Lead Scoring

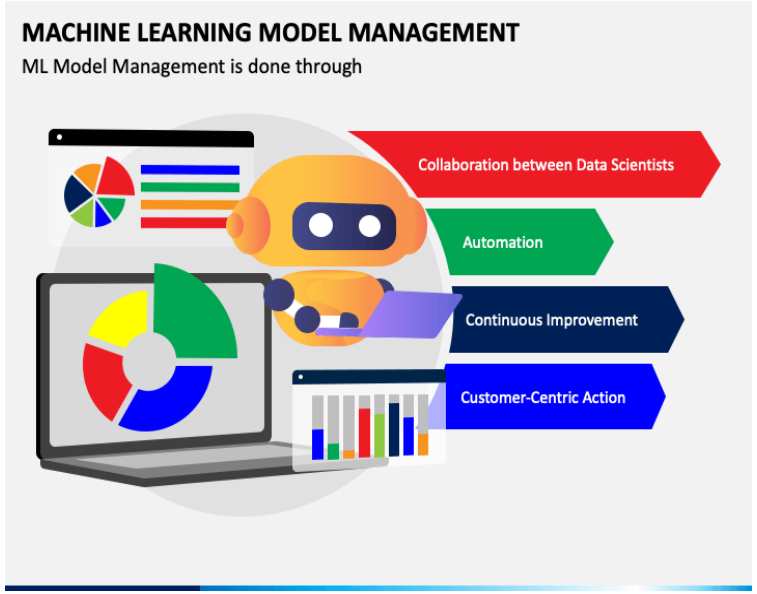
At X-Education

By:
Akshar Patel

- Data cleaning

- EDA

- Model



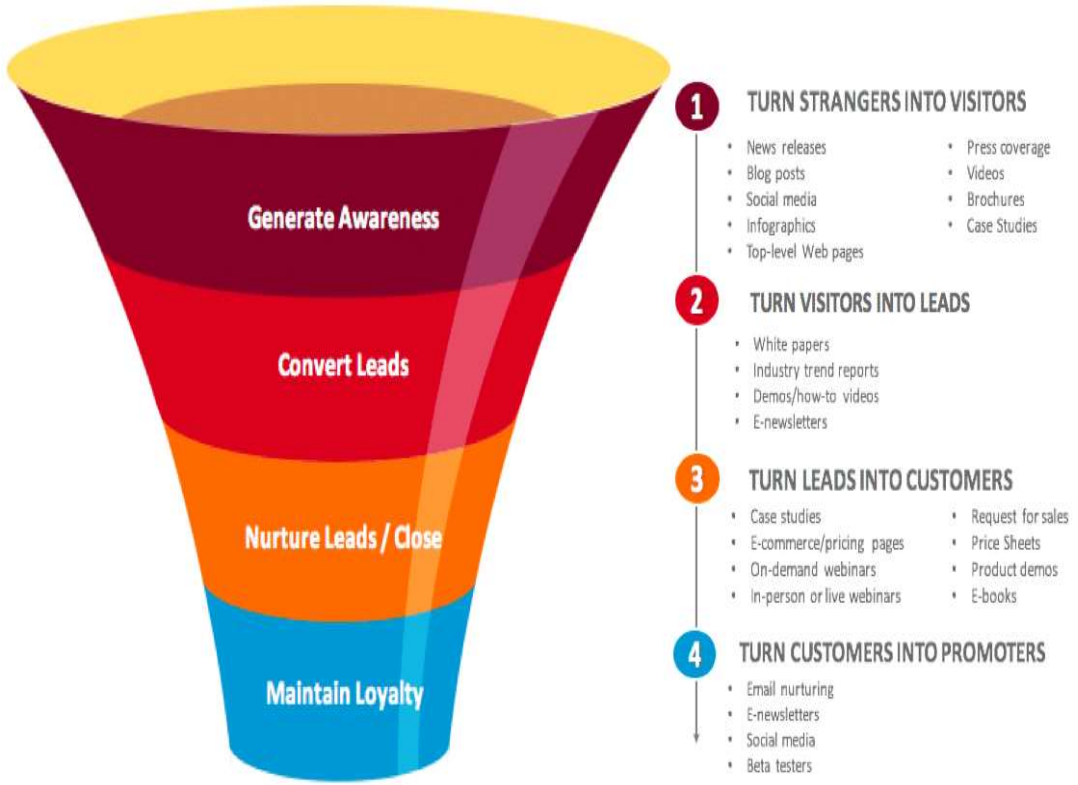
Problem statement:-

- X Education, an online course provider for industry professionals, faces a challenge with low lead conversion rates. Despite acquiring many leads through their website, search engine marketing, and referrals, only about 30% convert into paying customers.
- To address this, the company aims to identify 'Hot Leads'—the most promising potential customers. By focusing their sales efforts on these high-potential leads, X Education expects to significantly improve their lead conversion rate and overall sales efficiency.
- 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
- Data Preprocessing
- Data cleaning
- EDA
- Data Preparation

- Model building & Feature selection
- Model Evaluation
- Model Evaluation on test data



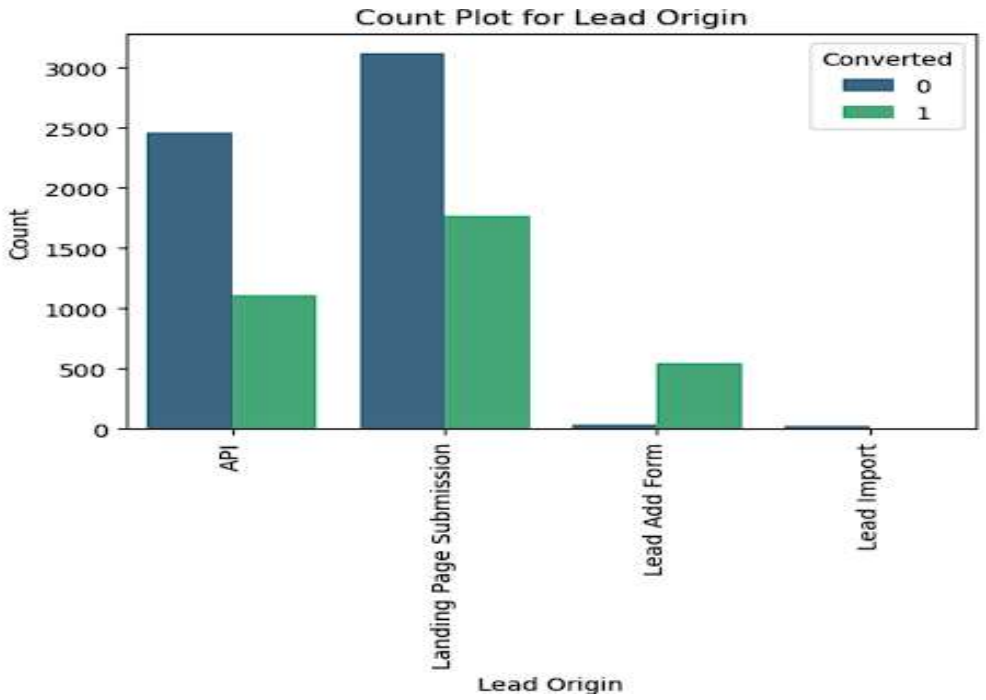
Stage:1
Data Cleaning:-

- The dataset contains 37 columns and approximately 9,000 rows, with a small number of missing values and no outliers.
- First, we applied a 30% cutoff to handle missing values and removed columns that were not useful for our analysis.
- Additionally, remove columns with selected data except for the 'specialization' column, and discard columns with index numbers, as these columns do not provide any valuable insights..
- Applied data imputation to columns with 15% null values after the cleaning process, filling in missing values with 'Unknown'.

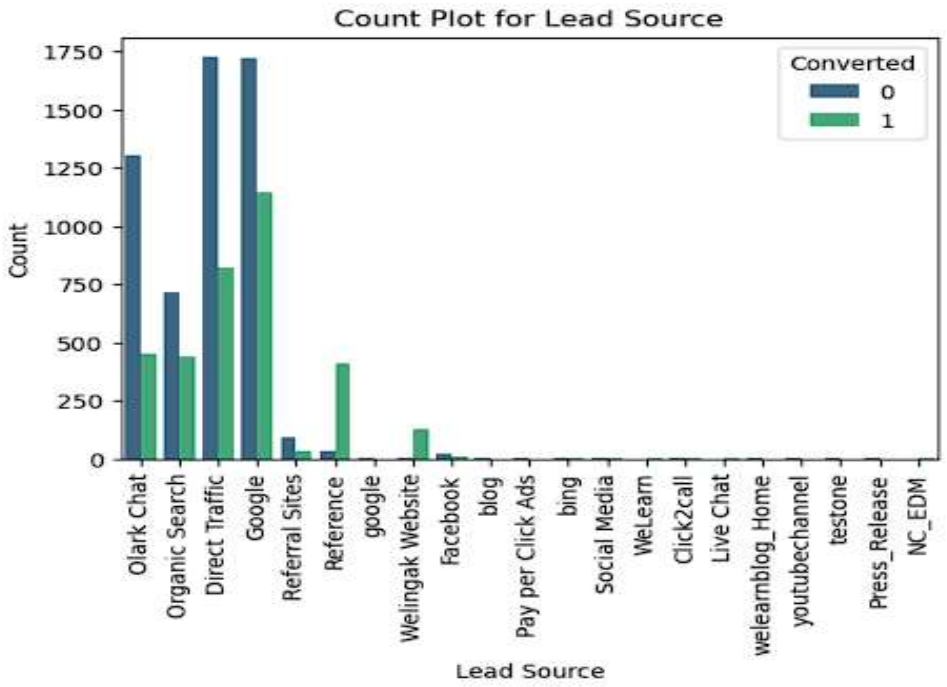
Stage:2
EDA:-

Lead Origin & Lead source

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

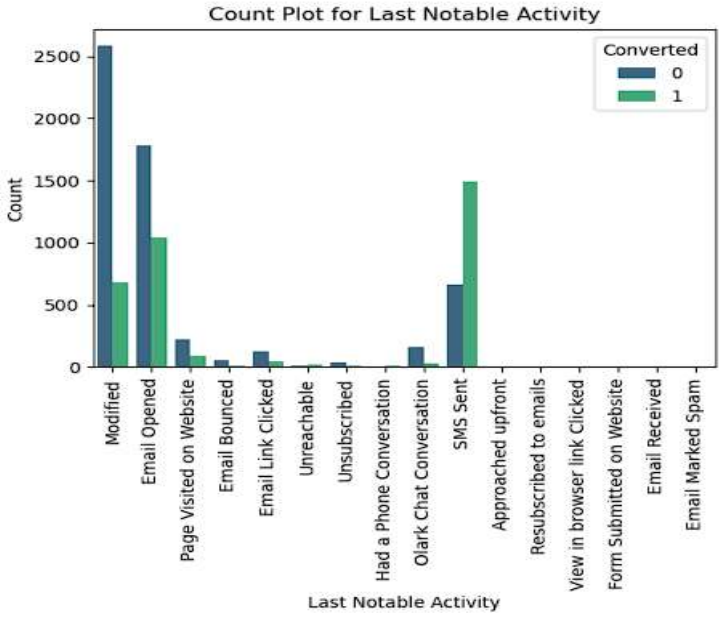


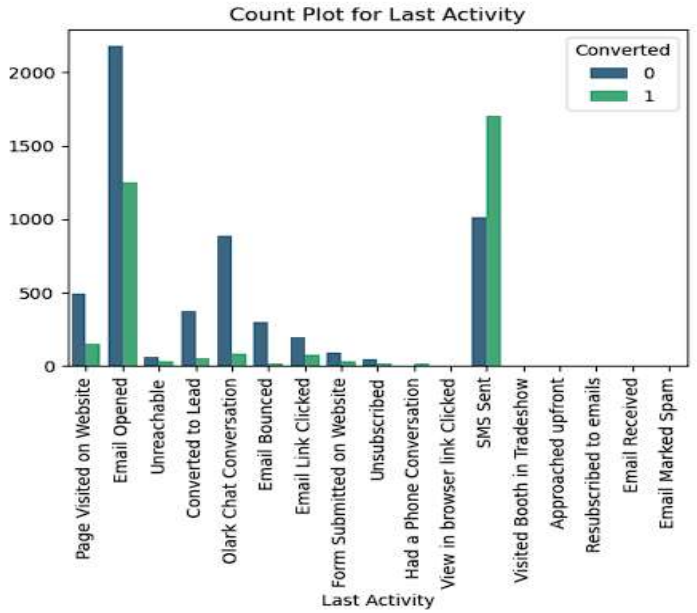
- 'Google' attracted more audience, but the 'Reference' and 'Welingakwebsite' has higher conversion rate.



Last Activity & Email:-

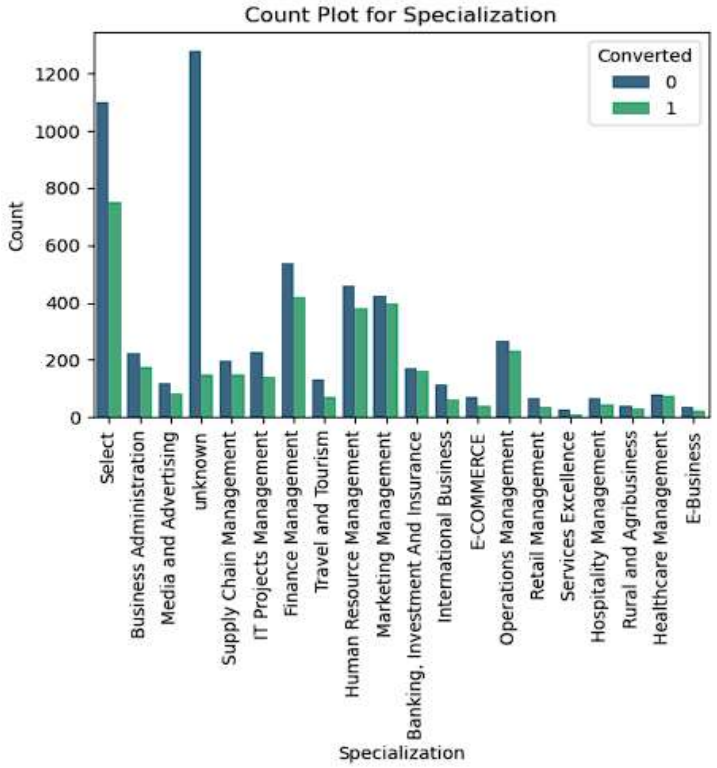
- Email Opened' shows the highest range of customers, but 'SMS Sent' has a higher conversion rate. Both columns have high 'SMS Sent'.



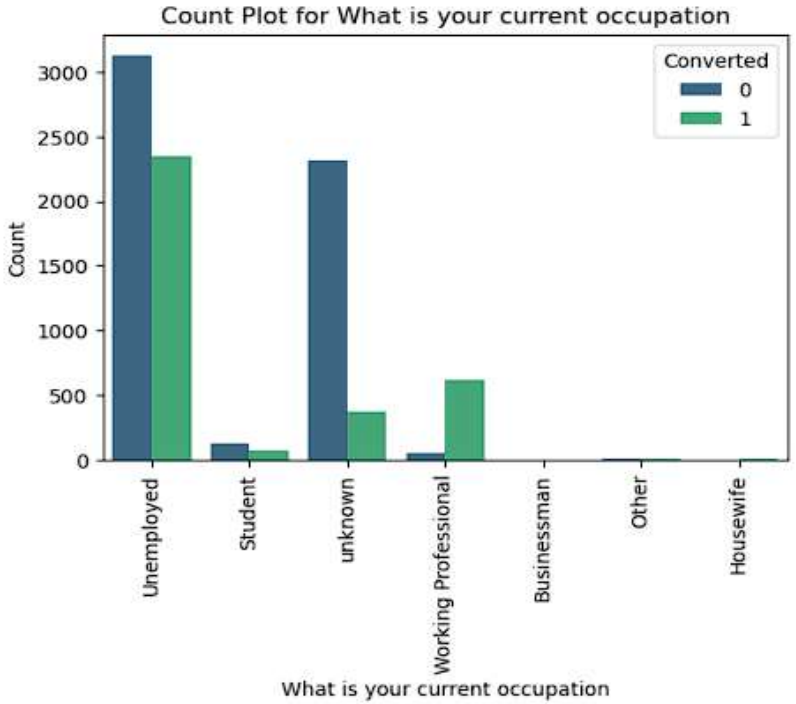


Specialization & Occupation:-

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

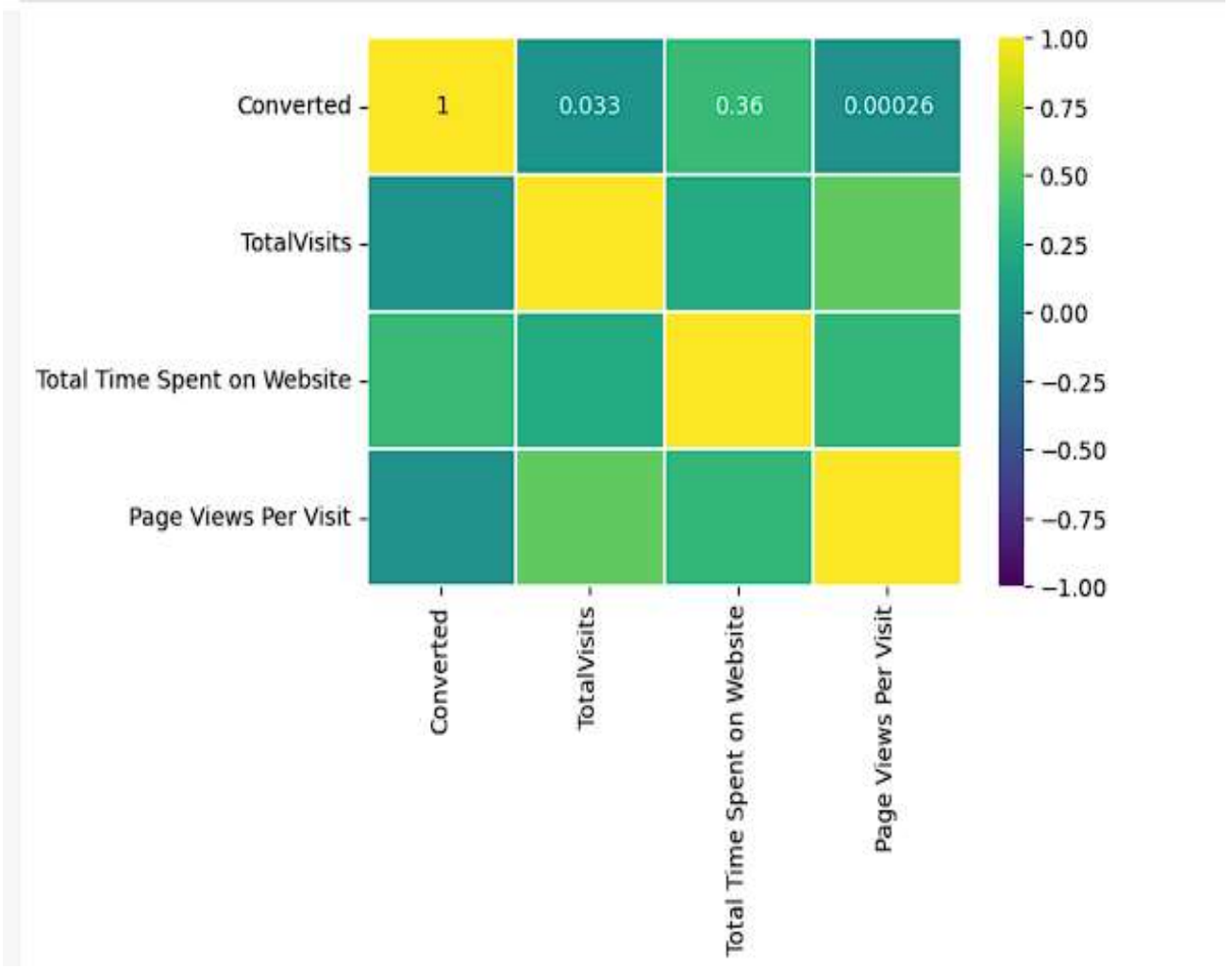


- Students who are approaching are mostly 'Unemployed', but the high conversion rate is in 'Working Professionals'.



Heat Map:-

- There is a strong positive correlation of 0.50 between TotalVisits and Page Views Per Visit, as well as a significant positive correlation of 0.40 between `Total Time Spent on Website` and `Converted`.
- At the beginning of the EDA, I dropped a column with the highest 'NO' values to achieve the highest accuracy in the model.
- At this stage, I have retained 98% of the rows for further analysis by choosing to drop some irrelevant columns instead of rows.

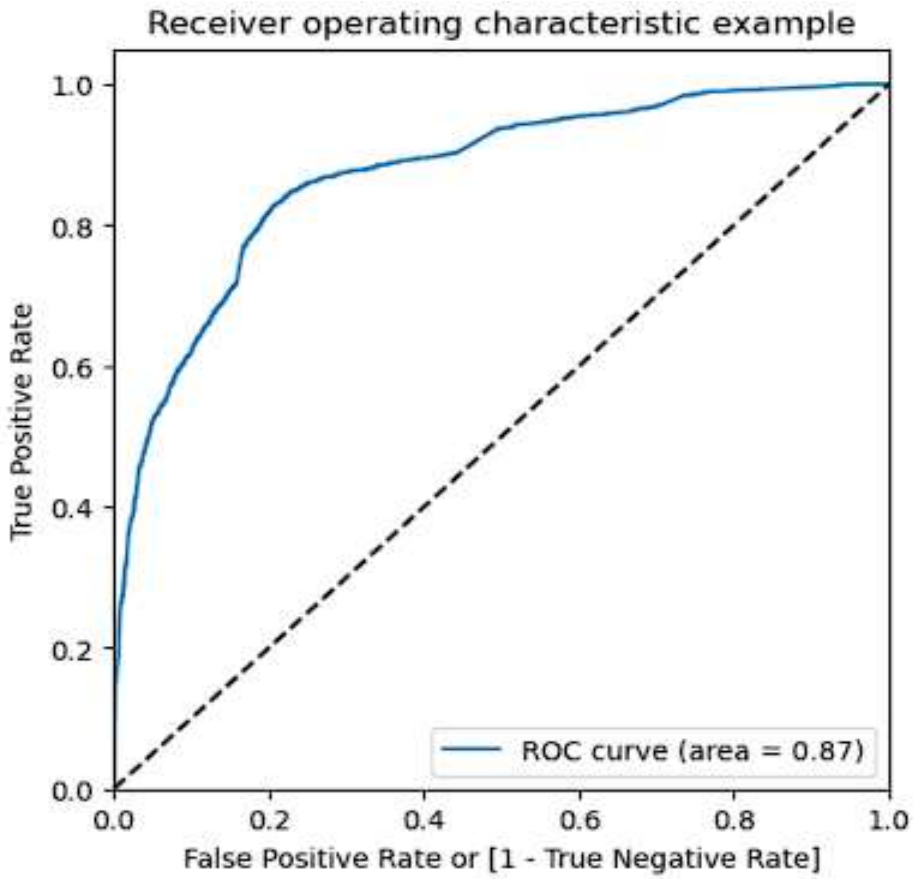


Stage:3
ML Model Building:-

- Data processing was performed after the EDA, where I split the data into 70% training and 30% testing sets, using a scaler and creating dummy variables for categorical columns.
- Utilized logistic regression algorithms to build the model since we are working with categorical data.
- Employed RFE and VIF to select the final features, ensuring that the P-value is below 0.05 and the VIF is below 5. Columns violating these criteria were dropped.
- Evaluated the model with the final selected features using metrics such as accuracy, specificity, sensitivity, and precision..
- Plotted the ROC curve, revealing an area under the curve (AUC) of 0.87, indicating that our model performs well.

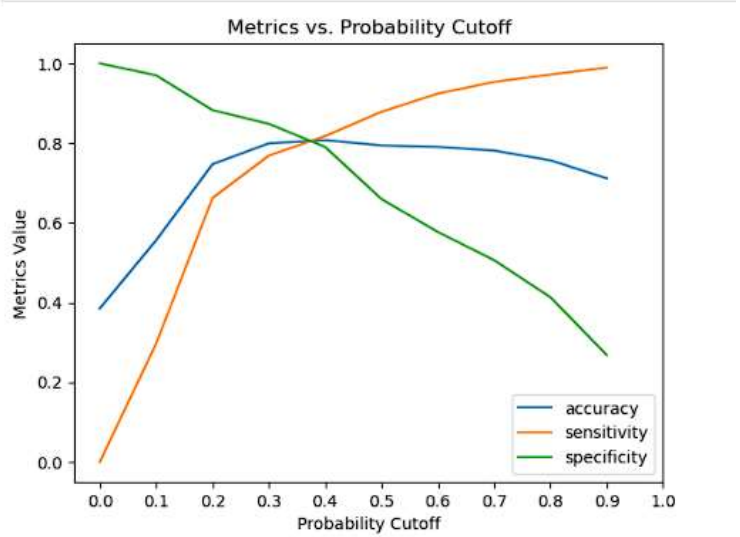
ROC:-

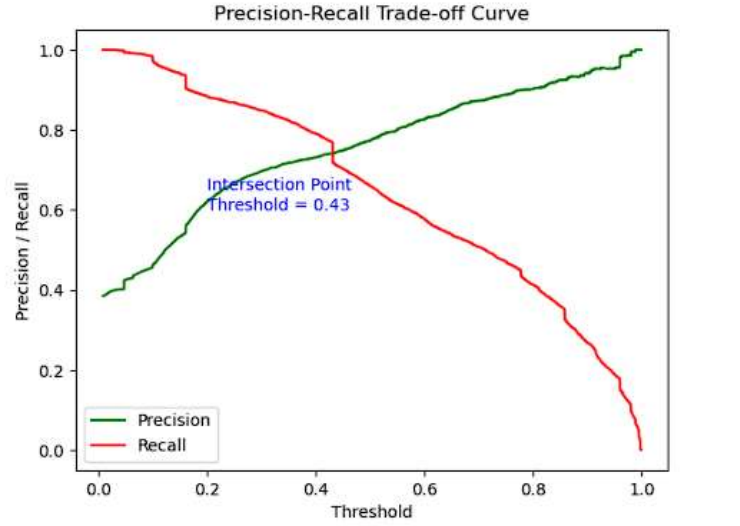
- The ROC curve illustrates the trade-off between a binary classifier's true positive rate and false positive rate across varying classification thresholds, aiding performance evaluation. From the ROC curve, the area under curve is 0.87 therefore our modelis good.



Cutoff:- (Train data)

- Evaluated the metrics and set the cutoff at 0.39. We then plotted the precision and recall graph, observing a slightly different cutoff. This suggests that the model's performance is stable across different datasets.
- On the training data, the confusion matrix is([[3178, 727], [501, 1945]]
- With, True Positives: 1945 True Negatives: 3178 False Positives: 727 False Negatives: 501





Test data Accuracy:-

For Train set we got the metrics as:

- **Accuracy = 0.806**
- **Sensitivity = 0.795**
- **Specificity = 0.813**
- **Precision = 0.727**
- **Recall = 0.795**

For Test set we got the metrics as:

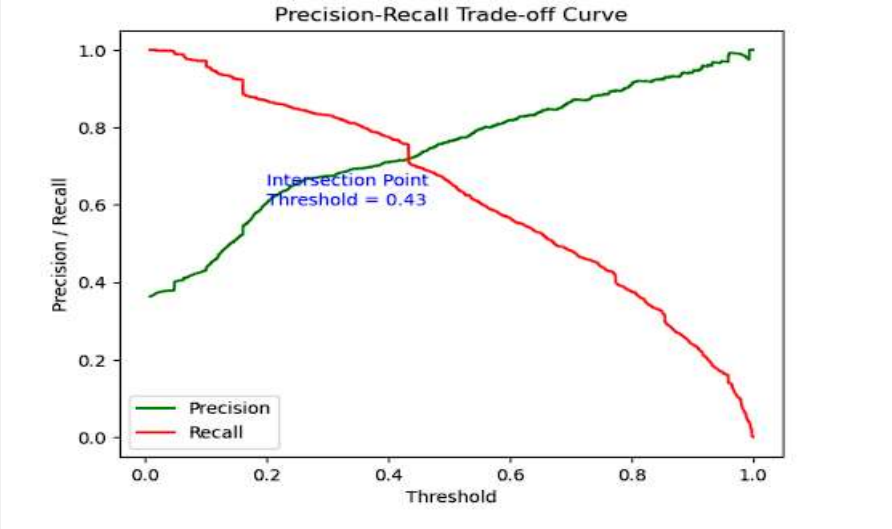
- **Accuracy = 0.802**
- **Sensitivity = 0.780**
- **Specificity = 0.814**
- **Precision = 0.706**
- **Recall = 0.780**

(test data)

INCOME PERCENTAGE

Achieved approximately 80% accuracy on the test data with an optimal cutoff of 0.39 for the model
This indicates that the model is neither overfitting nor underfitting and is performing well in generalizing to unseen data.

80%



Summary:-

- Focus on leads originating from `Reference` and `Welingak Website` and `Organic search` sources, as they have a higher likelihood of conversion.
- Prioritize leads who are `working professionals`, as they show a higher tendency to convert into customers.
- Reach out to leads who `spent more time on the website` or `page visited` as their engagement suggests a higher chance of conversion.
- Give special attention to leads from the `Olark Chat` source, as they exhibit a higher probability of conversion.
- Reach out to leads whose last activity was marked as `SMS Sent` as this activity indicates a better chance of conversion.
- Consider not contacting leads whose last activity was `Olark Chat Conversation` as these interactions are less likely to result in conversion.
- Be cautious with leads from `Landing Page Submission` as they tend to have lower conversion rates.
- Focus on leads with specializations like `Working professionals` and `Unemployed`, as they are more likely to seek better career opportunities and show a higher likelihood of conversion. Be selective with leads labelled as `Others`, `Businessman` or `Housewife` as these categories demonstrate a lower likelihood of conversion.
- Minimize outreach to leads who have opted for `Do Not Email` since they are less likely to convert.
- Also, Improve the design and content of your `lead capture forms` called `Lead form` to enhance data quality and lead scoring accuracy.

THANK YOU

AKSHARPATEL
MAIL:-patelakshar2025@gmail.com