X Education - Lead Scoring Case Study

IDENTIFICATION OF HOT LEADS TO FOCUS MORE ON THEM AND THUS ENHANCING THE CONVERSION RATIO FOR X EDUCATION

Group Members:
Dibyajit Dhara
Ritesh Patil

Problem Statement

X Education Company's Problem

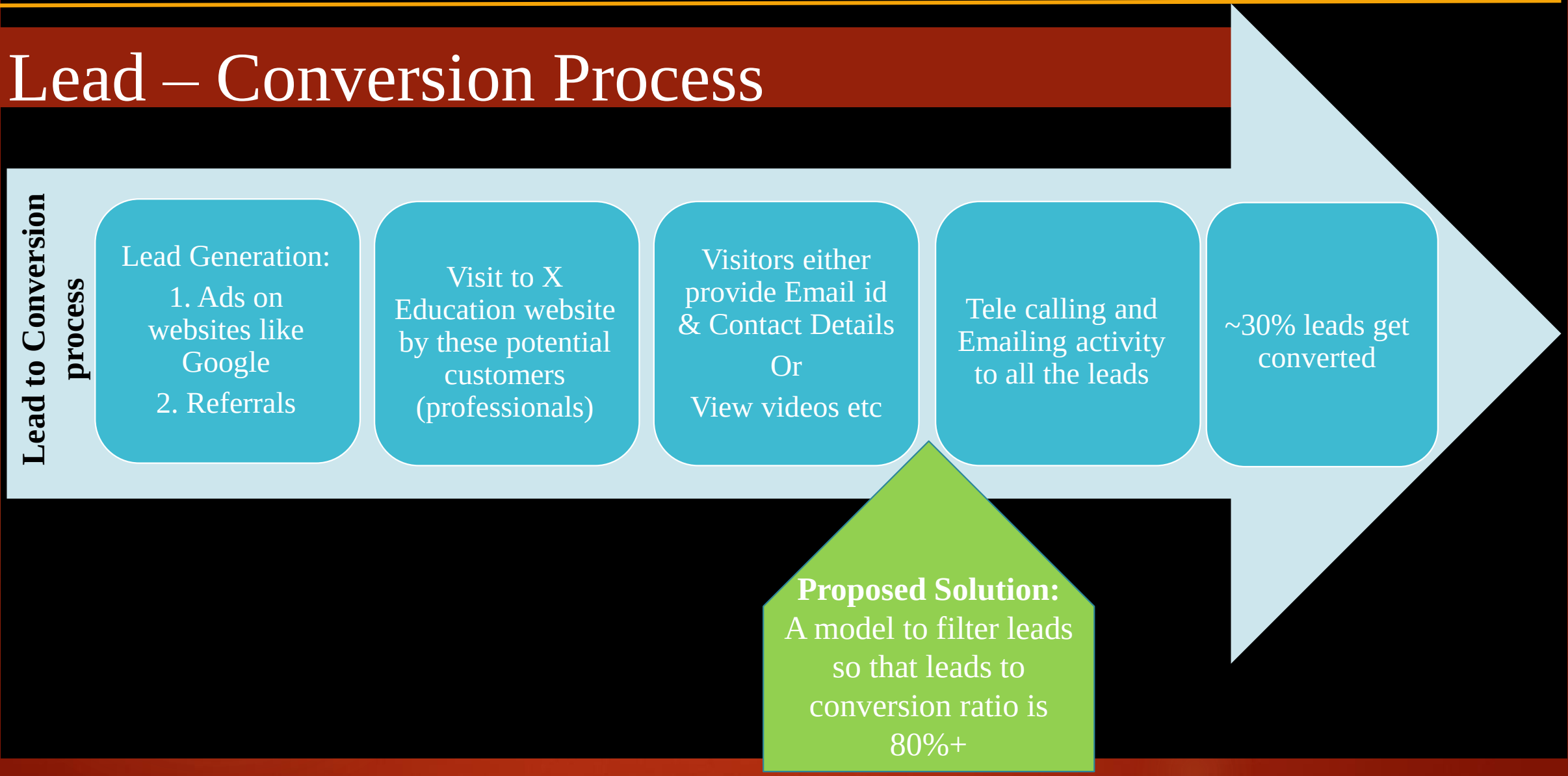
- X Education gets a lot of leads but its lead conversion rate is very poor.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'.
- If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Problem Statement

X Education Company's Problem

- We will help them to select the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- We are required to build a model wherein we need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance
- The CEO, in particular, has given a ballpark of the target lead conversion rate to be 80%.

Lead – Conversion Process



Proposed Solution

Selection of Hot Leads

Leads Clustering

We cluster the leads into certain categories based on their tendency or probability to convert, thus, getting a smaller section of hot leads to focus more on.

Communicating with Hot Leads

Focus Communication

Since we would have a smaller set of leads to have communication with, we might make more impact with effective communication.

Conversion of Hot Leads

Increase conversion

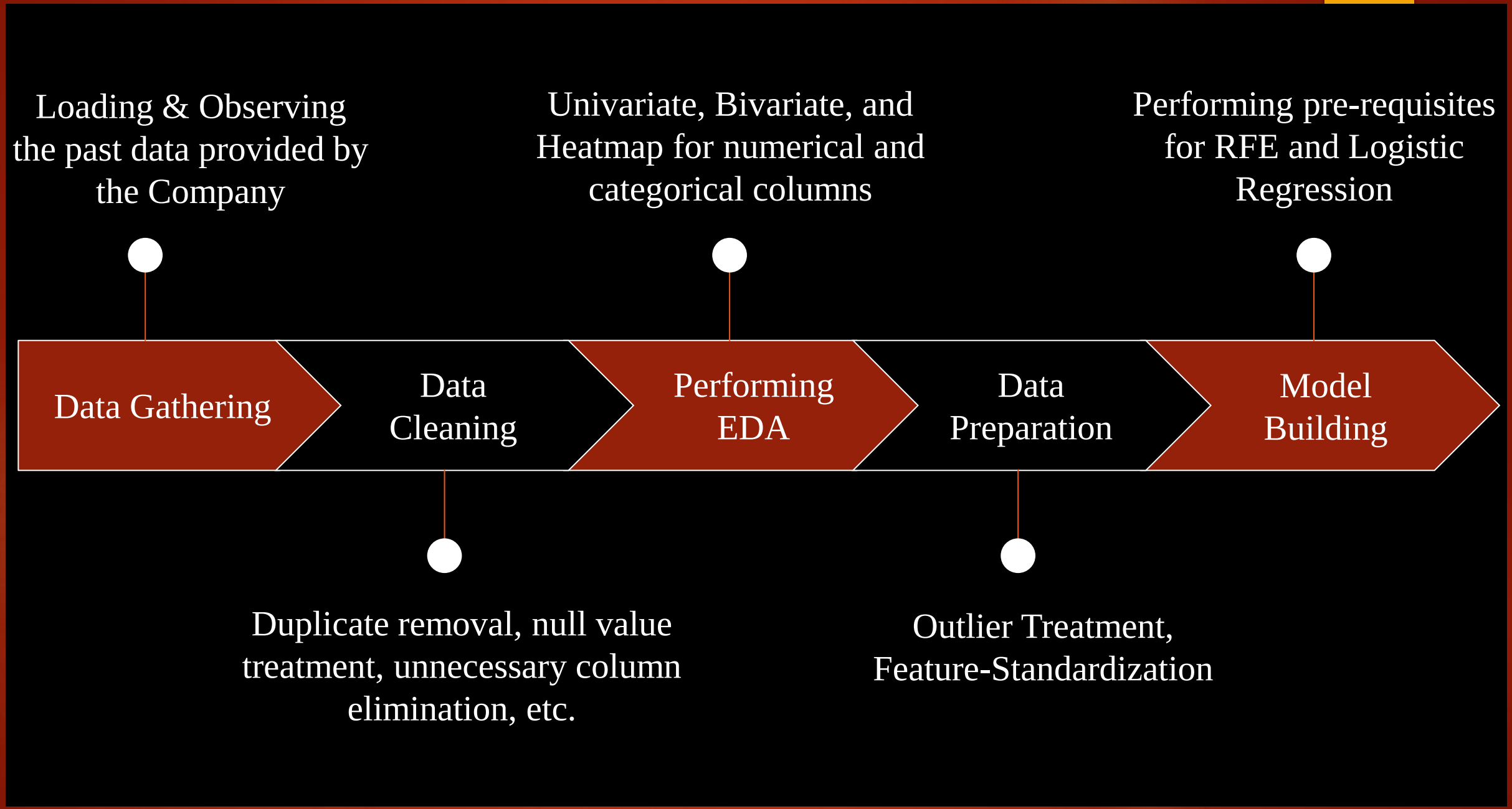
Since we focussed on hot leads, which were more probable to convert, we would have a better conversion rate, and hence we can achieve the 80% target.

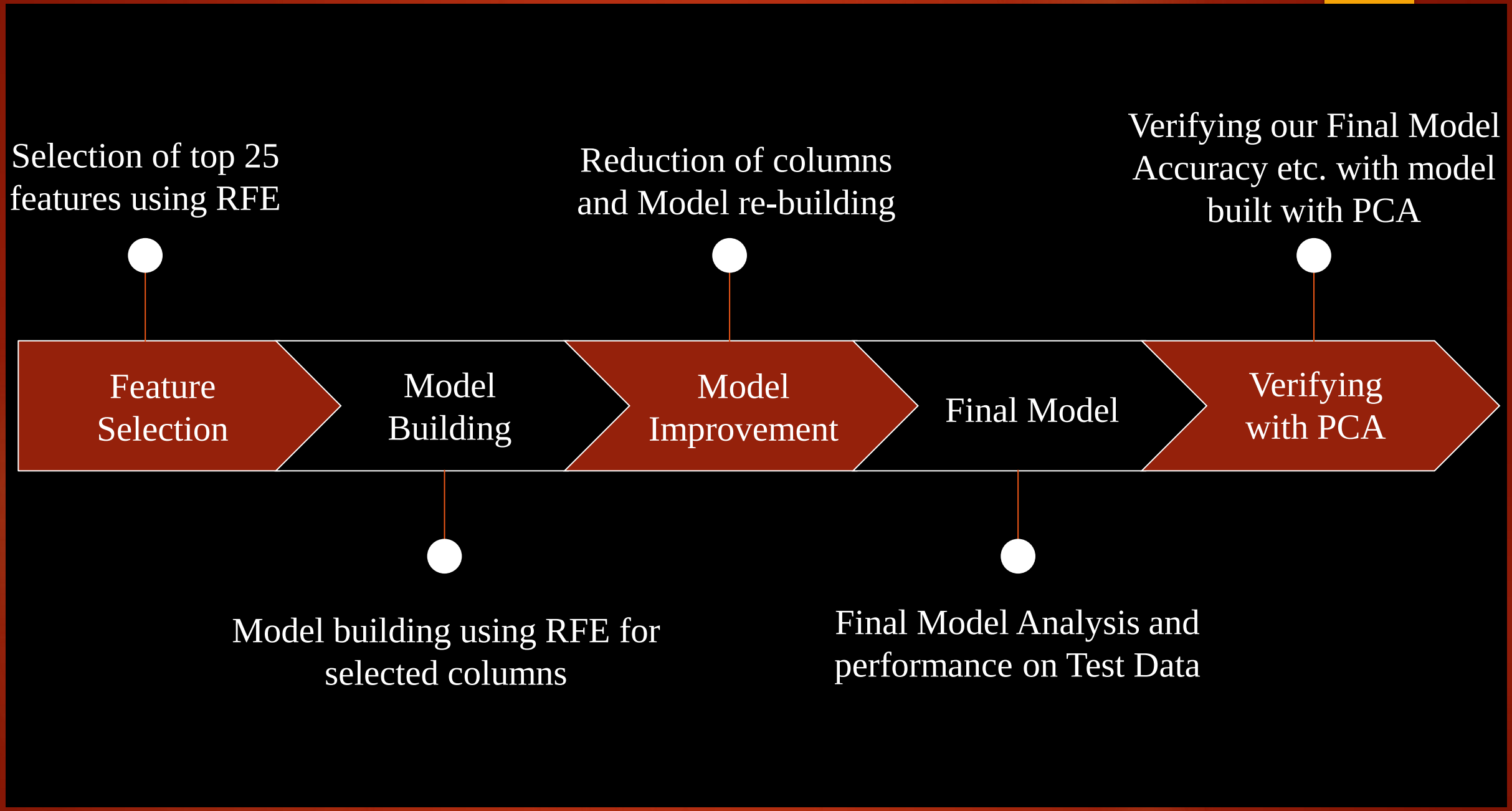
Solution

Selection of Hot Leads

- For our Problem Solution, the crucial part is to accurately identify hot leads.
- The more accurate we obtain the hot lead, the more chance we get of higher conversion ratio.
- Since we have a target of 80% conversion rate, we would want to obtain a high accuracy in obtaining hot leads.

Implementation

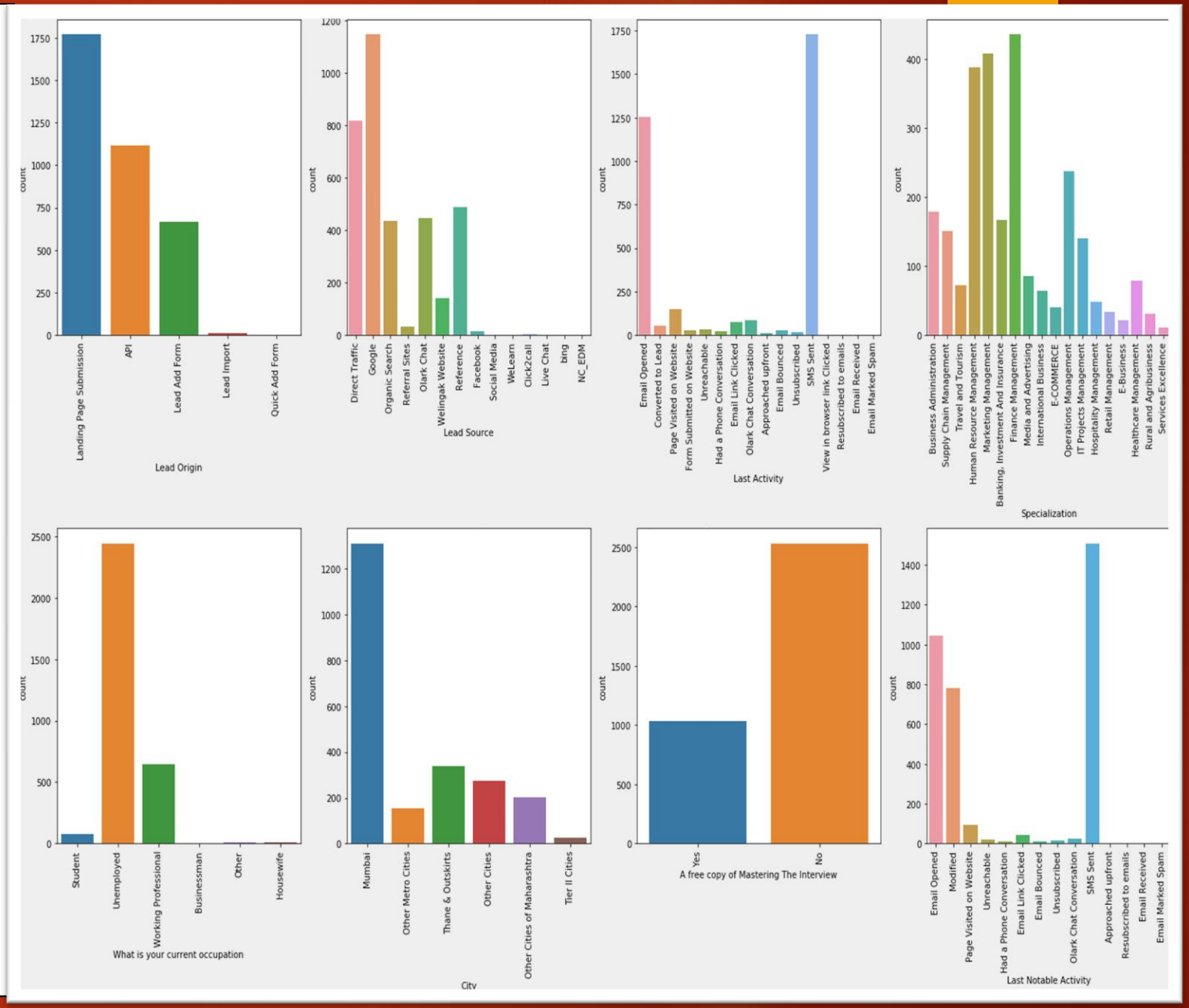

Loading & Observing the past data provided by the Company
Univariate, Bivariate, and Heatmap for numerical and categorical columns
Performing pre-requisites for RFE and Logistic Regression
Data Gathering
Data Cleaning
Performing EDA
Data Preparation
Model Building
Duplicate removal, null value treatment, unnecessary column elimination, etc.
Outlier Treatment, Feature-Standardization


Selection of top 25 features using RFE
Reduction of columns and Model re-building
Verifying our Final Model Accuracy etc. with model built with PCA
Feature Selection
Model Building
Model Improvement
Final Model
Verifying with PCA
Model building using RFE for selected columns
Final Model Analysis and performance on Test Data

Plots (Visualization)

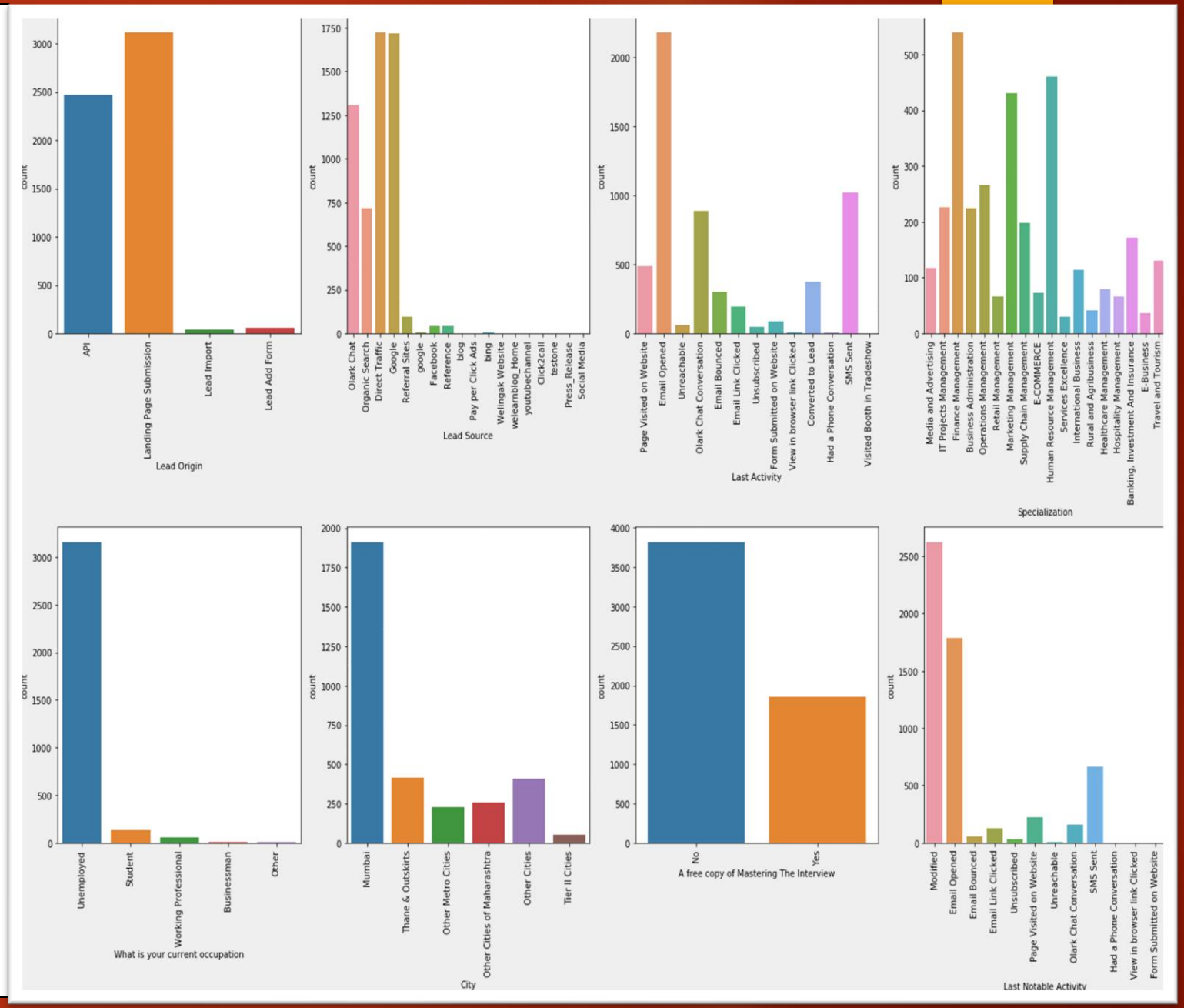
Facts about those whose leads are getting converted

1. Leads originated from website landing page submission.
2. Lead sourced from google.
3. Last activity done by the user is sent SMS same for last notable activity also.
4. Users are generally filling the form for 'Finance Management’.
5. Unemployed are mainly interested about x education company.
6. User belongs to Mumbai.
7. Those whose leads are converted are not interested about free copy of mastering the interview.

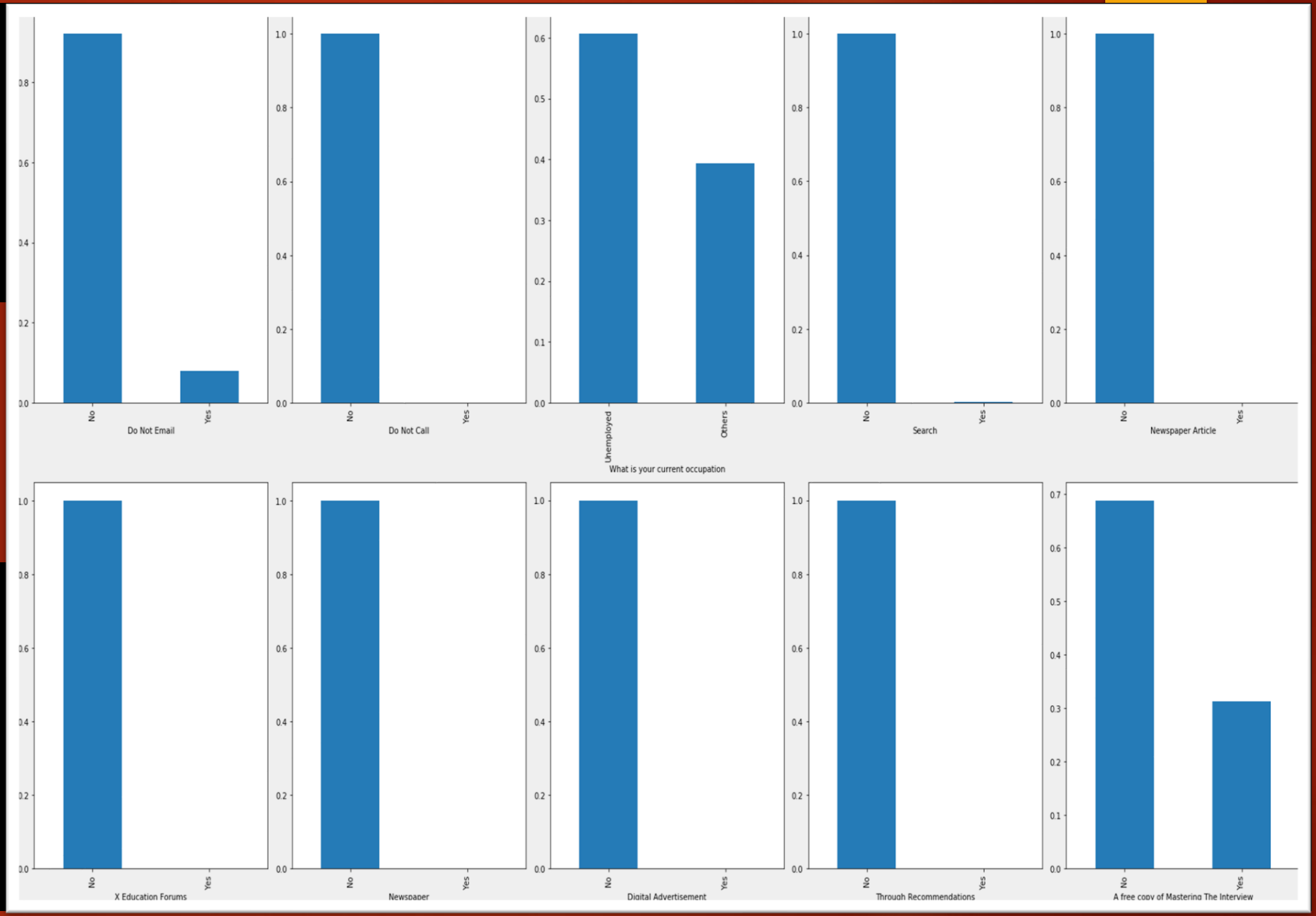


Facts about Them Whose leads are not getting converted

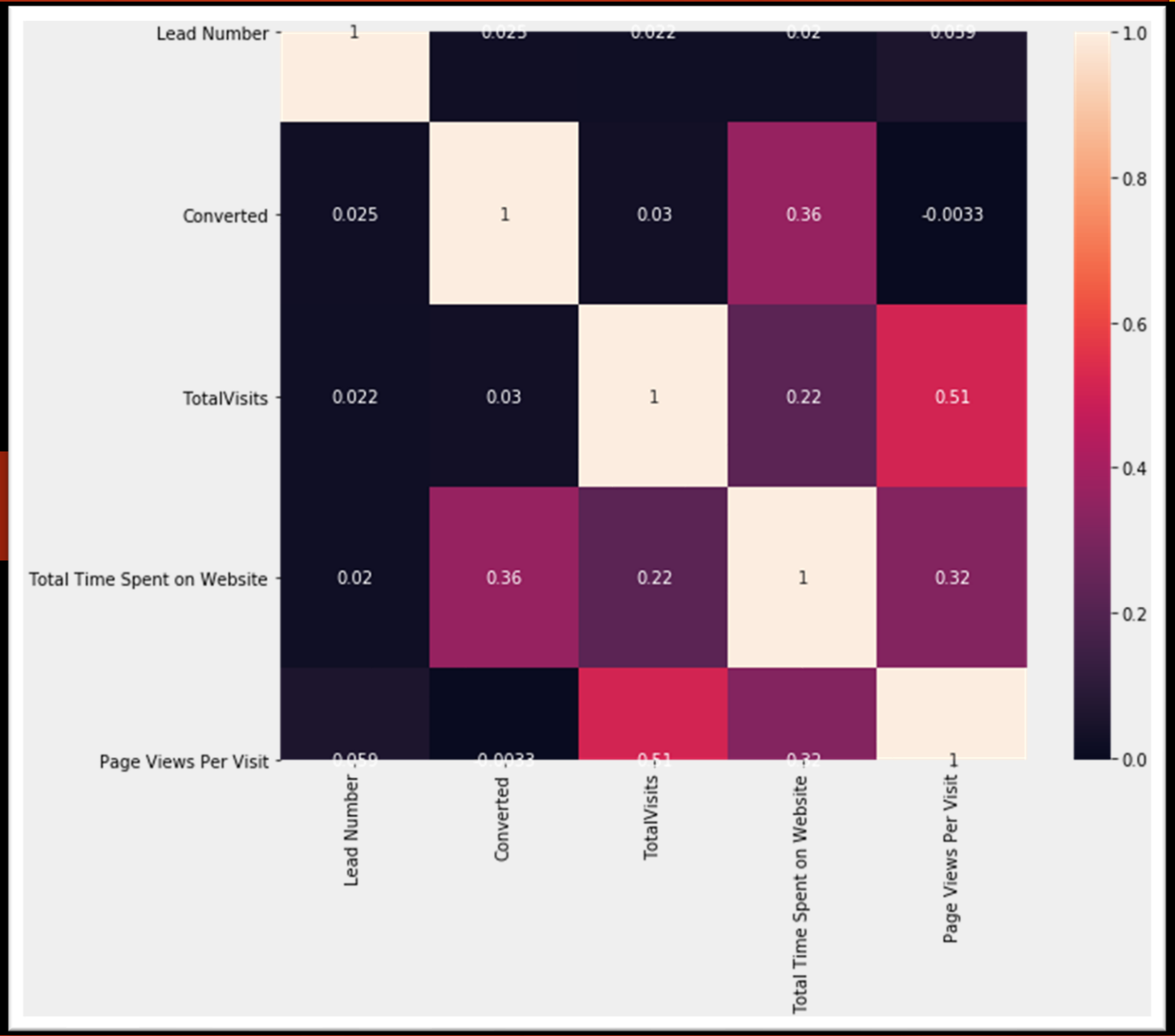
1. Leads originated from website landing page submission.
2. Lead sourced from Direct Traffic.
3. Last activity done by the user is email opened and last notable activity is modified.
4. Users are generally filling the form for 'Finance Management'.
5. Unemployed are mainly interested about x education company.
6. User belongs to Mumbai.
7. Those whose leads are converted are not interested about free copy of mastering the interview.



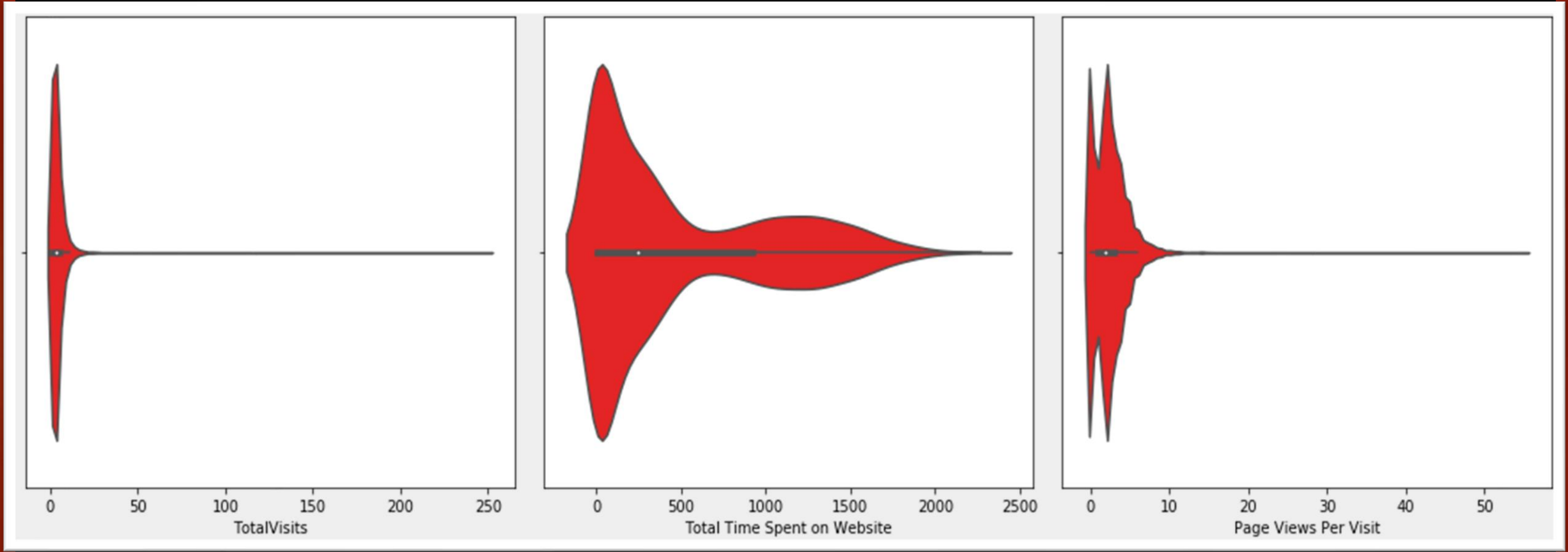
Visualization for 2 unique values



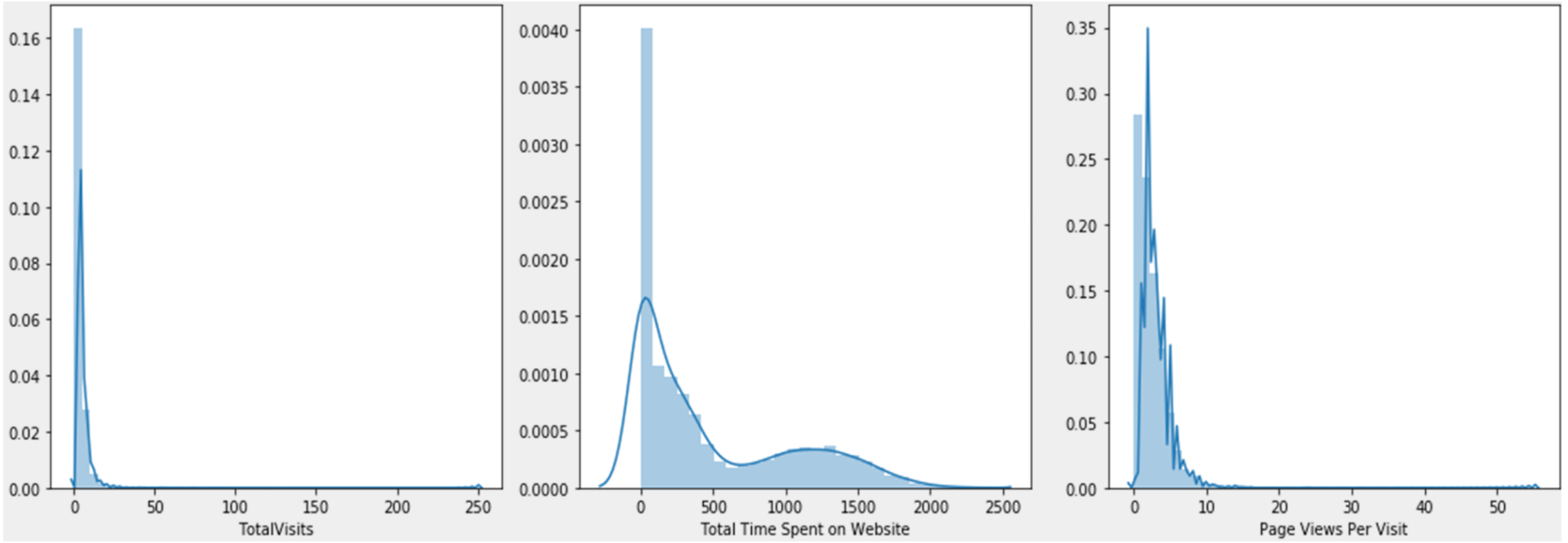
Heatmap


Lead Number
Converted
TotalVisits
Total Time Spent on Website
Page Views Per Visit
1
0.025
0.022
0.02
0.059
0.025
1
0.03
0.36
-0.0033
0.022
0.03
1
0.22
0.51
0.02
0.36
0.22
1
0.32
0.059
-0.0033
0.51
0.32
1
1.0
0.8
0.6
0.4
0.2
0.0

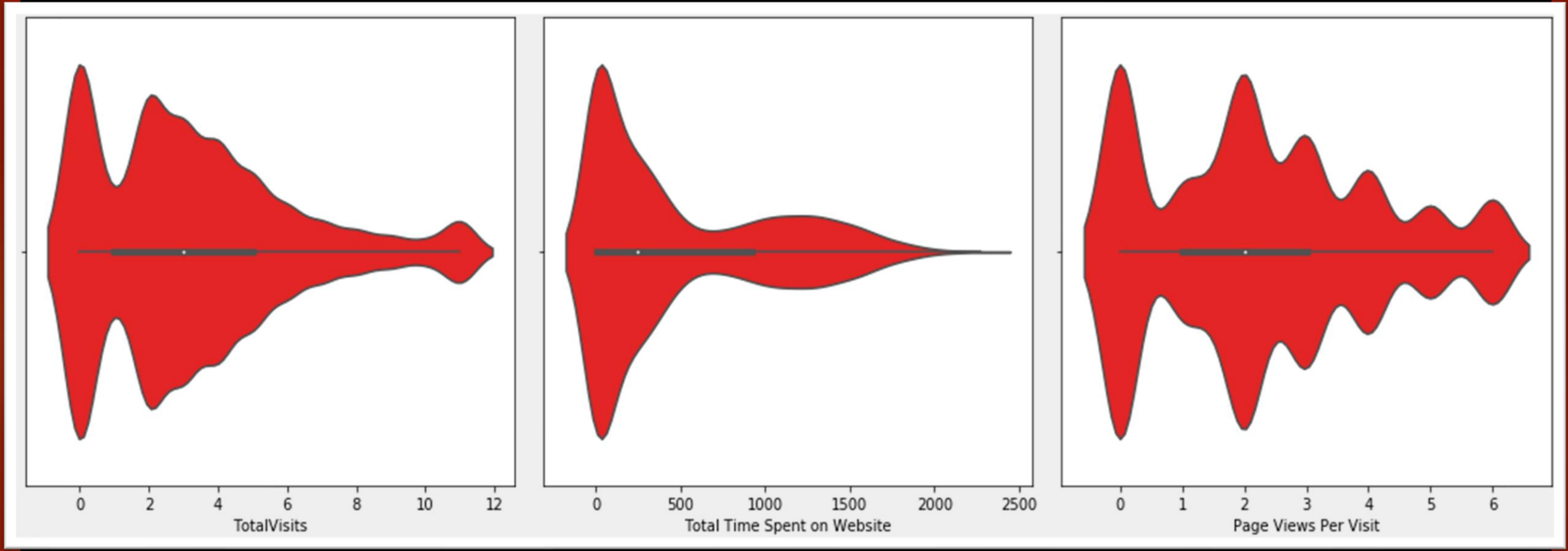
Visualisation of Outliers



Checking the Distribution of Data



Visualisation of Outliers After Removal



Inference / Conclusion

Model Analysis

Performance of our Final Model

- Overall accuracy on Train set: 79.03%.
- Sensitivity of our logistic regression model: 78.71%.
- Specificity of our logistic regression model: 79.23%.

Inferences from Model

Business Insights Derived from our Model

The top three variables in our model which contribute most towards the probability of a lead getting converted are:

- Lead Origin_Other Origins
- Lead Source_Olark Chat
- Total Time Spent on Website

Conclusion 1 (LR Model)

Our logistic regression model is decent and accurate enough, when compared to the model derived using PCA, with 79.03 % accuracy on test set, 78.71 % sensitivity and 79.23 % specificity.
We can vary these parameters by varying the cut-off value and thus predict hot leads based on scenarios like availability of extra resources and vice-versa.

Conclusion 2 (Recommendation)

X education company needs to focus on following key aspects to improve the overall conversion rate:

- Increase user engagement on their website since this helps in higher conversion
- Increase on sending SMS notifications since this helps in higher conversion
- Get total visits increased by advertising etc. Since this helps in higher conversion
- Improve the olark chat service since this is affecting the conversion negatively