LEAD SCORE CASE STUDY

By **Sumesh Bisoyi**

Suranjan Banerjee

Suparna Mahapatra

PROBLEM STATEMENT

- □ X Education is an online education company that offers courses to industry professionals.
- □ They market their courses on various websites, and interested people visit their website to browse and fill out forms.
- □ Those who provide their contact information are considered leads. The company gets leads through past referrals as well.
- □ The sales team makes calls and sends emails to convert the leads, but only around 30% of leads get converted.
- □ X Education aims to identify the most promising leads, also known as 'Hot Leads,' to improve their lead conversion rate. By focusing on these leads, the sales team can communicate more effectively and increase conversions
- □ X Education has asked to develop a model that assigns a lead score to each lead, indicating the likelihood of conversion. The CEO has set a target lead conversion rate of 80%.

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.

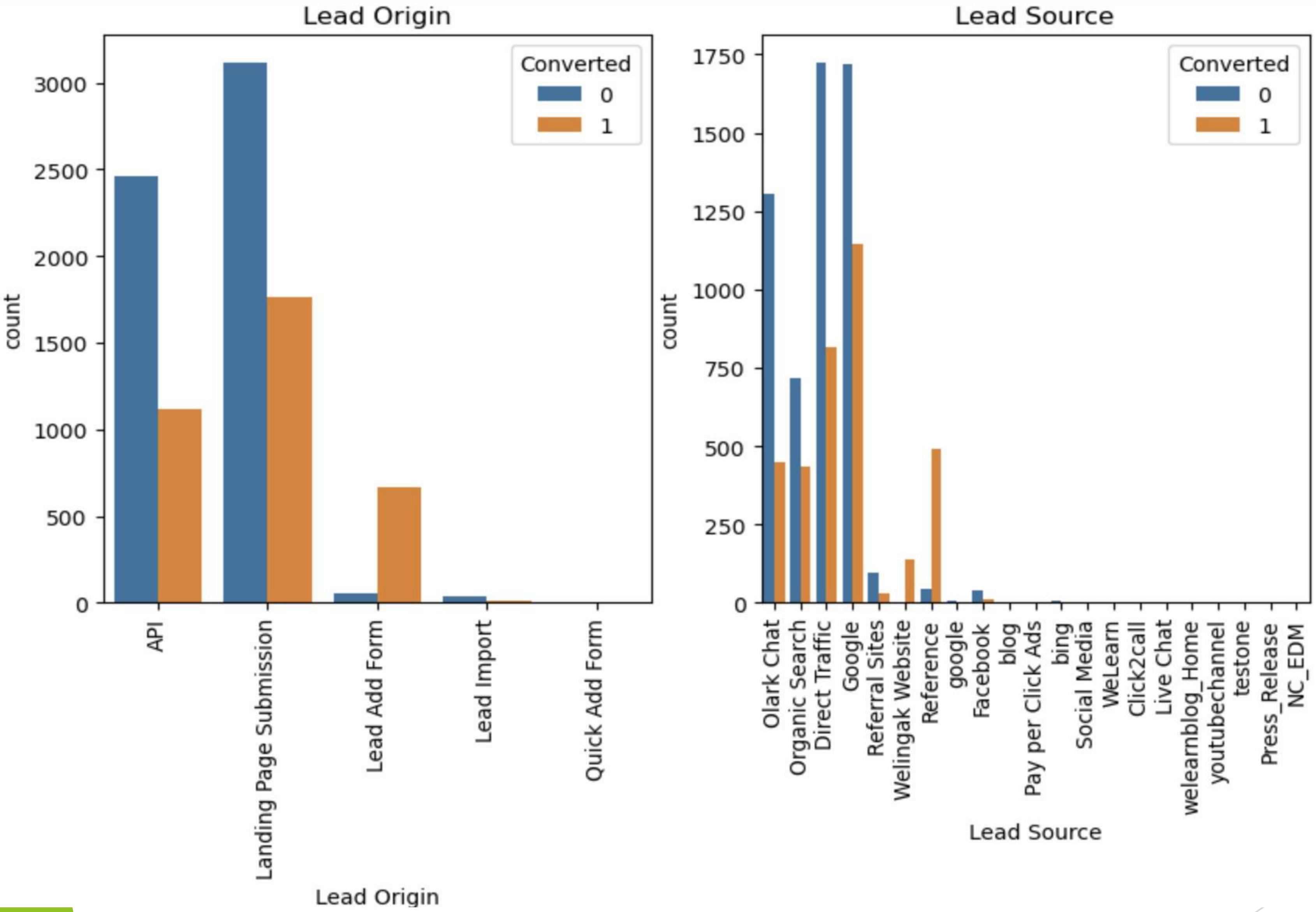
Steps taken

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

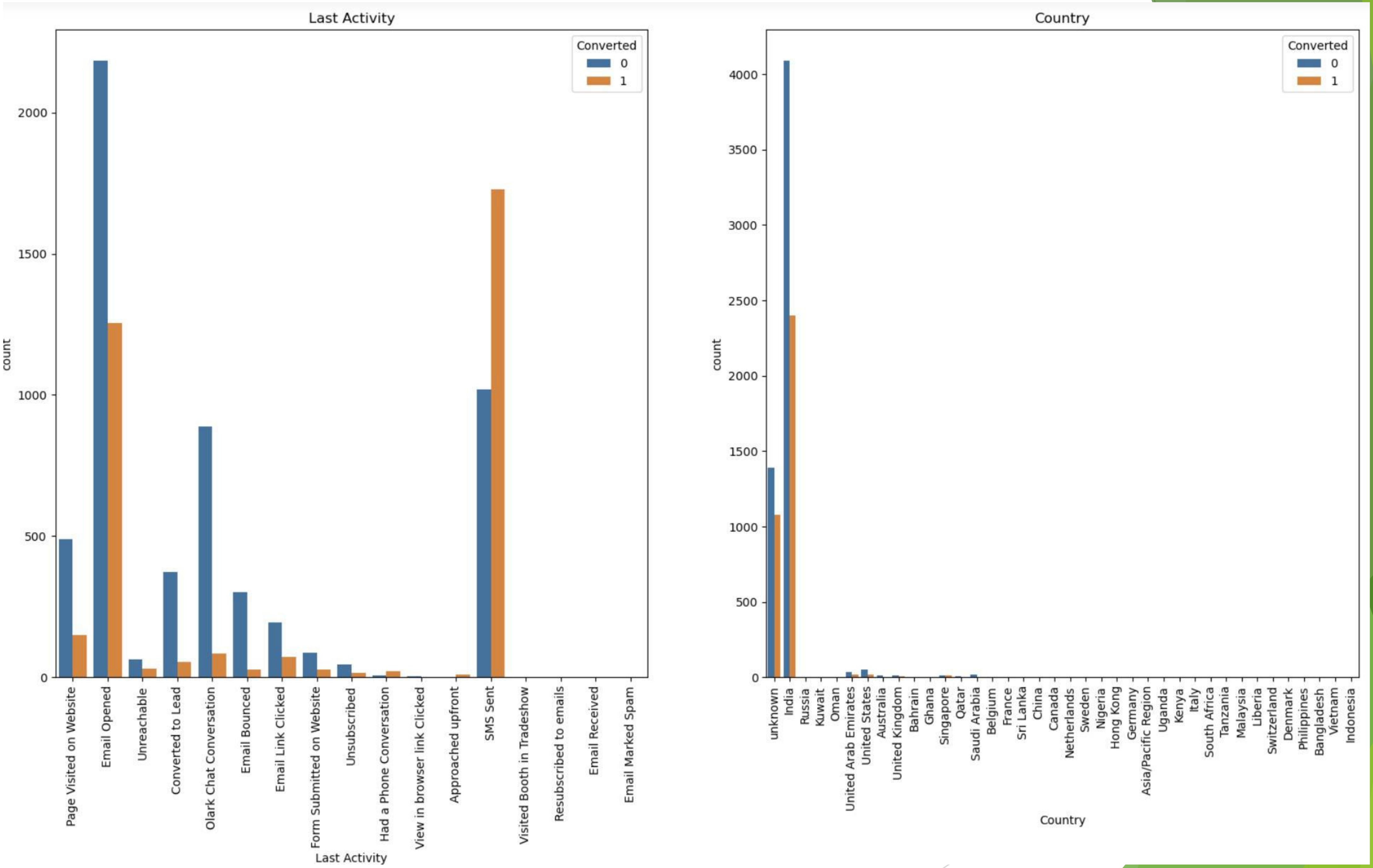
Data Manipulation

- □ 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” which 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’.

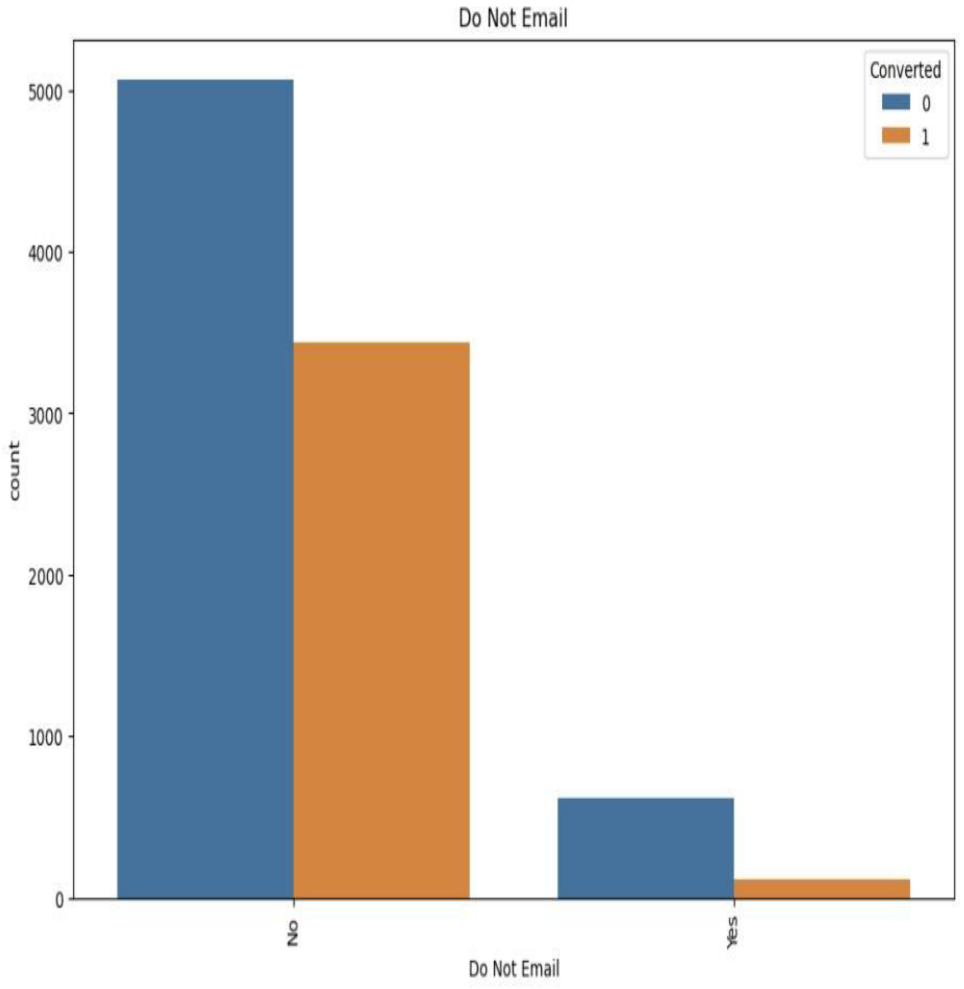
EDA

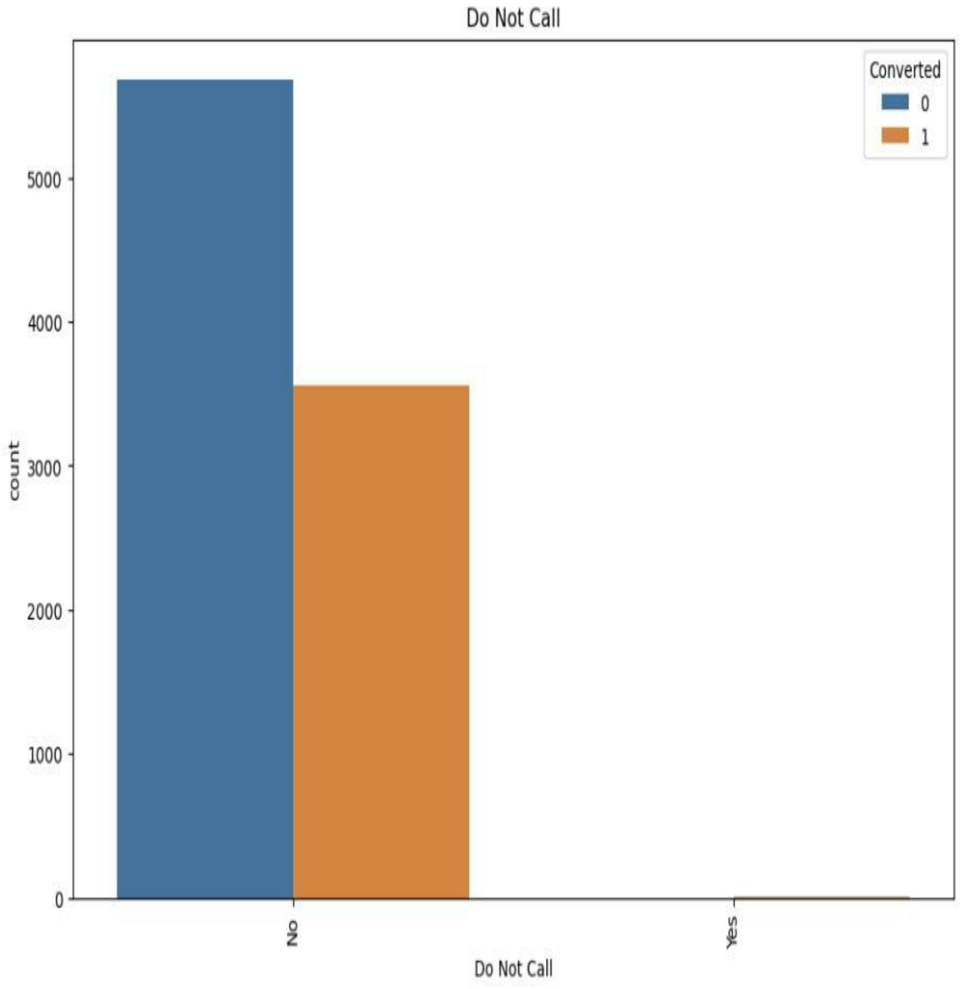


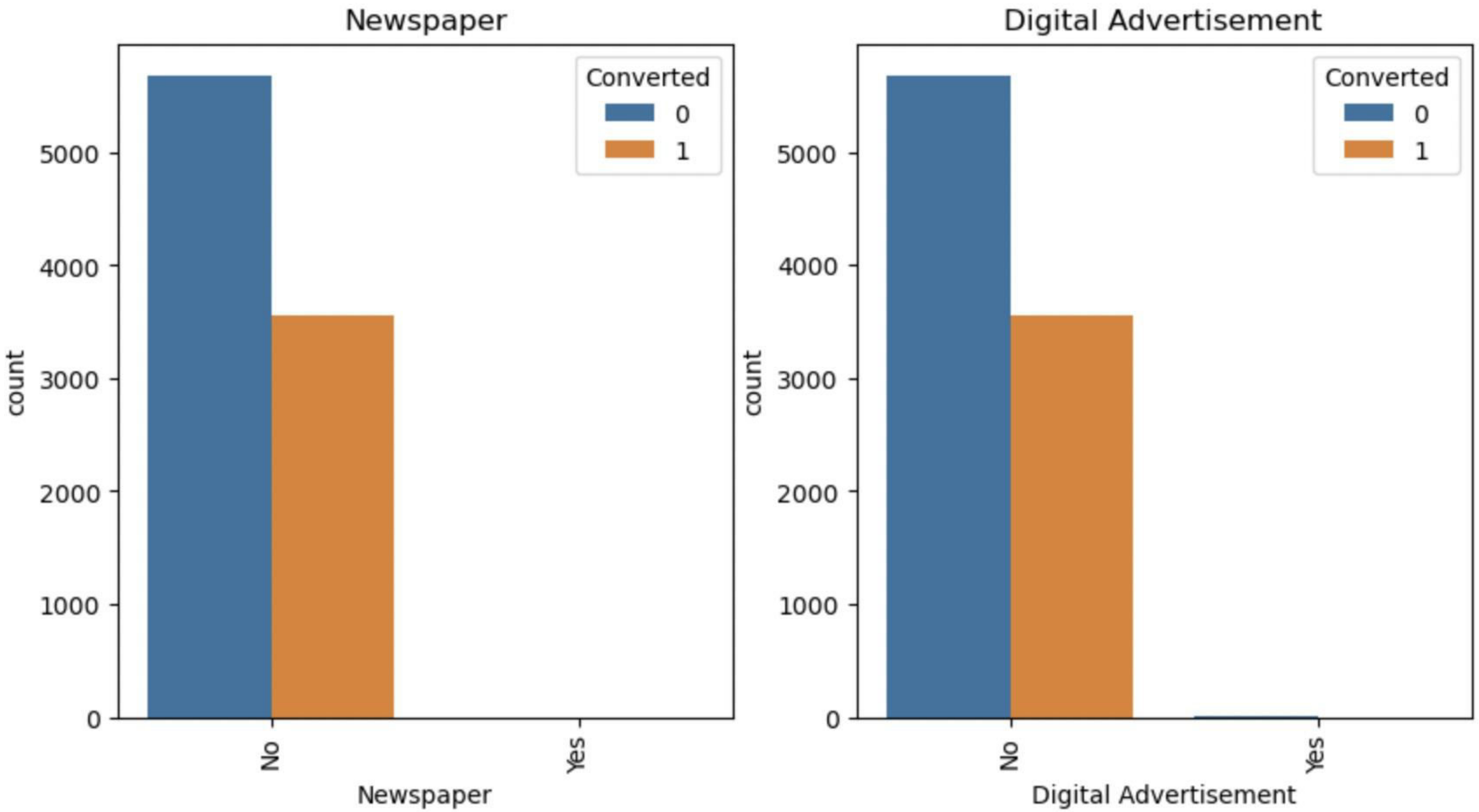
- We can see that lead source via google has good conversion rate .
- We can say same for the direct traffic as well.

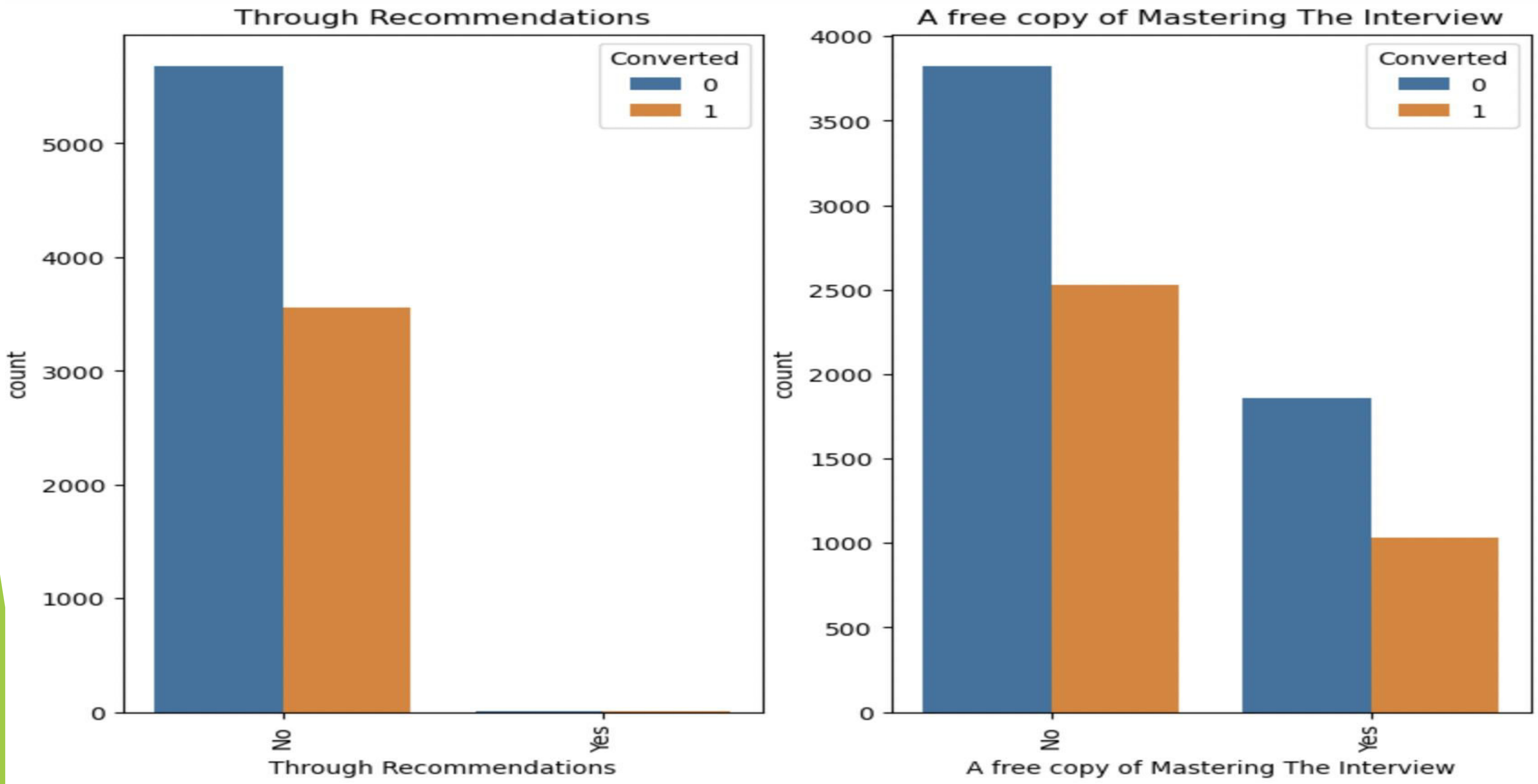

Last Activity

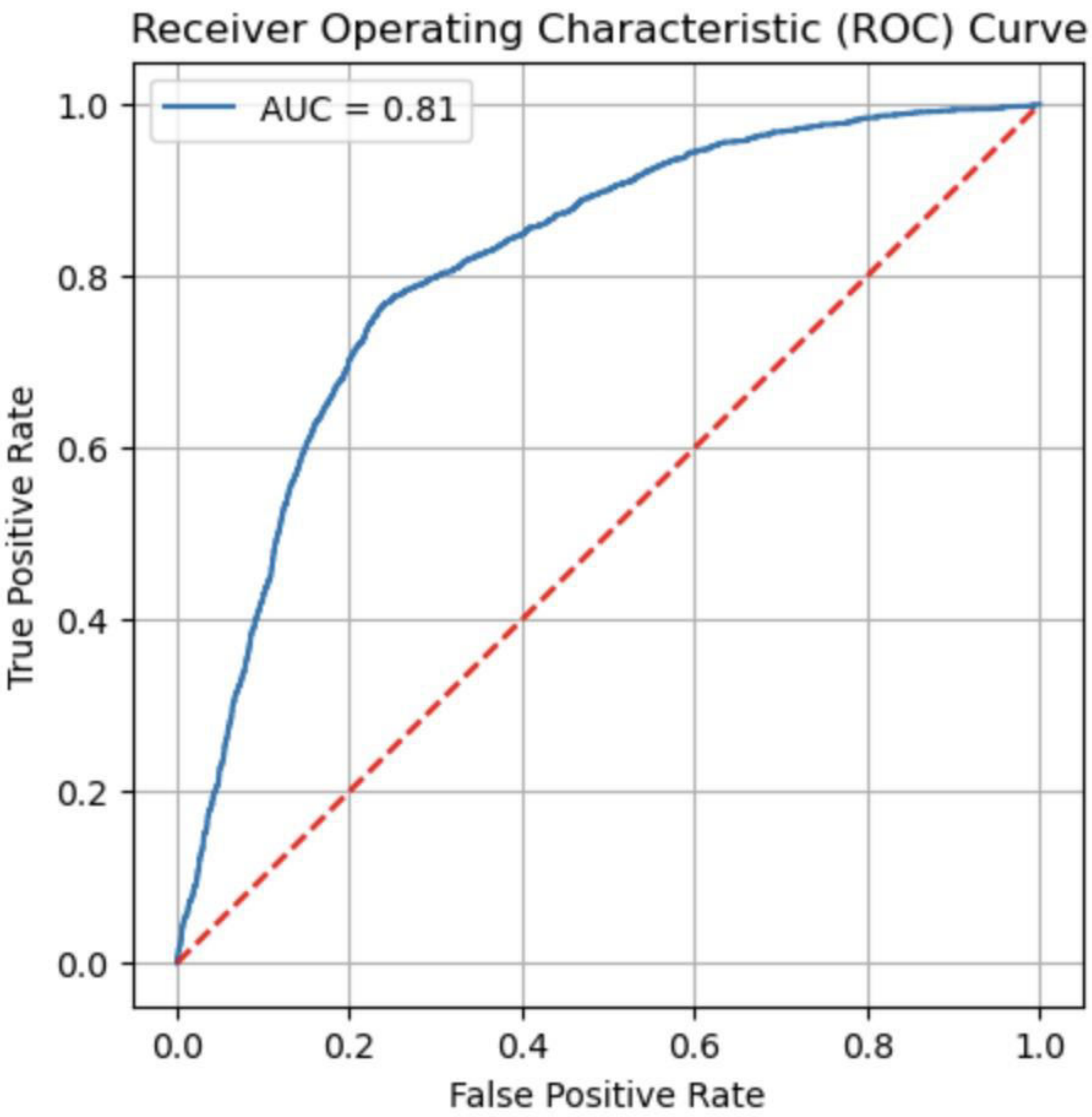
Country


Do Not Email
Converted
0
1
5000
4000
3000
2000
1000
0
count
No
Yes
Do Not Email


Do Not Call
Converted
0
1
5000
4000
3000
2000
1000
0
count
No
Yes
Do Not Call

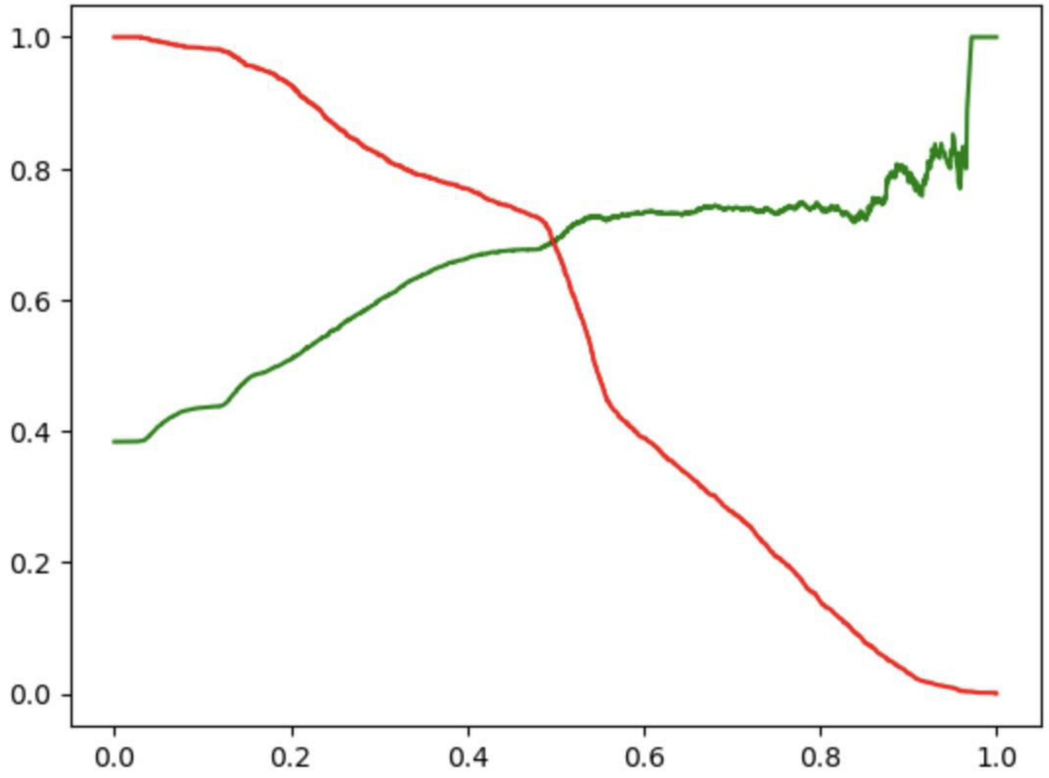

Newspaper
Converted
0
1
5000
4000
3000
2000
1000
0
count
No
Yes
Newspaper
Digital Advertisement
Converted
0
1
5000
4000
3000
2000
1000
0
count
No
Yes
Digital Advertisement


Through Recommendations
Converted
0
1
count
0
1000
2000
3000
4000
5000
No
Yes
Through Recommendations
A free copy of Mastering The Interview
Converted
0
1
count
0
500
1000
1500
2000
2500
3000
3500
4000
No
Yes
A free copy of Mastering The Interview



The area under ROC curve is 0.81, which is quite good!.

While for precision-recall tradeoff we got cutoff value of 0.5



Conclusion

- Hence variables that mattered most are as follows-
- Total Time Spent on Website
- Page Views Per Visit
- Do Not Email
- What is your current occupation_Working Professional
- Lead Origin_Landing Page Submission
- Specialization_Industry Specializations
- A free copy of Mastering The Interview
- TotalVisits