# LEAD SCORING CASE STUDY

## USING LOGISTIC REGRESSION

**Submitted By-**

**Swikriti Pradhan**
**Devi Shri Kailash Ganti**
**Surabhi Sharma**

## PROBLEM STATEMENT

- An education company named X Education sells online courses to industry professionals. Many professionals who are interested in the courses land on their website and browse for courses.
- Once these people land on the website and fill up a form providing their email address or phone number, they are classified to be a lead.
- Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted but **typical lead conversion rate at X education is around 30%.**

## OBJECTIVE

- To build a model wherein 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%.**

# ASSUMPTIONS

- Data given is reliable
- Data type of each column is correct
- Outliers in data are treated successfully
- Data set is complete and missing value can be removed or imputed for further analysis.

# APPROACH & METHODOLOGY

- **Data set is imported in jupyter notebook and shape of the data set- 9240 * 37**

- **MISSING VALUE TREATMENT AND IMPUTATION-**
  - Columns having more than 45% null values has been dropped except **the column 'Lead Quality', which has 51.6% missing values seems important.**
  - For categorical variables, we'll analyse the count/percentage plots.
    - **'Lead Quality' -** null values in this column can be imputed with the value 'Not Sure' as we can assume that not filling means the employee does not know or is not sure about the option.
    - **'Specializations'-**There are a lot of different specializations and it's not accurate to directly impute with the mode. So it is imputed with 'Other'
    - **'City'-**Around 60% of the City values are Mumbai. We can impute 'Mumbai' in the missing values.
    - **'Other categorical column'-** impute with the most frequent values i.e mode.
    - **'Lead Source' and 'Last Activity'-** imputing with the most frequent value is not accurate as the next most frequent value has similar frequency. Also, as these variables have very little missing values, it is better to drop the rows containing these missing values.

  - For numerical variable, we'll describe the variable and analyse the box plots.
    - **'Total Visit' and 'Page Views per Visit'-** the percentage of missing values for both of them are less than 2%, it is better to drop the rows containing missing values.

- **CHECKING OUTLIERS**
  - **Use of boxplots,to check for outliers in the data**
    - **'TotalVisits'**- the 95% quantile is 10 whereas the maximum value is 251. Hence, we should cap these outliers at 95% value.
    - **'Total Time Spent on Website'**- no significant outliers
    - **'Page Views Per Visit'**- the 95% quantile is 6 whereas the maximum value is 55. Hence, we should cap these outliers at 95% value.

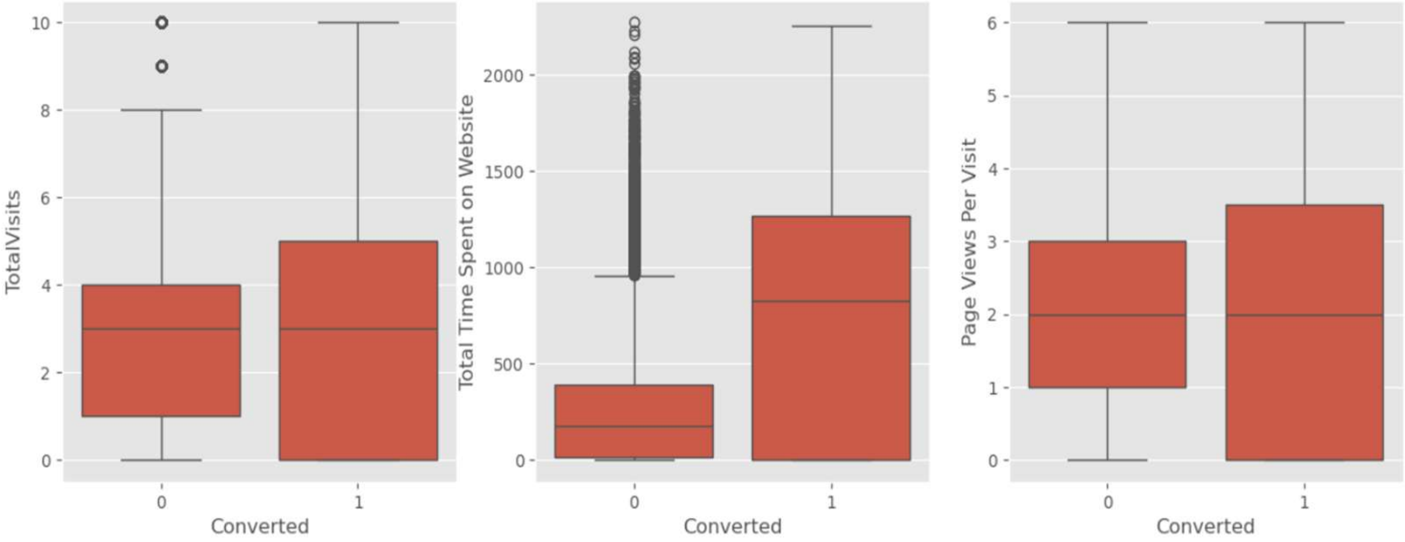
# APPROACH & METHODOLOGY

- DATA PREPARATION
  - DUMMY VARIABLE CREATION
    - **For categorical variables with multiple levels, we create dummy features (one-hot encoded).**
    - After dummy variable creations, no. of columns increased to 88.
  - TRAIN-TEST SPLIT
    - Data is splitted between Train data set and Test data set in ratio of 7:3 with random state of 100.
  - FEATURE SCALING
    - Standard Scalar applied to all numerical columns

- MODEL BUILDING
- MODEL EVALUATION

# ANALYSIS OF NUMERICAL & CATEGORICAL VARIABLES

- **VISUALISING NUMERICAL VARIABLES**



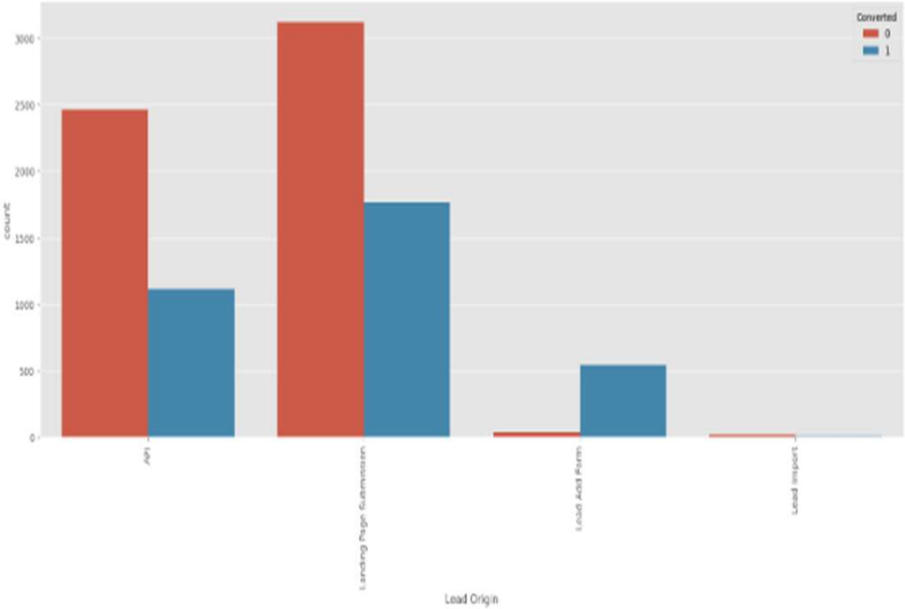
**Observations:**

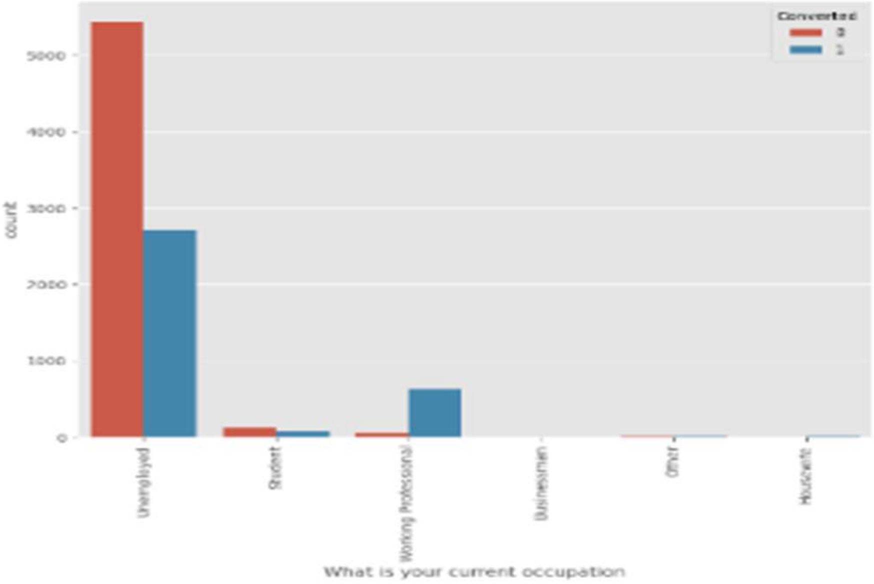
•'TotalVisits' has same median values for both outputs of leads. No conclusion can be drwan from this.

•People spending more time on the website are more likely to be converted. This is also aligned with our general knowledge.

•'Page Views Per Visit' also has same median values for both outputs of leads. Hence, inconclusive.

- **<u>VISUALISING CATEGORICAL VARIABLES</u>**
  - Lead Origin & What is your occupation

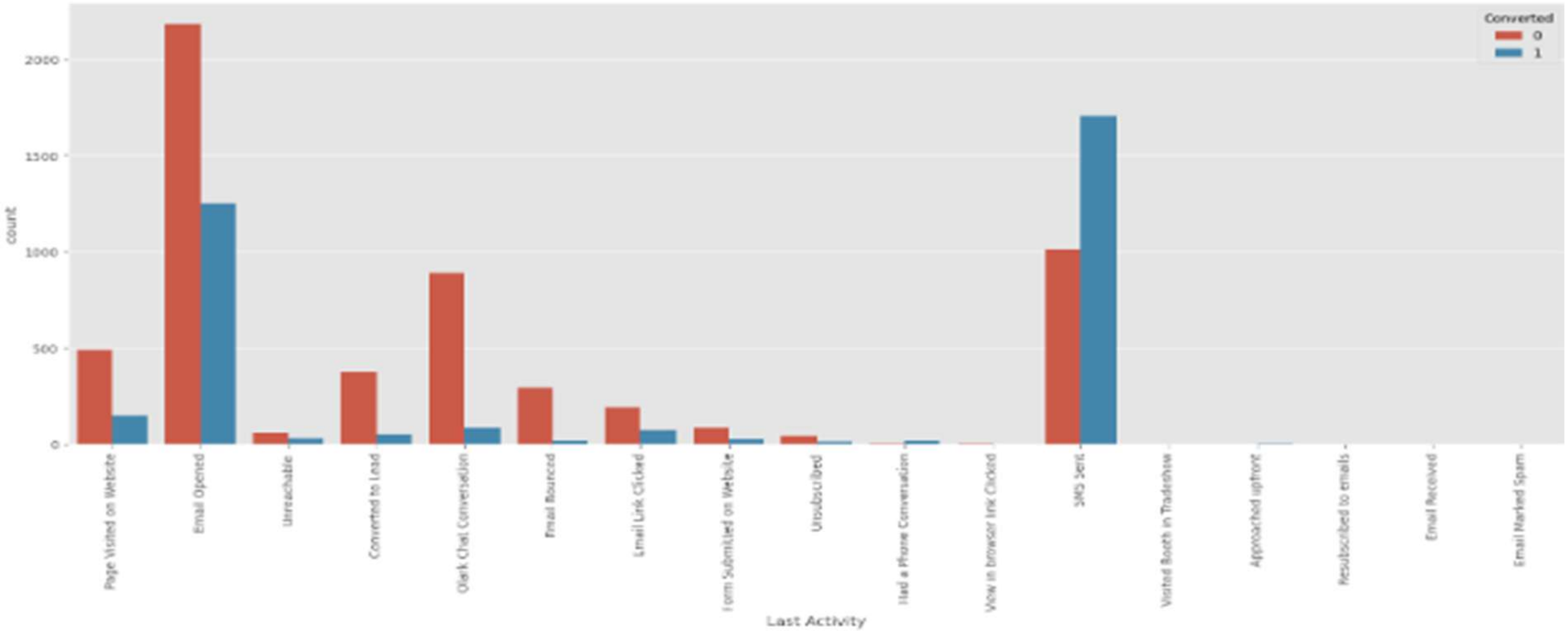




**Observations:**

- To improve overall performance, we should focus on increasing the conversion rate for 'API' and 'Landing Page Submission', both of which generate the most leads but have a conversion rate of around 30%.
- Additionally, we should work on boosting lead generation through the 'Lead Add Form', which has a strong conversion rate despite fewer leads.
- The highest conversion rate is for 'Working Professional'. High number of leads are generated for 'Unemployed' but conversion rate is low.

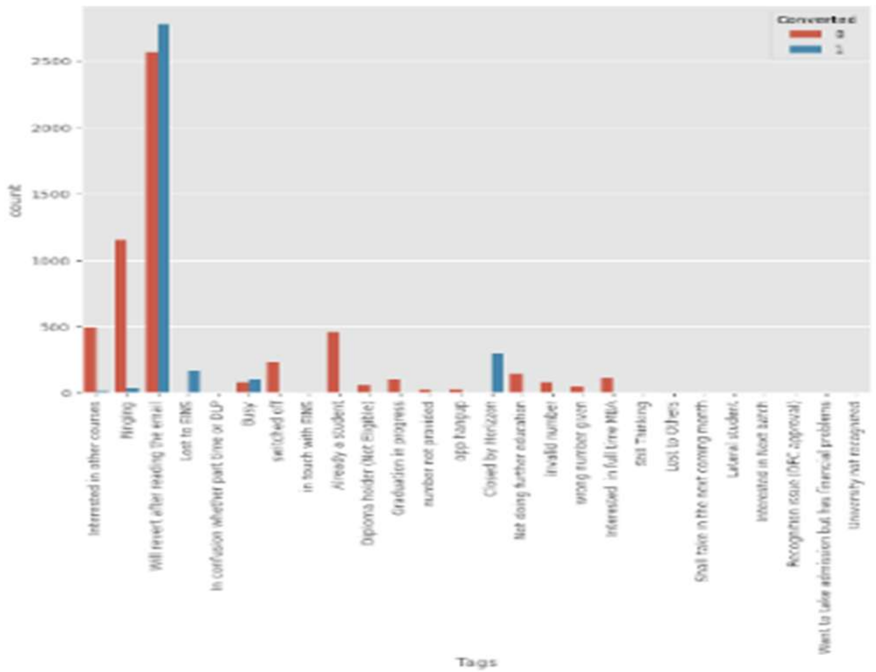
- **<u>VISUALISING CATEGORICAL VARIABLES  (contd..)</u>**
    - Last Activity

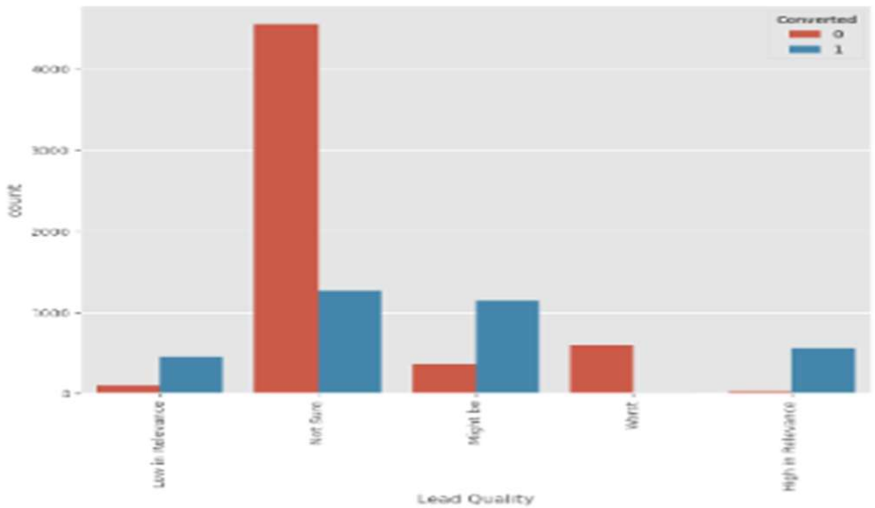


**Observations:**

- Highest number of lead are generated where the last activity is 'Email Opened' while maximum conversion rate is for the activity of 'SMS Sent'. Its conversion rate is significantly high.

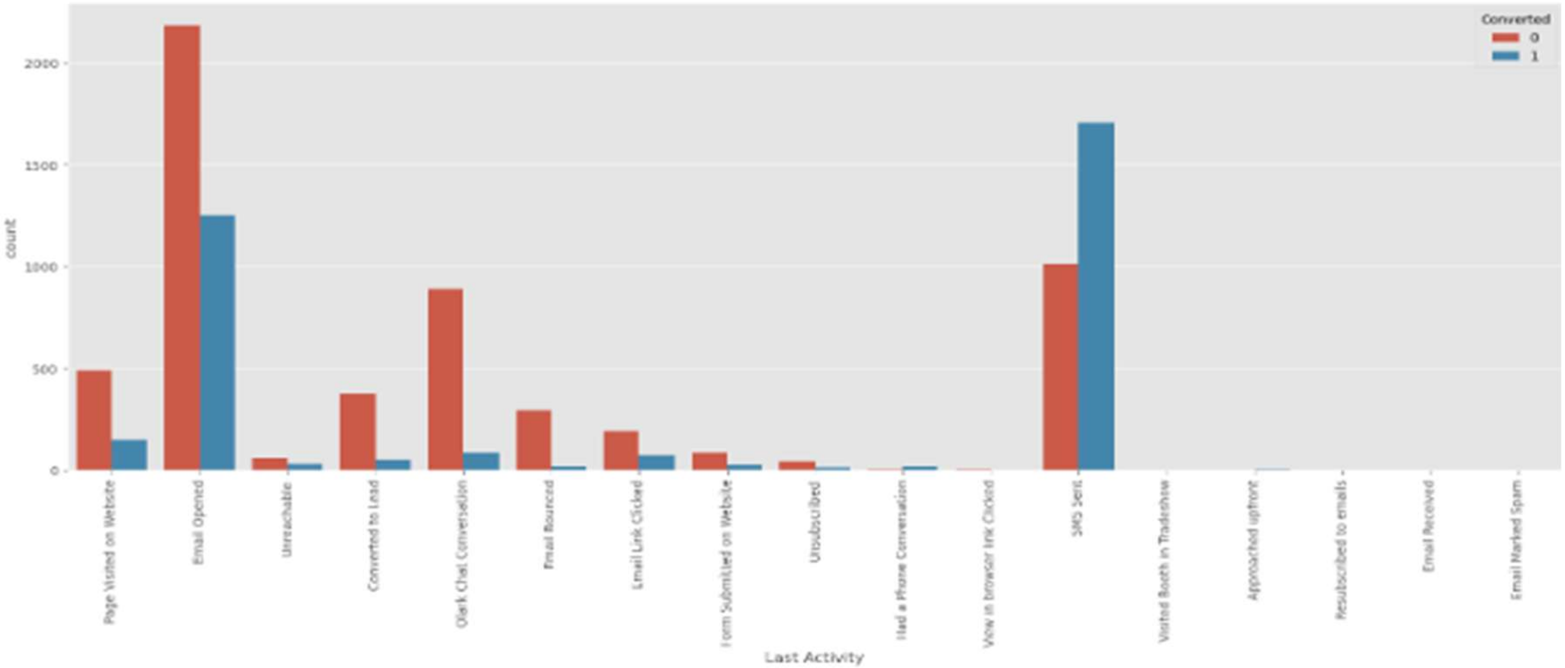
- **<u>VISUALISING CATEGORICAL VARIABLES (contd..)</u>**
    - Tags and Lead Quality





**Observations:**
•Most leads generated and the highest conversion rate are both attributed to the tag 'Will revert after reading the email'.
•In Lead quality, as expected, 'Might be' as the highest conversion rate while 'Worst' has the lowest.

- **<u>VISUALISING CATEGORICAL VARIABLES (contd..)</u>**
  - Last Activity



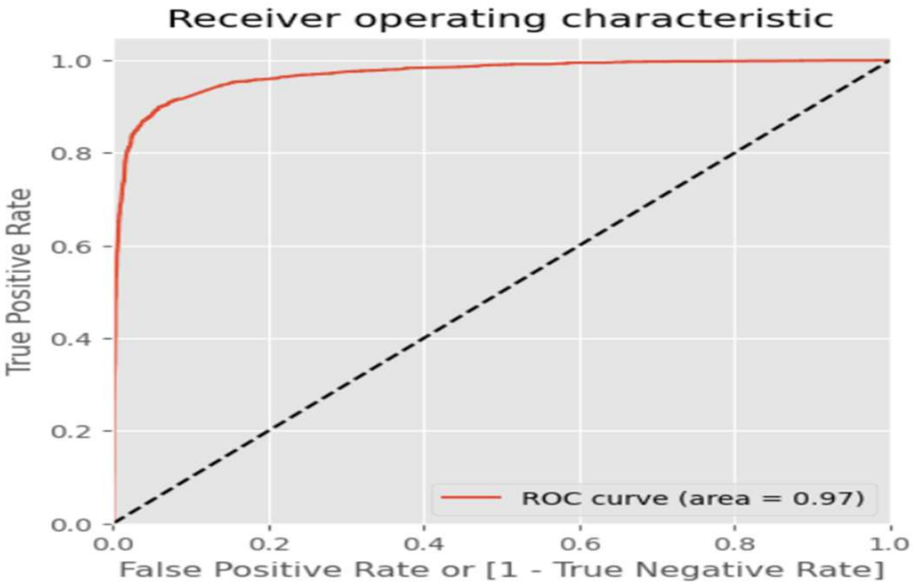
**Observations:**

- Highest number of lead are generated where the last activity is 'Email Opened' while maximum conversion rate is for the activity of 'SMS Sent'. Its conversion rate is significantly high.

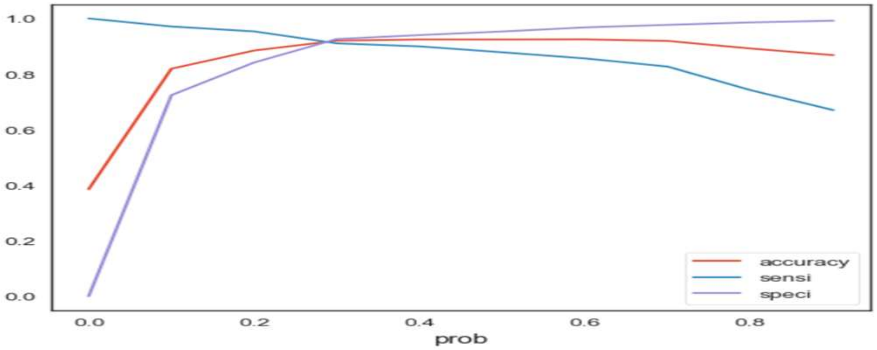
# BUILDING A MODEL

- **MODEL BUILDING**
  - FEATURE SELECTION USING RFE
    - RFE selected top 15 features
  - ASSESSING THE MODEL WITH STATSMODELS
    - Running the first model by using the features selected by RFE and evaluating p-value of all features
    - It is observed that ‘Tags_invalid_number’ & ‘Tags_number not provided’ has very high p-value. Hence it is insignificant and can be dropped
    - Running the second model by dropping above columns and it is observed that ‘Tags_wrong number’ has high p-value and can be dropped.
    - Running the third model by dropping above column and we also have to check VIFs (Variance Inflation Factors) of features to see if there's any multicollinearity present.
    - From VIF values and heat maps, we can see that there is not much multicollinearity present. All variables have a good value of VIF. These features seem important from the business aspect as well. So we need not drop any more variables and we can proceed with making predictions using this model only
  - ACCURACY OF FINAL MODEL
    - majorly handled all those variables with high p values
    - We have checked for colinearity among the feature variables
    - Training accuracy of 88.67% at a probability threshold of 0.05 is also very good.

- **MODEL EVALUATION**
  - ROC CURVE
    - tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
    - The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test
    - AUC (Area Under the Curve) = **0.96204**, indicating a highly accurate model.



- **MODEL EVALUATION**
  - FINAL OPTIMAL CUTOFF POINT
    - Optimal cutoff



  - CLASSIFICATION REPORT OF TRAIN DATA SET PREDICTION

| | Precision | Recall | F1-score | support |
|---|---|---|---|---|
| 0 | 0.97 | 0.84 | 0.90 | 3905 |
| 1 | 0.79 | 0.95 | 0.86 | 2446 |
| Accuracy | | | 0.88 | 6351 |

- **PREDICTION ON TEST SET**

  - AREA UNDER ROC CURVE- 0.96204
  - TRAINING ACCURACY IS 0.91
  - CLASSIFICATION REPORT OF TRAIN DATA SET PREDICTION

| | Precision | Recall | F1-score | support |
|---|---|---|---|---|
| 0 | 0.95 | 0.91 | 0.93 | 1734 |
| 1 | 0.85 | 0.92 | 0.88 | 989 |
| Accuracy | | | 0.91 | 2723 |

# CONCLUSION

- **KEY FINDINGS & BUSINESS INSIGHTS**
  - **IMPORTANT FEATURES FOR LEAD CONVERSION:**
    - 'Tags_Lost to EINS’
    - 'Tags_Closed by Horizzon’
    - 'Tags_Will revert after reading the email’
  - **ACTIONABLE INSIGHTS FOR THE SALES TEAM:**
    - Prioritize leads with high lead scores.
    - Focus on leads engaging through 'SMS Sent' as their conversion rate is high.
    - Improve conversion efforts on 'API' and 'Landing Page Submission' leads.
    - Increase efforts to attract 'Working Professionals,' as they convert the most.