LEAD SCORE CASE STUDY

GROUP MEMBERS

DEVENDRA NATH JHA

DISHA BHAISORA

VERSHA VERMA

PROBLEM STATEMENT

- X Education specializes in providing online courses tailored to meet the needs of industry professionals.
- Despite acquiring a significant number of leads, X Education struggles with a low lead conversion rate. For instance, out of 100 leads obtained in a day, only around 30 of them are successfully converted into customers.
- In order to streamline the process and improve efficiency, X Education aims to identify the most promising leads, often referred to as 'Hot Leads.'
- By successfully identifying these high-potential leads, X Education expects that the lead conversion rate will increase. This improvement can be attributed to the sales team's focused efforts on nurturing and communicating with the potential leads, rather than indiscriminately reaching out to everyone.

BUSINESS OBJECTIVE

- X Education seeks to identify the most promising leads among their prospects.
- To achieve this goal, they plan to develop a model that can accurately identify hot leads.
- Once the model is developed, X Education intends to deploy it for future use, enabling them to consistently identify and focus their efforts on the leads with the highest potential for conversion.

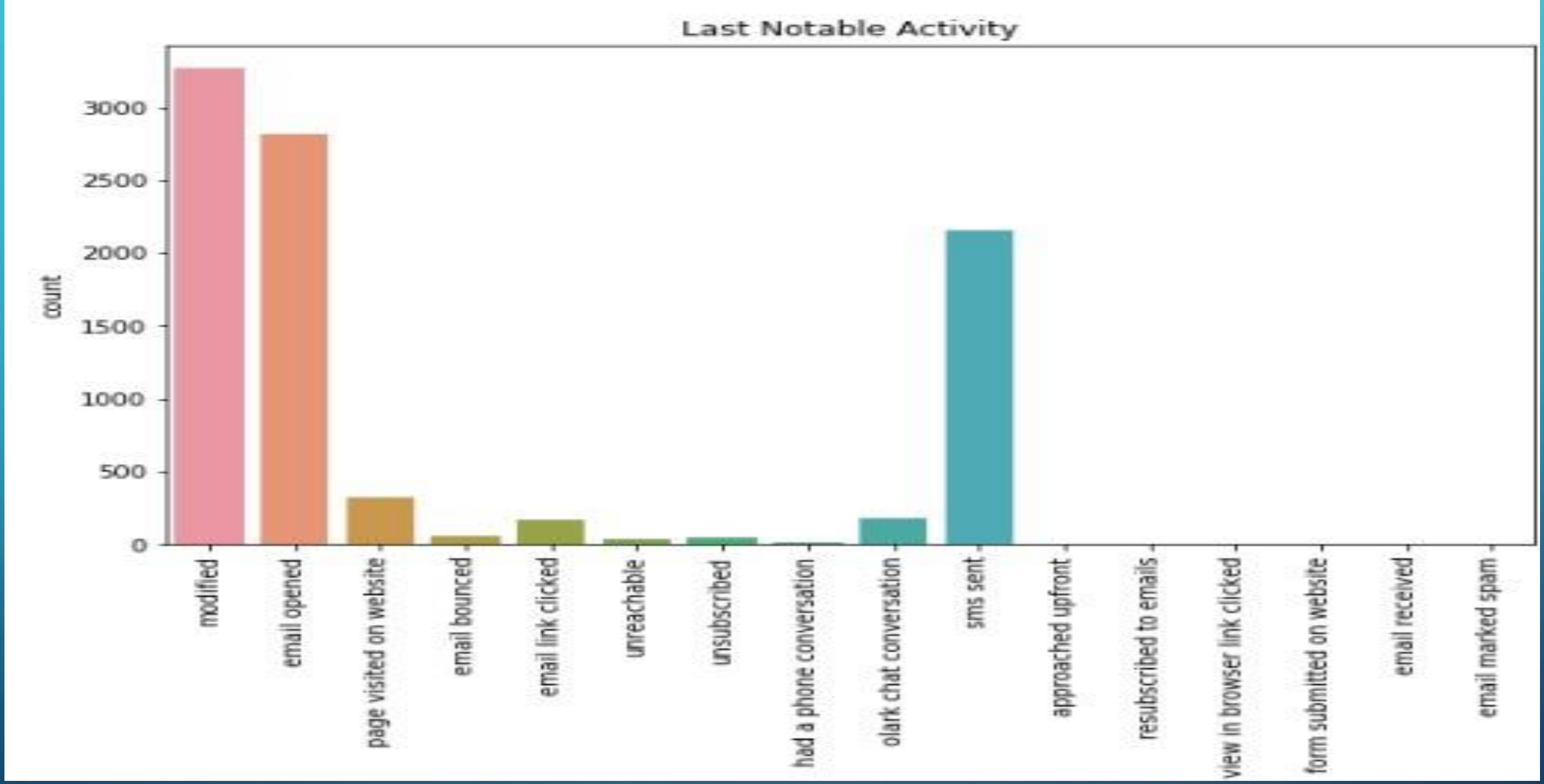
SOLUTION METHODOLOGY

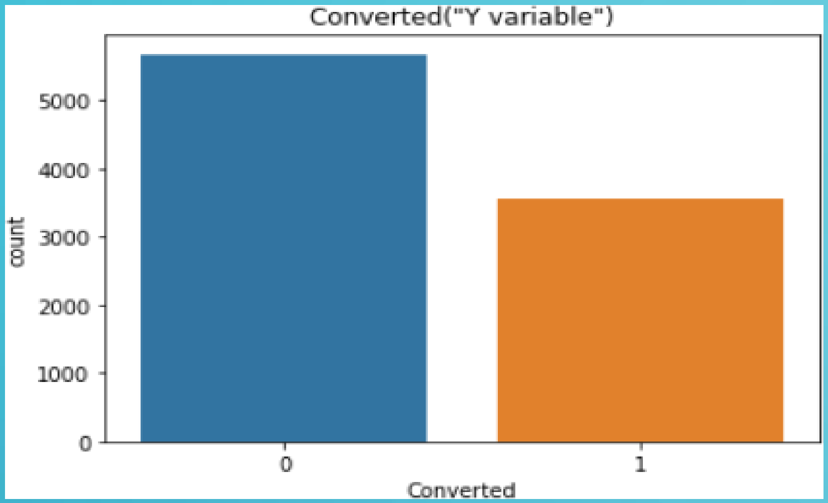
- **Data preprocessing: This involves data cleaning and manipulation tasks such as checking for and handling duplicate data, handling missing values (NA values), dropping columns with a large amount of missing values or that are not useful for analysis, and performing imputation of values when necessary.**
- **Exploratory Data Analysis (EDA): Involves analyzing the data through univariate analysis, which includes examining value counts and variable distributions, as well as bivariate analysis, which includes calculating correlation coefficients and identifying patterns between variables.**
- **Feature Scaling and Encoding: Involves performing feature scaling on numerical variables and encoding categorical variables using techniques like dummy variables or other encoding methods.**
- **Classification Technique: Logistic regression is utilized as the chosen classification technique for building and predicting the model.**
- **Model Validation: Involves validating the model's performance using appropriate evaluation metrics and techniques to ensure its reliability and accuracy.**
- **Model Presentation: The findings and insights from the model are presented, showcasing the key results and highlighting important features.**
- **Conclusions and Recommendations: A summary of the analysis is provided, along with conclusions drawn from the model's outcomes. Recommendations may also be offered based on the insights gained to guide decision-making or further actions.**

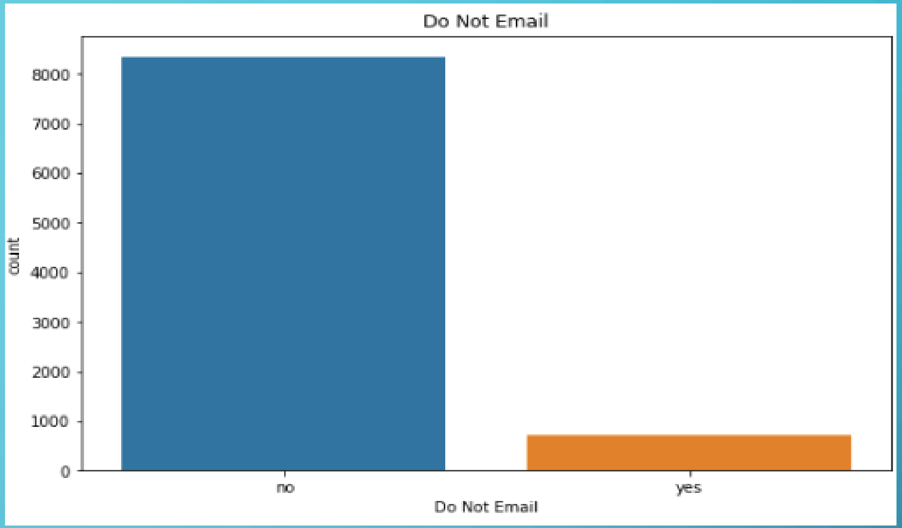
DATA MANIPULATION

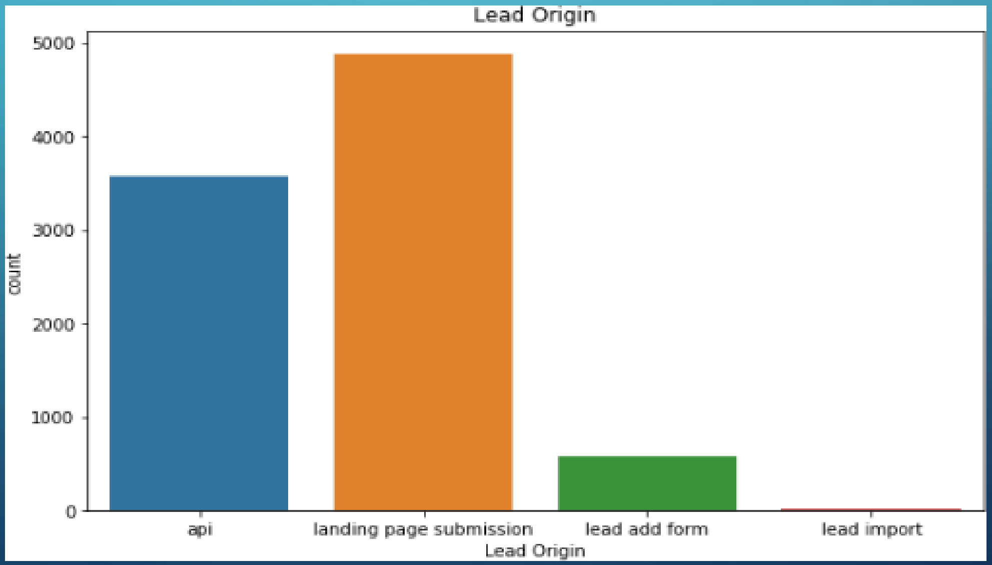
- The dataset contains a total of 37 rows and 9,240 columns.
- Single value features such as "Magazine," "Receive More Updates About Our Courses," "Update me on Supply," "Chain Content," "Get updates on DM Content," "I agree to pay the amount through cheque," etc., have been removed.
- Unnecessary columns like "Prospect ID" and "Lead Number" have been excluded from the analysis.
- After examining the value counts of some object-type variables, certain features with low variance have been dropped. These include "Do Not Call," "What matters most to you in choosing a course," "Search," "Newspaper Article," "X Education Forums," "Newspaper," "Digital Advertisement," etc.
- Columns with more than 35% missing values, such as "How did you hear about X Education" and "Lead Profile," have been removed.

EDA

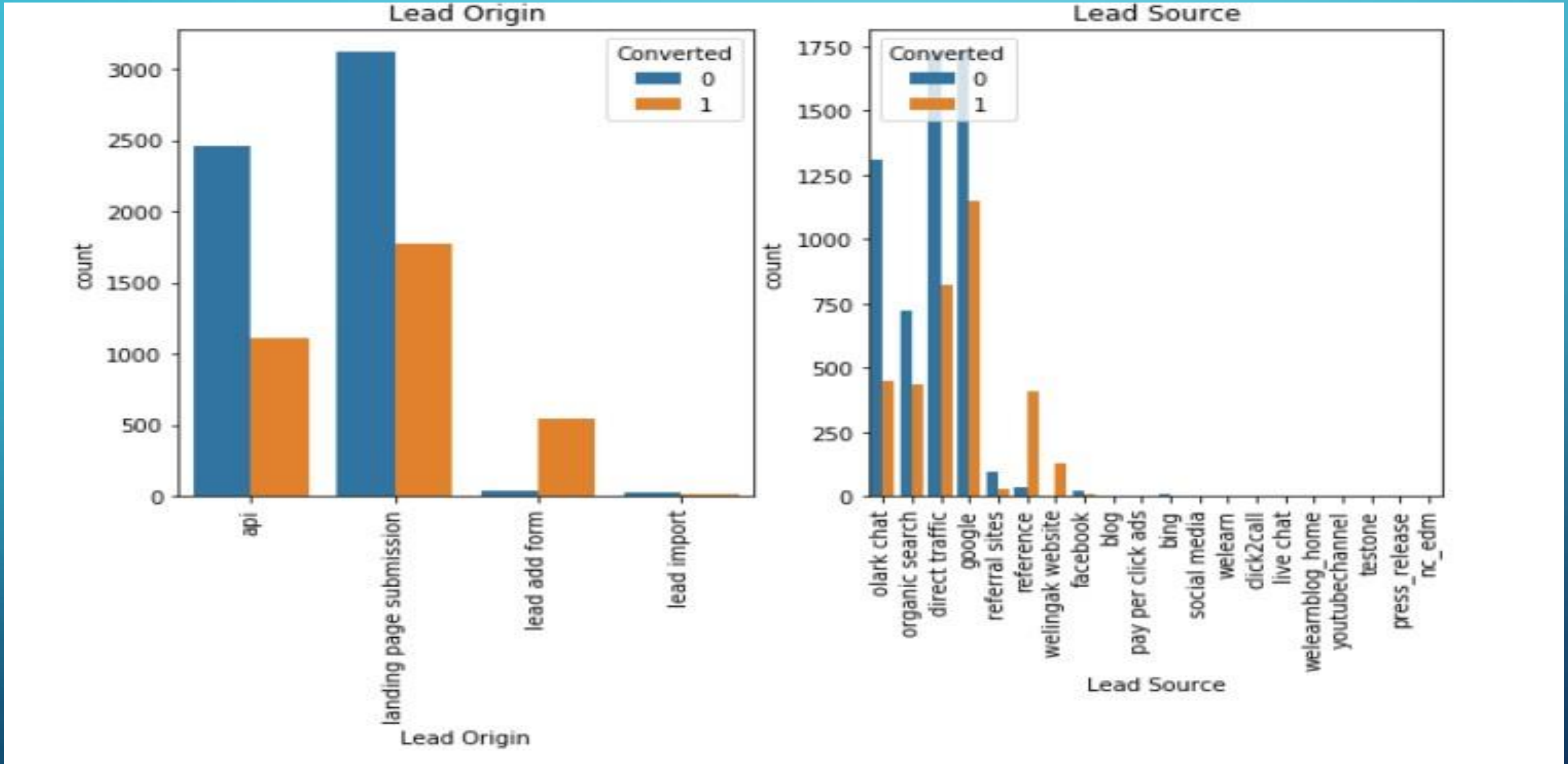

Last Notable Activity
count
3000
2500
2000
1500
1000
500
0
modified
email opened
page visited on website
email bounced
email link clicked
unreachable
unsubscribed
had a phone conversation
olark chat conversation
sms sent
approached upfront
resubscribed to emails
view in browser link clicked
form submitted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

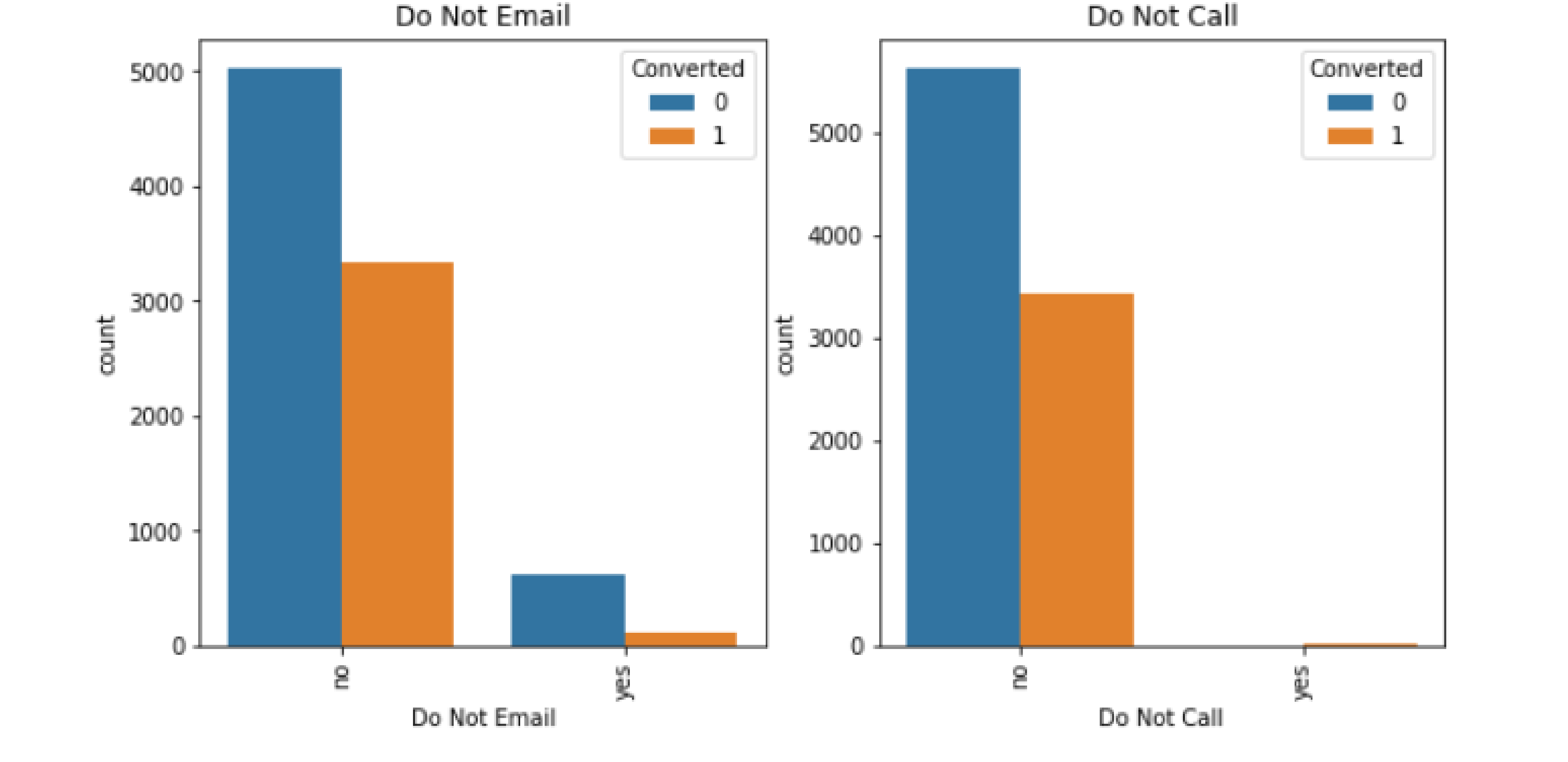

Converted("Y variable")
count
0
1000
2000
3000
4000
5000
0
1
Converted

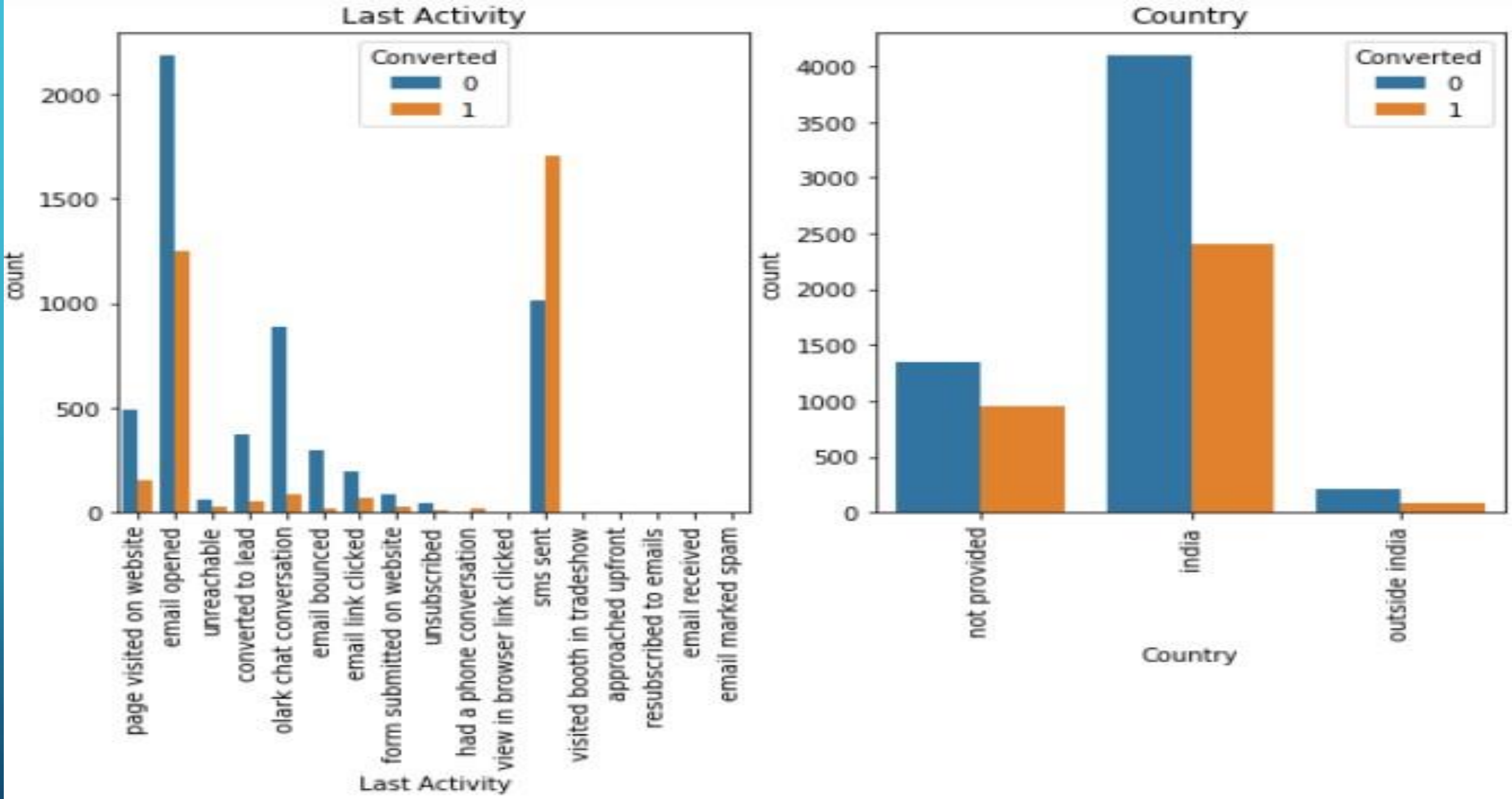

Do Not Email
count
0
1000
2000
3000
4000
5000
6000
7000
8000
no
yes
Do Not Email


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

CATEGORICAL VARIABLE RELATION




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 provided
india
outside india
Country

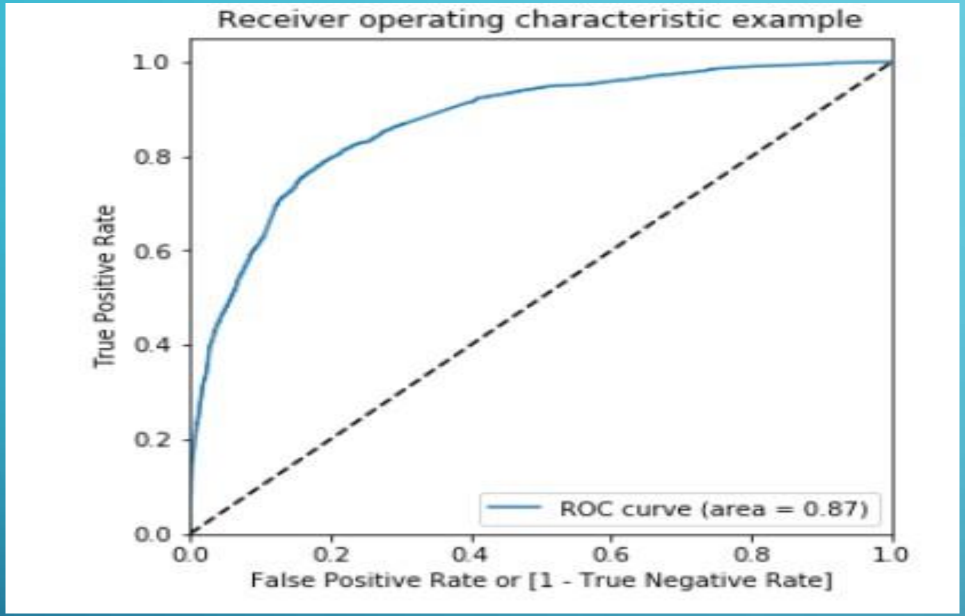
DATA CONVERSION

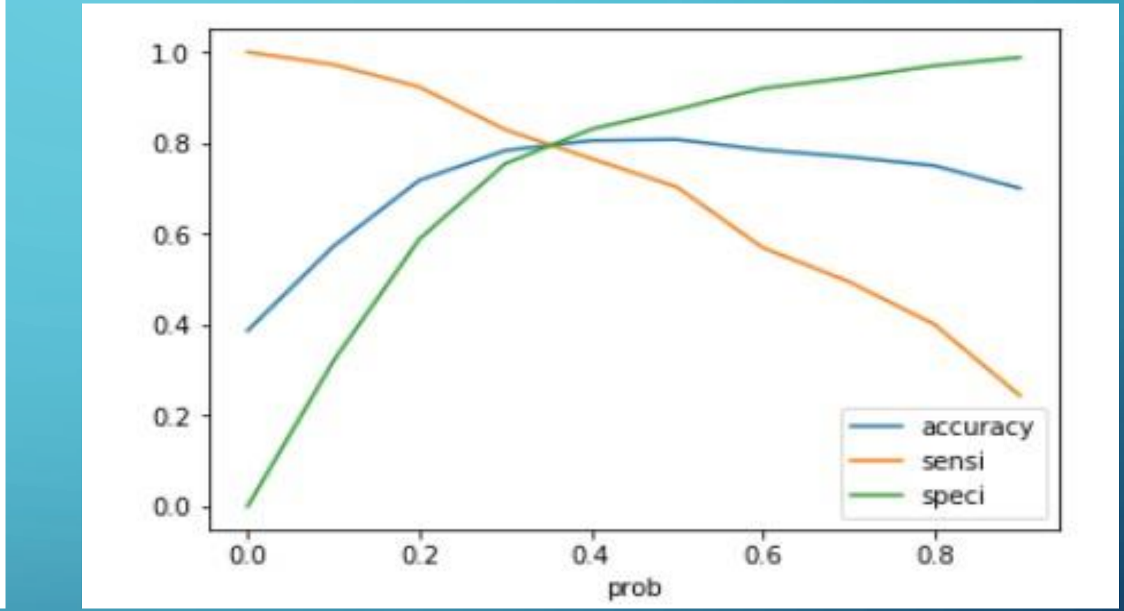
- The numerical variables have been normalized for analysis.
- Dummy variables have been created for object-type variables.
- The dataset used for analysis consists of 8,792 rows.
- The dataset used for analysis includes a total of 43 columns.

MODEL BUILDING

- The data has been split into training and testing sets as a fundamental step for regression analysis.
- The dataset was divided using a 70:30 ratio, with 70% of the data allocated for training and 30% for testing.
- Recursive Feature Elimination (RFE) was employed for feature selection.
- RFE was executed with the goal of selecting the top 15 variables as output.
- The model was built by eliminating variables with p-values greater than 0.05 and variance inflation factor (VIF) values greater than 5.
- Predictions were made on the test dataset using the built model.
- The overall accuracy of the predictions on the test dataset was determined to be 81%.

ROC CURVE





- **The process of finding the optimal cut-off point is undertaken.**
- **The optimal cut-off probability is the point where a balance is achieved between sensitivity and specificity.**
- **Upon examining the second graph, it is apparent that the optimal cut-off point is at 0.35.**

CONCLUSION

Based on the analysis, the variables that hold the most significance for potential buyers, in descending order, are as follows:

- **The total time spent on the website.**
- **The total number of visits.**
- **The lead source, prioritized in the following order: 1. Google, 2. Direct traffic, 3. Organic search, 4. Welingak website**
- **The last activity conducted, with preference given to: 1. SMS, 2. Olark chat conversation**
- **The lead origin being in the Lead add format.**
- **The potential buyer's current occupation as a working professional.**

Considering these findings, X Education is in a favorable position to excel as they have a high chance of persuading almost all potential buyers to change their minds and purchase their courses.

THANK YOU