# LEAD SCORE CASE STUDY

Submitted by :

Thanikachalam KS

# BUSINESS GOAL

- X Education requires assistance in identifying the most promising leads, i.e., the leads that are more likely to convert into paying customers.
- The company requires a reliable model that can assign a score to each lead based on their likelihood of converting into a potential customer.
- Additionally, the CEO has set a target lead conversion rate of approximately 80%.

# PROBLEM SOLVING METHODOLOGY

**Data Sourcing , Data Cleaning**

- Reading the data.
- Check and handle duplicate data.
- Check and handle missing values.
- Check and handle outlier and skewed data.
- Grouping the low frequency values.

**Exploratory Data Analysis**

- Visualizing the target variable.
- Visualizing the categorical and numerical variable.

**Data Preparation and Feature Selection**

- Getting dummies
- Splitting data into train and test set
- Feature Scaling of Numeric data using MinMaxScaler.
- Feature Selection Using RFE

**Model Building**

- Model building and optimal cutoff values is found.
- Calculate various metrics like accuracy, sensitivity etc.

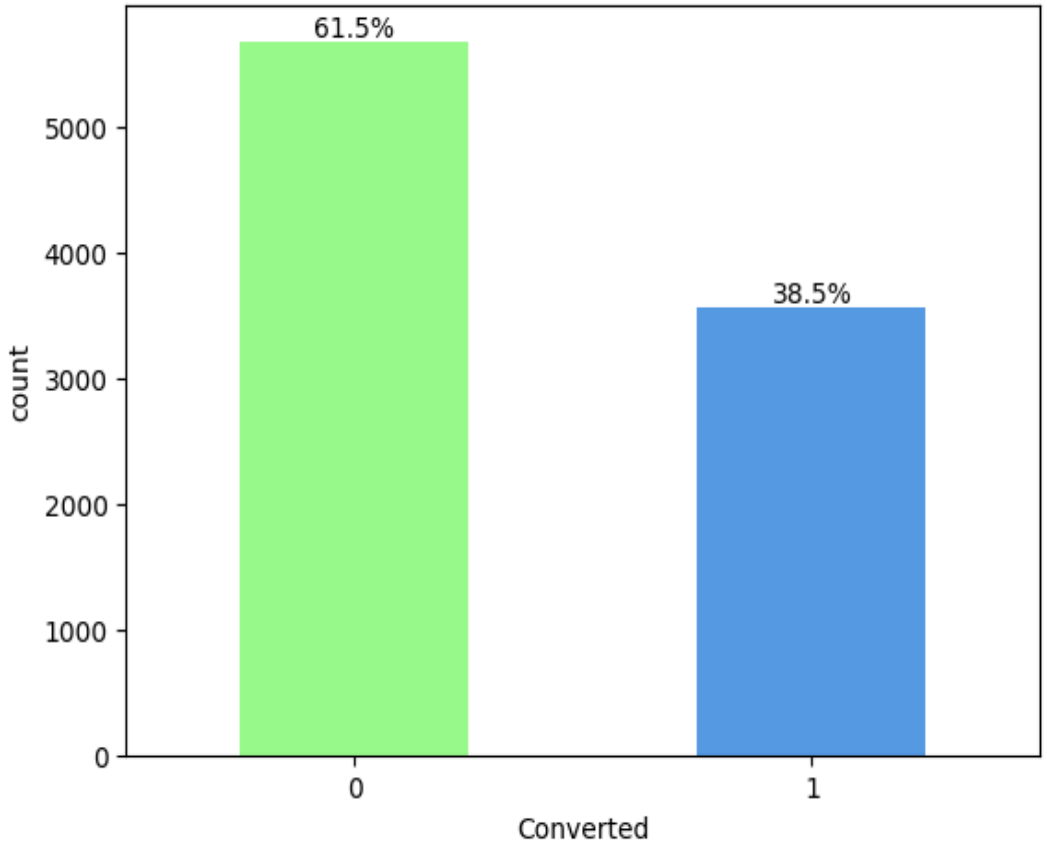
**Result**

- Prediction on test set and calculating lead score.
- Finding top 3 features that have a positive impact on predicting .

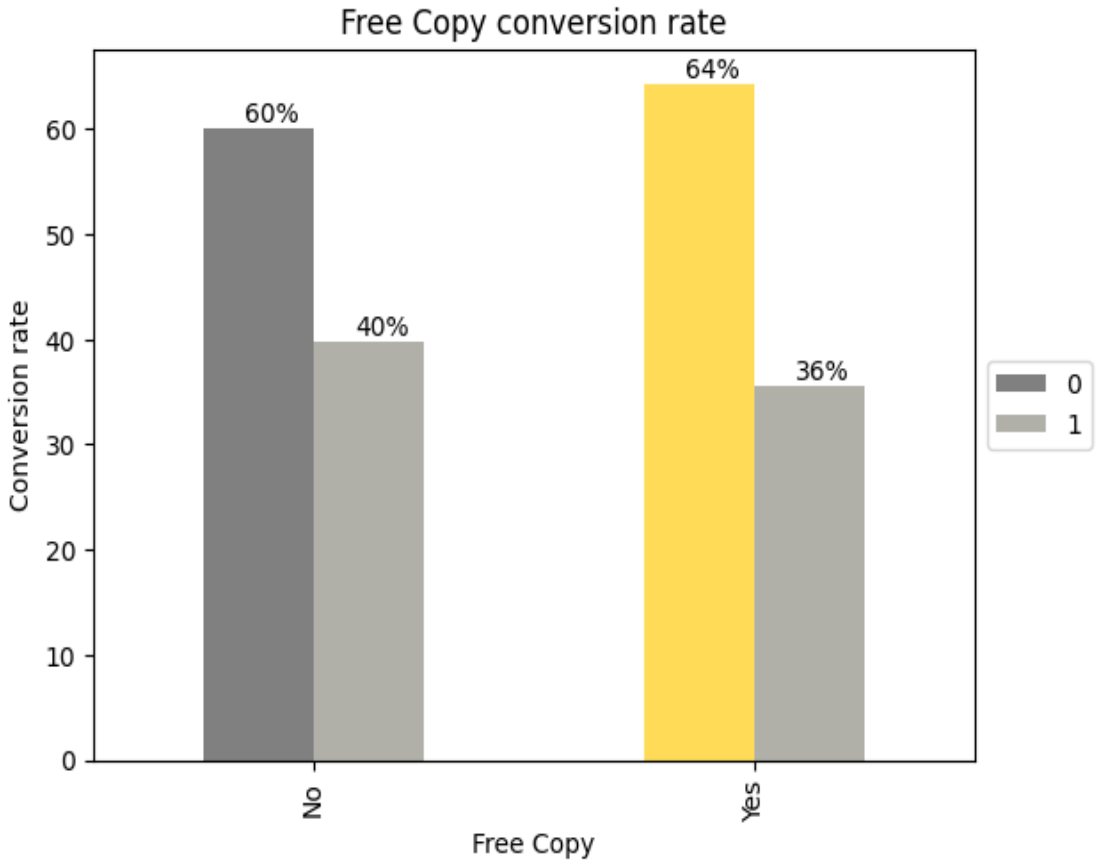
- Many of the categorical variables have a level called "Select", which is replaced with np.nan.
- During the process of data cleaning, columns with missing values greater than 40%, columns with only one unique value, unwanted columns, and skewed columns are dropped.
- As a result, the shape of the dataset is reduced from (9240, 37) to (9240, 11).
- To treat the outliers, the Capping/Winsorization method is used.
- For scaling the numerical columns, the MinMaxScaler method is used.
- The variables considered for building the model after data cleaning are Lead Origin, Lead Source, Do Not Email, Total Visits, Total Time Spent on Website, Page Views Per Visit, Last Activity, Specialization, Current Occupation, and Free Copy.

# Exploratory Data Analysis

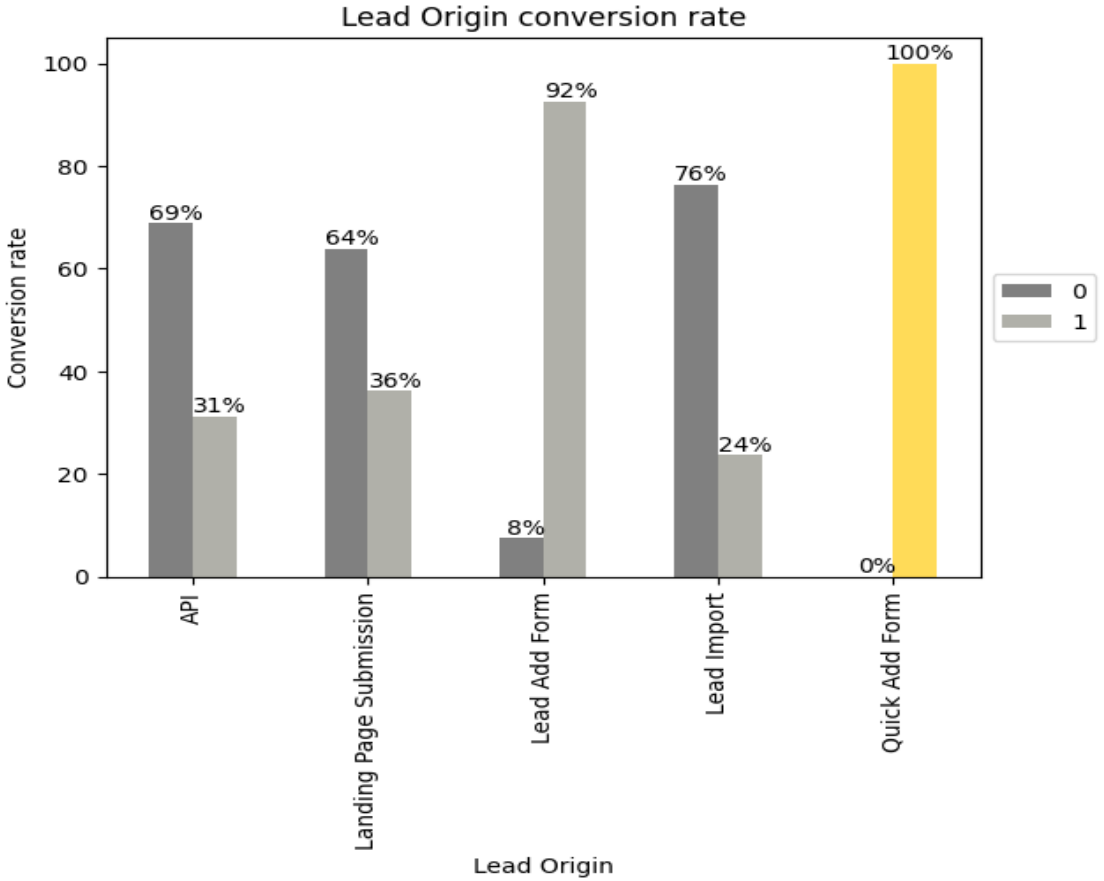
- Conversion rate is of 38.5%  meaning only 38.5% of the people have converted to leads.(Minority)
- While 61.5% of the people didn't convert to leads. (Majority).
- Data imbalance ratio : 1.59 : 1 (61.5 : 38.5)



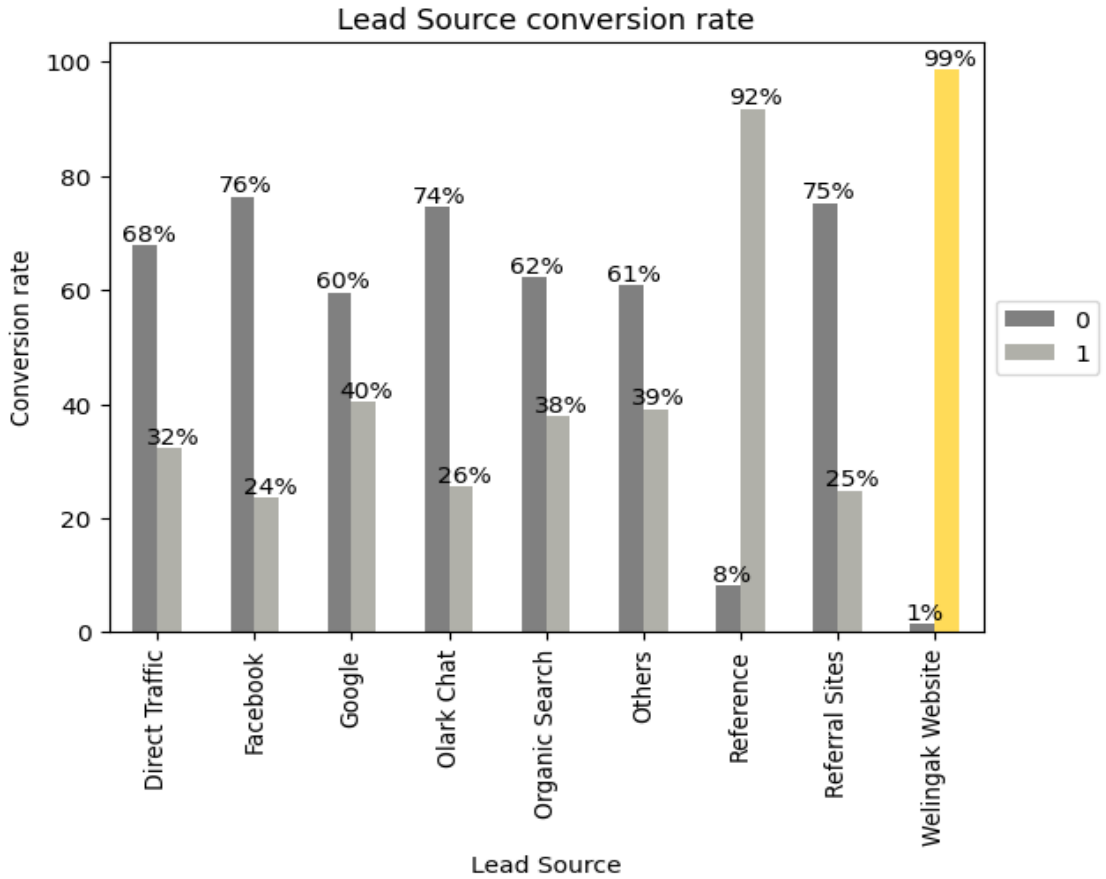

- More leads are from those who do not ask for free copy of Mastering Interviews. Can be focused for conversion.



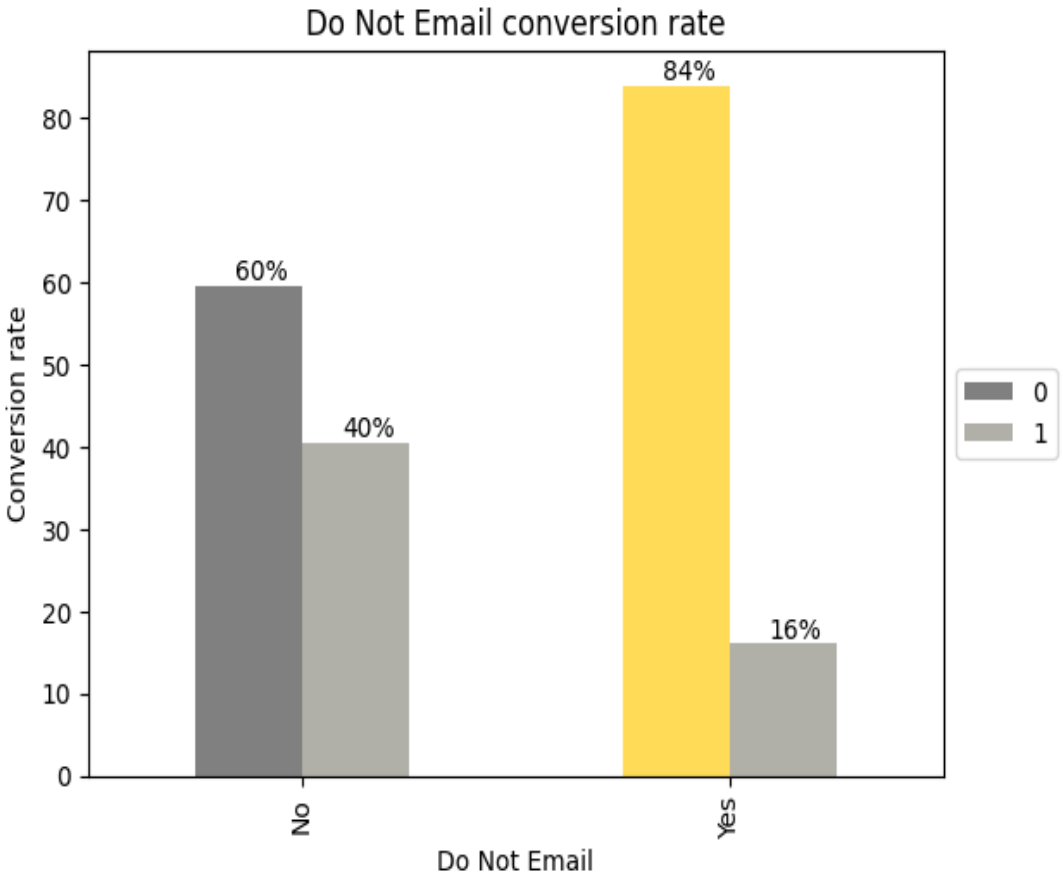
- Landing Page Submission identified 53% of customers, and API identified 39% of customers.
- Around 52% of all leads originated from Landing Page Submissions with a LCR of 36%. The API identified approximately 39% of customers with a LCR of 31%.
- Lead Add Form has a more than 90% conversion rate, but the count of lead are not very high. Lead Import and Quick Add From are very low in count so the conversation rate is high.



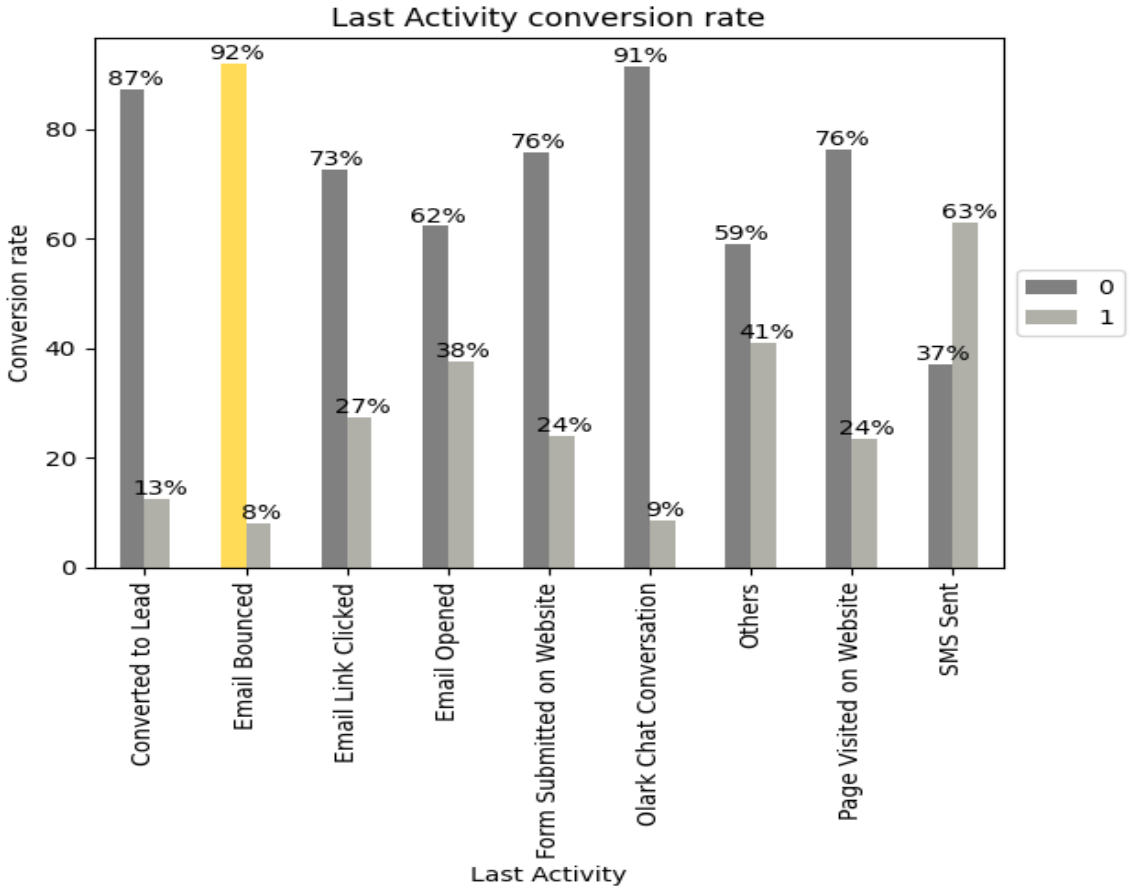

- Our data shows that 58% of our leads come from Google and Direct Traffic.
- Google has the highest LCR of 40% from 31% of our customers, followed by Direct Traffic with 32% LCR from 27% of customers.
- Organic Search provides 37.8% of LCR but only 12.5% of customers. Reference has the highest LCR of 91% but only 6% of customers through this source.



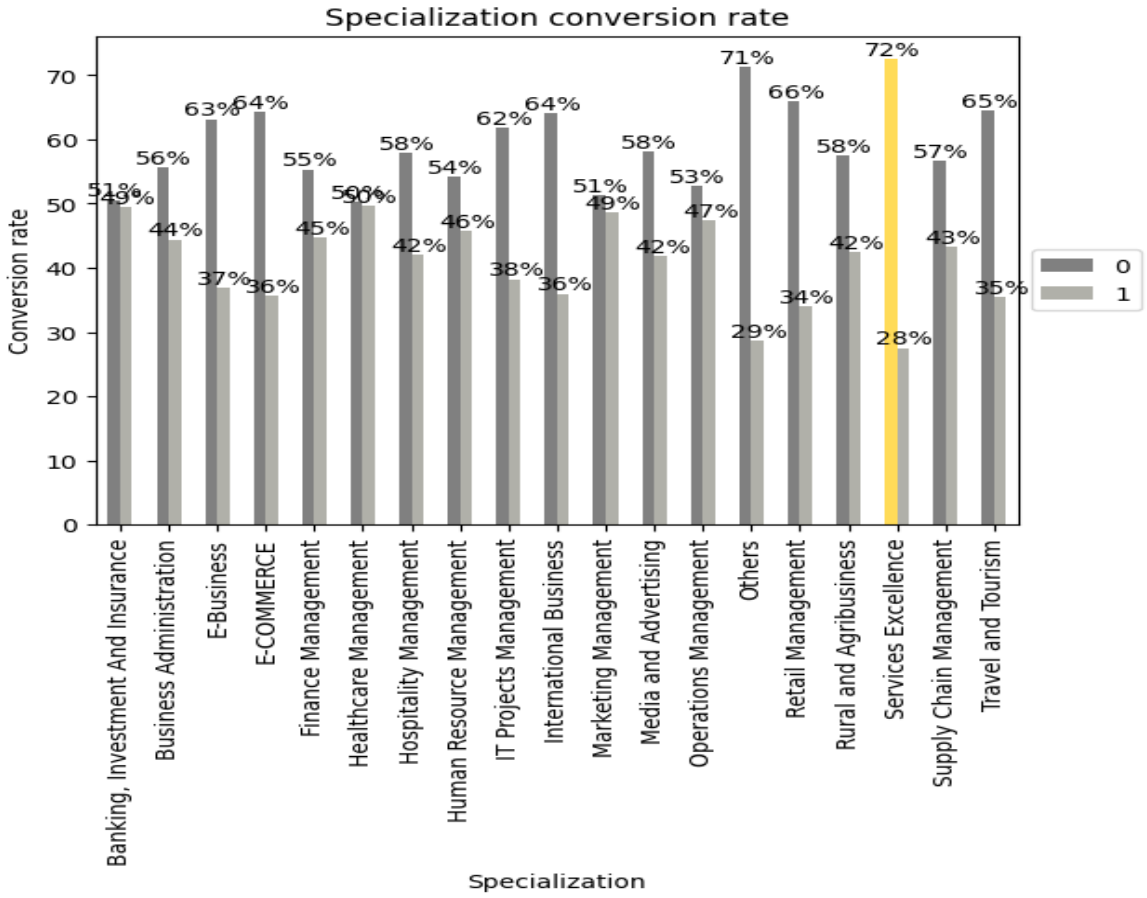
- 92% of the people have opted that they Do Not Email about the course.
- the people who have opted not to Email about the course have an LCR of 40% out of 92%.
- the people who have opted to receive emails about the course have LCR of 16% out of 10%.



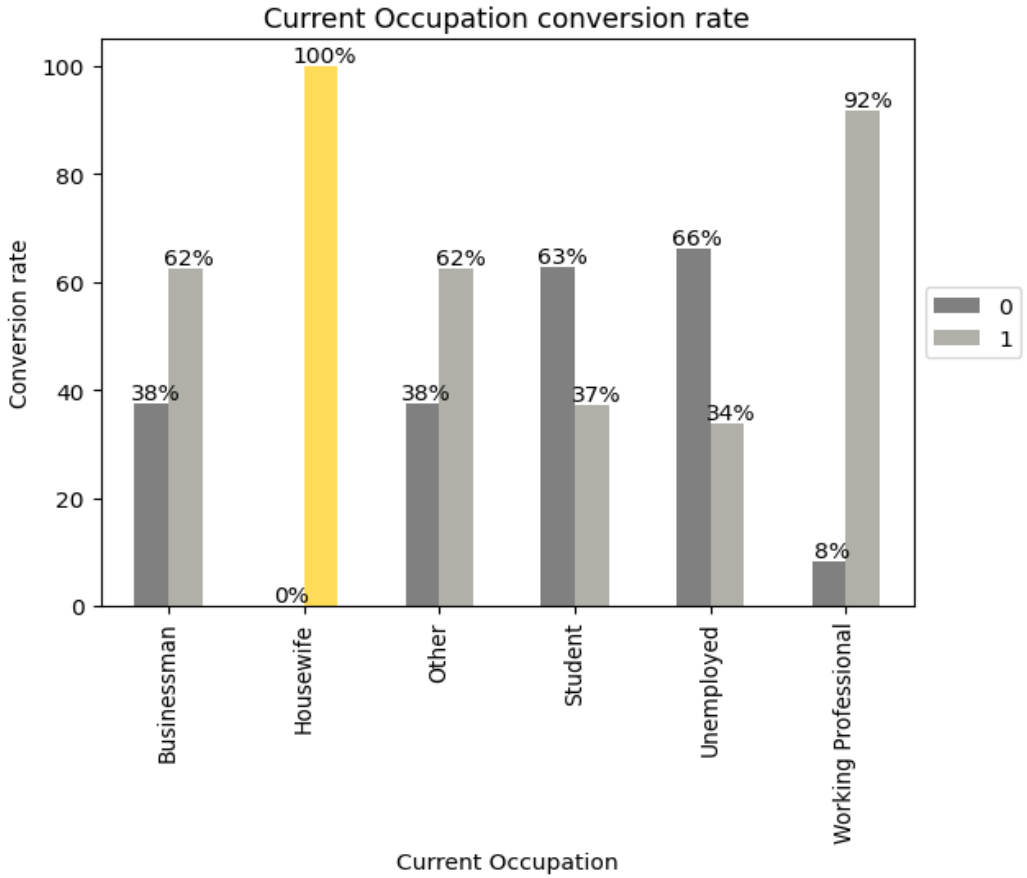

- 68% of customers contribute to SMS Sent and email Opened activities.
- SMS Sent has a high LCR of 63% with 30% contribution from last activities, Email Opened activity contributed 38% of last activities performed by the customers with 37% LCR.



- Most of the leads are from Other (Not specified) specialization but the conversion rate is poor.
- banking, Investment, and insurance, business, healthcare, HR, Marketing, supply chain management have LCR.

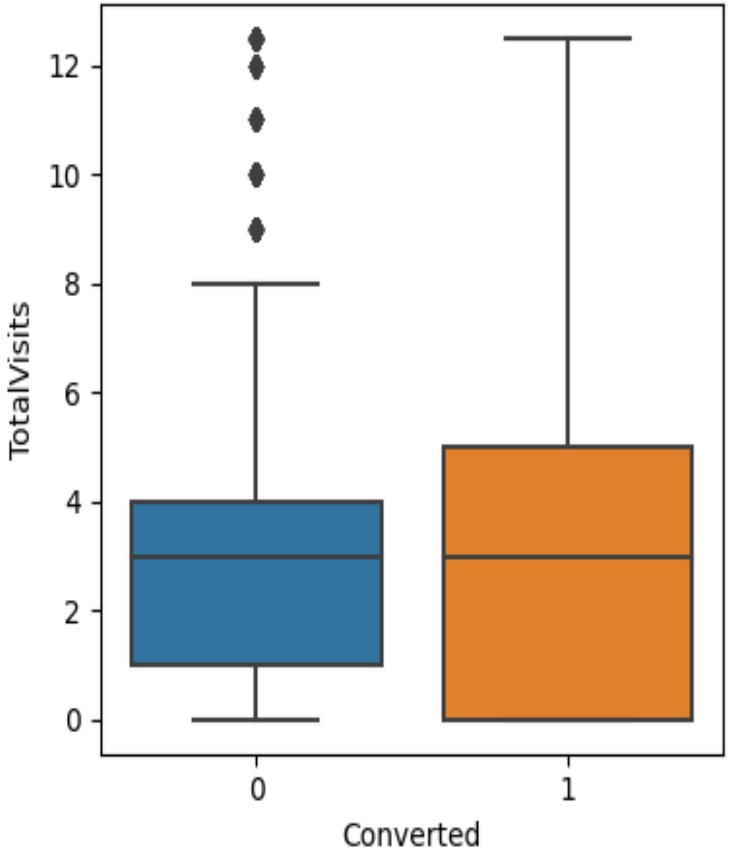


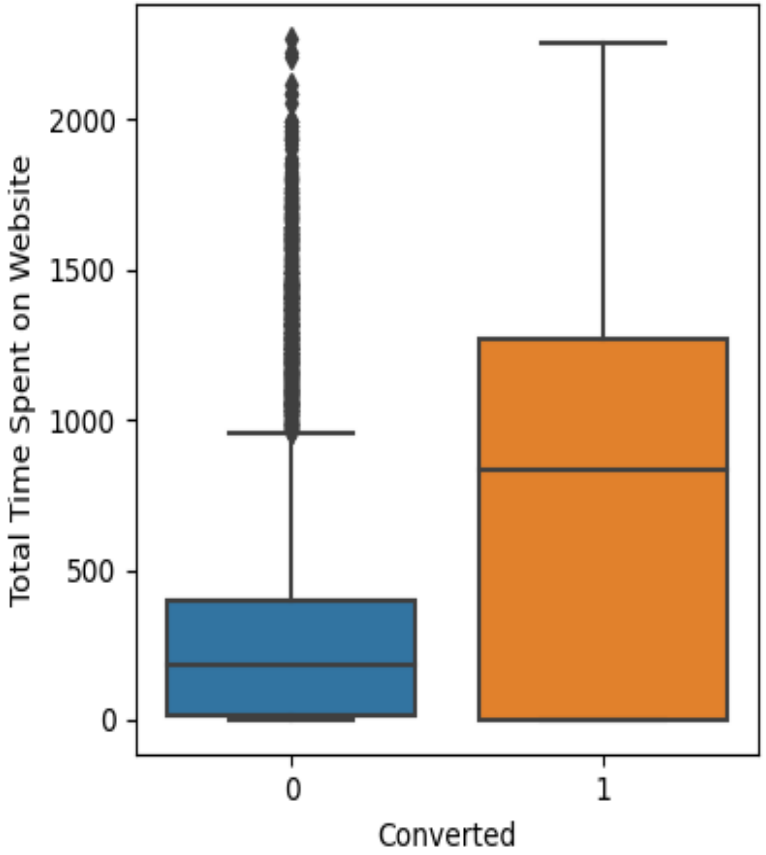

- Around 90% of the customers are unemployed with a LCR of 34%.
- while working professionals only make up 7.6% of the total customers with almost 92% LCR.

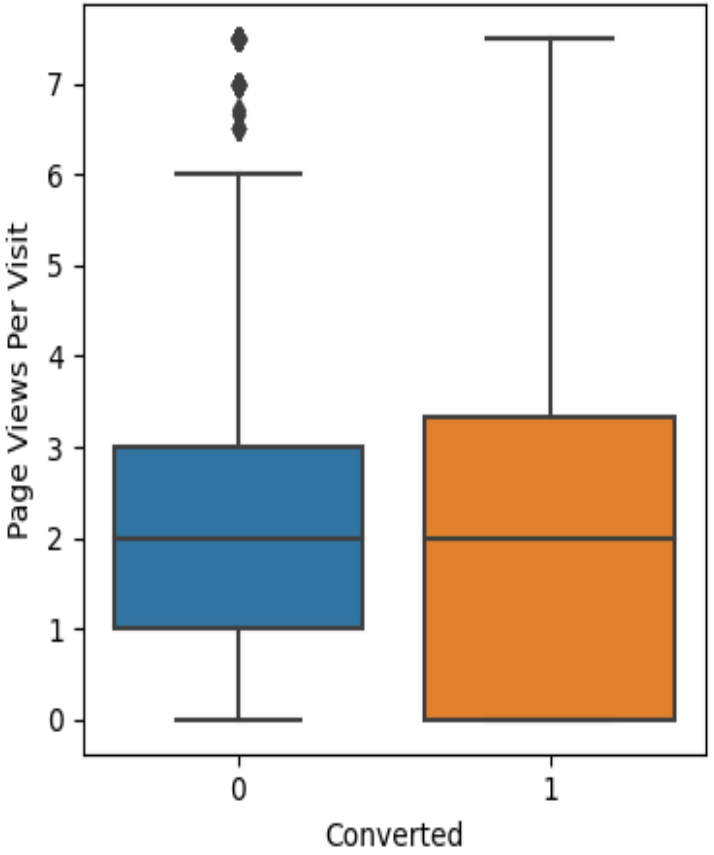


# Numerical variables vs target variable

- Customers who spent more ( >750sec )on the website are successfully converted to leads.
- There is a high chance if a customer visits above more than 3 times and views more than 3 pages are more likely to be leads.

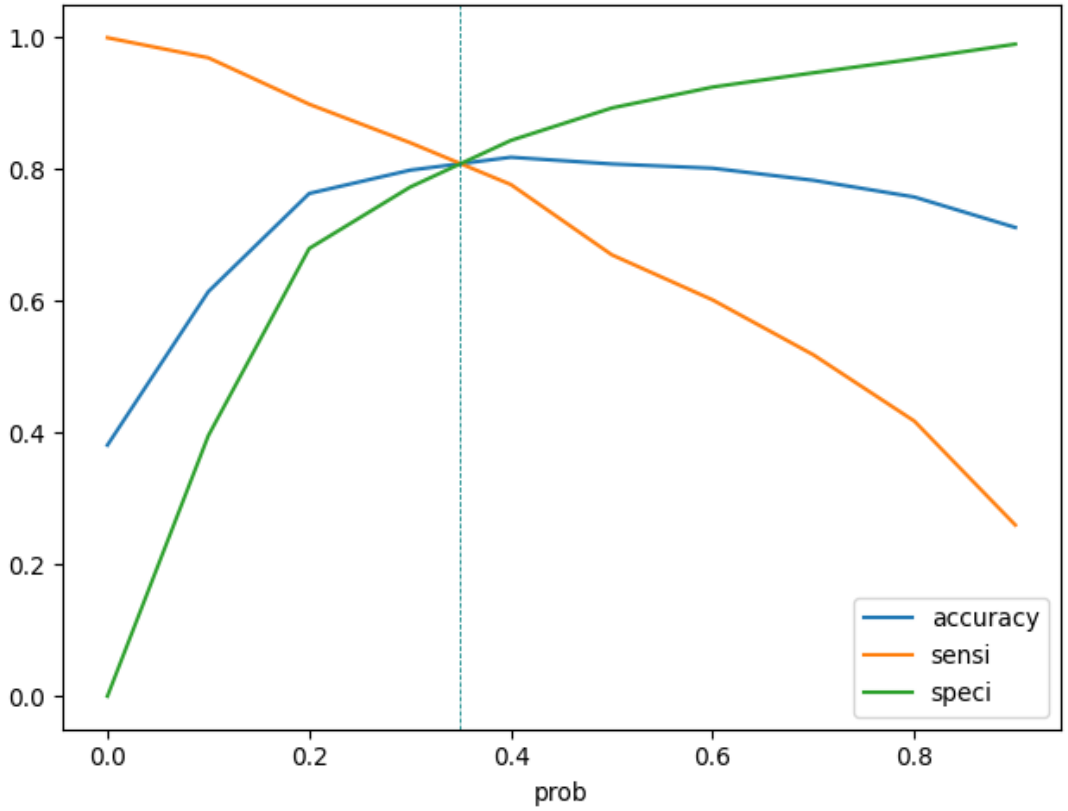


# Model Evaluation - Sensitivity and Specificity on Train Data Set

The graph depicts an optimal cut off is 0.35 based on Accuracy, Sensitivity and Specificity



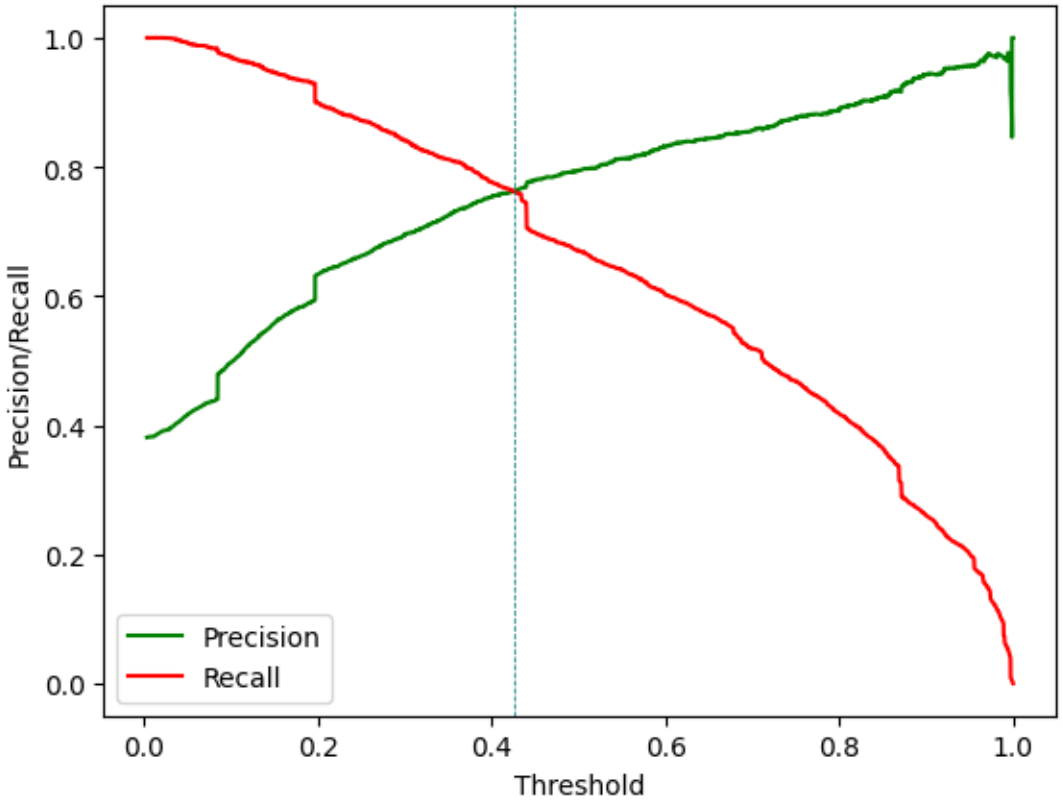

## Confusion Matrix

| | |
|---|---|
| 3242 | 760 |
| 465 | 2001 |

| | |
|---|---|
| True Negative: | 3242 |
| True Positive : | 2001 |
| False Negative : | 465 |
| False Positive : | 760 |
| Model Accuracy: | 0.8106 |
| Model Sensitivity: | 0.8114 |
| Model Specificity: | 0.8101 |
| Model Precision: | 0.7247 |
| Model Recall: | 0.8114 |
| Model True Positive Rate (TPR): | 0.8114 |
| Model False Positive Rate (FPR) : | 0.1899 |

# Model Evaluation- Precision and Recall on Train Dataset

The graph depicts an optimal cut off is 0.425 based on Accuracy, Sensitivity and Specificity




## Confusion Matrix

| | |
|---|---|
| 3417 | 585 |
| 586 | 1880 |

| | |
|---|---|
| True Negative: | 3417 |
| True Positive: | 1880 |
| False Negative: | 586 |
| False Positive: | 585 |
| Model Accuracy: | 0.819 |
| Model Sensitivity: | 0.7624 |
| Model Specificity : | 0.8538 |
| Model Precision : | 0.7627 |
| Model Recall: | 0.7624 |
| Model True Positive Rate (TPR): | 0.7624 |
| Model False Positive Rate (FPR): | 0.1462 |

- The precision-recall threshold cut-off of 0.425 yielded around 76% for True Positive Rate, Sensitivity, and Recall, while we need it to be close to 80% for our business objective.

- The sensitivity-specificity cut-off threshold of 0.35, on the other hand, achieved the required 80%. Therefore, we will consider the sensitivity-specificity view to determine the optimal cut-off for final predictions.

# Model Evaluation – Sensitivity and Specificity on Test Dataset

## Confusion Matrix

| | |
|---|---|
| 1354 | 323 |
| 223 | 872 |

| | |
|---|---|
| True Negative: | 1354 |
| True Positive: | 872 |
| False Negative: | 223 |
| False Positive: | 323 |
| Model Accuracy: | 0.803 |
| Model Sensitivity: | 0.7963 |
| Model Specificity: | 0.8074 |
| Model Precision: | 0.7297 |
| Model Recall: | 0.7963 |
| Model True Positive Rate (TPR): | 0.7963 |
| Model False Positive Rate (FPR): | 0.1926 |

# CONCLUSION

We have determined the following features that have the positive coefficients, and these features should be given priority in our marketing and sales efforts to increase lead conversion.

- Total Time Spent on Website
- Lead Origin in Lead Add Form
- Current Occupation in Working Professional
- Lead Source in Welingak Website
- Last Activity in SMS Sent
- Last Activity in Others
- Lead Source in Olark Chat
- Last Activity in Email Opened
- Total Visits

We have also identified features with negative coefficients that may indicate potential areas for improvement. These include:

- Specialization in Hospitality
- Do Not Email
- Lead Origin in Landing Page Submission
- Specialization in Others

# Recommendation

To increase our Lead Conversion Rates:

- Focus on features with positive coefficients for targeted marketing strategies.
- Prioritize calling leads who spend more time on the website.
- Tailor messaging to engage working professionals with higher conversion rates.
- Increase budget for Welingak Website advertising to attract more leads.
- SMS and email communication have higher conversion rates.
- Optimize communication channels based on lead engagement impact..

# THANK YOU