Lead Score Case Study

Pranjal Kumar Singh | Vikrant Vijayrao Shrimali | Meghana Annabeemoju

Problem Statement

- An education company named X Education sells online courses to industry professionals.
- The typical lead conversion rate at X education is around 30%. Now, although X Education gets a lot of leads, its lead conversion rate is very poor.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.
- If X Education can effectively identify this set of leads, the conversion rate should improve, as the sales team will be able to concentrate on engaging with high-potential leads rather than contacting everyone.

Business Objective

- Company wants to increase their business by selling courses to more professionals..
- For that they want to build a Model which identifies the hot leads(A lead having high probability of conversion)
- Use this modes for various use cases and future needs.

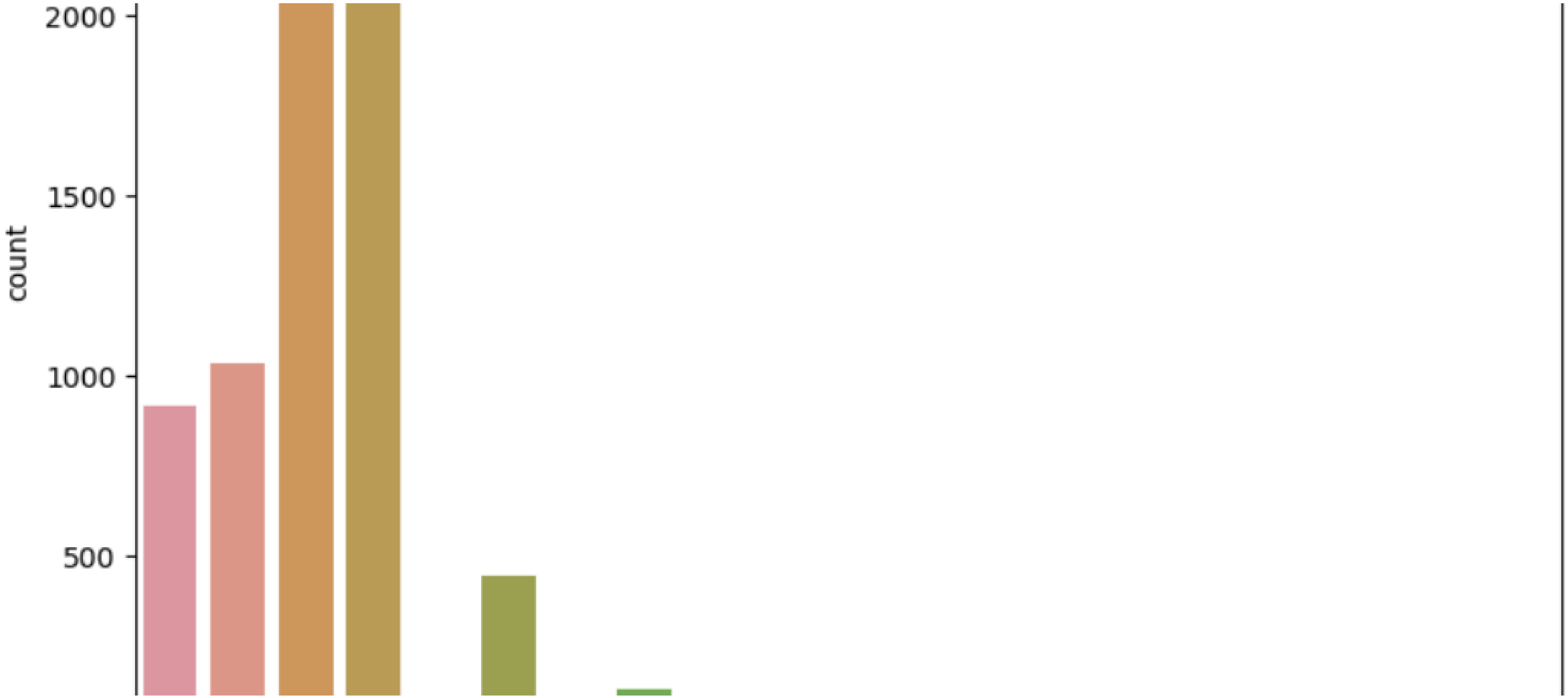
Solution Methodology

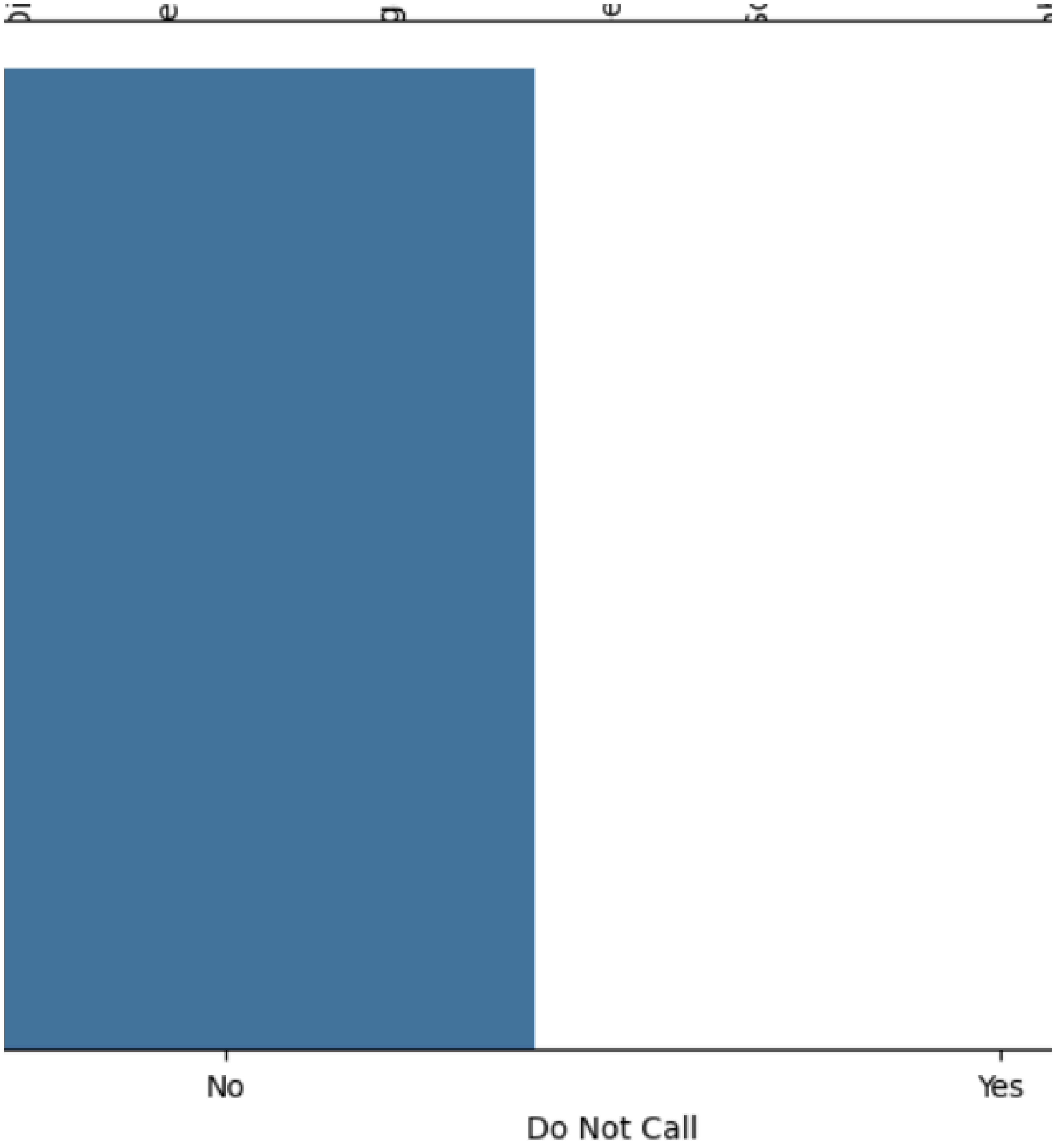
- **Data Cleaning and Manipulation:**
    - Check and handle duplicate data.
    - Check and handle NA values and missing values.
    - Drop columns that contain a large amount of missing values and are not useful for the analysis.
    - Impute missing values, if necessary.
    - Check and handle outliers in the data.
- **Exploratory Data Analysis (EDA):**
    - Univariate data analysis: value counts, distribution of variables, etc.
    - Bivariate data analysis: correlation coefficients and patterns between variables, etc.
- **Feature Engineering:**
    - Apply feature scaling and create dummy variables.
    - Encode categorical data.
- **Modeling:**
    - Use a classification technique, specifically logistic regression, for model building and prediction.
    - Validate the model.
- **Model Presentation:**
    - Present the model.
    - Provide conclusions and recommendations.

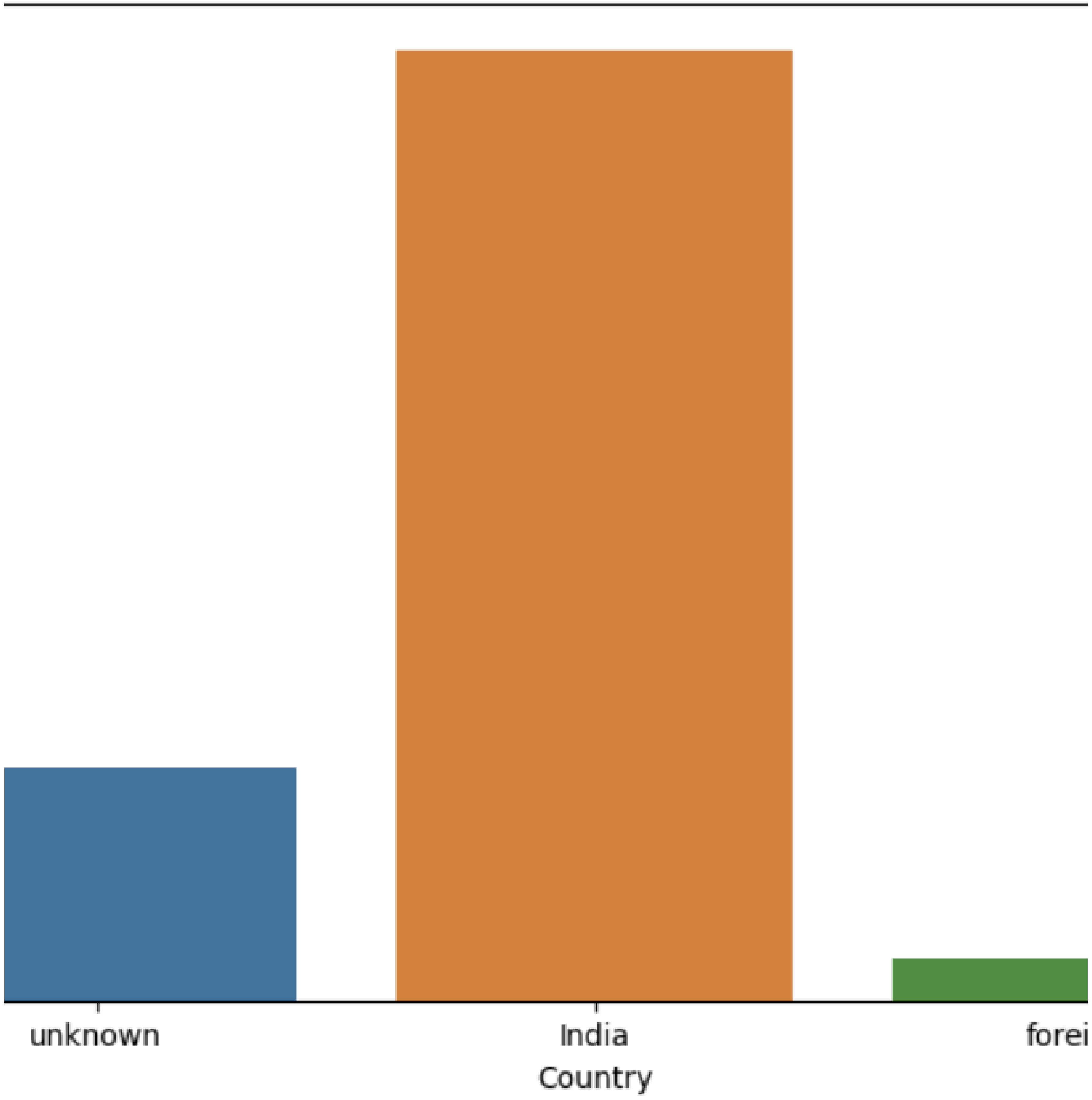
Data Cleaning & Manipulation

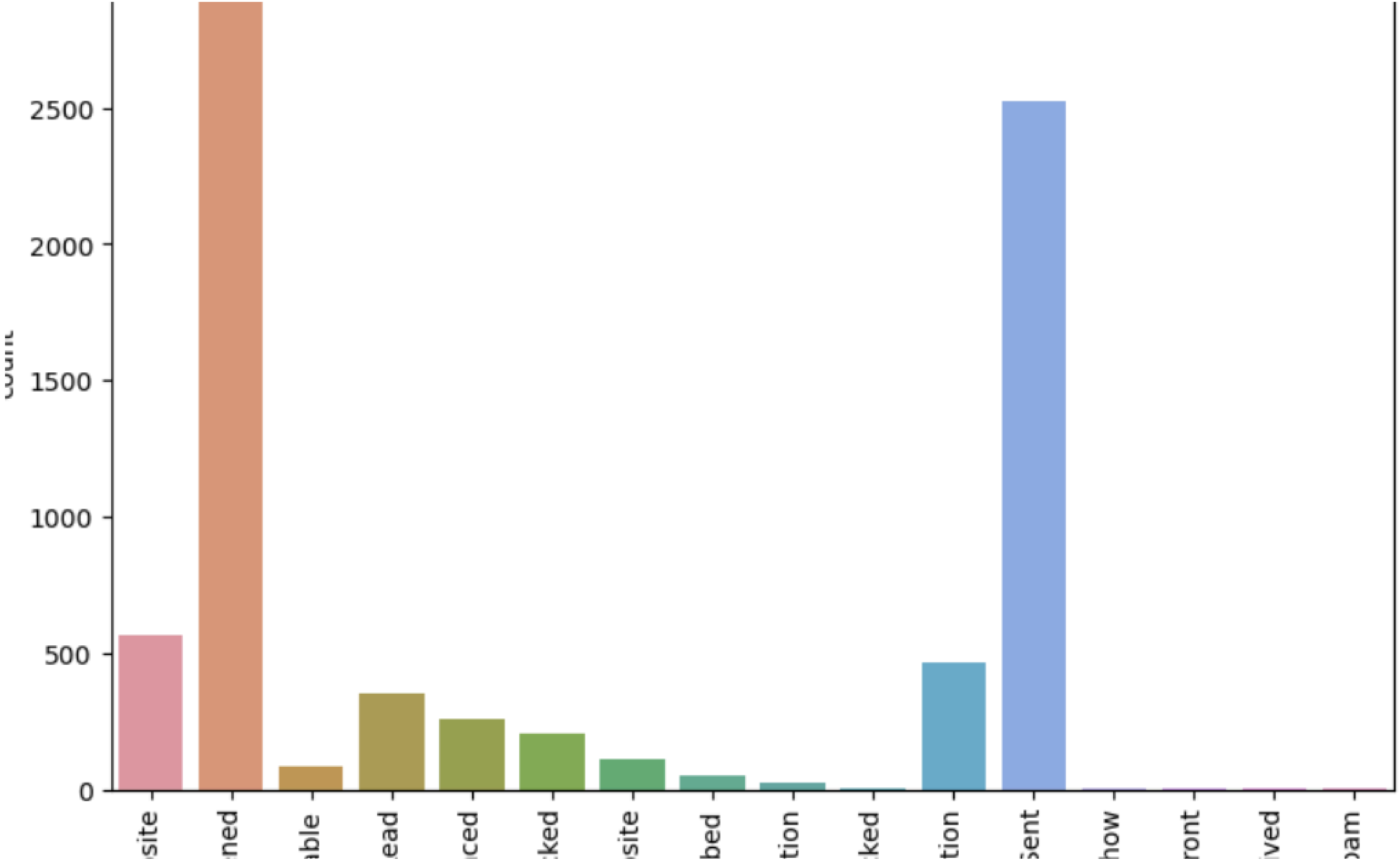
- Total number of rows: 37, total number of columns: 9,240.
- Removed columns having null values with more than 40%.
- Dropped columns such as ‘Tags’ , ‘Search’, ‘NewsPaper Article’, ‘XEducationForums’, ‘Newspaper’, ‘Lead Number’, ‘Digital Advertisement’, ‘Through recommendations’, ‘Asymmetrique Activity Index’, ‘Asymmetrique Profile Index’ etc since they don’t add any significance to the model.
- After checking the value counts for some categorical variables, replaced null values with unknown for the following features "What matters most to you in choosing course" , ‘what is your current occupation’, ‘country’, .

Analyzing Categorical Data

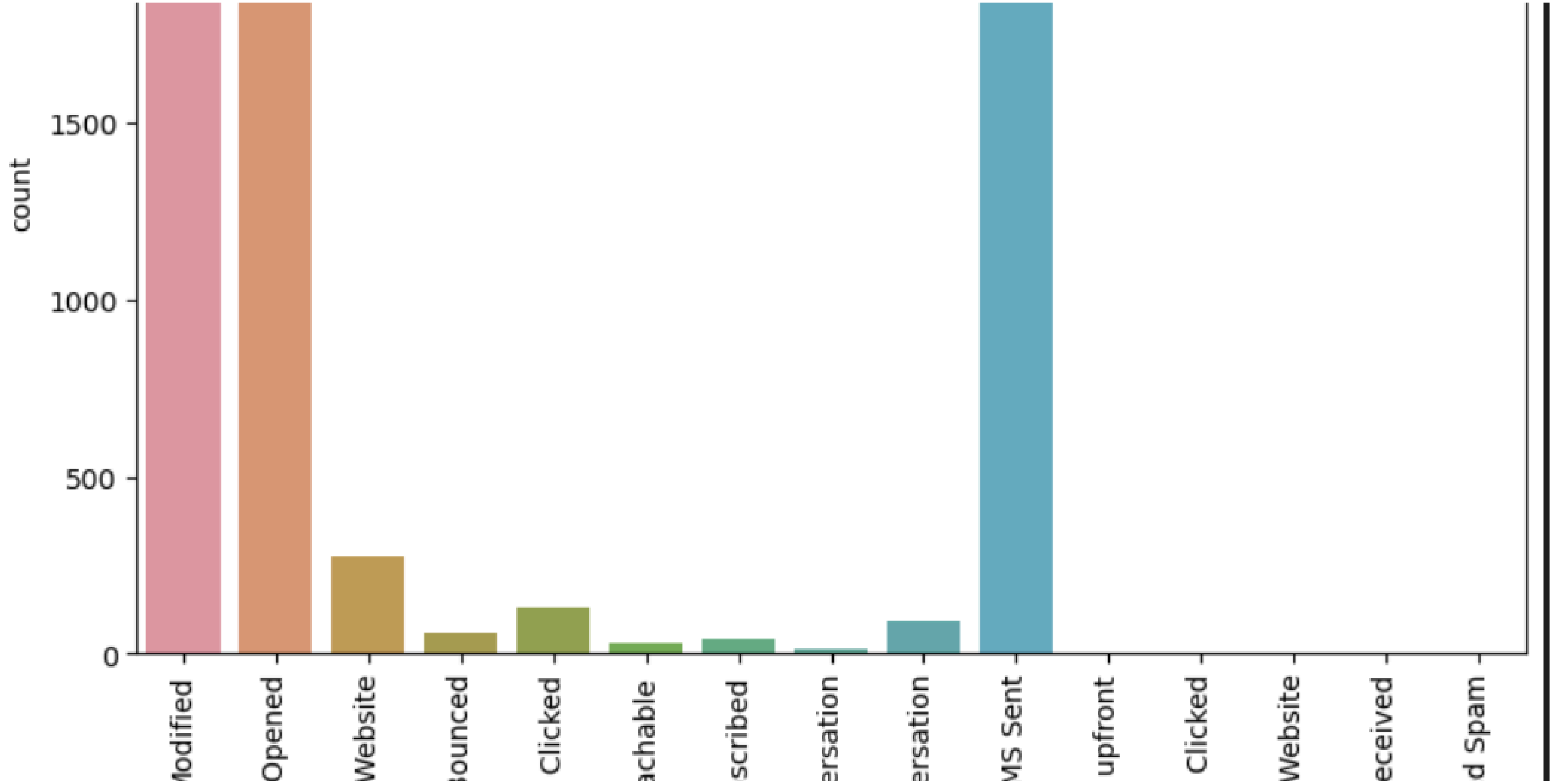


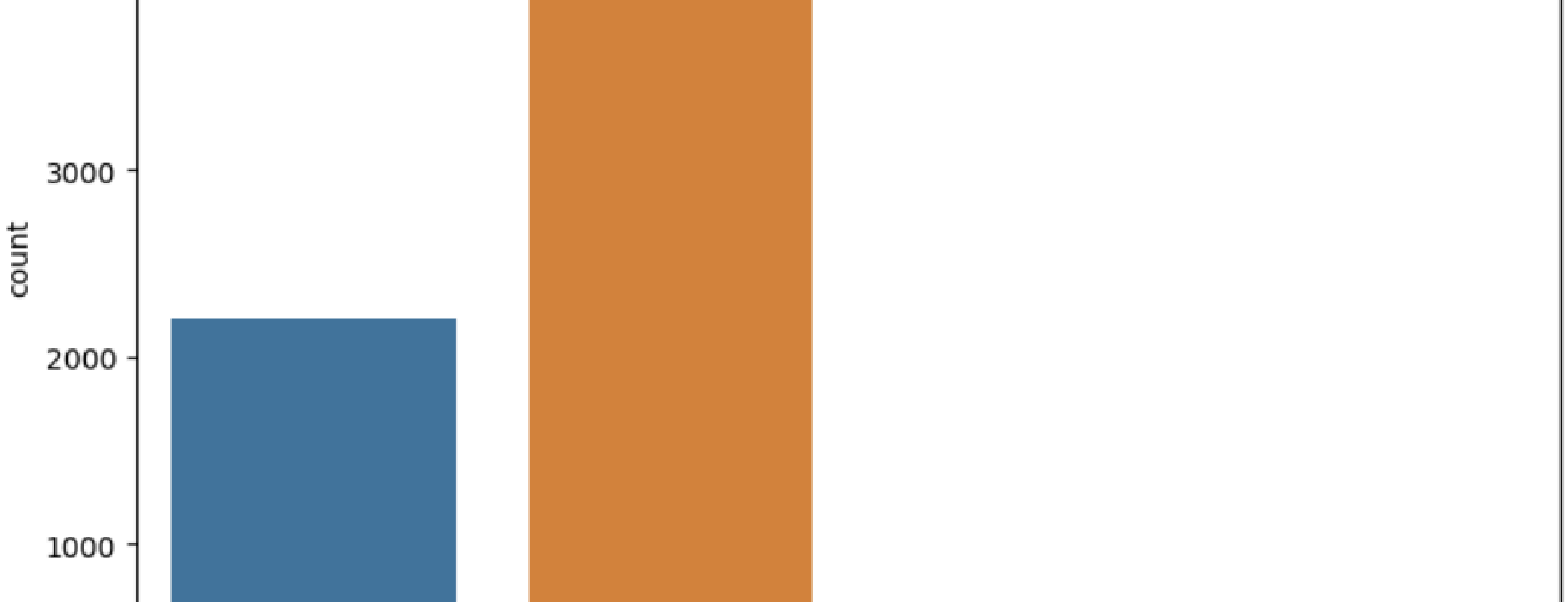

No
Yes
Do Not Call

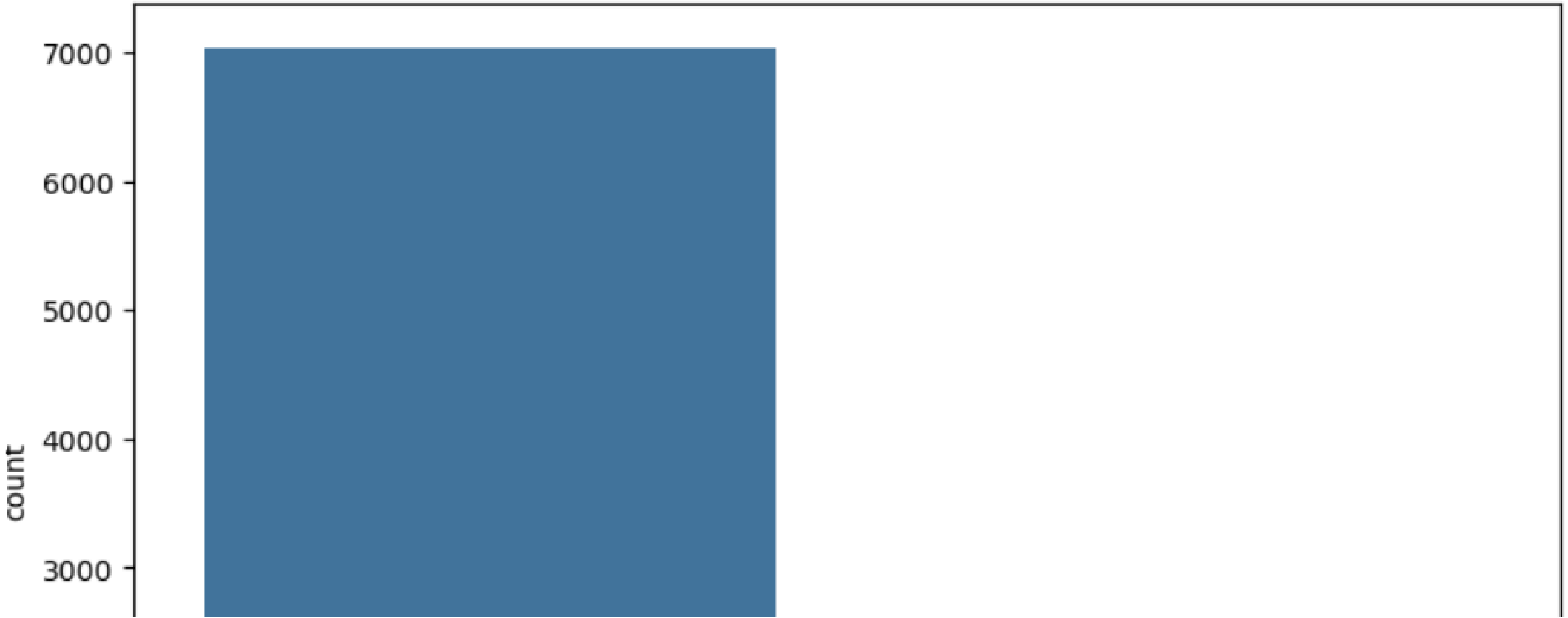

unknown
India
forei
Country


count
0
500
1000
1500
2000
2500
site
ned
able
ead
ced
ked
site
bed
tion
ked
tion
Sent
how
ront
ived
pam

EDA


count
1500
1000
500
0
Aodified
Opened
Website
Bounced
Clicked
achable
scribed
ersation
ersation
MS Sent
upfront
Clicked
Website
eceived
d Spam


3000
2000
1000
count


7000
6000
5000
4000
3000
count

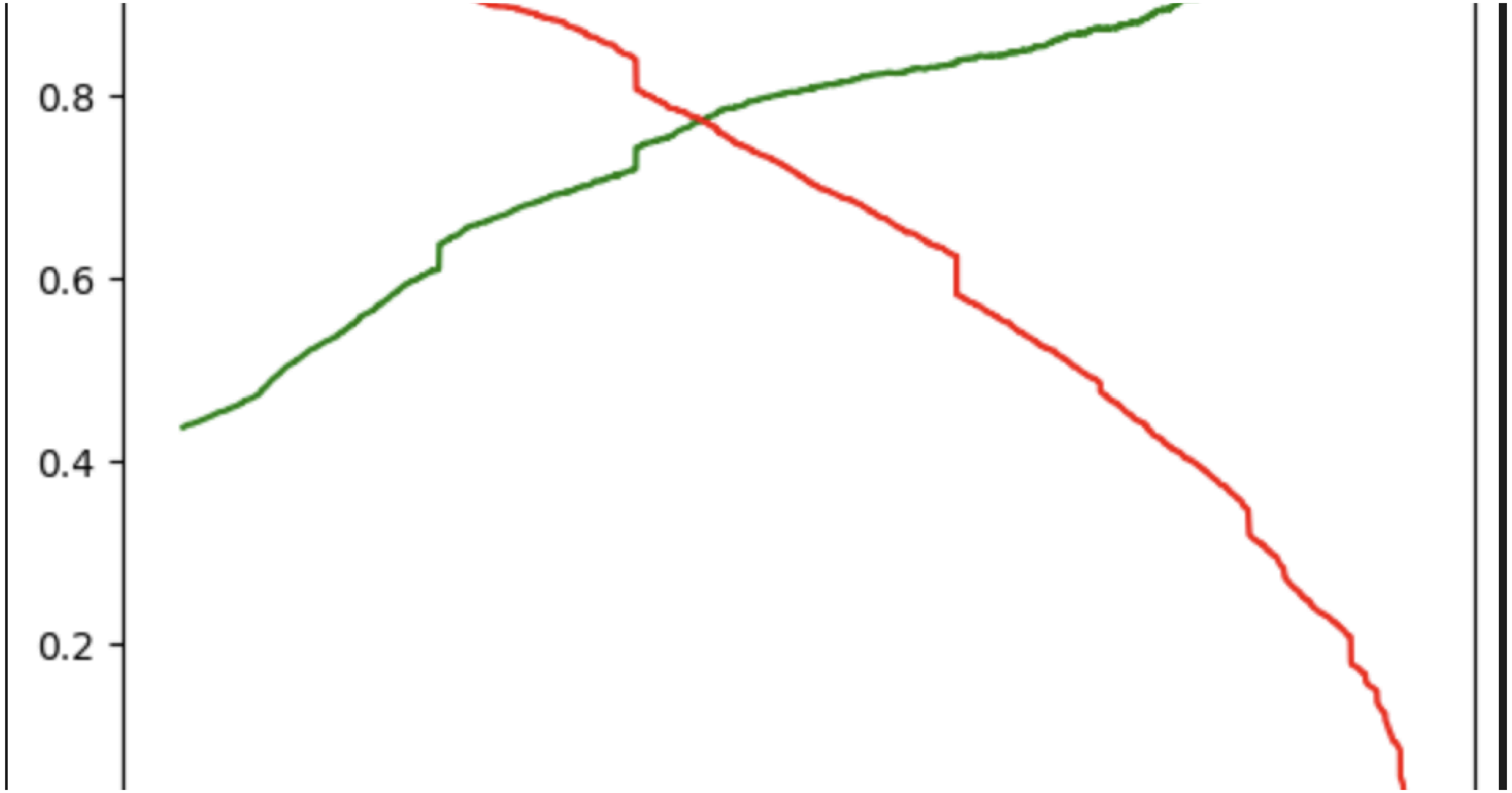
Data Conversion

- Reclassified geographical data into ‘India’, ‘foreigner’ and ‘unknown’
- Dummy variables have been created for categorical (object-type) variables.
- The dataset includes 8,792 rows and 43 columns for analysis.
- After removing all the unwanted data, we kept the data retension rate to 82%

Model Building

- Split the data into training and testing sets using a 72:28 ratio.
- Perform feature selection using Recursive Feature Elimination (RFE).
- Run RFE to select 15 variables as output.
- Build the model by removing variables with a p-value greater than 0.05 and a Variance Inflation Factor (VIF) greater than 5.
- Make predictions on the test dataset.
- Achieved an overall accuracy of 77%, precision of 80% and recall of 71%.

ROC Curve



- Identify the optimal cut-off point.
- The optimal cut-off probability is where sensitivity and specificity are balanced.
- From the second graph, the optimal cut-off is determined to be at 0.35.

Conclusion

The most signigicant variables for best lead prediction are

- *TotalVisits*
- *Total Time Spent on Website*
- *Page Views Per Visit*
- *Lead Origin_Lead Add Form*
- *Lead Source_Olark Chat*
- *Lead Source_Welingak Website*
- *Last Activity_Email Bounced*
- *WLast Activity_SMS Sent*
- *What is your current occupation_Working Profes.*
- *Lead Source_Olark Chat*
- *Last Activity_Email Bounced*
- *Last Notable Activity_Unreachable*
- *Last Notable Activity_Had a Phone Conversation*