Lead Scoring Case Study

By: Rupali Bansal

Problem Statement

1. X Education generates many leads through online courses for professionals but has a low lead conversion rate of around 30%.
2. The company seeks to identify the most promising leads to improve the efficiency of the sales team and increase conversions.
3. The goal is to assign a lead score to each lead so that the sales team can prioritize potential customers more effectively.
4. The CEO has set a target to raise the lead conversion rate to around 80%.

Business Objective:

1. Build a logistic regression model to predict the likelihood of lead conversion.
2. Assign a lead score between 0 and 100, where a higher score means a higher likelihood of conversion.
3. Optimize the sales team's focus on high-potential leads to improve the conversion rate and reduce unnecessary efforts.
4. Recommend actionable insights based on the model to maximize lead conversion.

Solution Methodology

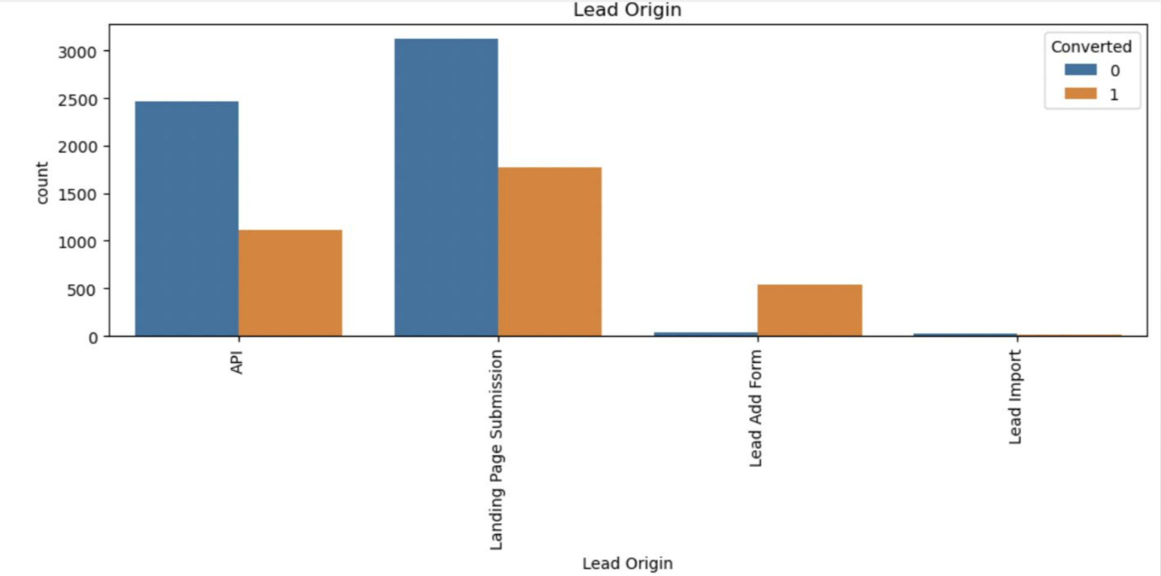
1. Data Cleaning and manipulation
2. Exploratory Data Analysis (EDA)
   - Univariate data analysis
   - Bivariate data analysis
3. Feature Engineering
   - Dummy Variable Creation
   - Feature Scaling
4. Model Building
5. Model Evaluation
6. Making Predictions on the test data

Data Cleaning

1. Converted 'Select' levels in categorical variables to null as they hold no useful information.
2. Checked for missing values and dropped records with low missing percentages (e.g., 'Lead Source', 'Total Visits').
3. Dropped columns with over 40% missing data to ensure data quality.
4. Removed unnecessary features like 'Prospect ID' and 'Lead Number', which didn’t contribute to the model.
5. Dropped highly skewed columns (e.g., 'What matters most to you in choosing a course').
6. Imputed missing values for key columns:
   - Replaced missing 'Country' values with the mode ('India').
   - For columns like 'Specialization' and 'City', filled missing values with 'Unknown'.
7. Converted appropriate columns from float to integer types for consistency.
8. Dropped columns with only one unique value or where a single value dominated the data.

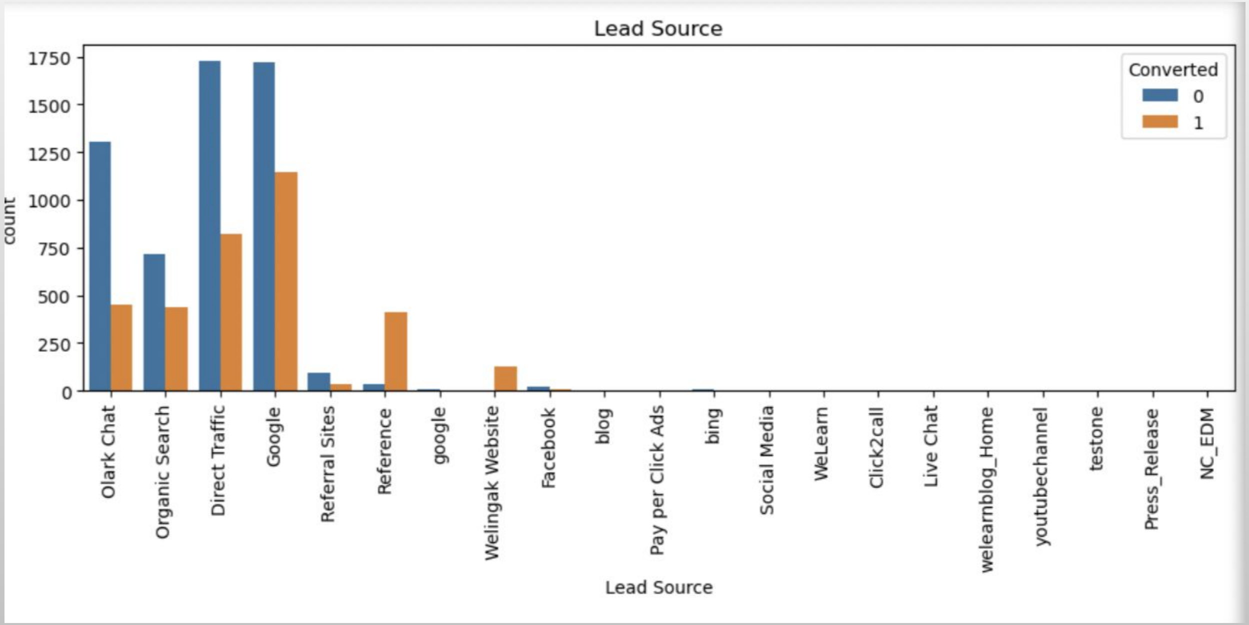
Exploratory Data Analysis (EDA)

- Analyzed distribution of features like lead source, lead origin, and website activity.
- I identified and removed outliers in crucial variables like 'Total Visits' and 'Page Views Per Visit'. Outliers can skew results and affect the model's performance, so cleaning them up was essential.
- Some columns, such as 'Do Not Email' and 'Tags', didn’t provide much useful information for the model. I decided to drop these features to keep the dataset clean and focused.



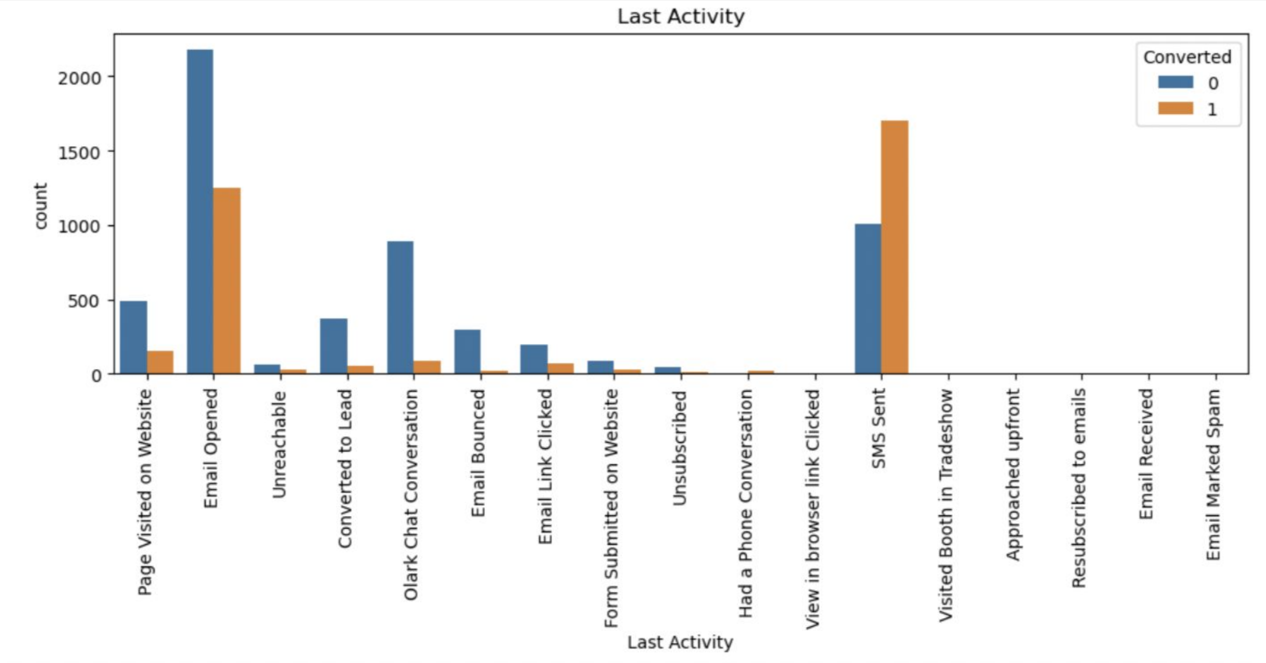
1. Leads from API and Landing Page Submissions have a 30-35% conversion rate, but their volume is relatively high.
2. In contrast, leads from Lead Add Forms boast a conversion rate of over 90%, though they are fewer in number.
3. Leads imported through Lead Import are quite rare.

To boost lead conversion, improve strategies for high-volume sources like API and Landing Page Submission, and increase leads from the high-conversion Lead Add Form.

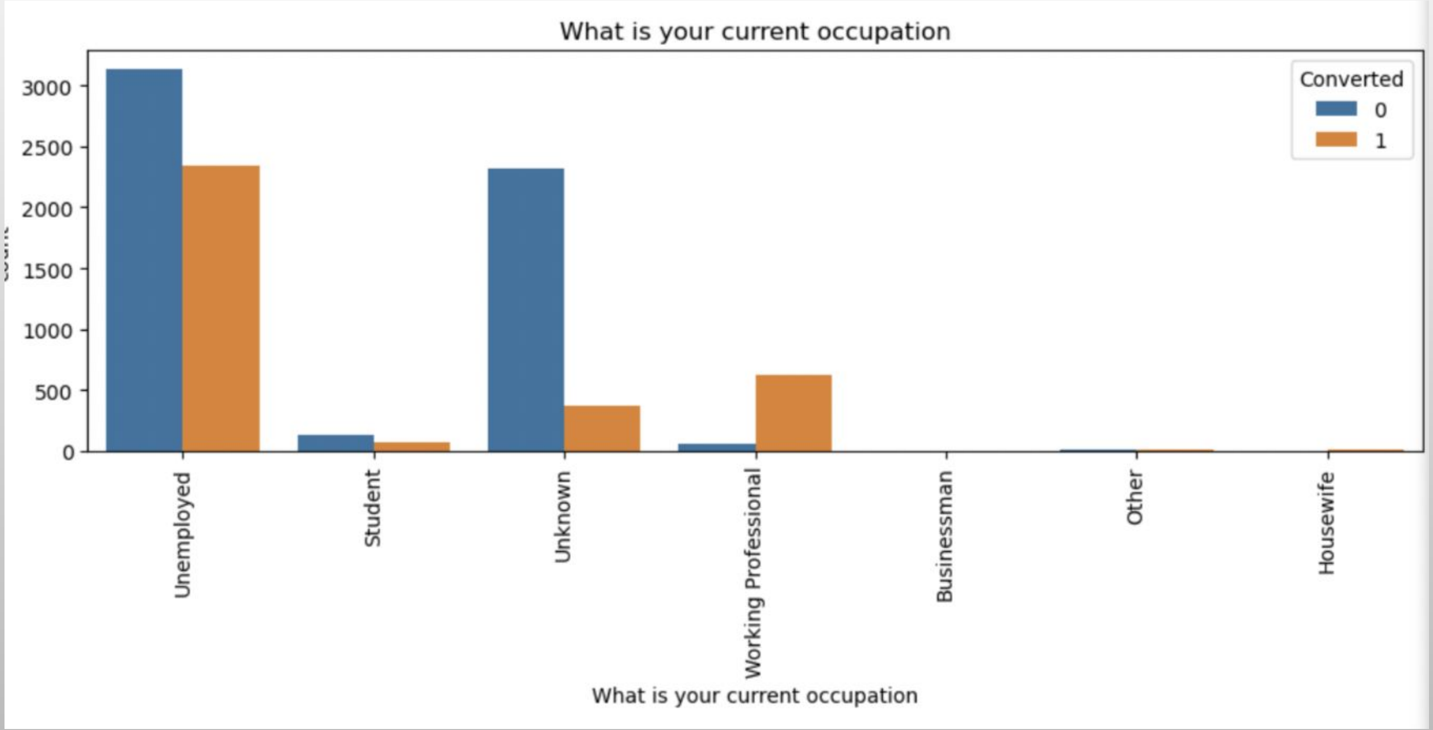


1. High Lead Volume: Google and Direct Traffic are the top sources for generating the most leads.
2. High Conversion Rates: Leads from referrals and the Welingak website have higher conversion rates.

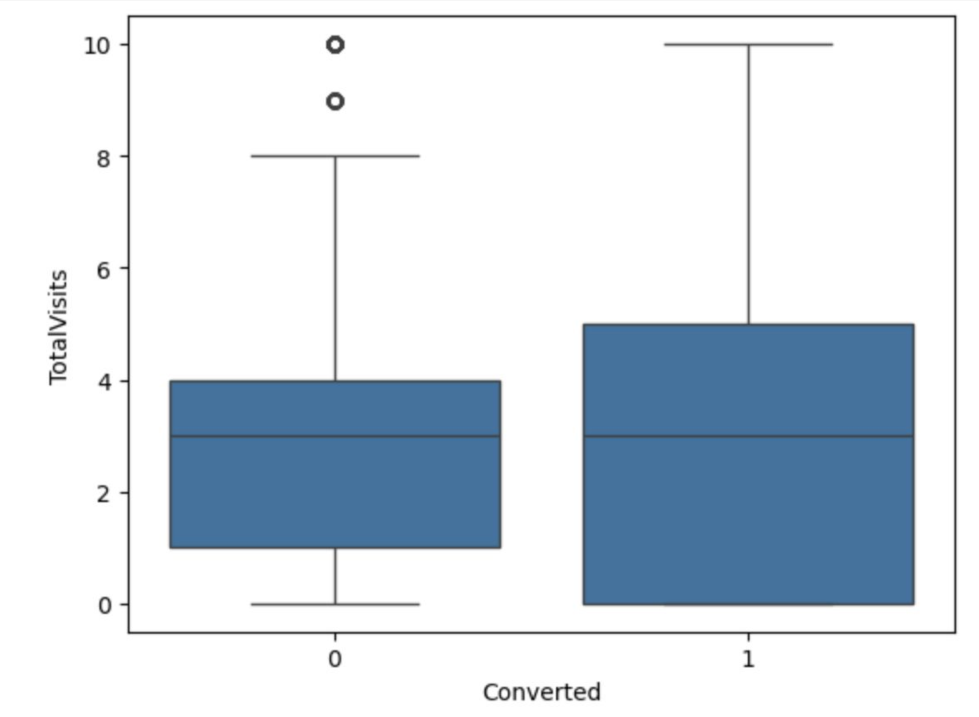
To boost the overall lead conversion rate, focus on improving conversions from Olark Chat, Organic Search, Direct Traffic, and Google leads, while increasing lead generation from Referrals and the Welingak website.



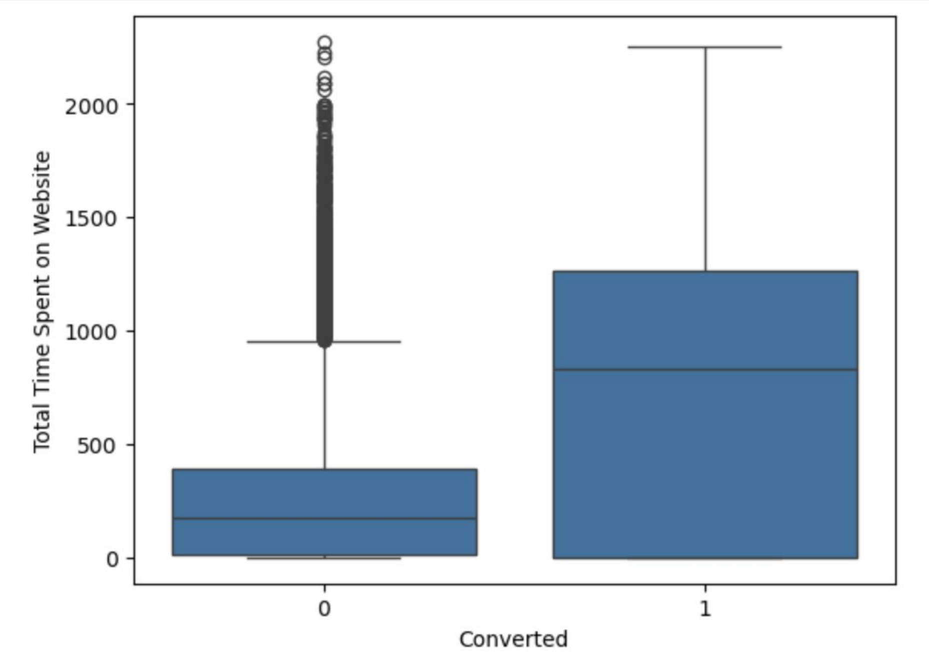
Most leads have "Email Opened" as their last activity, while those with "SMS Sent" as their last activity have a conversion rate of nearly 60%.



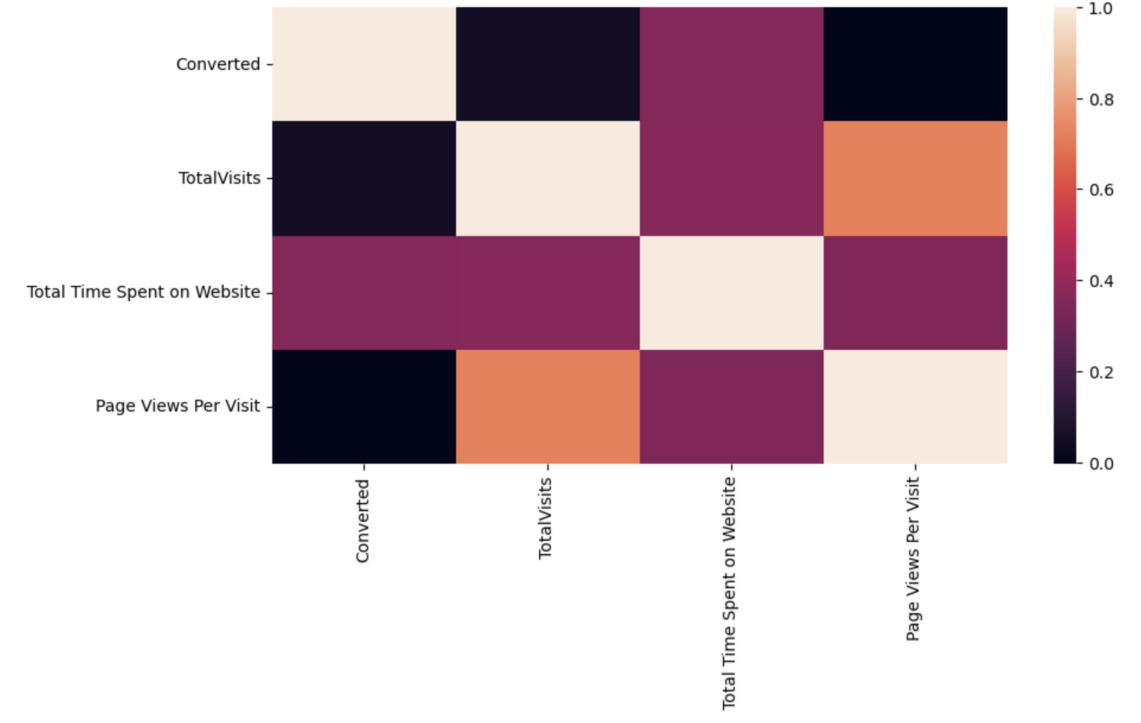
1. High Conversion Potential: Working Professionals have a higher likelihood of enrolling in the course.
2. Low Conversion Rate: Unemployed leads, although numerous, have a conversion rate of only 30-35%.



The median in total visits are identical for both converted and non-converted leads. No observations here



Leads who spend more time on the website are more likely to be converted.



There is no correlation between variables

Feature Engineering

1. Created Dummy Variables: Transformed categorical data into numerical format

2. Test-Train Split: Divided the dataset into training and testing sets to evaluate model performance.

3. Scaling the Train Dataset: Standardized or normalized the training data to ensure consistent feature scaling for improved model accuracy.

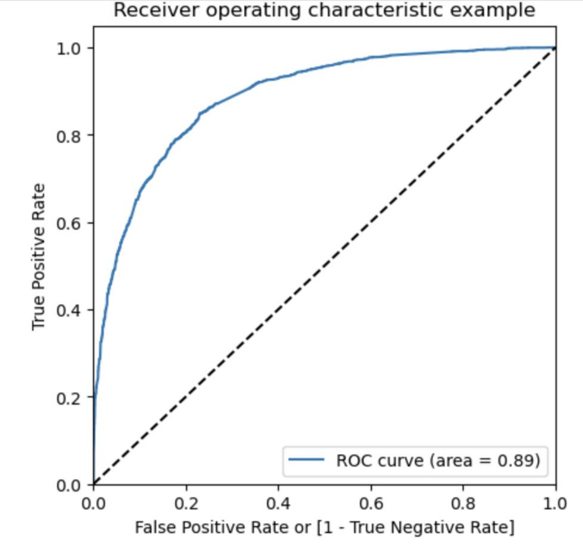
Model Building

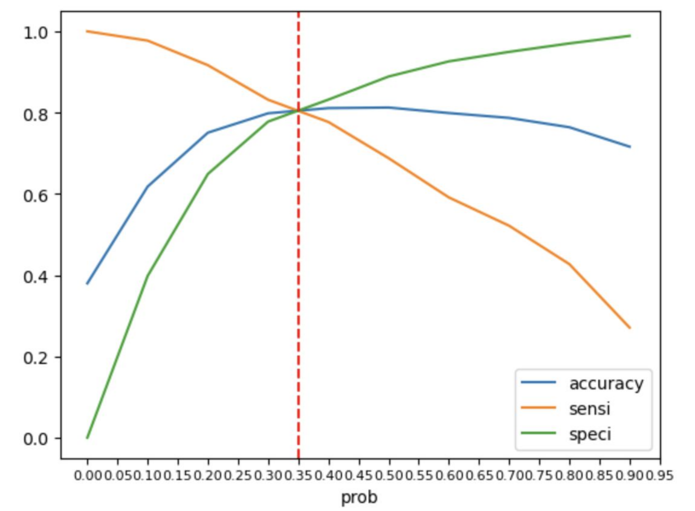
1. Feature Selection:
   - Used Recursive Feature Elimination (RFE) to identify the most relevant 15 features for the model.

2. Model Construction:
   - Built a logistic regression model using the selected features.

3. Handling Multicollinearity:
   - Removed features with high p-values and high Variance Inflation Factors (VIF) to address multicollinearity.

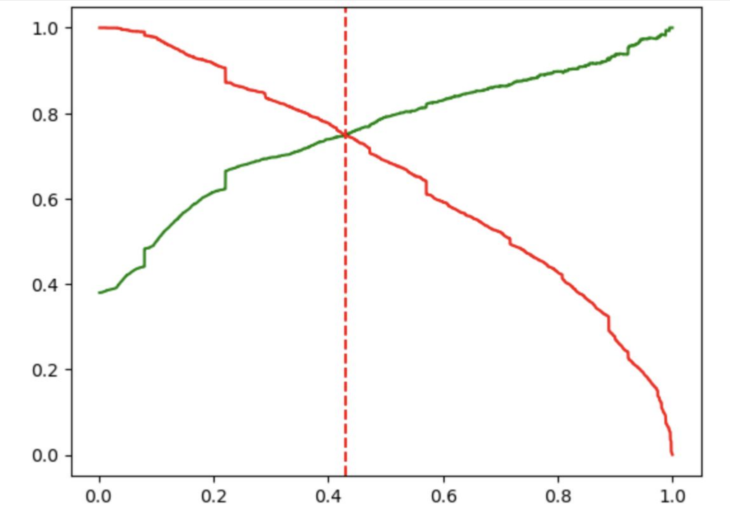
Model Evaluation - ROC curve





- The area under the ROC curve (AUC) is 0.89 which indicates the model's overall performance, with a higher AUC reflecting better model accuracy.
- Optimal Cutoff Point came as 0.35 where balanced sensitivity and specificity can be achieved.

Precision and recall tradeoff



Adjusted the classification threshold to 0.43, balancing precision and recall. Therefore, a prospect lead with a conversion probability higher than 43% can be considered a "hot lead."

Observations

Final Features list:

- Last Notable Activity_Had a Phone Conversation
- Lead Origin_Lead Add Form
- Lead Source_Welingak Website
- What is your current occupation_Working Professional
- Last Notable Activity_SMS Sent
- Total Time Spent on Website
- Lead Source_Olark Chat
- Lead Origin_Landing Page Submission

Train Dataset

- Accuracy : 80.8%
- Sensitivity : 74.7%
- Specificity : 84.6%

Test Dataset:

- Accuracy : 81%
- Sensitivity : 74.8%
- Specificity : 84.7%

Conclusion

Here are some actionable insights and recommendations based on the model:

1. **Focus on High-Conversion Sources:** Prioritize leads from “Lead Add Form”, “Welingak Website” and “Referrals”, as they have higher conversion rates. Increase marketing and lead generation efforts for these sources to capitalize on their high conversion potential.

2. **Refine Follow-Up Tactics:** Tailor follow-up communications based on lead activity, such as prioritizing leads with Had a Phone Conversation and SMS Sent as their last activity, which shows higher conversion potential.

3. **Target High-Value Occupations:** Concentrate efforts on “Working Professionals” who show a higher likelihood of enrolling. Adjust your marketing strategies to appeal more to this demographic.

4. **Enhance Engagement for Key Channels:** Optimize conversion strategies for “Olark Chat” and “Landing Page Submission” leads. Since these channels are important for lead generation, ensure they are effectively managed and supported with strong conversion tactics.

5. **Monitor Time Spent on Website:** Leads spending more “Total Time Spent on Website” are more likely to convert. Use this insight to identify and engage with high-interest leads who are investing time in exploring your offerings.

By focusing on these actionable insights, you can better allocate resources, improve lead engagement, and ultimately increase the overall lead conversion rate.