LEAD SCORE CASE STUDY

Ganesh Sali
Geo Abraham
Gautami

INTRODUCTION:

Professionals in the business can purchase online courses from X Education, an education company. The business advertises its classes on a number of websites and search engines, including Google.
After visiting the website, users can peruse the available courses, complete the course registration form, or watch some videos. These folks are categorized as leads when they complete a form with their phone number or email address. Additionally, the business receives leads from previous recommendations. After obtaining these leads, sales team members begin calling, emailing, and so on. At X education, the lead conversion rate is typically 30%.

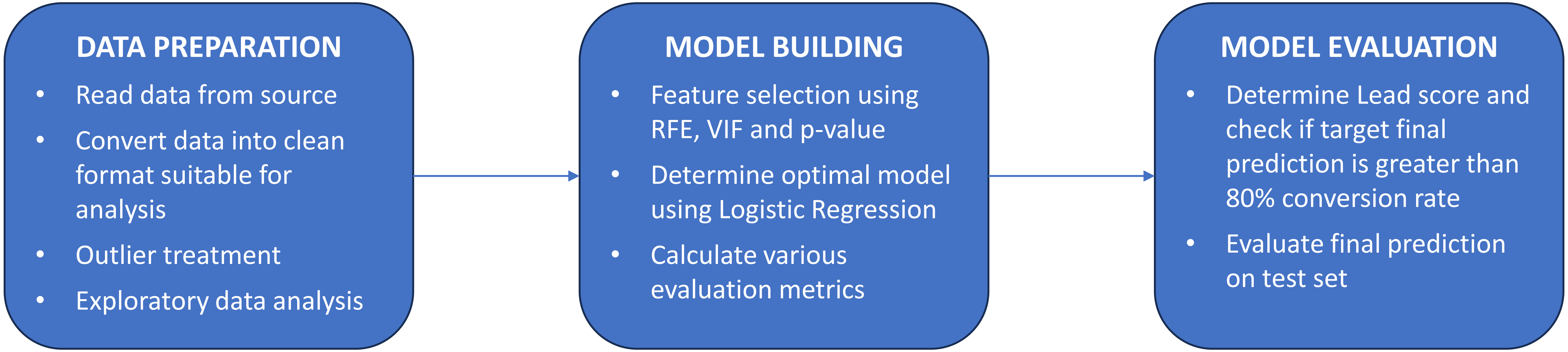
Problem Statement:

A company called X-Education provides professional online education courses and internet marketing through adverts. The company uses a variety of ways to obtain information, and it calls leads who inquire about a particular degree of schooling. Lead conversion is usually 30% of specific education. The company also uses specific criteria to identify Hot Leads. The ratio of leads converted to enrolments is lower. company provided Aim for 80% of the total enrolled.

BUSINESS GOAL:

"Hot Leads" are leads that have the highest potential and are sought after by the company. The business requires a model in which each lead is given a score, with the goal being to increase the conversion chance of a client with a higher lead score and decrease it for a customer with a lower lead score. Specifically, the CEO provided an approximate figure of 80% for the lead conversion rate.

METHODOLOGY



DATA PREPARATION

- Read data from source
- Convert data into clean format suitable for analysis
- Outlier treatment
- Exploratory data analysis

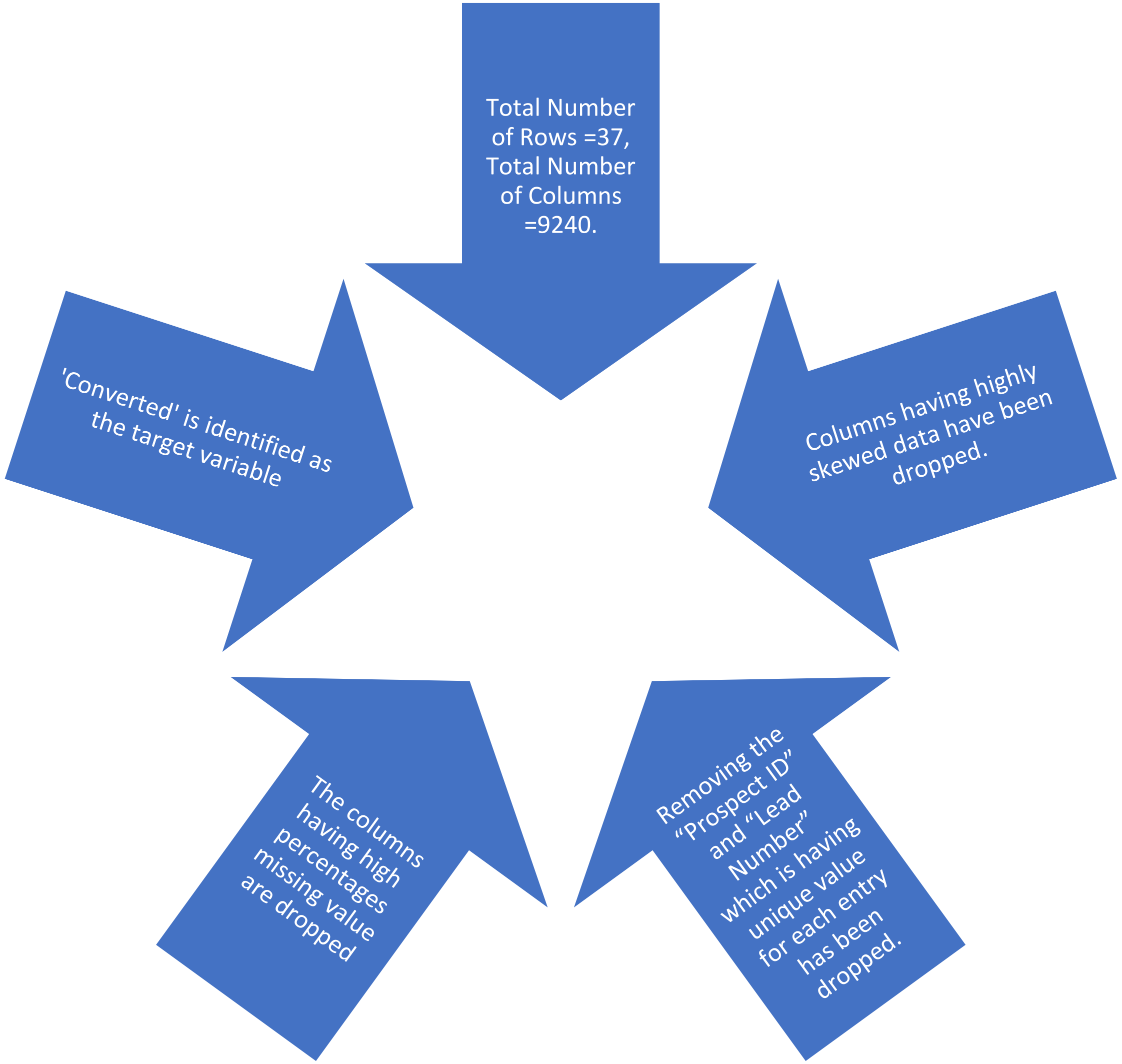
MODEL BUILDING

- Feature selection using RFE, VIF and p-value
- Determine optimal model using Logistic Regression
- Calculate various evaluation metrics

MODEL EVALUATION

- Determine Lead score and check if target final prediction is greater than 80% conversion rate
- Evaluate final prediction on test set

DATA MANIPULATION



Total Number of Rows =37, Total Number of Columns =9240.

Columns having highly skewed data have been dropped.

Removing the “Prospect ID” and “Lead Number” which is having unique value for each entry has been dropped.

The columns having high percentages missing value are dropped

'Converted' is identified as the target variable

EXPLORATORY DATA ANALYSIS – Unique Values

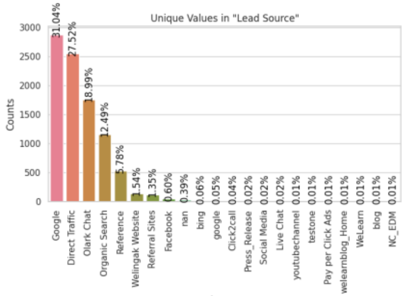


Figure 1

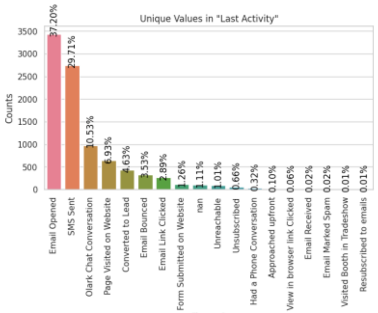


Figure 2

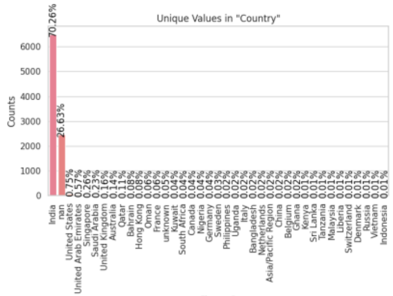


Figure 3

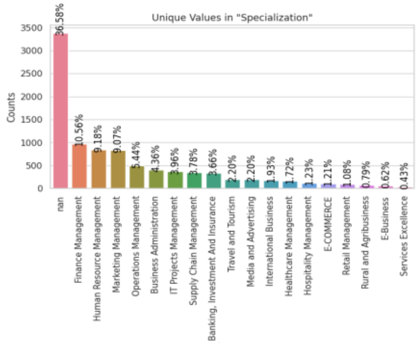


Figure 4

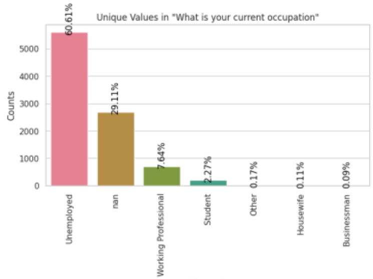


Figure 5

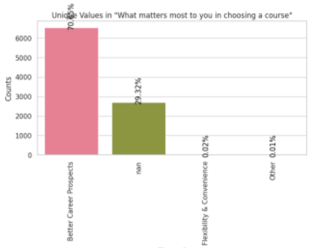


Figure 6

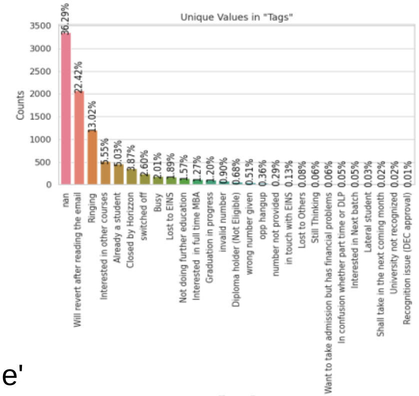


Figure 7

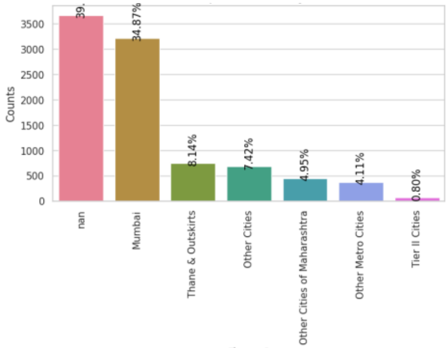


Figure 8

Figure 1: There are "NaN"(0.39%) and two 'Google' and 'google' values.

Figure 2: Last activity of leads are having missing values(1.11%).

Figure 3: 'Country' has 26.63% missing values.

Figure 4: The "Specialization" column contains 37% missing values.If the lead is a student, does not have a specialty, or if his specialty is not included among the alternatives, it is probable that he will leave this section blank. Thus, for this, we can make a new category called "Others."

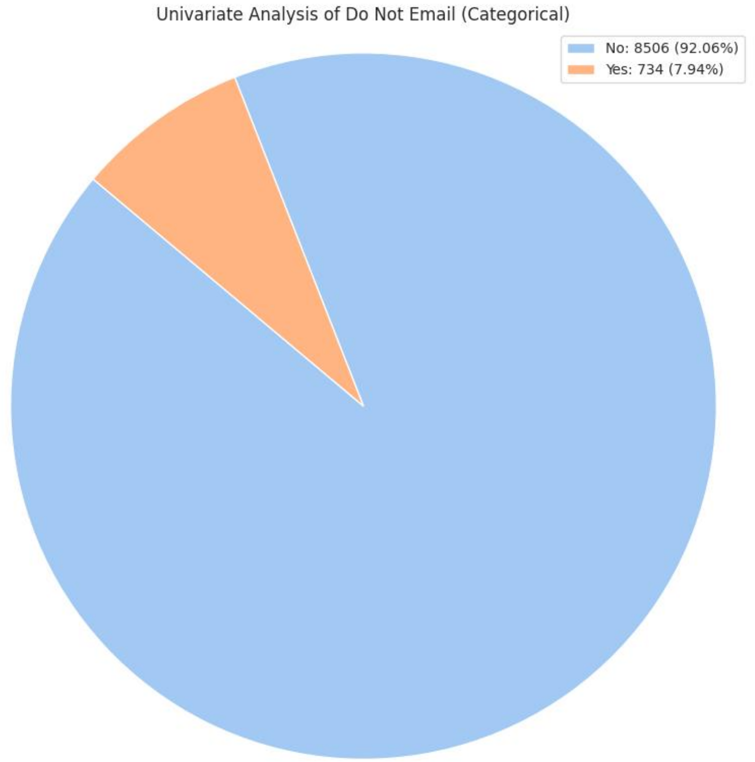
Figure 5: This column has 29.11% missing values.

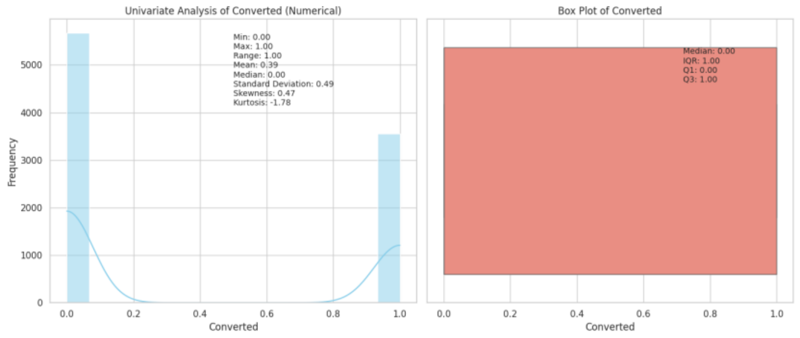
Figure 6: This column is highly skewed and has 29.32% missing values.This column can be removed

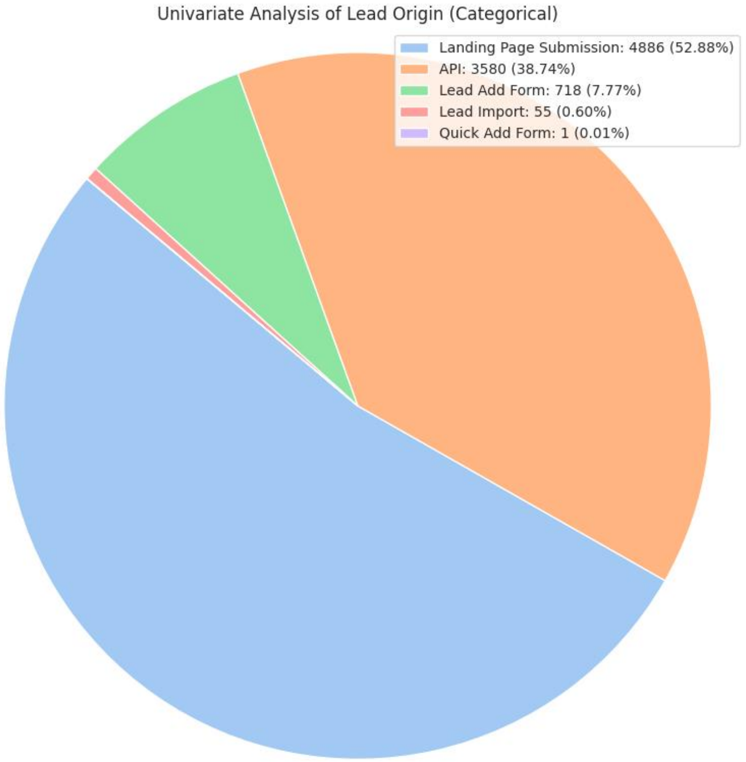
Figure 7: 'Tags' column has 36.29% missing values.

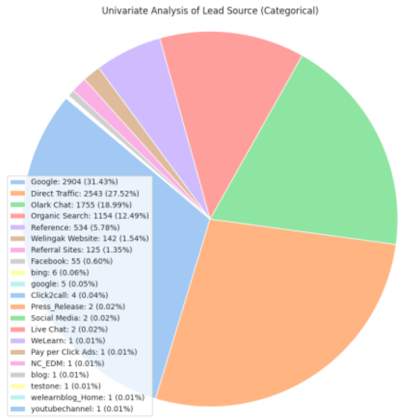
Figure 8: 'City' column has 39.71% missing values

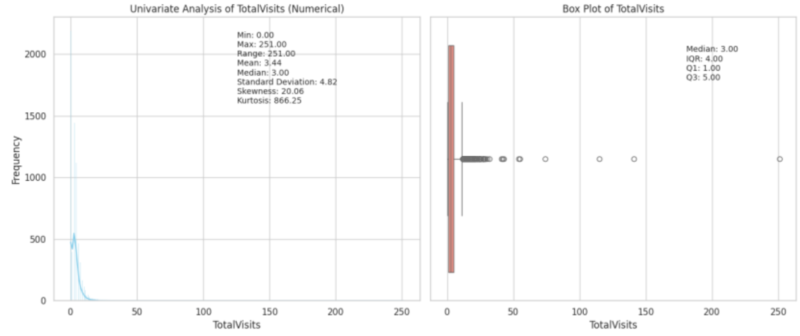
EXPLORATORY DATA ANALYSIS - Univariate



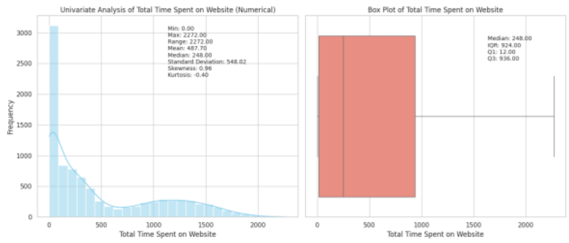


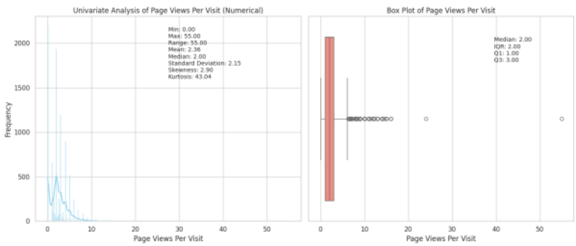


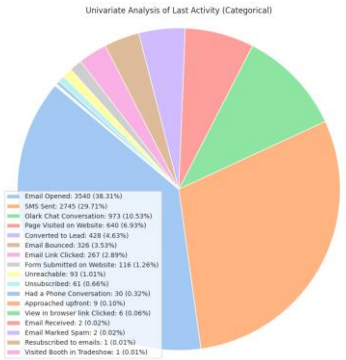


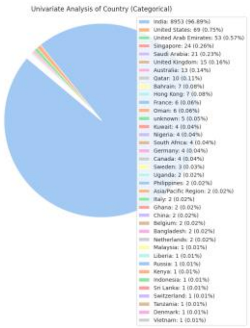


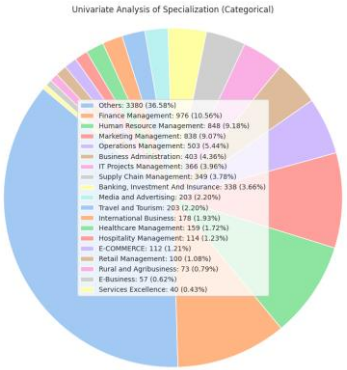
EXPLORATORY DATA ANALYSIS - Univariate

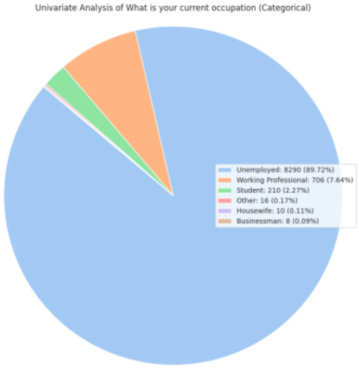


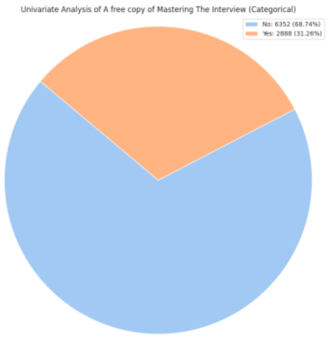




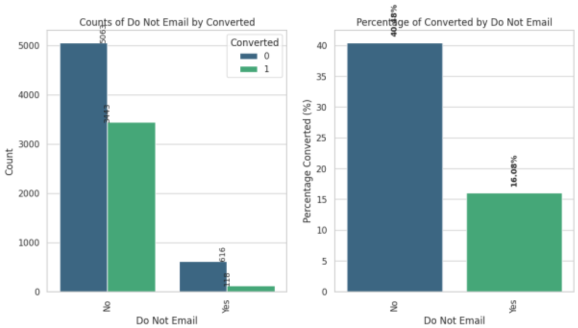


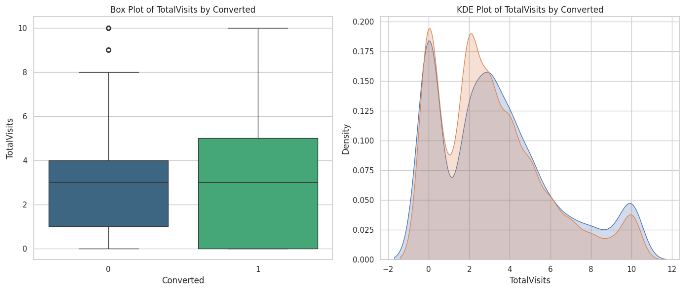


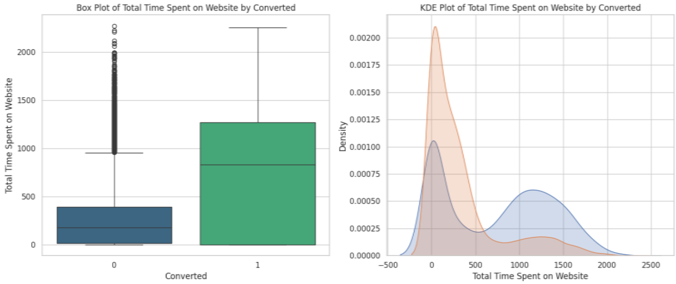


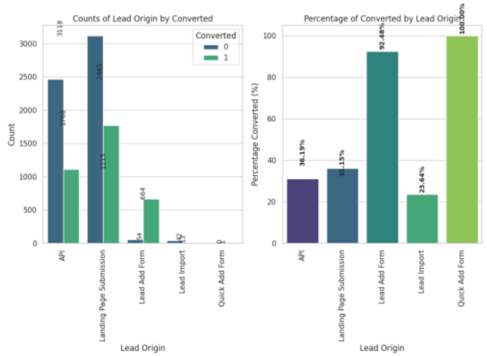


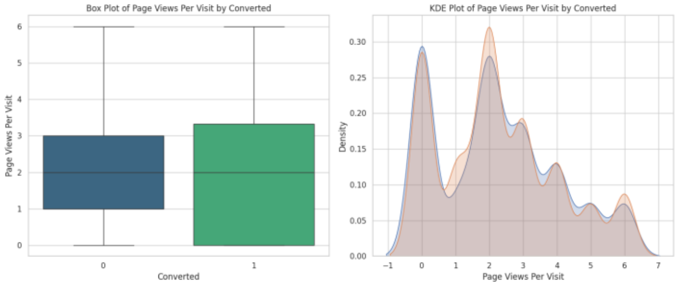
EXPLORATORY DATA ANALYSIS - Bivariate

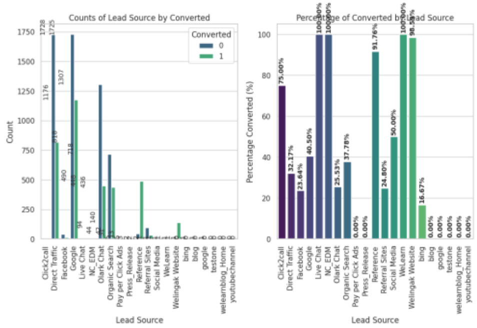




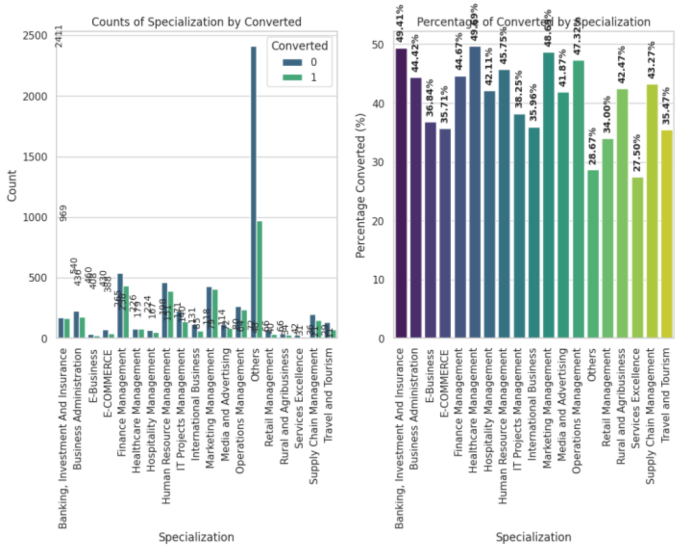


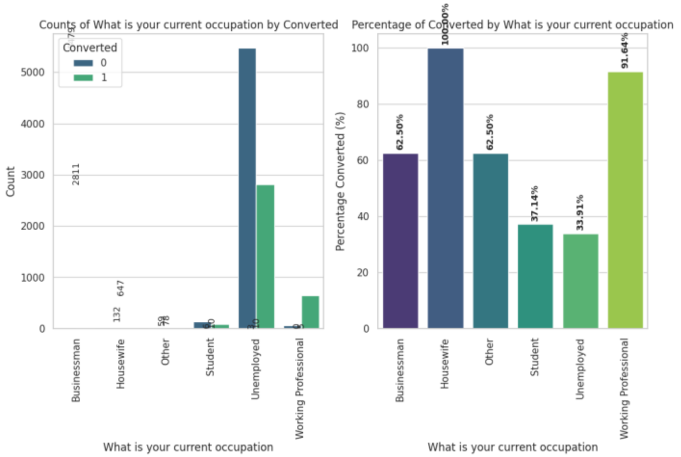


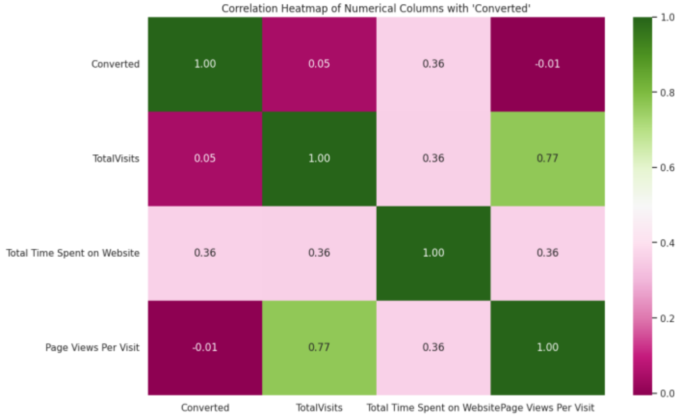


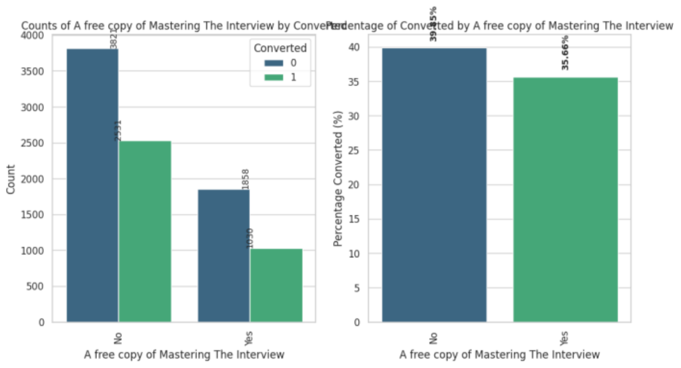


EXPLORATORY DATA ANALYSIS - Bivariate

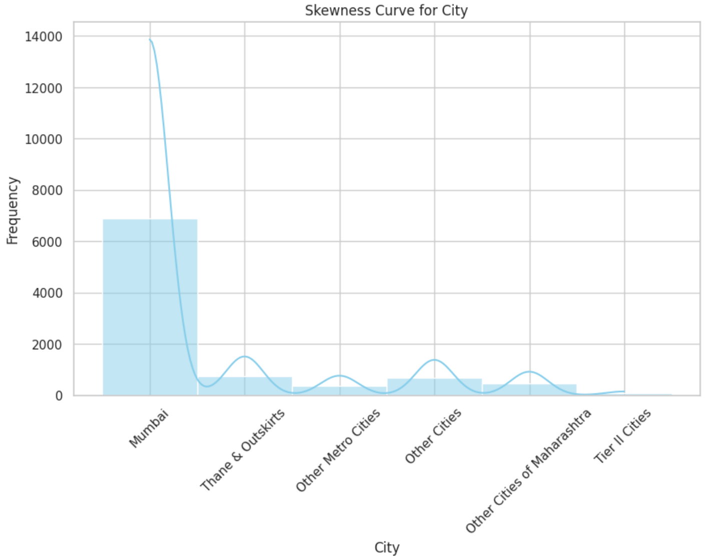


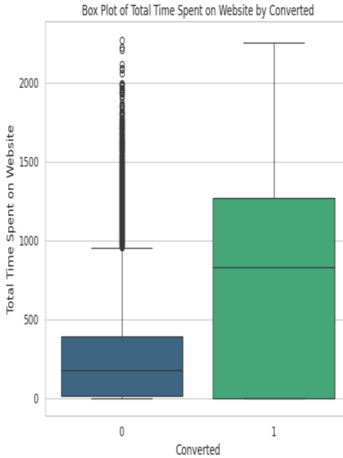


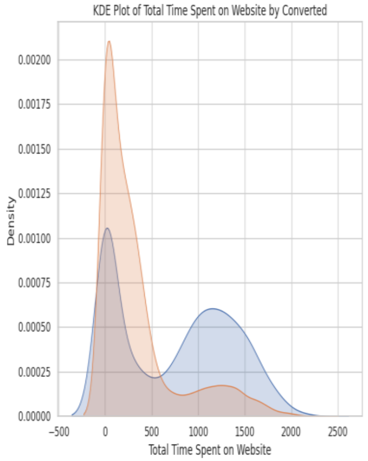


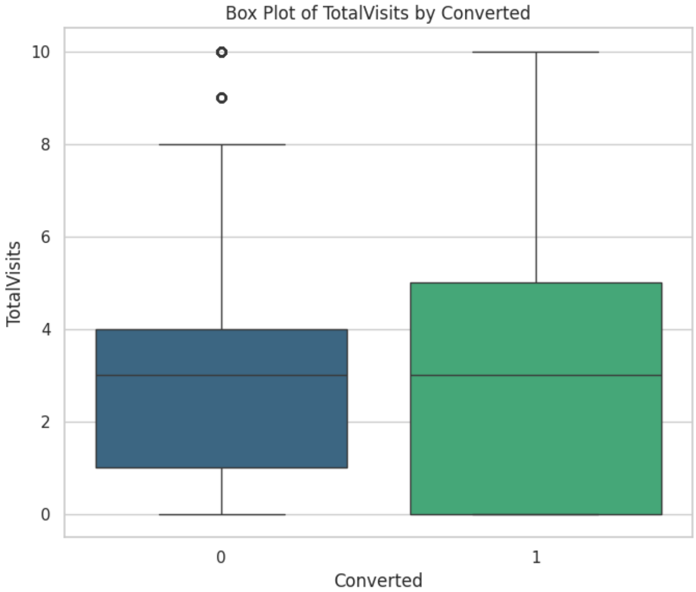


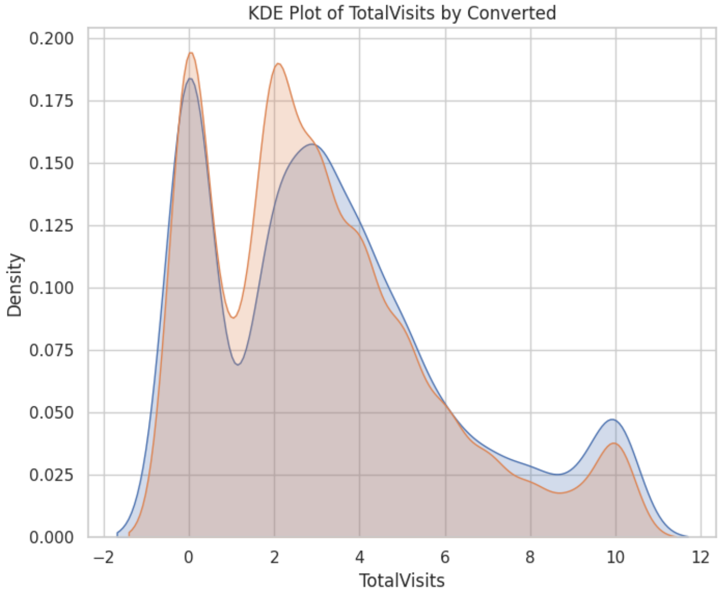
EXPLORATORY DATA ANALYSIS - Bivariate





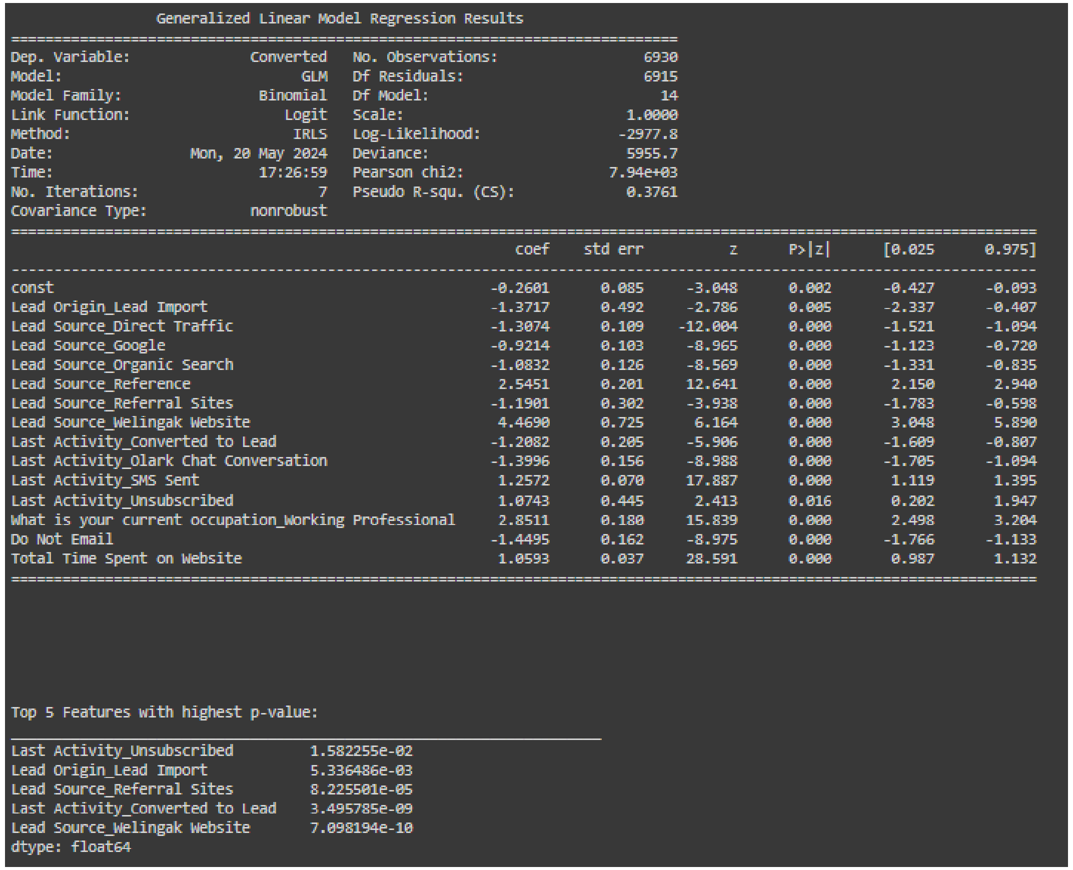






MODEL BUILDING

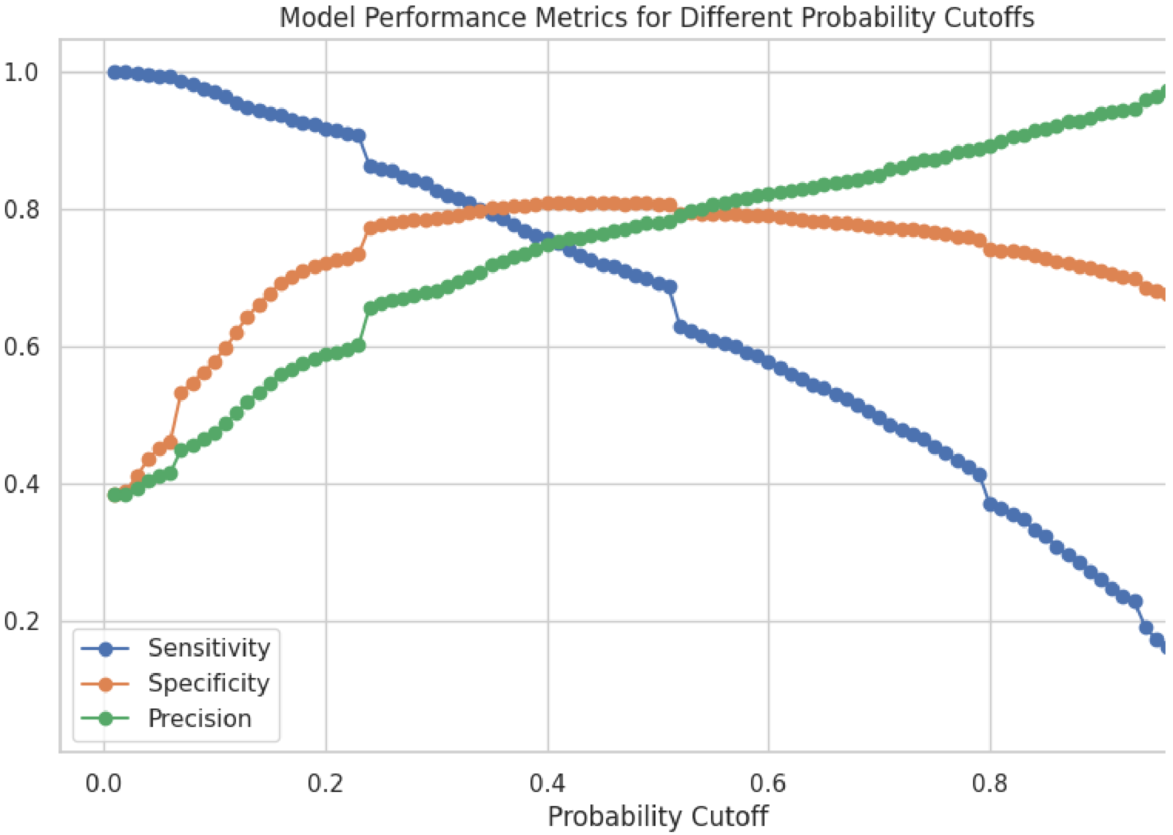
- Split Data into train and test data with ratio 75:25
- Using RFE to choose variables to be accepted or rejected
- Build the right model by removing variables whose p-value > 0.05 and VIF >5

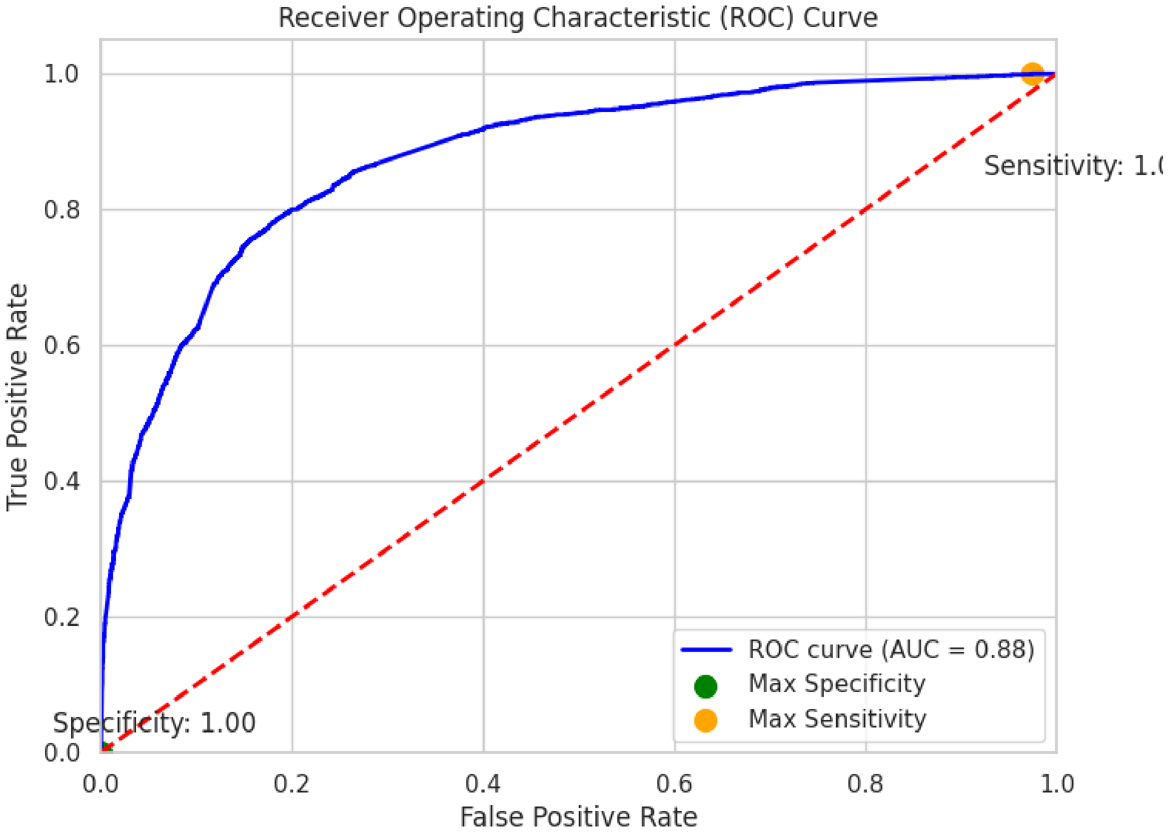


```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Converted   No. Observations:                 6930
Model:                            GLM   Df Residuals:                     6915
Model Family:                Binomial   Df Model:                           14
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -2977.8
Date:                Mon, 20 May 2024   Deviance:                       5955.7
Time:                        17:26:59   Pearson chi2:                 7.94e+03
No. Iterations:                     7   Pseudo R-squ. (CS):             0.3761
Covariance Type:            nonrobust
=====================================================================================================================
                                                          coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------------------------------------------
const                                                  -0.2601      0.085     -3.048      0.002      -0.427      -0.093
Lead Origin_Lead Import                                -1.3717      0.492     -2.786      0.005      -2.337      -0.407
Lead Source_Direct Traffic                             -1.3074      0.109    -12.004      0.000      -1.521      -1.094
Lead Source_Google                                     -0.9214      0.103     -8.965      0.000      -1.123      -0.720
Lead Source_Organic Search                             -1.0832      0.126     -8.569      0.000      -1.331      -0.835
Lead Source_Reference                                   2.5451      0.201     12.641      0.000       2.150       2.940
Lead Source_Referral Sites                             -1.1901      0.302     -3.938      0.000      -1.783      -0.598
Lead Source_Welingak Website                            4.4690      0.725      6.164      0.000       3.048       5.890
Last Activity_Converted to Lead                        -1.2082      0.205     -5.906      0.000      -1.609      -0.807
Last Activity_Olark Chat Conversation                  -1.3996      0.156     -8.988      0.000      -1.705      -1.094
Last Activity_SMS Sent                                  1.2572      0.070     17.887      0.000       1.119       1.395
Last Activity_Unsubscribed                              1.0743      0.445      2.413      0.016       0.202       1.947
What is your current occupation_Working Professional    2.8511      0.180     15.839      0.000       2.498       3.204
Do Not Email                                           -1.4495      0.162     -8.975      0.000      -1.766      -1.133
Total Time Spent on Website                             1.0593      0.037     28.591      0.000       0.987       1.132
=====================================================================================================================



Top 5 Features with highest p-value:

____________________________________________________________________
Last Activity_Unsubscribed         1.582255e-02
Lead Origin_Lead Import            5.336486e-03
Lead Source_Referral Sites         8.225501e-05
Last Activity_Converted to Lead    3.495785e-09
Lead Source_Welingak Website       7.098194e-10
dtype: float64
```

OPTIMAL CUT OFF PROBABILITY

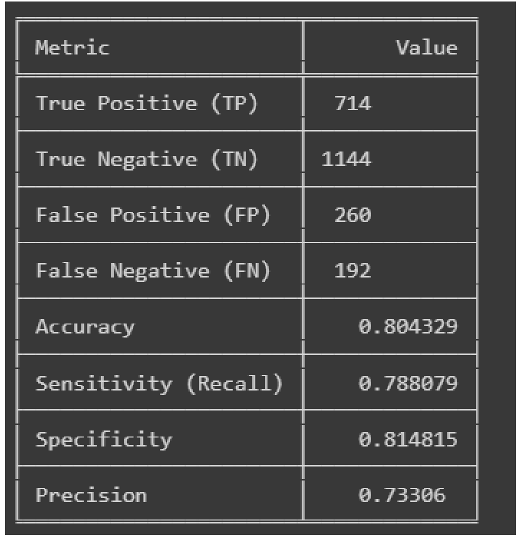
From graph, we are taking Overall Optimal Cutoff: 0.35





FINAL TRAINED MODEL

Evaluation metrics

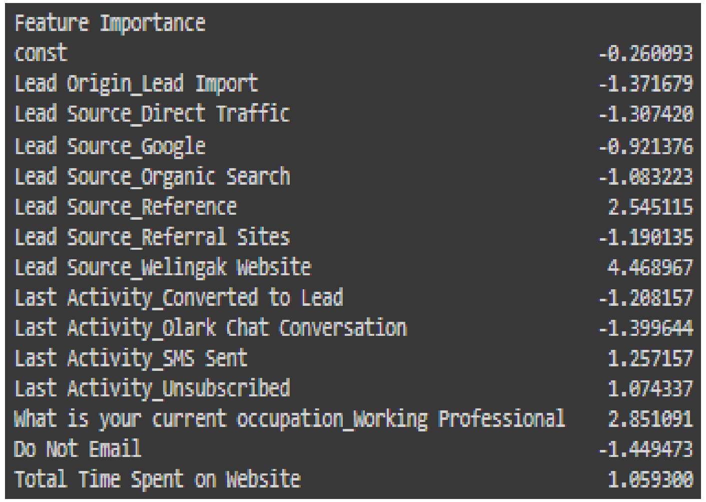


Metric	Value
True Positive (TP)	714
True Negative (TN)	1144
False Positive (FP)	260
False Negative (FN)	192
Accuracy	0.804329
Sensitivity (Recall)	0.788079
Specificity	0.814815
Precision	0.73306

Main Features that determine the ‘Conversion’ are:

- Lead Source_Welingak Website
- What is your current occupation_Working Professional
- Lead Source_Reference

Feature Importance



```
Feature Importance
const                                                   -0.260093
Lead Origin_Lead Import                                 -1.371679
Lead Source_Direct Traffic                              -1.307420
Lead Source_Google                                      -0.921376
Lead Source_Organic Search                              -1.083223
Lead Source_Reference                                    2.545115
Lead Source_Referral Sites                              -1.190135
Lead Source_Welingak Website                             4.468967
Last Activity_Converted to Lead                         -1.208157
Last Activity_Olark Chat Conversation                   -1.399644
Last Activity_SMS Sent                                   1.257157
Last Activity_Unsubscribed                               1.074337
What is your current occupation_Working Professional     2.851091
Do Not Email                                            -1.449473
Total Time Spent on Website                              1.059300
```

CONCLUSION

The logistic regression model is used to predict the probability of conversion of a customer.

While we have calculated both sensitivity-specificity as well as Precision-Recall metrics, we have considered optimal cut off on the basis of sensitivity-specificity for final prediction

Lead Score calculated shows the conversion rate of final predicted model is around 80.4329 % in test data as compared to 80.1876% in train data, as both scores are similar, proving a good model

In Business terms, this model has capability to adjust with the company's requirements in coming future