Lead_Scoring_Case_Study

Contents

- Background of X Education Company
- Problem Statement & Objective of the Study
- Suggested Ideas for Lead Conversion
- Analysis Approach
- Data Cleaning
- EDA
- Data Preparation
- Model Building (RFE & Manual fine tuning)
- Model Evaluation
- Recommendations

Background of X Education Company

● Industry professionals can purchase online courses from X Education, an education firm.
● A lot of professionals who are interested in the courses visit their website and look through the offerings on any given day.
The organization promotes its courses on multiple websites and search engines such as Google.
● After visiting the website, these individuals may peruse the available courses, complete the course registration form, or see some videos.
● These persons are categorized as leads when they complete a form with their phone number or email address.
● After these leads are obtained, sales team members begin calling, sending emails, etc.
During this procedure, a small percentage of leads convert, whilst the majority do not.
●The average conversion rate of leads at X education is around 30%.

Problem Statement & Objective of the Study

- Although X Education receives a lot of leads, its lead conversion rate is just about 30%, which is extremely low.
- X Education seeks to increase the efficiency of the lead conversion process by locating the "Hot Leads," or most prospective leads.
- Rather of calling everyone, their sales staff would prefer to know about this possible group of prospects with whom they may better communicate.

The study's goal is to assist X Education in identifying the most promising leads—that is, the leads with the highest likelihood of becoming paying clients.

- The organization mandates that we develop a model in which we must give each lead a score so that the clients who have higher lead scores are more likely to convert.

Suggested Ideas for Lead Conversion

- Leads are categorized according to how likely they are to convert.
- A targeted collection of hot leads is the end result.
- Our ability to communicate with fewer leads would enable us to make a bigger impression.
- Since our target conversion rate is 80%, we would like to acquire a high sensitivity in obtaining hot leads. • We would have a higher conversion rate and be able to hit the 80% objective since we focused on hot leads that were more likely to convert.

Analysis Approach

- Data Cleaning
- EDA
- Data Preparation
- Model Building
- Model Evaluation
- Predictions on test data
- Recommendation

Data Cleaning

● For certain categorical variables, the "Select" level denotes null values because no choice was selected by the customers from the list.

● Columns that had null values greater than 40% were removed.

● Value counts and specific considerations were used to handle missing values in category columns.

● Remove columns (tags, country) that don't provide any information or value to the study's goal.

A few categorical variables were imputable.

For certain variables, more categories were made.

● Only one response category was retained, and columns like Prospect ID and Lead Number that were useless for modeling were eliminated.

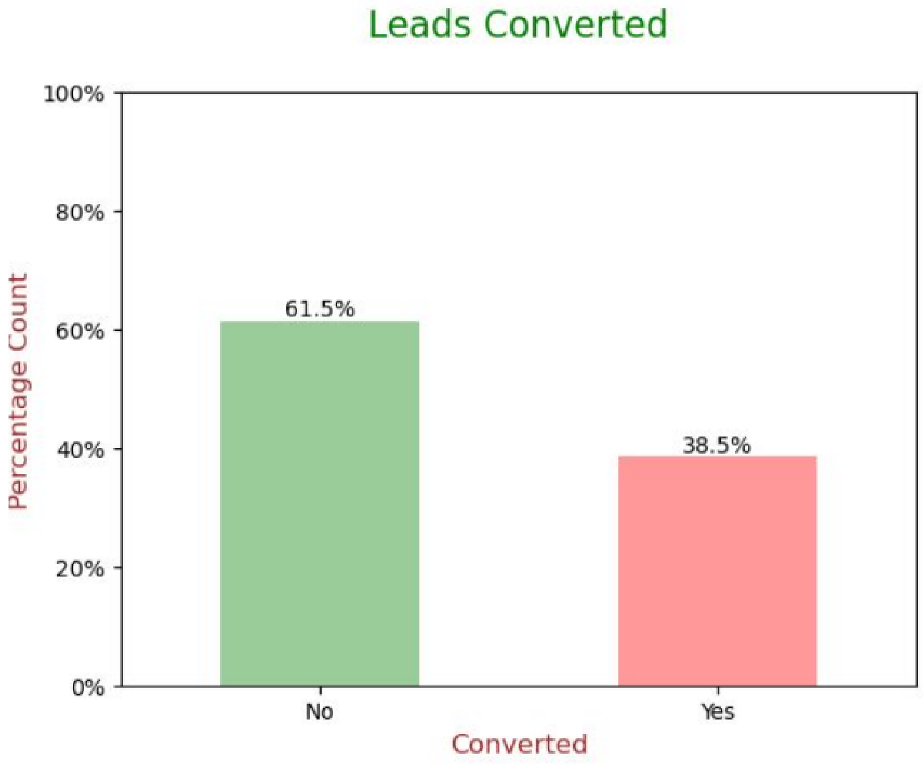
● After verifying the distribution, numerical data was imputed using the mode.

Data Cleaning

● Logistic regression models were tested for bias by checking and removing skewed category columns.
● Page views per visit and total visits outliers were handled and capped.
● Incorrect values were corrected, and information in some columns—like lead source—was standardized.
The low frequency values were categorized as "Others".
Categorical binary variables were mapped.
● Further cleansing procedures were carried out to guarantee the precision and quality of the data.
● Fixed invalid values and checked case styles, among other things, to standardize data in columns. (Google is the lead source.)

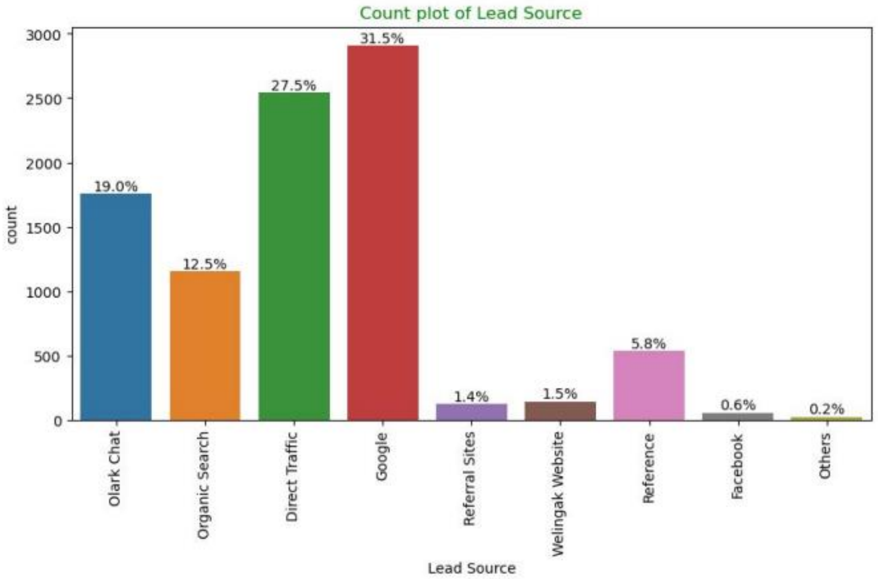
EDA

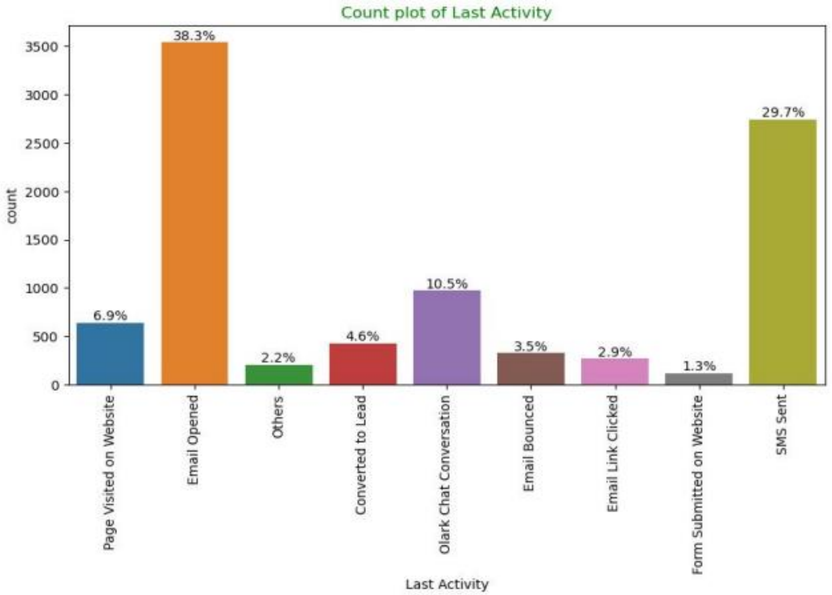
- The conversion rate is 38.5%, which indicates that just 38.5% of the visitors become leads.(Relatively minor)
- However, 61.5% of the visitors did not become leads



EDA

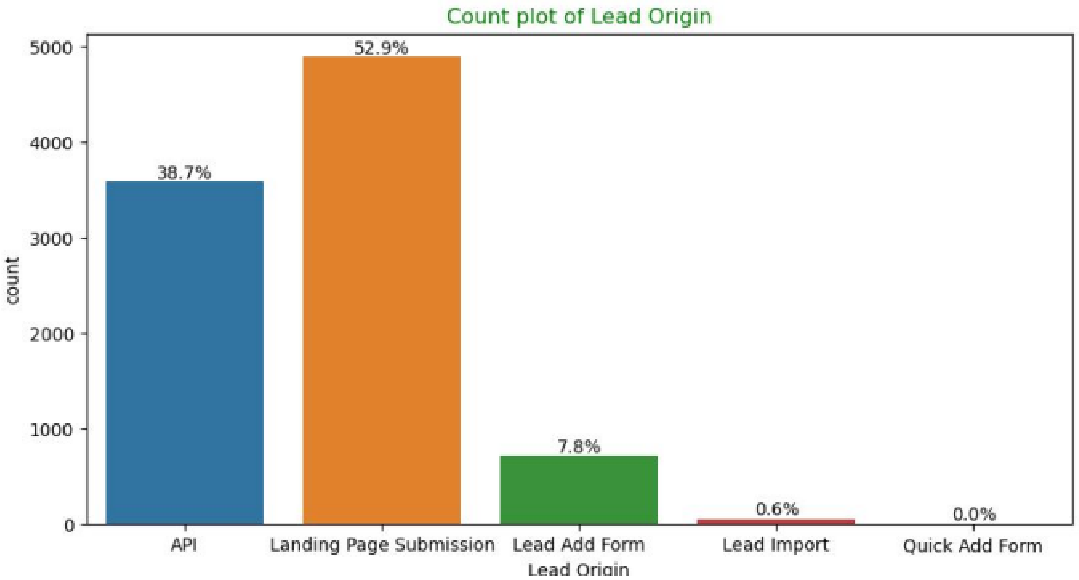
- Univariate Analysis – Categorical Variables

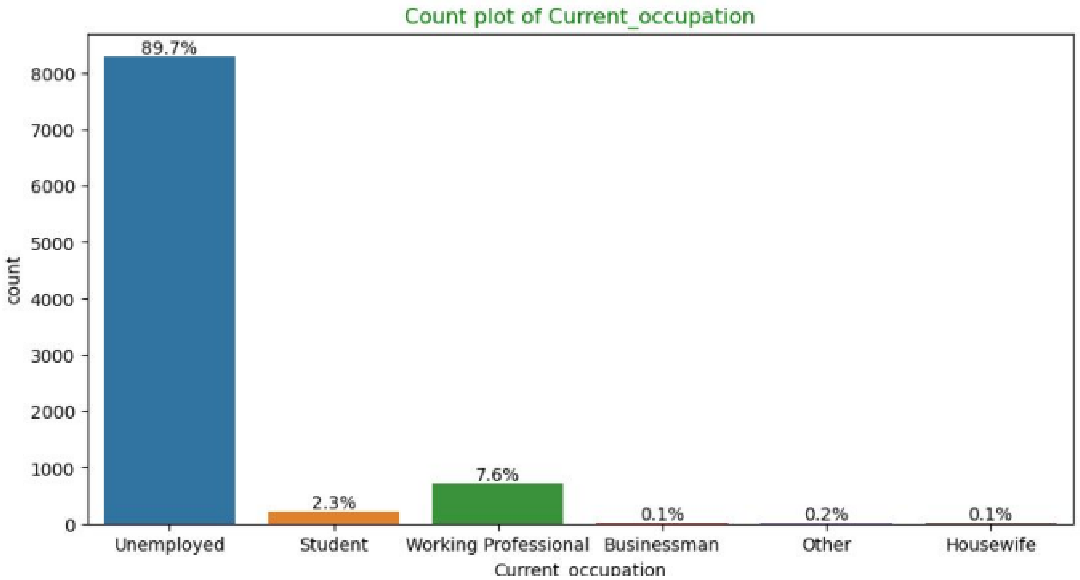




- 58% Lead source is from Google & Direct Traffic combined
- 68% of customers contribution in SMS Sent & Email Opened activities.

Univariate Analysis – Categorical Variables

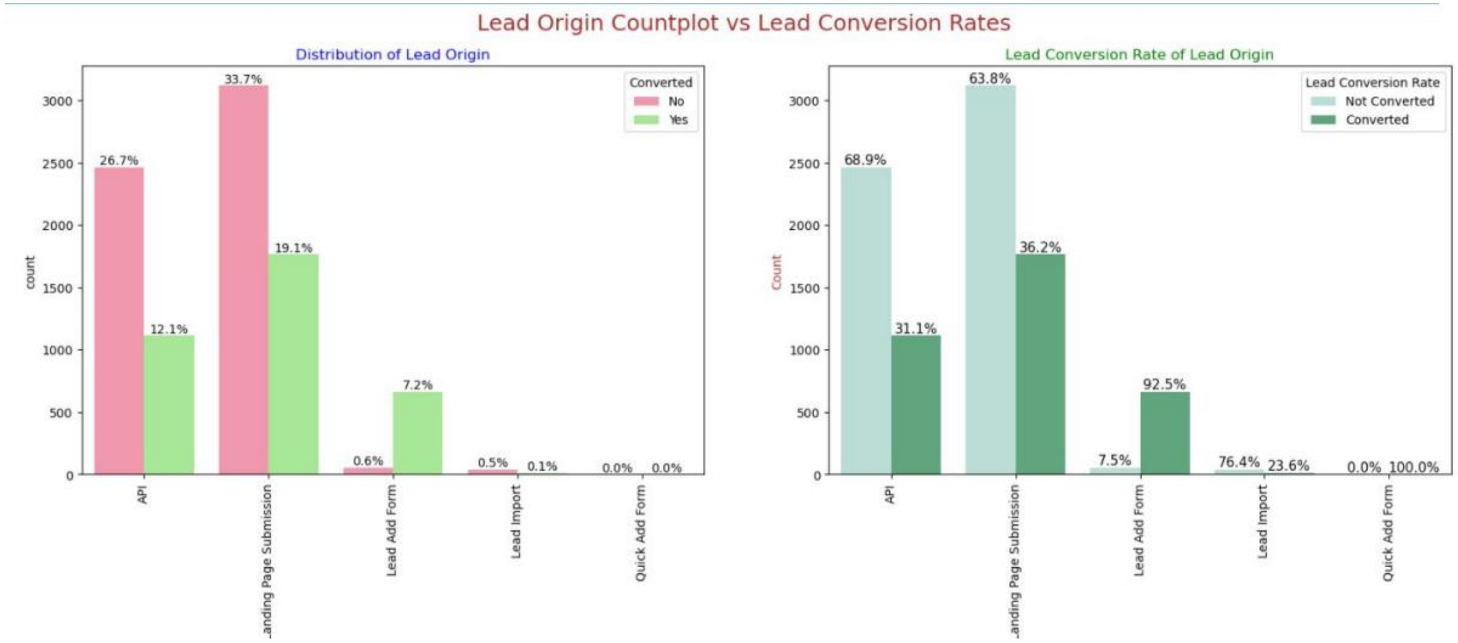




- "Landing Page Submission" identified 53% of customers, "API" identified 39%.
- Current_occupation has 90% of the customers as Unemployed.

EDA – Bivariate Analysis for Categorical Variables

Around 52% of all leads originated from "Landing Page Submission" with a lead conversion rate (LCR) of 36%. ● The "API" identified approximately 39% of customers with a lead conversion rate (LCR) of 31%.



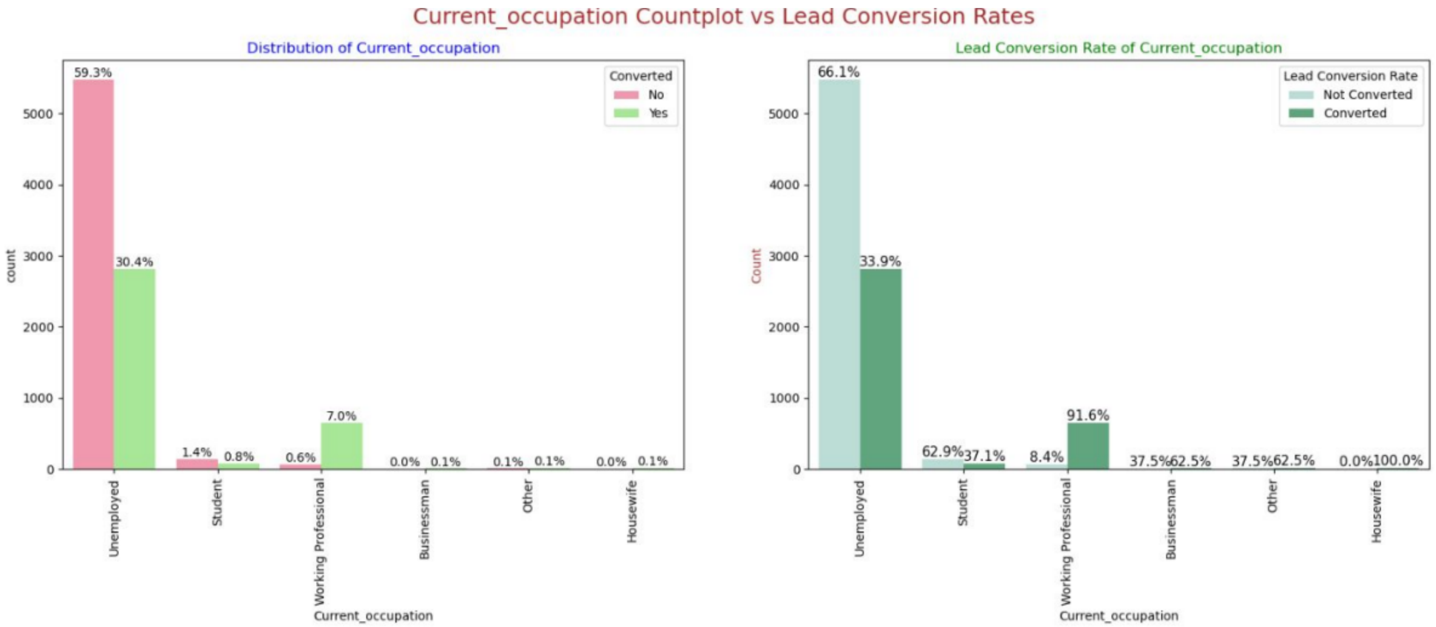
Lead Origin Countplot vs Lead Conversion Rates

Distribution of Lead Origin

Lead Conversion Rate of Lead Origin

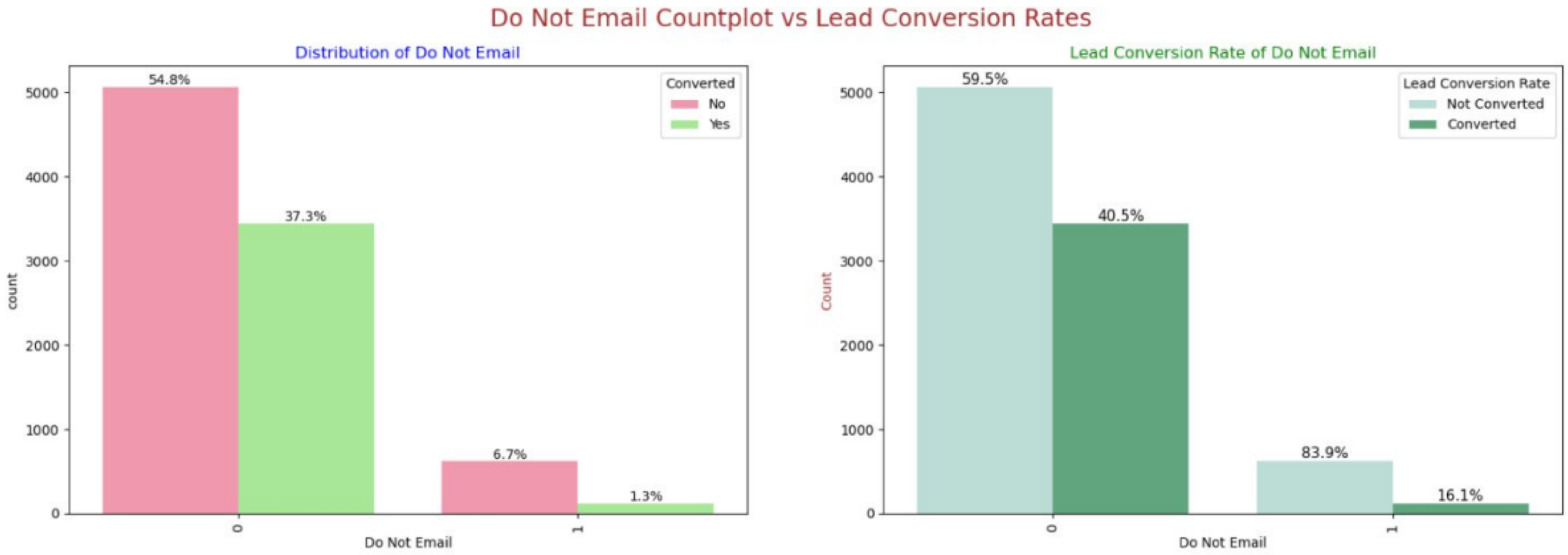
Bivariate Analysis for Categorical Variables

- Around 90% of the customers are Unemployed, with lead conversion rate (LCR) of 34%.
- While Working Professional contribute only 7.6% of total customers with almost 92% Lead conversion rate (LCR).



EDA – Bivariate Analysis for Categorical Variables

- 92% of the people has opted that they don't want to be emailed about the course & 40% of them are converted to leads.



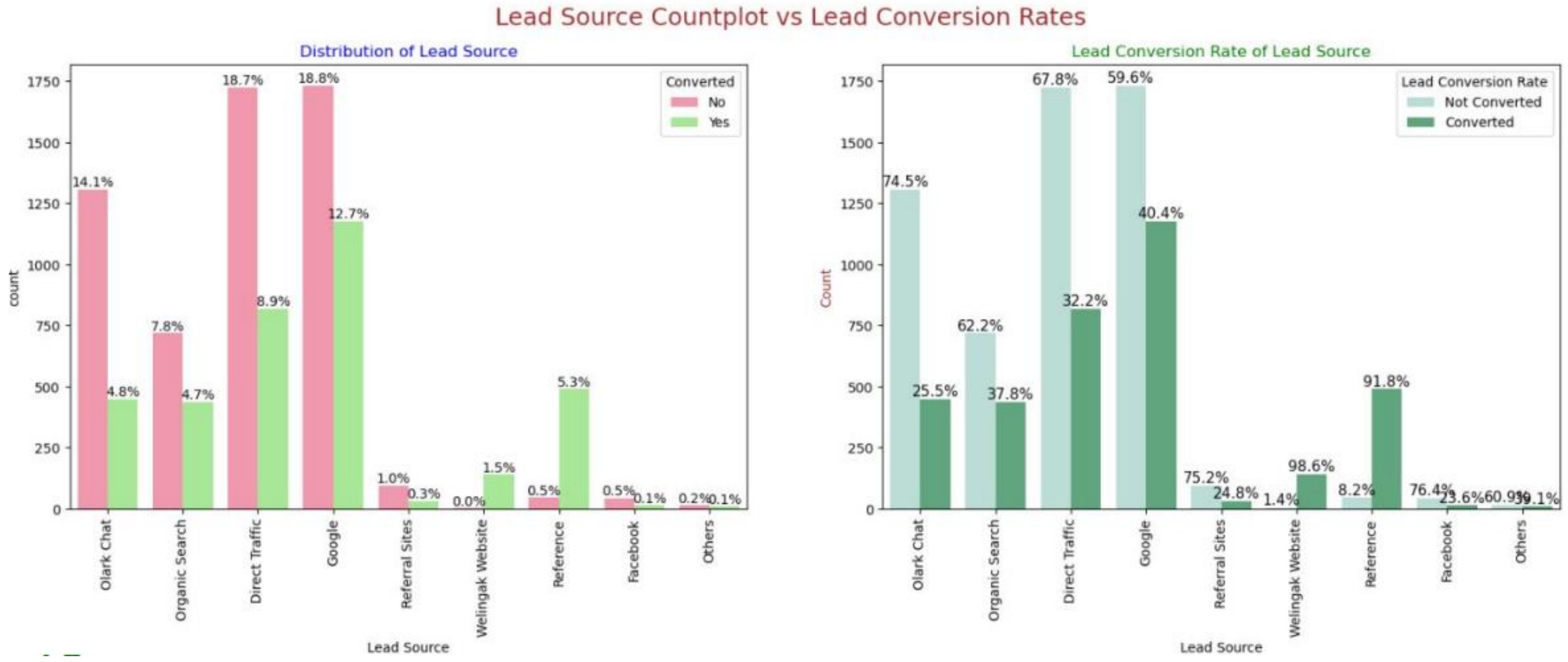
Do Not Email Countplot vs Lead Conversion Rates

Distribution of Do Not Email

Lead Conversion Rate of Do Not Email

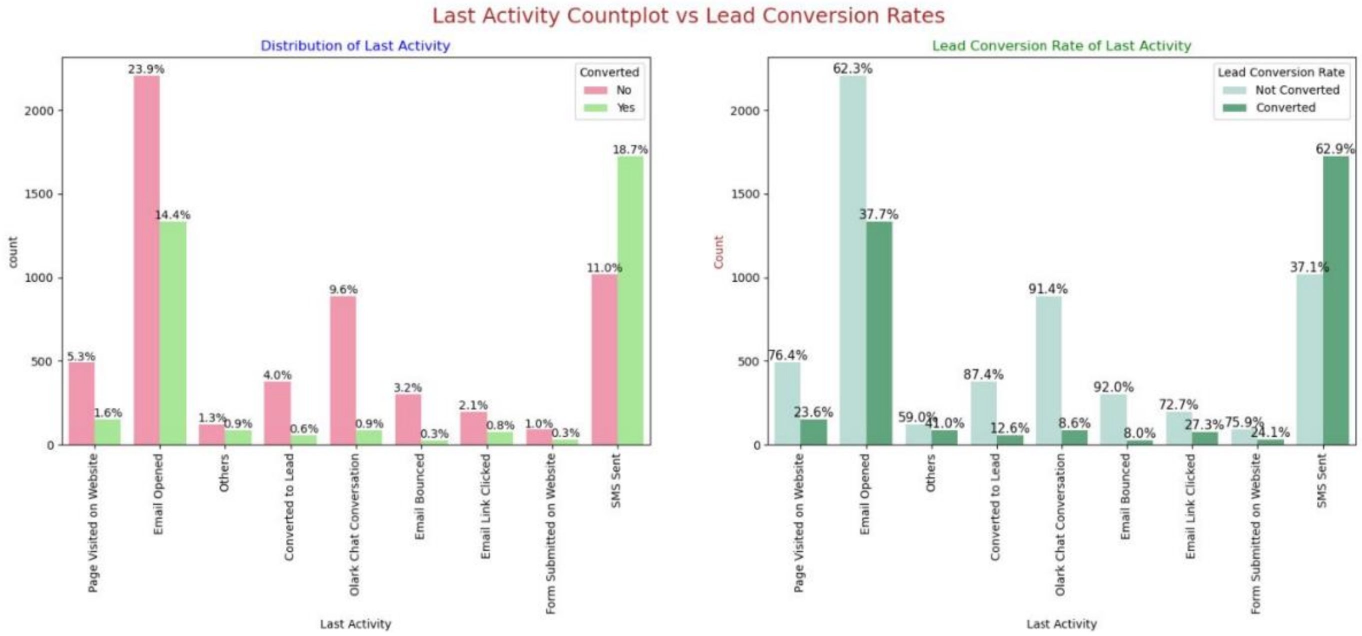
Bivariate Analysis for Categorical Variables

- Google has LCR of 40% out of 31% customers
- Direct Traffic contributes 32% LCR with 27% customers, which is lower than Google
- Organic Search also gives 37.8% of LCR, but the contribution is by only 12.5% of customers
- Reference has LCR of 91%, but there are only around 6% of customers through this Lead Source.



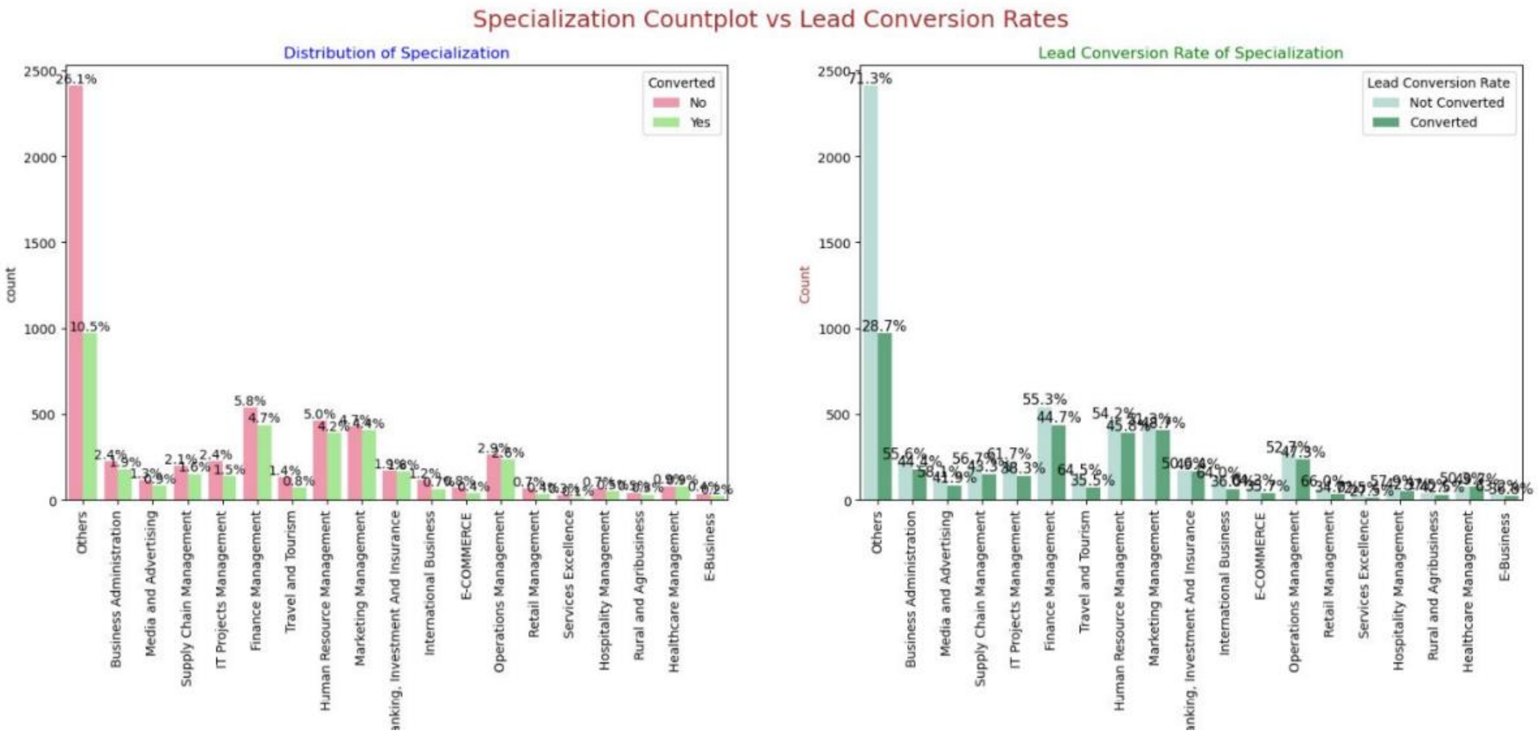
Bivariate Analysis for Categorical Variables

- ‘SMS Sent' has high lead conversion rate of 63% with 30% contribution from last activities
- 'Email Opened' activity contributed 38% of last activities performed by the customers, with 37% lead conversion rate.



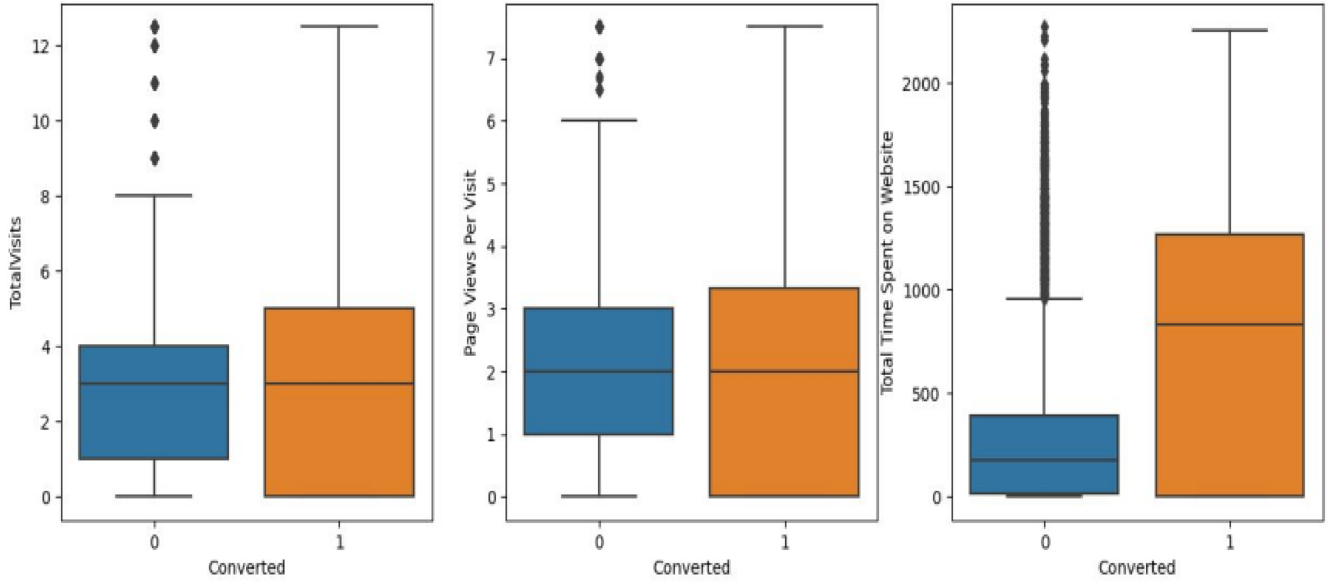
Bivariate Analysis for Categorical Variables

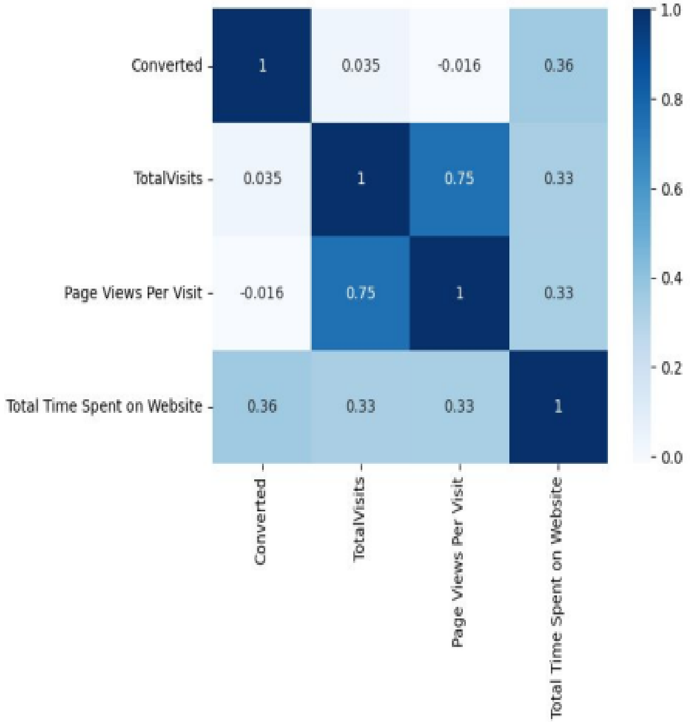
- Marketing Management, HR Management, Finance Management shows good contribution in Leads conversion than other specialization.



Bivariate Analysis for Numerical Variables

- Past Leads who spends more time on the Website have a higher chance of getting successfully converted than those who spends less time as seen in the box-plot





Data Preparation before Model building

- Binary level categorical columns were already mapped to 1 / 0 in previous steps
- Created dummy features (one-hot encoded) for categorical variables – Lead Origin, Lead Source, Last Activity, Specialization, Current_occupation
- Splitting Train & Test Sets ○ 70:30 % ratio was chosen for the split
- Feature scaling ○ Standardization method was used to scale the features
- Checking the correlations ○ Predictor variables which were highly correlated with each other were dropped (Lead Origin_Lead Import and Lead Origin_Lead Add Form).

Model Building

Feature Selection

- The data set has lots of dimension and large number of features.
- This will reduce model performance and might take high computation time.
- Hence it is important to perform Recursive Feature Elimination (RFE) and to select only the important columns.
- Then we can manually fine tune the model.
- RFE outcome ○ Pre RFE – 48 columns & Post RFE – 15 columns

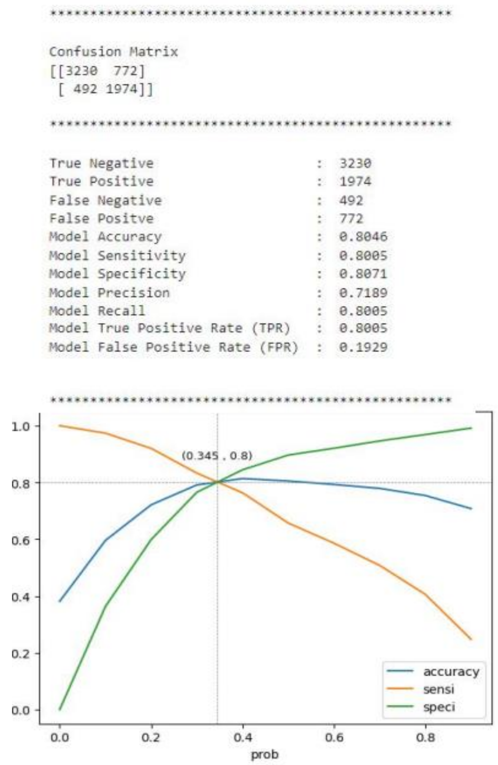
Model Building

- Manual Feature Reduction process was used to build models by dropping variables with p – value greater than 0.05.
- Model 4 looks stable after four iteration with:
- significant p-values within the threshold (p-values < 0.05) and
- No sign of multicollinearity with VIFs less than 5
- Hence, logm4 will be our final model, and we will use it for Model Evaluation which further will be used to make predictions.

TRAIN DATA SET

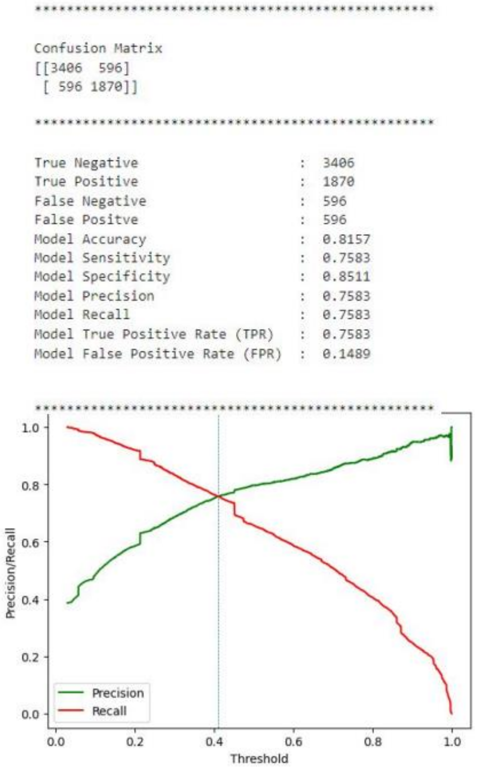
- It was decided to go ahead with 0.345 as cutoff after checking evaluation metrics coming from both plots

Confusion Matrix & Evaluation Metrics
with 0.345 as cutoff



```
**************************************************

Confusion Matrix
[[3230  772]
 [ 492 1974]]

**************************************************

True Negative                     :  3230
True Positive                     :  1974
False Negative                    :  492
False Positve                     :  772
Model Accuracy                    :  0.8046
Model Sensitivity                 :  0.8005
Model Specificity                 :  0.8071
Model Precision                   :  0.7189
Model Recall                      :  0.8005
Model True Positive Rate (TPR)    :  0.8005
Model False Positive Rate (FPR)   :  0.1929

**************************************************
```

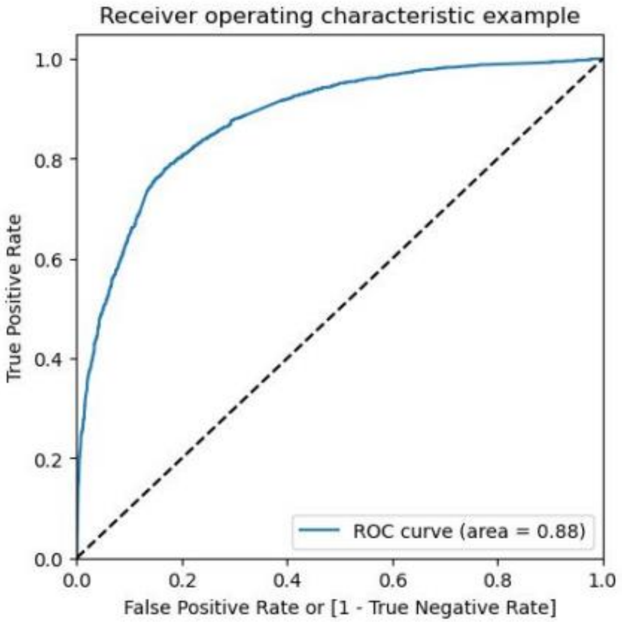
Confusion Matrix & Evaluation Metrics
with 0.41 as cutoff



```
**************************************************

Confusion Matrix
[[3406  596]
 [ 596 1870]]

**************************************************

True Negative                     :  3406
True Positive                     :  1870
False Negative                    :  596
False Positve                     :  596
Model Accuracy                    :  0.8157
Model Sensitivity                 :  0.7583
Model Specificity                 :  0.8511
Model Precision                   :  0.7583
Model Recall                      :  0.7583
Model True Positive Rate (TPR)    :  0.7583
Model False Positive Rate (FPR)   :  0.1489

**************************************************
```

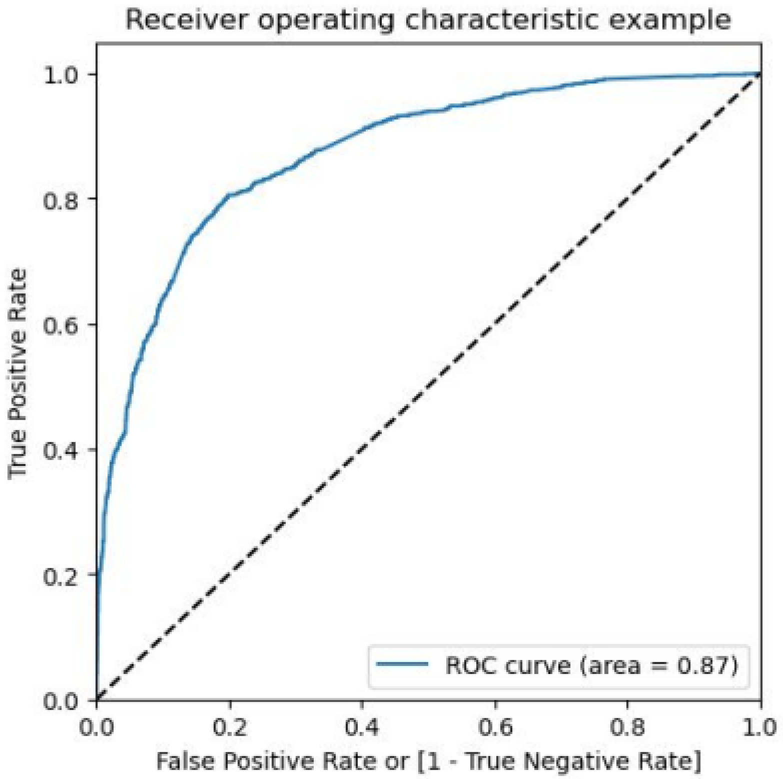
ROC Curve – Train Data Set

- Area under ROC curve is 0.88 out of 1 which indicates a good predictive model.
- The curve is as close to the top left corner of the plot, which represents a model that has a high true positive rate and a low false positive rate at all threshold values.



ROC Curve – Test Data Set

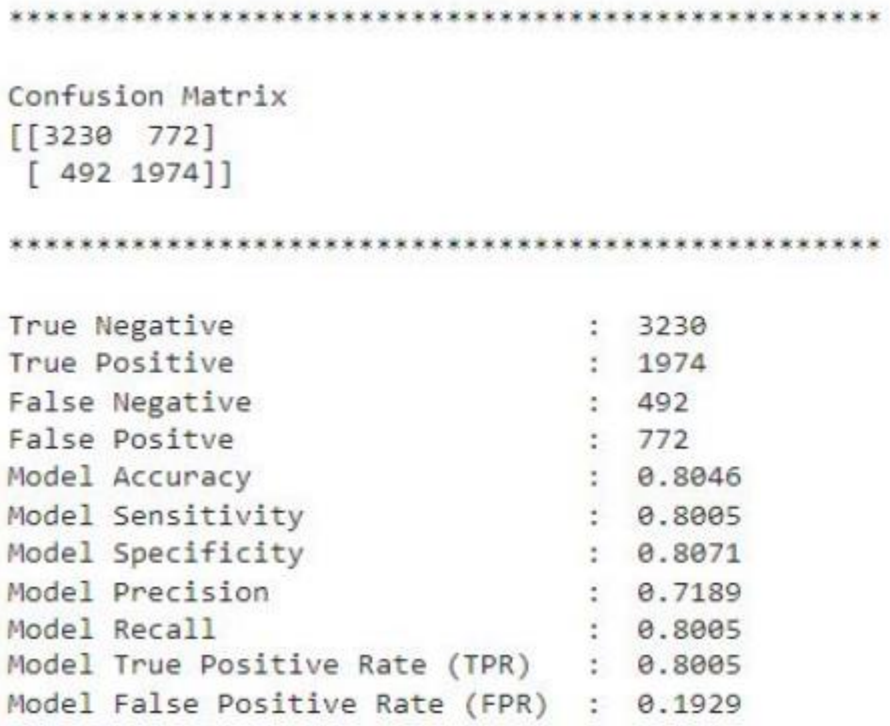
- Area under ROC curve is 0.87 out of 1 which indicates a good predictive model.
- The curve is as close to the top left corner of the plot, which represents a model that has a high true positive rate and a low false positive rate at all threshold values.



Model Evaluation

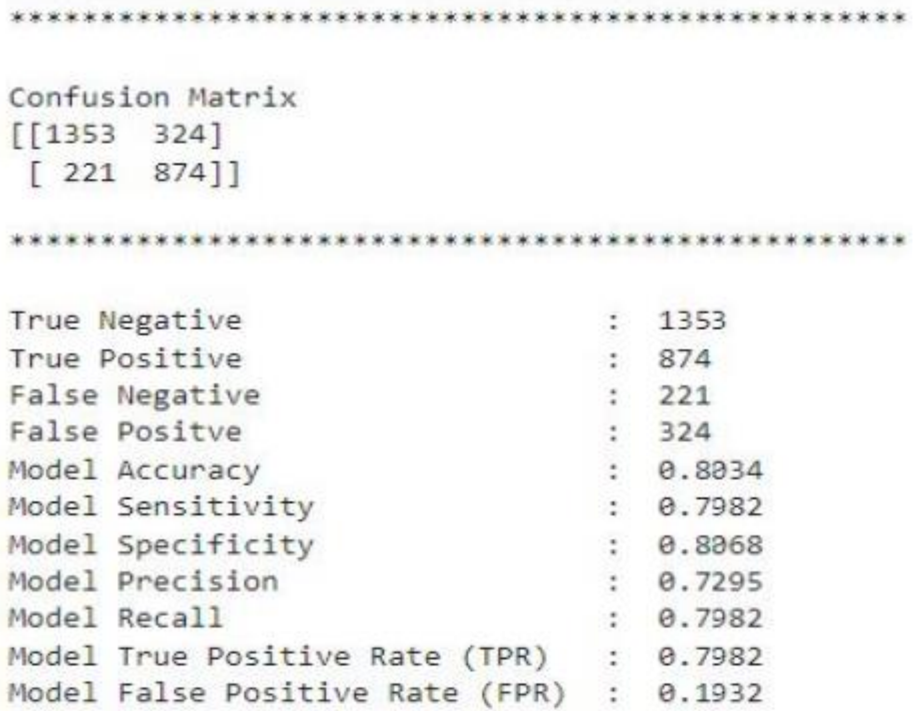
- Using a cut-off value of 0.345, the model achieved a sensitivity of 80.05% in the train set and 79.82% in the test set.
- Sensitivity in this case indicates how many leads the model identify correctly out of all potential leads which are converting
- The CEO of X Education had set a target sensitivity of around 80%.
- The model also achieved an accuracy of 80.46%, which is in line with the study's objectives.

Train Data Set



```
**************************************************

Confusion Matrix
[[3230  772]
 [ 492 1974]]

**************************************************

True Negative                     :  3230
True Positive                     :  1974
False Negative                    :  492
False Positve                     :  772
Model Accuracy                    :  0.8046
Model Sensitivity                 :  0.8005
Model Specificity                 :  0.8071
Model Precision                   :  0.7189
Model Recall                      :  0.8005
Model True Positive Rate (TPR)    :  0.8005
Model False Positive Rate (FPR)   :  0.1929
```

Test Data Set



```
**************************************************

Confusion Matrix
[[1353  324]
 [ 221  874]]

**************************************************

True Negative                     :  1353
True Positive                     :  874
False Negative                    :  221
False Positve                     :  324
Model Accuracy                    :  0.8034
Model Sensitivity                 :  0.7982
Model Specificity                 :  0.8068
Model Precision                   :  0.7295
Model Recall                      :  0.7982
Model True Positive Rate (TPR)    :  0.7982
Model False Positive Rate (FPR)   :  0.1932
```

Recommendation based on Final Model

- According to the problem description, X Education's expansion and success depend on a higher lead conversion rate. In order to do this, we have created a regression model that will enable us to pinpoint the key elements influencing lead conversion.
- In order to improve lead conversion, we have identified the attributes that have the highest positive coefficients. Our marketing and sales strategies should prioritize these features.
  Lead Source Reference: 2.93; Lead Source Welingak Website: 5.39
- The data includes the following: Current Occupation: Working Professional; Last Activity: 2.05; Last Activity: Others; Total Time Spent on Website: 1.05; Last Activity: Email Opened: 0.94.
- Olark Chat's Lead Source: 0.91
- Additionally, we've found characteristics with negative coefficients that can point to possible areas for development. Among them are:
- Hospitality Management Specialty: -1.09
- Specialization in Others: -1.20
- Lead Origin of Landing Page Submission: -1.26

Recommendation based on Final Model

- In order to raise our rates of lead conversion
- For focused marketing efforts, concentrate on traits with positive coefficients.
- Create plans to draw in quality leads from sources that provide the most leads.
- Adjust communication channels according to the impact of lead engagement.
- Use customized messages to engage professionals in the workforce.
- More money may be spent on advertising and other things on the Welingak website.
- Offers of rewards or discounts for supplying references that result in leads; promote supplying more references.
- Targeting working professionals aggressively is recommended due to their high conversion rate and improved financial standing, which allows them to pay larger costs.
- To determine what needs to be improved Examine negative coefficients in offerings related to specialization.
- Look for instances where the landing page submission process could be improved.