Lead Scoring Case Study

Joyita Sadhukhan

Problem Statement

- X Education sells online courses to industry professionals.
- X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.
- 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.

Business Objective

- X education wants to know most promising leads.
- For that they want to build a Model which identifies the hot leads.
- Deployment of the model for the future use.

Methodology

- Data cleaning and data manipulation.
  1. Check and handle duplicate data.
  2. Check and handle NA values and missing values.
  3. Drop columns, if it contains large amount of missing values and not useful for the analysis.
  4. Imputation of the values, if necessary.
  5.Check and handle outliers in data.

- EDA
  1. Univariate data analysis: value count, distribution of variable etc.
  2. Bivariate data analysis: correlation coefficients and pattern between the variables etc.

- Feature Scaling & Dummy Variables and encoding of the data.

- Classification technique: logistic regression used for the model making and prediction.

- Validation of the model.

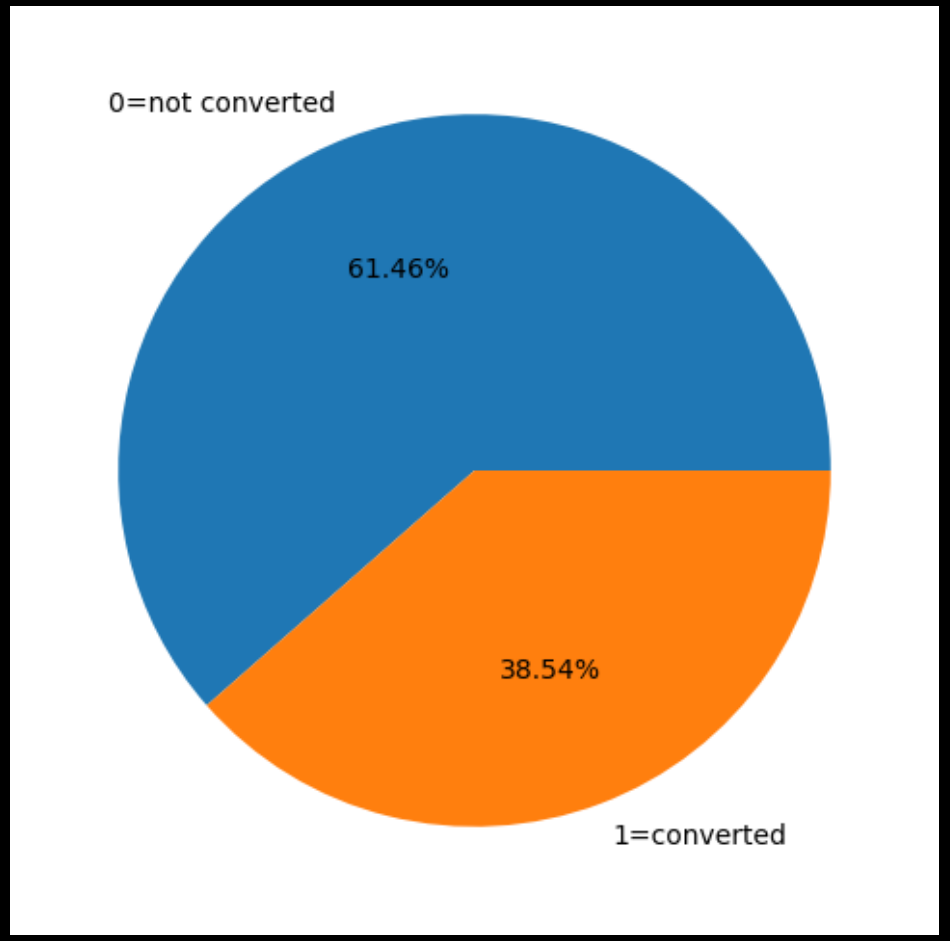
- Model presentation.

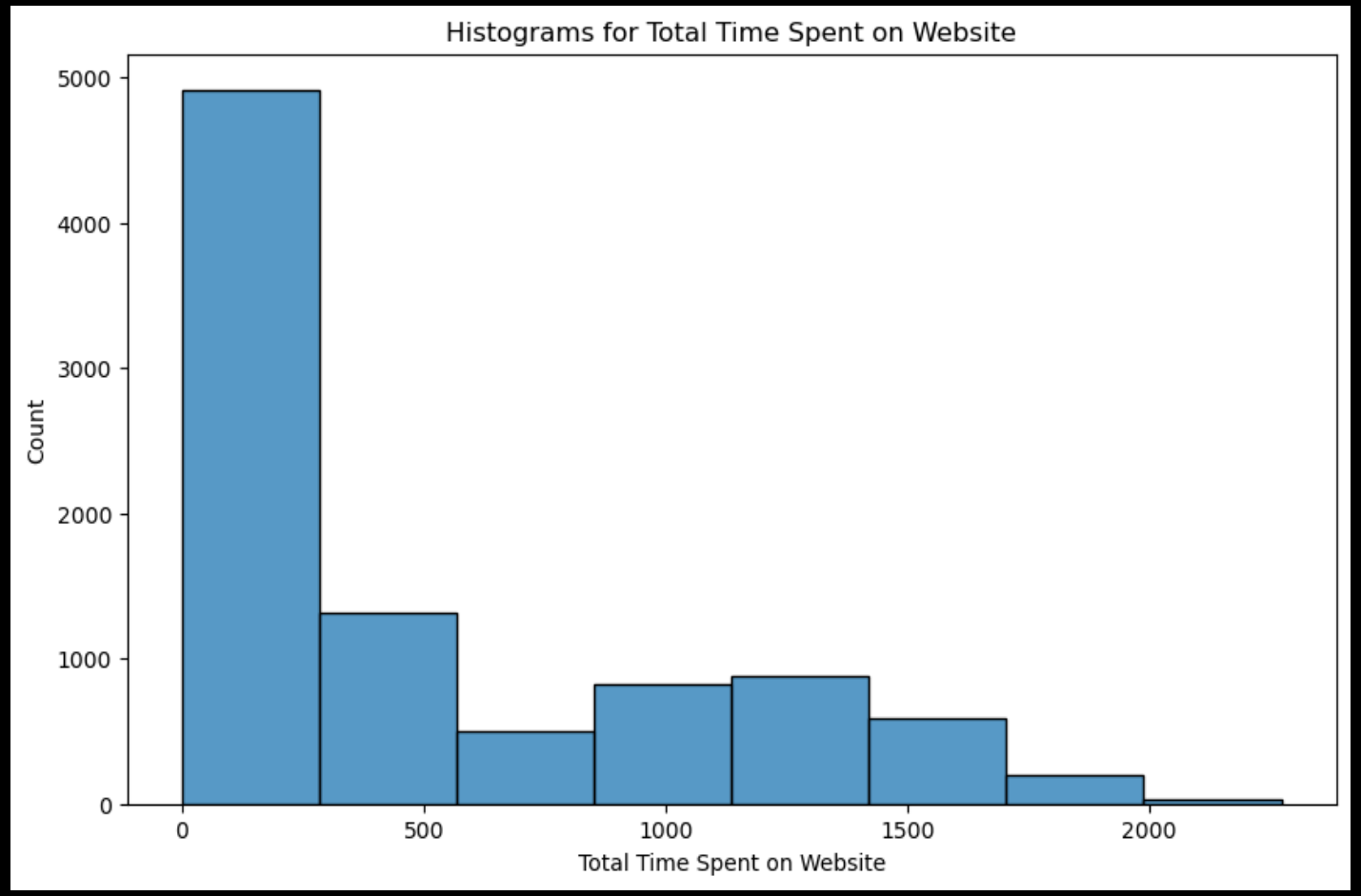
- Conclusion and Recommendation

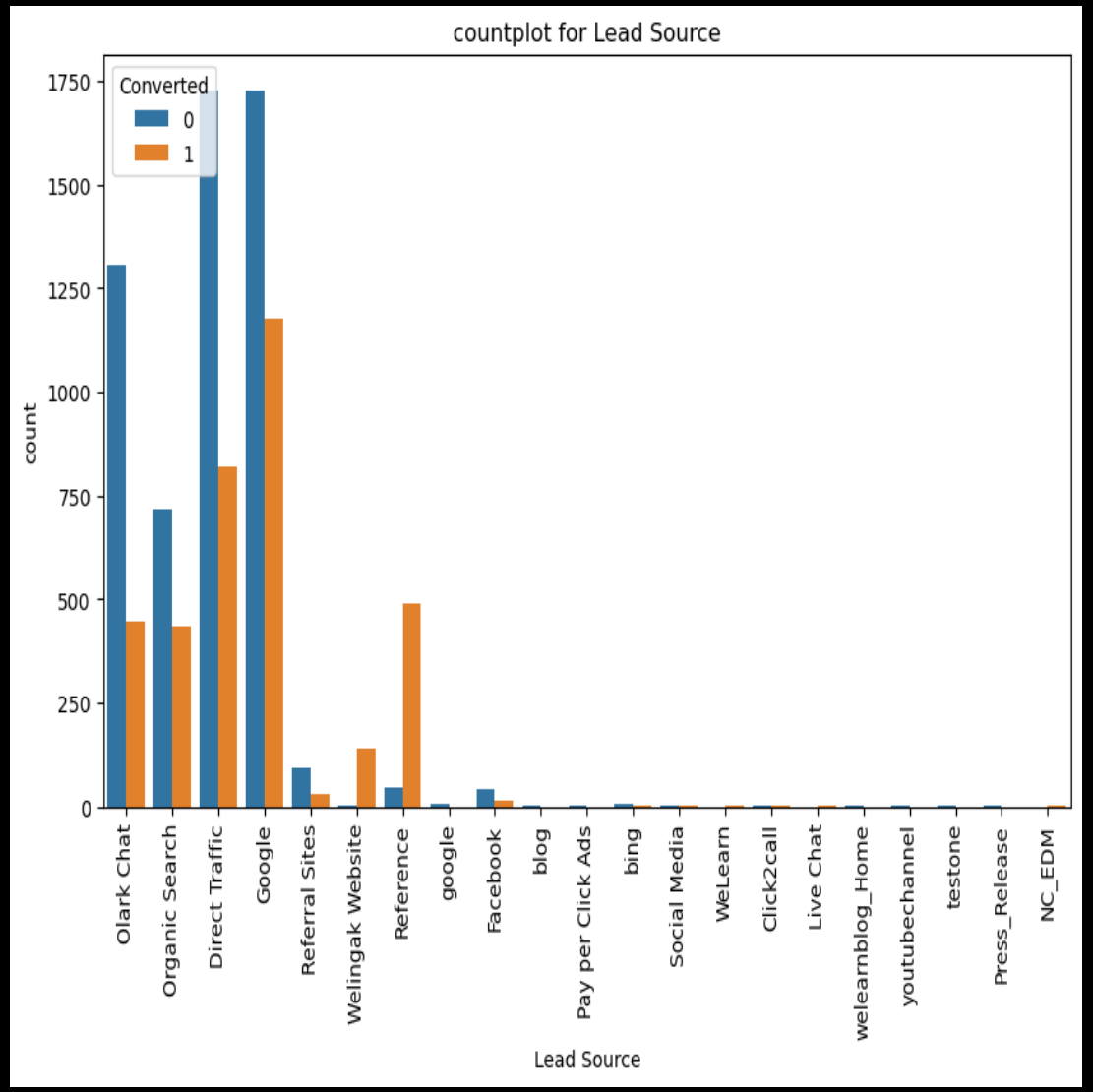
Data Manipulation

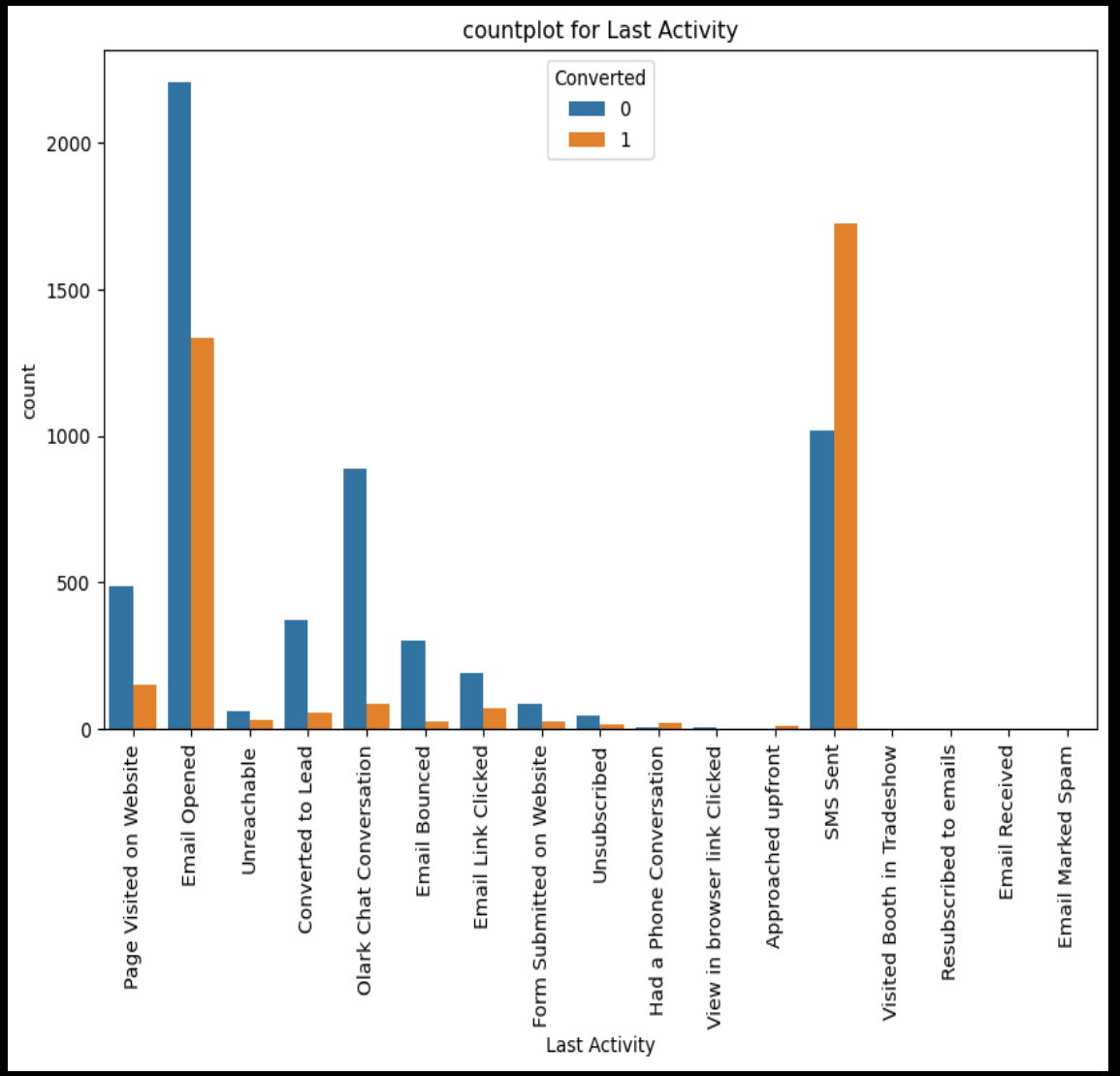
- Total Number of Rows =37, Total Number of Columns =9240.
- Single value features like “Magazine”, “Receive More Updates About Our Courses”, “Update me on Supply”, “Chain Content”, “Get updates on DM Content”, “I agree to pay the amount through cheque” etc. have been dropped.
- Removing the “Prospect ID” and “Lead Number” which is not necessary for the analysis.
- After checking the value counts for some of the object type variables, we find some of the features which has no enough variance, which we have dropped, the features are: “Do Not Call”, “What matters most to you in choosing course”, “Search”, “Newspaper Article”, “X Education Forums”, “Newspaper”, “Digital Advertisement” etc.
- Dropping the columns having more than 45% missing values such as ‘How did you hear about X Education’ and ‘Lead Profile’.

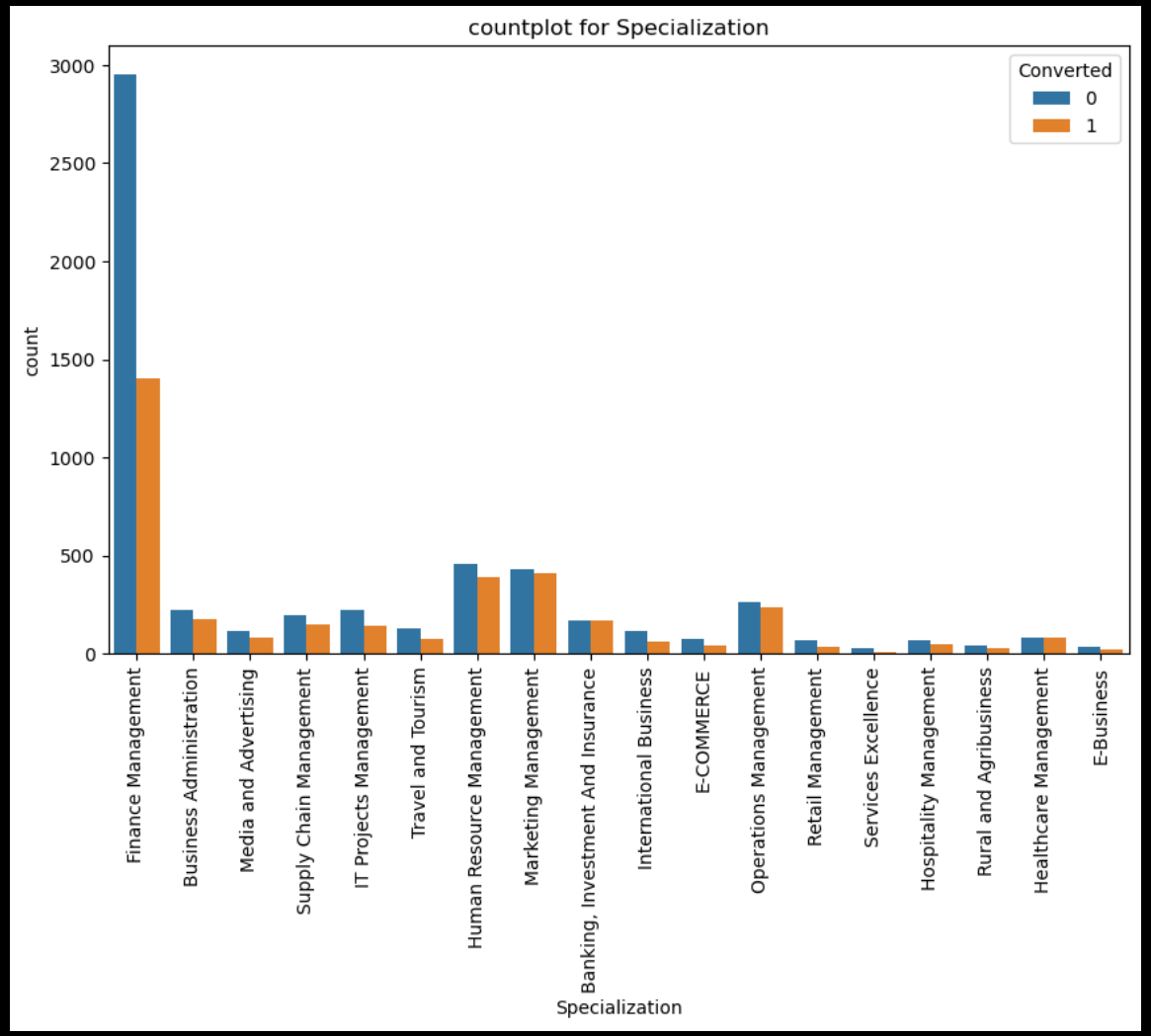
EDA

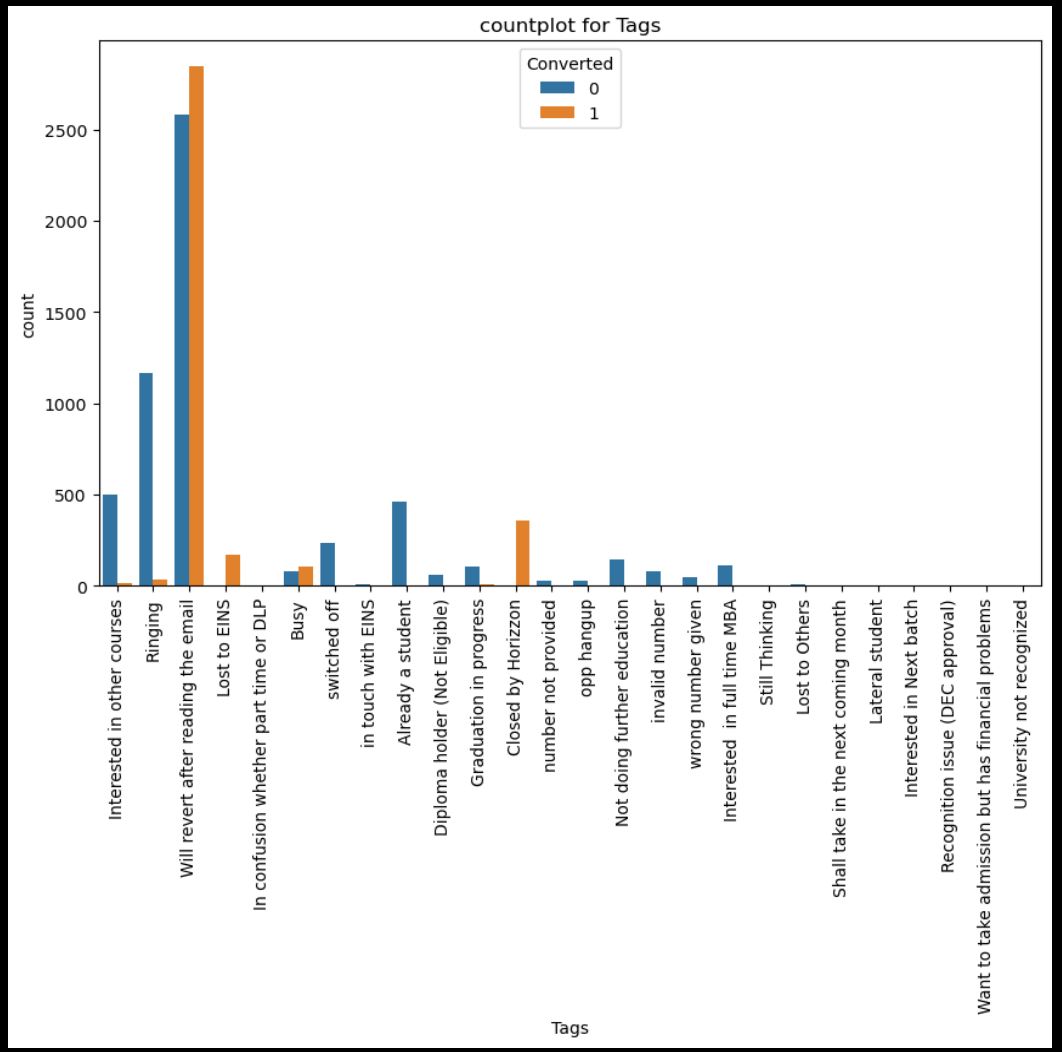

0=not converted
61.46%
38.54%
1=converted

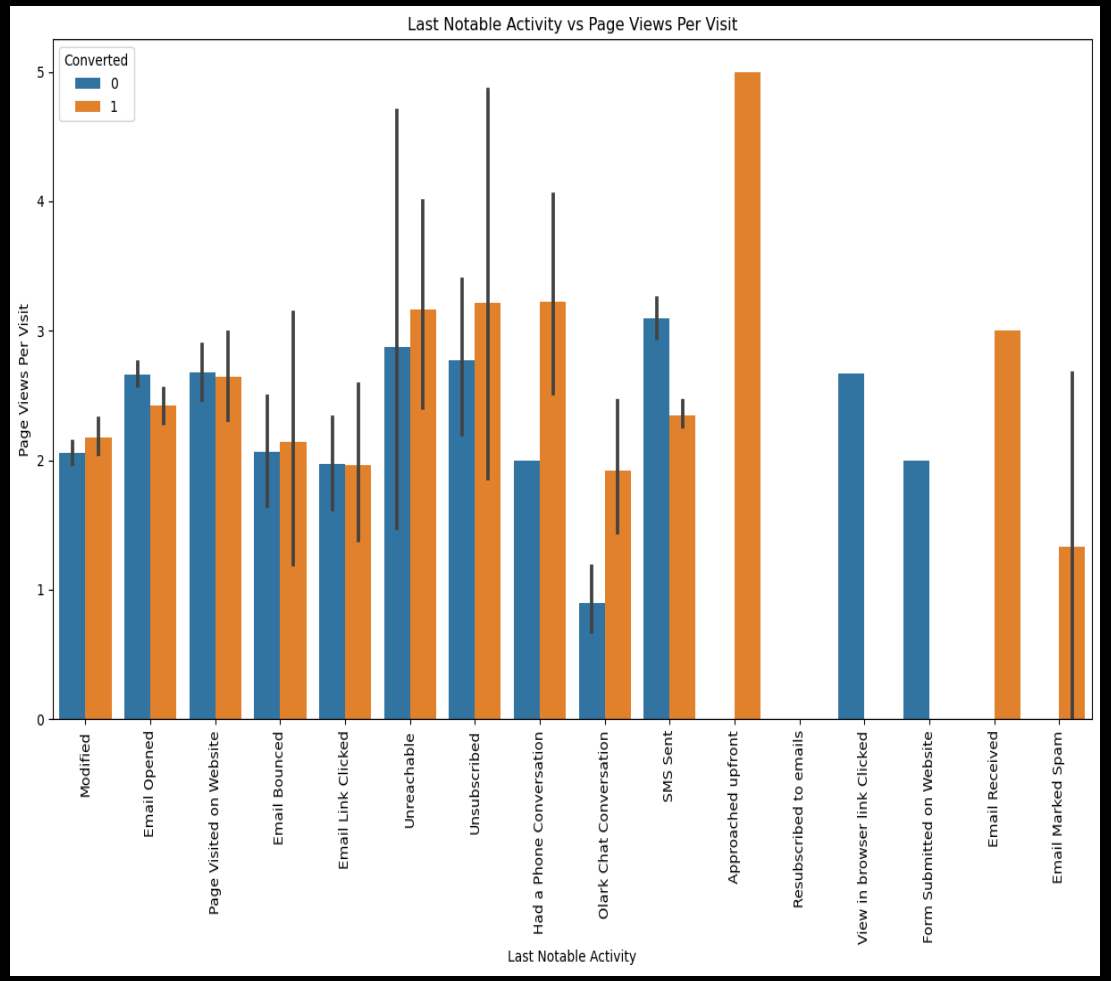

Histograms for Total Time Spent on Website
Count
5000
4000
3000
2000
1000
0
0
500
1000
1500
2000
Total Time Spent on Website

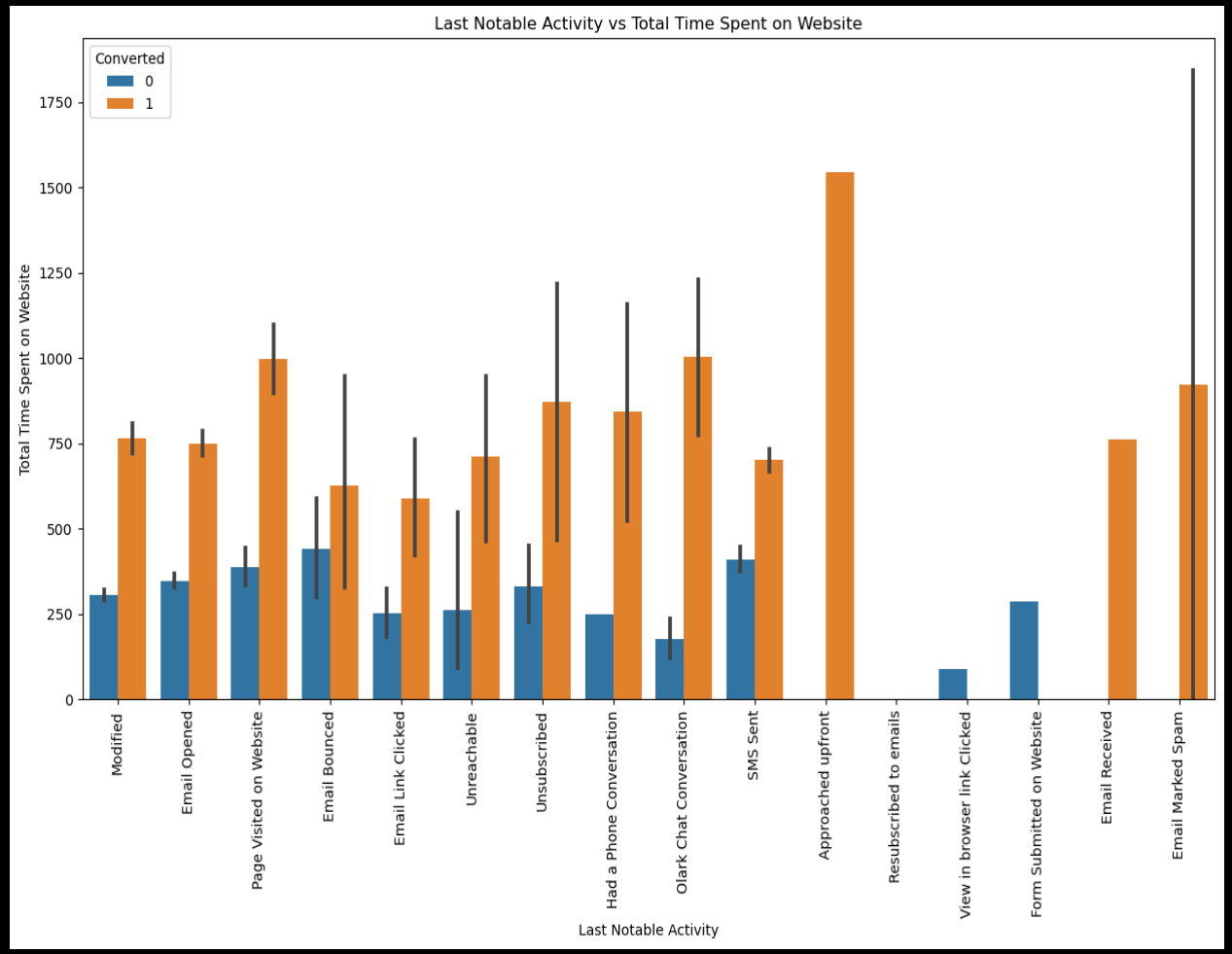

countplot for Lead Source
Converted
0
1
count
1750
1500
1250
1000
750
500
250
0
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Welingak Website
Reference
google
Facebook
blog
Pay per Click Ads
bing
Social Media
WeLearn
Click2call
Live Chat
welearnblog_Home
youtubechannel
testone
Press_Release
NC_EDM
Lead Source

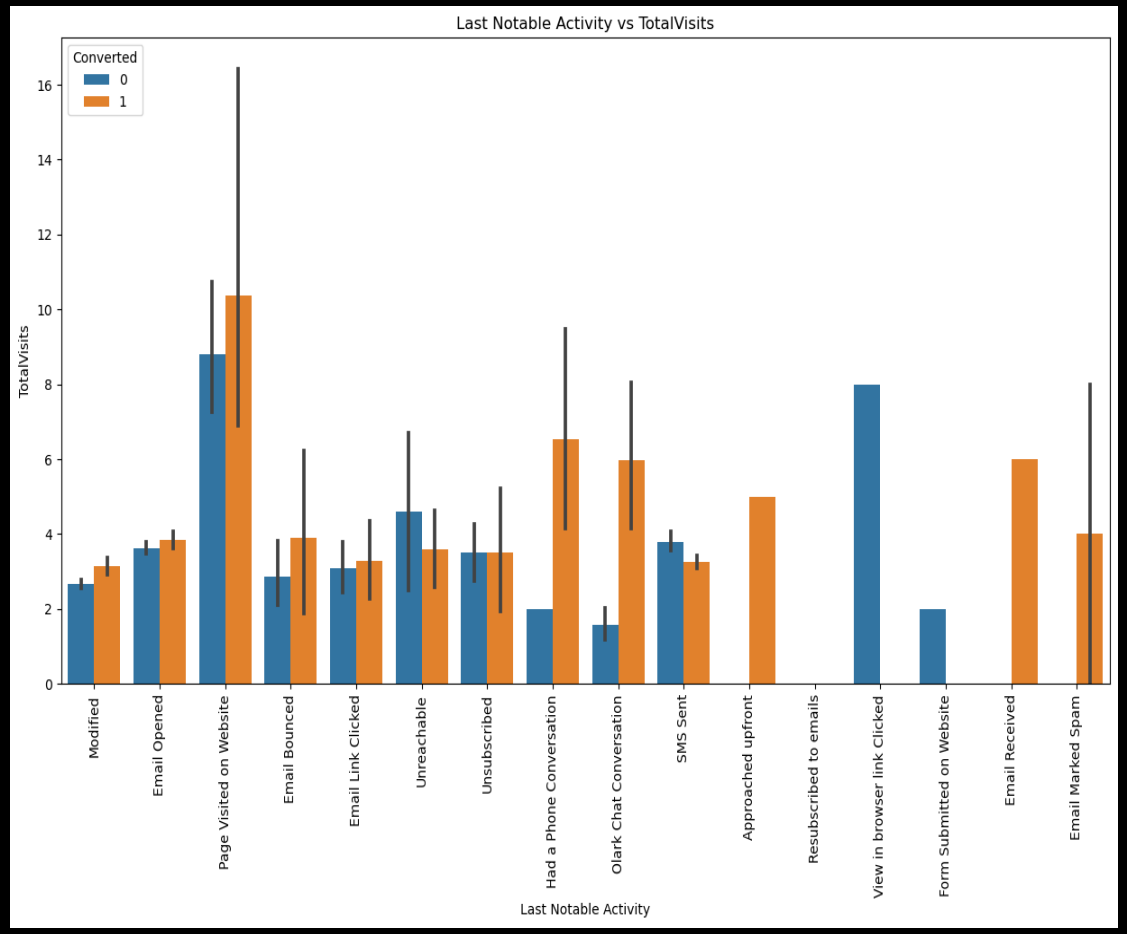

countplot for Last Activity
Converted
0
1
count
2000
1500
1000
500
0
Page Visited on Website
Email Opened
Unreachable
Converted to Lead
Olark Chat Conversation
Email Bounced
Email Link Clicked
Form Submitted on Website
Unsubscribed
Had a Phone Conversation
View in browser link Clicked
Approached upfront
SMS Sent
Visited Booth in Tradeshow
Resubscribed to emails
Email Received
Email Marked Spam
Last Activity

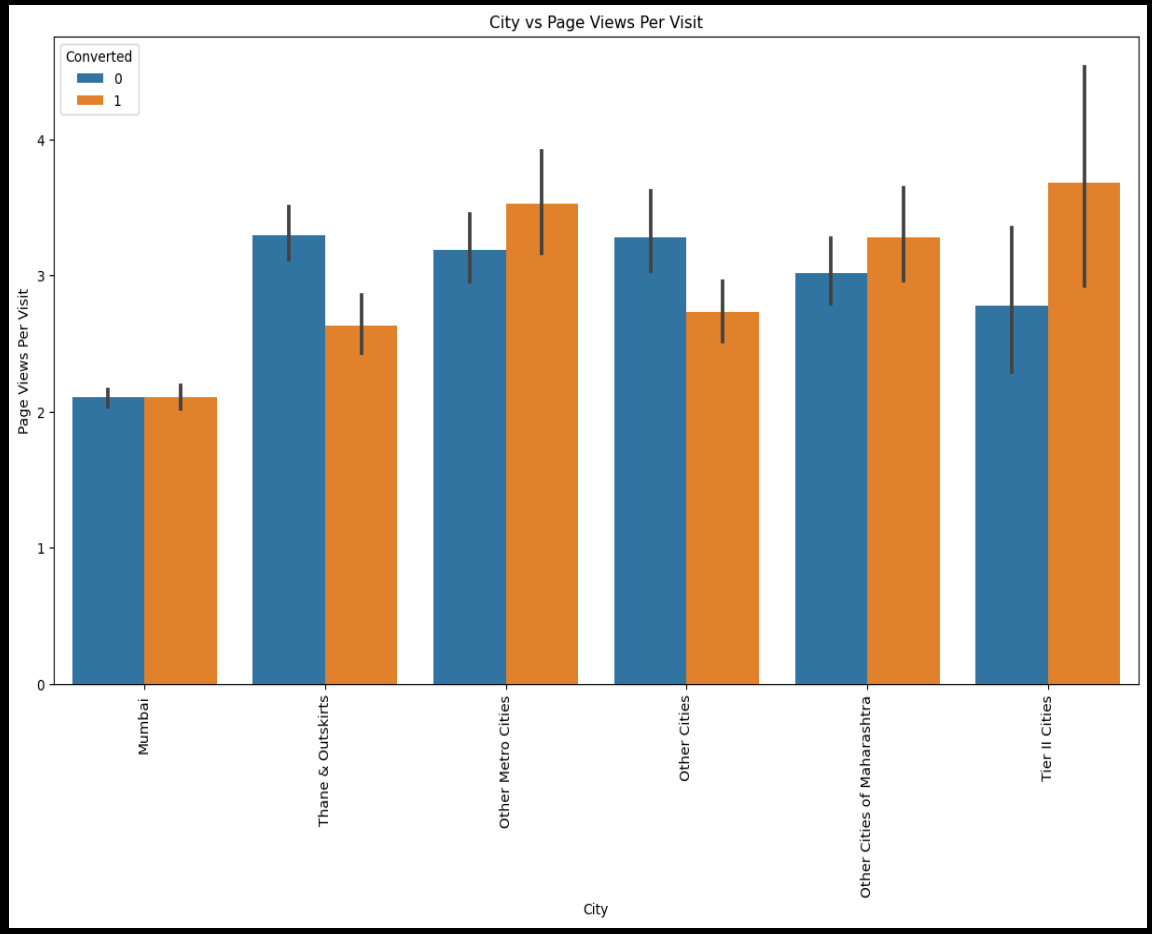

countplot for Specialization
Converted
0
1
count
3000
2500
2000
1500
1000
500
0
Finance Management
Business Administration
Media and Advertising
Supply Chain Management
IT Projects Management
Travel and Tourism
Human Resource Management
Marketing Management
Banking, Investment And Insurance
International Business
E-COMMERCE
Operations Management
Retail Management
Services Excellence
Hospitality Management
Rural and Agribusiness
Healthcare Management
E-Business
Specialization

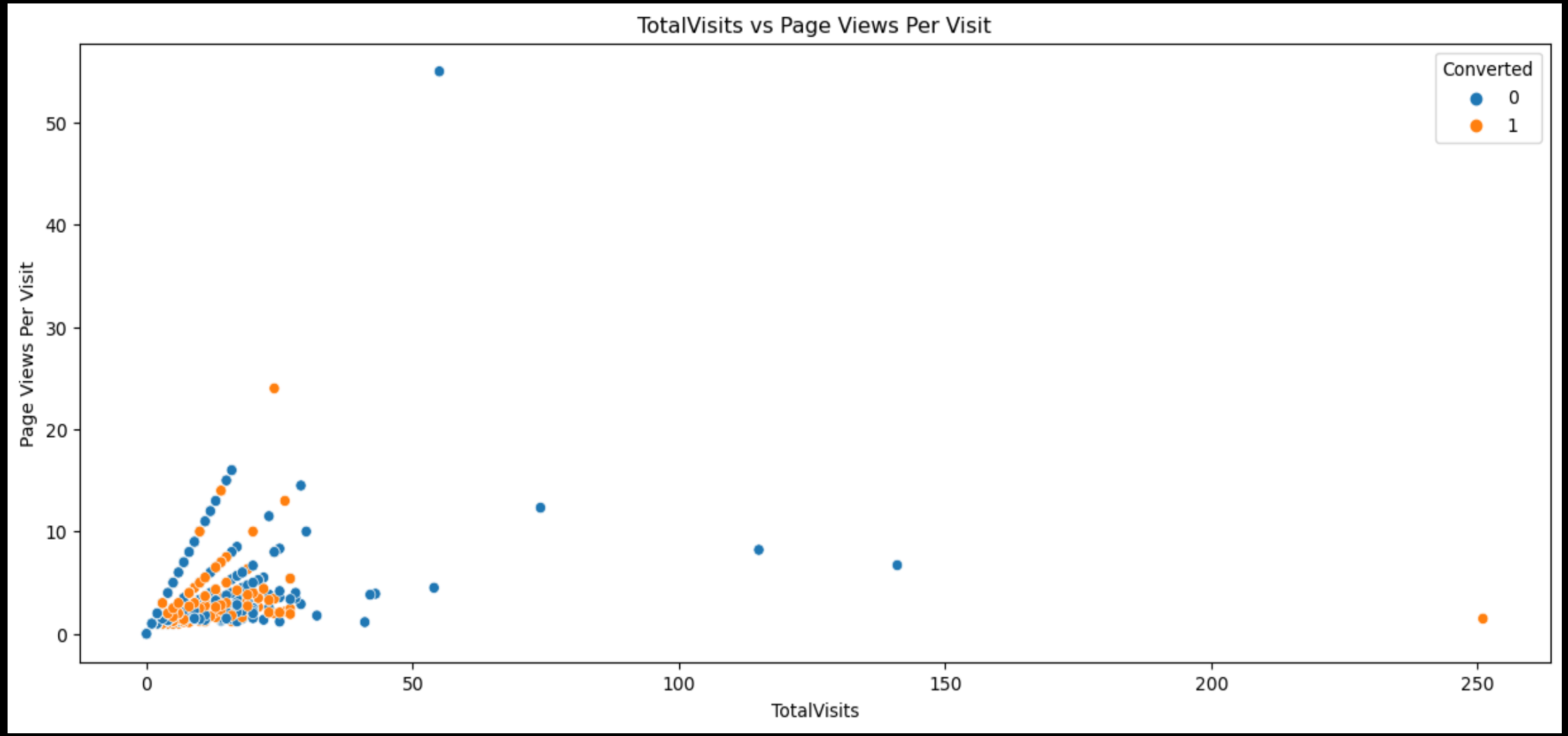

countplot for Tags
Converted
0
1
count
2500
2000
1500
1000
500
0
Interested in other courses
Ringing
Will revert after reading the email
Lost to EINS
In confusion whether part time or DLP
Busy
switched off
in touch with EINS
Already a student
Diploma holder (Not Eligible)
Graduation in progress
Closed by Horizzon
number not provided
opp hangup
Not doing further education
invalid number
wrong number given
Interested in full time MBA
Still Thinking
Lost to Others
Shall take in the next coming month
Lateral student
Interested in Next batch
Recognition issue (DEC approval)
Want to take admission but has financial problems
University not recognized
Tags

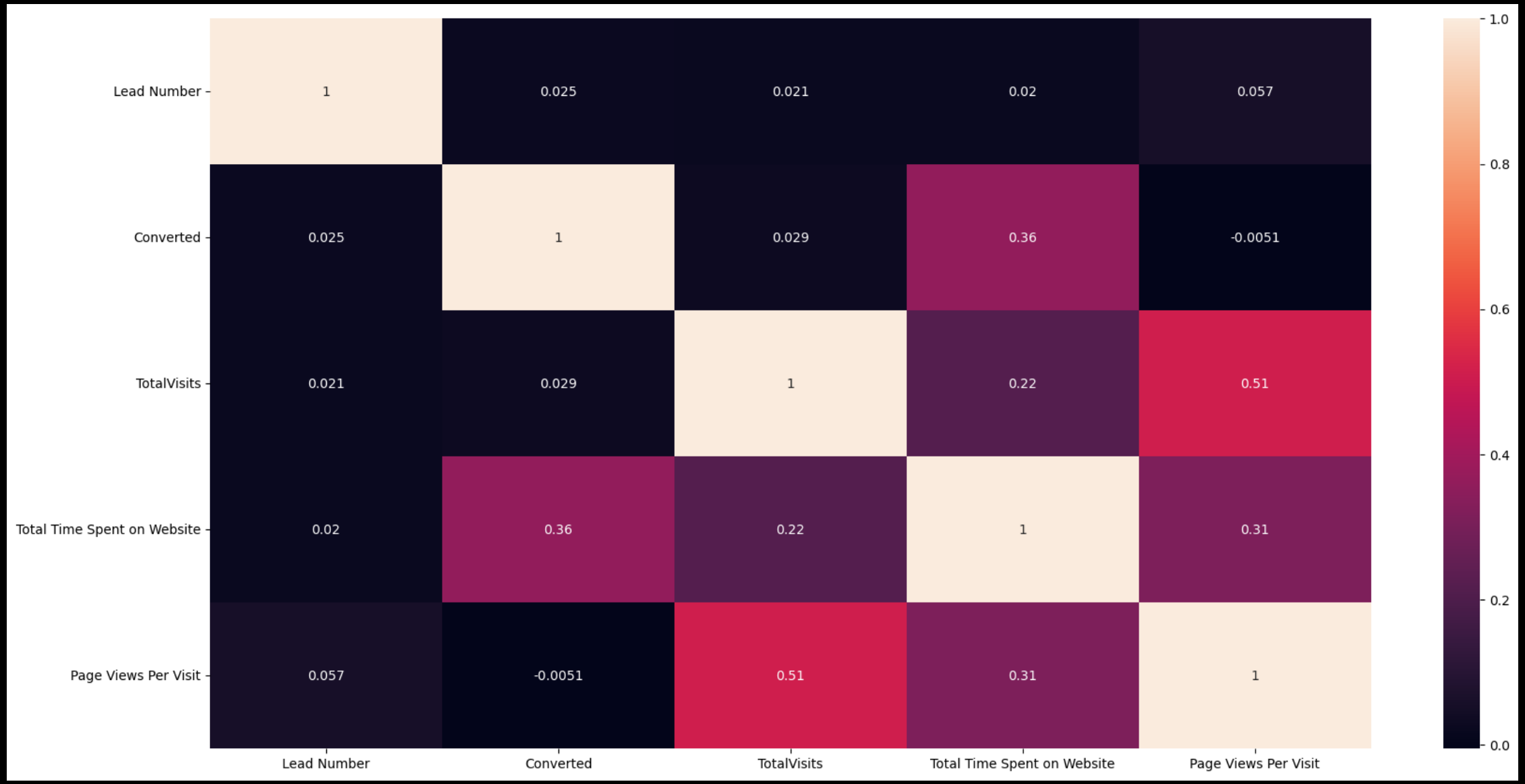

Last Notable Activity vs Page Views Per Visit
Converted
0
1
Page Views Per Visit
0
1
2
3
4
5
Modified
Email Opened
Page Visited on Website
Email Bounced
Email Link Clicked
Unreachable
Unsubscribed
Had a Phone Conversation
Olark Chat Conversation
SMS Sent
Approached upfront
Resubscribed to emails
View in browser link Clicked
Form Submitted on Website
Email Received
Email Marked Spam
Last Notable Activity


Last Notable Activity vs Total Time Spent on Website
Converted
0
1
Total Time Spent on Website
0
250
500
750
1000
1250
1500
1750
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

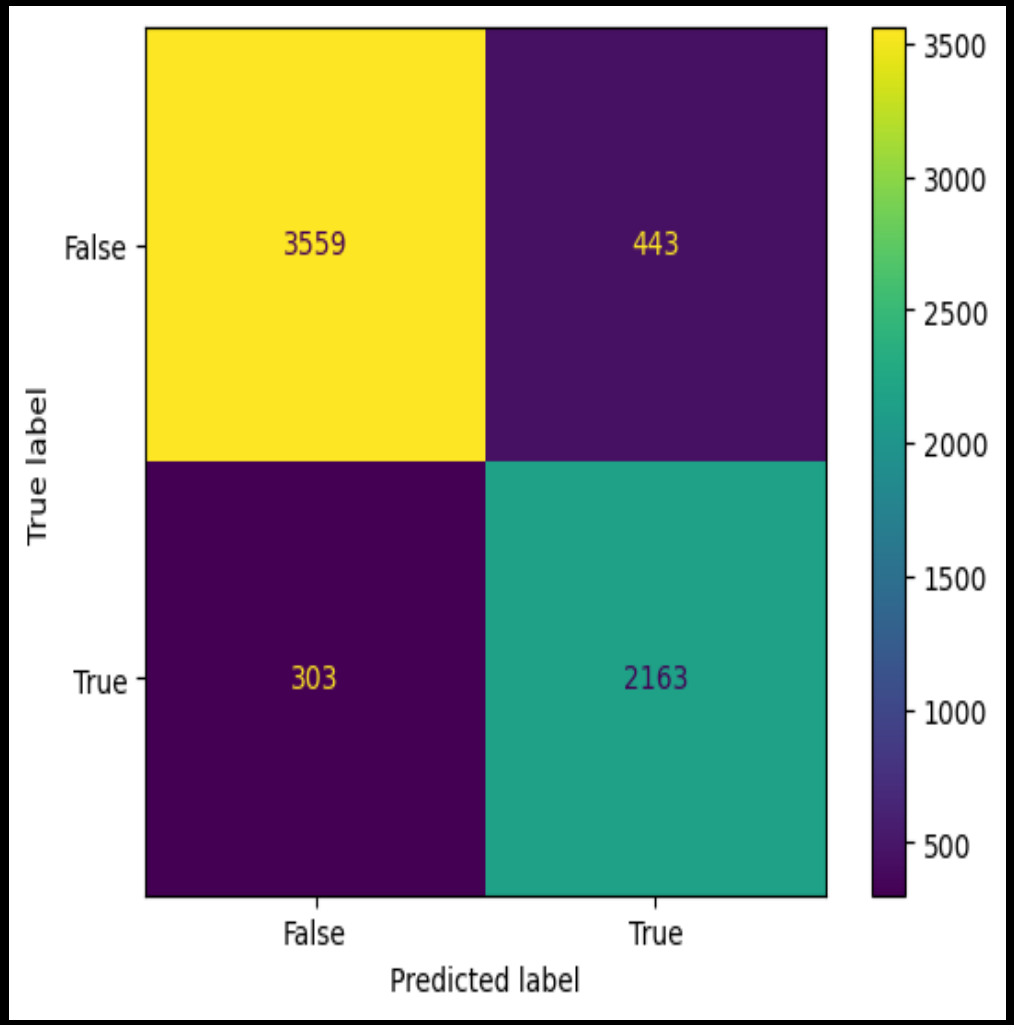

Last Notable Activity vs TotalVisits
Converted
0
1
TotalVisits
0
2
4
6
8
10
12
14
16
Modified
Email Opened
Page Visited on Website
Email Bounced
Email Link Clicked
Unreachable
Unsubscribed
Had a Phone Conversation
Olark Chat Conversation
SMS Sent
Approached upfront
Resubscribed to emails
View in browser link Clicked
Form Submitted on Website
Email Received
Email Marked Spam
Last Notable Activity


City vs Page Views Per Visit
Converted
0
1
Page Views Per Visit
0
1
2
3
4
Mumbai
Thane & Outskirts
Other Metro Cities
Other Cities
Other Cities of Maharashtra
Tier II Cities
City

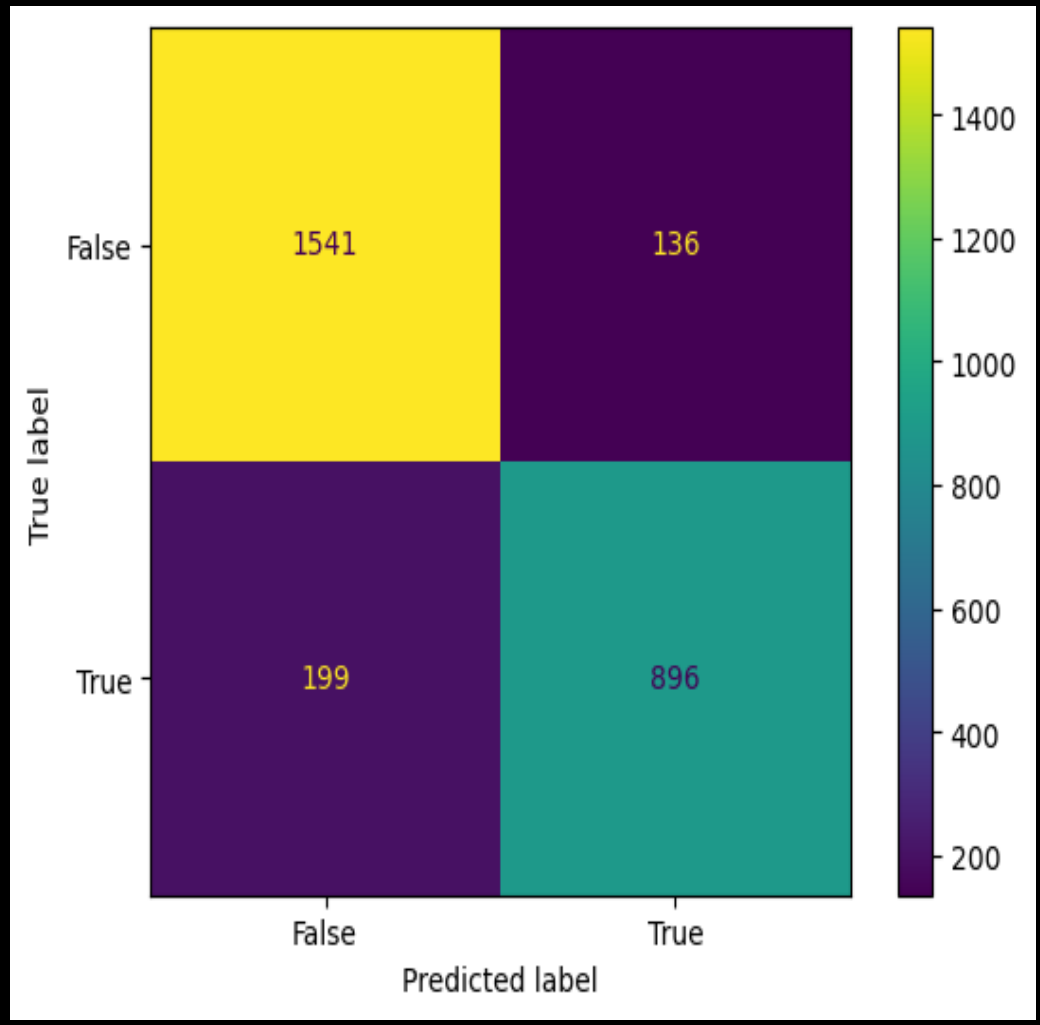

TotalVisits vs Page Views Per Visit
Converted
0
1
Page Views Per Visit
TotalVisits
0
10
20
30
40
50
0
50
100
150
200
250


Lead Number
Converted
TotalVisits
Total Time Spent on Website
Page Views Per Visit
1
0.025
0.021
0.02
0.057
0.025
1
0.029
0.36
-0.0051
0.021
0.029
1
0.22
0.51
0.02
0.36
0.22
1
0.31
0.057
-0.0051
0.51
0.31
1
Lead Number
Converted
TotalVisits
Total Time Spent on Website
Page Views Per Visit
1.0
0.8
0.6
0.4
0.2
0.0

Confusion Matrix for Train Data



Confusion Matrix for Test Data



Data Conversion

- Numerical Variables are Normalised
- Dummy Variables are created for object type variables.
- Total Rows for Analysis: 8792
- Total Columns for Analysis: 43
- With the help of feature Engineering we choose only top 10 columns

Model Building

- Splitting the Data into Training and Testing Sets.
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Use RFE for Feature Selection. Selected 10 features with RFE in order to build a model.
- Building Model by removing the variable whose p- value is greater than 0.05 and vif value is greater than 5.
- Predictions done on test data set .
- Overall accuracy approximately 88%.

Conclusion

- It was found that the variables that mattered the most in the potential buyers are:

	Features	VIF
0	const	3.586740
1	Total Time Spent on Website	1.105678
2	Lead Origin_Lead Add Form	1.261931
3	What is your current occupation_Working Profes...	1.094009
4	Tags_Busy	1.050880
5	Tags_Closed by Horizzon	1.314761
6	Tags_Lateral student	1.001699
7	Tags_Lost to EINS	1.053466
8	Tags_Will revert after reading the email	1.199001
9	Last Notable Activity_Had a Phone Conversation	1.002289
10	Last Notable Activity_SMS Sent	1.099063