Lead Scoring Case Study

By:
Balachandran Mathiyazhagan
&
Hema Priya

Problem Statement:

- X Education has a bad lead turn rate of about 30% despite receiving a lot of leads.
- By locating the most promising leads, also referred to as Hot Leads, X Education hopes to increase the effectiveness of the lead conversion process.
- Their sales team wants to be aware of this possible group of prospects, so rather than calling everyone, they will concentrate more on speaking with them.

Aim of the Case Study:

- In order to aid X Education in choosing the prospects that have the best chance of becoming paying clients.
- The business wants us to create a model in which we give each lead a lead score, with higher lead scores indicating a higher chance of conversion and lower lead scores indicating a lower chance of conversion.
- The CEO provided an approximate goal prospect conversion rate of 80%.

Summary of Contents:

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

An overview of X Education Company

- Industry professionals can purchase online classes from X education, a business that provides instruction.
- Many workers who are interested in the classes visit their website on any given day and search for courses.
- On numerous websites and search engines like Google, the business advertises its classes.
- Upon arriving at the website, these visitors may peruse the classes, submit a form for the course, or watch some videos.
- These individuals are categorised as leads when they fill out a questionnaire with their phone number or email address.
- Once these prospects are obtained, sales team members begin calling, sending emails, etc.
- Some prospects are converted during this procedure, but most are not.
- At X Education, the average prospect conversion rate is around 30%.

Ideas for Increasing Lead Conversion

Because we want to achieve an 80% conversion rate, we need to find fresh prospects with a high degree of awareness.

Grouping the Leads

- Based on their tendency, or probability to convert, leads are categorised.
- A targeted collection of hot prospects is produced as a consequence.

Improved Communication

- We might only need to contact a lesser number of prospects, which would give us more influence.

Increase Conversion

- Since we focused on hot prospects that were more likely to convert, we would have a higher conversion rate and be able to meet the 80% goal.

Analysis Approach:

Data Cleaning

- For some category variables, the "Select" level reflects null values because no consumers selected anything from the list.
- Over 40% of the empty entries in a column were removed.
- Categorical categories with missing values were managed using value counts and other factors.
- Remove any sections that don't provide any context or significance for the study's goal. (tags, country)
- A few category factors were imputed.
- For some factors, new groups were made.
- Prospect ID and Lead Number, as well as columns with only one type of answer, were eliminated from the model.
- After examining the distribution, numerical data was replaced with the mean.

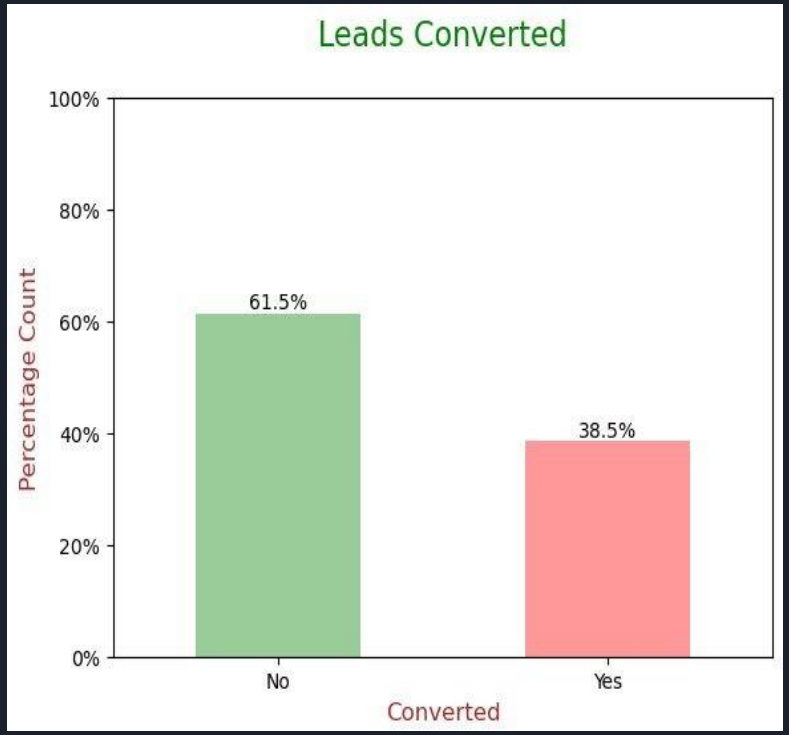
Data Cleaning

- To prevent prejudice in logistic regression models, skewed category entries were examined and removed.
- Page Views Per Visit and “TotalVisits” outliers were handled and limited.
- Data was standardised in some columns, such as lead source, and invalid numbers were rectified.
- Low frequency numbers were combined into a "Others" category.
- Variables with binary categorization were plotted.
- To guarantee the precision and integrity of the data, additional cleaning procedures were carried out.
  - By verifying casing styles, etc., invalid numbers were fixed and data in columns was standardised. (lead source has Google, google)

Exploratory Data Analysis

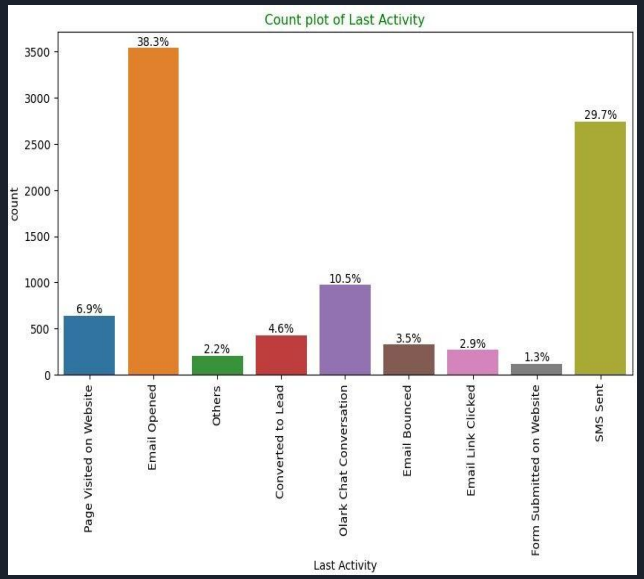
Data is imbalanced while the goal variable is being analyzed

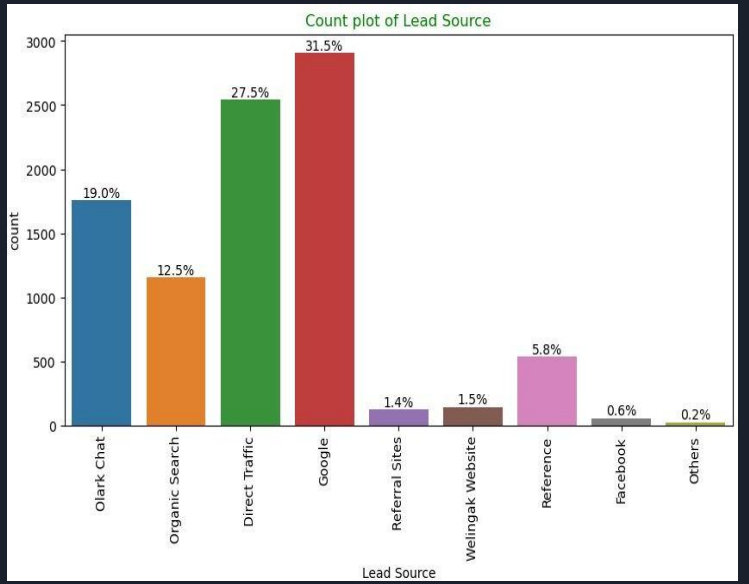
- Data is imbalanced while the goal variable is being analyzed
- A conversion percentage of 38.5% indicates that only 38.5% of the individuals have become leads.(Minority)
- While 61.5% of the visitors didn't become prospects. (Majority)



Exploratory Data Analysis

Univariate Analysis - Categorical Variables



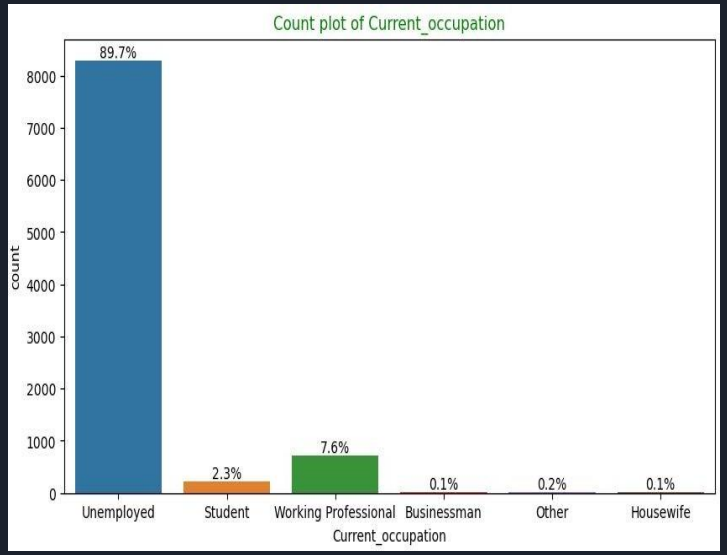


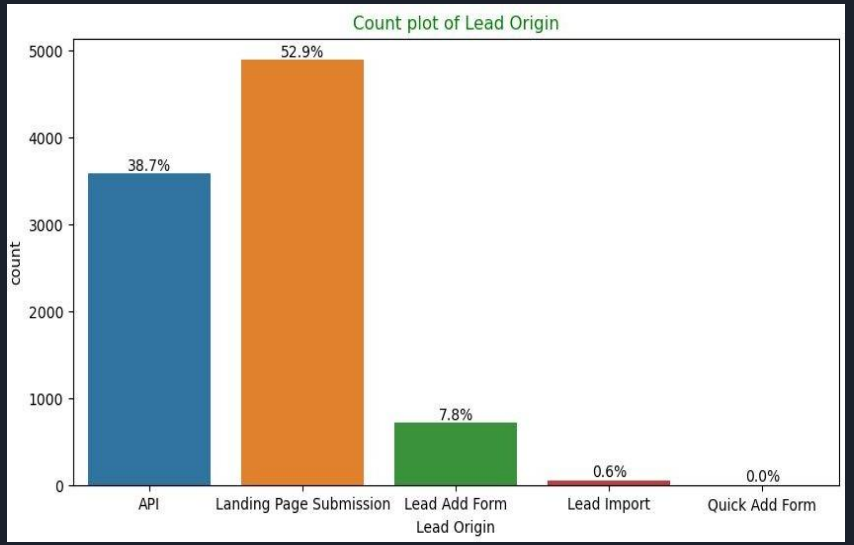
68% of consumers contributed to the last activity in SMS sent and email opened.

Lead Source: Google and direct traffic together account for 58% of the lead source.

Exploratory Data Analysis

Univariate Analysis - Categorical Variables



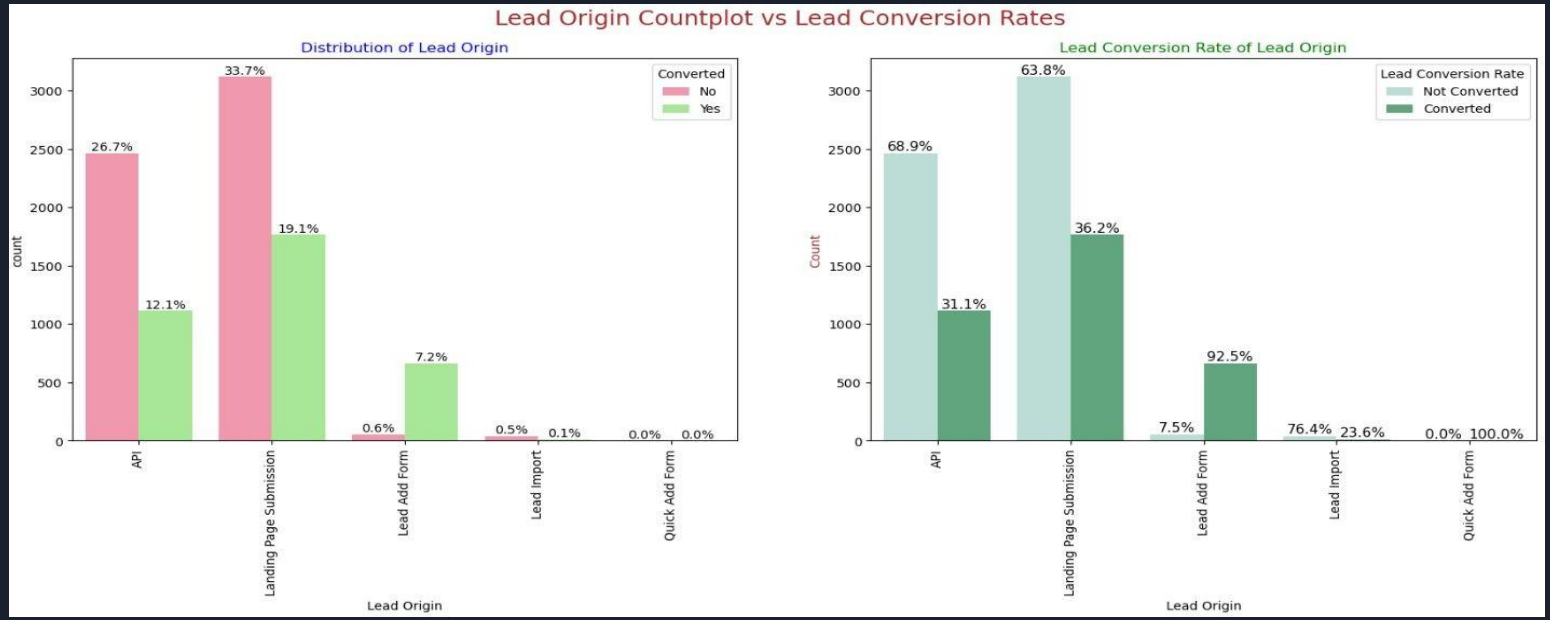


90% of the clients there are unemployed.

53% of users were recognised by "Landing Page Submission," while 39% were identified by "API."

Exploratory Data Analysis

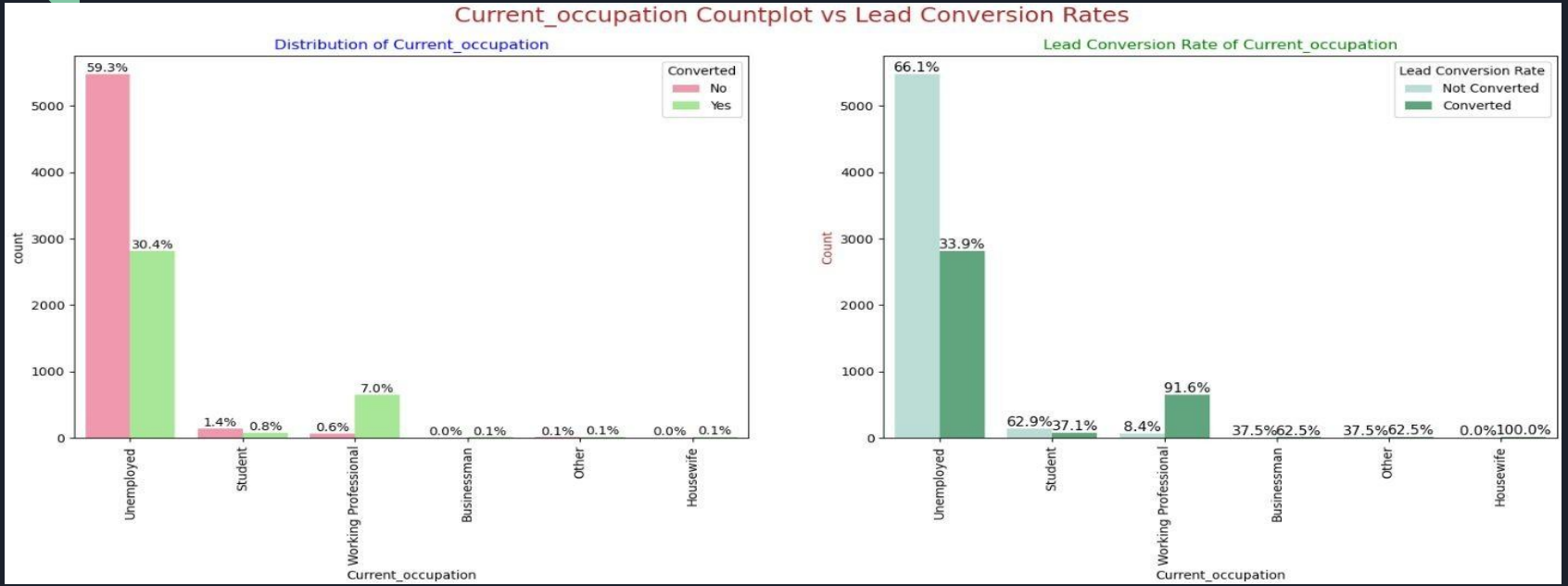
Bivariate Analysis - Categorical Variables



- The "Landing Page Submission" method generated about 52% of all leads, with a lead conversion rate (LCR) of 36%.
- A 31% lead conversion rate (LCR) and about 39% of consumers were found by the "API."

Exploratory Data Analysis

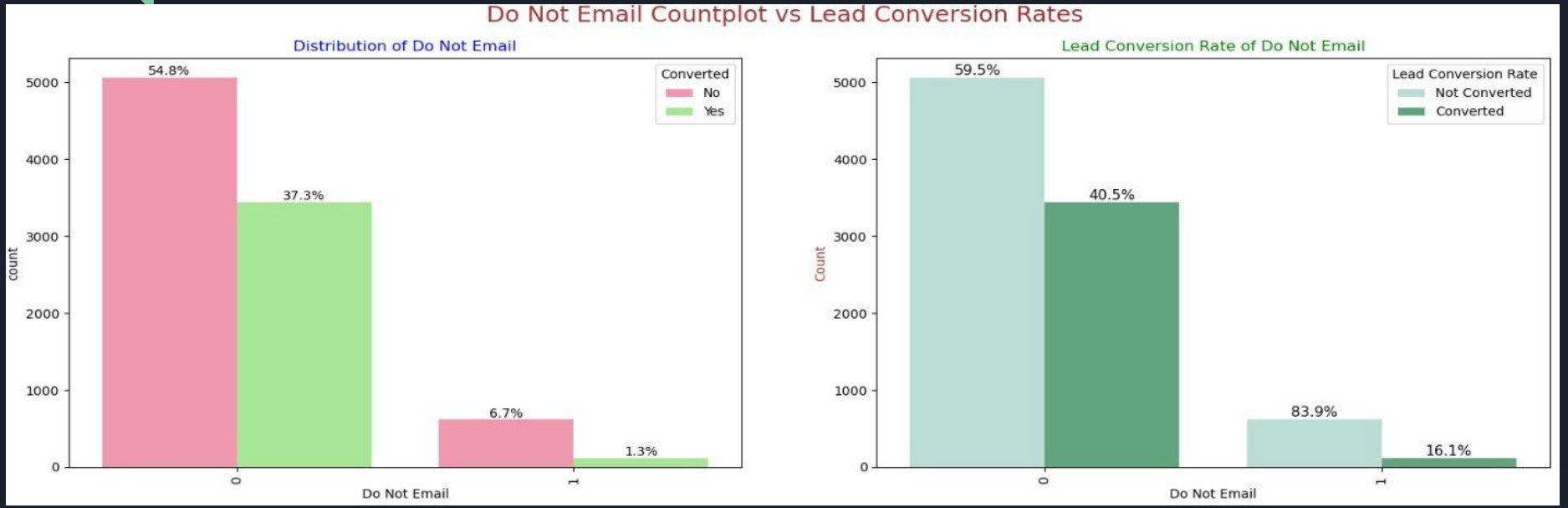
Bivariate Analysis - Categorical Variables



- 90% of clients are unemployed, and 34% of leads are converted into sales.
- Even though working professionals make up only 7.6% of all clients and have a lead response rate of almost 92% (LCR).

Exploratory Data Analysis

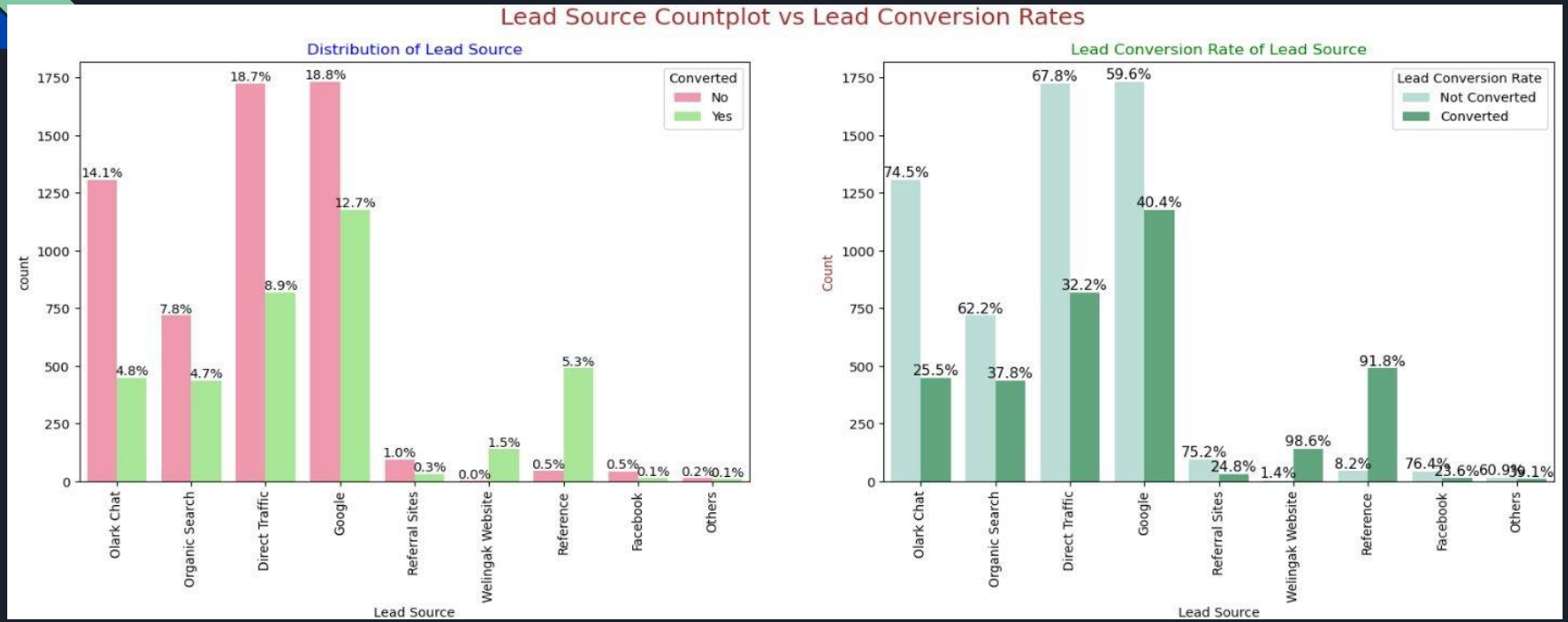
Bivariate Analysis - Categorical Variables



- 92% of the individuals have chosen not to receive emails about the training, and 40% of them have become leads.

Exploratory Data Analysis

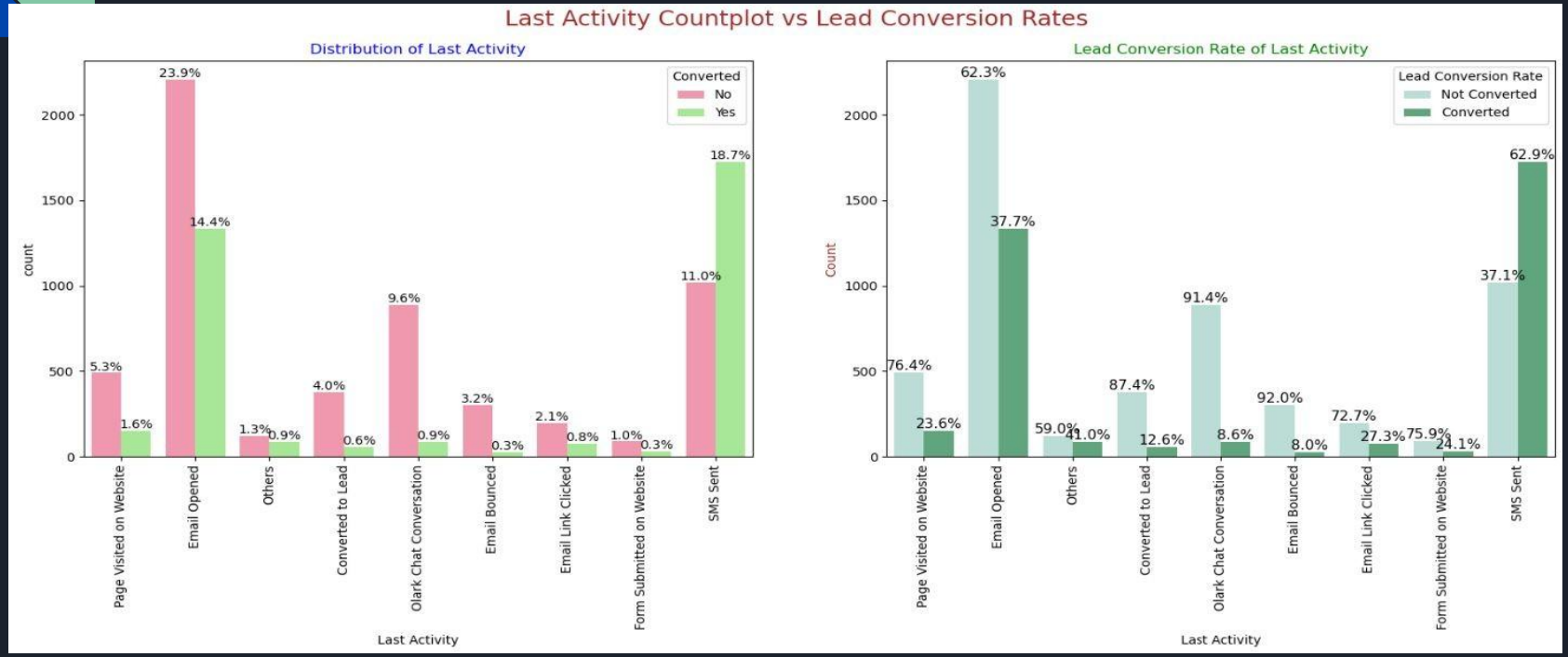
Bivariate Analysis - Categorical Variables



- LCR for Google is 40% out of 31% users.
- Less than Google, Direct Traffic provides 32% LCR with 27% clients.
- Additionally, Organic Search contributes 37.8% of LCR, but only 12.5% of consumers use it.
- Although Reference has an LCR of 91%, only about 6% of its clients come from this Lead Source.

Exploratory Data Analysis

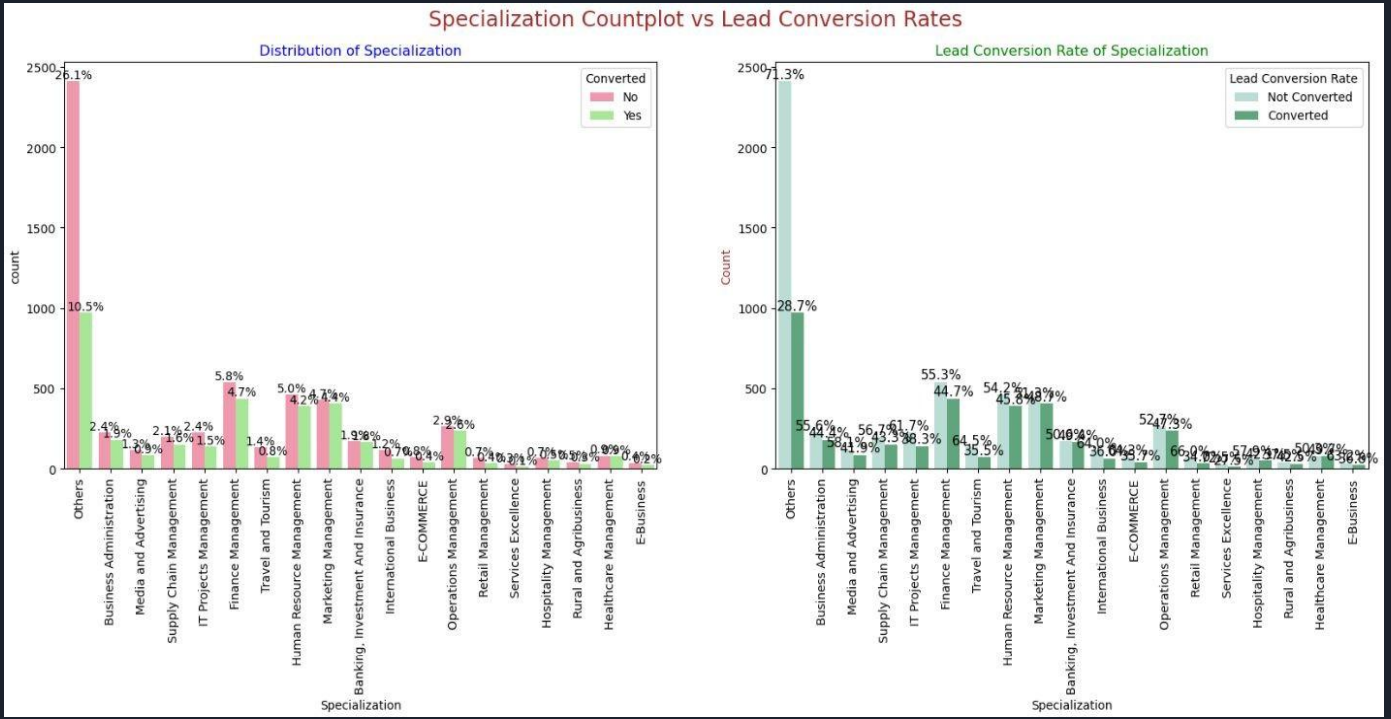
Bivariate Analysis - Categorical Variables



- A strong lead conversion rate of 63% for SMS Sent, with 30% of that coming from recent actions.
- 38% of the clients' most recent actions were "Email Opened"-related, with 37% of the lead.

Exploratory Data Analysis

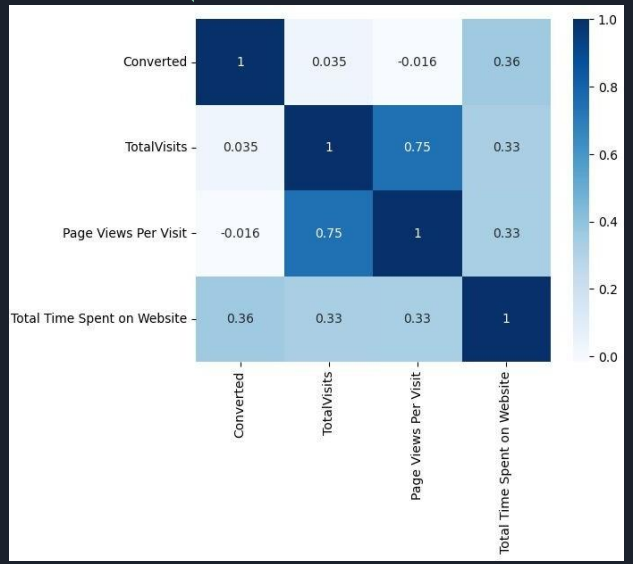
Bivariate Analysis - Categorical Variables

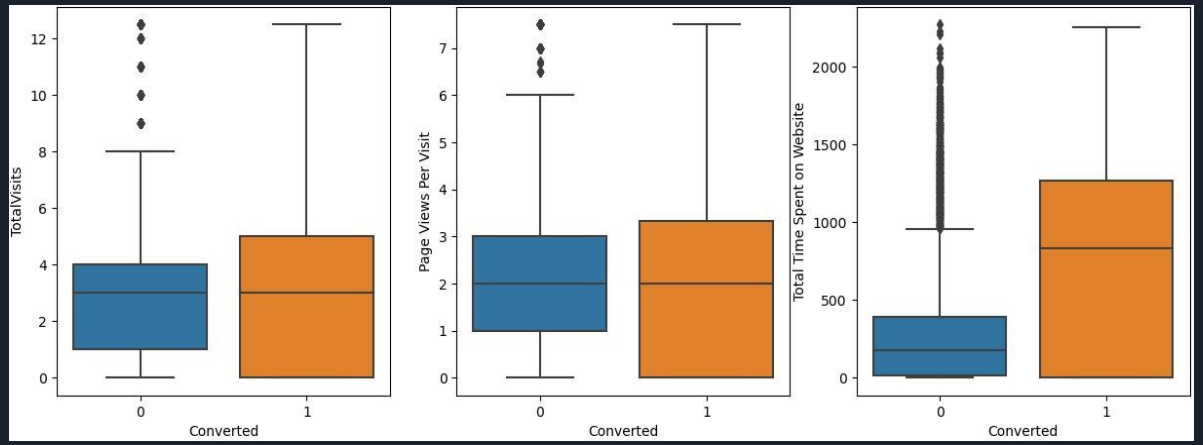


- The management of marketing, human resources, and finances contributes more favourably to the turn of leads than other specialisations.

Exploratory Data Analysis

Bivariate Analysis - Numerical Variables





According to the box-plot, past leads who spend more time on the website have a greater probability of being effectively converted than those who spend less time.

Data preparation prior to model construction :-

- ★ In earlier stages, binary level categorical categories were already assigned to 1 and 0.
- ★ For category factors such as Lead Origin, Lead Source, Last Activity, Specialization, and Current_occupation, fake features (one-hot encoded) were created.
- ★ Choosing a division ratio for the train and test sets was 70:30%.
- ★ Scaling of the features: The standardisation technique was used.
- ★ The Lead Origin_Lead Import and Lead Origin_Lead Add Form predictor factors were eliminated after the relationships were examined.

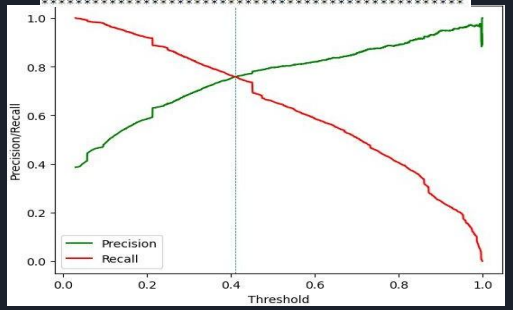
Model construction :-

- ★ There are numerous dimensions and characteristics in the data collection.
- ★ This could result in lengthy computations and lower model performance.
- ★ Recursive Feature Elimination (RFE) should be used, and only essential categories should be chosen.
- ★ After that, we can directly adjust the model.
- ★ Results of the RFE: 48 columns before and 15 columns after.
- ★ By eliminating variables with a p-value higher than 0.05, models were built manually through the feature reduction procedure.
- ★ After four iterations, Model 4 appears to be steady with
    - ○ significant p-values that are below the cutoff (p-values 0.05).
    - ○ There is no indication of multicollinearity for VIFs under 5.
- ★ As a result, logm4 will serve as our final model, and we'll use it for model evaluation before using it to make forecasts.

Evaluation of the Model

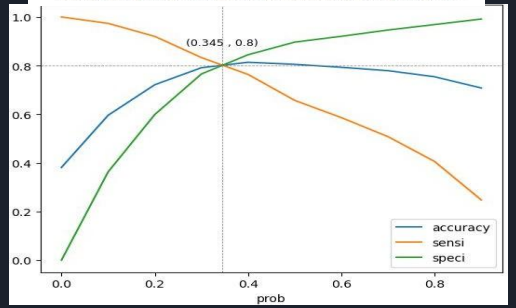
Train data set - 0.345 was chosen as the limit after assessment measures from both plots were examined.

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

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

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



Matrix of Confusion and Evaluation Metrics with Cutoff of 0.41

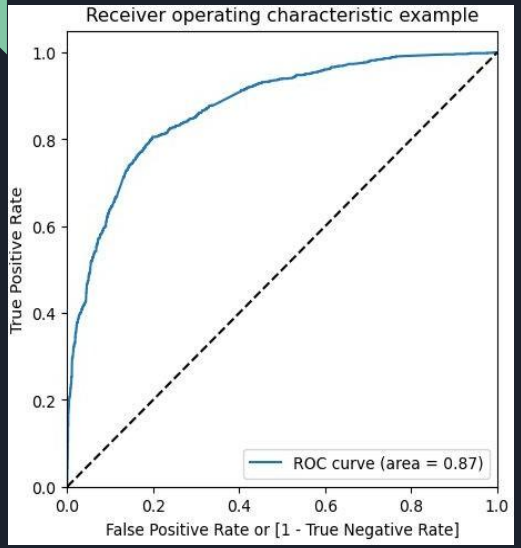
```
**************************************************

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

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



Matrix of Confusion and Evaluation Metrics with Cutoff of 0.345

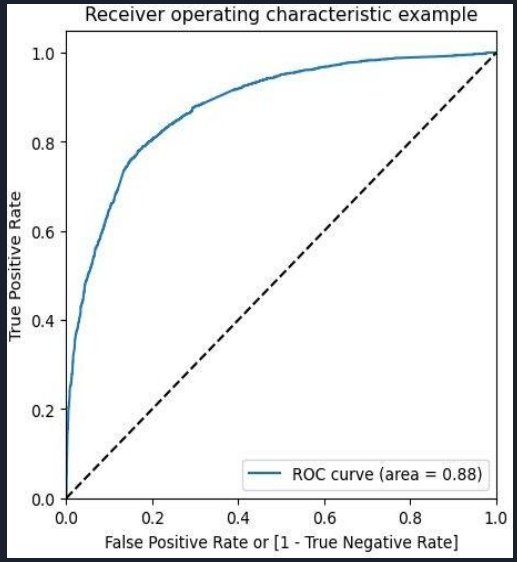
Evaluation of the Model

Train data set - 0.345 was chosen as the limit after assessment measures from both plots were examined.



Test Data Set for the ROC Curve:

- A decent forecasting model is one with a 0.87 out of 1 area under the ROC curve.
- A model with a high true positive rate and a low false positive rate at all cutoff values has a curve that is as near to the plot's upper-left corner as possible.



Test Data Set for the ROC Curve:

- A decent forecasting model is one with a 0.88 out of 1 area under the ROC curve.
- A model with a high true positive rate and a low false positive rate at all cutoff values has a curve that is as near to the plot's upper-left corner as possible.

Evaluation of the Model

Confusion Matrix & Metrics - Train 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
```

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

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
```

- The model obtained a sensitivity of 80.05% in the train set and 79.82% in the test set using a cut-off value of 0.345.
- In this situation, sensitivity refers to how many leads the model accurately predicts out of all possible leads that end up turning.
- A goal sensitivity of about 80% had been established by the CEO of X Education.
- Additionally, the model's precision, which was 80.46%, met the goals of the research.

Final Recommendations :-

- According to the issue statement, boosting lead conversion is essential for X Education's expansion and success. In order to accomplish this, we have created a regression model that can assist us in determining the most important variables that affect offer yield.
- The following features should be prioritised in our marketing and sales efforts to improve lead conversion, according to the characteristics we found to have the greatest positive coefficients.
  - Lead Source_Reference: 2.93.
  - Lead Source_Welingak Website: 5.39.
  - Last Activity_SMS Sent: 2.05.
  - Current_occupation_Working Professional: 2.67.
  - Total Time Spent on Website: 1.05.
  - Last Activity_Others: 1.25.
  - Lead Source_Olark Chat: 0.91.
  - Last Activity_Email Opened: 0.94.
- Additionally, we have found characteristics with negative coefficients that might point to areas that could use development. These consist of:
  - Specialization in Others: -1.20
  - Specialization in Hospitality Management: -1.09
  - Lead Origin of Landing Page Submission: -1.26

Final Recommendations :-

Increasing prospect turn rates is necessary:

- ➔ For tailored marketing tactics, pay attention to characteristics with high correlation coefficients.
- ➔ Create tactics to entice top-performing lead sources to send you high-quality leads.
- ➔ Adapt communication methods depending on the effect of lead engagement.
- ➔ Communicate with working people in a relevant way.
- ➔ On the Welingak website, additional spending can be made on things like ads.
- ➔ Rewards/discounts for supplying references that result in leads promote supplying more references.
- ➔ Working professionals should be actively pursued because they convert well and are more likely to have the money to pay higher fees.

The categories listed below can be enhanced:

- ➔ Review the specialisation options' negative factors.
- ➔ Look for places for growth in the landing page submission procedure.

Thank You