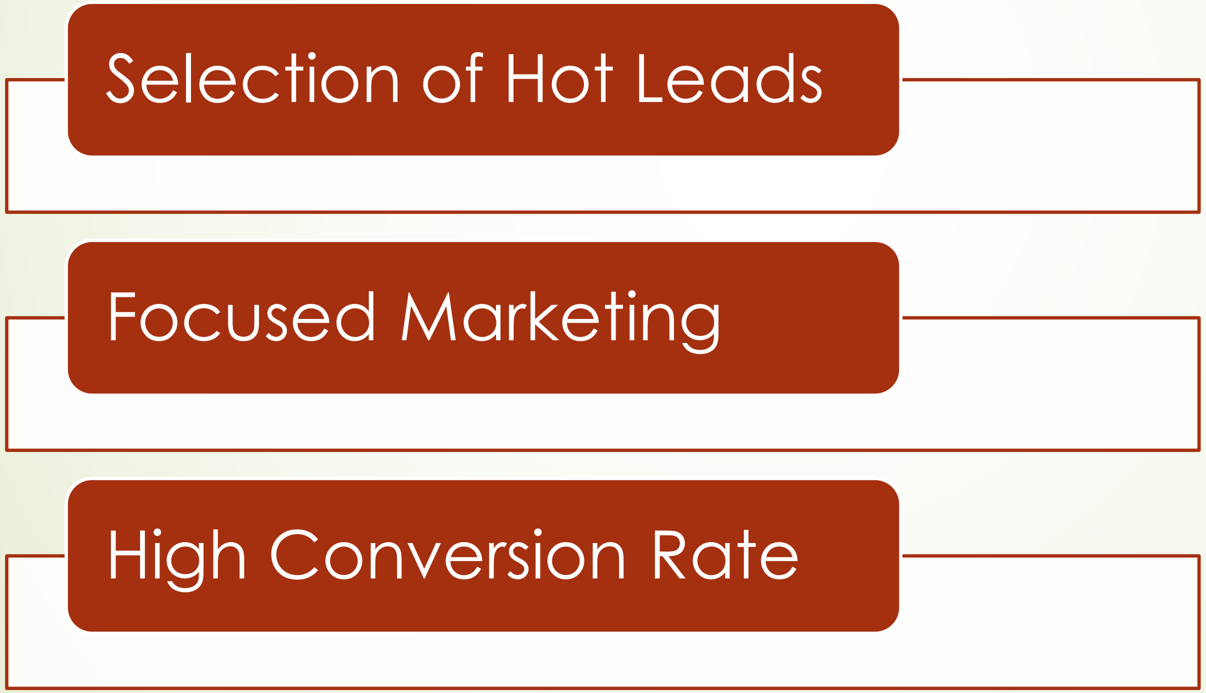
Lead Score Case Study

By

1. Varunprakash Shanmugam
2. E Usha Rani
3. Ushasree Vangala

Business Objective

In order to assist X Education in identifying the most favorable leads (Hot Leads), meaning the leads with the highest likelihood of becoming paying customers.

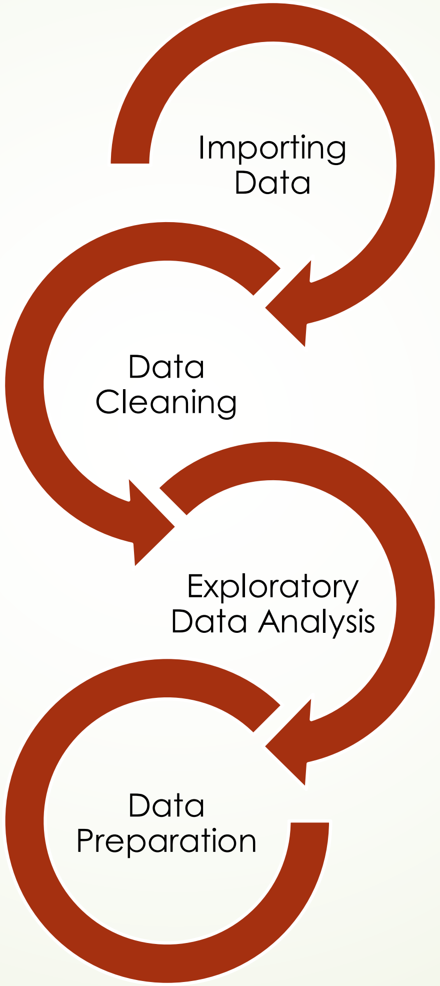


Selection of Hot Leads

Focused Marketing

High Conversion Rate

Solution Methodology



Solution Methodology

Here's a rephrased version of the steps involved in the data analysis process:

- Data Cleaning and Data Manipulation:
  1. Identify and address duplicate records.
  2. Detect and handle NA values and missing data.
  3. Eliminate columns with a significant amount of missing values that don't contribute to the analysis.
  4. If necessary, replace missing values through imputation.
  5. Identify and manage outliers in the dataset.

- Exploratory Data Analysis (EDA):
  1. Perform univariate data analysis, including value counts and variable distributions.
  2. Conduct bivariate data analysis, examining correlation coefficients and patterns between variables.

Solution Methodology

- Feature Preprocessing:
  - Normalize or scale features as needed.
  - Create dummy variables and encode categorical data for analysis.

- Classification Technique:
  - Utilize logistic regression for model development and predictions.

- Model Validation:
  - Validate the model to ensure its accuracy and reliability.

- Model Presentation:
  - Present the model and its findings effectively.

- Conclusions and Recommendations:
  - Summarize key insights and offer recommendations based on the analysis.

Data Manipulation

Here's a rephrased version of the data preprocessing steps:

- The dataset consists of 37 rows and 9,240 columns.

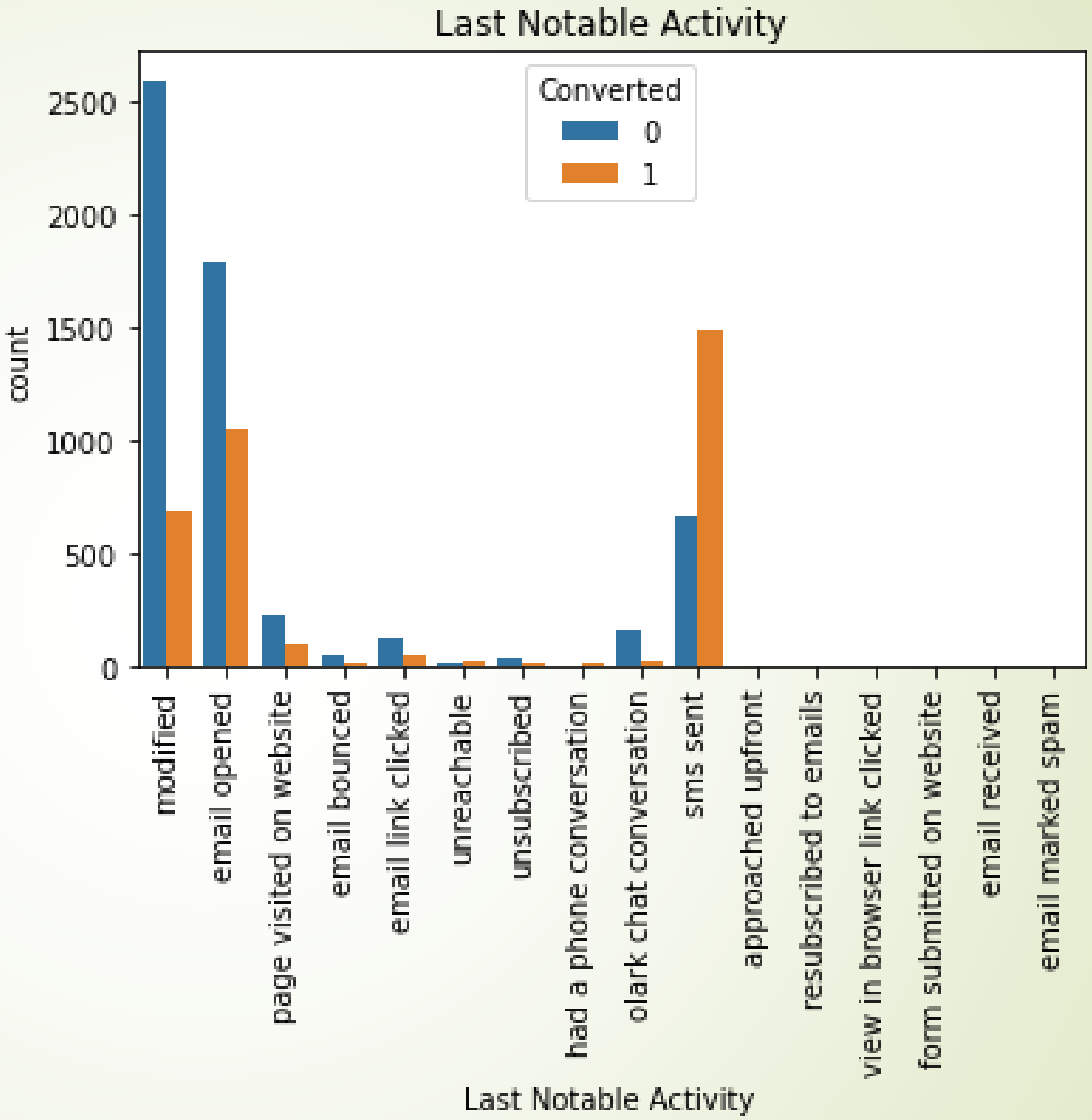
- Features with single constant values, such as "Magazine," "Receive More Updates About Our Courses," "Update me on Supply Chain Content," "Get updates on DM Content," "I agree to pay the amount through cheque," etc., have been eliminated.

- "Prospect ID" and "Lead Number," which are unnecessary for the analysis, have been removed.

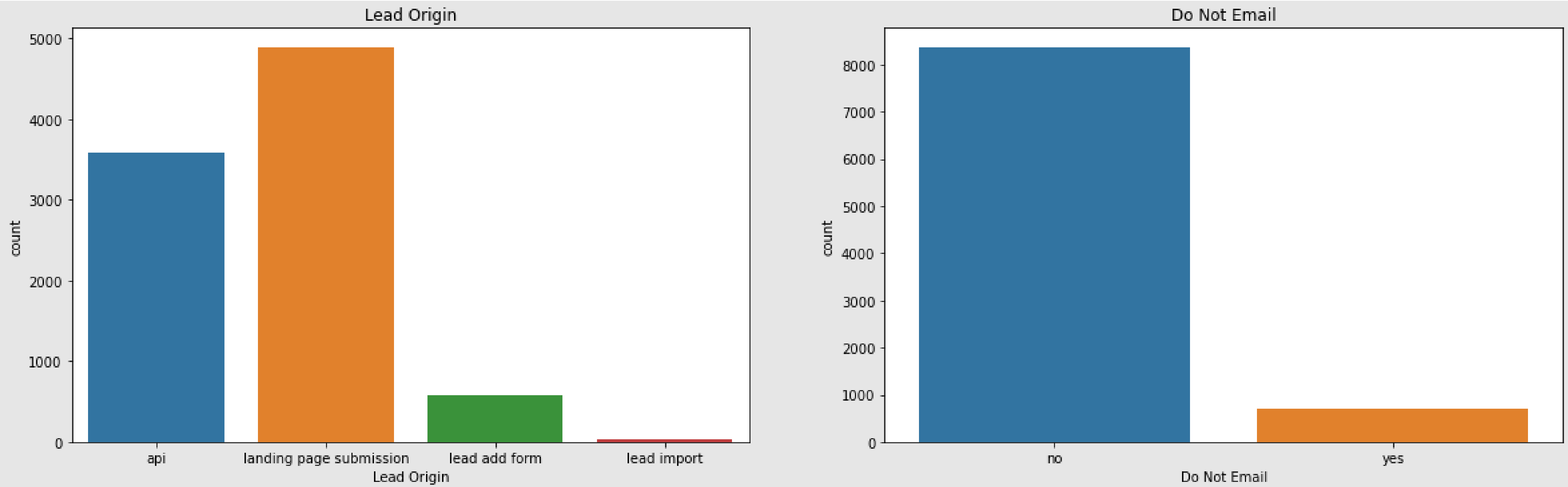
- Some object-type variables with limited variance, as observed from value counts, have been dropped. These features include "Do Not Call," "What matters most to you in choosing course," "Search," "Newspaper Article," "X Education Forums," "Newspaper," "Digital Advertisement," among others.

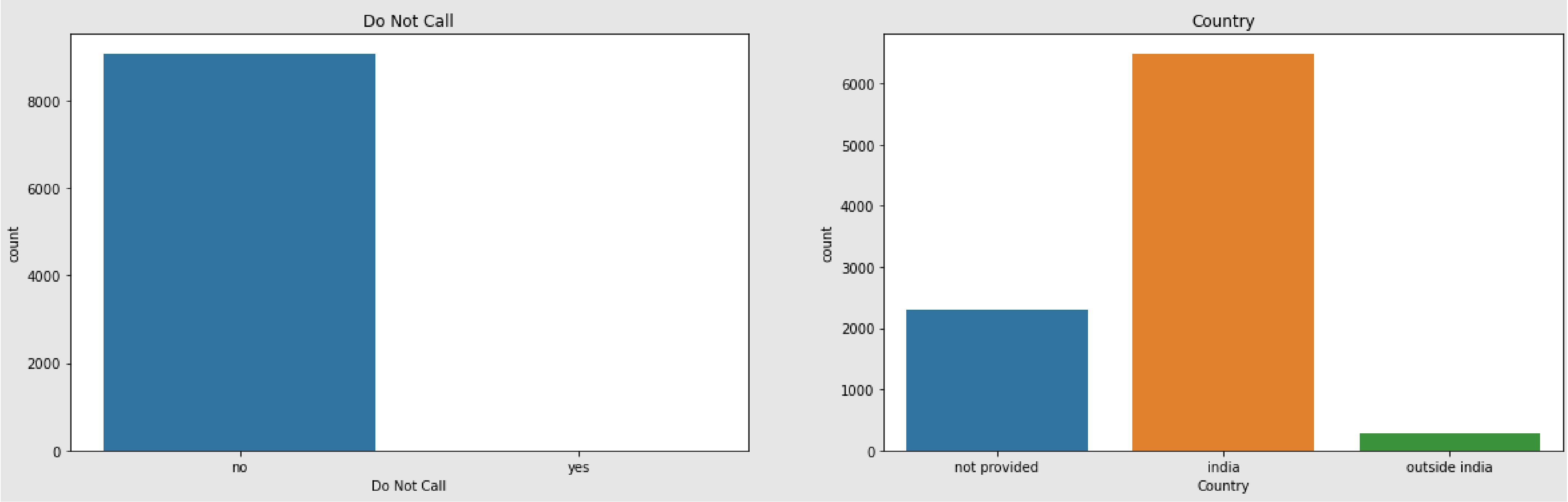
- Columns with more than 35% missing values, such as 'How did you hear about X Education' and 'Lead Profile,' have been excluded from the analysis.

- Data Visualization



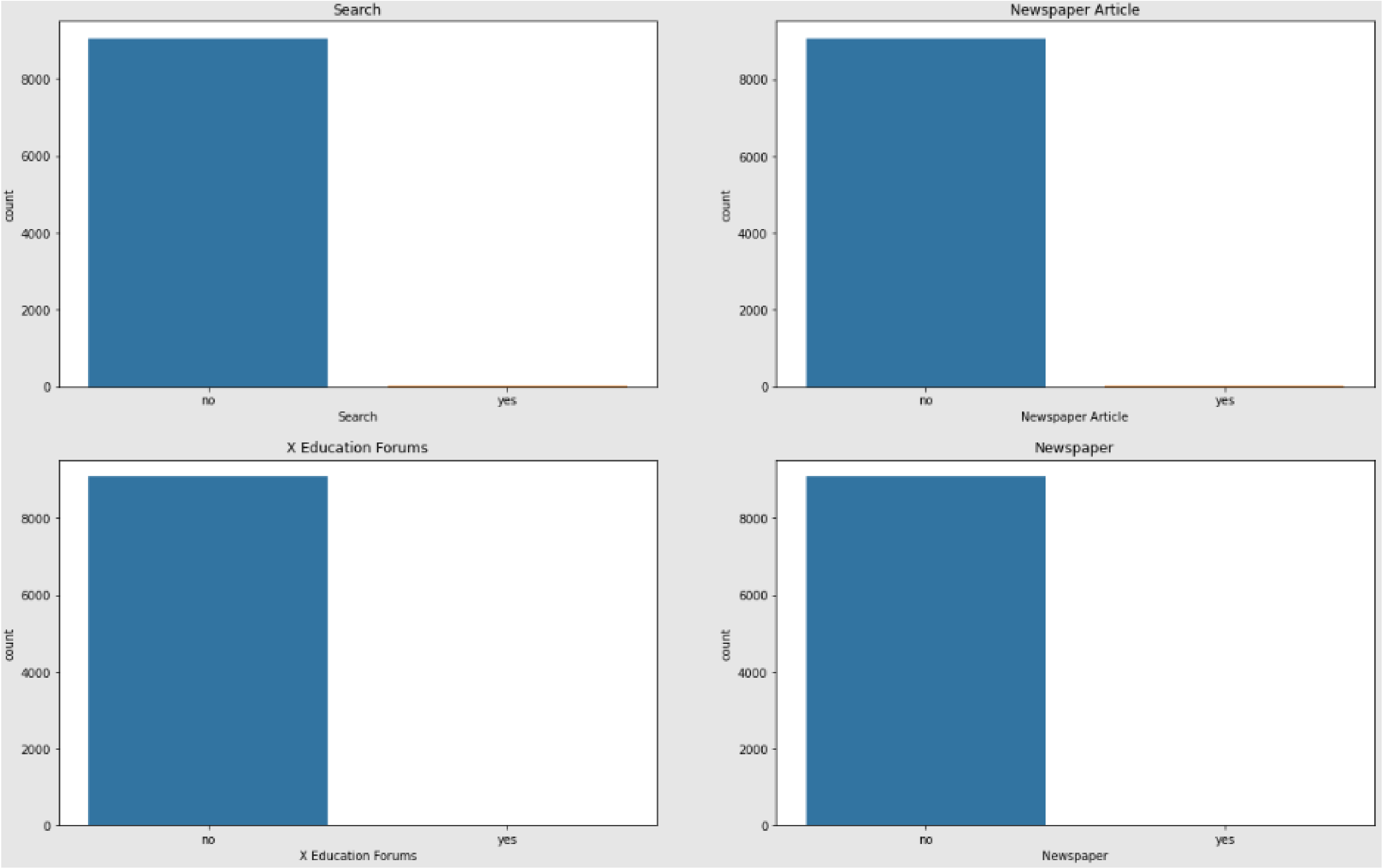
Highest conversion rate is for the last notable activity ‘SMS Sent


Lead Origin
count
5000
4000
3000
2000
1000
0
api
landing page submission
lead add form
lead import
Lead Origin
Do Not Email
count
8000
7000
6000
5000
4000
3000
2000
1000
0
no
yes
Do Not Email



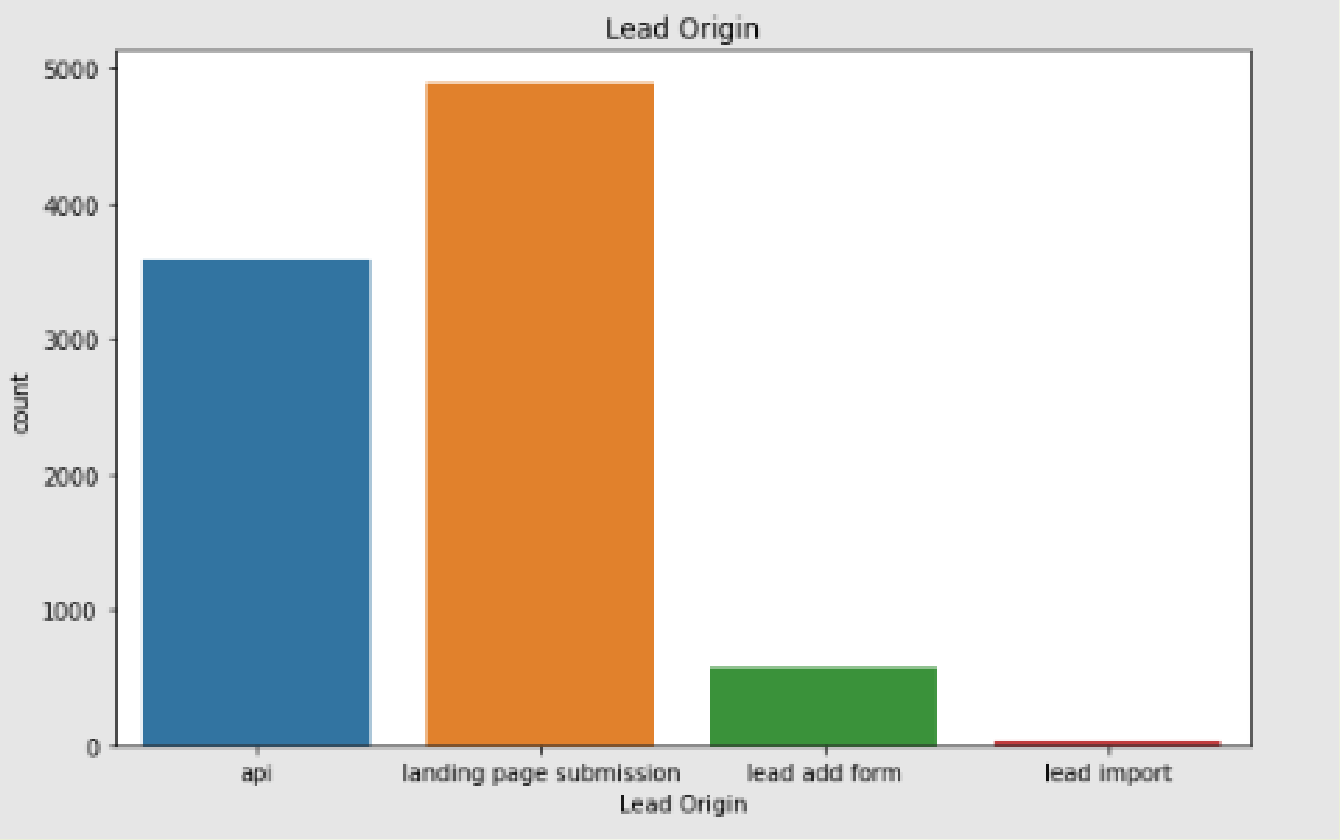
Do Not Call

Country


Search
8000
6000
4000
2000
0
count
no
yes
Search
Newspaper Article
8000
6000
4000
2000
0
count
no
yes
Newspaper Article
X Education Forums
8000
6000
4000
2000
0
count
no
yes
X Education Forums
Newspaper
8000
6000
4000
2000
0
count
no
yes
Newspaper

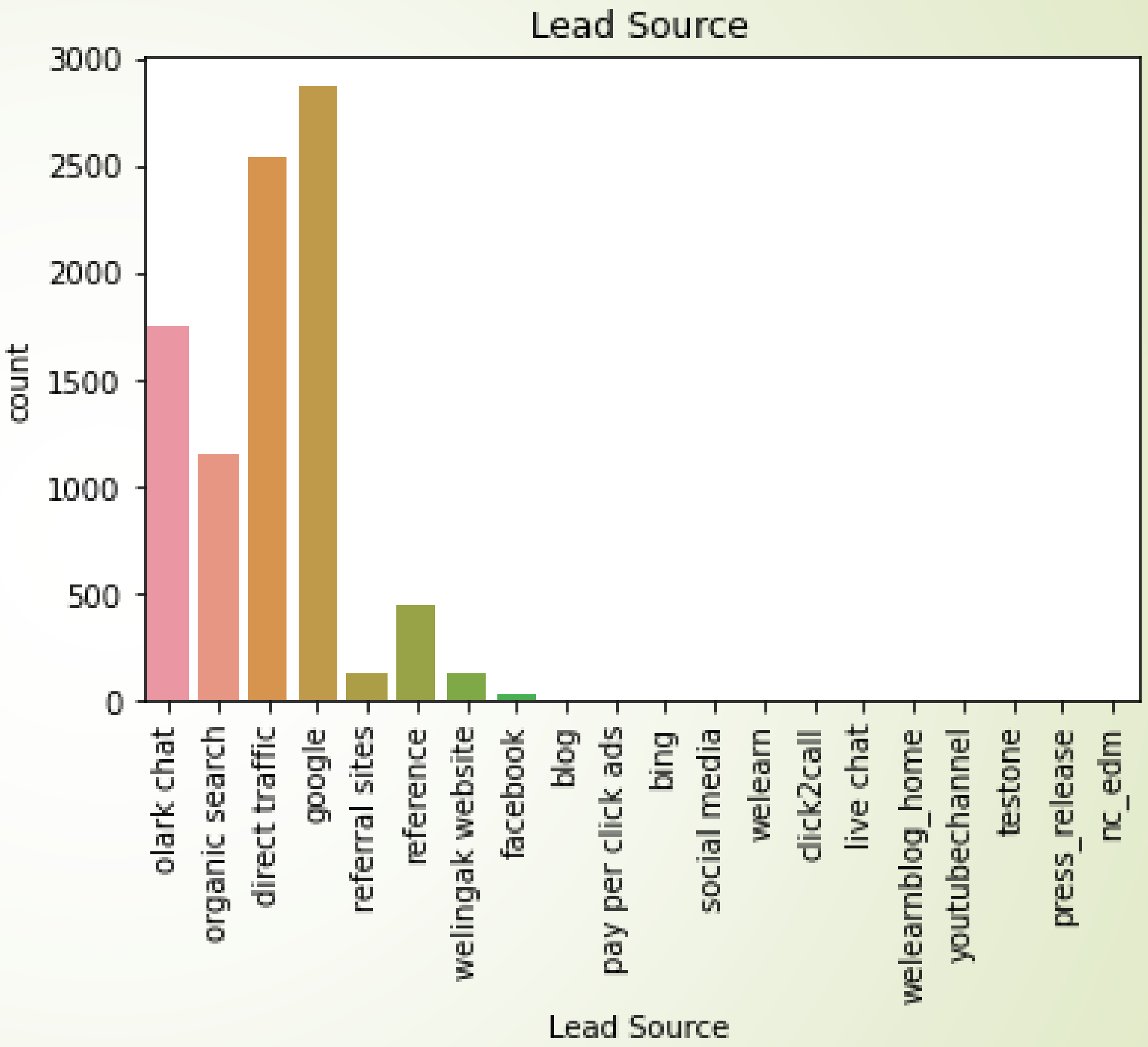
'API' and 'Landing Page Submission' generate the highest number of leads but exhibit lower conversion rates, whereas 'Lead Add Form' generates fewer leads but boasts a notably higher conversion rate.

The goal is to enhance the conversion rates for 'API' and 'Landing Page Submission' while increasing lead generation through the utilization of 'Lead Add Form.'

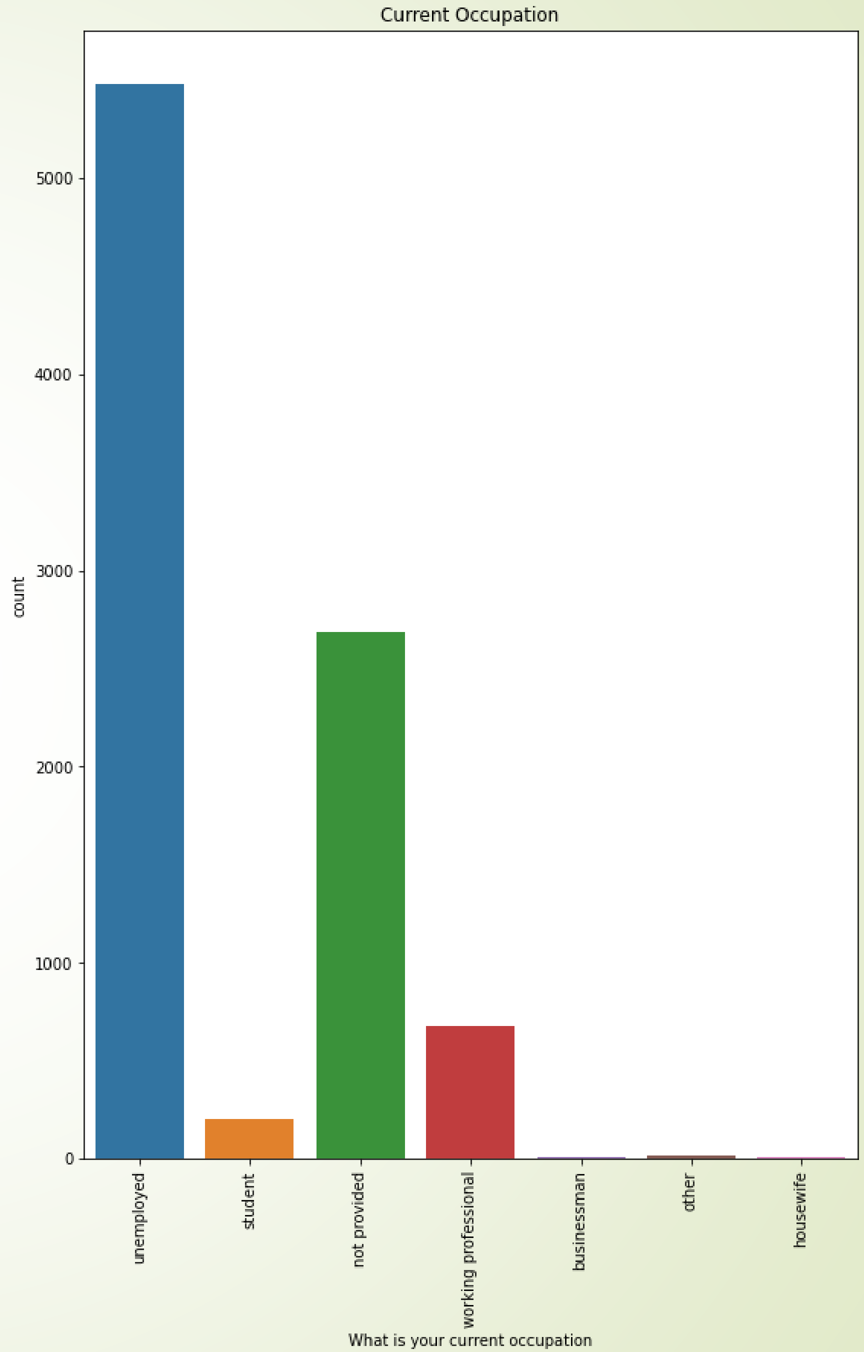


- Exceptionally high conversion rates are observed for lead sources 'Reference' and 'Welingak Website.’

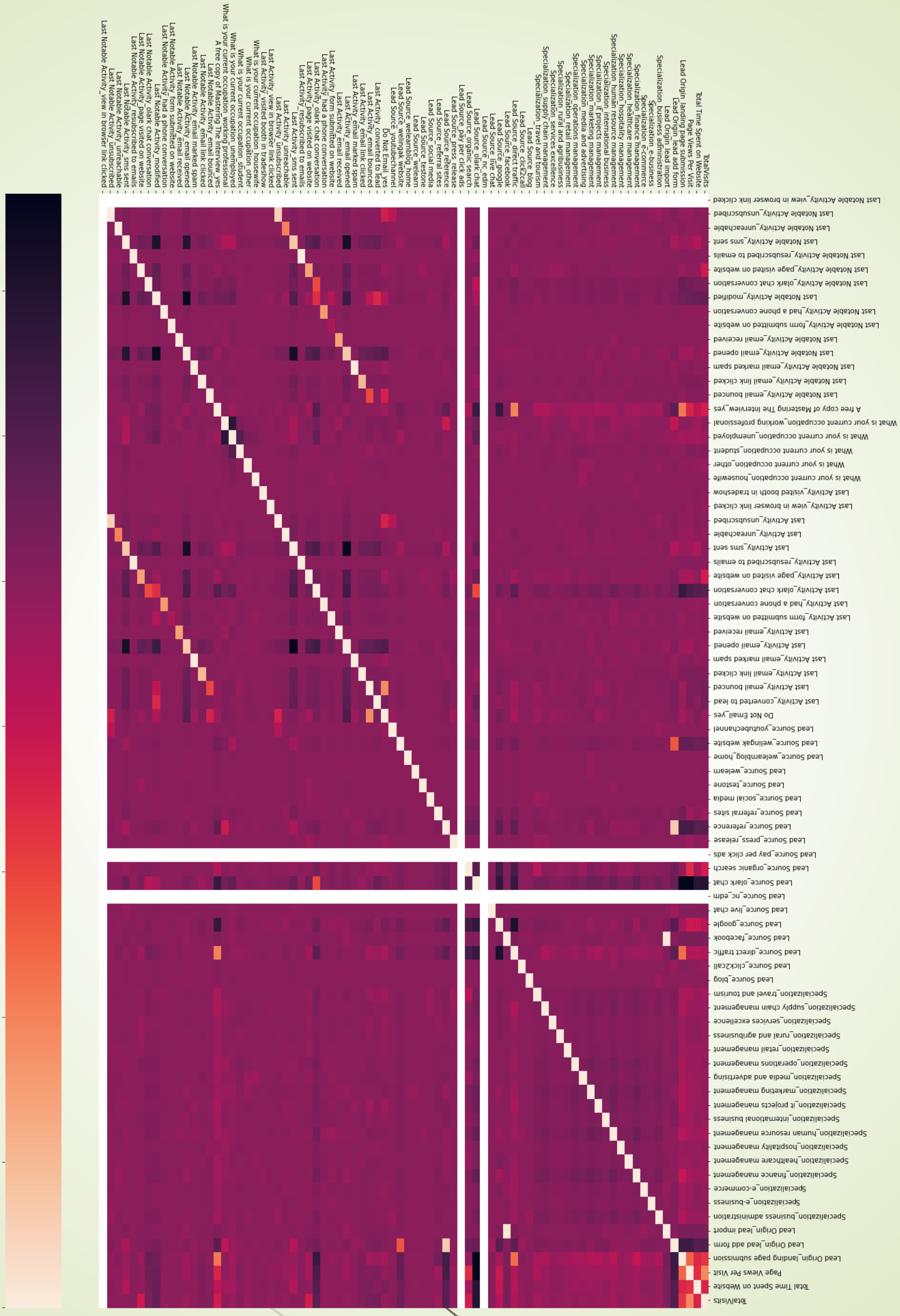
- The majority of leads originate from 'Direct Traffic' and 'Google.'



Individuals with a working professional background have the highest likelihood of conversion.



The associations between features in the ultimate model are insignificant.



Data Conversion

- Numerical variables have been standardized.
- Dummy variables have been generated for object-type variables.
- The dataset for analysis contains 8,792 rows and 43 columns.

Model Building

Data Splitting for Training and Testing

- The initial step in regression analysis involves splitting the data into training and testing sets, with a chosen ratio of 70:30.

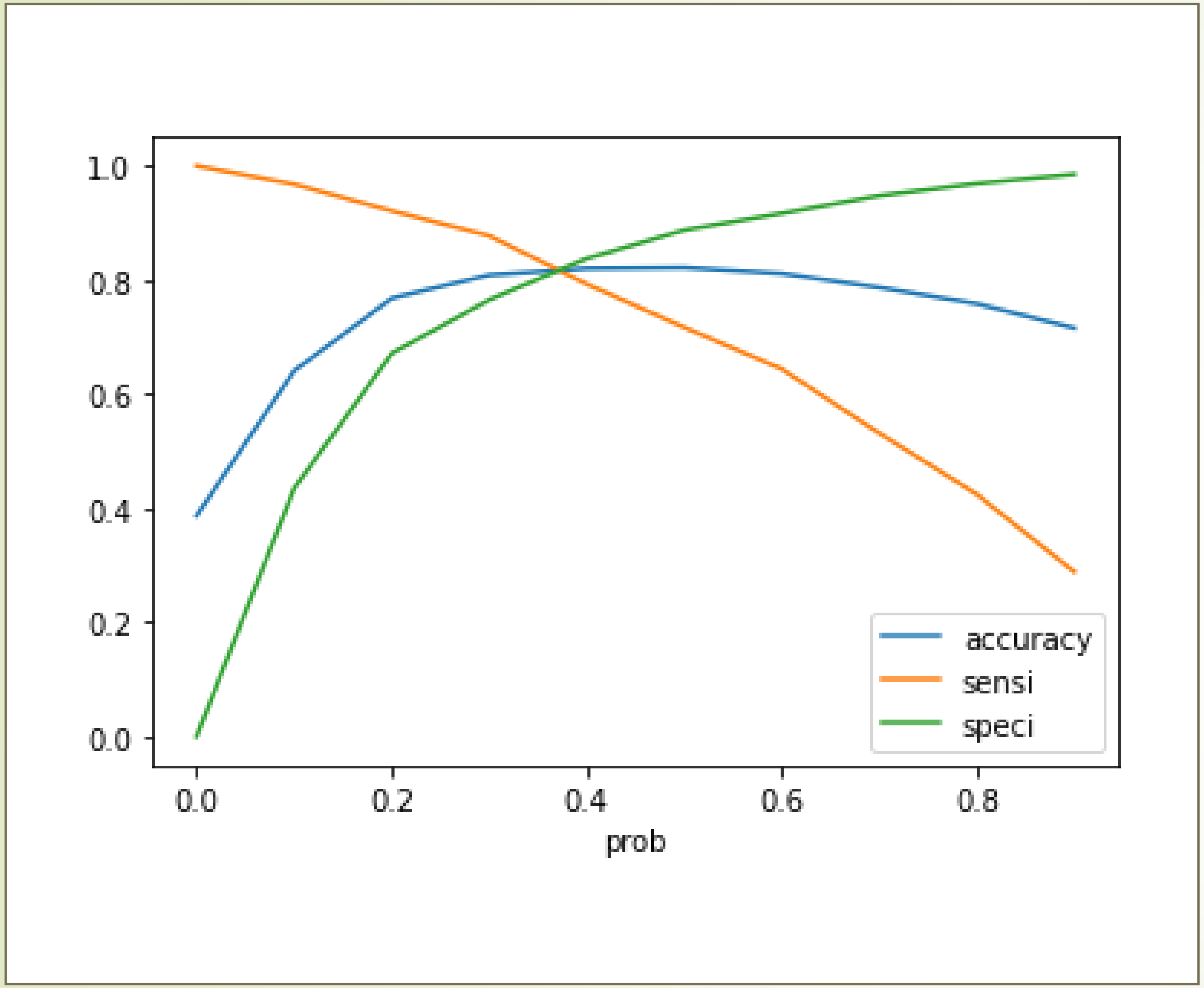
- Feature Selection Using Recursive Feature Elimination (RFE)

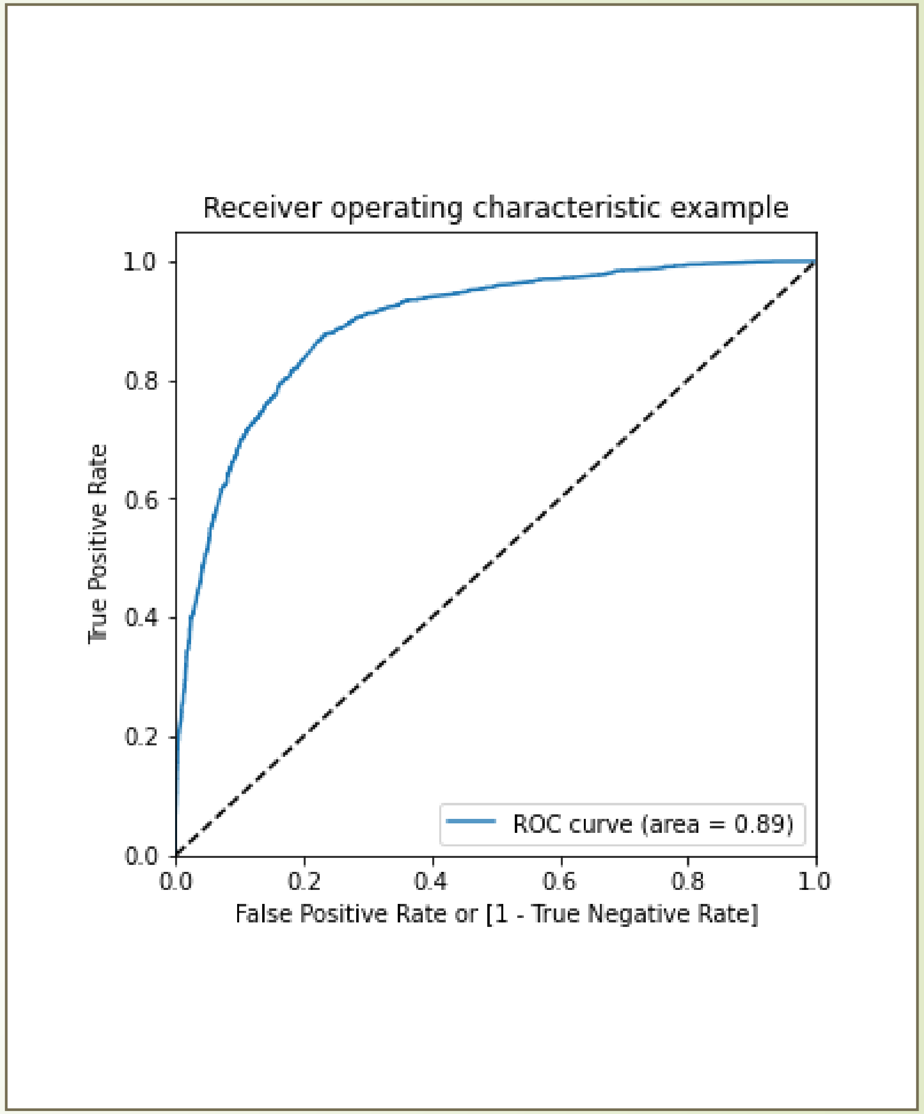
- RFE was executed to select the top 15 variables as output.

- Model Development: Variables were removed from the model if their p-value exceeded 0.05 or their VIF value exceeded 5.

- Predictions were made on the test dataset.

- The overall accuracy of the model achieved an 81% accuracy rate.


1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
prob
accuracy
sensi
speci


Receiver operating characteristic example
True Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate or [1 - True Negative Rate]
ROC curve (area = 0.89)

ROC Curves

ROC Curves

Determining the Optimal Threshold

- The optimal threshold probability is the point where a balance between sensitivity and specificity is achieved.

- As depicted in the second graph, it's evident that the optimal threshold is situated at 0.35.

Recommendations

Analyzing the Data Visualizations, Prioritize the Following:

- Enhancing conversion rates for categories that generate substantial leads.

- Increasing lead generation for categories with already high conversion rates.

Consider the Significance of Features in the Model and Their Influence on Conversion Probability—whether positive or negative.

Adjust the Probability Threshold Value for Identifying Potential Leads as per specific business requirements.

Conclusion

The key determinants among potential buyers, ranked in descending order of importance, are as follows:

- Total time spent on the website.
- Total number of visits.
- Lead source, with priority given to:
a. Google
b. Direct traffic
c. Organic search
d. Welingak website
- Last activity, with emphasis on:
a. SMS
b. Olark chat conversation
- Lead origin in Lead add format.
- Current occupation as a working professional.

With these insights in mind, X Education has a significant opportunity to effectively engage nearly all potential buyers, persuading them to reconsider and enroll in their courses, leading to flourishing outcomes.