# LEAD SCORING CASE STUDY

## ANALYSIS OF EDUCATION X LEADS TO PREDICT HOTLEADS' CONVERSION

# PROBLEM STATEMENT

**Identifying Hot leads and their conversion**

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| When an individual interested in our courses completes a form with their email address or phone number, they are classified as a lead. | The most promising leads, commonly known as "Hot Leads," are your best opportunities for success. | After acquiring leads, the sales team begins making calls and sending emails. Some of these leads get converted, but most do not. | The typical lead conversion rate at X Education is approximately 30%. | The company aims to increase its lead conversion rate to approximately 80%. |

# BUSINESSOBJECTIVE

Objective of the Case study

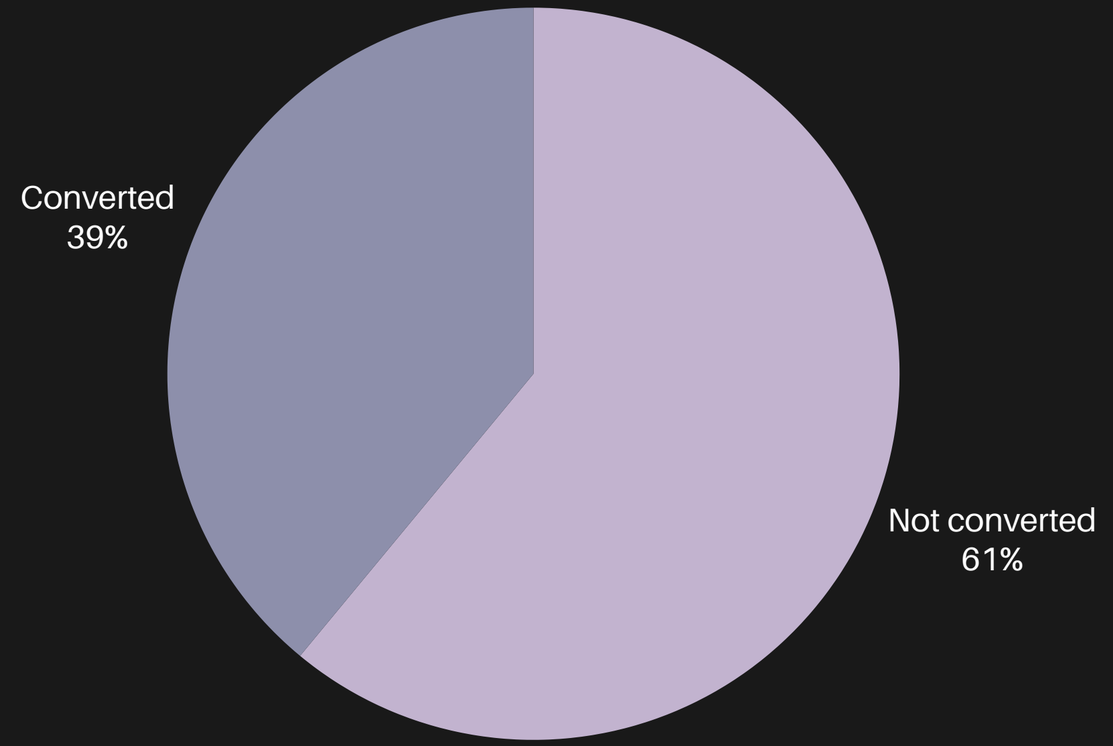
1. Develop a logistic regression model to assign lead scores, with a higher score indicating a hotter lead.
2. Provide insights and recommendations along with answers to company problems identified, using the logistic regression model.

# DATA PROVIDED

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

Target column: **CONVERTED**
1 – Lead is converted to a customer
2 – Lead is not converted to a customer.

The ratio of converted to non-converted in the data is 1:1.6.

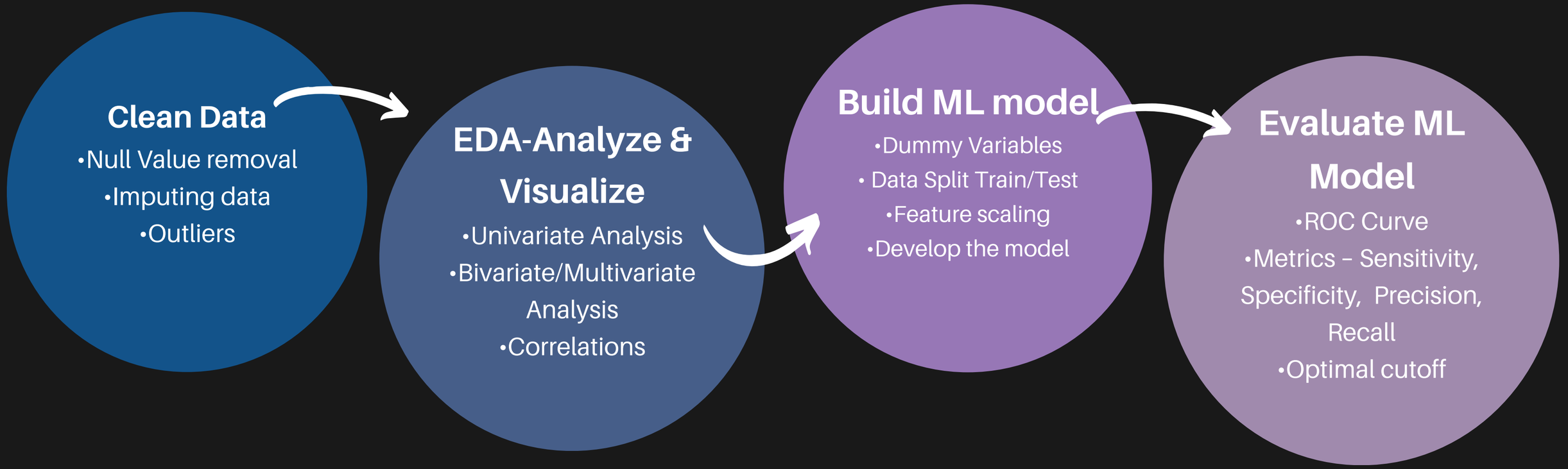


# ML MODEL BUILDING

(Logistic Regression)

Logistic regression, also known as logit regression, estimates the parameters of a logistic model, specifically the coefficients in linear or nonlinear combinations.

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



# DATA CLEANING & PREPARATION

Dropping columns

Columns with 30% or more missing values were removed because they contained significant gaps.

Categorical columns where most or all rows had the same values were dropped, as they do not contribute to the analysis.

For columns with fewer missing values, only the rows containing null values were removed.

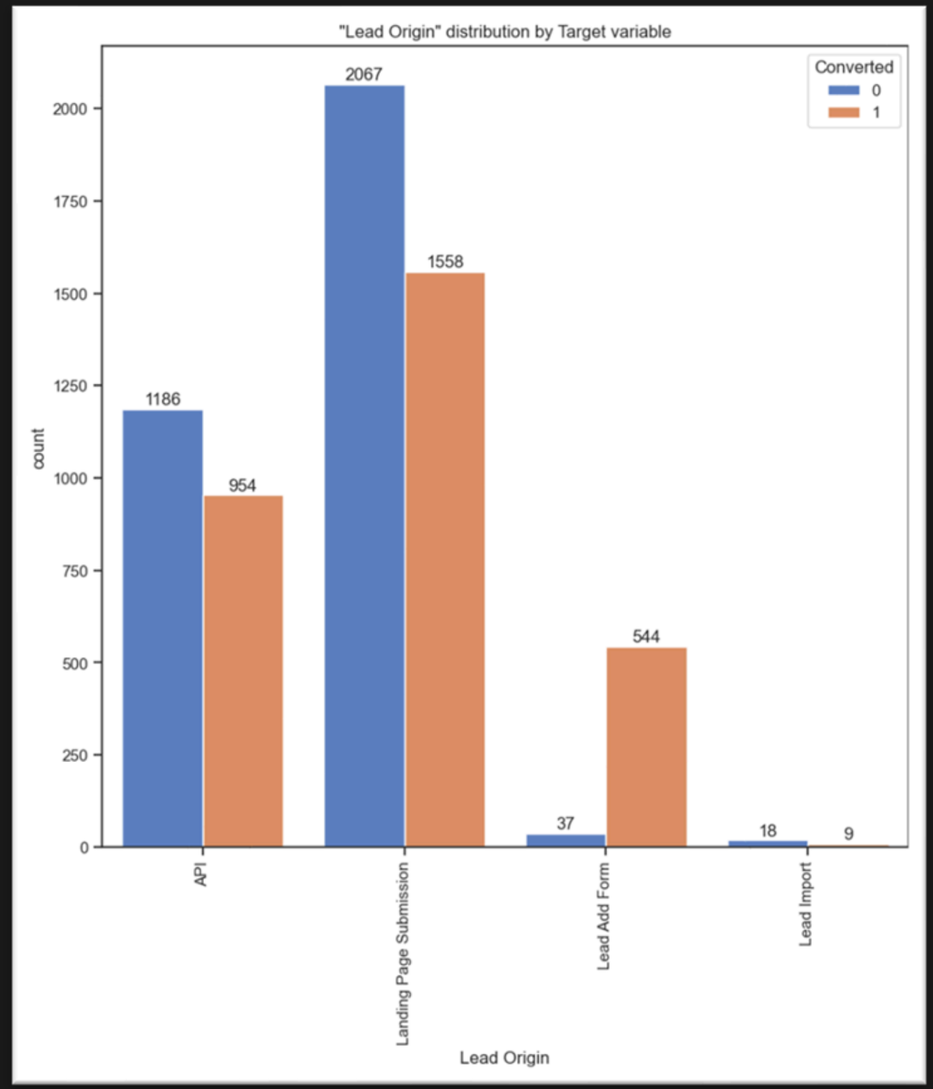
After data cleanup, there were 6,373 rows and 12 columns left from an original 9,240 rows and 37 columns, indicating that nearly 30% of the data was removed during the cleanup.

# EDA- ANALYZE & VISUALIZE

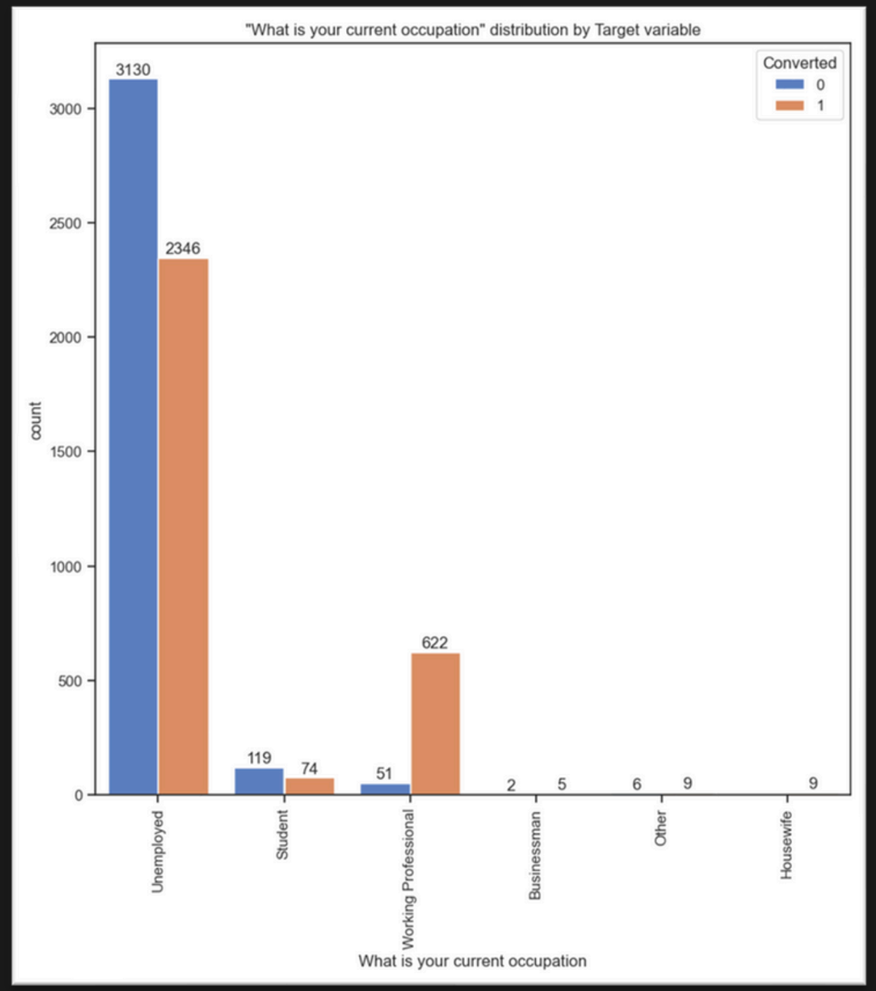
The identifier that indicates how a customer was recognized as a lead, such as via API or landing page submission.

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

The 'Lead Add Form' and 'References' sections have a high conversion rate.



# EDA- ANALYZE & VISUALIZE




## Occupation

This indicates whether the customer is a student, unemployed, or employed.

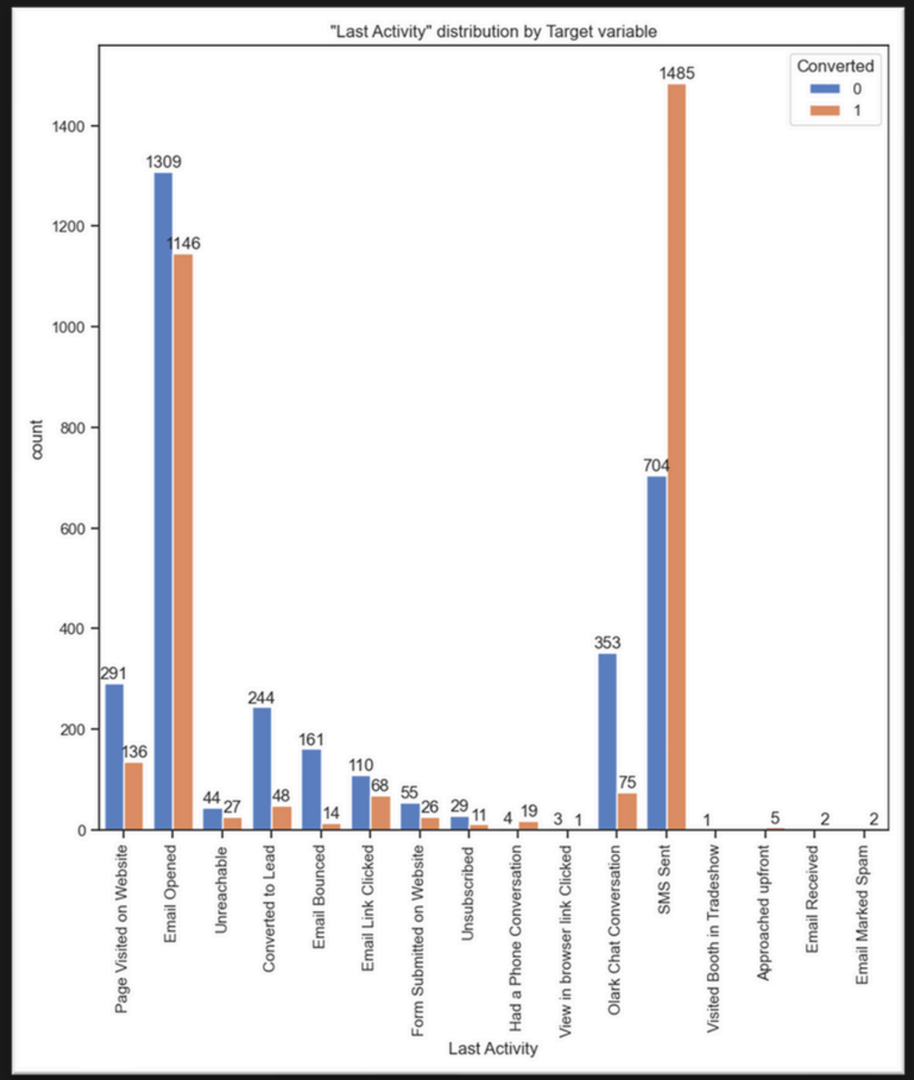
While most of the leads are unemployed, they have a conversion rate of only 42%.

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

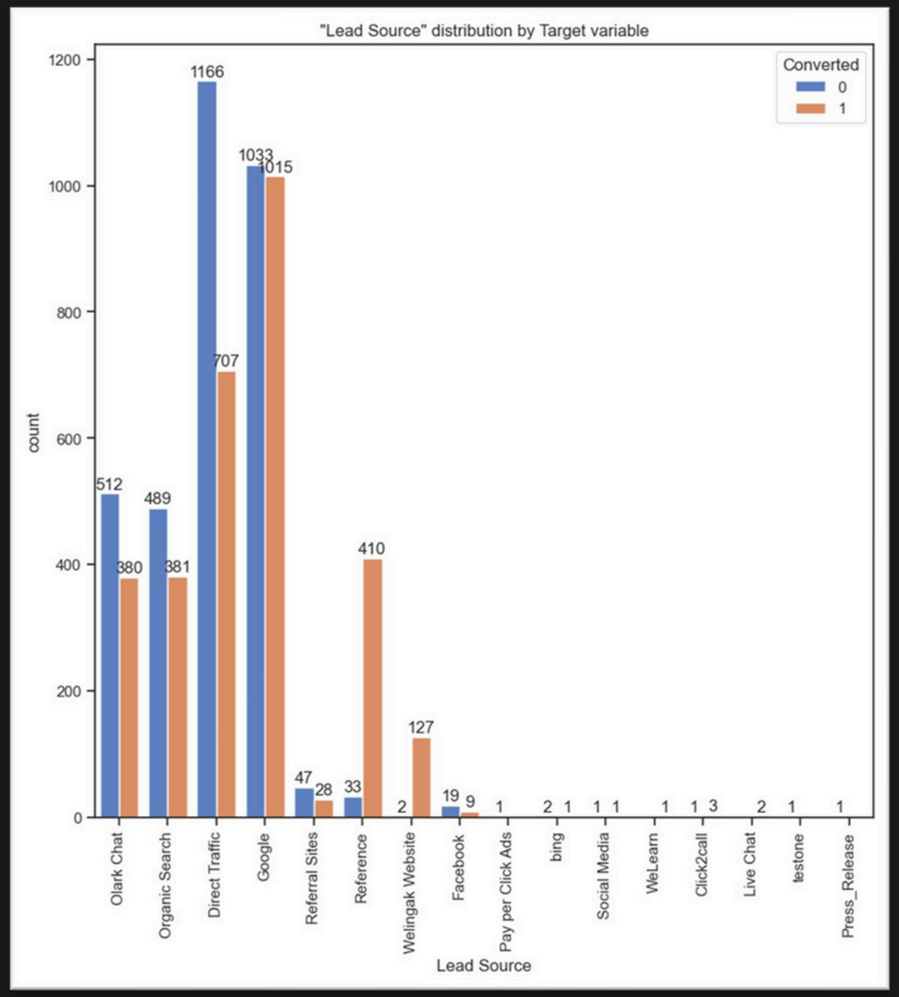
# EDA- ANALYZE & VISUALIZE

| Last Activity | Conversion Rate % |
|---|---|
| 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 |

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



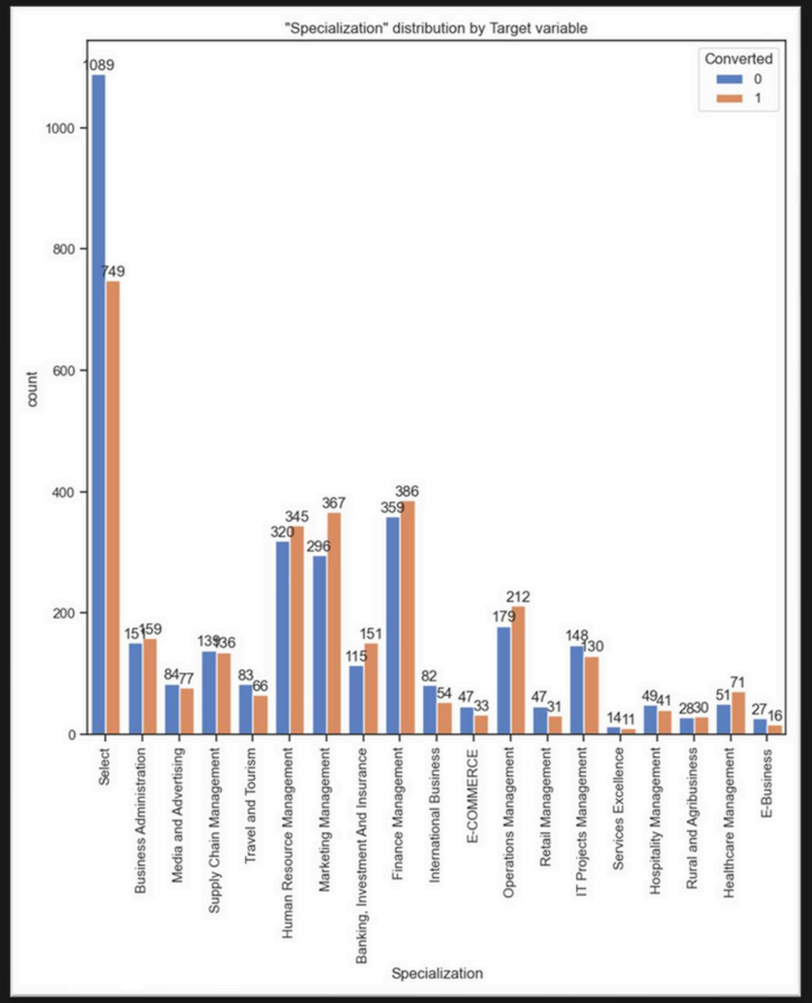
# EDA- ANALYZE & VISUALIZE




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

| Lead Source | Conversion Rate % |
|---|---|
| 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 & VISUALIZE

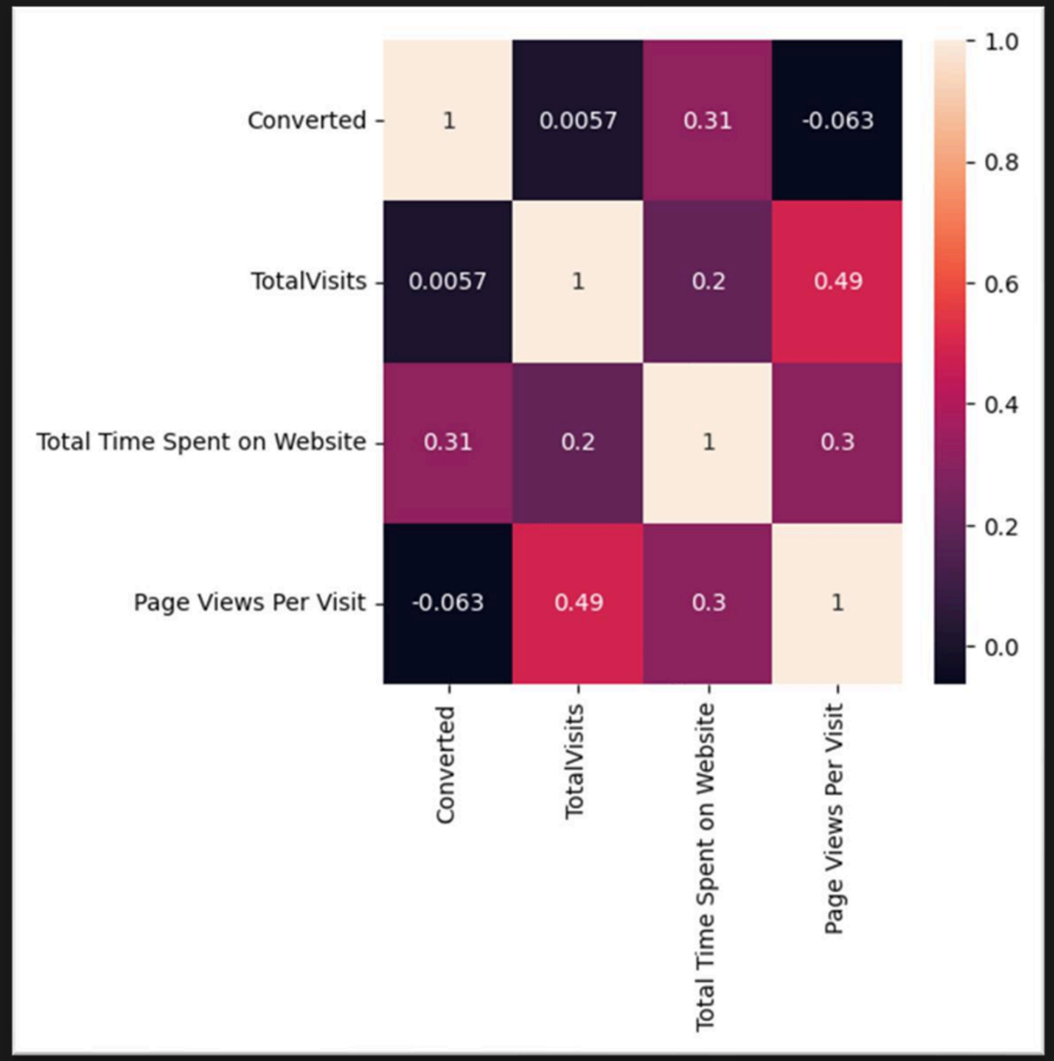



The industry domain in which the customer worked before.

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

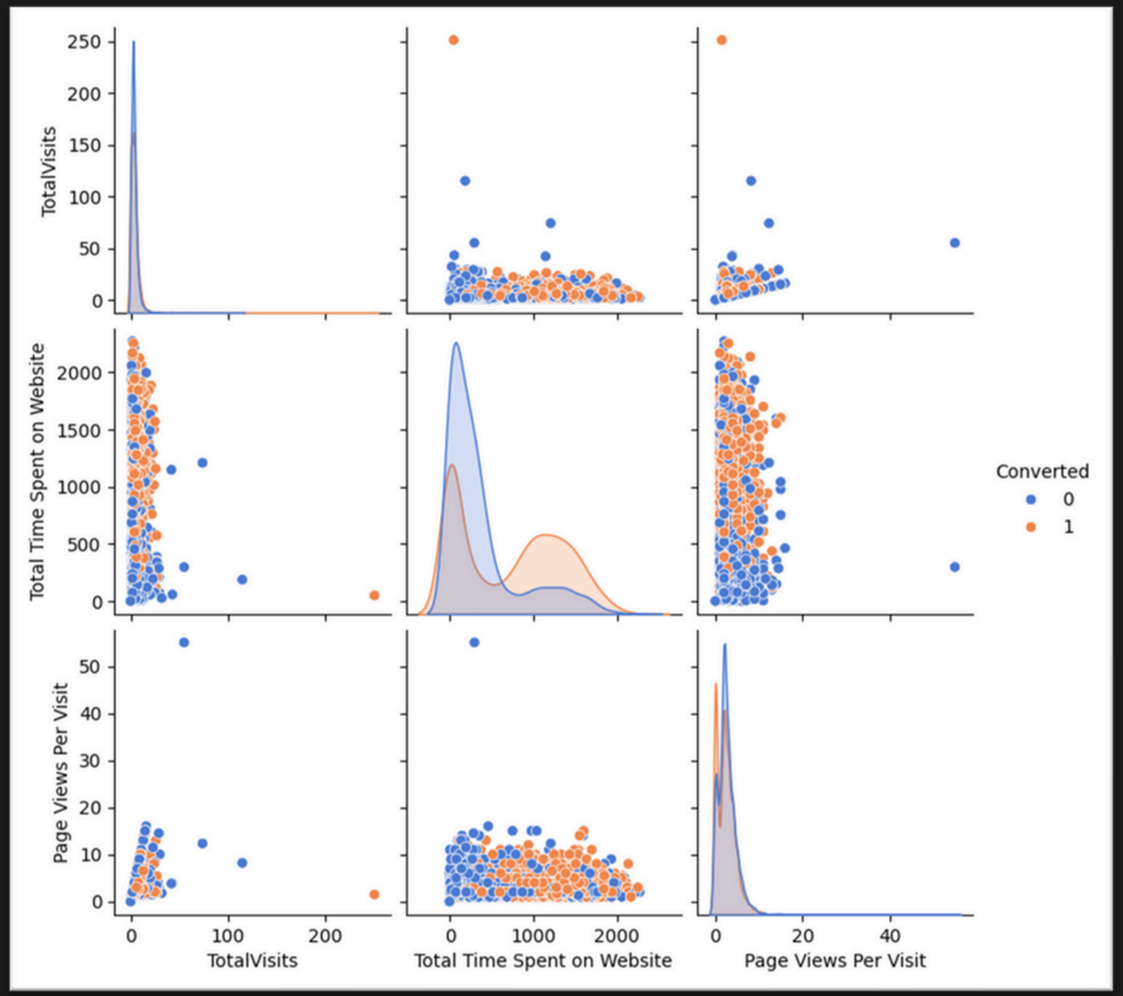
| Specialization | Conversion Rate % |
|---|---|
| Healthcare Management | 58.196721 |
| Banking, Investment And Insurance | 56.766917 |
| Marketing Management | 55.354449 |
| Operations Management | 54.219949 |
| Human Resource Management | 51.879699 |
| Finance Management | 51.812081 |
| Rural and 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 & VISUALIZE




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

# EDA- ANALYZE & VISUALIZE




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

# EDA- ANALYZE & VISUALIZE

Generalized Linear Model Regression Results

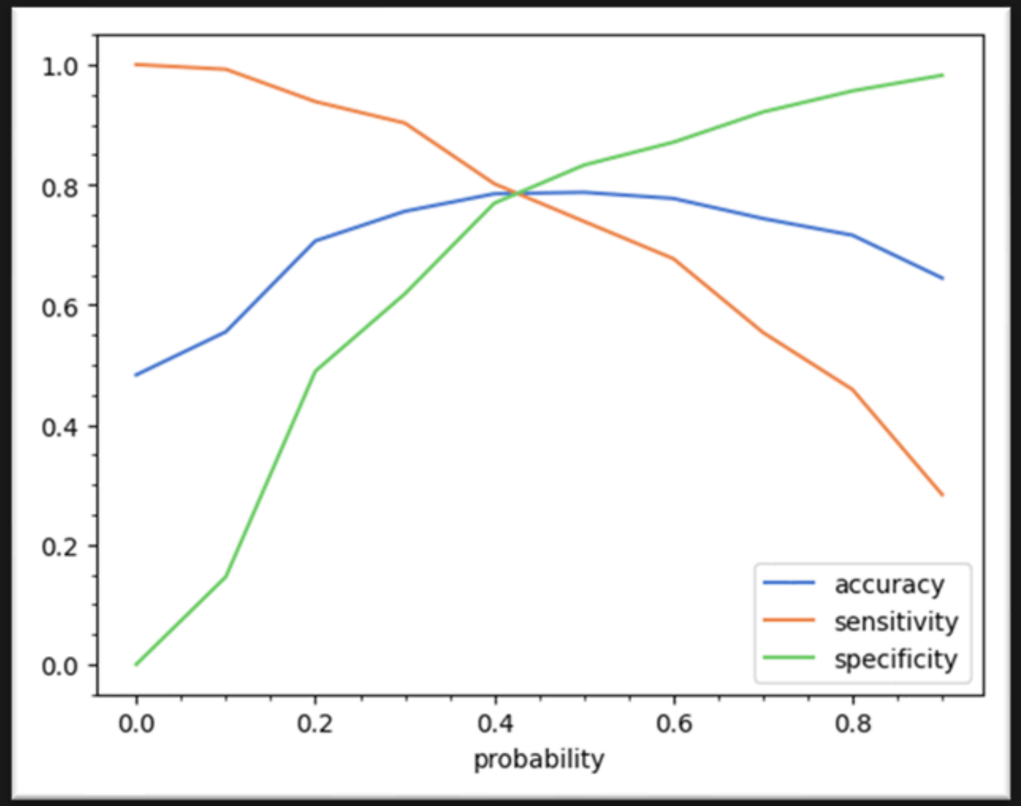
| | | | |
|---|---|---|---|
| Dep. Variable: | Converted | No. Observations: | 4461 |
| Model: | GLM | Df Residuals: | 4449 |
| Model Family: | Binomial | Df Model: | 11 |
| Link Function: | Logit | Scale: | 1.0000 |
| Method: | IRLS | Log-Likelihood: | -2050.4 |
| Date: | Fri, 15 Nov 2024 | Deviance: | 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 |

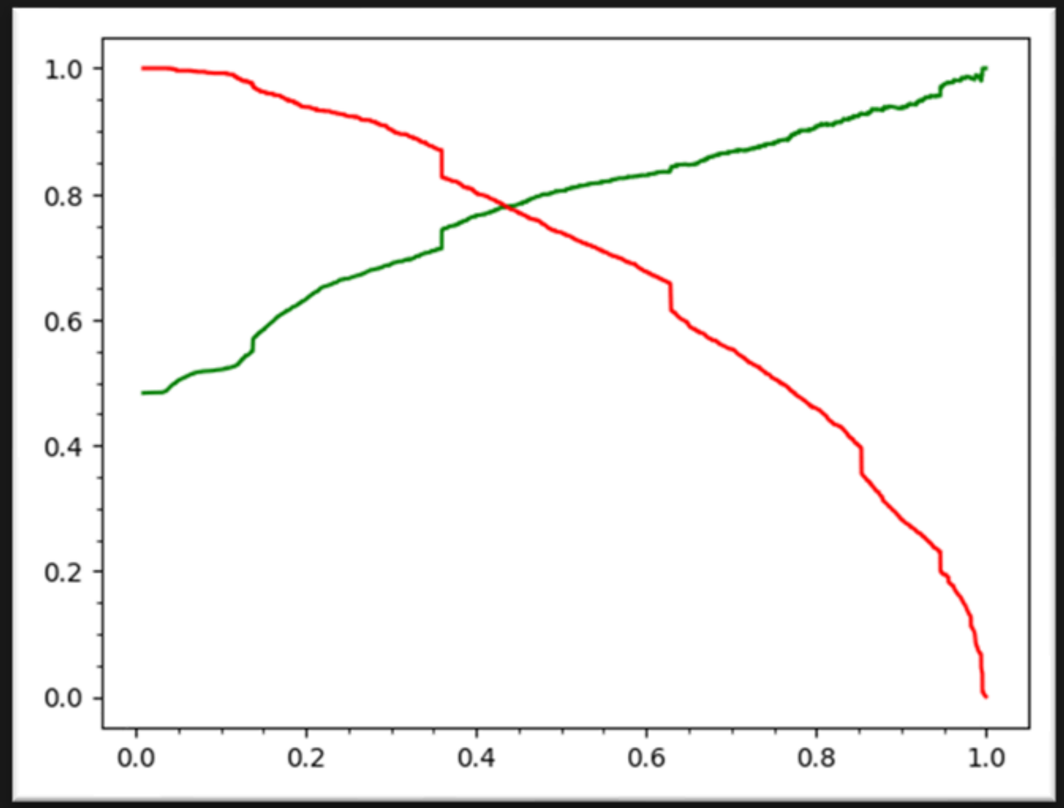
| Features | VIF |
|---|---|
| Total Time Spent on Website | 1.65 |
| Last Activity_SMS Sent | 1.49 |
| TotalVisits | 1.36 |
| Lead Source_Olark Chat | 1.22 |
| What is your current occupation_Working Profes... | 1.22 |
| Last Activity_Olark Chat Conversation | 1.19 |
| Lead Source_Reference | 1.14 |
| 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 |

- 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

# EDA- ANALYZE & VISUALIZE




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




•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%

## Test Data set

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

# INSIGHTS

• About 70% of leads indicated that they were interested in taking the courses to further their careers.

Website metrics like page visits and time spent were key indicators of lead conversion.

•A logistic regression model was created to predict lead conversion.

The model identified an optimal conversion prediction threshold of 0.42.

# RECOMMENDATIONS

Enhance data collection by making critical fields mandatory to prevent unusable entries.

Enhance the website's user experience and content to increase engagement.

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

# Thank You

Radha Padwal
Himanshu Tulsani
Sagarmay Biswas