Lead Scoring Case Study

(Optimizing Lead Conversion: A Data-Driven Approach)

- Building a logistic regression model to assign lead scores between 0 and 100 to potential leads.
- Addressing additional problems provided by the company and adjusting the model accordingly.
- Documenting recommendations incorporating the logistic regression model and handling potential future changes.

Submitted by : (Ayushi Rathore, Jyothirmay Barua and Ishika Bansal)

Contents

- Problem statement
- Solution Strategy
- EDA
- Correlation analysis
- Model Evaluation
- Observations and Insights
- Conclusion

Problem Statement :

- An X Education needs assistance in identifying the most promising leads for conversion into paying customers.
- Development of a lead scoring model to differentiate leads based on conversion likelihood.
- Target lead conversion rate set at approximately 80% by the CEO.

Goals and Business Objectives:

1. Build a logistic regression model to assign lead scores between 0 and 100 for each lead.
- Higher scores indicate a higher likelihood of conversion (hot leads), while lower scores signify lower conversion chances (cold leads).

2. Adaptability to address future changes in company requirements.
- Ability to adjust the model based on evolving needs provided in a separate document.
- Ensure inclusion of adjustments in the final presentation for recommendations.

Solution Strategy

- **Data Import and Inspection:**
- **Importing and examining the dataset.**

- **Data Preparation:**
- **Preparing the data for analysis.**

- **Exploratory Data Analysis (EDA):**
- **Conducting exploratory analysis to gain insights.**

- **Dummy Variable Creation:**
- **Creating dummy variables as needed.**

- **Test-Train Split:**
- **Splitting the data into training and testing sets.**

- **Feature Scaling:**
- **Scaling the features for consistency.**

- **Correlation Analysis:**
- **Analyzing correlations between variables.**

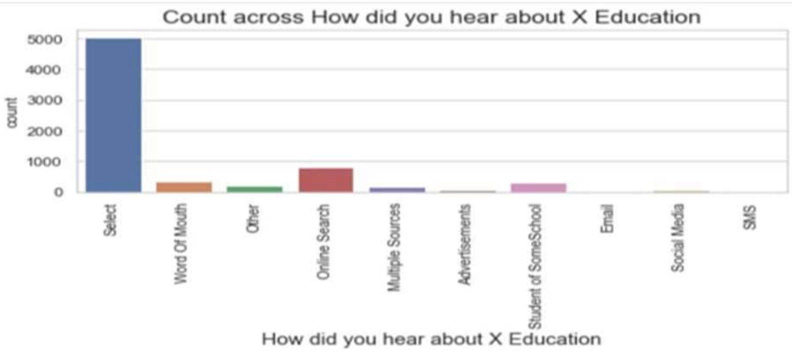
- **Model Building:**
- **Utilizing Recursive Feature Elimination (RFE) alongside R-squared, VIF, and p-values for model selection.**

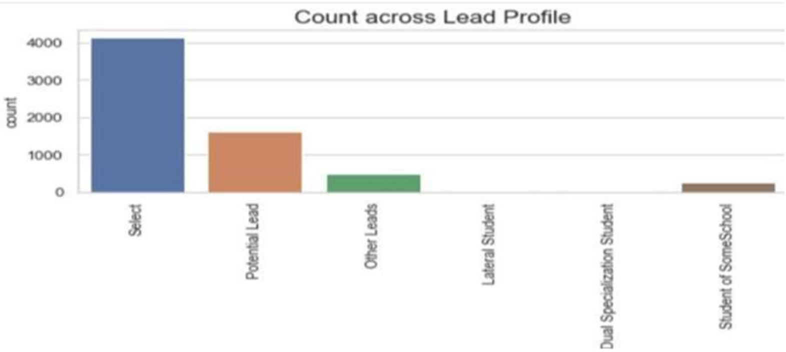
- **Model Evaluation:**
- **Evaluating the performance of the built model.**

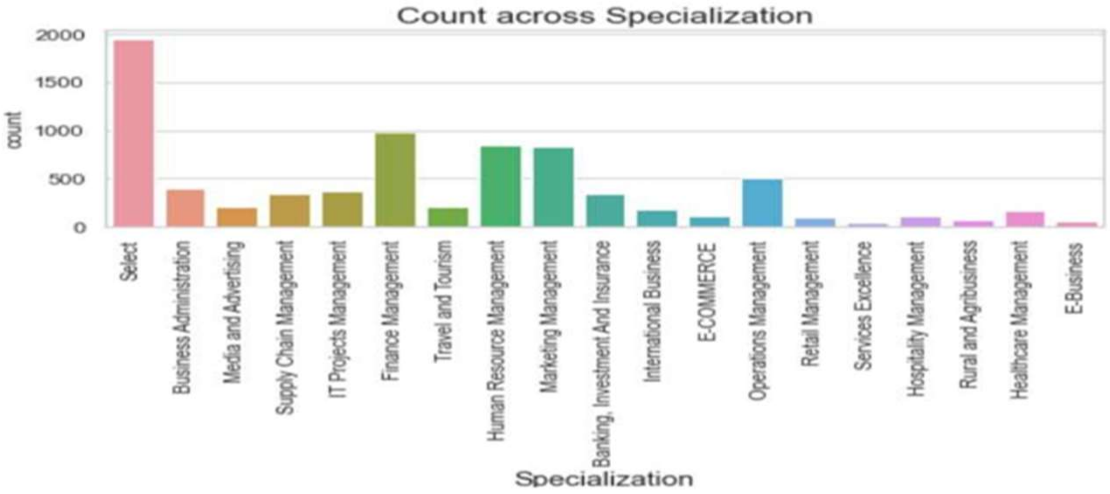
- **Prediction on Test Set:**
- **Making predictions on the test dataset.**

EDA – Data Cleaning

Handling the 'Select' Level in Certain Columns: Managing and addressing the presence of the 'Select' level within specific columns.

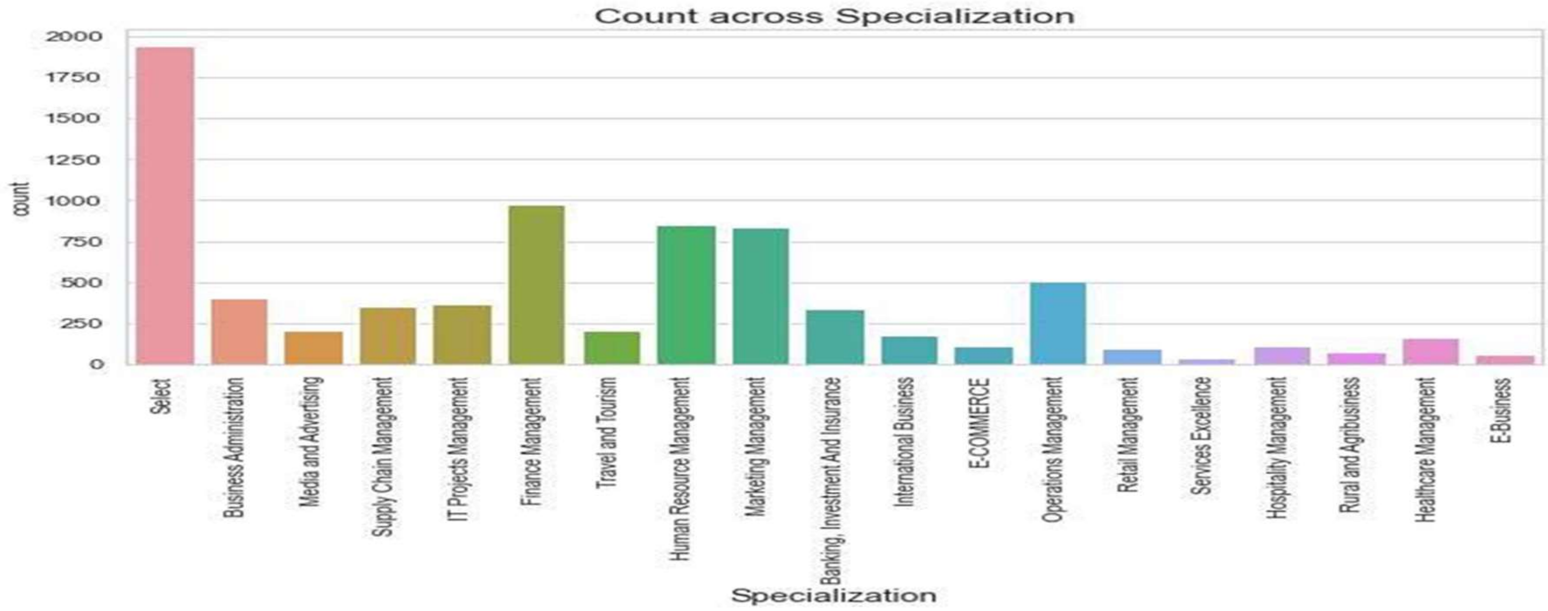






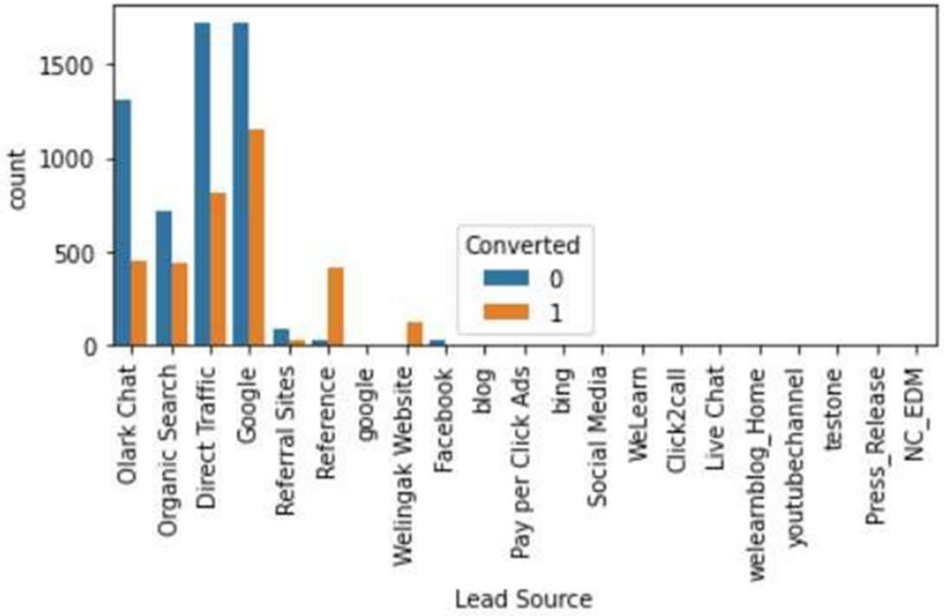
Specialization

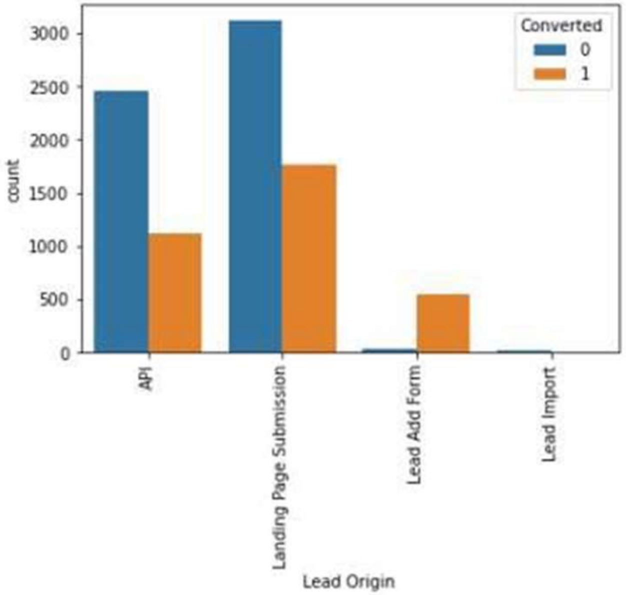
Leads originating from HR, Finance, and Marketing management specializations display a significantly heightened probability of converting into paying customers.



Lead Source & Lead origin

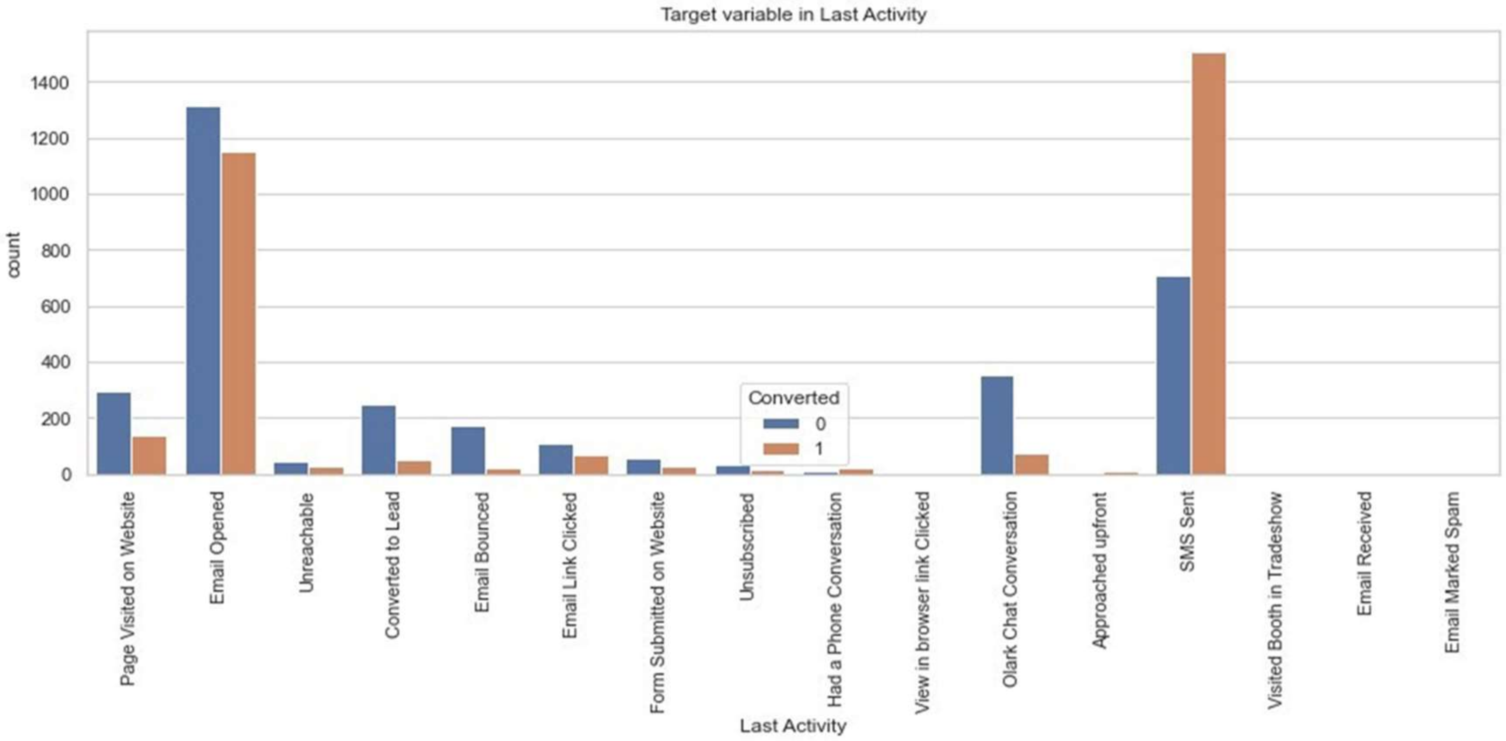
Leads originating from distinguished sectors such as HR, Finance, and Marketing management demonstrate a notably elevated tendency to convert into paying customers, showcasing a substantial potential for conversion.





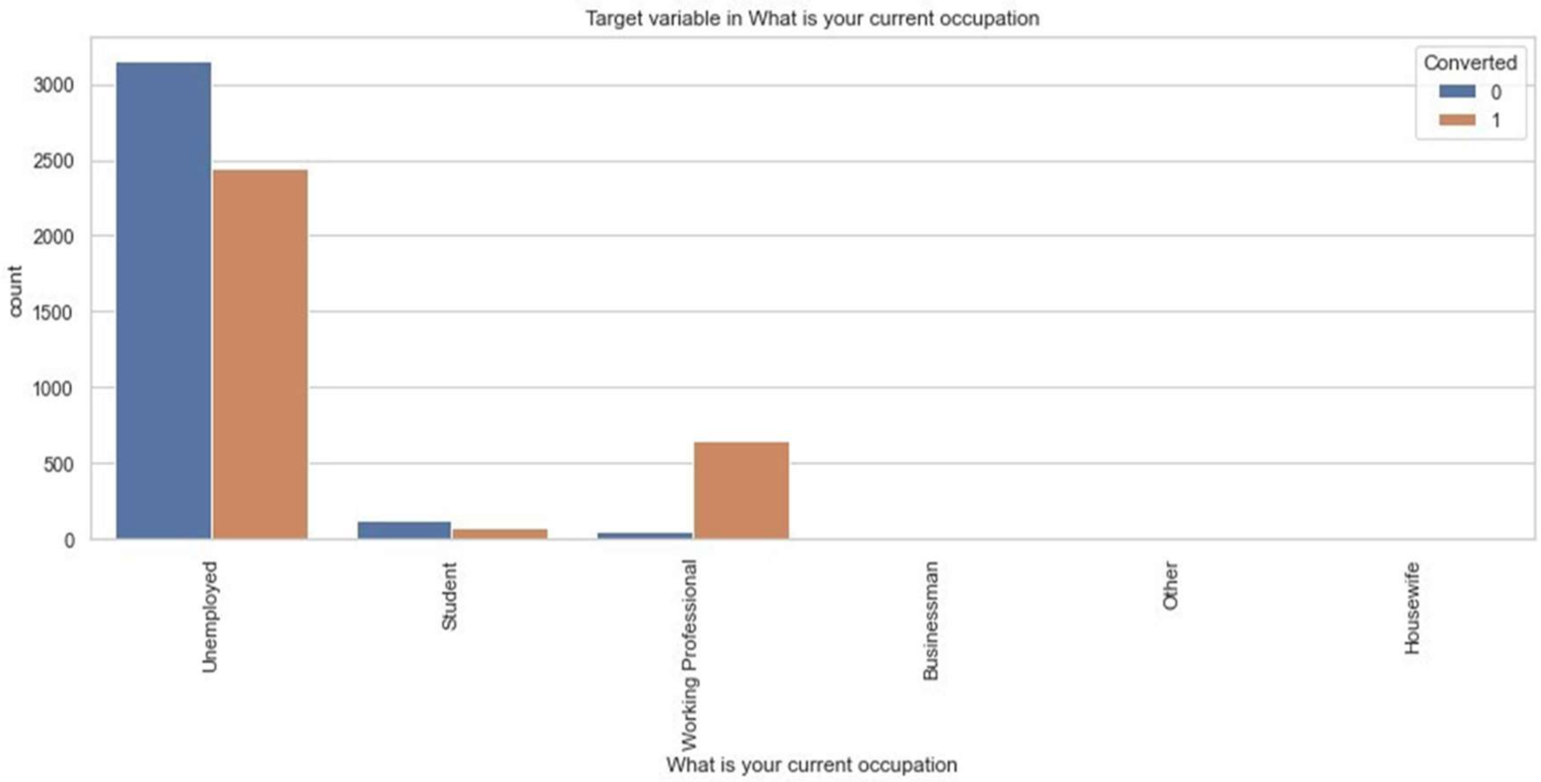
Last lead Activity

Leads that engage by opening emails show a heightened probability of conversion, while the likelihood of conversion is similarly enhanced by sending SMS communications to these leads.



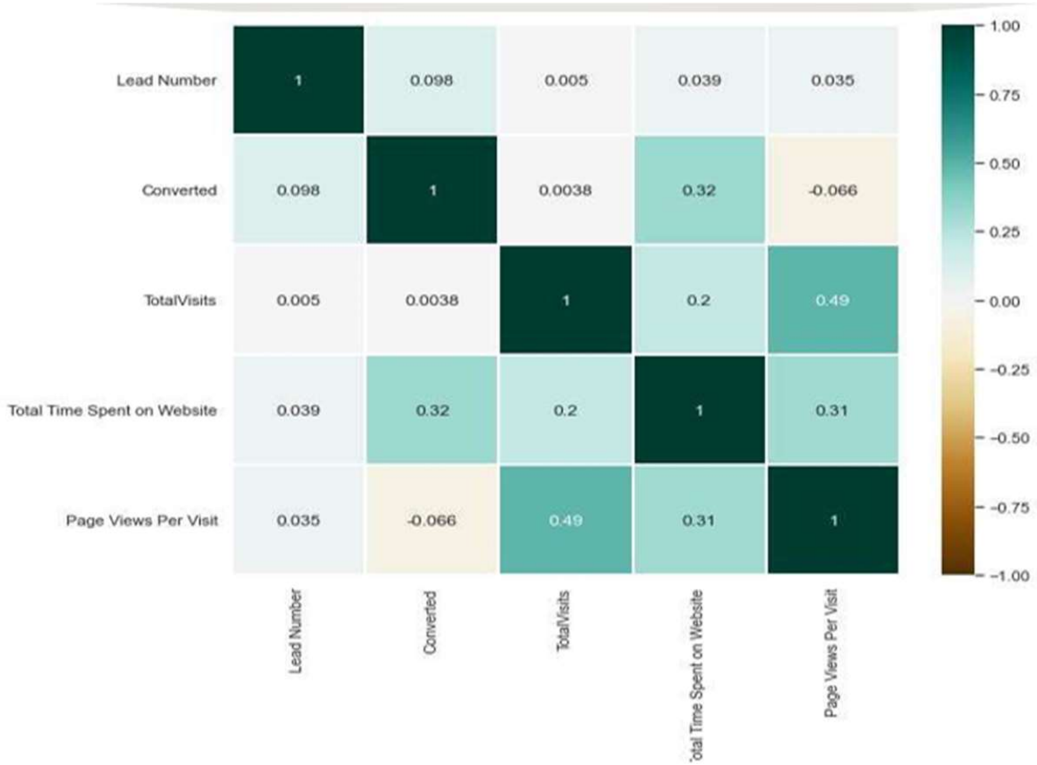
Last What is Your Occupation

Leads categorized as unemployed display a greater level of interest in enrolling in the course compared to leads in other employment statuses.



Correlation analysis

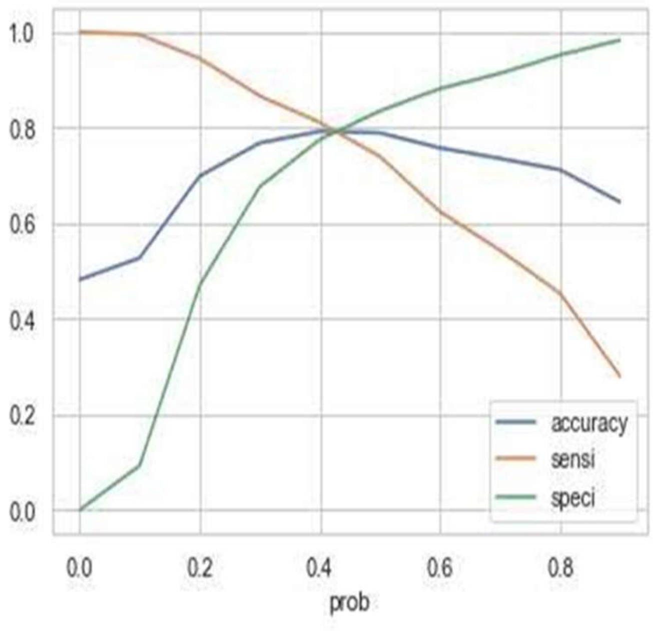
There is a lack of correlation observed among the variables under study, indicating independence and non-association between them.

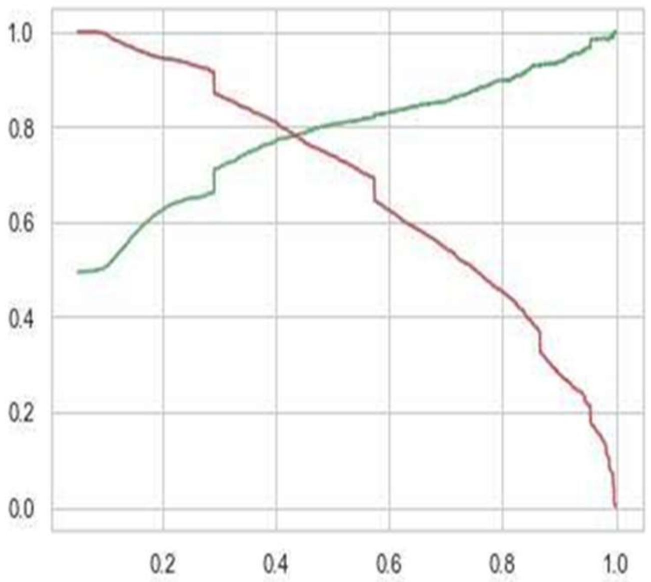


	Lead Number	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit
Lead Number	1	0.098	0.005	0.039	0.035
Converted	0.098	1	0.0038	0.32	-0.066
TotalVisits	0.005	0.0038	1	0.2	0.49
Total Time Spent on Website	0.039	0.32	0.2	1	0.31
Page Views Per Visit	0.035	-0.066	0.49	0.31	1

Model Evaluation (ROC Curve)

Given a tradeoff value of 0.42 between Precision and Recall, it is advisable to classify Prospect Leads with a Conversion Probability surpassing 42% as prime candidates, representing a significant likelihood of conversion.





Observations and Insights

In the training dataset:

- Accuracy is at 80%, indicating the overall correctness of the model.
- Sensitivity stands at 77%, reflecting the model's ability to correctly identify positive instances.
- Specificity is recorded at 80%, illustrating the model's capacity to accurately pinpoint negative instances.

In the testing dataset:

- The accuracy rate is 80%, demonstrating the overall correctness of predictions made by the model.
- Sensitivity is noted at 77%, representing the model's ability to accurately identify positive instances.
- Specificity is at 80%, indicating the model's accuracy in recognizing negative instances..

The finalized features list includes:

- **Lead Source: Olark Chat**
- **Specialization: Others**
- **Lead Origin: Lead Add Form**
- **Lead Source: Welingak Website**
- **Total Time Spent on Website**
- **Lead Origin: Landing Page Submission**
- **Current Occupation: Working Professionals**
- **Do Not Email**
- **These selected features have been identified and considered essential for further analysis and modeling in the current context..**

Conclusion & Key Takeaways:

(Insights on Lead Conversion Analysis)

- The conversion rates for leads from API and Landing page submissions hover around 30% to 35%, in line with industry averages. Conversely, the conversion rates are notably lower for Lead Add form and Lead import, suggesting a need for increased focus on leads originating from API and Landing page submissions.

- The primary lead source is Google/direct traffic, with references and the Welingak website boasting the highest conversion ratios.

- A positive correlation is observed between the time spent by leads on the website and their likelihood to convert.

- The most prevalent activity among leads is opening emails, yet the highest conversion rate stems from SMS interactions. While the majority of leads are unemployed, the conversion rates are notably higher among working professionals.