Lead Score Case Study

Submitted by :

Naval Kishore

Mukul Pant

Muthukumar Natarajan

Lead Score Case Study

- **Problem Statement** :

- X Education sells online courses to industry professionals. 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. When these 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%.

- **Business Goal**:

- X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- The company needs a model wherein you a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with 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%.

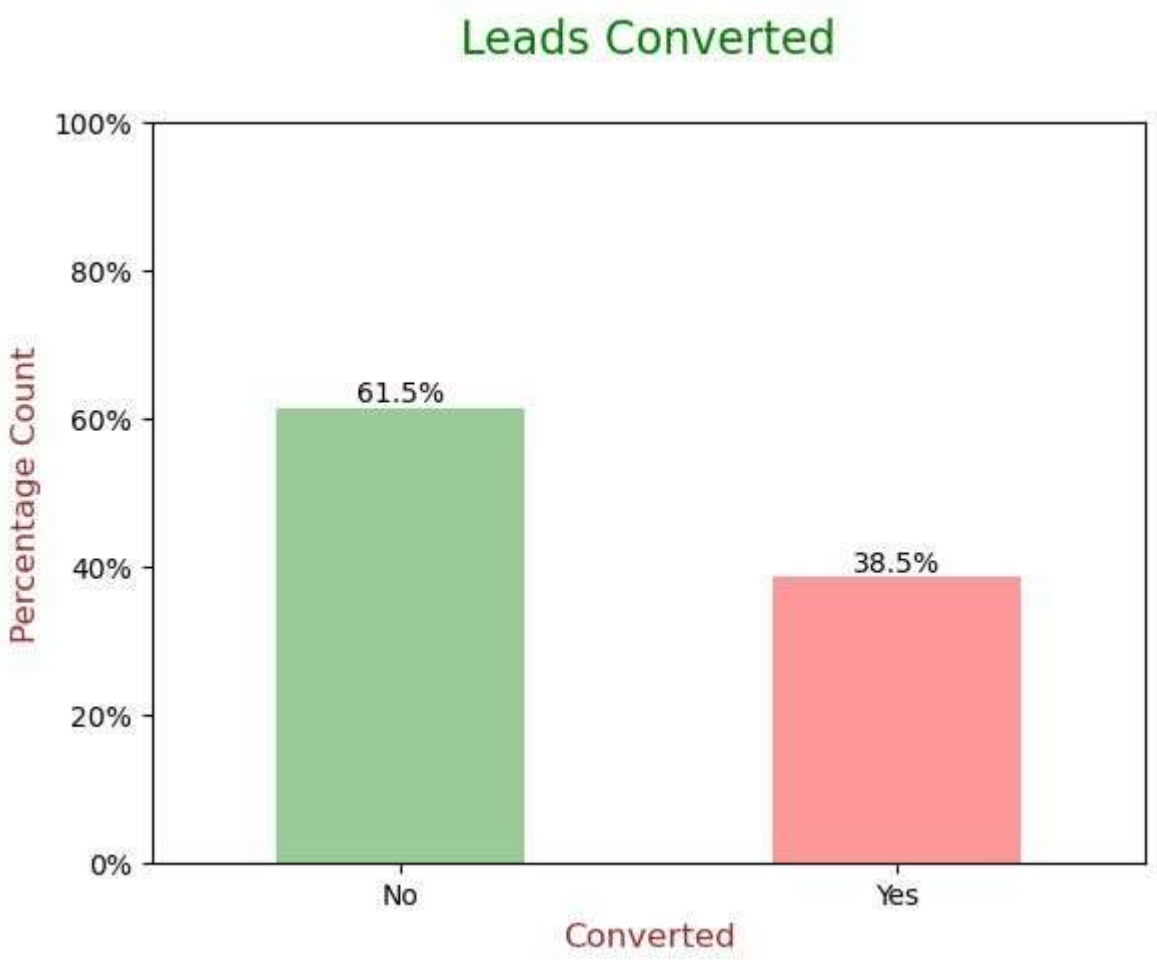
Analysing Approach

- Importing the data for analysis
- Sanity Check.
- Exploratory Data Analysis.
- Feature Scaling.
- Splitting the data into Test and Train dataset.
- Logistic Regression model and calculate Lead Score.
- Evaluating the model by using different metrics.
- Applying the best model in Test data based on the Sensitivity and Specificity Metrics.

Data Cleaning

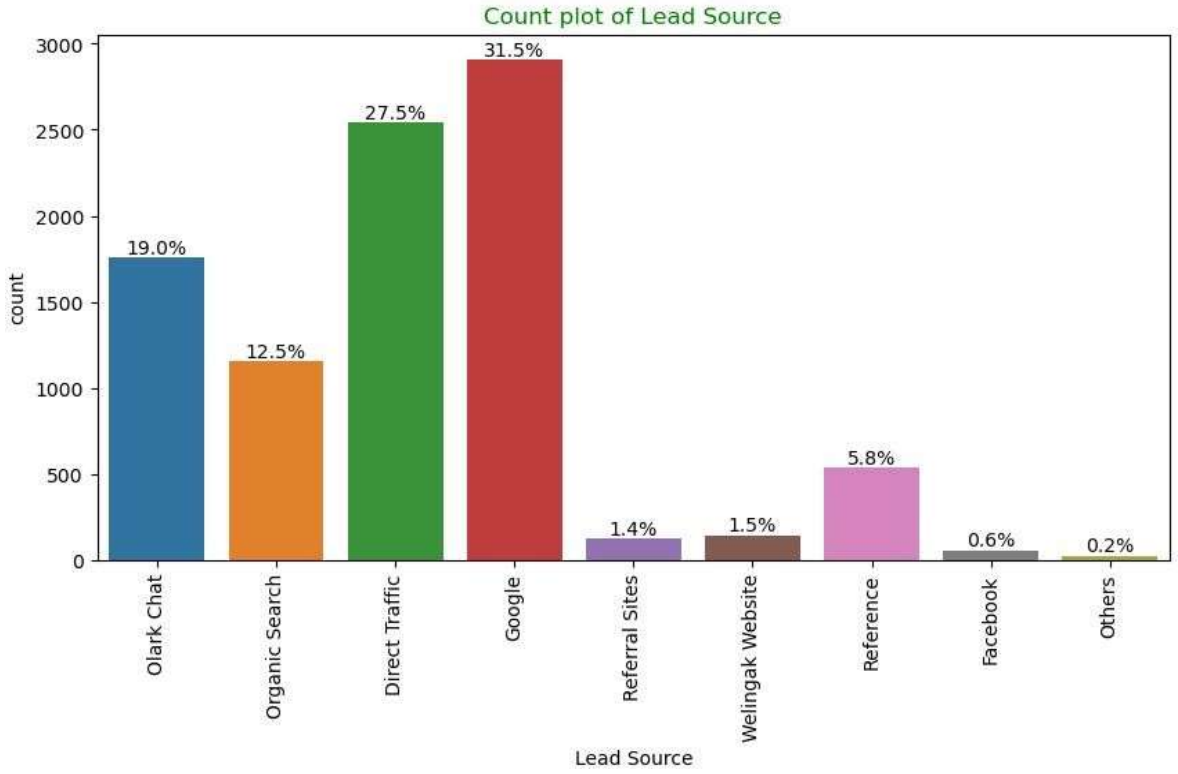
- "**Select**" level represents null values for some categorical variables, as customers did not choose any option from the list.
- Columns with over 40% null values were dropped.
- Missing values in categorical columns were handled based on value counts and certain considerations.
- Drop columns that don't add any insight or value to the study objective (tags, country)
- Additional categories were created for some variables.
- 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.
- Outliers in TotalVisits and Page Views Per Visit were treated and capped.

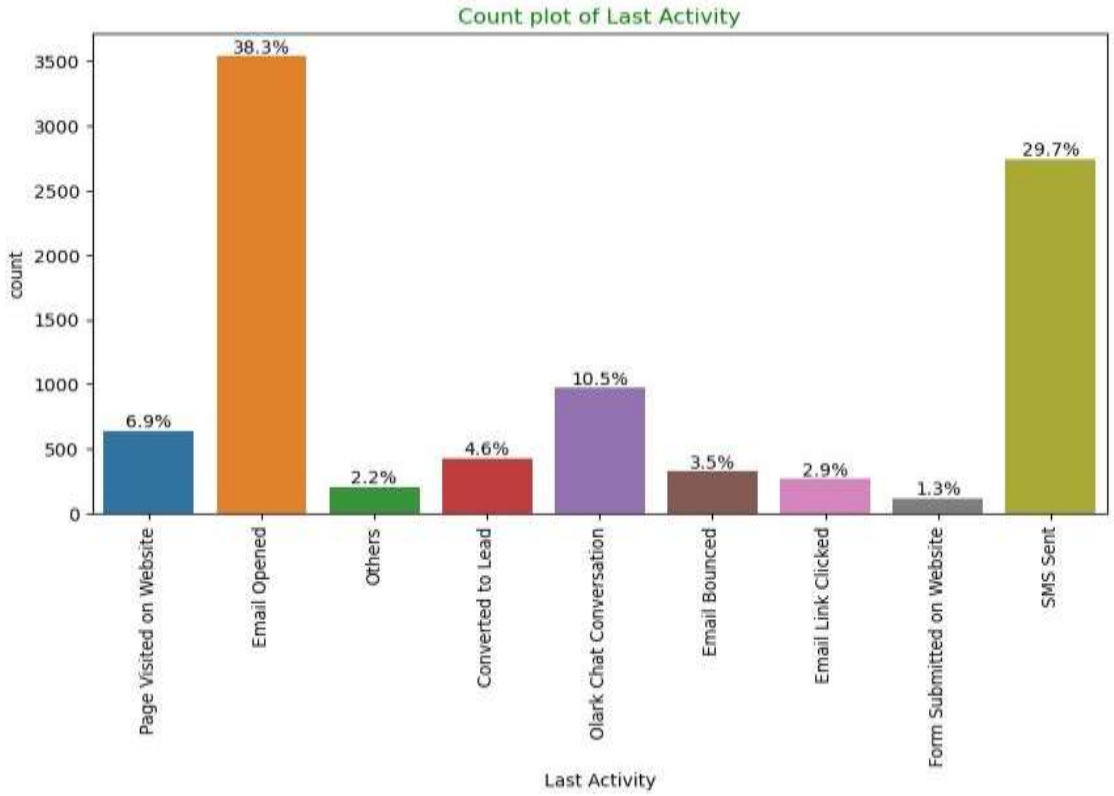
Exploratory Data Analysis



- Conversion rate is of 38.5%, meaning only 38.5% of the people have converted to leads.(Minority)
- While 61.5% of the people didn't convert to leads. (Majority)

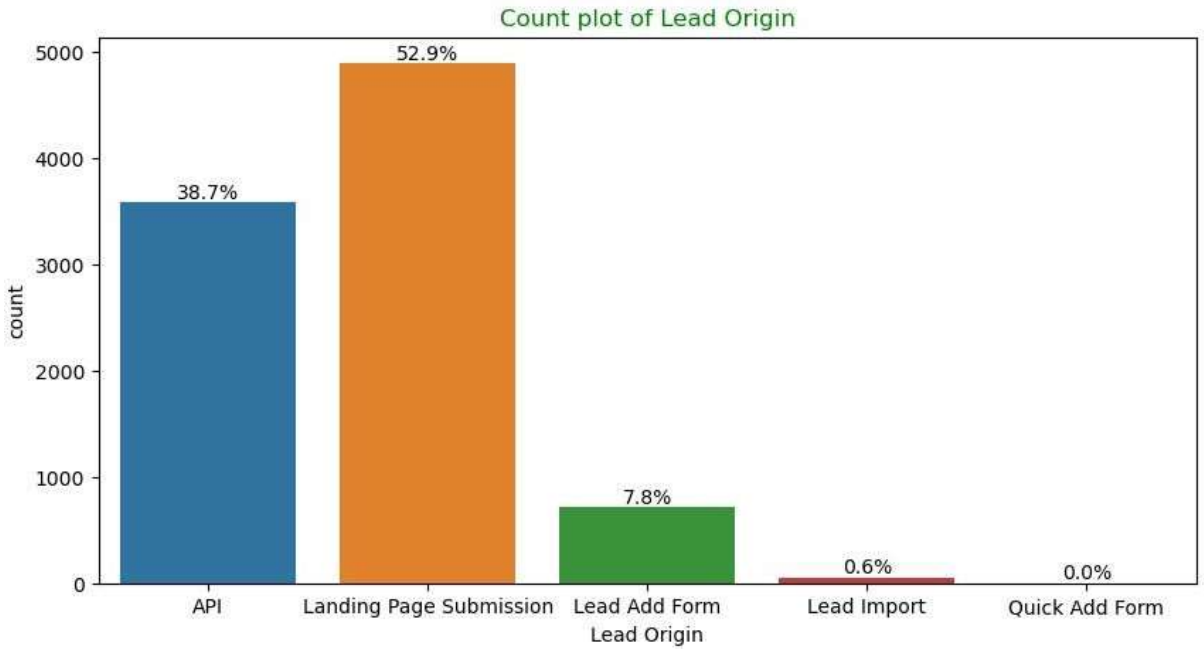
Univariate Analysis – Categorical Variables

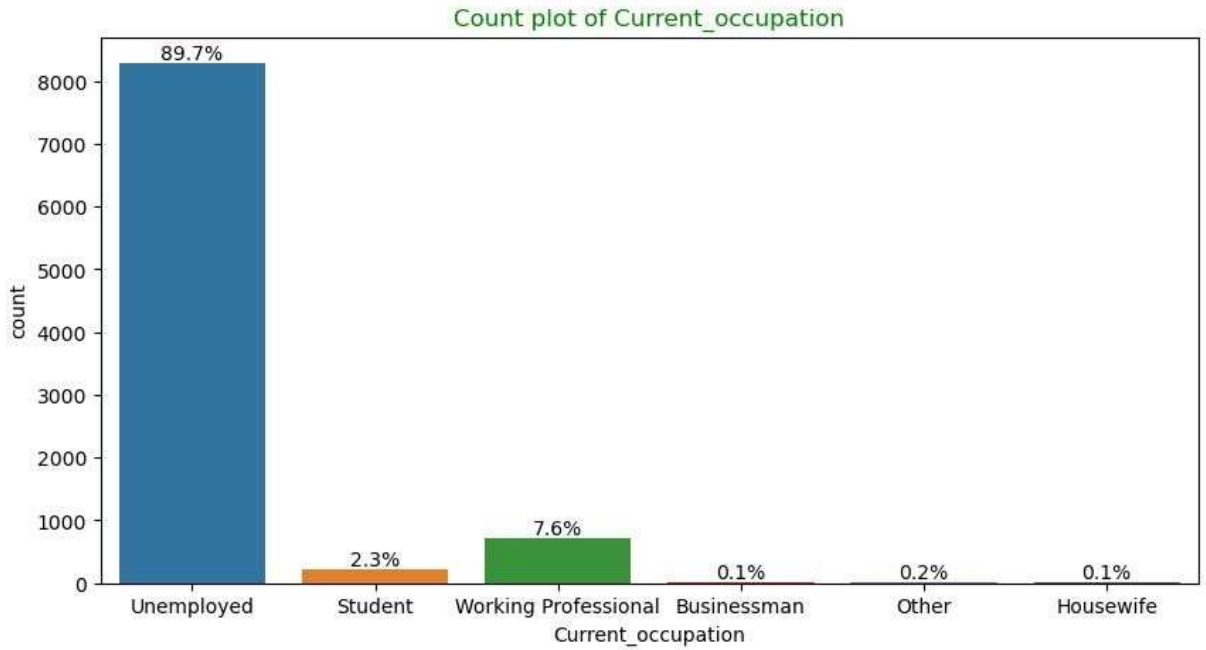




- **Lead Source:** 58% Lead source is from Google & Direct Traffic combined.
- **Last Activity:** 68% of customers contribution in SMS Sent & Email Opened activities.

Univariate Analysis – Categorical Variables

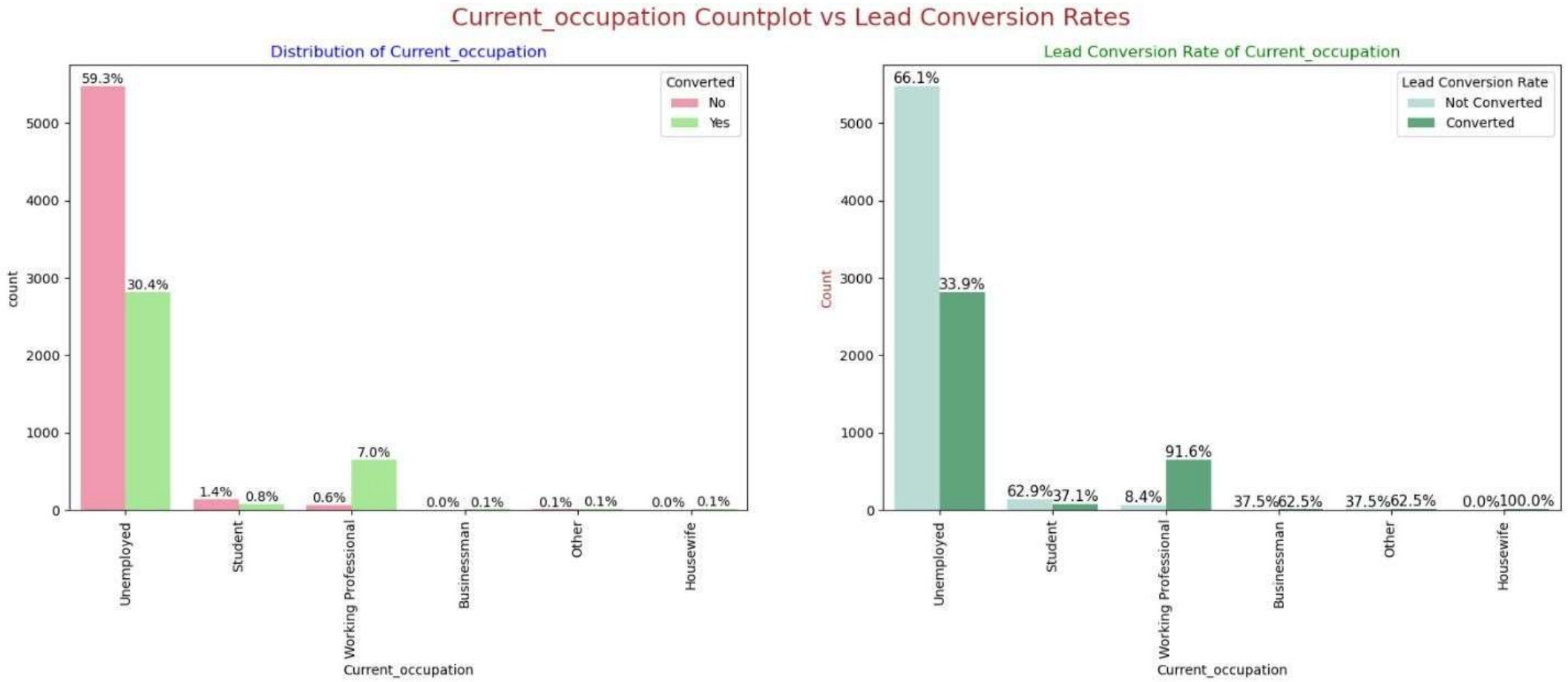




- **Lead Origin:** "Landing Page Submission" identifies 53% of customers, "API" identifies 39%.

- **Current_occupation:** It has 90% of the customers as Unemployed.

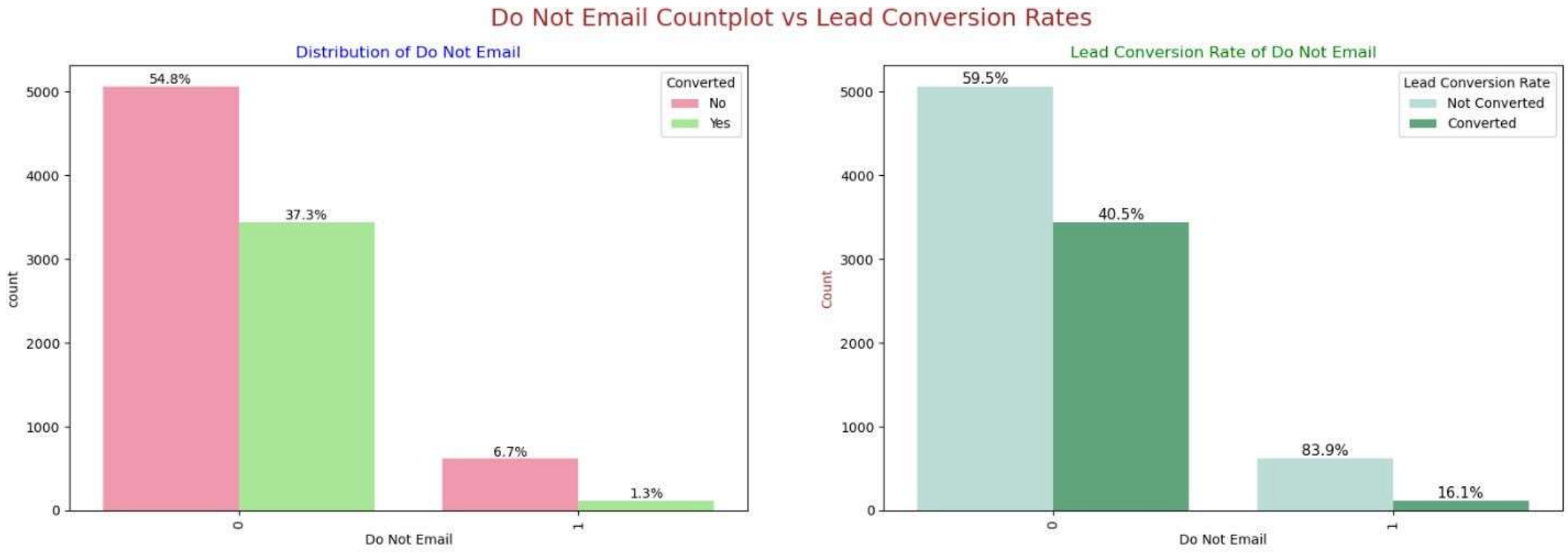
Bivariate Analysis for Categorical Variables



Current_occupation:

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

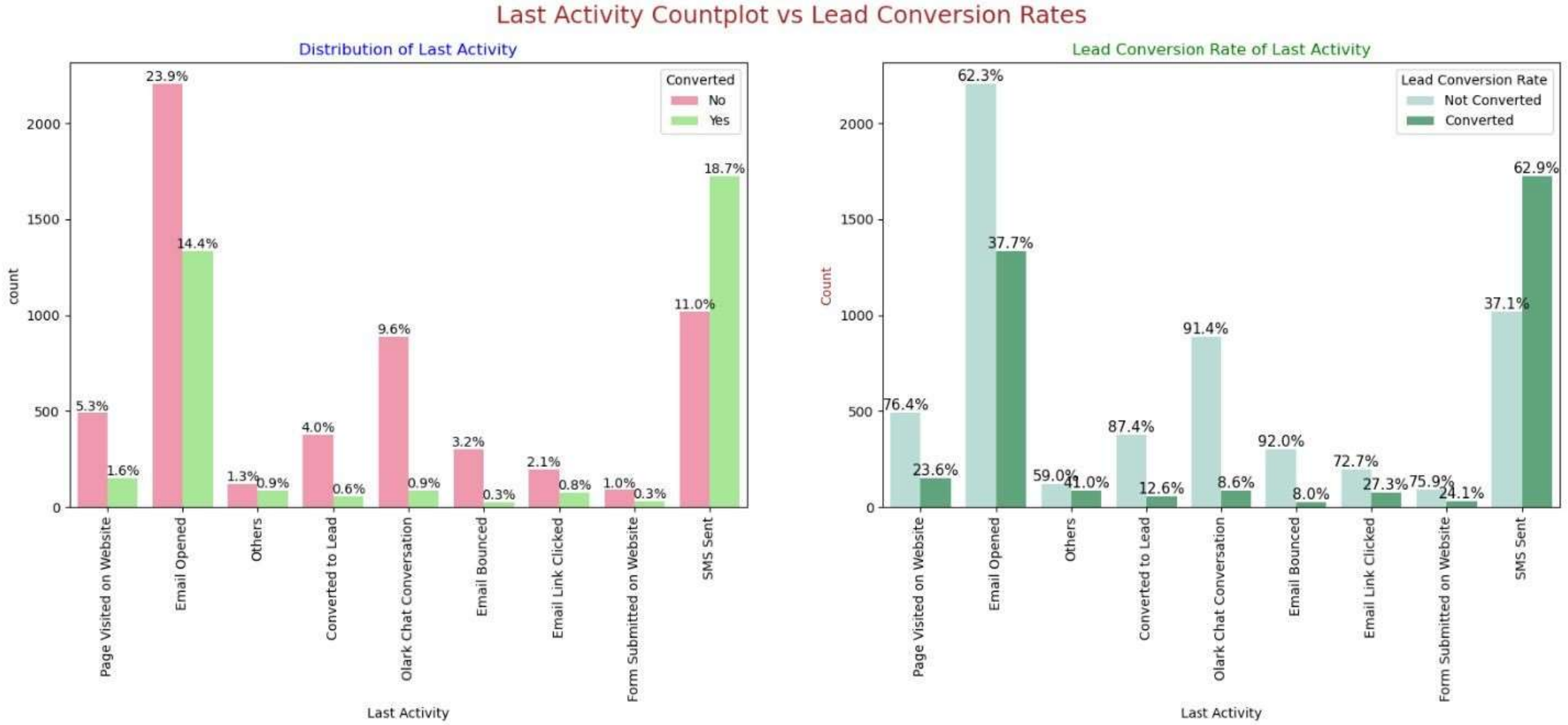
Bivariate Analysis for Categorical Variables



Do Not Email:

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

Bivariate Analysis for Categorical Variables



Last Activity:

- '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.

Data Preparation

- 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 Evaluation

Train Data Set

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
```

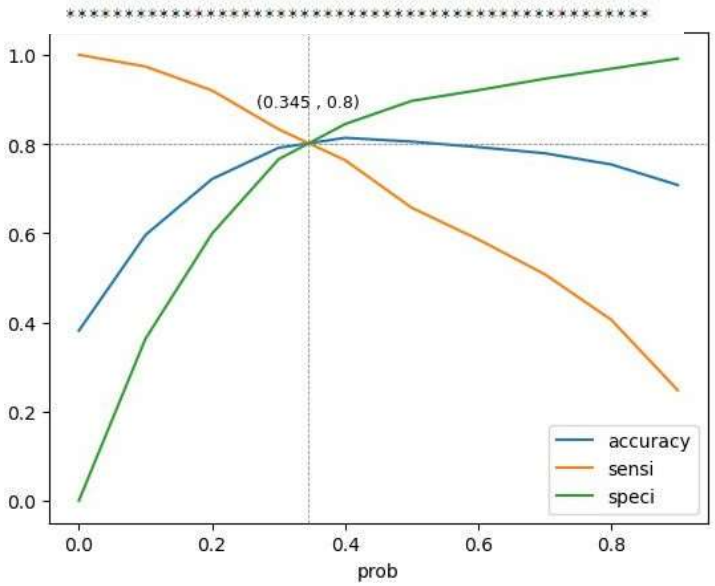
Confusion Matrix & Evaluation Metrics with 0.41 as cutoff

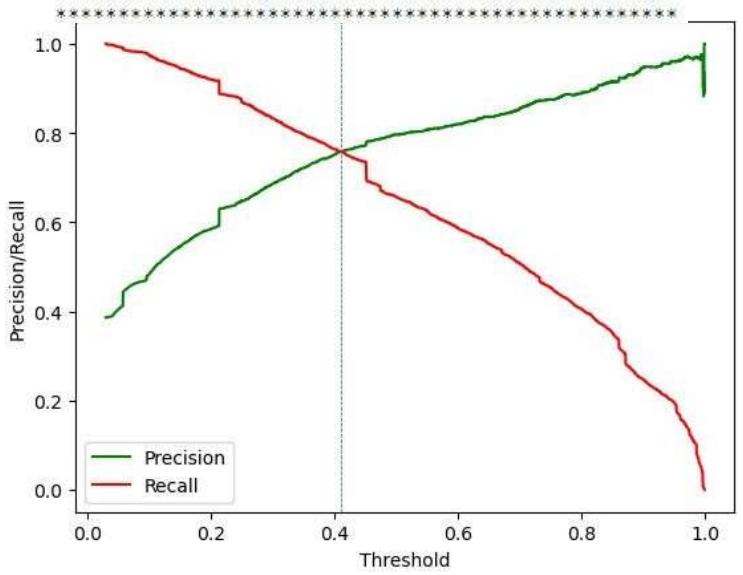
```
**************************************************

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
```





Model Evaluation

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
```

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
```

- 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.

Recommendation based on Final Model

- As per the problem statement, increasing lead conversion is crucial for the growth and success of X Education. To achieve this, we have developed a regression model that can help us identify the most significant factors that impact lead conversion.
- Accuracy, Sensitivity and Specificity values of test set are around 81%, 79% and 82% which are approximately closer to the respective values calculated using trained set.
- The top 3 variables that contribute for lead getting converted in the model are
  - Total time spent on website
  - Lead Add Form from Lead Origin
  - Had a Phone Conversation from Last Notable Activity
- Engage **working professionals** with tailored messaging.
- Working professionals to be aggressively targeted as they have high conversion rate and will have better financial situation to pay higher fees too.

- Analysing negative coefficients in specialization offerings.
- Review landing page submission process for areas of improvement.