Lead Scoring Case Study

Lead conversion for X Education

Soumya Mandapaka
Suraj Das
Amit Kumar Singh

Problem Statement

Identifying Leads for conversion

An education company named X Education sells online courses to industry professionals

- The company markets its courses on several websites and search engines like Google. When interested people fill up a form providing their email address or phone number, they are classified to be a lead.
- The company wishes to identify the most potential leads, also known as ‘Hot Leads’.
- The typical lead conversion rate at X education is around 30%. The CEO has given a ballpark of the target lead conversion rate to be around 80%.

Lead Conversion Process
(Demonstrated as a funnel)

Business Objective

Business Objective of the Case study

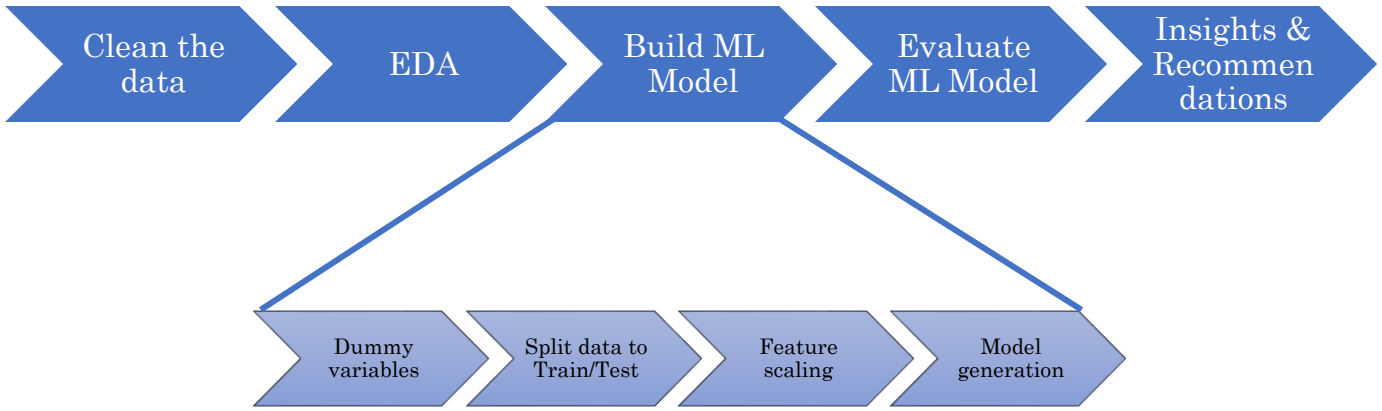
- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads.
  - A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

- Address problems presented by the company using the model and provide recommendations

Process Overview

This case study is an **Machine Leaning model building** exercise which primarily constitutes of the following steps

1. Data Cleaning and Preparation
2. Exploratory Data Analysis (EDA)
3. Model Building
    1. Creating of dummy variables
    2. Splitting the data to Train/Test
    3. Feature scaling
    4. Model generation
    5. Model Evaluation
4. Inference
5. Recommendations



Data Provided

- The data gathered by the company is provided as the dataset - '**leads.csv**'
- This dataset contains the information of the leads as filled in the forms online.
- The data includes a target variable that indicates whether a client has converted or not

Data Description

Property	Comments
Number of rows (observations)	There are **9240 rows**
Number of columns (parameters)	There are **37 columns**
Missing values	There are 6 columns with more than 30% missing values
Special values	Some columns has an value 'Select' which denotes that the lead has not selected a valid option

Data Cleaning

At a high level data cleaning involves

- Identify the columns with higher null values and drop them
- If the null rows are fewer, remove the only the rows that contains null information but retain the columns
- Identify columns with skewed data (i.e. most of the values are the same) and drop them
- Identify the best way to address the invalid values in the data

Missing data

- 6 Columns had 30% or more missing values and were dropped as they have significant gaps.
- For 3 columns with fewer null values, only the rows containing nulls were removed
- 17 columns were dropped as most of them had a single value or all of them had the same value

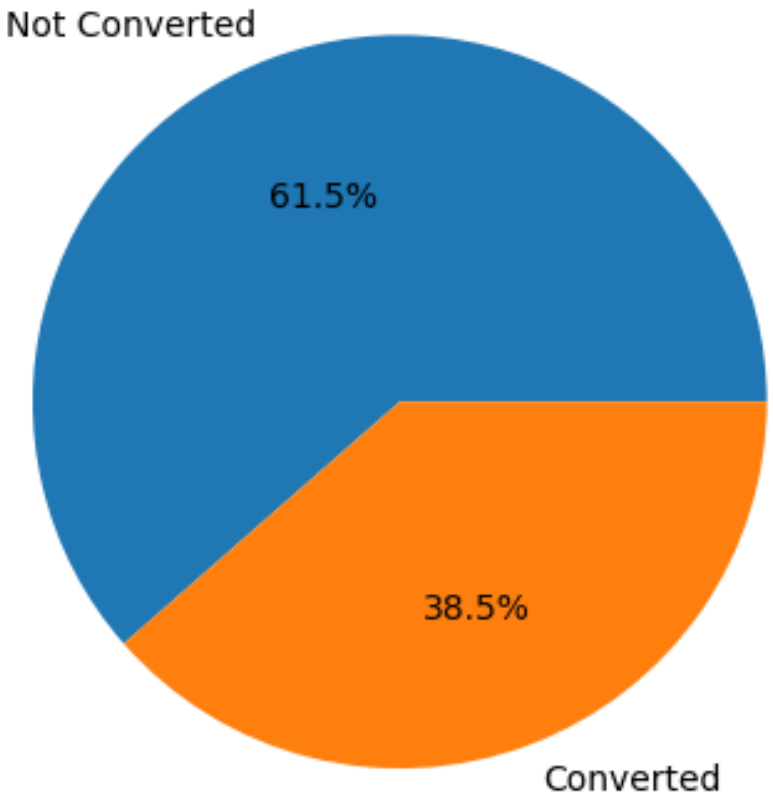
Data Analysis

Data Imbalance

In the provided data set, based on the pie chart it can be seen that

- Converted percentage is 38.5%
- Non-Converted percentage is 61.5%

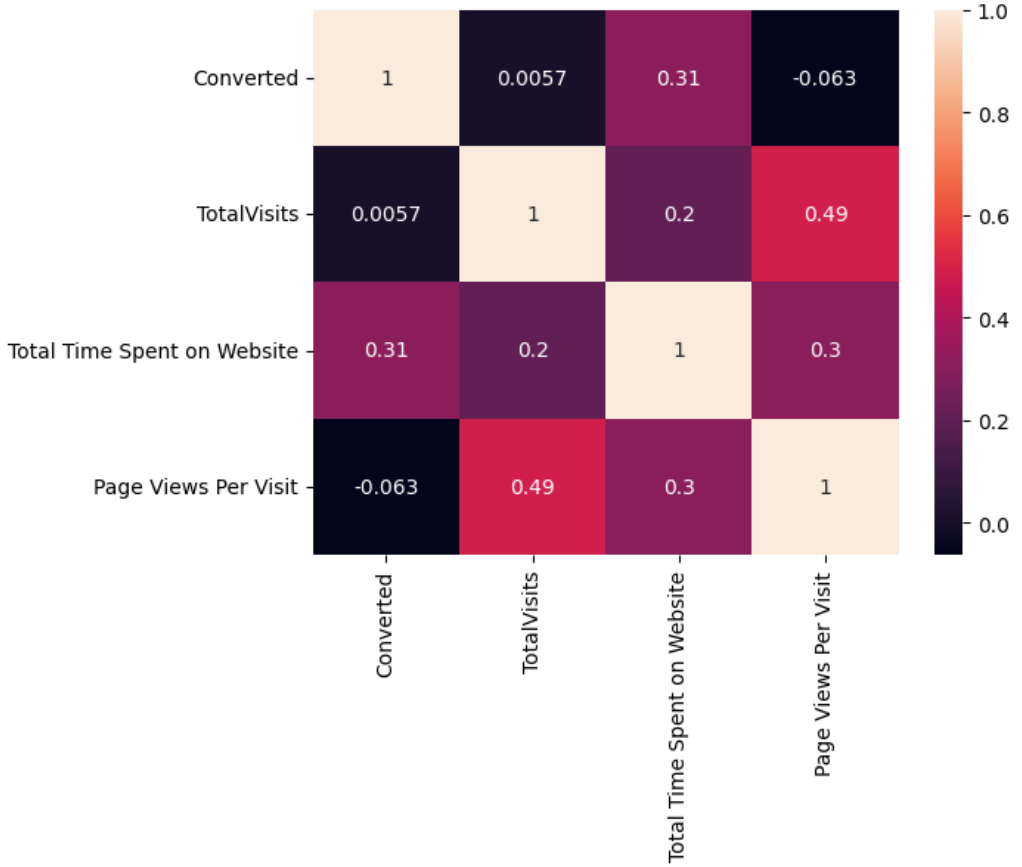
Data imbalance ratio for Non-converted with with respect to Converted is 1.6:1 (approx.)

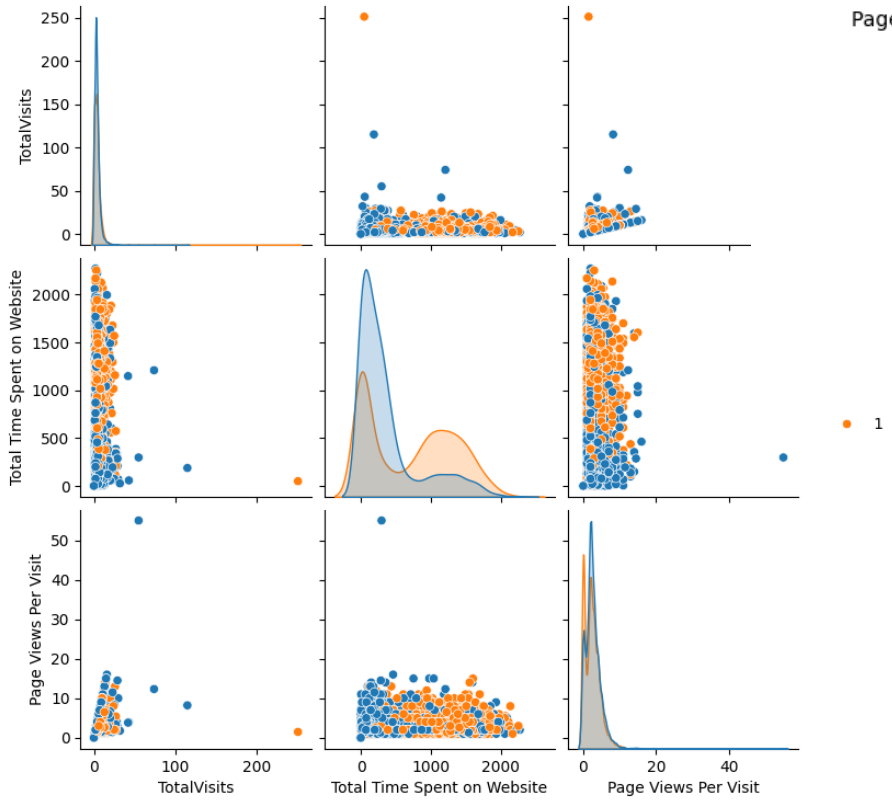


Data Analysis

Heatmap and Pair plot

Only **Total Time spent on Website** seem to have a noticeable linear correlation with the target variable **Converted** which makes sense

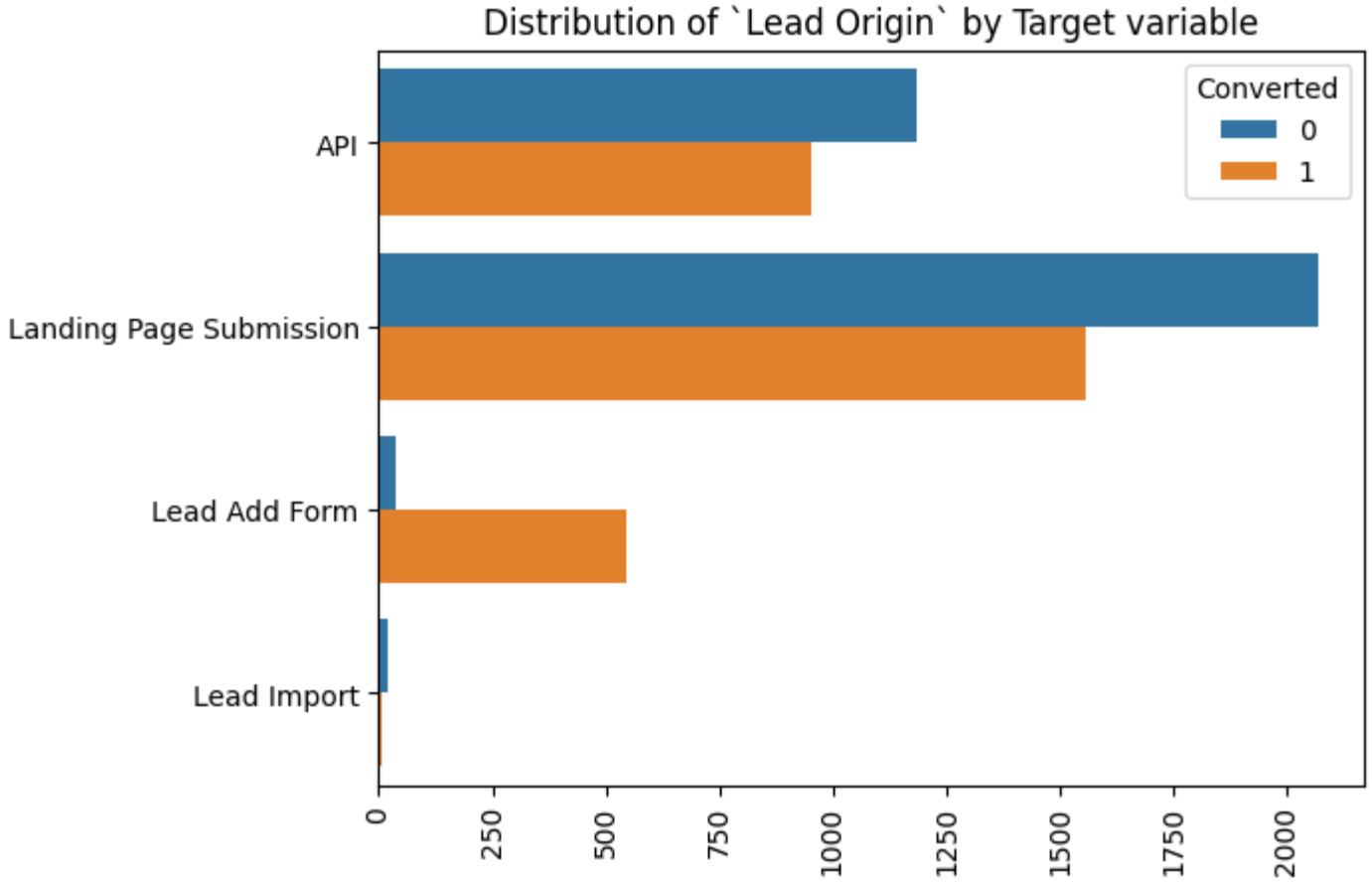




Data Analysis

Lead Origin

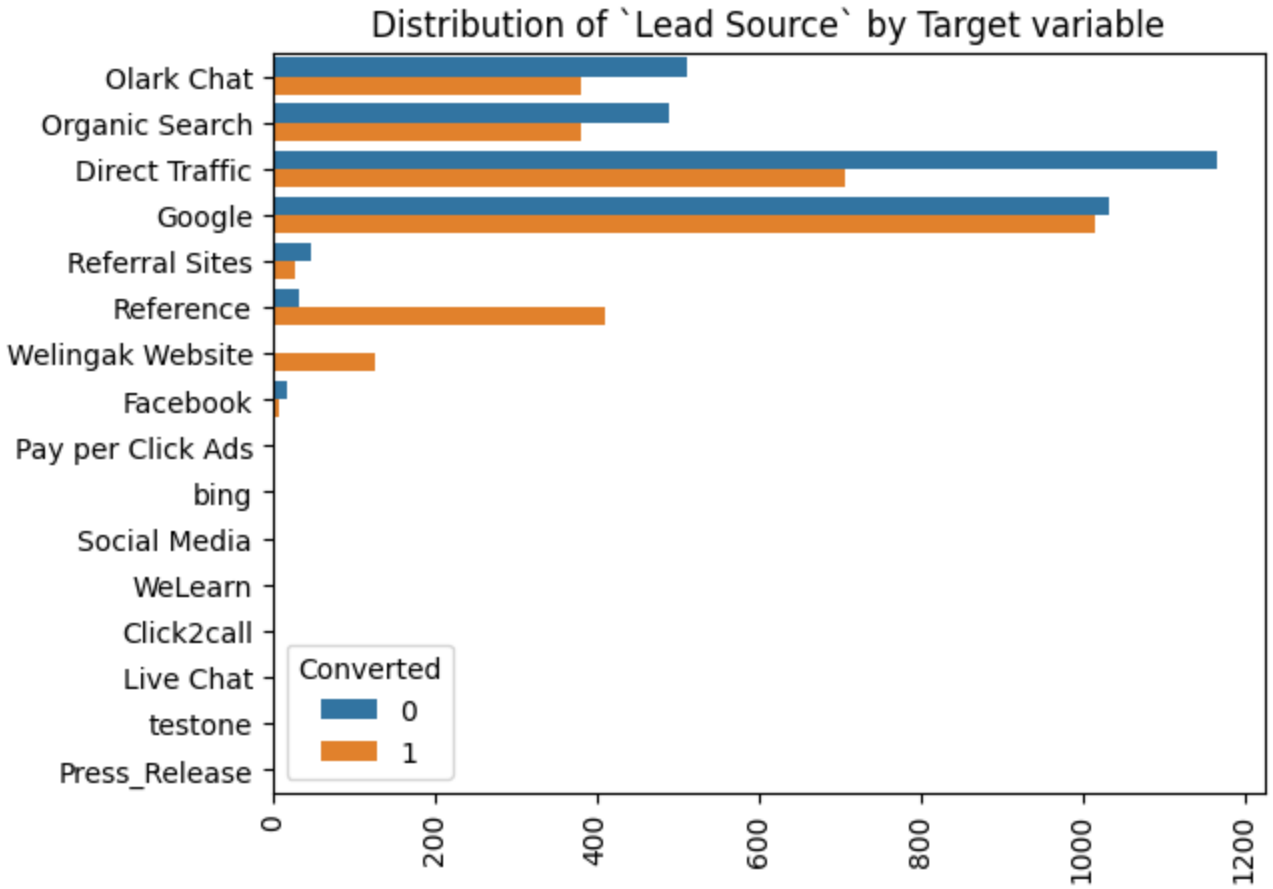
	Converted	Not Converted	Rate
Lead Add Form	544	37	0.94
API	954	1186	0.45
Landing Page Submission	1558	2067	0.43
Lead Import	9	18	0.33



Data Analysis

Lead Source

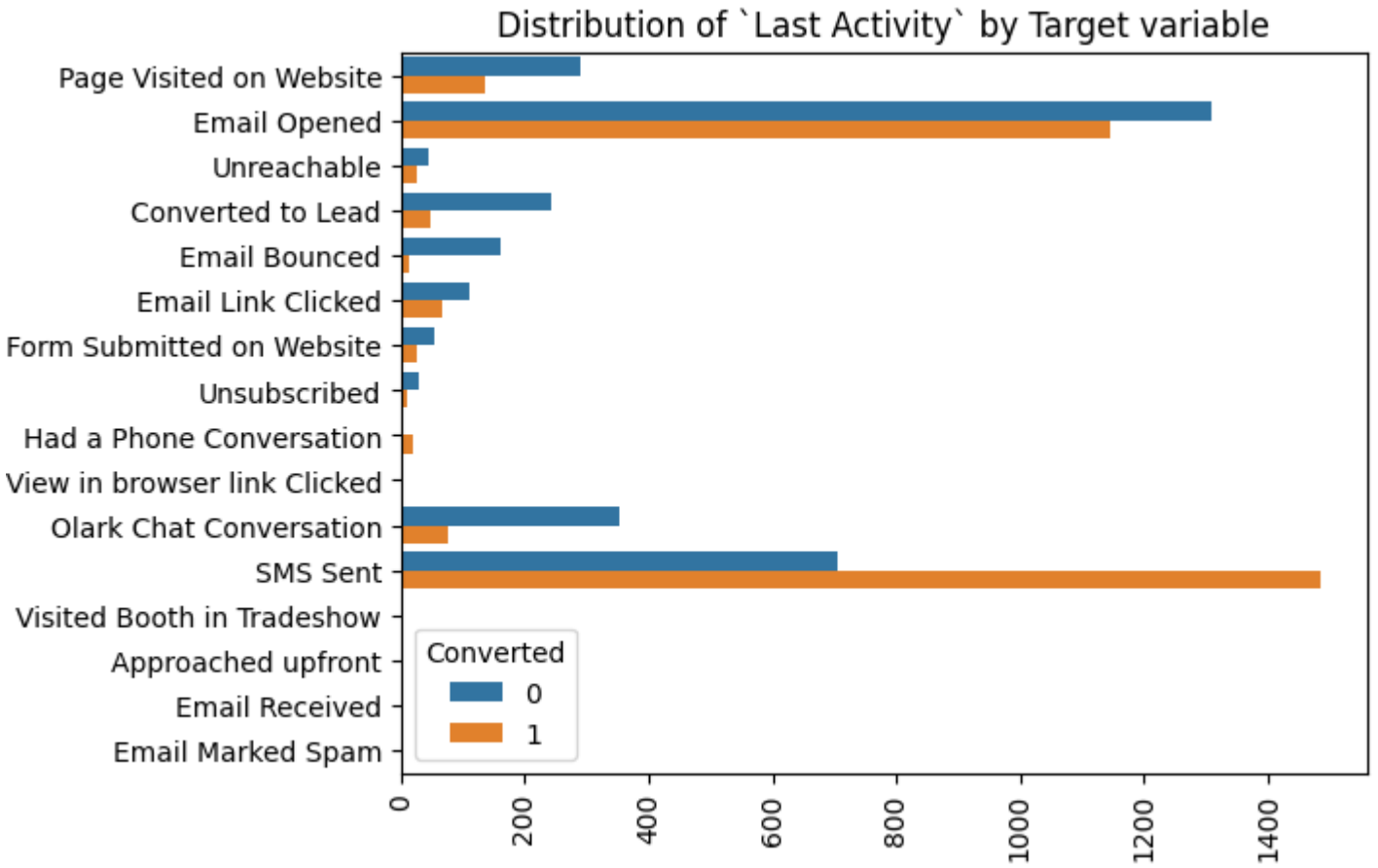
	Converted	Not Converted	Rate
Welingak Website	127	2	0.98
Reference	410	33	0.93
Click2call	3	1	0.75
Social Media	1	1	0.50
Google	1015	1033	0.50
Organic Search	381	489	0.44
Olark Chat	380	512	0.43
Direct Traffic	707	1166	0.38
Referral Sites	28	47	0.37
bing	1	2	0.33
Facebook	9	19	0.32
Live Chat	2	NaN	NaN
Pay per Click Ads	NaN	1	NaN
Press_Release	NaN	1	NaN
WeLearn	1	NaN	NaN
testone	NaN	1	NaN



Data Analysis

Last Activity

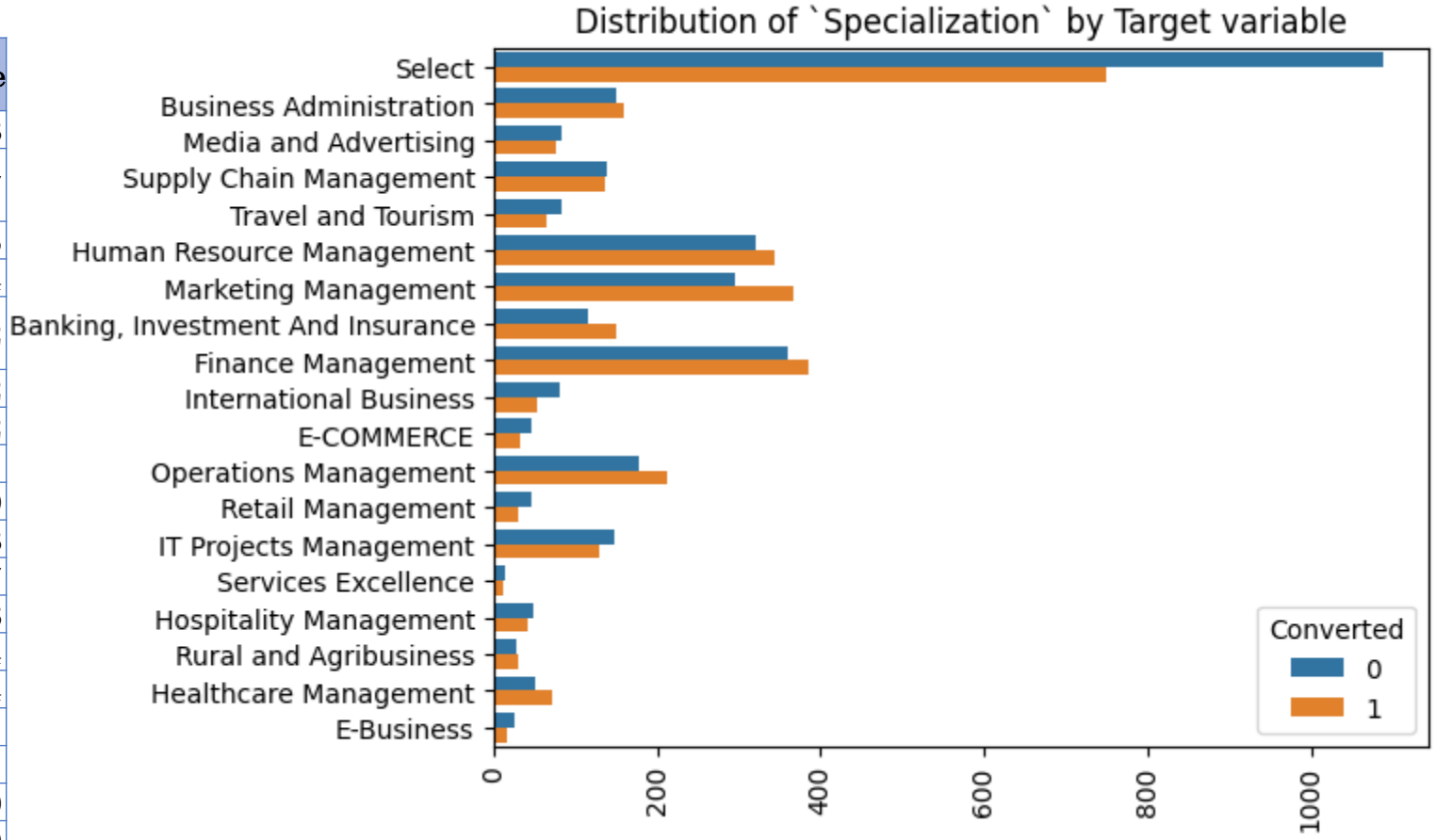
	Converted	Not Converted	Rate
Had a Phone Conversation	19	4	0.83
SMS Sent	1485	704	0.68
Email Opened	1146	1309	0.47
Email Link Clicked	68	110	0.38
Unreachable	27	44	0.38
Form Submitted on Website	26	55	0.32
Page Visited on Website	136	291	0.32
Unsubscribed	11	29	0.28
View in browser link Clicked	1	3	0.25
Olark Chat Conversation	75	353	0.18
Converted to Lead	48	244	0.16
Email Bounced	14	161	0.08
Approached upfront	5	NaN	NaN
Email Marked Spam	2	NaN	NaN
Email Received	2	NaN	NaN
Visited Booth in Tradeshow	NaN	1	NaN



Data Analysis

Specialization

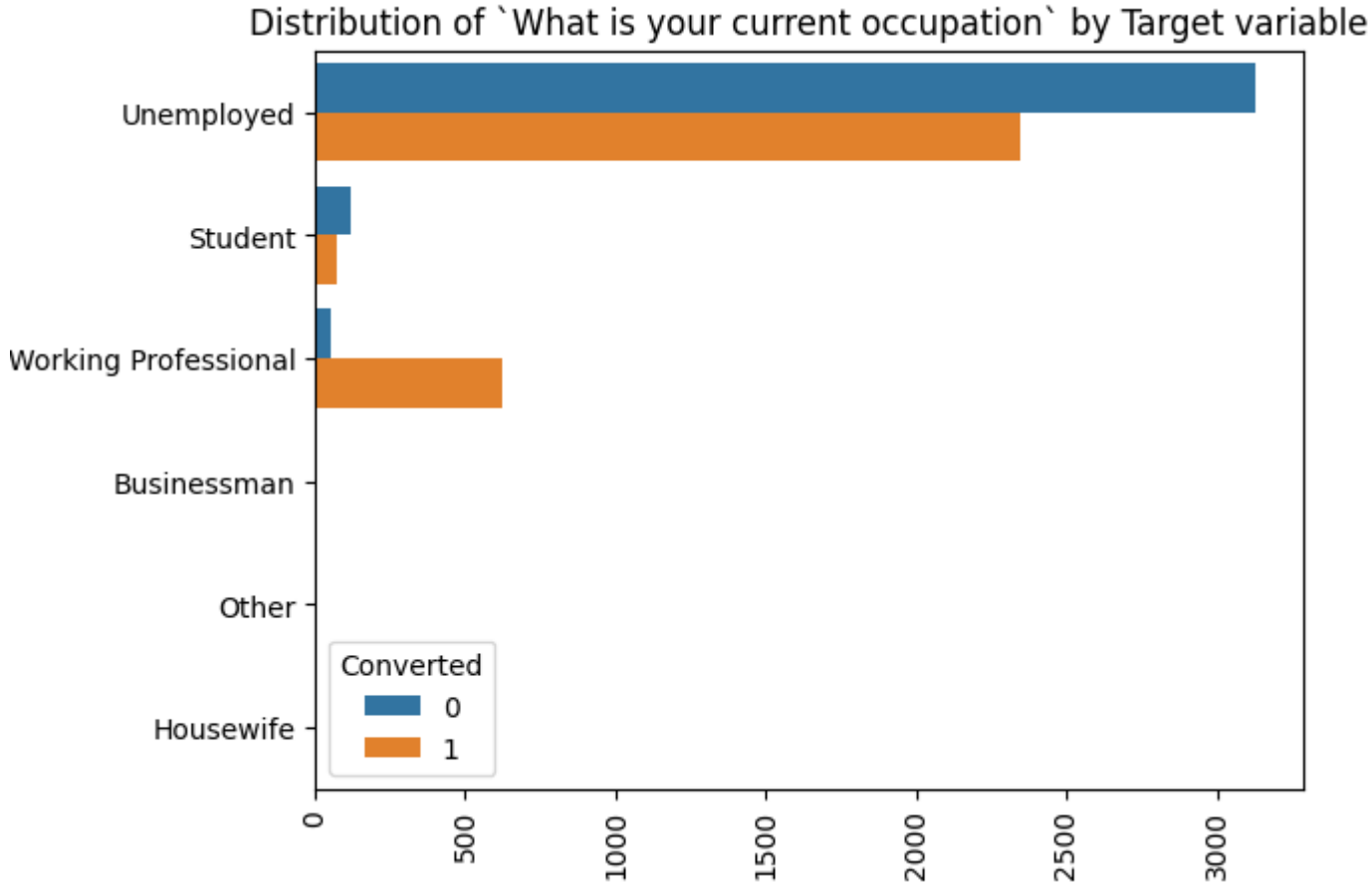
	Converted	Not Converted	Rate
Healthcare Management	71	51	0.58
Banking, Investment And Insurance	151	115	0.57
Marketing Management	367	296	0.55
Operations Management	212	179	0.54
Human Resource Management	345	320	0.52
Finance Management	386	359	0.52
Rural and Agribusiness	30	28	0.52
Business Administration	159	151	0.51
Supply Chain Management	136	139	0.49
Media and Advertising	77	84	0.48
IT Projects Management	130	148	0.47
Hospitality Management	41	49	0.46
Travel and Tourism	66	83	0.44
Services Excellence	11	14	0.44
E-COMMERCE	33	47	0.41
Select	749	1089	0.41
Retail Management	31	47	0.40
International Business	54	82	0.40
E-Business	16	27	0.37



Data Analysis

Current Occupation

	Converted	Not Converted	Rate
Working Professional	622	51	0.92
Businessman	5	2	0.71
Other	9	6	0.60
Unemployed	2346	3130	0.43
Student	74	119	0.38
Housewife	9	NaN	NaN



Model Building

Logistic Regression Model

Summary of the final model

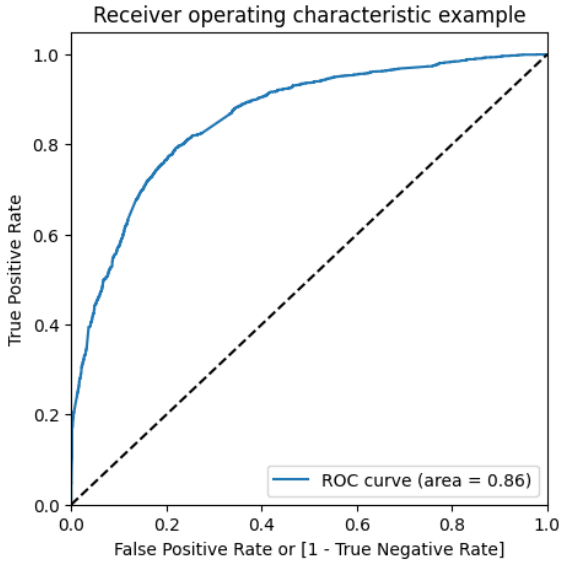
```
 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Converted   No. Observations:                 4461
Model:                            GLM   Df Residuals:                     4447
Model Family:                Binomial   Df Model:                           13
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -2094.0
Date:                Fri, 15 Nov 2024   Deviance:                       4188.1
Time:                        04:23:27   Pearson chi2:                 4.60e+03
No. Iterations:                     7   Pseudo R-squ. (CS):             0.3601
Covariance Type:            nonrobust
==============================================================================================================
                                                 coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------------------------
const                                         -1.9383      0.112    -17.285      0.000      -2.158      -1.719
TotalVisits                                    2.7258      0.592      4.602      0.000       1.565       3.887
Total Time Spent on Website                    4.1910      0.181     23.109      0.000       3.836       4.546
Page Views Per Visit                          -1.2300      0.432     -2.848      0.004      -2.076      -0.384
Lead Origin_Lead Add Form                      3.6094      0.240     15.044      0.000       3.139       4.080
Lead Source_Olark Chat                         1.3427      0.140      9.624      0.000       1.069       1.616
Lead Source_Welingak Website                   1.8585      0.752      2.471      0.013       0.385       3.332
Do Not Email_Yes                              -1.5632      0.191     -8.177      0.000      -1.938      -1.189
Last Activity_Converted to Lead               -0.9346      0.225     -4.163      0.000      -1.375      -0.495
Last Activity_Had a Phone Conversation         2.5539      0.903      2.830      0.005       0.785       4.323
Last Activity_Olark Chat Conversation         -1.1728      0.182     -6.442      0.000      -1.530      -0.816
Last Activity_SMS Sent                         0.9932      0.083     11.945      0.000       0.830       1.156
What is your current occupation_Working Professional     2.6613      0.198     13.455      0.000       2.274       3.049
Last Notable Activity_Unreachable              3.1375      1.066      2.943      0.003       1.048       5.227
==============================================================================================================
```

Final list of features and their VIF values

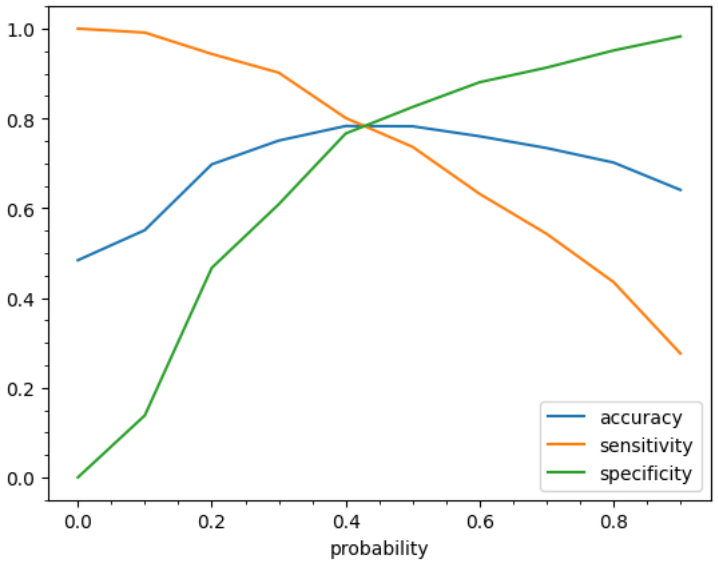
Features	VIF
Page Views Per Visit	4.15
TotalVisits	3.52
Total Time Spent on Website	2.05
Last Activity_SMS Sent	1.57
Lead Origin_Lead Add Form	1.49
Lead Source_Welingak Website	1.32
Lead Source_Olark Chat	1.22
What is your current occupation_Working Profes...	1.21
Last Activity_Olark Chat Conversation	1.19
Do Not Email_Yes	1.06
Last Activity_Converted to Lead	1.03
Last Activity_Had a Phone Conversation	1.01
Last Notable Activity_Unreachable	1.01

Model Evaluation – Train Data

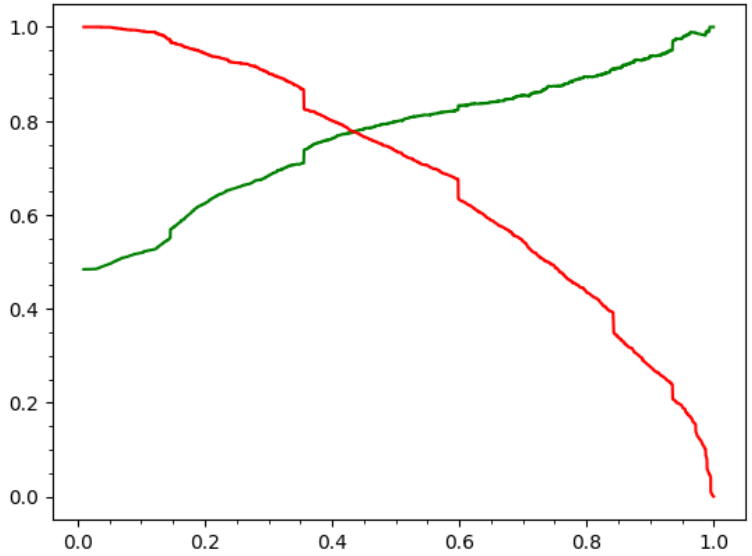
ROC Curve



Optimal Cutoff Point



Precision vs Recall



Optimal Cutoff Value identified	0.43
Accuracy	0.78
Precision	0.77
Recall	0.78
Sensitivity	0.78
Specificity	0.79

Model Prediction – Test Data

Optimal Cutoff Value identified	0.43
Accuracy	0.80
Precision	0.79
Recall	0.79
Sensitivity	0.79
Specificity	0.81

Final Feature List

Feature	Description
TotalVisits	The total number of visits made by the customer on the website.
Total Time Spent on Website	The total time spent by the customer on the website.
Page Views Per Visit	Average number of pages on the website viewed during the visits.
Lead Origin_Lead Add Form	The origin identifier with which the customer was identified to be a lead. Includes API, Landing Page Submission, etc.
Lead Source_Olark Chat	The source of the lead. Includes Google, Organic Search, Olark Chat, etc.
Lead Source_Welingak Website	The source of the lead. Includes Google, Organic Search, Olark Chat, etc.
Do Not Email_Yes	An indicator variable selected by the customer wherein they select whether of not they want to be emailed about the course or not.
Last Activity_Converted to Lead	Last activity performed by the customer. Includes Email Opened, Olark Chat Conversation, etc.
Last Activity_Had a Phone Conversation	Last activity performed by the customer. Includes Email Opened, Olark Chat Conversation, etc.
Last Activity_Olark Chat Conversation	Last activity performed by the customer. Includes Email Opened, Olark Chat Conversation, etc.
Last Activity_SMS Sent	Last activity performed by the customer. Includes Email Opened, Olark Chat Conversation, etc.
What is your current occupation_Working Professional	Indicates whether the customer is a student, umemployed or employed.
Last Notable Activity_Unreachable	The last notable activity performed by the student.

Insights and Recommendation

- A significant proportion of leads originated from **India**, with **Mumbai** as the primary city.
- Many fields in the dataset contained **invalid placeholder** values (e.g., "Select"), which rendered them unusable.
- The leads were **evenly distributed** across various specializations, indicating the course catalog appeals to diverse professional backgrounds.
- A large share of leads came from **unemployed** individuals, followed by **working professionals**.
- **Page visits** and **time spent** website metrics were strong predictors of lead conversion.
- Communication modes such as phone calls and **SMS** remain dominant in engaging leads.

To enhance lead conversions and overall effectiveness of the course offerings

- Maintain an **informative website**
- Enhancing the website's content and **user experience** (UX) for sustained success.
- Improve **mandatory data field** completion (to avoid entries like ‘Select’)
- Offer more **attractive pricing** for targeted demographics