Lead Scoring Case Study

SHEETAL SAXENA
RUPALI
ABHINEET SEHGAL

Problem Statement

An education company named 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.

The company requires us to build a model wherein we need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.

The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Objectives

- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.
- There are some more problems presented by the company which the model should be able to adjust to if the company's requirement changes in the future so need to handle these as well.
- Need to make recommendations based on the logistic regression model.

Data Understanding

- The provided *Leads* dataset has around 9000 data points. This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. which may or may not be useful in ultimately deciding whether a lead will be converted or not.

- The target variable, in this case study, is the column 'Converted' which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn't converted.

Case Study Approach

I. Data Preparation, Cleaning and EDA

II. Test-train Split and Scaling

III. Model Building

IV. Model Evaluation

V. Predictions on the Test Dataset

VI. Conclusion

Data Preparation, Cleaning & Exploratory Data Analysis

STEP 1: IMPORTING DATA

STEP 2: INSPECTING THE DATAFRAME

STEP 3: DATA CLEANING

STEP 4: UNIVARIATE ANALYSIS

STEP 5: BIVARIATE ANALYSIS

STEP 6: MULTIVARIATE ANALYSIS

STEP 7: DATA PREPARATION

Importing and Inspecting the data

- There are 9240 rows and 37 columns in the dataframe.
- There are 30 categorical and 7 numeric columns in the dataframe.

Data Cleaning

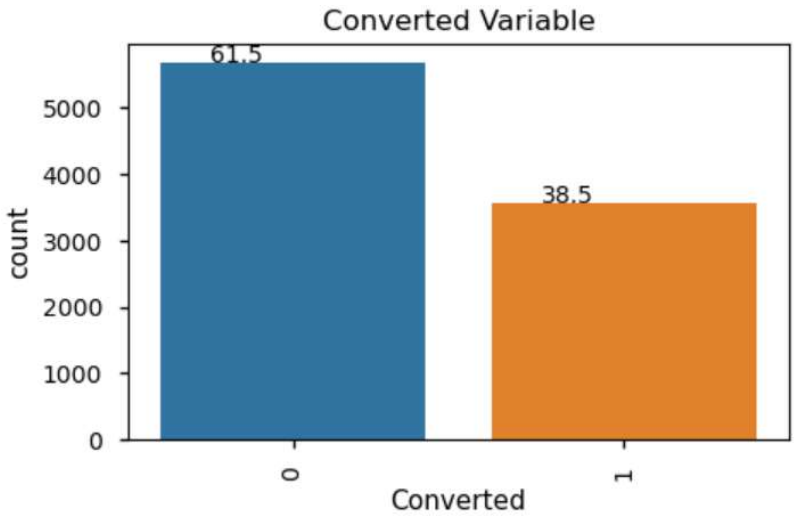
- Missing Values Analysis (Percentage Wise)

Lead Quality	51.59
Asymmetrique Profile Score	45.65
Asymmetrique Activity Score	45.65
Asymmetrique Profile Index	45.65
Asymmetrique Activity Index	45.65
Tags	36.29
Lead Profile	29.32
What matters most to you in choosing a course	29.32
What is your current occupation	29.11
Country	26.63
How did you hear about X Education	23.89
Specialization	15.56
City	15.37
TotalVisits	1.48
Page Views Per Visit	1.48
Last Activity	1.11
Lead Source	0.39

Data Cleaning

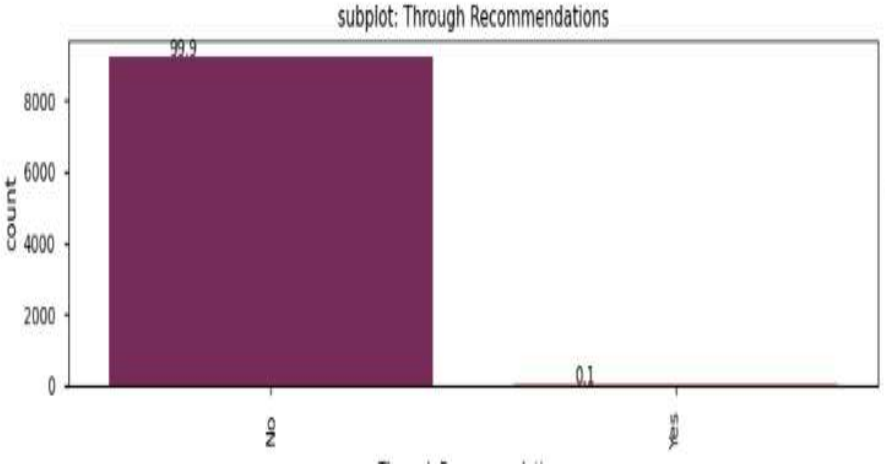
- After analyzing the null value percentage of all columns present in Leads Dataset, 75 percentile of columns have null percentage less than 45%.
- Remaining 25 percentile columns have more than 45% of null values.
- We have dropped the columns having Missing Values Percentage more than 45%.
- The 'Select' values for the variables *Lead Profile, How did you hear about X Education, Specialization* and *City* have been handled by treating them as Unknown. Later in Data Preparation Stage, we have dropped dummy variable corresponding to Unknown values.
- The Missing/Null Value handling is done as per the Null Value Percentage and the datatype of the variable.
- For the Numeric Variables, the data imputation is done by median for that variable.

Univariate Analysis

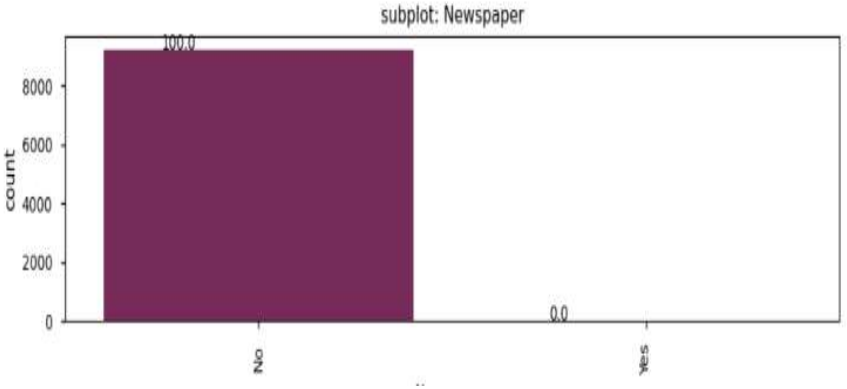


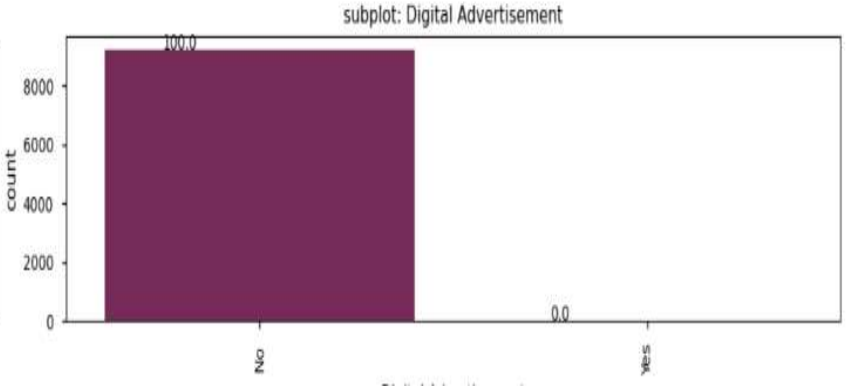
- As we can infer from the graph , 38.5% of leads have been converted.
- 61.5% of the entire Leads data have not been successfully converted.

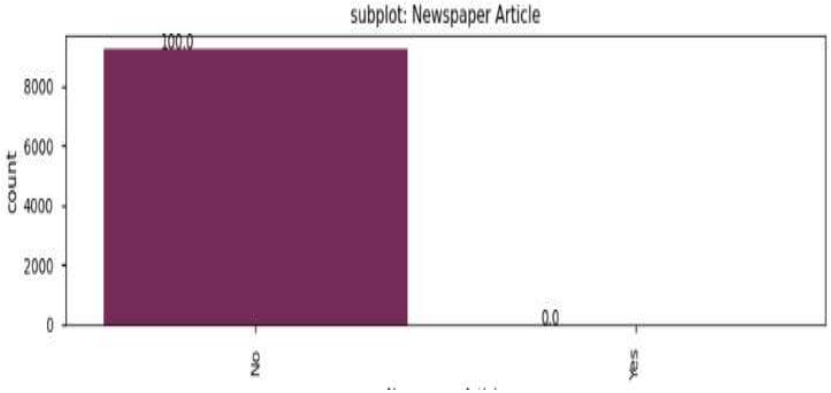
Univariate Analysis

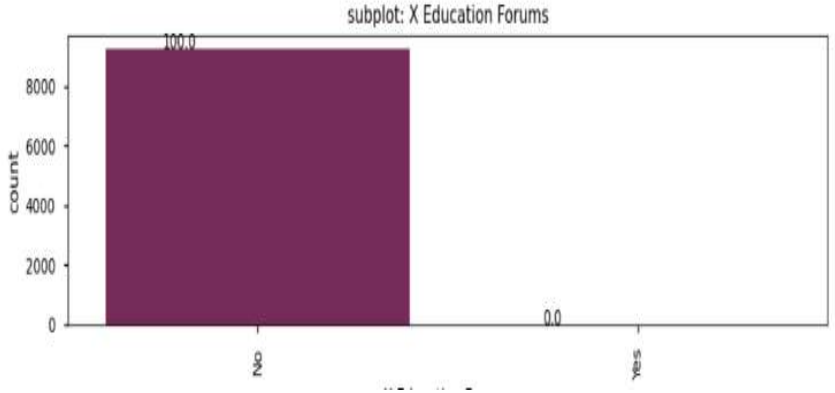


- *Through Recommendations* Variable have highly skewed data and have constant values almost across all rows


subplot: Newspaper
count
100.0
0.0
No
Yes


subplot: Digital Advertisement
count
100.0
0.0
No
Yes


subplot: Newspaper Article
count
100.0
0.0
No
Yes


subplot: X Education Forums
count
100.0
0.0
No
Yes

Univariate Analysis

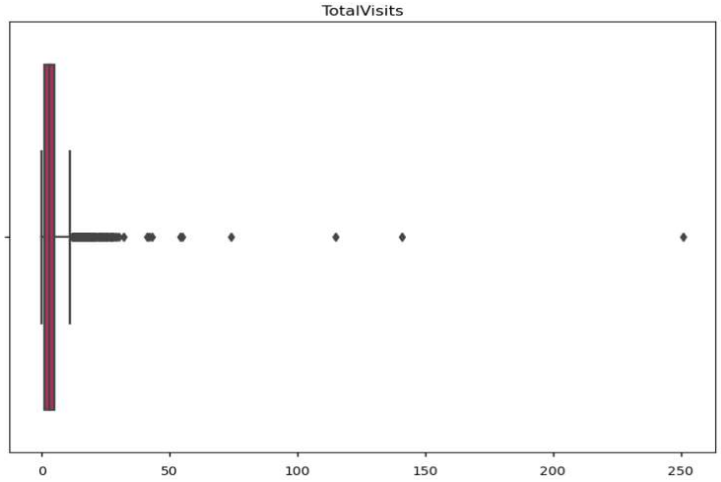
The following columns have highly skewed data and have constant values almost across all rows:

- *Through Recommendations*
- *Digital Advertisement*
- *Newspaper*
- *Newspaper Article*
- *X Education Forums*
- *Search*
- *Do not Call*
- *Do not Email*

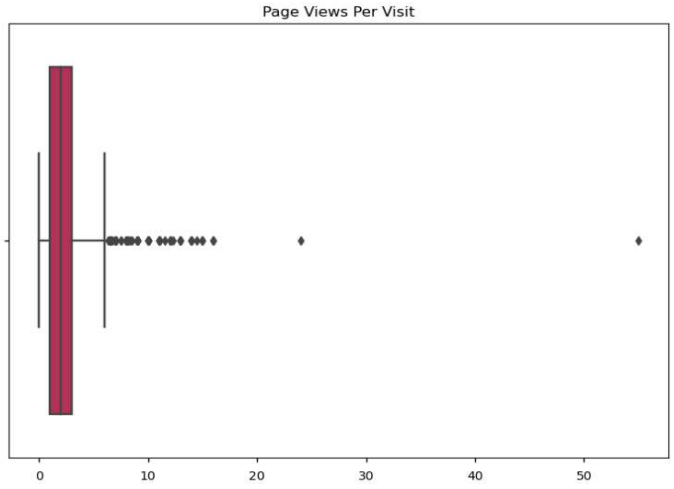
We can drop these columns as they will not add much value to the model

Univariate Analysis - Outliers

TotalVisits Variable

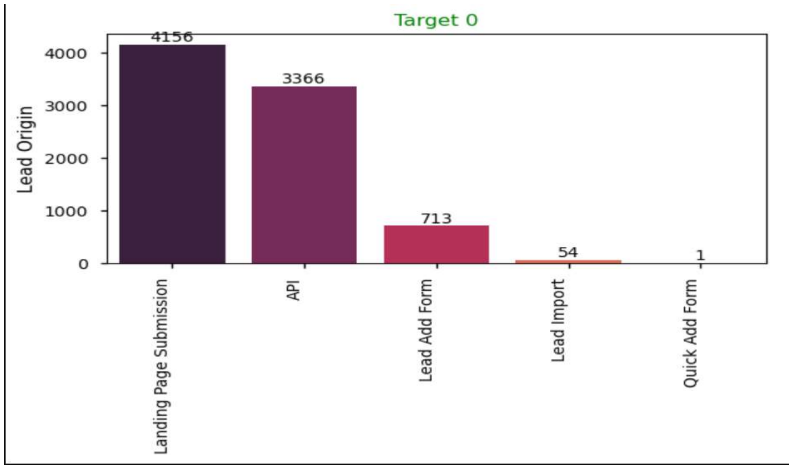


Page Views Per Visit Variable



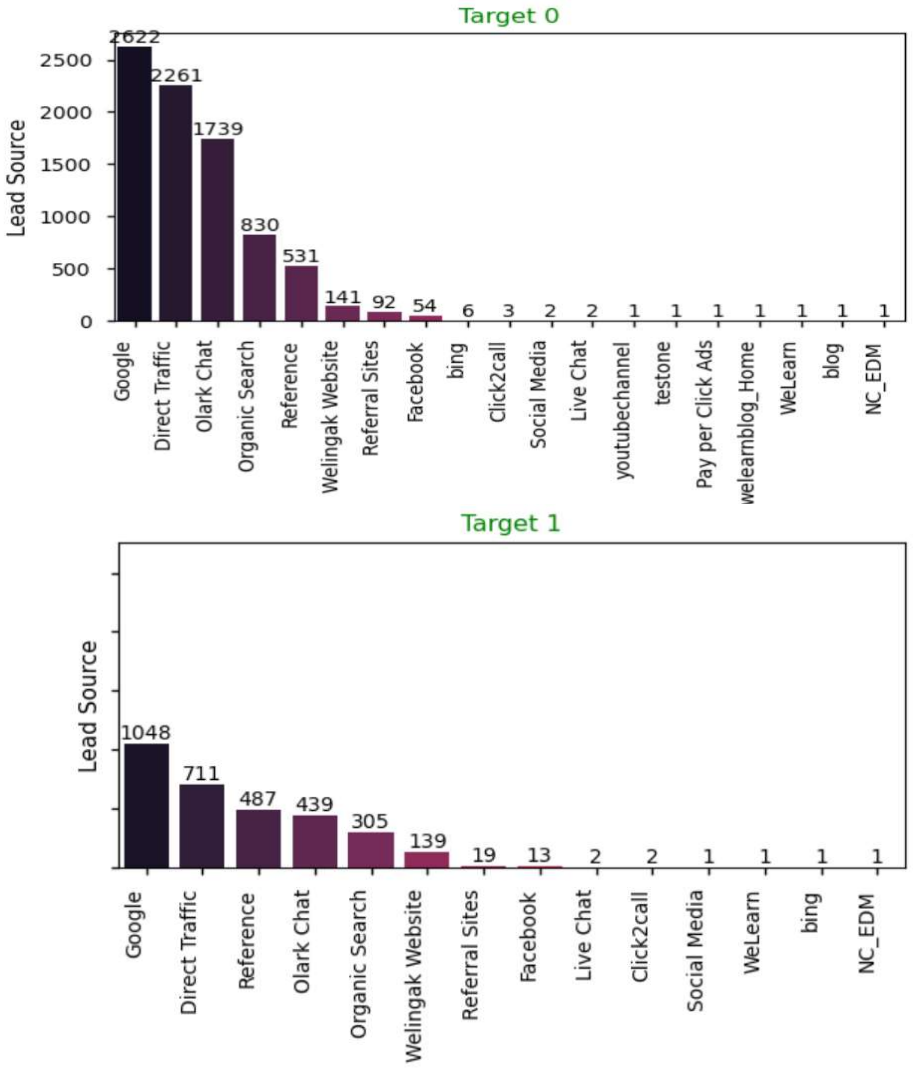
Bivariate Analysis

- For the leads which got converted, Lead origin was maximum through Page submission and the least was Quick Add form
- For the leads which could not get converted , lead origin was maximum through Page submission and least through Quick Add form



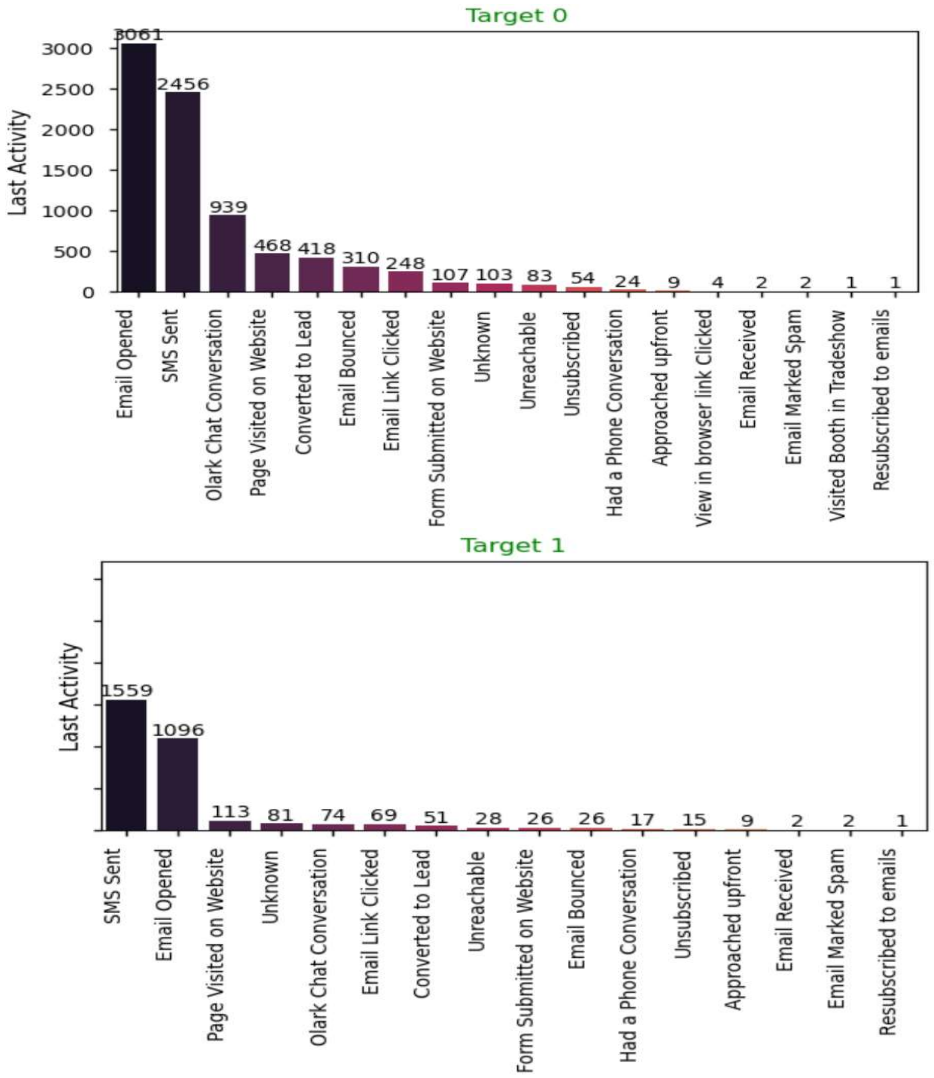
Bivariate Analysis

For Lead Source variable, maximum traffic is received via Google in case of both Lead Conversion and Lead Non-Conversion



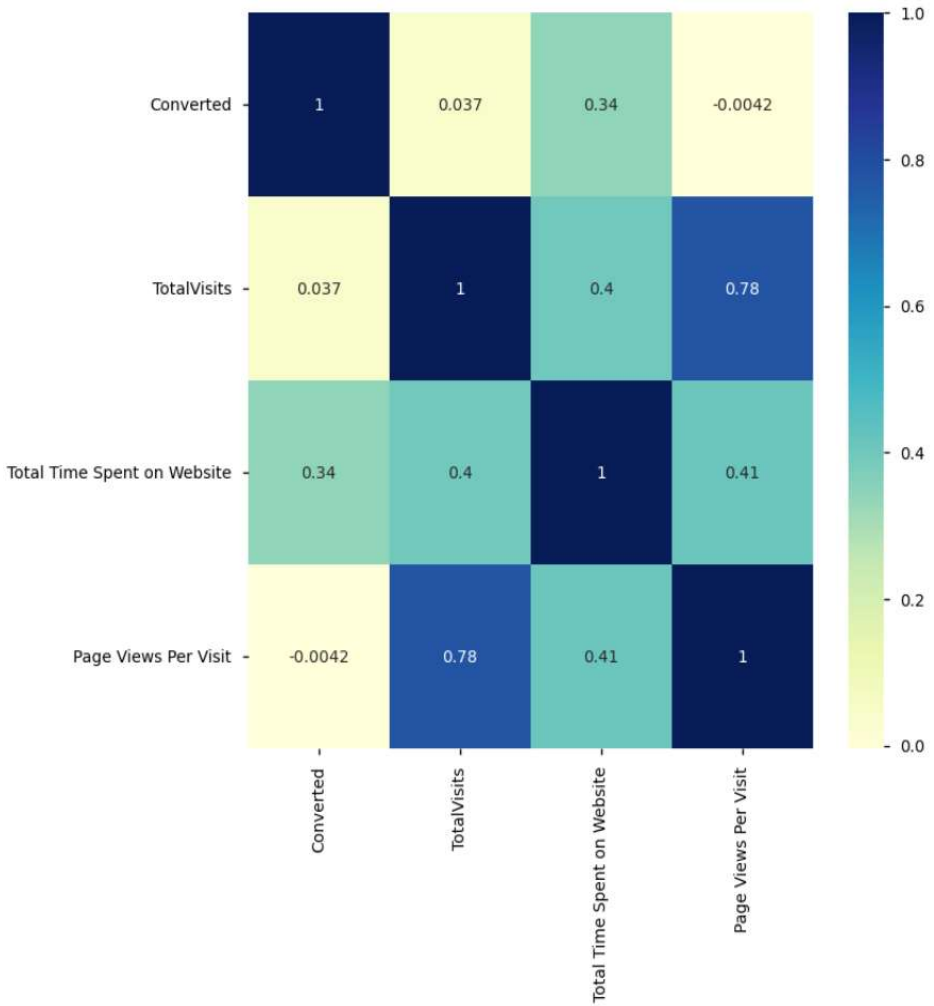
Bivariate Analysis

- For Lead Non-Conversion Cases, the maximum Last Activity is Email Opened
- For Lead Conversion Cases, the maximum Last activity is SMS Sent.



Multivariate Analysis

- There is a strong correlation between Total Visits and Pages Views Per visit.
- There is a correlation between Total Visits and Converted.
- There is a negative correlation between Pages Views Per visit and Converted.



Test-train Split and Scaling

STEP 1: TEST-TRAIN SPLIT

STEP 2: FEATURE SCALING

Test-train Split and Scaling

- Step 1: Test-Train Split

  The dataset is split into Training and Testing Data in the ratio of 70:30 .

- Step 2: Feature Scaling
  - The Feature Scaling is done by using *StandardScalar* function.
  - For training data, *fit_transform* function is used.
  - For testing data, *transform* function is used.

Model Building

STEP 1: FEATURE ELIMINATION BASED ON CORRELATIONS

STEP 2: FEATURE SELECTION USING RFE (COARSE TUNING)

STEP 3: MANUAL FEATURE ELIMINATION (USING P-VALUES AND VIFS)

Model Building

- Step 1: Feature elimination based on correlations

  After analysis of the heatmap, the dummy following variables having high correlation (Considering Absolute correlation > 0.8):

  - Lead Origin_API and Lead Origin_Landing Page Submission
  - Lead Origin_Lead Add Form and Lead Source_Reference
  - Notable Activity_Email Opened and Last Activity_Email Opened
  - Notable Activity_SMS Sent and Last Activity_SMS Sent

  We have dropped one variable from each pair.

Model Building

- Step 2: Feature selection using RFE (Coarse Tuning)

  We have selected top 15 Variables for our model using RFE

- Step 3: Manual feature elimination (using p-values and VIFs)

  We have eliminated the variables having VIF > 5 or p-value > 0.05 from our model

Model Evaluation

STEP 1: ACCURACY

STEP 2: SENSITIVITY AND SPECIFICITY

STEP 3: OPTIMAL CUT-OFF USING ROC CURVE

STEP 4: PRECISION AND RECALL

Accuracy & Confusion Matrix

```
## Accuracy Score for the model
accuracy(lm_pred2)
```

```
0.9103911769774254
```

```
## Confusion Matrix
confusion = confusion(lm_pred2)
```

```
confusion
```

```
array([[3448,  101],
       [ 419, 1835]], dtype=int64)
```

Sensitivity & Specificity

```
TP = confusion[1,1] # true positive
TN = confusion[0,0] # true negatives
FP = confusion[0,1] # false positives
FN = confusion[1,0] # false negatives
```

```
## Sensitivity Calculation
sensitivity = TP / float(TP+FN)
sensitivity
```

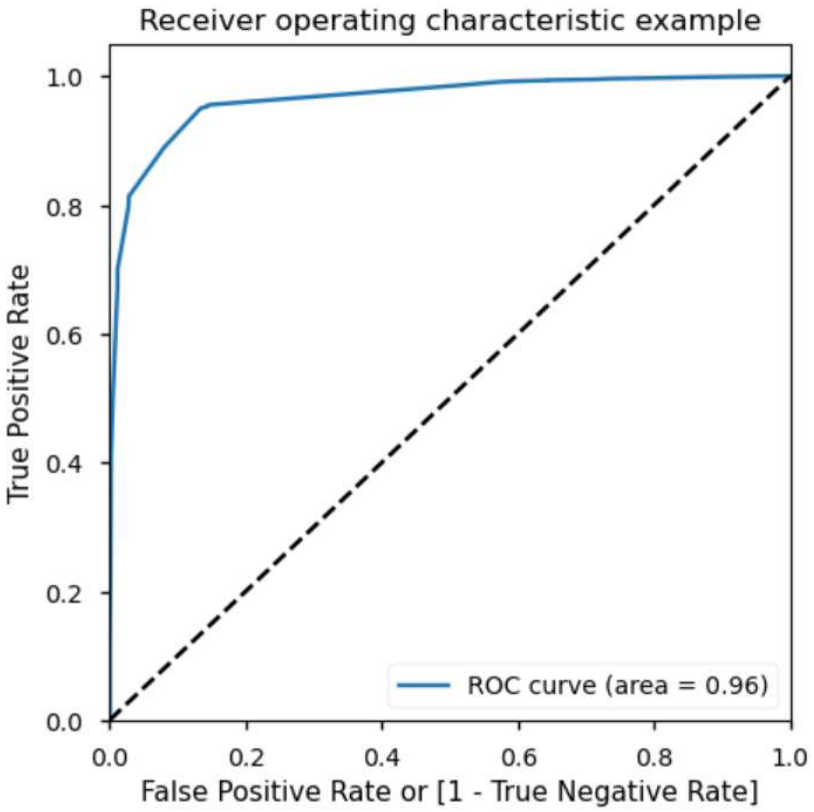
```
0.8141082519964508
```

```
## Specificity Calculation
specificity = TN / (TN+FP)
specificity
```

```
0.9715412792335869
```

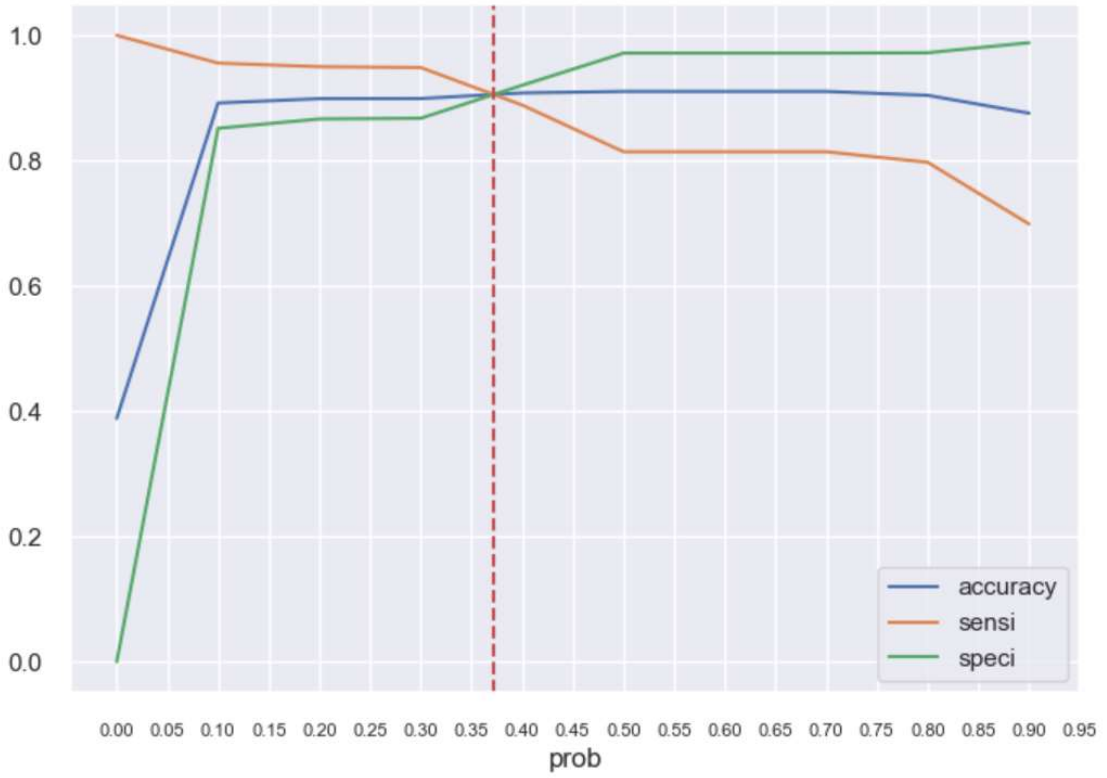
ROC curve

Area under ROC Curve = 0.96



ROC curve – Optimal Cutoff Point

The Optimal Cutoff value is 0.372.



Precision and Recall

```
## confusion matrix
confusion
```

```
array([[3448,  101],
       [ 419, 1835]], dtype=int64)
```

```
precision = confusion[1,1]/( confusion[0,1] + confusion[1,1])
precision
```

```
0.9478305785123967
```

```
recall = confusion[1,1]/( confusion[1,0] + confusion[1,1])
recall
```

```
0.8141082519964508
```

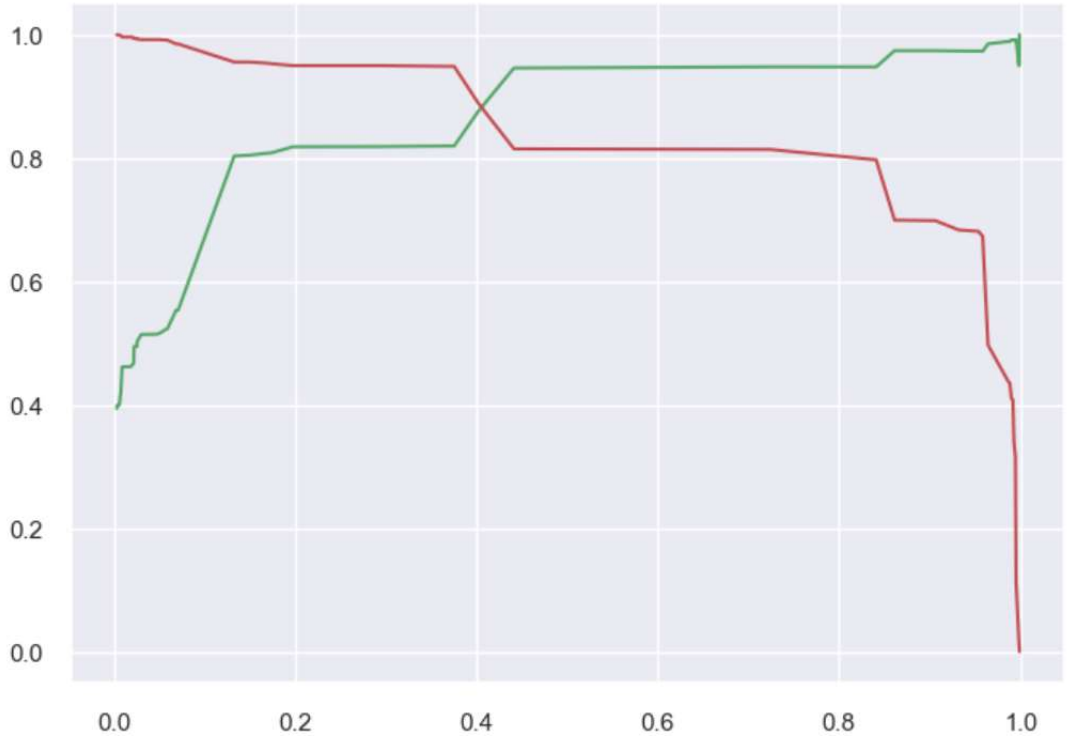
```
precision_score(lm_pred2.Conversion ,lm_pred2.Predicted)
```

```
0.9478305785123967
```

```
recall_score(lm_pred2.Conversion ,lm_pred2.Predicted)
```

```
0.8141082519964508
```

Precision Recall Curve



Predictions on the Test Dataset

THE PREDICTIONS ARE MADE FOR TEST DATASET ON THE BASIS OF DEVELOPED LOGISTIC REGRESSION TRAINING MODEL.

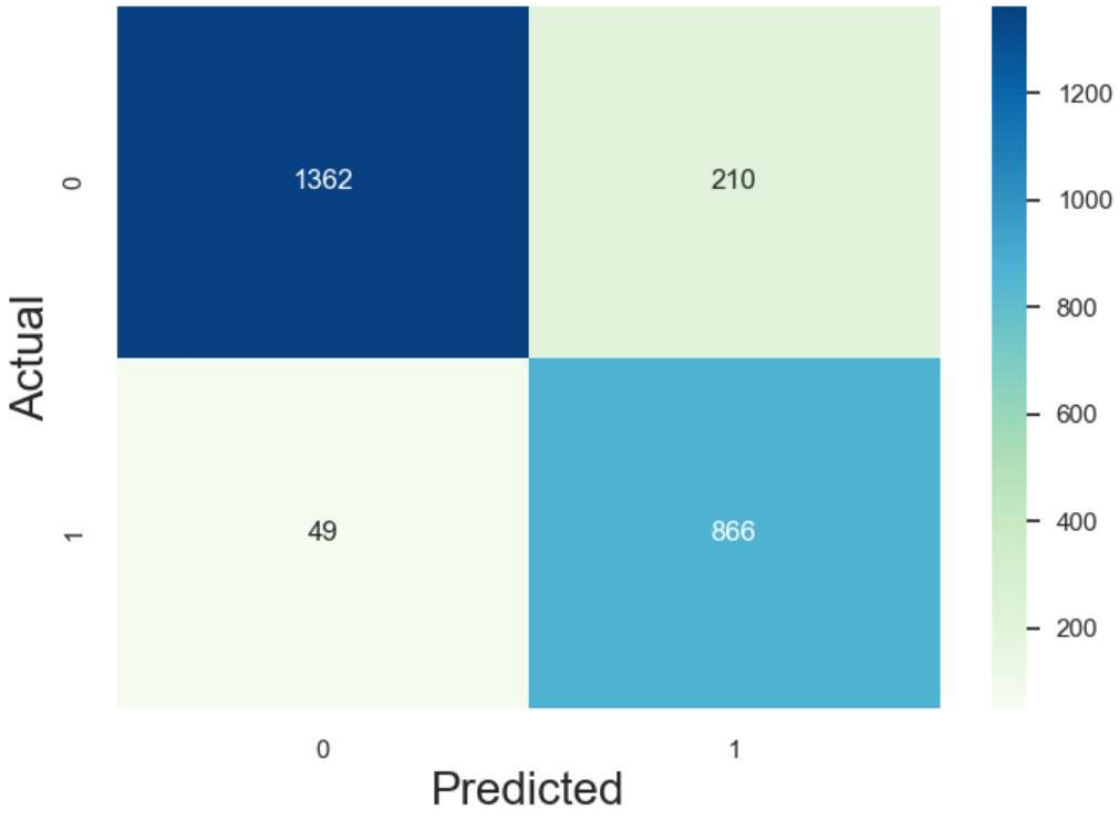
Test Data Prediction Accuracy

```
## we use the sensitivity , specificity and accuracy cutoff , for this model it is 0.372
y_pred_final['final_Predicted'] = y_pred_final.Conversion_Probability.map(lambda x: 1 if x > 0.372 else 0)
```

```
metrics.accuracy_score(y_pred_final.Converted, y_pred_final.final_Predicted)
```

```
0.895858464012867
```

Test Data Confusion Matrix

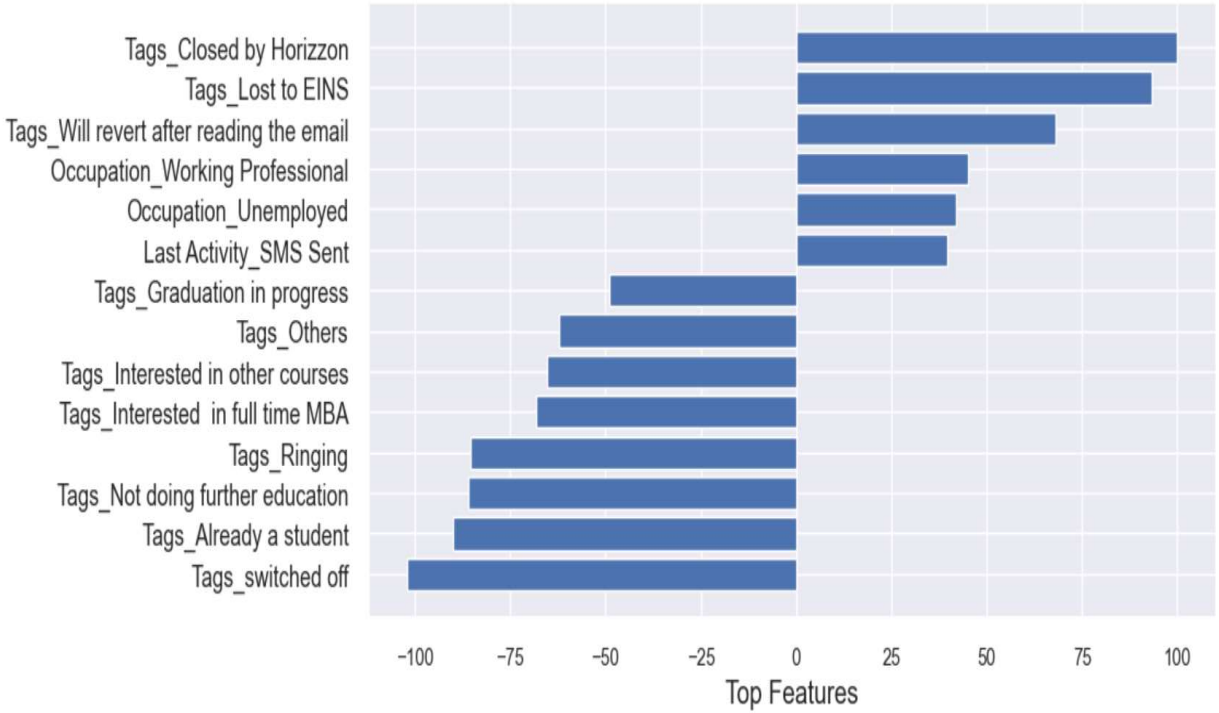


Conclusion

FINAL OBSERVATION

RECOMMENDATIONS

Final Observation – Top Features



Final Observation – Evaluation Metrics for Test Data

Metrics	Value
Accuracy	0.90
Sensitivity	~ 0.95
Specificity	0.87
Precision	0.80
Recall	0.95

Recommendations

To improve the potential lead conversion rate, X-Education will have to mainly focus important features which are responsible for good conversion rate. Some of them are as follows:

- ***Occupation_Working Professional:*** The leads whose occupation is 'Working Professional' have higher lead conversion rate. The company should focus on working professionals and try to get more number of leads.
- ***Last Activity_SMS Sent:*** Leads whose last activity is sms sent can be potential leads for company.
- ***Occupation_Unemployed:*** The leads whose Occupation is 'Unemployed' have a good conversion rate. Company can target this feature to get more leads as in order to get employed people want to upskill themselves.
- ***Tags will revert after reading the mail*** - This feature also has a good Conversion rate .