X Education - Lead Scoring Case Study

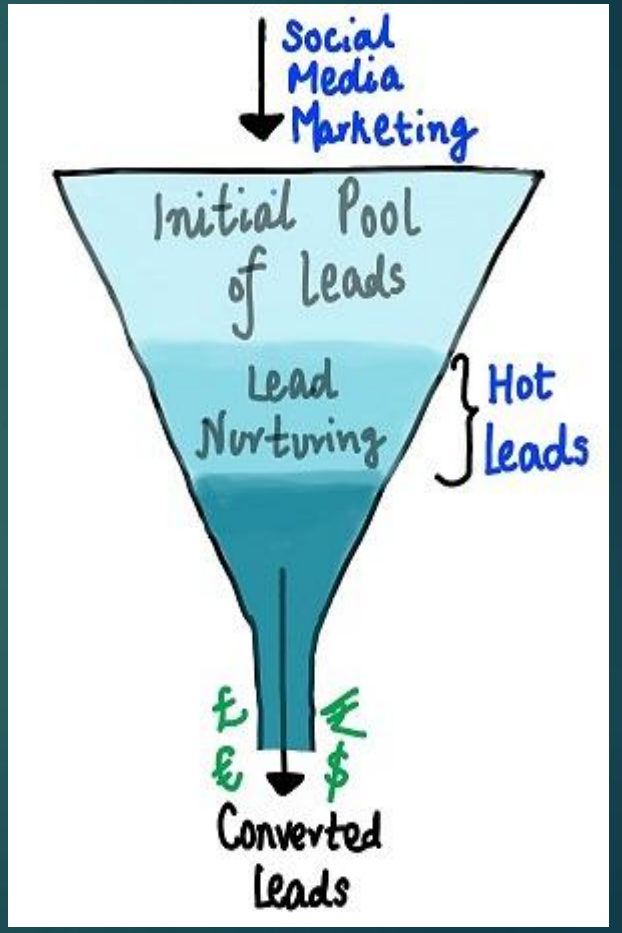
Filtration of convertible leads out of total leads pool so the conversion rate of X education can be improved from 30% to 80%

We are going to perform the below steps in the following order.

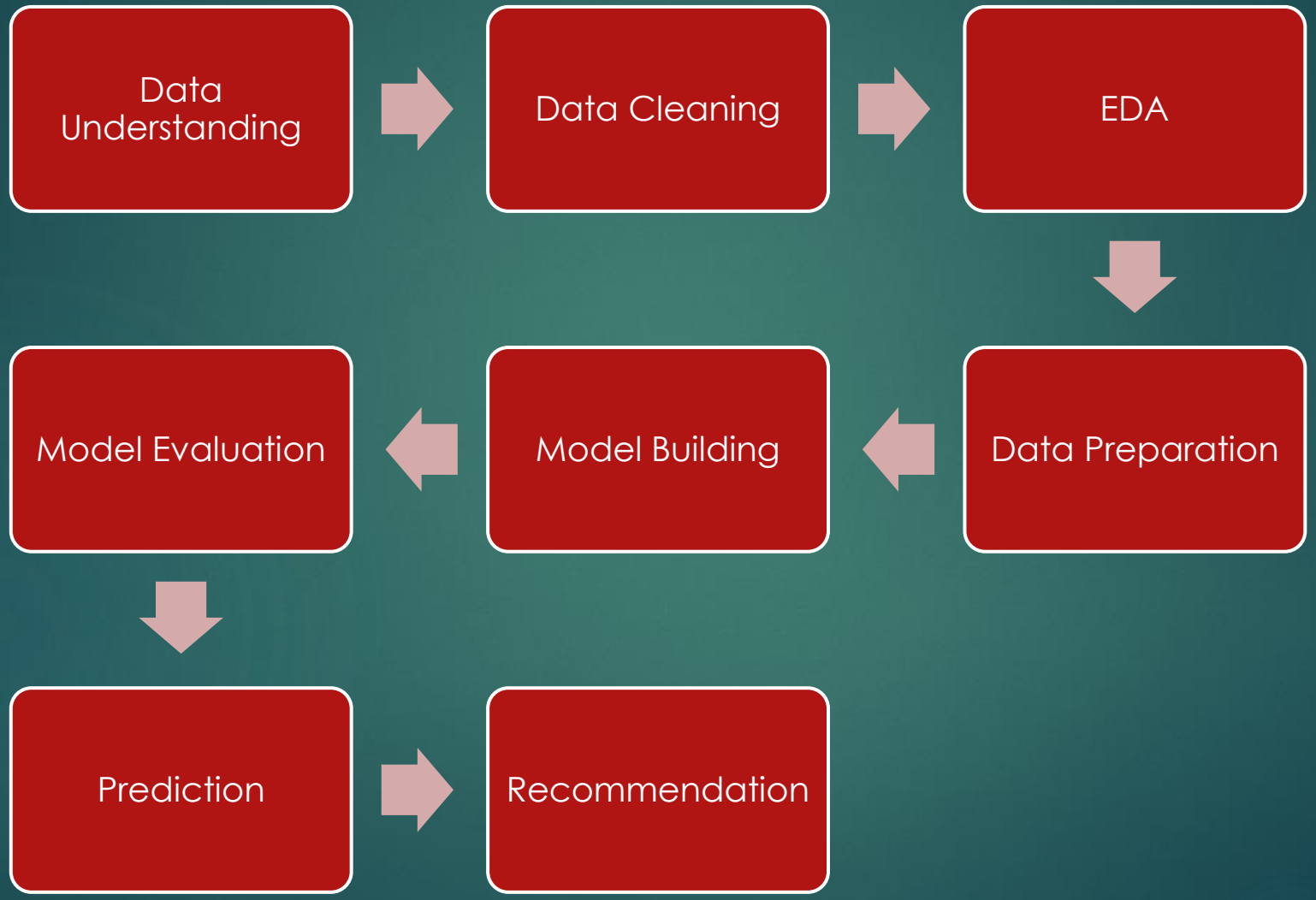
- Step 1: Importing Libraries & Data
- Step 2: Reading & Understanding the Data
- Step 3: Data Cleaning
- Step 4: Data Analysis
- Step 5: Data Preparation
- Step 6: Train-Test Split
- Step 7: Feature Scaling
- Step 8: Model Building
- Step 9: Model Evaluation
- Step 10: Making Predictions on test set

Business Objective

- To Increase the leads conversion from 30% to at least 80% by filtering out potential leads.
- To create a model that can rank all the leads based on leads score so sales team can prioritize them.



Steps of Analysis



Data Understanding and Cleaning

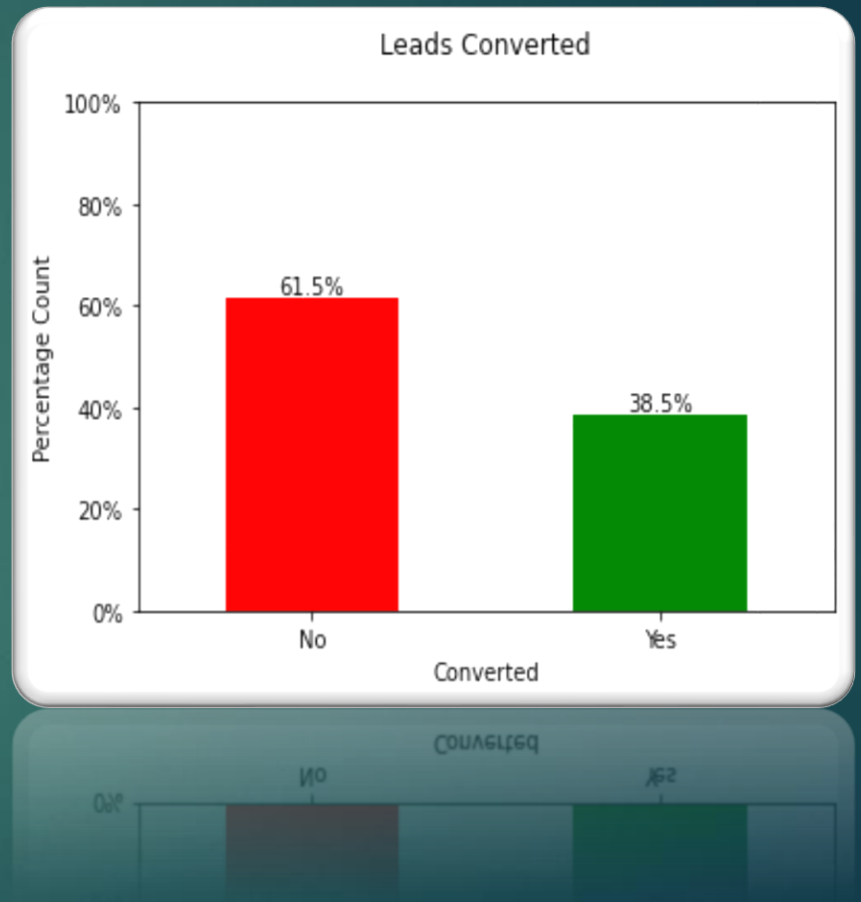
- The data have Approx. 9k records with 37 columns.
- During Data cleaning process we discovered 7 columns having more than 40% of missing value. These columns were dropped.
- Columns such as Lead number and Prospect ID were also dropped as they had unique values for every entry indicating it was an ID column proving no value add in model building.
- Imputation was done for categorical columns having missing values.
- In numerical columns mode imputation was used for handling missing values.
- Categorical columns having skewed data were dropped as they might bias our logistic regression model.

Data Understanding and Cleaning

- Winsorization method was used for outliers treatment.
- Other cleansing steps
  - Standardized invalid values.
  - Encoding was done for categorical variables.
  - Standard Casing was done for categorical variables.

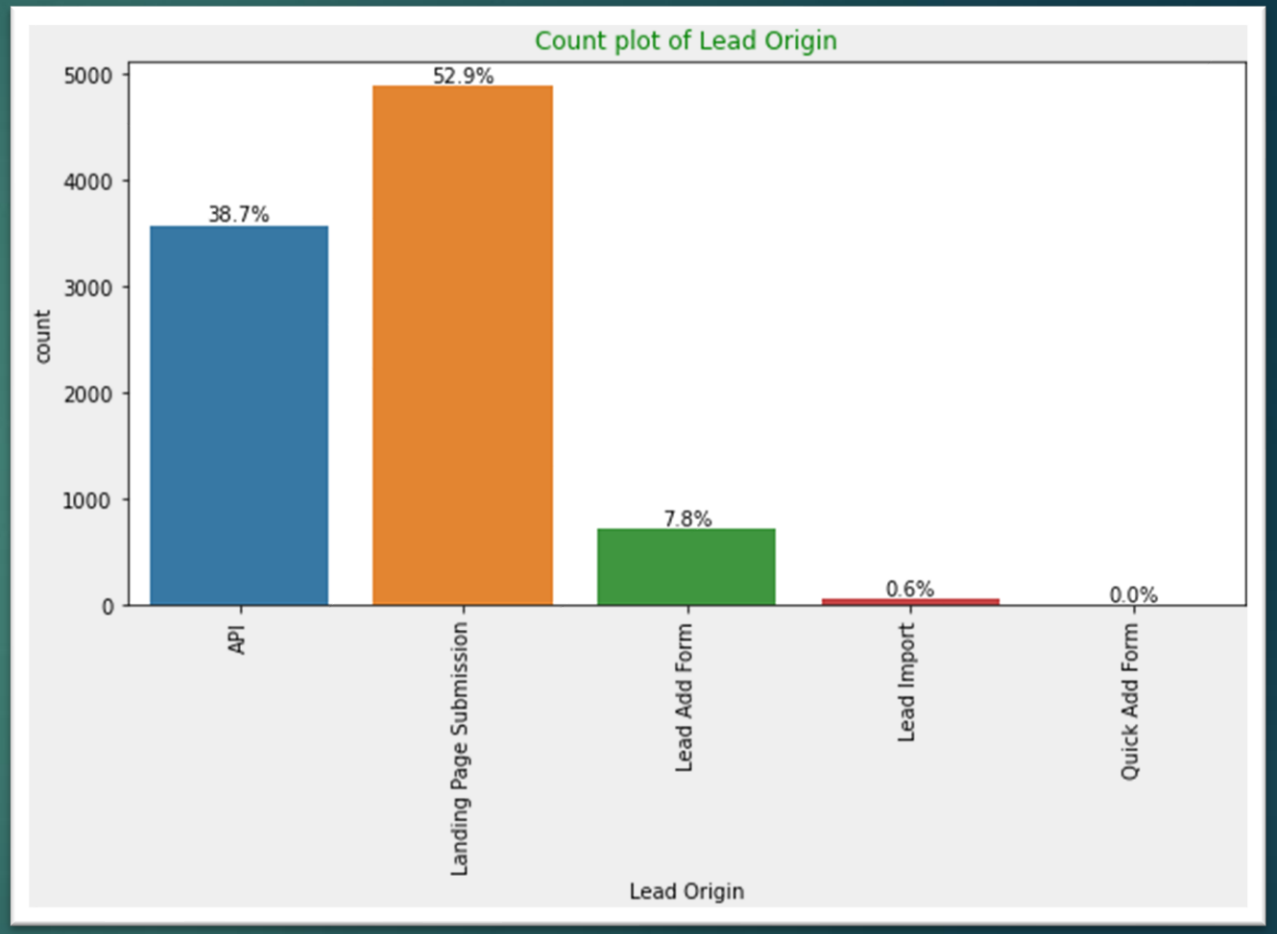
Exploratory Data Analysis (EDA)

- Univariate Analysis
  - 38.5% of the people have converted to leads, While 61.5% of the people didn't convert to leads.
  - This Explains the data is in line with the problem statement given by X Education.



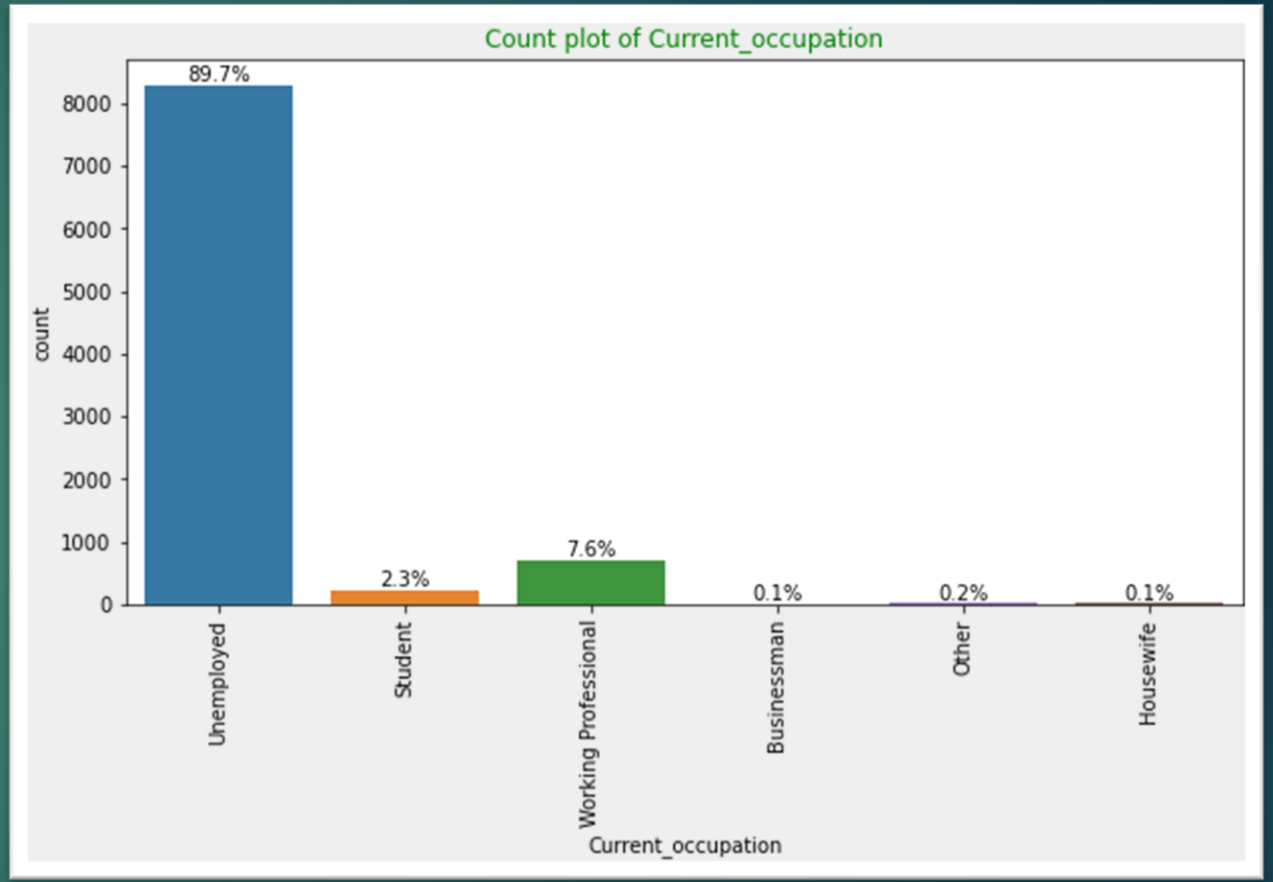
Exploratory Data Analysis (EDA)

- Univariate Analysis
  - The source for 53% of customers is "Landing Page Submission," while "API" accounts for 39%.



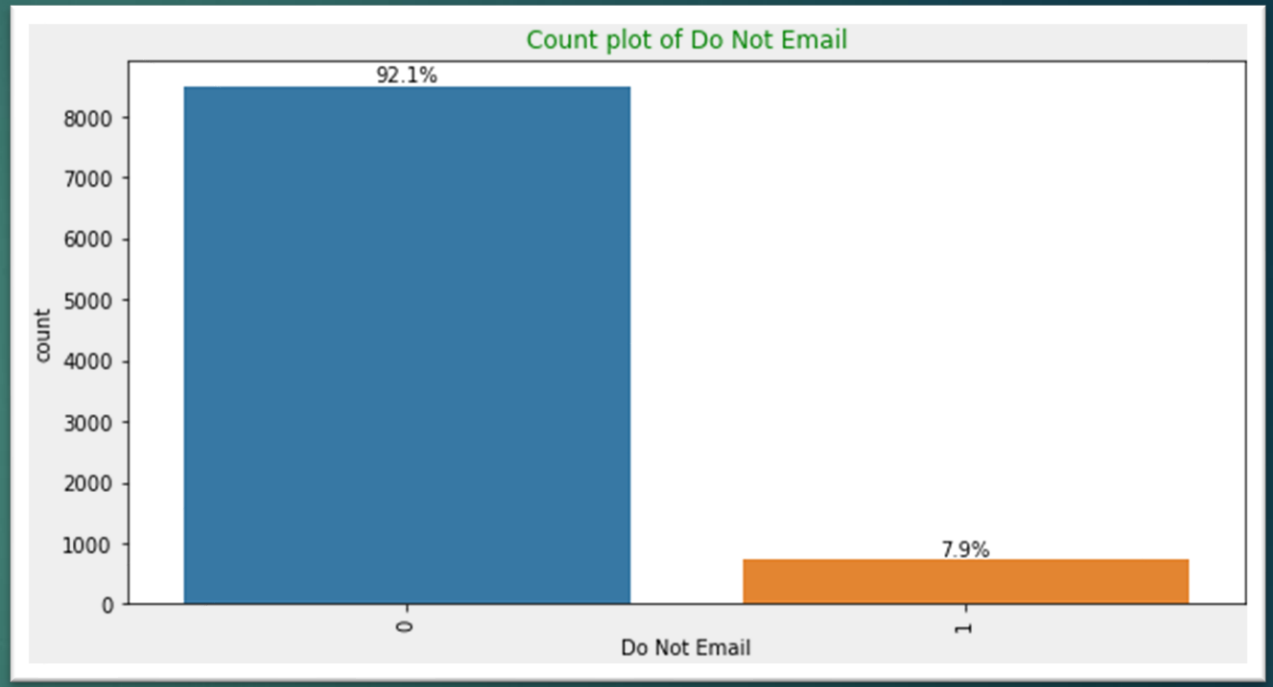
Exploratory Data Analysis (EDA)

- Univariate Analysis
  - Approximately 90% of customers are classified as "Unemployed“.
  - Rest of the categories are minimal.



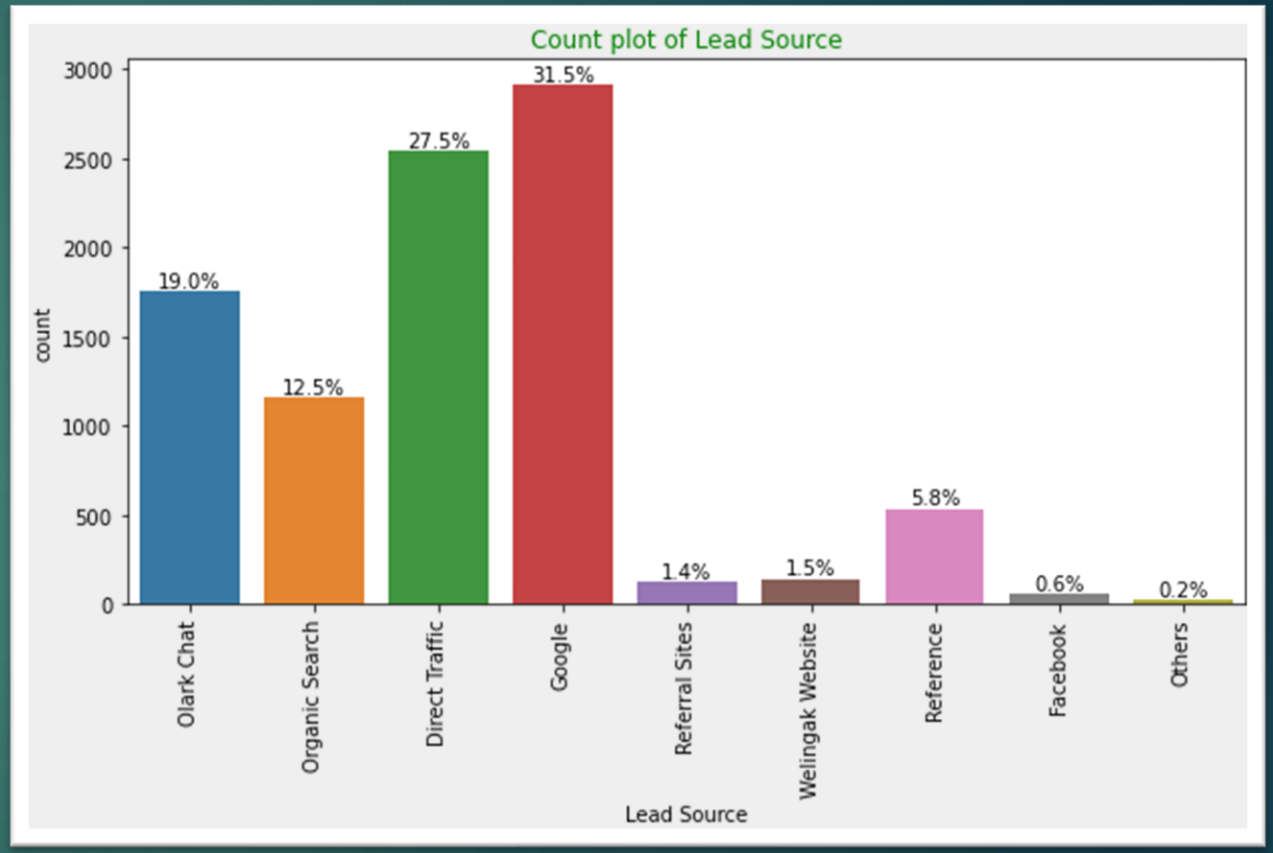
Exploratory Data Analysis (EDA)

- Univariate Analysis
  - Around 92% of individuals have indicated their preference not to receive emails regarding the course.
  - A collective approach should be given to change this numbers.



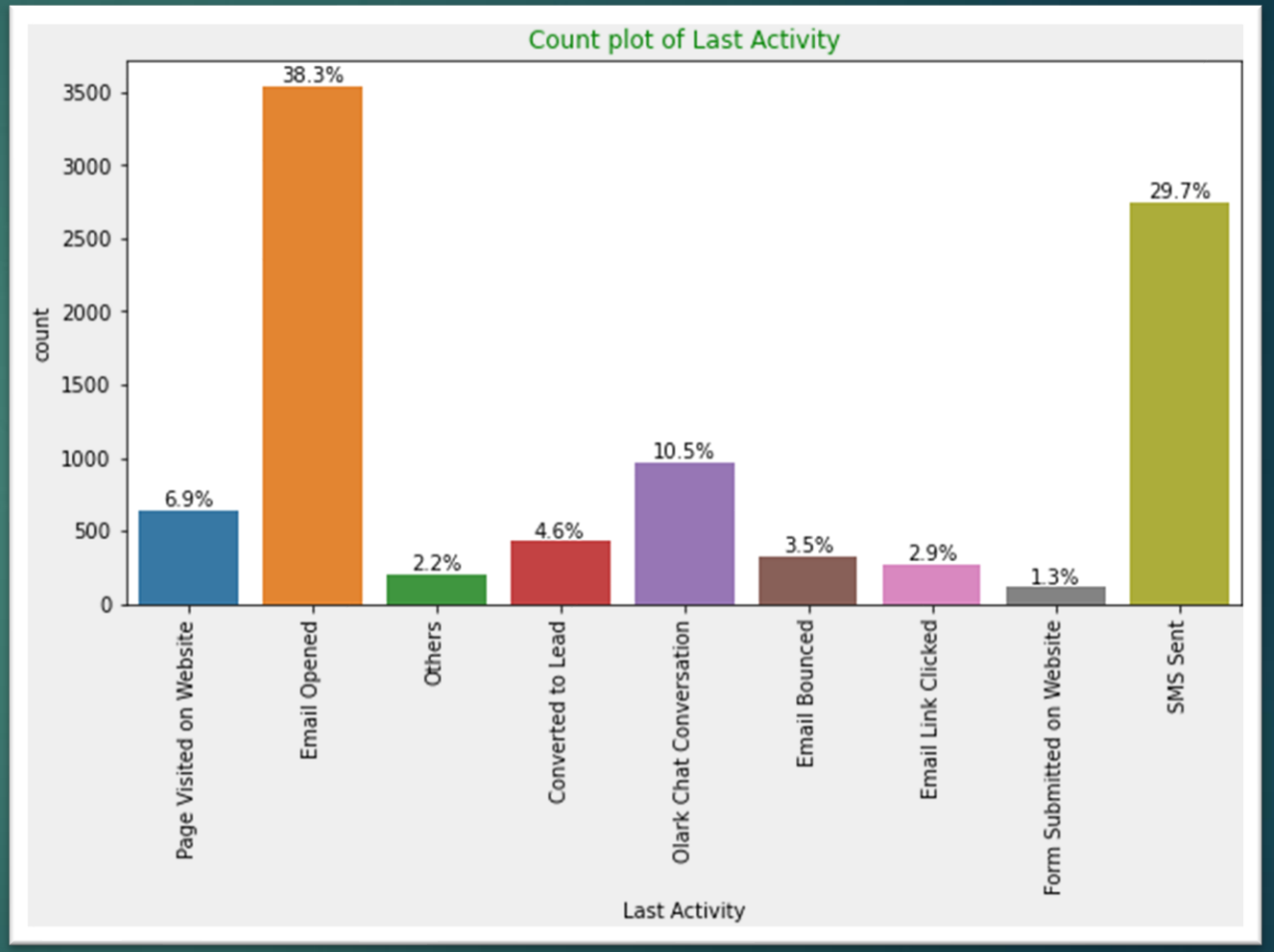
Exploratory Data Analysis (EDA)

- Univariate Analysis
  - The combination of Google and Direct Traffic contributes to 58% of the lead sources.



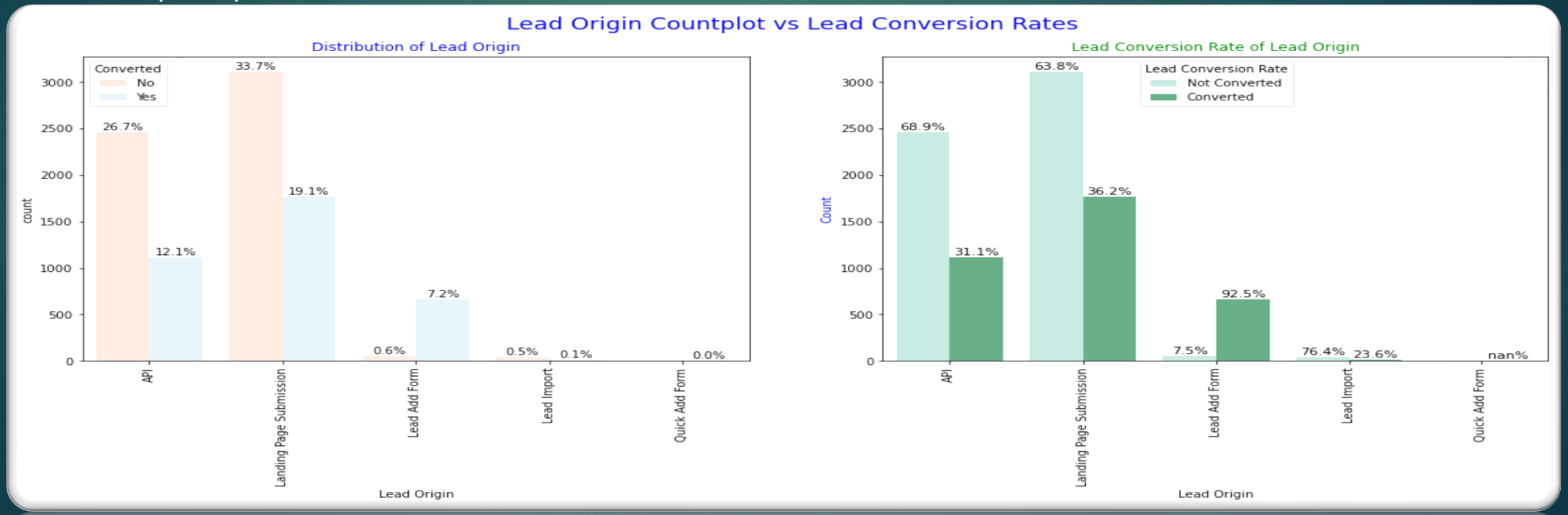
Exploratory Data Analysis (EDA)

- Univariate Analysis
  - The majority of customer interactions, approximately 68%, involve "SMS Sent" and "Email Opened" activities.



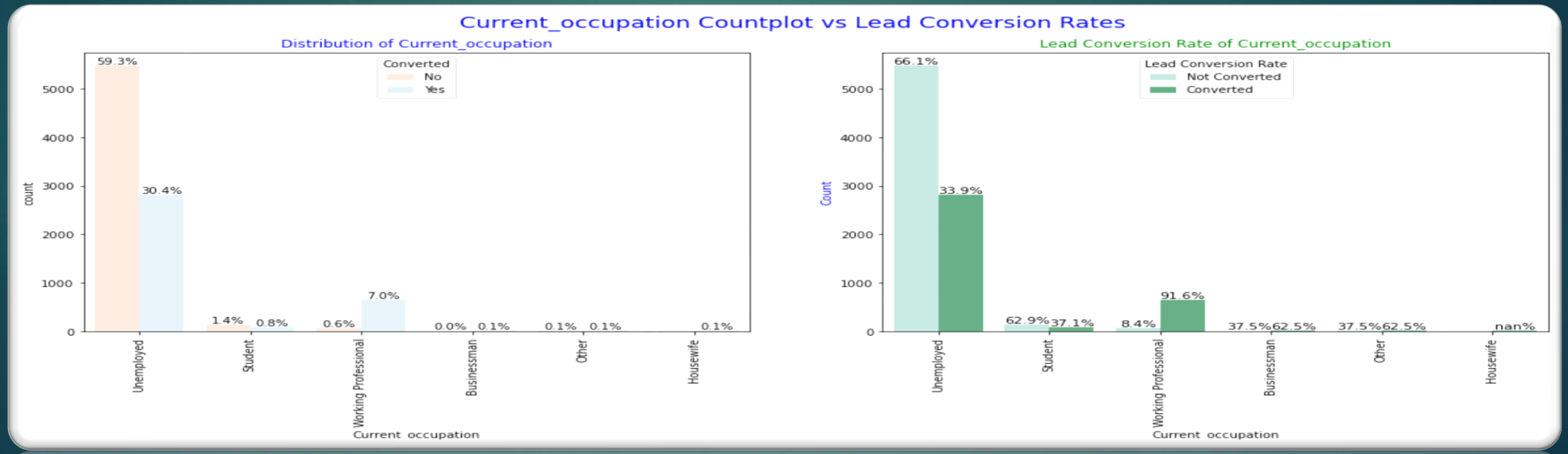
Exploratory Data Analysis (EDA)

- Bivariate Analysis
  - The majority of leads, approximately 52%, were generated through "Landing Page Submission," and these leads had a lead conversion rate (LCR) of 36%. Another source, the "API," accounted for around 39% of customers, with a lead conversion rate (LCR) of 31%.



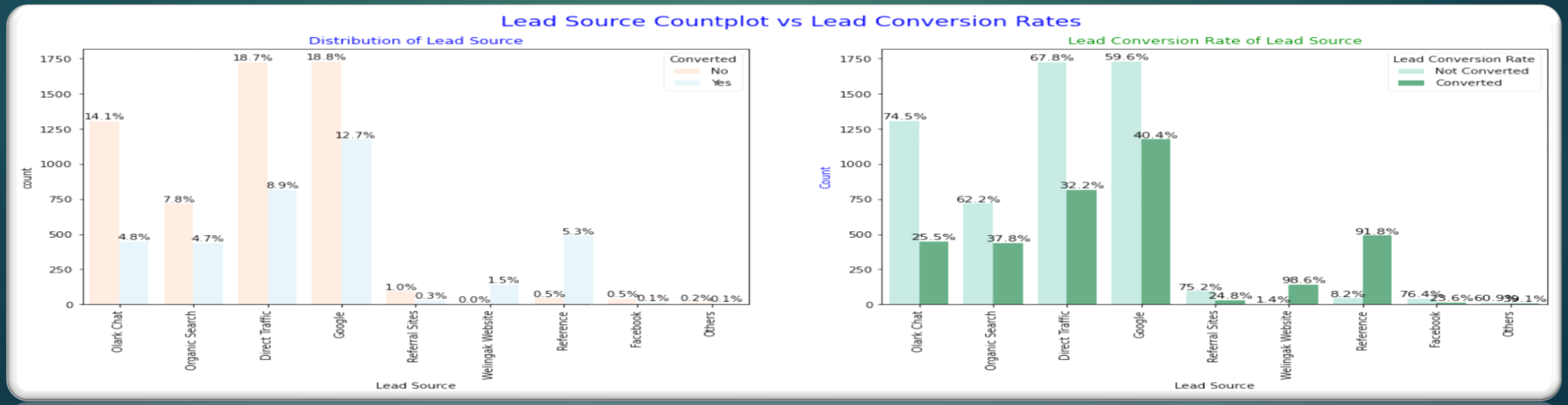
Exploratory Data Analysis (EDA)

- Bivariate Analysis
  - Unemployed individuals comprised approximately 90% of the customer base, and they had a lead conversion rate (LCR) of 34%.
  - Working Professionals constituted only 7.6% of the total customers but had an impressive lead conversion rate (LCR) of almost 92%.



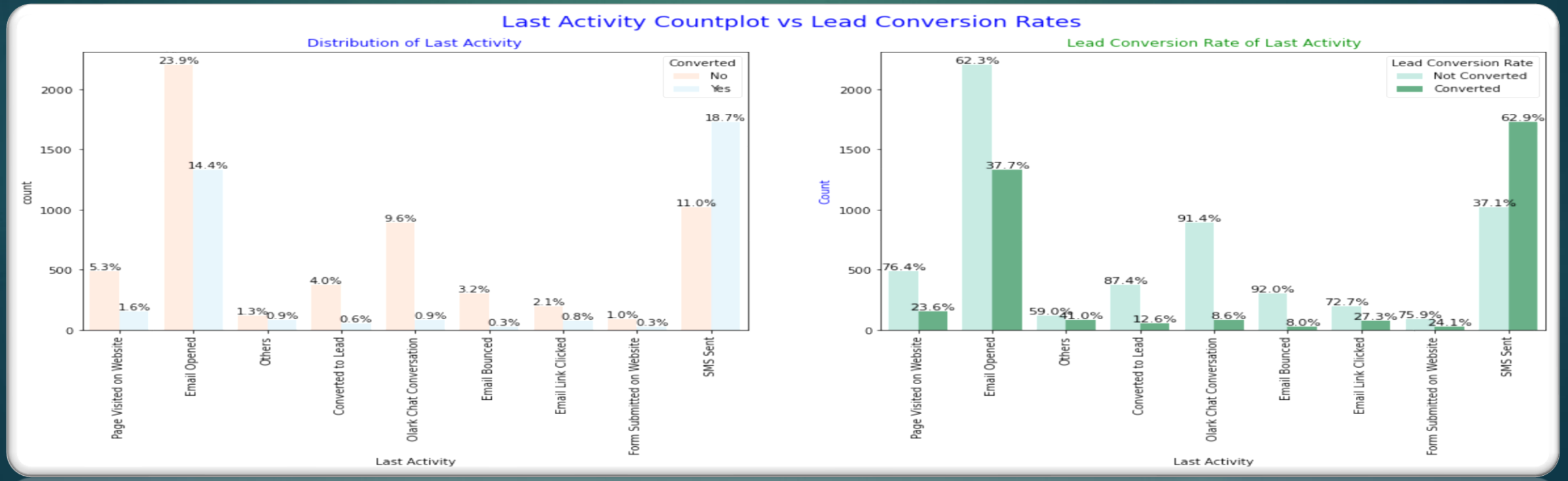
Exploratory Data Analysis (EDA)

- Bivariate Analysis
  - Among the customers, Google has a lead conversion rate (LCR) of 40%, accounting for 31% of the total customers.
  - Direct Traffic, although contributing to 27% of the customers, has a slightly lower LCR of 32% compared to Google.
  - Organic Search, on the other hand, yields a higher LCR of 37.8%, but it only represents 12.5% of the customer base.
  - Reference is an interesting case with a high LCR of 91%, but it only contributes to approximately 6% of the customers.

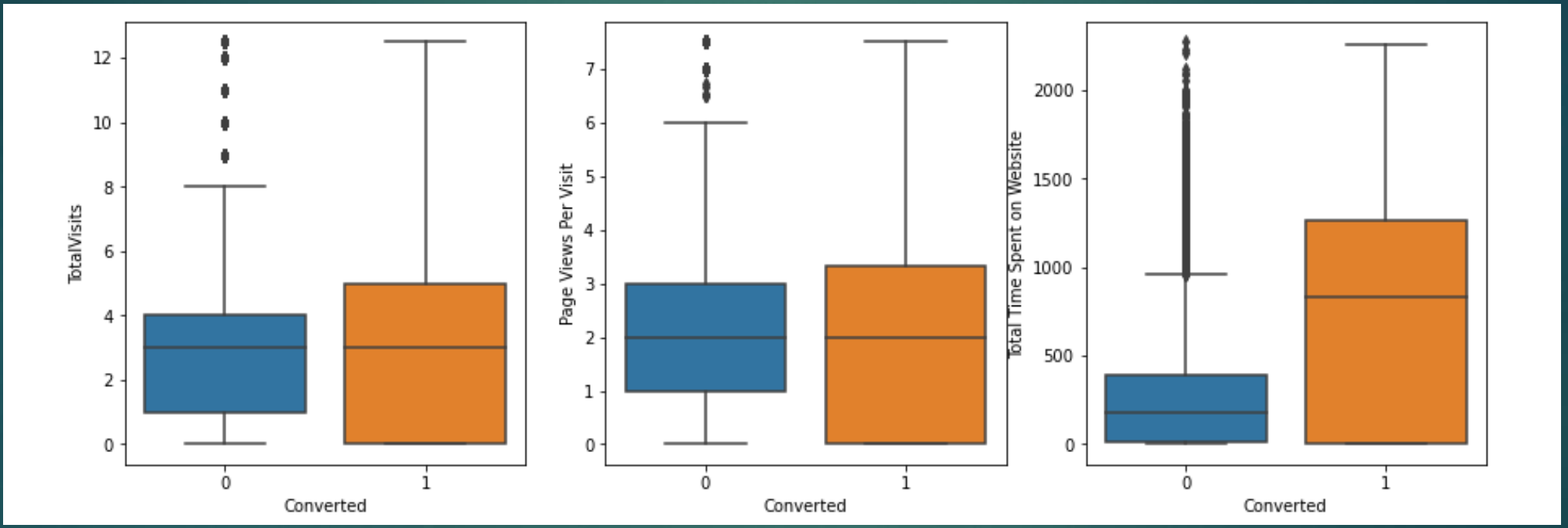


Exploratory Data Analysis (EDA)

- Bivariate Analysis
  - The "SMS Sent" activity stands out with a significant lead conversion rate of 63%, and it is attributed to 30% of the last activities performed by the customers. "Email Opened" represents 38% of the last activities and has a respectable lead conversion rate of 37%.



EDA – Bivariate Analysis for Numerical Variables



Inferences:
Past Leads who spends more time on Website are successfully converted than those who spends less as seen in the boxplot

Data Preparation

- Mapping binary categorical columns to 1 / 0.
- Created dummy features for categorical variables – Lead Origin, Lead Source, Last Activity, Specialization, Current_occupation.
- Splitting Train & Test Sets.
- Feature scaling - Standardization method was used to scale the features.
- Checking the correlations.
- Dropping highly correlated predictor variables.

Checking for highly correlated variables.

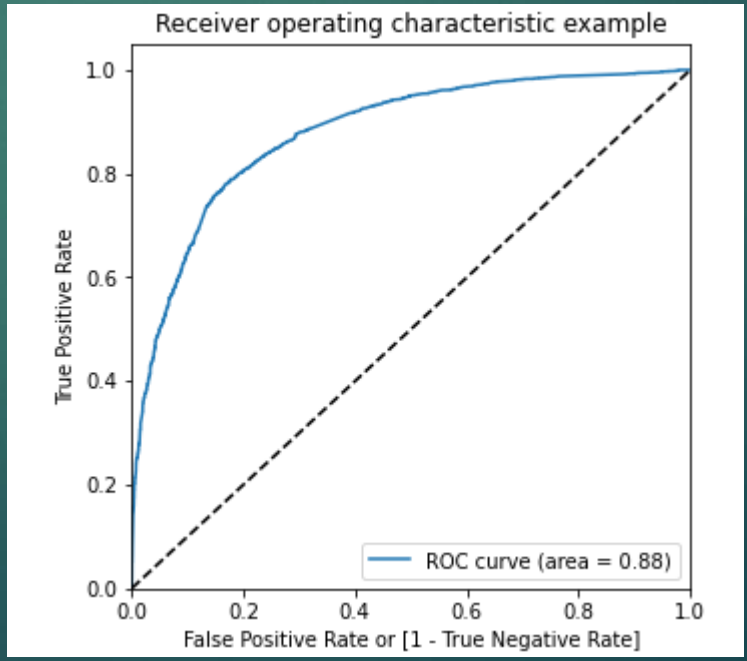
Model Building

- Feature Selection Using Recursive Feature Elimination RFE.
  - select top 15 features.

- Manual Feature reduction using P value and VIF
- Fine Tuning the model – based on P values we dropped few features and rebuild the model further to arrive at final model Model4.

Model Evaluation

- Confusion Matrix
- Accuracy
- Sensitivity and Specificity
- Threshold determination using ROC & Finding Optimal cutoff point
- Precision and Recall
- Plotting the ROC Curve



Final Conclusion

- **Train - Test Comparison**
- **Train Data Set:**
  - **Accuracy:** 80.46%
  - **Sensitivity:** 80.05%
  - **Specificity:** 80.71%
- **Test Data Set:**
  - **Accuracy:** 80.34%
  - **Sensitivity:** 79.82%
  - **Specificity:** 80.68%
- **Observation:**
  - The evaluation metrics show close similarity, indicating consistent performance of the model across different evaluation metrics in both the training and test datasets.
  - Utilizing a cut-off value of 0.345, the model attained a sensitivity of 80.05% in the training set and 79.82% in the test set.
  - Sensitivity, in this context, denotes the accurate identification of leads from the total potential leads that convert.
  - The target sensitivity set by the CEO of X Education was approximately 80%.

Model parameters

- The final Logistic Regression Model has 12 features
- **Top 3 features that contributing positively to predicting hot leads in the model are:**
  - **Lead Source_Welingak Website**
  - **Lead Source_Reference**
  - **Current_occupation_Working Professional**
- **NOTE:** The optimal cutoff probability is determined to be 0.345. Any converted probability exceeding this threshold will be classified as a converted lead, while probabilities lower than 0.345 will be classified as not converted lead

Final Recommendations

- **To increase our Lead Conversion Rates:**
  - Emphasize features that have positive coefficients for targeted marketing strategies.
  - Create tactics to attract high-quality leads from the most successful lead sources.
  - Engage working professionals through personalized messaging.
  - Optimize communication channels based on their impact on lead engagement.
  - Allocate additional budget for advertising and other initiatives on the Welingak website.
  - Offer incentives or discounts for referring leads that convert, and encourage providing more references.
  - Implement aggressive targeting strategies towards working professionals due to their high conversion rates and better financial capacity to afford higher fees.

Areas of improvement:

- Examine the adverse coefficients found in the available specialization courses.
- Assess the submission process on the landing page to identify areas that can be enhanced.