Lead Scoring Case study

by akshay tula

Problem Statement

X Education aims to enhance lead conversion rates by identifying 'Hot Leads' using a dataset of 9000 leads and attributes like Lead Source, Total Time Spent on Website, etc. The objective is to build a model assigning lead scores to prioritize higher conversion potential, aiming to increase the rate from 30% to around 80%. Crucially, managing categorical variables, especially the 'Select' level resembling null values, is vital for precise analysis and modeling accuracy.

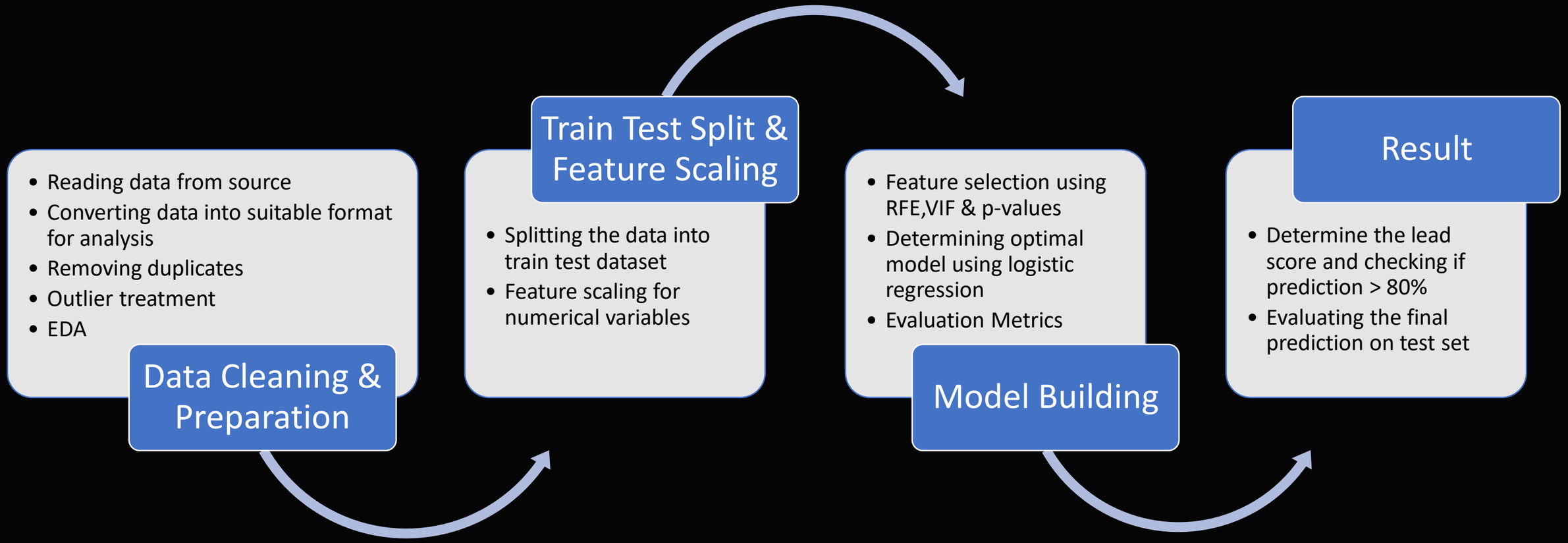
Business Goals

1. The company aims to pinpoint 'Hot Leads'—the most promising potential leads.
2. t requires a model to assign lead scores, ensuring higher scores correspond to increased conversion chances while lower scores signify lower conversion rates.
3. CEO has set an ambitious 80% lead conversion target.

Strategy Applied

1. Data Cleaning and imputing missing values
2. Exploratory Data Analysis: Univariate, Bivariate & Multivariate Analysis
3. Feature scaling and dummy variable creation
4. Logistic Regression Model building & RFE
5. Model Evaluation: Specificity, Sensitivity, Precision & Recall
6. Conclusion & necessary Recommendations

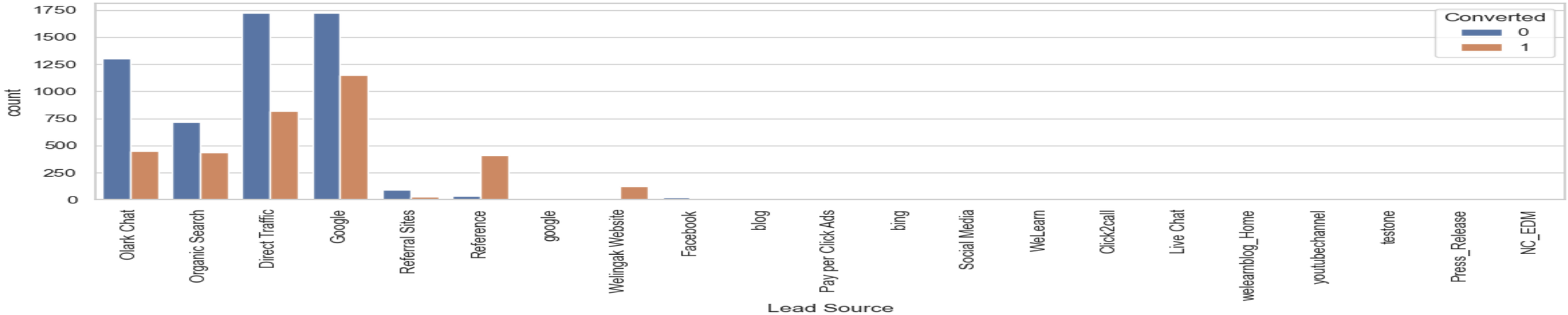
Problem Solving Methodology

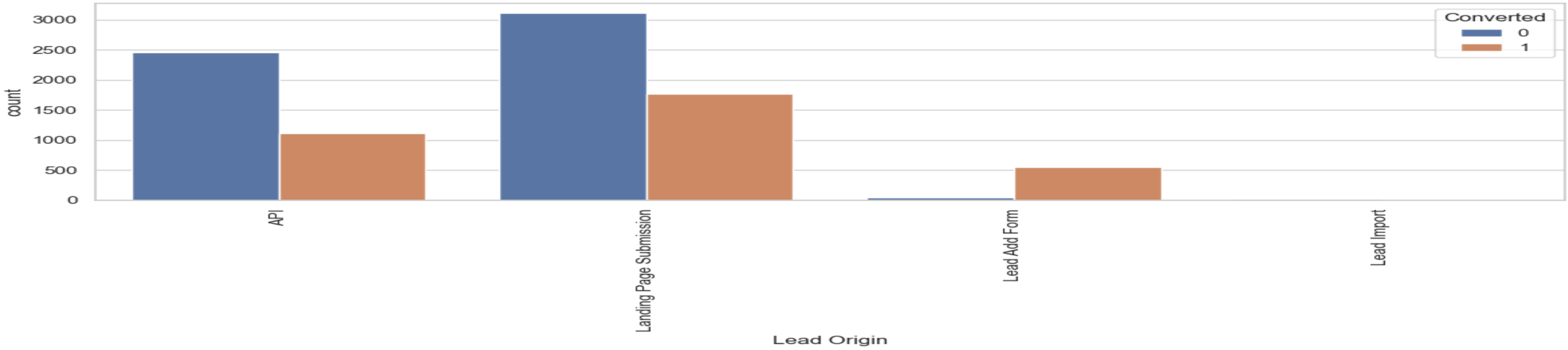


Data Conversion Mechanism

1. Converting the variables with values Yes/No to 1/0
2. Converting the ‘Select’ values with NAN’s
3. Dropping the column having >70% null values
4. Dropping unnecessary columns
5. Dropping the rows as the null values<2%

Exploratory Data Analysis

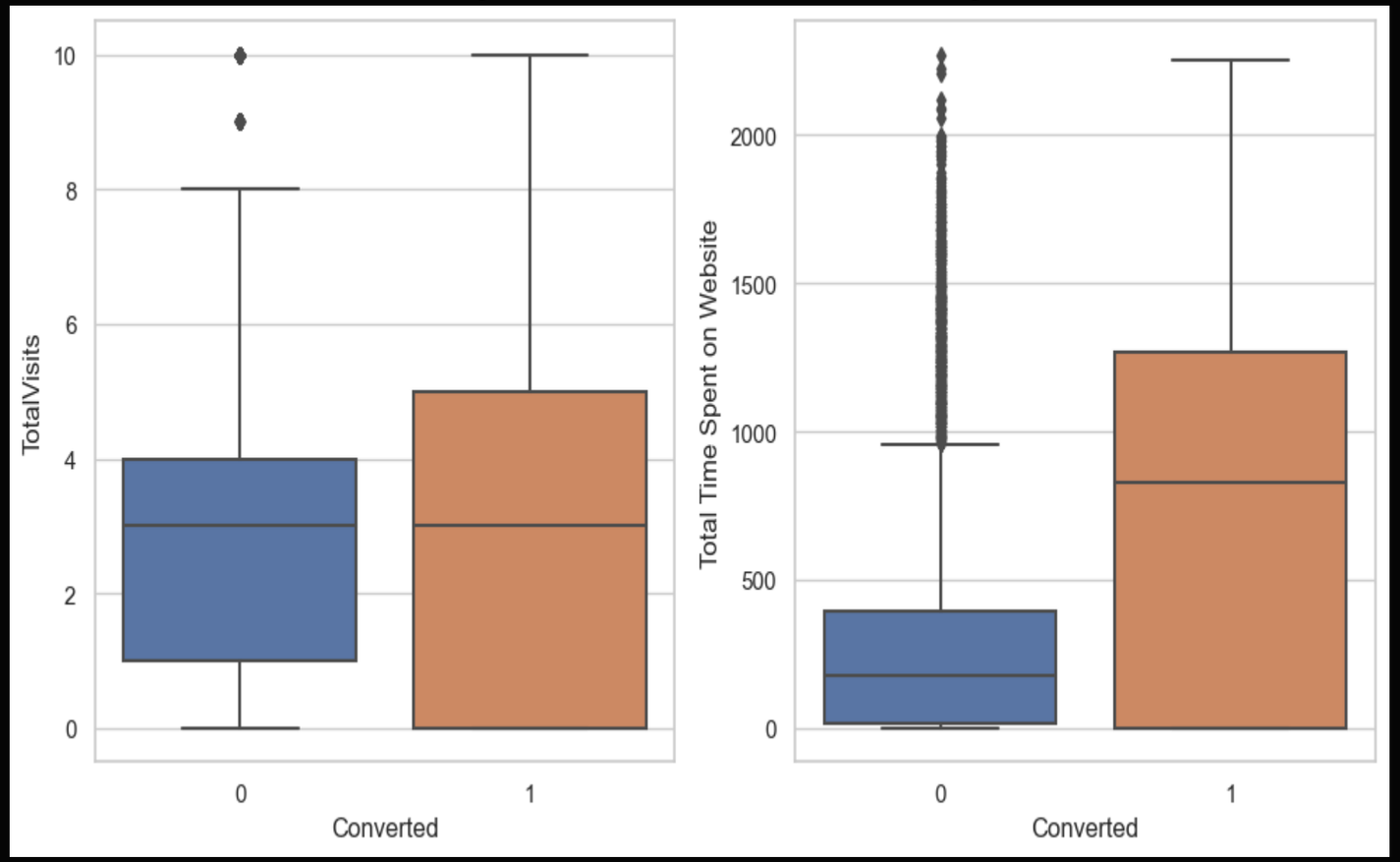


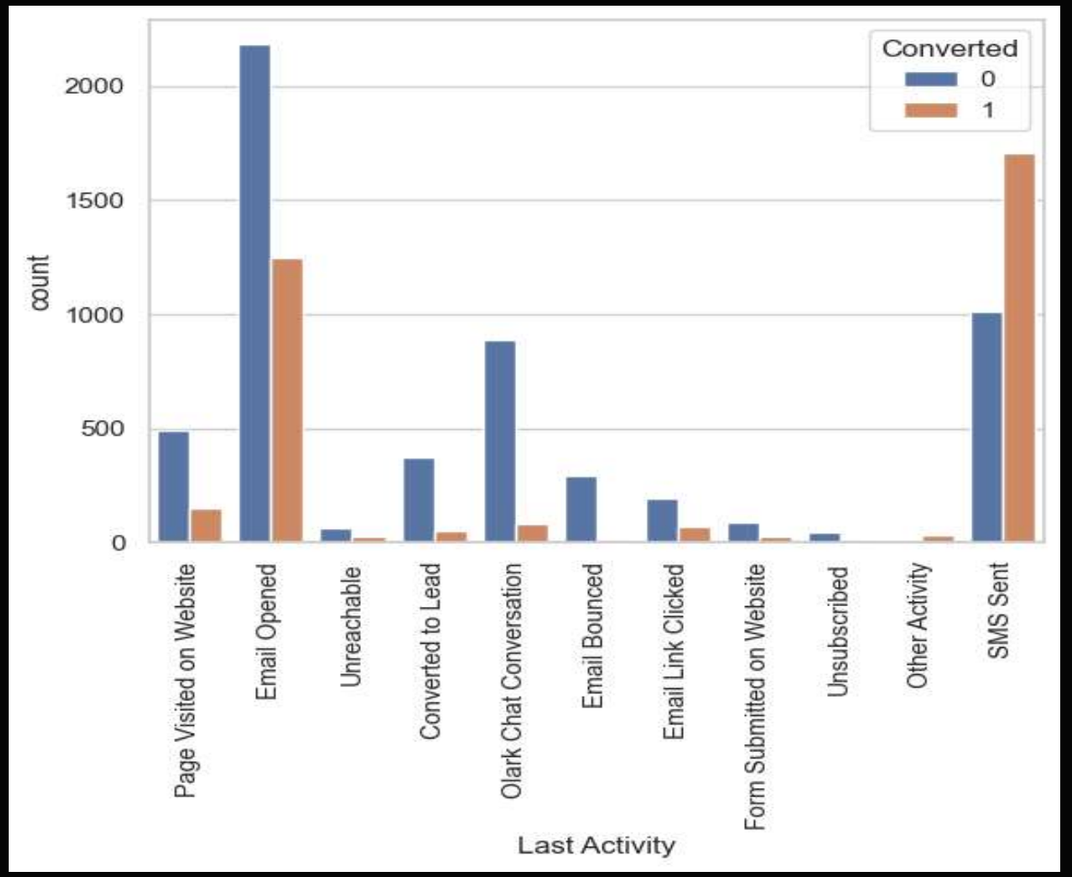


- Lead Conversion Rate is around 30%
- Count of leads from Google & direct traffic is maximum

- Conversion of leads from Reference & Welingak website is maximum
- API & landing page submissions has low conversion approx. 30%, however it has considerable page visits

Exploratory Data Analysis

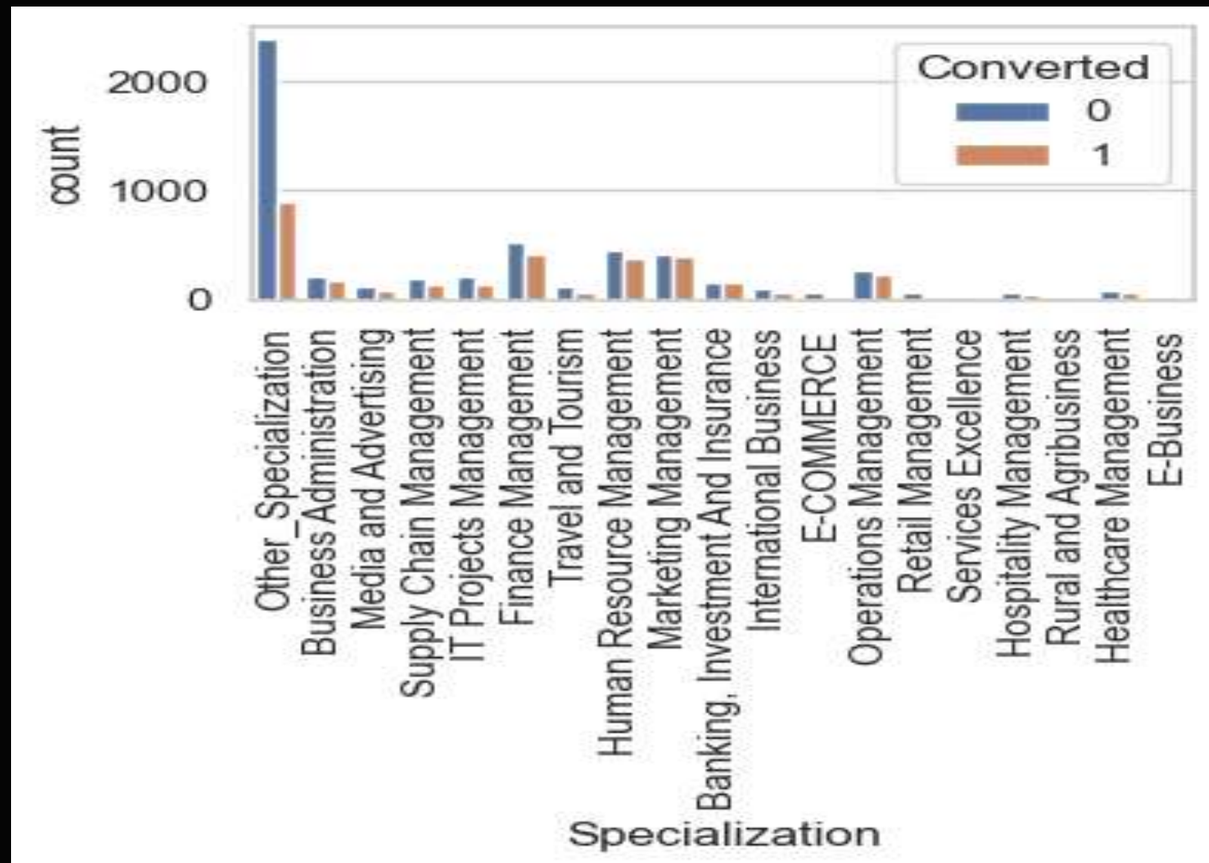


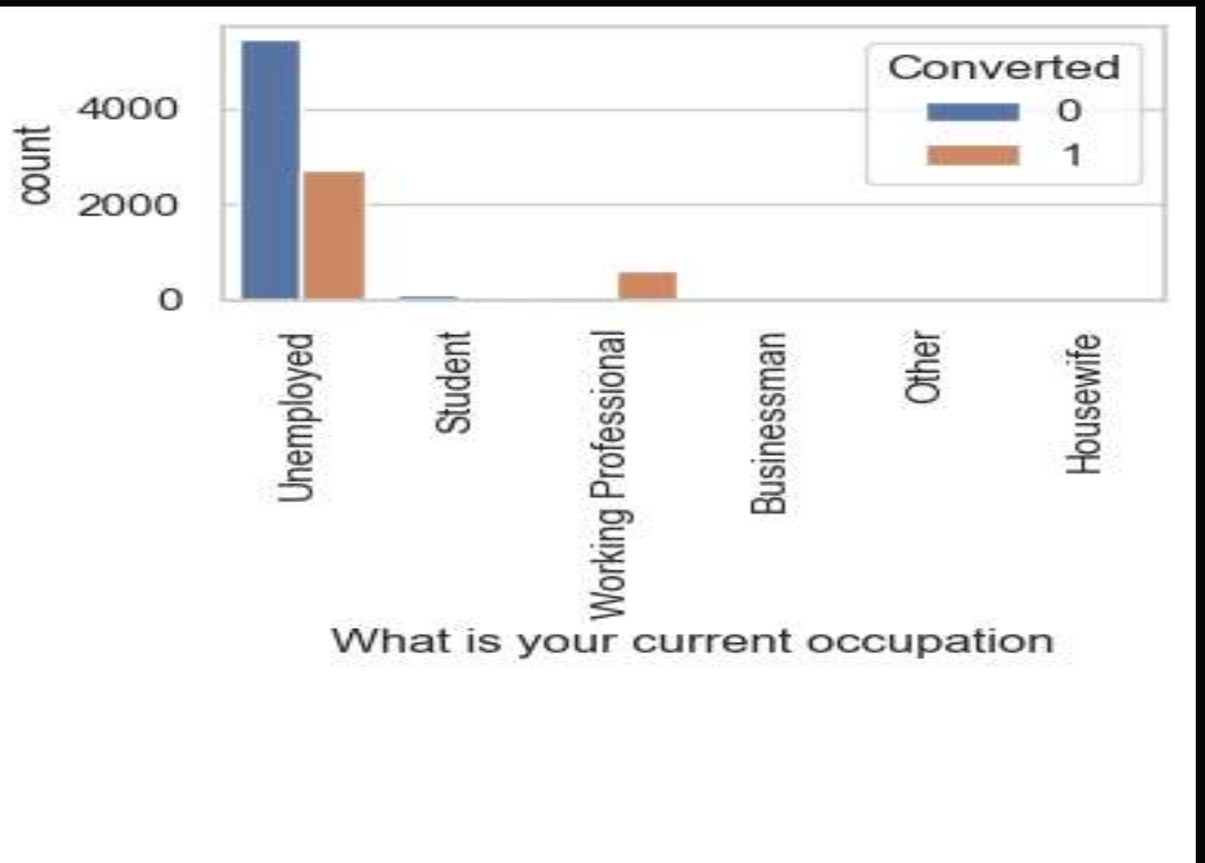


- Median of both the conversions are same hence nothing conclusive can be inferred
- Users spending more time on website are more likely to get converted

- Count of last activity as ‘Email Opened’ is maximum
- Conversion rate of SMS sent as last activity is maximum

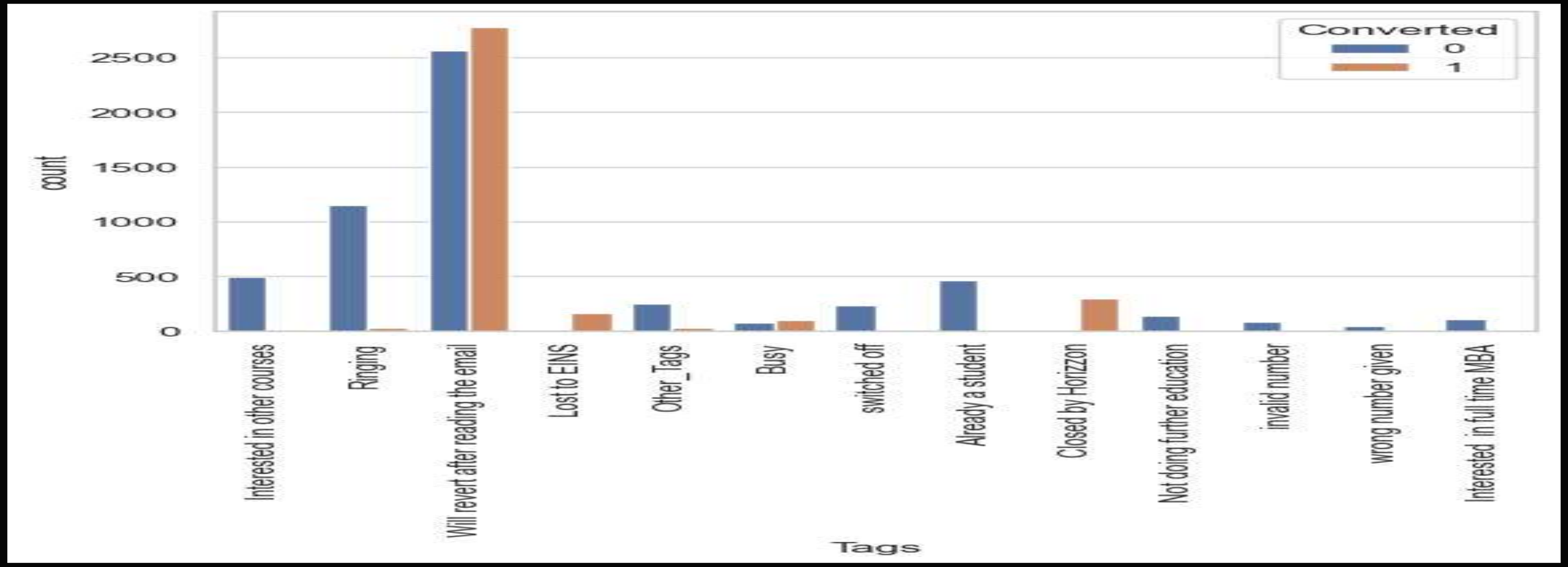
Exploratory Data Analysis





- No of unemployed leads are significantly high as compared to other occupations
- Working Professionals have high conversion rate

Exploratory Data Analysis



- Users having a tag with ‘Will revert after reading email’ has a high conversion rate
- Users having a tag replying with ‘Closed by horizon’ have decent conversion rate

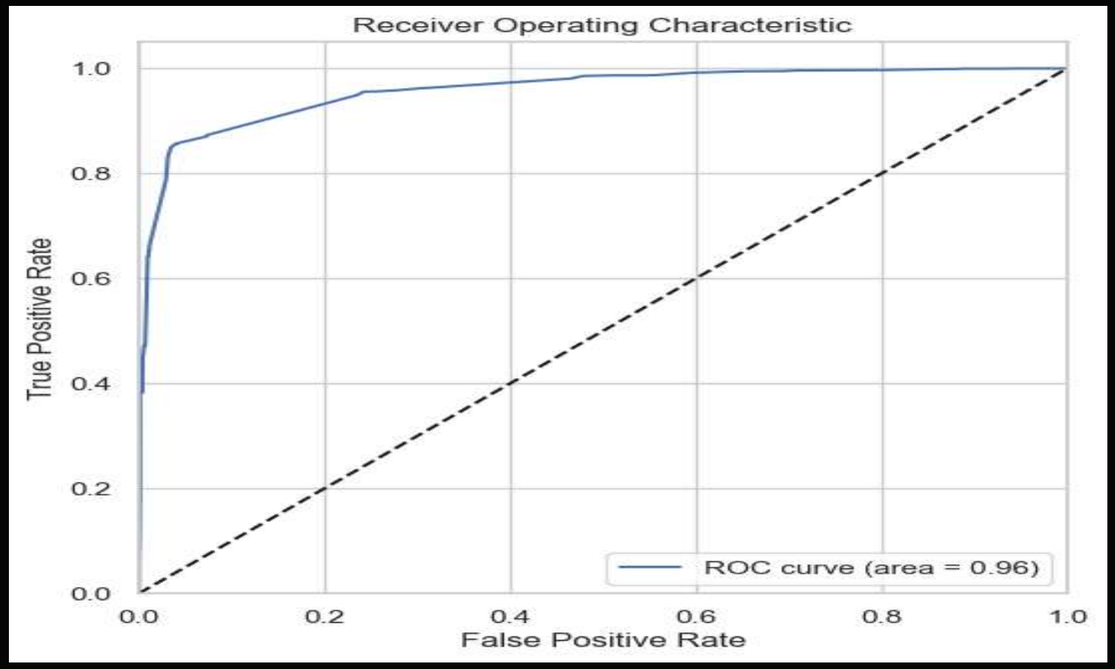
Model Building

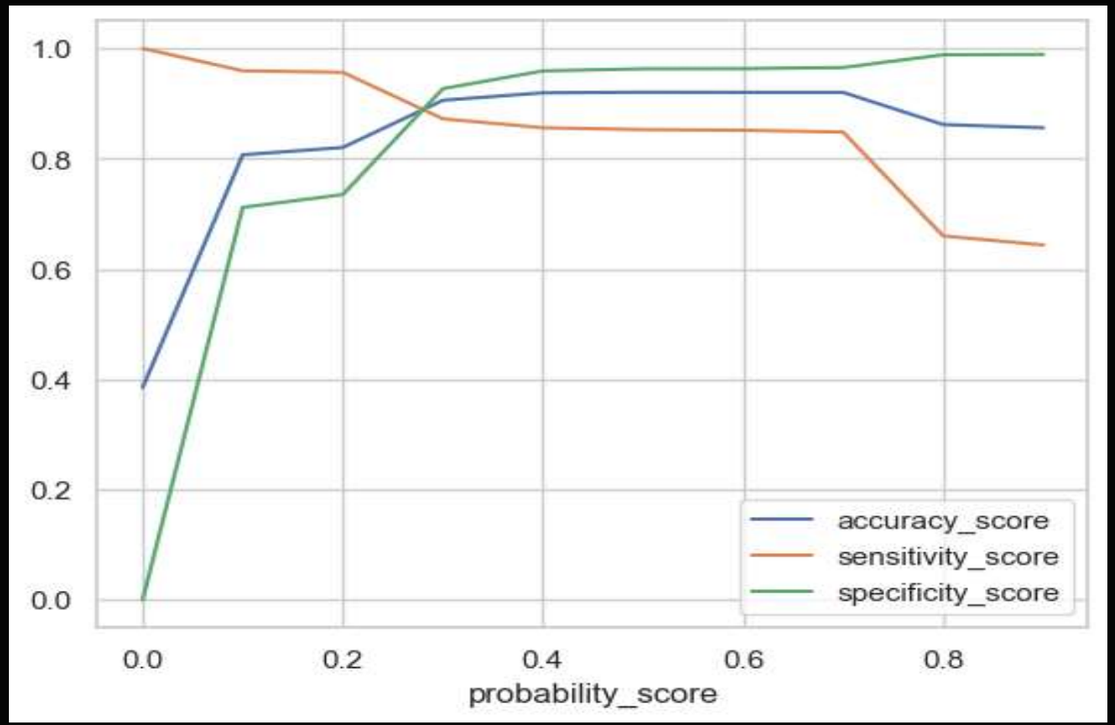
Strategy Applied

- Splitting the data into train test split ratio of 70:30
- Used RFE to choose top 15 variables
- Build model by removing variables with p-values

  >0.05 & VIF >5

- Overall accuracy of the model is 92%





Model Evaluation

Strategy Applied

- Calculated accuracy, sensitivity & specificity for various cutoffs from 0.1 to 0.9
- As per the graph & looking at the other scores, it is observed that optimal point is 0.27

	probability_score	accuracy_score	sensitivity_score	specificity_score	precision_score
0.0	0.0	0.385136	1.000000	0.000000	0.385136
0.1	0.1	0.807117	0.959526	0.711652	0.675785
0.2	0.2	0.820343	0.956664	0.734955	0.693333
0.3	0.3	0.905999	0.872445	0.927017	0.882183
0.4	0.4	0.919540	0.856092	0.959283	0.929427
0.5	0.5	0.920642	0.852821	0.963124	0.935426
0.6	0.6	0.920328	0.851594	0.963380	0.935759
0.7	0.7	0.920328	0.848324	0.965429	0.938914
0.8	0.8	0.861912	0.659853	0.988476	0.972875
0.9	0.9	0.856086	0.643500	0.989245	0.974010

Train Data- Confusion Matrix

Predicted Actual	Not Converted	Converted
Not Converted	2,987	918
Converted	124	2,322

Model Performance

Accuracy	83.59%
Sensitivity	94.9%
Specificity	76.49%
Precision	71.66%

Model Prediction

```
----------------------Feature Importance--------------------
const                                                -1.248649
Do Not Email                                         -1.180501
Lead Origin_Lead Add Form                             0.908052
Lead Source_Welingak Website                          3.218160
Last Activity_SMS Sent                                1.927033
Tags_Busy                                             3.649486
Tags_Closed by Horizzon                               8.555901
Tags_Lost to EINS                                     9.578632
Tags_Ringing                                         -1.771378
Tags_Will revert after reading the email              3.831727
Tags_switched off                                    -2.336683
Lead Quality_Not Sure                                -3.479228
Lead Quality_Worst                                   -3.943680
Last Notable Activity_Modified                       -1.682075
Last Notable Activity_Olark Chat Conversation        -1.304940
dtype: float64
```

Top Features

Test Data- Confusion Matrix

Predicted Actual	Not Converted	Converted
Not Converted	1,303	431
Converted	71	918

Model Performance

Accuracy	81.56%
Sensitivity	92.82%
Specificity	75.14%
Precision	68.05%

Conclusions

Prediction

- Logistic Regression model is used to predict the probability of conversion of a user
- While both sensitivity-specificity as well as precision-recall metrics has been utilized, however optimal cut-off is based on sensitivity-specificity for final prediction

Lead Scores

- For Train Data- 95%
- For Test Data- 92%

Top 3 Variables

- Tags_Lost to EINS
- Tags_Closed by Horizzon
- Tags_Will revert after reading the email