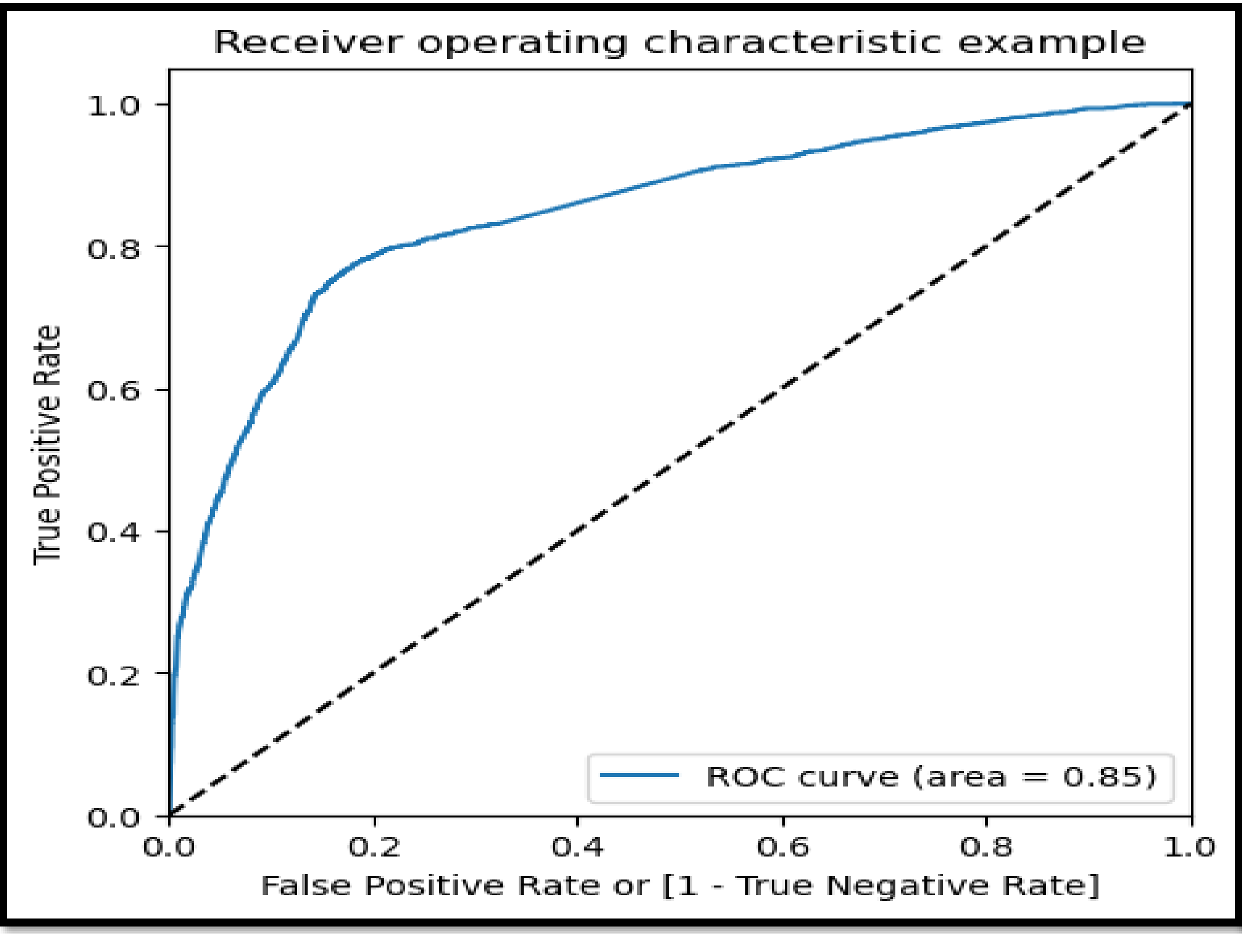
Specificity Score: 74.96

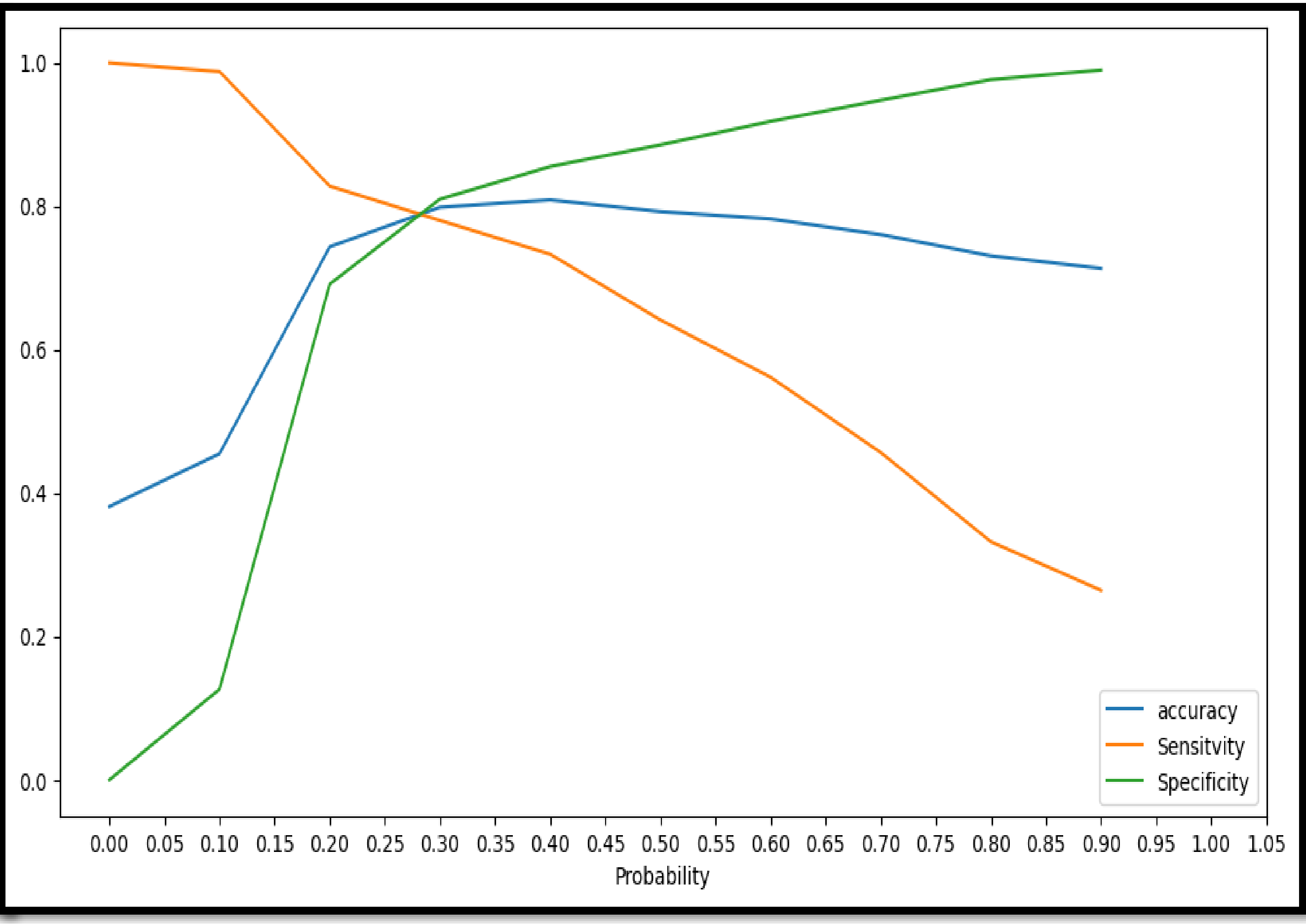
Precision Score: 67.62

Recall Score: 80.09

ROC Curve

Optimal Cut-off curve





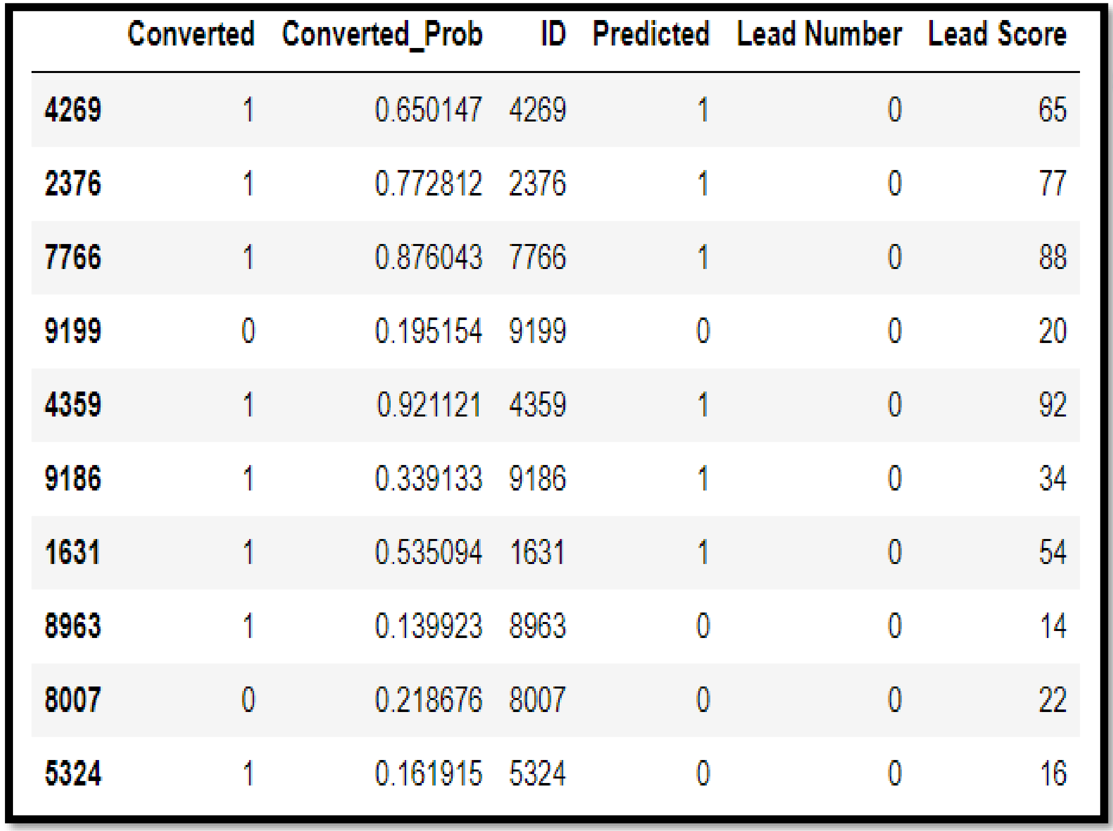
- The area under ROC curve is 0.85 which is close to 1 which indicates the above model can be used for prediction

- From the above graph, it is visible that the optimal cut-off is around 0.25 so we used this value to predict the results

Conclusion

- After analyzing the datasets we came to the conclusion that the model we have created is predicting almost accurately
- This model is flexible in nature which means can be modified to meet the needs of the company in the future.
- To increase the conversion rate sales team should focus on the leads mentioned below

Leads To Be Focused	Leads Not To Be Focused
References	Students
Working Professionals	Unemployed
Leads Tagged as “Will revert after reading the email”	Leads Tagged as “Ringing” or “Switched off”
SMS Sent	



	Converted	Converted_Prob	ID	Predicted	Lead Number	Lead Score
4269	1	0.650147	4269	1	0	65
2376	1	0.772812	2376	1	0	77
7766	1	0.876043	7766	1	0	88
9199	0	0.195154	9199	0	0	20
4359	1	0.921121	4359	1	0	92
9186	1	0.339133	9186	1	0	34
1631	1	0.535094	1631	1	0	54
8963	1	0.139923	8963	0	0	14
8007	0	0.218676	8007	0	0	22
5324	1	0.161915	5324	0	0	16

THANK
YOU