LEAD SCORE CASE STUDY

Group members
ANANNA PATRA
ANIKET SHAMBHARKAR
SHIVA KUMAR

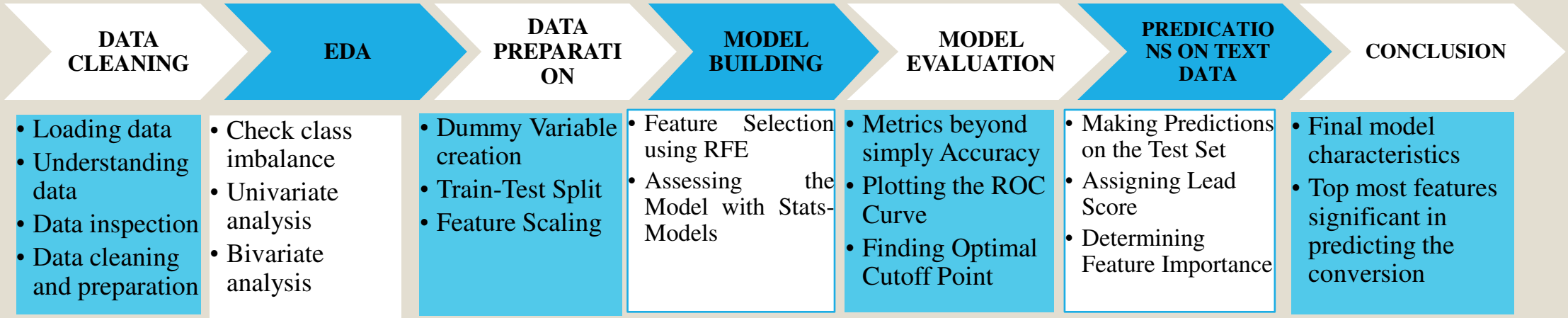
PROBLEM STATEMENT

- X education sells online courses to industry professionals.
- They acquire leads through their website where professionals browse for courses and fill out forms.
- After acquiring leads, the sales team contacts them through calls and emails.
- The typical lead conversion rate is around 30%, meaning only 30 out of 100 leads are converted.
- The company wants to improve efficiency by identifying hot leads, those with the highest potential for conversion.
- Current lead generation methods are not effectively helping with conversions.
- The goal is to increase the conversion rate by targeting leads more effectively.

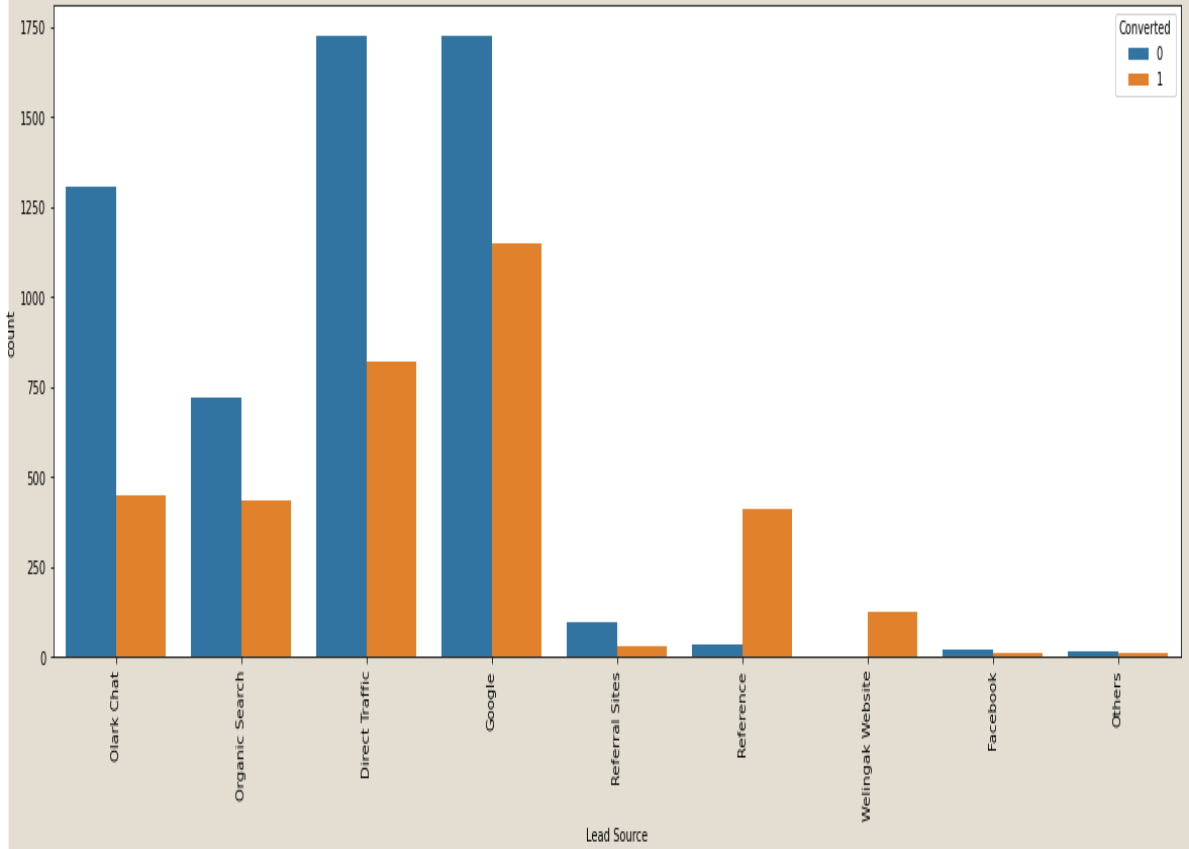
BUSINESS STRATEGY

- X education aims for an 80% lead conversion rate.
- They seek to develop a lead scoring model ranging from 0 to 100 to identify hot leads.
- The model should consider future constraints such as peak time actions and manpower utilization.
- After achieving the target, approaches for sustaining success need to be outlined.
- The CEO is focused on improving lead conversion efficiency through strategic modeling.

ANALYSIS APPROACH

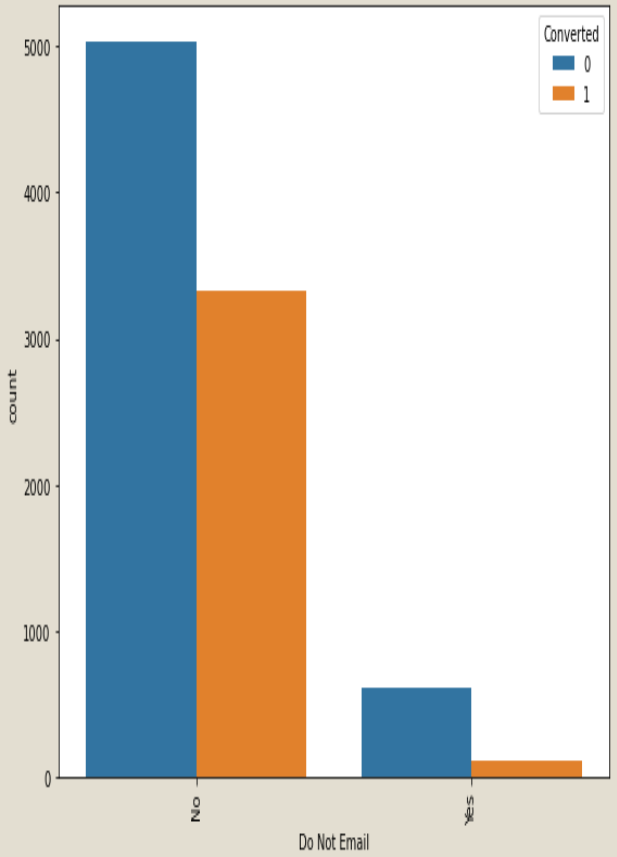


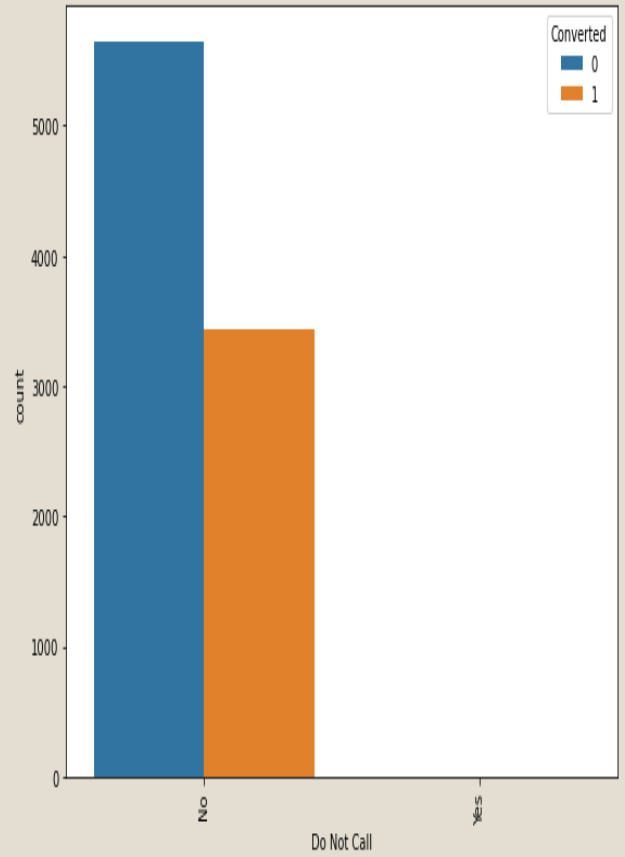
EXPLORATORY DATA ANALYSIS

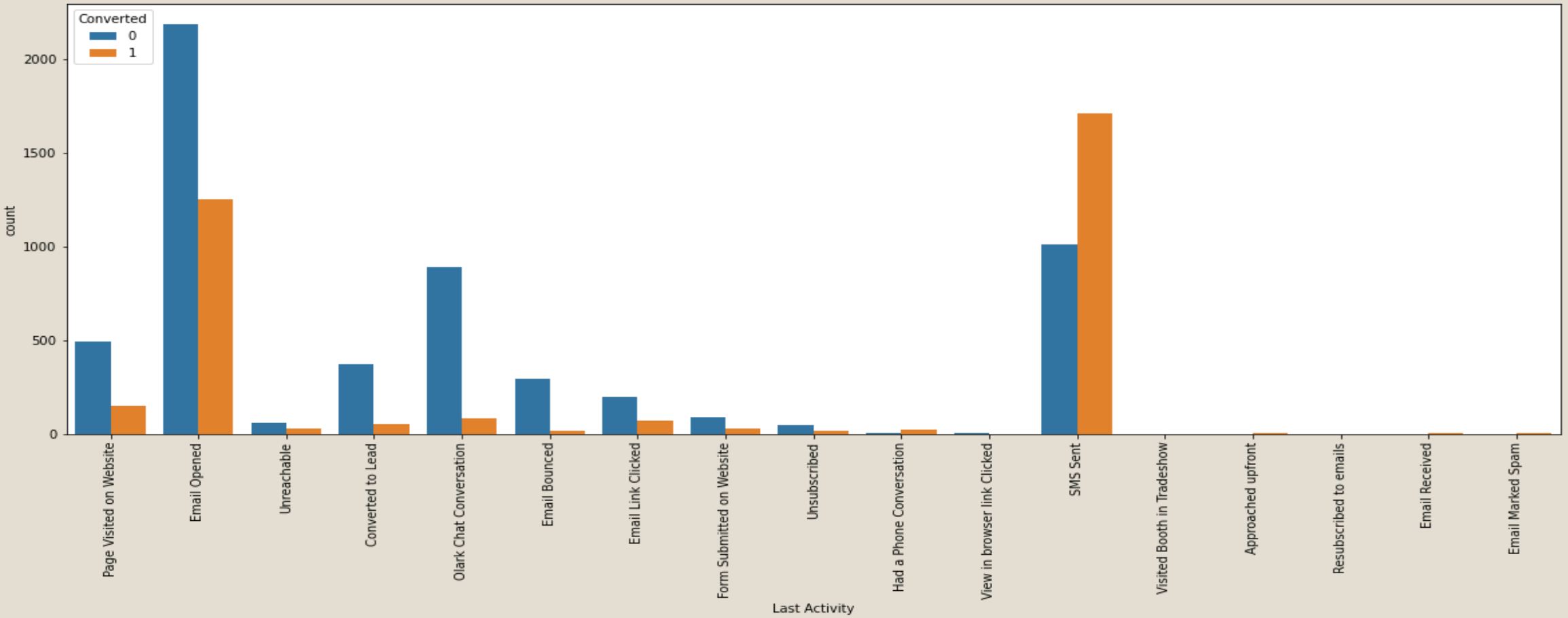


- **Google searches have resulted in more successful conversions compared to other methods, while referrals have shown a higher rate of successful conversions.**

- **Observations for Do Not Email and Do Not Call : As one can expect, most of the responses are 'No' for both the variables which generated most of the leads.**

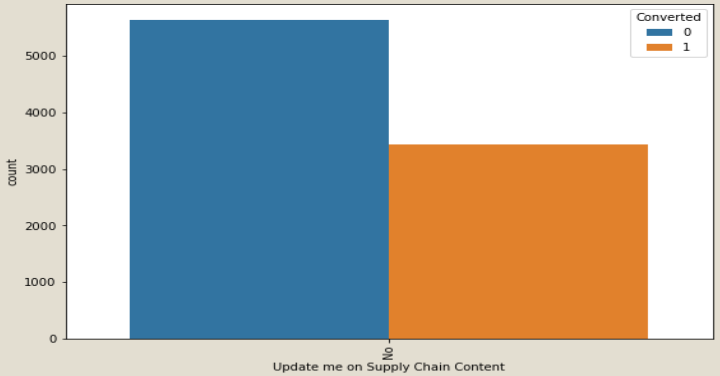


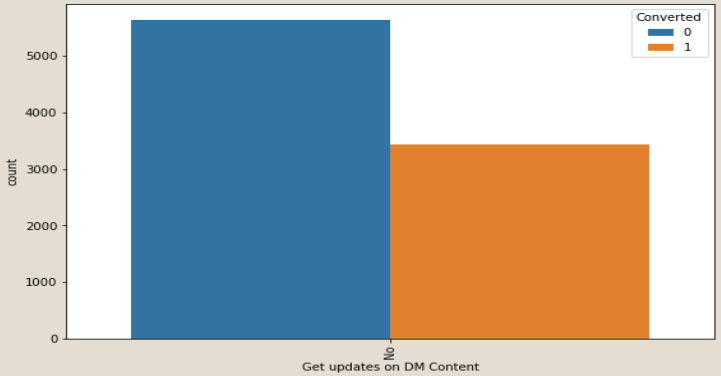


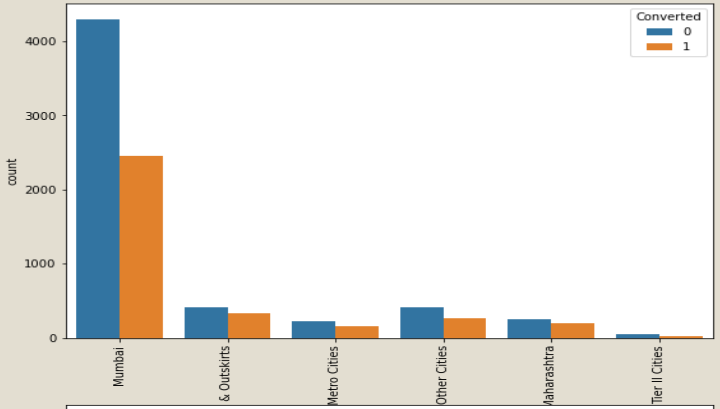


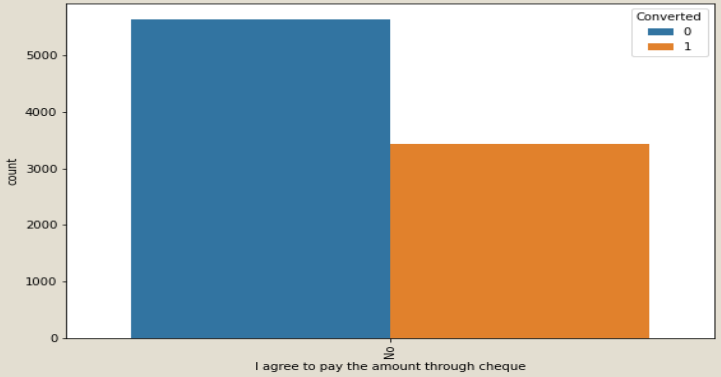
Observations for Last Activity :

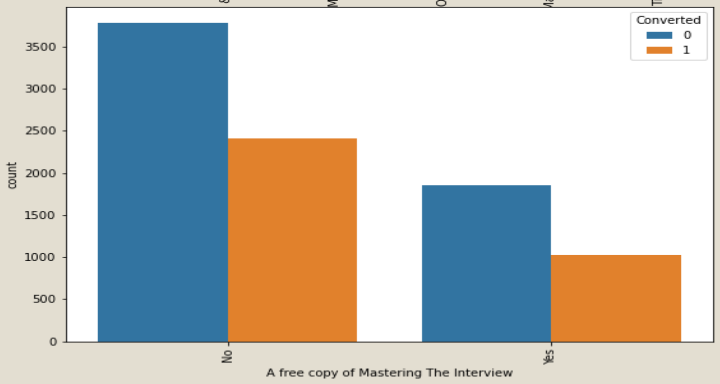
Highest number of lead are generated where the last activity is 'Email Opened' while maximum conversion rate is for the activity of 'SMS Sent'. Its conversion rate is significantly high. Categories after the 'SMS Sent' have almost negligible effect. We can aggregate them all in one single category.

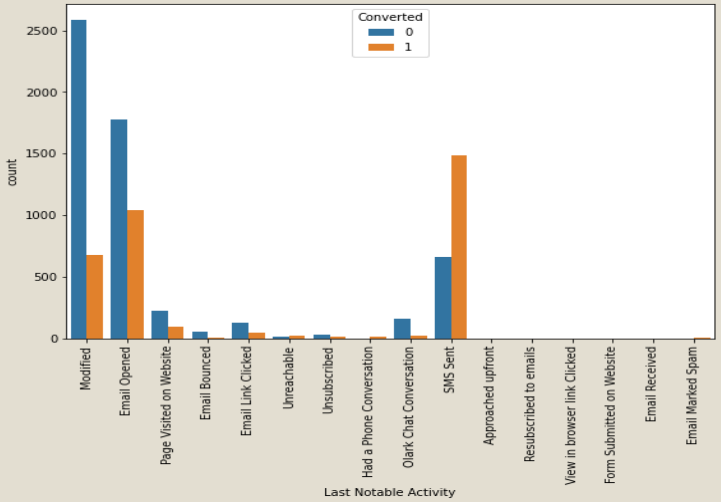












Observations for Update me on Supply Chain Content, Get updates on DM Content, City, I agree to pay the amount through cheque, A free copy of Mastering The Interview, and Last Notable Activity :

Most of these variables are insignificant in analysis as many of them only have one significant category 'NO'.

In City, most of the leads are generated for 'Mumbai'.

In 'A free copy of Mastering The Interview', both categories have similar conversion rates.

In 'Last Notable Activity', we can combine categories after 'SMS Sent' similar to the variable 'Last Activity'. It has most generated leads for the category 'Modified' while most conversion rate for 'SMS Sent' activity.

DATA PREPARATION

1) Binary-level categorical columns were previously encoded as 1s and 0s.
2) We generated additional features using one-hot encoding for categorical variables such as Lead Origin, Lead Source, Last Activity, Specialization, and Current Occupation.
3) The dataset was split into training and testing sets with a ratio of 70:30.
4) Feature scaling was performed using the standardization method to ensure uniformity in scale across features. We have 37.85% conversion rate.
5) Correlation analysis was conducted to identify and eliminate predictor variables with high correlations, such as Lead Origin_Lead Import and Lead Origin_Lead Add Form.

MODEL BUILDING

Recursive Feature Elimination (RFE) reduced the dataset from 48 to 15 columns, focusing on the most influential features.

Manual feature reduction further refined the model by dropping variables with p-values greater than 0.05.

Model 4 emerged as stable after four iterations, with significant p-values (< 0.05) and VIFs less than 5.

The final model, logm4, will be used for Model Evaluation and predictions, ensuring efficiency and accuracy.

MODEL EVALUATION

Metrics beyond simply Accuracy

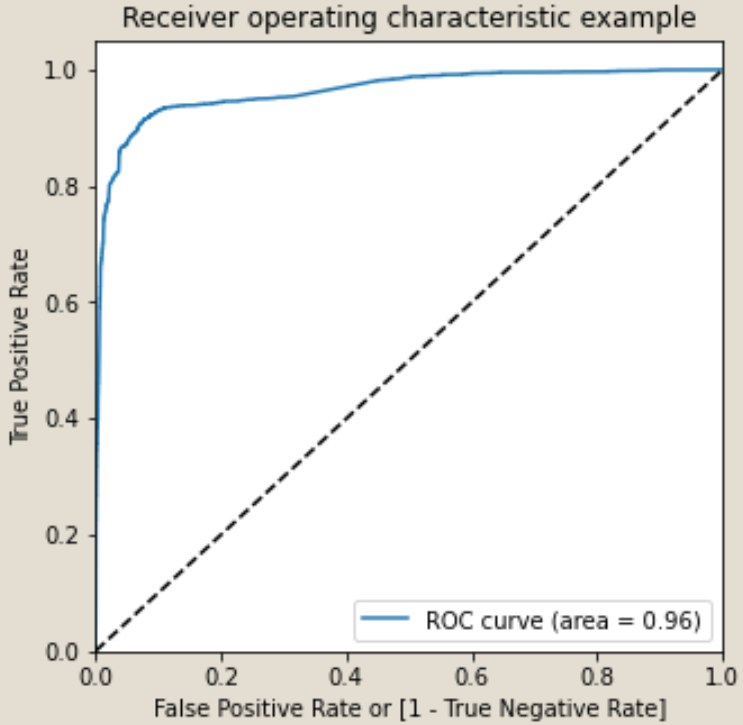
```
Confusion Matrix:
[[3701  204]
 [ 297 2149]]
Training Accuracy:  0.9211147850732169
```

```
Sensitivity:  0.8785772690106296
Specificity:  0.9477592829705506
False postive rate - predicting the lead conversion when the lead does not convert:  0.0522407170294942
Positive predictive value:  0.9133021674458138
Negative predictive value:  0.9257128564282141
```

1. Plotting the ROC Curve

An ROC curve demonstrates several things:

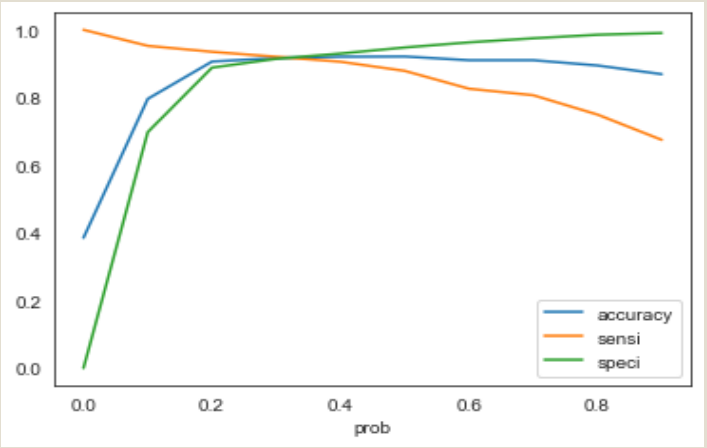
- **It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).**
- **The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.**
- **The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.**



- **Area under curve (auc) is approximately 0.96 which is very close to ideal auc of 1.**

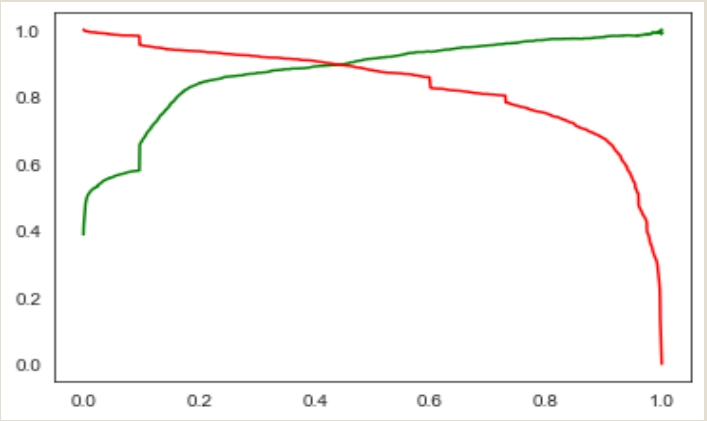
2. Finding Optimal Cutoff Point

- **Optimal cutoff probability is the prob where we get balanced sensitivity and specificity.**



- **From the curve above, 0.2 is the optimum point to take as a cutoff probability.**

3. Precision and Recall Tradeoff



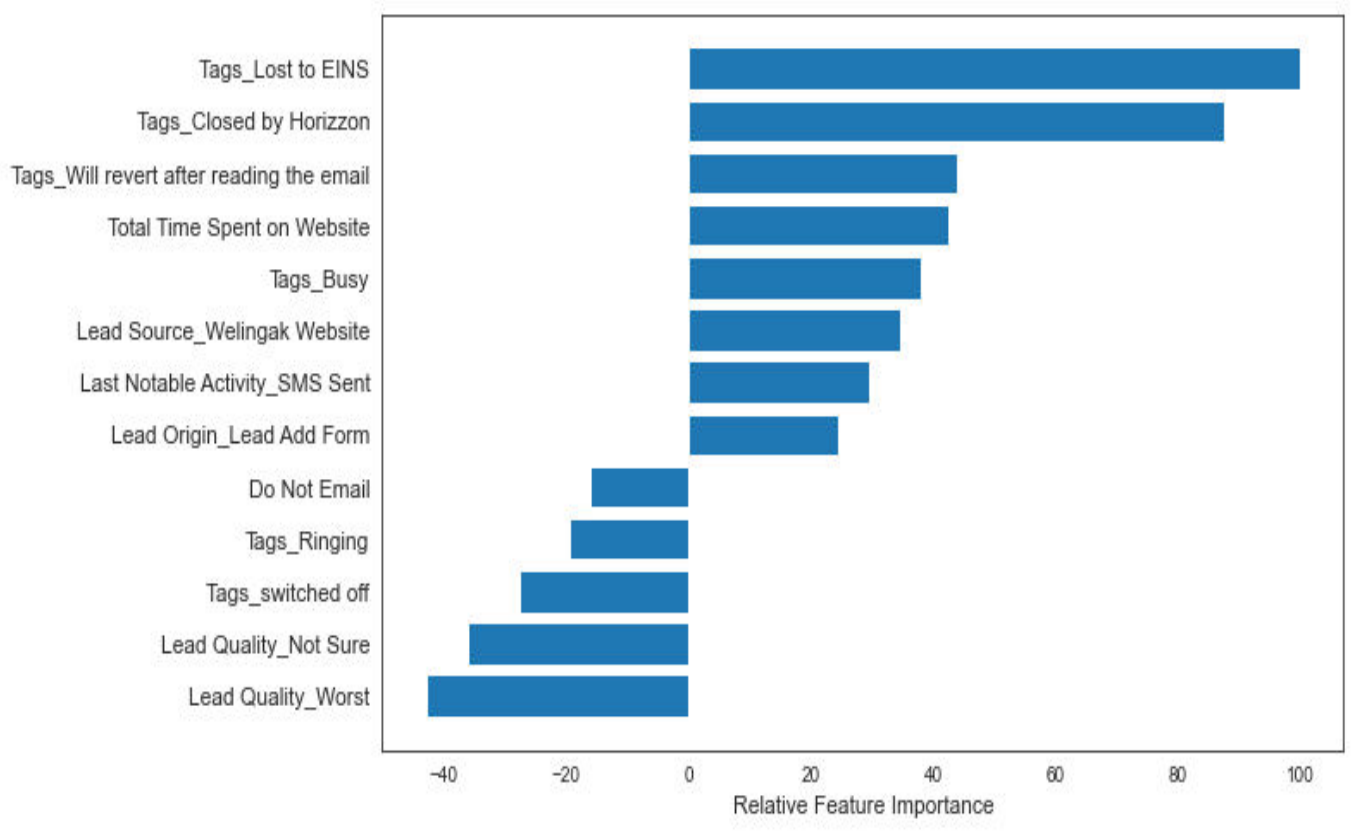
- **From the curve above, 0.2 is the optimum point to take as a cutoff probability using Precision-Recall. We can check our accuracy using this cutoff too.**

OBSERVATIONS

- **TRAIN DATA**
- **Accuracy-92%**
- **Recall-87%**
- **Precision-91%**
- **Sensitivity-90%**
- **Specificity-92%**

- **TEST DATA**
- **Accuracy-89%**
- **Recall-91%**
- **Precision-89%**
- **Sensitivity-92%**
- **Specificity-88%**

DETERMINING FEATURE IMPORTANCE



Do Not Email -11
Total Time Spent on Website -10
Lead Origin_Lead Add Form- 9
Lead Source_Welingak Website- 7
Tags_Busy 0
Tags_Closed by Horizzon -2
Tags_Lost to EINS- 12
Tags_Ringing- 3
Tags_Will revert after reading the email -4
Tags_switched off- 1
Lead Quality_Not Sure- 8
Lead Quality_Worst -5
Last Notable Activity_SMS Sent -6

CONCLUSION

Final Model

I. All p-values are very close to zero.
II. VIFs for all features are very low, indicating minimal multicollinearity.
III. The overall testing accuracy is 89.86% at a probability threshold of 0.11, which is excellent.

Features with Positive Impact on Conversion Probability:

1. Tags_Lost to EINS
2. Tags_Closed by Horizzon
3. Tags_Will revert after reading the email
4. Tags_Busy
5. Lead Source_Welingak Website
6. Last Notable Activity_SMS Sent
7. Lead Origin_Lead Add Form

Features with Negative Impact on Conversion Probability:

1. Lead Quality_Worst
2. Lead Quality_Not Sure
3. Tags_switched off
4. Tags_Ringing
5. Do Not Email


thank
you