upGrad

Lead Scoring Case Study

DETECTION OF HOT LEADS TO CONCENTRATE MORE OF MARKETING EFFORTS ON THEM, IMPROVING CONVERSION RATES FOR X EDUCATION

Team Members: Raksha More, K Shilpa & Sagar

Table Content

Background of X Education Company

- An education company named X Education sells online courses to industry professionals.
- Many professionals who are interested in the courses land on their website and browse for courses.
- The company markets its courses on several websites and search engines like Google.
- Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos.
- 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
- Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.
- Although X Education gets a lot of leads, its lead conversion rate is very poor.
- X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- The company requires you to build a model wherein you 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%.

Problem Statement & Objective of the Study

- The X Education wants to make lead conversion process more efficient by identifying the most potential leads.
- The X Education gets a lot of leads, its lead conversion rate is very poor at around 30%.
- Rather than making calls to everyone their sales team want to know these potential set of leads, which they will be focusing more on communicating.
- To help X Education select the most promising leads, i.e., the leads that are most likely to convert into paying customers.
- 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 has given a ballpark of the target lead conversion rate to be around 80%.

Suggested Ideas for Lead Conversion

Leads Grouping

- Leads are grouped based on their propensity or likelihood to convert.
- This results in a focused group of hot leads.

Better Communication

- We could have a smaller pool of leads to communicate with, which would allow us to have a greater impact.

Boost Conversion

- We would have a greater conversion rate and be able to hit the 80% objective since we concentrated on hot leads that were more likely to convert.

We would want to obtain a high sensitivity in obtaining hot leads Since we have a target of 80% conversion rate.

Analysis Approach

- **Data Cleaning:-** Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled. If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct. There is no one absolute way to prescribe the exact steps in the data cleaning process because the processes will vary from dataset to dataset. But it is crucial to establish a template for your data cleaning process so you know you are doing it the right way every time.
- **EDA:-** Exploratory Data Analysis (EDA) is an approach that is used to analyze the data and discover trends, patterns, or check assumptions in data with the help of statistical summaries and graphical representations
- **Data preparation:-** Data preparation is the process of preparing raw data so that it is suitable for further processing and analysis. Key steps include collecting, cleaning, and labeling raw data into a form suitable for machine learning (ML) algorithms and then exploring and visualizing the data. Data preparation can take up to 80% of the time spent on an ML project. Using specialized data preparation tools is important to optimize this process.
- **Model Building :-**In this phase data science team needs to develop data sets for training, testing, and production purposes. These data sets enable data scientist to de;-velop analytical method and train it, while holding aside some of data for testing the model.

- **Model Evaluation :-** Model evaluation is the process of using different evaluation metrics to understand a machine learning model's performance, as well as its strengths and weaknesses. Model evaluation is important to assess the efficacy of a model during initial research phases, and it also plays a role in model monitoring**.**
- **Prediction on Test Data :-** Predictive Test Selection is a new approach to Test Impact Analysis that uses machine learning to dynamically select which tests to run based on the characteristics of a code change. To achieve this, historic test results and information about the changes that were tested are used to train a machine learning model.
- **Recommdation:-** A recommendation system is a subclass of Information filtering Systems that seeks to predict the rating or the preference a user might give to an item. In simple words, it is an algorithm that suggests relevant items to users. Eg: In the case of Netflix which movie to watch, In the case of e-commerce which product to buy, or In the case of kindle which book to read, etc.

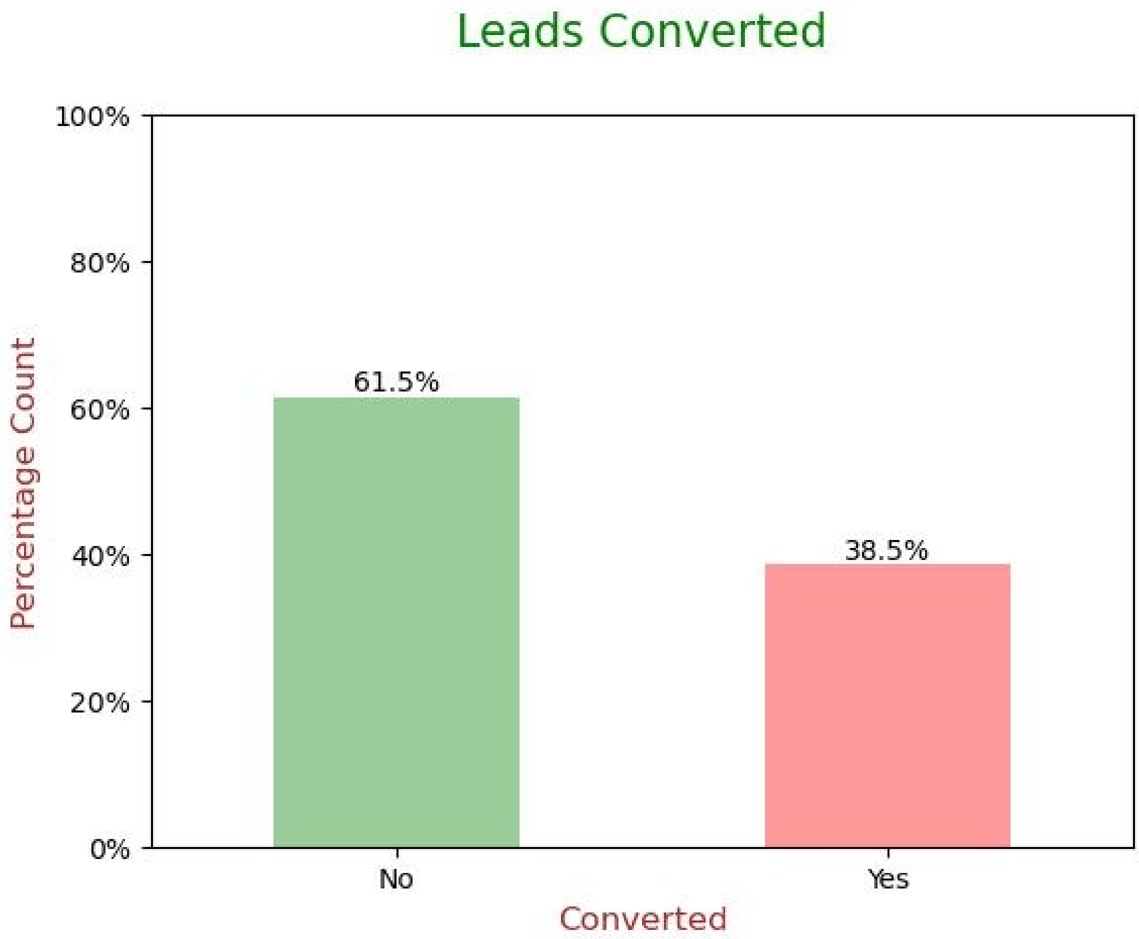
Data Cleaning Approach

- As customers did not choose any option from the list, "Select" level represents null values for some categorical variables.
- The unused columns with over 40% null values were dropped.
- Based on value counts and certain considerations, missing values in categorical columns were handled.
- Drop columns that don't add any insight or value to the study objective (tags, country)

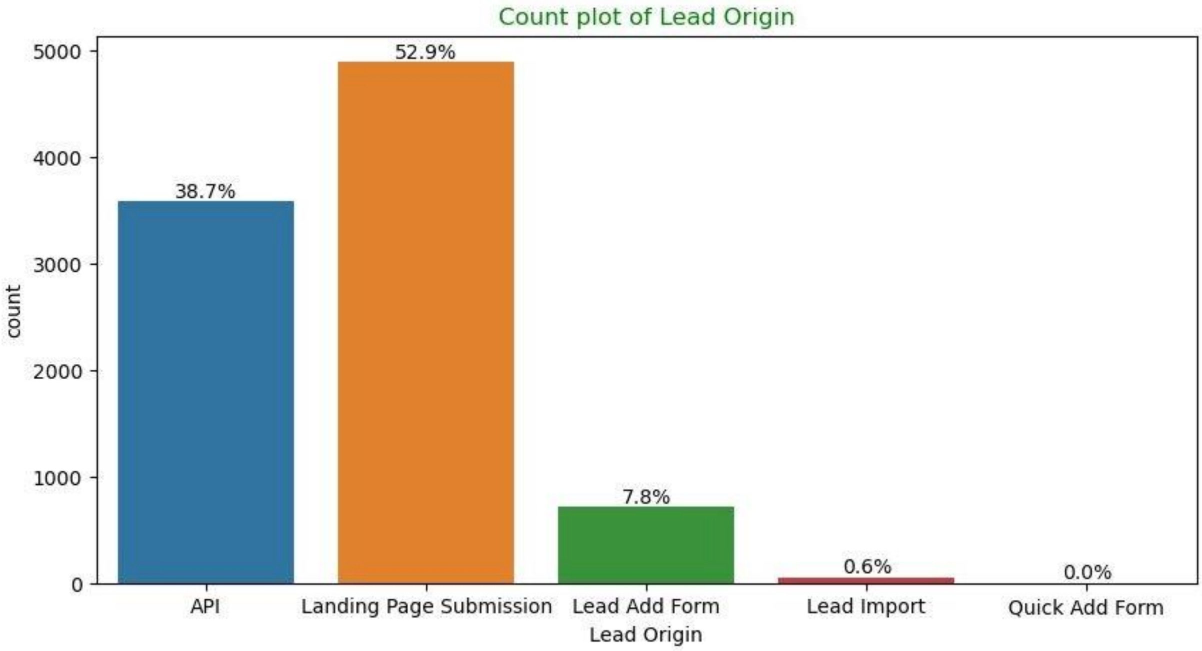
- Imputation was used for some categorical variables.
- For some variables, additional categories were created.
- Columns with no use for modeling (Prospect ID, Lead Number) or only one category of response were dropped.
- Numerical data was imputed with mode after checking distribution.
- Skewed category columns were checked and dropped to avoid bias in logistic regression models.
- Outliers in Total Visits and Page Views Per Visit were treated and capped.
- Invalid values were fixed and data was standardized in some columns, such as lead source.
- Low frequency values were grouped together to "Others".
- Binary categorical variables were mapped to zero and one's.
- Other cleaning activities were performed to ensure data quality and accuracy.
- Fixed Invalid values & Standardizing Data in columns by checking casing styles, etc.

EDA Exploratory Data Analysis

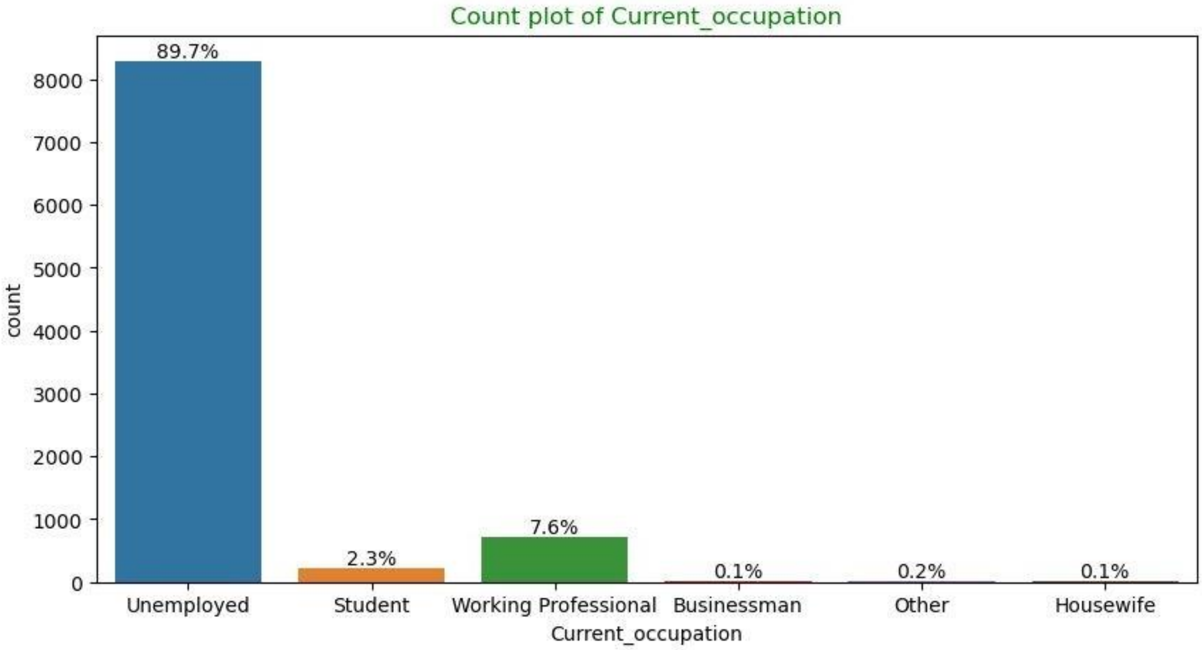
- ➢ Data is imbalanced while analyzing target variable
- Only 38.5% of the people have converted to leads which is conversion rate is of 38.5% and it's a minority.
- While 61.5% of the people didn't convert to leads and it's a majority.



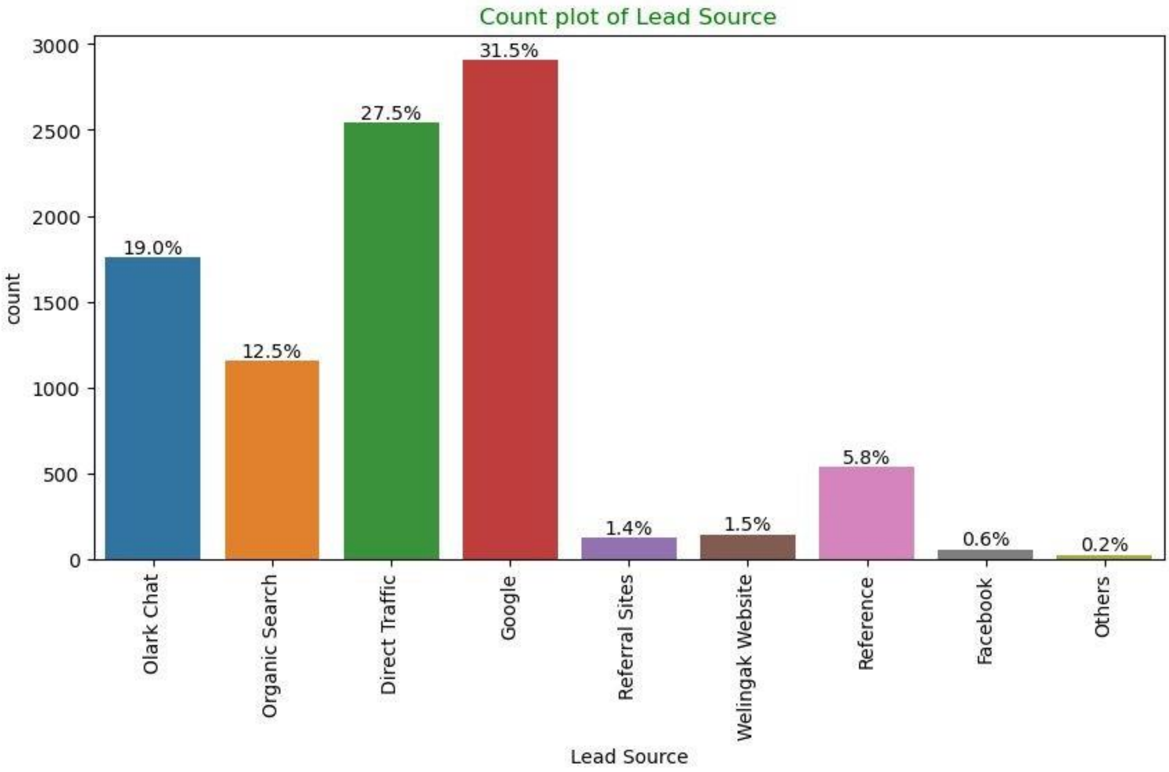
Univariate Analysis – Categorical Variables



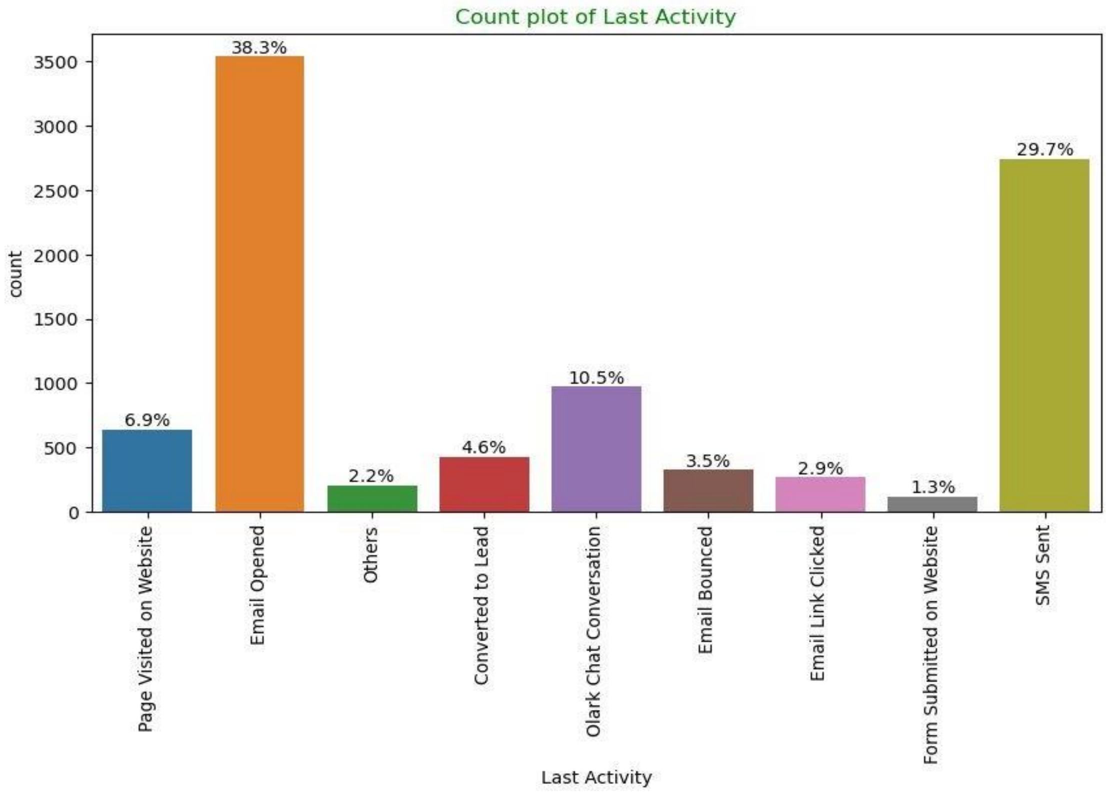
As per above analysis we observe that Lead Origin: "Landing Page Submission" identified 53% of customers, "API" identified 39%.



As per above analysis we observe that Current_occupation has 90% of the customers as Unemployed.

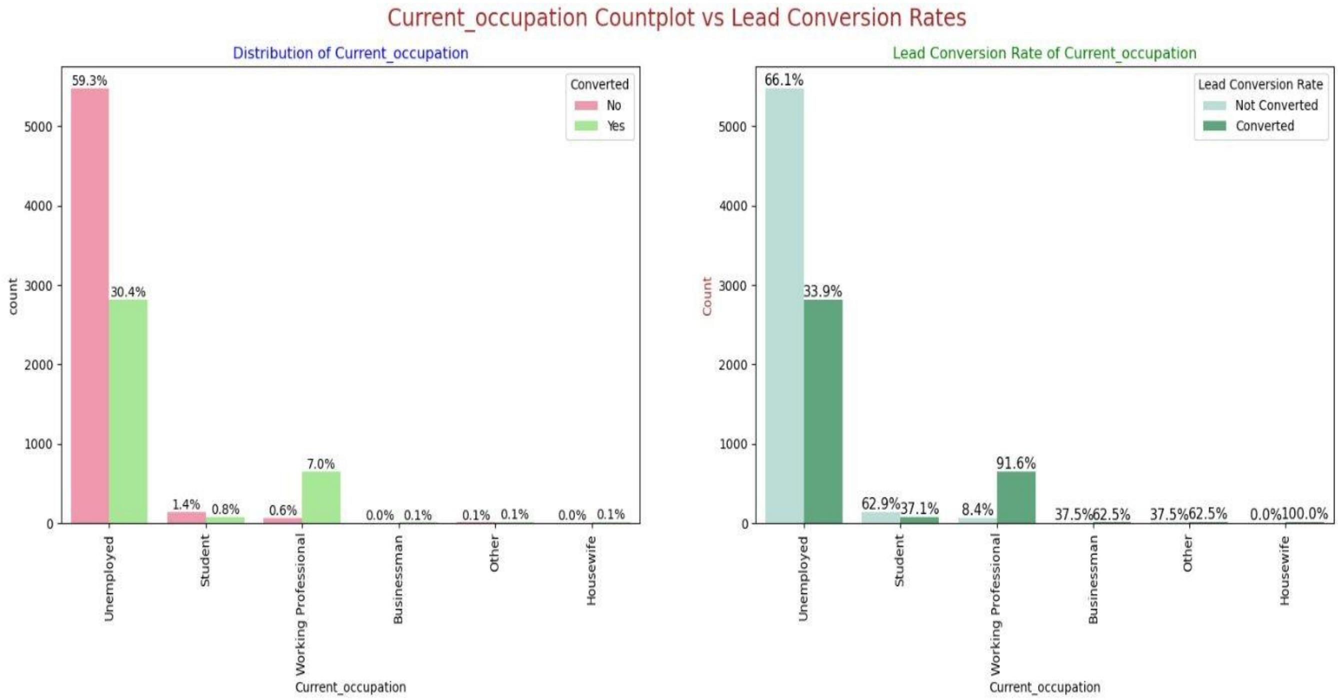


As per above analysis we observe that 58% Lead source is from Google & Direct Traffic combined.



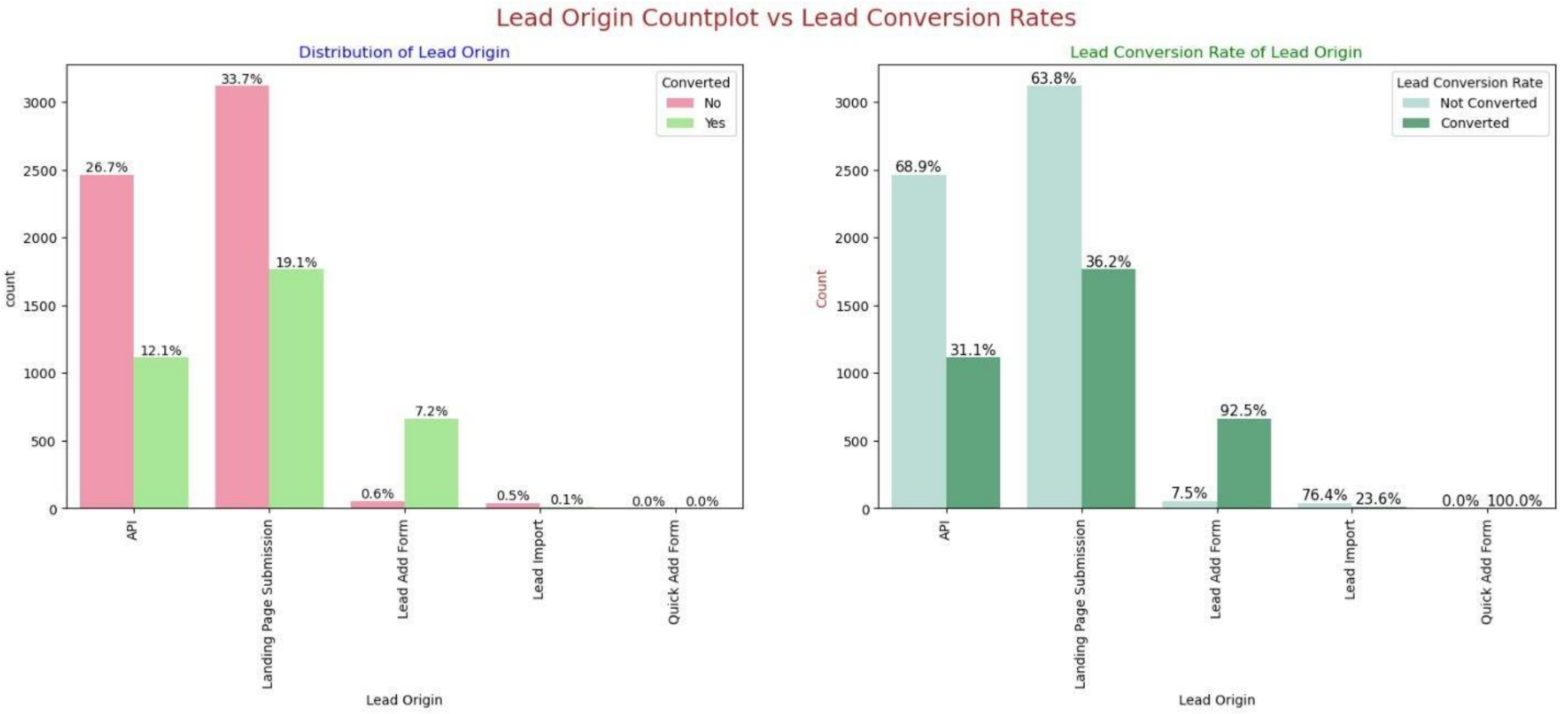
As per above analysis we observe that68% of customers contribution in SMS Sent & Email Opened activities.

- ➢ Bivariate Analysis for Categorical variable



As per above analysis we found below observation

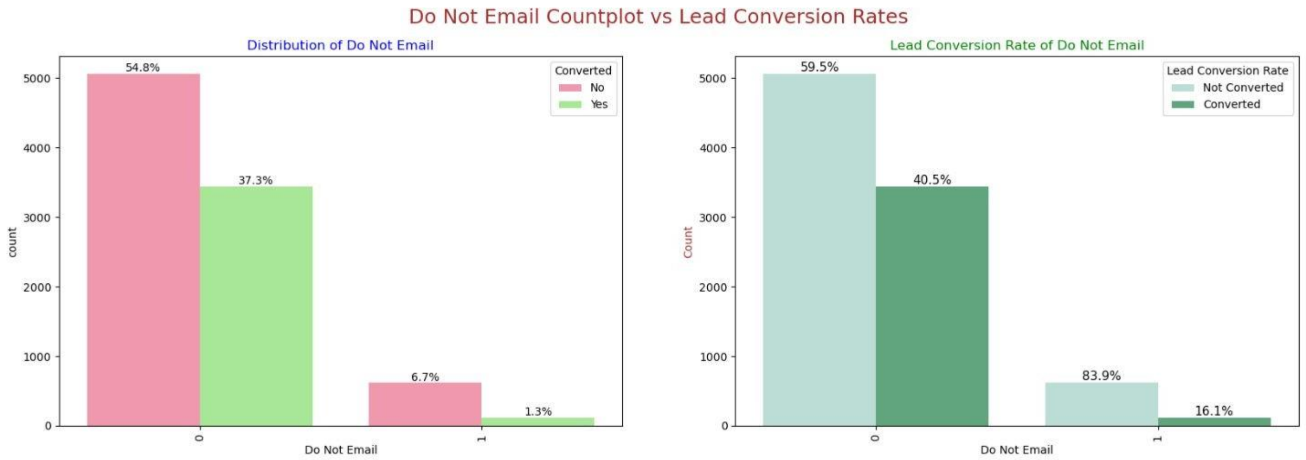
- Around 90% of the customers are Unemployed, with lead conversion rate (LCR) of 34%.
- While Working Professional contribute only 7.6% of total customers with almost 92% Lead conversion rate (LCR).

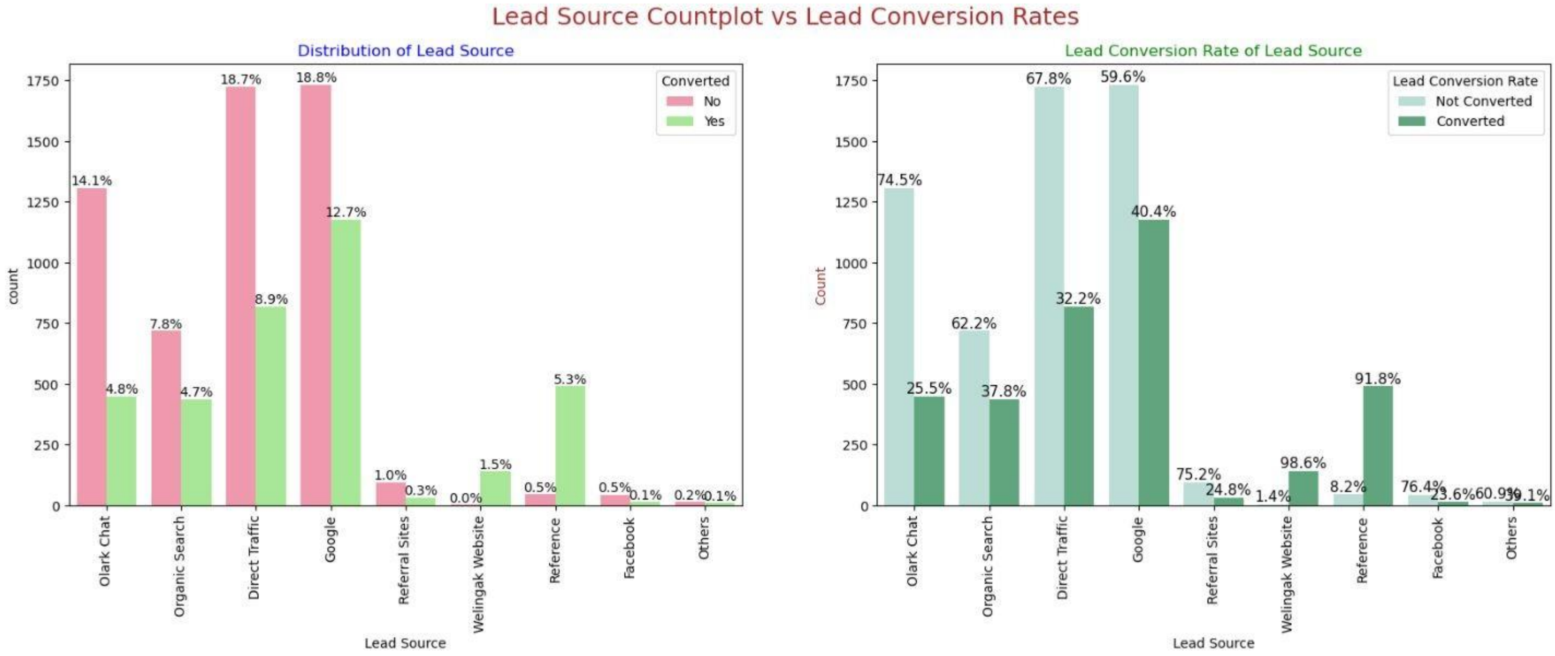


As per above analysis we found below observation

- 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%.

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

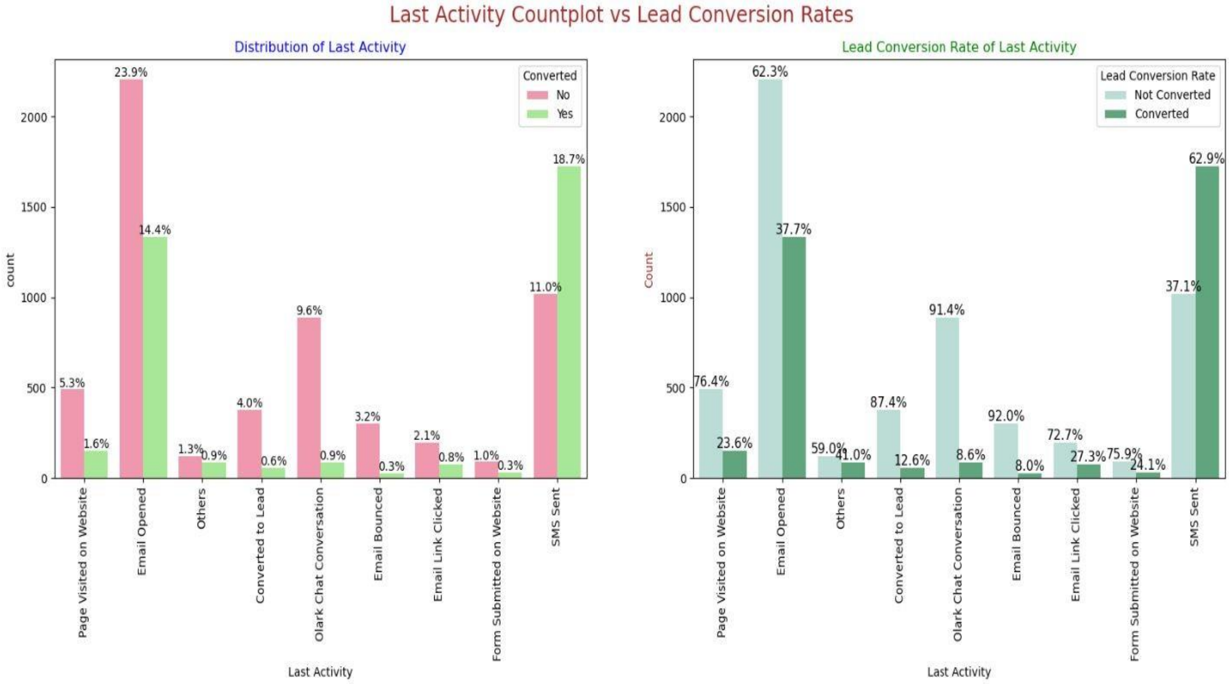




As per above analysis we observe below points:

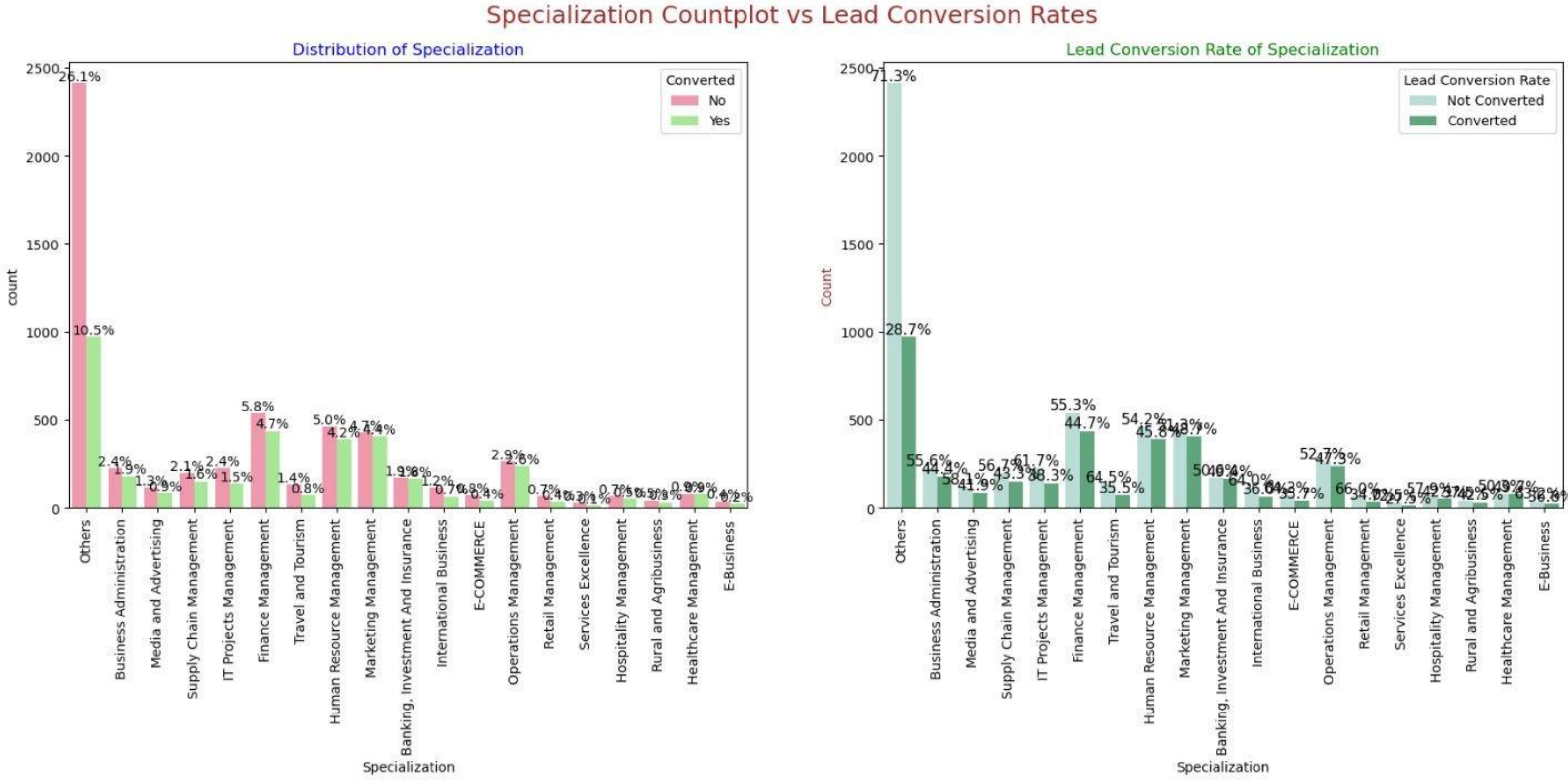
- Google has LCR of 40% out of 31% customers,
- Direct Traffic contributes 32% LCR with 27% customers, which is lower than Google,
- Organic Search also gives 37.8% of LCR, but the contribution is by only 12.5% of customers,

- Reference has LCR of 91%, but there are only around 6% of customers through this Lead Source.

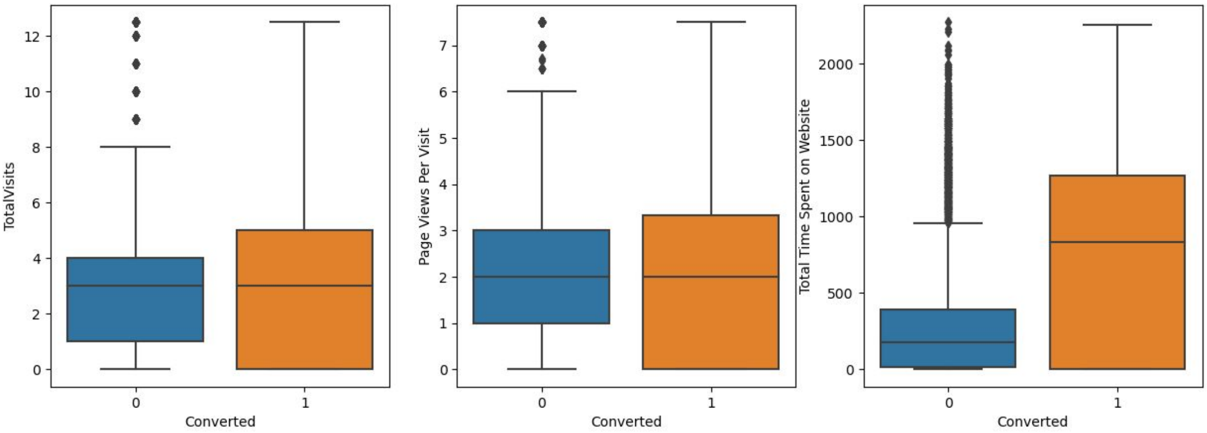


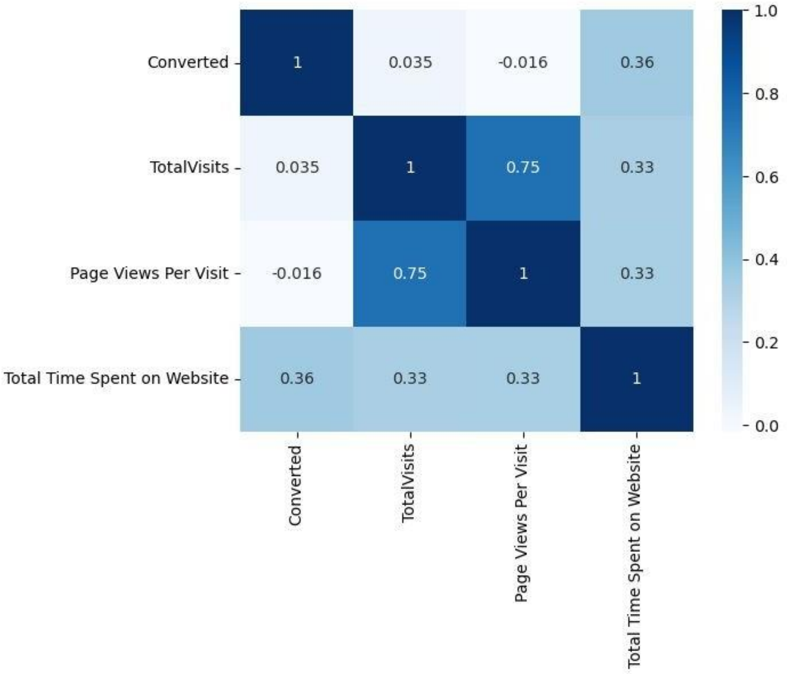
As per above analysis we observe below points:

- '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.



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





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

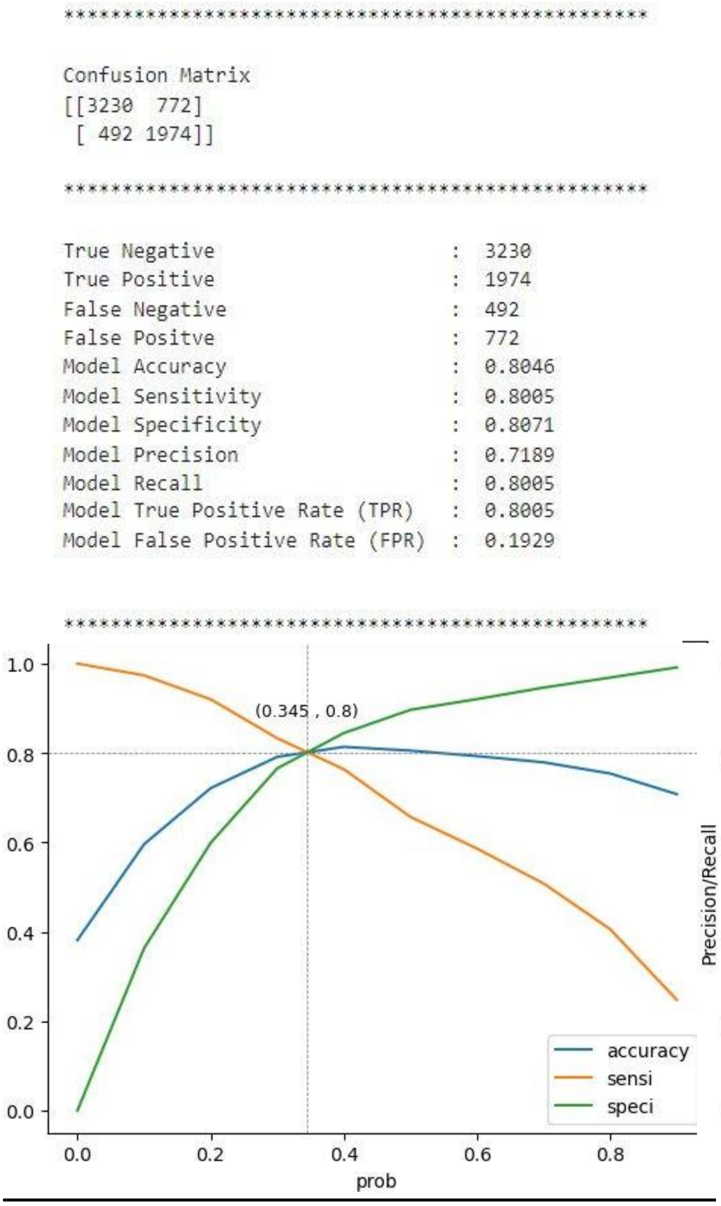
- Created dummy features (one-hot encoded) for categorical variables – Lead Origin, Lead Source, Last Activity, Specialization, Current_occupation.
- Binary level categorical columns were already mapped to 1 or 0 in previous steps.
- Splitting Train & Test Sets-70:30 % ratio was chosen for the split.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

Feature Selection

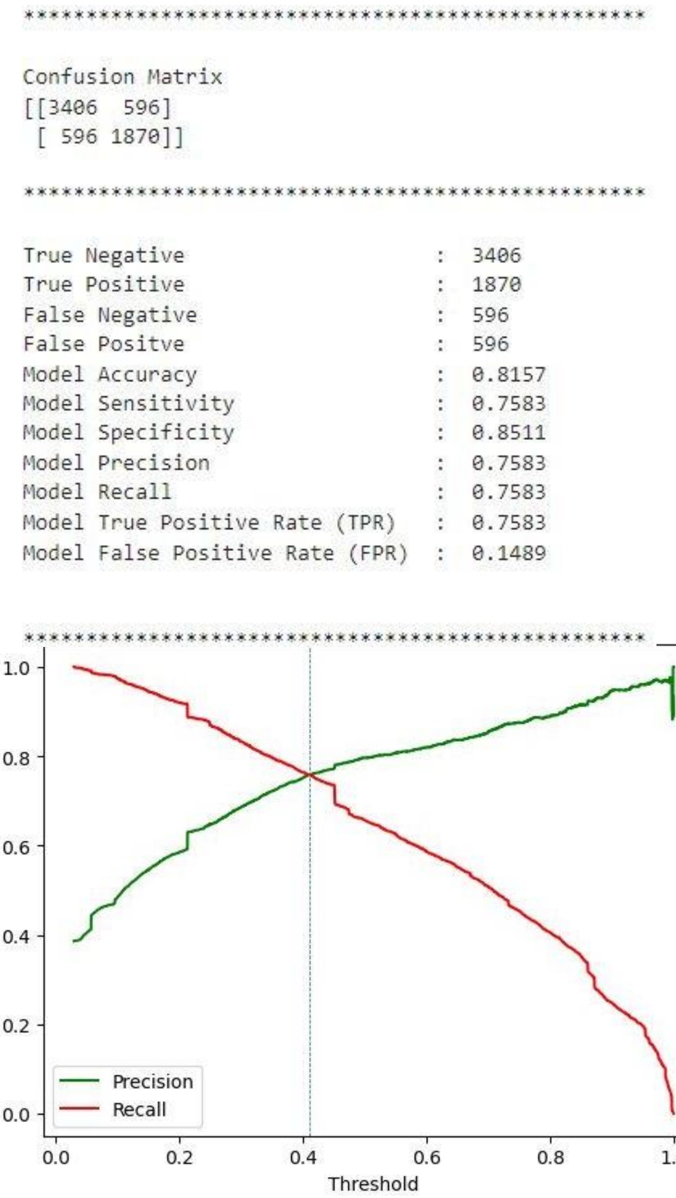
- The data set has lots of dimension and large number of features.
- This will reduce model performance and might take high computation time.
- Hence it is important to perform Recursive Feature Elimination (RFE) and to select only the important columns.
- Then we can manually fine tune the model.
- RFE outcome-Pre RFE – 48 columns & Post RFE – 15 columns
- 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:
    a. significant p-values within the threshold (p-values < 0.05) and
    b. 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.

Model Evalution

- **Train Dataset**



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

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

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



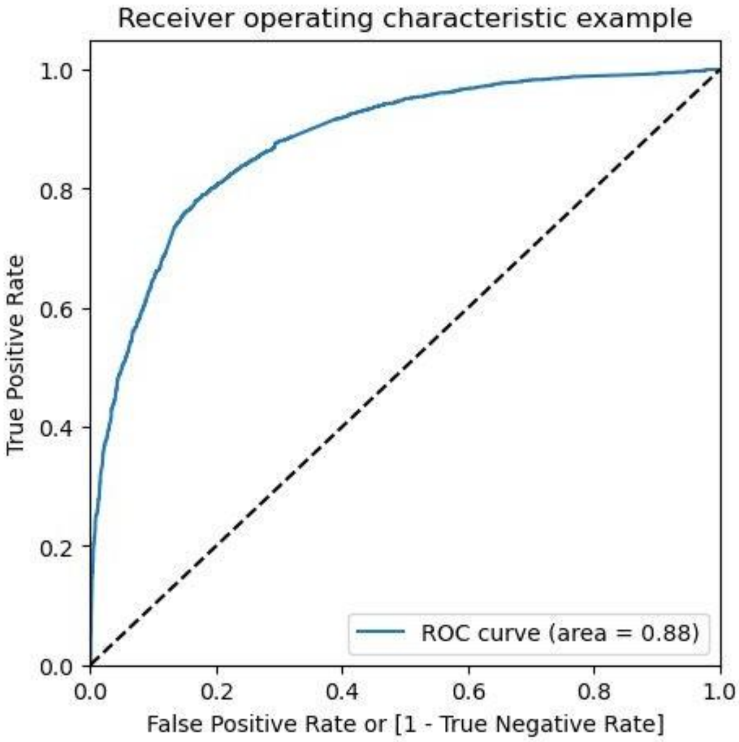
```
**************************************************

Confusion Matrix
[[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

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

Confusion Matrix & Evaluation Metrics with 0.345 as cutoff

Confusion Matrix & Evaluation Metrics with 0.41 as cutoff

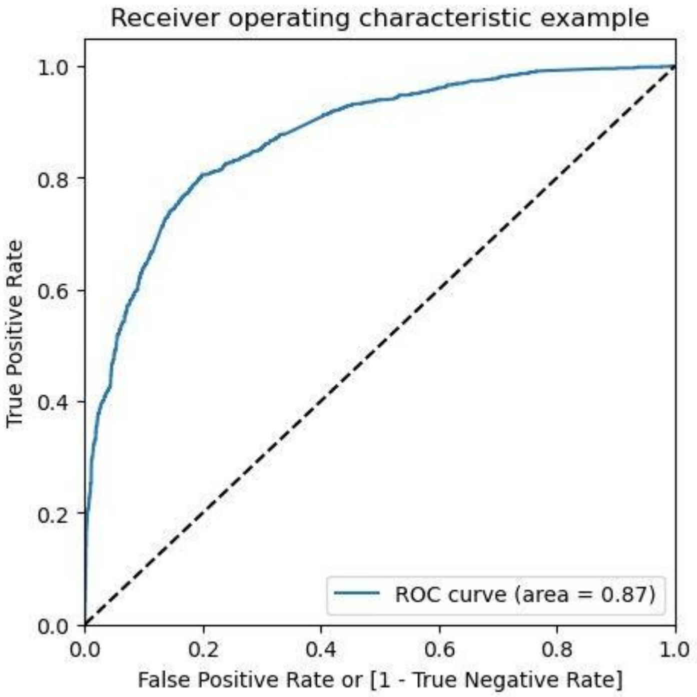
It was decided to go ahead with 0.345 as cutoff after checking evaluation metrics coming from both plots

ROC Curve – Train Data Set



- It's indicates a good predictive model because area under ROC curve is 0.88 out of 1.
- 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.

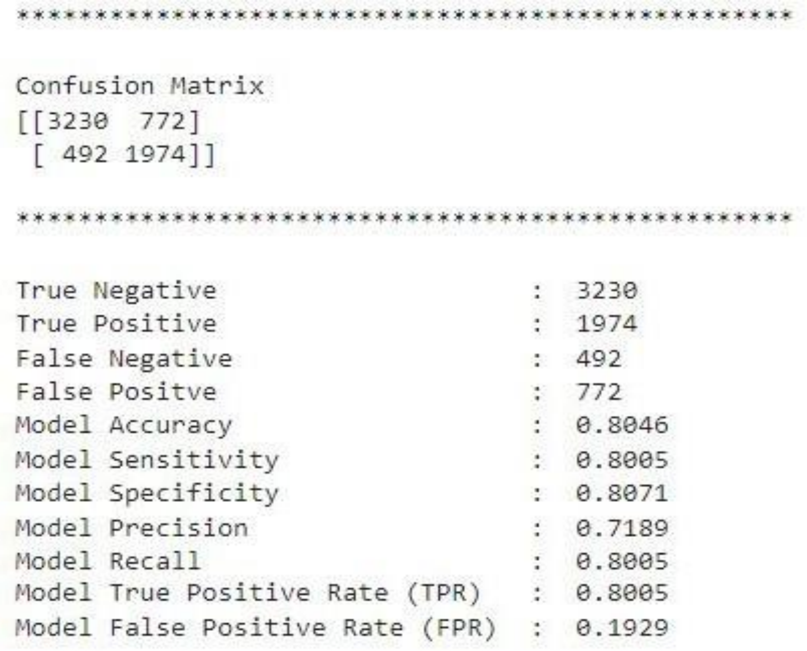
ROC sCurve – Test Data Set



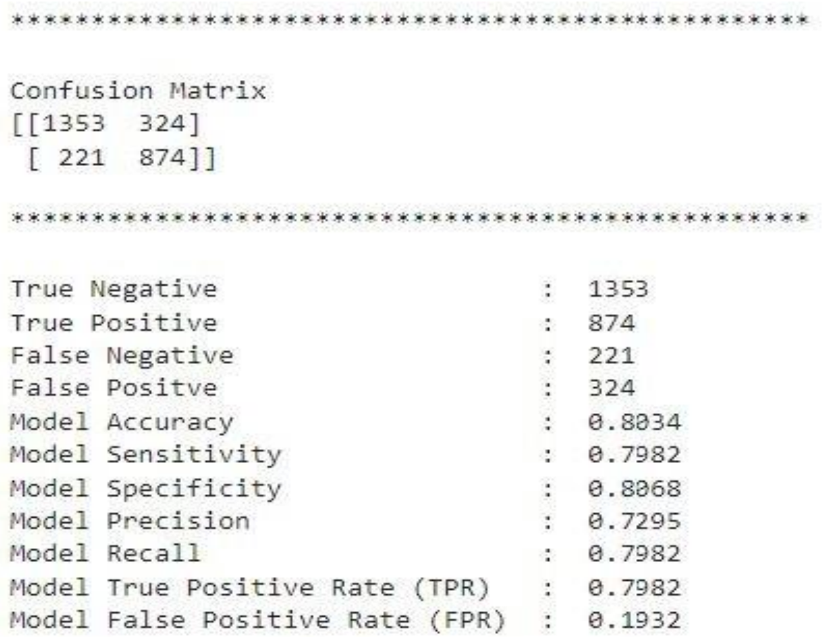
- Area under ROC curve is 0.87 out of 1 which indicates a good predictive model.
- The curve is as close to the top left corner of the plot, which represents a model that has a high true positive rate and a low false positive rate at all threshold values

Confusion Matrix & Metrics

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
```



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

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
```

- While 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.
- How many leads the model identify correctly out of all potential leads which are converting is a part of sensitivity
- Around 80% need to set a target sensitivity for X Education CEO.
- The model also achieved an accuracy of 80.46%, which is in line with the study's objectives.

Recommendations

As per the problem statement, increasing lead conversion is crucial for the growth and success of X Education. For that we have developed a regression model that can help us identify the most significant factors that impact lead conversion.

We have analysis the following features that have the highest positive coefficients, and these features should be given priority in our marketing and sales efforts to increase lead conversion.

- Lead Source_Welingak Website: 5.39
- Lead Source_Reference: 2.93
- Current_occupation_Working Professional: 2.67
- Last Activity_SMS Sent: 2.05
- Last Activity_Others: 1.25
- Total Time Spent on Website: 1.05
- Last Activity_Email Opened: 0.94
- Lead Source_Olark Chat: 0.91

We have also identified features with negative coefficients that may indicate potential areas for improvement.

These include:

- Specialization in Hospitality Management: -1.09
- Specialization in Others: -1.20
- Lead Origin of Landing Page Submission: -1.26

<u>To increase our Lead Conversion Rates</u>

- Focus on features with positive coefficients for targeted marketing strategies.
- Create strategies to attract high-quality leads from top-performing lead sources.
- Optimize communication channels based on lead engagement impact.
- Engage working professionals with tailored messaging.
- More budget/spend can be done on Welingak Website in terms of advertising, etc.
- Incentives/discounts for providing reference that convert to lead, encourage providing more references.
- Working professionals to be aggressively targeted as they have high conversion rate and will have better financial situation to pay higher fees too.