Lead Scoring Case Study

Group:

1. Sindhu Manakame
2. Prateek Singh

Problem Statement and Business Objective

Problem Statement

- An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.
- The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. 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%

Business Objective

- The company requires to build a model to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.
- The company's senior leadership wants to achieve the target lead conversion rate of around 80%.

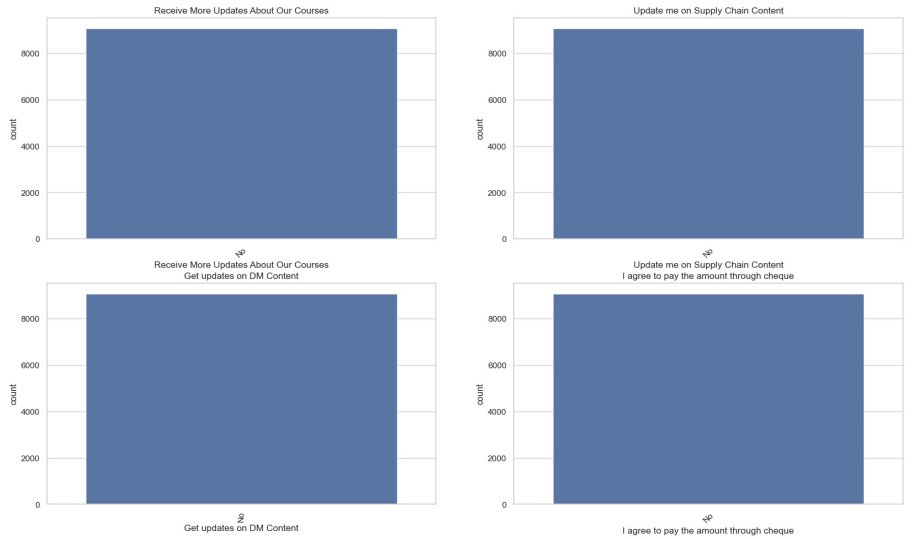
Solution Approach

- Understanding the problem and business objective
- Sourcing the data
- Understanding the data and its structures thoroughly
- Handling Data Quality issues (if any)
- Exploratory Data Analysis
- Data Preparation
- Logistic Model building
- Model Evaluation
- Inferences

Data Sourcing and Cleaning

- Data Sourced from CSV file and understood the business significance of each columns.
- Based on analysis, columns with missing values >35% were dropped, this including the columns with values as ‘Select’ which are treated as missing.
- ID columns which are not relevant for analysis were dropped
- Categorical columns which had specific values < 2% - 3% overall were categorised as ‘Others’ category for the better analysis.
- The Boolean columns which had single value for all records were dropped as they don’t add value in further analysis (please see fig 1.)

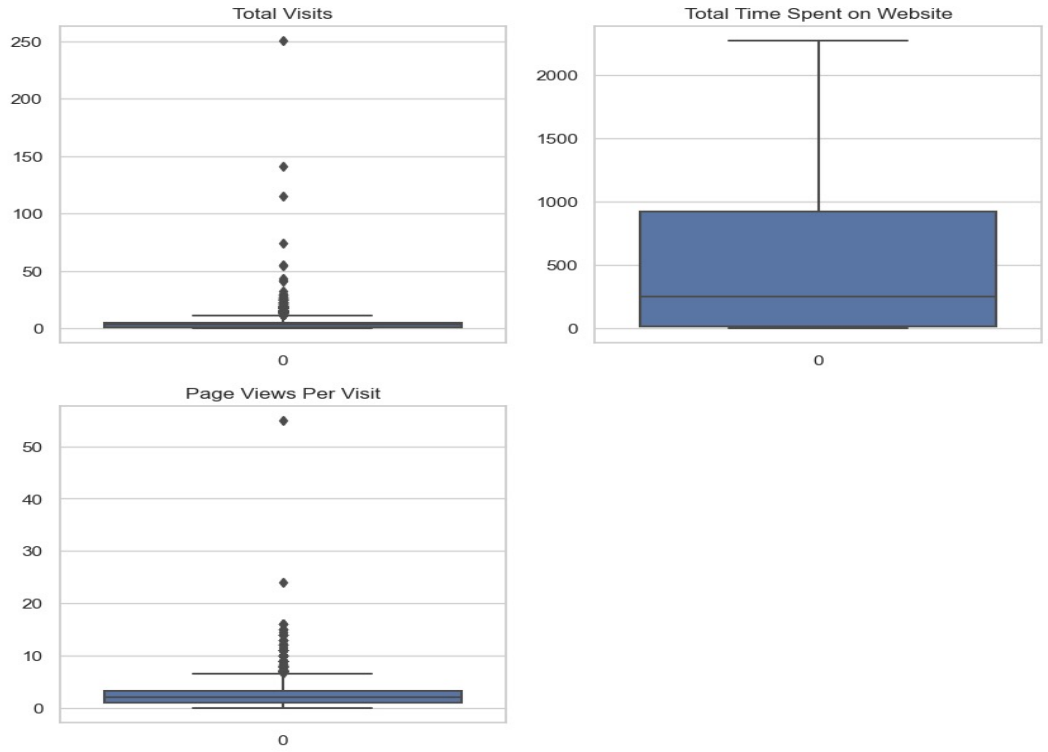
Fig. 1 Count plots depicting. Columns which has only only ‘No’ values.



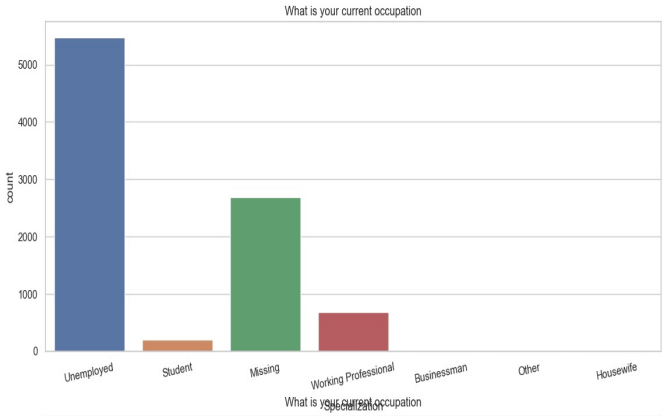
Data Cleaning - Outlier Handling

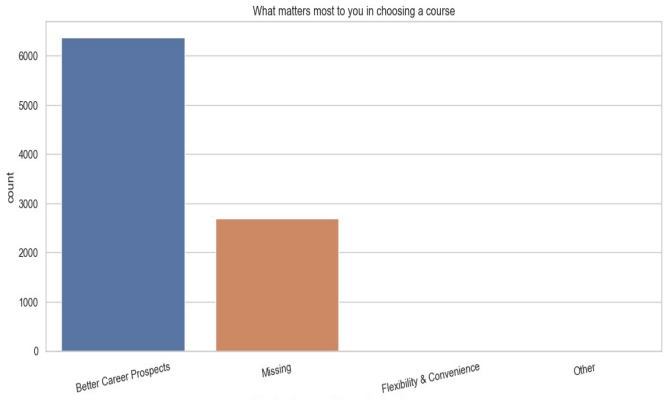
- It was observed that ‘Total Visits’ and ‘Page Views Per Visit’ columns have outliers, as shown in Fig.2.
- It was also observed that rows with outliers had conversion potential.
- Hence, the outliers were imputed to median values so that it doesn’t impact the models and important information is not removed.

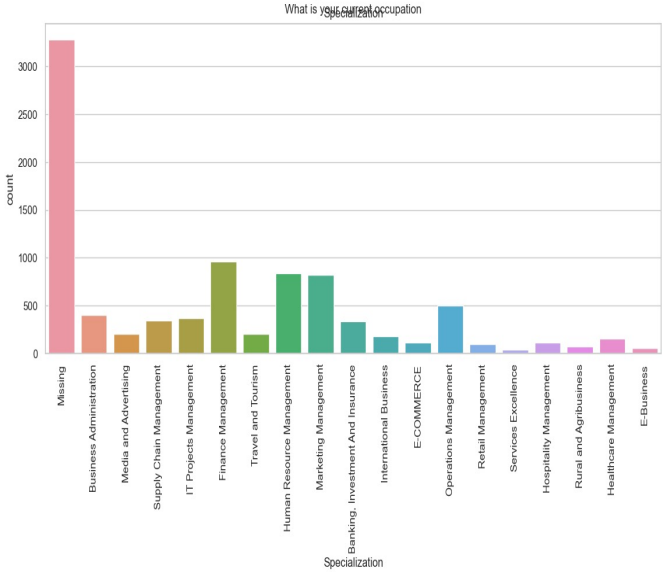
Fig. 2 Box plots depicting. Columns which with outliers

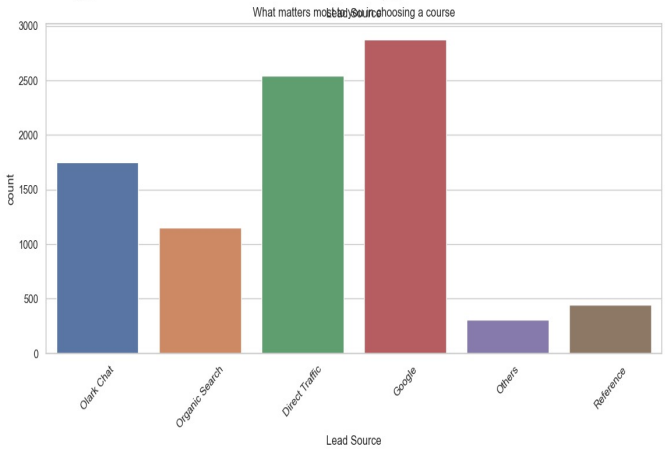


Univariate Analysis









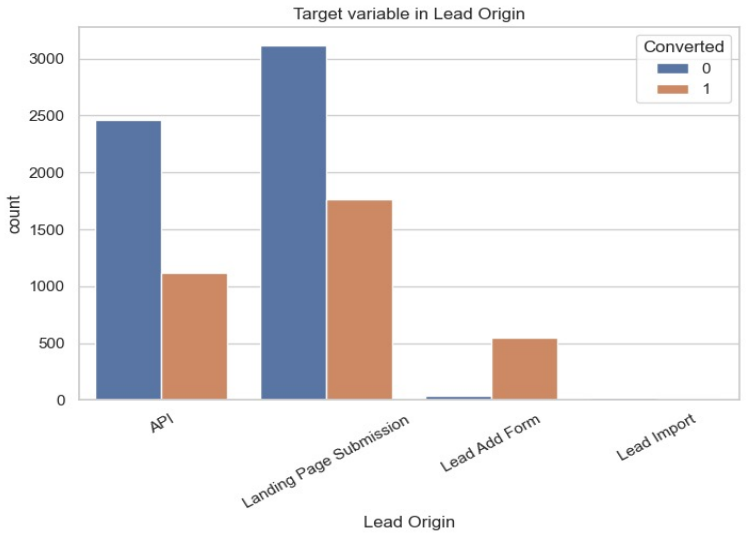
From the plots;

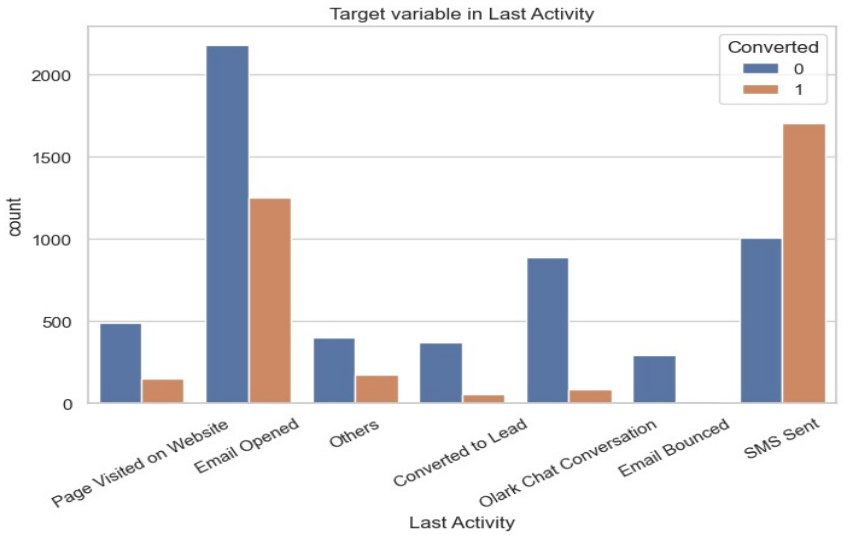
it is observed that following people are more likely to enquire about the courses or potential leads

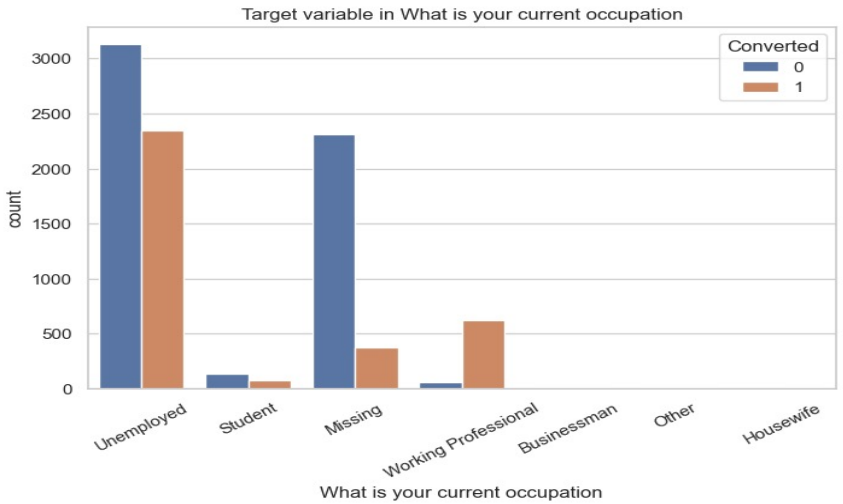
- Unemployed
- Looking for better career prospects

- The Source for most of the leads is ‘Google’
- Most leads chose not to put their ‘Specialization’

Bivariate Analysis



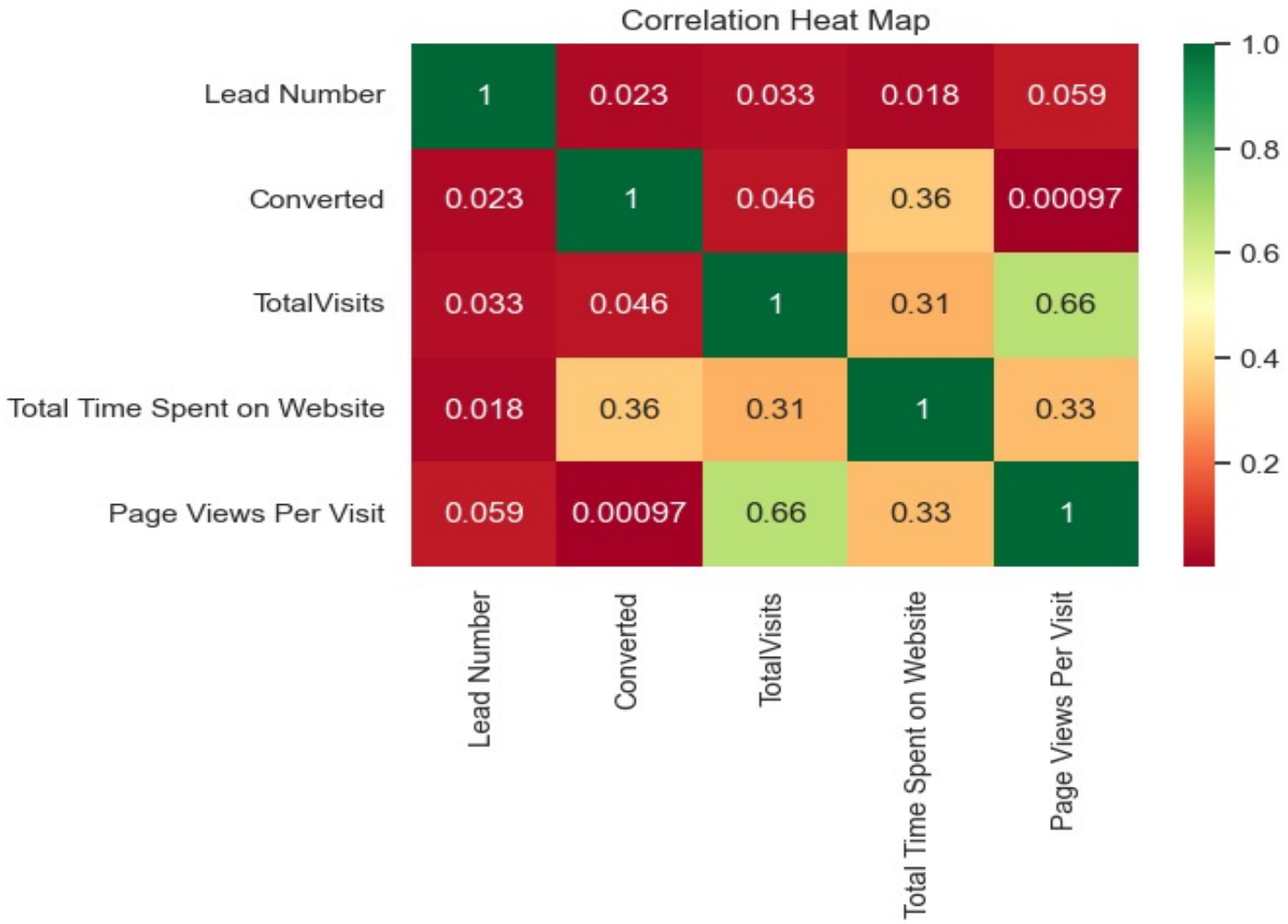




From the plots it is observed that;

- Leads with origin as ‘Landing Page Submission’ are most likely to get converted.
- Sending SMS was the last activity of most of the leads converted.
- They also chose ‘No’ for ‘Do Not Email’
- Key point to note is, unemployed people are looking for upskilling, so company should target them more.

Correlation Between Numerical Variables



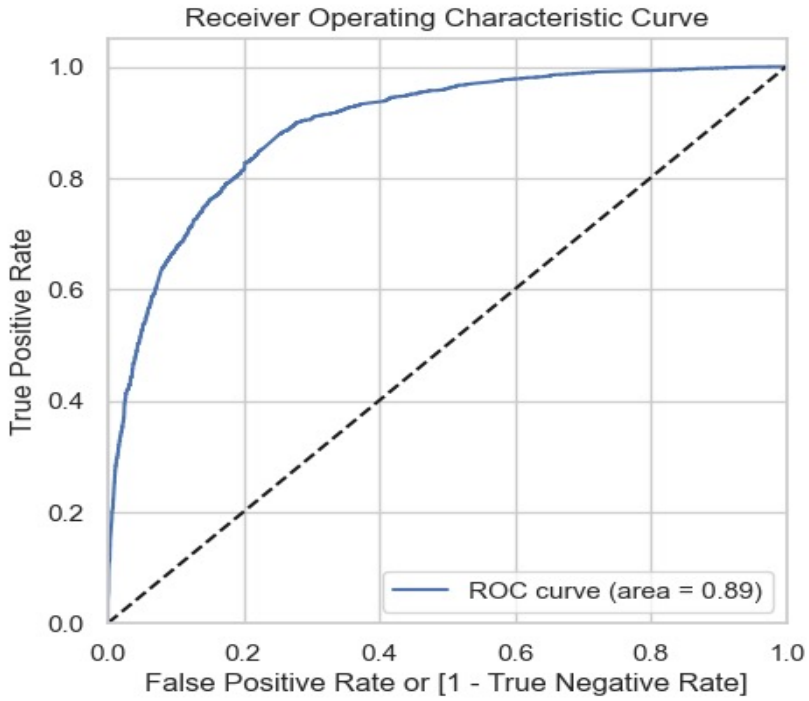
- It is observed that there is reasonable high correlation between 'Page Views Per Visit' and 'TotalVisits'.

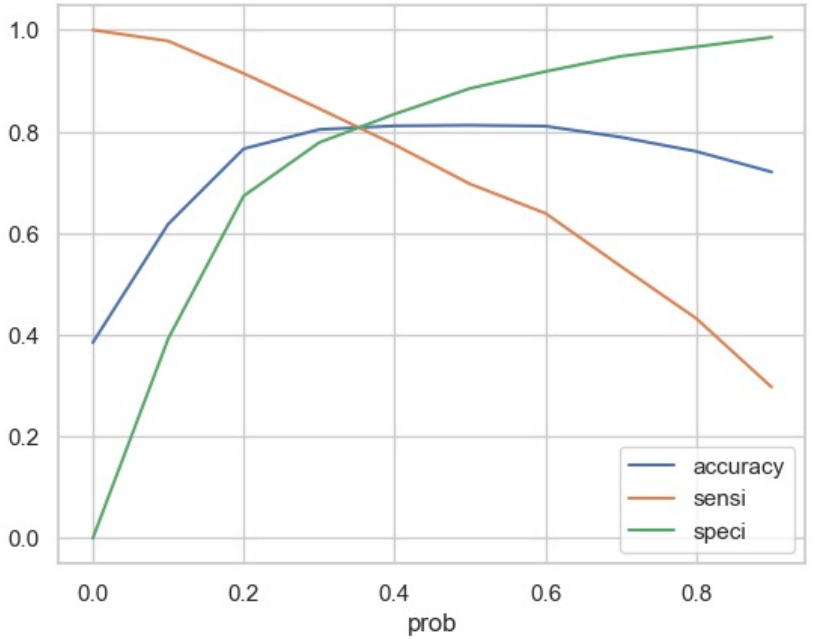
Data Preparation & Model Building

- Top 15 Relevant and Optimal features/variable selected using RFE method.
- Creation of dummy variables for Categorical columns
- Splitting test and train data sets
- Scaled train data variables using Mix Max Scaler
- Built the model using Stats Model, checked the p values and VIF for the variables
- Eliminated the variables with p value > 0.05 and VIF > 5 by recursively model building.
- Built the final model with the relevant and significant variables

Model Evaluation

- Evaluated the model by creating Confusion Matrix
- Plotted the ROC Curve to find the area under the curve which is 0.89 as shown below.
- Plotted accuracy, specificity and sensitivity to get Optimal Cutoff Point





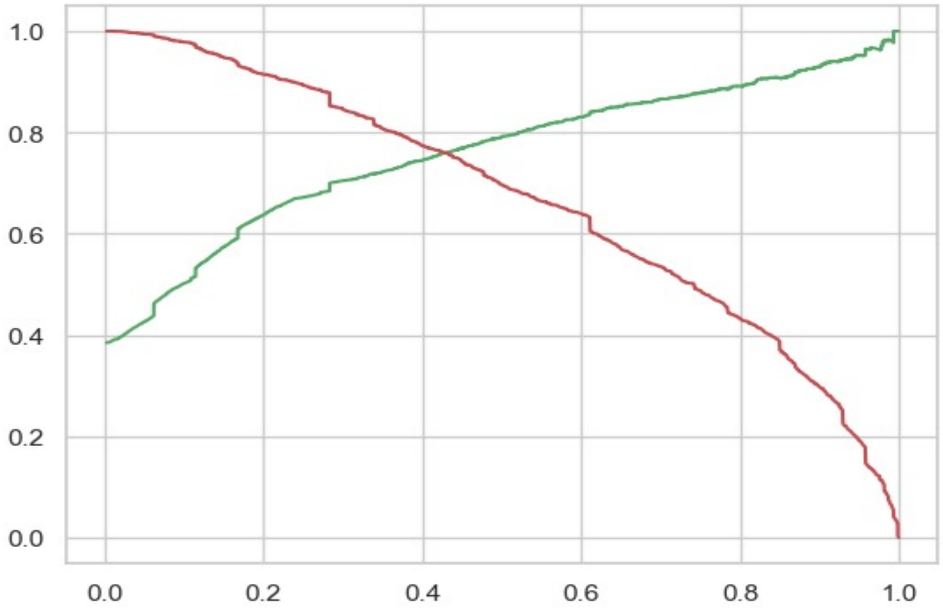
From the curve above, 0.35 is the optimum point to take it as a cut-off probability

- Optimal cut-off point is that probability where we get balanced sensitivity and specificity.
- The cut-off value of 0.35 has given accuracy, sensitivity and specificity above 80% both on train and test data

Model Evaluation

Finding optimal cut-off based on Precision and Recall

Plotting the trade-off between Precision and Recall



Above graph shows 0.42 as the cut-off considering precision and recall

- The cut-off value of 0.42 has given accuracy of 81%
- Recall and Precision approx. 75% and 76% respectively on both test and train data.

Observations

Below listed features show the conversion possibilities.

- TotalVisits
- Total Time Spent on Website
- Lead Origin_Landing Page Submission
- Lead Origin_Lead Add Form
- Lead Source_Olark Chat
- Do Not Email_Yes
- Last Activity_Email Opened
- What is your current occupation_Working Professional

Recommendations and Conclusions

- It is observed that following people are more likely to enquire about the courses or are potential conversion leads
  - Unemployed
  - Looking for better career prospects
- The Source for most of the leads is ‘Google’, followed by ‘Direct Traffic’.
- Most leads who are converted chose not to put their ‘Specialization’.
- Leads with origin as ‘Landing Page Submission’ are most likely to get converted.
- Sending SMS was the last activity of most of the leads converted, followed by ones who had opened the emails sent by Sales/Marketing teams.
- They also chose ‘No’ for ‘Do Not Email’
- Key point to note is, unemployed people are looking for upskilling, so company should target Unemployed and Working Professionals, as they are more likely to get converted.