LEAD SCORE CASE STUDY

Submitted by :

Murali Chandra Pamujula

Akhila Kolan

Ravuri Venkata Naga Karthik

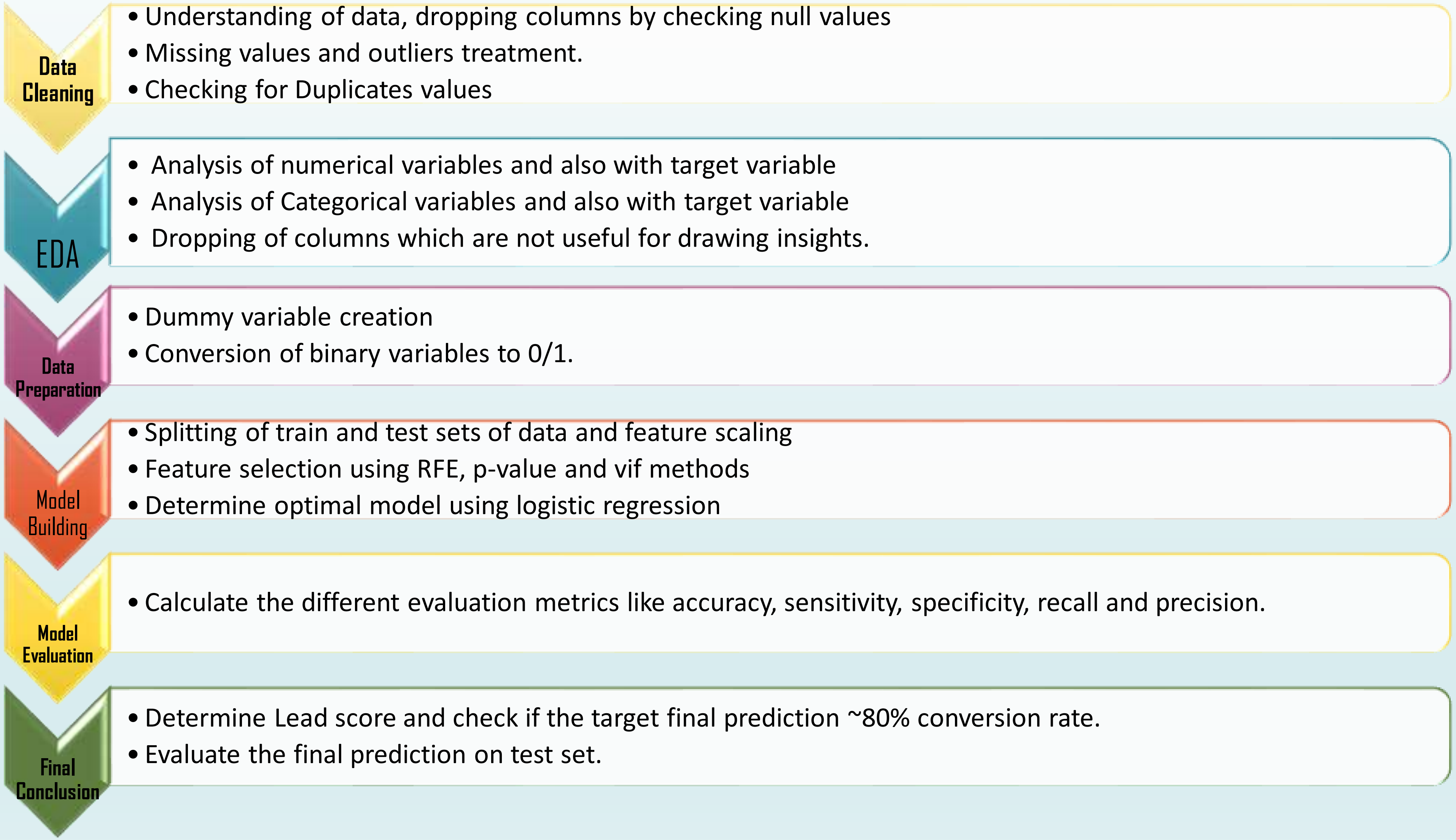
Problem Statement:

- An X Education Company sells online courses and it needs help to select the most promising leads, i.e. the leads that are most likely to be converted to paying learners so that their marketing team may be optimally utilized.
- To make this process more efficient, the company wishes to identify the most potential leads , also known as ‘Hot Leads’.
- If they successfully identify the Hot Leads, the conversion rate will increase as mentioned by the CEO and sales team can focus on the alternative methods rather than making calls to everyone.

BUSINESS OBJECTIVE :

- The company wants to know the most promising leads to increase the conversion rate.
- Assign the leads with lead score, so that they can focus on the leads with high lead score which in turn increase the conversion rate.
- To Build a model, so that the Final model should be able to adjust if the company's requirement changes in the future.

Approach followed:



Data Cleaning

- Understanding of data, dropping columns by checking null values
- Missing values and outliers treatment.
- Checking for Duplicates values

EDA

- Analysis of numerical variables and also with target variable
- Analysis of Categorical variables and also with target variable
- Dropping of columns which are not useful for drawing insights.

Data Preparation

- Dummy variable creation
- Conversion of binary variables to 0/1.

Model Building

- Splitting of train and test sets of data and feature scaling
- Feature selection using RFE, p-value and vif methods
- Determine optimal model using logistic regression

Model Evaluation

- Calculate the different evaluation metrics like accuracy, sensitivity, specificity, recall and precision.

Final Conclusion

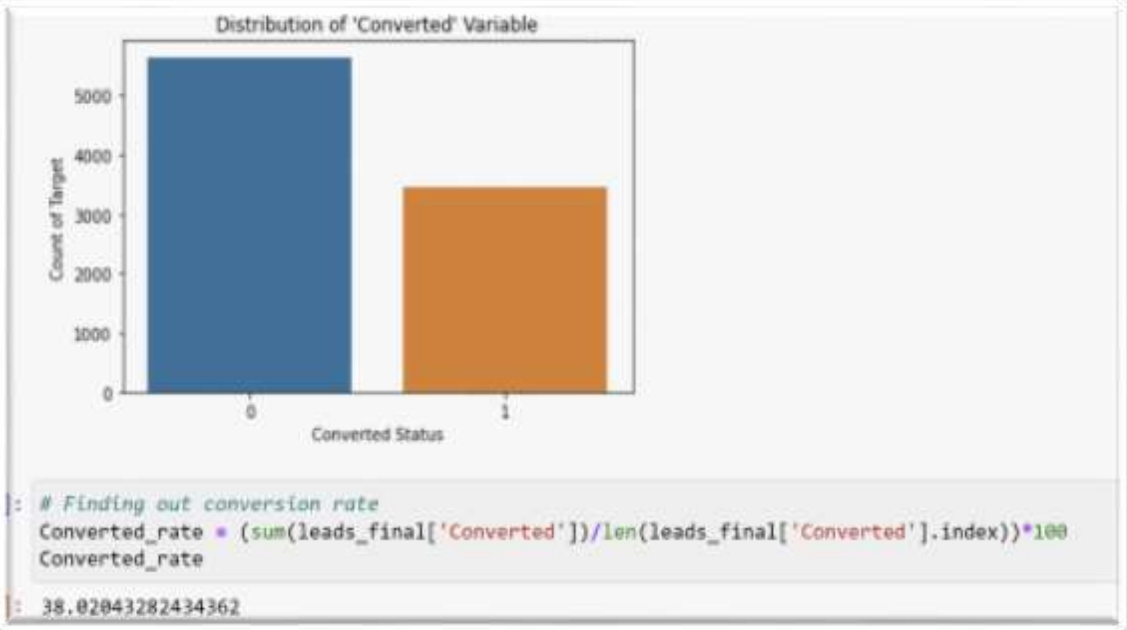
- Determine Lead score and check if the target final prediction ~80% conversion rate.
- Evaluate the final prediction on test set.

Data Cleaning :

- ❑ Initially there are 9240 rows and 37 columns in the leads dataset.
- ❑ There are no duplicate rows.
- ❑ There are missing values in some of the columns , which are dropped by taking a threshold of 35% of missing values.
- ❑ Imputed the categorical columns with the mode value for the missing values.
- ❑ Columns with less than 2% of missing values, the corresponding rows are dropped.
- ❑ Outliers are there in two numerical columns, which are treated by flooring and capping method.

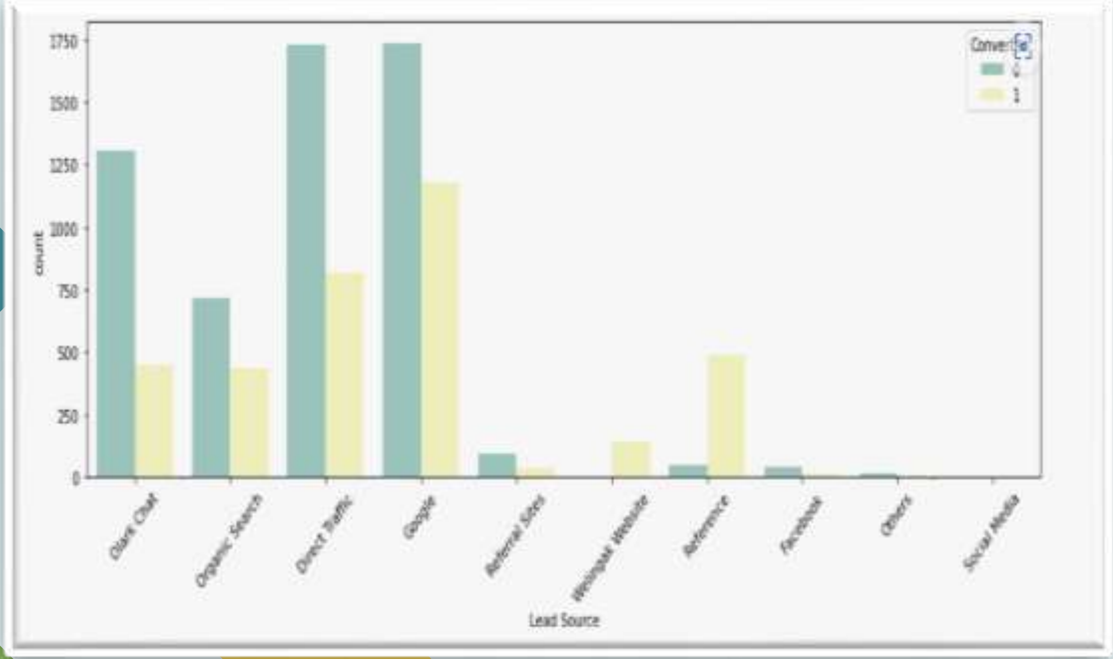
Exploratory Data Analysis :

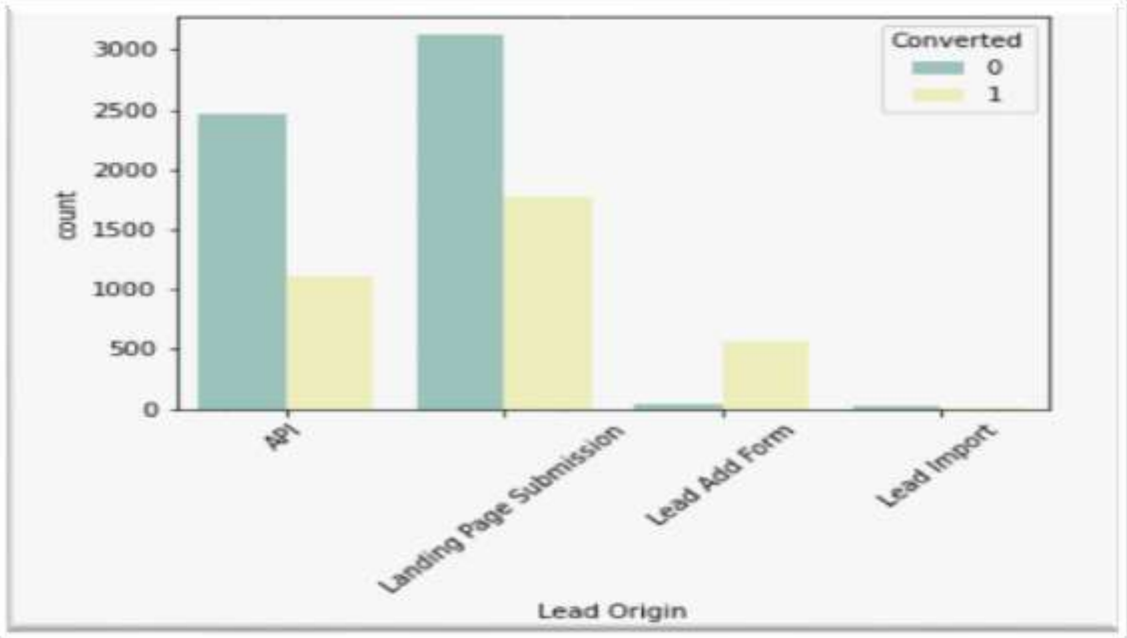
- ➢ The conversion rate is around 38%.
- ➢ The conversion rate of leads from reference and weilingak website is maximum.
- ➢ API and landing page submission have low conversion rate but leads are high.
- ➢ Lead add form origin has less leads but conversion rate is high.



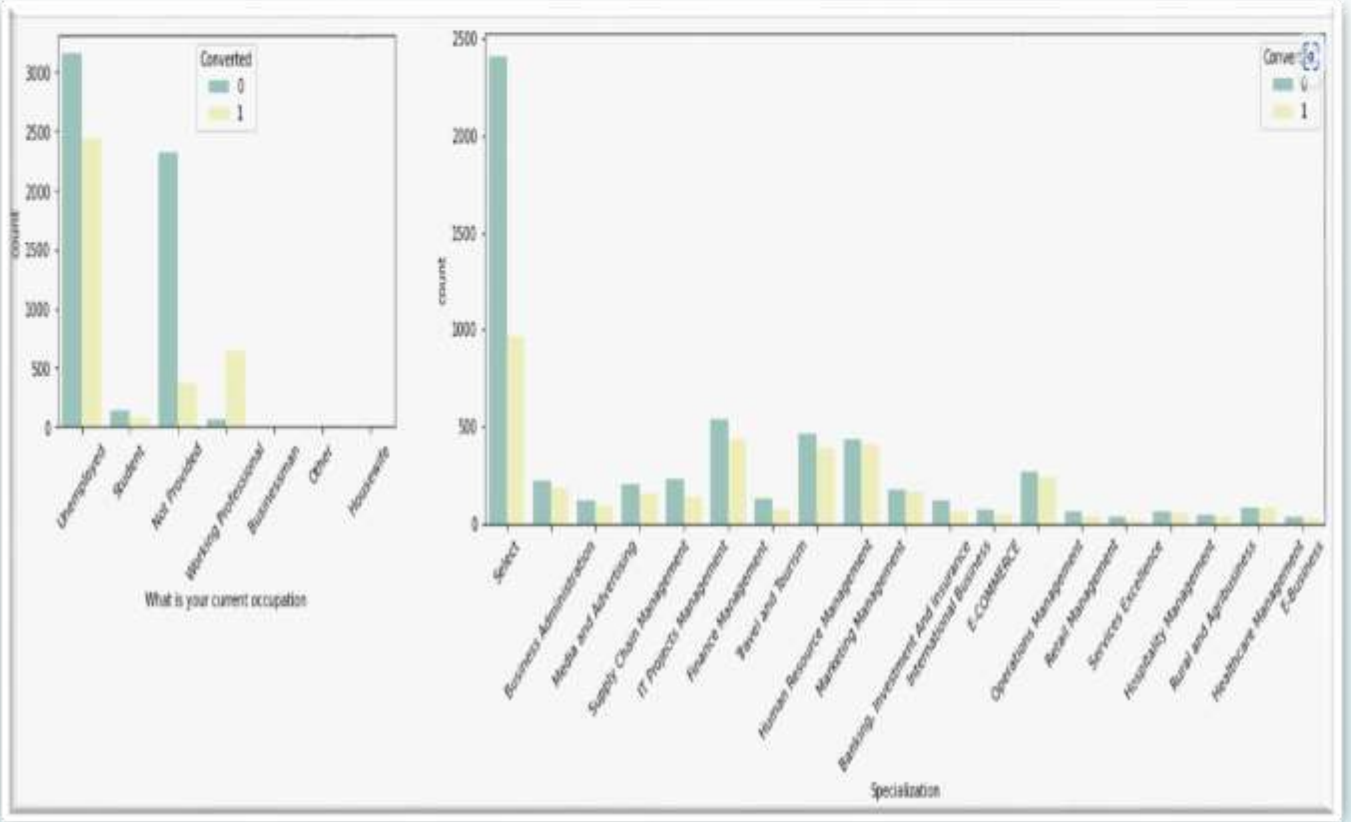
```
# Finding out conversion rate
Converted_rate = (sum(leads_final['Converted'])/len(leads_final['Converted'].index))*100
Converted_rate
```

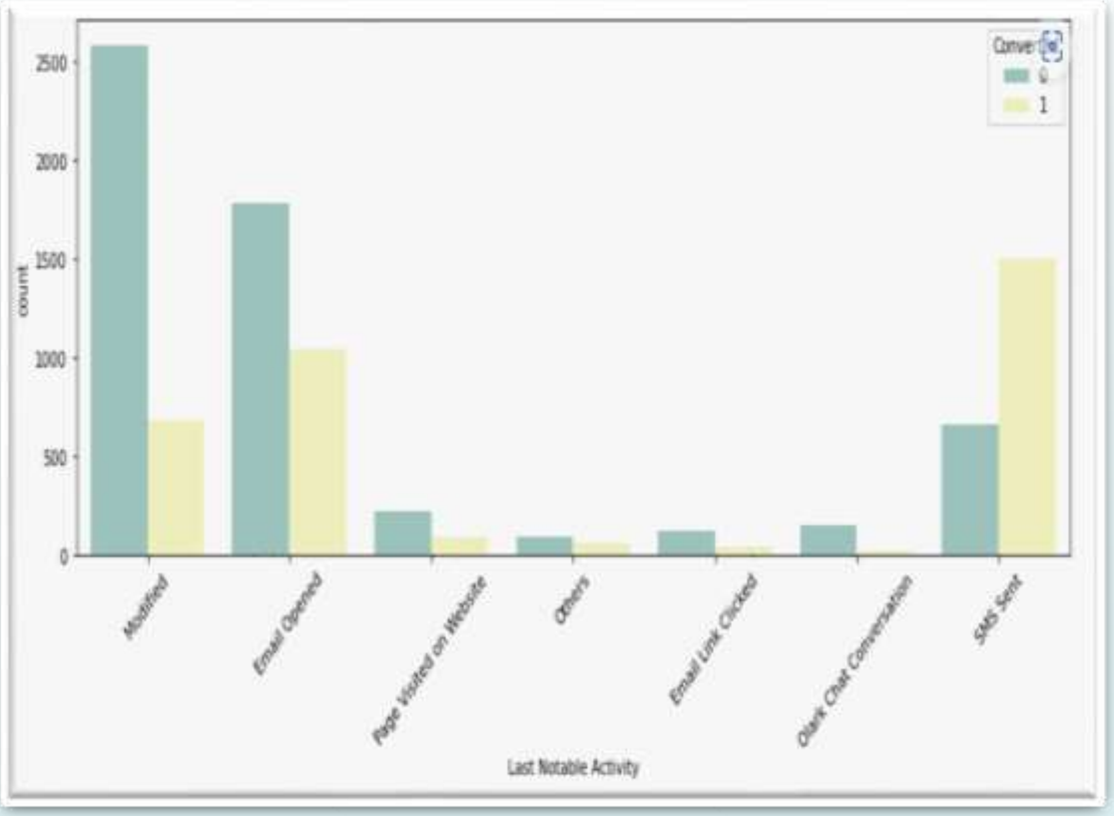
```
38.02043282434362
```





Exploratory Data Analysis :

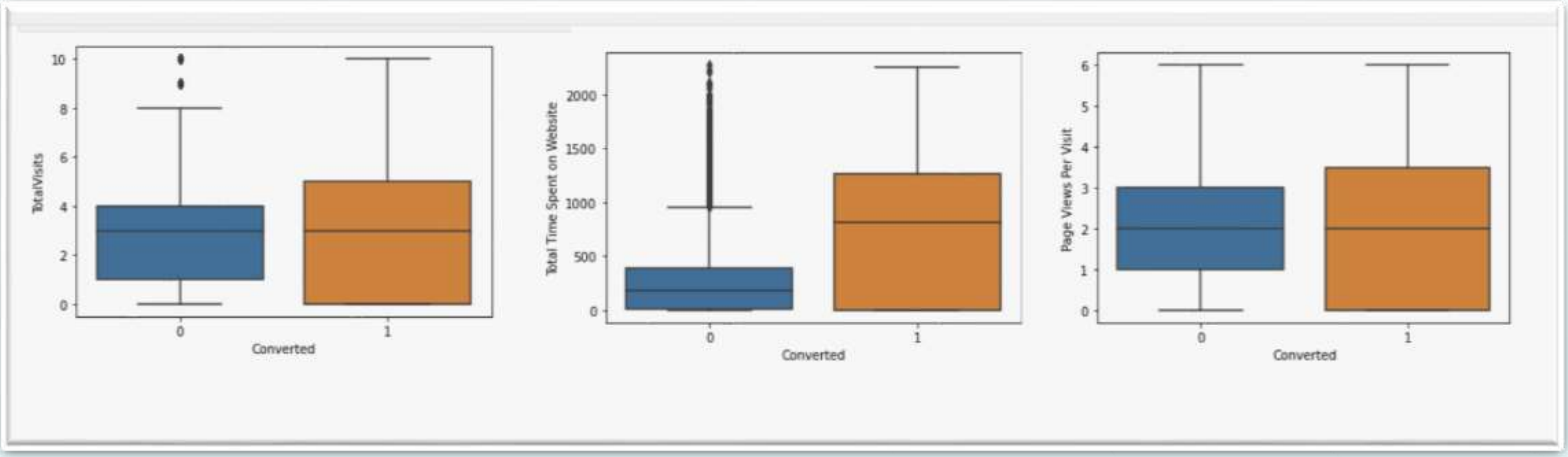




- From the above graph of specialization, no particular inference can be drawn.
- The conversion rate of working professionals from current occupation is maximum.
- Unemployed Leads are the maximum under current occupation

- SMS sent under last activity has good conversion rate.
- Modified and email opened has the maximum leads.

Exploratory Data Analysis :



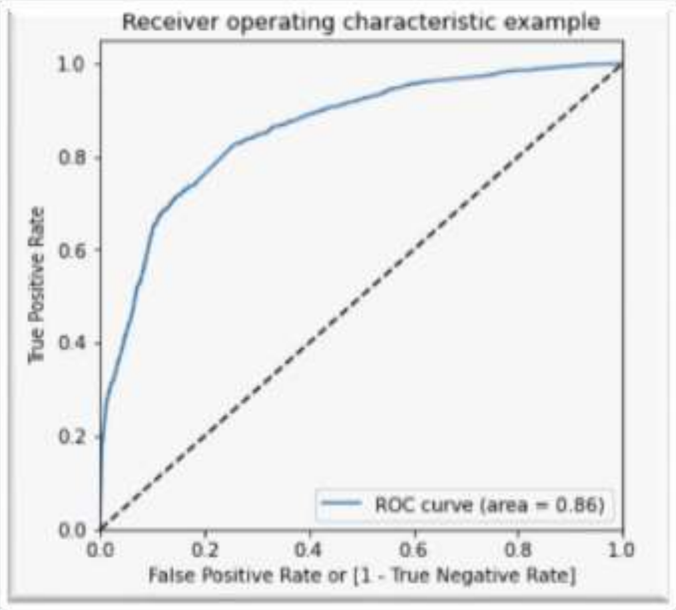
- As the median of the both converted and not converted in case of Total visits and Page views per visit columns are same, so no particular inference can be drawn
- Users spending more time on the website are likely to get converted.
- Few columns after the EDA are dropped as there is no inference is drawn.

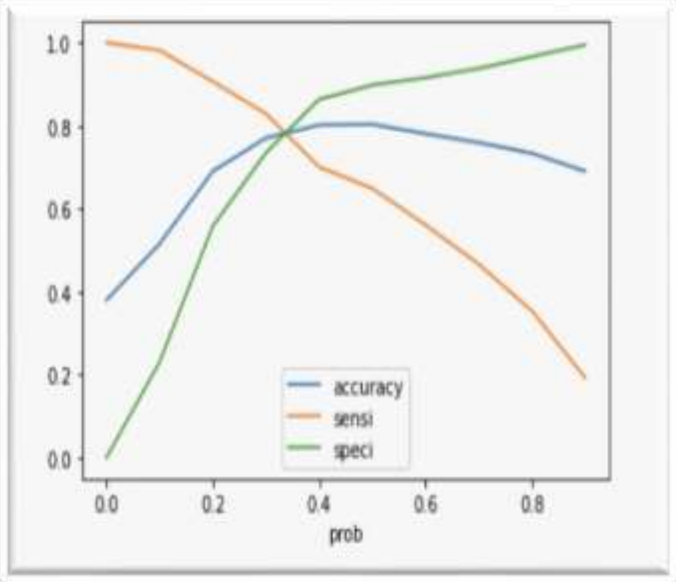
Data Preparation :

- ❑ Converted the binary variables(YES/NO) to 0/1.
- ❑ Dummy variables are created for the categorical columns.
- ❑ Total number of rows for analysis : 23 rows
- ❑ Total number of columns for analysis : 9103 columns.

Model Building :

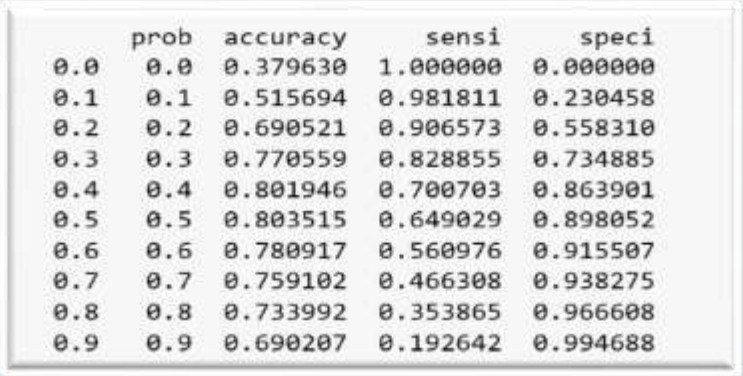
- ❑ Splitting the data into training and test sets.
- ❑ The train_test split ratio is taken as 70:30.
- ❑ The Features are scaled using a standard scaler.
- ❑ By using RFE , 15 features are selected.
- ❑ Models are build by removing the variables whose p-value > 0.05 and also checked VIF(which should be less than 5).
- ❑ ROC curve and optimal cutoff curve are drawn.
- ❑ Optimal cutoff probability can be seen from the graph as around 0.37





Model Evaluation :

- □ Calculated the accuracy, sensitivity and specificity for various probability cutoffs from 0.1 to 0.9.
- □ Confusion matrix for the training and test data are calculated.



```
     prob  accuracy     sensi     speci
0.0   0.0  0.379630  1.000000  0.000000
0.1   0.1  0.515694  0.981811  0.230458
0.2   0.2  0.690521  0.906573  0.558310
0.3   0.3  0.770559  0.828855  0.734885
0.4   0.4  0.801946  0.700703  0.863901
0.5   0.5  0.803515  0.649029  0.898052
0.6   0.6  0.780917  0.560976  0.915507
0.7   0.7  0.759102  0.466308  0.938275
0.8   0.8  0.733992  0.353865  0.966608
0.9   0.9  0.690207  0.192642  0.994688
```

TRAIN DATA-Confusion Matrix

PREDICTED ACTUAL	NOT CONVERTED	CONVERTED
NOT CONVERTED	2905	1048
CONVERTED	414	2005

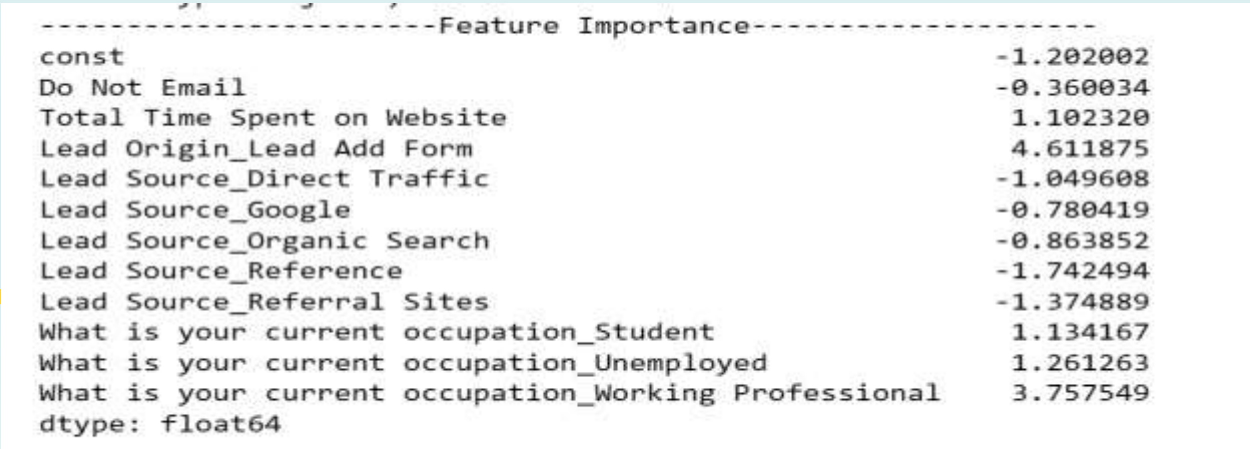
ACCURACY	77%
SPECIFICITY	73%
SENSITIVITY	83%
PRECISION SCORE	65.6%
RECALL SCORE	83%

Model Prediction :

TEST DATA-Confusion Matrix

PREDICTED ACTUAL	NOT CONVERTED	CONVERTED
NOT CONVERTED	1252	437
CONVERTED	177	865

ACCURACY	77.5%
SPECIFICITY	74%
SENSITIVITY	83%
PRECISION SCORE	66%
RECALL SCORE	83%



```
-----------------------Feature Importance-------------------
const                                                  -1.202002
Do Not Email                                           -0.360034
Total Time Spent on Website                             1.102320
Lead Origin_Lead Add Form                               4.611875
Lead Source_Direct Traffic                             -1.049608
Lead Source_Google                                     -0.780419
Lead Source_Organic Search                             -0.863852
Lead Source_Reference                                  -1.742494
Lead Source_Referral Sites                             -1.374889
What is your current occupation_Student                 1.134167
What is your current occupation_Unemployed              1.261263
What is your current occupation_Working Professional    3.757549
dtype: float64
```

Conclusion:

After the Final model, the features that need to be focused on are :

- When the lead origin is Lead Add Format.
- When the current occupation is working professional or student or unemployed.
- When the lead source was :

  a. Google b. Direct Traffic c. Organic Search d. Reference
- Users spending more time spent on website .
- For the final model, the conversion rate is obtained as around 80%(as asked by CEO) by which we can say it’s a good model.

So X Education company can now focus on these features to convert the potentials leads for buying their courses.

THANK YOU