# Lead Scoring Case Study using logistic regression

By: Satya Ranjan, Srushti Keskar, Swapnil Kapoor

# Problem statement

X Education offers online courses to industry professionals and generates many leads through websites and search engines. However, its lead conversion rate is low, around 30%. To improve this, the company wants to identify "Hot Leads" with a higher potential to convert. By focusing on these leads, the sales team can increase conversion efficiency. The company aims to develop a model that assigns a lead score to each lead, prioritizing those with higher scores for better conversion chances, with a target conversion rate of 80%.

# Business Goal

### Model Development and Purpose

The company seeks to build a model that assigns a lead score to each lead, indicating its potential for conversion. Higher lead scores signify higher chances of conversion. The goal is to achieve an 80%+ conversion rate.

### Identification of Hot Leads

X Education aims to identify the most promising or "hot" leads. A model will be developed to assist in recognizing these leads, with an emphasis on deployment for future use.

### Challenge 3

X needs a model that assigns a score from 0 to 100 to each lead that helps identifying hot leads and improve conversion rates. The model should also be adaptable to peak time actions and optimal resource allocation.

# Business Goal

## Flexibility for Future Constraints

The model must not only predict lead conversion but also handle future constraints like peak times, manpower and adjustments after achieving targets. These aspects should be incorporated into the model to meet future requirements.

## Logistic Regression Approach

A logistic regression model will be used. The model should also be flexible enough to adjust to changing company needs. These adjustments should be documented and included in the final presentation with recommendations.

# Analysis Strategy

1. Importing and inspecting the data
2. Data Cleaning and Preparation
3. Exploratory data analysis
4. Correlation
5. Model Building and Evaluation
6. Prediction on the test set
7. Precision Recall-Review curve
8. Observations
9. Conclusion

# Importing and Inspecting the data

# Importing and inspecting the data

- Once the data is imported in the pandas dataframe, we get to see that the data comprises 9240 rows spread across 37 columns.
- Out of these 37 columns there are only 7 columns with numerical values. Rest of the columns are of categorical type.

# Data Cleaning and Preparation

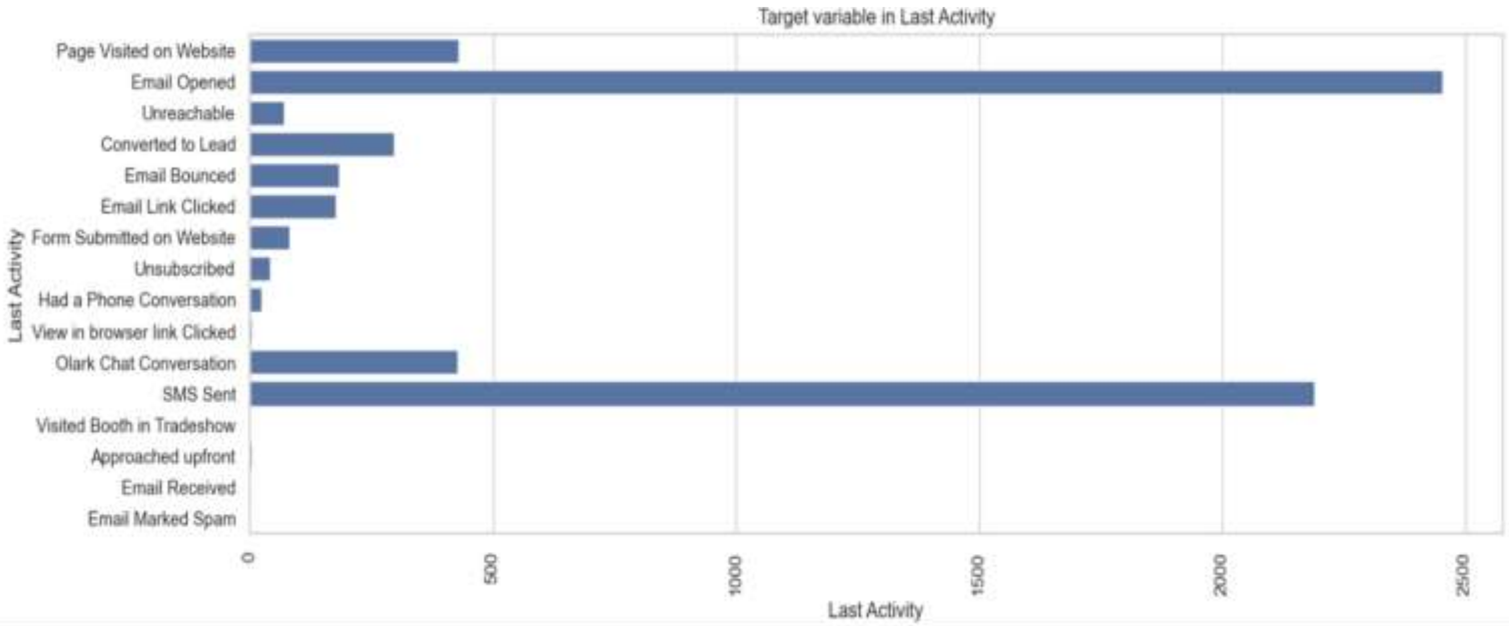
## Data Cleaning and Preparation

- While checking for the null values in the data it seems there are quite a few columns containing null values
- We dropped the columns with more than 30% missing or null values which brought down the number of columns to 31
- Some of the categorical columns had a single value and the count of that value is equal to the number of row, which means all the users have selected the same option so keep these is also not relevant. So, dropped these columns too.
- Now we are left with 26 columns for analysis shown below
- There are some columns with Select as one of the values, we replace that with Unknown
- Now the data looks ready for analysis

# Exploratory data analysis

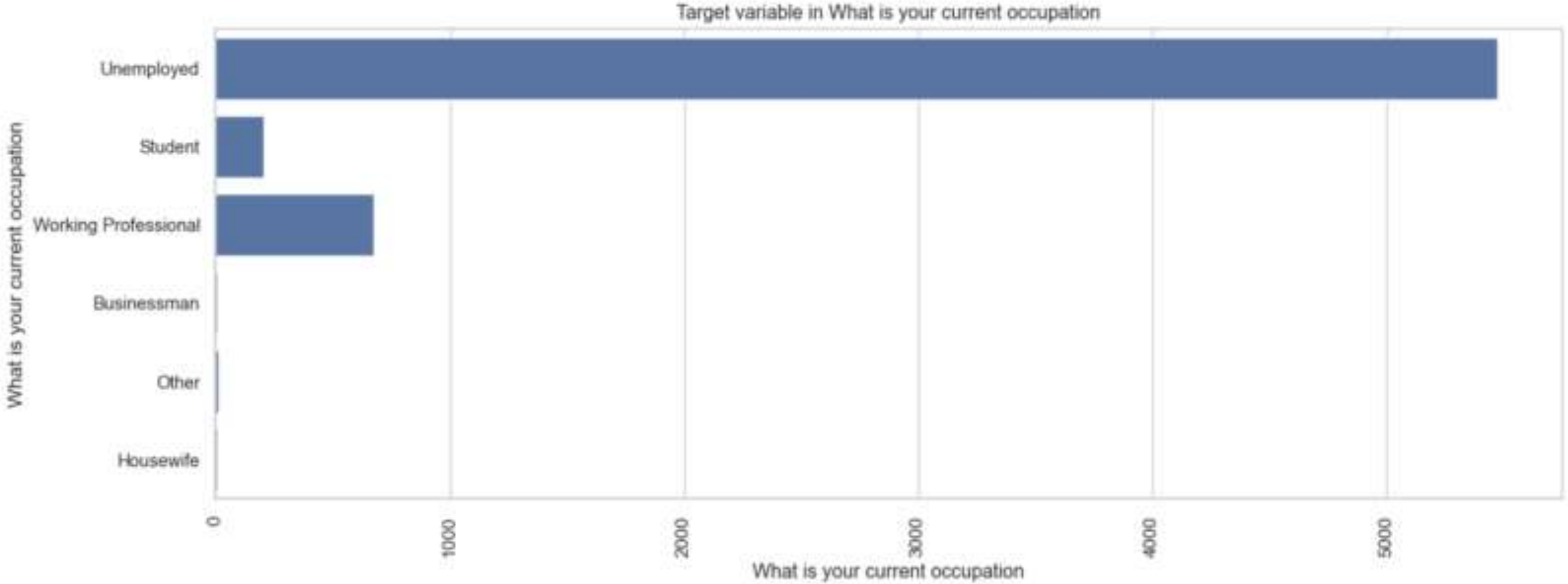
# Exploratory data analysis: Highlights from Bi-Variate Analysis

- Most leads respond to emails as shown by the last activity data



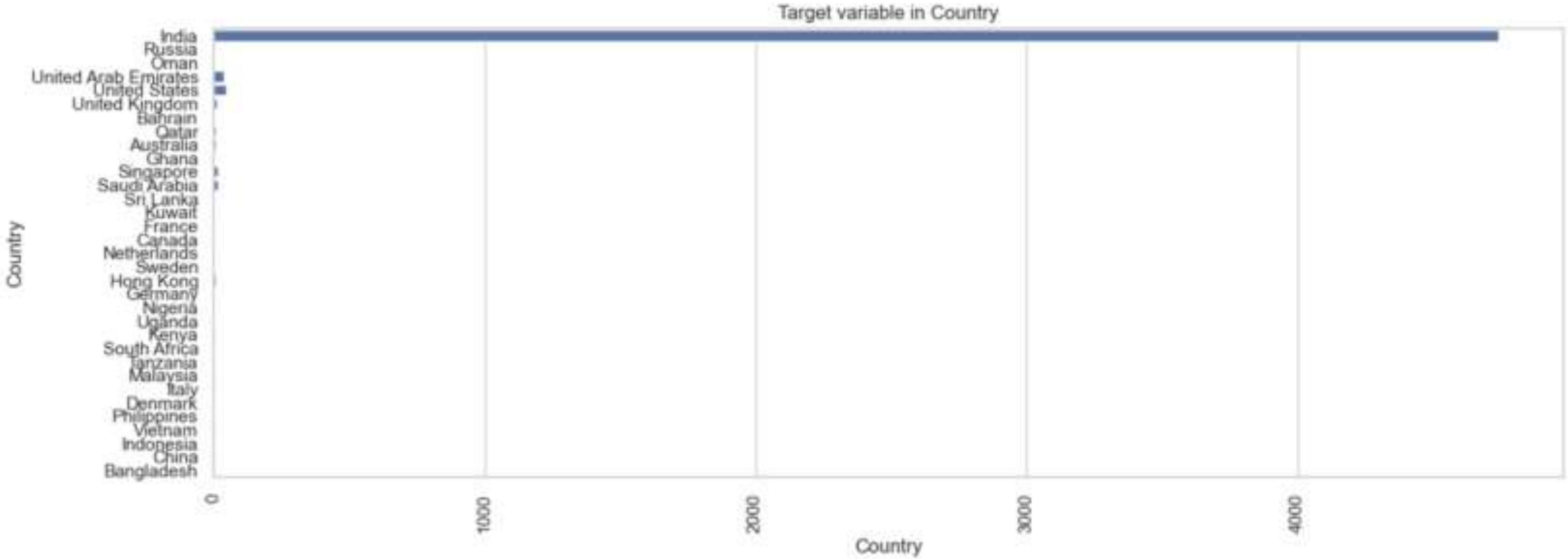
# Exploratory data analysis: Highlights from Bi-Variate Analysis

- Unemployed people are most interested in taking the course



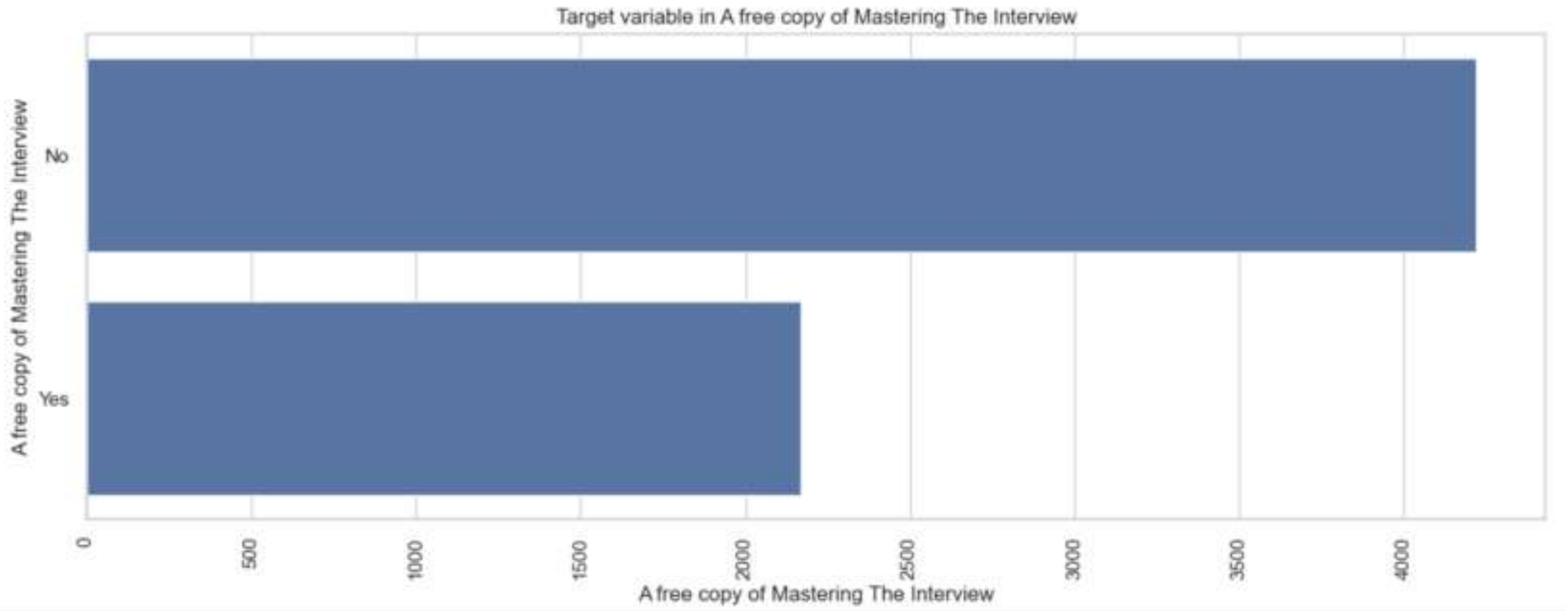
# Exploratory data analysis: Highlights from Bi-Variate Analysis

- Almost all the leads are from India



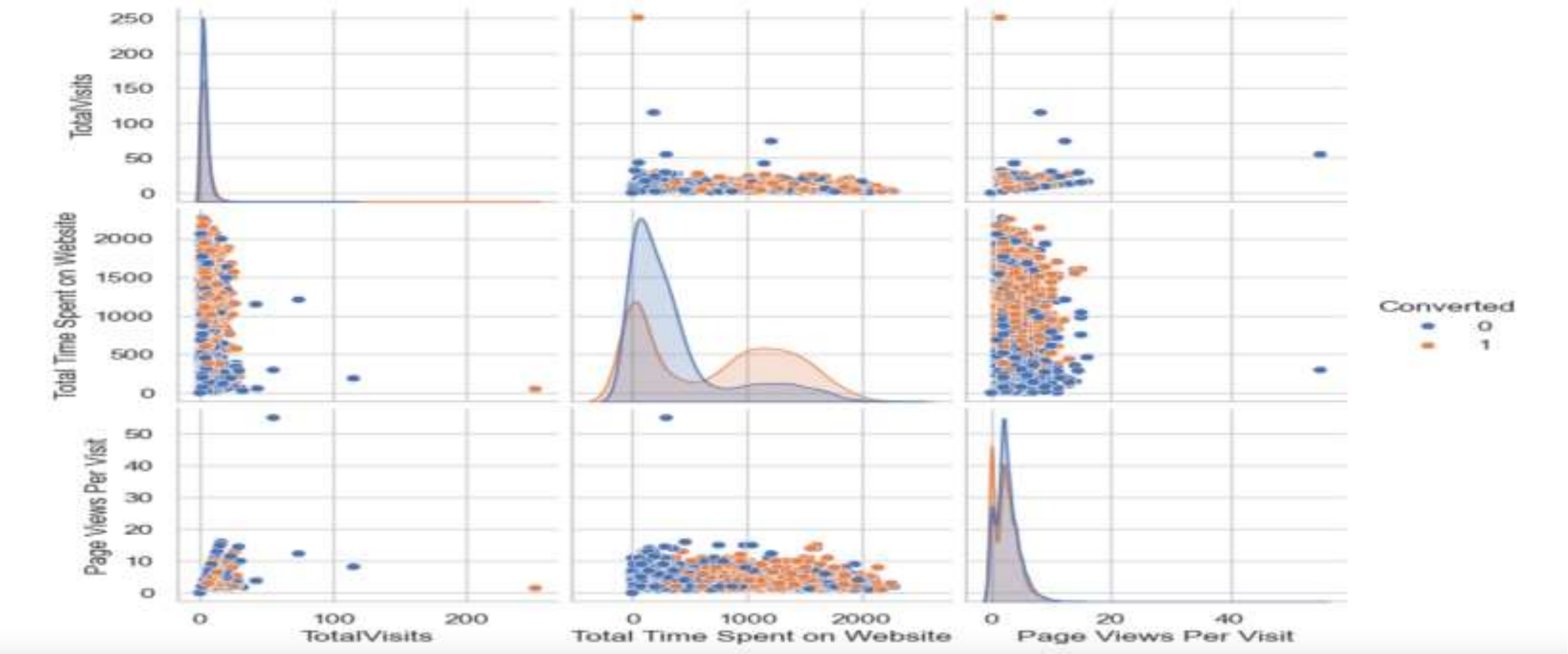
# Exploratory data analysis: Highlights from Bi-Variate Analysis

- Higher number of people did not opt for a free copy of interview prep book



# Exploratory data analysis: Highlights from Bi-Variate Analysis

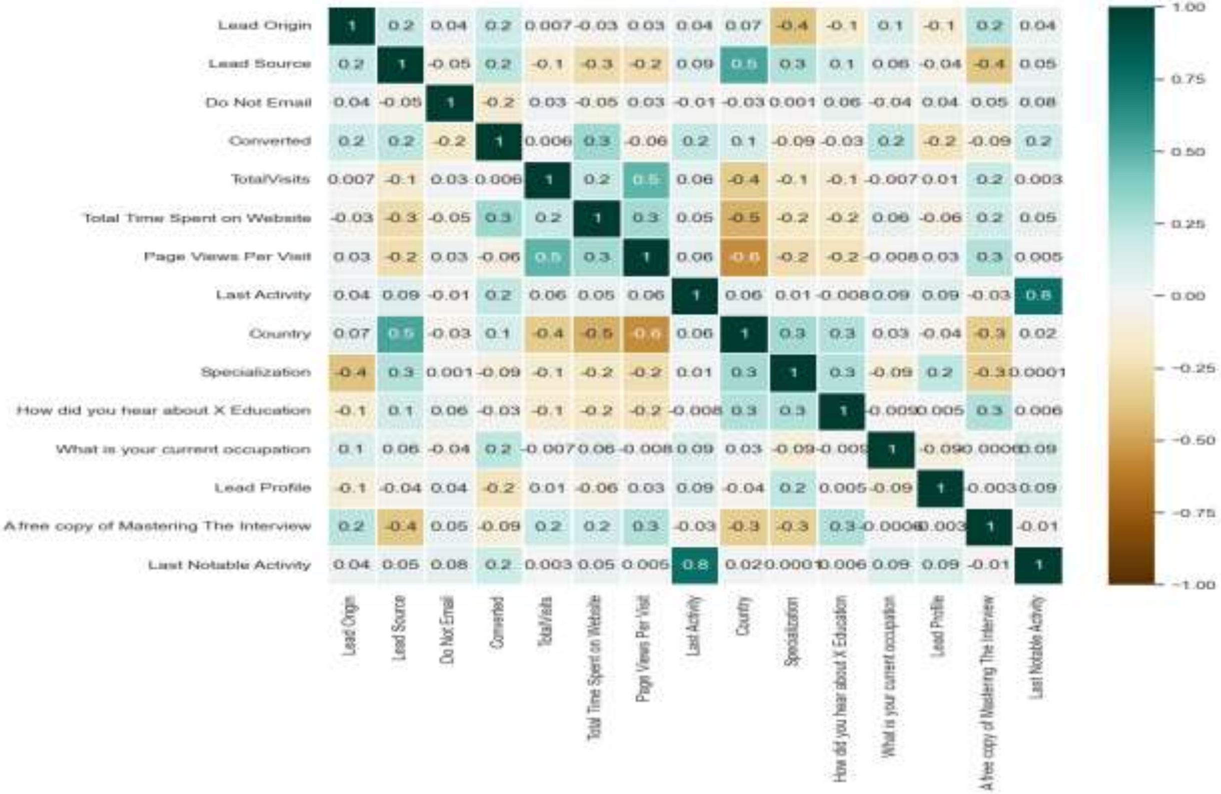
- Below graph shows that the more time people spend on the website and the more pages they visit, the higher their rate of conversion is.



# Correlation

# Correlation

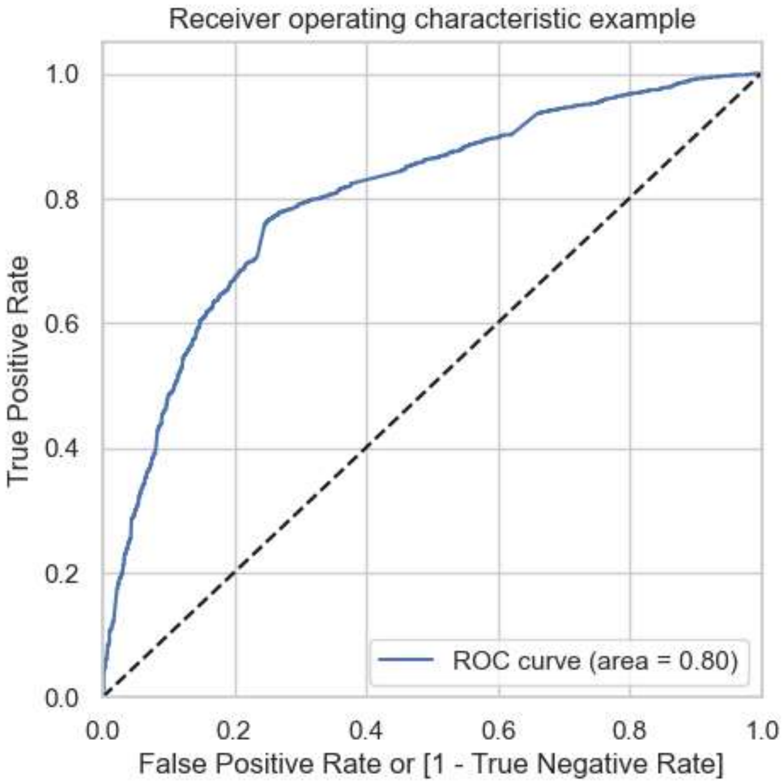
- The heat map shows the correlation between the variables with the target variable



# Model building and Evaluation

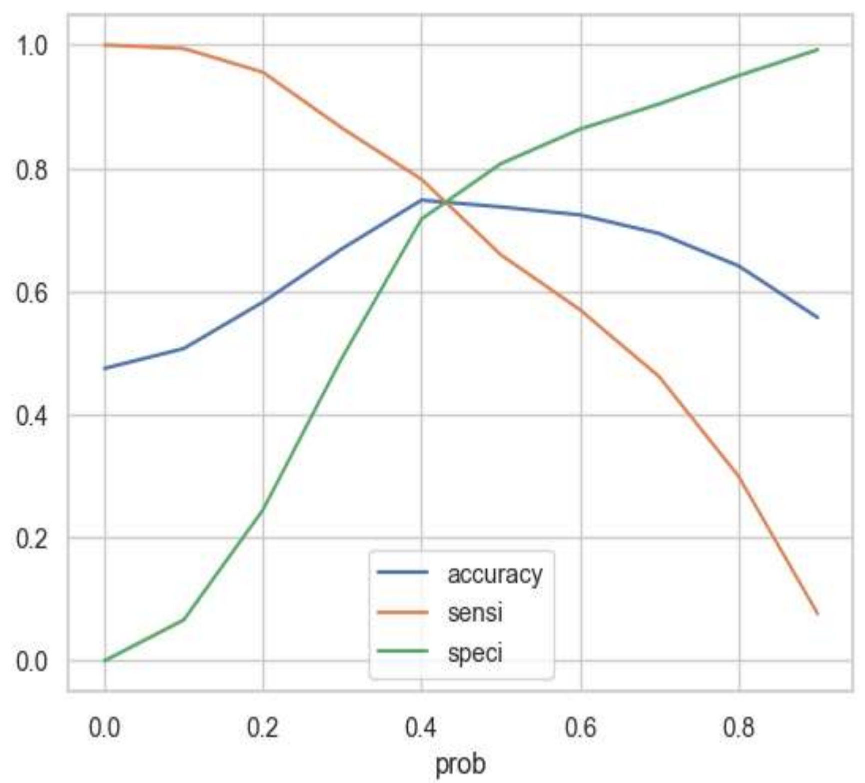
# Model building and Evaluation

- We created 3 models with this data. With the use of Model 1 and 2 we dropped columns with the highest VIF values. Dropped 'What is your current occupation' after Model-1 and ‘Lead Profile’ after Model-2
- By Model-3 we had decent VIF and p-values so we used this model for evaluation
- The ROCE curve from model 3 shows a good area, which is 0.8, under the curve



# Model building and Evaluation

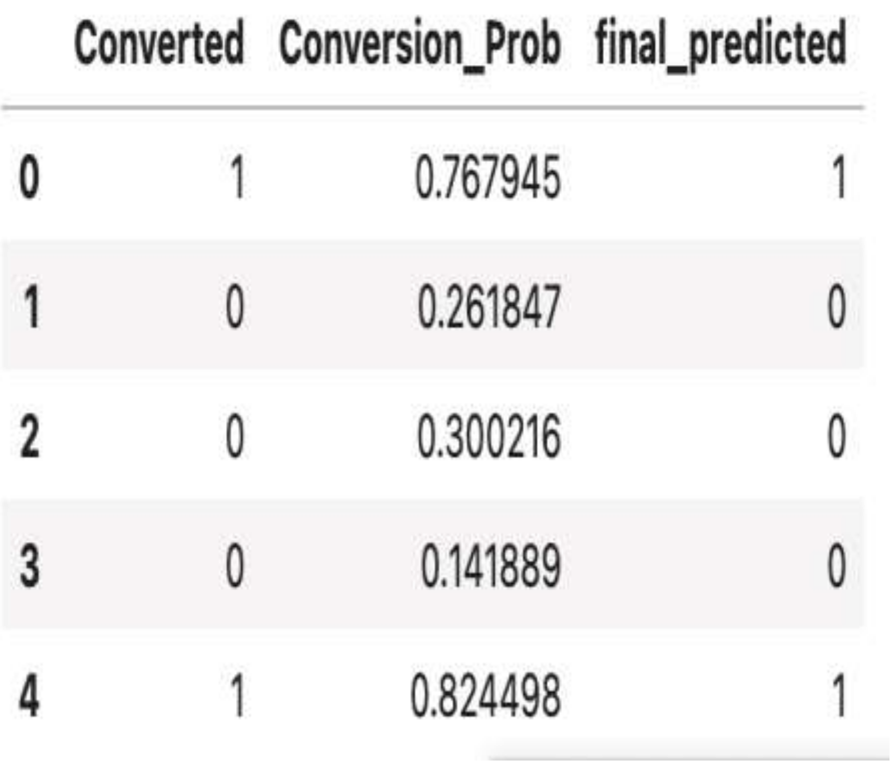
- Using the confusion matrix we identified that 0.43 is the perfect point for cutoff. The plot below shows
- This helped in identifying the Sensitivity and Specificity as 0.7631826741996234 and 0.7484035759897829 respectively with 0.7554214173932484 as accuracy



# Prediction on the test set

# Prediction on the test set

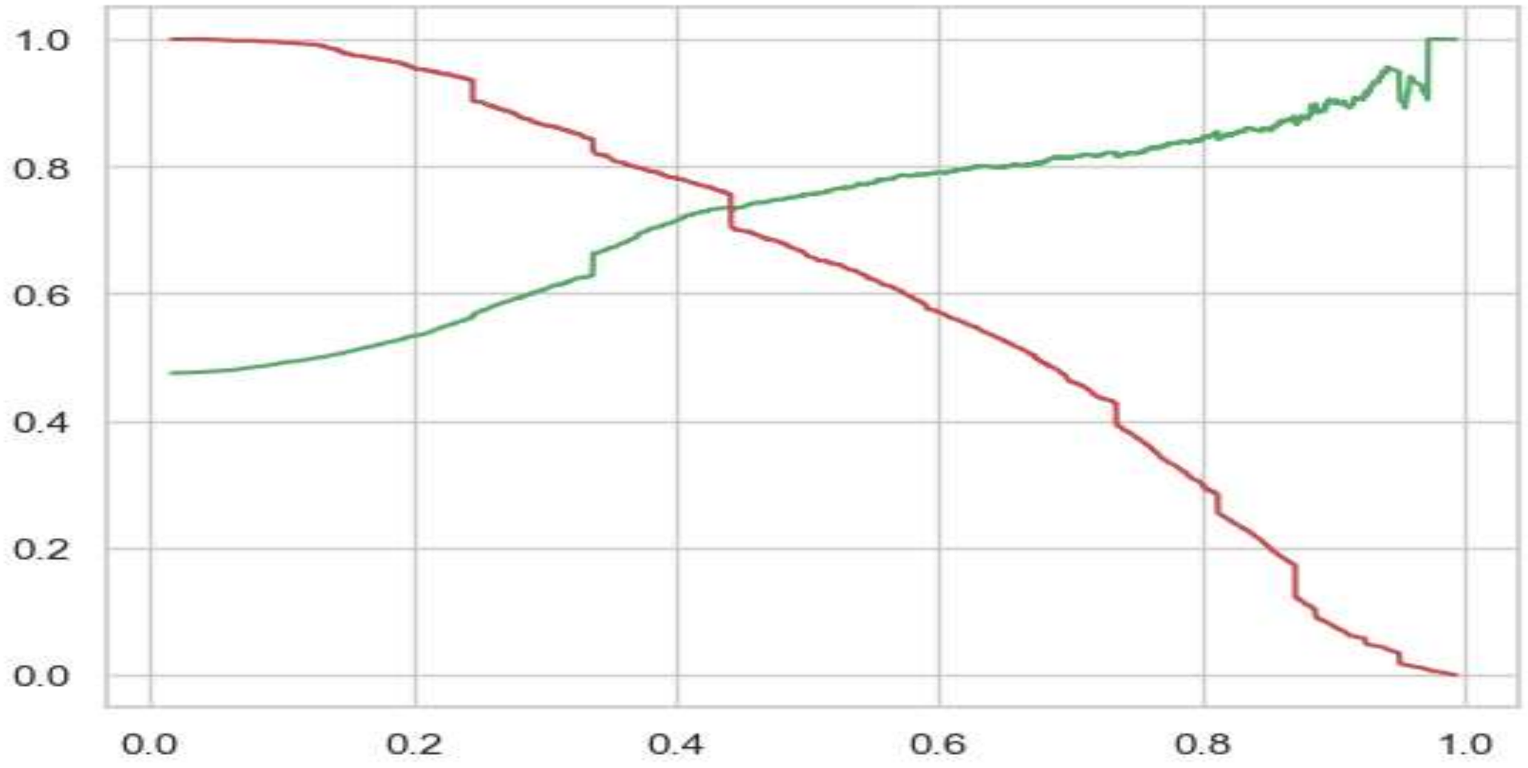
- Using the model, the predictions made were correct when tested on the test set
- The accuracy, sensitivity and specificity score were 0.745046923879407, 0.7598299681190224 and 0.7308085977482088 respectively.



|   | Converted | Conversion_Prob | final_predicted |
|---|---|---|---|
| **0** | 1 | 0.767945 | 1 |
| **1** | 0 | 0.261847 | 0 |
| **2** | 0 | 0.300216 | 0 |
| **3** | 0 | 0.141889 | 0 |
| **4** | 1 | 0.824498 | 1 |

# Precision Recall-Review curve

# Precision Recall-Review curve



# Observations

# Observations

- The accuracy of the test and train data matched perfectly
- The accuracy, sensitivity and specificity scores for both, test and train data were as follows
    - Accuracy = 75% (0.7450469238790407)
    - Sensitivity = 76% (0.7598299681190224)
    - Specificity = 73% (0.7308085977482088)

- The list of features selected for the final model for evaluation is below
    - Lead Origin
    - Lead Source
    - Do Not Email
    - Total Time Spent on Website
    - Page Views Per Visit
    - What is your current occupation
    - Lead Profile
    - Last Notable Activity

# Conclusion

# Observations

- Lead conversions are highest from landing page submissions, Google redirects and direct traffic. While leads spending more time on the website and employed professionals have a higher probability of conversion
- Time spent on the website, number of visits, source of lead (Google, Direct traffic, Organic search, Welingak website), and activities like SMS have turned out to be the key source of influencing leads
- Data suggests the importance of website engagement and specific user segments. Targeted marketing efforts on these metrics will boost sales and conversion.
- The logistic regression model accurately predicts lead conversions with 75% accuracy in both test and training data. So, this will prove to be a good model for predicting and boosting conversion and so the sales too.