LEAD SCORE CASE STUDY

Group Members
1. Aashka Vijapura
2. Luqman

PROBLEM STATEMENT

X Education sells online courses to industry professionals and 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.

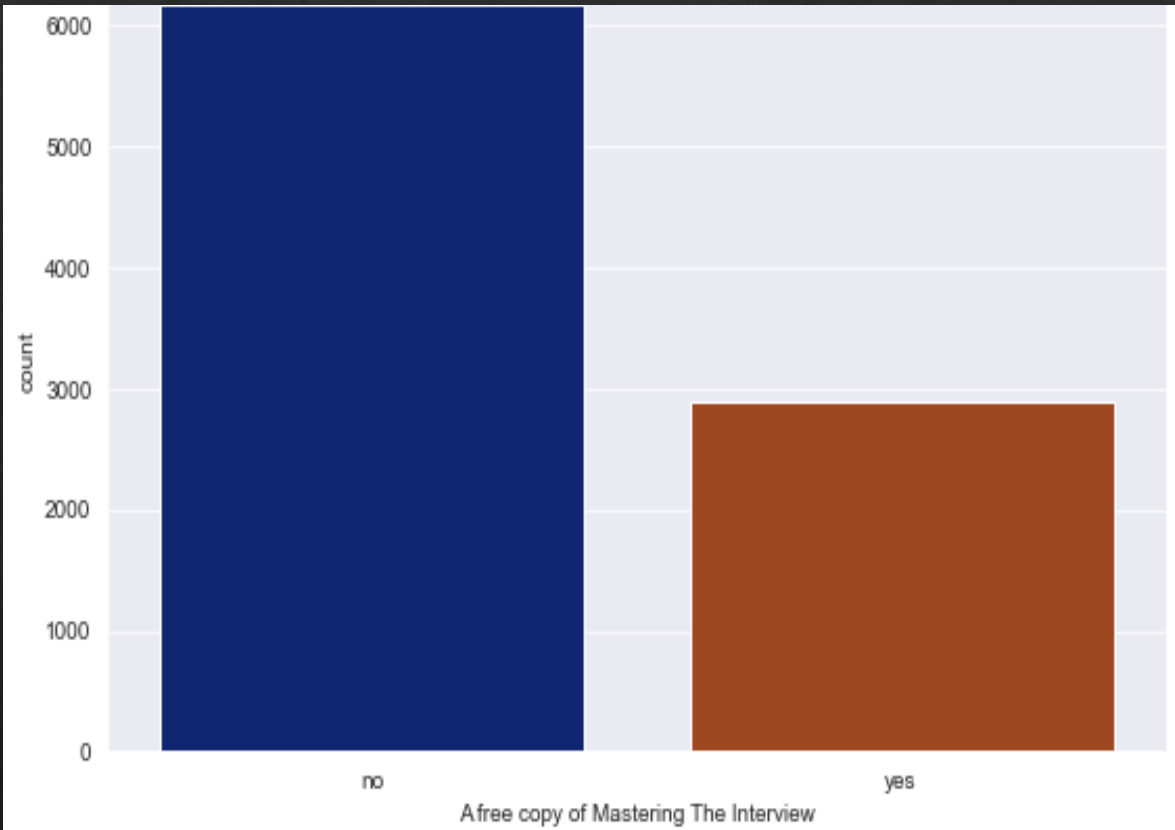
Steps to Build Model

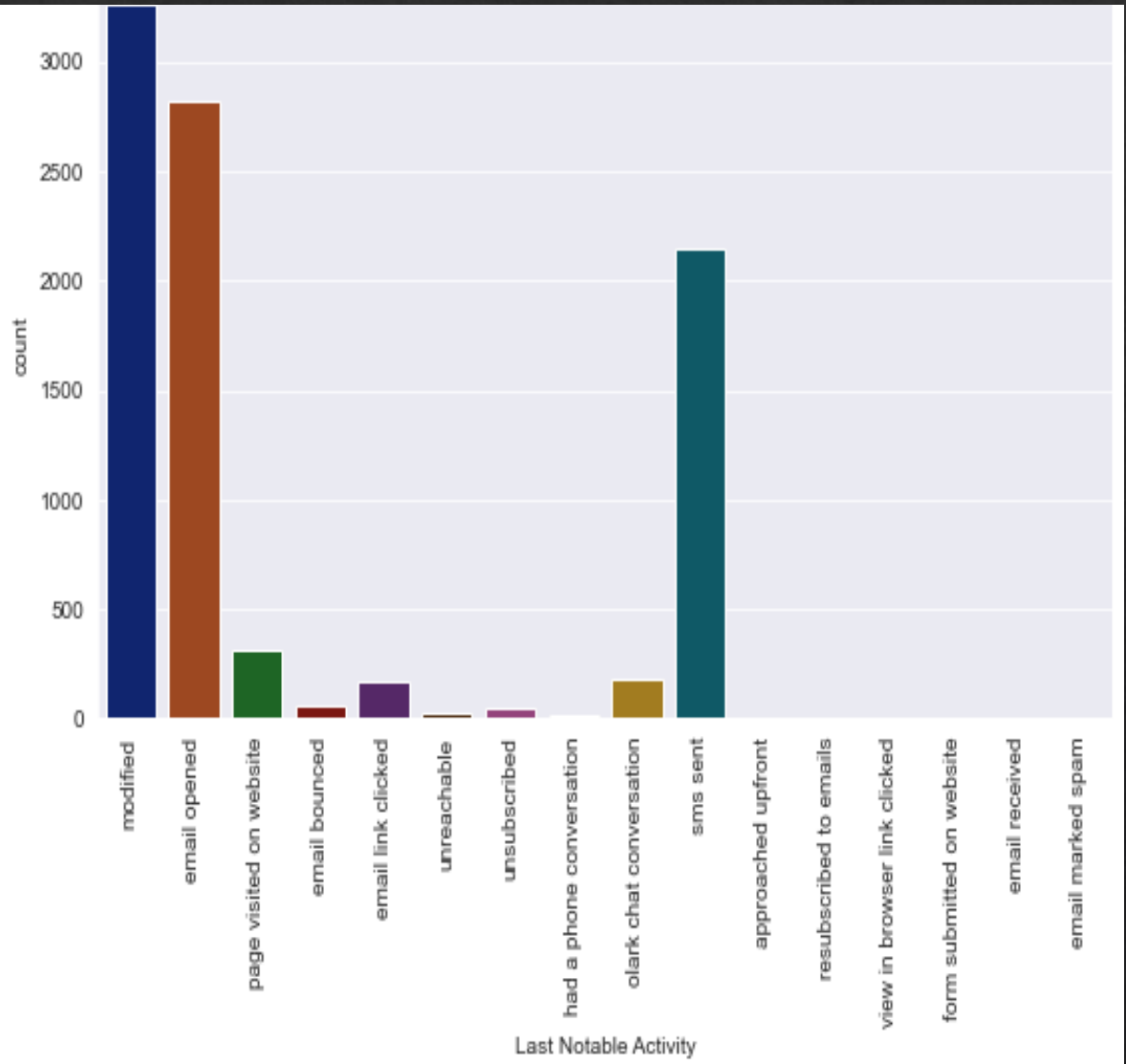
- Data cleaning and data manipulation.
  - Check and handle duplicate data.
  - Check and handle NA values and missing values.
  - Drop columns, if it contains large amount of missing values and not useful for the analysis.
  - Imputation of the values, if necessary.
  - Check and handle outliers in data.
- EDA
  - Univariate data analysis: value count, distribution of variable etc.
  - Bivariate data analysis: correlation coefficients and pattern between the variables etc.
- Feature Scaling & Dummy Variables and encoding of the data.
- Classification technique: logistic regression to build model and prediction.
- Evaluation of the model.
- Model presentation.
- Conclusions and recommendations.

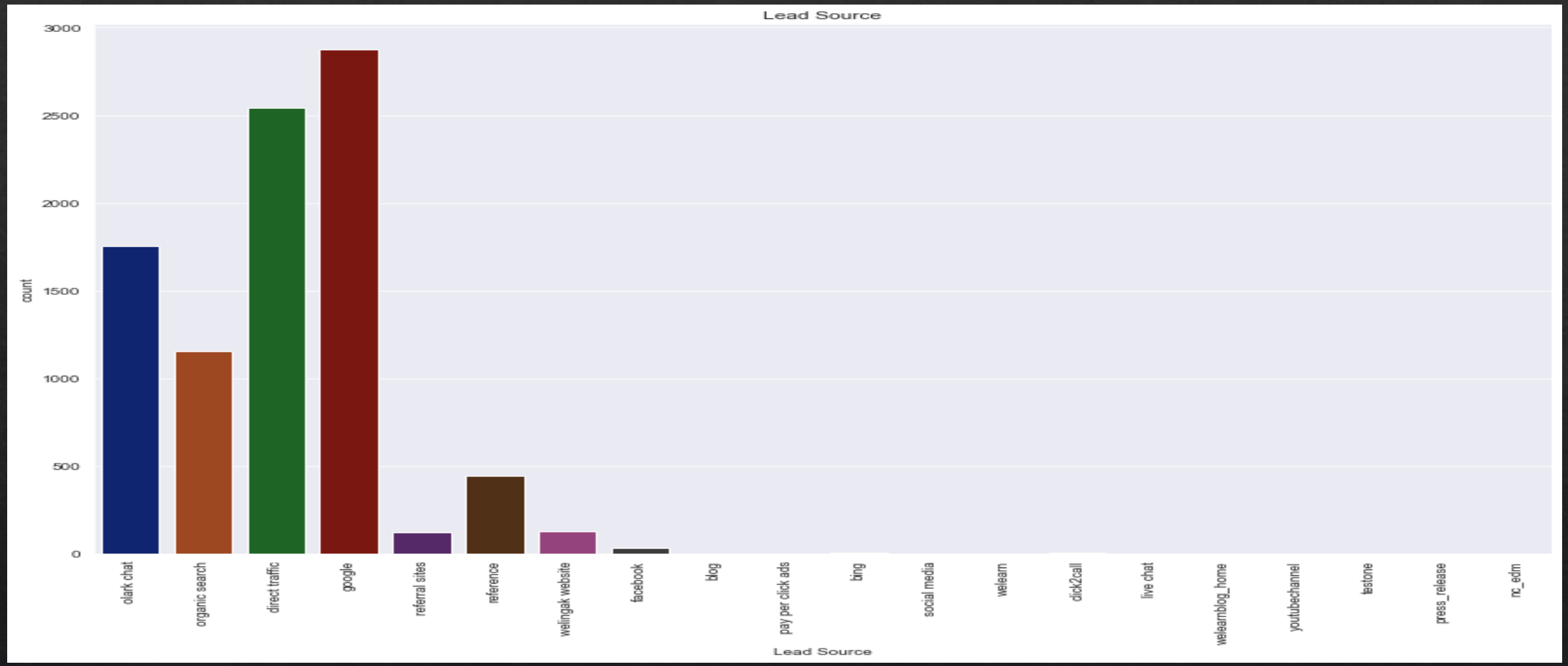
Data Preprocessing

- Total Number of Rows =37, Total Number of Columns =9240.
- Single value features like “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 dropped.
- Removing the “Prospect ID” and “Lead Number” since it is not necessary for the analysis.
- After checking for the value counts for some of the object type variables, we find some of the features which has no enough variance, which we have dropped, the features are: “Do Not Call”, “What matters most to you in choosing course”, “Search”, “Newspaper Article”, “X Education Forums”, “Newspaper”, “Digital Advertisement” etc.
- Dropping the columns having more than 35% as missing value such as ‘How did you hear about X Education’ and ‘Lead Profile’.

Exploratory Data Analysis

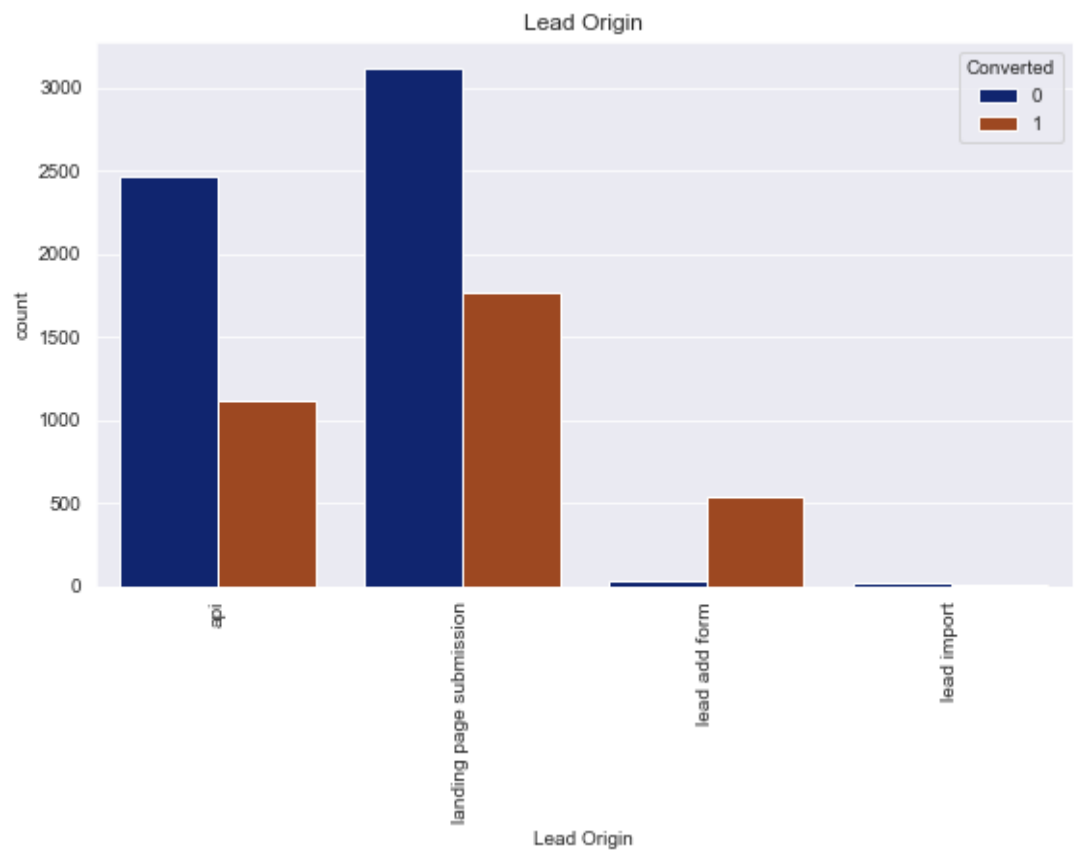


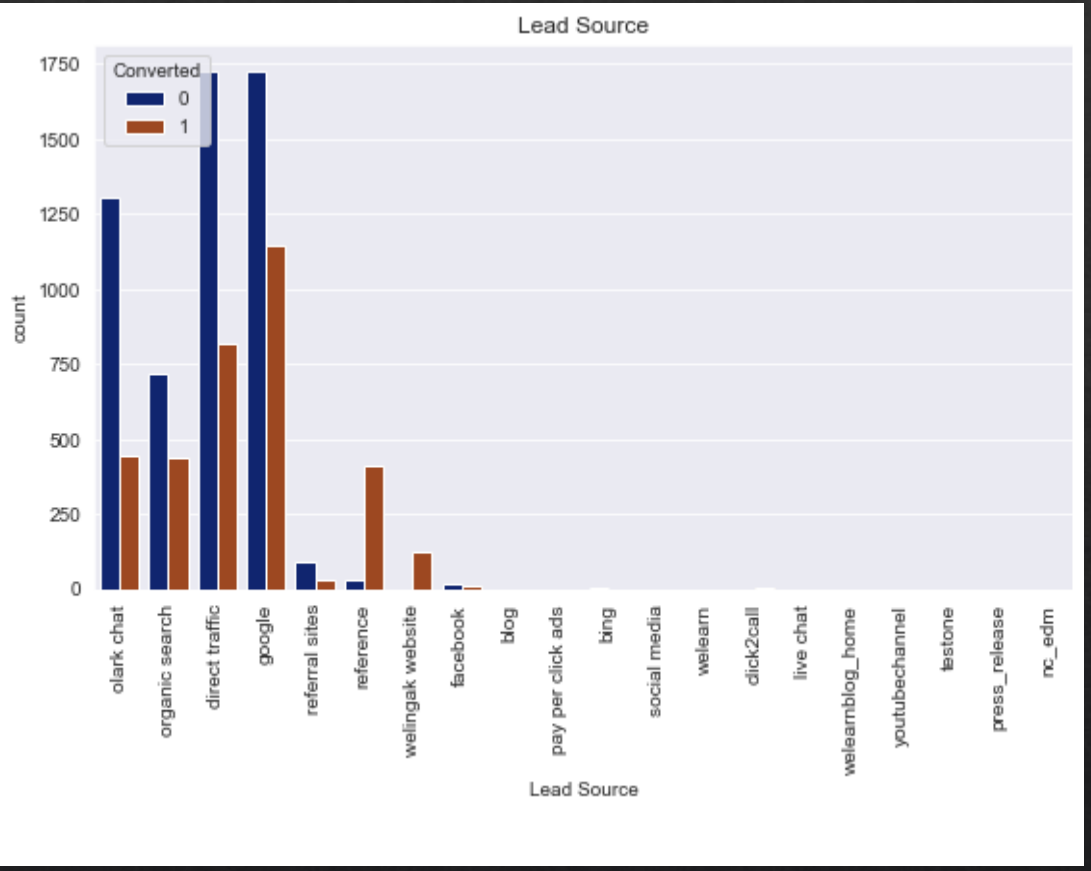




The highest number of leads are through google platform, direct traffic ,olark chat and organic search.

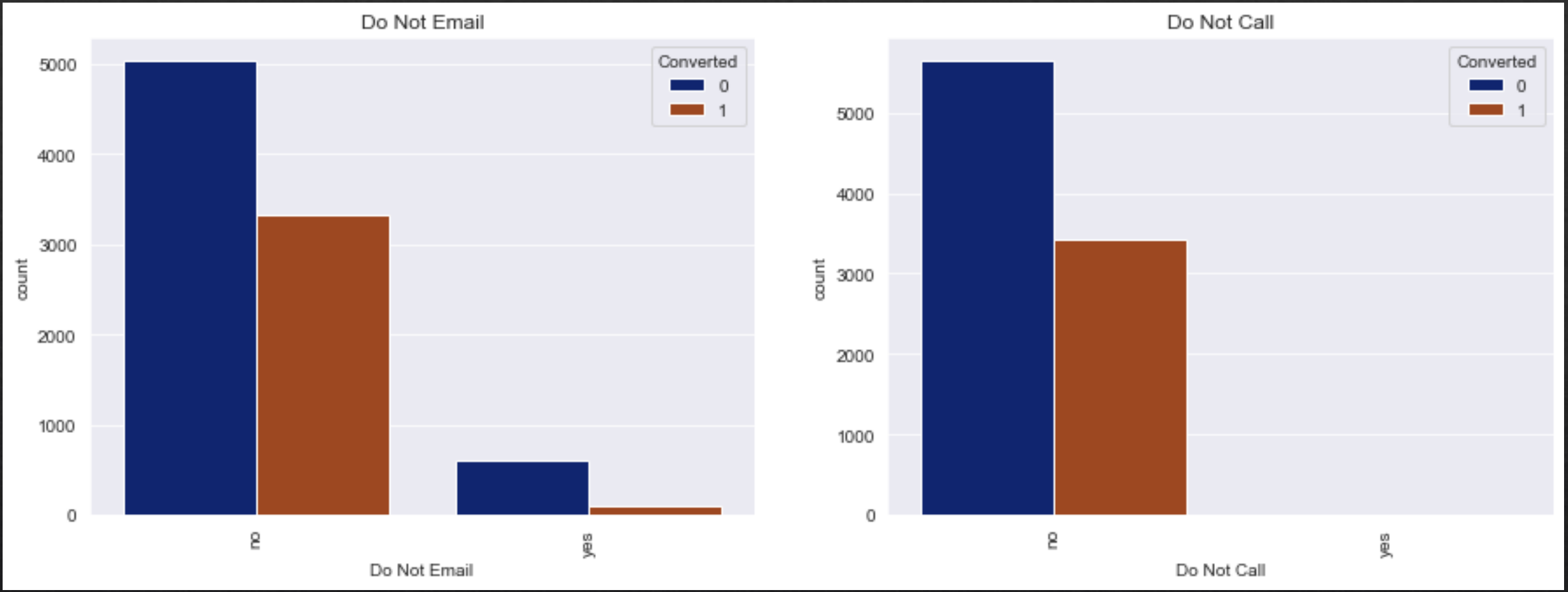
Categorical Analysis





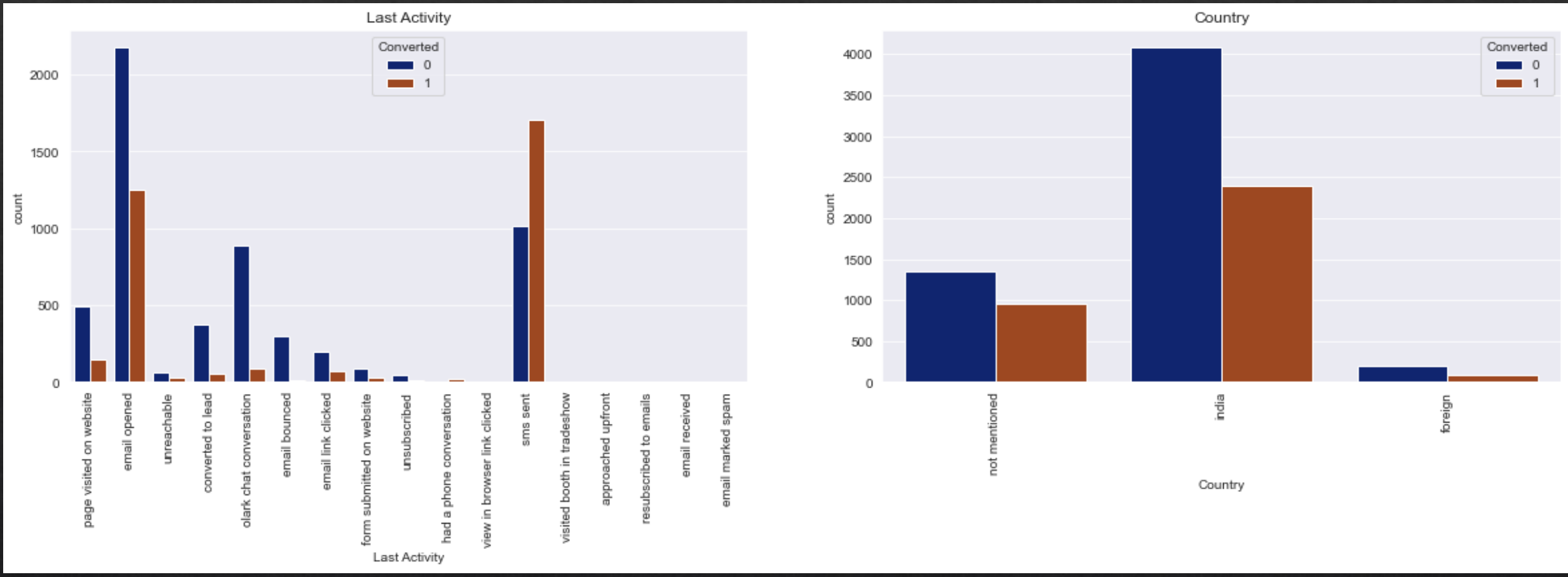
1. Most number of leads were found through landing page submission and lead add form has more number of leads converted than number of leads lost.

2. The number of conversions where more than the leads lost for these two sources reference and welingak website.



1. The leads who have opted for email did not convert as much as the leads who have not opted for email.

2. None of the leads have not opted for call option.



1.The leads whose last activity was SMS sent converted the most.

2.The highest number of leads converted are from India.

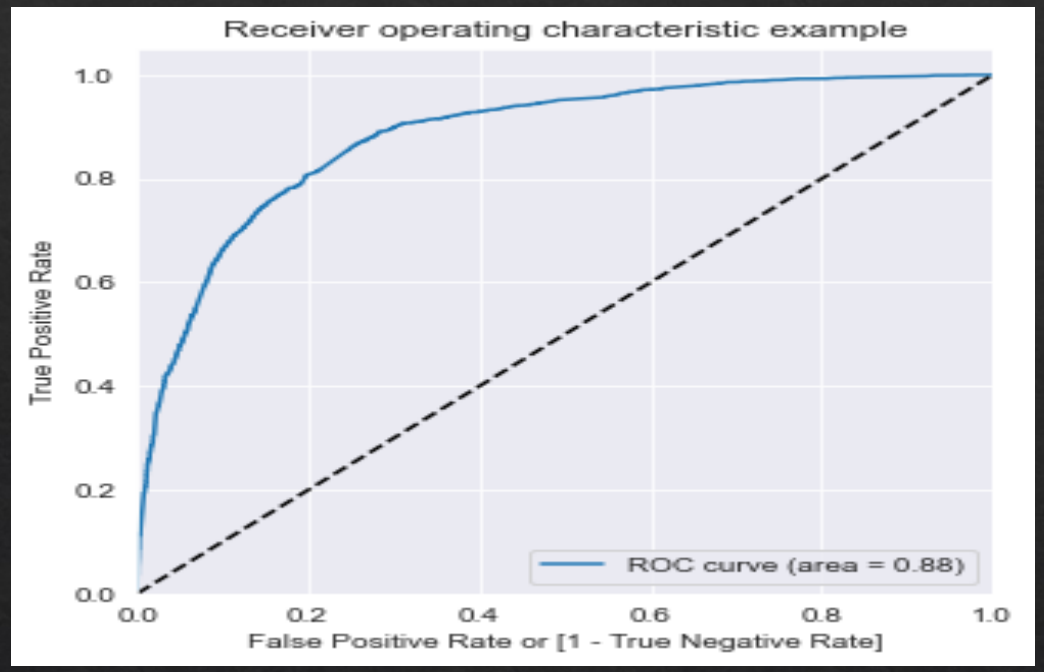
Data Standardization and Dummy variable creation

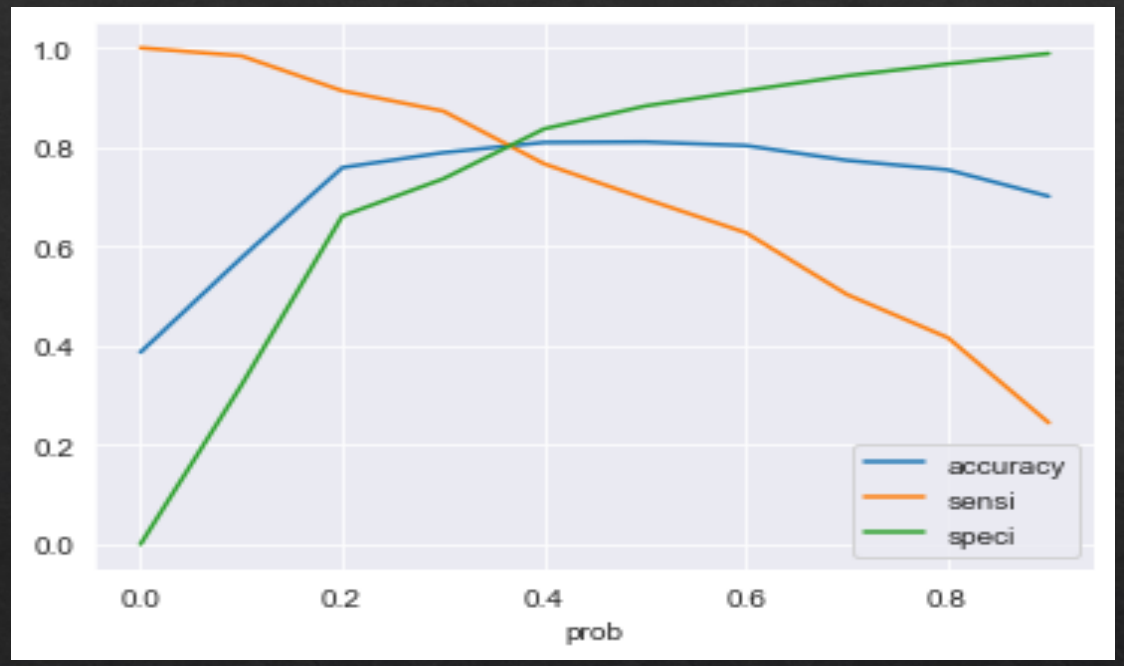
- Numerical Variables are Normalised
- Dummy Variables are created for object type variables
- Total Rows for Analysis: 8792
- Total Columns for Analysis: 43

Model Building

- Splitting the Data into Training and Testing Sets (70% train data set and 30% test data set).
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Use RFE for Feature Selection
- Running RFE with 15 variables as output
- Building Model by removing the variable whose p- value is greater than 0.05 and vif value is greater than 5
- Predictions on test data set
- Overall accuracy 81%

To choose Optimal Cutoff





- **Here we find the Optimal Cut off Point from the above plotted graph.**
- The optimal cut off probability is where we get balanced sensitivity and specificity.
- Higher the area under the curve of ROC better the model
- From the second graph it is visible that the optimal cut off is at 0.35.

Conclusion

It was found that the variables that mattered the most in the potential buyers are :

1. Total number of visits.

2. The total time spend on the Website.

3. When the lead source was:
    a. Google
    b. Direct traffic
    c. Organic search

4. When the last activity was:
    a. SMS
    b. Olark chat conversation

5. When the lead origin is Lead add form.

6. When their current occupation is as a working professional, student, unemployed or other.

7. The option given by the company to get an email from company is an important variable.

8. The Last Notable Activity where student is unreachable.