LEAD SCORE CASE STUDY

Submitted by:

Sweta Kumari

Shwetha G

Suparsha Das

PROBLEM STATEMENT

INTRODUCTION:

An education company, X Education sells online courses to industry professionals. The company markets its courses on various websites and search engines such as Google

Once people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals

Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. The typical lead conversion rate at X education is around 30%

BUSINESS GOALS:

Company wishes to identify the most potential leads, also known as "**Hot Leads**"

The company needs a model wherein a lead score is assigned to each of the leads such that the customer with higher lead score have a higher conversion chance and customer with lower lead score have a lower conversion chance

The CEO, in particular, has given a ballpark number for the lead conversion rate i.e. **80%**

OVERALL APPROACH

1. DATA UNDERSTANDING AND EXPLORATION

2. EXPLORATORY DATA ANALYSIS : UNIVARIATE , BIVARIATE and MULTIVARIATE ANALYSIS

3. DATA PREPARATION(FEATURE SCALING AND DUMMY VARIABLE CREATION)

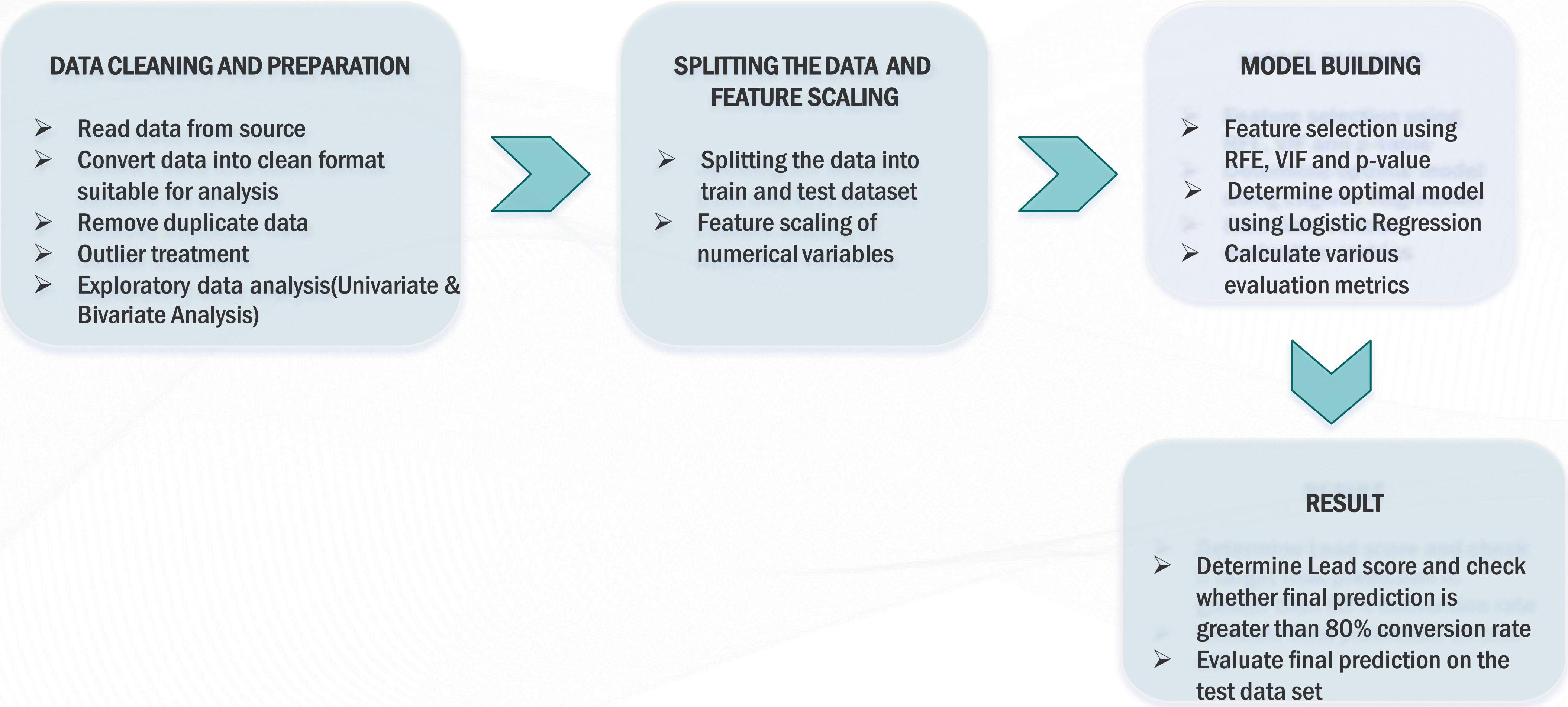
4. LOGISTIC REGRESSION MODEL BUILDING

5. MODEL EVALUATION : SPECIFICITY , SENSITIVITY, PRECISION and RECALL

6. CONCLUSION AND RECOMMENDATION

UpGrad

PROBLEM SOLVING METHODOLOGY



DATA CLEANING AND PREPARATION

- Read data from source
- Convert data into clean format suitable for analysis
- Remove duplicate data
- Outlier treatment
- Exploratory data analysis(Univariate & Bivariate Analysis)

SPLITTING THE DATA AND FEATURE SCALING

- Splitting the data into train and test dataset
- Feature scaling of numerical variables

MODEL BUILDING

- Feature selection using RFE, VIF and p-value
- Determine optimal model using Logistic Regression
- Calculate various evaluation metrics

RESULT

- Determine Lead score and check whether final prediction is greater than 80% conversion rate
- Evaluate final prediction on the test data set

DATA CONVERSION

1. CONVERTING THE VARIABLE WITH VALUES **YES/NO** to **1/0s**

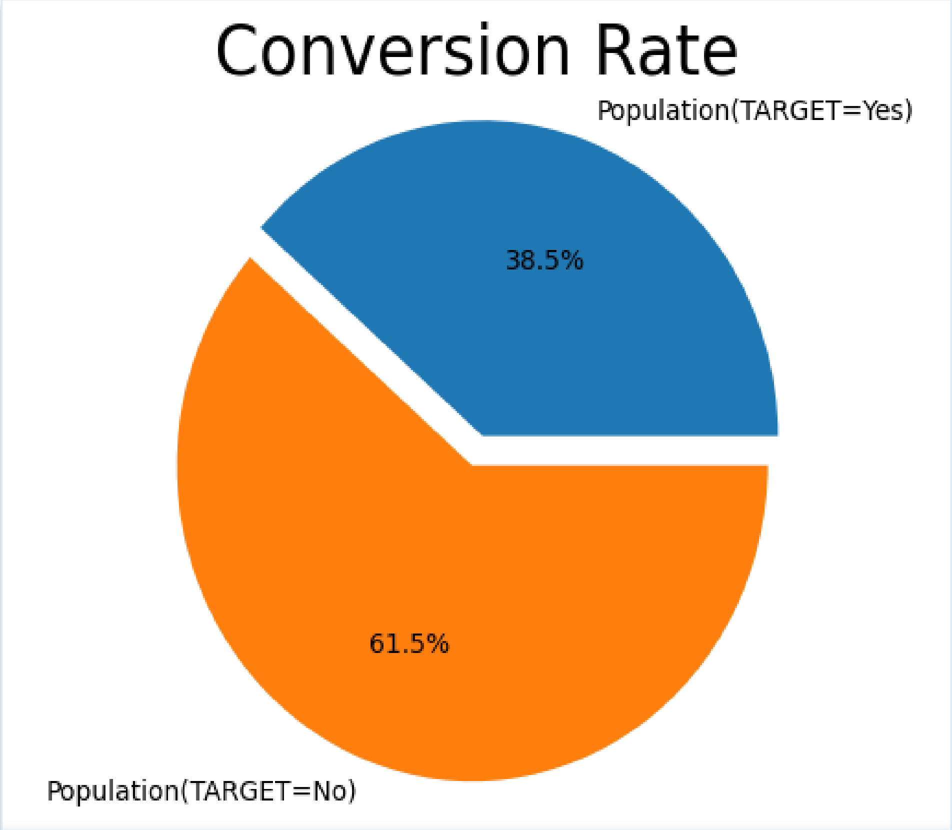
2. CONVERTING THE '**SELECT**' VALUES WITH **NaNs**

3. DROPIING THE COLUMNS HAVING **>70%** OF NULL VALUES

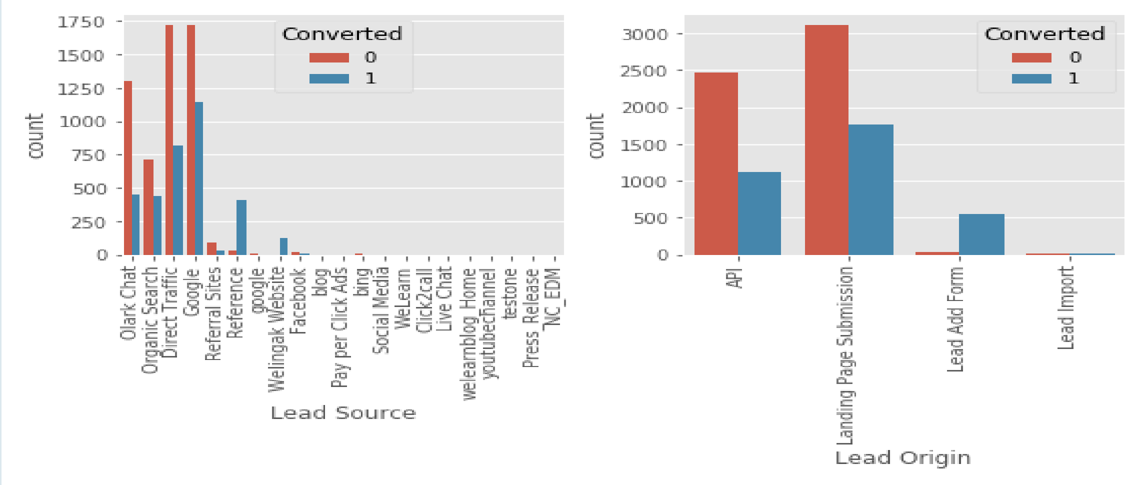
4. DROPPING UNNECESSARY COLUMNS

5. DROPPING THE ROWS AS THE NULL VALUES WERE **<2%**

EXPLORATORY DATA ANALYSIS

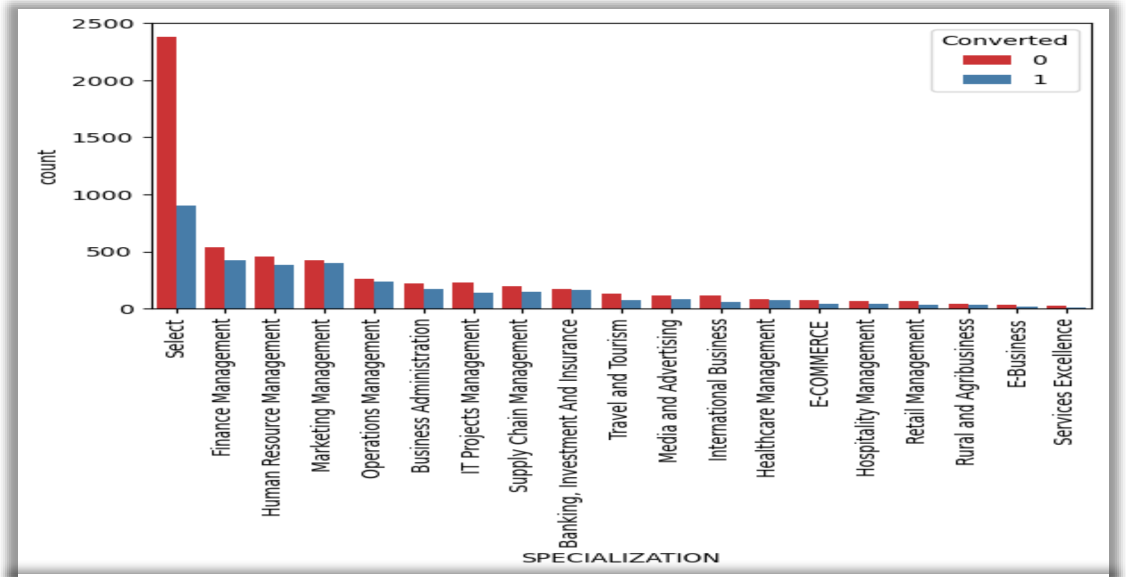


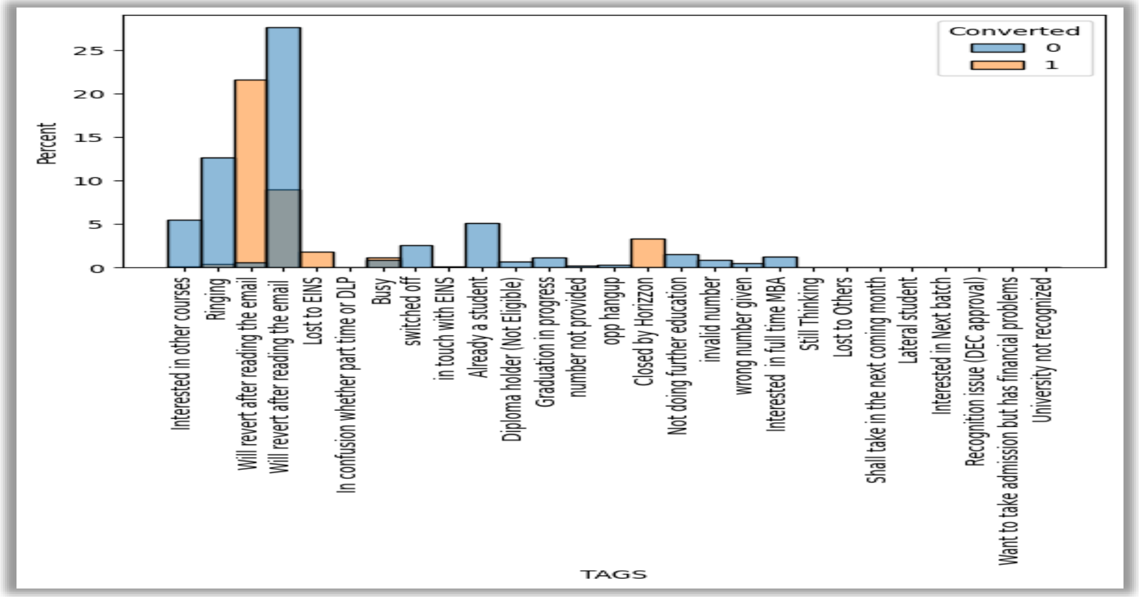
Conversion rate as depicted in the fig is approx. 39%

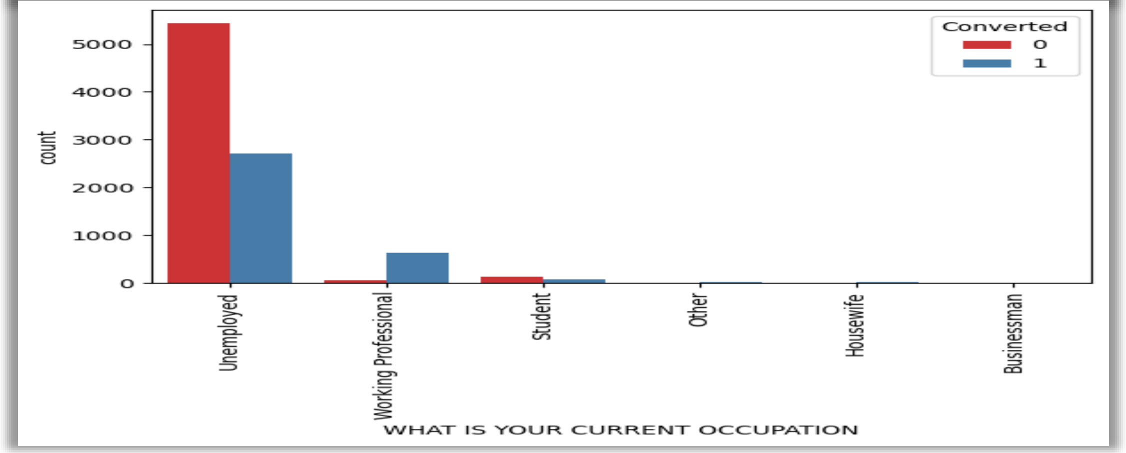


- ➢ The conversion rate of the leads from Reference and Welingak Website is maximum
- ➢API and Landing Page Submission has less conversion rate(~30%) but counts of the leads from them are considerable
- ➢The count of leads from the Lead Add Form is pretty low but the conversion rate is very high

EXPLORATORY DATA ANALYSIS

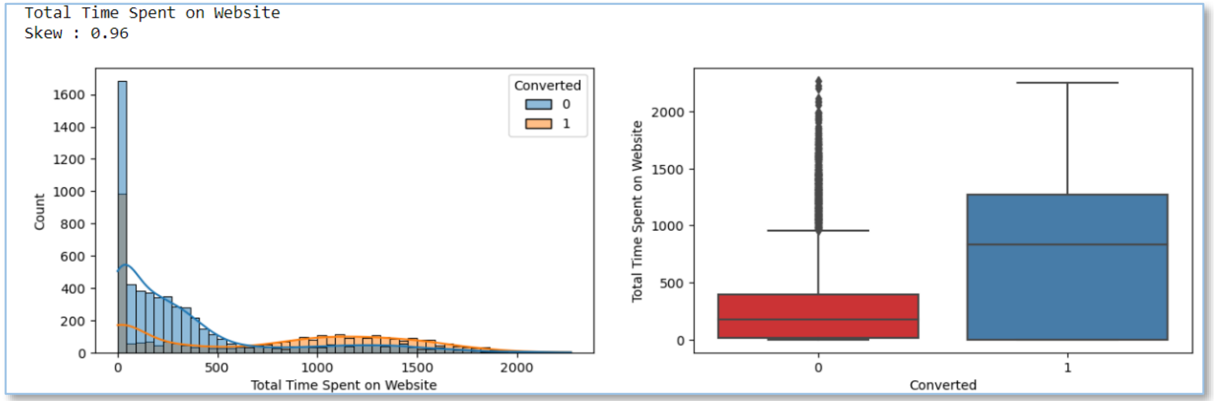




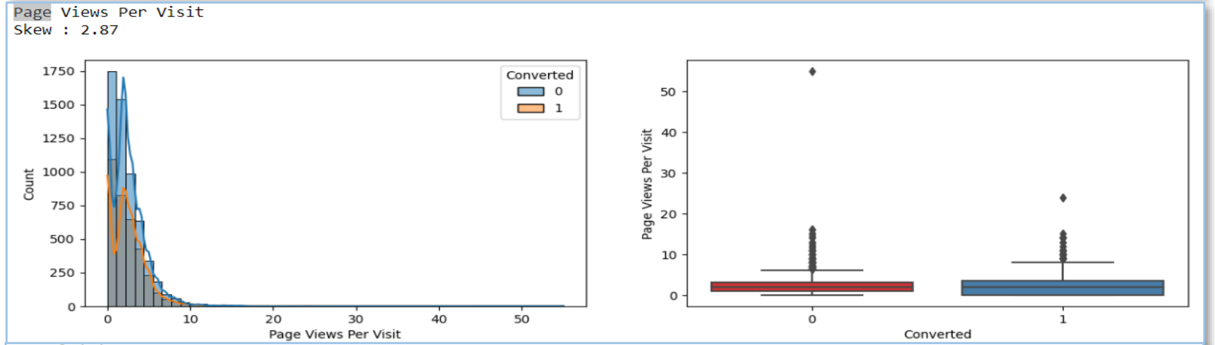


- Looking at above plot, no inference can be made for specialization.
- Working Professional has high conversion rate
- Un employed leads are more than other category.
- “Will revert after reading email” has high percentage of being converted.

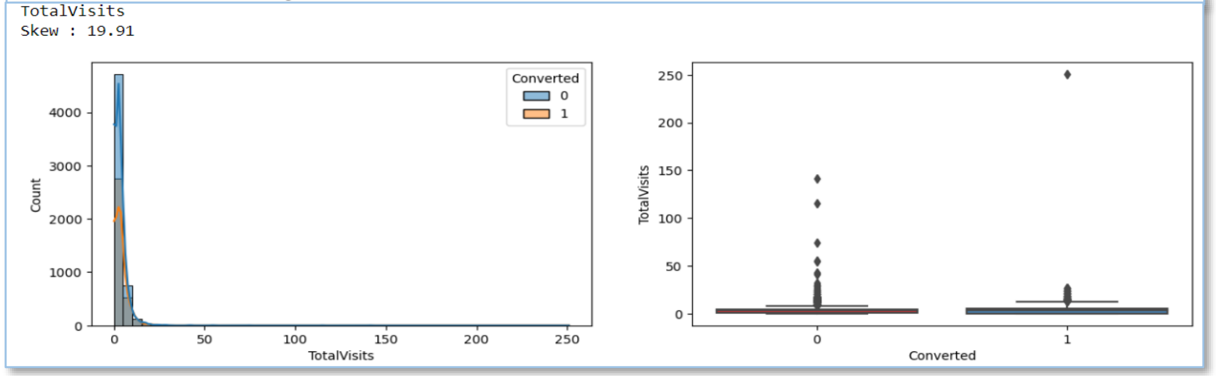
EXPLORATORY DATA ANALYSIS



```
Total Time Spent on Website
Skew : 0.96
```



```
Page Views Per Visit
Skew : 2.87
```



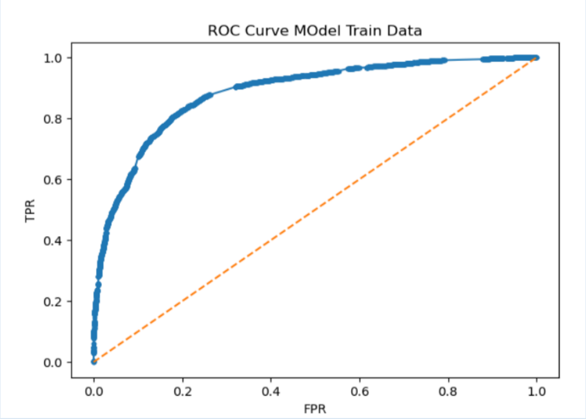
```
TotalVisits
Skew : 19.91
```

- ➢ The Hist plot depicts that the values are right skewed.
- ➢ Box plot of the “Pages Visits per View”& “Total Visits” shows that the median of both converted & Non-converted is same. Nothing conclusive can be said with the information.
- ➢ Users spending most time in websites are likely to be converted

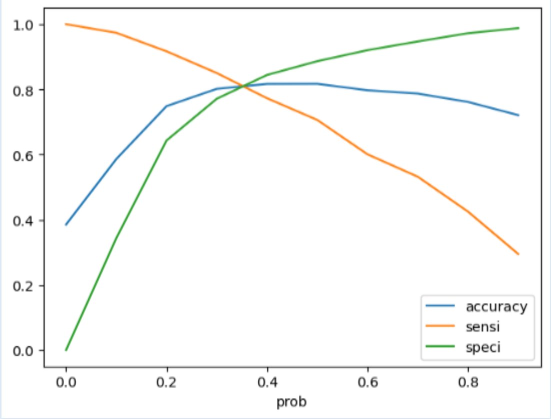
MODEL BUILDING

- SPLITTING THE DATA INTO TEST AND TRAINING SETS
- WE HAVE CHOSEN THE TRAIN_TEST SPLIT RATIO AS 70:30
- USING RFE TO CHOOSE TOP 20 VARIABLES
- BUILD MODEL BY REMOVING THE VARIABLES WHOSE p-VALUE > 0.05 AND VIF > 5
- PREDICTIONS ON TEST DATASET
- OVERALL ACCURACY IS 81.6%
- OPTIMAL CUT OFF POINT OF 0.37 AS DEPICTED IN THE FIGURE.

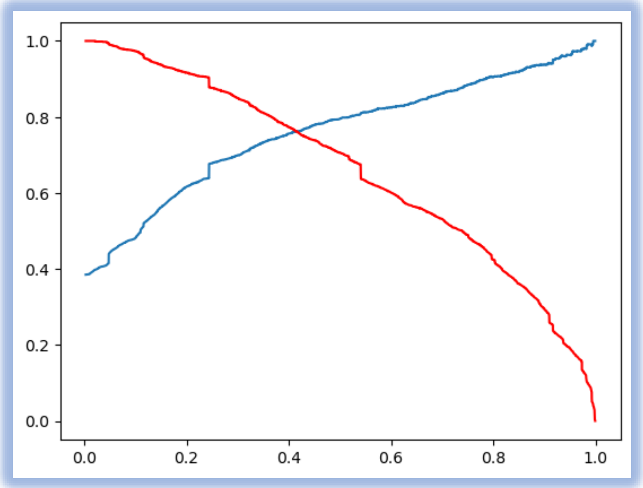
ROC AUC CURVE TRAIN DATA



OPTIMAL CUT-OFF



TRADE OFF CURVE PRECISION RECALL

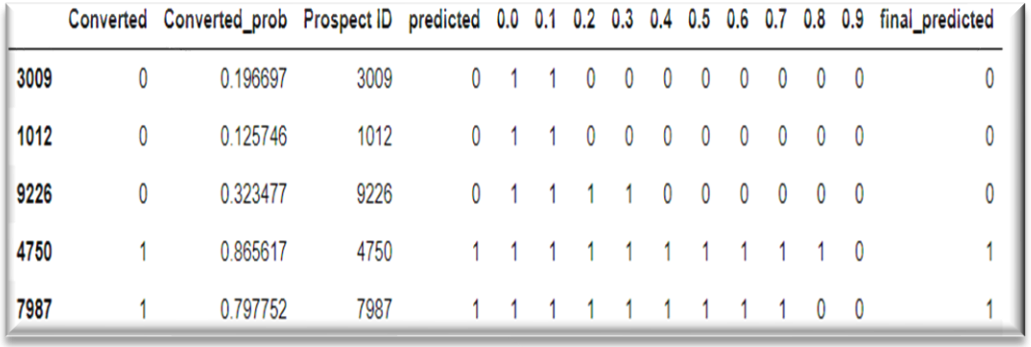


MODEL EVALUATION

- Calculated **ACCURACY, SPECIFICITY,SENSITIVITY** for various probability cut off btw 0.1 to 0.9.

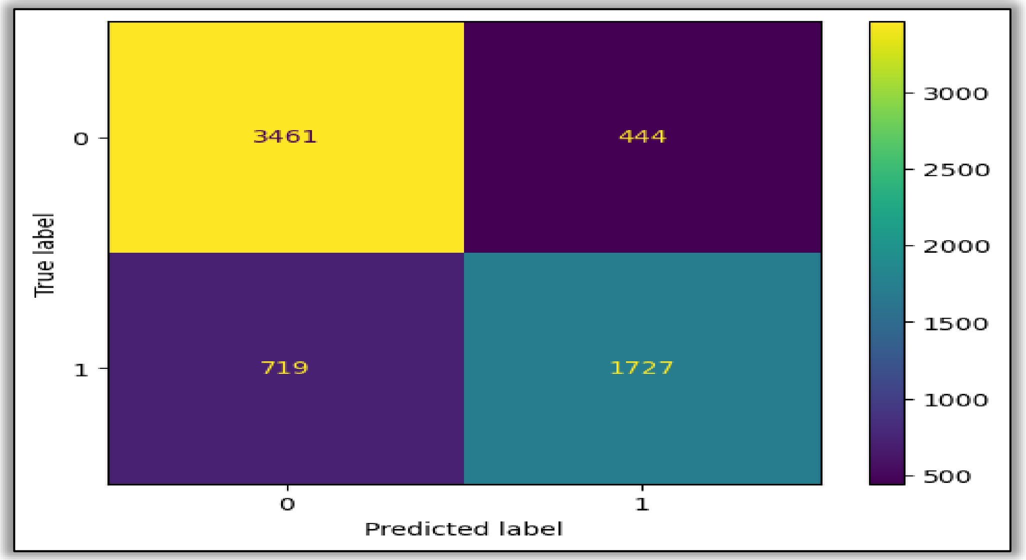
- As per the graph, it can be seen that the optimal cut off point is **0.37**

PREDICTED ACTUAL	NOT CONVERTED	CONVERTED
NOT CONVERTED	3461	444
CONVERTED	719	1727



	Converted	Converted_prob	Prospect ID	predicted	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	final_predicted
3009	0	0.196697	3009	0	1	1	0	0	0	0	0	0	0	0	0
1012	0	0.125746	1012	0	1	1	0	0	0	0	0	0	0	0	0
9226	0	0.323477	9226	0	1	1	1	1	0	0	0	0	0	0	0
4750	1	0.865617	4750	1	1	1	1	1	1	1	1	1	1	0	1
7987	1	0.797752	7987	1	1	1	1	1	1	1	1	1	0	0	1

CONFUSION MATRIX TRAIN DATA



METRICS SCORE(TRAIN DATA)

ACCURACY	83.59%
PRECISION	71.6%
SENSITIVITY	79.6%
SPECIFICITY	82.6%

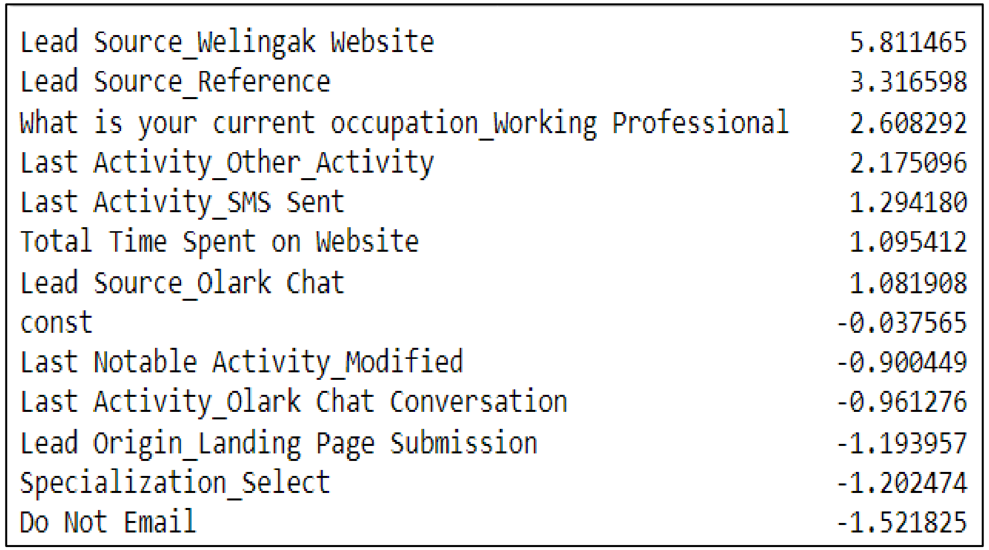
MODEL EVALUATION

- **ACCURACY, SENSITIVITY AND SPECIFICITY** FOR VARIOUS PROBABILITY CUTOFFS FROM **0.1 TO 0.9**
- AS PER THE GRAPH AND LOOKING AT THE OTHER SCORES, IT CAN BE SEEN THAT THE **OPTIMAL POINT IS 0.37**

CONFUSION MATRIX TEST DATA

PREDICTED ACTUAL	NOT CONVERTED	CONVERTED
NOT CONVERTED	1468	266
CONVERTED	234	755

FEATURE IMPORTANCE



```
Lead Source_Welingak Website                          5.811465
Lead Source_Reference                                 3.316598
What is your current occupation_Working Professional  2.608292
Last Activity_Other_Activity                          2.175096
Last Activity_SMS Sent                                1.294180
Total Time Spent on Website                           1.095412
Lead Source_Olark Chat                                1.081908
const                                                -0.037565
Last Notable Activity_Modified                       -0.900449
Last Activity_Olark Chat Conversation                -0.961276
Lead Origin_Landing Page Submission                  -1.193957
Specialization_Select                                -1.202474
Do Not Email                                         -1.521825
```

METRICS SCORE (TEST DATA)

ACCURACY	81.6%
PRECISION	73.9%
SENSITIVITY	76.3%
SPECIFICITY	84.6%

CONCLUSION

The logistic regression model is used to predict the probabillty of conversion of a customer.

While we have calculated both **sensitivity-specificity** as well as **Precision-Recall** metrics, we have considered optimal cut off on the basis of **sensitivity-specificity** for final prediction

Lead Score calculated shows the conversion rate of final predicted model is around **80% in test data** & **train data set**

In Business terms, this model has capability to adjust with the company's requirements in coming future

TOP variables that contributes for lead getting converted in the model are:

- Lead Source_Welingak Website
- Lead Source_Reference
- What is your current occupation_Working Professional

Hence Overall this model seems to be good