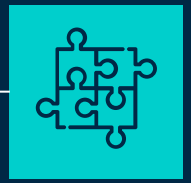
# LEAD SCORING CASE STUDY

-KSHAMA SHETYE (DSC31)
-RAUNAK BASU (DSC31)

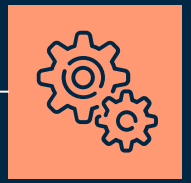
# TABLE OF CONTENTS



01

## INTRODUCTION

Problem Statement and Goals of the Analysis



02

## ANALYSIS AND INFERENCES

Data understanding and EDA

03

## LOGISTIC REGRESSION MODEL

Model building and evaluation

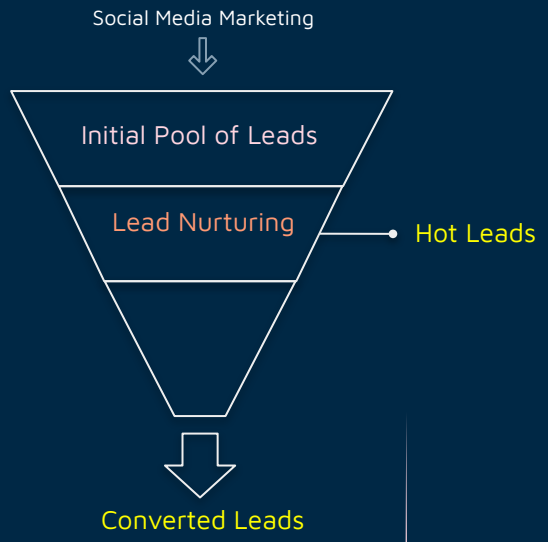
# INTRODUCTION - PROBLEM STATEMENT & GOALS

- X Education markets the courses on several websites and search engines which direct potential leads to their website
- When people fill up a form providing their information, they are classified to be a **lead**. Moreover, the company also gets leads through past referrals

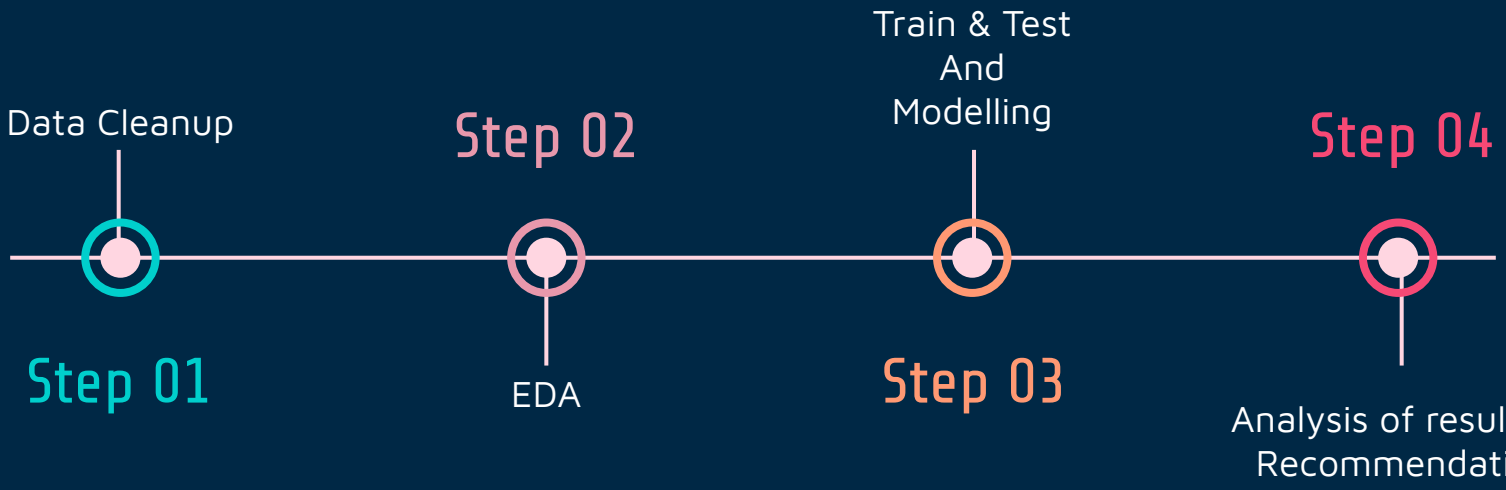
- Once these leads are acquired, sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not.

- Main **goal** is to identify the potential leads (Hot Leads) by building a model and assigning a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance, ultimately improving upon the target lead conversion rate.

- The CEO has given a ballpark of target lead conversion rate to be around 80%.

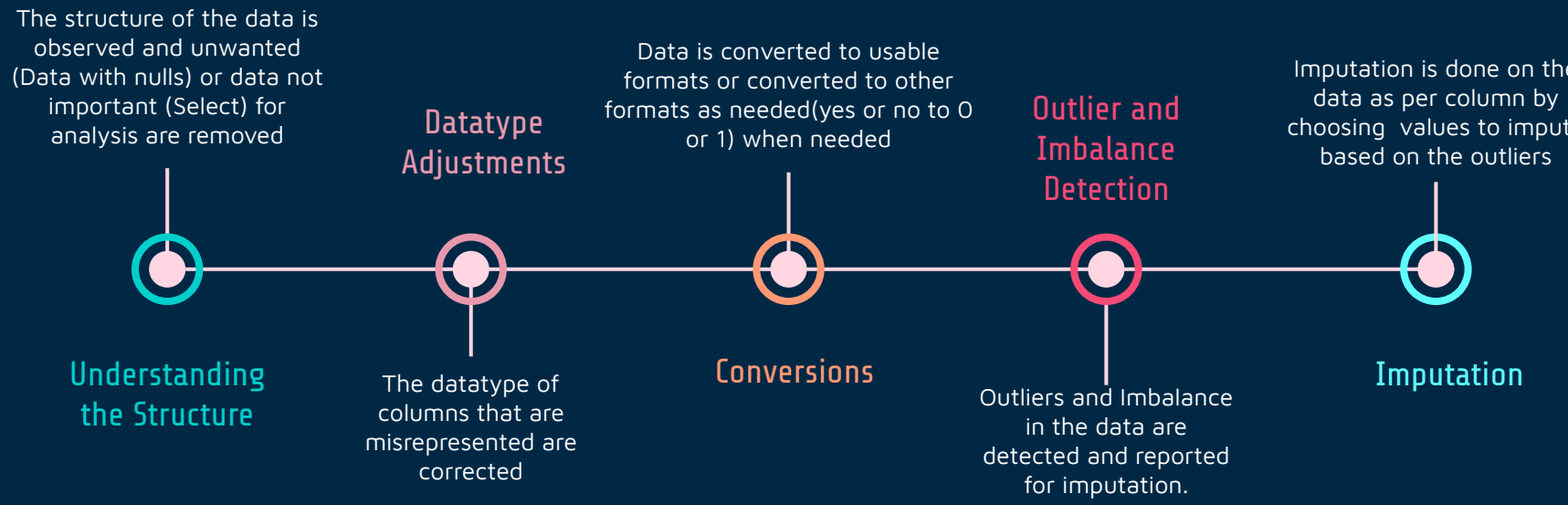


# MODELLING METHODOLOGY



# DATA CLEANING

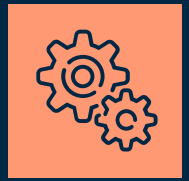
- Before any data is used, it needs to be cleaned and treated to get the best possible views for analysis.
- Following steps were performed on the data set.



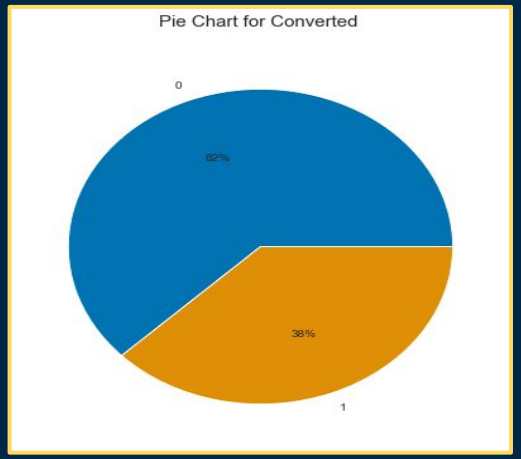

**Plotting and analysis is done after cleaning the data.**

Suitable scale and further imputation is chosen for very specific cases where removing the data may cause loss of valuable important information

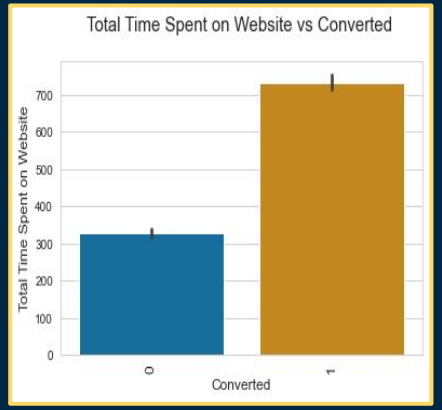
# ANALYSIS AND INFERENCES



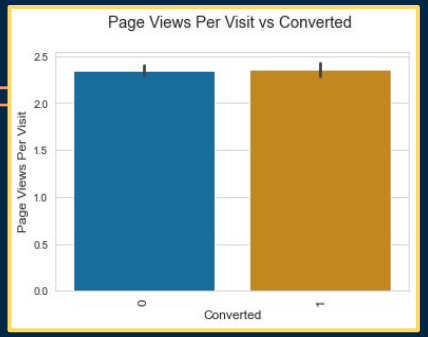
# EXPLORATORY DATA ANALYSIS (EDA)
## Numeric Variables



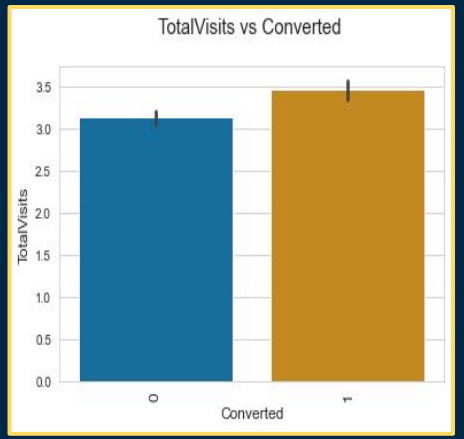

- The Conversion rate is 38%.




- People who spent a lot of time on the website had a much higher number of conversions.

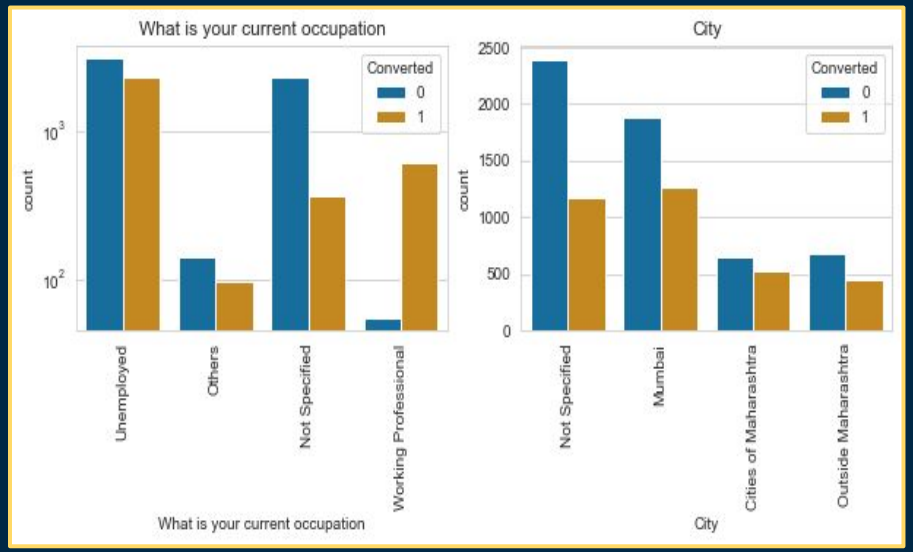


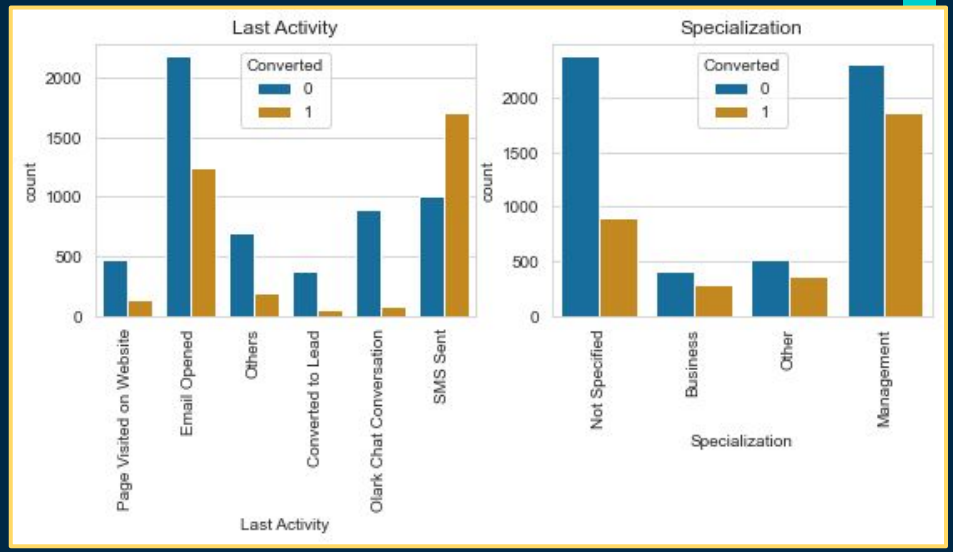

- Page views per visit doesn't seem to have much effect on conversion.




- Those who visited often have a slightly higher chance of converting.

# EXPLORATORY DATA ANALYSIS (EDA)
## Categorical Variables

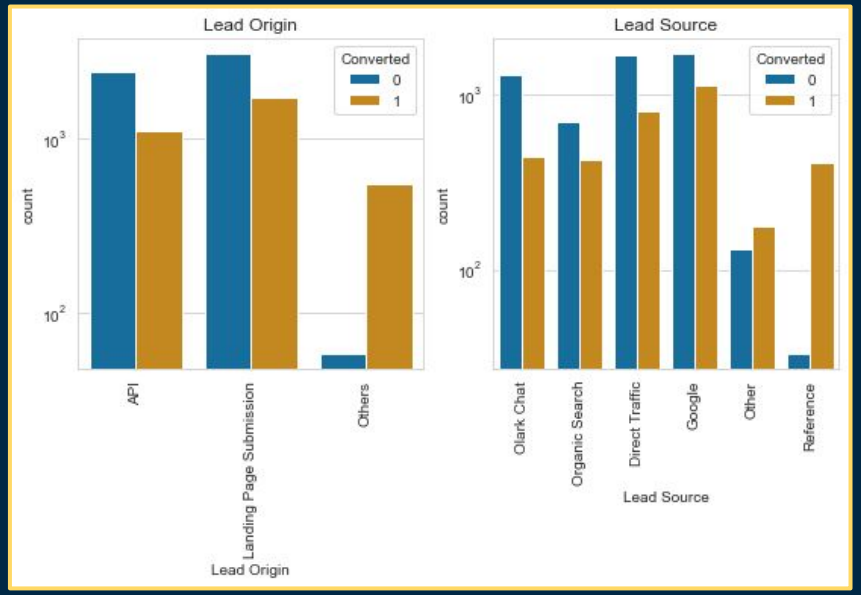


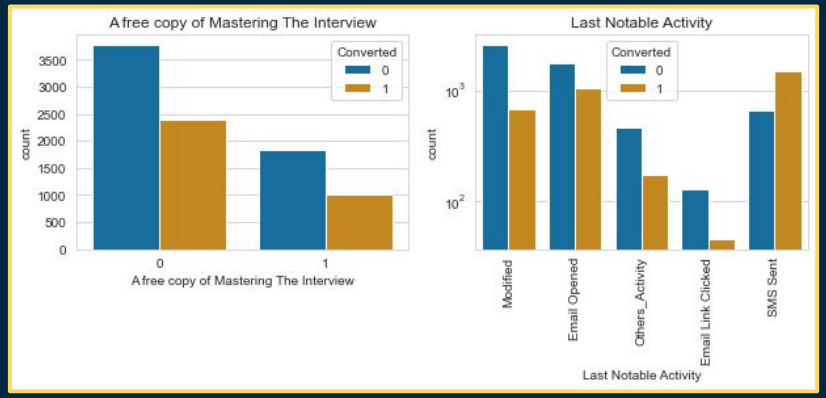




- Working professionals have the highest conversion rates , Unemployed have the highest number of leads.
- A large portion of leads do-not specify there location, Mumbai has the most leads from a single city.

- Those who sent the SMS , have the highest conversion rate, followed by those who opened the email.
- Applicants from Management specialization have the the highest conversion rate.

# EXPLORATORY DATA ANALYSIS (EDA)
## Categorical Variables




- Others have the highest conversion rates, Landing Page Submission have the highest number of leads.
- A large portion are from Google and Reference have the highest conversion count.
- In Last Activity, Sms Sent have highest conversion count.
- The conversion count is high for those who don't want free copy.

# LOGISTIC REGRESSION MODEL

# TRAIN & TEST AND MODELLING

## TRAIN & TEST

- For categorical variables having multiple levels, one-hot encoding was performed in creating dummy variables.
- The Lead_score dataframe is splitted in the ratio 70:30, where train set = 0.7 & test set = 0.3
- Rescaling numerical variables using Normalization (Min-Max-Scaler) technique to bring units of coefficient in same scale for modelling.
- Train set = (6310,29)  & Test set =(2705,29).

## MODELLING

- Recursive Feature Elimination is used to detect the 15 best variables for the logistic regression model.

- Summary and VIF is generated to detect insignificant and highly correlated data. Such variables are carefully removed one at a time while monitoring the p and VIF values.

- The model is finalised when p-value < 0.05 & VIF's are within acceptable range i.e. below 5.

# FINALISED MODEL

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | -2.2584 | 0.120 | -18.860 | 0.000 | -2.493 | -2.024 |
| Total Time Spent on Website | 4.5109 | 0.166 | 27.187 | 0.000 | 4.186 | 4.836 |
| Last Activity_Email Opened | 0.7724 | 0.099 | 7.810 | 0.000 | 0.579 | 0.966 |
| Lead Source_Direct Traffic | -1.7079 | 0.108 | -15.819 | 0.000 | -1.919 | -1.496 |
| Lead Source_Google | -1.2510 | 0.104 | -12.083 | 0.000 | -1.454 | -1.048 |
| Lead Source_Organic Search | -1.3544 | 0.126 | -10.740 | 0.000 | -1.602 | -1.107 |
| Lead Source_Reference | 2.3430 | 0.233 | 10.043 | 0.000 | 1.886 | 2.800 |
| Last Activity_SMS Sent | 1.8113 | 0.102 | 17.832 | 0.000 | 1.612 | 2.010 |
| What is your current occupation_Unemployed | 1.1355 | 0.081 | 14.040 | 0.000 | 0.977 | 1.294 |
| What is your current occupation_Working Professional | 3.5683 | 0.194 | 18.383 | 0.000 | 3.188 | 3.949 |
| Last Notable Activity_Modified | -0.6713 | 0.085 | -7.943 | 0.000 | -0.837 | -0.506 |

| | Features | VIF |
|---|---|---|
| 7 | What is your current occupation_Unemployed | 2.67 |
| 3 | Lead Source_Google | 2.48 |
| 0 | Total Time Spent on Website | 2.34 |
| 2 | Lead Source_Direct Traffic | 2.25 |
| 1 | Last Activity_Email Opened | 1.97 |
| 6 | Last Activity_SMS Sent | 1.95 |
| 4 | Lead Source_Organic Search | 1.55 |
| 9 | Last Notable Activity_Modified | 1.42 |
| 8 | What is your current occupation_Working Profes... | 1.37 |
| 5 | Lead Source_Reference | 1.31 |

- The above model is finalised with p-value < 0.05 & VIF's are within acceptable range i.e. below 5.
- Top 3 variables are based on coefficient:-
  - ❑ Total Time Spent on Website
  - ❑ What is your current occupation_Working Professional
  - ❑ Lead Source_Reference

0.1169994879672299

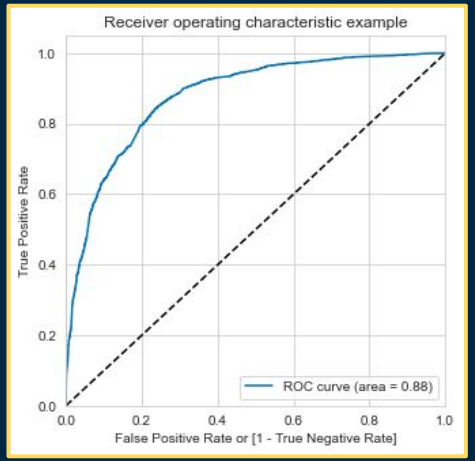
# MODEL METRICS – TRAIN DATA SET (Default Cutoff=0.5)

Confusion Matrix

| | |
|---|---|
| 3449 | 457 |
| 796 | 1608 |

- Metrics at default cutoff of 0.5

| | |
|---|---|
| Accuracy | 80.14% |
| Sensitivity/Recall | 66.89% |
| Specificity | 88.30% |
| False Positive Rate | 11.70% |

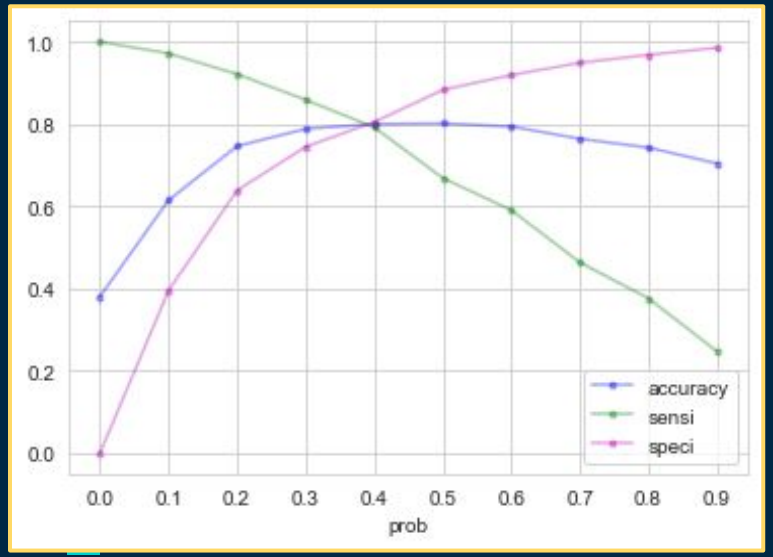



- ROC Curve area of 0.88 was obtained from the model

# MODEL METRICS – TRAIN DATA SET (OPTIMISED CUTOFF)

Confusion Matrix

| | |
|---|---|
| 3144 | 762 |
| 500 | 1904 |




Optimal cutoff was set at 0.4 from the plot.

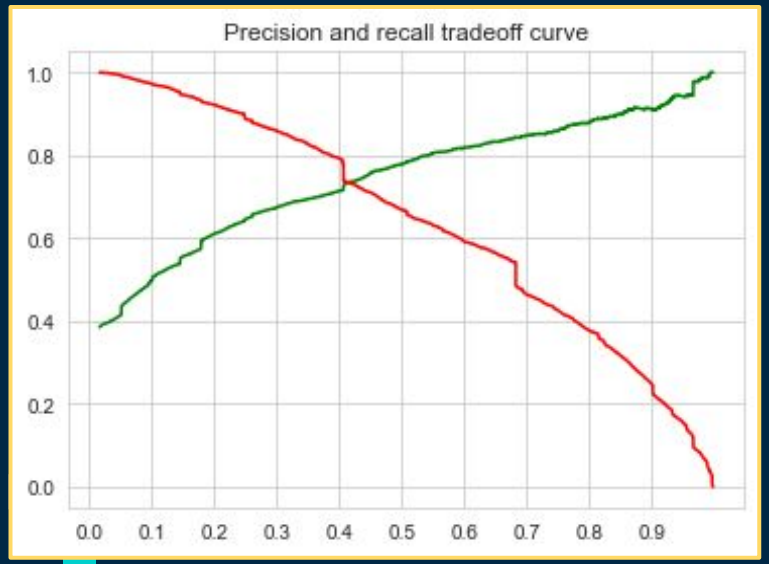
| | |
|---|---|
| Accuracy | 80.00% |
| Sensitivity/Recall | 79.20% |
| Specificity | 80.49% |
| Precision | 71.42% |
| False Positive Rate | 19.51% |

- The threshold for the was changed to 0.4 to get optimal metrics.

# MODEL METRICS – TEST DATA SET (OPTIMISED CUTOFF)

Confusion Matrix

| | |
|---|---|
| 1371 | 327 |
| 197 | 810 |




Metrics of Test data at cutoff 0.4

| | |
|---|---|
| Accuracy | 80.63% |
| Sensitivity/Recall | 80.44% |
| Specificity | 80.74% |
| Precision | 71.24% |
| False Positive Rate | 19.51% |

- The behaviour of the model can be adjusted according to the needs of the company by changing the cutoff values to vary Recall and Precision.

# RECOMMENDATIONS

- The cutoff for our analysis was set at 0.40 to get a baseline optimal recall and precision value however during implementation of the model it can be adjusted to suit the current needs and capabilities of the company
- Accuracy, Sensitivity, Specificity of test data set are 81%, 80% & 81% which are very close to respective values of train set.
- Target lead conversion rate for train set = 79.2% & test set = 80.4% on the final model.
- Recall for the model can be improved by dropping the cutoff below the 0.40 threshold.
- Following variables can have the highest impact on the conversion-
    - Total TIme Spent on Website
    - What is your current occupation
        - Working Professional
        - Unemployed
    - Lead Source
        - Reference
        - Google
    - Last Activity
        - SMS Sent
        - Email Opened

# RECOMMENDATIONS

- In order to get a more aggressive lead conversion rate, we can reduce the cutoff of the model below the 0.4 (optimal cutoff).

Cutoff = 0.3

Or

Cutoff = 0.2

| | prob | accuracy | sensi | speci | preci | f_score |
|---|---|---|---|---|---|---|
| 0.0 | 0.0 | 0.380983 | 1.000000 | 0.000000 | 0.380983 | 0.551756 |
| 0.1 | 0.1 | 0.614897 | 0.971714 | 0.395289 | 0.497233 | 0.657843 |
| 0.2 | 0.2 | 0.746434 | 0.921381 | 0.638761 | 0.610866 | 0.734660 |
| 0.3 | 0.3 | 0.788273 | 0.858985 | 0.744752 | 0.674396 | 0.755580 |
| 0.4 | 0.4 | 0.800000 | 0.792013 | 0.804916 | 0.714179 | 0.751085 |

- Or situations where the focus of sales team needs to be shifted and we need to be more sure about the customers approached , we can increase the threshold above 0.4 (optimal cutoff).

Cutoff = 0.5

Or

Cutoff = 0.6

| | prob | accuracy | sensi | speci | preci | f_score |
|---|---|---|---|---|---|---|
| 0.0 | 0.0 | 0.380983 | 1.000000 | 0.000000 | 0.380983 | 0.551756 |
| 0.1 | 0.1 | 0.614897 | 0.971714 | 0.395289 | 0.497233 | 0.657843 |
| 0.2 | 0.2 | 0.746434 | 0.921381 | 0.638761 | 0.610866 | 0.734660 |
| 0.3 | 0.3 | 0.788273 | 0.858985 | 0.744752 | 0.674396 | 0.755580 |
| 0.4 | 0.4 | 0.800000 | 0.792013 | 0.804916 | 0.714179 | 0.751085 |
| 0.5 | 0.5 | 0.801426 | 0.668885 | 0.883001 | 0.778692 | 0.719624 |
| 0.6 | 0.6 | 0.793978 | 0.590682 | 0.919099 | 0.817972 | 0.685990 |
| 0.7 | 0.7 | 0.764184 | 0.463810 | 0.949053 | 0.848554 | 0.599785 |
| 0.8 | 0.8 | 0.742789 | 0.376456 | 0.968254 | 0.879495 | 0.527236 |
| 0.9 | 0.9 | 0.704279 | 0.247504 | 0.985407 | 0.912577 | 0.389398 |

# THANK YOU!