Lead Scoring Case Study by using Logistic Regression

Submitted by :

1) Jyoti Kadam
2) Mohammed Omar
3) Aditya Shukla

Contents :

Problem Statement :

- X Education, an online course provider for industry professionals, witnesses frequent visits to its website by interested individuals searching for courses.

- Courses are promoted across diverse online platforms, including search engines like Google. Upon landing on the website, visitors might browse available courses, fill out course forms, or engage with informational videos.
  Leads are generated when visitors provide their contact details through forms or referrals. Once acquired, the sales team initiates outreach via calls, emails, etc., aiming to convert these leads into customers.

- Despite acquiring numerous leads, X Education struggles with a low lead conversion rate, approximately 30%. For instance, out of 100 daily leads, only around 30 get converted.

- To optimize efficiency, the company aims to pinpoint 'Hot Leads,' the most promising prospects. Identifying these leads is crucial, as a focus on these potential leads can substantially increase the lead conversion rate by directing sales efforts more precisely.

Business Objective :

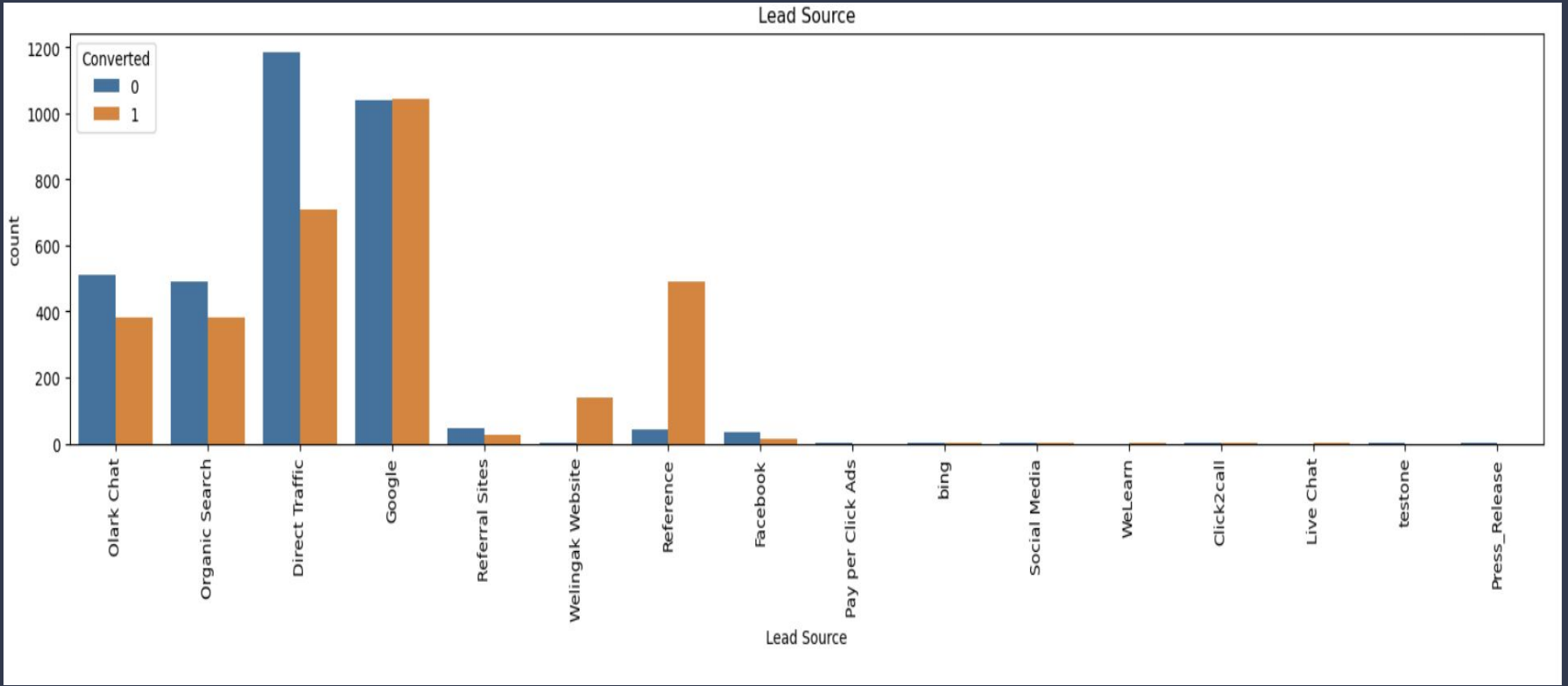
- Develop a model to assign lead scores, indicating potential paying customers based on their likelihood of conversion.
- Prioritize higher lead scores for prospects with increased conversion potential and lower scores for those with reduced chances of conversion.
- The CEO's target is an approximate lead conversion rate of 80%.

Approach :

- Importing and Merging
- Inspecting the data frame
- Data Cleaning
- EDA
- Data Processing
- Model building
- Evaluation
- Determining Threshold

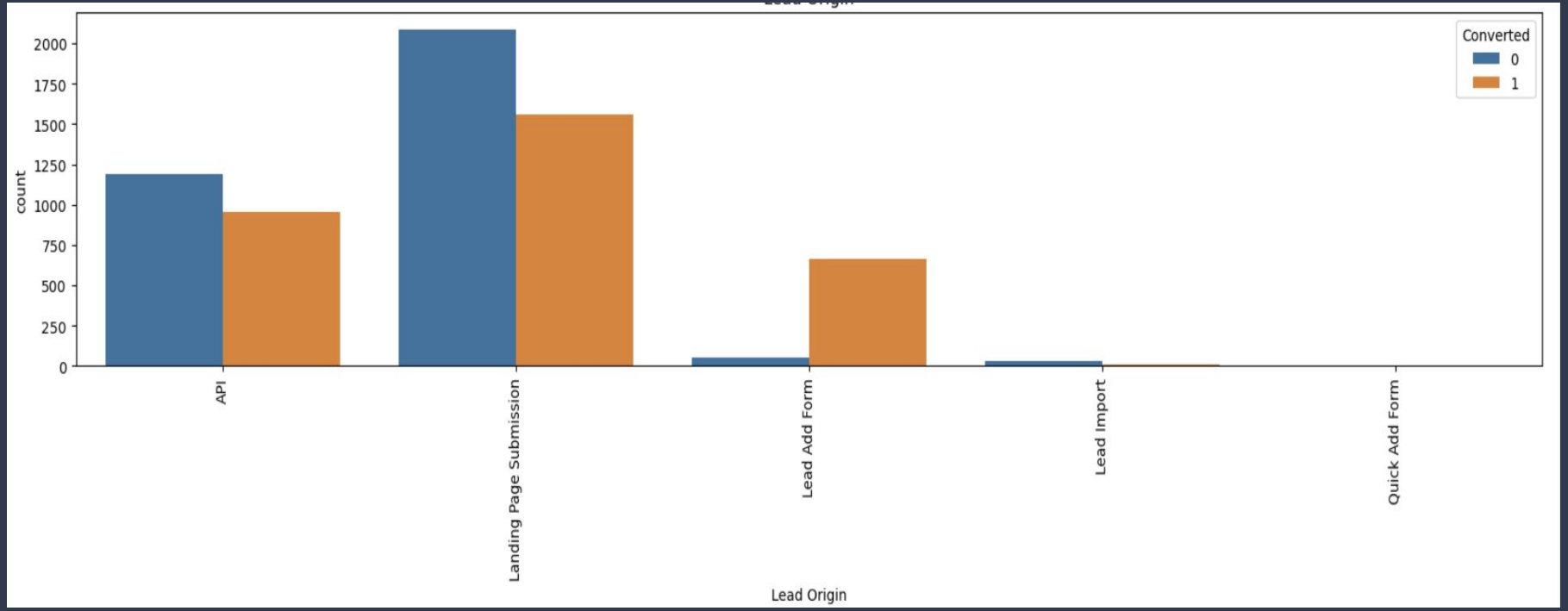
EDA : Lead Source

Lead Source the leads generated through Direct-Traffic and Google has the high chance of conversation



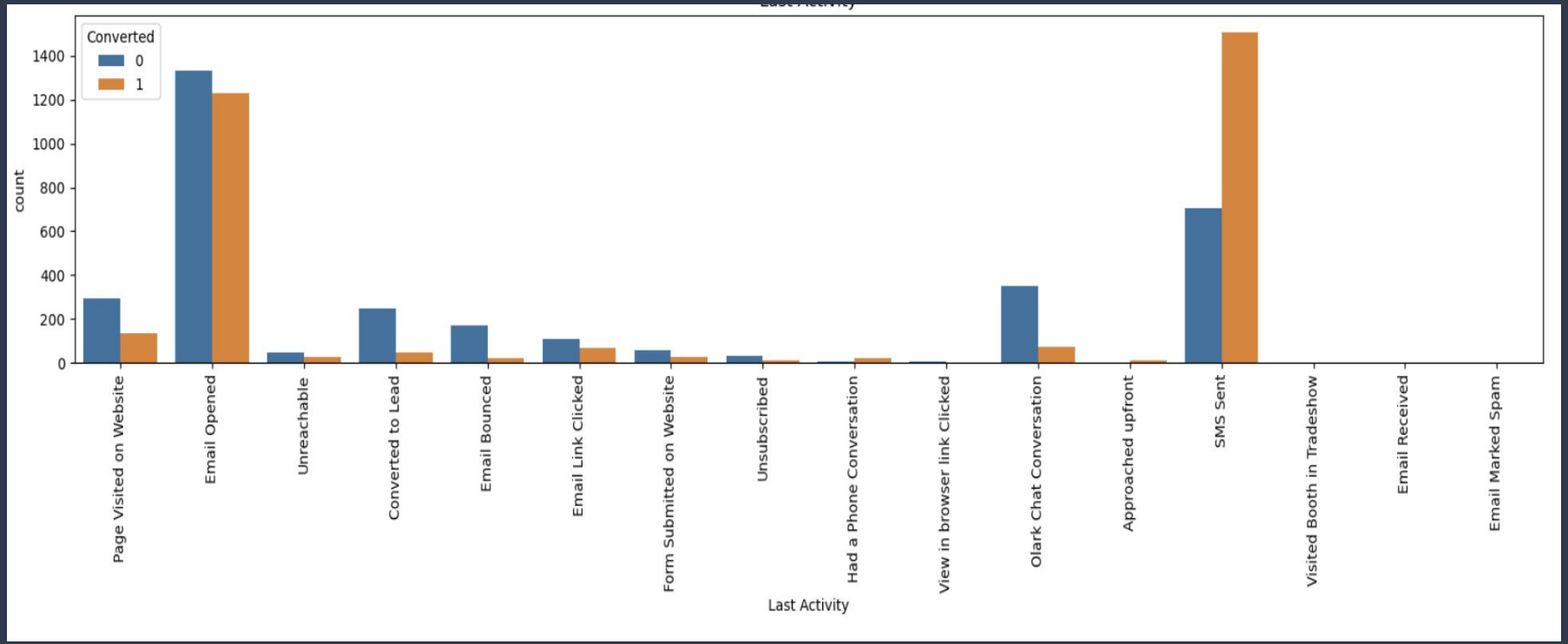
Lead Origin :

In Lead Origin category, a significant number of leads come from landing page submission.



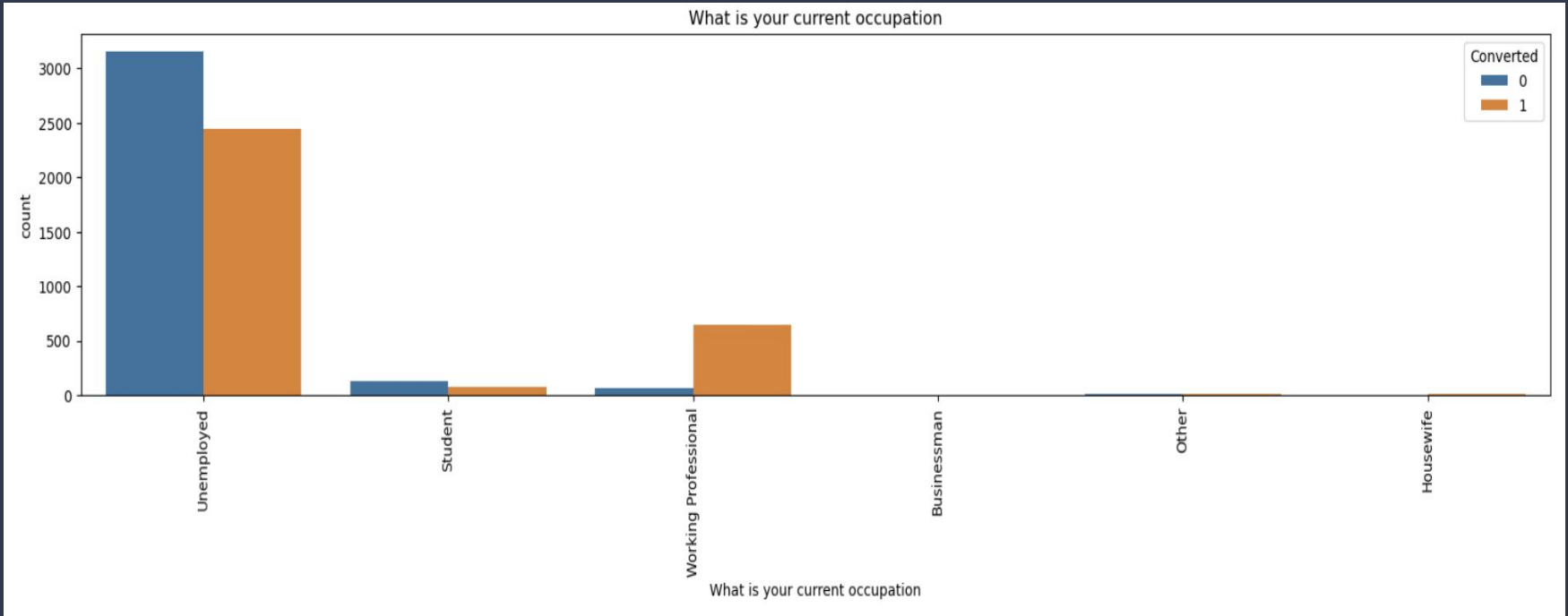
Last Lead Activity :

Leads that originate from opened emails show a higher likelihood of conversion, and a similar advantageous impact can be observed for leads originating from SMS interactions.



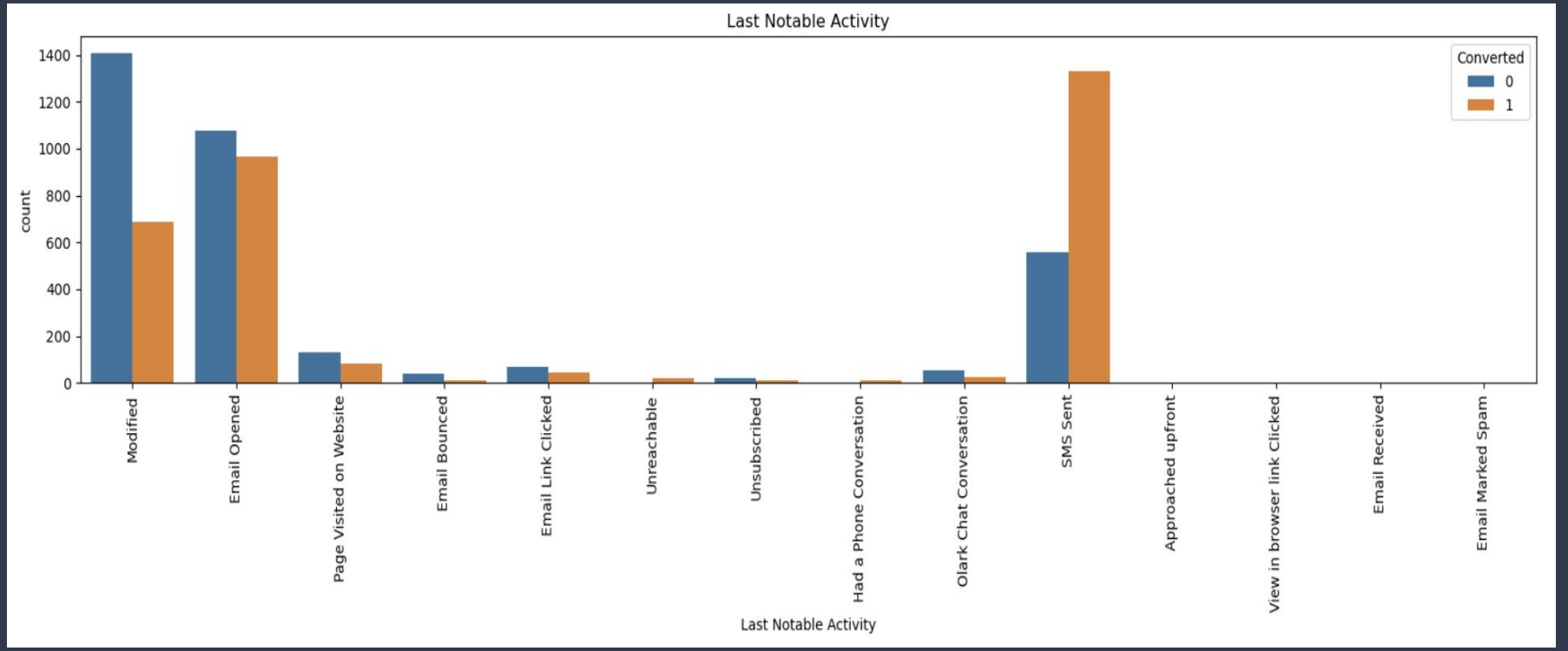
What is your Current Occupation :

Individuals not currently employed display a greater inclination to enroll in the course compared to those in other employment categories.



Last Notable Activity :

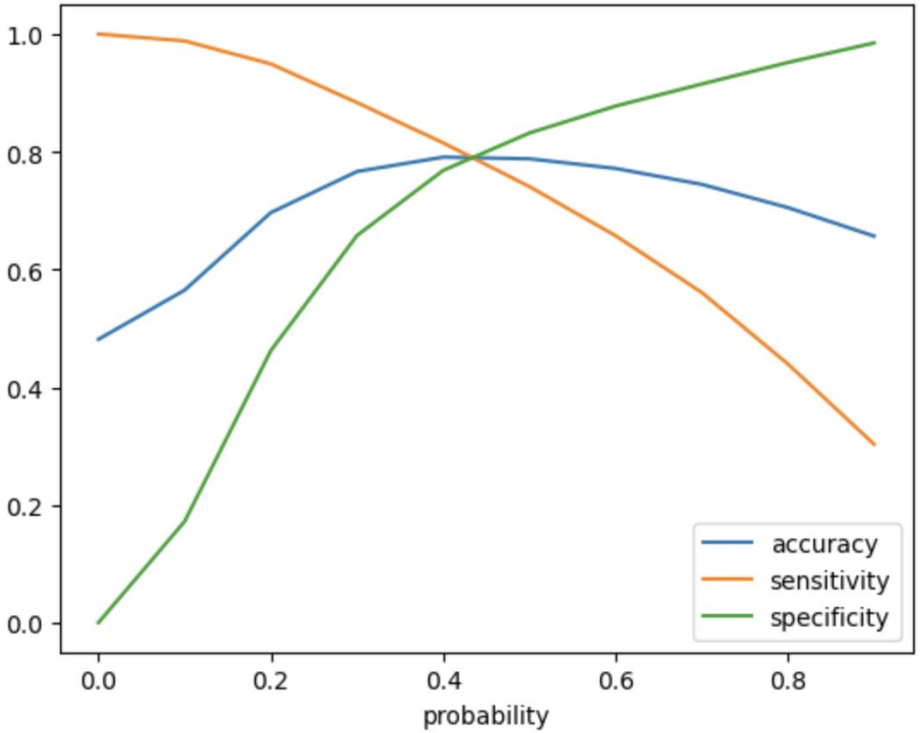
"Modified" stands as the most recent notable activity for the majority of users, followed by the email opened.



Model Evaluation

Graphs shows changes in sensitivity, Specificity, and Accuracy with changes in Probability threshold values

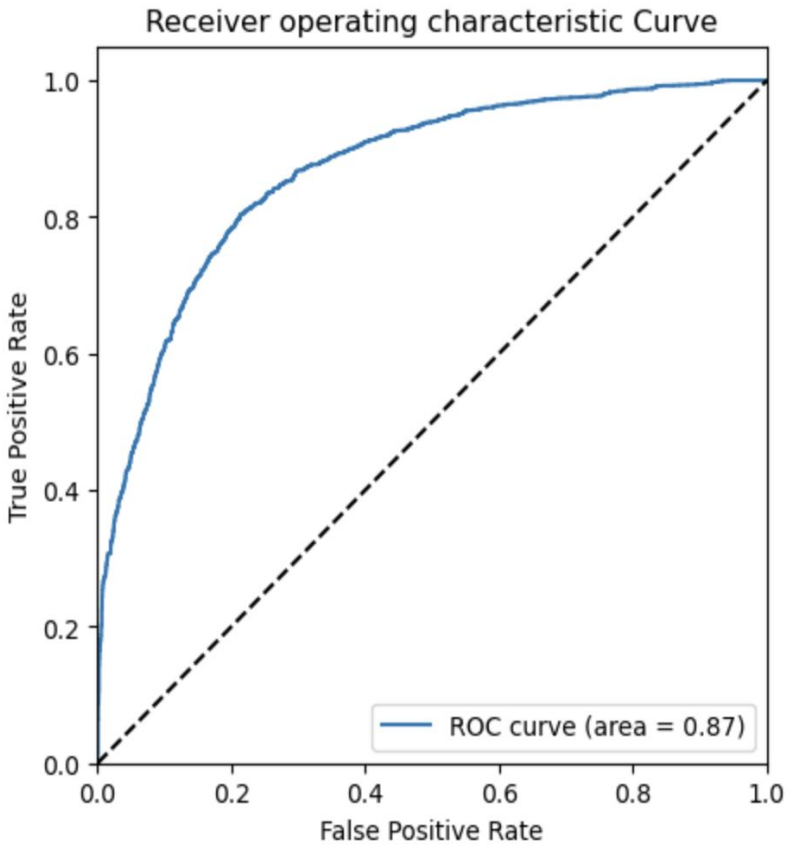
Optimal threshold = 0.43



ROC Curve

Area under ROC curve is 0.87 out of 1 which indicates good Predictive Model

The Curve is close to top left corner of the plot. Which represent a model that has high true positive rate and a low false positive rate at all threshold values



Final Features List

- Lead Number
- Do Not Email
- TotalVisits
- Total Time Spent on Website
- Page Views Per Visit
- Lead Origin_Lead add form
- Lead Source_Olark Chat
- Lead Source_ Welingak Website
- Last Acivity_SMS Sent
- Current occupation_Working Professionals

Conclusion:

- Conversion rates around 30-35% are common, particularly for API and Landing page submissions. Lead Add form and Lead import show lower rates, indicating a need to prioritize leads from API and Landing pages.
- Google and direct traffic yield the most leads. Better conversion ratios are seen through referrals and the Welingak website.
- Leads spending more time on the website tend to convert better.
- 'Email opened' is the most frequent last activity, while 'SMS Sent' results in the highest conversions. Unemployed individuals dominate the leads, but working professionals show the highest conversion rates.