# Hot Leads Model

Authors: Uddeshya Kumar

# Problem StateMent

X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you 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, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

## Goals of the Case Study

There are quite a few goals for this case study:

- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.
- There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

# OverAll Approach

1. Import and Inspect DataSet
2. Data Preparations ( Encoding Categorical variables &, Handling of Null Values )
3. EDA ( Univariate Analysis , Outlier Detection , Checking Data Imbalance )
4. Dummy Variables Creation
5. Test-Train Split
6. Feature Scaling
7. Coorelations check
8. Model Building ( feature selection using (RFE, P-value), Improving model further according to value adjusted r2, VIF )
9. Model Evaluation with different metrics Sensitivity &, Specificity

## Data Cleaning and Preparation

- Read Data
- Convert Data into clean format
- Null Values handling
- Outlier Treatment
- Exploratory Data Analysis

## Splitting Data And Feature Scaling

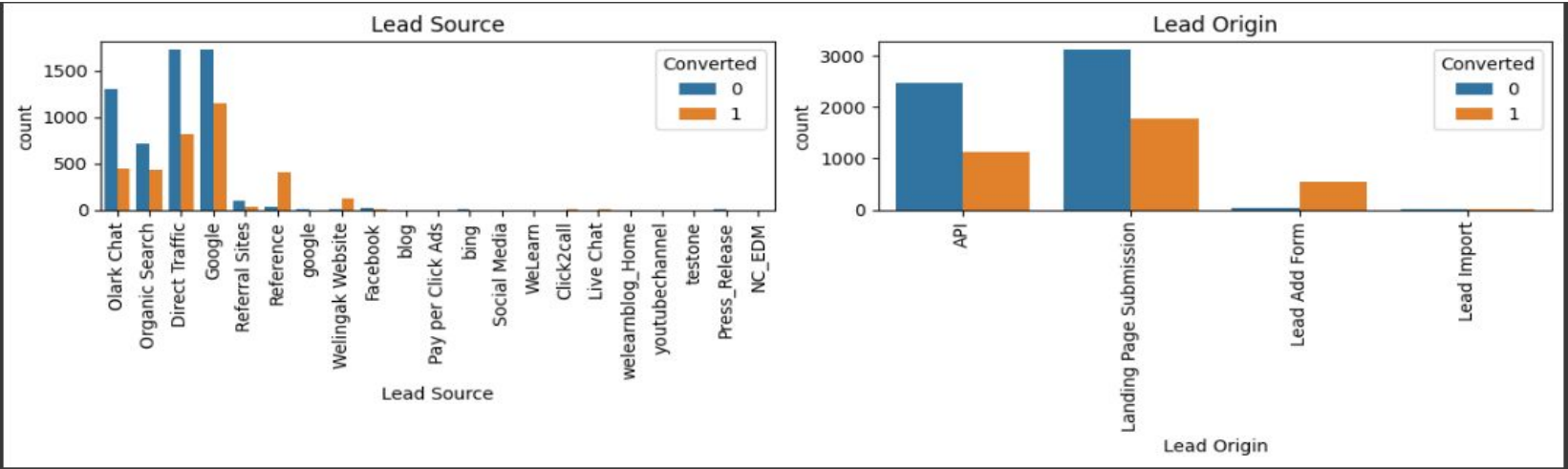
- Splitting Data into train and test dataset.
- Feature Scaling of Numerical variables
- Dummy variables

## Model Building

- Feature Selection using RFE, VIF and p-value
- Determine optimal Model using logistic Regression
- Calculate various evaluation metrics.

## Result

- Determine Lead Score and check if target final prediction is greater than 80 % conversion rate.
- Evaluate the final prediction on test Set.
- Calculate the precision and sensitivity



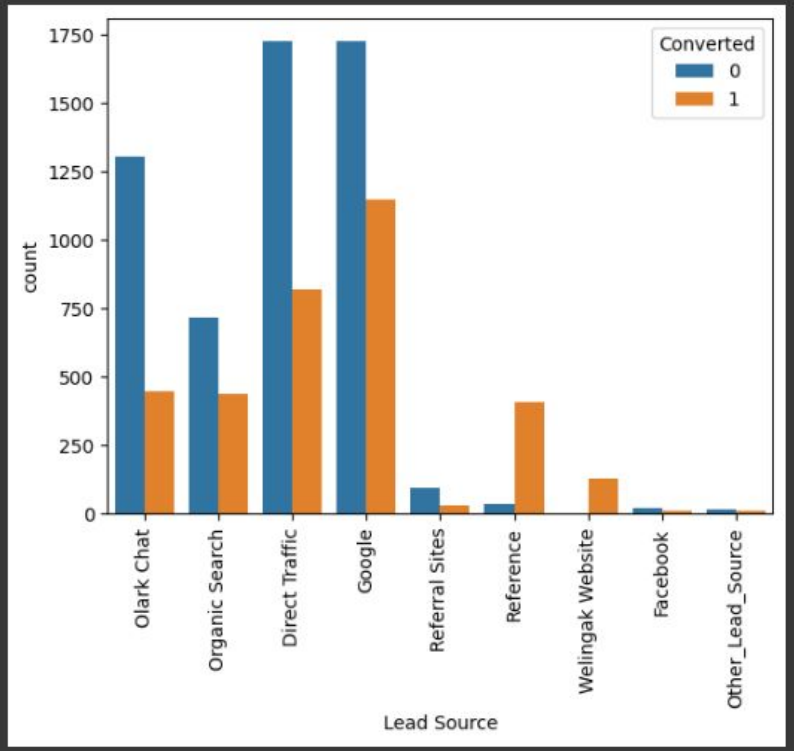

- Observations

API and Landing Page submission has less conversion rate but counts of leads are More.

The count of leads from Add Form is pretty low but conversion rate is High.

Lead import has very less count as well as conversion rate can be ignored

To improve overall Lead conversion rate, we need to focus on inc. the conversion rate of 'API' &, 'Landing Page Submission' –> inc the no of leads in Lead Add Form



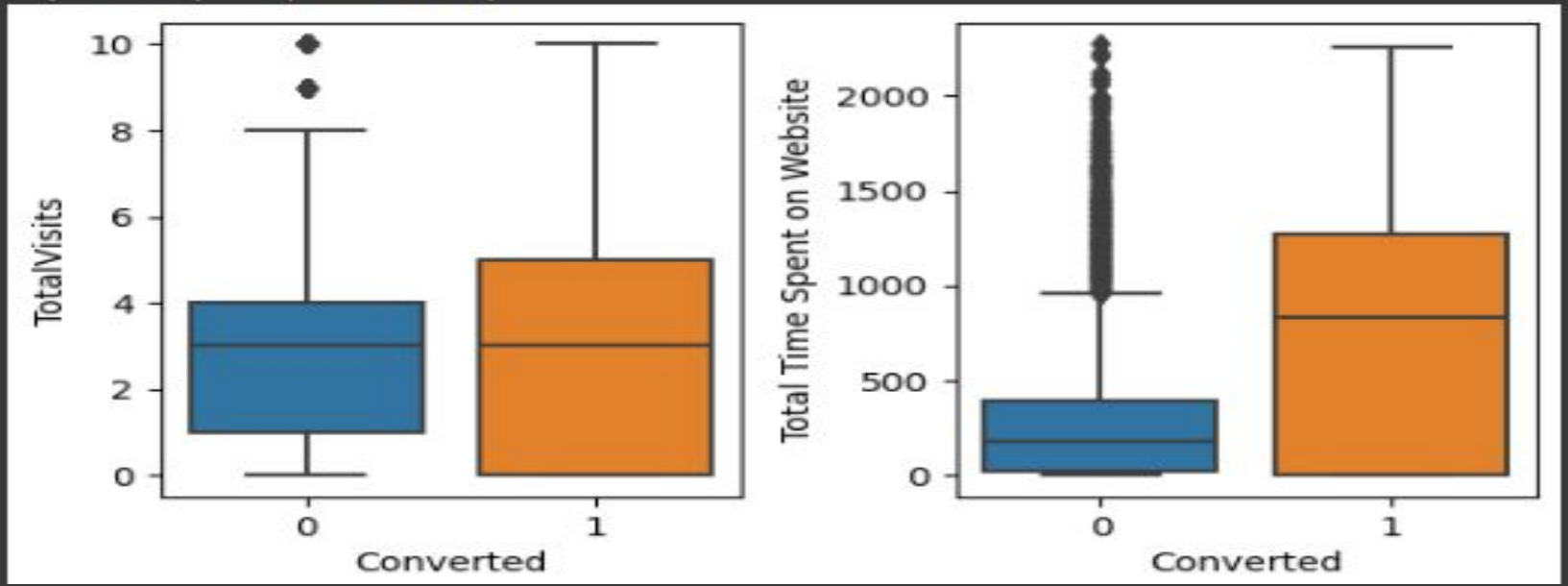

## OBSERVATION

Count of leads from Google and Direct Traffic is max.

The Conversion rate of the leads from Reference and Welingak website is maximum

To Improve overall lead conversion rate, Focus is needed on increasing the conversion rate of 'Google','Olark Chat'.'Organic Search','Direct traffic' &, also increasing the number of leads from 'Reference' &, 'Welingak Website'

```
plt.subplot(2, 2, 1+1)
```



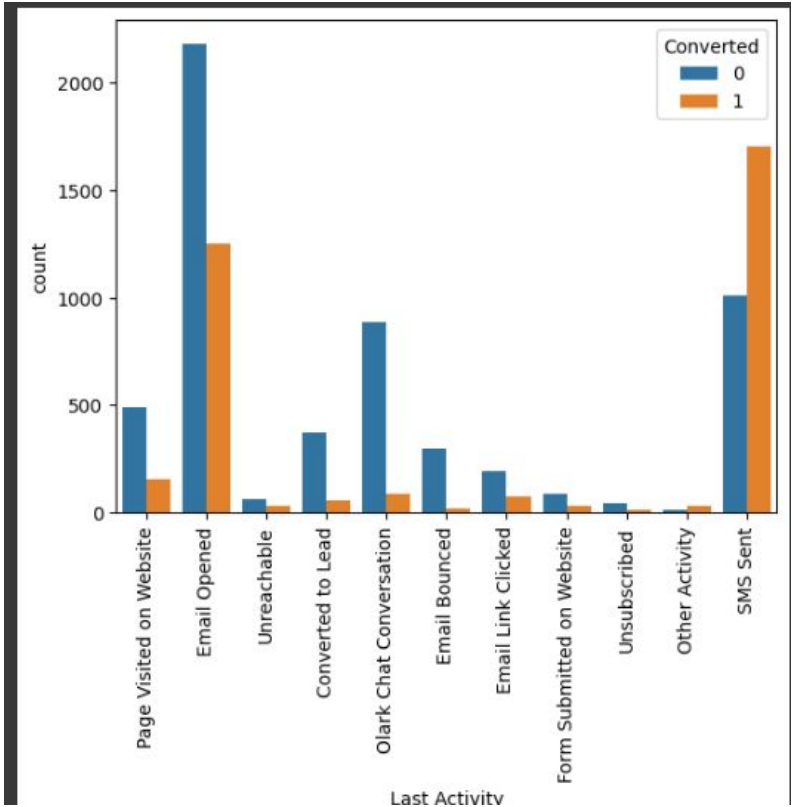

- Observation:

  The median of both the conversion and non-conversion are same &, hence nothing can be concluded .

  User is spending time on website are more likely to be converted

  Recomendation

  –> Websites can be more appealing so as to increase the time of the users on website



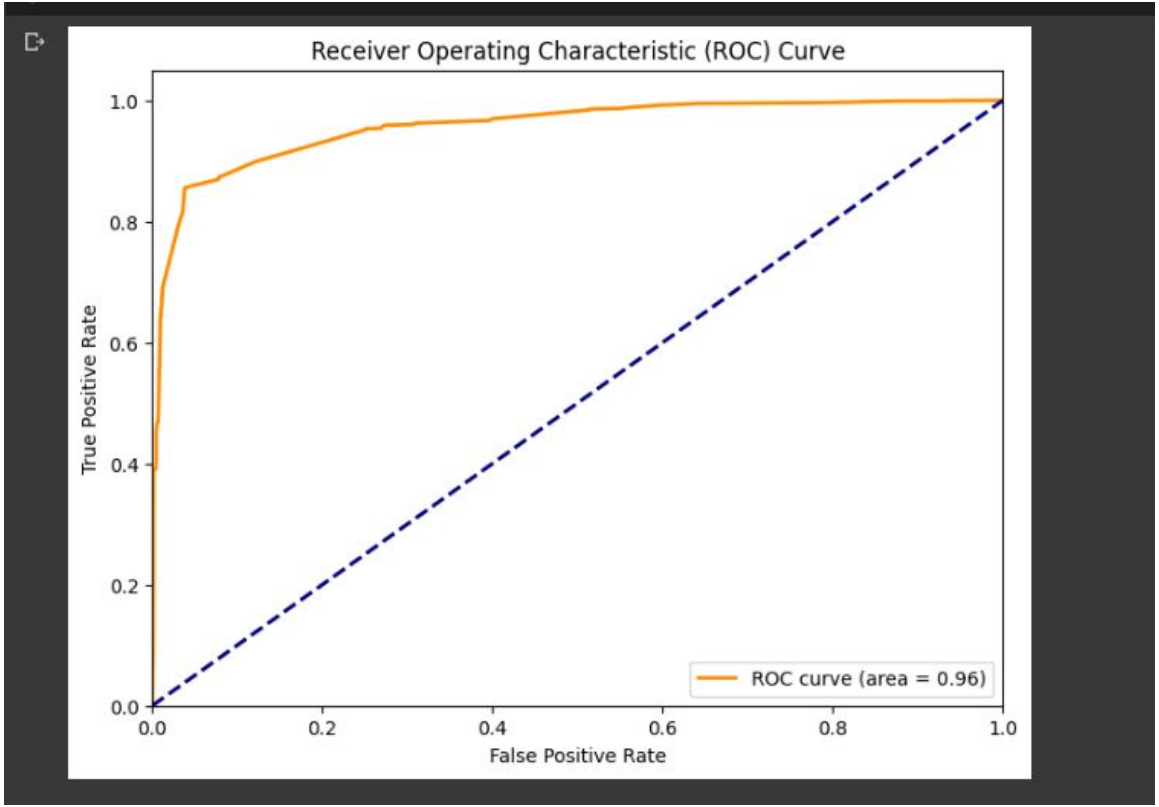

▾ Observation

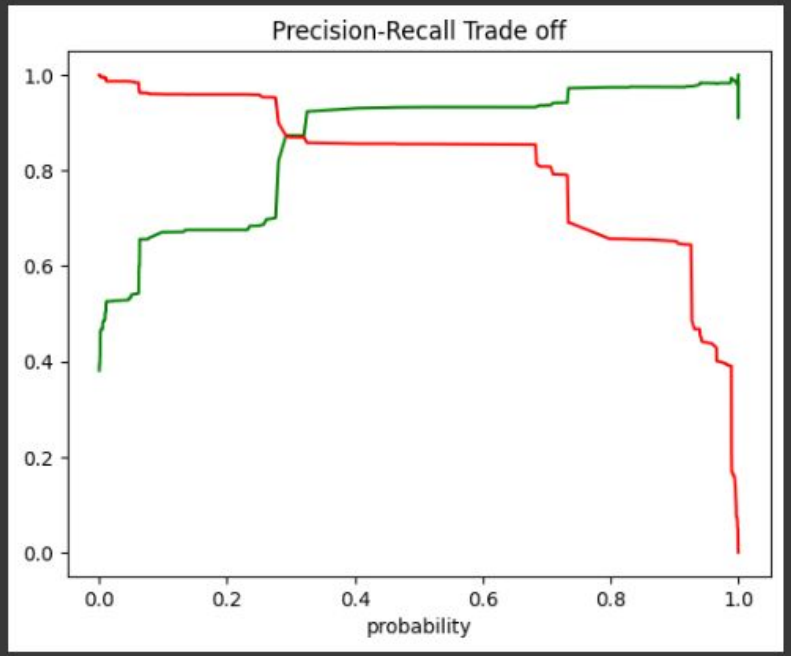
The Count of "Email Opened" is maxm.

The Conversion rate of SMS sent as last activity is maxm

Recommendation

Need to Focus on increasing the conversion rate of those having last activity as Email opened by making call to those leads ans also try to increase the count if the ines having last activity as SMS sent


Receiver Operating Characteristic (ROC) Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.96)




Observations

In Sensitivity-Specificity-Accuracy plot 0.27 probability looks optimal. In Precision-Recall Curve 0.3 looks optimal.

We are taking 0.27 is the optimum point as a cutoff probability and assigning Lead Score in training data.

# Recommendation

- Need to Focus on increasing the conversion rate of those having last activity as Email opened by making call to those leads and also try to increase the count if the having last activity as SMS sent.

- Websites can be more appealing so as to increase the times on website