LEAD SCORE CASE Study

Group Member Name

1. Vishwaani Ravula
2. Sheetal Dandekar
3. Divya Raja

AGENDA

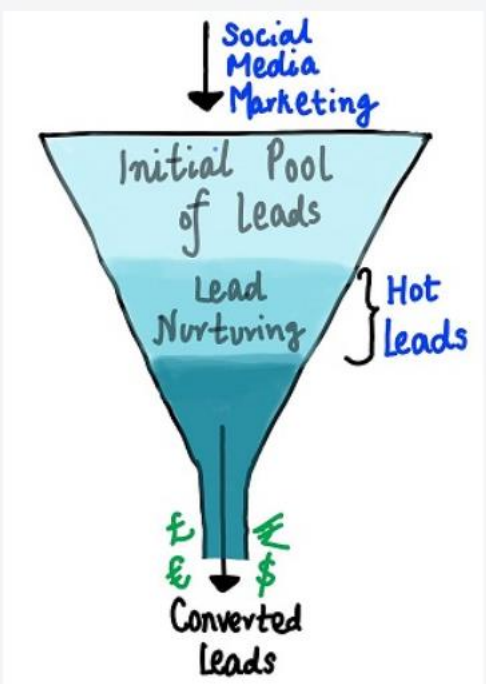
- Problem Statement
- Business Objective
- Solution Methodology
- Data Manipulation
- EDA
- Categorical Variable Relation
- Data Conversion
- Model Building
- ROC Curve

Problem Statement

- X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses. 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 potential leads rather than making calls to everyone.
- The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Business Objective

- X Education wants to know the most promising leads.
- For that, they want to build a Model which identifies the hot leads.
- Deployment of the model for the future use. Education



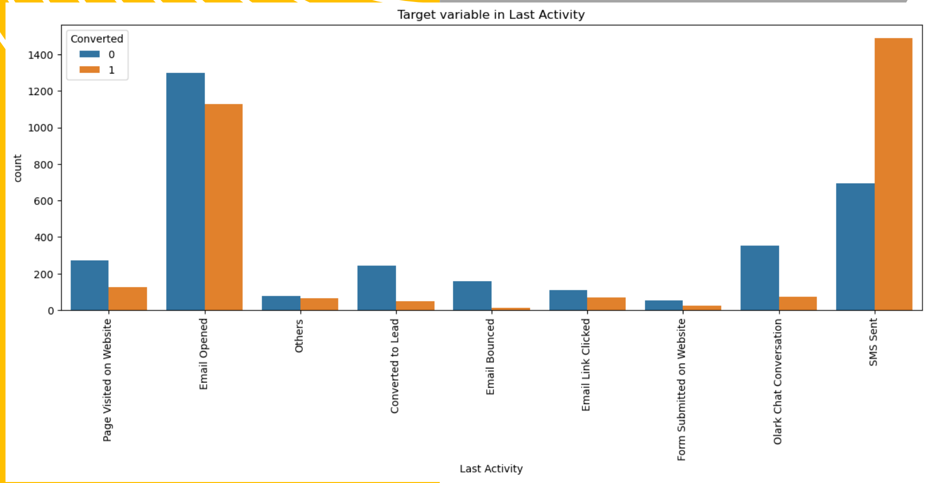
Solution Methodology

- Data cleaning and data manipulation
  1. Check/handle all duplicate data.
  2. Check/handle NA values and missing values.
  3. Drop columns, if they contain a large amount of missing values and are not useful for the analysis.
  4. Check and handle outliers in data.
- EDA
  1. Univariate data analysis
  2. Bivariate data analysis
  3. Feature Scaling and dummy Variables and encoding of the data.
  4. Classification technique: logistic regression is used for the model making and prediction.
  5. Validation of the model.
  6. Model presentation.
  7. Conclusions and recommendations.

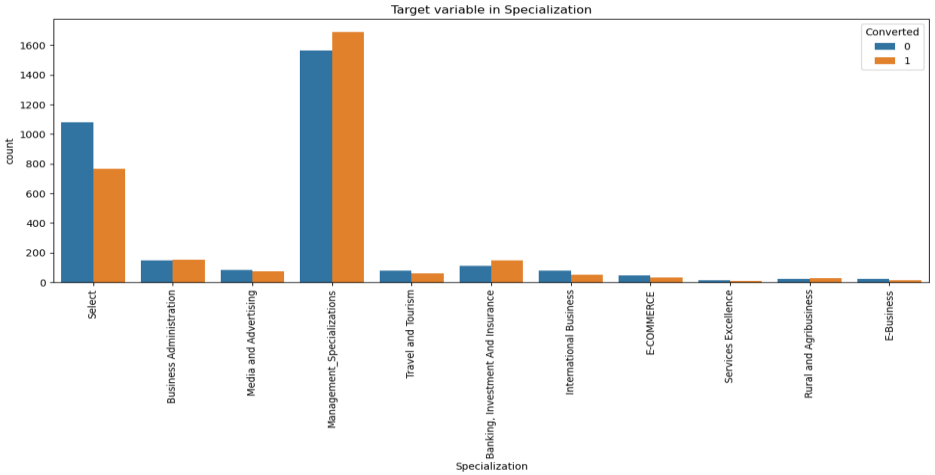
Data Manipulation

1. Total Number of Rows =37, Total Number of Columns =9240.
2. Dropping the “Prospect ID” and “Lead Number” which are not necessary for the analysis.
3. After checking for the value counts for some of the object type variables, we find some of the features which have no enough variance, which we have dropped, some of the features are: “Do Not Call”, “Search”, “Newspaper Article”, “X Education Forums”, “Newspaper”, “Digital Advertisement” etc.
4. Dropping the columns having more than 30% as missing values.

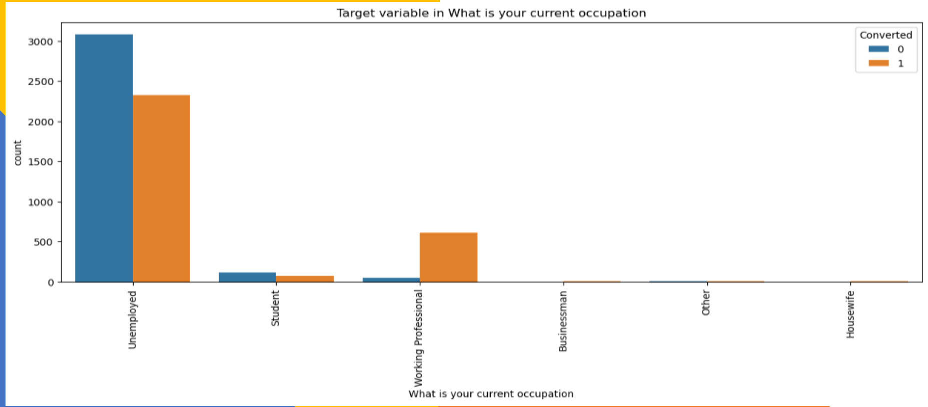
VISUALISING THE CATEGORICAL VARIABLES



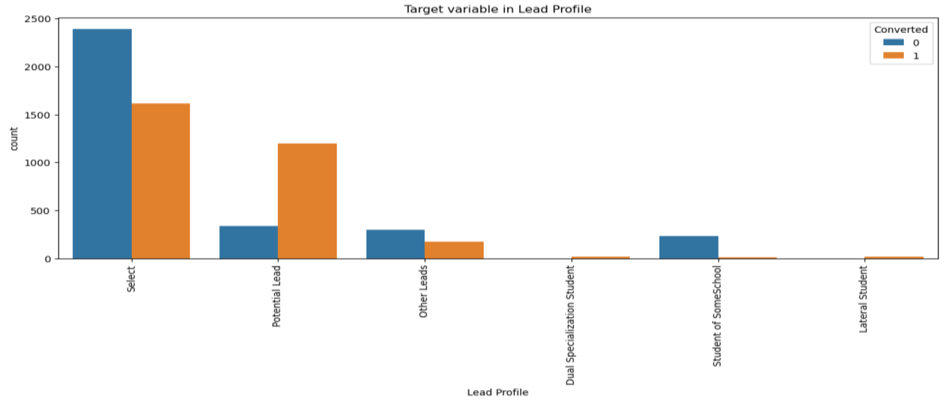
People who have interacted with a SMS being sent in showing an interest seems to be enrolling more in the courses They have a higher conversion ratio.



People who are identified as Potential leads and followed up seems to be converting as a customer to the website

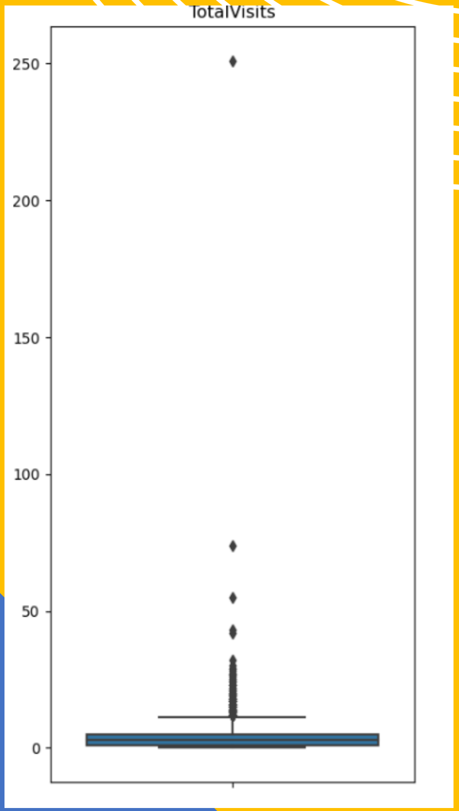


Working Professionals seem to be in the higher conversion ratio compared to the other categories of the people



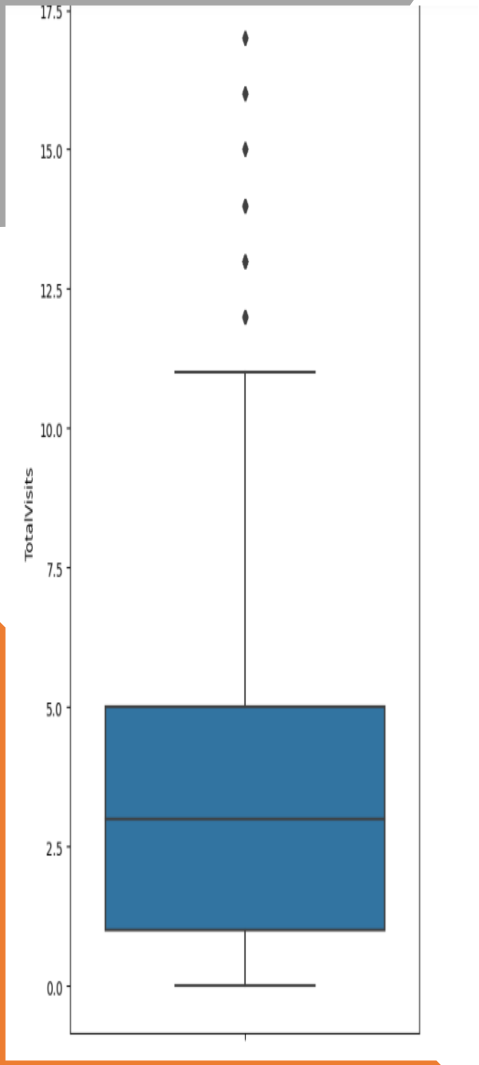
People who have been referred have a higher chance of converting into a customer

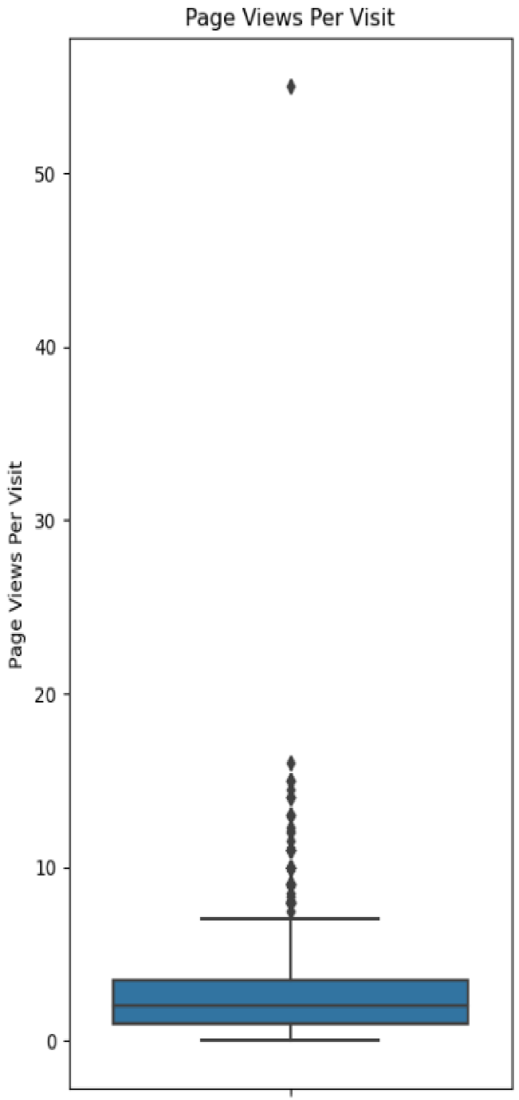
OUTLIER TREATMENT



Before

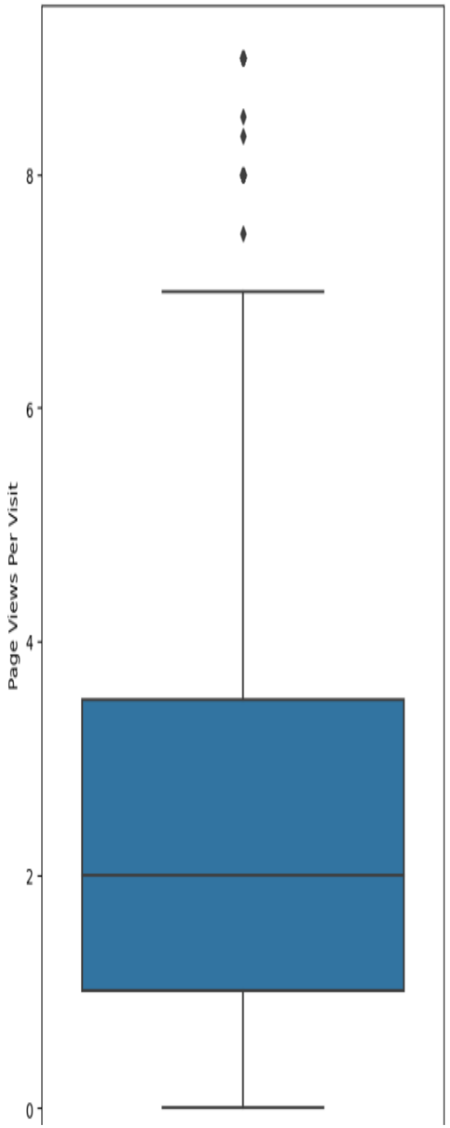
AFTER





Before

AFTER



EDA

Creating a Dummy Variables for the Categorical Variables

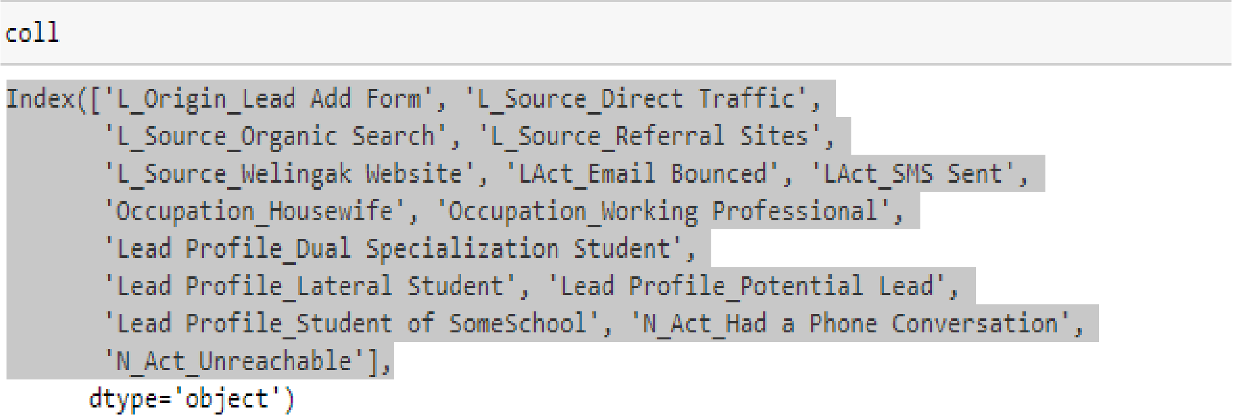
* In order to make it easier for analysis, we will be converting the categorical variables into dummy variables so that it holds easy as to what sort of variable has a much better influence on the target variable

* Converting the Leads Origin, Lead Source, Last Activity, Specialization, How did you hear about X Education, What is your current occupation and the Lead Profile categorical variables into dummy variables and then dropping these.

TEST-TRAIN SPLIT LOGISTIC REGRESSION

- Dividing the entire dataset into test and train set for logistic regression.
- We will be using the Standard Scaler to scaling the values down to comparable values for further correlation and other such values.
- The X-train size is [4404,64] and the y-train size is [4404,]

RFE METHOD

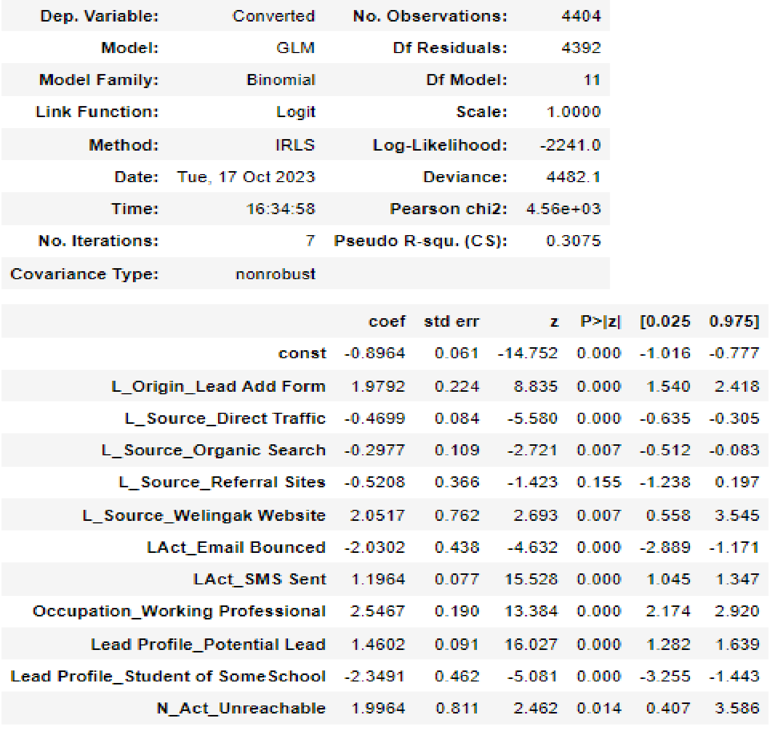


```
coll
```

```
Index(['L_Origin_Lead Add Form', 'L_Source_Direct Traffic',
       'L_Source_Organic Search', 'L_Source_Referral Sites',
       'L_Source_Welingak Website', 'LAct_Email Bounced', 'LAct_SMS Sent',
       'Occupation_Housewife', 'Occupation_Working Professional',
       'Lead Profile_Dual Specialization Student',
       'Lead Profile_Lateral Student', 'Lead Profile_Potential Lead',
       'Lead Profile_Student of SomeSchool', 'N_Act_Had a Phone Conversation',
       'N_Act_Unreachable'],
      dtype='object')
```

- Columns that have been chosen by the RFE method

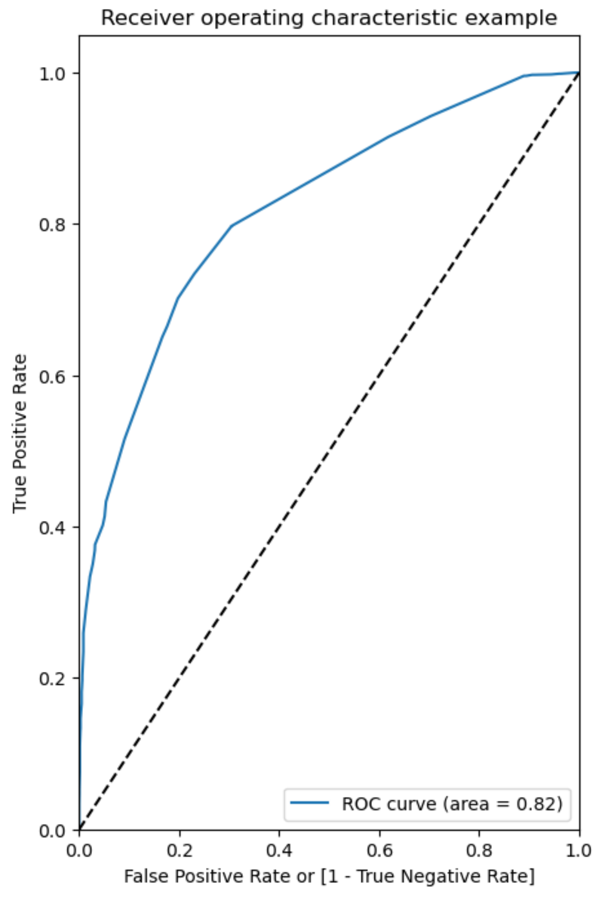
After running a few iterations and removing the feature variables which have a high p-value and the VIF values, we are down to the Model 5 where we have finalised the model on which we will be training the train and test set.



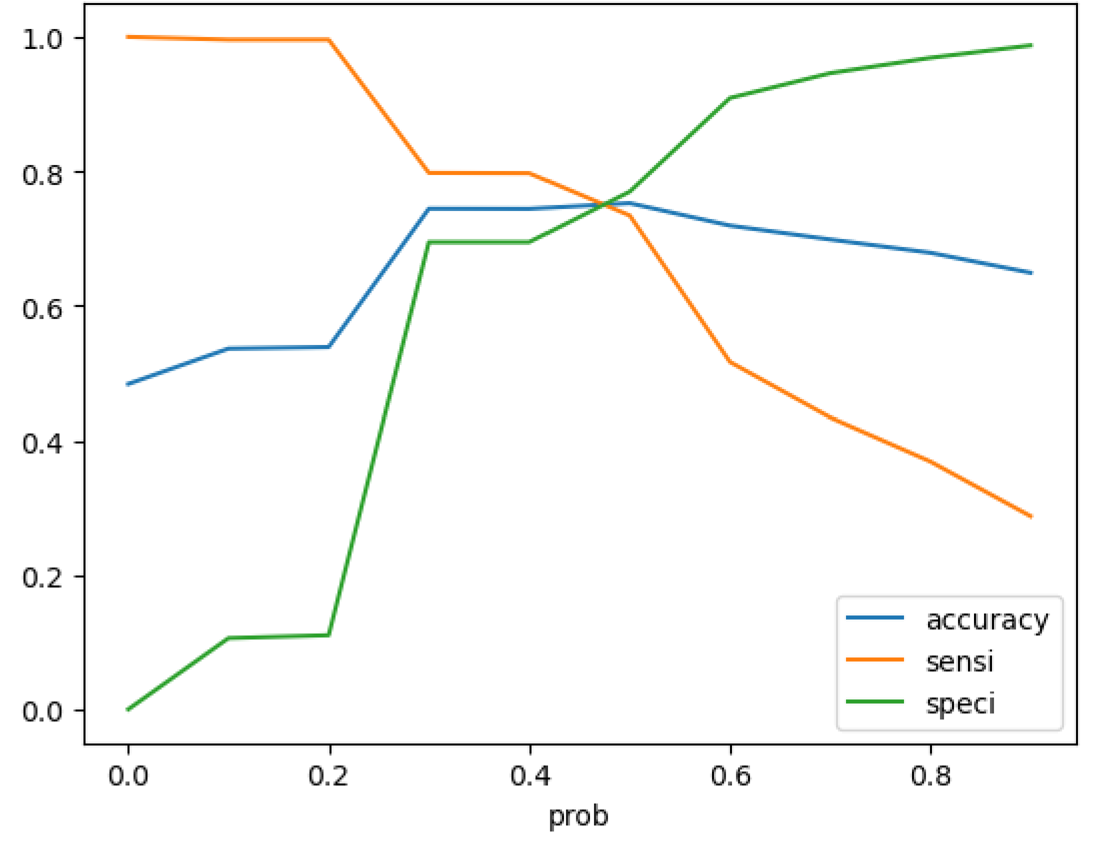
```
     Dep. Variable:         Converted      No. Observations:      4404
            Model:               GLM          Df Residuals:       4392
     Model Family:          Binomial              Df Model:         11
    Link Function:             Logit                 Scale:     1.0000
           Method:              IRLS        Log-Likelihood:    -2241.0
             Date:  Tue, 17 Oct 2023              Deviance:     4482.1
             Time:          16:34:58          Pearson chi2:   4.56e+03
   No. Iterations:                 7   Pseudo R-squ. (CS):      0.3075
 Covariance Type:          nonrobust

                                        coef  std err        z  P>|z|  [0.025  0.975]
                              const  -0.8964    0.061  -14.752  0.000  -1.016  -0.777
             L_Origin_Lead Add Form   1.9792    0.224    8.835  0.000   1.540   2.418
            L_Source_Direct Traffic  -0.4699    0.084   -5.580  0.000  -0.635  -0.305
            L_Source_Organic Search  -0.2977    0.109   -2.721  0.007  -0.512  -0.083
            L_Source_Referral Sites  -0.5208    0.366   -1.423  0.155  -1.238   0.197
          L_Source_Welingak Website   2.0517    0.762    2.693  0.007   0.558   3.545
                 LAct_Email Bounced  -2.0302    0.438   -4.632  0.000  -2.889  -1.171
                      LAct_SMS Sent   1.1964    0.077   15.528  0.000   1.045   1.347
    Occupation_Working Professional   2.5467    0.190   13.384  0.000   2.174   2.920
        Lead Profile_Potential Lead   1.4602    0.091   16.027  0.000   1.282   1.639
 Lead Profile_Student of SomeSchool  -2.3491    0.462   -5.081  0.000  -3.255  -1.443
                  N_Act_Unreachable   1.9964    0.811    2.462  0.014   0.407   3.586
```

ROC Curve and the measurement of the factors

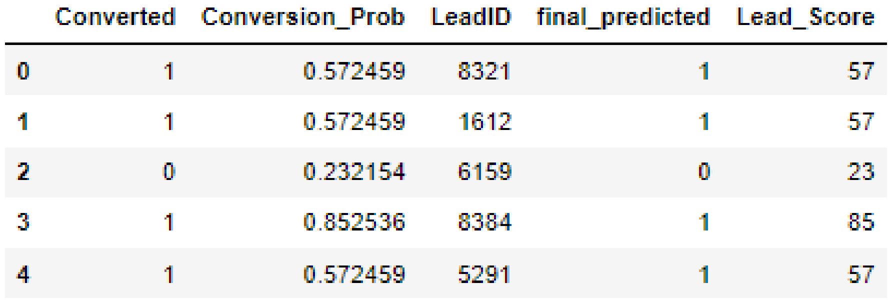


The area under the ROC curve is 0.82

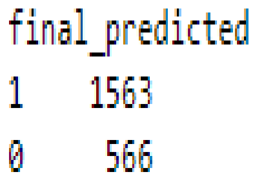


The optimal cut-off value for the three graphs is 0.48

Conversion Probability



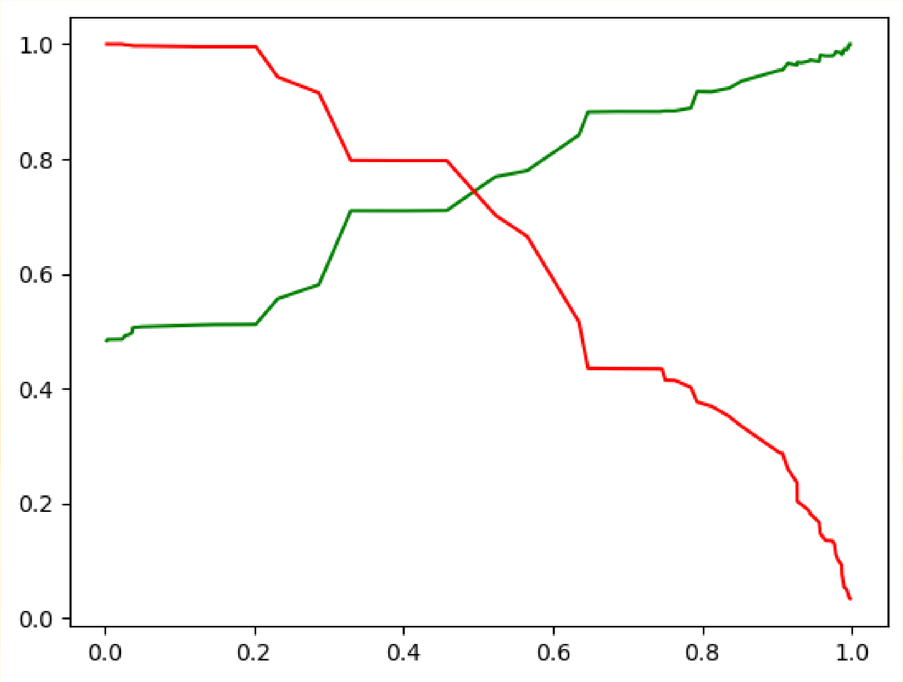
	Converted	Conversion_Prob	LeadID	final_predicted	Lead_Score
0	1	0.572459	8321	1	57
1	1	0.572459	1612	1	57
2	0	0.232154	6159	0	23
3	1	0.852536	8384	1	85
4	1	0.572459	5291	1	57

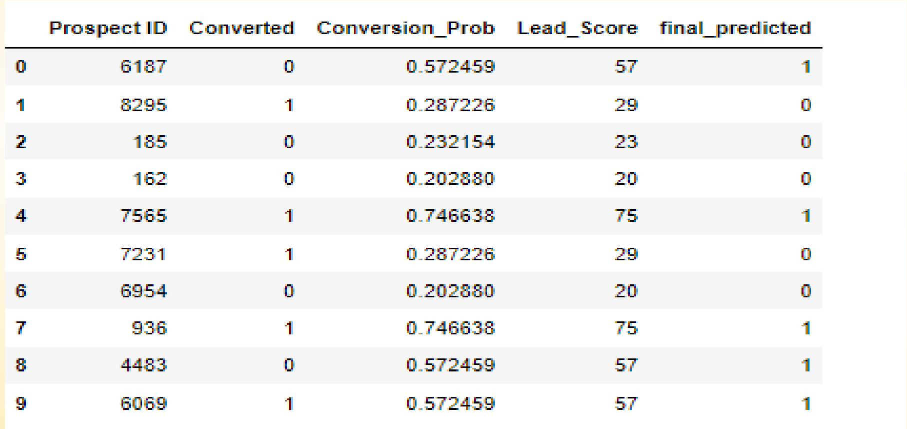


```
final_predicted
1     1563
0      566
```

The data has been converted and the lead IDs have been assigned and further after predicting the score and calculating the conversion probability, we are able to assign the Lead Score for each of the Lead IDs

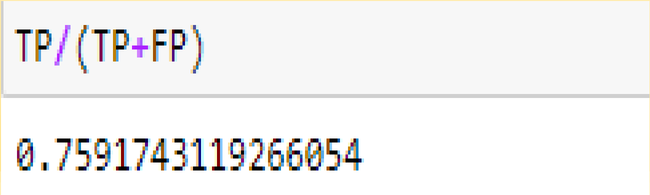
Predictions on the Test Set





	Prospect ID	Converted	Conversion_Prob	Lead_Score	final_predicted
0	6187	0	0.572459	57	1
1	8295	1	0.287226	29	0
2	185	0	0.232154	23	0
3	162	0	0.202880	20	0
4	7565	1	0.746638	75	1
5	7231	1	0.287226	29	0
6	6954	0	0.202880	20	0
7	936	1	0.746638	75	1
8	4483	0	0.572459	57	1
9	6069	1	0.572459	57	1

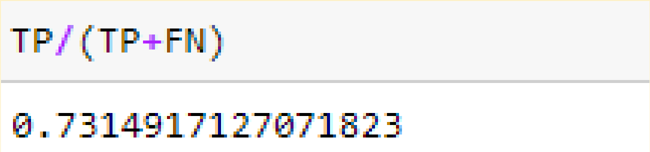
Conversion Probability Table on the Test Set



```
TP/(TP+FP)
```

```
0.7591743119266054
```

Precision



```
TP/(TP+FN)
```

```
0.7314917127071823
```

Recall

After predicting it on the test set , we have an accuracy of 76% coupled with a precision of 76% and a recall value of 73%.

THANK YOU