Lead Scoring Case Study

Submitted by

Parijaat Sunil
Pratikshit Gaur
Shreya Pattan

Understanding the Problem

- Problem Statement
  - While X education gets a lot of leads, the conversion rate of such leads is low. Currently, the conversion rate of the leads is approximately 30%
  - To make their selling process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’
  - The company management wants to create a lead scoring system which rates leads from 0 to 100 based on likelihood of conversion
  - The company would like to increase the conversion rate of leads to 80%

- Solution Objective
  - Our proposed solution must be able to assign every lead with a score based on the available features
  - The proposed solution will be based on logistic regression

Proposed Solution Methodology

- We have arrived at our proposed solution using the below steps:
  - Understanding the data: shape, data types, number of missing values
  - Data cleaning including
    - Handling of missing values
    - Handling values marked ‘Select’
    - Dropping irrelevant columns
  - EDA
    - Univariate data analysis
    - Bivariate data analysis
    - Multivariate data analysis
  - Data preparation
    - Dummy variable creation for categorical data
    - Train test split of data
    - Feature Scaling
  - Creation of Model
    - RFE
    - Manual model building
    - VIF Analysis
  - Evaluation of Model
    - Accuracy, Sensitivity and Specificity
    - ROC curve
    - Finding the optimal cutoff point
    - Precision and recall tradeoff analysis
  - Final prediction on test set
  - Calculation of lead scores and listing of final factors

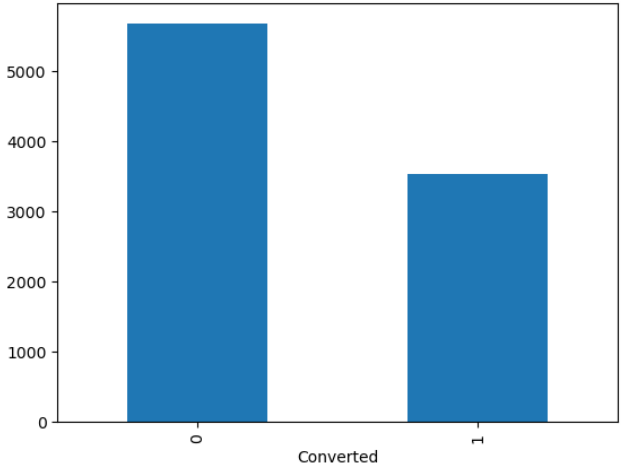
Understanding and cleaning the data

- Available data had 37 columns and 9240 rows initially
- Dropped columns which had unique values and single value
- Replaced ‘Select’ values with NaN values
- Dropped columns with more than 40% of missing data
- Dropped Country column as it was extremely skewed
- Imputed values in ‘What is your current occupation’, ‘Specialization’, and ‘City’
- Dropped Tags column as the data had many ambiguous values
- Standardizing columns having binary Yes/No data with 1/0

- **Data available after data cleaning: 37 columns and 9204 rows**

Exploratory Data Analysis - Univariate

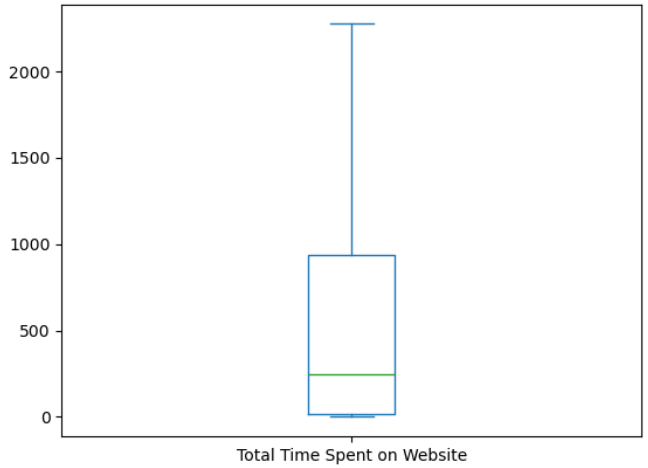
Converted



Inferences:

- The current conversion ratio for leads is 38%
- The target conversion ratio is 80%

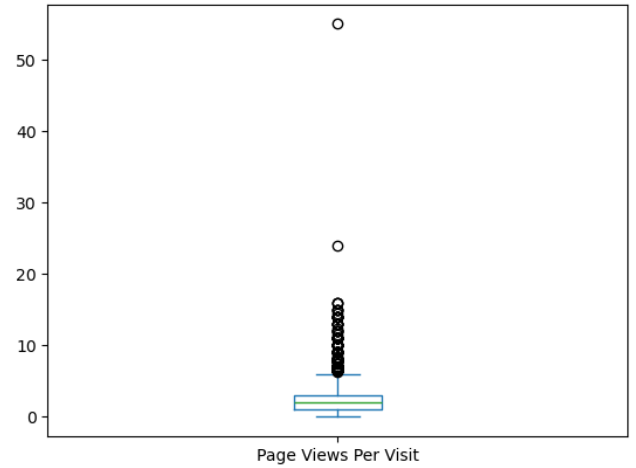
Total Time Spent on Website



Inferences:

- There are no outliers in the data
- The median time spent on the company website is 250 seconds

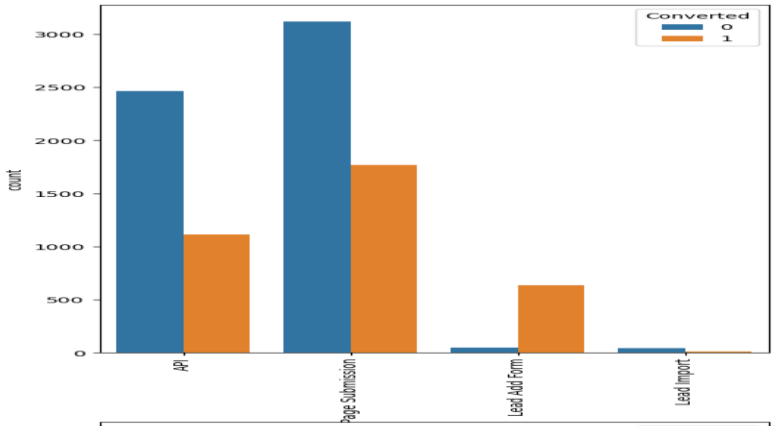
Page Views Per Visit

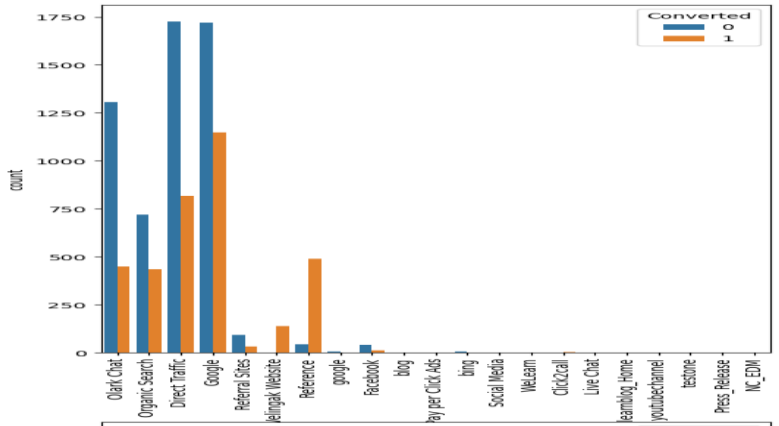


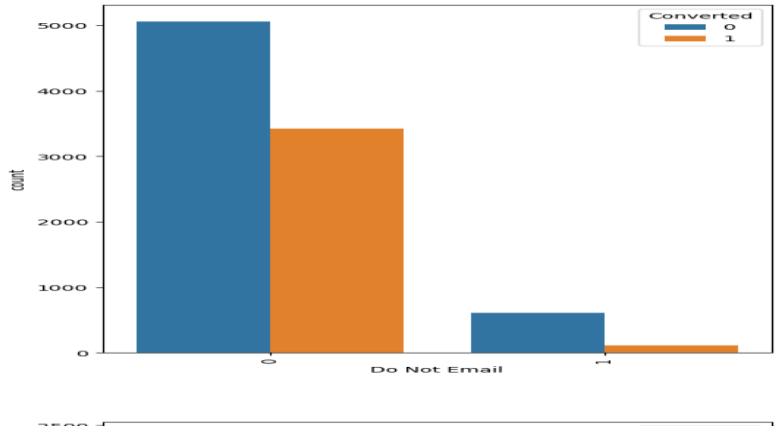
Inferences:

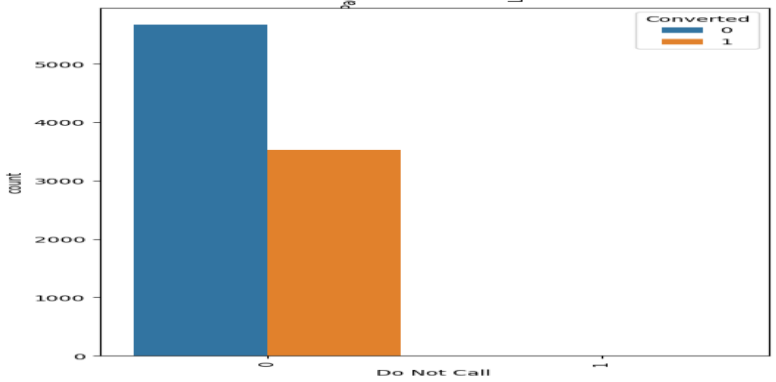
- There are many outliers in the data
- The median page views per visit on the company website is 2 pages

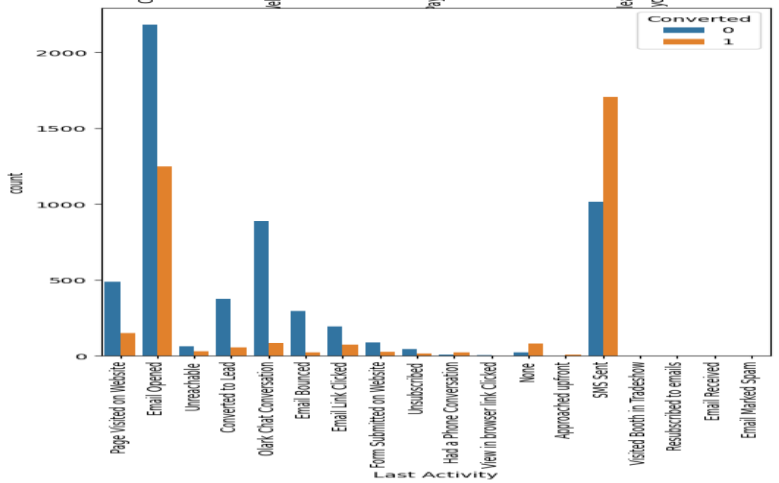
Exploratory Data Analysis – Bivariate (1/3)

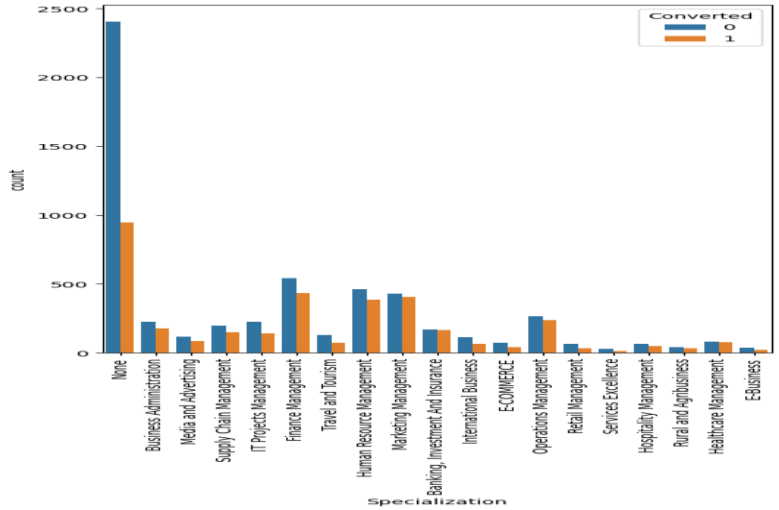








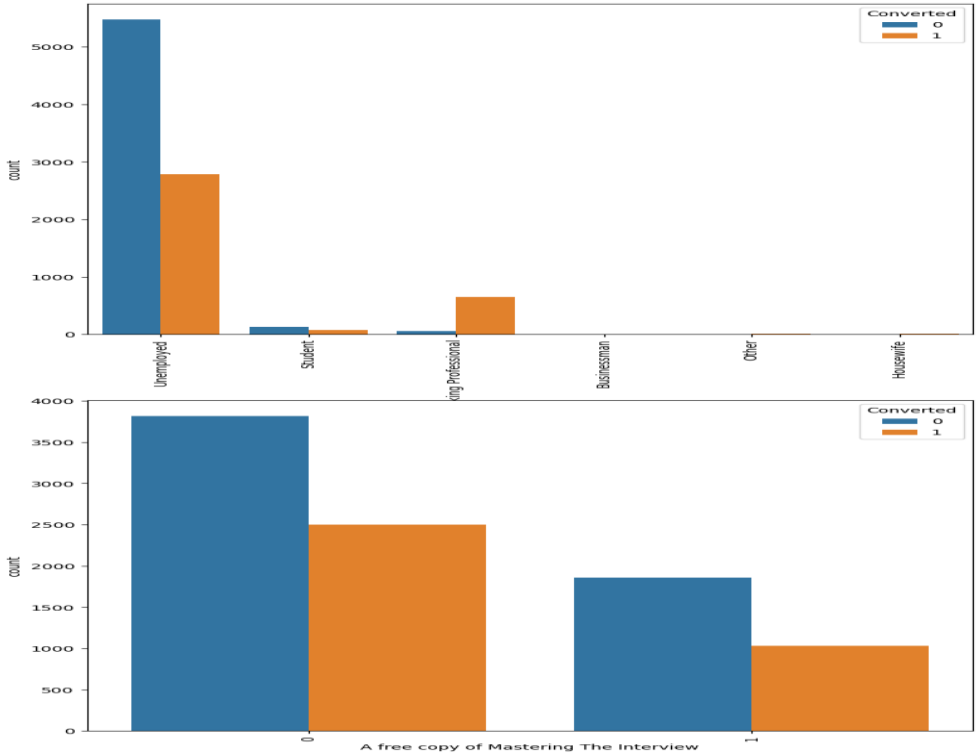


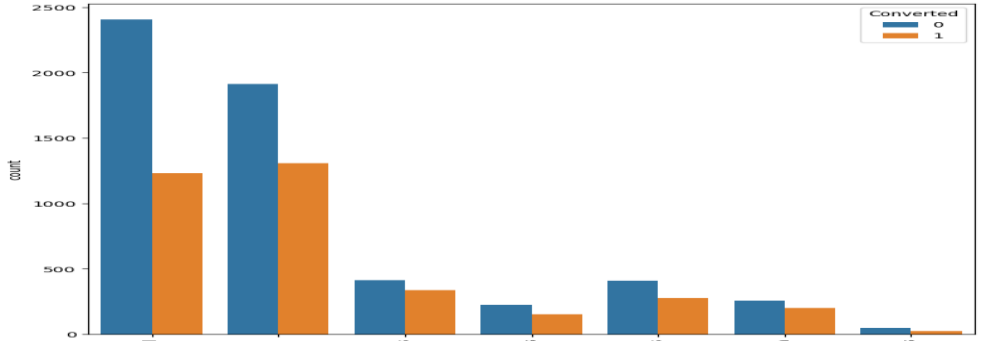


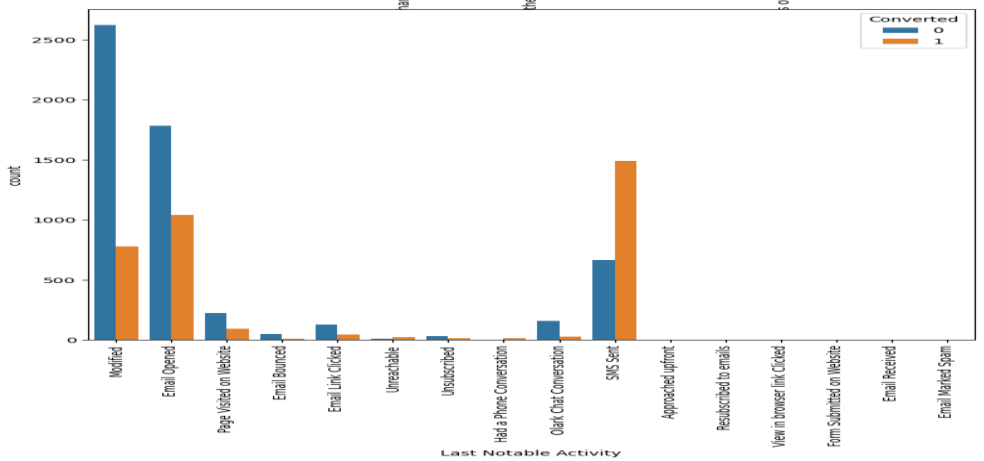
Inferences:

- There is a high rate of conversion from users filling the Lead Add Form, or coming in through references
- Users who check the 'Do not Email' or 'Do not Call' option have lower conversion rates

Exploratory Data Analysis – Bivariate (2/3)



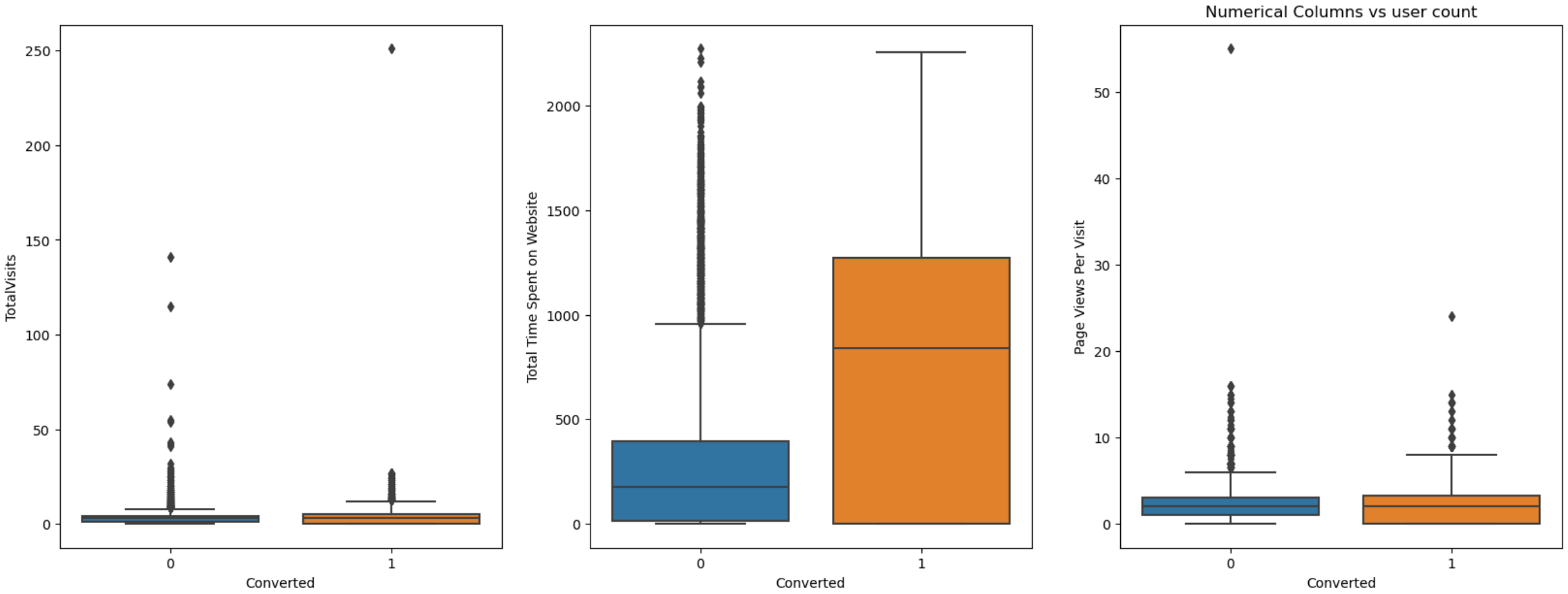




Inferences:

- Working professionals have a high conversion rate
- Other categories do not seem to a high correlation with conversion rate

Exploratory Data Analysis – Bivariate (3/3)



Inferences:

- Users who have spent more time on the company website have a higher rate of conversion
- Total visits and number of pages per visit doesn't seem to have a large affect on rate of conversion

Data Preparation

- Creation of dummy variables for categorical columns
- Creation of train test split (used 75% of data for train and 25% of data for test)
- Usage of standard scaler to standardize the numerical data columns

- **Total size of the data after data preparation is 96 columns and 9204 rows**

Model Building

- After RFE, the final model was arrived at after 3 iterations

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6903
Model:	GLM	**Df Residuals:**	6890
Model Family:	Binomial	**Df Model:**	12
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2955.4
Date:	Tue, 21 May 2024	**Deviance:**	5910.9
Time:	19:06:58	**Pearson chi2:**	7.06e+03
No. Iterations:	7	**Pseudo R-squ. (CS):**	0.3776
Covariance Type:	nonrobust		

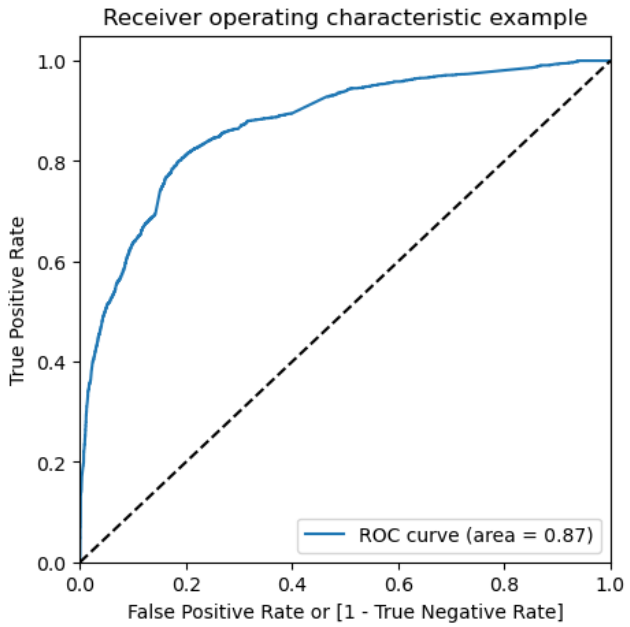
	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.6266	0.063	-9.973	0.000	-0.750	-0.503
Do Not Email	-1.3777	0.164	-8.404	0.000	-1.699	-1.056
Total Time Spent on Website	1.0111	0.035	29.026	0.000	0.943	1.079
City_Other Metro Cities	0.2677	0.154	1.739	0.082	-0.034	0.569
City_Thane & Outskirts	0.2582	0.118	2.191	0.028	0.027	0.489
Lead Origin_Landing Page Submission	-0.6130	0.073	-8.393	0.000	-0.756	-0.470
Lead Origin_Lead Add Form	3.0927	0.192	16.090	0.000	2.716	3.469
Lead Source_Welingak Website	2.7692	1.026	2.699	0.007	0.758	4.780
Last Activity_SMS Sent	1.3088	0.070	18.590	0.000	1.171	1.447
What is your current occupation_Working Professional	2.5381	0.169	14.994	0.000	2.206	2.870
Last Notable Activity_Modified	-1.0047	0.074	-13.606	0.000	-1.149	-0.860
Last Notable Activity_Olark Chat Conversation	-1.4625	0.322	-4.536	0.000	-2.094	-0.831
Last Notable Activity_Unreachable	1.8701	0.510	3.670	0.000	0.871	2.869

	Features	VIF
4	Lead Origin_Landing Page Submission	1.74
5	Lead Origin_Lead Add Form	1.47
7	Last Activity_SMS Sent	1.34
9	Last Notable Activity_Modified	1.28
6	Lead Source_Welingak Website	1.27
8	What is your current occupation_Working Profes...	1.17
0	Do Not Email	1.12
3	City_Thane & Outskirts	1.12
1	Total Time Spent on Website	1.10
2	City_Other Metro Cities	1.08
10	Last Notable Activity_Olark Chat Conversation	1.00
11	Last Notable Activity_Unreachable	1.00

- Both the p-values and VIFs seem to be decent enough for all the variables
- We can use this model to make our predictions using this final set of features.

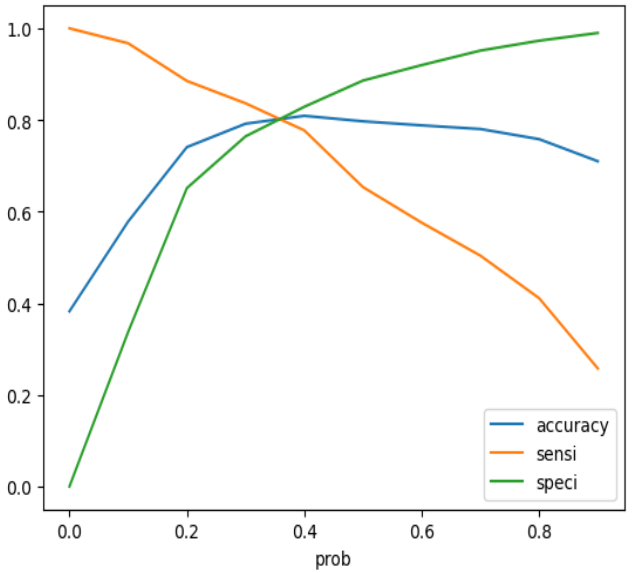
Model Evaluation

ROC Curve



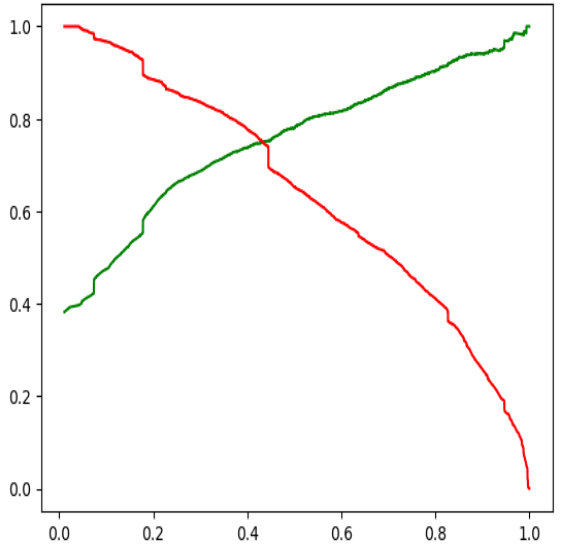
The area under the curve of the ROC is 0.87, which is a good indicator

Optimal cutoff



The optimal cutoff is at probability value of 0.37

Precision Recall Tradeoff



The optimal cutoff is at probability value of 0.43

Evaluation parameters

- Selected 0.43 as the optimal cutoff for conversion probability
- Parameters of model on train set are
  - Accuracy: 81%
  - Sensitivity: 75%
  - Specificity: 75%
- Parameters of model on test set are
  - Accuracy: 81%
  - Sensitivity: 78%
  - Specificity: 84%

Conclusion

Lead scores of all leads were calculated

Lead_Ref	Calculated Lead Score
0	7
1	43
2	67
3	7
4	38
...	...
9235	48
9236	40
9237	14
9238	59
9239	63

Final list of coefficients of most important features

const	-0.626640
Do Not Email	-1.377654
Total Time Spent on Website	1.011128
City_Other Metro Cities	0.267691
City_Thane & Outskirts	0.258205
Lead Origin_Landing Page Submission	-0.613007
Lead Origin_Lead Add Form	3.092750
Lead Source_Welingak Website	2.769202
Last Activity_SMS Sent	1.308847
What is your current occupation_Working Professional	2.538086
Last Notable Activity_Modified	-1.004739
Last Notable Activity_Olark Chat Conversation	-1.462540
Last Notable Activity_Unreachable	1.870134

The below columns are used to predict if the lead is likely to be converted with approximately 81% accuracy. The below list is in descending order of correlation

- **Feature Correlation:**
  - Lead Origin_Lead Add Form
  - Lead Source_Welingak Website
  - What is your current occupation_Working Professional
  - Last Notable Activity_Unreachable
  - Last Notable Activity_Olark Chat Conversation
  - Do Not Email
  - Last Activity_SMS Sent
  - Total Time Spent on Website
  - Last Notable Activity_Modified
  - Lead Origin_Landing Page Submission
  - City_Other Metro Cities
  - City_Thane & Outskirts

Thank You!