LEAD SCORING CASE STUDY

Submitted by

Kumar Kanishka

Understanding the Problem

- Problem Statement
  - X education generates numerous leads, but their conversion rate is currently low at approximately 30%.
  - To streamline their sales process, the company aims to pinpoint the most promising leads, termed as 'Hot Leads'.
  - Management plans to implement a lead scoring system, ranking leads from 0 to 100 based on their likelihood of conversion.
  - The company's goal is to boost the lead conversion rate to 80%.
- Solution Objective
  - Our proposed solution must be able to assign every lead with a score based on the available features
  - Logistic Regression will the basis for our proposed solution

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

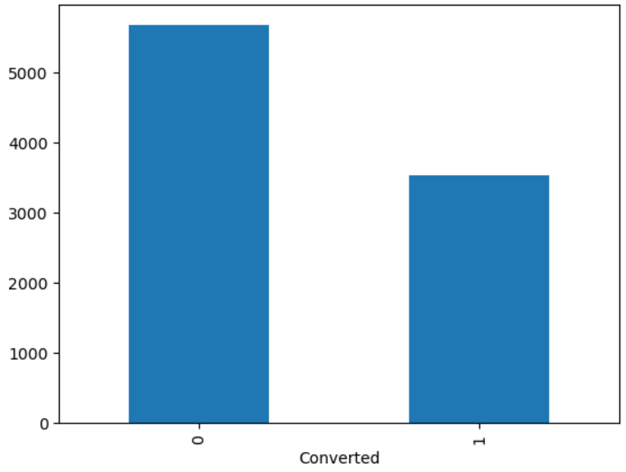
Understanding and cleaning the data

- 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
- 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

- **After cleaning, Data available: 37 columns and 9204 rows**

Exploratory Data Analysis - Univariate

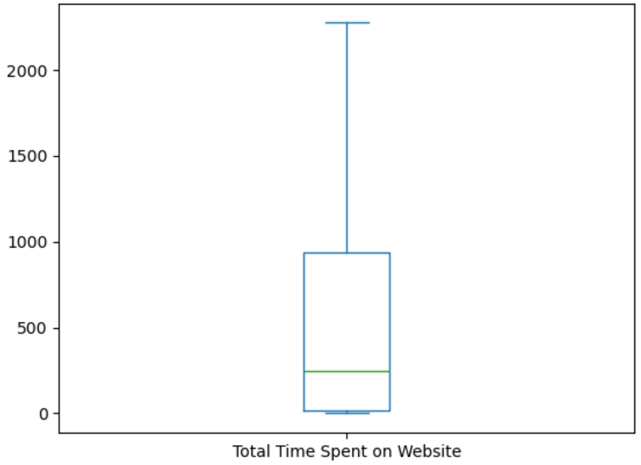
Converted



Inferences:

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

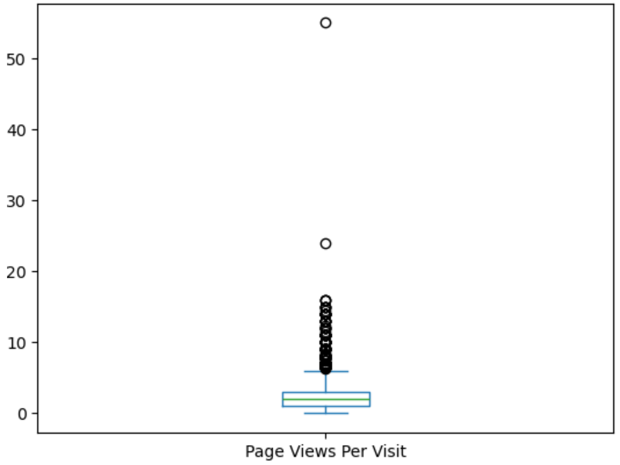
Total Time Spent on Website



Inferences:

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

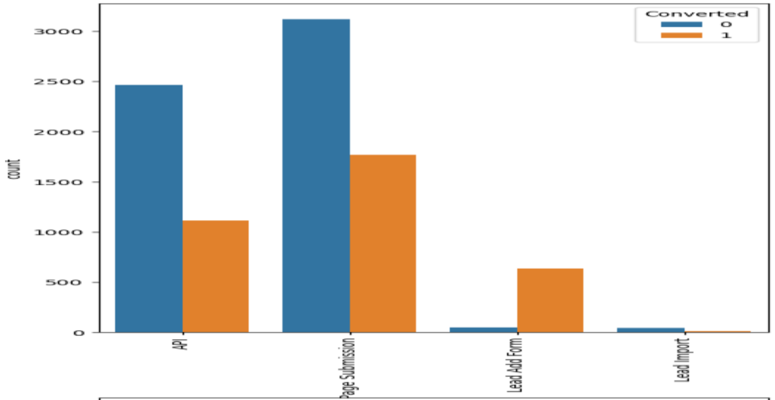
Page Views Per Visit

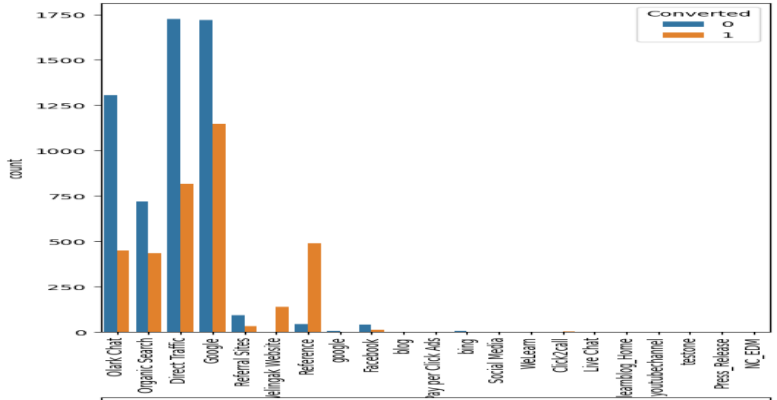


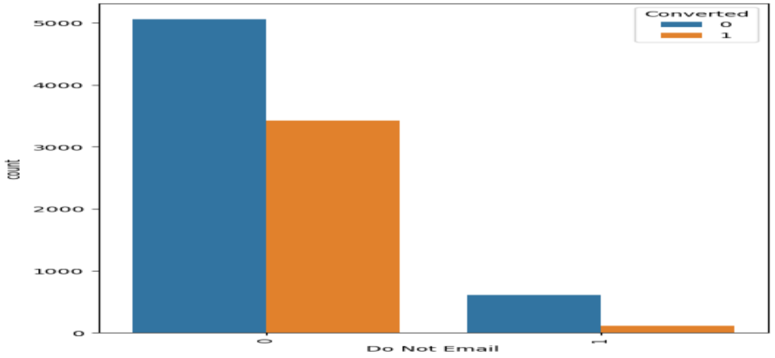
Inferences:

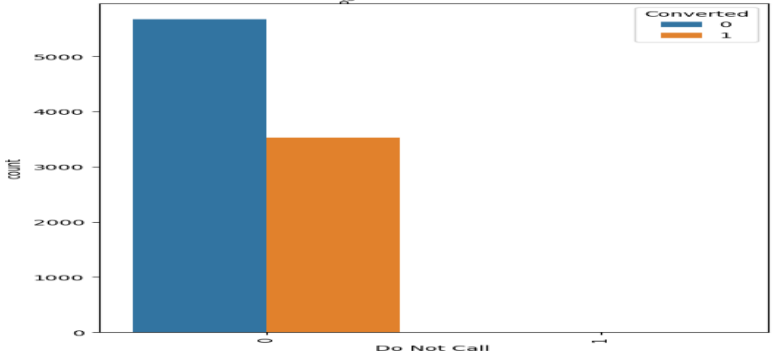
- There are many outliers in the data
- The median page views per visit on the company website is 2 pages

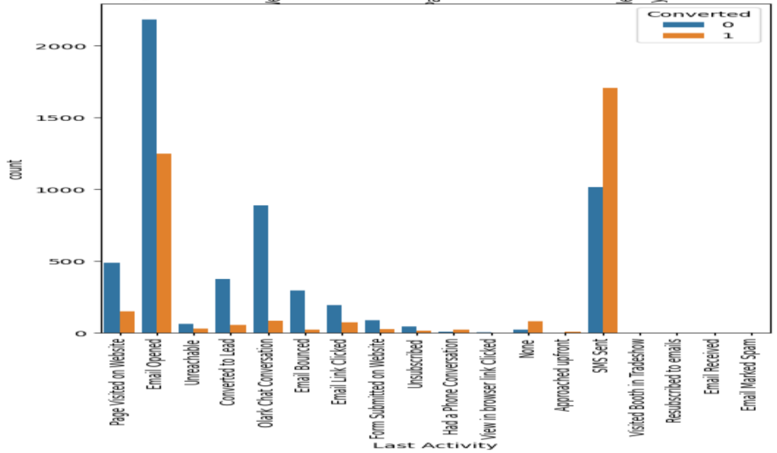
Exploratory Data Analysis – Bivariate (1/3)

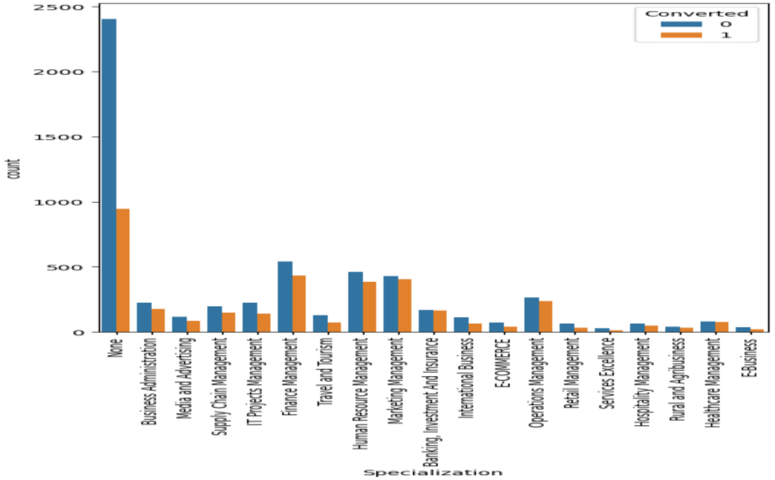








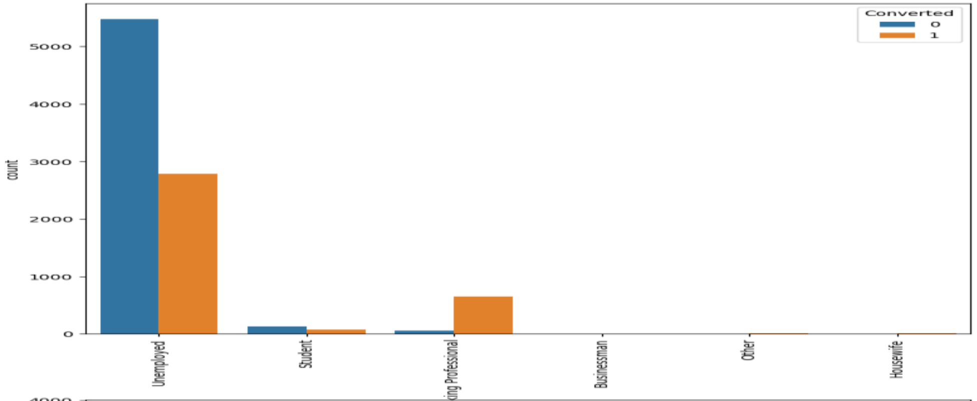


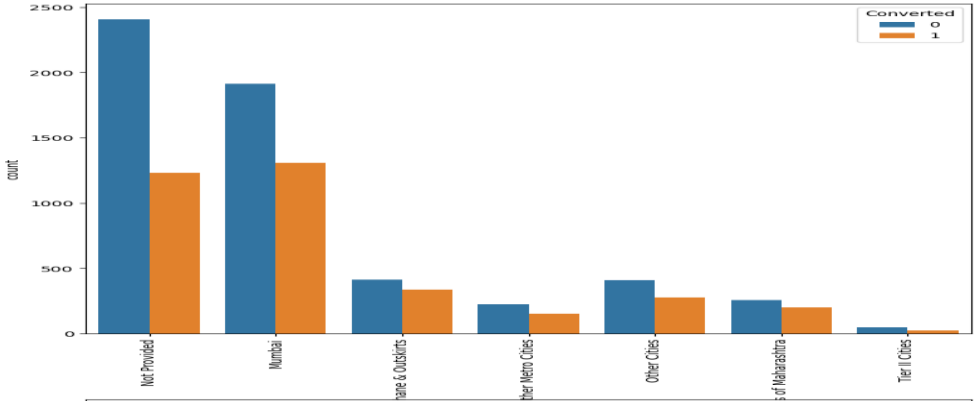


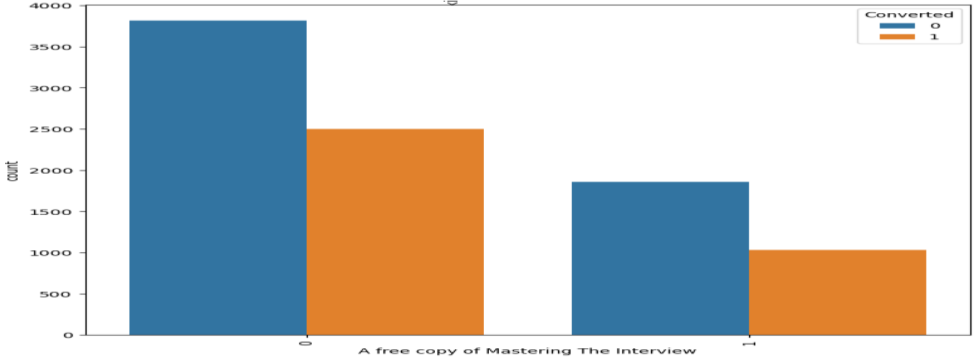
Inferences:

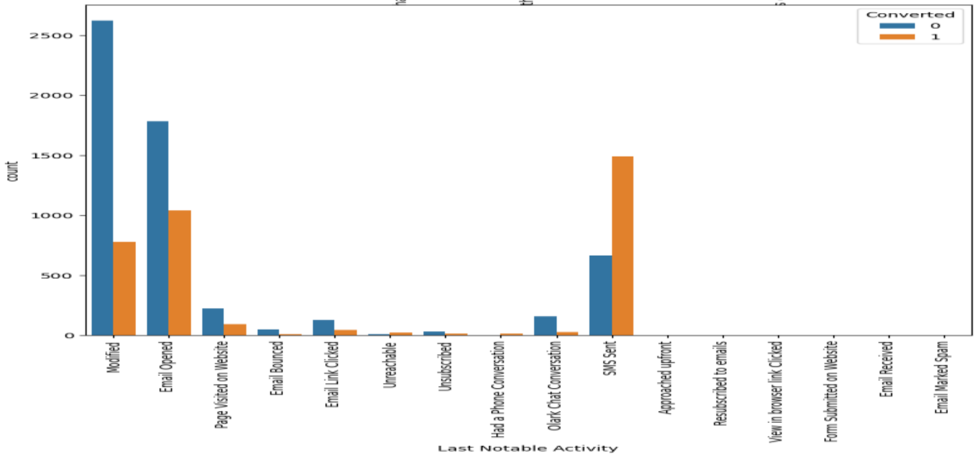
- Users who fill out the Lead Add Form or come through references show a high conversion rate
  Users opting for 'Do not Email' or 'Do not Call' have lower conversion rates

Exploratory Data Analysis – Bivariate (2/3)





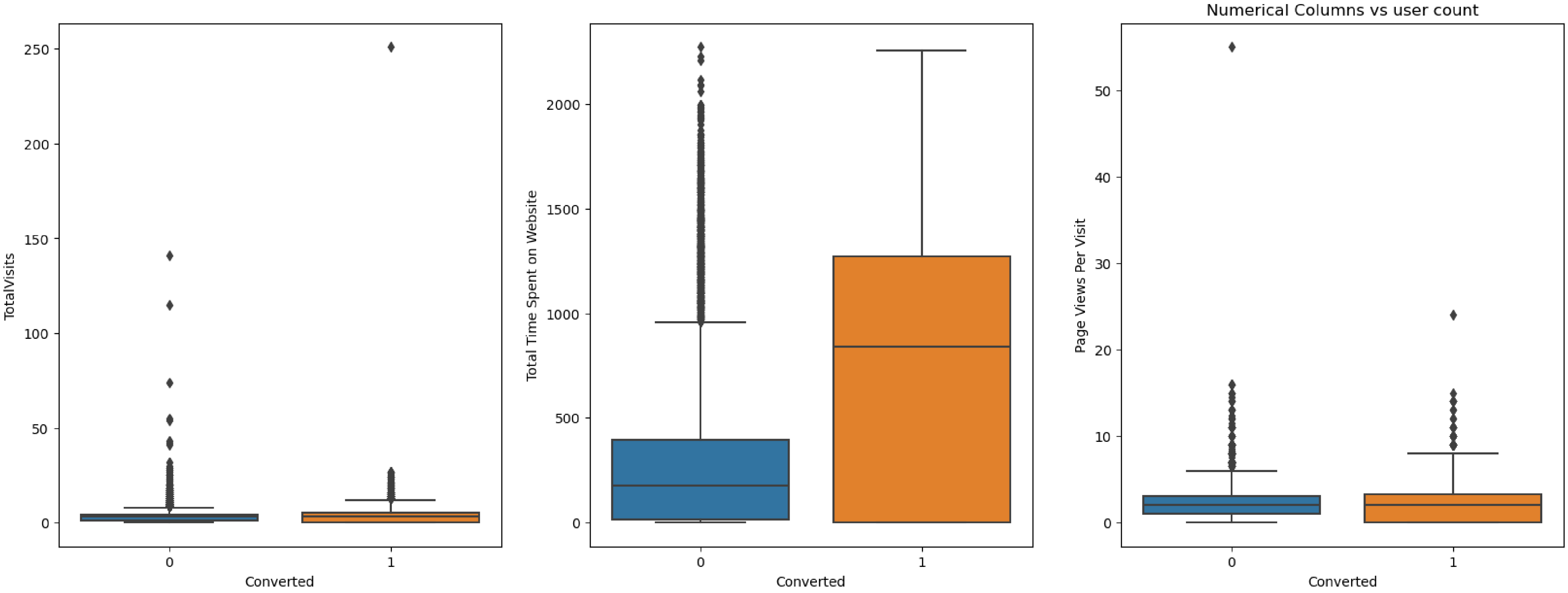




Inferences:

- 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 spend more time on the company website tend to convert at a higher rate
- Total visits and number of pages per visit do not significantly impact the conversion rate

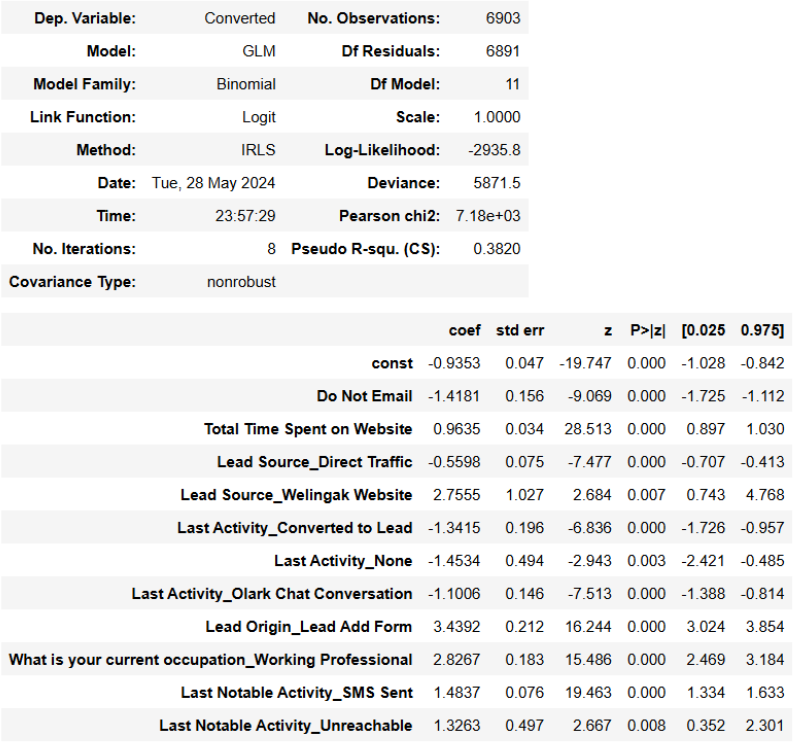
Data Preparation

- 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**

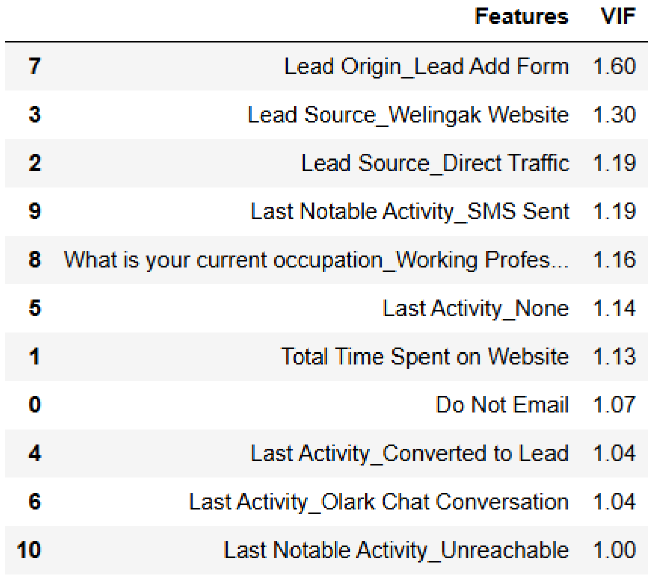
Model Building

- After RFE, the final model was arrived at after 4 iterations



```
   Dep. Variable:        Converted   No. Observations:      6903
          Model:              GLM       Df Residuals:      6891
   Model Family:         Binomial           Df Model:        11
  Link Function:            Logit              Scale:    1.0000
         Method:             IRLS     Log-Likelihood:   -2935.8
           Date:  Tue, 28 May 2024           Deviance:    5871.5
           Time:          23:57:29        Pearson chi2:  7.18e+03
 No. Iterations:                8  Pseudo R-squ. (CS):    0.3820
Covariance Type:        nonrobust

                                                  coef  std err        z  P>|z|  [0.025  0.975]
                                      const    -0.9353    0.047  -19.747  0.000  -1.028  -0.842
                               Do Not Email    -1.4181    0.156   -9.069  0.000  -1.725  -1.112
                  Total Time Spent on Website   0.9635    0.034   28.513  0.000   0.897   1.030
                   Lead Source_Direct Traffic  -0.5598    0.075   -7.477  0.000  -0.707  -0.413
                 Lead Source_Welingak Website   2.7555    1.027    2.684  0.007   0.743   4.768
               Last Activity_Converted to Lead -1.3415    0.196   -6.836  0.000  -1.726  -0.957
                           Last Activity_None  -1.4534    0.494   -2.943  0.003  -2.421  -0.485
         Last Activity_Olark Chat Conversation -1.1006    0.146   -7.513  0.000  -1.388  -0.814
                     Lead Origin_Lead Add Form  3.4392    0.212   16.244  0.000   3.024   3.854
What is your current occupation_Working Professional 2.8267 0.183 15.486 0.000  2.469   3.184
                 Last Notable Activity_SMS Sent 1.4837    0.076   19.463  0.000   1.334   1.633
              Last Notable Activity_Unreachable 1.3263    0.497    2.667  0.008   0.352   2.301
```



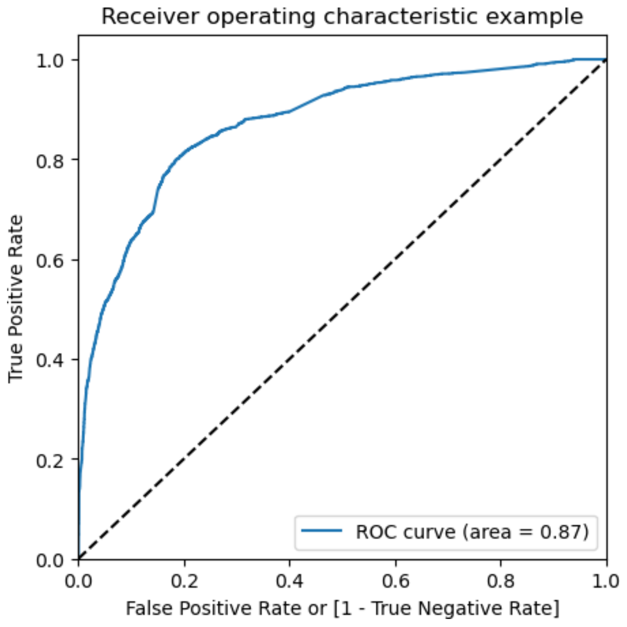
	Features	VIF
7	Lead Origin_Lead Add Form	1.60
3	Lead Source_Welingak Website	1.30
2	Lead Source_Direct Traffic	1.19
9	Last Notable Activity_SMS Sent	1.19
8	What is your current occupation_Working Profes...	1.16
5	Last Activity_None	1.14
1	Total Time Spent on Website	1.13
0	Do Not Email	1.07
4	Last Activity_Converted to Lead	1.04
6	Last Activity_Olark Chat Conversation	1.04
10	Last Notable Activity_Unreachable	1.00

Results:

- **The p-values and VIFs for all variables appear satisfactory**
- **Finally, we will proceed to use this model for making predictions based on the final set of features.**

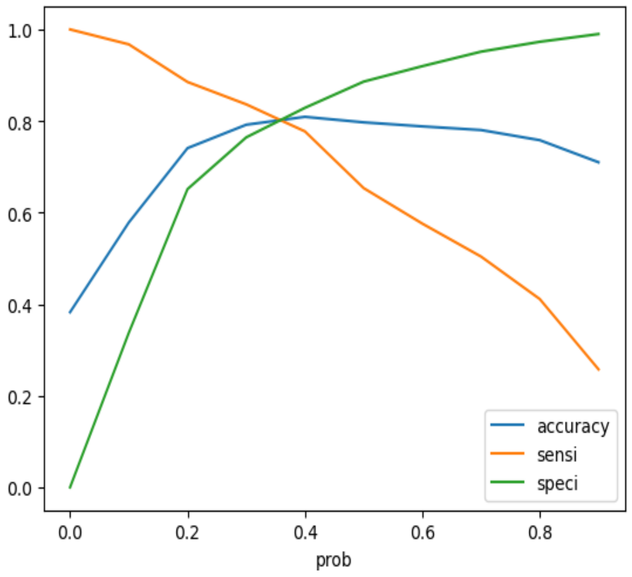
Model Evaluation

ROC Curve



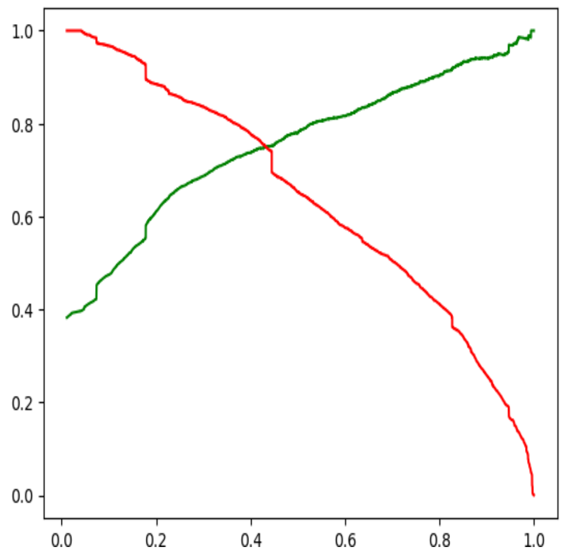
The area under the curve of the ROC is 0.87, which is a good indicator

Optimal cutoff



The optimal cutoff is at probability value of 0.37

Precision Recall Tradeoff



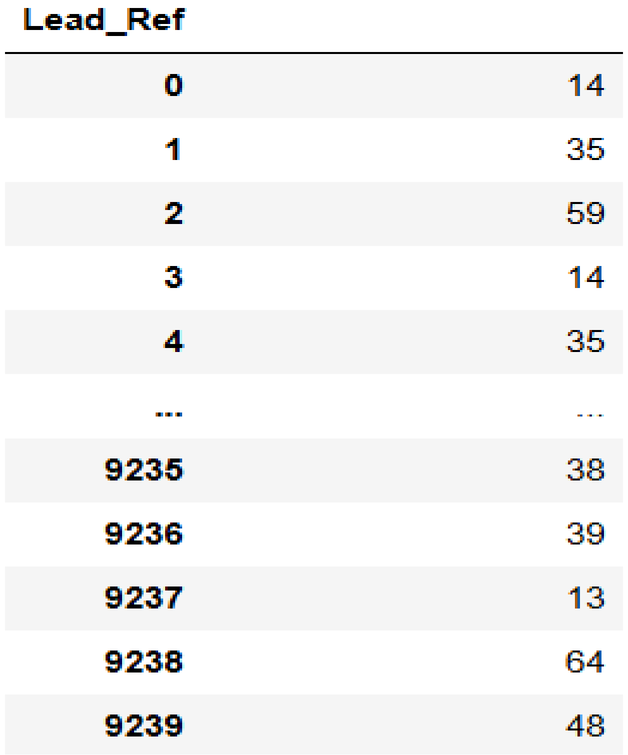
The optimal cutoff is at probability value of 0.43

Evaluation parameters

- Selected 0.43 as the optimal cutoff for conversion probability
- Parameters of model on train set are
  - Accuracy: 80%
  - Sensitivity: 74%
  - Specificity: 72%
- Parameters of model on test set are
  - Accuracy: 80%
  - Sensitivity: 74%
  - Specificity: 83%

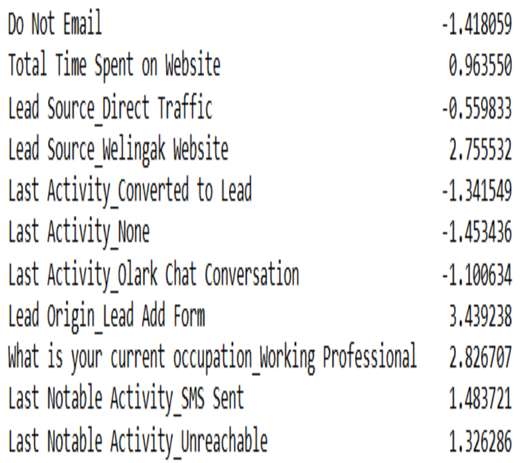
Conclusion

Lead scores of all leads were calculated



Lead_Ref	
0	14
1	35
2	59
3	14
4	35
...	...
9235	38
9236	39
9237	13
9238	64
9239	48

Final list of coefficients of most important features



```
Do Not Email                                      -1.418059
Total Time Spent on Website                        0.963550
Lead Source_Direct Traffic                        -0.559833
Lead Source_Welingak Website                       2.755532
Last Activity_Converted to Lead                   -1.341549
Last Activity_None                                -1.453436
Last Activity_Olark Chat Conversation             -1.100634
Lead Origin_Lead Add Form                          3.439238
What is your current occupation_Working Professional 2.826707
Last Notable Activity_SMS Sent                     1.483721
Last Notable Activity_Unreachable                  1.326286
```

The below columns are used to predict if the lead is likely to be converted with approximately 80% accuracy.

- **Feature Correlation:**
  - Do Not Email
  - Total Time Spent on Website
  - Lead Source_Direct Traffic
  - Lead Source_Welingak Website
  - Last Activity_Converted to Lead
  - Last Activity_None
  - Last Activity_Olark Chat Conversation
  - Lead Origin_Lead Add Form
  - What is your current occupation_Working Professional
  - Last Notable Activity_SMS Sent
  - Last Notable Activity_Unreachable

Thank You.