LEAD SCORE CASE STUDY

SHANTANU SINGH

Introduction

Objective:

- Develop a predictive model to assign lead scores between 0 and 100 for prioritizing potential leads.

Scope:

- Data cleaning
- Exploratory data analysis(EDA)
- Logistic regression
- Modeling Evaluation

Business Context:

- X Education: Sells online courses to industry professionals
- .Current Challenge: Low lead conversion rate (~30%).
- Goal: Identify 'Hot Leads' to improve conversion rates by focusing sales efforts on potential leads.

Solution Methodology

Steps:

- Data cleaning and manipulation
- Handle duplicates and missing values
- Drop irrelevant columns
- Impute necessary values
- Manage outliers

EDA

- Univariate and bivariate analysis

Data Transformation

- Feature scaling and encoding

Model Building

- Logistic regression

Model Validation and Presentation.

Data Cleaning

Initial Dataset:

- 9240 rows, 37 columns

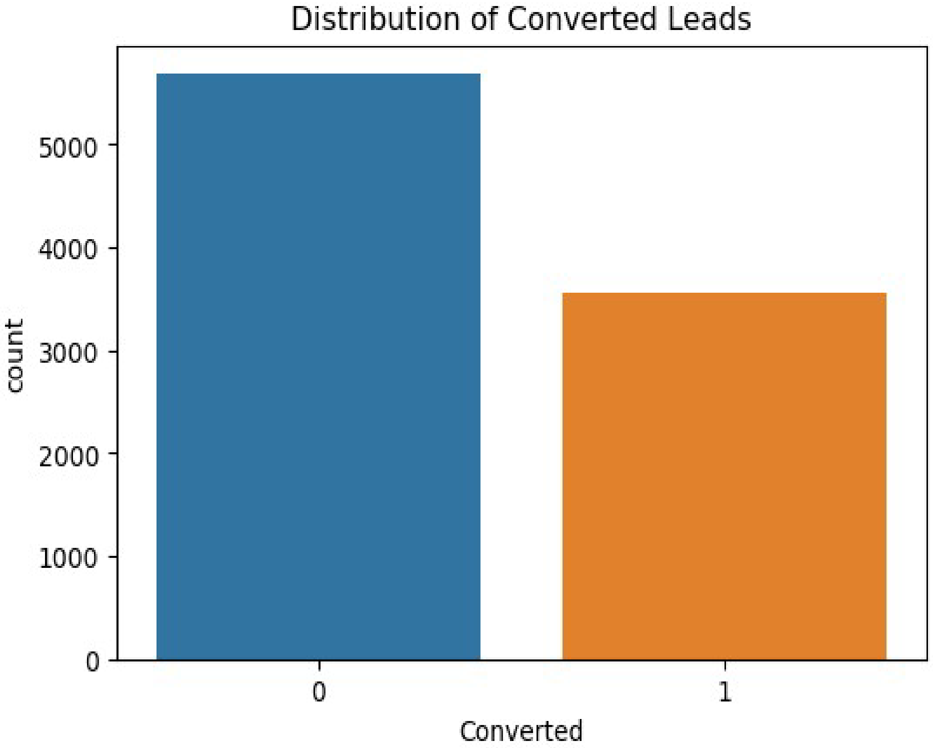
Cleaning Steps:

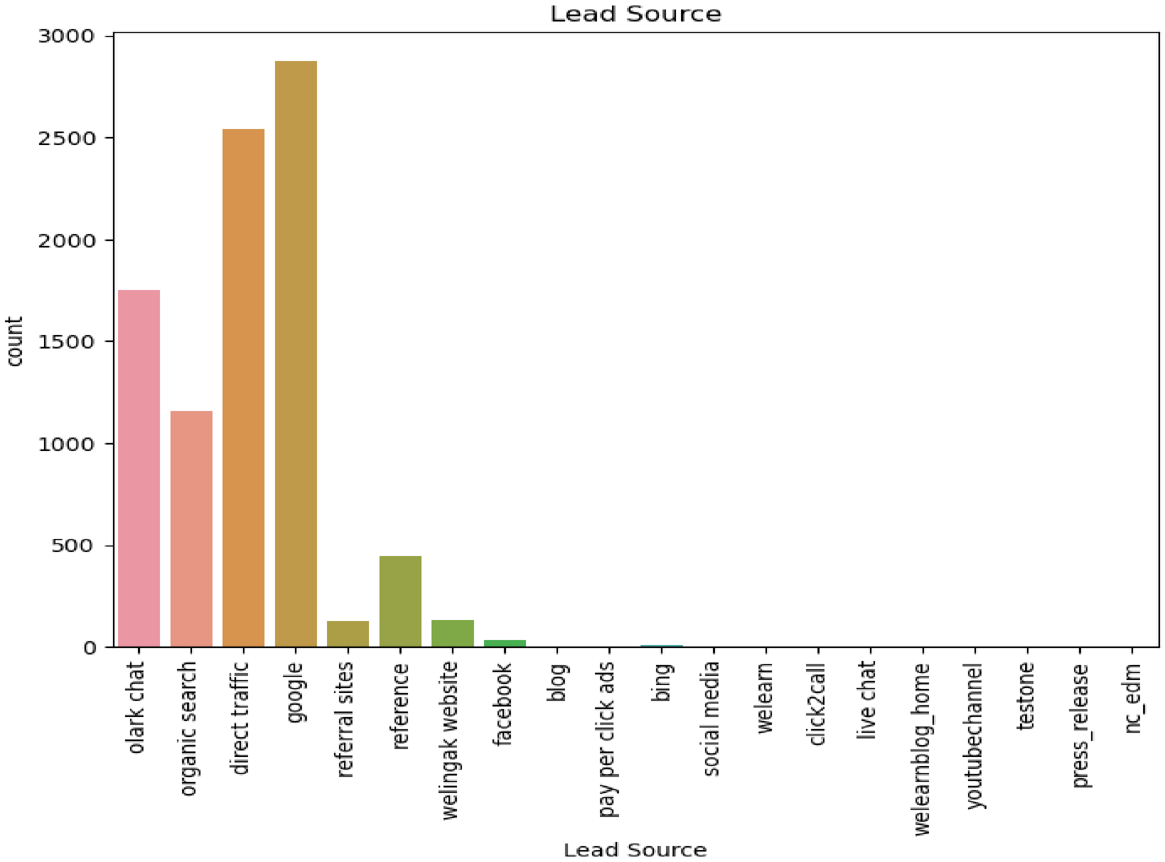
- Removed columns with single values or irrelevant information.
- Dropped columns with over 40% missing values.
- Imputed missing values for key features.

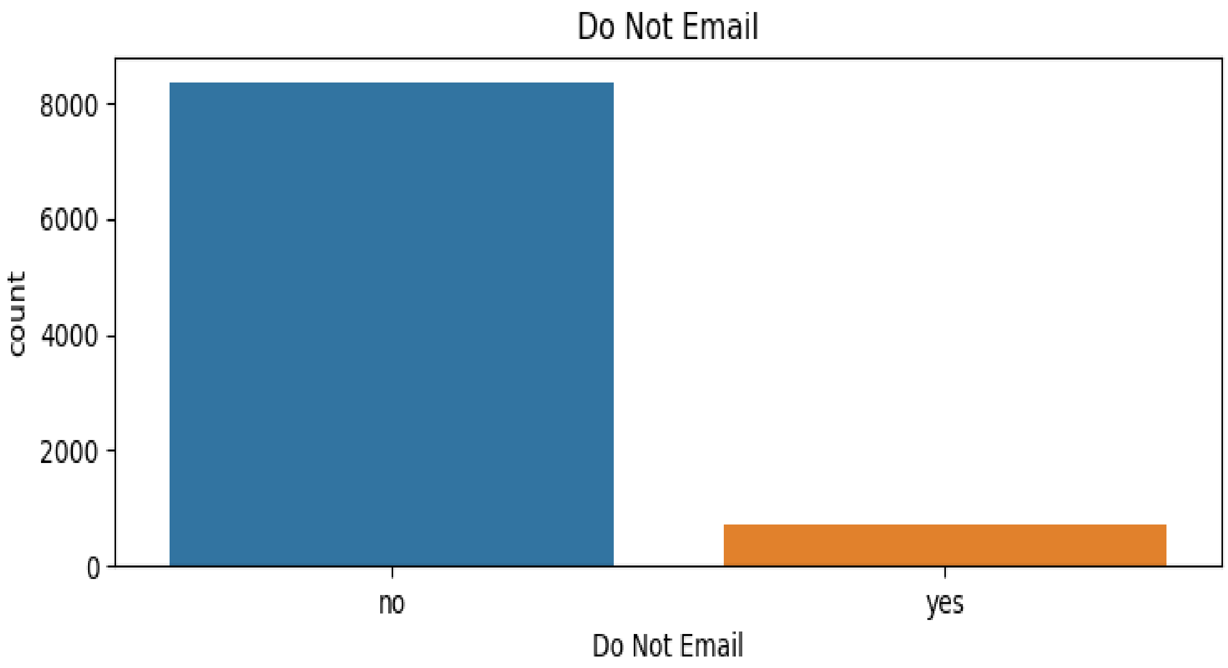
Result:

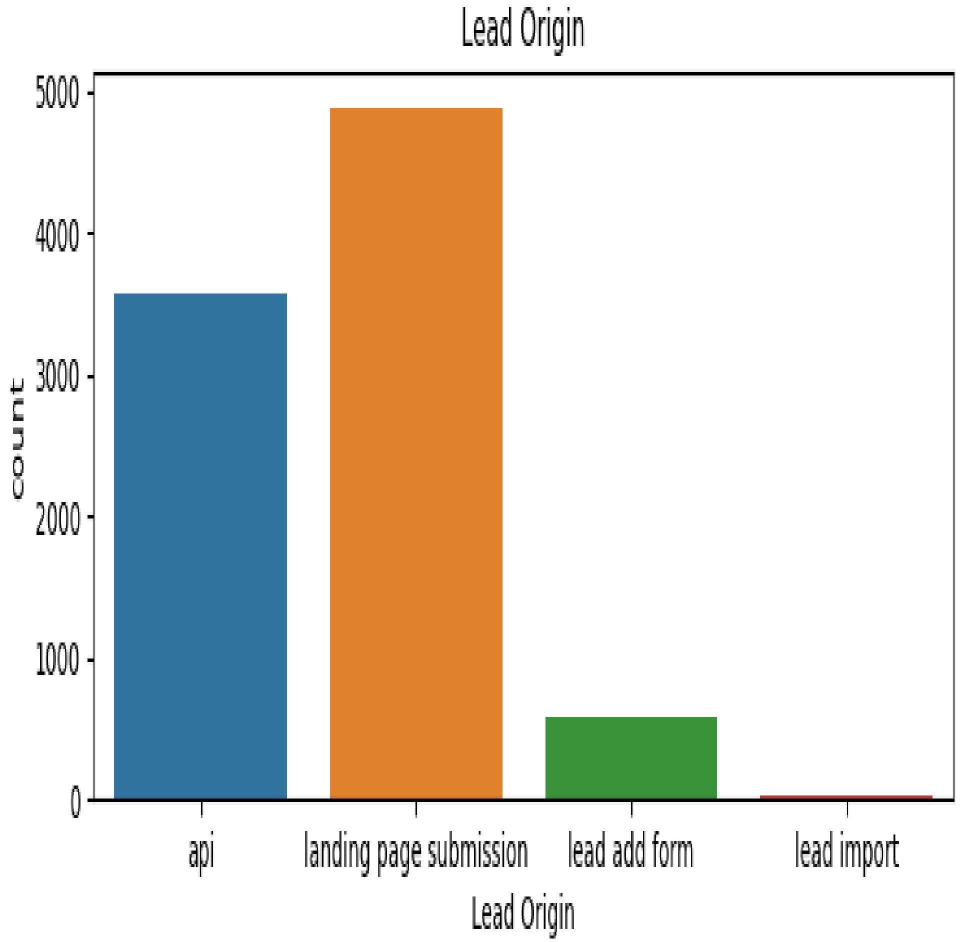
- Clean dataset with essential features retained.

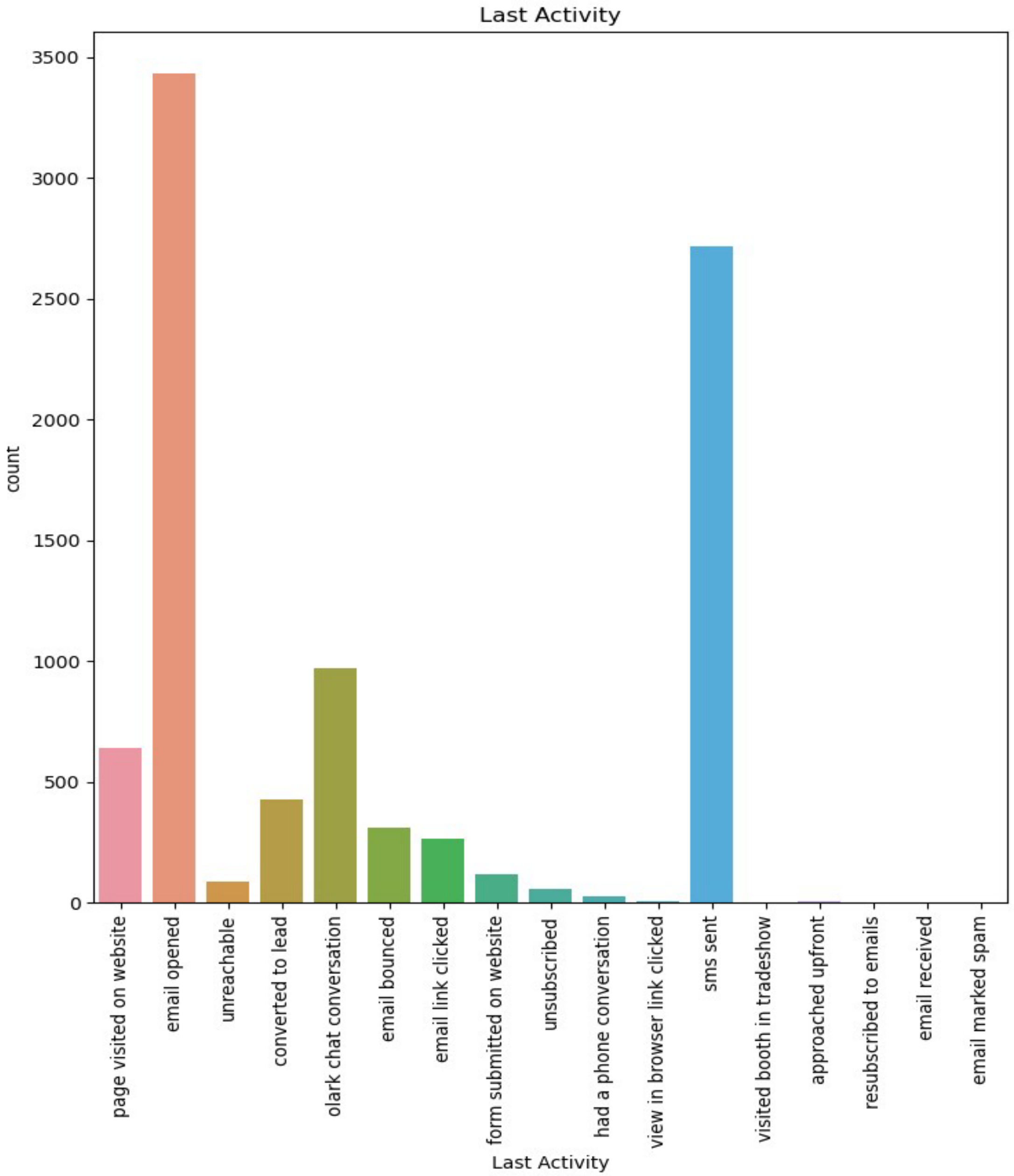
Exploratory Data Analysis (EDA)



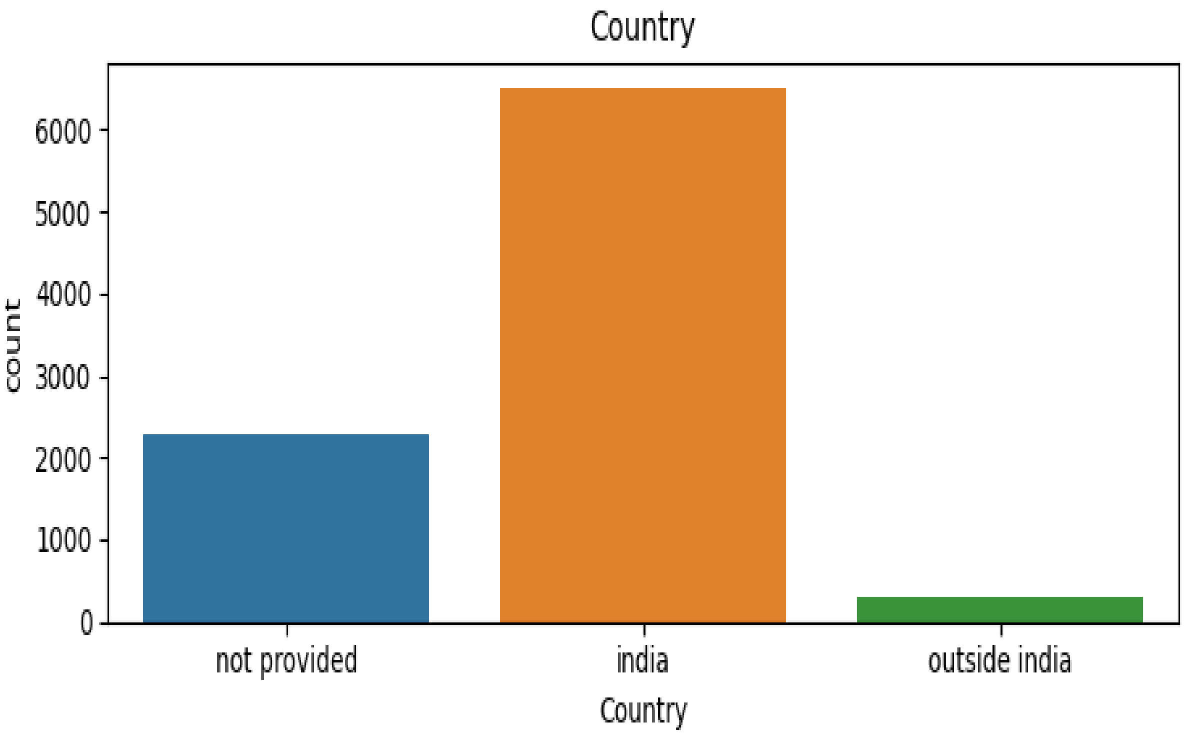



Do Not Email
count
8000
6000
4000
2000
0
no
yes
Do Not Email

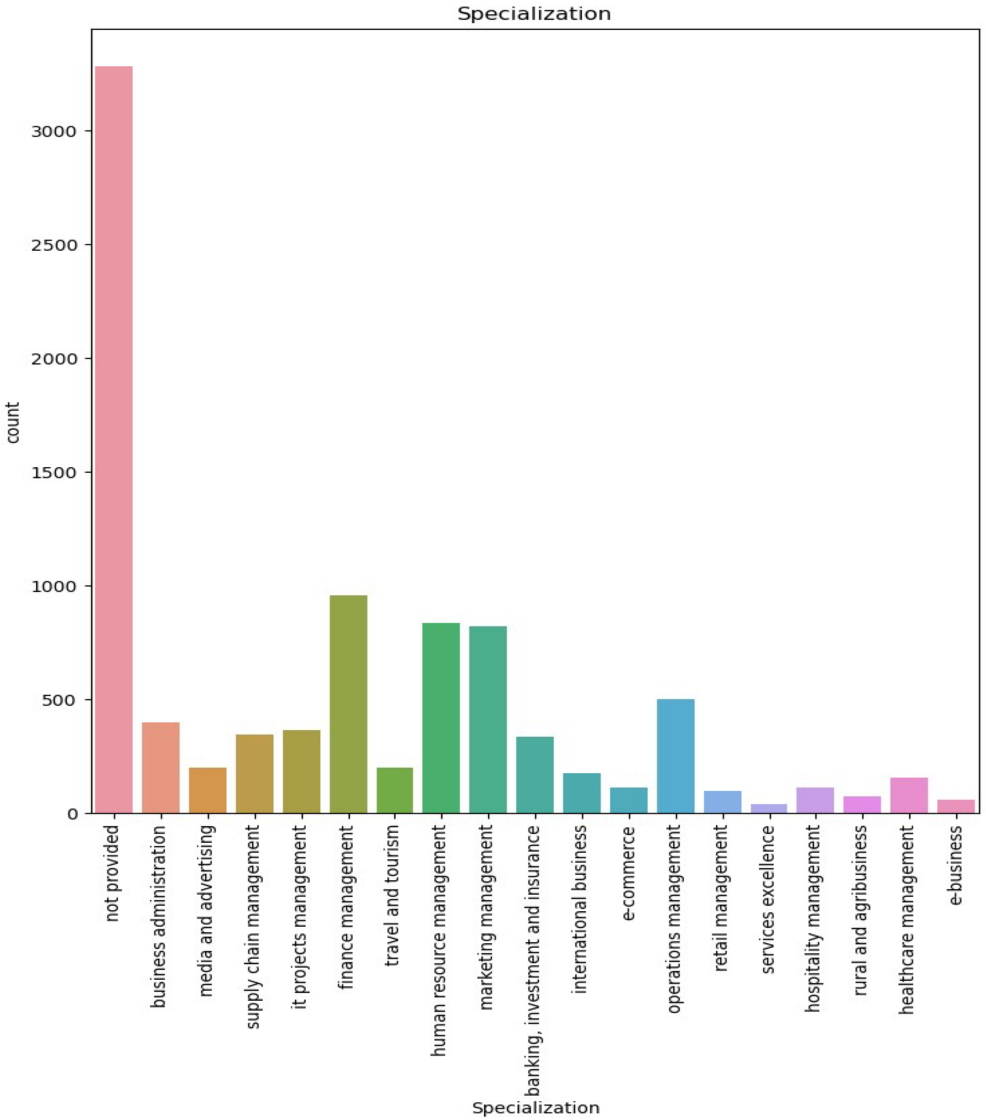

Lead Origin
count
5000
4000
3000
2000
1000
0
api
landing page submission
lead add form
lead import
Lead Origin



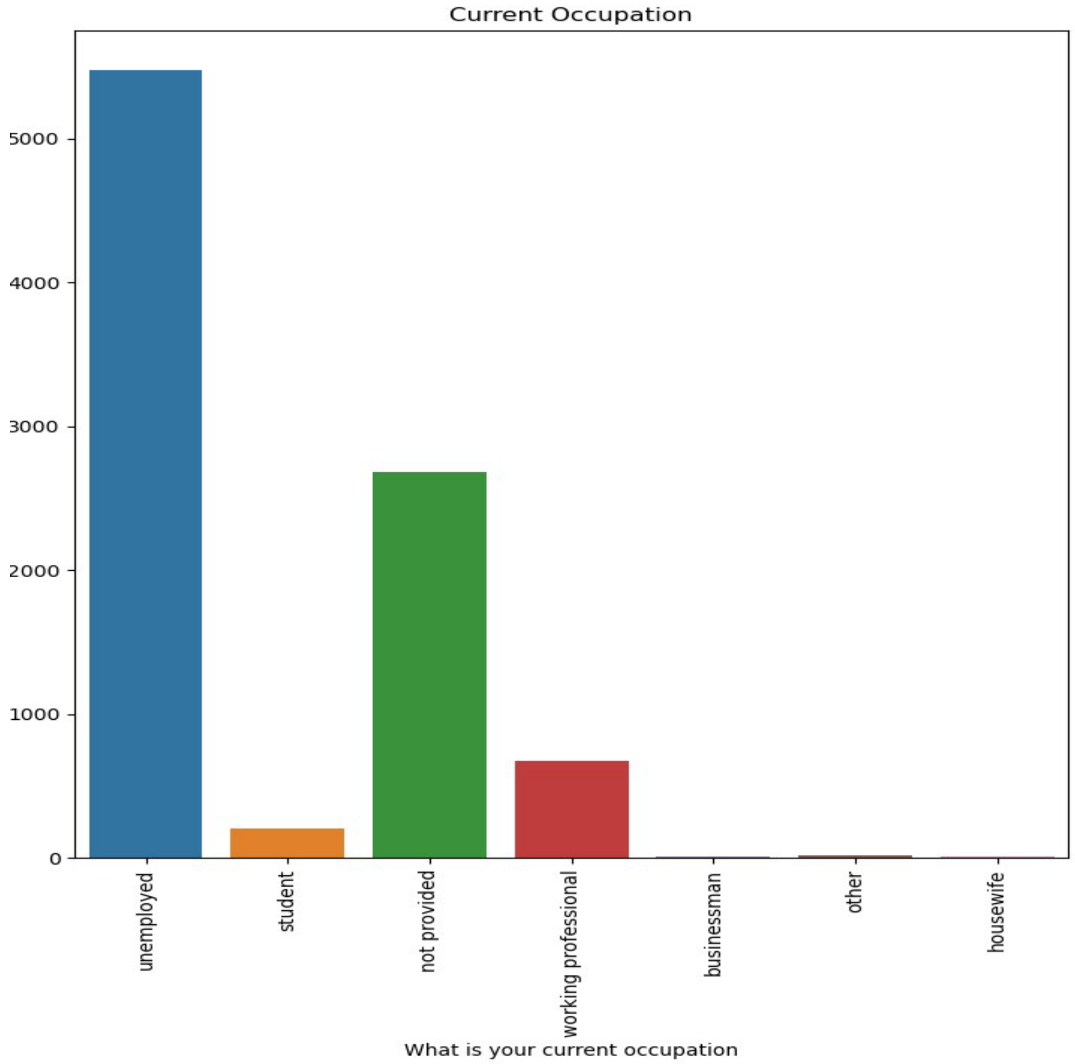
Last Activity
count
3500
3000
2500
2000
1500
1000
500
0
page visited on website
email opened
unreachable
converted to lead
olark chat conversation
email bounced
email link clicked
form submitted on website
unsubscribed
had a phone conversation
view in browser link clicked
sms sent
visited booth in tradeshow
approached upfront
resubscribed to emails
email received
email marked spam
Last Activity



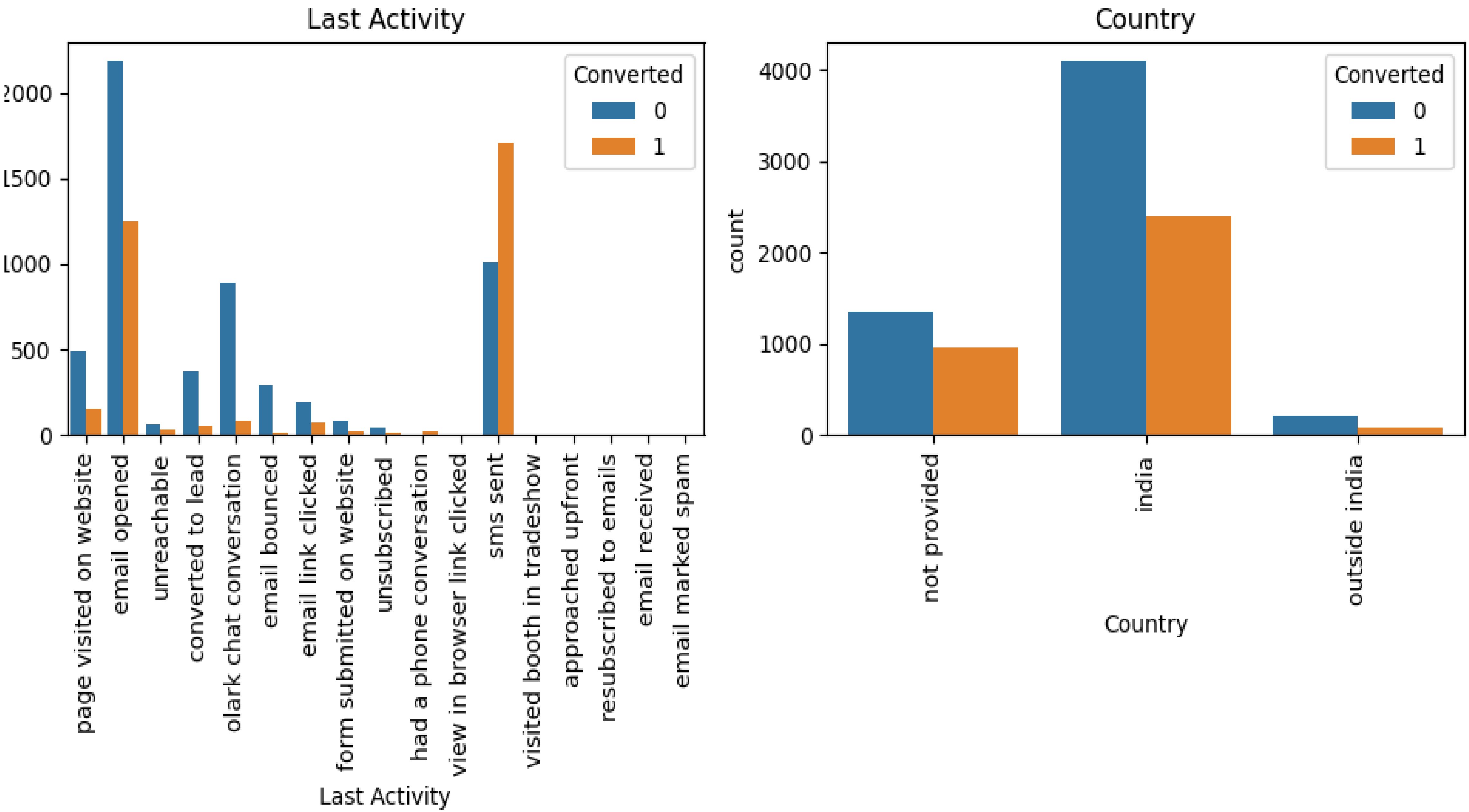
Country
count
6000
5000
4000
3000
2000
1000
0
not provided
india
outside india
Country

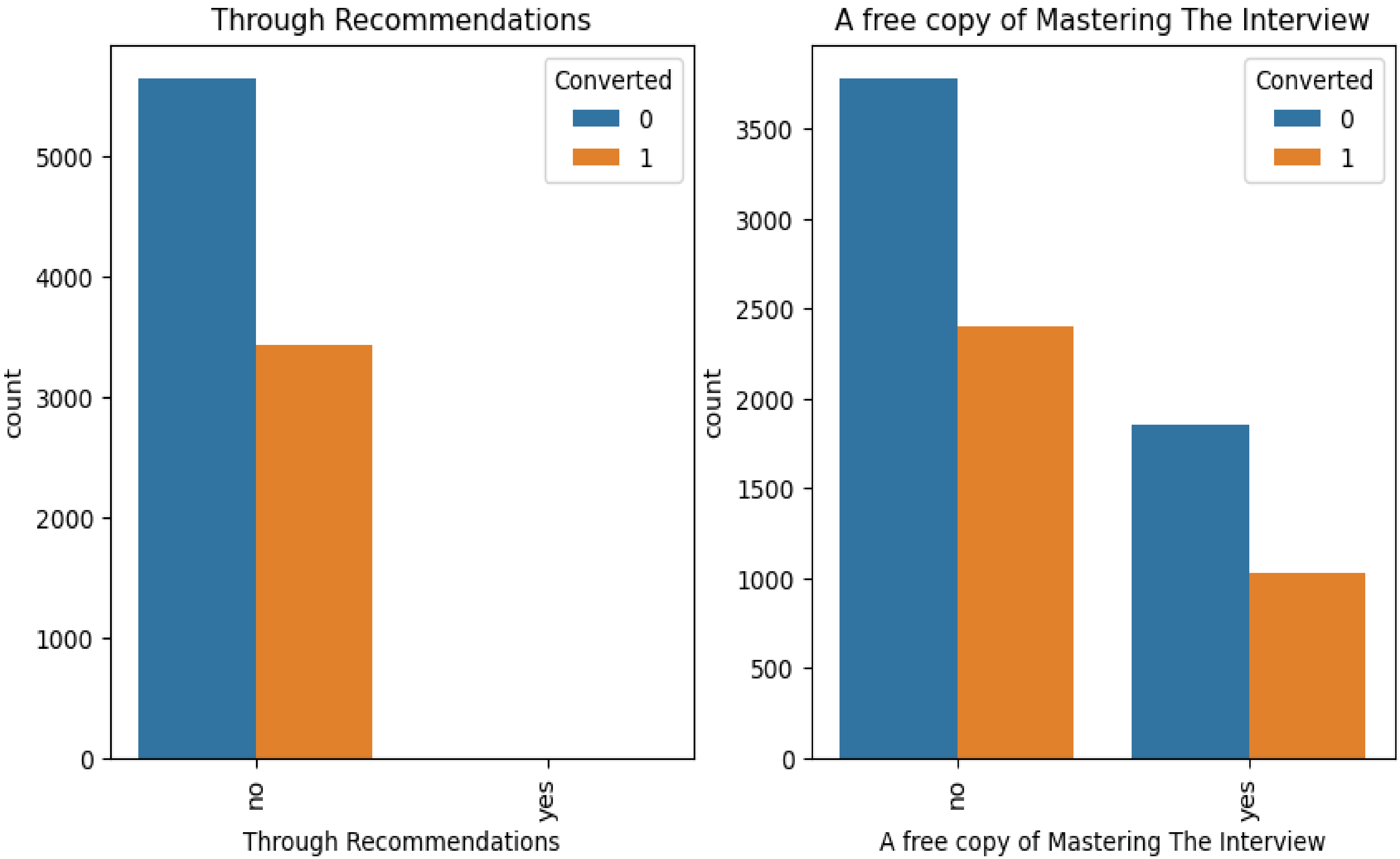


Specialization



Current Occupation


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


Through Recommendations
Converted
0
1
count
0
1000
2000
3000
4000
5000
no
yes
Through Recommendations
A free copy of Mastering The Interview
Converted
0
1
count
0
500
1000
1500
2000
2500
3000
3500
no
yes
A free copy of Mastering The Interview

Model Building-Logistic Regression

Process:

Split data into training (70%) and testing (30%) sets.

Performed Recursive Feature Elimination (RFE) for feature selection.

Model: Logistic regression to predict the probability of conversion.

Model Evaluation

 Splitting the Data into Training and Testing Sets

 The first basic step for regression is performing a train-test split, we have chosen 70:30

ratio.

 Use RFE for Feature Selection

 Running RFE with 15 variables as output

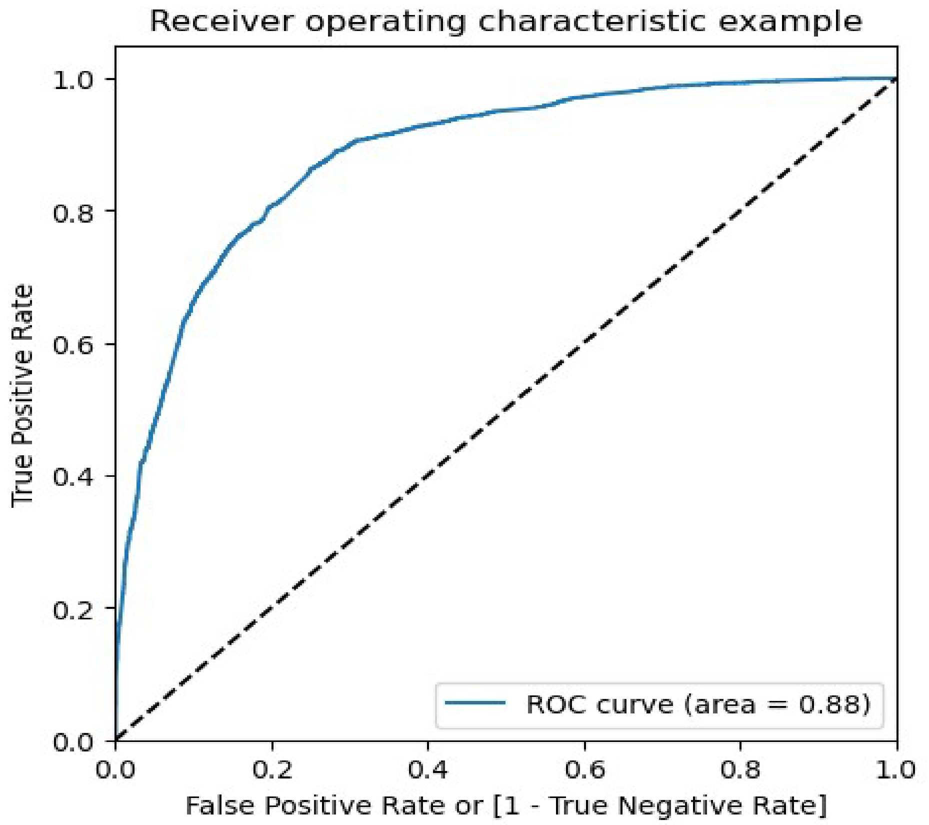
 Building Model by removing the variable whose p- value is greater than 0.05 and vif

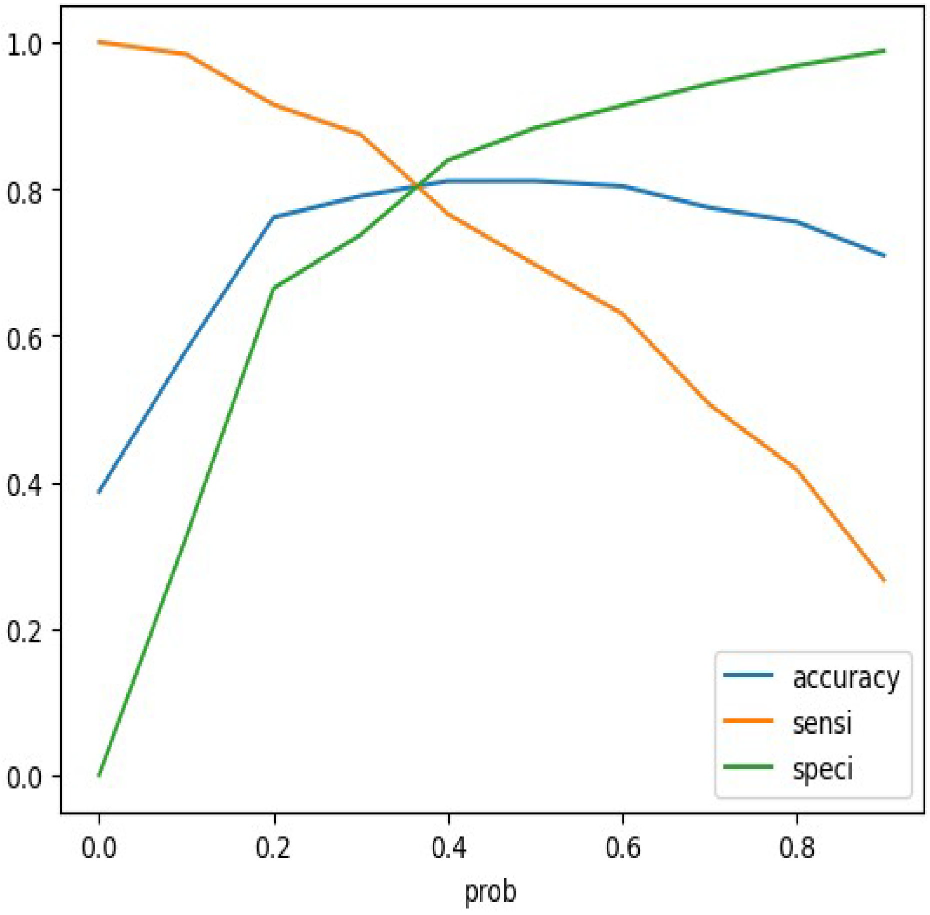
value is greater than 5

 Predictions on test data set

 Overall accuracy 81%

ROC Curve





1. Finding Optimal Cut off Point
2. Optimal cut off probability is that
3. probability where we get balanced sensitivity and specificity.
4. From the second graph it is visible that the optimal cut off is at 0.35.

Summary:

Successfully navigated through data cleaning, exploratory analysis, outlier handling, and model building stages.

Demonstrated significant correlations between certain features (e.g., Total Time Spent on Website) and lead conversion.

Developed a robust logistic regression model with high predictive accuracy and ROC AUC score.

Final Thoughts:

The project highlights the potential to significantly improve conversion rates and operational efficiency through data-driven lead management strategies.

Continued dedication to innovation and adaptation will ensure sustained improvement in lead conversion and business growth.