Lead score case study

Logistic Regression

By,

Venkata Pavan Kumar Alapati
Bhavitha Bojja
Lavanya Koripadu

Problem statement

X Education is a company that offers online courses tailored for professionals in various industries. The courses offered by the company are promoted across various websites and search engines, including Google.

Despite generating a substantial number of leads, X Education struggles with a low lead conversion rate. X Education typically experiences a lead conversion rate of approximately 30% through the whole process of turning leads to customers by approaching them.

In order to enhance efficiency, the company aims to pinpoint the most promising prospects referred to as 'Hot Leads'. By successfully discerning this group of leads, the conversion rate is anticipated to increase. The implementation process for lead generation attributes lacks effectiveness in contributing to conversions.

Business Goal

The company requires to build a model.

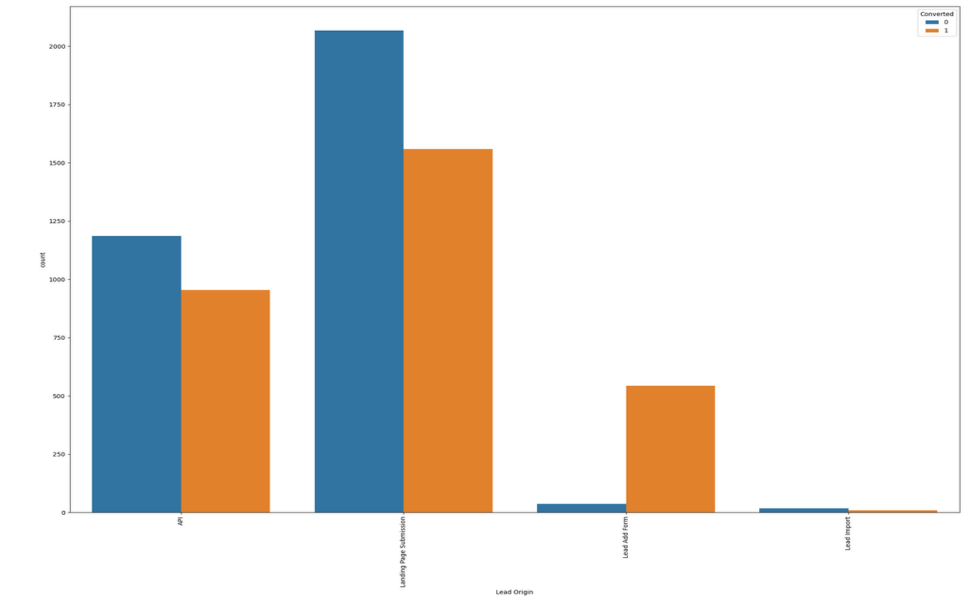
Allocate a lead score to each lead, ensuring that customers with higher scores are more likely to convert, while those with lower scores have a reduced likelihood of conversion.

Desired lead conversion rate is 80%.

Strategy

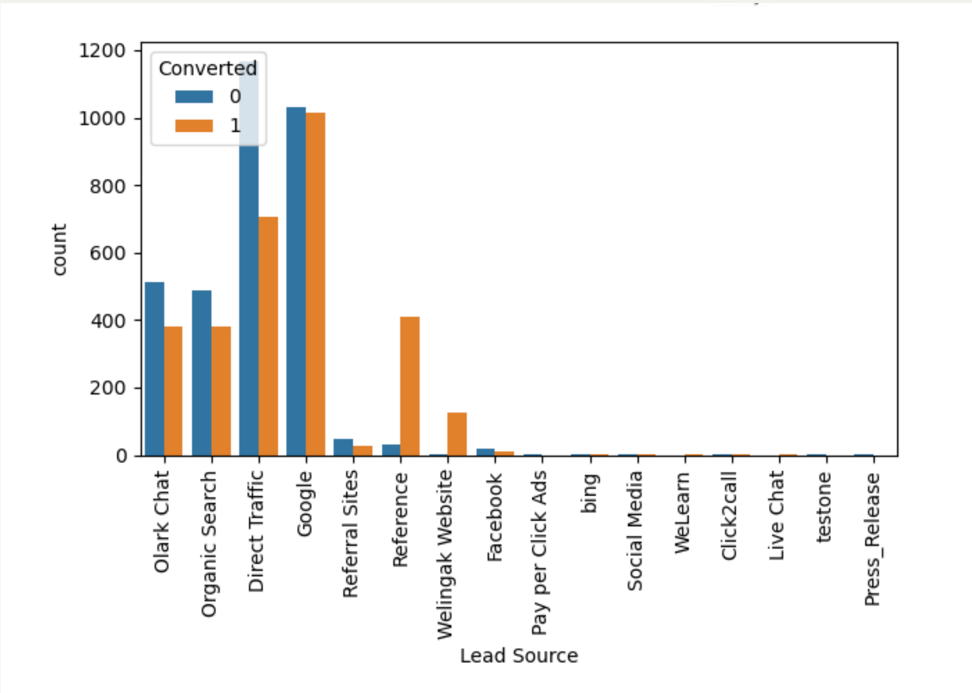
- Import Data
- Clean and prepare the data for further analysis
- Exploratory data analysis for selecting helpful attributes for conversion
- Selecting features
- Prepare the data for model building
- Build a logistic regression model
- Test the model on train set
- Evaluate model for different metrics
- Test the model on test set
- Measure the accuracy of model and other metrics for evaluation
- Assign a lead score for Hot leads

Exploratory Data Analysis



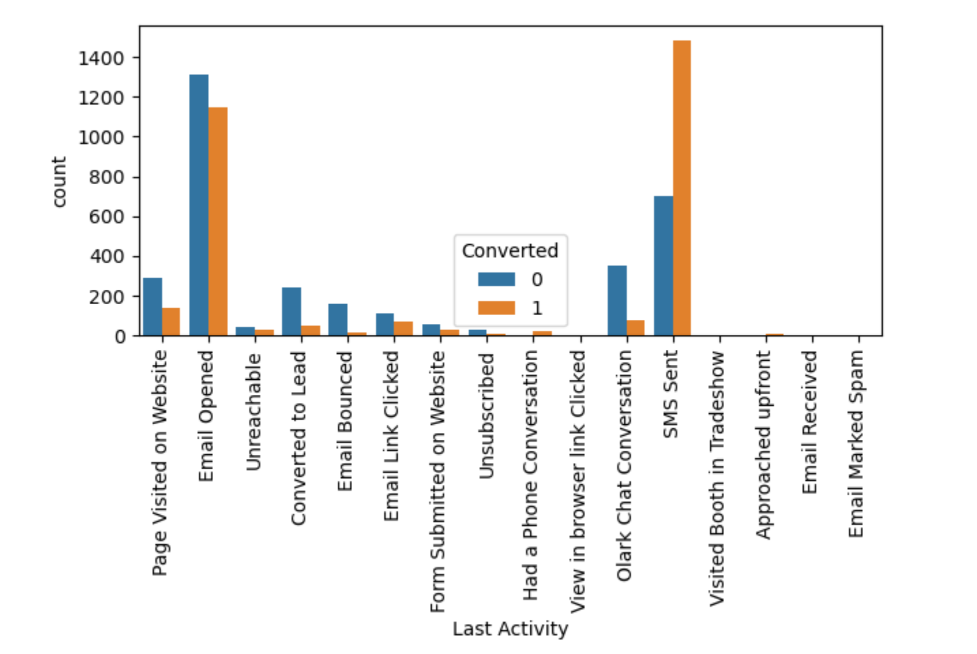
Lead Origin

'Landing Page Submission' has greatest source of audience, but conversion rate is higher in 'Lead Add Form'.



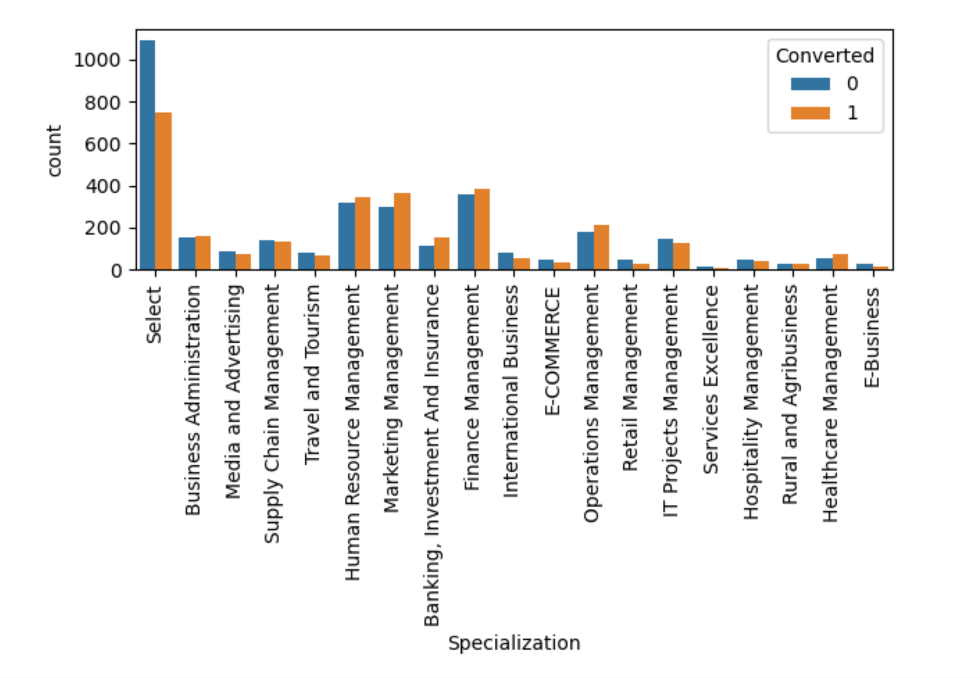
Lead Source

'Google' attracted more audience, but the 'Reference' and 'Welingak website' has higher conversion rate.



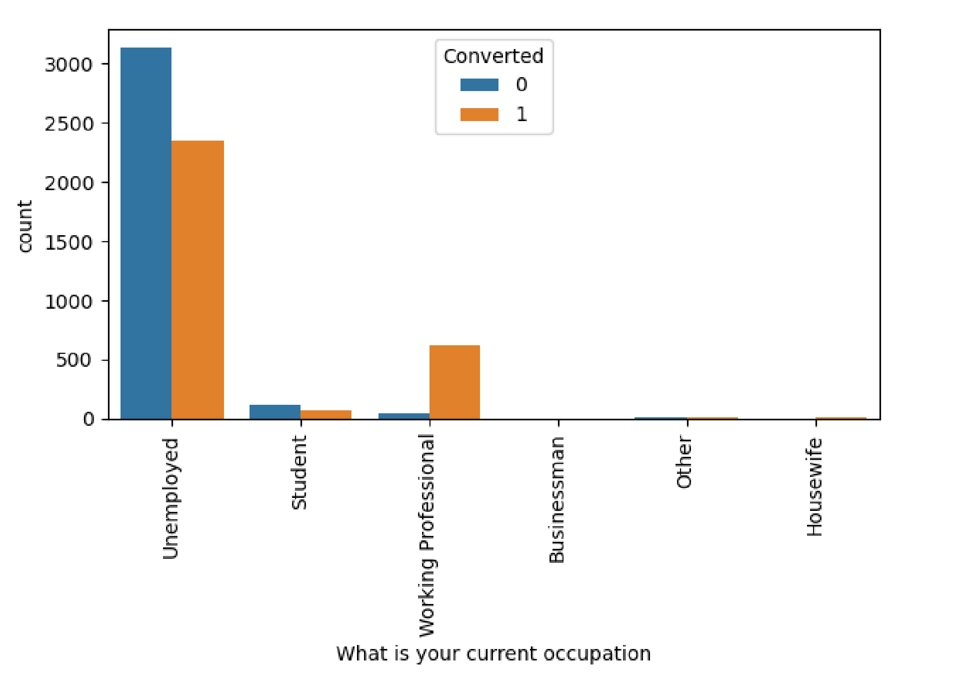
Last Activity

'Email Opened' has the highest range of customers, but the conversion rate is high in 'SMS Sent'.



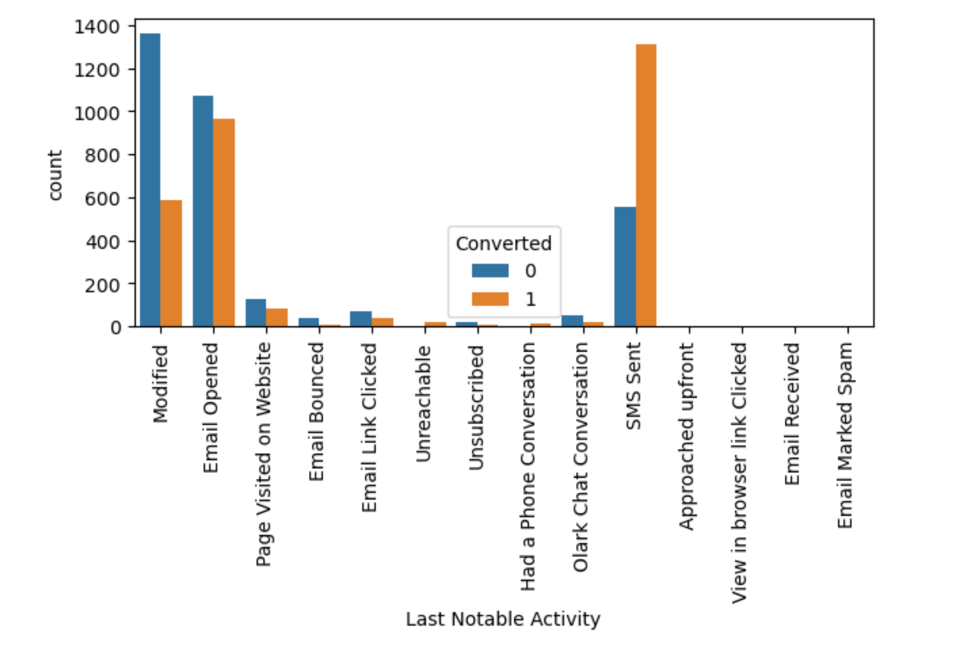
Specialization

Majority of people didn't select any specialization. so, from the remaining, we can see 'Marketing Management' has the highest conversion rate.



What is your current occupation

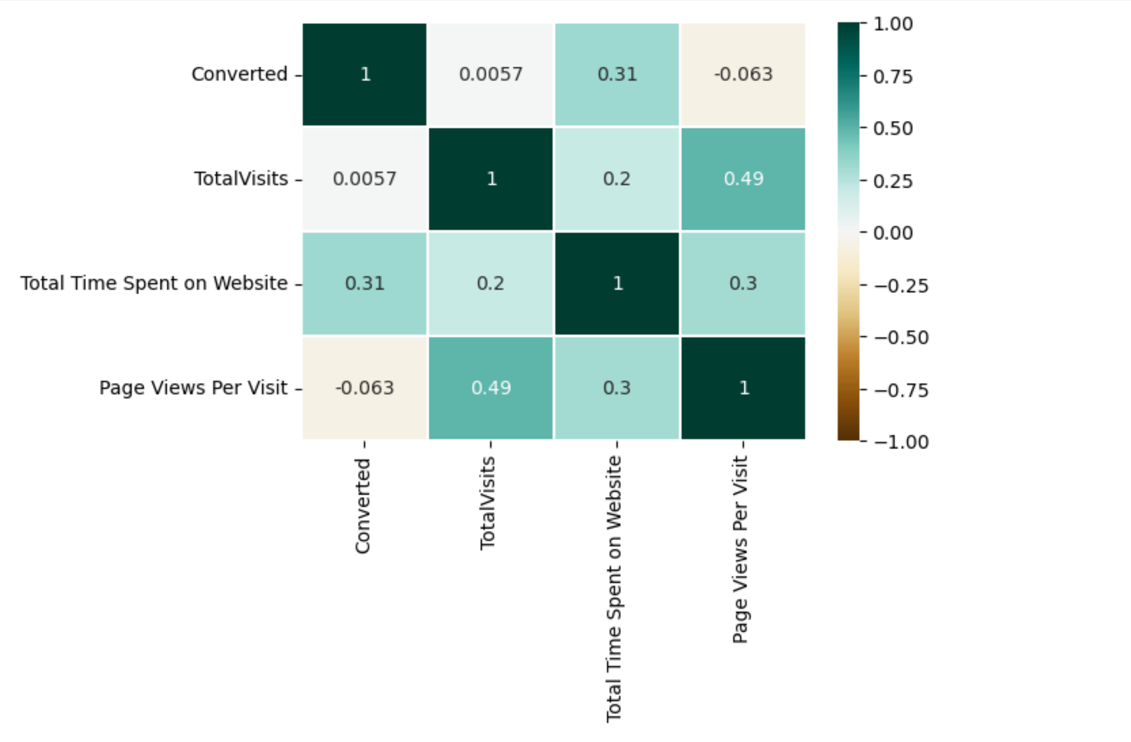
Students who are approaching are mostly 'Unemployed', but the high conversion rate is in 'Working Professionals'.



Last Notable Activity

Highest is seen in 'Email Opened', but the conversion rate is higher in 'SMS Sent'.

Heat Map



	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit
Converted	1	0.0057	0.31	-0.063
TotalVisits	0.0057	1	0.2	0.49
Total Time Spent on Website	0.31	0.2	1	0.3
Page Views Per Visit	-0.063	0.49	0.3	1

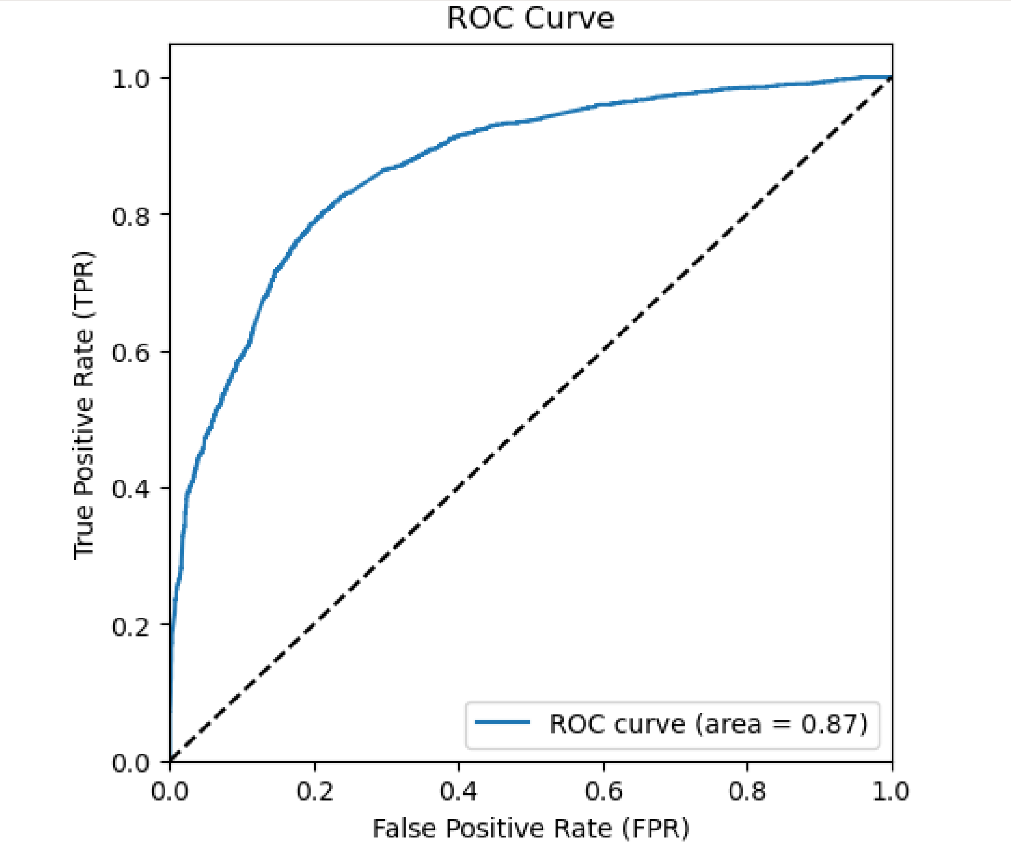
- `Page views Per Visit` and `Converted` has highest negative correlation of `-0.063`

- `Total Visits` and `Page Views Per Visit` shows the highest correlation of `0.49`

- `Total Time Spent on Website` and `Converted` shows` 0.31` correlation.

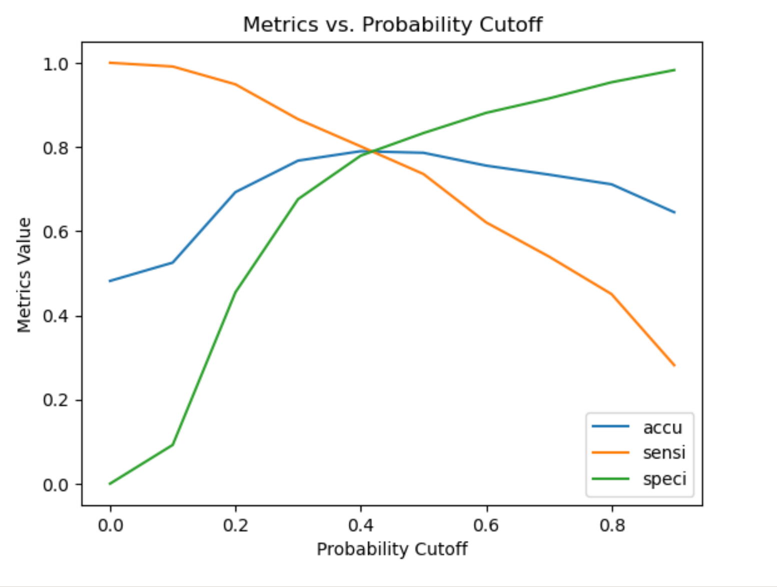
Model Building

- Splitting the data into Train and test sets
- Scaling the variables
- Build the first model
- Use RFE to eliminate less relevant variables
- Build the next model
- Eliminate variables based on p-values
- Check VIF value for all the existing columns
- Predict the train set
- Evaluate accuracy and other metrics
- Predict test set
- Precision and recall analysis



The ROC curve illustrates the trade-off between a binary classifier's true positive rate and false positive rate across varying classification thresholds, aiding performance evaluation. From the ROC curve, the area under curve is 0.87 therefore our model is good.

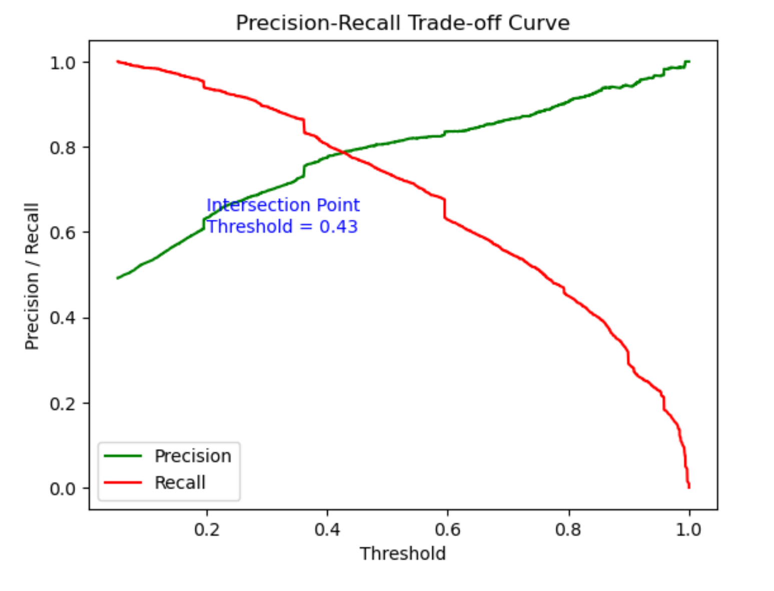
Model Evaluation – Train set



confusion matrix

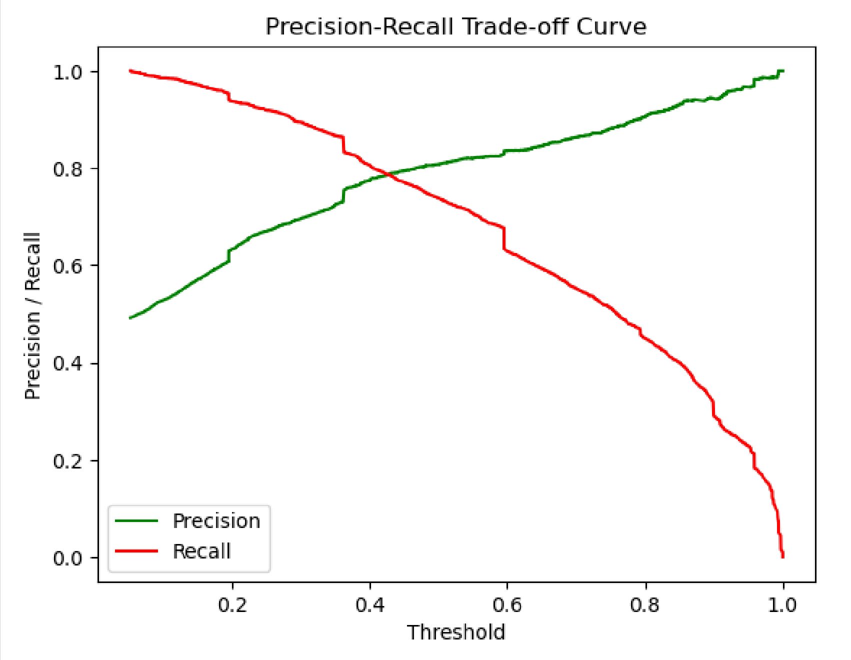
1859	453
461	1688

Accuracy = 79.5%
Sensitivity = 78.5%
Specificity = 80.4%



- Precision= 78.8%
- Recall = 78.5%

Model Evaluation – Test set



Test threshold has been set as 0.43

confusion matrix

805	191
205	711

- Precision= 78.8%
- Recall = 77.6%

Accuracy = 79.2%
Sensitivity = 77.6%
Specificity = 80.8%

Summary

1. Target High-Conversion Sources: Focus on leads originating from "Welingak Website" and "Reference" sources, as they have a higher likelihood of conversion.

2. Prioritize Working Professionals: Prioritize leads who are working professionals, as they show a higher tendency to convert into customers.

3. Engage High-Engagement Visitors: Reach out to leads who spent more time on the website, as their engagement suggests a higher chance of conversion.

4. Tap into Chat Leads: Give special attention to leads from the "Olark Chat" source, as they exhibit a higher probability of conversion.

5. Leverage SMS-Sent Activity: Reach out to leads whose last activity was marked as "SMS Sent," as this activity indicates a better chance of conversion.

6. Avoid Olark Chat Conversations: Consider not contacting leads whose last activity was "Olark Chat Conversation," as these interactions are less likely to result in conversion.

7. Assess Lead Origin: Be cautious with leads from "Landing Page Submission" as they tend to have lower conversion rates.

8. Specialization Matters: Be selective with leads whose specialization is labeled as "Others," as they demonstrate a lower likelihood of conversion.

9. Avoid "Do Not Email" Leads: Minimize outreach to leads who have opted for "Do Not Email" since they are less likely to convert.

These recommendations are based on the analysis of the model and can assist in optimizing lead engagement strategies for higher conversion rates.