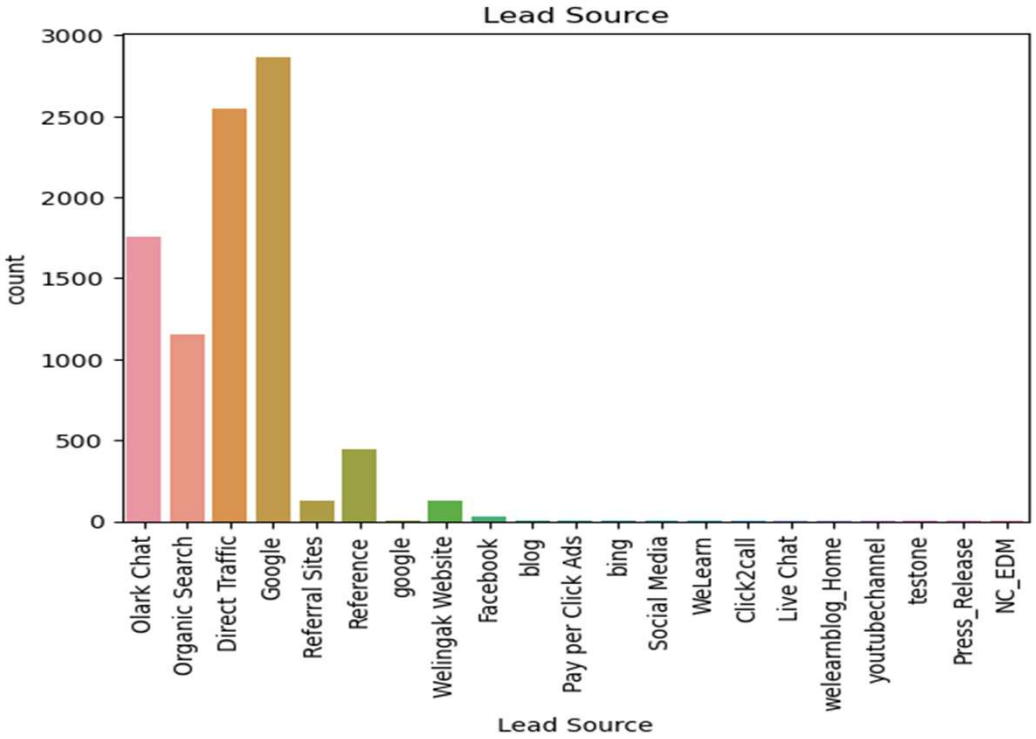
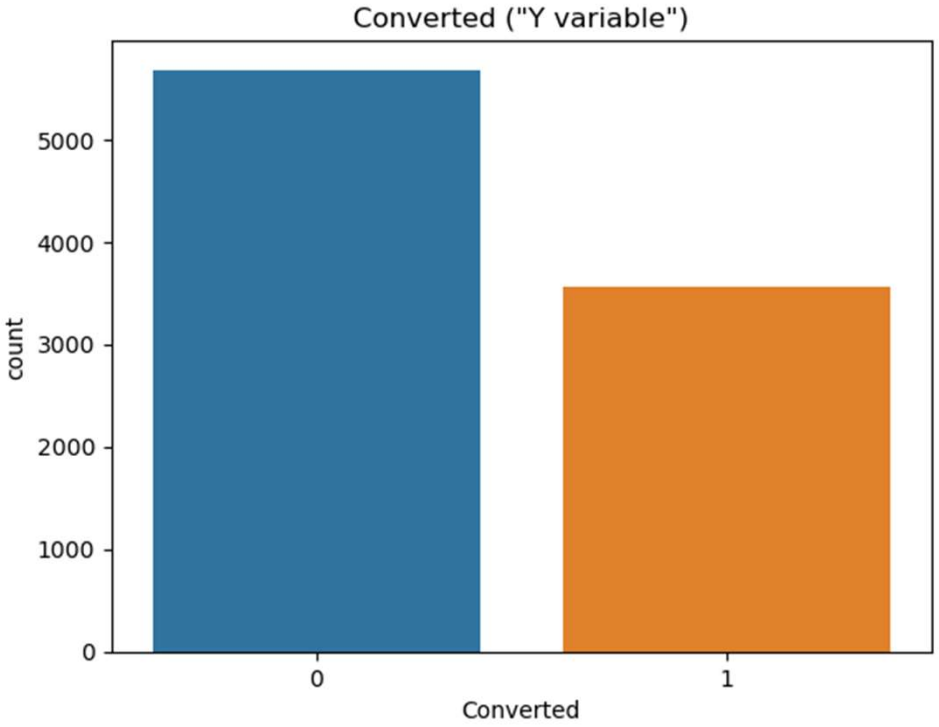
Issue at Hand:

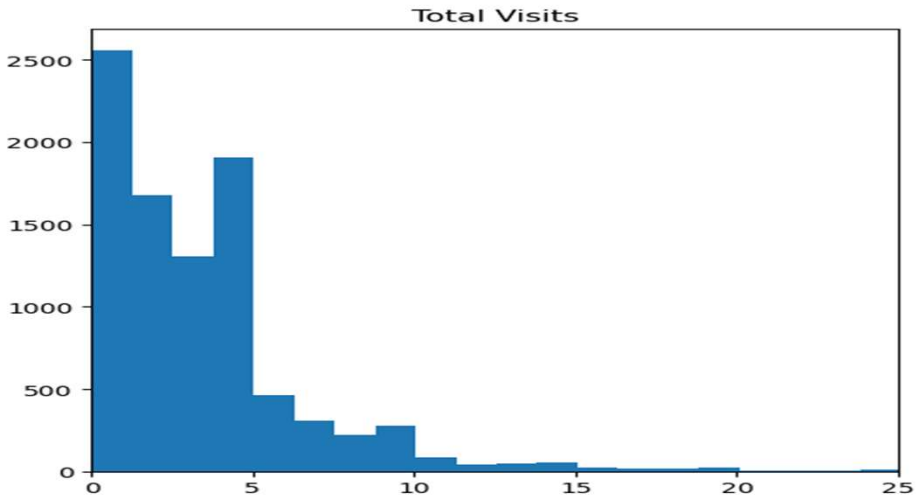
X Education seeks assistance in identifying the most promising leads—those with the highest likelihood of transitioning into paying customers. The company tasks us with constructing a model to assign a lead score to each lead, where higher scores indicate a greater chance of conversion, and lower scores suggest a lower likelihood. The CEO has set a target lead conversion rate of approximately 80%.

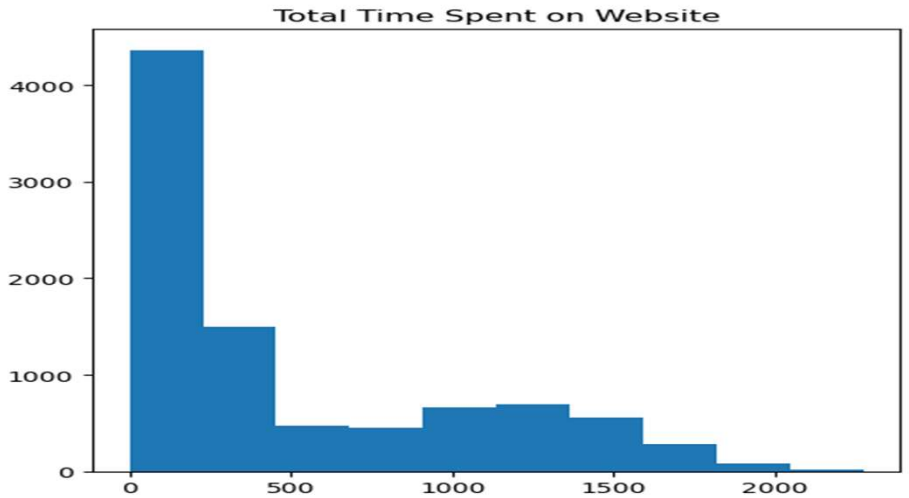
Data Cleanup : We embarked on a tidying mission, addressing null values and substituting those ambiguous 'select' options with nulls. Additionally, we categorized some nulls as 'not provided' to minimize data loss. Furthermore, we streamlined location information to 'India', 'Outside India', and 'not provided'.

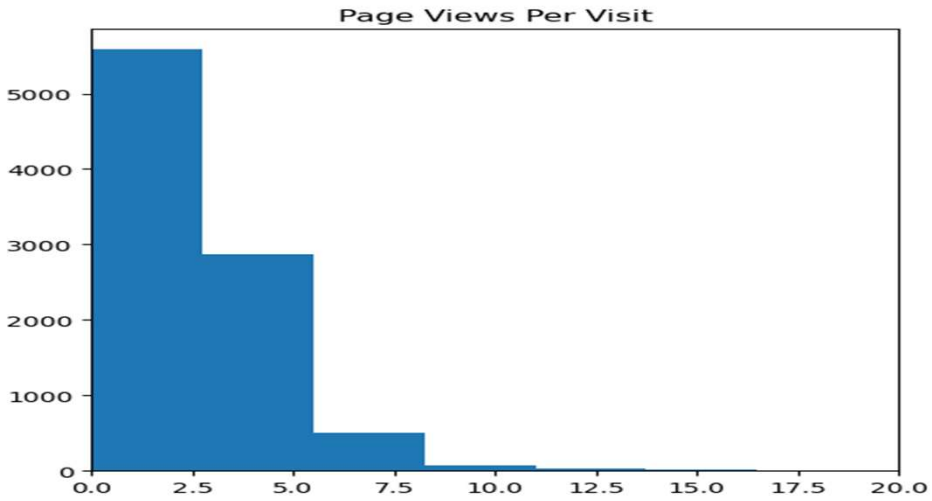
EDA

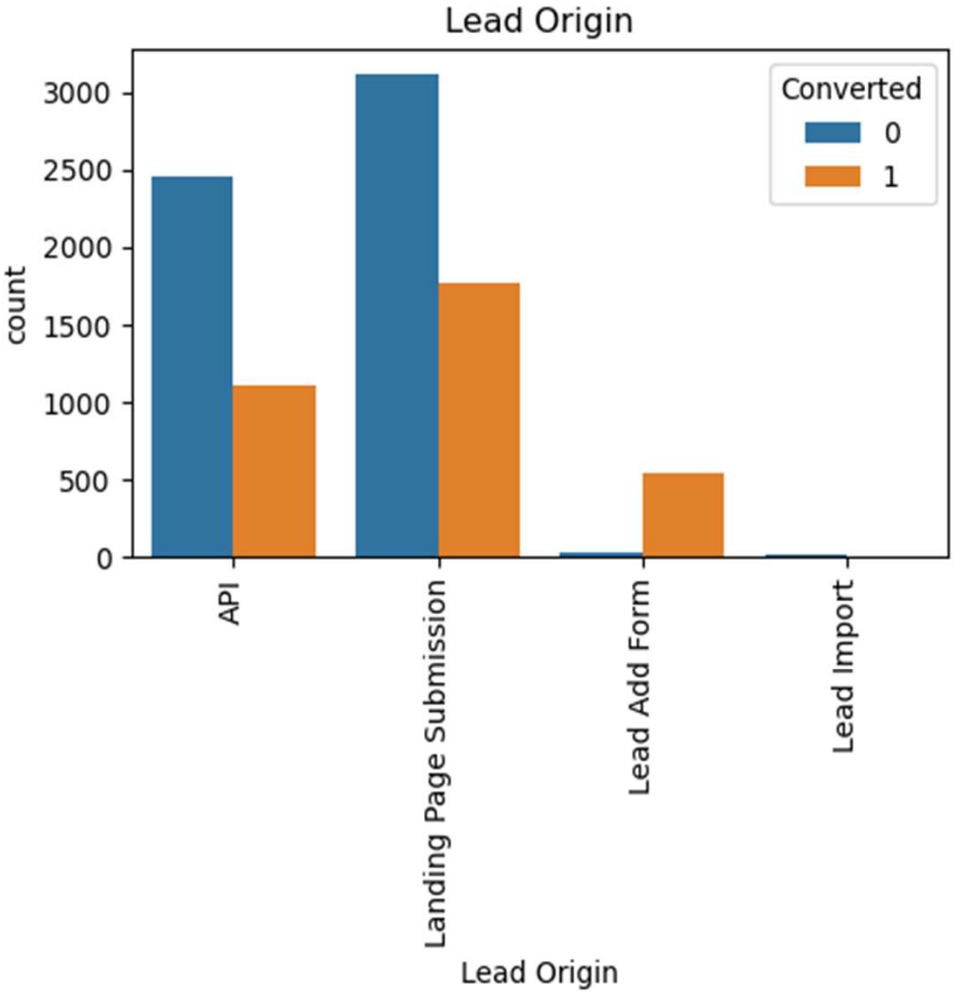

Lead Source
count
3000
2500
2000
1500
1000
500
0
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Reference
google
Welingak Website
Facebook
blog
Pay per Click Ads
bing
Social Media
WeLearn
Click2call
Live Chat
weleamblog_Home
youtubechannel
testone
Press_Release
NC_EDM
Lead Source

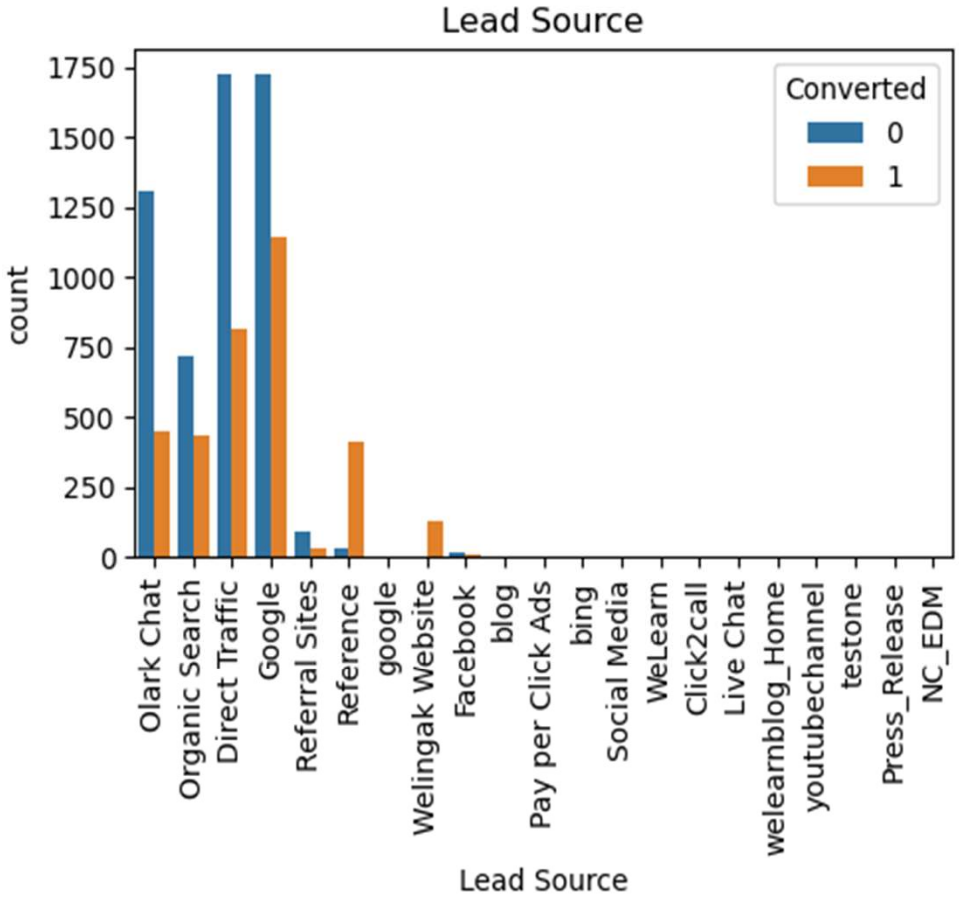

Converted ("Y variable")
count
5000
4000
3000
2000
1000
0
0
1
Converted

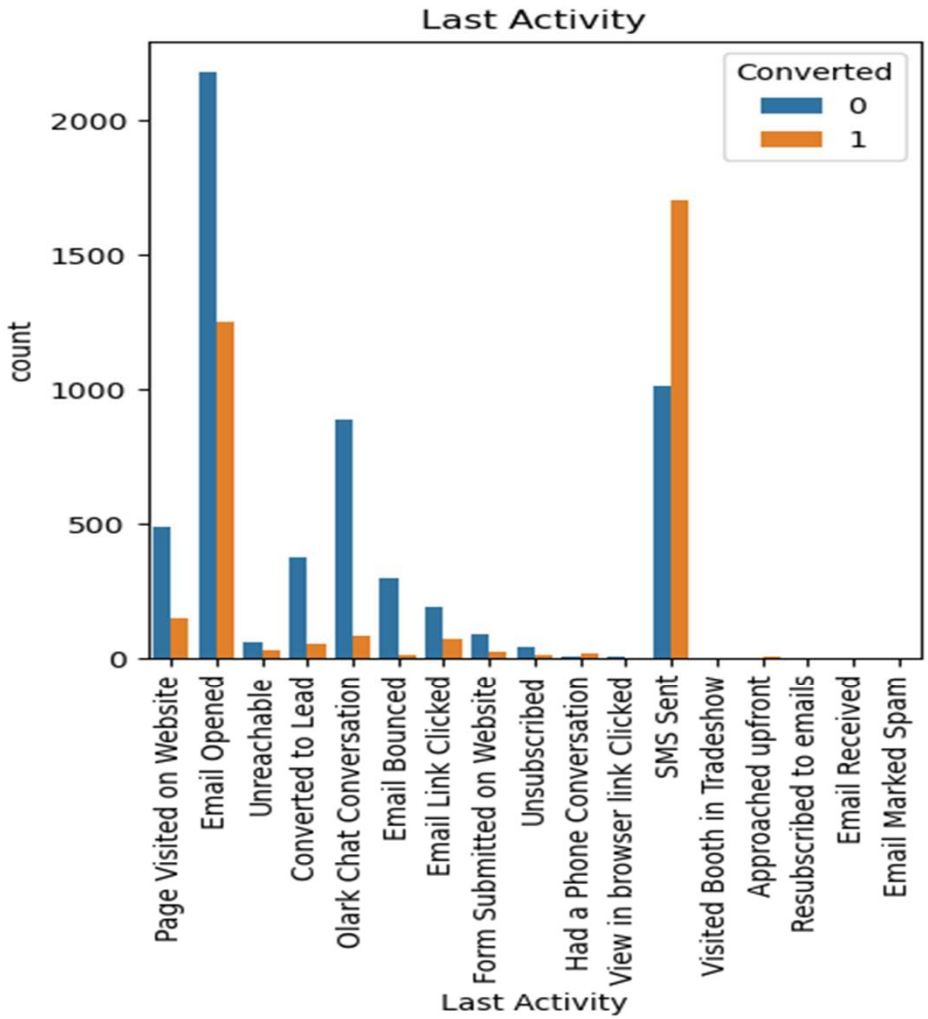

Total Visits
2500
2000
1500
1000
500
0
0
5
10
15
20
25

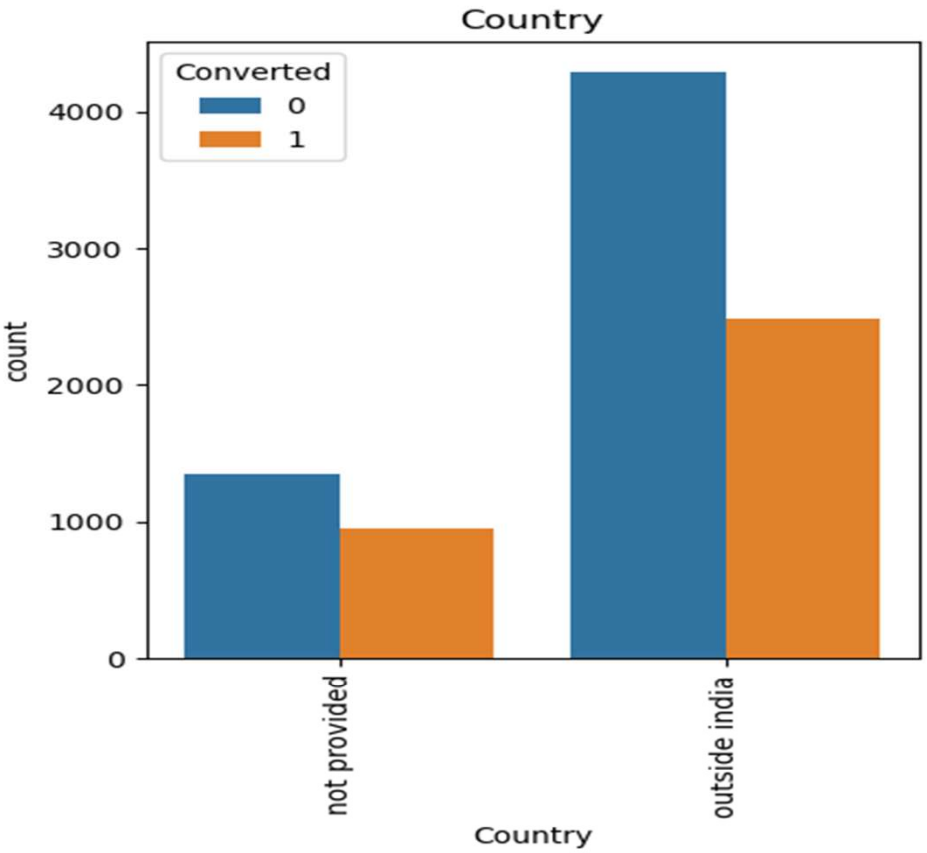

Total Time Spent on Website
4000
3000
2000
1000
0
0
500
1000
1500
2000

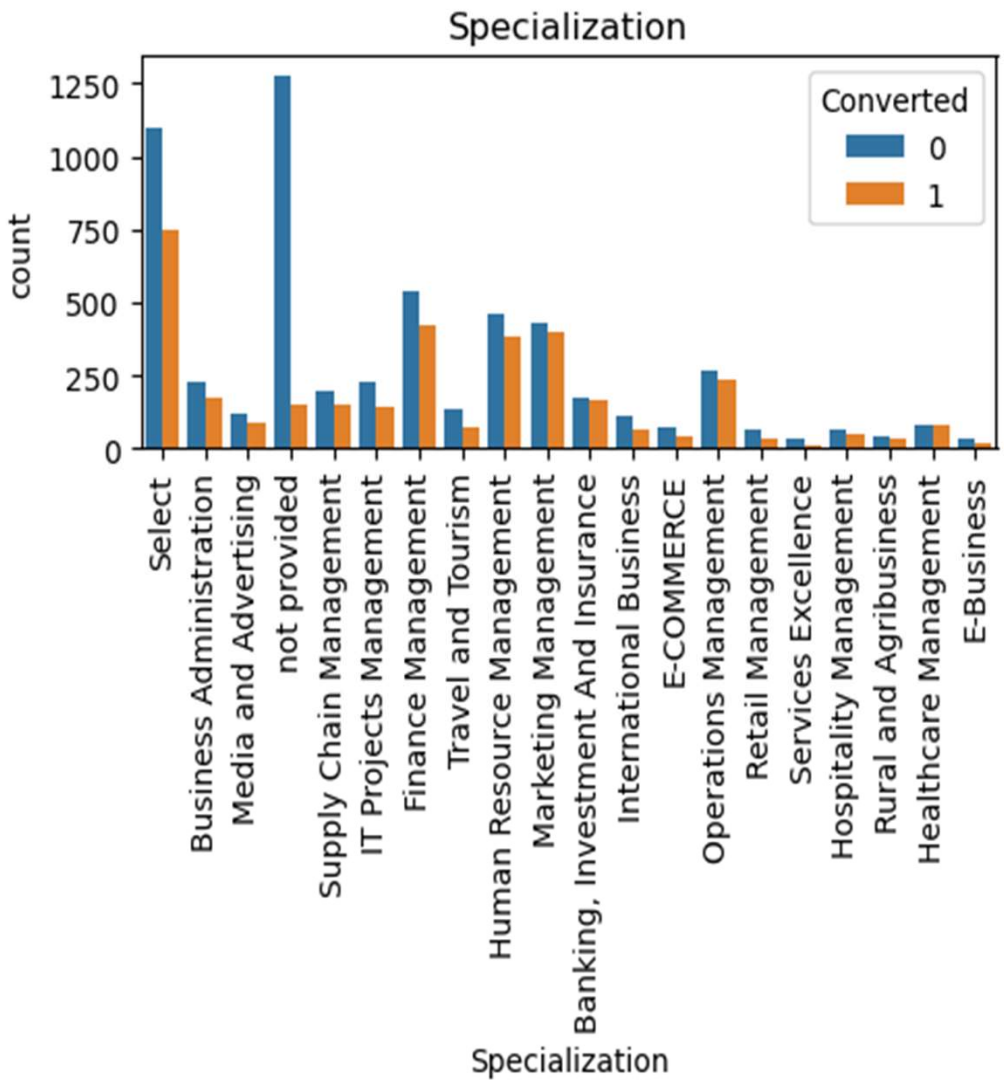

Page Views Per Visit
5000
4000
3000
2000
1000
0
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0

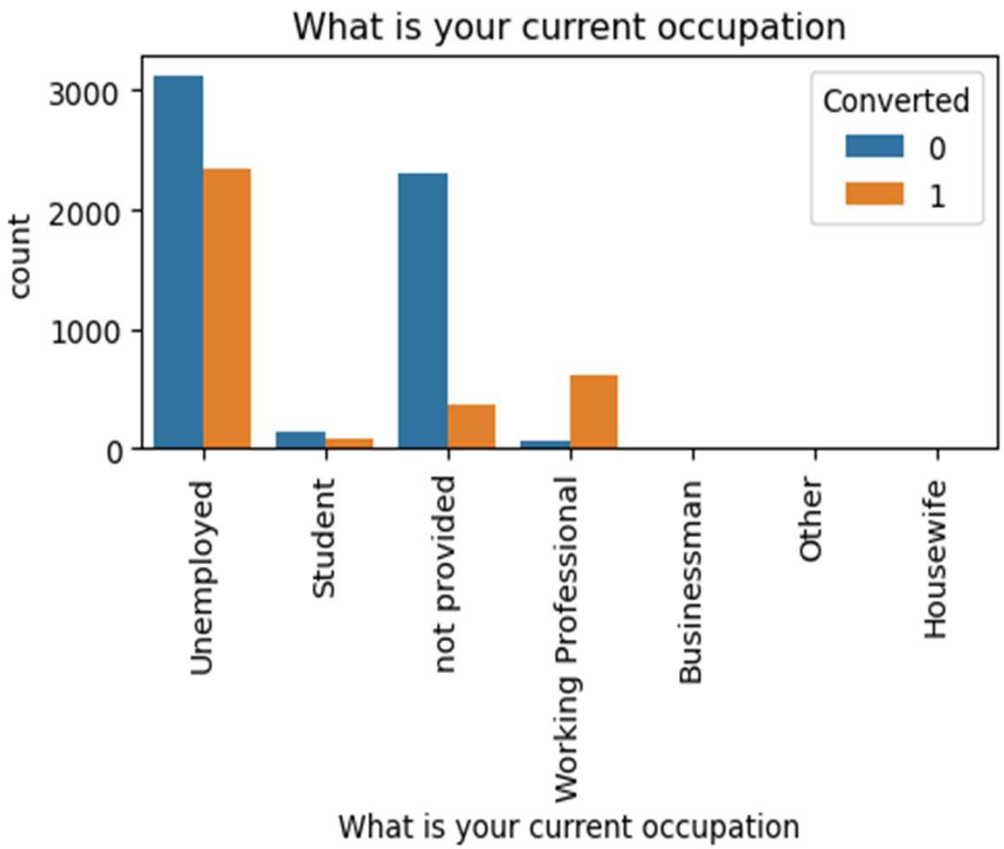

Lead Origin
Converted
0
1
count
3000
2500
2000
1500
1000
500
0
API
Landing Page Submission
Lead Add Form
Lead Import
Lead Origin

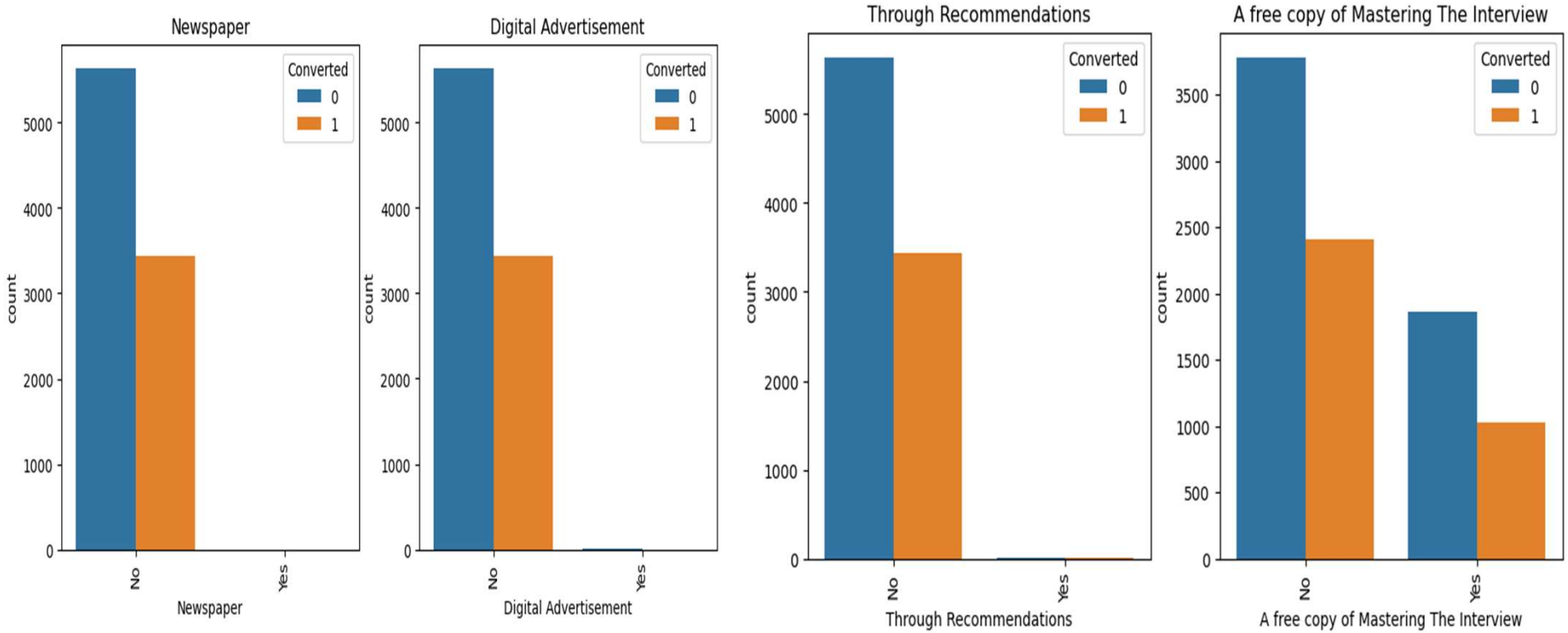

Lead Source
Converted
0
1
count
1750
1500
1250
1000
750
500
250
0
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Reference
google
Welingak Website
Facebook
blog
Pay per Click Ads
bing
Social Media
WeLearn
Click2call
Live Chat
welearnblog_Home
youtubechannel
testone
Press_Release
NC_EDM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


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
SMS Sent
Visited Booth in Tradeshow
Approached upfront
Resubscribed to emails
Email Received
Email Marked Spam
Last Activity


Country
Converted
0
1
count
4000
3000
2000
1000
0
not provided
outside india
Country


Specialization
Converted
0
1
count
1250
1000
750
500
250
0
Select
Business Administration
Media and Advertising
not provided
Supply Chain Management
IT Projects Management
Finance 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


What is your current occupation
Converted
0
1
count
3000
2000
1000
0
Unemployed
Student
not provided
Working Professional
Businessman
Other
Housewife
What is your current occupation


Newspaper
Converted
0
1
count
5000
4000
3000
2000
1000
0
No
Yes
Newspaper
Digital Advertisement
Converted
0
1
count
5000
4000
3000
2000
1000
0
No
Yes
Digital Advertisement
Through Recommendations
Converted
0
1
count
5000
4000
3000
2000
1000
0
No
Yes
Through Recommendations
A free copy of Mastering The Interview
Converted
0
1
count
3500
3000
2500
2000
1500
1000
500
0
No
Yes
A free copy of Mastering The Interview

Other Analysis :

- Dummy Variables : With 'not provided' entries out of the picture, we forged ahead and generated dummy variables. Numeric data underwent scaling using MinMaxScaler to level the playing field.
- Data Splitting : To ensure a fair assessment, we partitioned the data into 70% for training and 30% for testing purposes.
- Model Development : Employing Recursive Feature Elimination (RFE), we pinpointed the top 15 contenders. Subsequently, through manual intervention guided by Variance Inflation Factor (VIF) and p-values, we curated the final lineup, retaining variables with VIF < 5 and p-value < 0.05.
- Evaluation Metrics : Armed with a confusion matrix, we embarked on a quest for the optimal cutoff value, employing the mystical powers of ROC curves. Our endeavors culminated in achieving commendable accuracy, sensitivity, and specificity rates, all hovering around the 80% mark

Conclusion :

The heavyweight contenders influencing potential buyer behavior include:

1. Total time spent on the website.
2. Total number of visits.
3. Lead sources: - Google