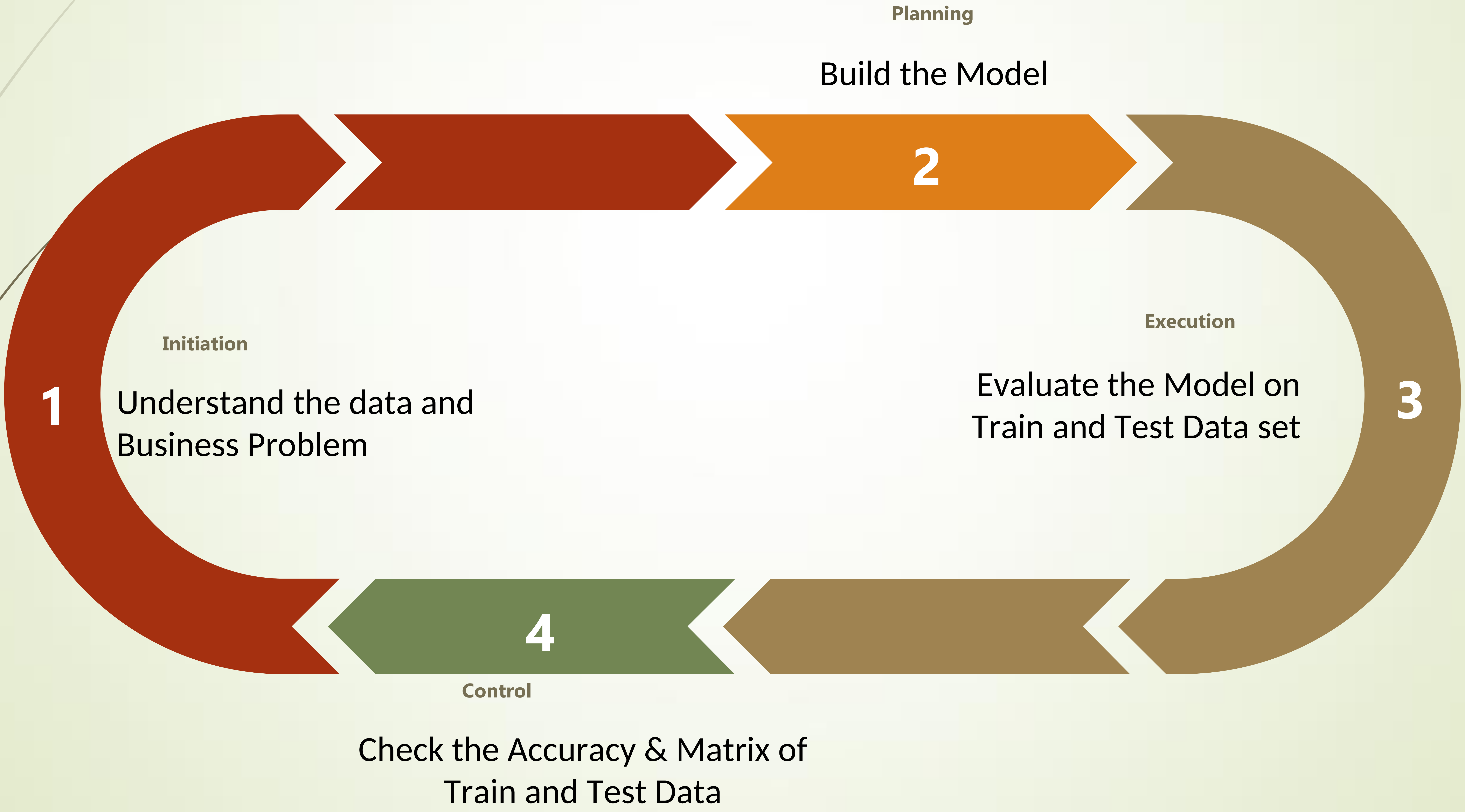
LEAD SCORE CASE STUDY

To Build a Logistic Regression Model that will enable X Education to find out from the list the top Hot convertible leads with the help of lead score



Planning

Build the Model

Initiation

Understand the data and Business Problem

Execution

Evaluate the Model on Train and Test Data set

Control

Check the Accuracy & Matrix of Train and Test Data

Process to Follow

- Create a Logistic Regression Model that will predict the Lead conversion probabilities for each lead.
- To calculate the threshold point that will predict whether a lead will be treated as convert or not convert.
- Multiply the lead conversion Probability to arrive at the Lead score that would be arranged in order of highest lead score to the lowest.

Problem Solving Steps

01 **Reading the Dataset**

Importing necessary Libraries and reading the dataset to work upon

02 **Understanding the data .**

Check for Shape, datatypes and informatics of the data.

03 **EDA**

Dropping columns with more than 45% missing values, replace NAN, Outliers univariate & Bivariate Analysis

04 **Creation of Dummies and Scaling Train Test Split**

After all the adjustments to the data we proceed with Data Split (Train Test and Scaling.

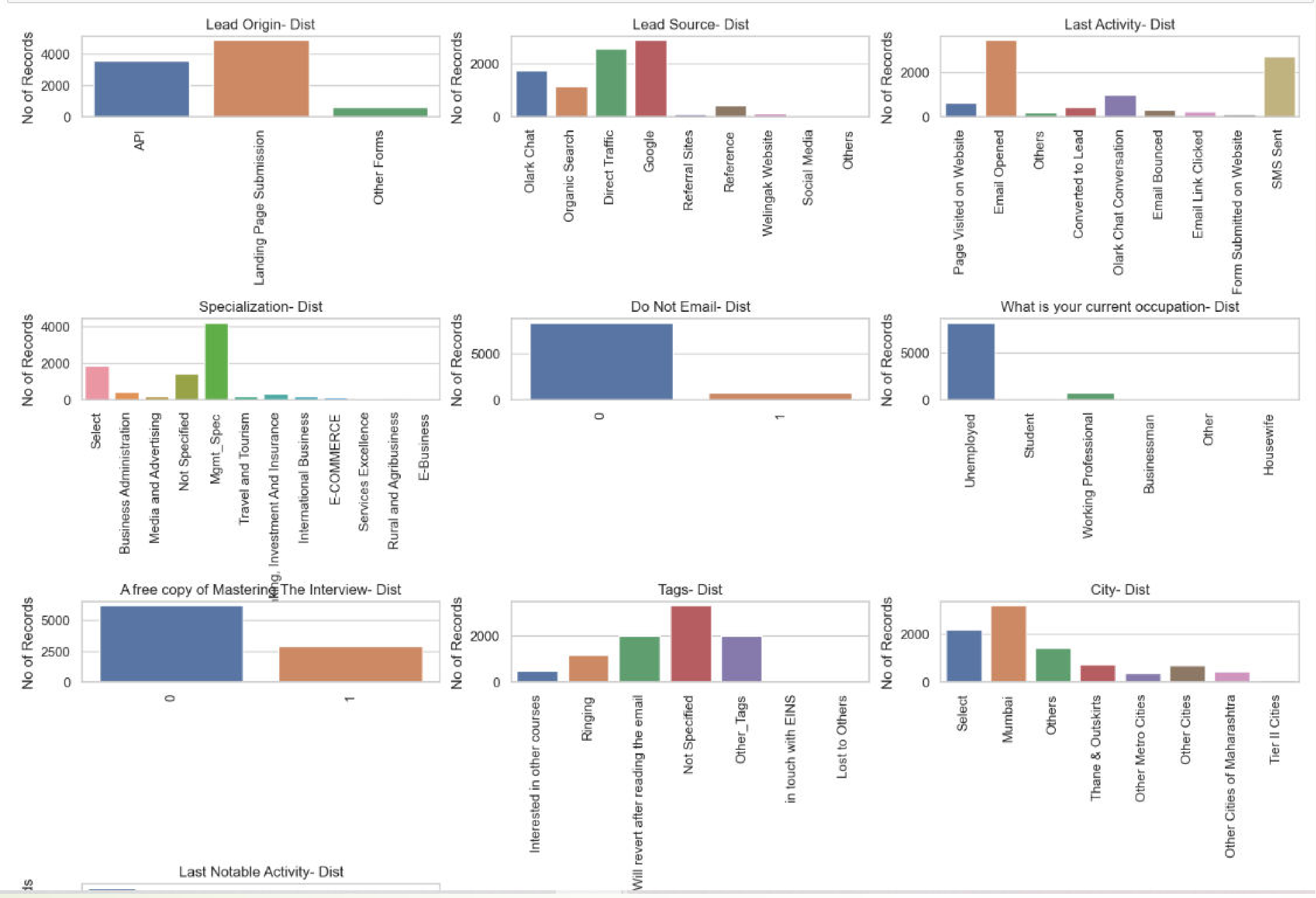
05 **Creation of Dummies and Scaling Train Test Split**

After all the adjustments to the data we proceed with Data Split (Train Test and Scaling.

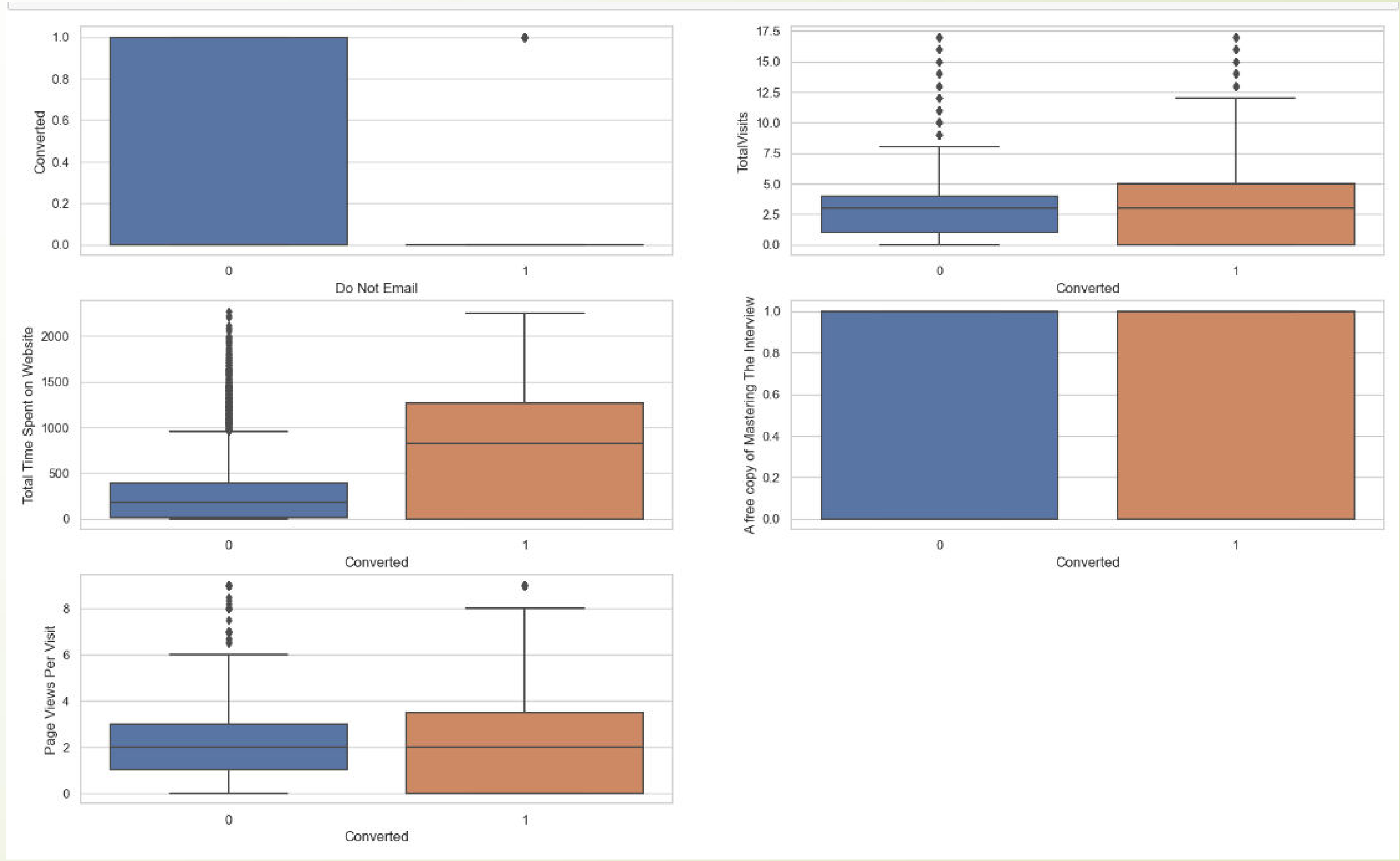
06 **Evaluate the model on RFE & VIF Confusion Matrix**

Perform statical , RFE VIF for feature elimination and check the confusion Matrix and Accuracy score by building 2-3 Models

Univariate Analysis of Categorical

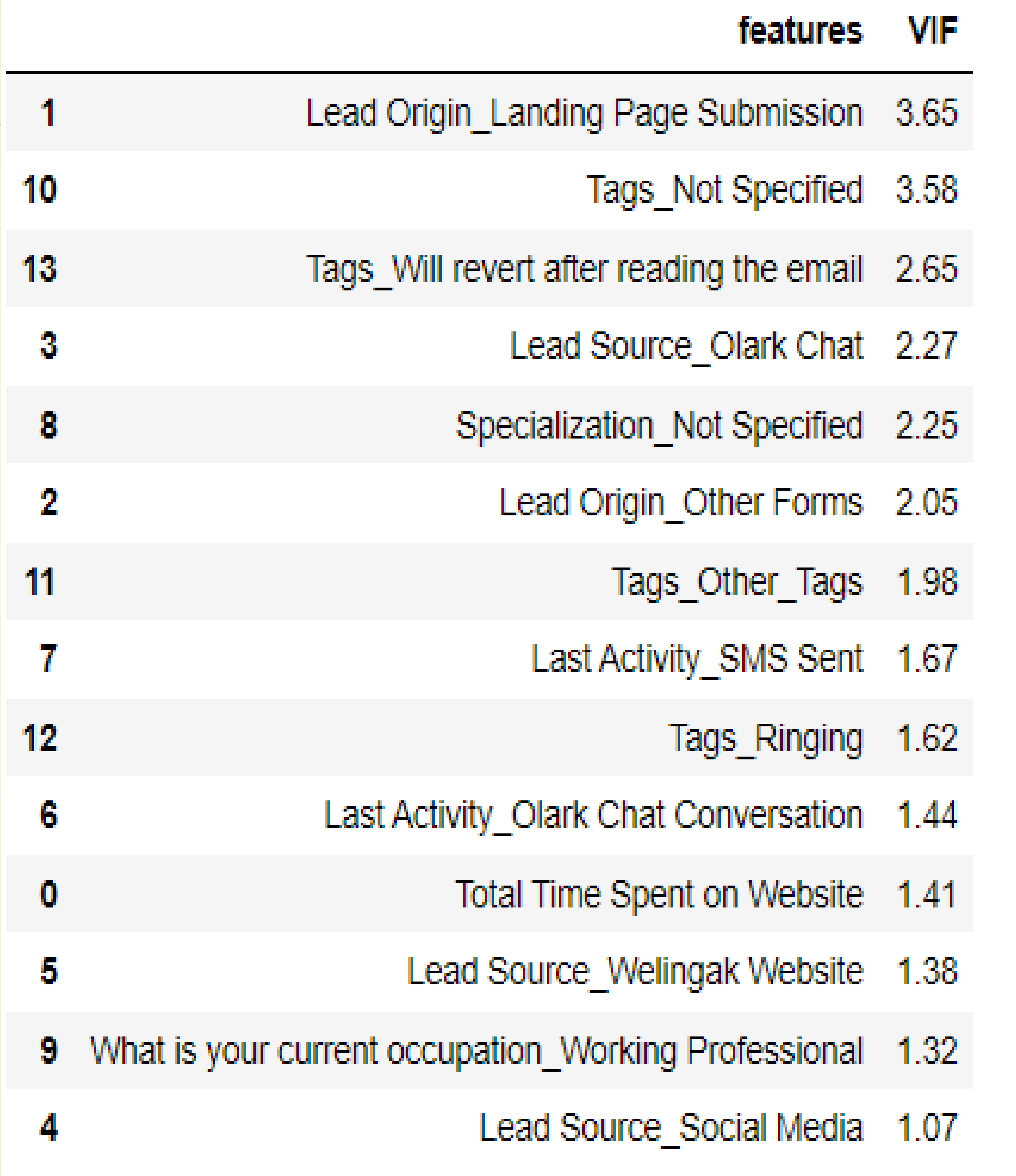


Shows the distribution of categories within each column and the count of each categories that influence the lead



Shows the distribution of numerical variables within dataset.

Building the Model



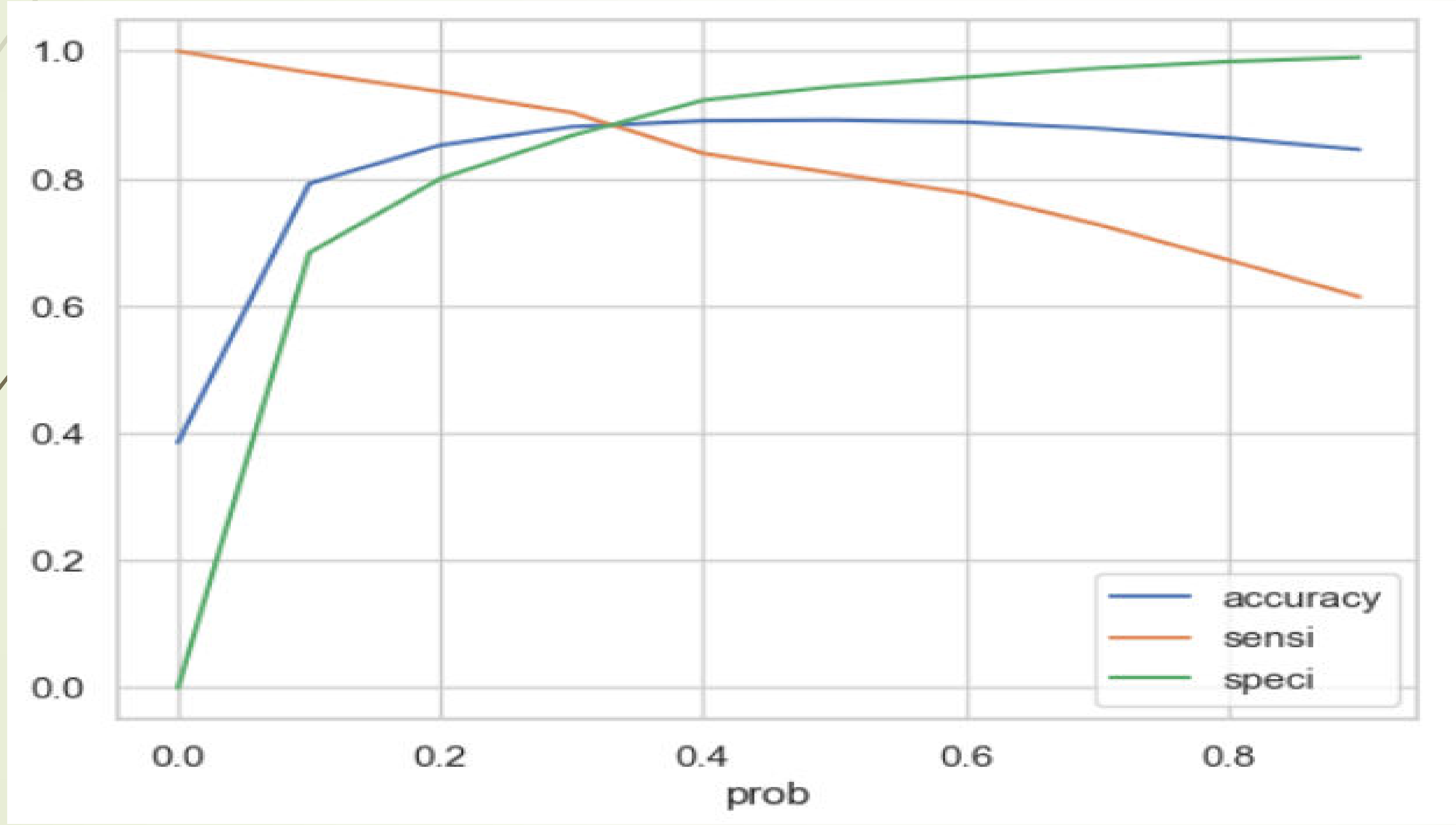
	features	VIF
1	Lead Origin_Landing Page Submission	3.65
10	Tags_Not Specified	3.58
13	Tags_Will revert after reading the email	2.65
3	Lead Source_Olark Chat	2.27
8	Specialization_Not Specified	2.25
2	Lead Origin_Other Forms	2.05
11	Tags_Other_Tags	1.98
7	Last Activity_SMS Sent	1.67
12	Tags_Ringing	1.62
6	Last Activity_Olark Chat Conversation	1.44
0	Total Time Spent on Website	1.41
5	Lead Source_Welingak Website	1.38
9	What is your current occupation_Working Professional	1.32
4	Lead Source_Social Media	1.07

- Generalized Linear Models from Stats Library is used to Build Logistic Regression Model.
- The Model is built Initially with a minimum of 15 features that are selected by RFE.
- The features where the p value is greater than 0.05 and VIF greater than 5 are dropped and the model is run again to check if dropping these features has significantly impacted the accuracy score which should not be the case.
- The Final model is built using 12 features and tested for its accuracy.

Evaluation of Model

(Sensitivity, Specificity on Train Data)

The graph depicts an optimal cut off of 0.3 based on Accuracy, Sensitivity & Specificity



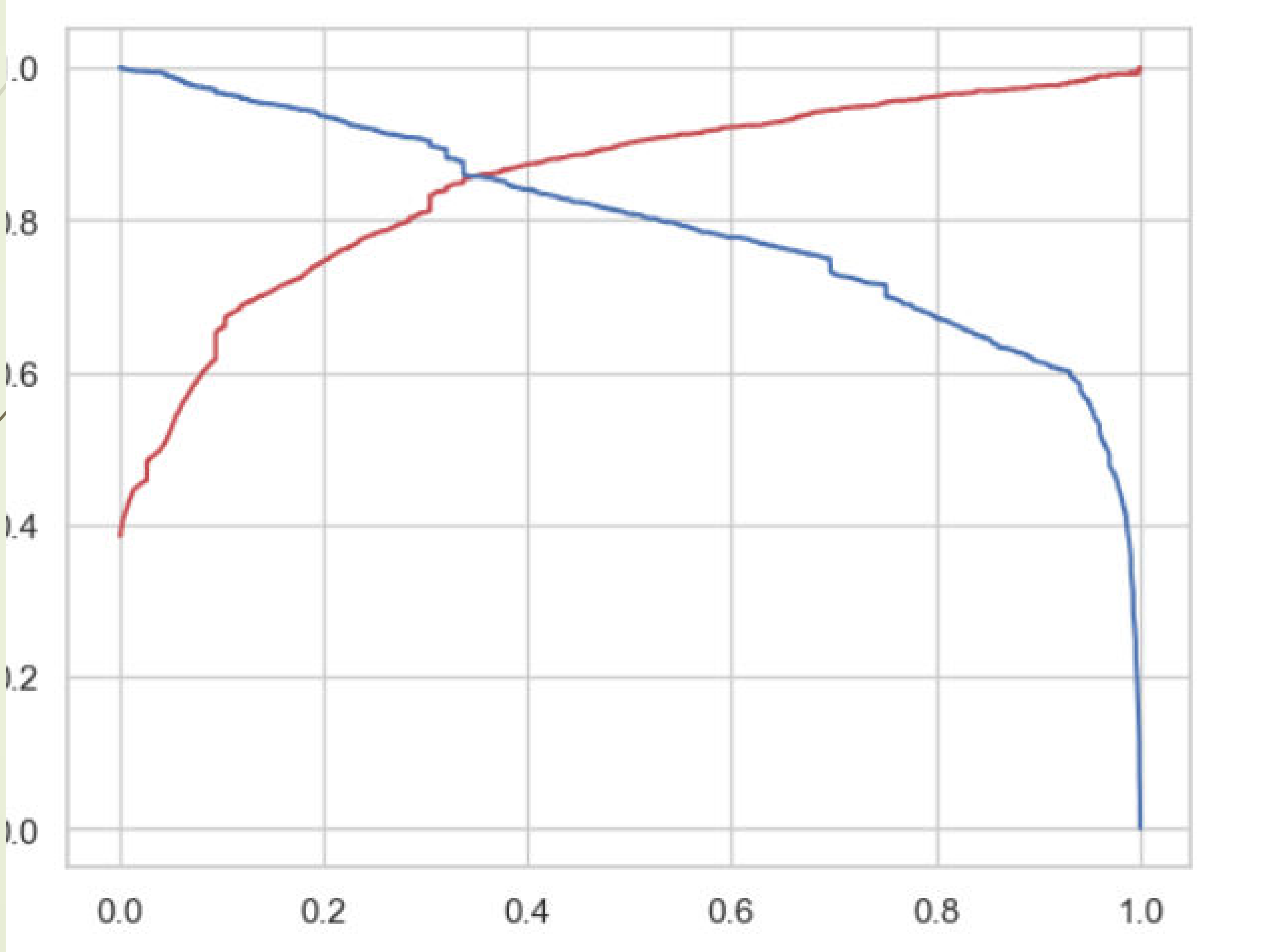
Confusion Matrix

- Accuracy : 88.13%
- Sensitivity : 90.35%
- Specificity : 86.73%
- False Positive Rate : 13.27%
- Positive Predictive Value : 81.01%
- Negative Predictive Value : 93.43%

Evaluation of Model

(Precision and Recall on Train Data)

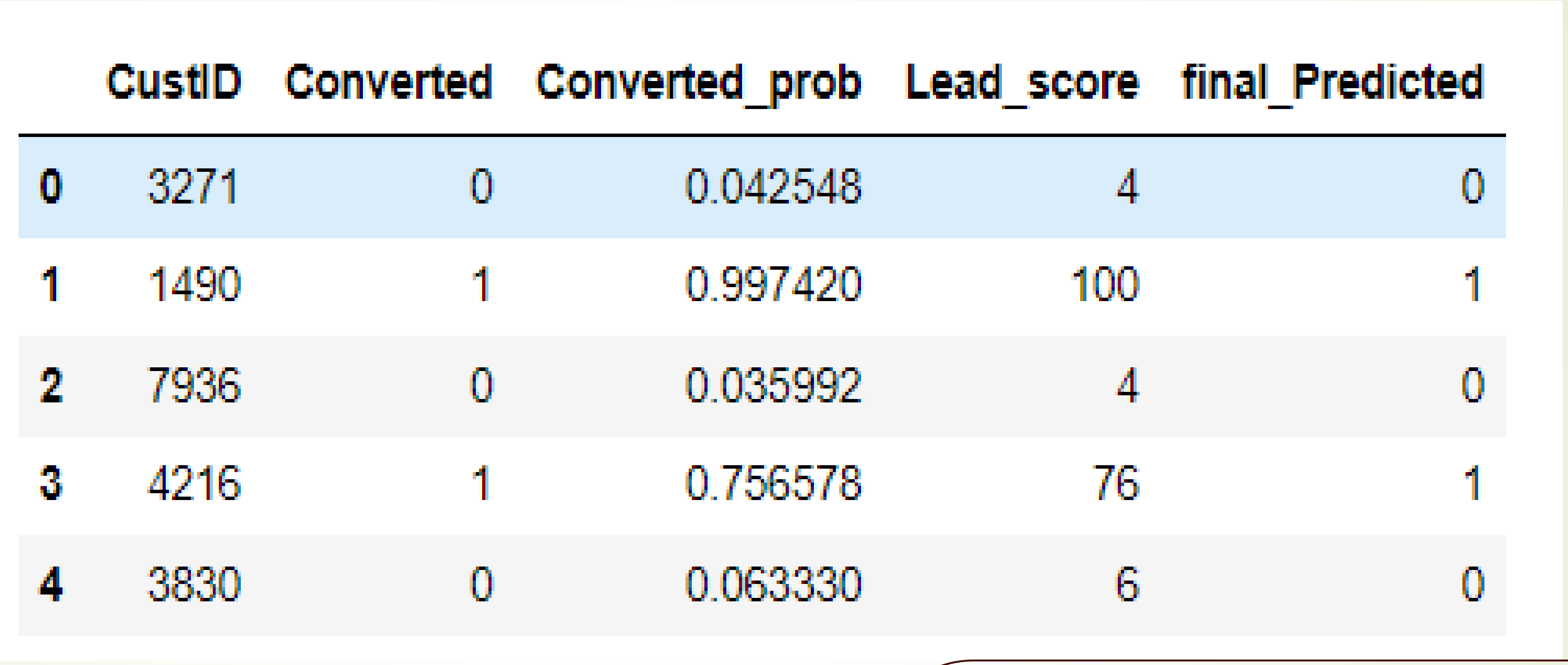
The graph depicts an optimal cut off of 0.35 based on Precision and Recall



▪ Precision	: 81.01%
▪ Recall	: 90.35%

Evaluation of Model

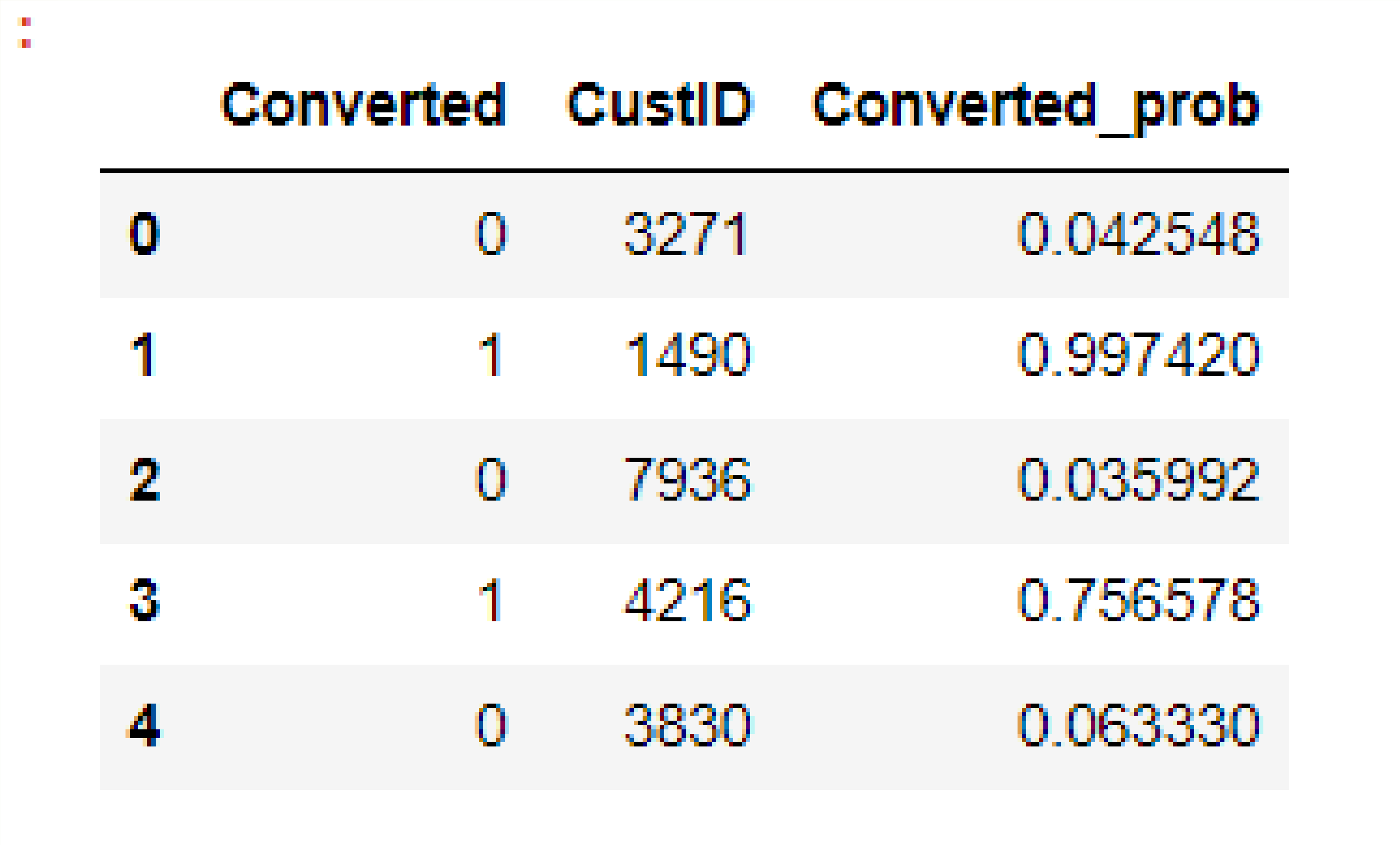
(Sensitivity and Specificity on Test Data)



	CustID	Converted	Converted_prob	Lead_score	final_Predicted
0	3271	0	0.042548	4	0
1	1490	1	0.997420	100	1
2	7936	0	0.035992	4	0
3	4216	1	0.756578	76	1
4	3830	0	0.063330	6	0

- Accuracy : 86.63%
- Sensitivity : 88.89%
- Specificity : 85.35%

Conversion Probability - Prediction

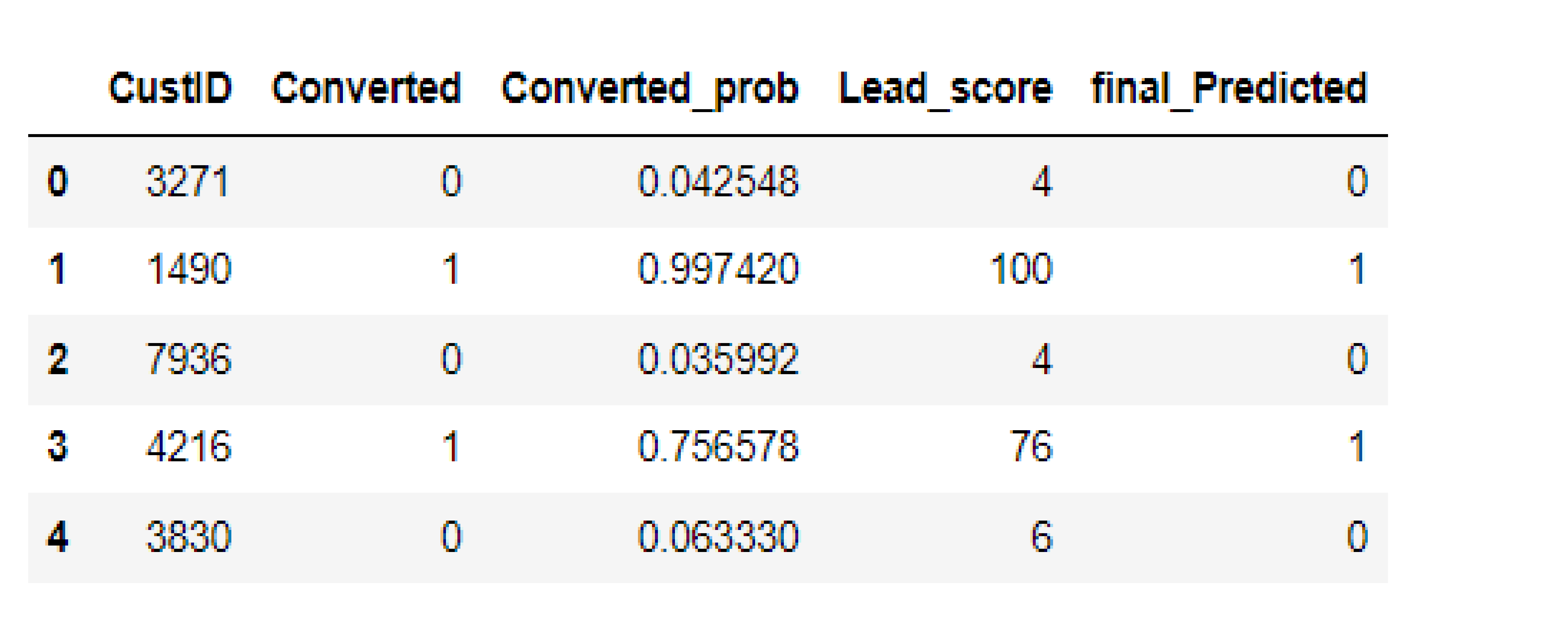


	Converted	CustID	Converted_prob
0	0	3271	0.042548
1	1	1490	0.997420
2	0	7936	0.035992
3	1	4216	0.756578
4	0	3830	0.063330

- Created a Data Frame with the actual converted flag and Predicted probabilities

- Checking the top 5 records with the converted probabilities.

Final Conversion Probability - Prediction

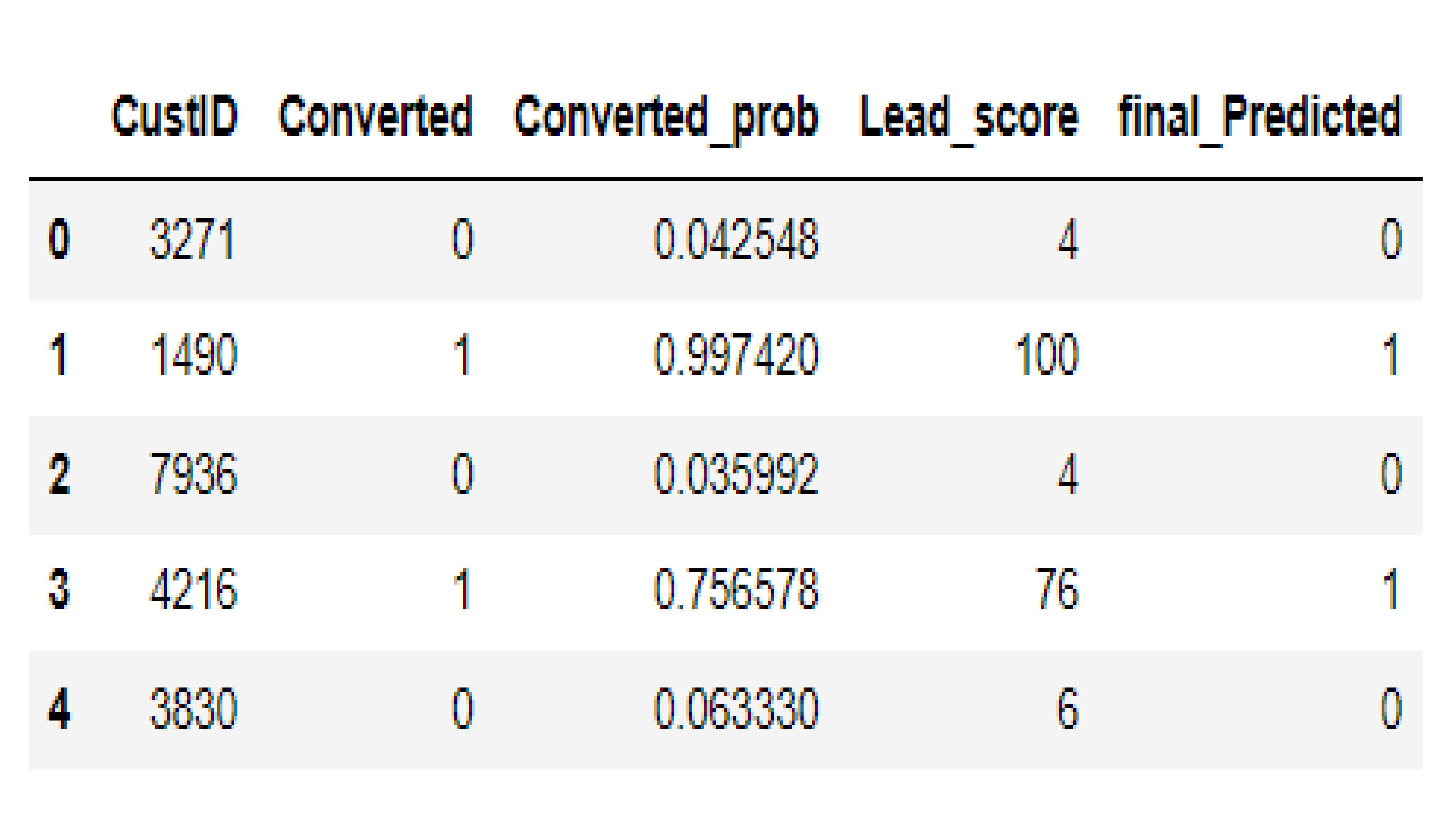


	CustID	Converted	Converted_prob	Lead_score	final_Predicted
0	3271	0	0.042548	4	0
1	1490	1	0.997420	100	1
2	7936	0	0.035992	4	0
3	4216	1	0.756578	76	1
4	3830	0	0.063330	6	0

- Created a final_Predicted column that shows that if the converted_prob > 0.3 to be taken as 1 else it would be taken as 0.
- Checking the top 5 records with the final_Predicted column.

Final Conversion Probability - Prediction

Lead Score = 100 * Converted_prob



	CustID	Converted	Converted_prob	Lead_score	final_Predicted
0	3271	0	0.042548	4	0
1	1490	1	0.997420	100	1
2	7936	0	0.035992	4	0
3	4216	1	0.756578	76	1
4	3830	0	0.063330	6	0

Lead score is calculated for all the leads present in the Original Data Frame.

- The train and test dataset is concatenated to get the entire list of leads available.
- The conversion probability is multiplied by 100 to obtain the lead score for each lead.
- Higher the lead score high is the changes of the lead getting converted and vice-versa
- Since we have used 0.3 as our final probability threshold for deciding if a lead will be converted or not, any value with a lead of 30% and above will have a value of 1 in the final_Predicted column.

Conclusion

- As we have checked both Sensitivity & Specificity as well as Precision and Recall Metrics, we have considered the optimal threshold on Sensitivity and Specificity for calculating the final Prediction.

- The Accuracy Sensitivity and Specificity values of the test sets are close to that of the Train data set we can conclude that the Model is working on the test data in the same way as its working on the train data.

- The lead score conversion rate on the final predicted model is around 88% (train data set) and 86% (test data set)

- Hence we can conclude that the model is good to be implemented.

Recommendation

- We will choose a lower threshold value for conversion probability. This will ensure the Sensitivity rating is very high which in turn will make sure almost all leads that are likely to convert are rightly identified.
- In such a scenario the sales team can proceed with calling the candidate to initiate the converts.
- The sales force can thus work towards the leads with higher score of more than 25% and for the rest automated calls will take care.
- If we choose a higher threshold for conversion Probability the specificity rating will be higher that will eliminate the lead that are almost on the brink i.e. probability of them getting converted as well as they might not be selected.
- In such scenario the sales team might not have to make unnecessary calls and might focus on collating and analyzing the reviews from past group.
- In offseason the sales team can target leads with score of more than 60 or 70 and for rest we can have automated communication set.

THANK YOU