CASE STUDY

On

Lead Conversion Analysis To Maximize the Conversion Rate

Submitted By: Hitesh Mehta, Mohammed Mirza Ahmed, M Meghana

upGrad & IIITB | Data Science Program - March 2023

Lead Conversion Analysis To Maximize the Conversion Rate

Problem Statement:

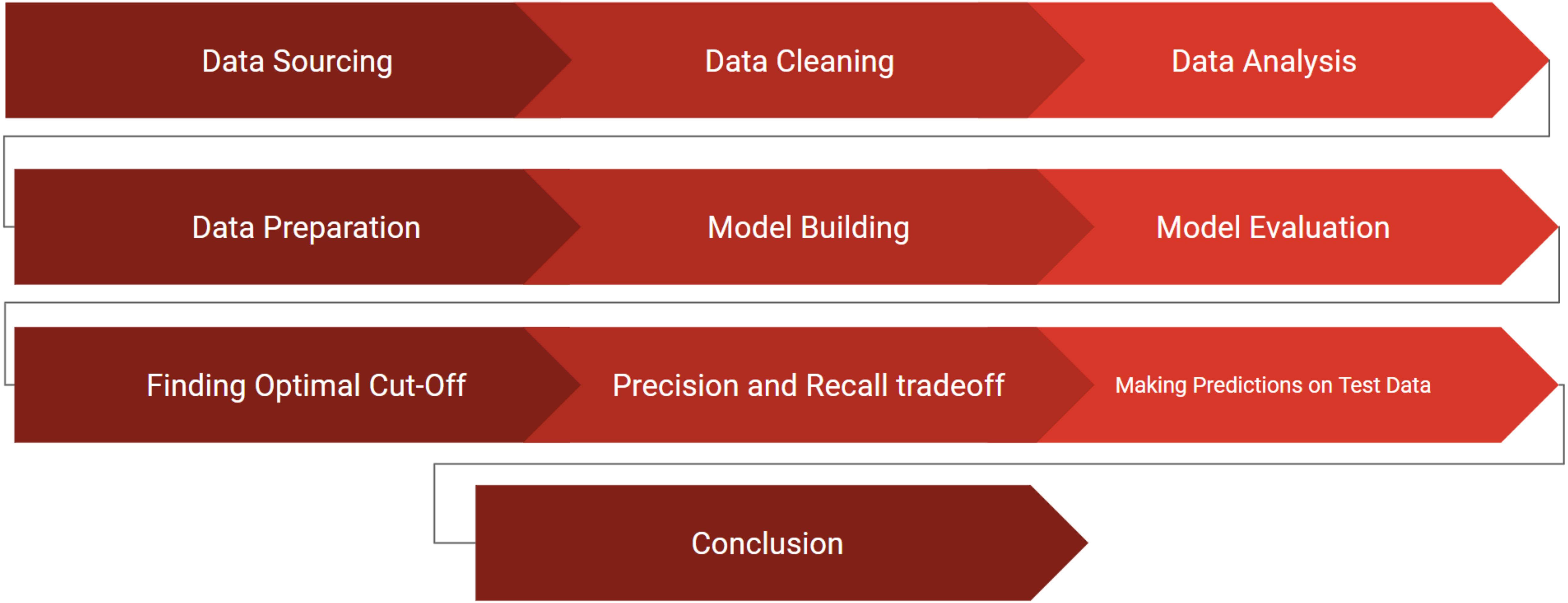
This exercise is about an education company who is having a current conversion rate of around 30% from the entire population of incoming leads, wants to focus on leads having maximum potential to convert and increase the conversion rate to around 80%

This case study aims to build a logistic model to identify patterns/parameters to help target the leads having the maximum potential for conversion and assign a lead score accordingly to each lead. Higher the score, more the potential for conversion.

Following data set have been provided:

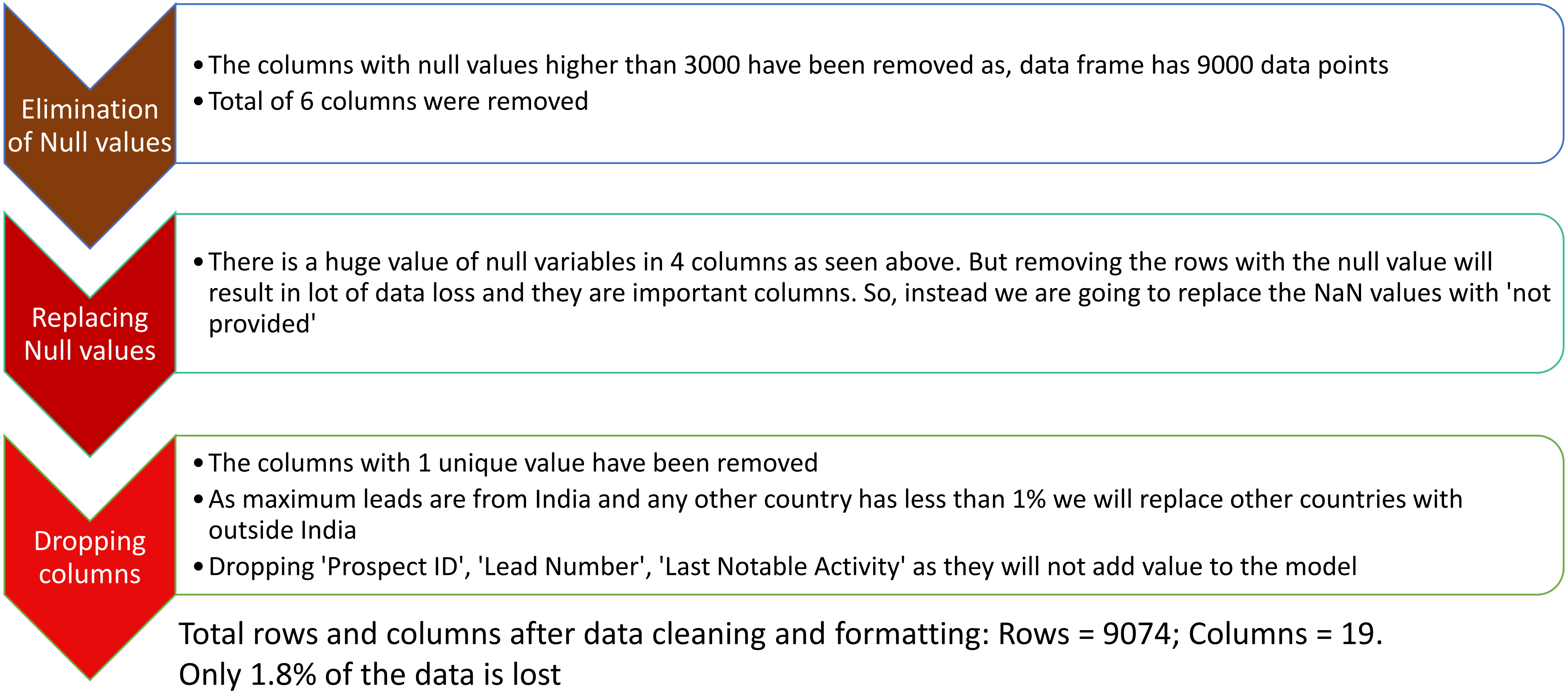
‘Leads.csv’ contains all the information of the leads coming to the company.

Steps for Data Analysis:



Data cleaning and Data Formatting

Total rows and columns before data cleaning and formatting were: Rows = 9240; Columns = 37



Elimination of Null values

- The columns with null values higher than 3000 have been removed as, data frame has 9000 data points
- Total of 6 columns were removed

Replacing Null values

- There is a huge value of null variables in 4 columns as seen above. But removing the rows with the null value will result in lot of data loss and they are important columns. So, instead we are going to replace the NaN values with 'not provided'

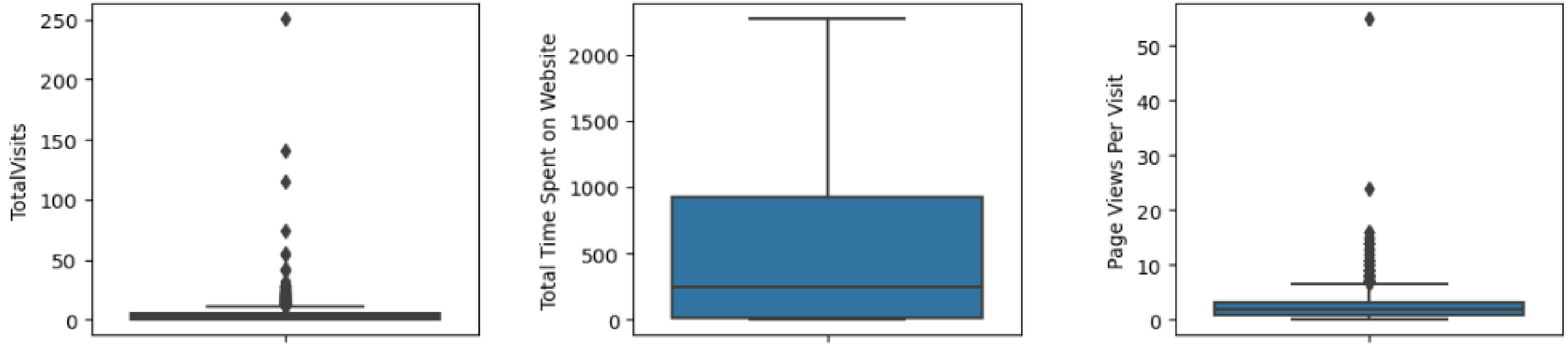
Dropping columns

- The columns with 1 unique value have been removed
- As maximum leads are from India and any other country has less than 1% we will replace other countries with outside India
- Dropping 'Prospect ID', 'Lead Number', 'Last Notable Activity' as they will not add value to the model

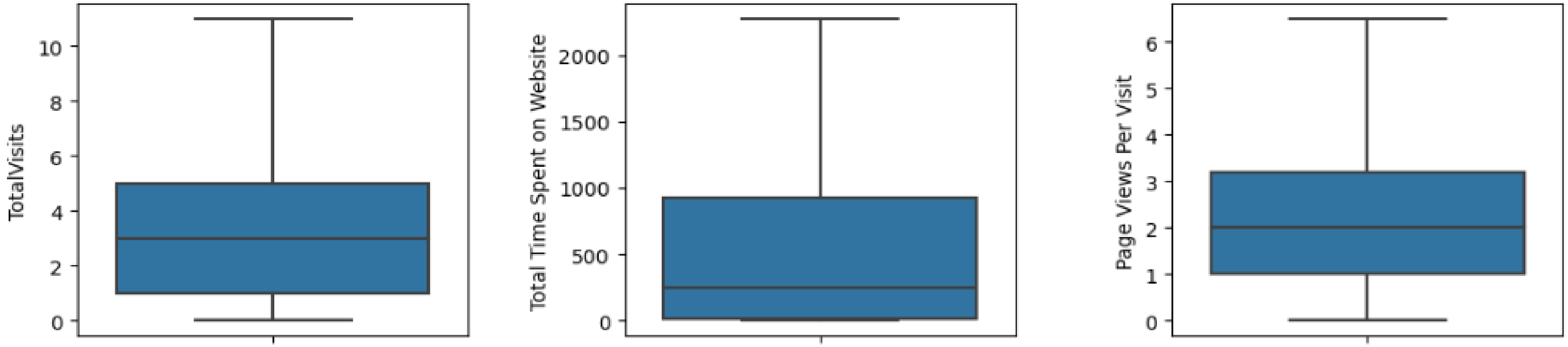
Total rows and columns after data cleaning and formatting: Rows = 9074; Columns = 19.
Only 1.8% of the data is lost

Dealing with Outliers

With outliers



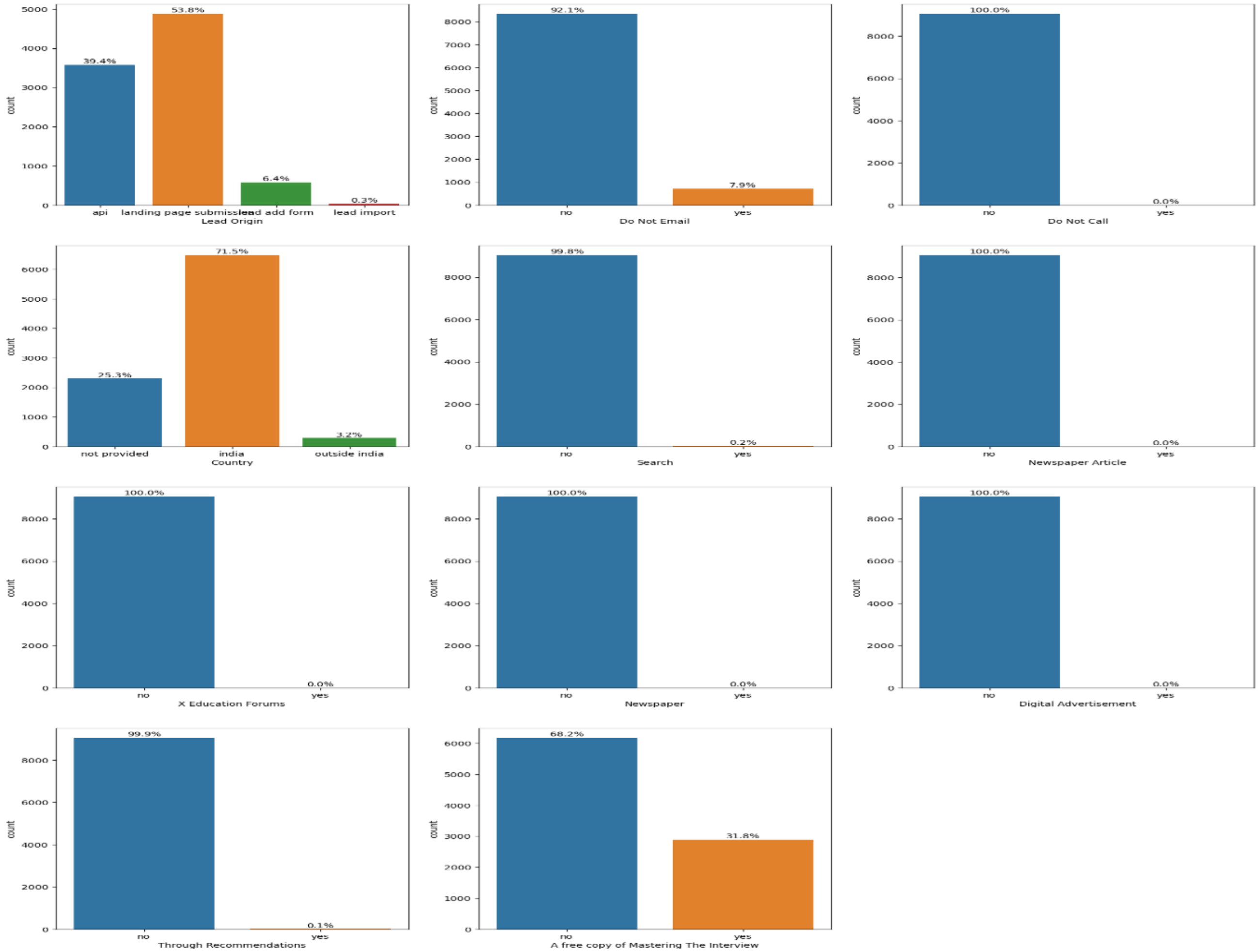
After removal outliers



Outliers were found in following columns:
“TotalVisits”
“Page Views Per Visit”

Univariate Analysis

Analyzing the categorical variables

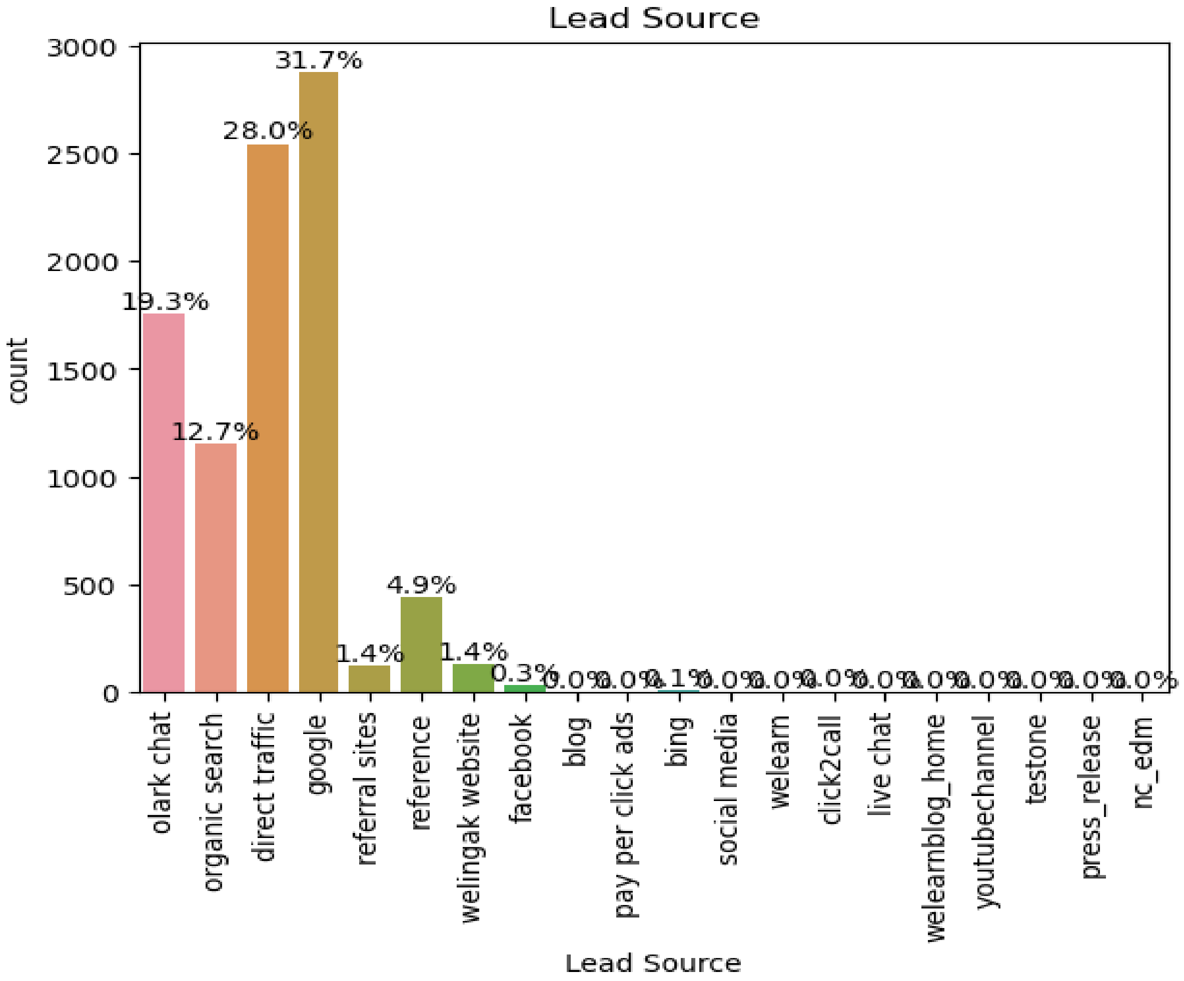


Lead Origin: "Landing Page Submission" identified 53% customers, "API" identified 39%.

Do Not Email: 92% of the people has opted that they don’t want to be emailed about the course.

Univariate Analysis

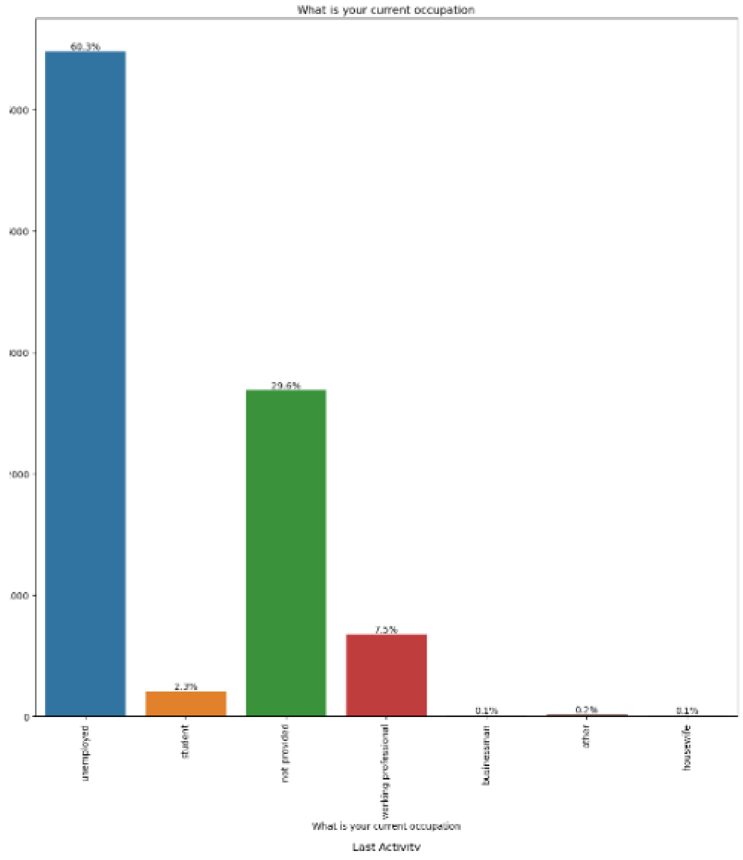
Analyzing the categorical variables

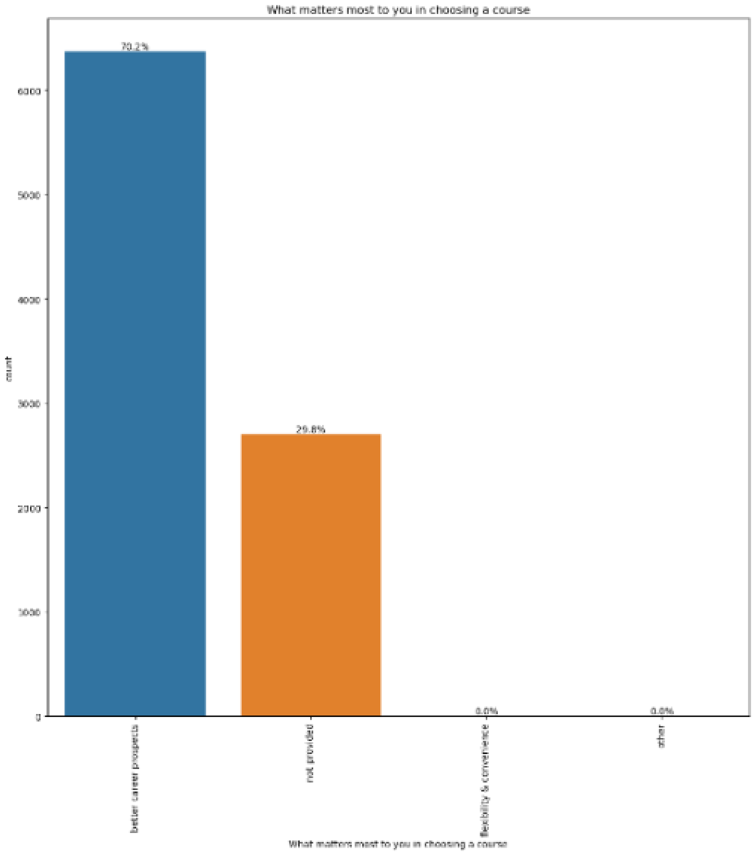


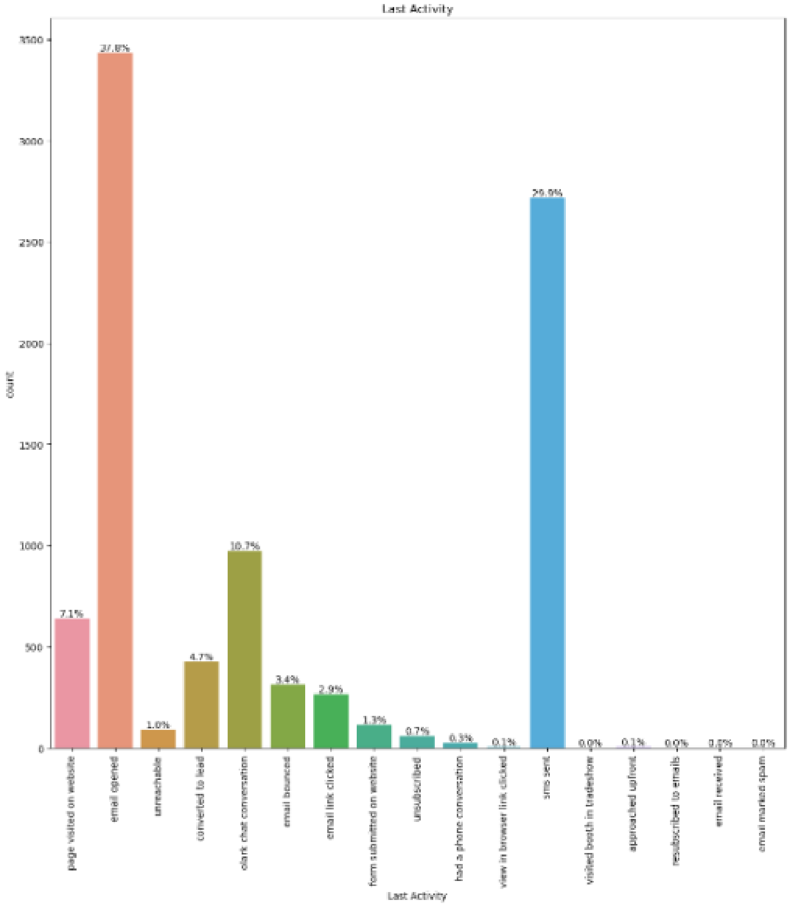
Lead Source: 58% Lead source is from Google & Direct Traffic combined

Univariate Analysis

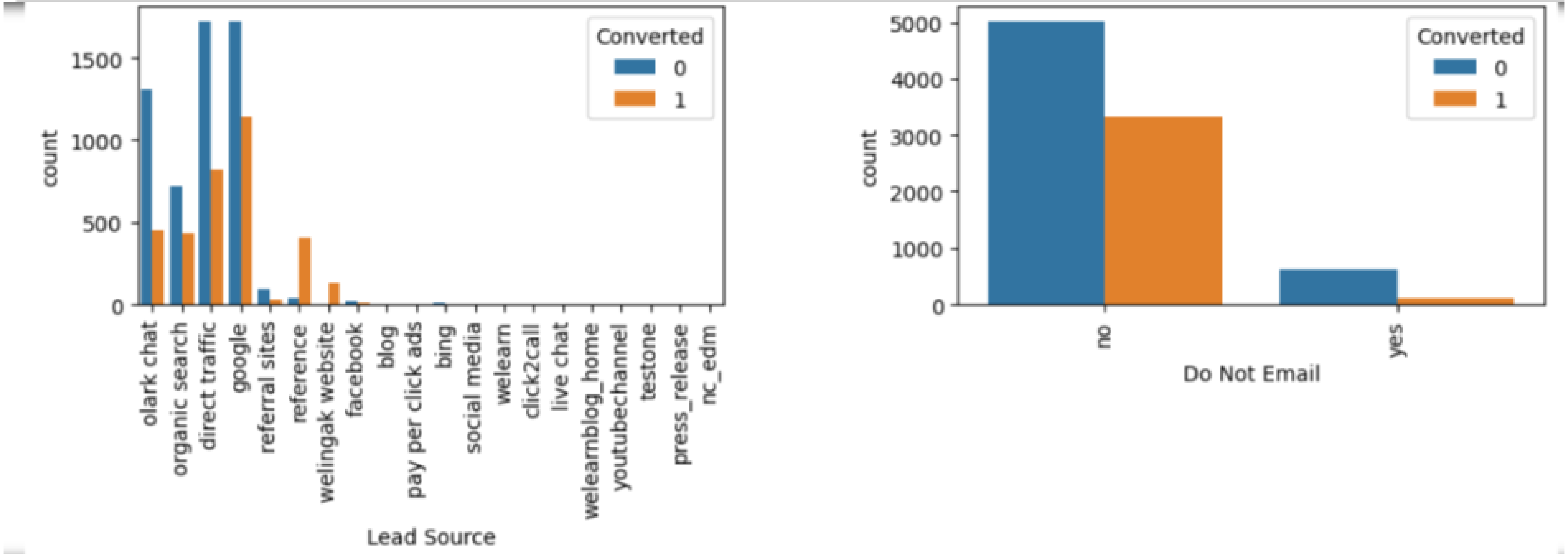
Last Activity: 68% of customers contribution in SMS Sent & Email Opened activities



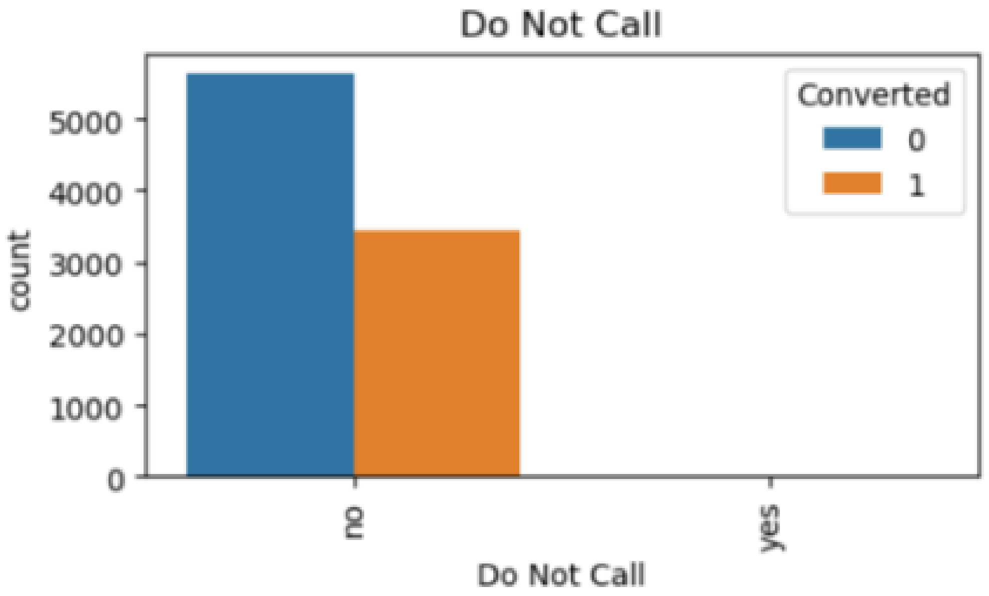


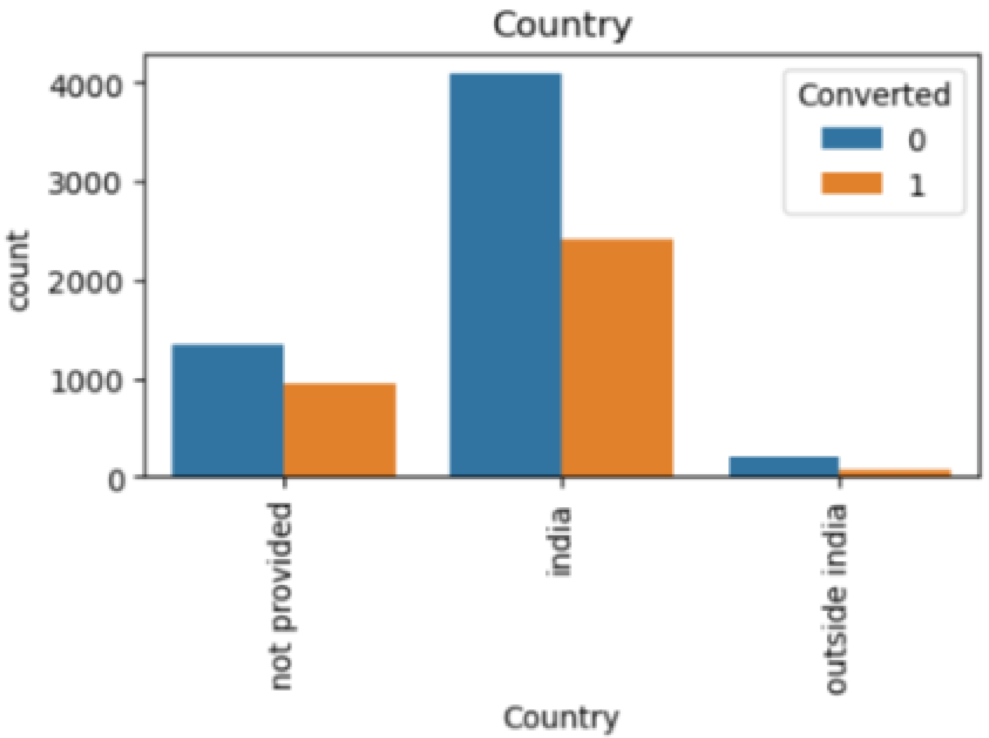


Bivariate Analysis

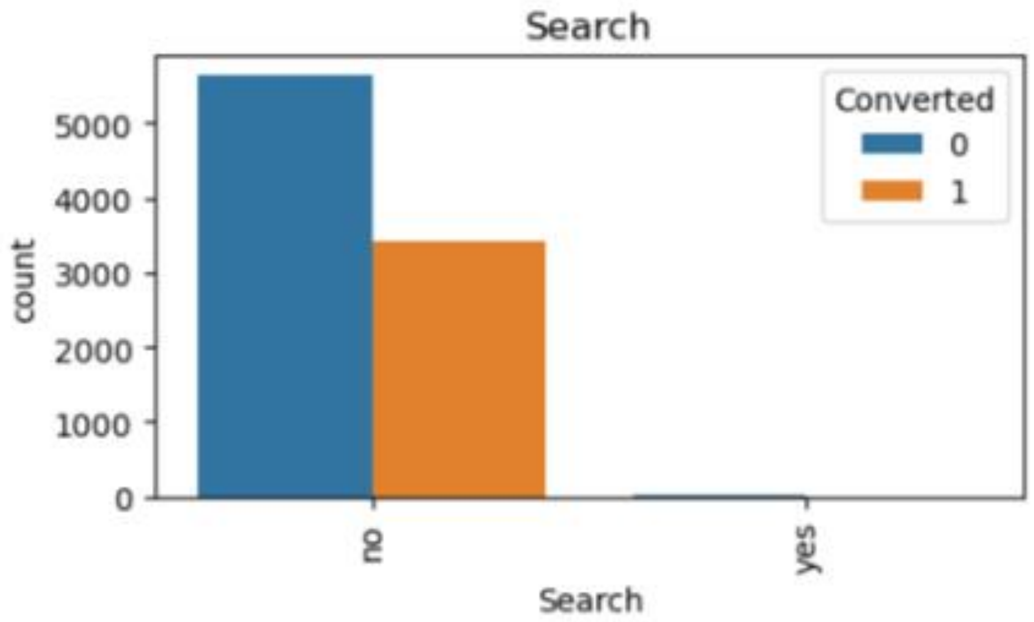


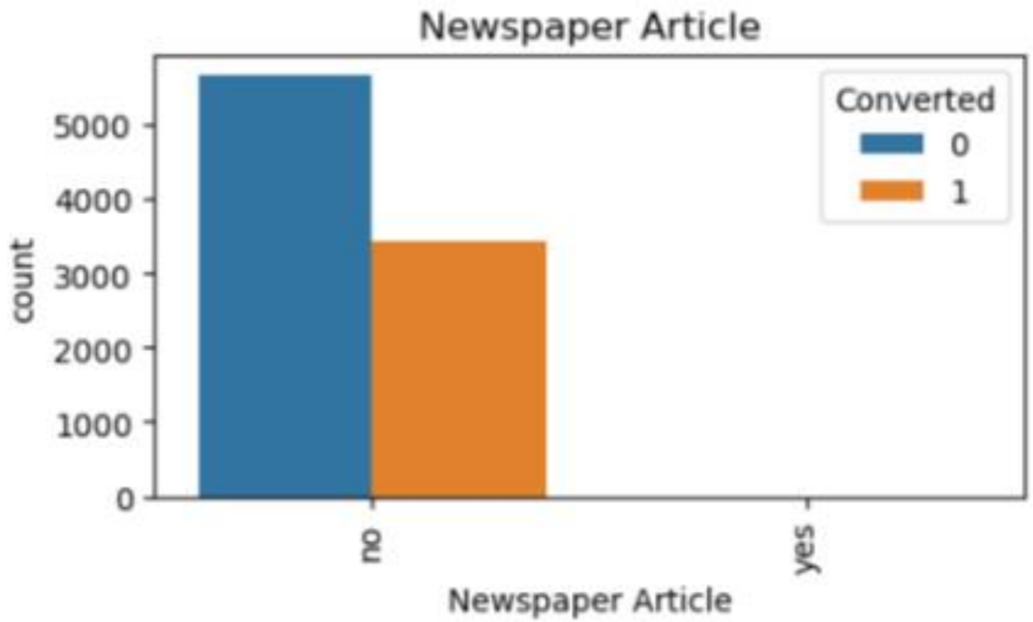
Bivariate Analysis



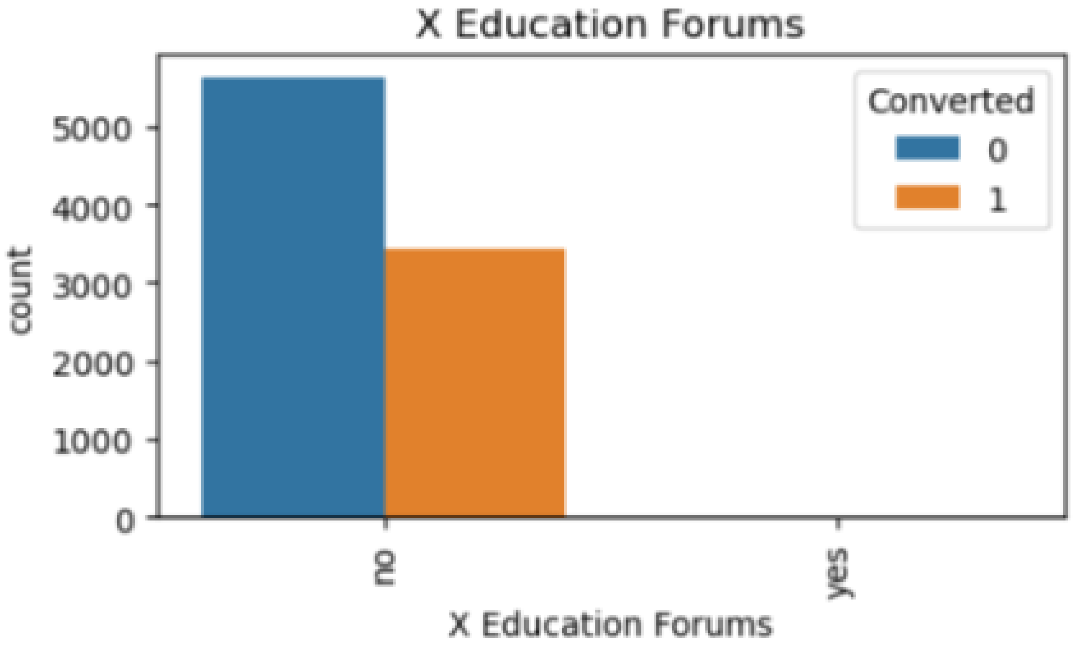


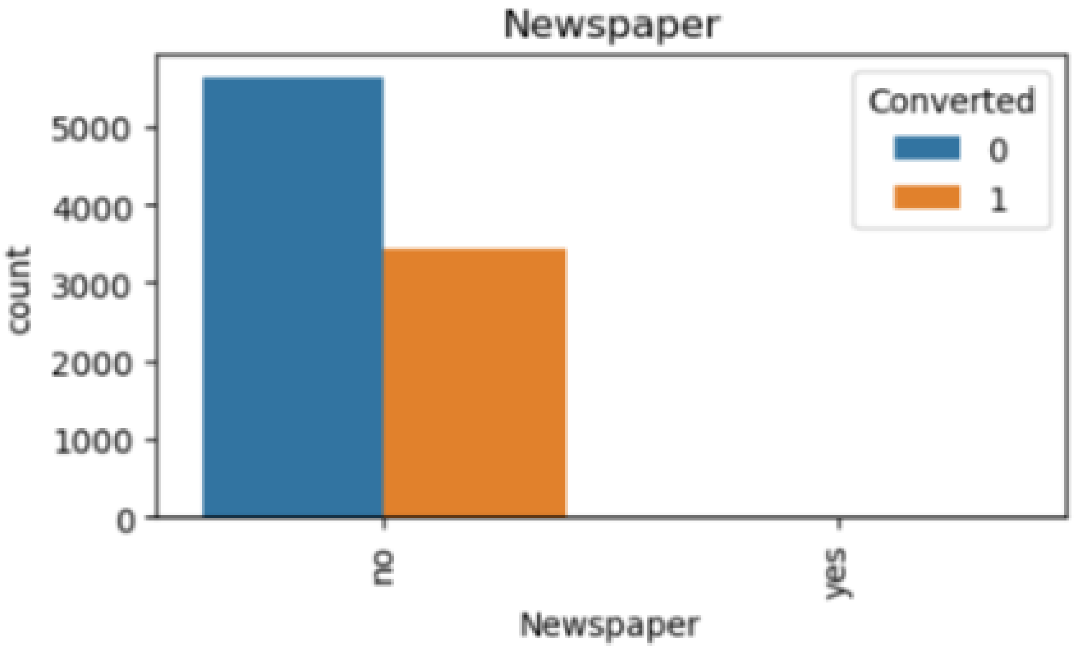
Bivariate Analysis



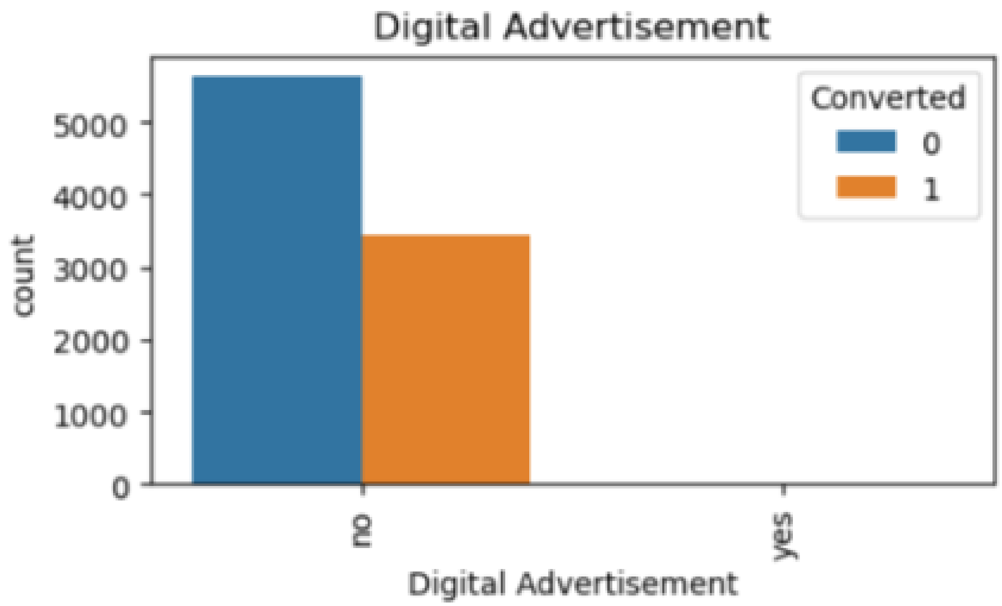


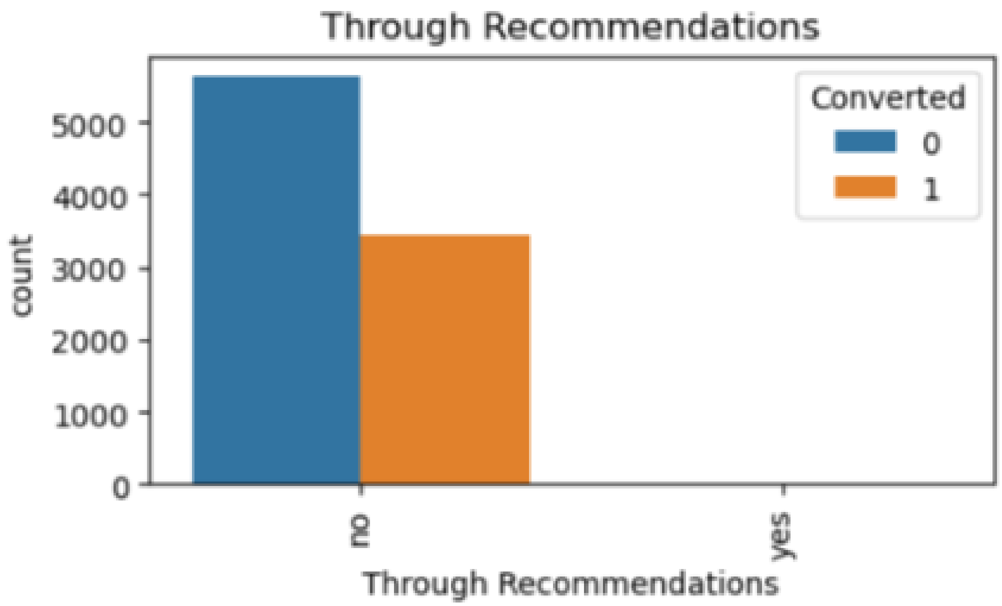
Bivariate Analysis



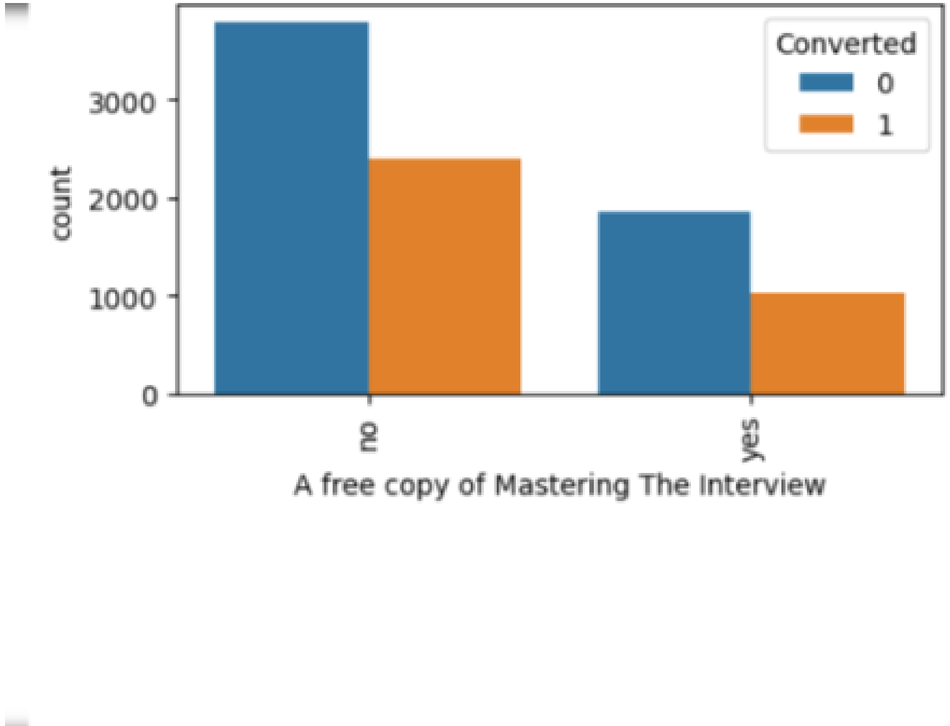


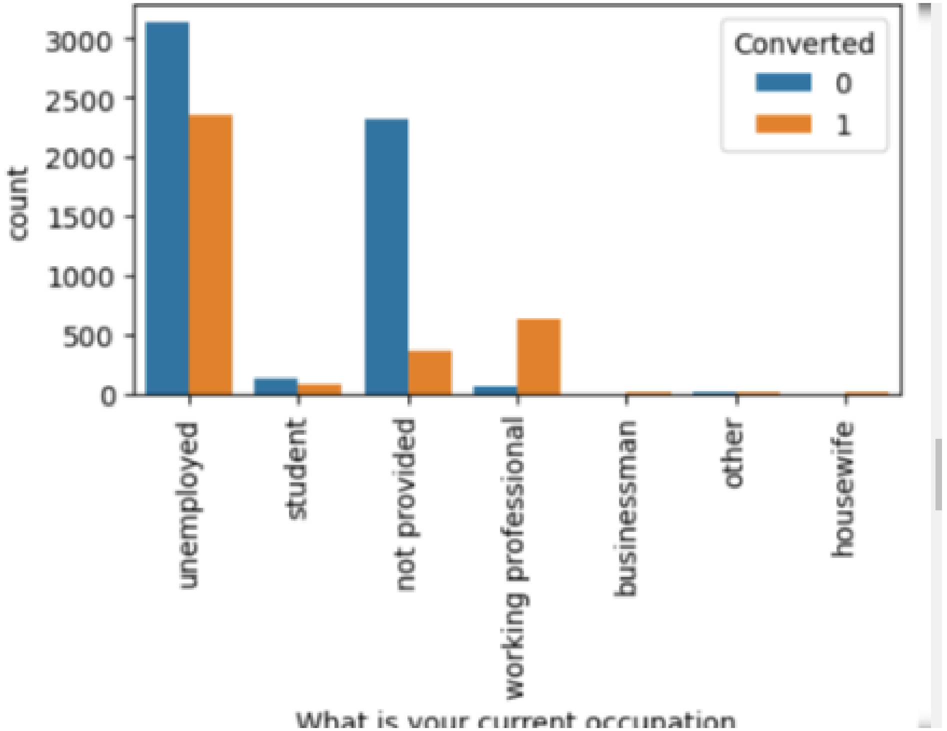
Bivariate Analysis



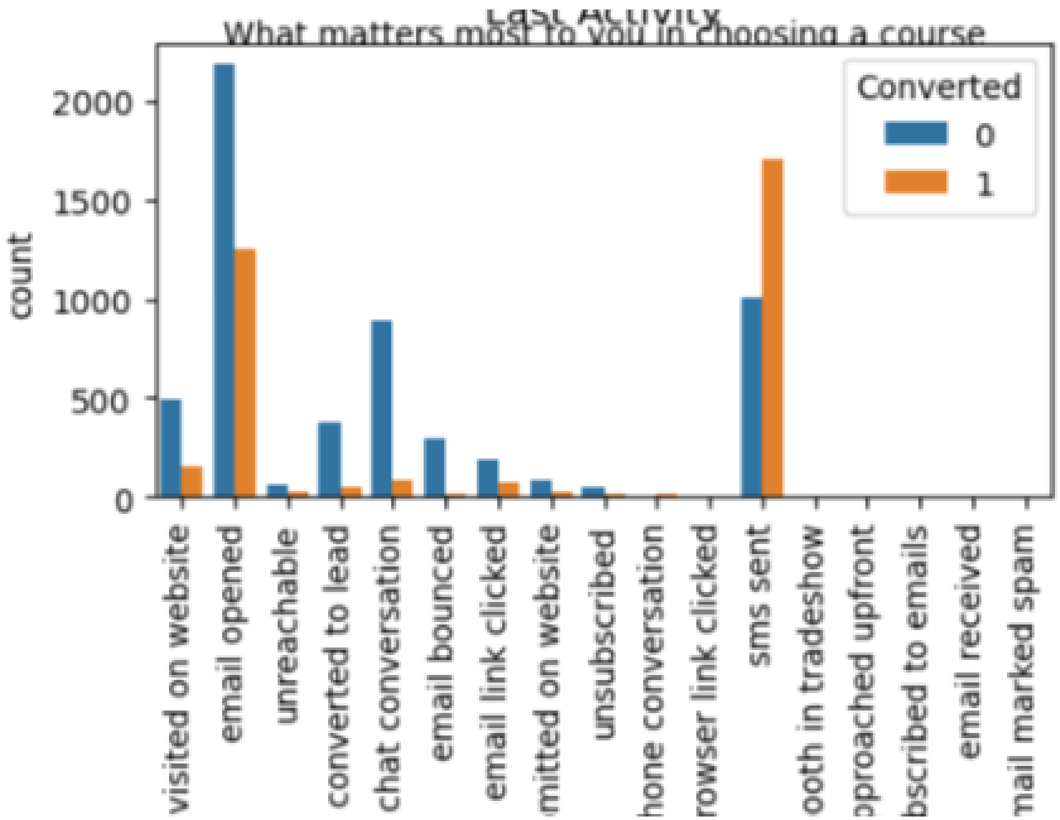


Bivariate Analysis





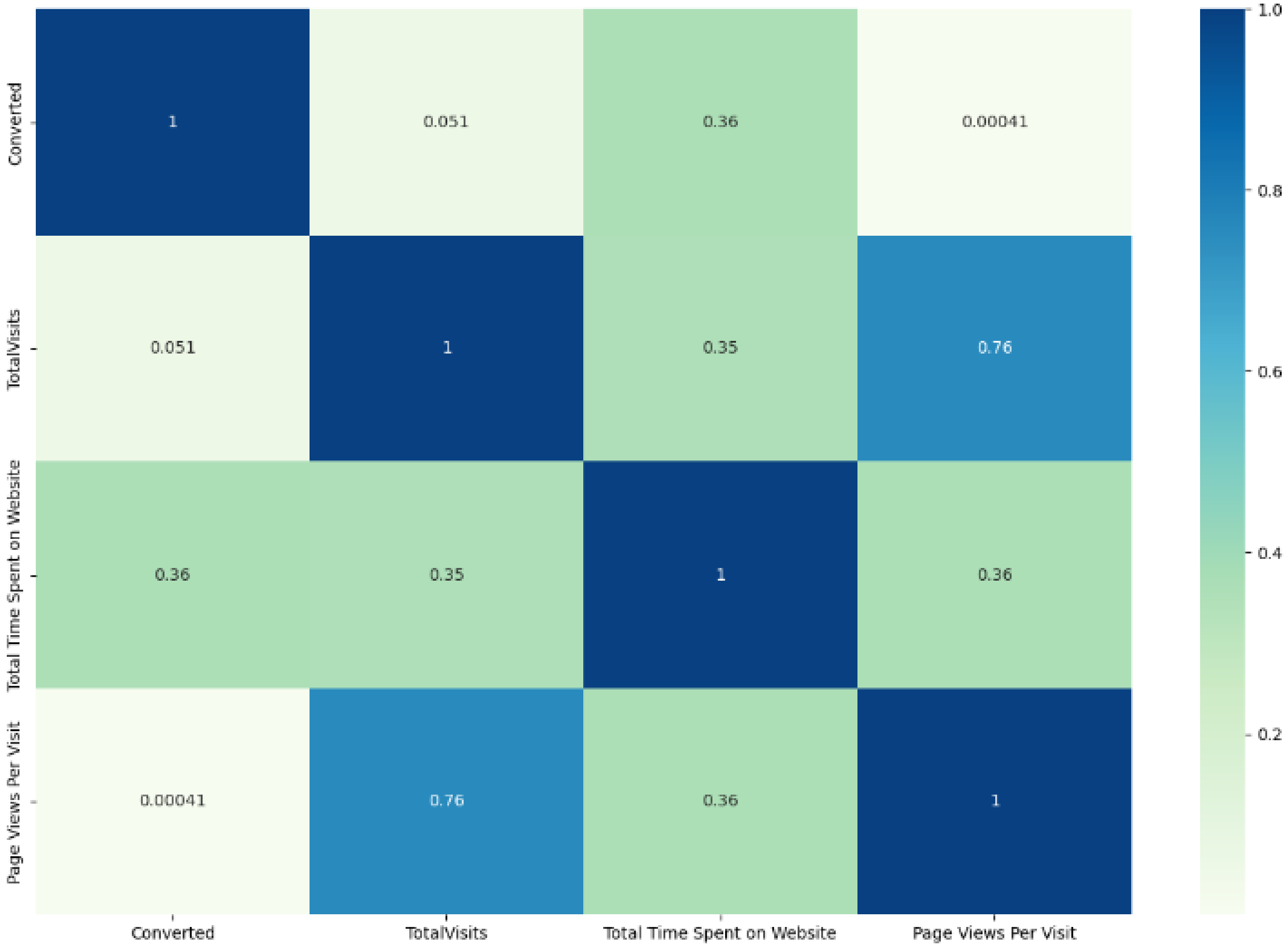
Bivariate Analysis



Following insights can be drawn from the plots:

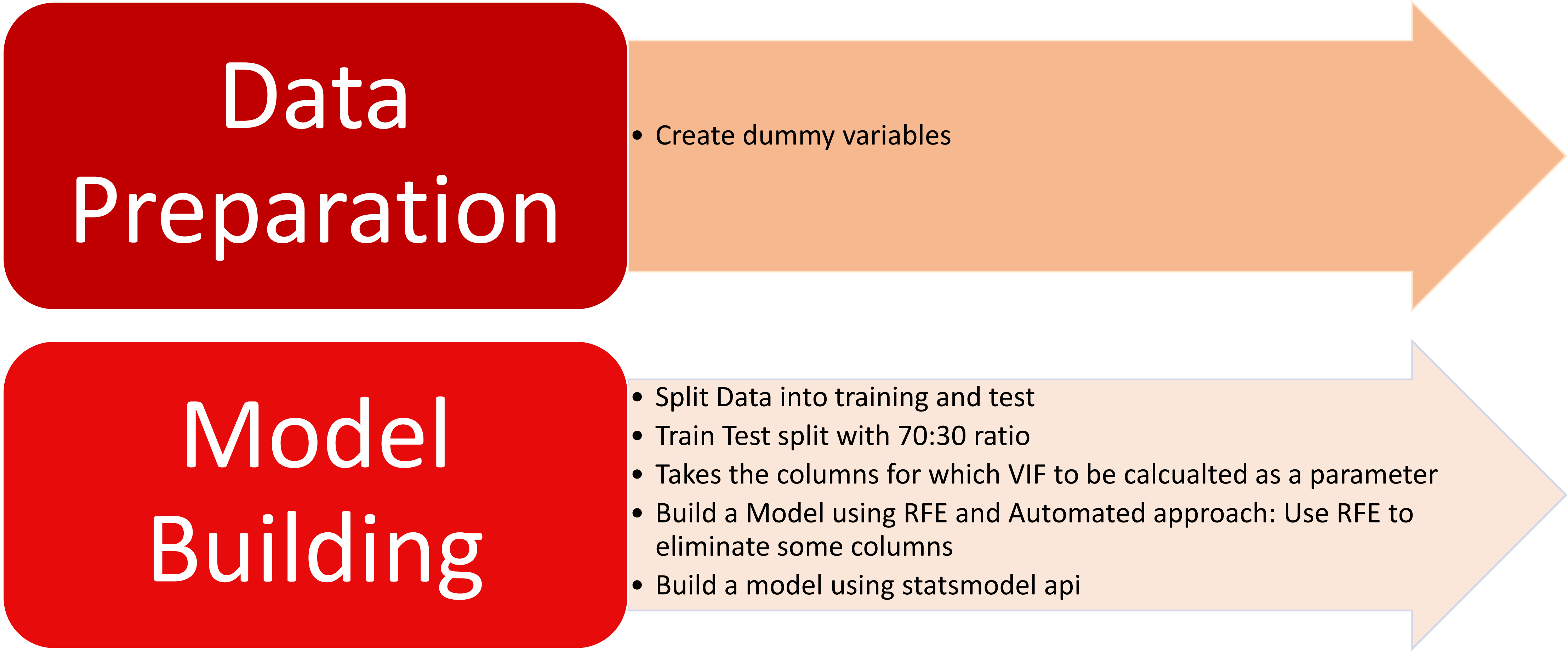
Lead Origin: Around 52% of all leads originated from "Landing Page Submission" with a highest lead conversion rate. The "API" identified approximately 39% of customers with a second highest lead conversion rate

Bivariate Analysis



From the correlation Total Time Spent on Website seems to be highly correlated with conversion positively.

Logistic Regression Model



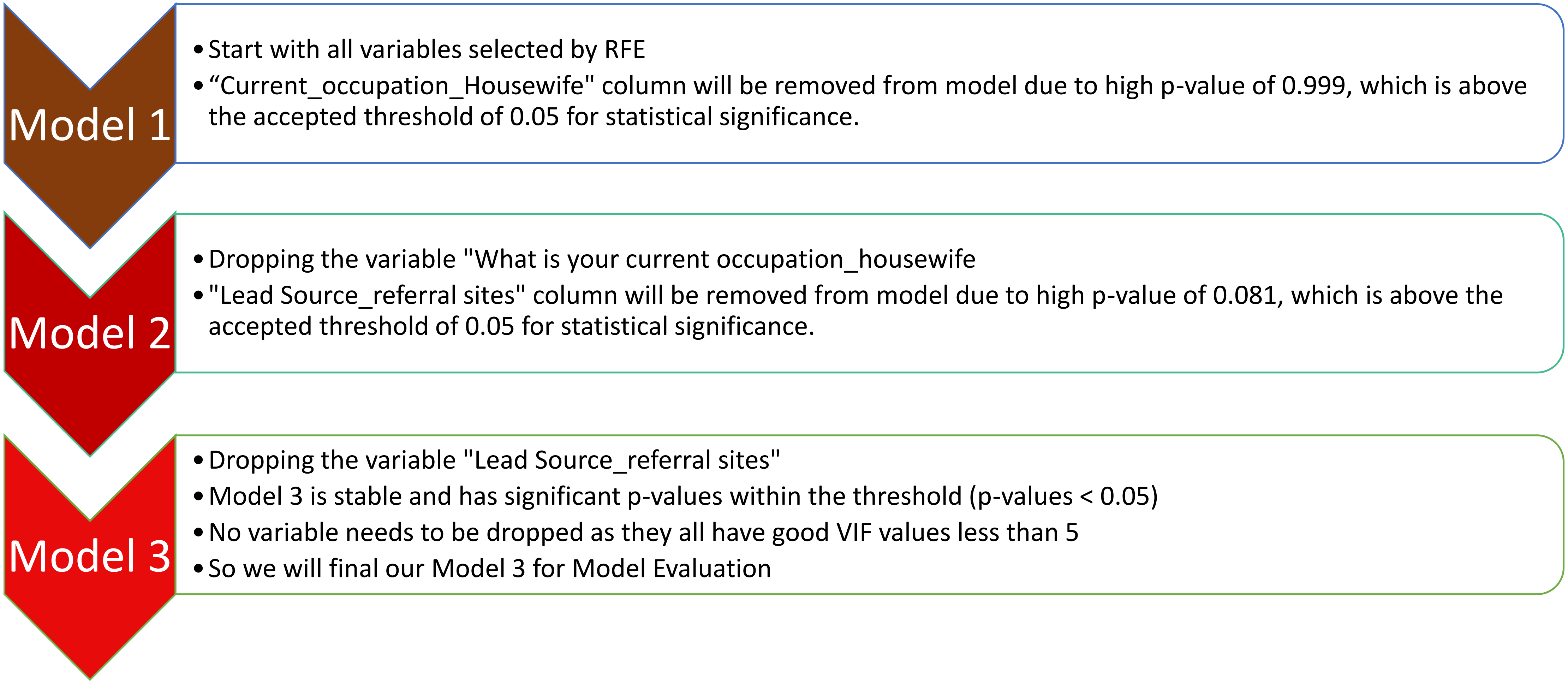
Data Preparation

- Create dummy variables

Model Building

- Split Data into training and test
- Train Test split with 70:30 ratio
- Takes the columns for which VIF to be calcualted as a parameter
- Build a Model using RFE and Automated approach: Use RFE to eliminate some columns
- Build a model using statsmodel api

Logistic Regression Model



Model 1

- Start with all variables selected by RFE
- “Current_occupation_Housewife" column will be removed from model due to high p-value of 0.999, which is above the accepted threshold of 0.05 for statistical significance.

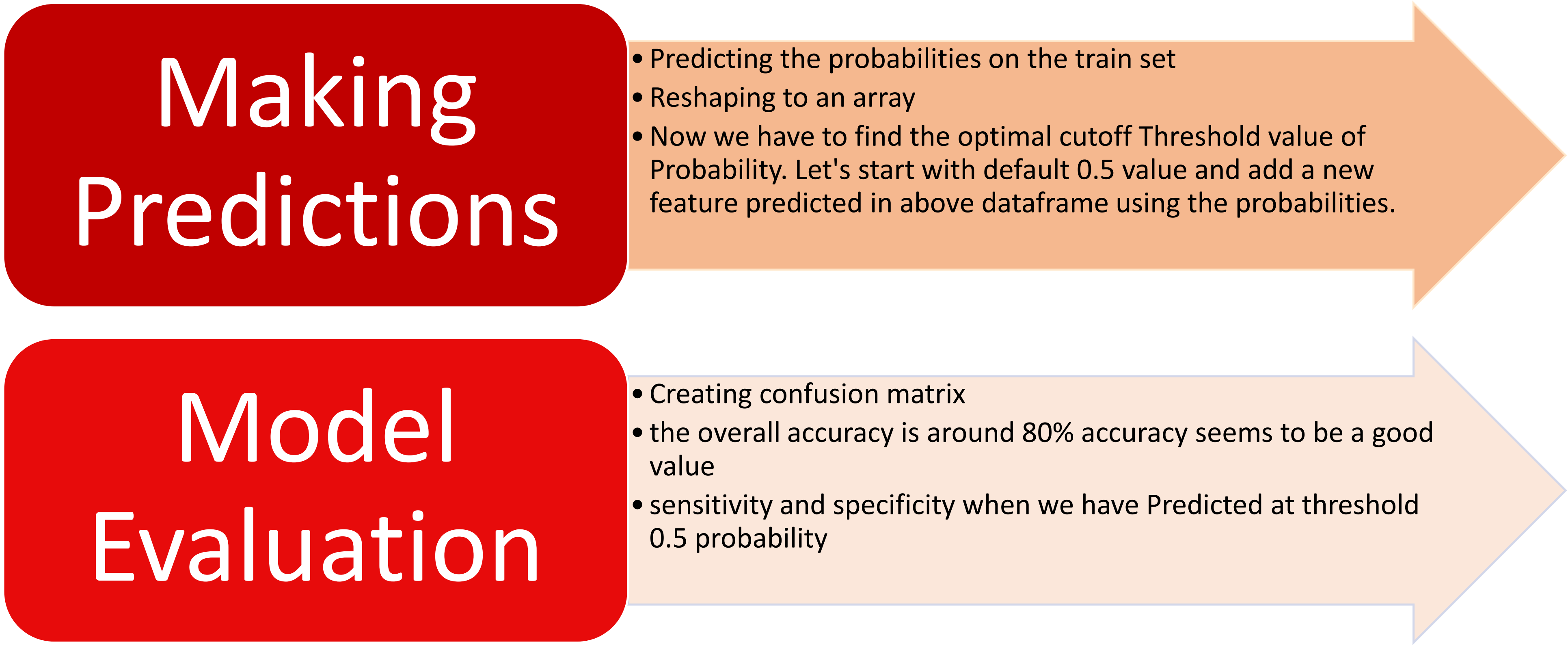
Model 2

- Dropping the variable "What is your current occupation_housewife
- "Lead Source_referral sites" column will be removed from model due to high p-value of 0.081, which is above the accepted threshold of 0.05 for statistical significance.

Model 3

- Dropping the variable "Lead Source_referral sites"
- Model 3 is stable and has significant p-values within the threshold (p-values < 0.05)
- No variable needs to be dropped as they all have good VIF values less than 5
- So we will final our Model 3 for Model Evaluation

Logistic Regression Model



Making Predictions

- Predicting the probabilities on the train set
- Reshaping to an array
- Now we have to find the optimal cutoff Threshold value of Probability. Let's start with default 0.5 value and add a new feature predicted in above dataframe using the probabilities.

Model Evaluation

- Creating confusion matrix
- the overall accuracy is around 80% accuracy seems to be a good value
- sensitivity and specificity when we have Predicted at threshold 0.5 probability

<u>Results</u>:

1. Sensitivity: 0.6684
2. Specificity: 0.8840

<u>Results</u>:

3. false positive rate - predicting conversion when customer does not have converted : 0.1159
4. positive predictive value: 0.7749
5. negative predictive value: 0.8171

Conclusion:

For Test set Accuracy : 80.34% Sensitivity : 79.78% ≈ 80% Specificity : 80.68%.
These metrics are very close to train set, so out final model is performing with good consistency on both Train & Test set

Detailed analysis in combination with other variables shows that the variables that mattered the most in potential conversions are (in the descending order):

- Total Time Spent on the website.
- Current occupation - Working Professional.
- Lead Origin - Lead Add Form.
- Lead Source - Welingak Website.
- Current occupation - Other.
- Last Activity - Had a Phone Conversation.

Thus, it is advisable to prioritize leads based on the above characteristics, as these positively contribute to the conversion rate.

Conclusion:

On the other hand, a key highlight after performing logistic regression on data was that some features were identified that had a negative correlation with the lead conversion. These variables are:

- Lead Source - referral sites
- Lead Source - direct traffic
- Last Activity - olark chat conversation
- Do Not Email - yes

Thus, it is advisable to not approach those prospects with the above features