Lead Scoring Case Study

Analysis of Education X Leads to predict Hot Leads’ conversion

Surya Adhikarla
Uppalapadu Sai
Irfan Syed

Problem Statement

Identifying Hot leads and their conversion

X Education company sells online courses to industry professionals.

- When someone interested in the courses fill up a form providing their email address or phone number, they are classified to be a lead.
- The most potential leads, also known as ‘Hot Leads’
- Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not.
- The typical lead conversion rate at X education is around 30%.
- The company wants to improve the lead conversion rate to around 80%

Business Objective

Objective of the Case study

- Build a logistic regression model to assign a lead score. Higher the score means the lead is hot.

- Provide insights and recommendations along with some answers to company problems identified, using the logistic regression model.

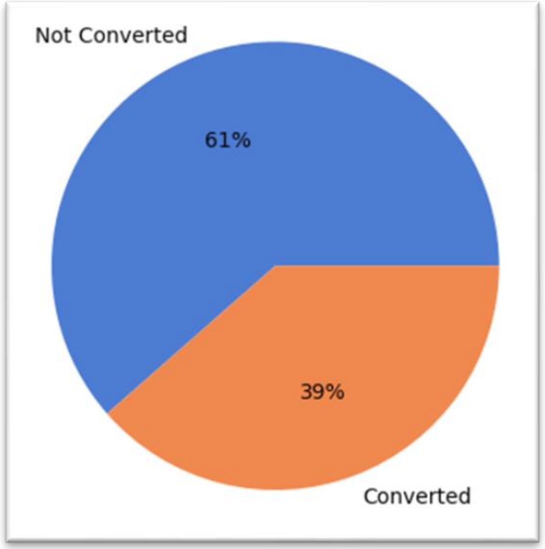
Data Provided

‘leads.csv' contains all the information of the leads gathered by the company.

Target column: CONVERTED
0 – Lead is converted to customer
1 – Lead is not converted to customer

Property	Comments
Size of Data	9240 rows & 37 columns
Missing information	There are columns with 52% to 15% missing values with 9 having 30% or more nulls
Special values	Many of the categorical variables have a level called 'Select’ (which means the lead did not select any option in the form)

The converted vs Non-converted ratio in the data is 1:1.6

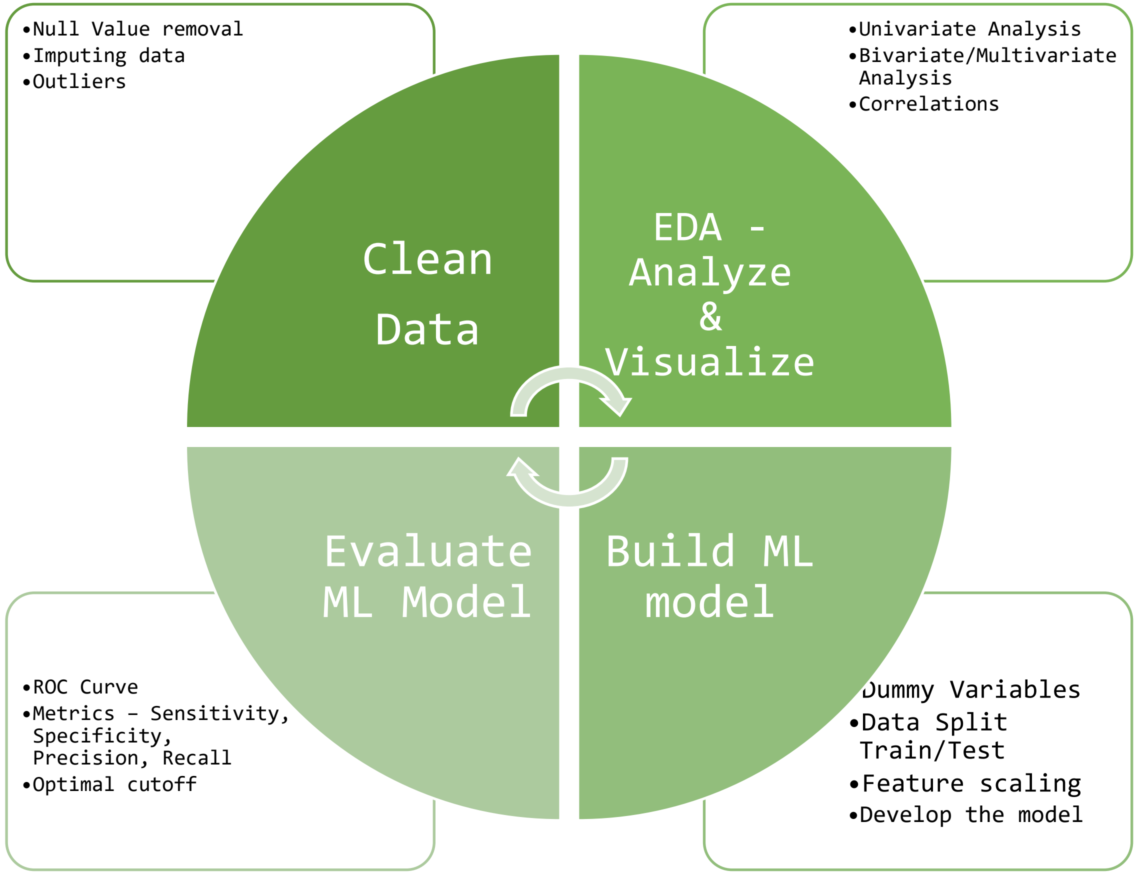


ML model building (Logistic Regression)

- **Logistic regression** (or logit regression) estimates the parameters of a logistic model (the coefficients in the linear or non linear combinations).

- In binary logistic regression there is a single binary dependent variable, coded by an indicator variable, where the two values are labeled "0" and "1“

Source: [Wikipedia]()



- Null Value removal
- Imputing data
- Outliers

Clean Data

EDA - Analyze & Visualize

- Univariate Analysis
- Bivariate/Multivariate Analysis
- Correlations

Evaluate ML Model

- ROC Curve
- Metrics – Sensitivity, Specificity, Precision, Recall
- Optimal cutoff

Build ML model

- Dummy Variables
- Data Split Train/Test
- Feature scaling
- Develop the model

Data Cleaning and Preparation

Dropping columns

- Columns with 30% or more missing values were dropped as they have significant gaps.

- Categorical columns with most or all of the rows having same values were dropped as they cannot contribute to the analysis.

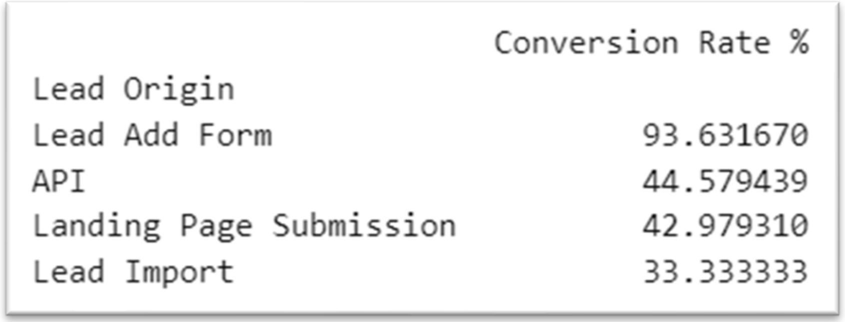
- For Columns with fewer missing values, only the rows with null values were removed.

- Post-Clean up the data had 6373 rows and 12 columns out of 9240 rows and 37 columns (Almost 30% of data was removed during the clean up).

EDA - Analyze and Visualize

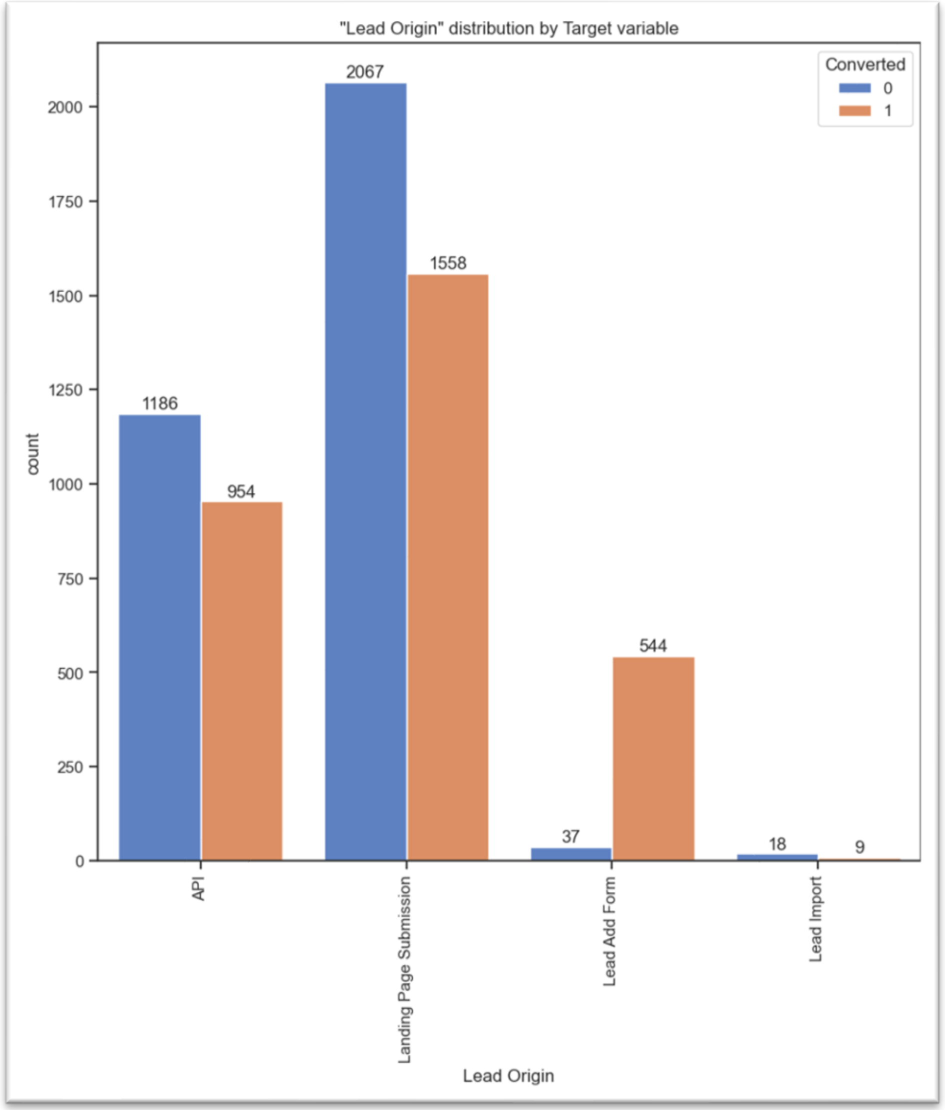
Lead Origin

The origin identifier with which the customer was identified to be a lead. Includes API, Landing Page Submission, etc.



```
                         Conversion Rate %
Lead Origin
Lead Add Form                    93.631670
API                              44.579439
Landing Page Submission          42.979310
Lead Import                      33.333333
```

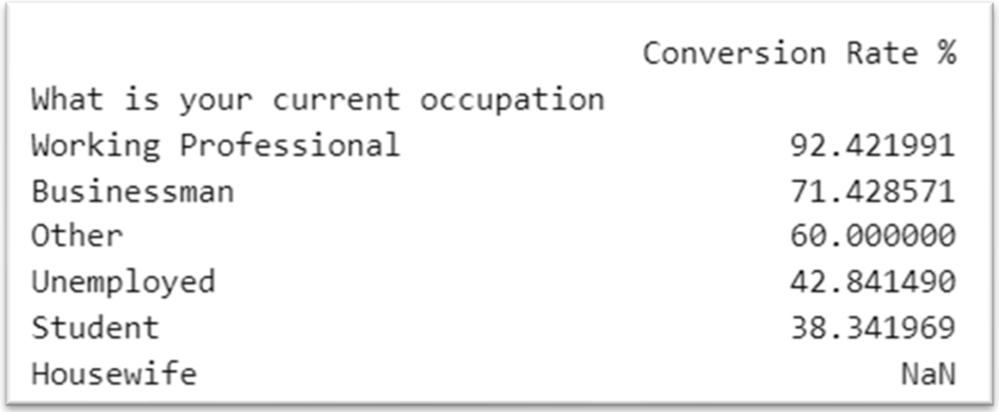
'Lead Add form' and 'References' have high conversion rate



EDA - Analyze and Visualize

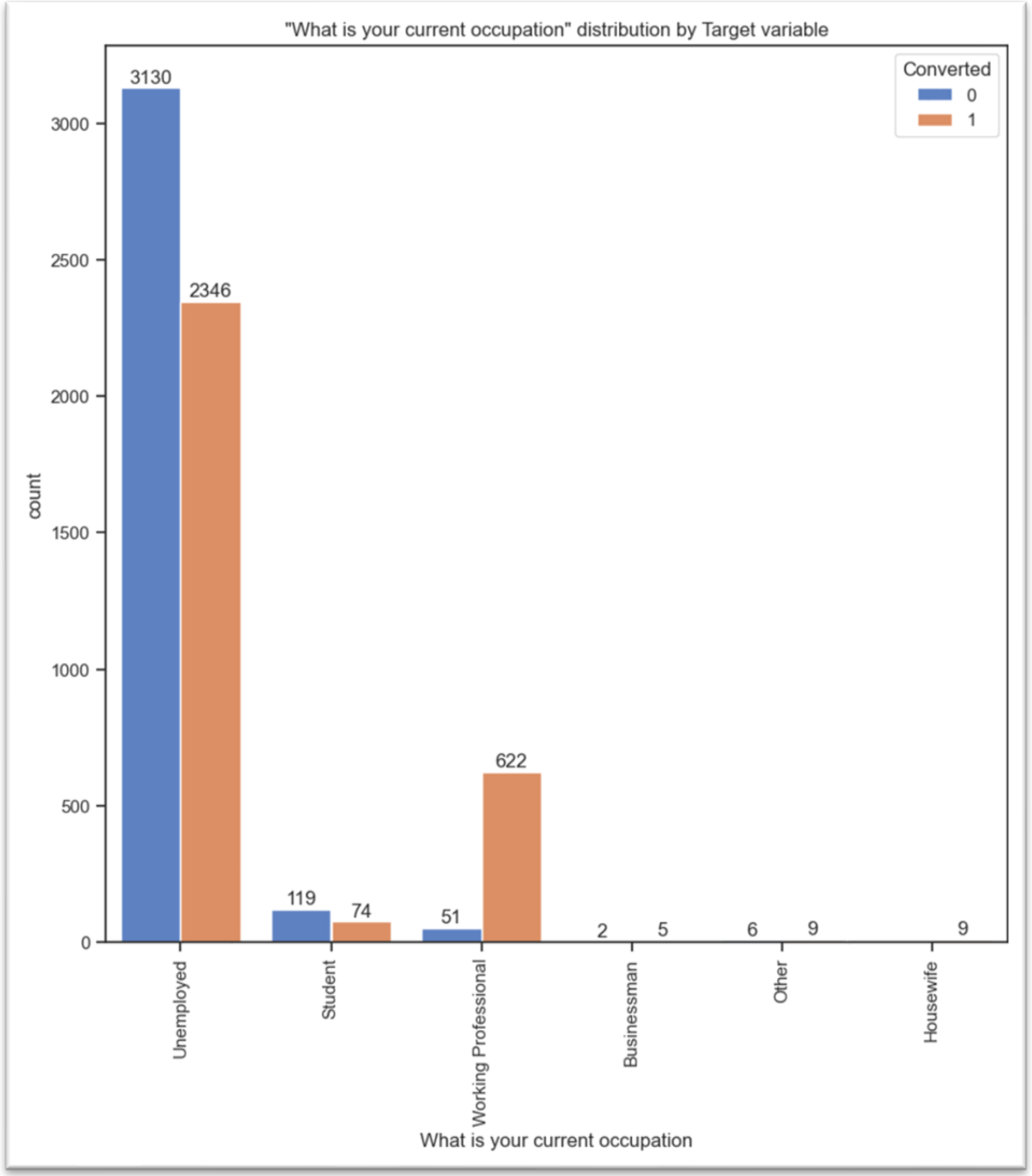
Occupation

Indicates whether the customer is a student, umemployed or employed.



```
                                  Conversion Rate %
What is your current occupation
Working Professional                      92.421991
Businessman                               71.428571
Other                                     60.000000
Unemployed                                42.841490
Student                                   38.341969
Housewife                                       NaN
```

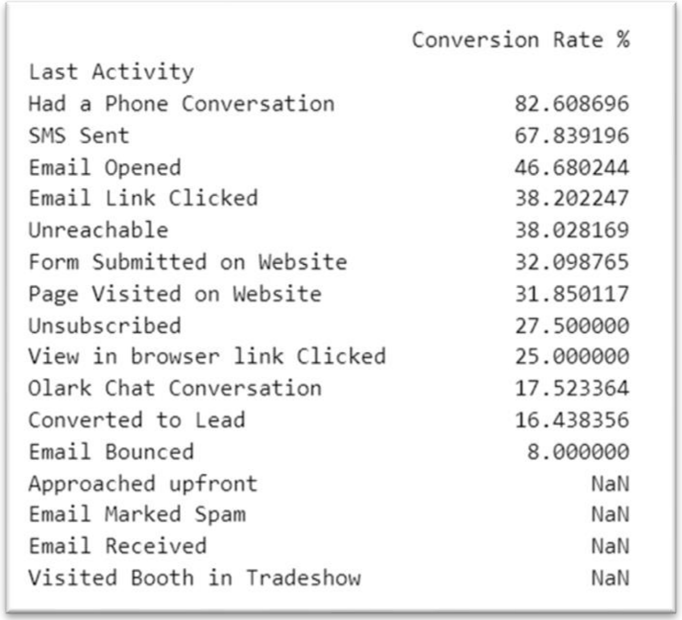
Although most of the leads are Unemployed they only have a conversion rate of 42%



EDA - Analyze and Visualize

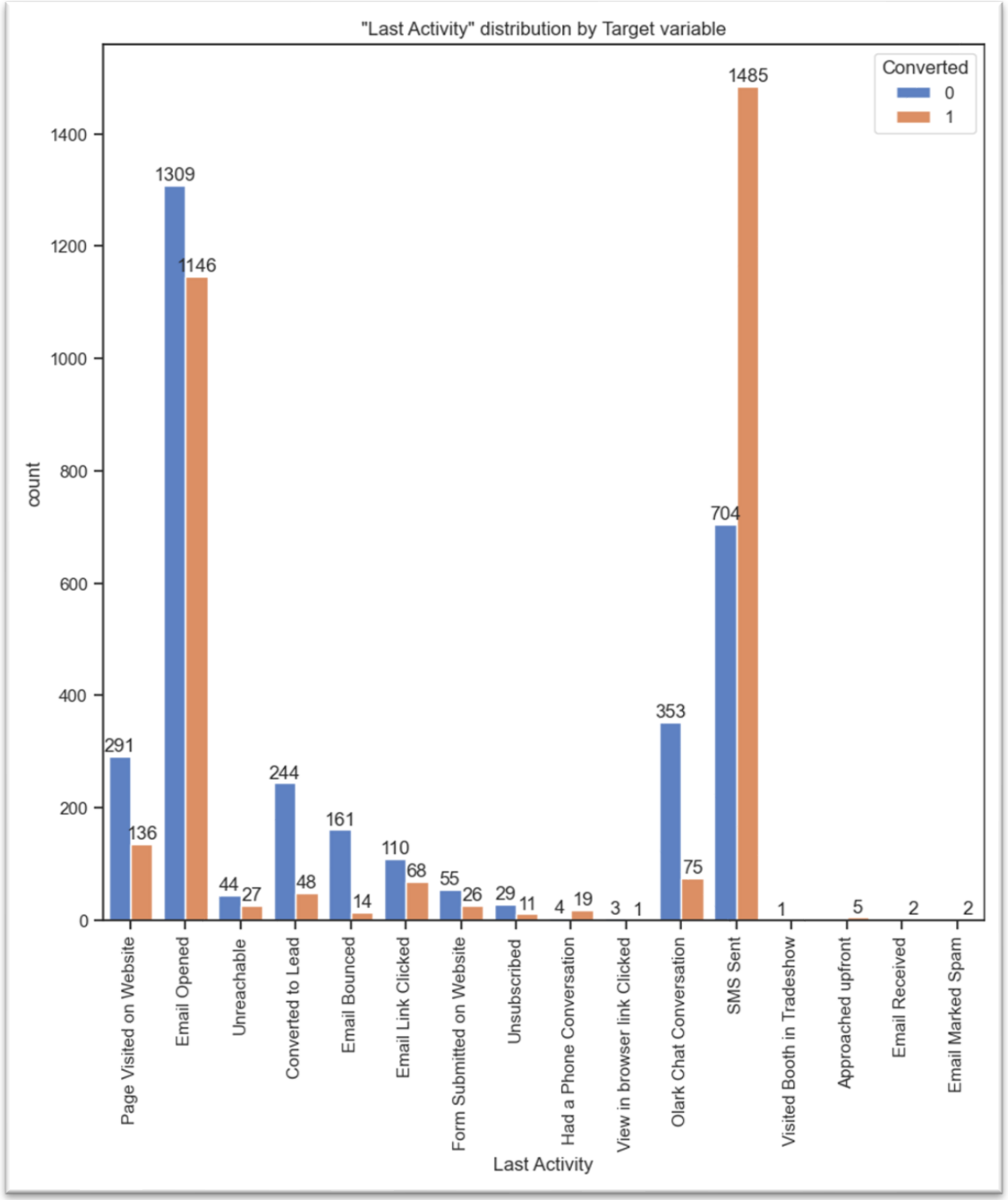
Last Activity

Last activity performed by the customer. Includes Email Opened, Olark Chat Conversation, etc.



```
                              Conversion Rate %
Last Activity
Had a Phone Conversation              82.608696
SMS Sent                              67.839196
Email Opened                          46.680244
Email Link Clicked                    38.202247
Unreachable                           38.028169
Form Submitted on Website             32.098765
Page Visited on Website               31.850117
Unsubscribed                          27.500000
View in browser link Clicked          25.000000
Olark Chat Conversation               17.523364
Converted to Lead                     16.438356
Email Bounced                          8.000000
Approached upfront                          NaN
Email Marked Spam                           NaN
Email Received                              NaN
Visited Booth in Tradeshow                  NaN
```

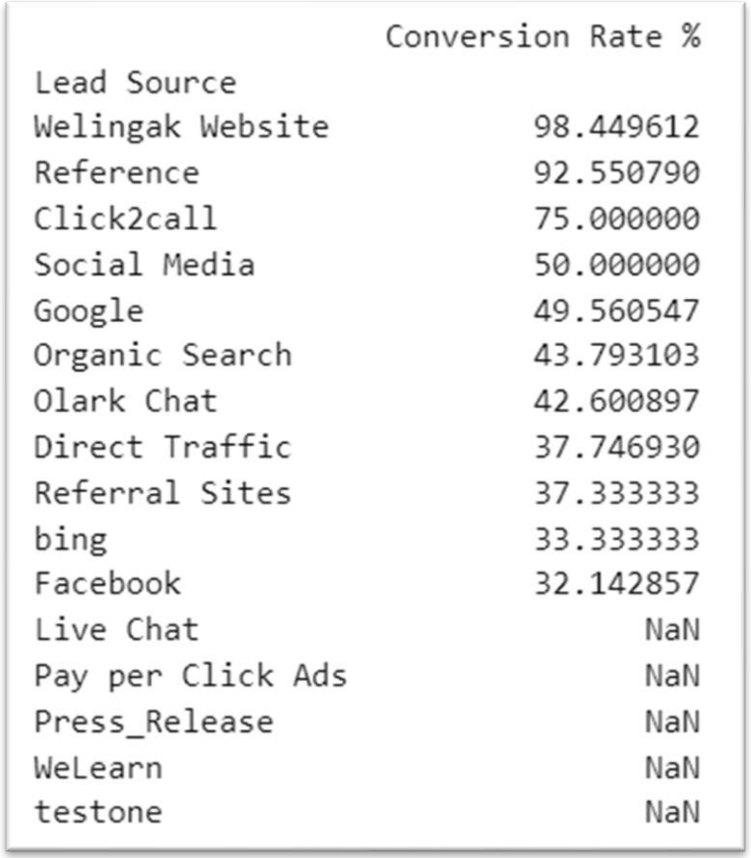
The most common last activity for converted customers seems to be SMS and Email



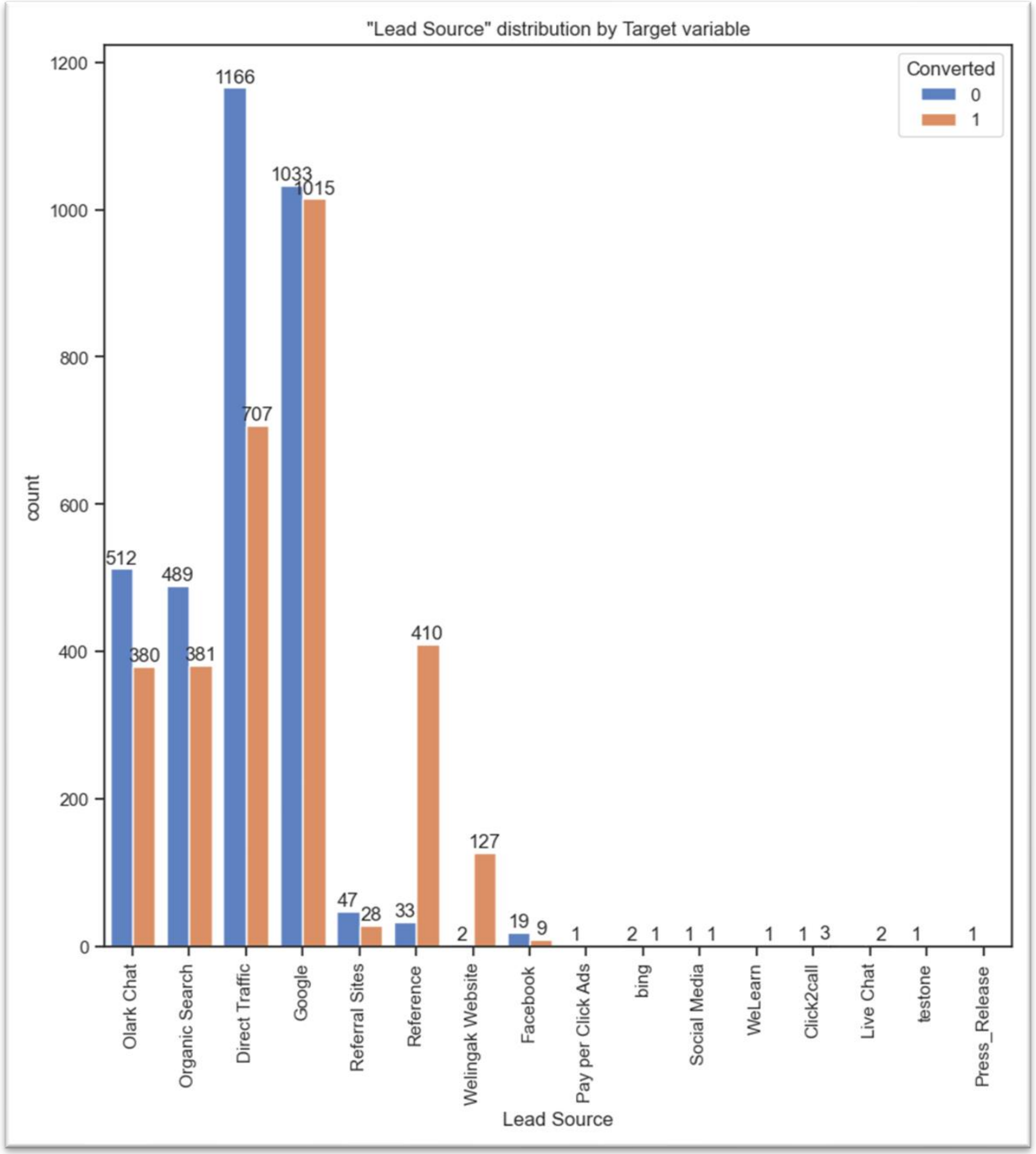
EDA - Analyze and Visualize

Lead Source

The source of the lead. Includes Google, Organic Search, Olark Chat, etc.



```
                   Conversion Rate %
Lead Source
Welingak Website           98.449612
Reference                  92.550790
Click2call                 75.000000
Social Media               50.000000
Google                     49.560547
Organic Search             43.793103
Olark Chat                 42.600897
Direct Traffic             37.746930
Referral Sites             37.333333
bing                       33.333333
Facebook                   32.142857
Live Chat                        NaN
Pay per Click Ads                NaN
Press_Release                    NaN
WeLearn                          NaN
testone                          NaN
```



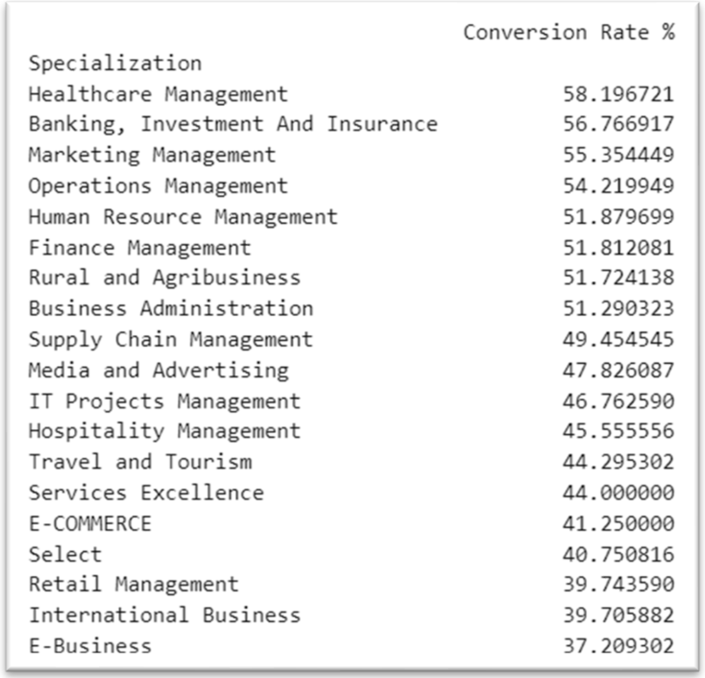
EDA - Analyze and Visualize

Specialization

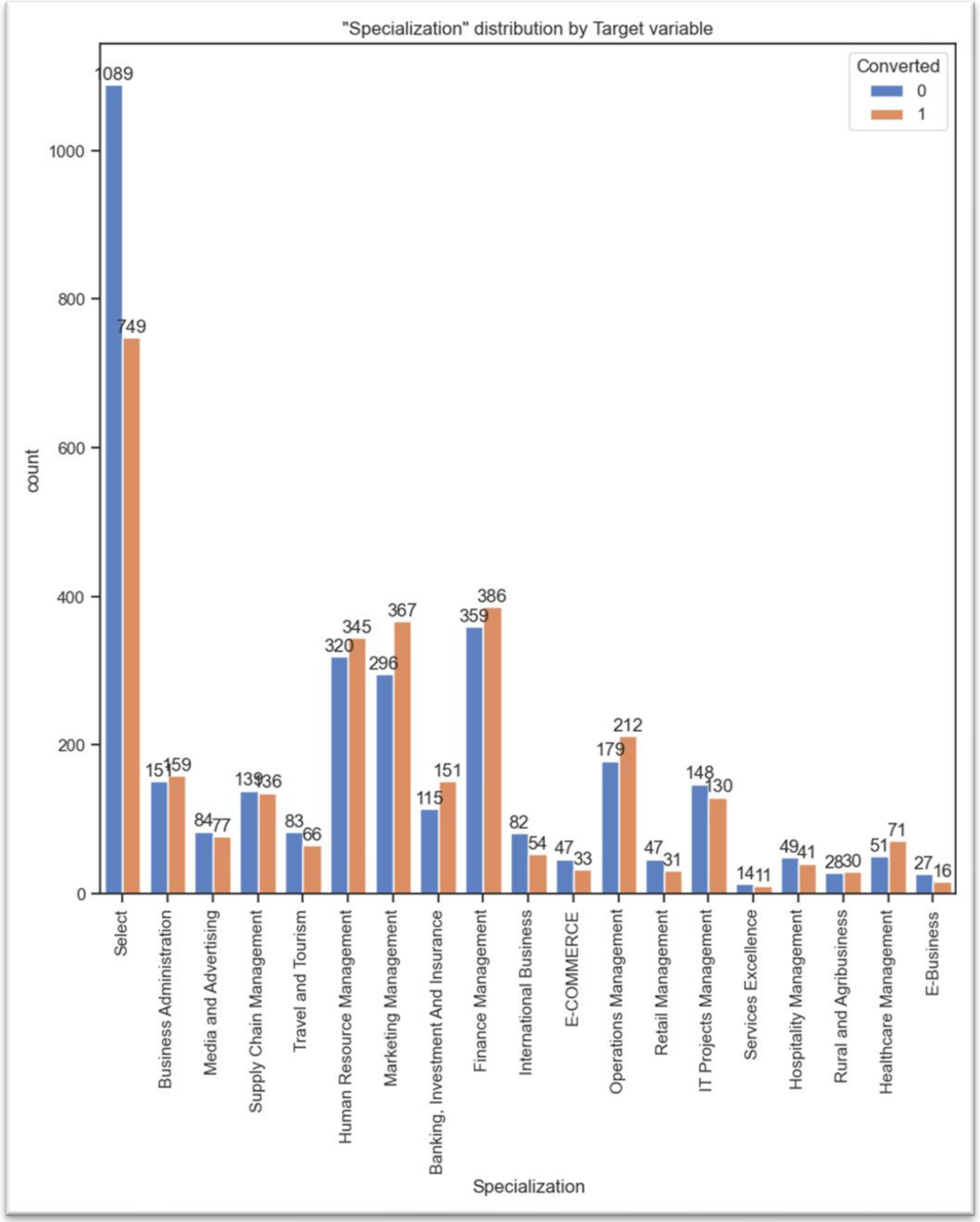
The industry domain in which the customer worked before.

Note: Select => The customer has not selected this option while filling form

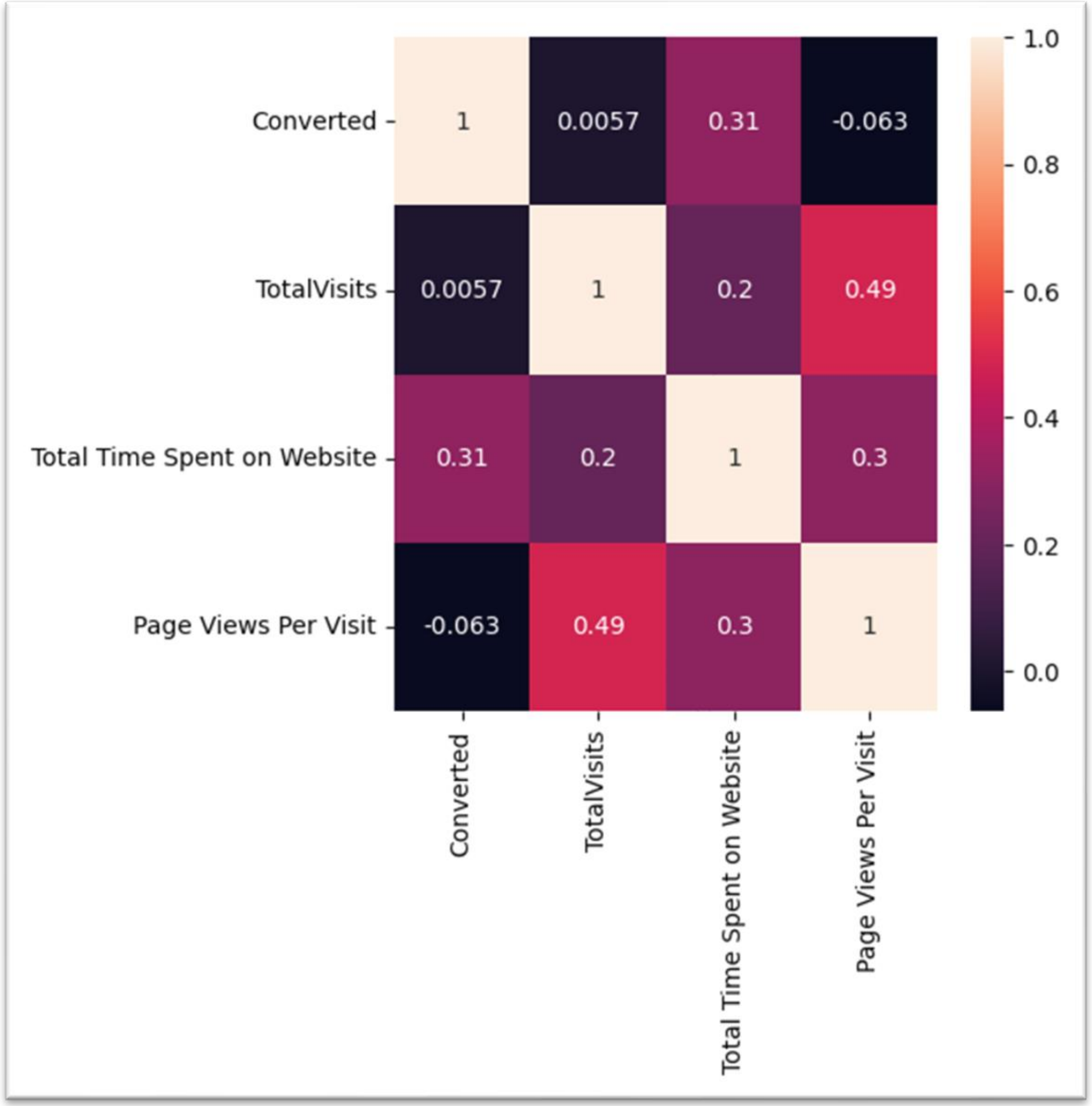
The conversion rate seems to be evenly distributed across all Specializations within 40% - 50%



```
                                   Conversion Rate %
Specialization
Healthcare Management                      58.196721
Banking, Investment And Insurance          56.766917
Marketing Management                       55.354449
Operations Management                      54.219949
Human Resource Management                  51.879699
Finance Management                         51.812081
Rural and Agribusiness                     51.724138
Business Administration                    51.290323
Supply Chain Management                    49.454545
Media and Advertising                      47.826087
IT Projects Management                     46.762590
Hospitality Management                     45.555556
Travel and Tourism                         44.295302
Services Excellence                        44.000000
E-COMMERCE                                 41.250000
Select                                     40.750816
Retail Management                          39.743590
International Business                     39.705882
E-Business                                 37.209302
```



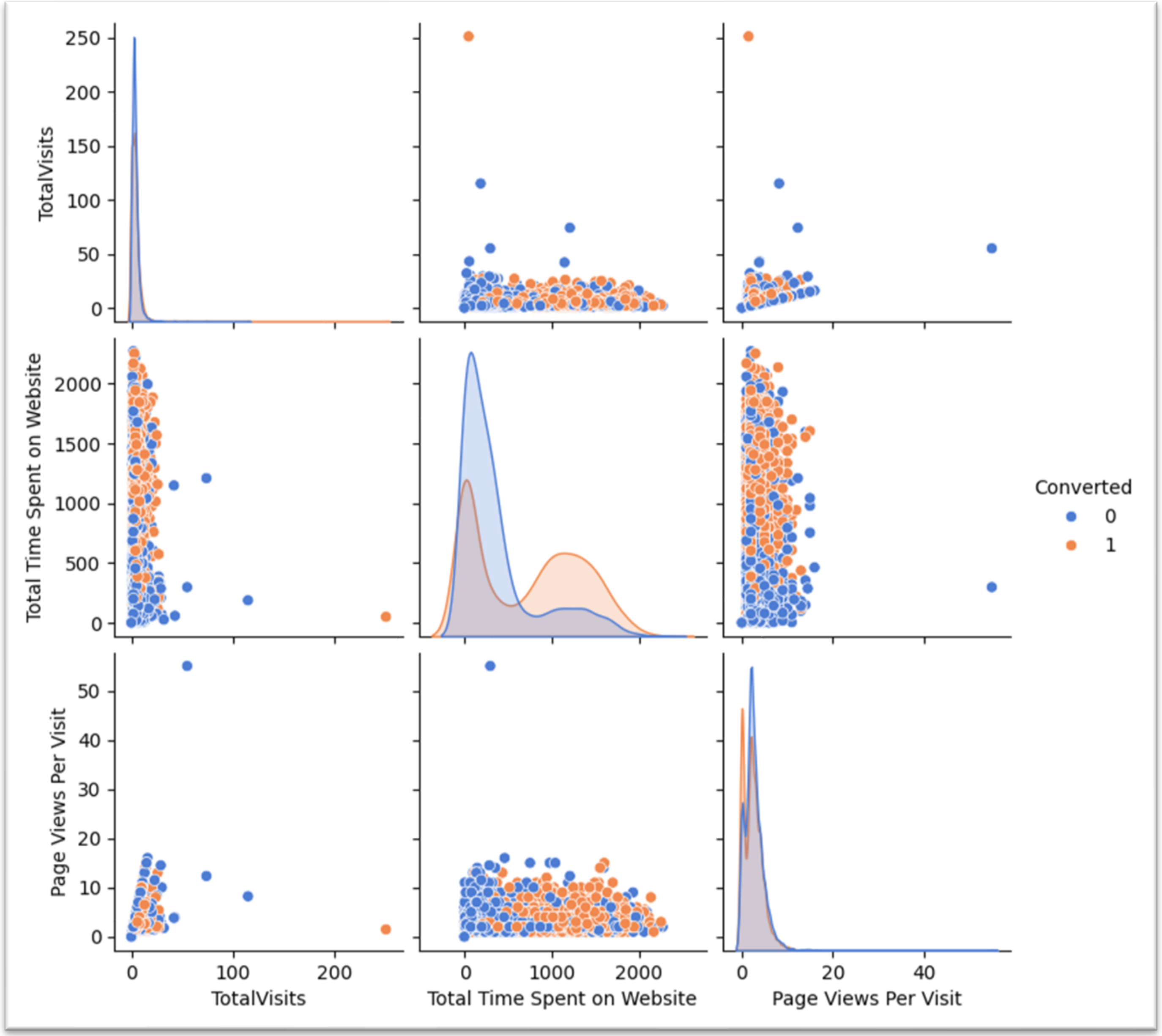
EDA - Analyze and Visualize



Target variable Converted seems to have a small linear correlation only with Total Time spent on Website

EDA - Analyze and Visualize

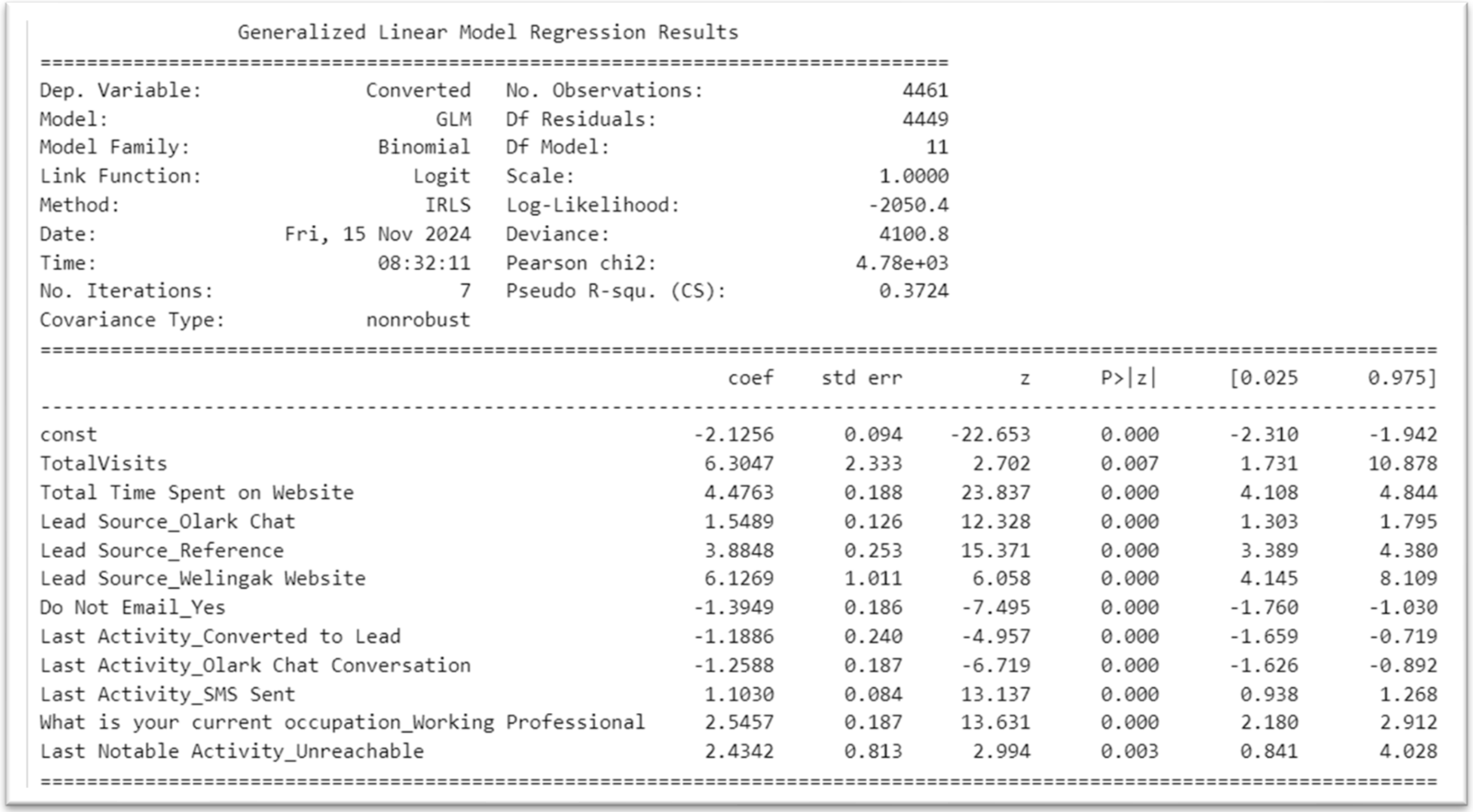
Target variable Converted seems to have a small linear correlation only with Total Time spent on Website



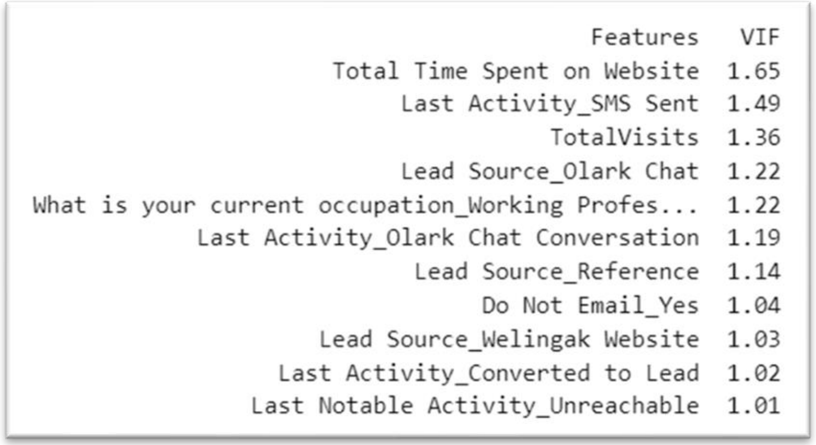
Build ML Model

Logistic Regression model

- Used RFE to identify 15 features
- Used manual elimination by reviving VIF and p-values to arrive at the final model
- Final model has 11 features
- All VIF values are < 5



```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Converted   No. Observations:                 4461
Model:                            GLM   Df Residuals:                     4449
Model Family:                Binomial   Df Model:                           11
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -2050.4
Date:                Fri, 15 Nov 2024   Deviance:                       4100.8
Time:                        08:32:11   Pearson chi2:                 4.78e+03
No. Iterations:                     7   Pseudo R-squ. (CS):             0.3724
Covariance Type:            nonrobust
========================================================================================================================
                                                          coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------------------------------
const                                                  -2.1256      0.094    -22.653      0.000      -2.310      -1.942
TotalVisits                                             6.3047      2.333      2.702      0.007       1.731      10.878
Total Time Spent on Website                             4.4763      0.188     23.837      0.000       4.108       4.844
Lead Source_Olark Chat                                  1.5489      0.126     12.328      0.000       1.303       1.795
Lead Source_Reference                                   3.8848      0.253     15.371      0.000       3.389       4.380
Lead Source_Welingak Website                            6.1269      1.011      6.058      0.000       4.145       8.109
Do Not Email_Yes                                       -1.3949      0.186     -7.495      0.000      -1.760      -1.030
Last Activity_Converted to Lead                        -1.1886      0.240     -4.957      0.000      -1.659      -0.719
Last Activity_Olark Chat Conversation                  -1.2588      0.187     -6.719      0.000      -1.626      -0.892
Last Activity_SMS Sent                                  1.1030      0.084     13.137      0.000       0.938       1.268
What is your current occupation_Working Professional    2.5457      0.187     13.631      0.000       2.180       2.912
Last Notable Activity_Unreachable                       2.4342      0.813      2.994      0.003       0.841       4.028
========================================================================================================================
```

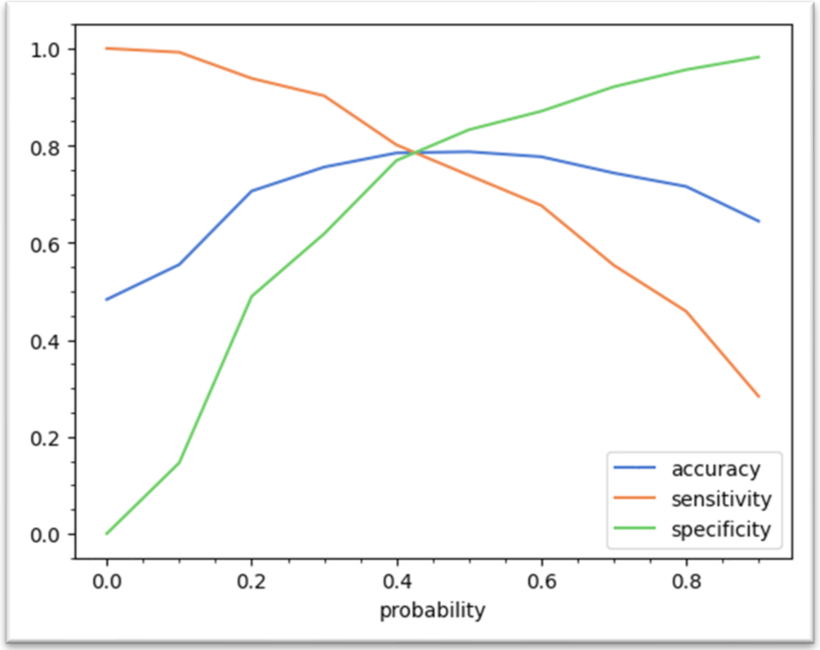


```
                                         Features   VIF
                      Total Time Spent on Website  1.65
                           Last Activity_SMS Sent  1.49
                                      TotalVisits  1.36
                           Lead Source_Olark Chat  1.22
What is your current occupation_Working Profes...  1.22
            Last Activity_Olark Chat Conversation  1.19
                            Lead Source_Reference  1.14
                                 Do Not Email_Yes  1.04
                     Lead Source_Welingak Website  1.03
                  Last Activity_Converted to Lead  1.02
                Last Notable Activity_Unreachable  1.01
```

Evaluate ML Model

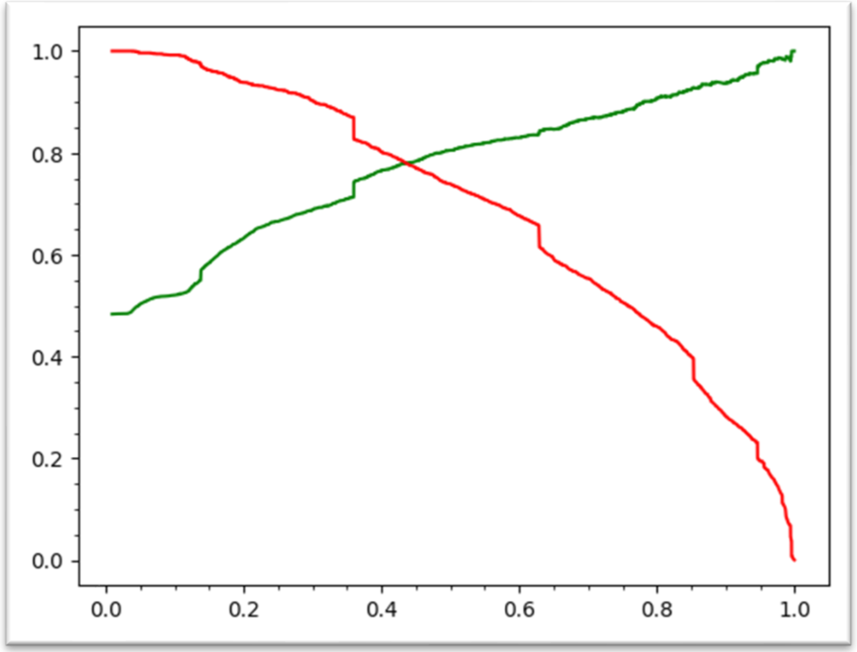
Optimal Cutoff

- The optimal cutoff seemed to be at 0.42 where accuracy , sensitivity and specificity are almost the same



Precision/Recall Tradeoff

- The precision vs recall plot also seems to intersect at approx. 0.42



Evaluate ML Model - Metrics

Train Data set

- Accuracy = 78.7%
- Precision = 77.2%
- Recall = 79.2%
- Sensitivity = 79.2%
- Specificity = 78.2%

Confusion matrix

Actual	Predicted	
	Not Converted	Converted
Not Converted	1803	502
Converted	449	1707

Test Data set

- Accuracy = 78.7%
- Precision = 77.2%
- Recall = 79.2%
- Sensitivity = 79.2%
- Specificity = 78.2%

Confusion matrix

Actual	Predicted	
	Not Converted	Converted
Not Converted	799	204
Converted	194	715

Insights

- Approximately 70% of leads expressed interest in the courses to advance their careers.

- Website metrics, such as page visits and time spent, were critical predictors of lead conversion.

- A logistic regression model was developed to predict lead conversion.

- The model determined an optimal probability threshold of 0.42 for predicting conversions.

Recommendations

- Improve data collection processes by making critical fields mandatory to avoid unusable entries.

- Maintain and enhance the website’s user experience (UX) and content to boost engagement.

- Unemployed leads formed a significant portion of the dataset but had a conversion rate of only 42%. Revisiting course pricing and commitment levels could enhance their appeal.

Final Feature List
TotalVisits
Total Time Spent on Website
Lead Source_Olark Chat
Lead Source_Reference
Lead Source_Welingak Website
Do Not Email_Yes
Last Activity_Converted to Lead
Last Activity_Olark Chat onversation
Last Activity_SMS Sent
What is your current occupation_Working Professional
Last Notable Activity_Unreachable