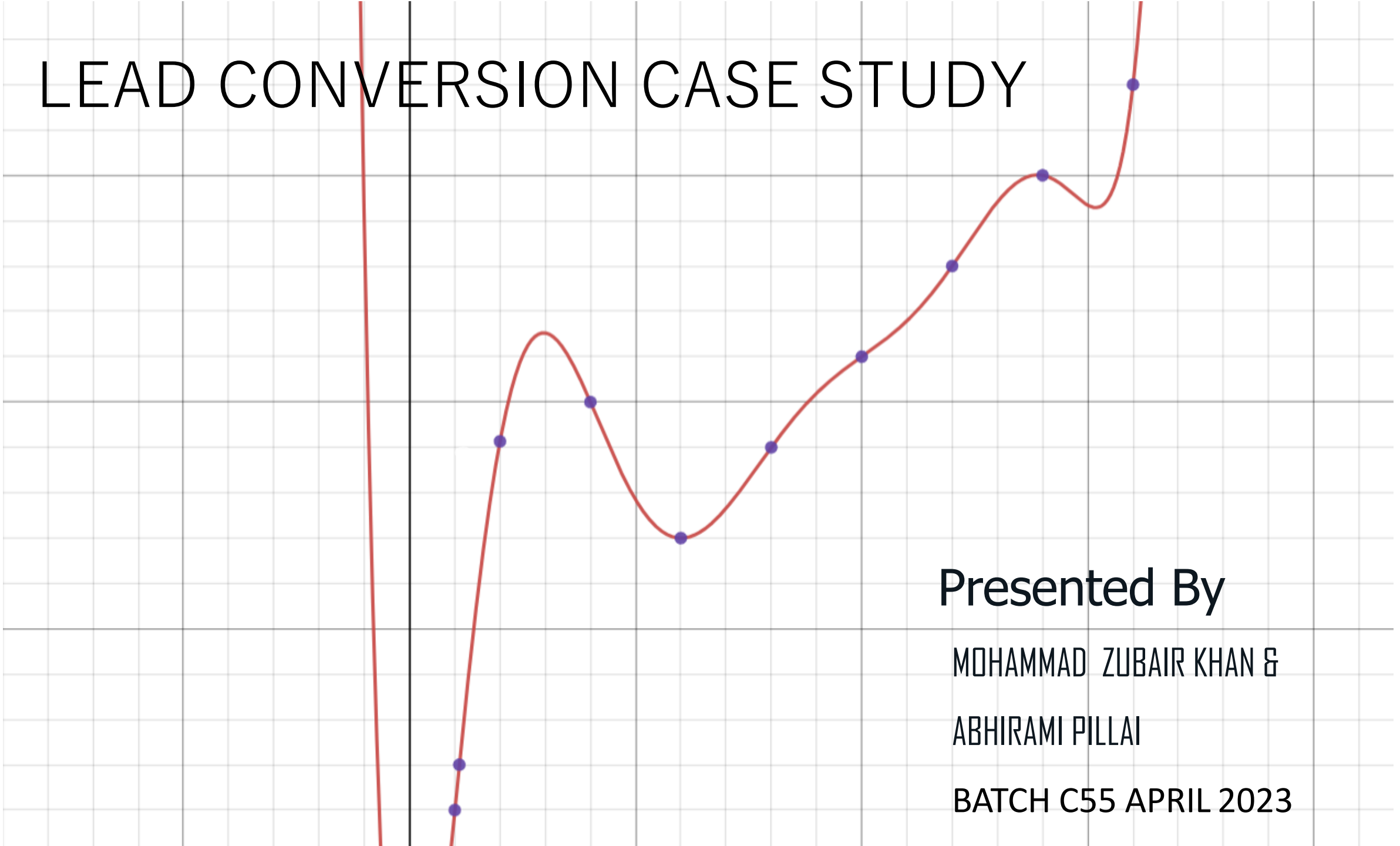
LEAD CONVERSION CASE STUDY



Presented By

MOHAMMAD ZUBAIR KHAN &

ABHIRAMI PILLAI

BATCH C55 APRIL 2023

Introduction

Objective

The objective of this analysis is to assist X Education, an online course provider, in optimizing its lead conversion process. X Education sells courses to industry professionals and faces a challenge with its lead conversion rate, which currently stands at around 30%. The company aspires to improve this rate and identify the most promising leads, referred to as 'Hot Leads,' with a target lead conversion rate of approximately 80%.

Steps Involved

1. **Loading Data**
2. **Inspecting Data**
3. **Data Cleaning and Analysis (Outliers, nulls, datatypes)**
4. **Checking the Data Imbalance**
5. **Categorical Univariate analysis**
6. **Numerical Univariate Analysis**
7. **Bivariate : Numerical – Categorical Analysis**
8. **Bivariate: Numerical – Numerical Analysis**
9. **Correlations**
10. **Analysing "previous_application.csv"**
11. **Conclusions**

LOADING DATA

- **Data Sources**

1. Leads data : **Leads.csv.**
2. Data dictionary : **Leads Data Dictionary.xlsx.**

EDA ANALYSIS

- **Data Frame Inspection**

- Exploring and Understanding Data

- Examining a few records from the dataset using methods such as .shape, .info(), and .describe().
- Data Cleaning
- Handling Data Type and Inconsistent Data

- Assessing the percentage of null values in the data frame, ordered in descending order.
- Examining the number of columns with null values.
- Removing columns with null values exceeding 40%.
- Imputing columns with null values less than or equal to 2% using the mode value for numeric columns, except for continuous numeric columns where the median value was employed. And then going case by case for all columns that had between 2% nulls to 40% nulls.

- Furthermore, I transformed the values in the columns which had Yes and No values in them. From ‘Yes' and ‘No' to 1 and 0, respectively, for ease of analysis.

Analysis

Data Types Conversion of Variables

During our examination of the data frame, we identified several columns that are categorized as "object" data types. To optimize our analysis and prepare data for logistic regression modeling, we decided to convert and obtain dummy columns with one-hot encoding.

Considering the large number of columns present in the data frame, we will proceed by eliminating the columns that are not required for our subsequent analysis. This will help streamline the dataset and focus on the relevant columns for our investigation.

Outlier Analysis

I created boxplots for relevant of these columns.

- Outliers are evident in both the variables "TotalVisits" and "Page Views Per Visit," indicating the need for outlier treatment. Additionally, it's noteworthy that the values are significantly skewed above the median in the "Total Time Spent on Website.“

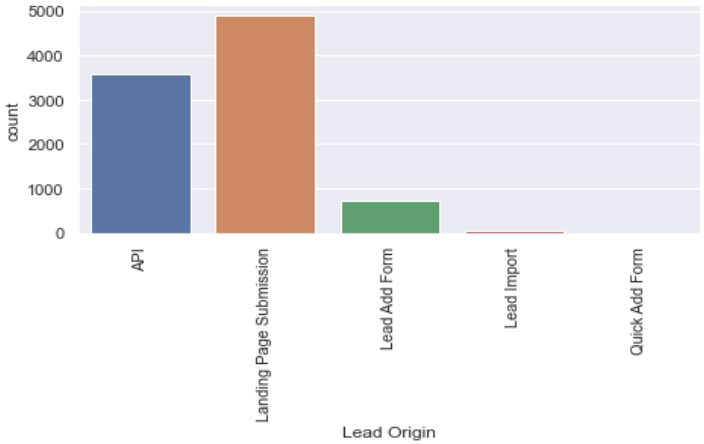
I have provided visual representations of these observations through the utilization of boxplots in .ipynb

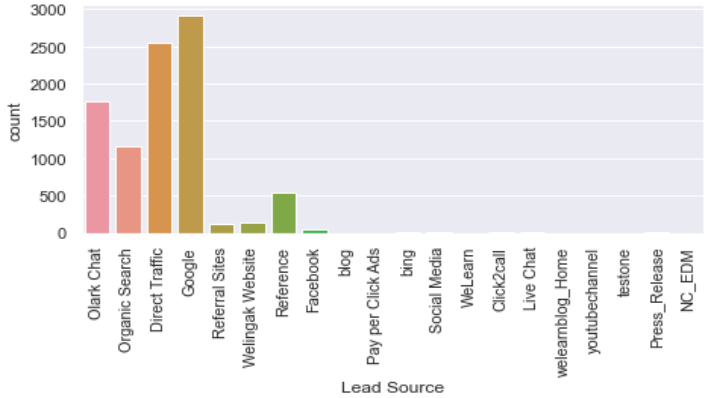
Data Imbalance

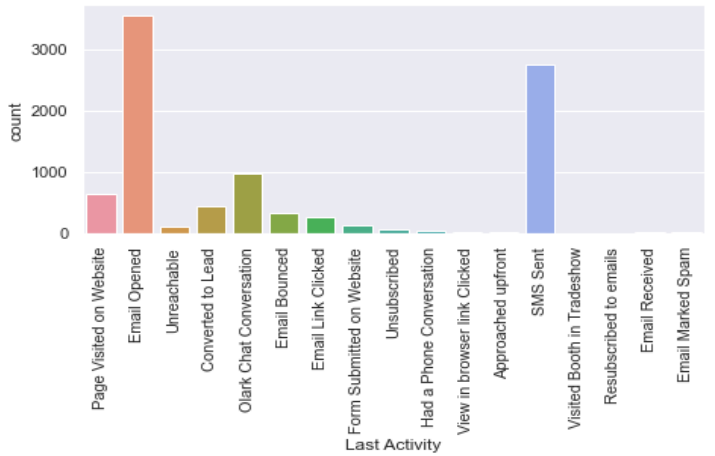
Checking Data Imbalance for Target Variable

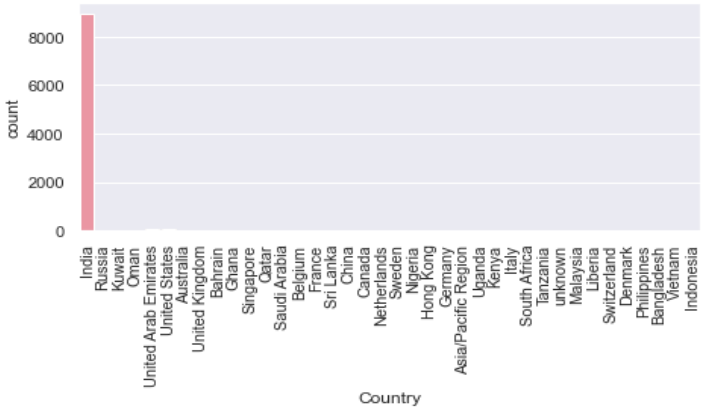
38% is the Conversion rate and the data imbalance of Target variable

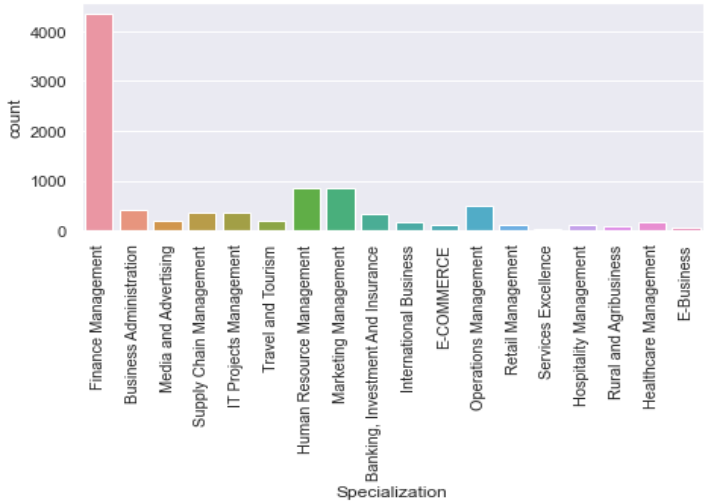
Univariate Analysis Categorical

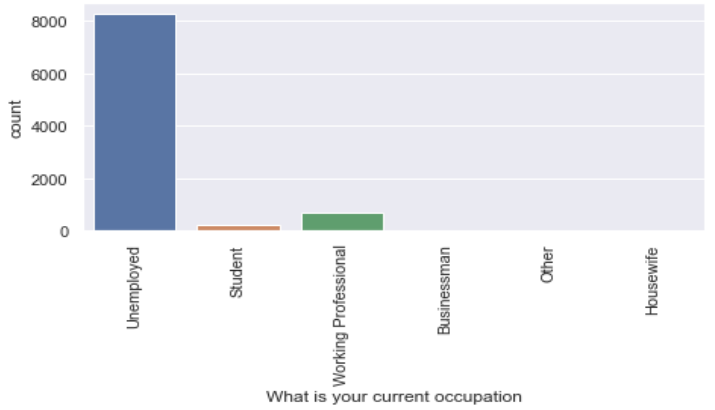

count
API
Landing Page Submission
Lead Add Form
Lead Import
Quick Add Form
Lead Origin

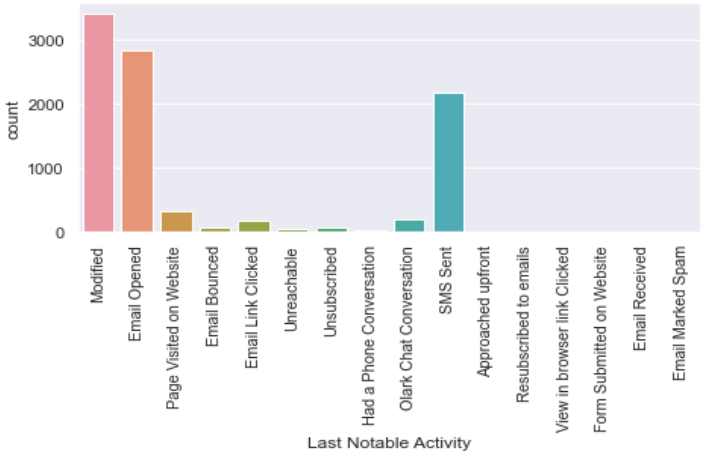

count
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Welingak Website
Reference
Facebook
blog
Pay per Click Ads
bing
Social Media
WeLearn
Click2call
Live Chat
welearnblog_Home
youtubechannel
testone
Press_Release
NC_EDM
Lead Source


count
Page Visited on Website
Email Opened
Unreachable
Converted to Lead
Olark Chat Conversation
Email Bounced
Email Link Clicked
Form Submitted on Website
Unsubscribed
Had a Phone Conversation
View in browser link Clicked
Approached upfront
SMS Sent
Visited Booth in Tradeshow
Resubscribed to emails
Email Received
Email Marked Spam
Last Activity

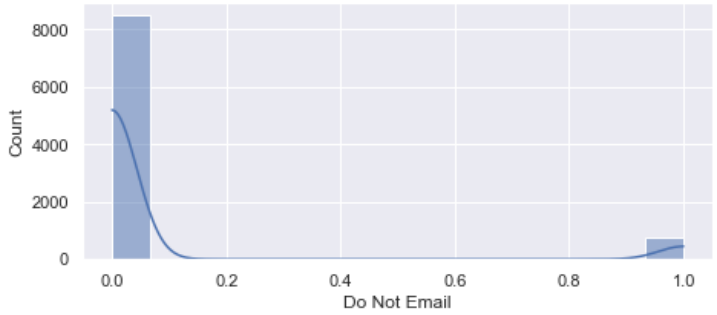

count
India
Russia
Kuwait
Oman
United Arab Emirates
United States
Australia
United Kingdom
Bahrain
Ghana
Singapore
Qatar
Saudi Arabia
Belgium
France
Sri Lanka
China
Canada
Netherlands
Sweden
Nigeria
Hong Kong
Germany
Asia/Pacific Region
Uganda
Kenya
Italy
South Africa
Tanzania
unknown
Malaysia
Liberia
Switzerland
Denmark
Philippines
Bangladesh
Vietnam
Indonesia
Country

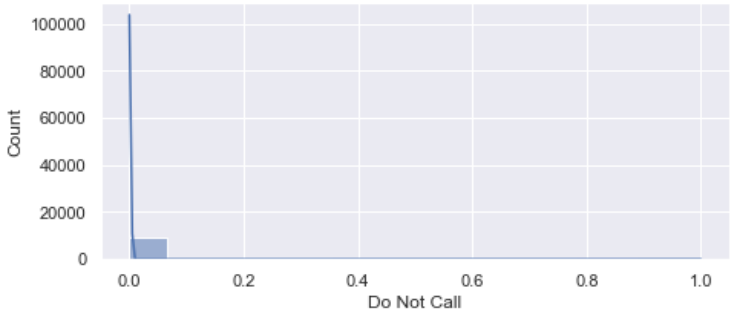

count
Finance Management
Business Administration
Media and Advertising
Supply Chain Management
IT Projects Management
Travel and Tourism
Human Resource Management
Marketing Management
Banking, Investment And Insurance
International Business
E-COMMERCE
Operations Management
Retail Management
Services Excellence
Hospitality Management
Rural and Agribusiness
Healthcare Management
E-Business
Specialization

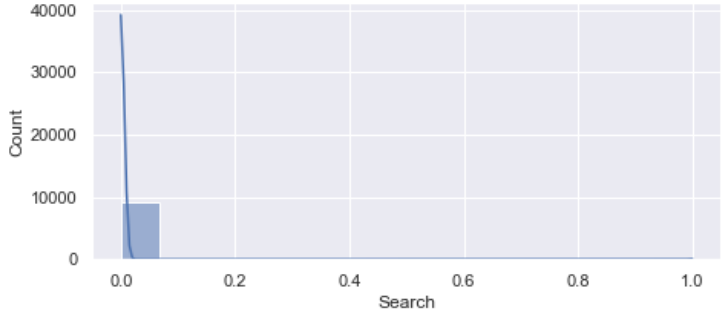

count
Unemployed
Student
Working Professional
Businessman
Other
Housewife
What is your current occupation

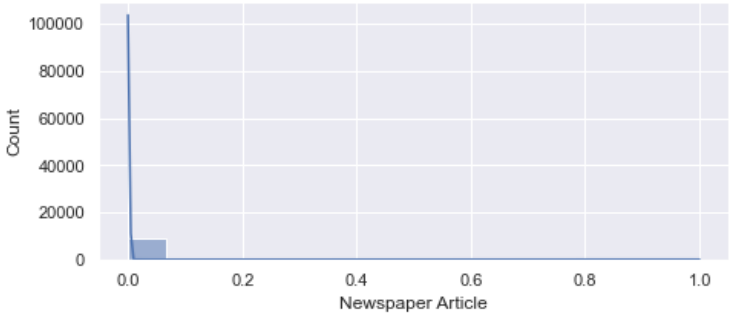

count
Modified
Email Opened
Page Visited on Website
Email Bounced
Email Link Clicked
Unreachable
Unsubscribed
Had a Phone Conversation
Olark Chat Conversation
SMS Sent
Approached upfront
Resubscribed to emails
View in browser link Clicked
Form Submitted on Website
Email Received
Email Marked Spam
Last Notable Activity

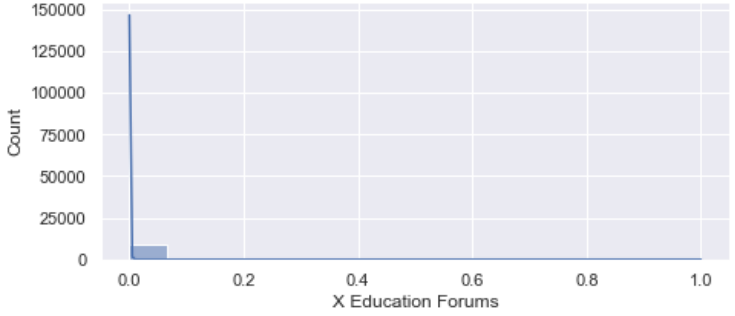
Univariate Analysis Numerical

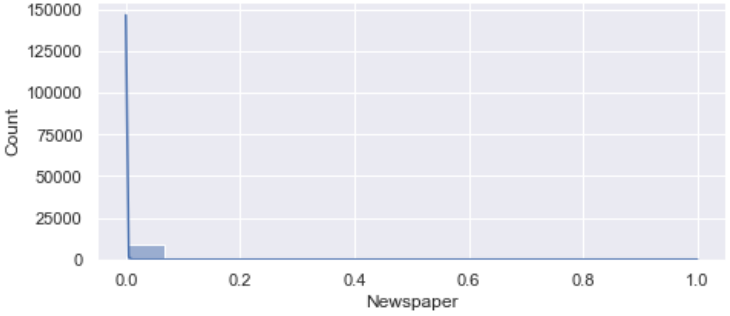

Count
Do Not Email

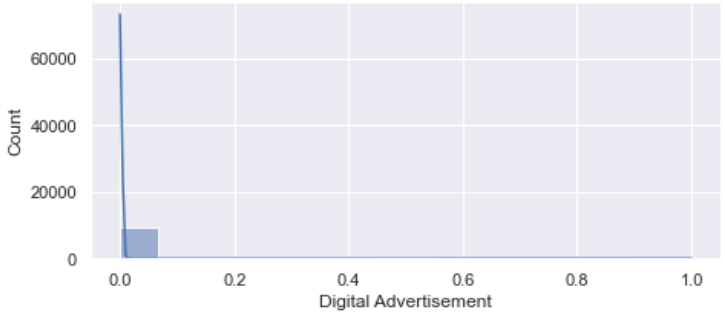

Count
Do Not Call

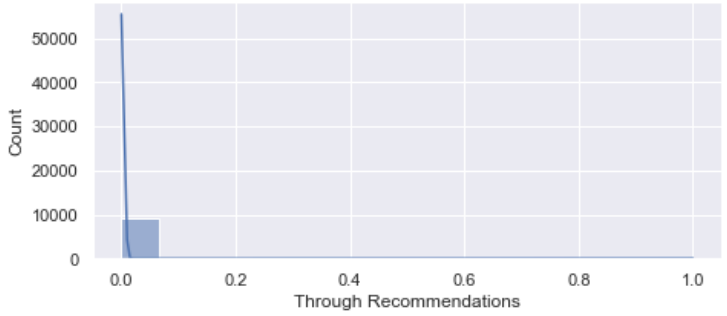

Count
Search

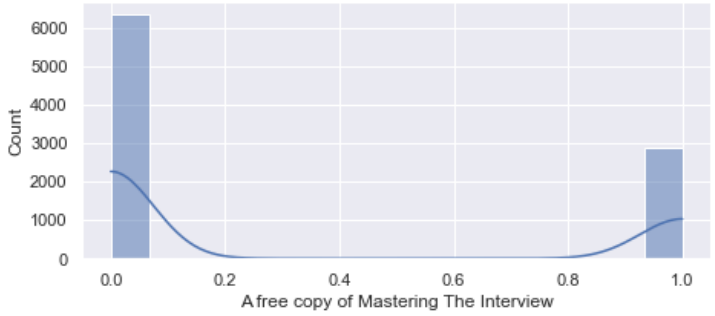

Count
Newspaper Article


Count
X Education Forums

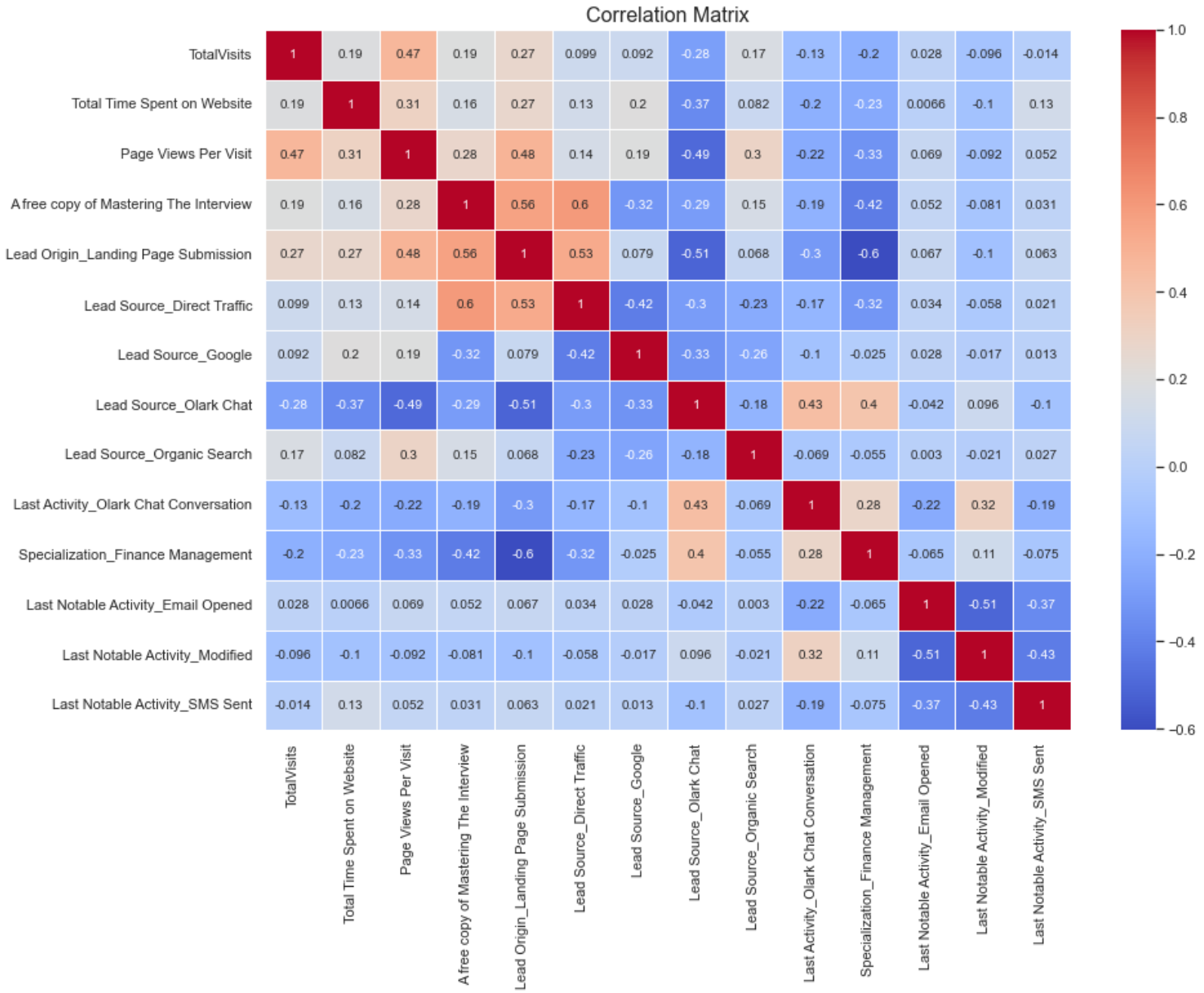

Count
Newspaper


Count
Digital Advertisement

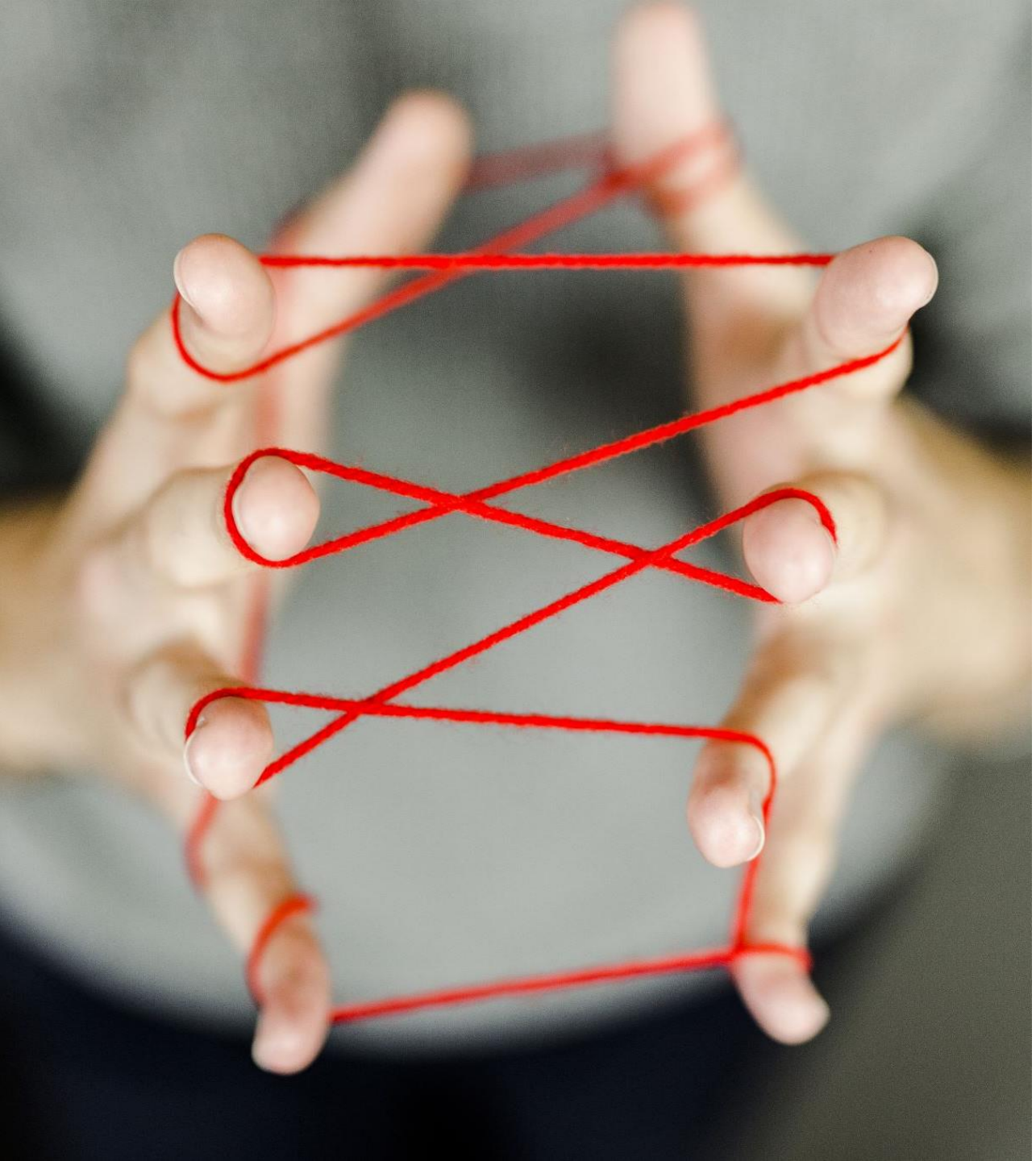

Count
Through Recommendations


Count
A free copy of Mastering The Interview

Correlations



Correlation



Conclusions

Final Insights

The Features that are highly positively influencing the lead conversion.

- Total Time Spent on Website
- Last Notable Activity_SMS Sent
- Last Notable Activity_Email Opened

The Features that are least significant in the lead conversion.

- Lead Source Direct Traffic
- Lead Source Google

The Accuracy was around 80% and the lead score is 100 multiplied by the conversion_prob

THANK YOU