Lead Scoring Case Study

By - Sanjay Yadav

Business Understanding

X Education is an education company that provides online courses to professionals in various industries. They attract a significant number of potential customers who visit their website to browse the available courses. When these visitors provide their email address or phone number on the website, they are considered as leads. Additionally, X Education also receives leads through referrals from their past customers. The sales team then proceeds to contact the leads through phone calls and emails, among other methods, in an attempt to convert them into paying customers. However, only a fraction of the leads get converted, with the typical conversion rate standing at 40%.

X Education has conducted an analysis to identify the characteristics of leads that are more likely to convert into paying customers. With this knowledge, they aim to increase their conversion rate to 80%.

Data Summary

We have been provided with a leads dataset from the past with around 9000 data points. This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. which may or may not be useful in ultimately deciding whether a lead will be converted or not. The target variable, in this case, is the column ‘Converted’ which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn’t converted.

Objective

Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads.

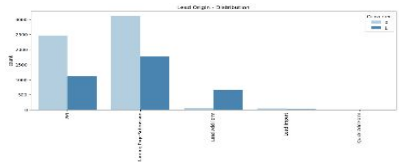
A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

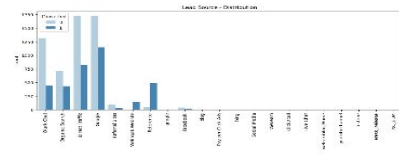
EDA

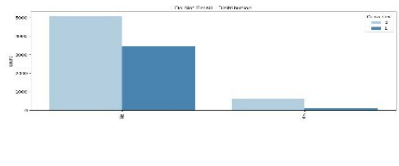
Steps Performermed:

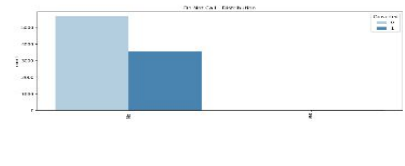
Data cleaning - removing columns with more than 30% missing values, replacing values with median for numerical columns and mode for categorical columns, checking for proper data types, making the columns positive with numerical attributes for proper analysis, replacing features with less than 100 value counts by ‘others’ category in particular feature.

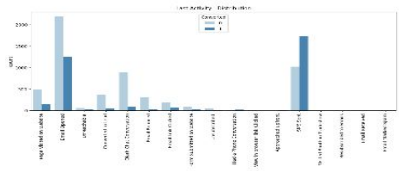
Distribution of categorical features

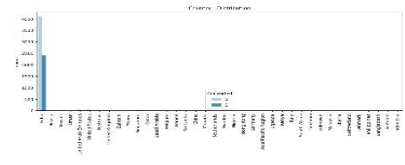

Lead Origin - Distribution
Converted
0
1
API
Landing Page Submission
Lead Add Form
Lead Import
Quick Add Form

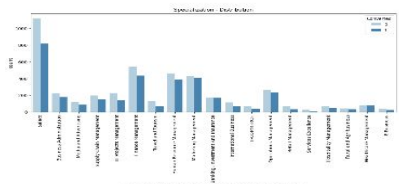

Lead Source - Distribution
Converted
0
1
Google
Direct Traffic
Olark Chat
Organic Search
Reference
Welingak Website
Referral Sites
Facebook
bing
Click2call
Press_Release
Social Media
Live Chat
youtubechannel
testone
Pay per Click Ads
welearnblog_Home
WeLearn
blog
NC_EDM

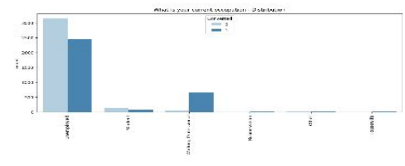

Do Not Email - Distribution
Converted
0
1
No
Yes

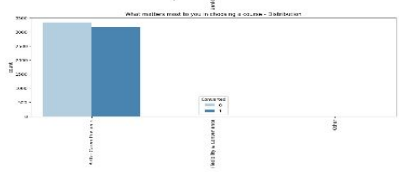

Do Not Call - Distribution
Converted
0
1
No
Yes

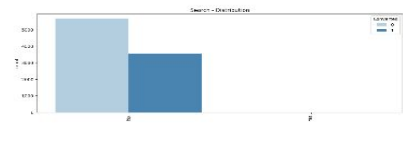

Last Activity - Distribution
Converted
0
1
Email Bounced
Email Link Clicked
Email Marked Spam
Email Opened
Form Submitted on Website
Had a Phone Conversation
Olark Chat Conversation
Page Visited on Website
Resubscribed to emails
SMS Sent
Unreachable
Unsubscribed
View in browser link Clicked

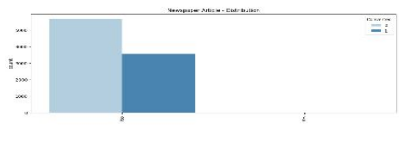

Country - Distribution
Converted
0
1
Australia
Bahrain
Bangladesh
Belgium
Canada
China
Denmark
Germany
Ghana
Hong Kong
India
Indonesia
Italy
Kenya
Kuwait
Malaysia
Netherlands
Nigeria
Oman
Philippines
Qatar
Russia
Saudi Arabia
Singapore
South Africa
Sri Lanka
Sweden
Switzerland
Tanzania
Uganda
United Arab Emirates
United Kingdom
United States
Vietnam
unknown

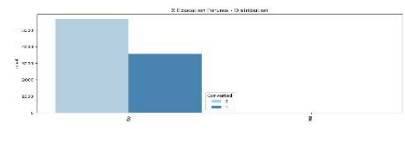

Specialization - Distribution
Converted
0
1
Banking, Investment And Insurance
Business Administration
E-COMMERCE
E-Business
Finance Management
Healthcare Management
Hospitality Management
Human Resource Management
IT Projects Management
International Business
Marketing Management
Media and Advertising
Operations Management
Retail Management
Rural and Agribusiness
Services Excellence
Supply Chain Management
Tourism and Travel Management
Travel and Leisure
unknown

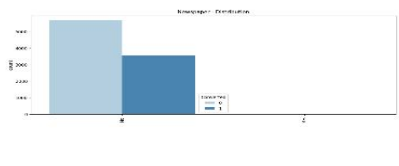

What is your current occupation - Distribution
Converted
0
1
Unemployed
Working Professional
Student
Businessman
Other
Housewife

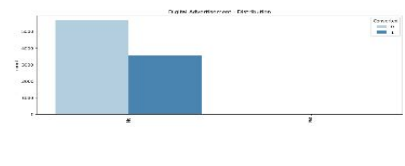

What matters most to you in choosing a course - Distribution
Converted
0
1
Better Career Prospects
Flexibility & Convenience
Other

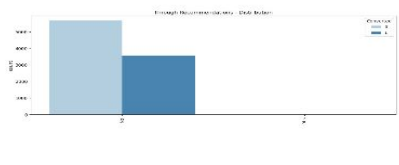

Search - Distribution
Converted
0
1
No
Yes

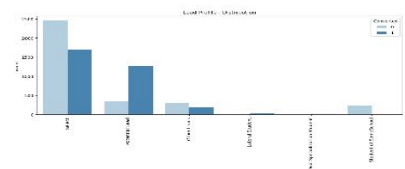

Newspaper Article - Distribution
Converted
0
1
No
Yes

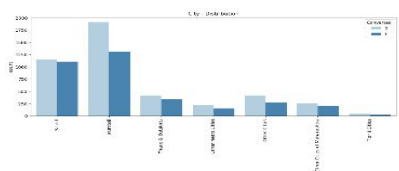

X Education Forums - Distribution
Converted
0
1
No
Yes

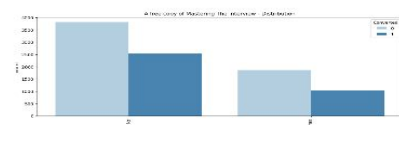

Newspaper - Distribution
Converted
0
1
No
Yes

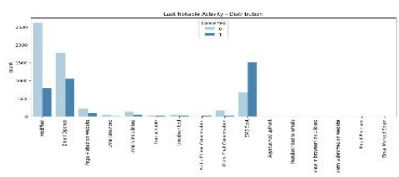

Digital Advertisement - Distribution
Converted
0
1
No
Yes


Through Recommendations - Distribution
Converted
0
1
No
Yes

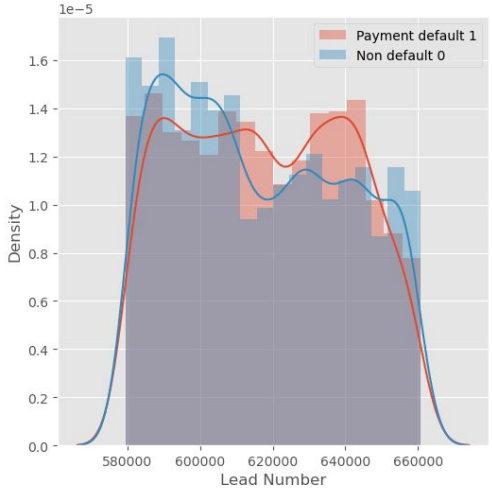

Lead Profile - Distribution
Converted
0
1
Dual Specialization Student
Lateral Student
Potential Lead
Student of SomeSchool
Top Student

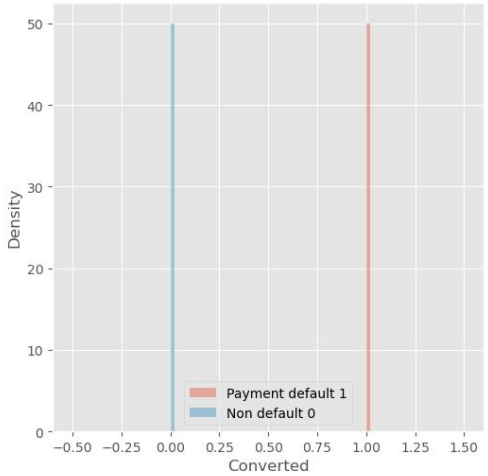

City - Distribution
Converted
0
1
Mumbai
Other Cities
Other Cities of Maharashtra
Select
Thane & Outskirts

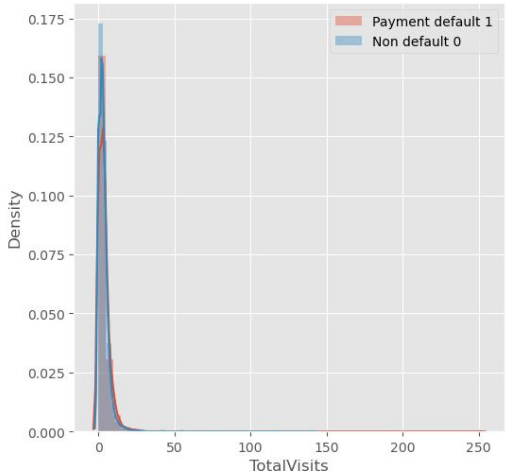

A free copy of Mastering The Interview - Distribution
Converted
0
1
No
Yes

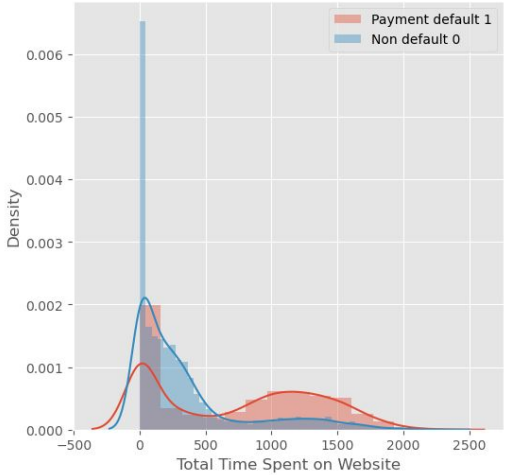

Last Notable Activity - Distribution
Converted
0
1
Approached upfront
Email Bounced
Email Link Clicked
Email Marked Spam
Email Opened
Form Submitted on Website
Had a Phone Conversation
Modified
Olark Chat Conversation
Page Visited on Website
Resubscribed to emails
SMS Sent
Unreachable
Unsubscribed
View in browser link Clicked

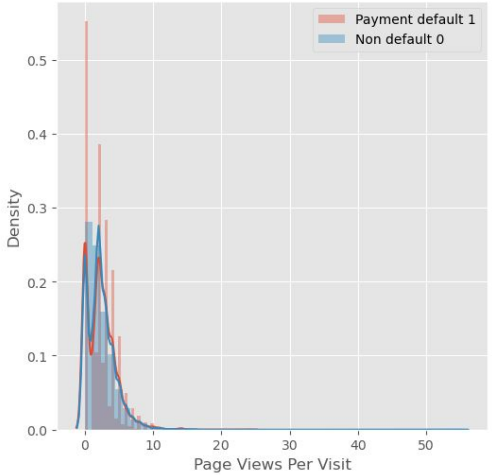
Distribution of Numerical Features


1e−5
Payment default 1
Non default 0
Density
Lead Number

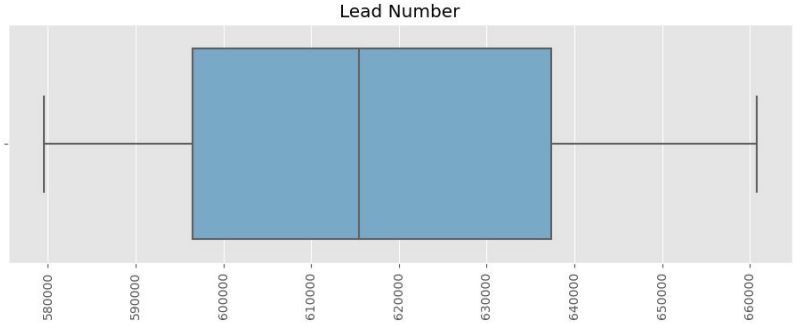

Payment default 1
Non default 0
Density
Converted

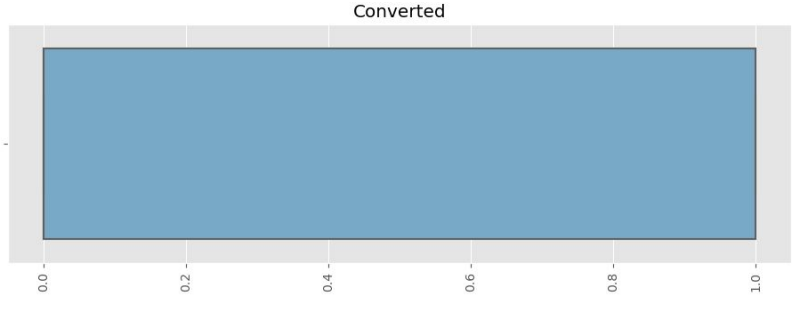

Payment default 1
Non default 0
Density
TotalVisits

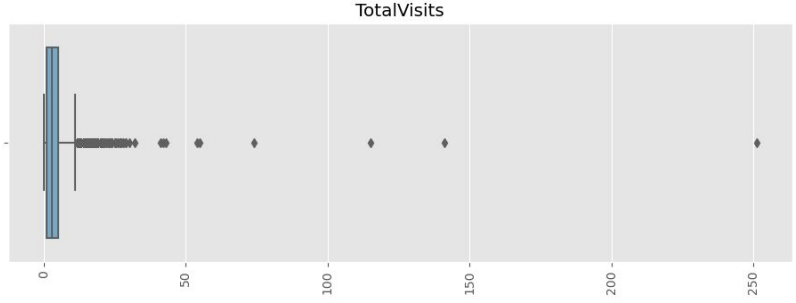

Payment default 1
Non default 0
Density
Total Time Spent on Website

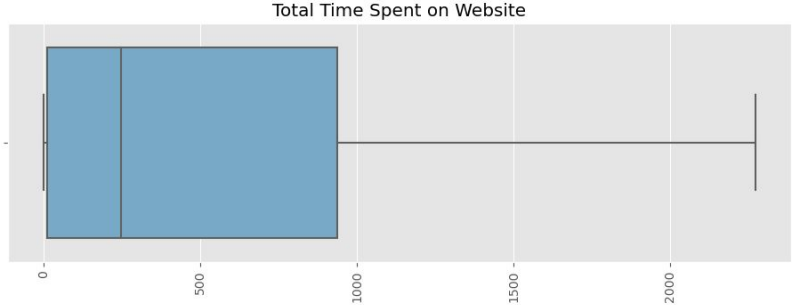

Payment default 1
Non default 0
Density
Page Views Per Visit

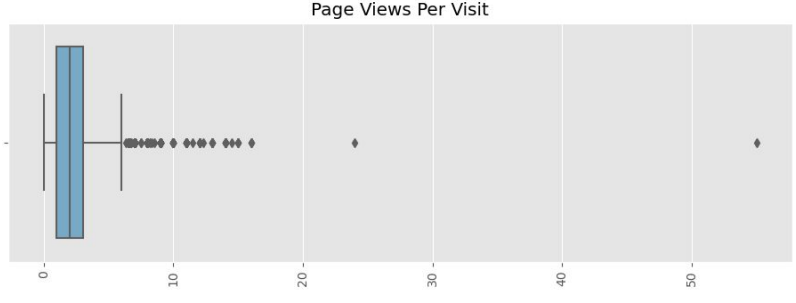
Boxplot of Numerical Features


Lead Number
580000
590000
600000
610000
620000
630000
640000
650000
660000

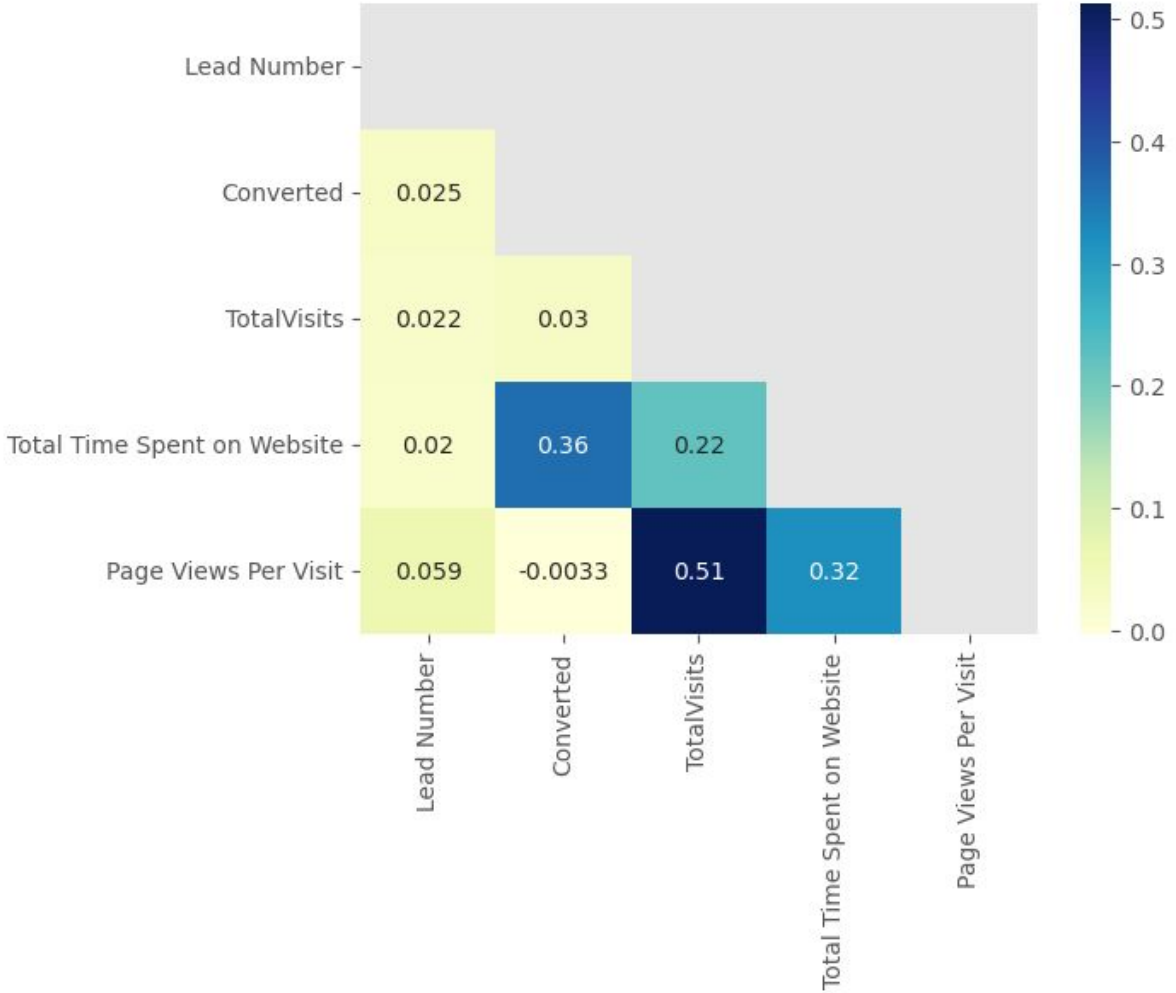

Converted
0.0
0.2
0.4
0.6
0.8
1.0


TotalVisits
0
50
100
150
200
250

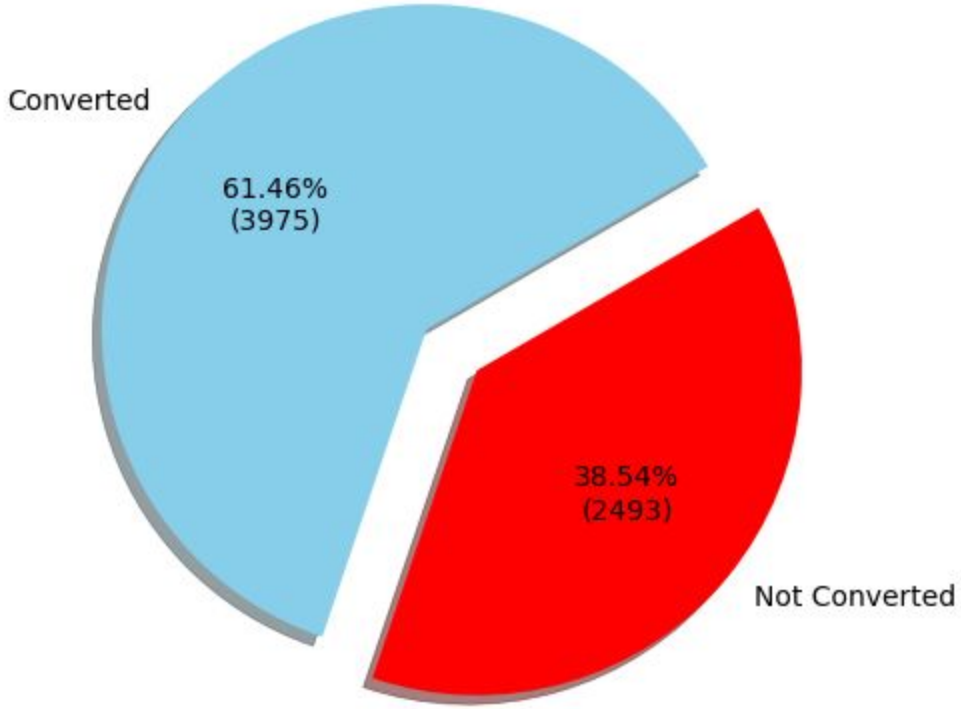

Total Time Spent on Website
0
500
1000
1500
2000


Page Views Per Visit
0
10
20
30
40
50

Correlation analysis for numerical features


Lead Number
Converted
TotalVisits
Total Time Spent on Website
Page Views Per Visit
0.025
0.022
0.03
0.02
0.36
0.22
0.059
-0.0033
0.51
0.32
Lead Number
Converted
TotalVisits
Total Time Spent on Website
Page Views Per Visit
0.5
0.4
0.3
0.2
0.1
0.0

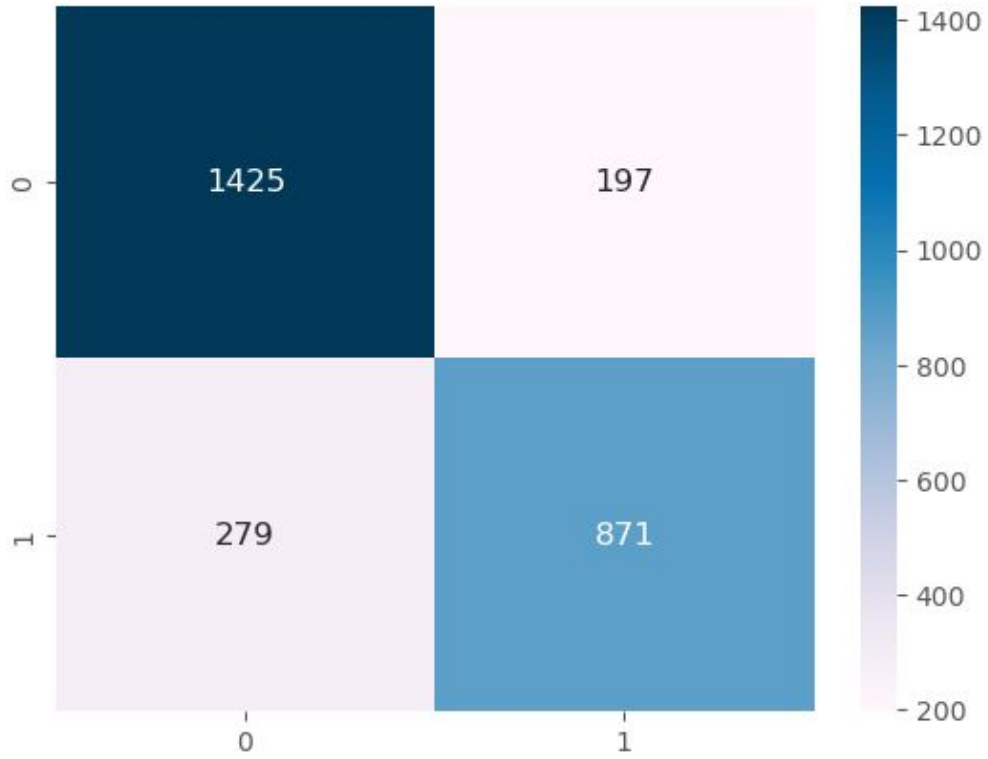
Class Imbalance For Final Data


Converted
61.46%
(3975)
38.54%
(2493)
Not Converted

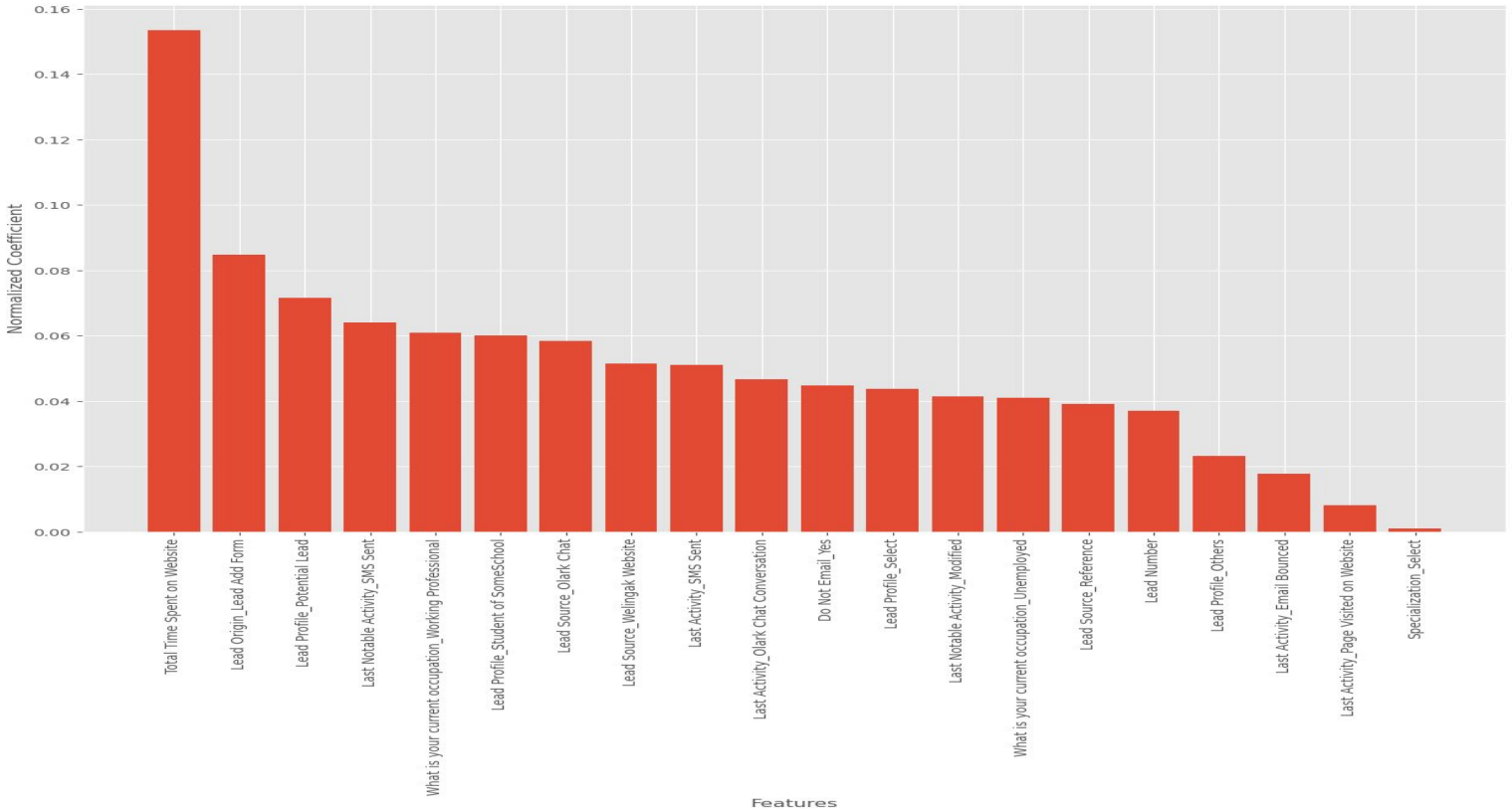
Model Evaluation

```
Accuracy on train data :  0.831
Accuracy on test data :  0.828
Precision on test data :  0.816
Recall on test data :  0.785
ROC score on test data :  0.818
```

Confusion Matrix



Feature Importance



Recommendations

- Individuals who are unemployed tend to prioritize courses that offer improved career opportunities.

- Targeting working professionals could be the next step in attracting individuals interested in courses that enhance career prospects.

- Utilizing Google search engine can significantly increase business for the company, as the lead conversion rate is higher compared to other sources, followed by referrals.

- Those who respond to email inquiries are more likely to enroll in courses.