# LEAD SCORE CASE STUDY

Submitted by:

**Ivaturi Sri Jahnavi Archita**

**Apeksha M Angadi**

# PROBLEM STATEMENT

- X Education is an education company that offers online courses to industry professionals. Each day, numerous professionals interested in these courses visit the company's website to explore the available options.
- The company promotes its courses across various websites and search engines, such as Google. When these individuals arrive on the website, they may browse the course offerings, complete a form for more information, or watch instructional videos. If they provide their email address or phone number through a form, they are categorized as leads. Additionally, the company receives leads from past customer referrals.
- Once leads are obtained, the sales team engages with them through phone calls, emails, and other communication methods. Through this process, some leads convert into customers, though the majority do not. Typically, X Education experiences a lead conversion rate of around 30%.

# BUSINESS GOAL

- X Education has tasked you with selecting the most promising leads, those most likely to become paying customers.
- To achieve this, you are required to build a model that assigns a lead score to each lead. Leads with higher scores should have a greater likelihood of conversion, while those with lower scores should have a lower chance.
- The CEO has set an approximate target lead conversion rate of around 80%.

# METHODOLOGY

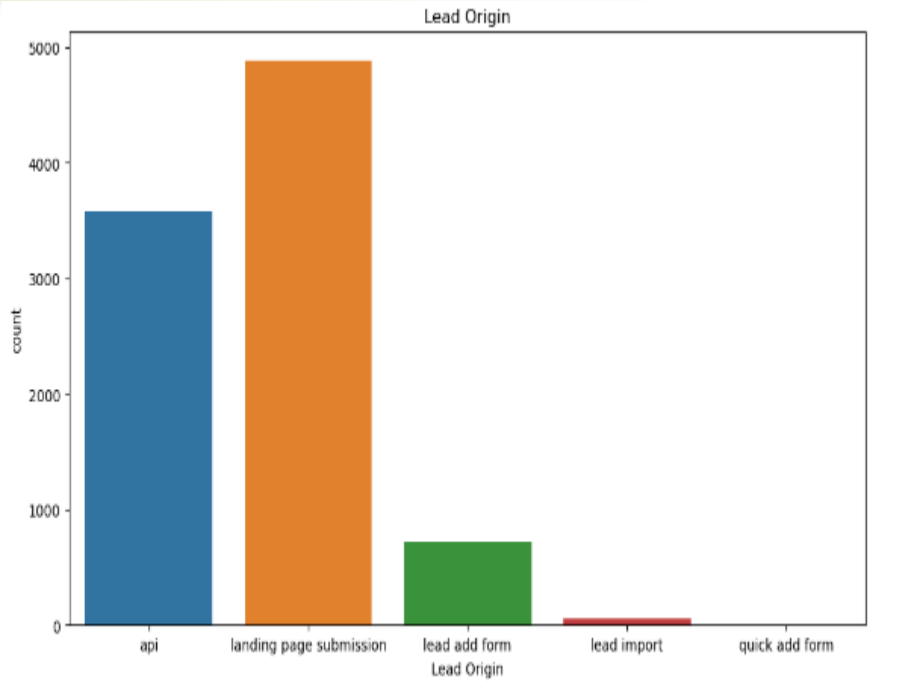
- Importing the data
- Data Cleaning
- Exploratory Data Analysis.
- Data preparation
- Splitting the data into Test and Train dataset
- Feature scaling
- Building of Model
- Evaluation of Model
- Model prediction and Lead Score Calculation

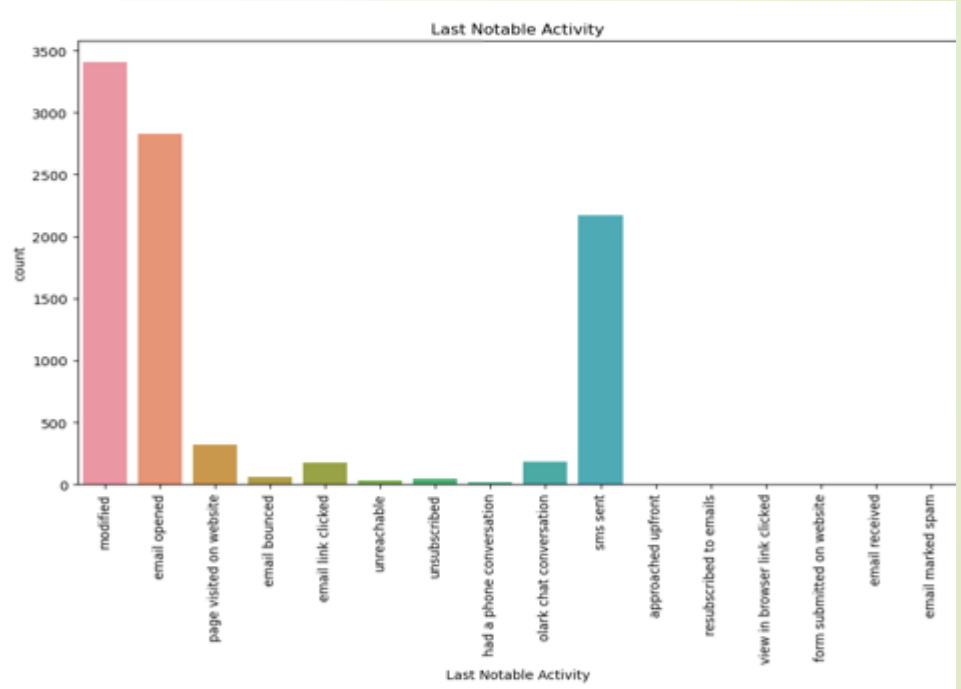
# Data Manipulation

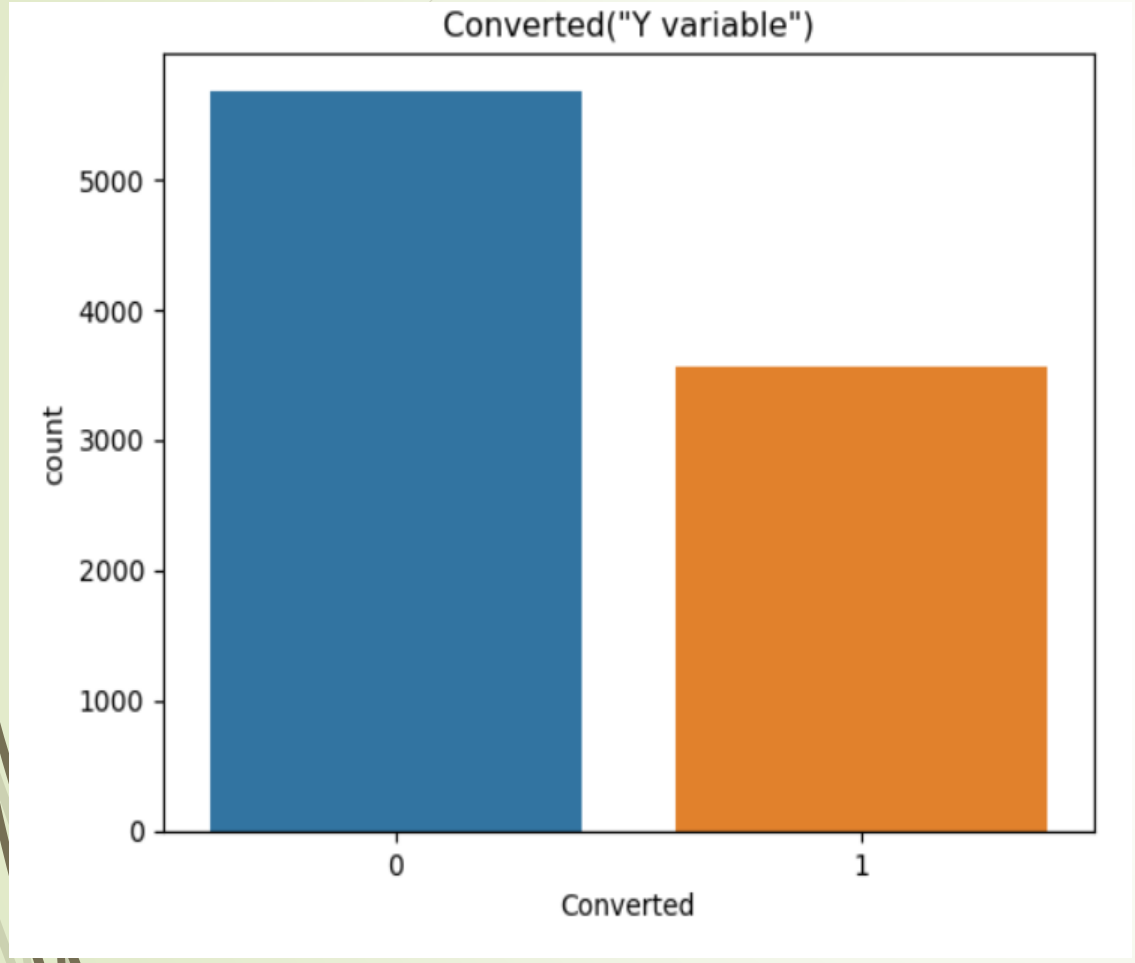
- The dataset consists of 9,240 rows and 37 columns.
- Features with single values, such as “Magazine," "Receive More Updates About Our Courses," "Update me on Supply Chain Content," "Get updates on DM Content," and “I agree to pay the amount through cheque,” etc., have been excluded from further processing.
- Columns with 35% or more null values, such as 'Asymmetrique Profile Index,’ 'Asymmetrique Activity Index,’ 'Asymmetrique Activity Score,’ 'Asymmetrique Profile Score,’ 'Lead Profile,’ 'Tags,’ 'Lead Quality,’ and 'How did you hear about X Education,’ have been removed.
- The columns “Prospect ID” and “Lead Number” have also been dropped, as they are not necessary for further analysis.

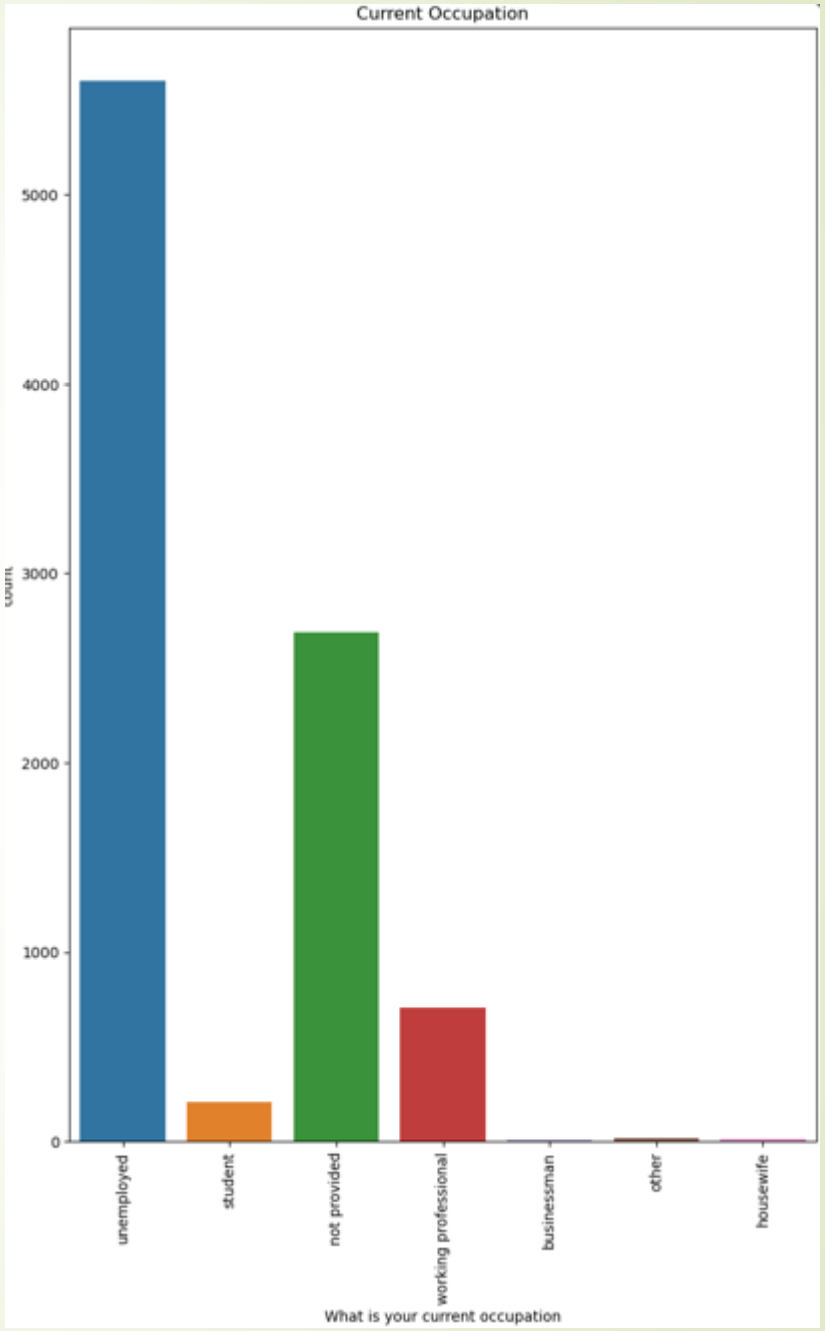
# Exploratory Data Analysis

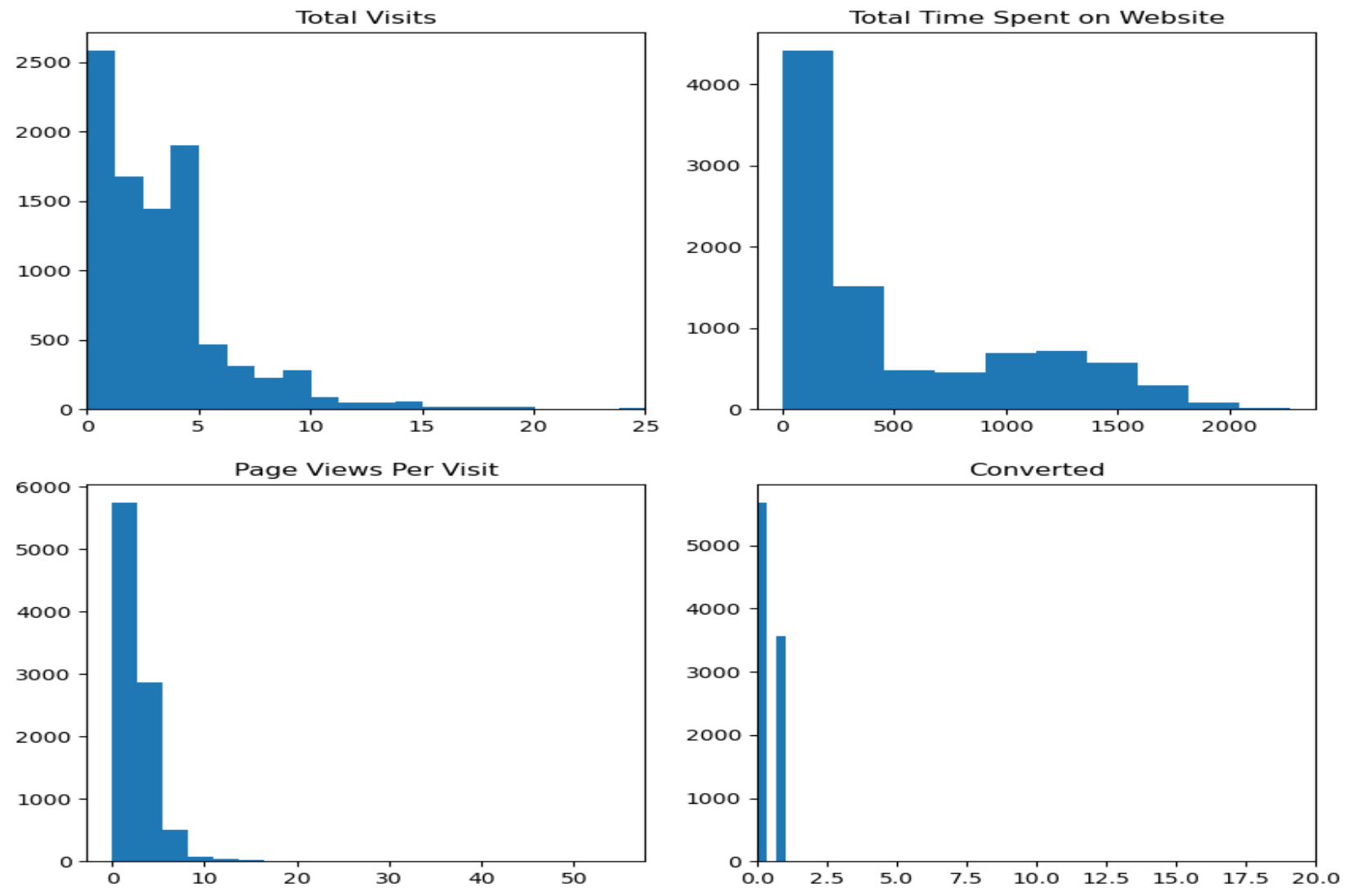
The univariate analysis on the categorical and numerical variables was performed and various insights were found:



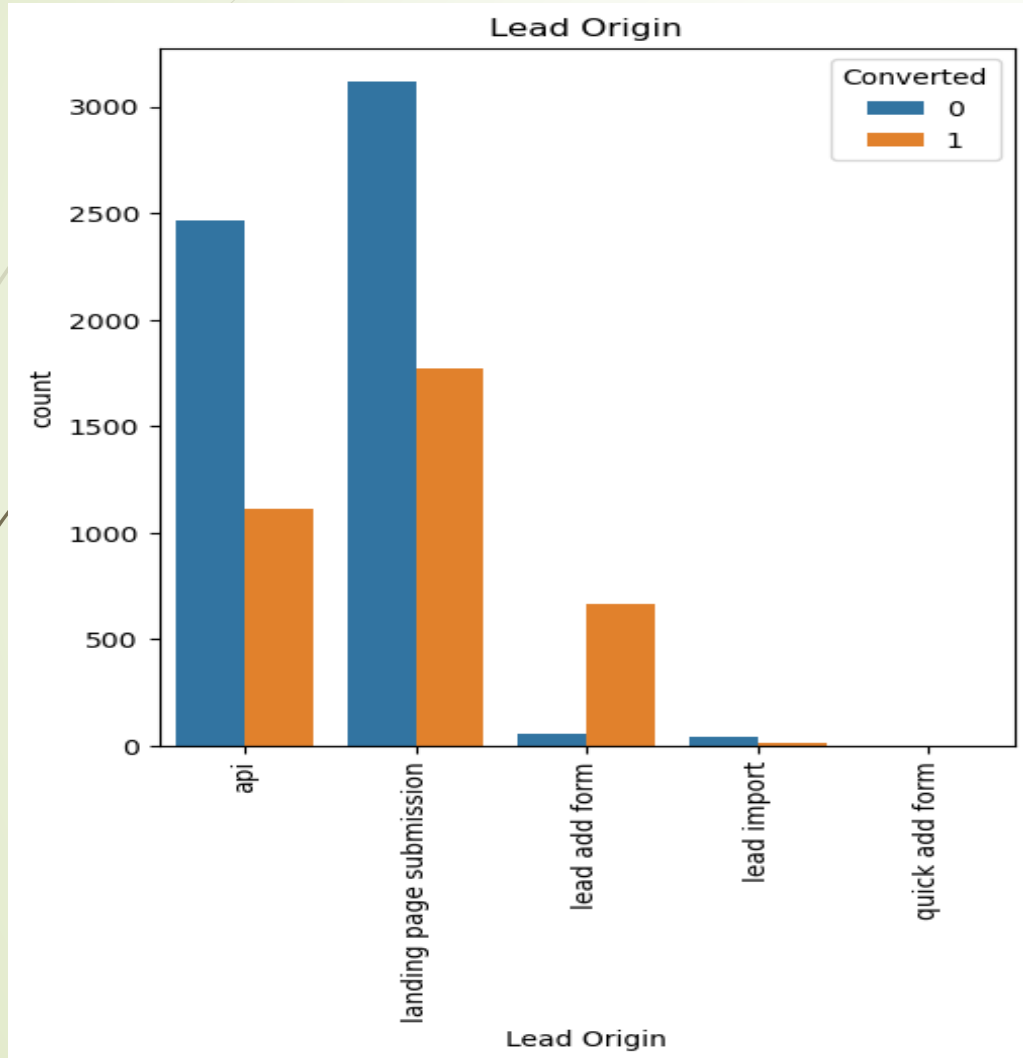

Converted("Y variable")
count
5000
4000
3000
2000
1000
0
0
1
Converted

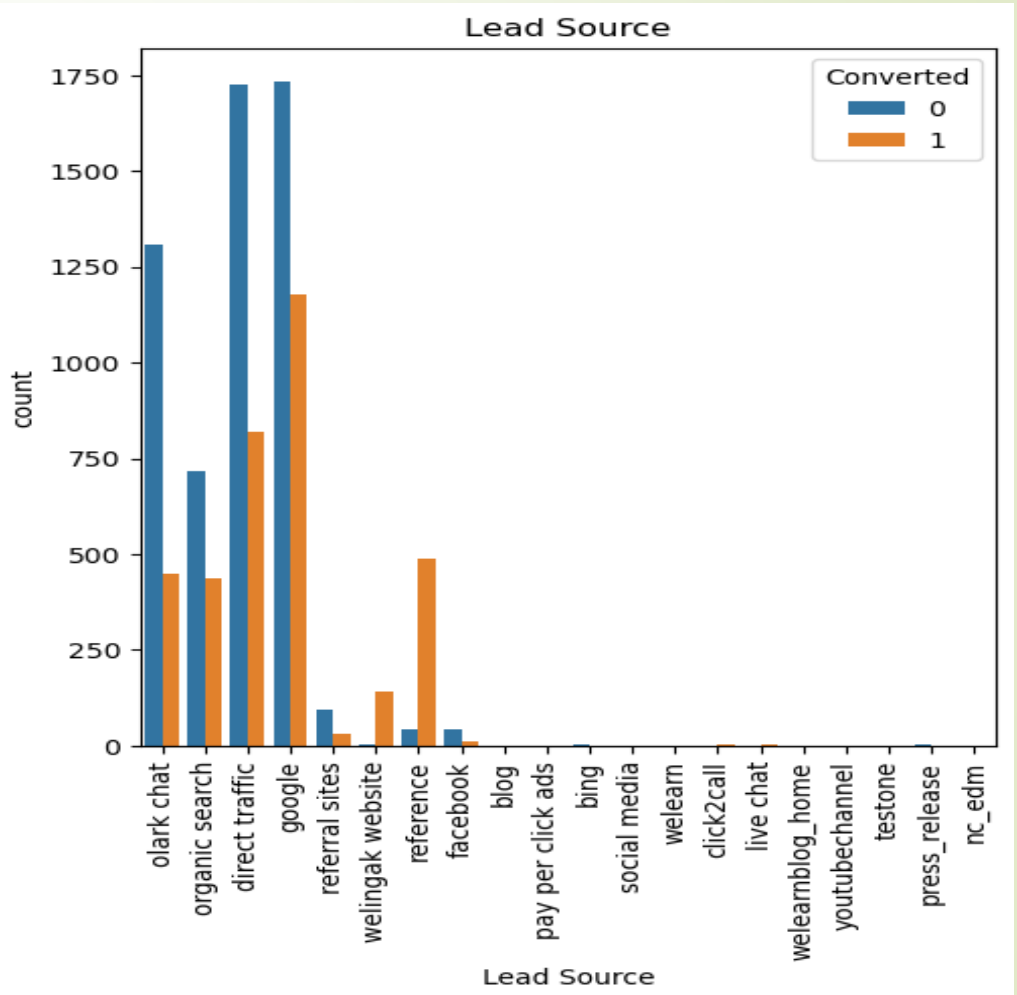

Current Occupation
count
5000
4000
3000
2000
1000
0
unemployed
student
not provided
working professional
businessman
other
housewife
What is your current occupation

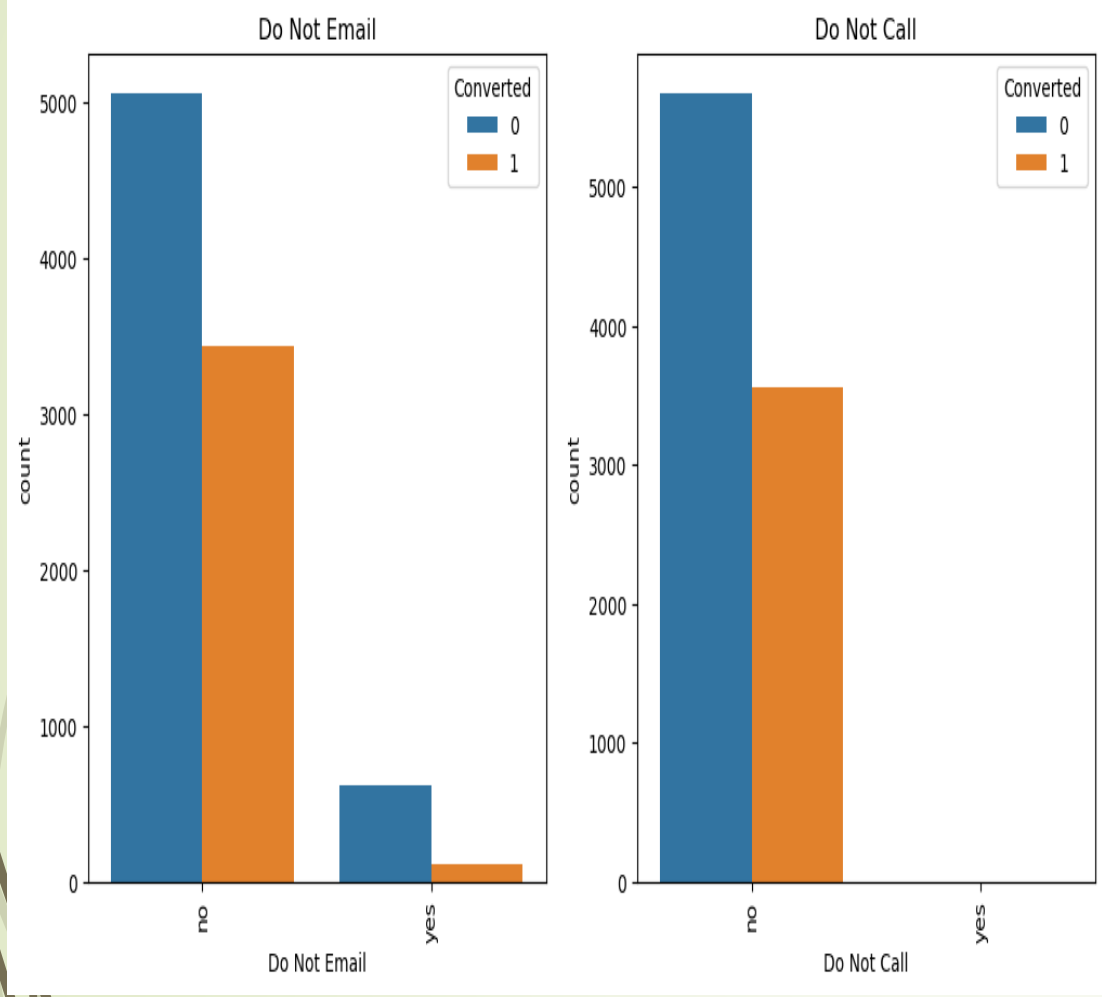


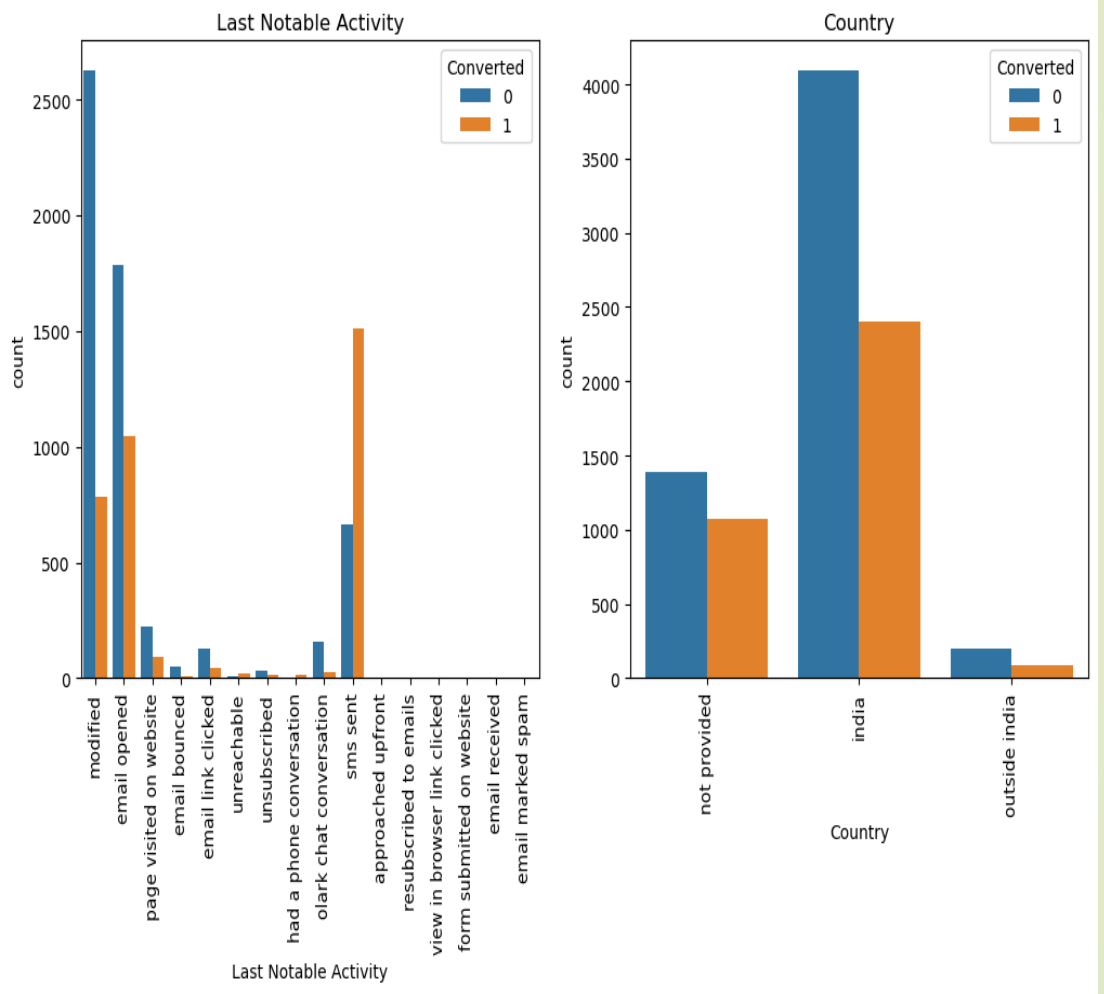

These are the numerical variables that had the most significant impact on the conversion rate: the higher the number of visits or the time spent on the website, the greater the likelihood of the lead being converted.

# Relating the categorical variables to “Converted” :



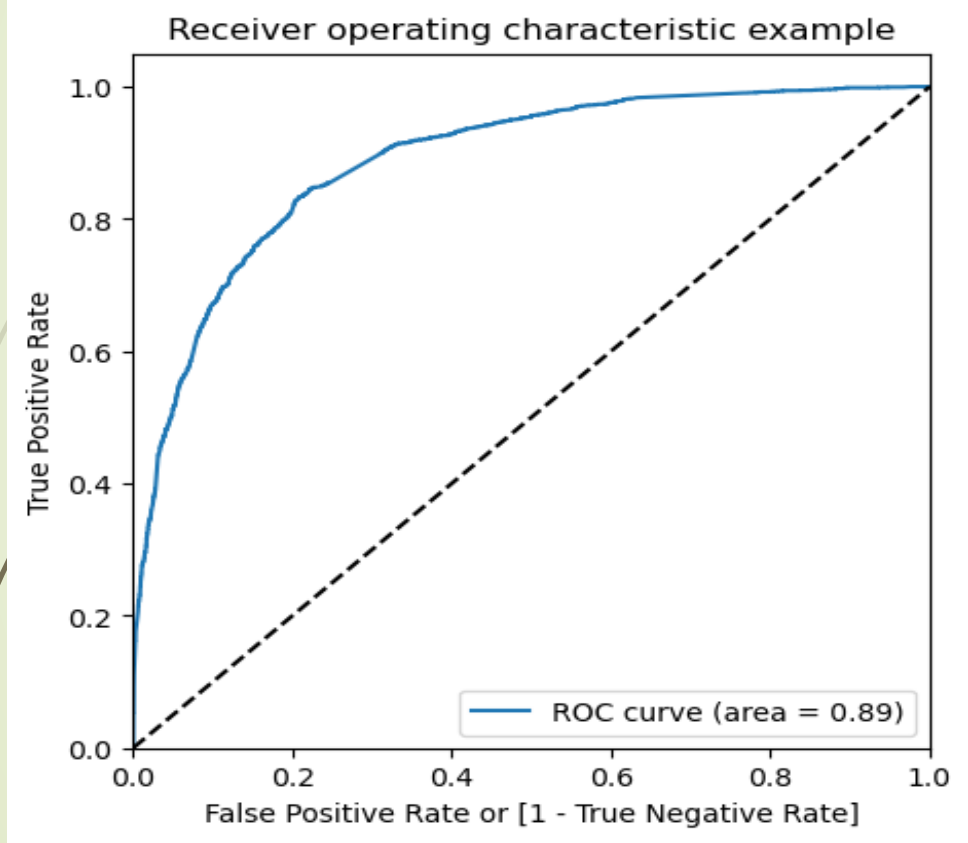

Do Not Email
Converted
0
1
count
5000
4000
3000
2000
1000
0
no
yes
Do Not Email
Do Not Call
Converted
0
1
count
5000
4000
3000
2000
1000
0
no
yes
Do Not Call

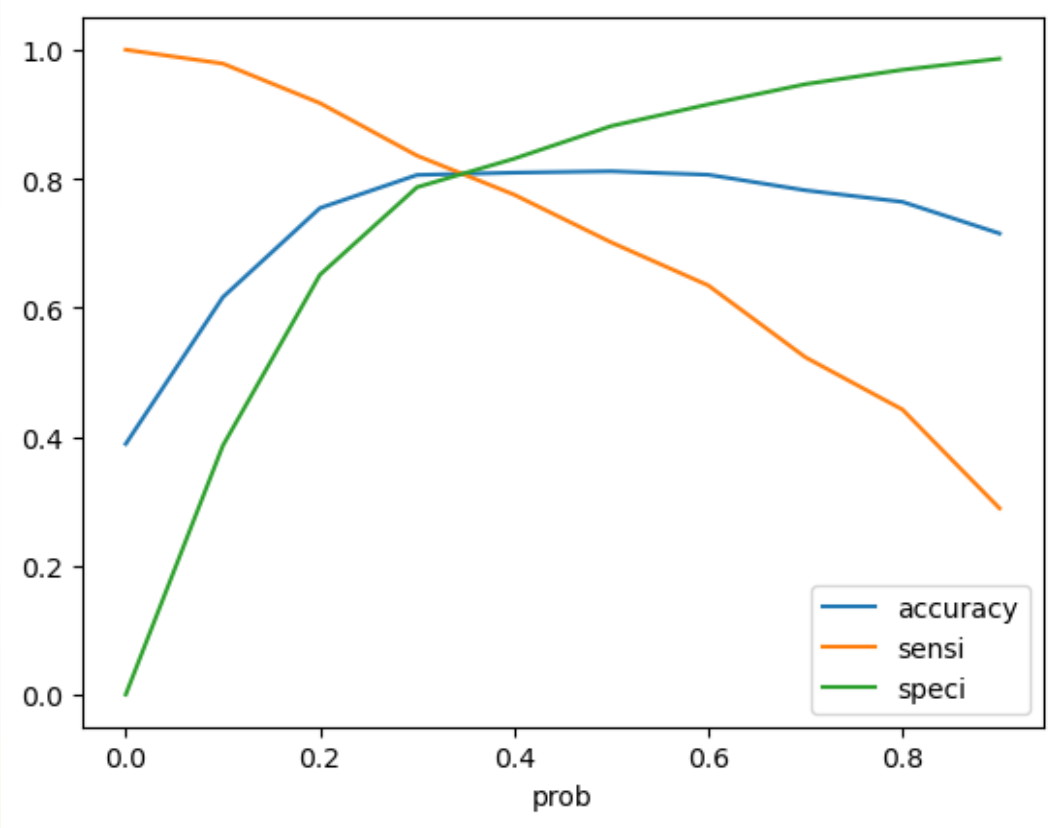

Last Notable Activity
Converted
0
1
count
2500
2000
1500
1000
500
0
modified
email opened
page visited on website
email bounced
email link clicked
unreachable
unsubscribed
had a phone conversation
olark chat conversation
sms sent
approached upfront
resubscribed to emails
view in browser link clicked
form submitted on website
email received
email marked spam
Last Notable Activity
Country
Converted
0
1
count
4000
3500
3000
2500
2000
1500
1000
500
0
not provided
india
outside india
Country

# Data Conversion & Model Building

- Numerical variables have been normalized.
- Dummy variables have been created for object-type variables.
- The data has been split into training and testing sets.
- For the first step of regression, the dataset was divided into 70% for training and 30% for testing.
- RFE (Recursive Feature Elimination) was used for feature selection.
- The RFE process was run with 15 variables selected as output.
- The model achieved approximately 81% accuracy, with a sensitivity of around 70% and a specificity of about 88%.

# ROC CURVE




Using the graph above the optimal cut off was found out to be 0.35.

# Conclusion:

- If a lead visits the website multiple times, they are more likely to convert and purchase the product.
- The amount of time a lead spends on the website is also crucial; more time indicates a higher likelihood of conversion.
- The source of the lead significantly impacts the conversion rate, with sources like Google, direct traffic, organic search, and the Welingak website being particularly influential.
- The last activity of the lead, such as an SMS or Olark chat conversation, can provide insight into their level of engagement.
- If the lead's current occupation is "working professional," it suggests that they may have a strong interest in the product for career advancement, indicating a higher potential for conversion.

# THANK YOU