Lead Score Case Study

GROUP MEMBERS

1.AMIT KUMAR

2.AMIT KUMAR SHARMA

3.AMOKSHA SHARMA

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.

Solution Methodology

Data cleaning and manipulation involves several steps:

1. Checking and handling duplicate data.
2. Handling NA values and missing values.
3. Dropping columns with a large amount of missing values that are not useful for analysis.
4. Imputing values if necessary for missing data.
5. Checking and handling outliers in the data.

Following data cleaning, Exploratory Data Analysis (EDA) is performed, which includes:

1. Univariate data analysis: examining value counts, distributions of variables, etc.
2. Bivariate data analysis: exploring correlation coefficients and patterns between variables.

Further steps include feature scaling, creation of dummy variables, and encoding of the data for classification techniques such as logistic regression, which is used for model building and prediction.

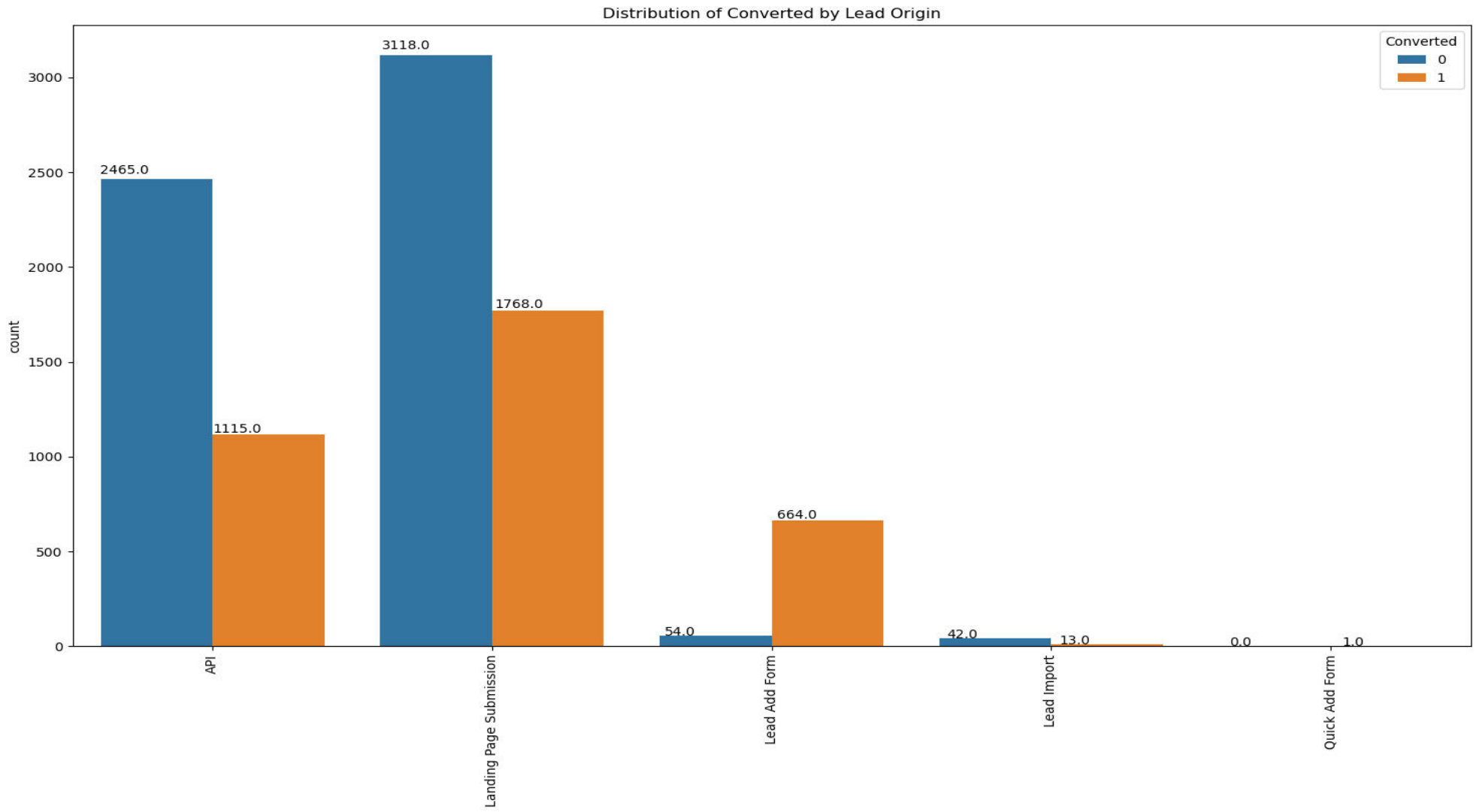
After building the model, it is validated, and the results are presented. Finally, conclusions and recommendations are drawn based on the analysis.

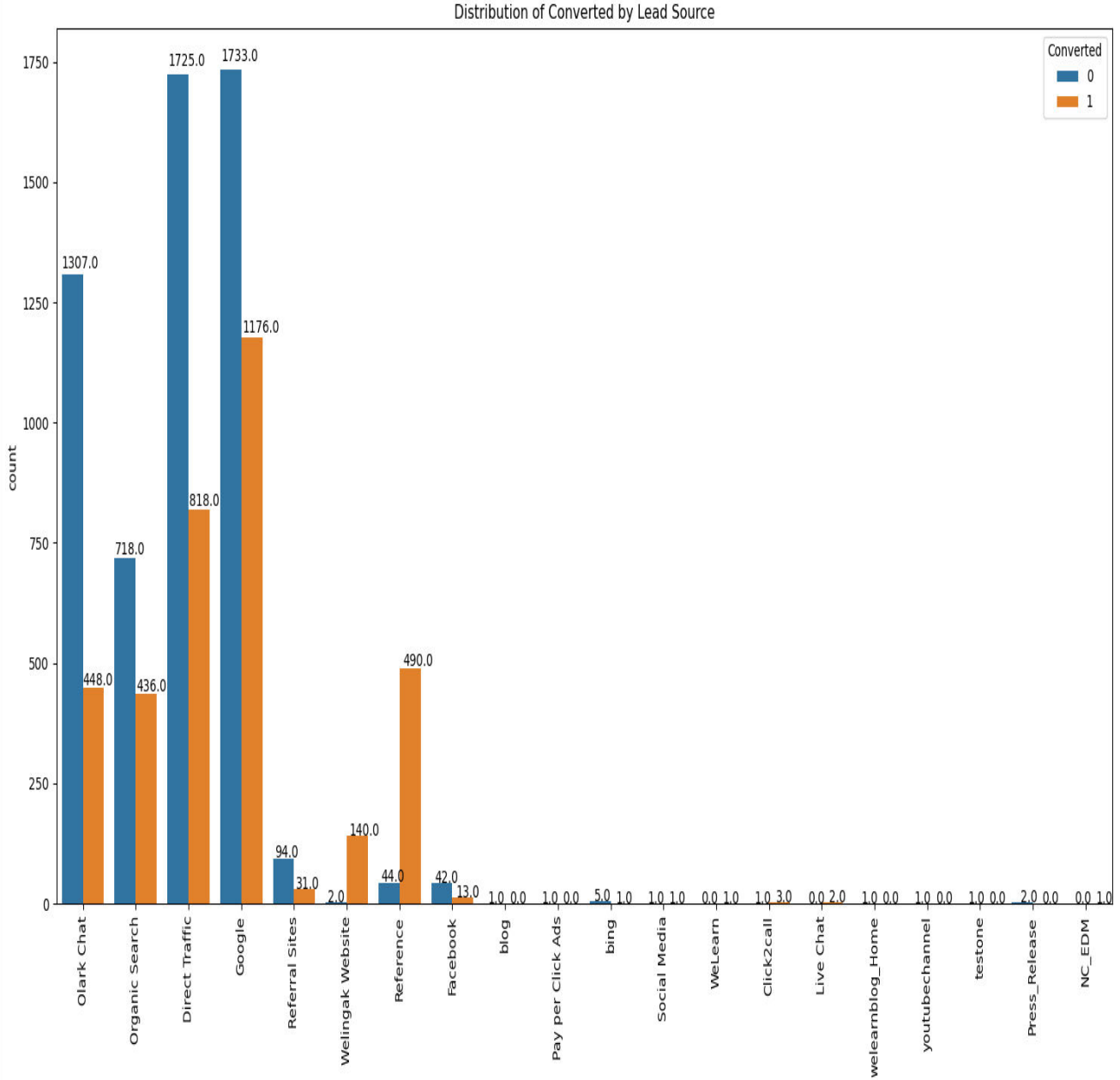
Data Manipulation

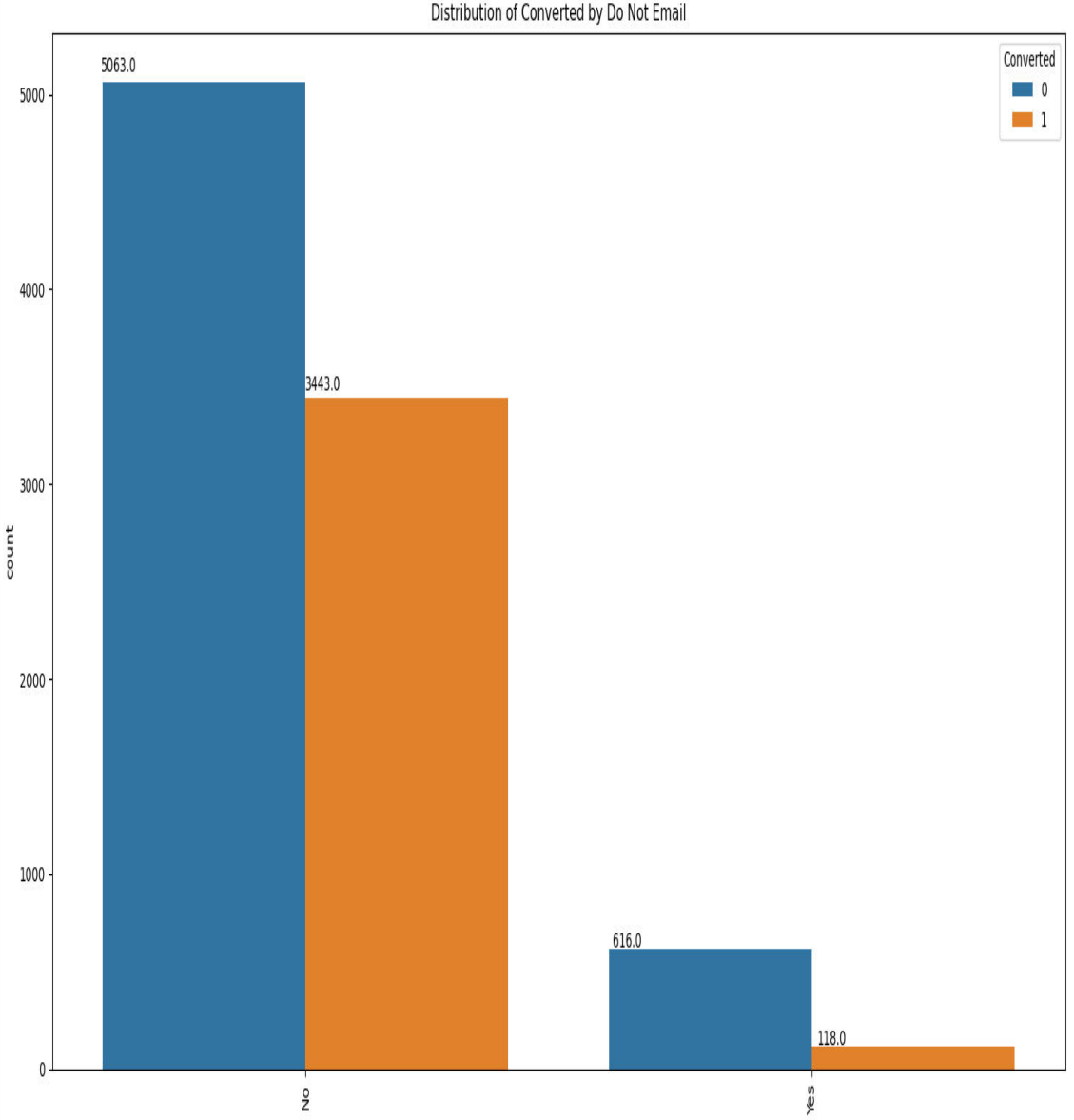
- The dataset contains 9240 rows and 37 columns.
- 'Select' values were converted to null as they are considered equivalent to null values.
- 'Prospect ID' and 'Lead Number' columns contain unique values and were dropped as they serve as unique identifiers.
- Columns with missing values exceeding 30%, except 'Specialization', were dropped for later imputation as it appears significant.
- Columns with highly imbalanced or skewed data were removed as they are unlikely to contribute significantly.
- 'Last Notable Activity' column was removed as it duplicates information in 'Last Activity' and is redundant for analysis.

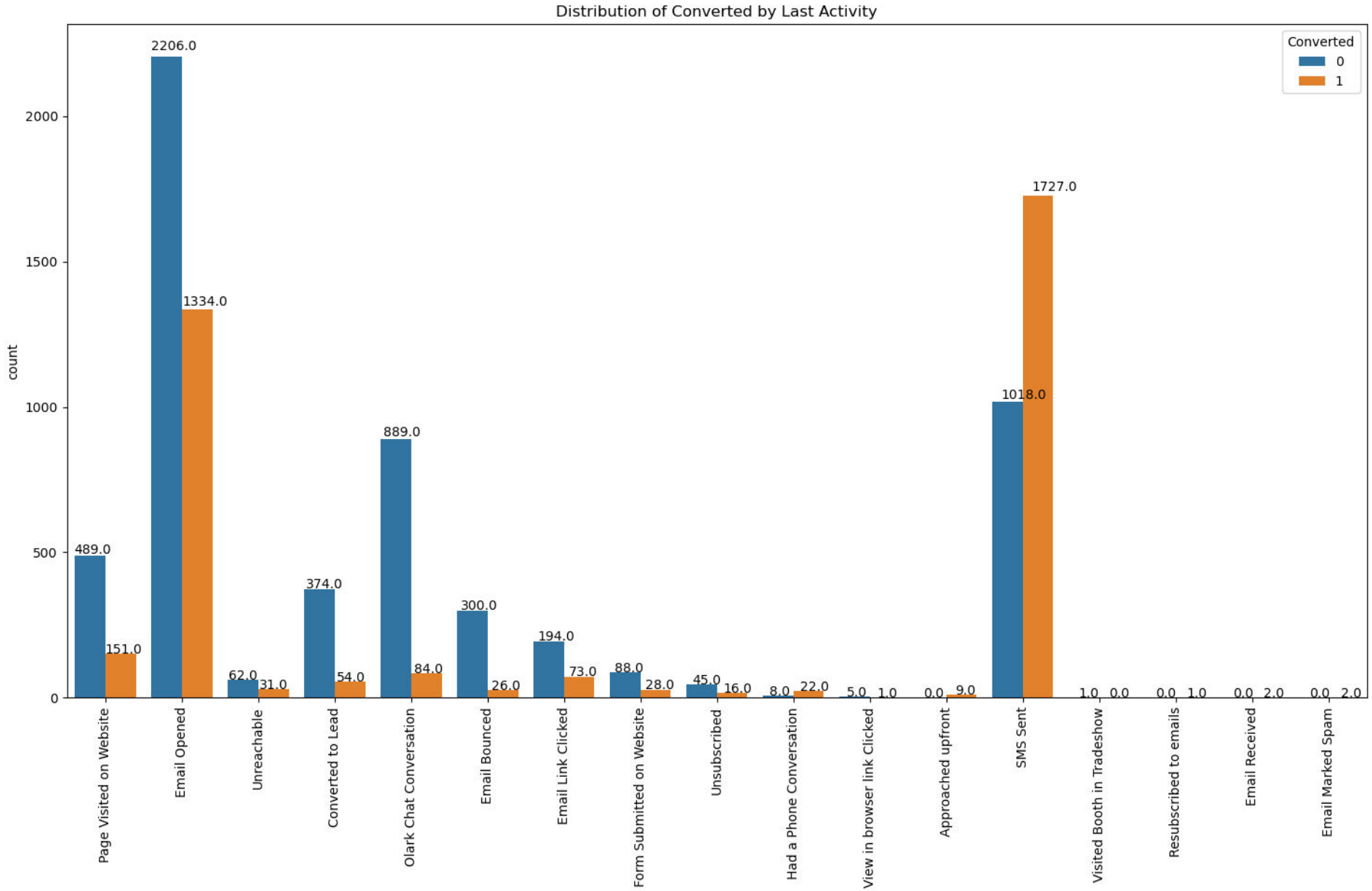
EDA

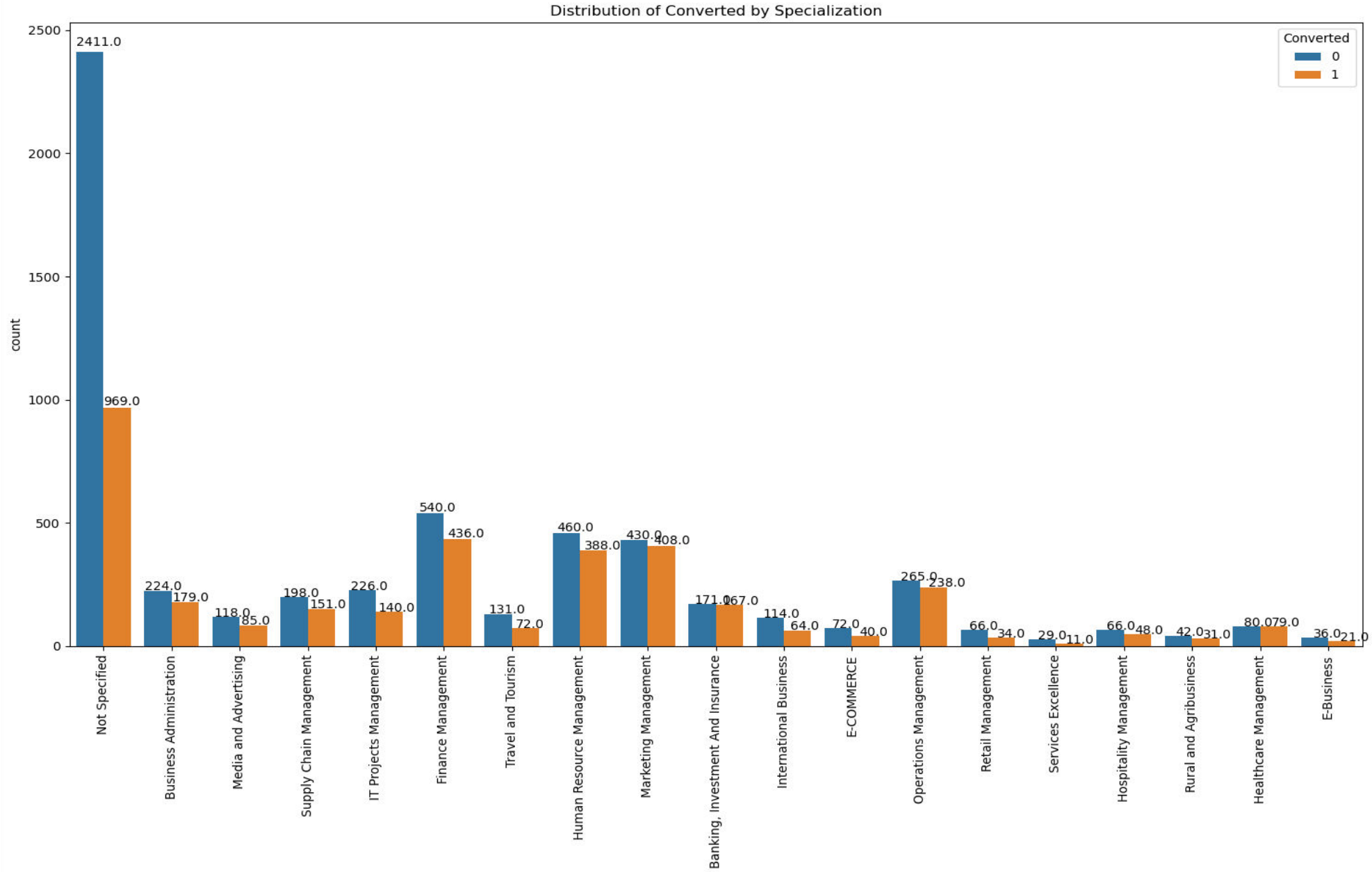
Univariate Analysis

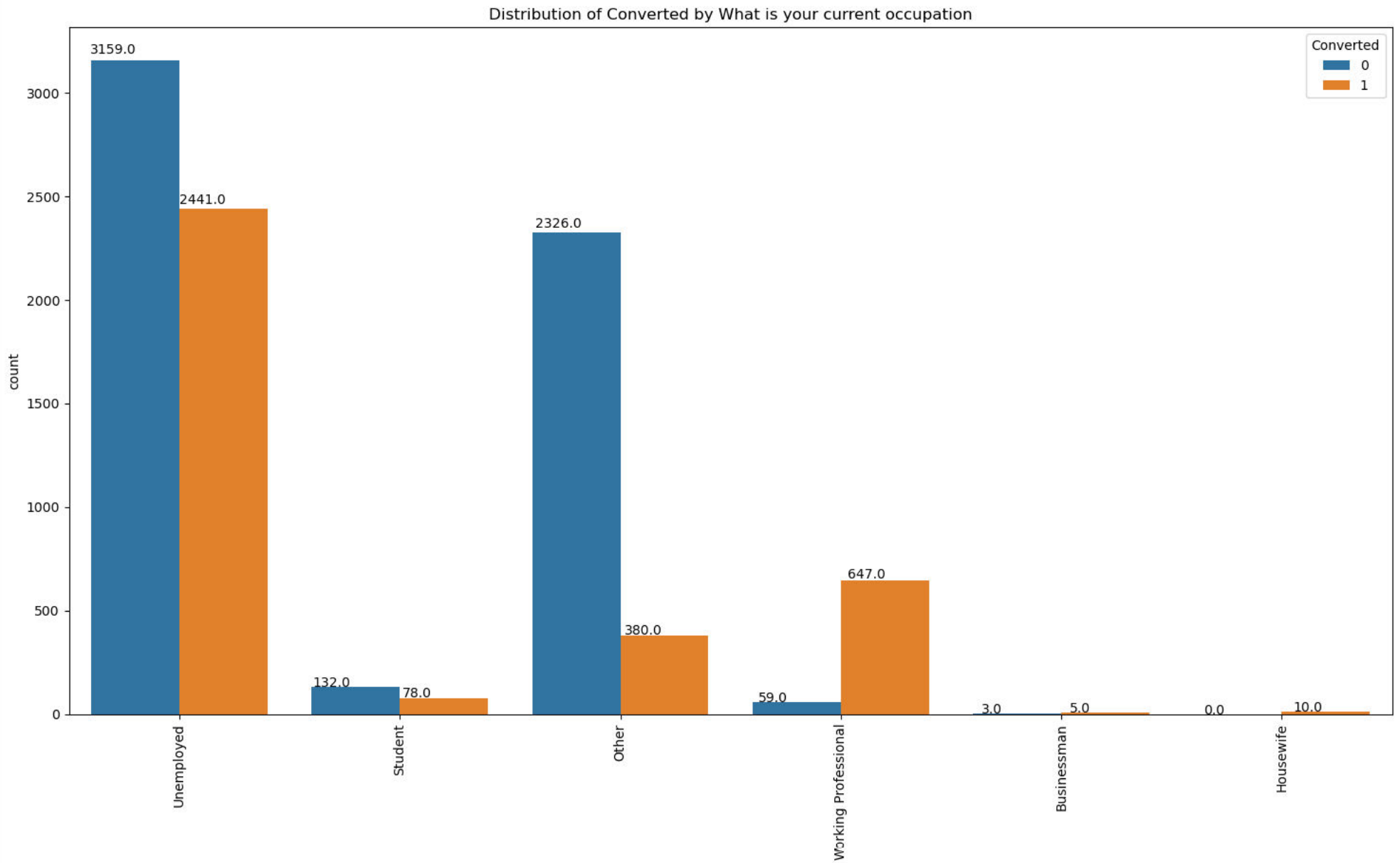



Distribution of Converted by Lead Source
Converted
0
1
count
1307.0
448.0
718.0
436.0
1725.0
818.0
1733.0
1176.0
94.0
31.0
2.0
140.0
44.0
490.0
42.0
13.0
1.0 0.0
1.0 0.0
5.0 1.0
1.0 1.0
0.0 1.0
1.0 3.0
0.0 2.0
1.0 0.0
1.0 0.0
1.0 0.0
2.0 0.0
0.0 1.0
Olark Chat
Organic Search
Direct Traffic
Google
Referral 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


Distribution of Converted by Do Not Email
Converted
0
1
count
5063.0
3443.0
616.0
118.0
No
Yes


Distribution of Converted by Last Activity
Converted
0
1
count
2000
1500
1000
500
0
2206.0
1334.0
1727.0
1018.0
889.0
489.0
151.0
374.0
54.0
300.0
26.0
194.0
73.0
88.0
28.0
62.0
31.0
45.0
16.0
8.0
22.0
5.0
1.0
0.0
9.0
1.0
0.0
0.0
1.0
0.0
2.0
0.0
2.0
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


Distribution of Converted by Specialization
Converted
0
1
count
2500
2000
1500
1000
500
0
2411.0
969.0
224.0
179.0
118.0
85.0
198.0
151.0
226.0
140.0
540.0
436.0
131.0
72.0
460.0
388.0
430.0
408.0
171.0
167.0
114.0
64.0
72.0
40.0
265.0
238.0
66.0
34.0
29.0
11.0
66.0
48.0
42.0
31.0
80.0
79.0
36.0
21.0
Not Specified
Business Administration
Media and Advertising
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


Distribution of Converted by What is your current occupation
Converted
0
1
count
3159.0
2441.0
2326.0
380.0
132.0
78.0
59.0
647.0
3.0
5.0
0.0
10.0
Unemployed
Student
Other
Working Professional
Businessman
Housewife

Bivariate Analysis

The analysis reveals significant correlations between certain variables. Notably, 'TotalVisits' and 'Page Views per Visit' exhibit a strong correlation of 0.75, indicating that as the total number of visits increases, so does the average number of page views per visit. Additionally, 'Total Time Spent on Website' shows a moderate correlation of 0.36 with the target variable 'Converted'. This suggests that increased time spent on the website may have a positive impact on conversion rates, although the relationship is not as strong as the correlation observed between 'TotalVisits' and 'Page Views per Visit'.

EDA Observations

1. The conversion rate for 'API' stands at approximately 31%, while for 'Landing Page Submission', it is around 36%. Additionally, 'Lead Add Form' shows a higher number of conversions compared to unsuccessful ones, and 'Lead Import' has a lower count.

2. To enhance the overall lead conversion rate, the focus should be on improving the conversion of leads originating from API and Landing Page Submission sources. Additionally, efforts should be made to generate more leads from the Lead Add Form.

3. The plot indicates that Google and Direct traffic generate the maximum number of leads. Moreover, the conversion rate for 'Reference' and 'Welingak Website' leads is high. To improve the overall lead conversion rate, emphasis should be placed on improving lead conversion from Olark Chat, organic search, direct traffic, and Google leads, while also increasing leads from reference and Welingak website.

4. The plot further highlights that people who opt for the mail option are becoming more leads.

5. Analysis of the plot reveals that the conversion rate for the last activity of 'SMS Sent' is approximately 63%, with 'Email Opened' being the highest last activity among leads.

6. Among different occupations, 'Unemployed' leads generate a higher number of leads and have a conversion rate of approximately 45%, while the conversion rate is higher for 'Working Professionals'.

7. The count plot of 'Specialization' shows that 'Management' specialization generates the highest number of leads overall, with the 'Other' category also contributing significantly to lead generation.

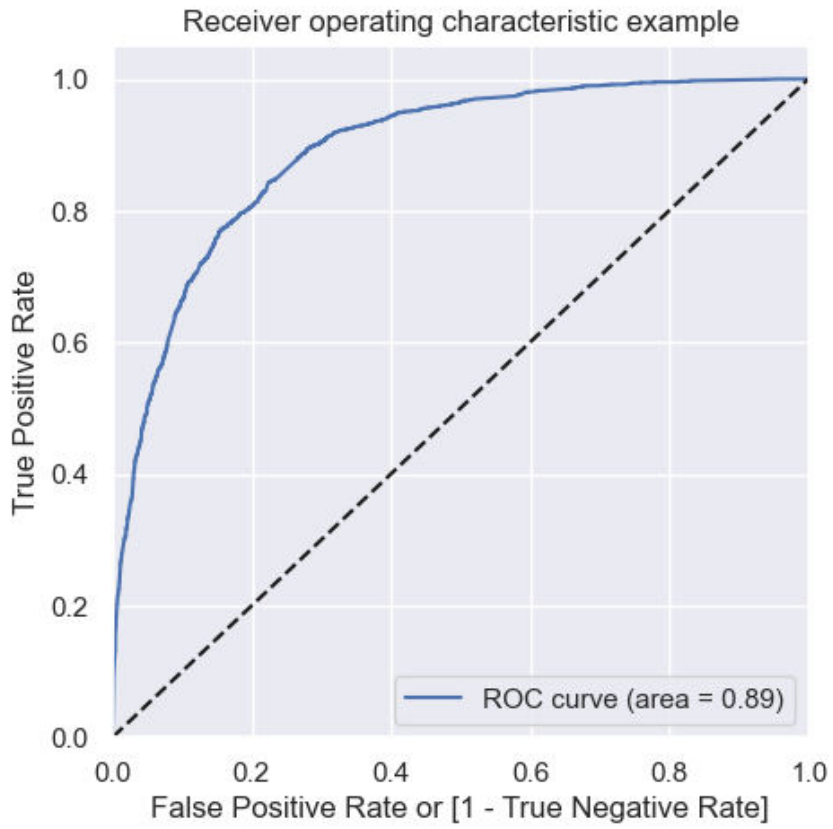
Data Preparation

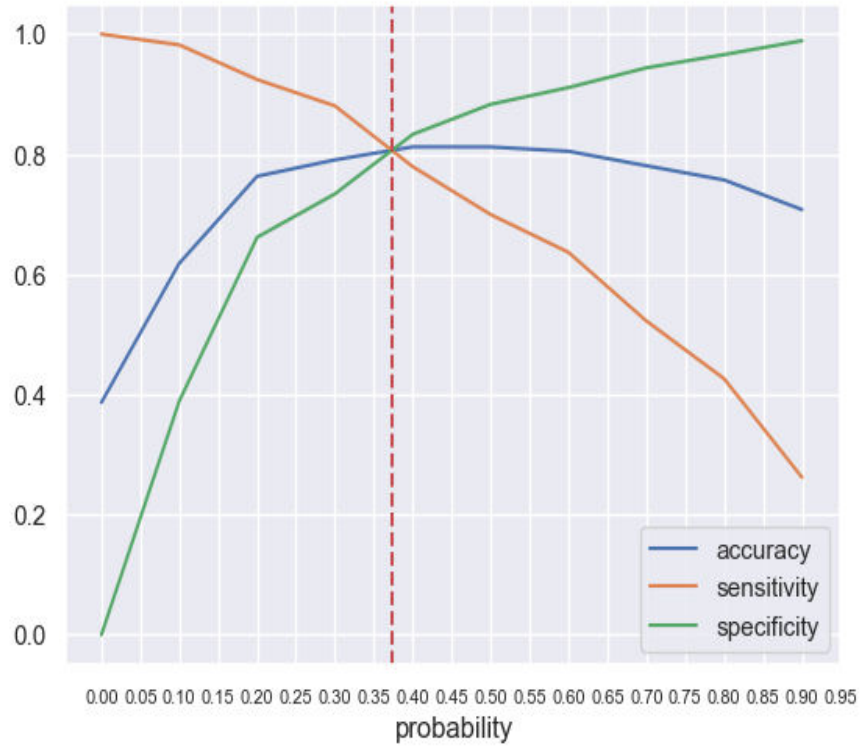
- Numerical variables have been normalized.
- Dummy variables have been created for object type variables.
- Total rows for analysis: 9090 .
- Total columns for analysis: 68.

Model Building

- The data has been split into training and testing sets, with a 70:30 ratio.
- Recursive Feature Elimination (RFE) has been used for feature selection, resulting in 15 variables as output.
- A model has been built by removing variables with p-values greater than 0.05 and VIF values greater than 5.
- Predictions have been made on the test dataset, achieving an overall accuracy of 80%.

ROC CURVE





- The optimal cut-off point is the probability at which we achieve a balance between sensitivity and specificity.
- Based on the second graph, the optimal cut-off point is observed to be at 0.374.

Conclusion

- To improve the potential lead conversion rate, X-Education should focus on important features that significantly influence conversion rates:
- Lead Source_Welingak Website: The conversion rate is higher for leads originating from the 'Welingak Website'. Thus, the company should concentrate on optimizing this website to attract more potential leads.
- Lead Origin_Lead Add Form: Leads generated through the 'Lead Add Form' exhibit a higher conversion rate. Therefore, X-Education should prioritize strategies to capture more leads through this channel.
- What is your current occupation_Working Professional: Leads identified as 'Working Professionals' have a higher likelihood of conversion. X-Education should target this demographic to increase the number of potential leads.
- Last Activity_SMS Sent: Leads with the last activity recorded as SMS Sent tend to have a higher conversion rate. The company should emphasize SMS marketing to engage potential leads effectively.
- Total Time Spent on Website: Leads spending more time on the website demonstrate a higher propensity to convert. Hence, X-Education should focus on improving website engagement to attract and retain potential leads.