Lead Score Case Study

Sandeep Yadav
Ragini Maurya
Tarun Kumar Pal

Problem Description

- X Education focuses in offering online courses designed with industry experts in mind, with the goal of improving their professions and skill sets.
- X Education receives a lot of leads every day, but only 30% of those leads end up as sales.
- The organization wants to find "Hot Leads" so that it can increase productivity and increase conversion rates.
- The sales staff can concentrate on high-potential prospects by identifying "Hot Leads," which will raise the lead conversion rate.

Objective

- X Education seeks to identify the most promising leads to improve conversion rates.
- X Education aims to build and deploy a model to identify hot leads, enabling the sales team to focus on high-potential prospects and improve future lead conversion rates.

Solution Approaches

➢ Data cleaning :

Check and Handle Duplicates: Identify and remove duplicate records.

Check and Handle NA/Missing Values: Identify columns with NA/missing values.

Drop Columns: Remove columns with a high percentage of missing values and low relevance.

Imputation: Impute missing values where necessary using mean, median, or mode.

Check and Handle Outliers: Identify outliers using statistical methods and handle them appropriately to ensure data integrity.

- Exploratory Data Analysis (EDA)

  Univariate: Perform univariate data analysis by examining value counts, and distribution of each variable to understand their individual characteristics.

  Bivariate: Conduct bivariate data analysis by calculating correlation coefficients and examining patterns between variables to identify relationships and dependencies.

- Apply feature scaling for numerical features, create dummy variables for categorical data, and encode them to prepare the dataset for modeling.
- Use logistic regression as the classification technique for model building and prediction, aiming to classify and predict lead conversions.
- Model presentation.
- Conclusions.

Data Construction

The dataset initially contains 37 rows and 9,240 columns. Several preprocessing steps were performed to clean the data:

Dropped Single Value Features: Features with single values such as “Magazine,” “Receive More Updates About Our Courses,” “Update me on Supply,” and others like “Asymmetrique Profile Index,” “Asymmetrique Activity Index,” “Asymmetrique Activity Score,” “Asymmetrique Profile Score,” “Lead Profile,” “Tags,” “Lead Quality,” and “City” were removed, as they provide no useful information for analysis.

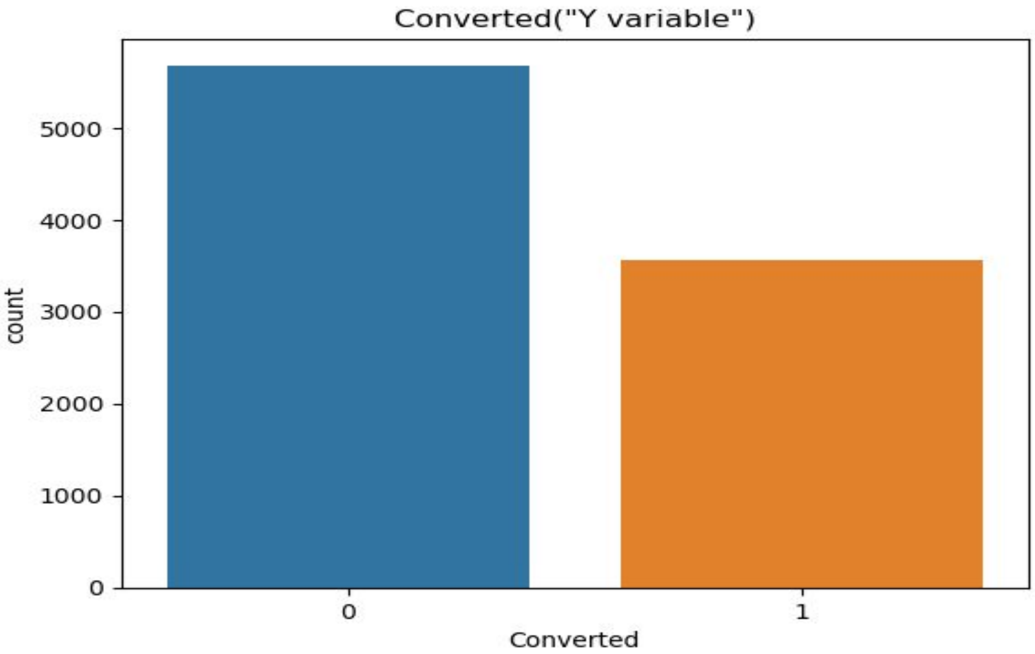
Removed Irrelevant Columns: Columns like “Prospect ID” and “Lead Number” were discarded as they were not necessary for the analysis.

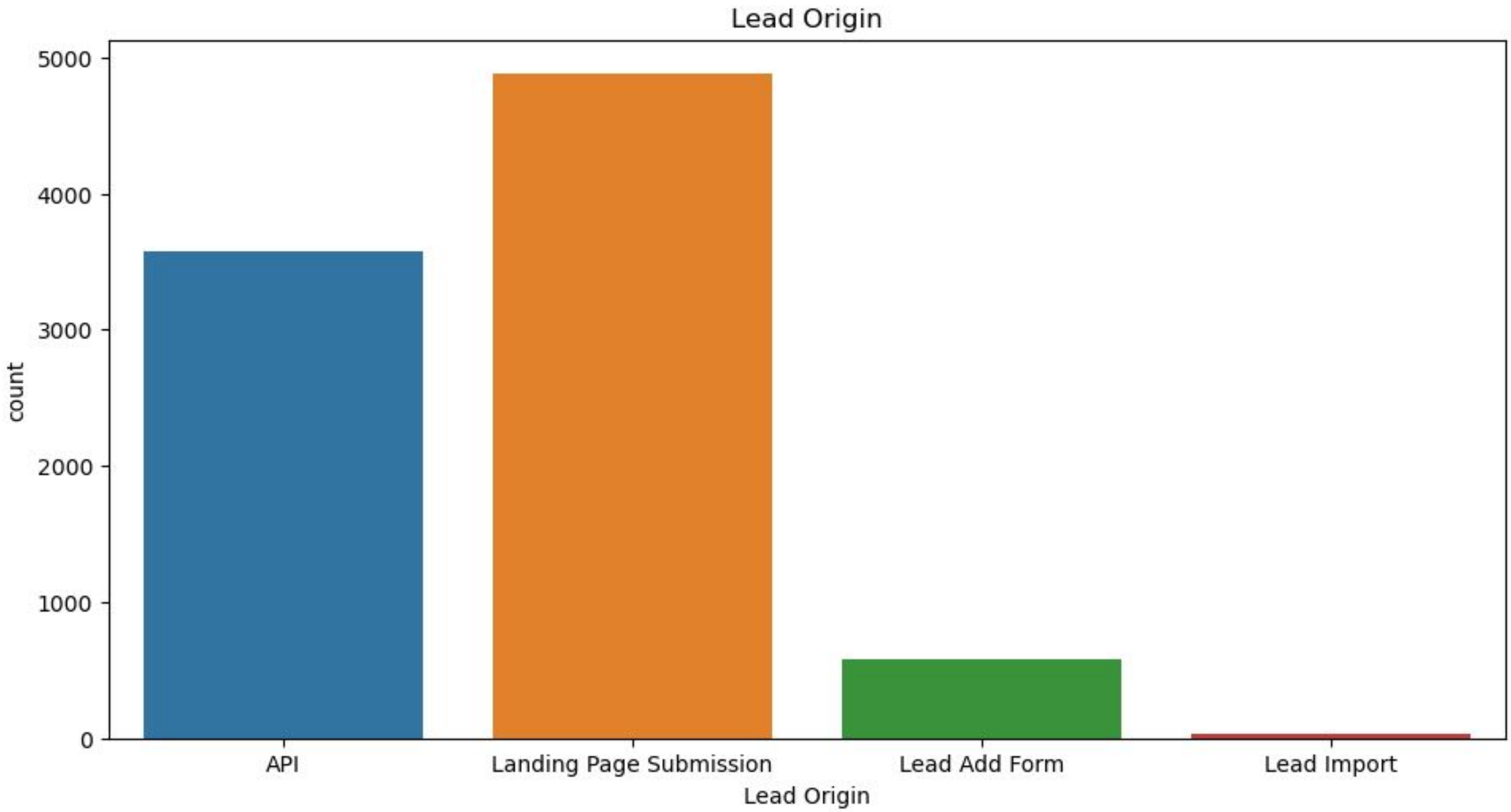
Features with Low Variance: Variables with minimal variance, such as “Do Not Call,” “What matters most to you in choosing course,” “Search,” “Newspaper Article,” “X Education Forums,” “Newspaper,” and “Digital Advertisement,” were dropped.

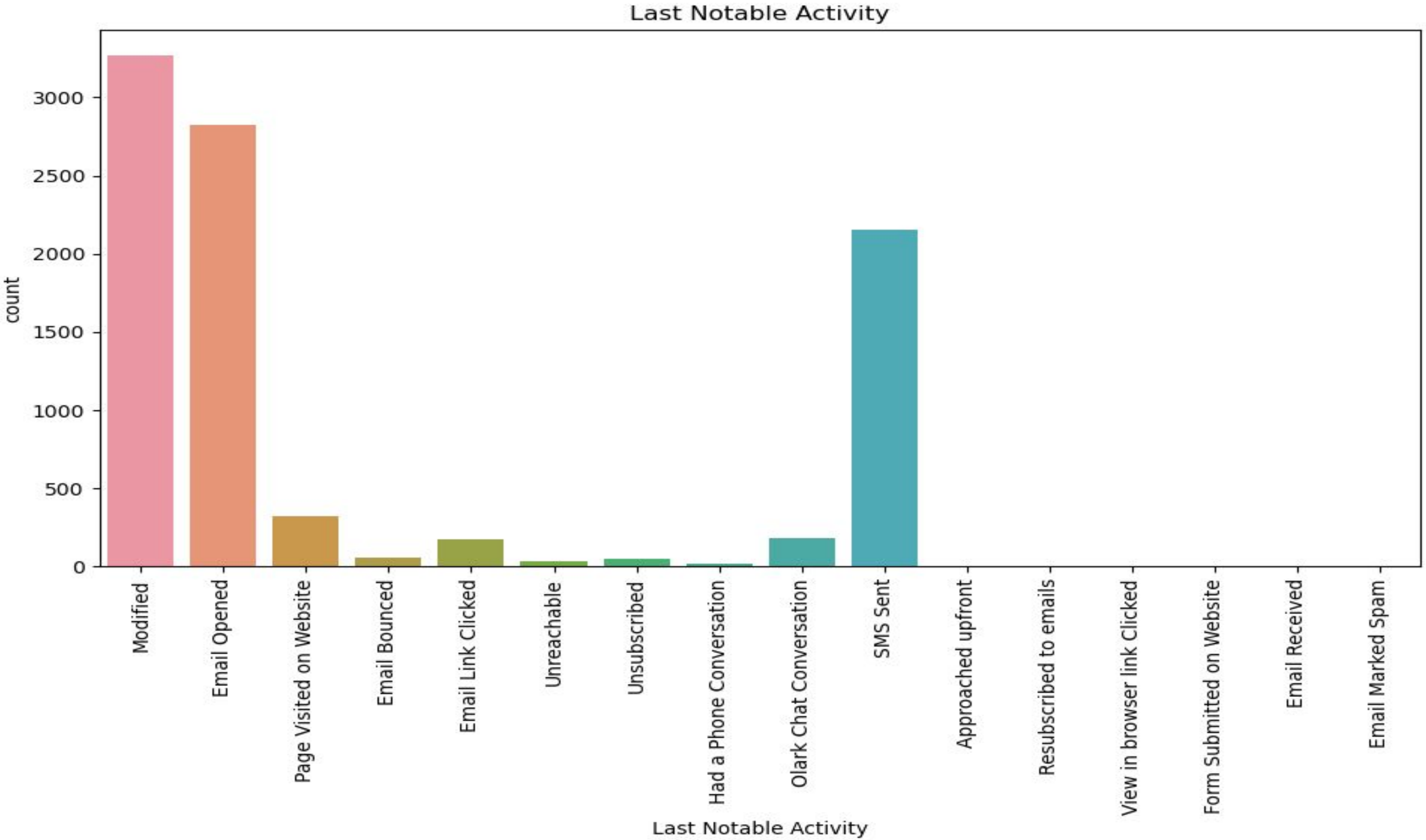
High Missing Values: Columns with over 35% missing values, including ‘How did you hear about X Education’ and ‘Lead Profile,’ were removed to improve data quality and relevance.

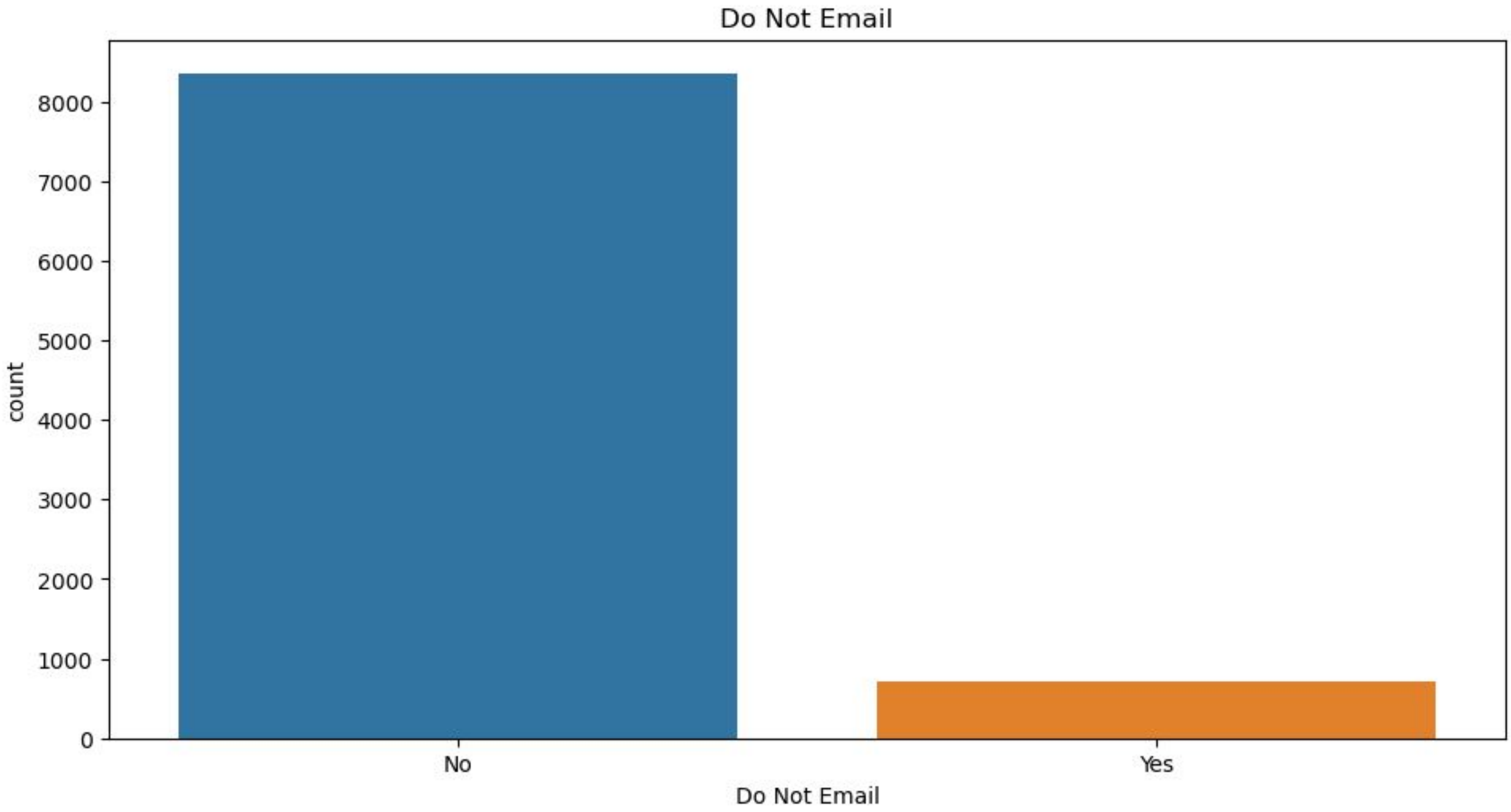
Exploratory Data Analysis (EDA)

Understanding the dataset's structure, uncovering patterns, and preparing data for further analysis or modeling.

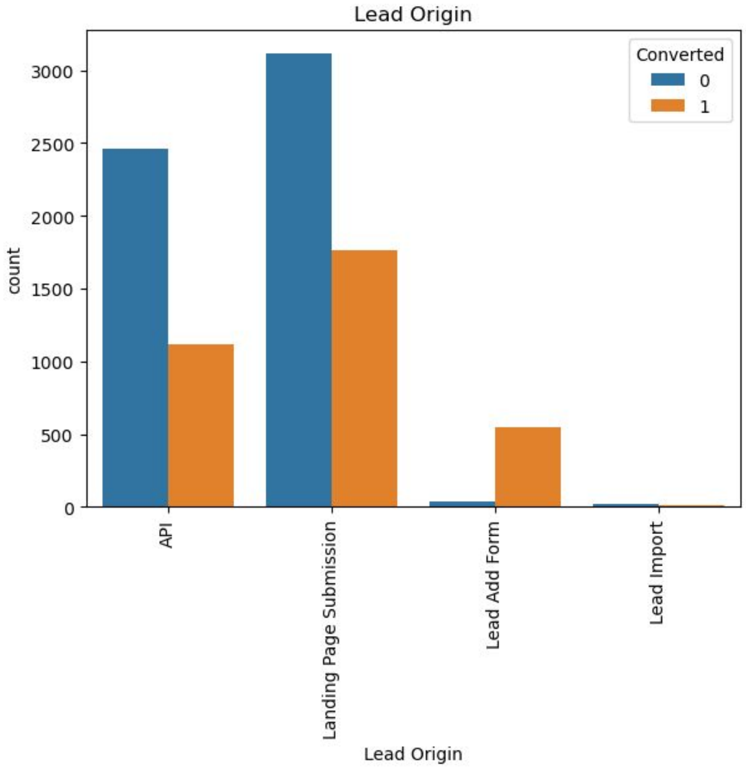


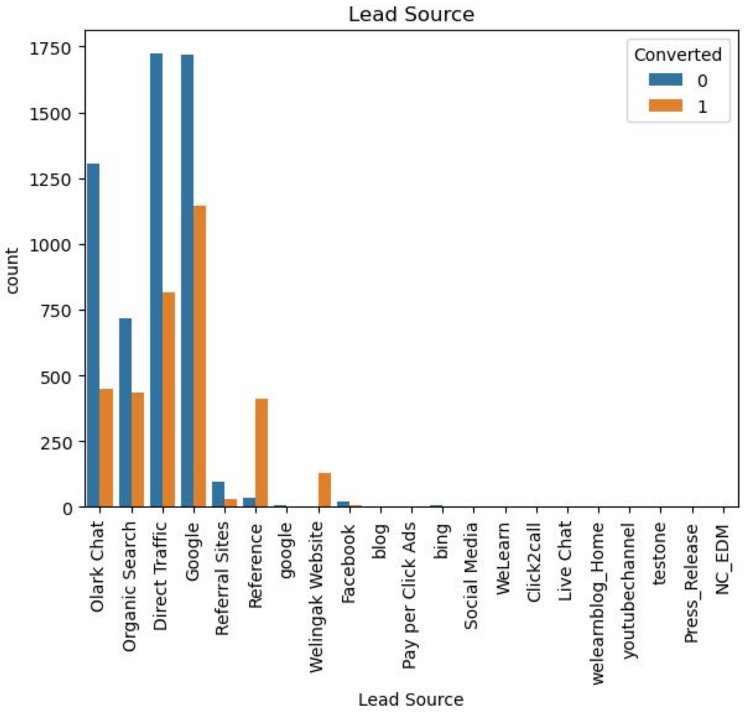

Lead Origin
count
5000
4000
3000
2000
1000
0
API
Landing Page Submission
Lead Add Form
Lead Import
Lead Origin

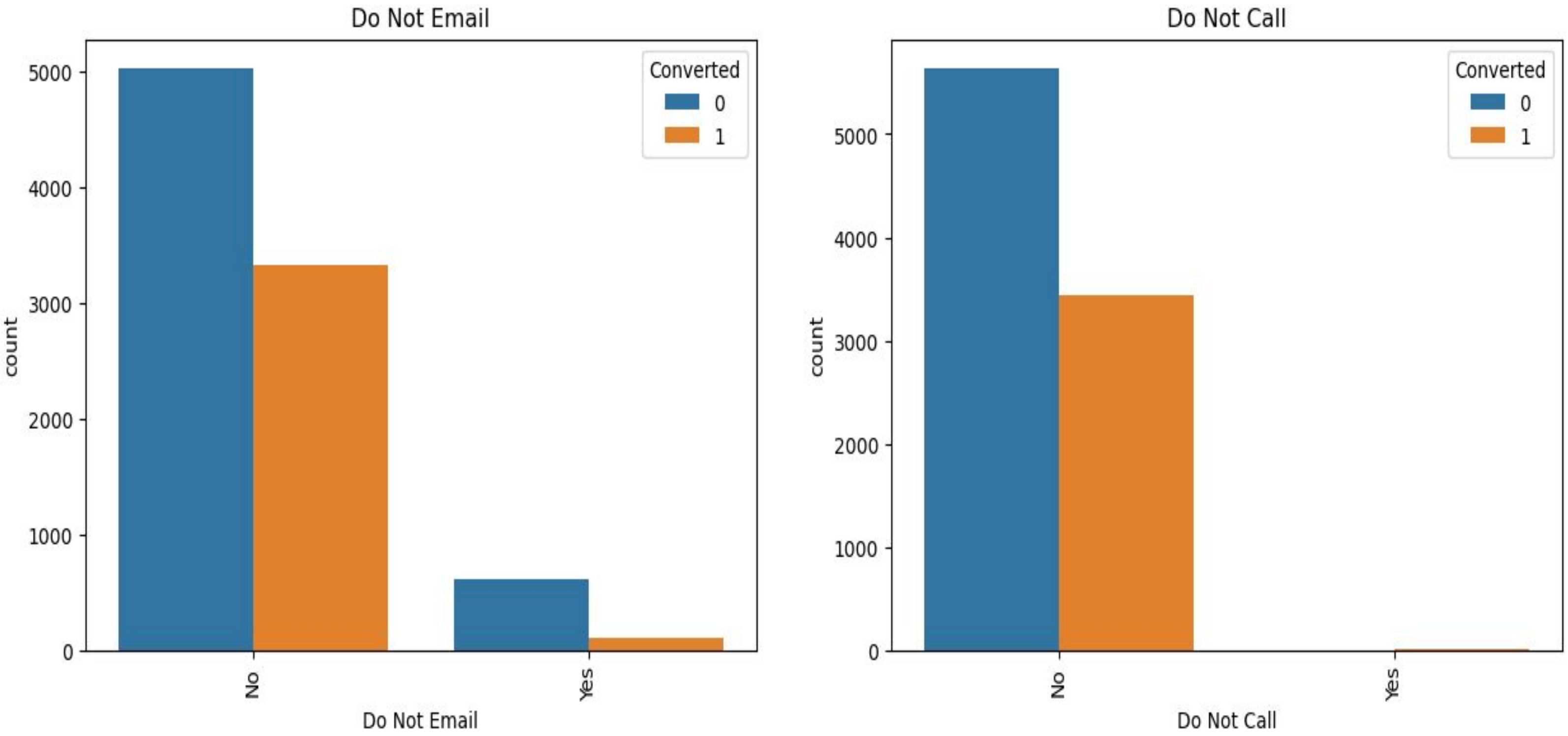

Last Notable Activity
count
3000
2500
2000
1500
1000
500
0
Modified
Email Opened
Page Visited on Website
Email Bounced
Email Link Clicked
Unreachable
Unsubscribed
Had a Phone Conversation
Olark Chat Conversation
SMS Sent
Approached upfront
Resubscribed to emails
View in browser link Clicked
Form Submitted on Website
Email Received
Email Marked Spam
Last Notable Activity

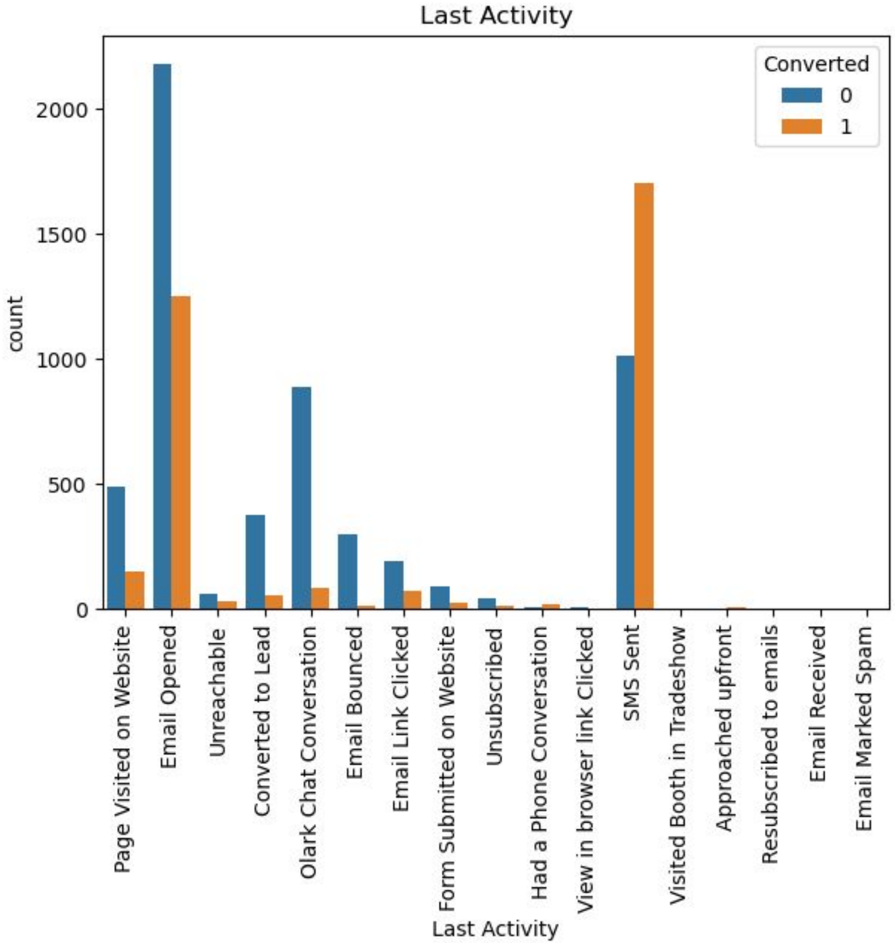

Do Not Email
count
8000
7000
6000
5000
4000
3000
2000
1000
0
No
Yes
Do Not Email

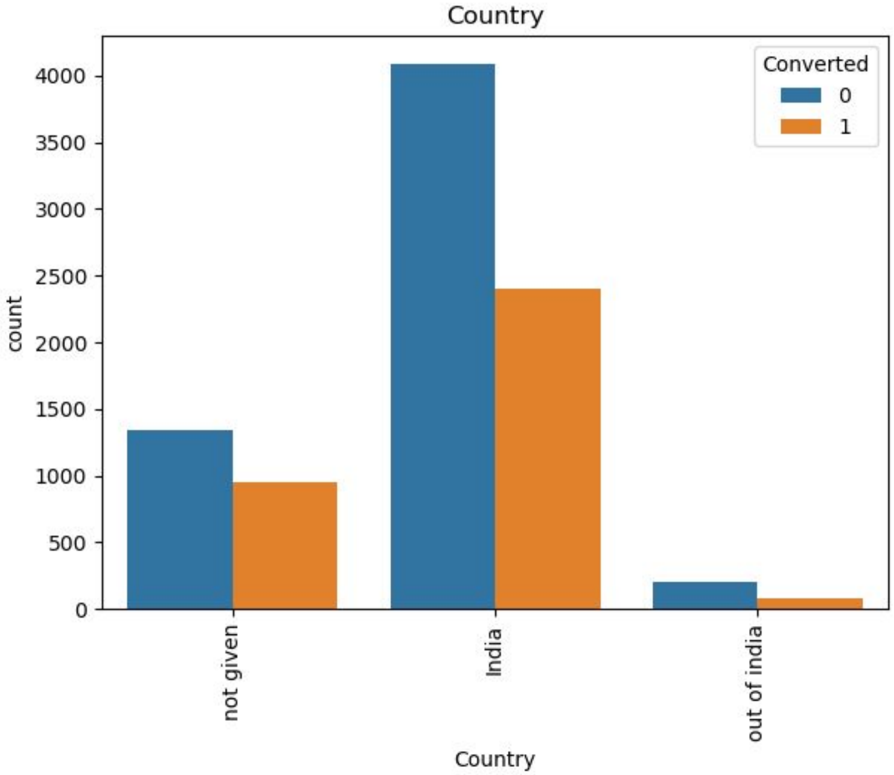
Categorical Variable Between Converted

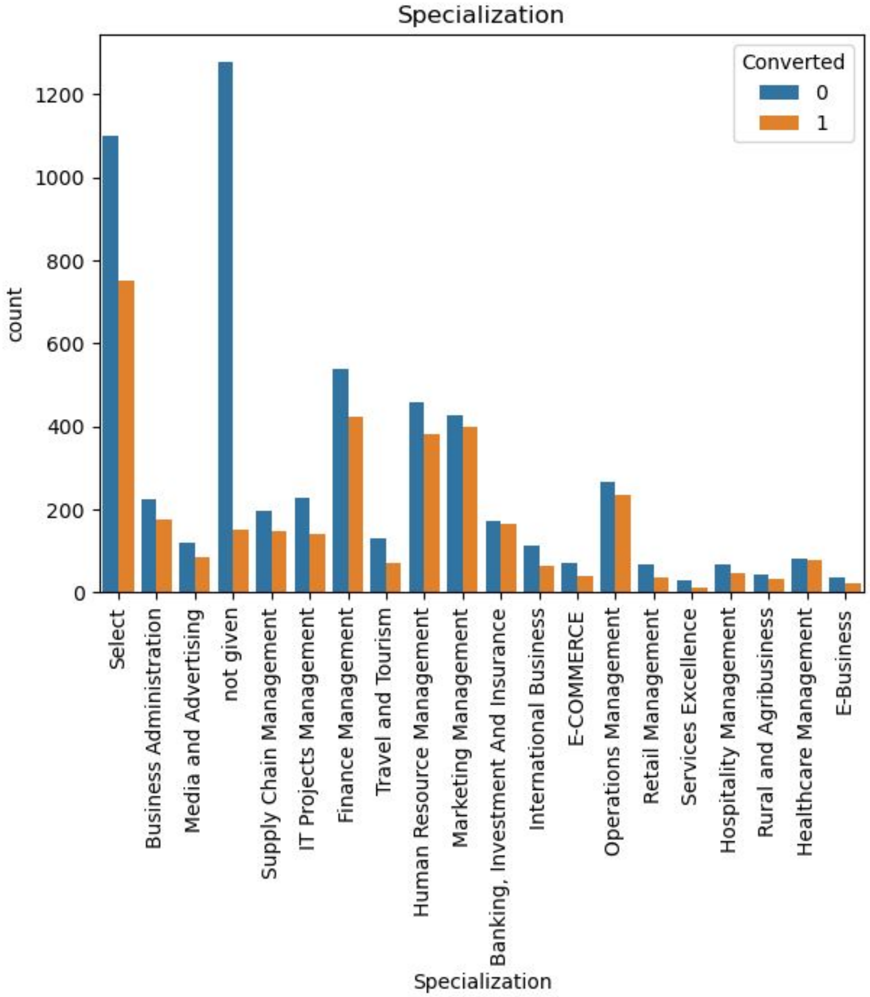


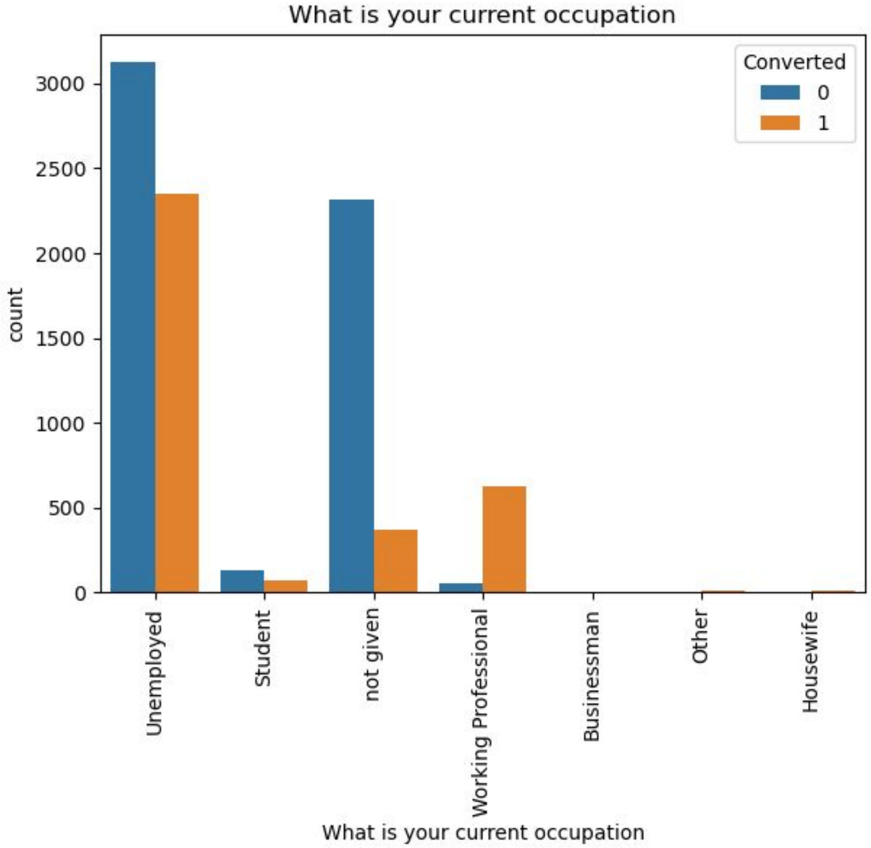


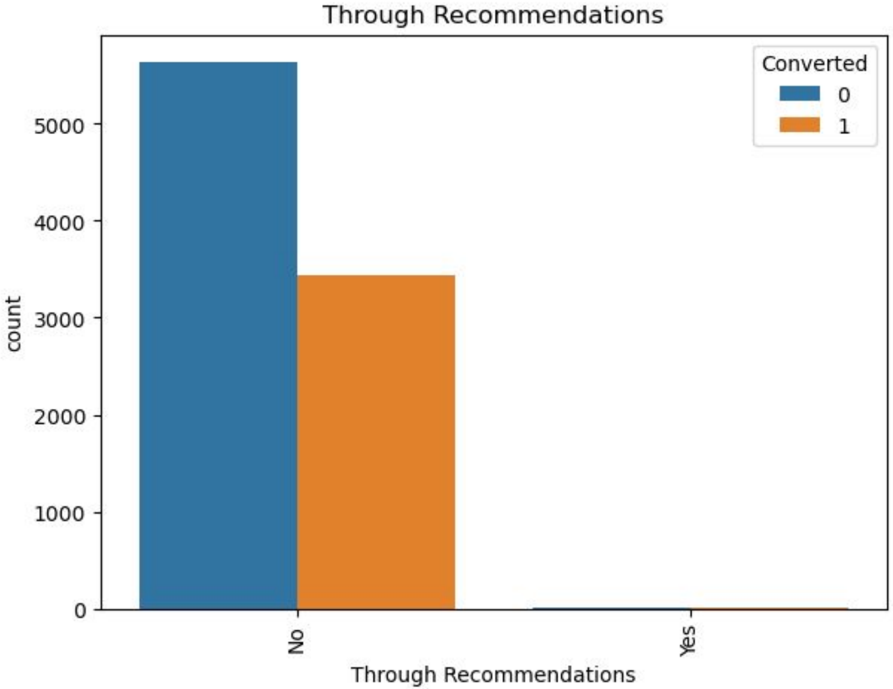

Do Not Email
Converted
0
1
count
5000
4000
3000
2000
1000
0
No
Yes
Do Not Email
Do Not Call
Converted
0
1
count
5000
4000
3000
2000
1000
0
No
Yes
Do Not Call

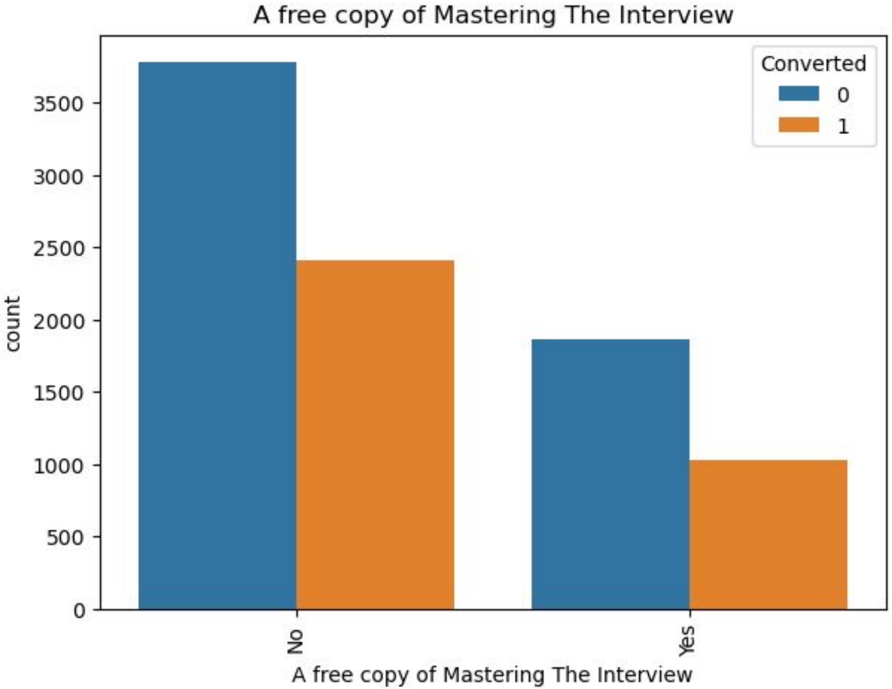

Last Activity
Converted
0
1
count
2000
1500
1000
500
0
Page Visited on Website
Email Opened
Unreachable
Converted to Lead
Olark Chat Conversation
Email Bounced
Email Link Clicked
Form Submitted on Website
Unsubscribed
Had a Phone Conversation
View in browser link Clicked
SMS Sent
Visited Booth in Tradeshow
Approached upfront
Resubscribed to emails
Email Received
Email Marked Spam
Last Activity


Country
Converted
0
1
count
4000
3500
3000
2500
2000
1500
1000
500
0
not given
India
out of india
Country


Specialization
Converted
0
1
count
0
200
400
600
800
1000
1200
Select
Business Administration
Media and Advertising
not given
Supply Chain Management
IT Projects Management
Finance Management
Travel and Tourism
Human Resource Management
Marketing Management
Banking, Investment And Insurance
International Business
E-COMMERCE
Operations Management
Retail Management
Services Excellence
Hospitality Management
Rural and Agribusiness
Healthcare Management
E-Business
Specialization


What is your current occupation
Converted
0
1
count
0
500
1000
1500
2000
2500
3000
Unemployed
Student
not given
Working Professional
Businessman
Other
Housewife
What is your current occupation


Through Recommendations
Converted
0
1
5000
4000
3000
2000
1000
0
count
No
Yes
Through Recommendations


A free copy of Mastering The Interview
Converted
0
1
3500
3000
2500
2000
1500
1000
500
0
count
No
Yes
A free copy of Mastering The Interview

Data Transformation

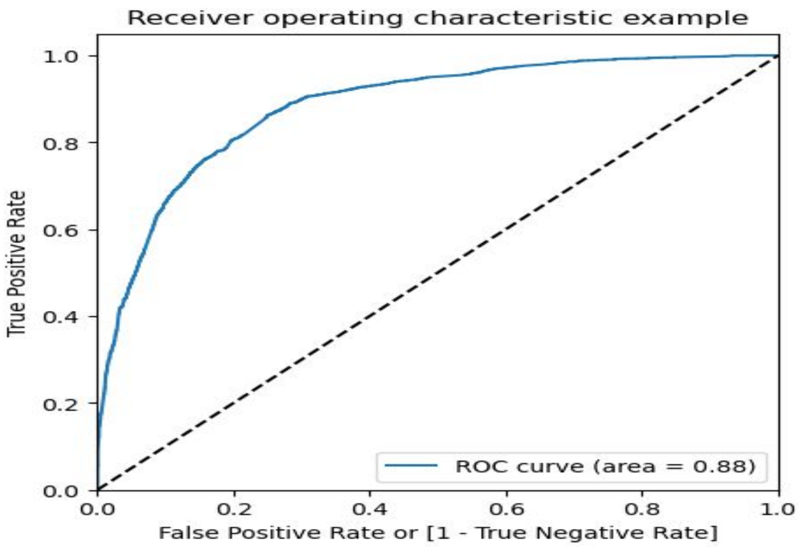
- ➢ Standardize numerical variables.
- ➢ Dummy variables are created for object-type variables.
- ➢ After preprocessing, the dataset contains 8,792 rows and 43 columns.
- ➢ Thorough preprocessing ensures the data is clean and structured.
- ➢ The data is ready for building and evaluating machine learning models.
- ➢ Normalizing numerical data improves model accuracy.
- ➢ Converting categorical variables into a suitable format enhances model reliability.
- ➢ This process helps in predicting and analyzing potential leads for X Education.

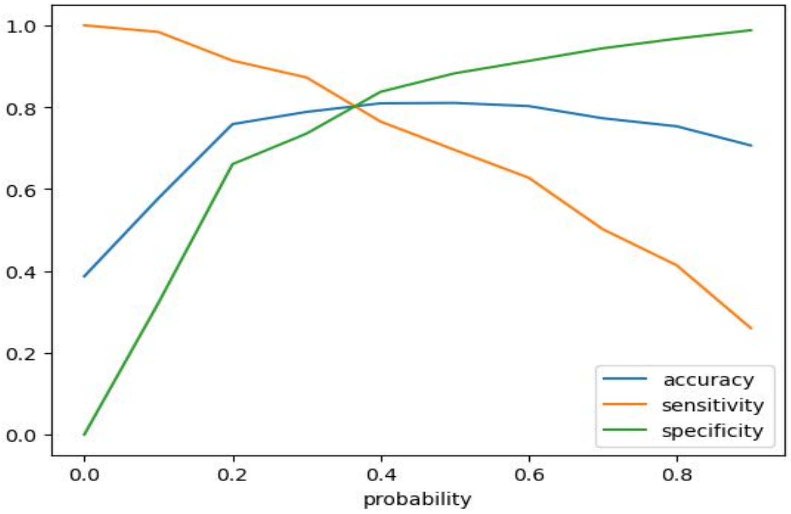
Model Building

- **Train-Test Split**: Perform a train-test split with a 70:30 ratio.
- **Feature Selection**: Use Recursive Feature Elimination (RFE).
- **RFE Output**: Select 15 variables using RFE.
- **Model Building**: Remove variables with p-value > 0.05 and VIF > 5.
- **Predictions**: Make predictions on the test data set.
- **Accuracy**: Achieve an overall accuracy of 80%.

ROC Curve

- **Optimal Cut Off Point**: Identify the optimal cut off probability.
- **Balanced Metrics**: Look for the probability where sensitivity and specificity are balanced.
- **Graph Analysis**: From the second graph, the optimal cut off is 0.35.





Conclusion

- Total Time Spent on Website:

The total time a lead spends on the website is the most crucial factor. Longer engagement typically indicates higher interest and potential for conversion.

- Total Number of Visits:

The frequency of visits by a lead reflects their level of interest. More visits generally correlate with a higher likelihood of conversion, as it shows persistent engagement.

- Lead Source:

Google: Leads originating from Google search are highly valuable, indicating targeted interest.
Direct Traffic: Direct traffic signifies strong brand recognition or prior engagement.
Organic Search: Leads from organic search show a genuine interest driven by content relevance.
Welingak Website: Specific website sources like Welingak also play a significant role, indicating targeted interest.

- ➢ Last Activity:

SMS: Recent SMS interactions are crucial, as they are often timely and personal.

Olark Chat Conversation: Conversations through chat platforms like Olark are indicative of high engagement and interest.

- ➢ Lead Origin:

Leads from the "Lead add format" are more likely to convert, suggesting that the format of the lead capture influences potential.

- ➢ Current Occupation:

Leads who are working professionals generally have a higher likelihood of conversion, likely due to their immediate need for further education or career advancement.

By prioritizing these factors, X Education can strategically target potential buyers and significantly improve their conversion rates.