Lead Scoring Case Study

Submitted By: Harpreet Kaur, Kritika Joshi, Justy D varughese

Submission Date: 23rd July, 2024

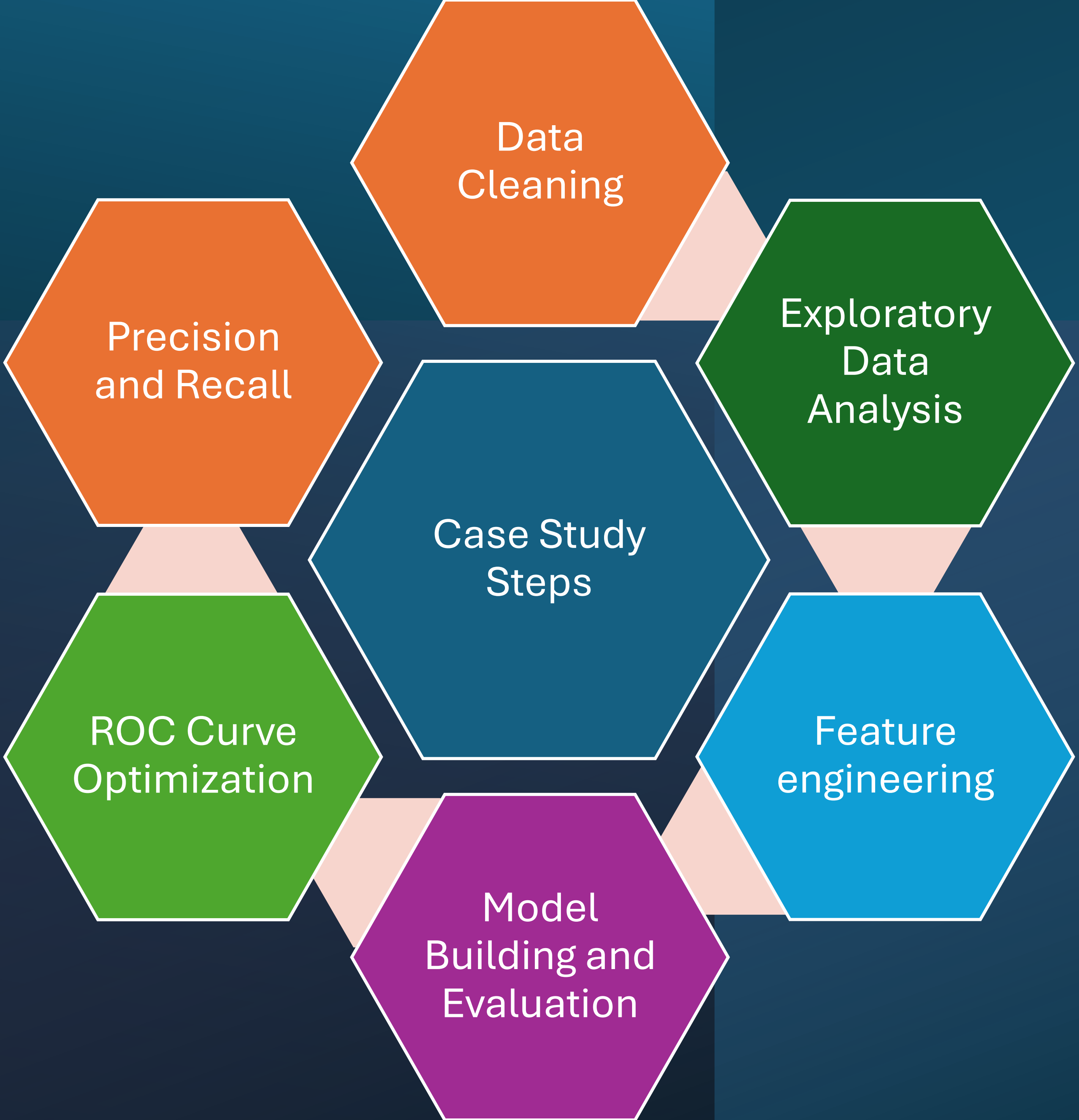
Problem Statement

X Schooling offers online courses to industry experts. Several websites and search engines, including Google, are used by the company to promote its courses. These individuals might browse the courses, fill out a form for the course, or watch some videos once they get to the website. A lead is a person who provides their email address or phone number when they fill out a form. Additionally, the company receives leads from previous referrals. Employees on the sales team begin following up on these leads with phone calls, emails, and other forms of communication. Some of the leads are converted through this process, but the majority are not. At X education, the typical lead conversion rate is around 30%.

Case Study Aim

X Education requires assistance selecting the most promising leads, also known as leads that have the greatest potential to become paying customers. A model that assigns a lead score to each lead so that customers with a higher lead score have a greater chance of conversion and customers with a lower lead score have a lower chance of conversion is needed by the business. In particular, the CEO has stated that the target lead conversion rate should be around 80%.

Case Study Steps Overview



Data Cleaning and Preparation Steps

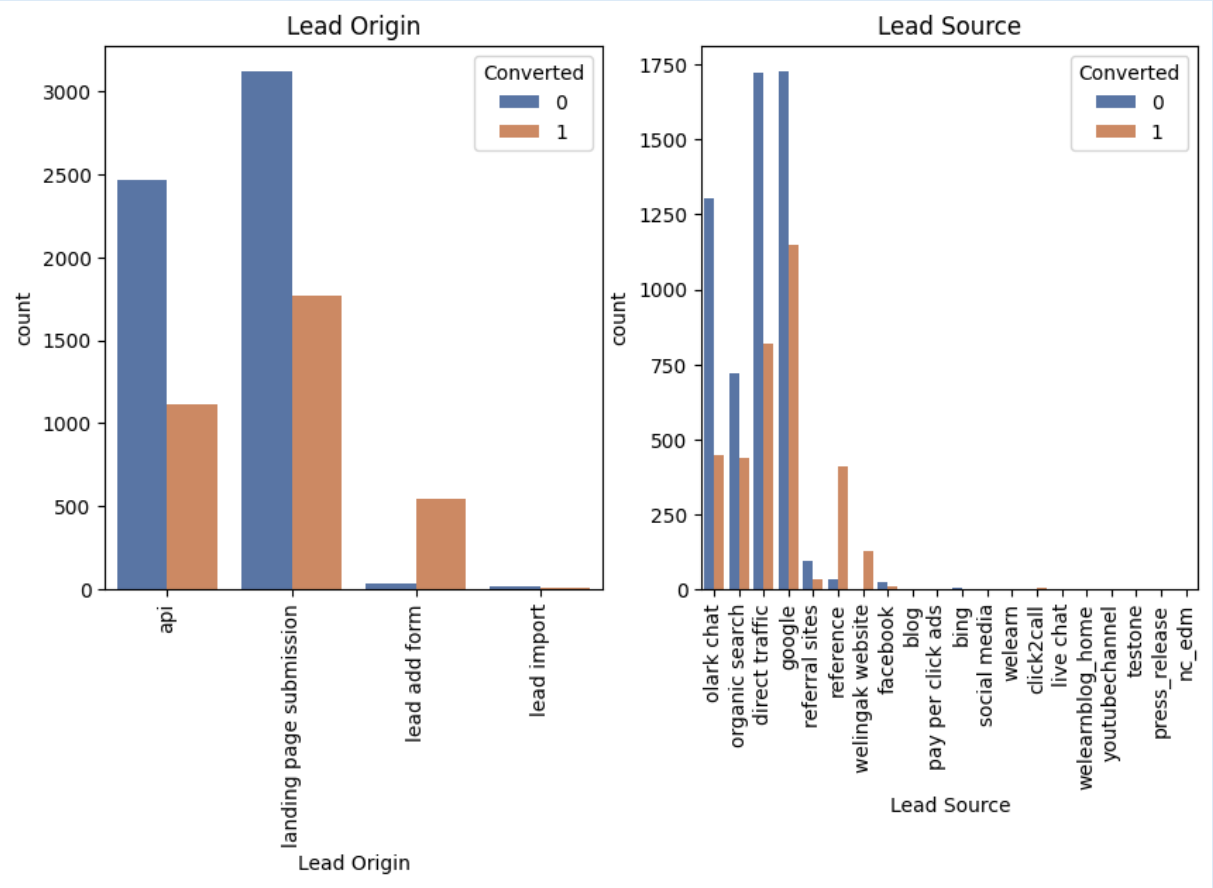
- **Addressing Select Category**: Instances of the “Select” category were replaced with null values due to their lack of meaningful information.
- **Filling Missing Values**: Since we had limited data in the dataset dropping the variables with high number of missing values did not seemed like a good approach. Therefore instead of getting rid of them filled missing values with “Unknown” values.
- **Removing Irrelevant Columns**: Columns like "Prospect ID" and "Lead Number" were removed as they did not impact model performance.

Exploratory Data Analysis

EDA was crucial in understanding the data and building a robust model:

- **Univariate Analysis**:
  - **For categorical variables:** Count plots were created to visualize distributions, revealing that most leads originated from landing page submissions.
  - **For numerical variables:** Histograms illustrated the distribution of numerical features.
- **Relation with Target Variable “Converted”**:
  - **For categorical columns:** Count plots helped assess how different categories related to lead conversion.
  - **For numerical columns:** Heatmaps were used to understand feature correlations with the target variable.

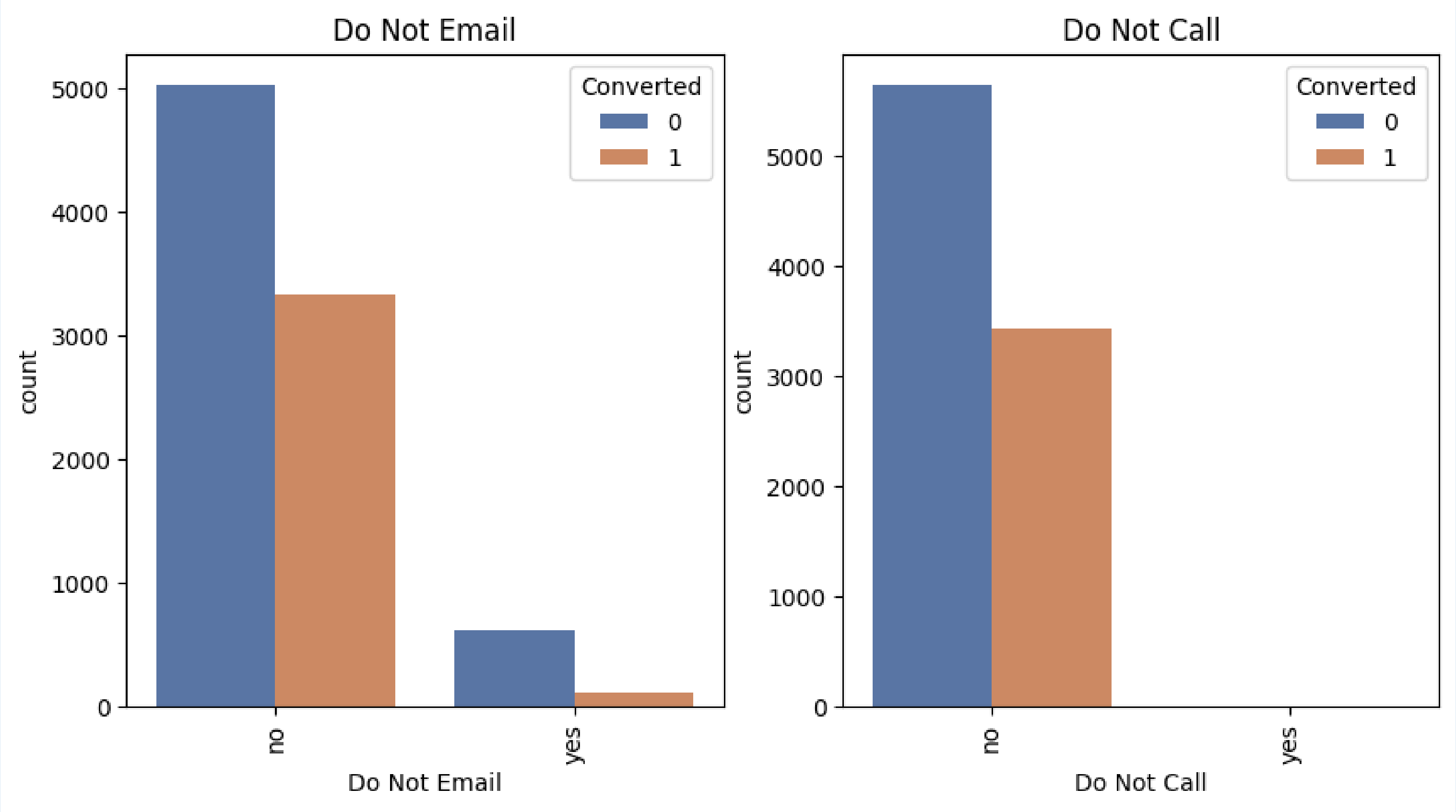
Relation of Categorical columns with Target variable "Converted"



1.There is a high non conversion rate when the lead was identified through Landing Page submission.

2. In terms of lead source, organic search, google and olark chart are leading categories.

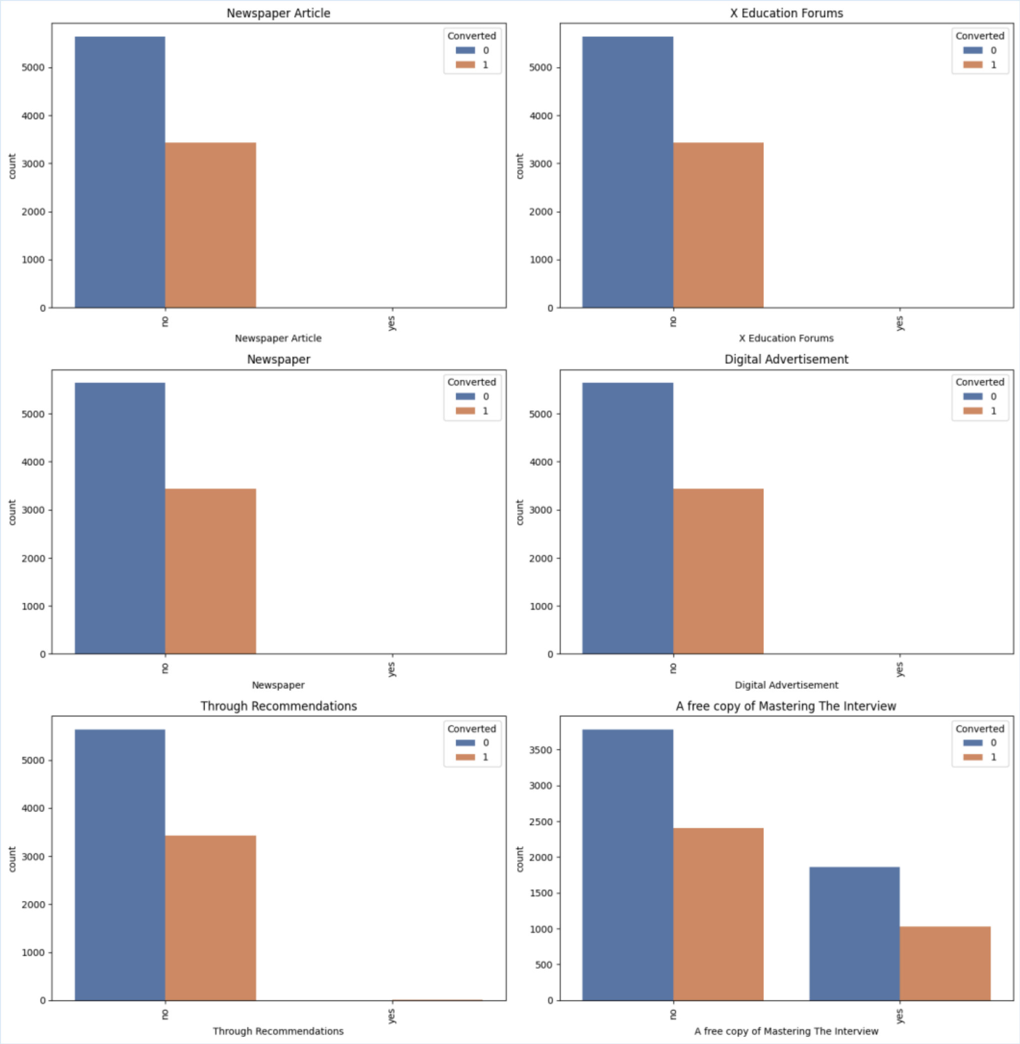
Contd.



This indicates people that leads that don't want to be contacted have very high chance of not being converted into Leads. This suggests lead is not interested in undertaking the course.

Contd.

Potential leads who have seen ads about the X Education from Newspaper, Newspaper article and through recommendations have high chance of getting converted into a potential lead.



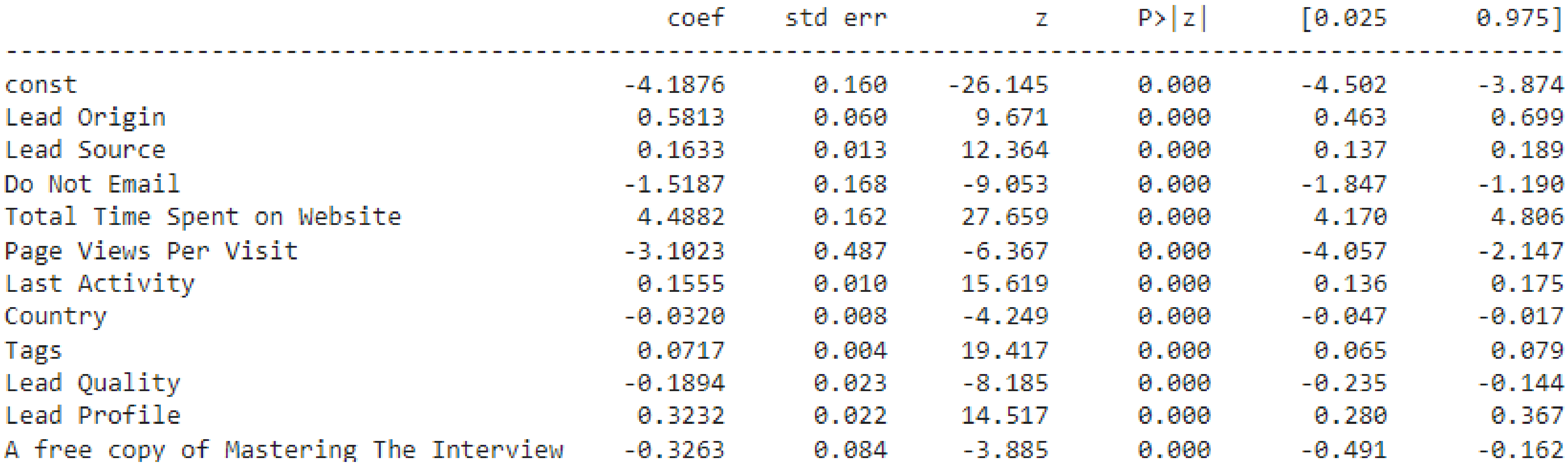
Feature Engineering

a. Label Encoding: Now in this case we were dealing with data that has a lot of categories so creating dummy variables for each column was not the approach we followed instead we went ahead and did Label Encoding to assign a unique integer value to each category.

b. Scaling of Numerical feature: Since features like Total time spent on website would have larger values than other numerical feature like Total Visits therefore to give each feature equal importance we have done the scaling.

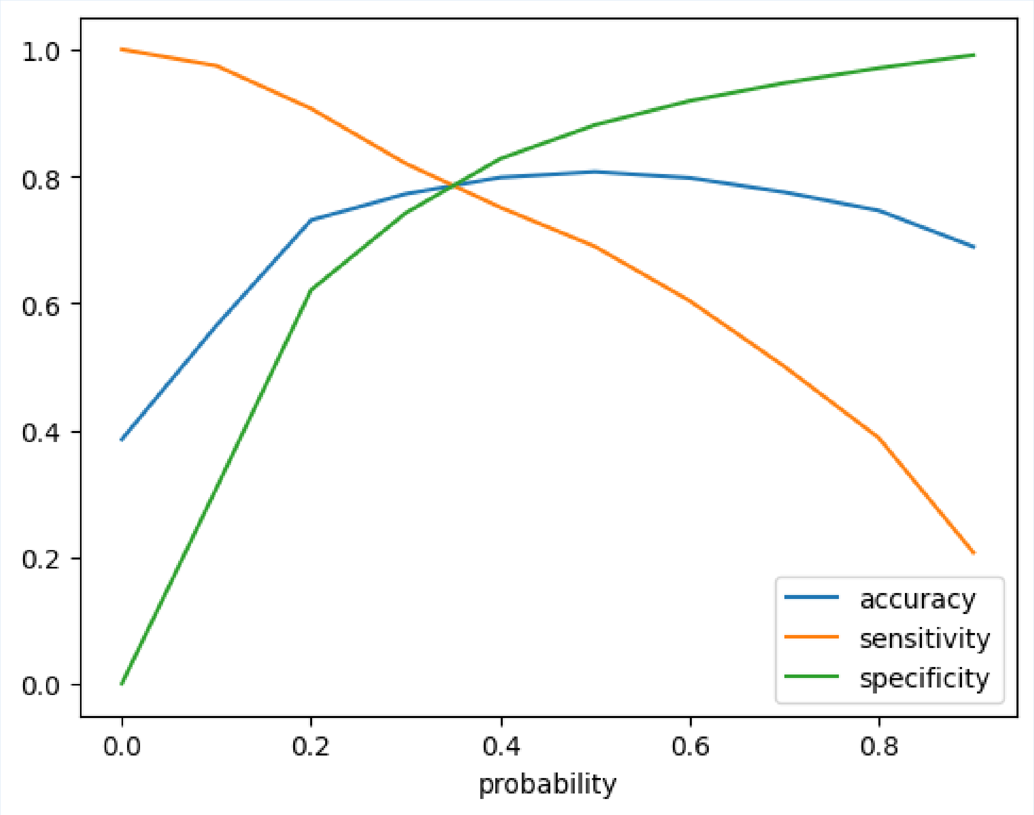
Final Model details

- For selecting features of the model we have used Recursive Feature Elimination from sklearn.
- Final model had the below variables along with their coefficients:



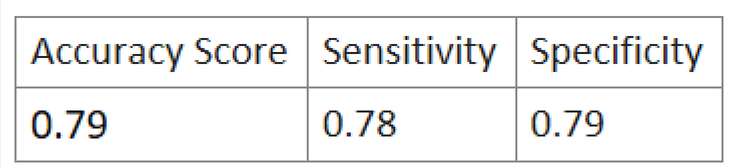
```
                                      coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------------------------
const                              -4.1876      0.160    -26.145      0.000      -4.502      -3.874
Lead Origin                         0.5813      0.060      9.671      0.000       0.463       0.699
Lead Source                         0.1633      0.013     12.364      0.000       0.137       0.189
Do Not Email                       -1.5187      0.168     -9.053      0.000      -1.847      -1.190
Total Time Spent on Website         4.4882      0.162     27.659      0.000       4.170       4.806
Page Views Per Visit               -3.1023      0.487     -6.367      0.000      -4.057      -2.147
Last Activity                       0.1555      0.010     15.619      0.000       0.136       0.175
Country                            -0.0320      0.008     -4.249      0.000      -0.047      -0.017
Tags                                0.0717      0.004     19.417      0.000       0.065       0.079
Lead Quality                       -0.1894      0.023     -8.185      0.000      -0.235      -0.144
Lead Profile                        0.3232      0.022     14.517      0.000       0.280       0.367
A free copy of Mastering The Interview   -0.3263      0.084     -3.885      0.000      -0.491      -0.162
```

Model Evaluation - Sensitivity and Specificity using ROC Curve



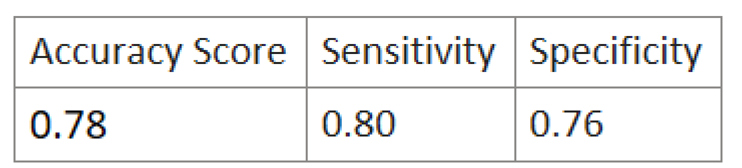
With cut off as 0.35 we have accuracy, sensitivity and specificity between 79-80% percent.

Training Data set metrics



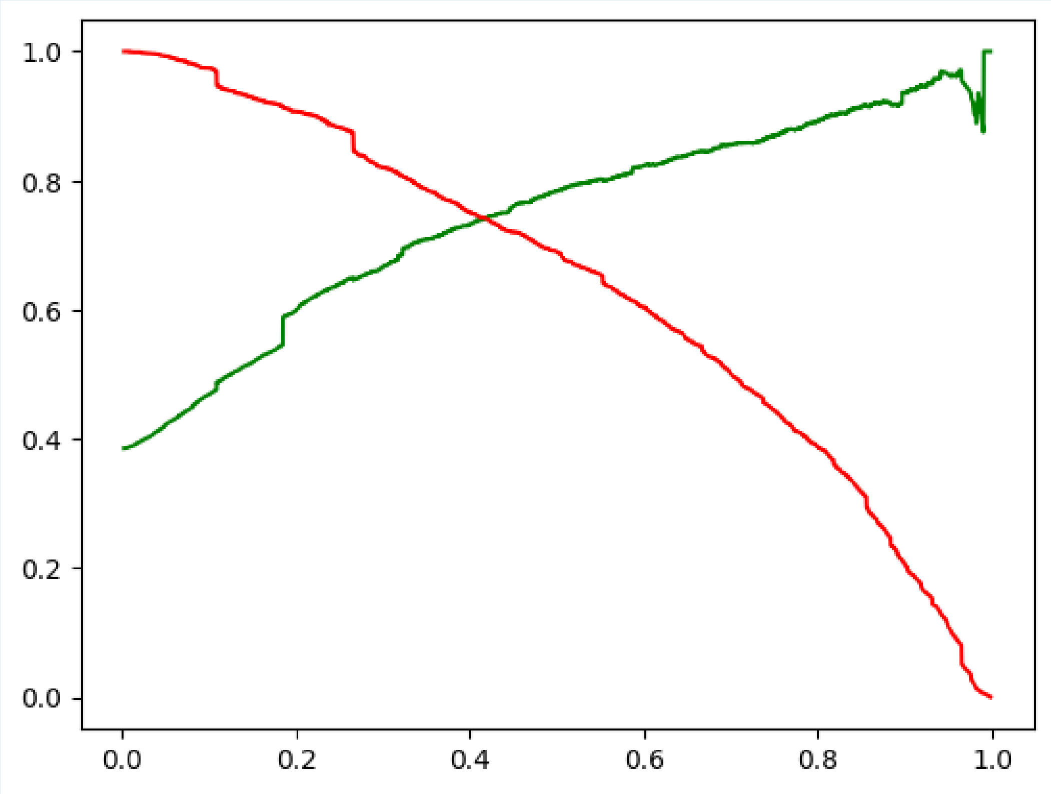
Accuracy Score	Sensitivity	Specificity
0.79	0.78	0.79

Testing Data set metrics



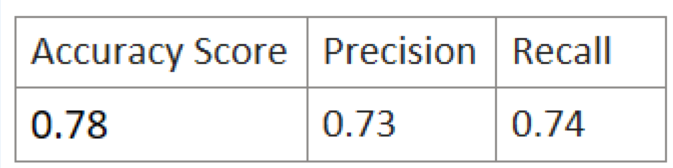
Accuracy Score	Sensitivity	Specificity
0.78	0.80	0.76

Model Evaluation – Precision and Recall



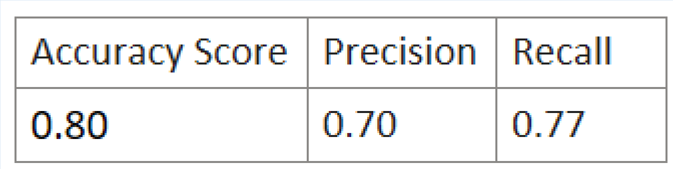
Training Data set metrics

Training Data set metrics



Accuracy Score	Precision	Recall
0.78	0.73	0.74

Testing Data set metrics



Accuracy Score	Precision	Recall
0.80	0.70	0.77

Summary

It was discovered that the most influential factors for potential buyers, in descending order of importance, are:

1. The total time spent on the website.
2. Page view per visit
3. The lead source, specifically when it is:

- ➢ Google
- ➢ Direct Traffic
- ➢ Welingak website
- ➢ Organic search

4. The last activity, particularly when it is:

- ➢ SMS
- ➢ Olark chat conversation

5. The lead origin, especially when it is a landing page submission.
6. Lead quality and profile also play an important role considering High relevance leads are potential leads and will be converted to hot leads.

Thank you