# Lead Scoring Case Study

Presented by:
Santhosh Rajasingaram
Tejaswini Dhurshetty

Data: July 2022

# Problem Statement

X Education offers online courses to industry professionals. The company markets its courses on several websites and search engines like Google. Although 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 improve the efficiency of this process, the organisation aims to identify the most promising prospects, commonly known as "Hot Leads”. If this group of leads is successfully identified, the lead conversion rate should increase as the sales staff will concentrate more on connecting with the prospective leads instead of calling indiscriminately.

# Business Goal

Identify the most promising leads such that customers with higher lead score have a higher conversion chance.

Build an analytical model that can estimate the lead score for the new customer based learning from the past data.

Enhance the usability of the model to reduce the number of unnecessary follow ups and use the human resources efficiently.

# Data

For the construction of the analytical model, a dataset of historical leads including around 9000 data points is supplied. In this instance, the target variable is the column 'Converted,' which indicates whether a previous lead was converted or not. Many of the categorical variables have a "Select" level that must be managed.

# Solution Approach

Load the leads dataset and ensure the format

Clean the Dataset

Explore and Analyse the Dataset

Prepare the Dataset

Split the data into Training and Test Set

Standardize the Training Data

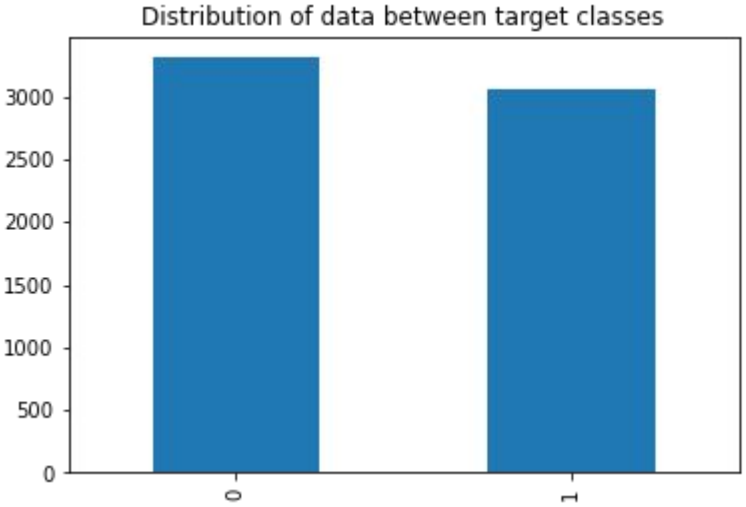
Select the Features to that can contribute to the model accuracy

Build a Logistic regression model to assign a lead score between 0 1nd 100

Evaluate the Model Performance

# Load & Describe Dataset

The leads dataset has 9240 Rows and 37 Columns including the target. The target has 2 classes 0 and 1 which represents not converted and converted respectively. The Plot shows that the dataset is balanced between two classes.



# Data Cleaning

Remove columns with constant value

Remove columns with unique values

Remove columns with more number of constant values

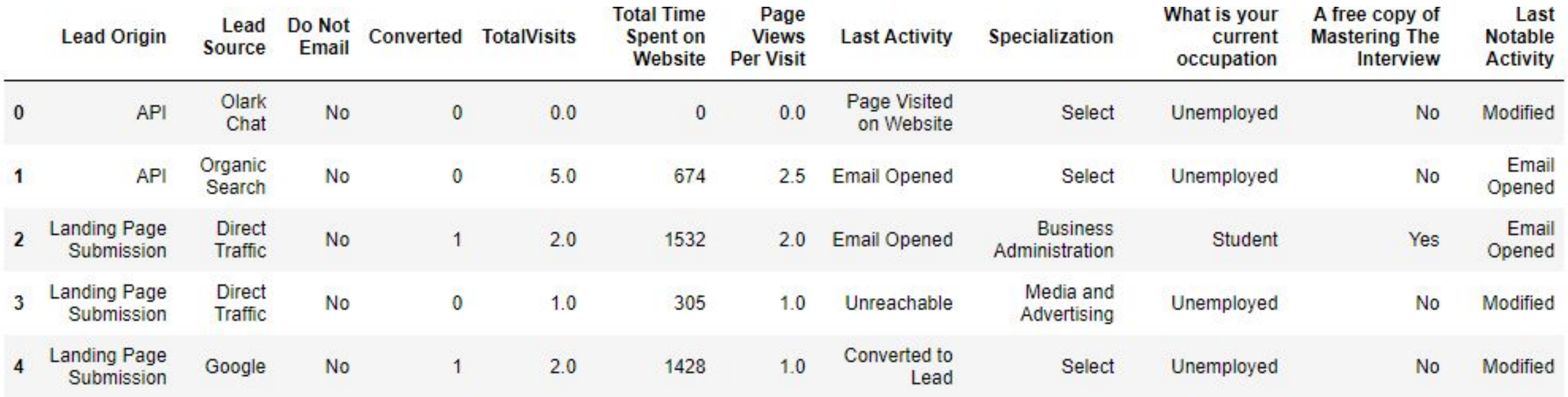
Identify missing values

Drop columns having high percentage of missing values

Impute/Drop columns with less percentage of missing values

# Post Data Cleaning

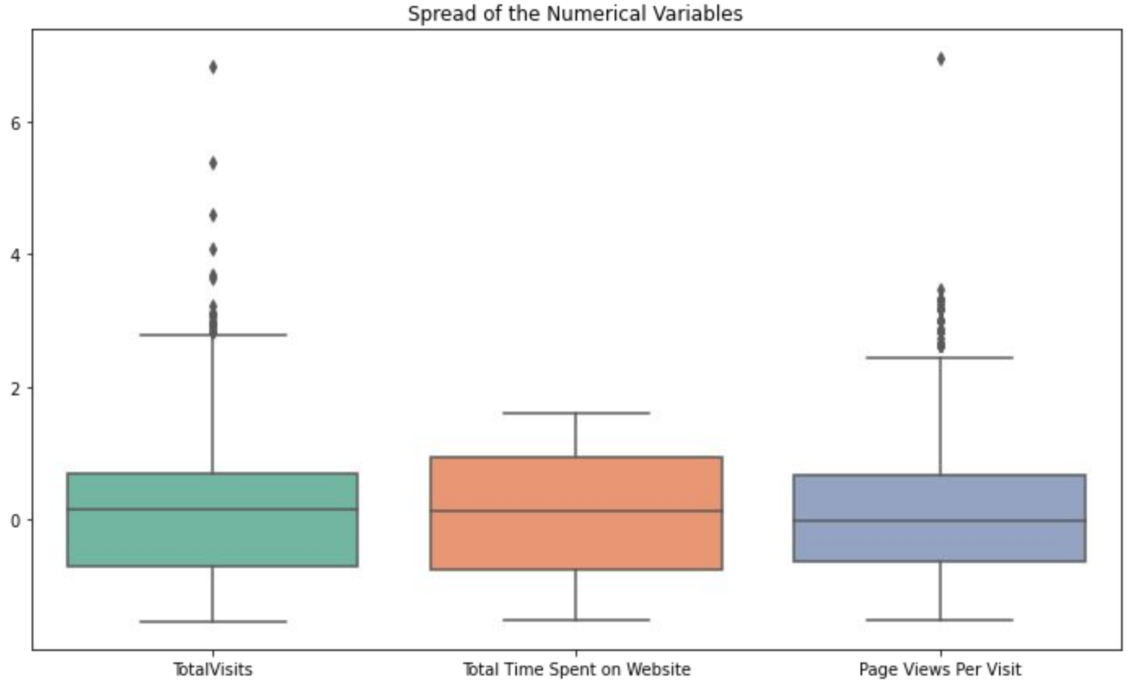
The cleaned dataset contains 6373 rows and 12 columns.



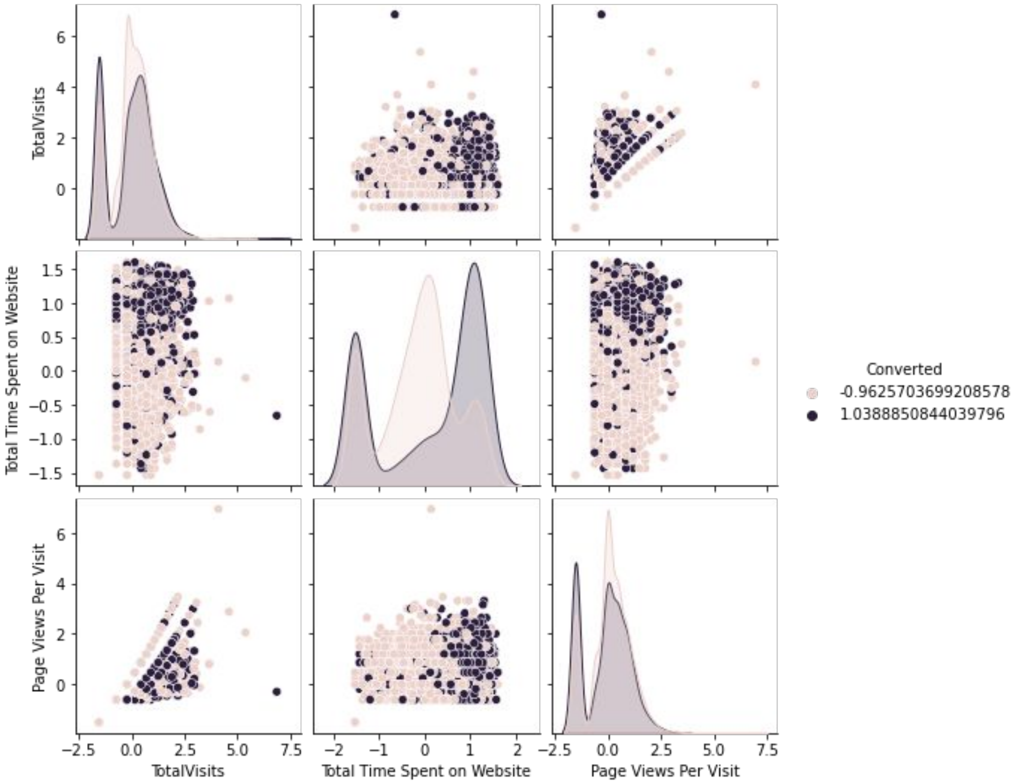
| | Lead Origin | Lead Source | Do Not Email | Converted | TotalVisits | Total Time Spent on Website | Page Views Per Visit | Last Activity | Specialization | What is your current occupation | A free copy of Mastering The Interview | Last Notable Activity |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | API | Olark Chat | No | 0 | 0.0 | 0 | 0.0 | Page Visited on Website | Select | Unemployed | No | Modified |
| 1 | API | Organic Search | No | 0 | 5.0 | 674 | 2.5 | Email Opened | Select | Unemployed | No | Email Opened |
| 2 | Landing Page Submission | Direct Traffic | No | 1 | 2.0 | 1532 | 2.0 | Email Opened | Business Administration | Student | Yes | Email Opened |
| 3 | Landing Page Submission | Direct Traffic | No | 0 | 1.0 | 305 | 1.0 | Unreachable | Media and Advertising | Unemployed | No | Modified |
| 4 | Landing Page Submission | Google | No | 1 | 2.0 | 1428 | 1.0 | Converted to Lead | Select | Unemployed | No | Modified |

# Outlier Analysis

Observing the spread of the numerical variable shows there are no considerable outliers



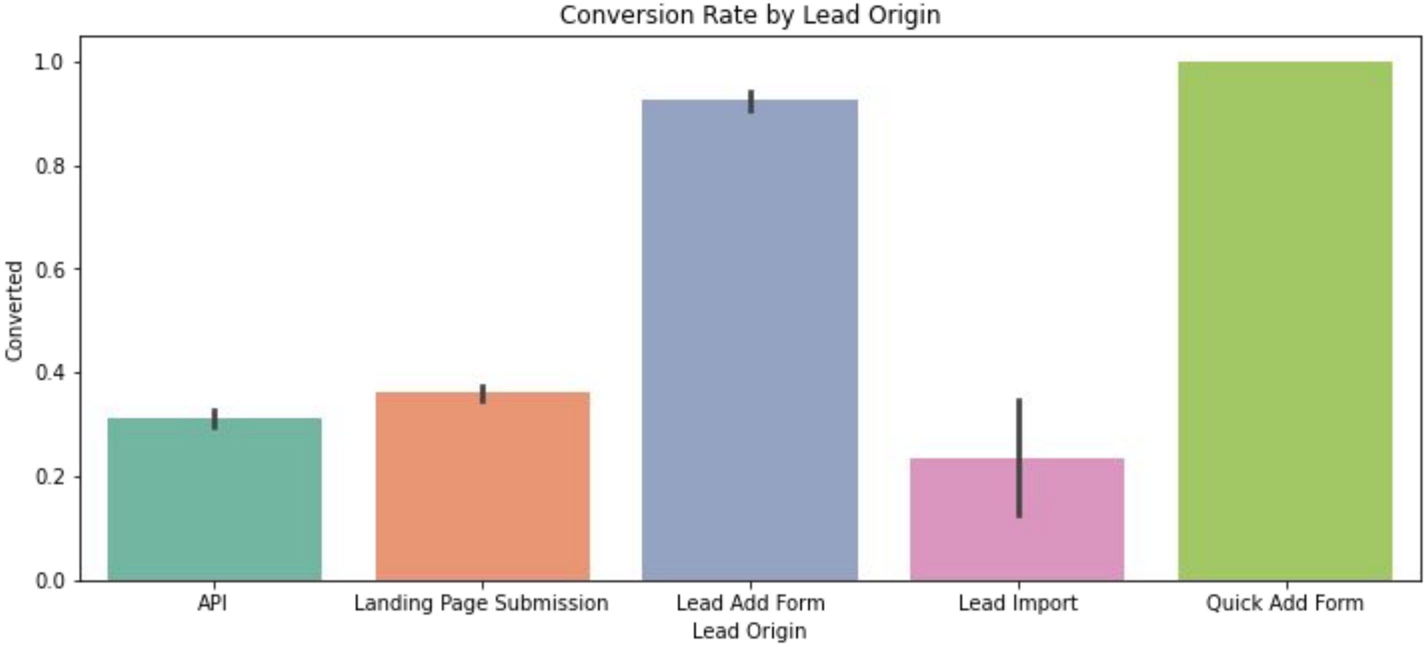
# Numerical Variable Analysis



On visualizing the pair plot of the numerical variable, there is a distinguishable pattern emerges between the two classes of Target (Converted and Not Converted). Hence these numerical variables are potential predictors.

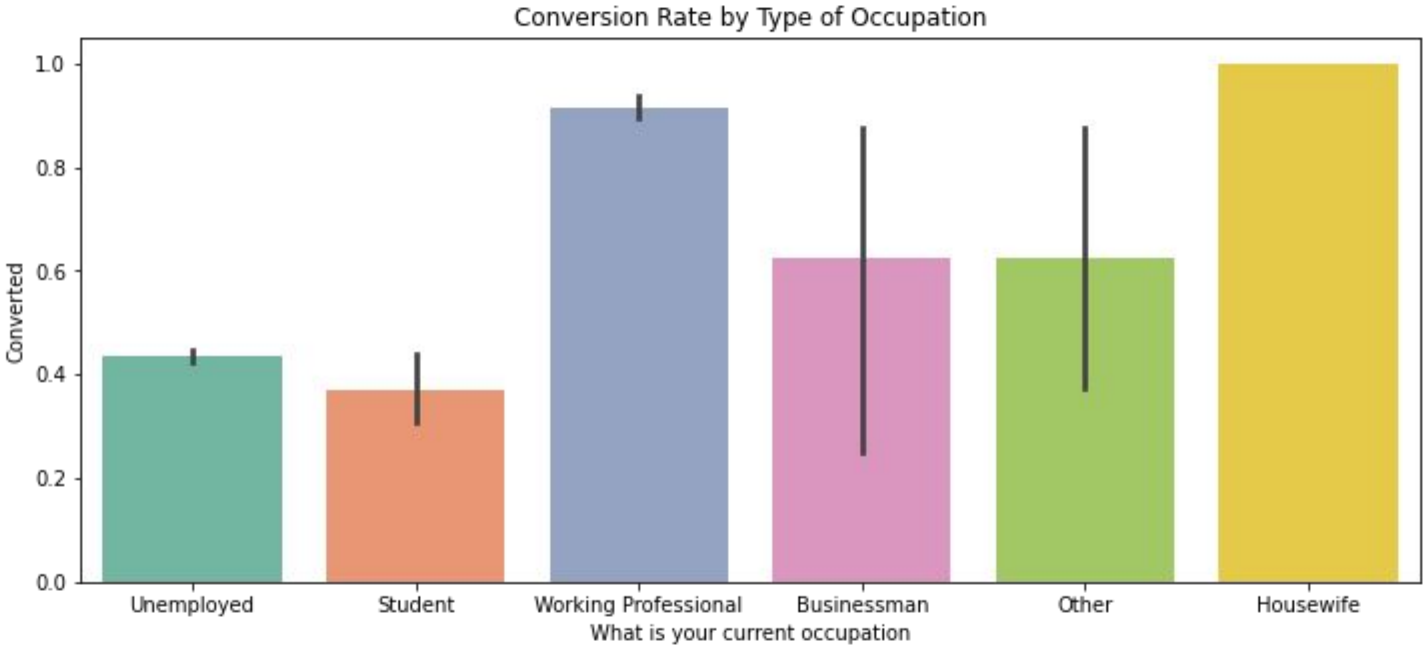
# Categorical Variable Analysis - I

The plot shows that the leads originated from the Add Forms has a good conversion rate



# Categorical Variable Analysis - II

The plot shows that the working professionals and housewives have slightly higher conversion rate than people of other occupations



# Data Preparation

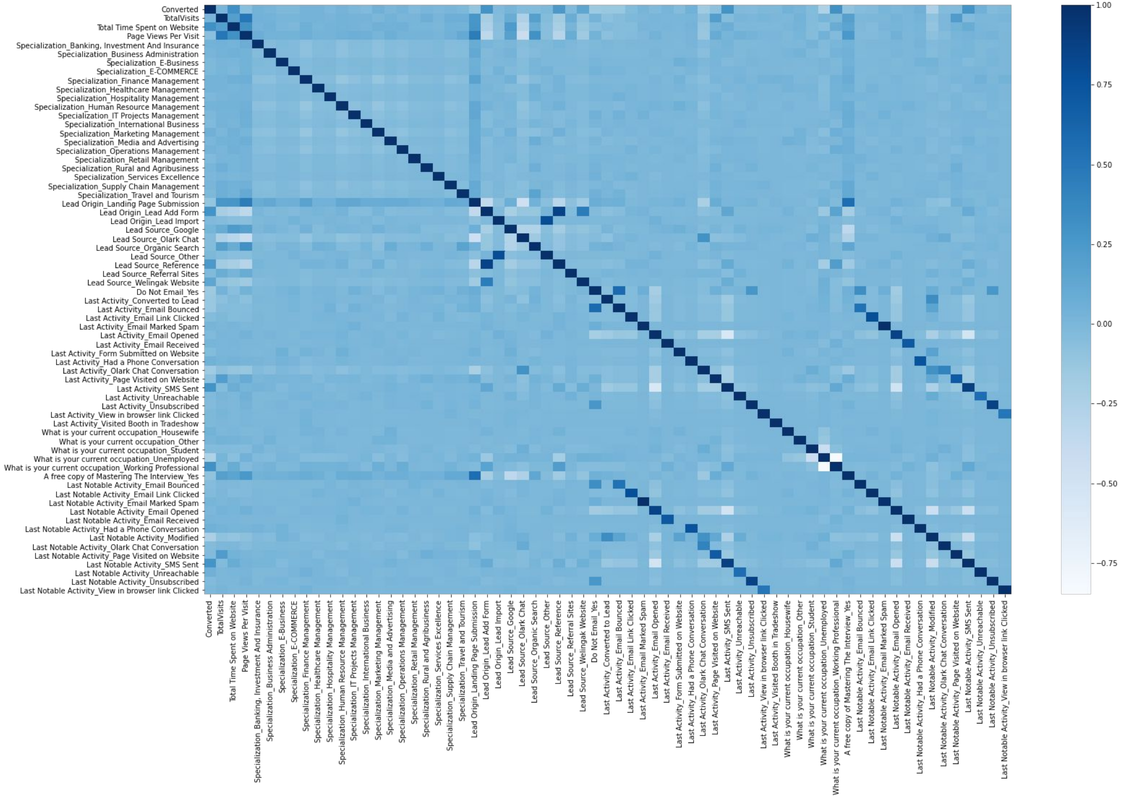
Group Categories having less number of values

Encode categorical Variables to one-hot encoding

Split dataset into Train and Test set into 7:3 Ratio

Normalize the Train dataset

# Observe Correlation



# Build Model

Do automated feature elimination using reduce the number of week predictors in the input dataset.

Build initial model using selected set of features

Analyse the statistical significance of the feature remove the insignificant feature

Iteratively do the feature elimination manually until there are no more insignificant feature to eliminate

Build the model with the final set of features

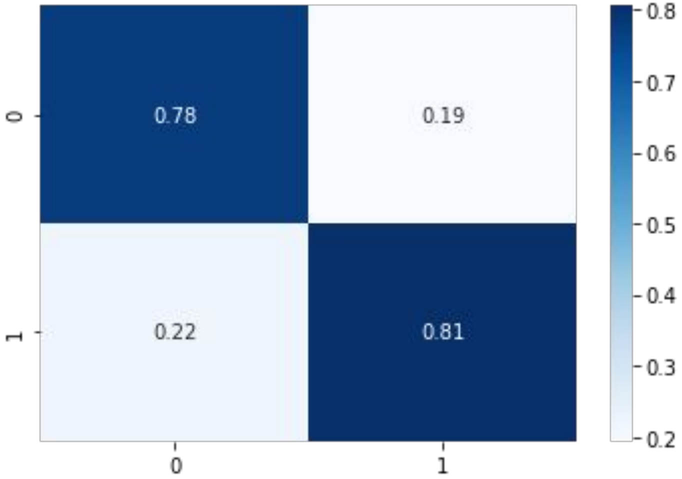
# Model Evaluation

Initially the model is evaluated with a default probability cutoff of 0.5.

Accuracy - 79%

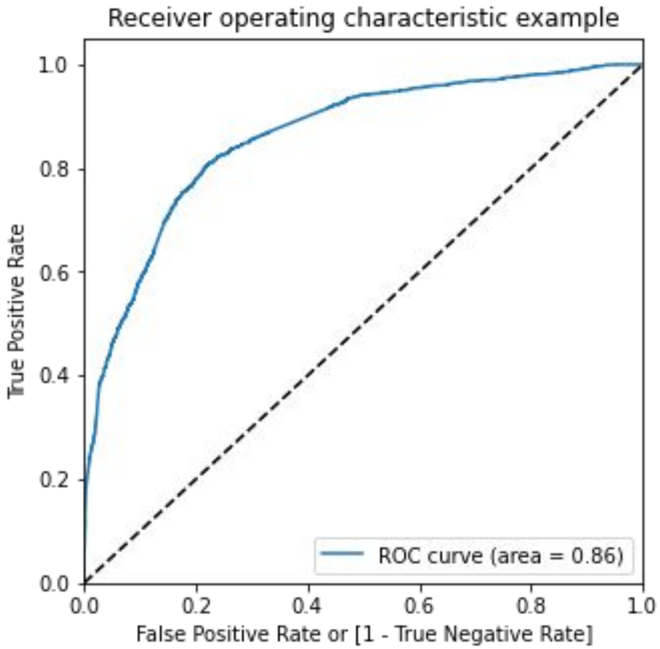
Sensitivity - 74%

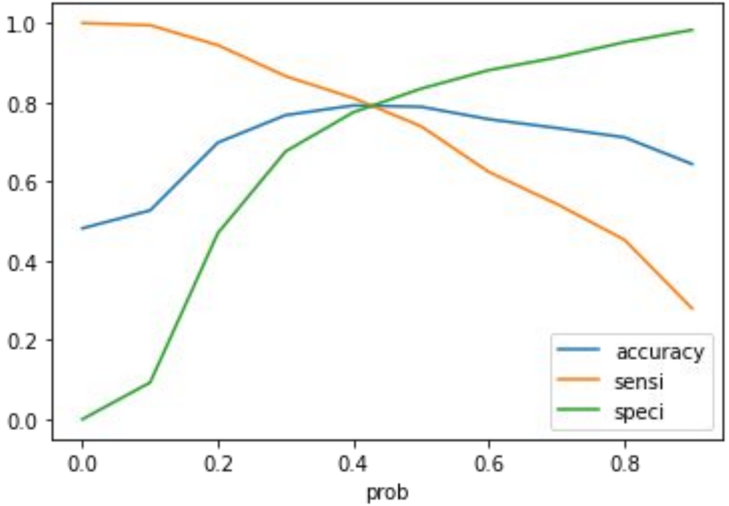
Specificity - 83%



Confusion Matrix

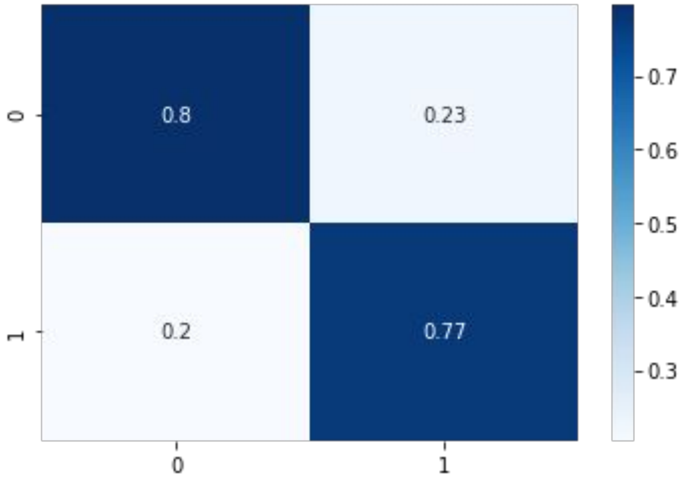
# Optimize the Probability Cutoff





The best probability cutoff can be determined by the intersection point of sensitivity and specificity curve. Here it is 0.42.

# Post Optimization Performance



Confusion Matrix

Model is evaluated with an optimal probability cutoff of 0.42.

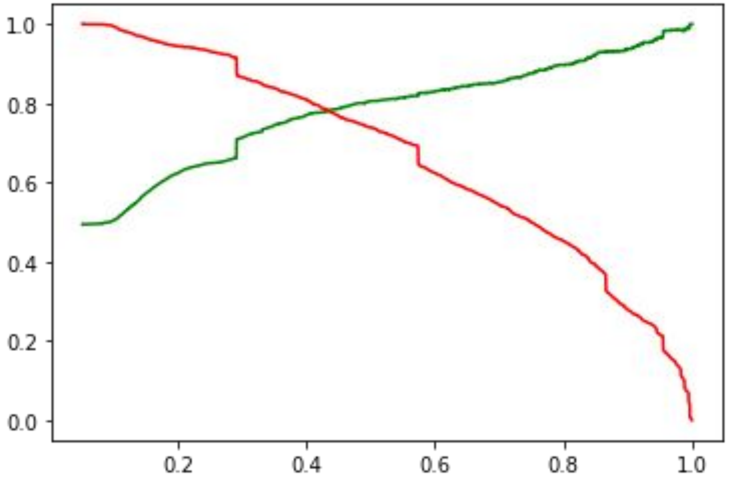
Sensitivity - 78%

Specificity - 79%

The model is much better with almost even sensitivity and specificity

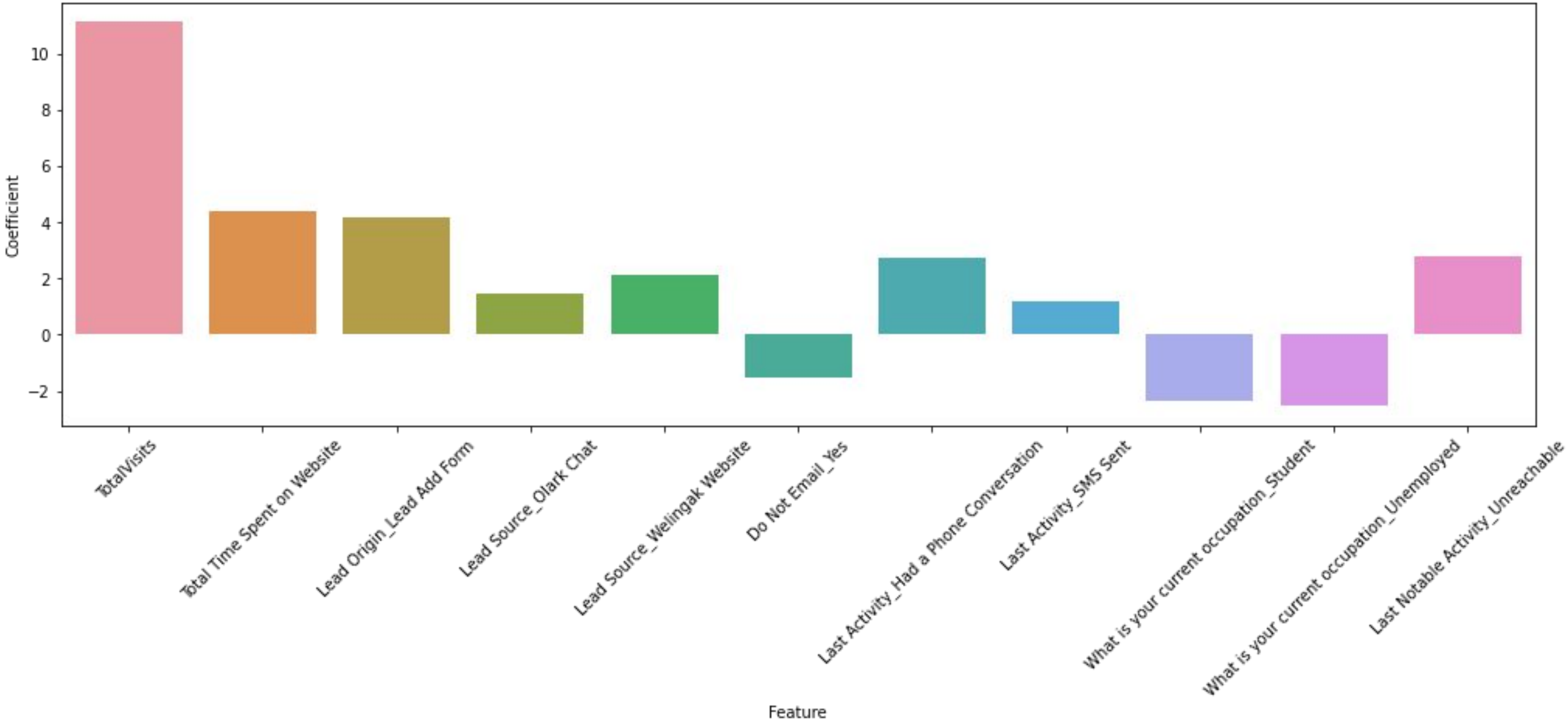
# Precision Recall Score

The optimal threshold can also be determined using Precision-Recall Curve which in this case is also around 4.2



Precision Vs Recall

# Feature Importance



# Results

Three features that contribute the most to the lead conversion probability.

- Total Visits
- Total Time Spent on Website
- Lead Origin Lead Add Form

Two features that negatively influence the likelihood of a lead conversion.

- What is your current occupation Student
- What is your current occupation Unemployed

# Summary

The constructed model is capable of calculating the lead score of consumers based on a specified set of features. Not only is there value for the business in developing the model, but also in using it. Therefore, the organisation should deploy the model, predict potential customers, and make the results easily accessible to staff in order to operate a prosperous business. In addition, the model should be trained frequently and augmented with additional features to maximise the accuracy of its predictions. The business analysts must evaluate several perspectives from which the model might be used to address business challenges.

# Thank You