LEAD SCORING CASE STUDY

BY:

MANOJ KUMAR DARA

MEENAKSHY M

TABLE OF CONTENTS

- Problem Statement & Goal of the Study
- Analysis Approach
- Data Cleaning
- EDA
- Model Building (RFE & Manual fine tuning)
- Model Evaluation
- Conclusion of the Analysis

PROBLEM STATEMENT & GOAL OF THE STUDY

Problem Statement:

- X Education gets a lot of leads, its lead conversion rate is very poor at around 30%
- X Education wants to make lead conversion process more efficient by identifying the most potential leads, also known as Hot Leads
- Their sales team want to know these potential set of leads, which they will be focusing more on communicating rather than making calls to everyone.

Goal of the Study:

- To help X Education select the most promising leads, i.e., the leads that are most likely to convert into paying customers.
- The company requires us to build a model wherein we need to assign a lead score to each of the leads. such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.
- The CEO has given a ballpark of the target lead conversion rate to be around 80%.

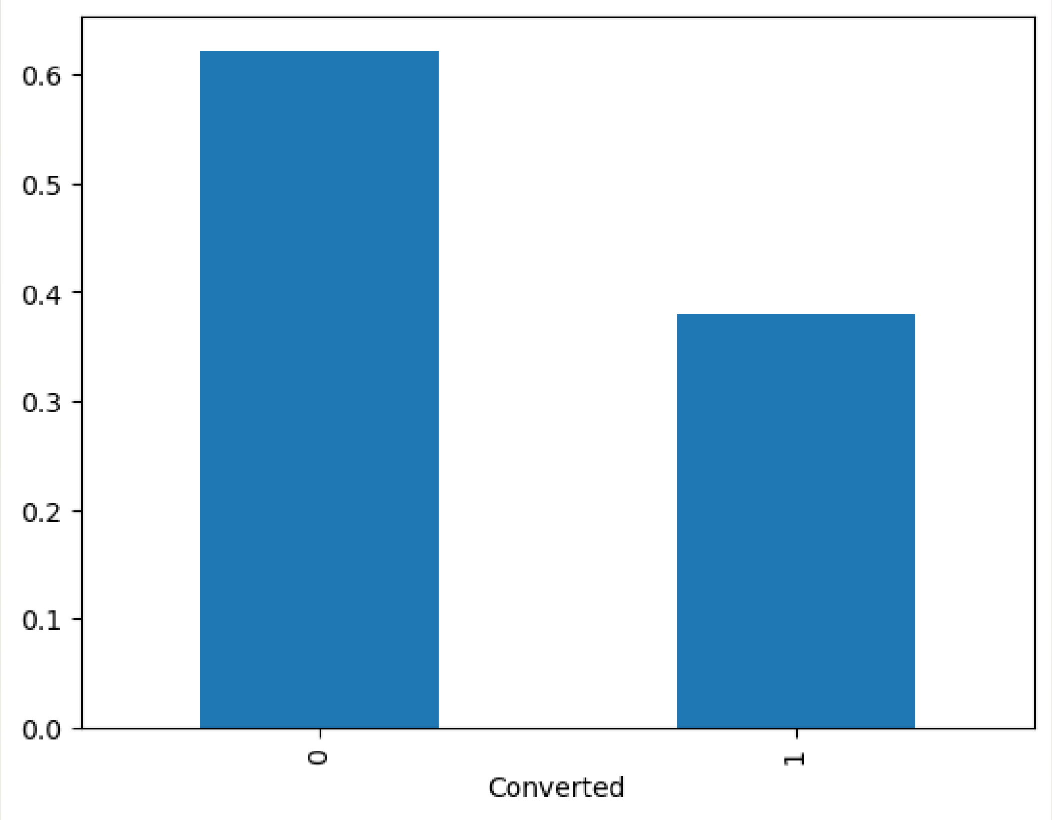
ANALYSIS APPROACH

1. Data cleaning and Imputing missing values
2. Exploratory Data Analysis - Univariate Analysis
3. Feature Scaling and Dummy Variable Creation
4. Logistic Regression Model Building
5. Model Evaluation - Sensitivity, Specificity, Precision, Recall

DATA CLEANING

- **"Select"** represents null values for some categorical variables, as customers did not choose any option from the list which has been changed to “**unknown**”.
- Columns with over 40% null values were dropped.
- Drop columns that did not add any insight or value to the study objective (Example: City, Country).
- Imputation was used for some columns.(Example: Lead Quality)
- Columns with no use for modeling (Prospect ID, Lead Number) or only one category of response were dropped.

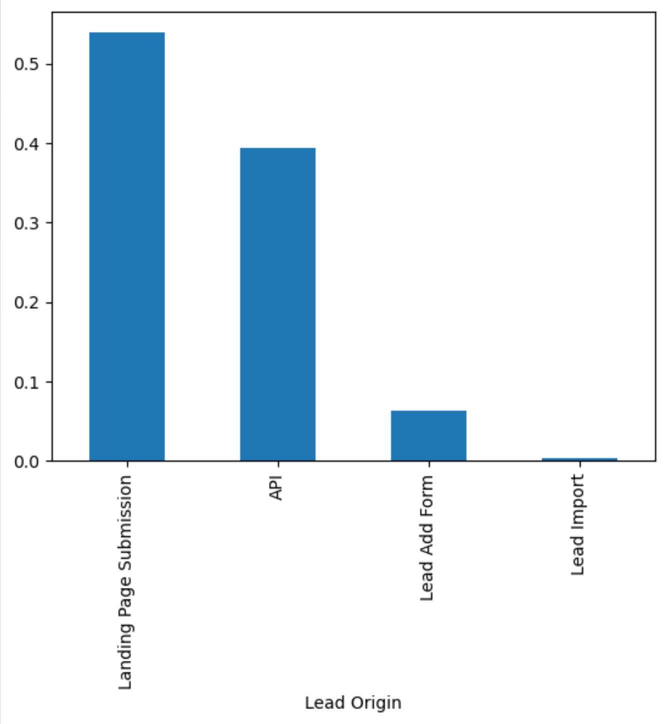
EDA



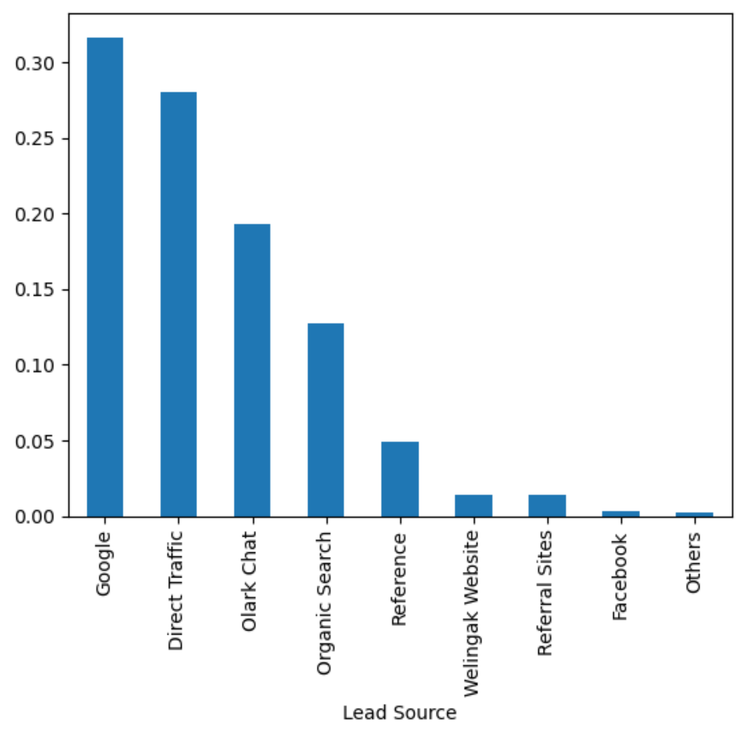
On checking the Datapoints we found that there is a data imbalance on checking the target variable. Below are the observations from the target variable:

- Conversion rate is of 38%, meaning only 38% of the people have converted to leads.(Minority)
- While 62% of the people didn't convert to leads.(Majority)

UNIVARIATE ANALYSIS

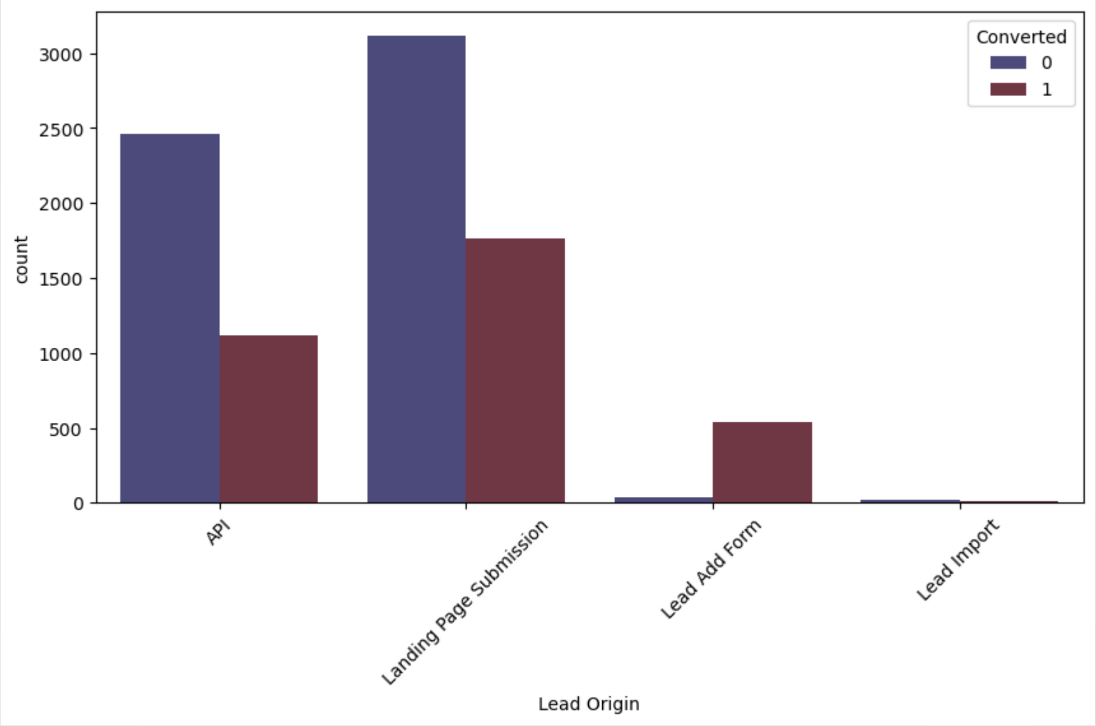


Lead Origin: "Landing Page Submission" identified 53% of customers, "API" identified 39%.



Lead Source: 58% Lead source is from Google & Direct Traffic combined.

BIVARIATE ANALYSIS

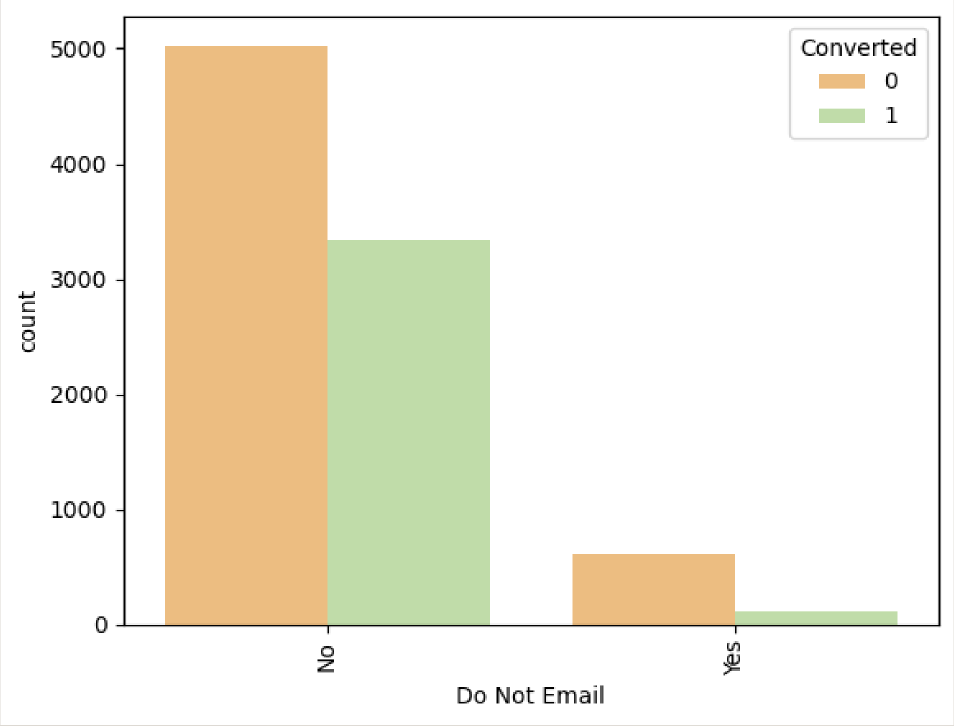


Lead Origin:

● Around 52% of all leads originated from *"Landing Page Submission"* with a **lead conversion rate (LCR) of 36%**.

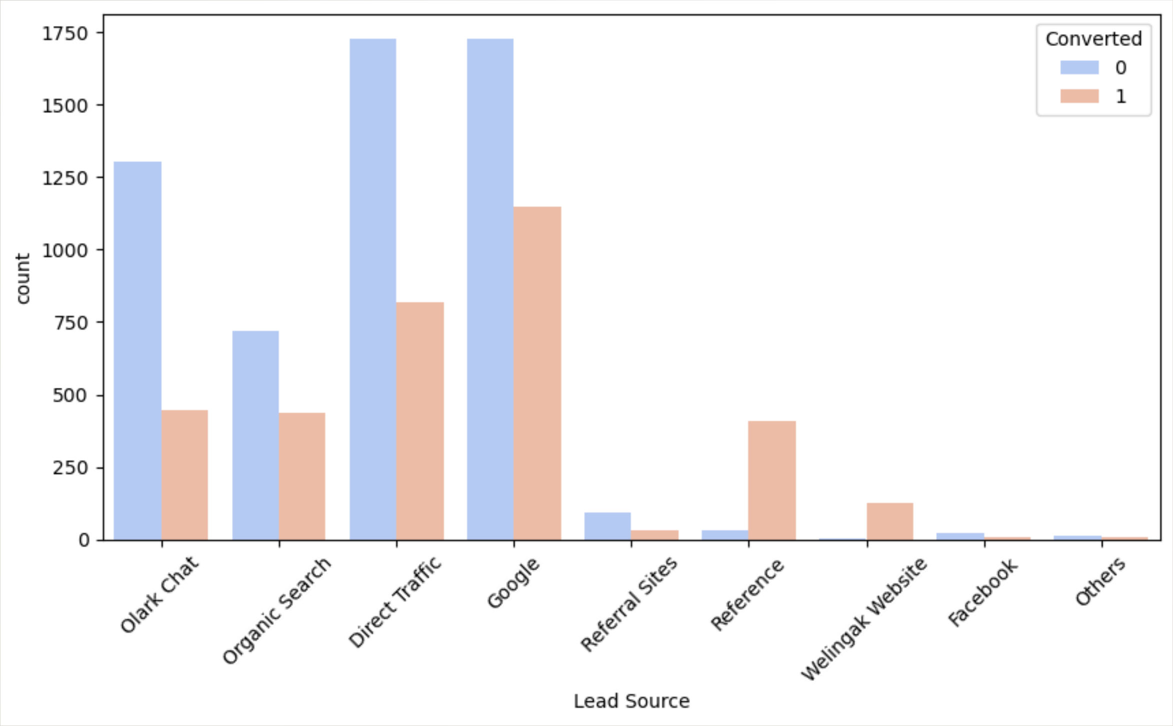
● The *"API"* identified approximately 39% of customers with a **lead conversion rate (LCR) of 31%**.

BIVARIATE ANALYSIS



Do Not Email:
92% of the people have opted that they don't want to be emailed about the course & 40% of them are converted to leads.

BIVARIATE ANALYSIS



Lead Source:

● *Google* has **LCR of 39%** out of 31% customers.

● *Direct Traffic* contributes **32% LCR** with 27% customers, which is lower than Google.

● *Organic Search* also gives **37.8% of LCR**, but the contribution is by only 12.5% of customers,.

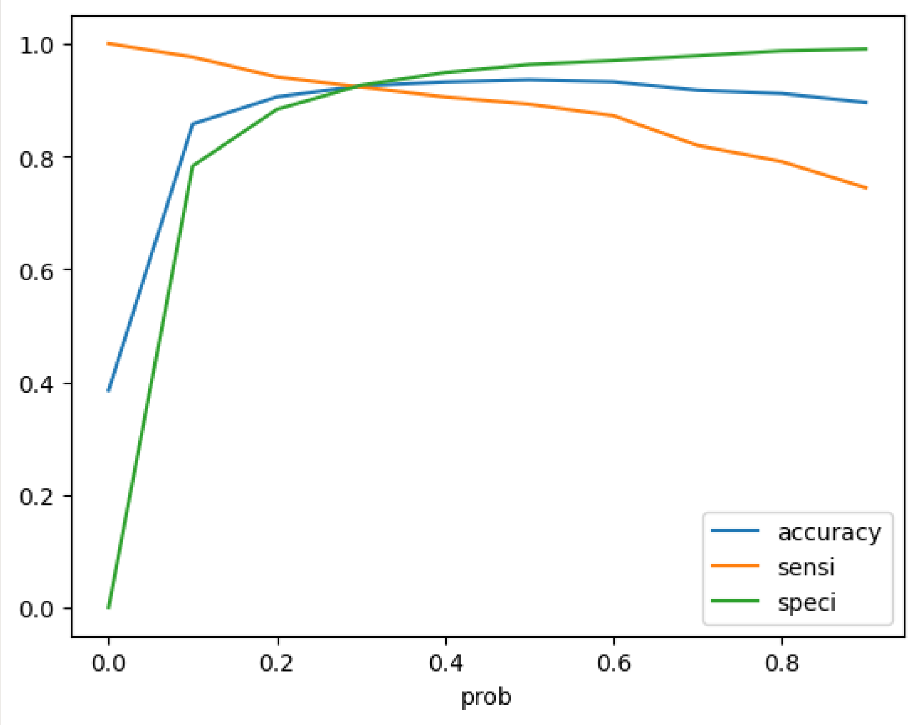
● *Reference* has **LCR of 92%**, but there are only around 6% of customers through this Lead Source.

MODEL BUILDING

Feature Selection

- The data set has lots of dimension and large number of features. Starting with 37 different columns.
- This will reduce model performance and might take high computation time.
- Hence it is important to perform **Recursive Feature Elimination** (RFE) and to select only the important columns.
- RFE Outcome
  - Pre RFE – 108 columns & Post RFE – 20 columns
- We can manually fine tune the models by dropping variables with P-value greater than 0.05 and VIF value less than 5.
- Hence **logm6** will be our final model, and we will use it for Model Evaluation which further will be used to make predictions.

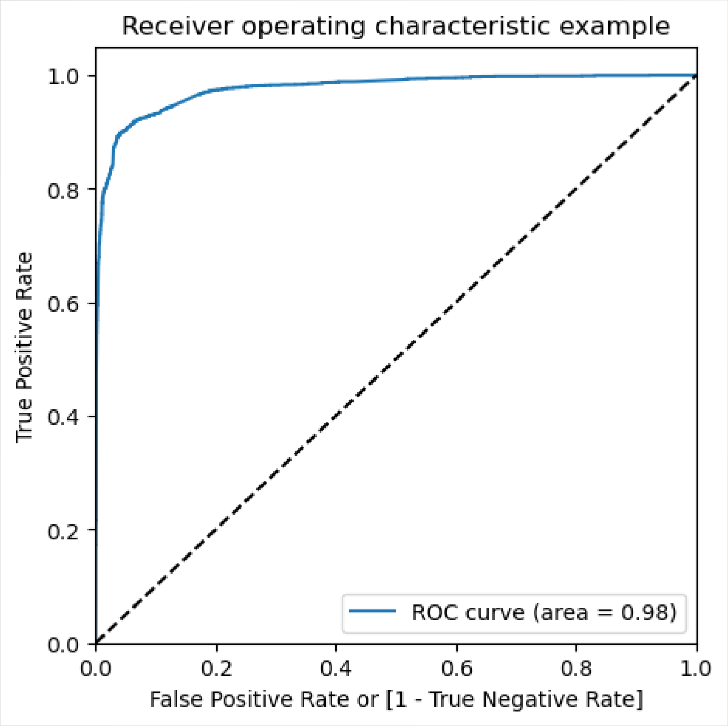
MODEL EVALUATION



Cutoff:

It was decided to go ahead with 0.30 as cutoff after checking evaluation metrics coming from both plots

MODEL EVALUATION



ROC Curve – Train Data Set:

Area under ROC curve is 0.98 out of 1 which indicates a good predictive model.

The curve is as close to the top left corner of the plot, which represents a model that has a high true positive rate and a low false positive rate at all threshold values.

MODEL EVALUATION

- Using a cut-off value of 0.30, the model achieved a **sensitivity 92% in the train set** and **90% in the test set.**
- Sensitivity in this case indicates how many leads the model identify correctly out of all potential leads which are **sensitivity of around 92%.**
- **Accuracy of 92%** , which is in line with the study's objectives.

Train Data Set:

- Accuracy: 92%
- Sensitivity: 92%
- Specificity : 93%

Test Data Set:

- Accuracy: 91%
- Sensitivity: 90%
- Specificity: 92%

CONCLUSION OF THE ANALYSIS

- The company should make calls which are closed by Horizzon CRM Tools first.
- The Company should make calls to the leads coming from "Tags_Will revert after reading the email".
- The company should make calls to the leads coming from the lead sources "Welingak Websites" and "Reference" as these are more likely to get converted.
- The company should make calls to the leads who spent "more time on the websites" as these are more likely to get converted.
- The company should make calls to the leads coming from the lead sources "Olark Chat" as these are more likely to get converted.
- The company should not make calls to the leads whose lead origin is "Landing Page Submission" as they are not likely to get converted.
- The company should not make calls to the leads “Last Notable Activity” is less.
- The company should not make calls to the leads whose lead “Quality worst”.