Lead Score Case study
Batch : DSC52 (Jan-2023)

- GROUP MEMBERS :
- O V M SARMA
- PAVITHRA T
- PARTH GUPTHA

Problem statement

- X Education sells online courses to industry professionals .
- X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'.
- If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Business Objective:

- X education wants to know most promising leads.
- For that they want to build a Model which identifies the hot leads. Deployment of the model for the future use.

Solution Methodology

Data Preprocessing and Analysis :

- Read the data.
- Check and handle duplicate data.
- Check and handle NA values and missing values.
- Drop rows if missing value<1%
- Drop columns, if it contains missing values more than 40%.
- Imputation of the values, if necessary.
- Check and handle outliers in data.

Exploratory Data Analysis :

- Univariate data analysis: value count, distribution of variable etc.
- Bivariate data analysis: correlation coefficients and pattern between the variables etc.
- Performed Feature Scaling & Dummy Variable encoding of the data.

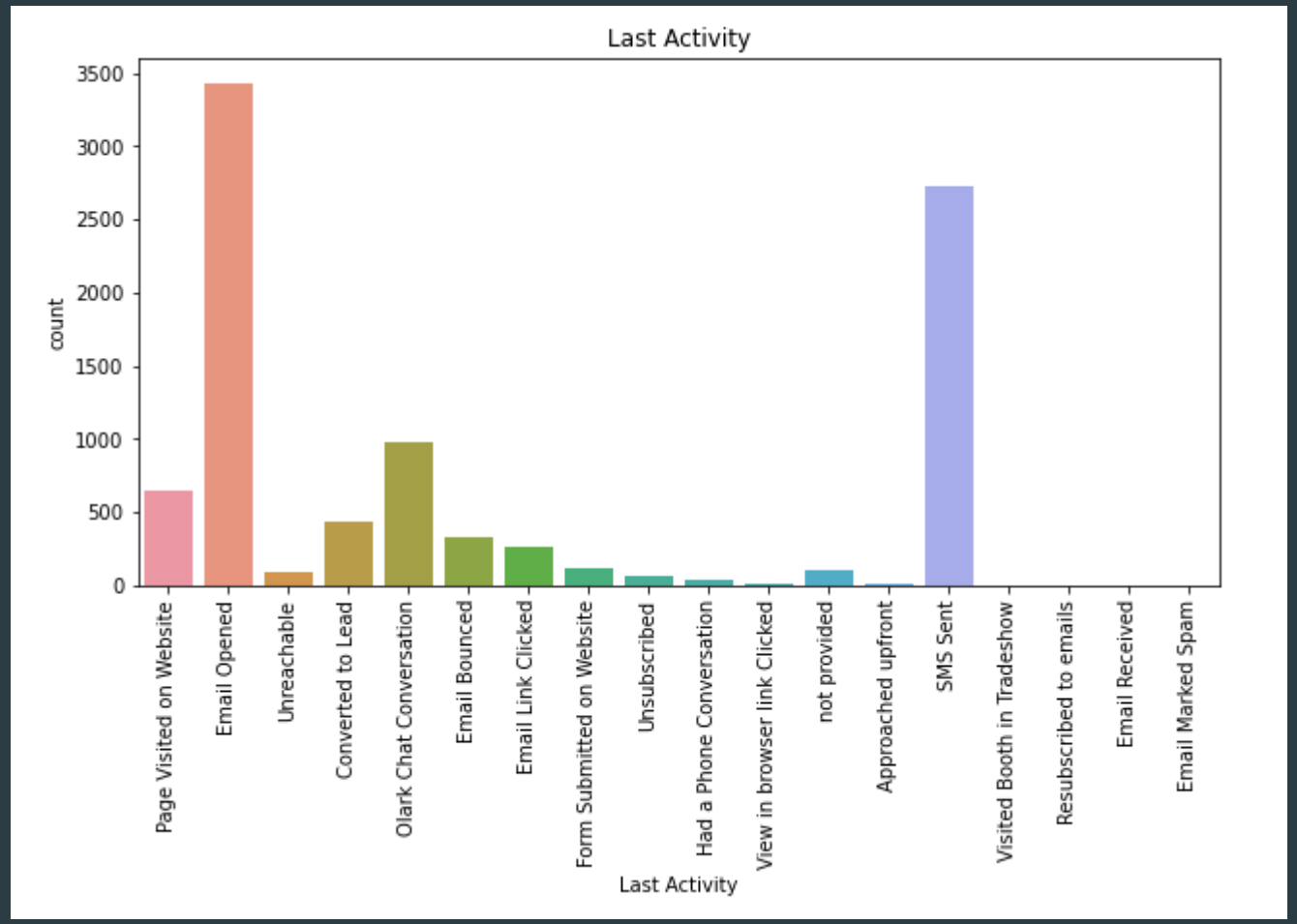
Model Development:

- Logistic Regression method used for developing the model.
- Performed Validation of the model.
- Model presentation.
- Conclusions and recommendations.

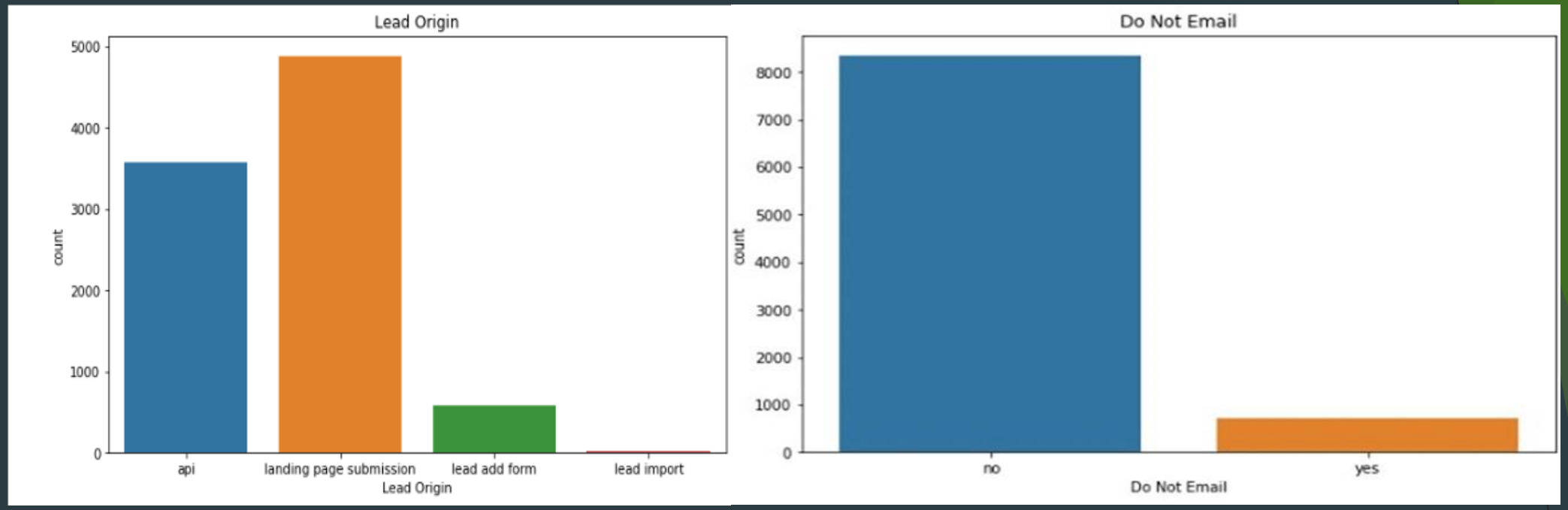
Data Manipulation :

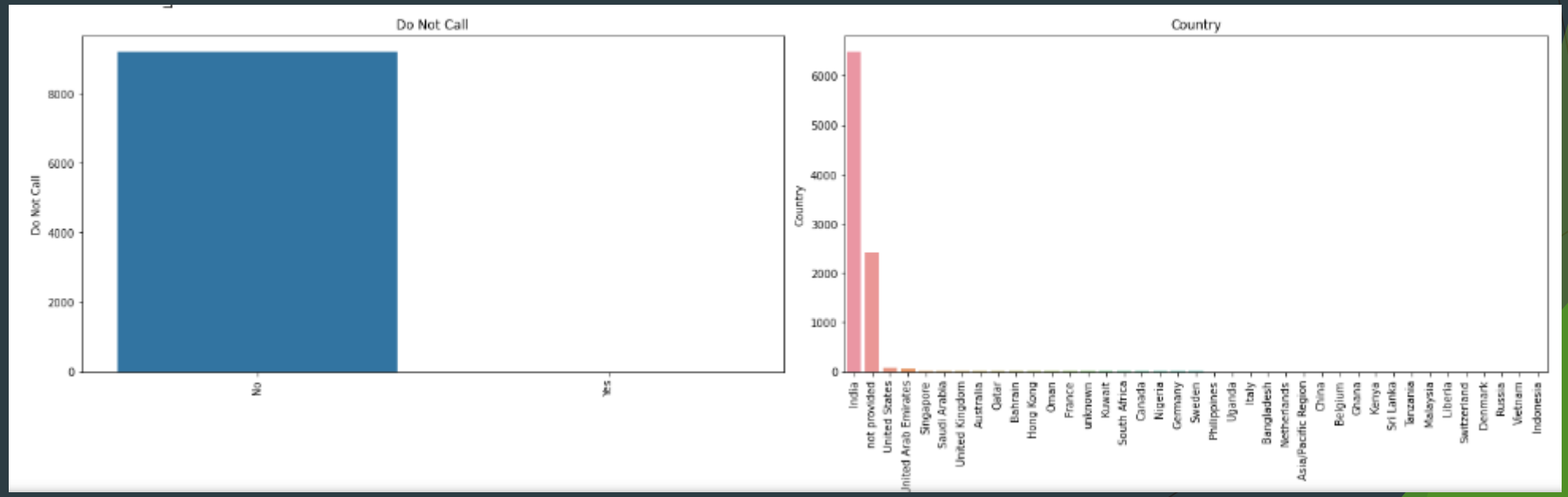
- Total Number of Rows = 37, Total Number of Columns = 9240.
- Dropped row where missing value of Lead Source=0.3
- Dropped 07 columns having more than 45% of missing value
- After these two drop operations, data size is : 9204 rows and 30 columns
- In most columns, selection was not done by the user from dropdown and hence values are mentioned as "Selection". We replaced them with "not selected"
- Checking the data for data imbalance resulted in the following values:
  - Target 0: 5672 and Target 1:3532 and hence data is slightly imbalanced

EDA (Exploratory Data Analysis)



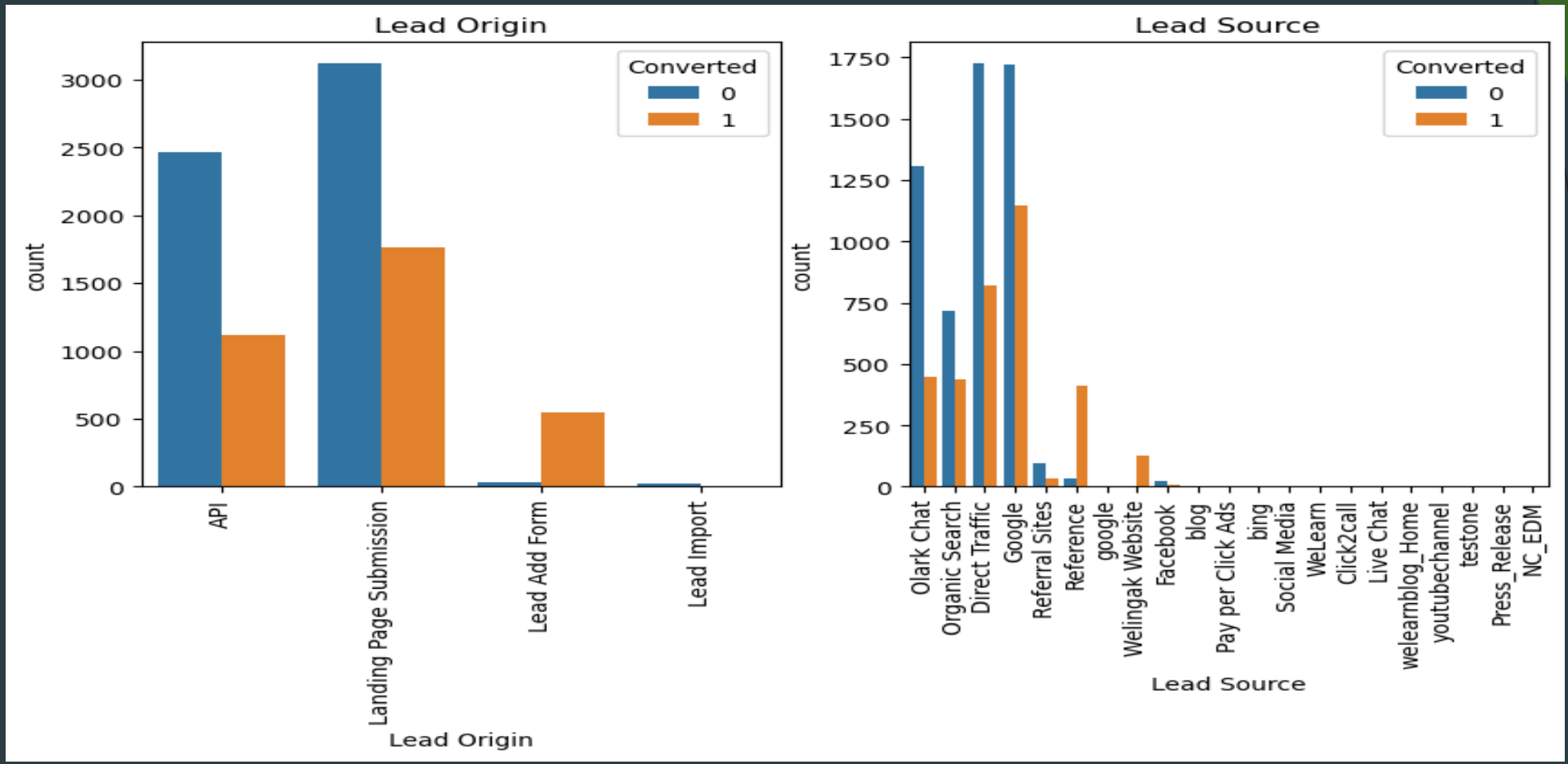
Last activity we can see that are mails and SMS which can be used as medium for communication


Lead Origin
count
Lead Origin
api
landing page submission
lead add form
lead import
Do Not Email
count
Do Not Email
no
yes

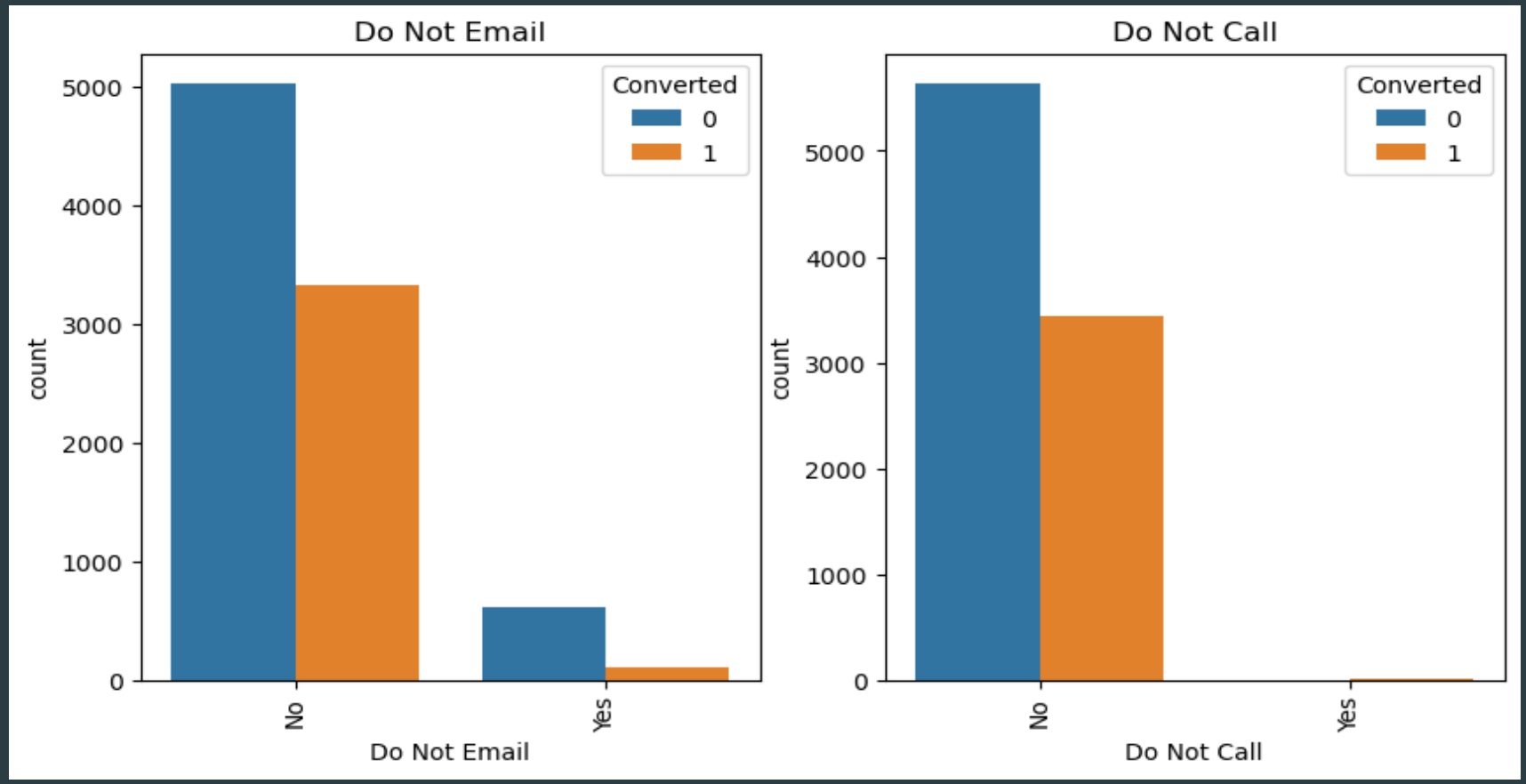

Do Not Call
Do Not Call
No
Yes
Country
Country
India
not provided
United States
United Arab Emirates
Singapore
Saudi Arabia
United Kingdom
Australia
Qatar
Bahrain
Hong Kong
Oman
France
unknown
Kuwait
South Africa
Canada
Nigeria
Germany
Sweden
Philippines
Uganda
Italy
Bangladesh
Netherlands
Asia/Pacific Region
China
Belgium
Ghana
Kenya
Sri Lanka
Tanzania
Malaysia
Liberia
Switzerland
Denmark
Russia
Vietnam
Indonesia

Categorical Variable Relation

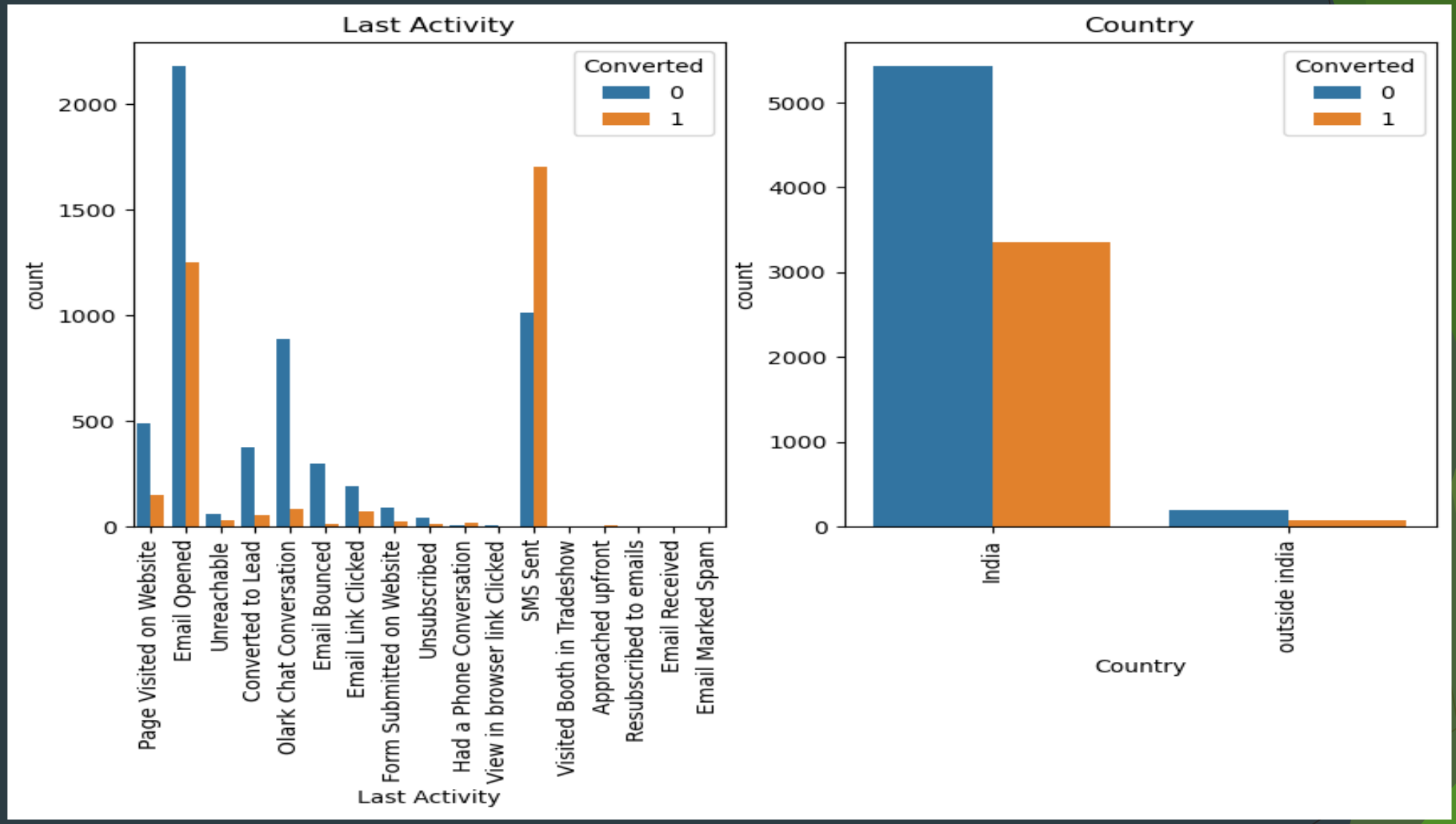
We can observe that there is a good correlation between Total time spent on website and the target variable



API and Landing Page submission helps in identifying the Lead score



Chances of Converting is more if they say Yes to mail and Call



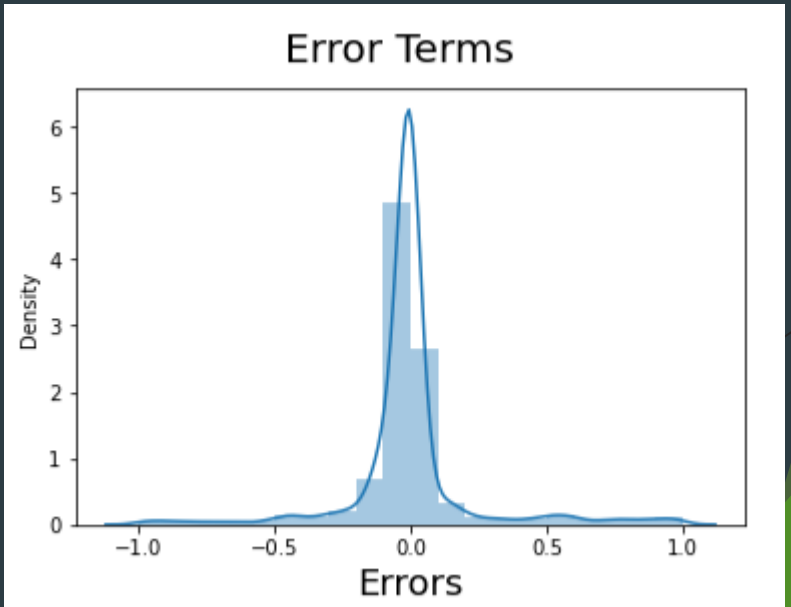
SMS and email helps in conversion and Indians are more interested in Course than other country candidates

Data Preparing for building the model

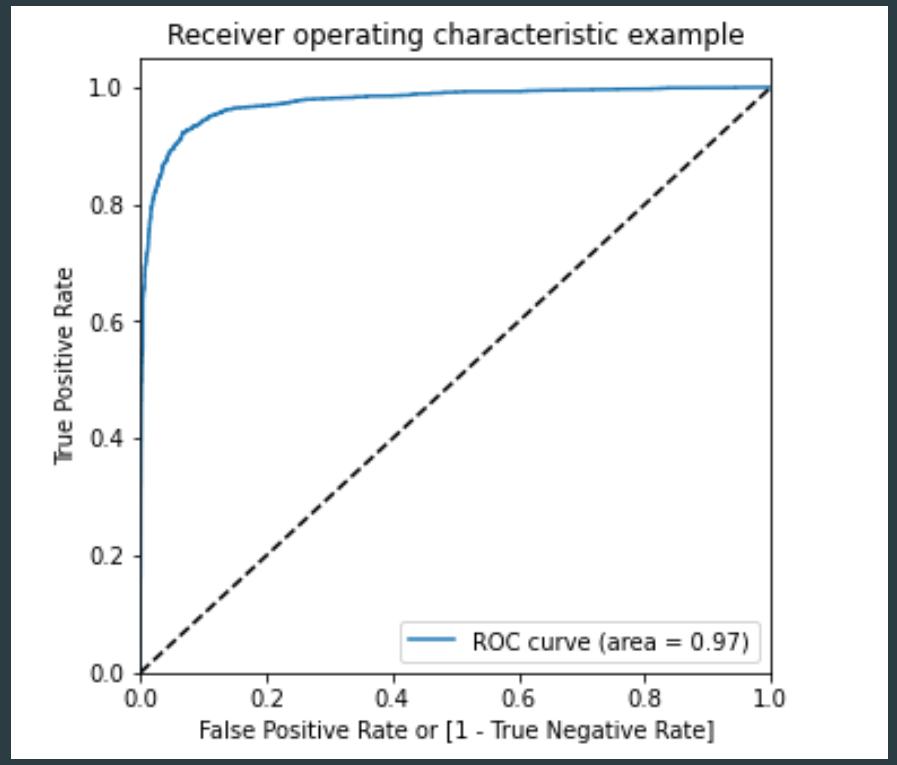
- Binary mapping is done for Categorical Variables with Yes/No values
- Created dummy variables for the remaining categorical variables and deleted the Primary columns
- dropped the Prospect ID as it does not hold any value that helps in taking decision
- checked for Cross Correlation
- drop highly correlated features
- We did split the data with 80% for training and 20% for testing
- Performed feature scaling for the variables ['Total Visits','Total Time Spent on Website', 'Page Views Per Visit']

Model Building

- Splitting the Data into Training and Testing Sets.
- The first basic step for regression is performing a train-test split, we have chosen 80:20 ratio.
- Used RFE for Feature Selection to find VIF value.
- Built Model by removing the variable whose p- value is greater than 0.05 and VIF value is greater than 5.
- Prediction is made on the test data by using the developed model
- Obtained accuracy for the training data is 92.6% with sensitivity 87% and Specificity of 95.8%
- Obtained accuracy for the test data is 93% with sensitivity 88% and Specificity of 96%
- Error Distribution is normal as shown



ROC Curve



We can see that max area is covered under ROC

Conclusion

- It was found that the variables that mattered the most in the potential buyers are :
  - The total time spend on the Website.
  - Total number of visits.

  - When the lead source was:
    - Google
    - Direct traffic
    - Organic search
    - Welingak website

  - When the last activity was:
    - SMS
    - Olark chat conversation
    - Opened email
  - When the lead origin is Lead add format
  - When their current occupation is as a working professional/unemployed.

Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses

- When their current occupation is as a working professional.
  Keeping these in mind the X Education can flourish as they have a very high chance to get
  almost all the potential buyers to change their mind and buy their courses

Thank You