Lead Score Case Study

BY: TANMAY SRIVASTAVA

JAYANTA BOSE

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

Methods to Find The Solution

The method to find the solution to the problem and to achieve Business objective are:

- Data Cleaning and Data manipulation: In the Data Cleaning and Data Manipulation we done the following steps:

a) Checked duplicate data and handled them.

b) Checked and handled missing values and NA values.

c) Dropped columns if they are not useful for analysis and if they contain large amount of missing values.

d) Imputation of missing values if necessary.

e) Checked and handled the outliers.

- EDA: In the EDA had dine the following steps:

a) Univariate data analysis: value count, distribution of variable etc.

b) Bivariate data analysis: correlation coefficients and pattern between the variables etc.

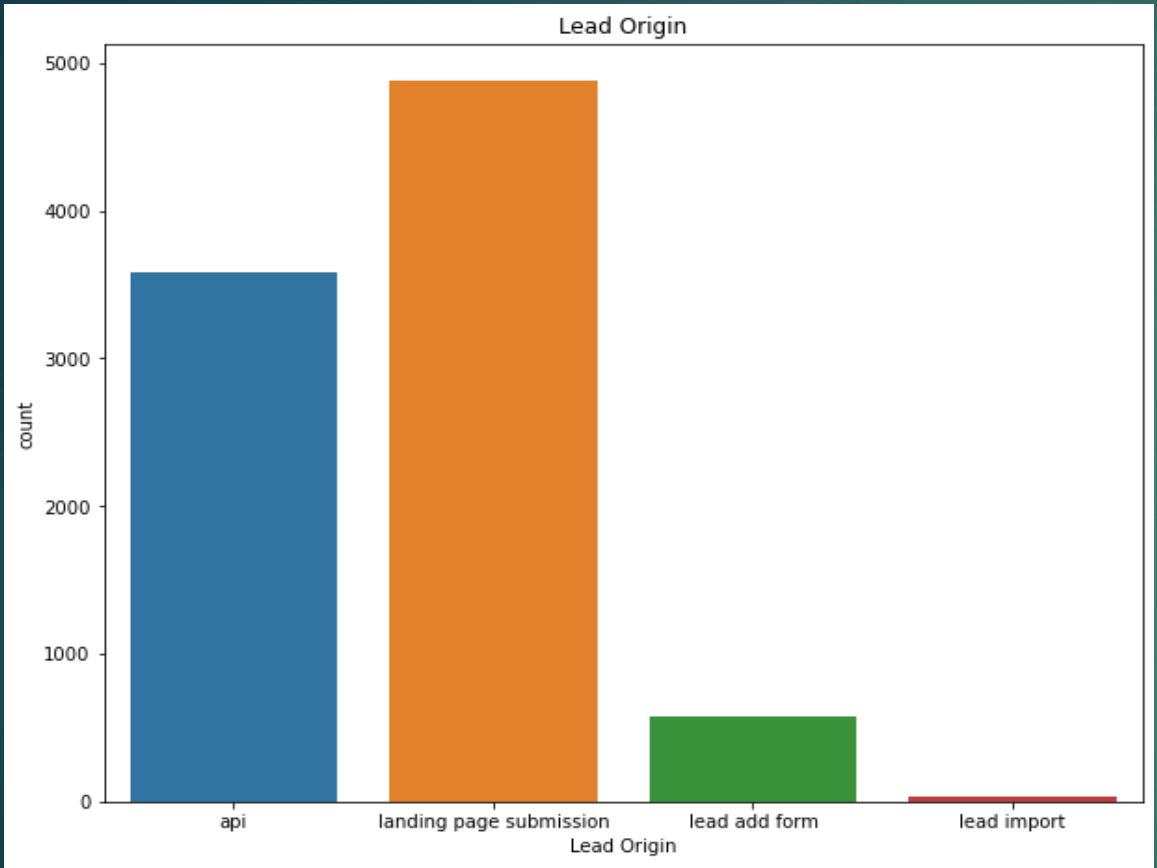
- Feature Scaling & Dummy Variables and encoding of the data.
- Classification technique: logistic regression used for the model making and prediction.
- Validation of the model.
- Model presentation.
- Conclusions and recommendations.

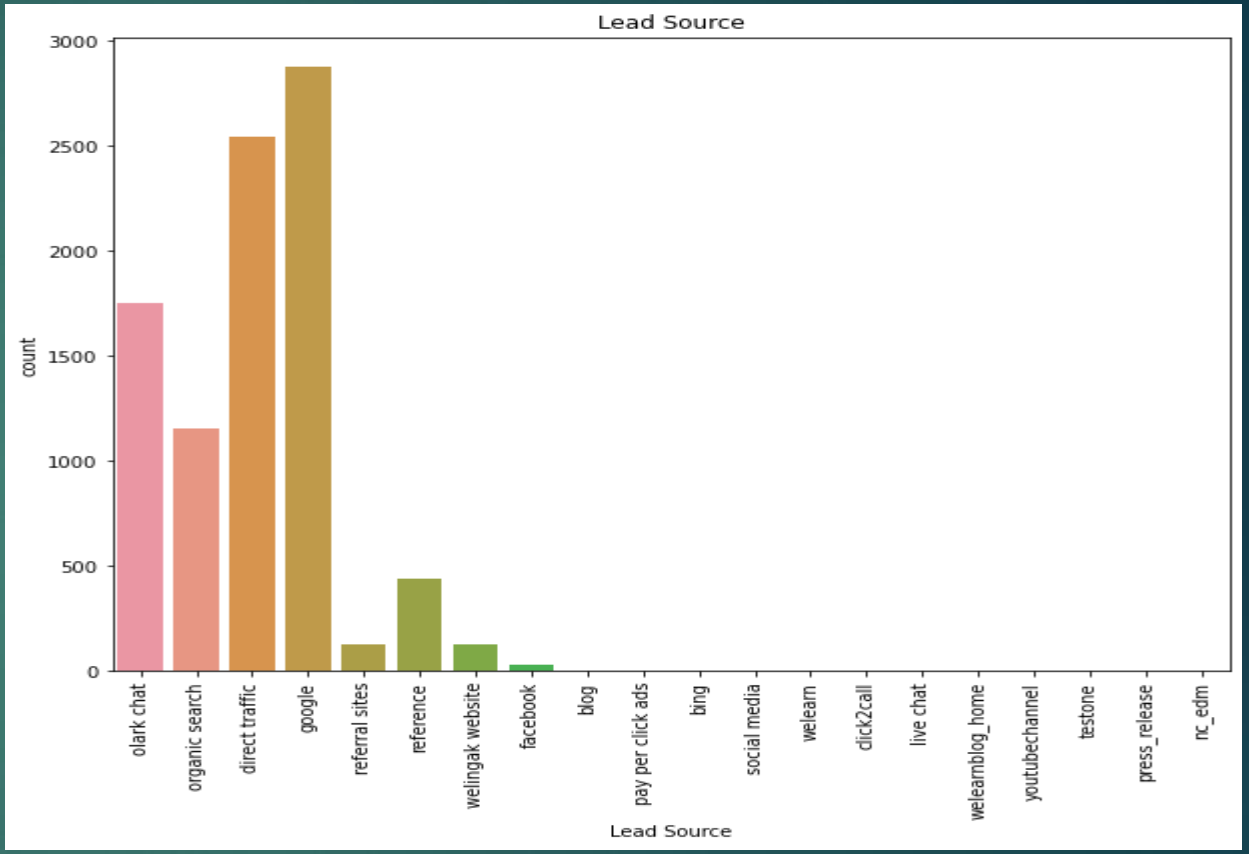
Data Manipulation

- Total Number of Rows is equal to 9240, Total Number of Columns is equal to 37.
- 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’

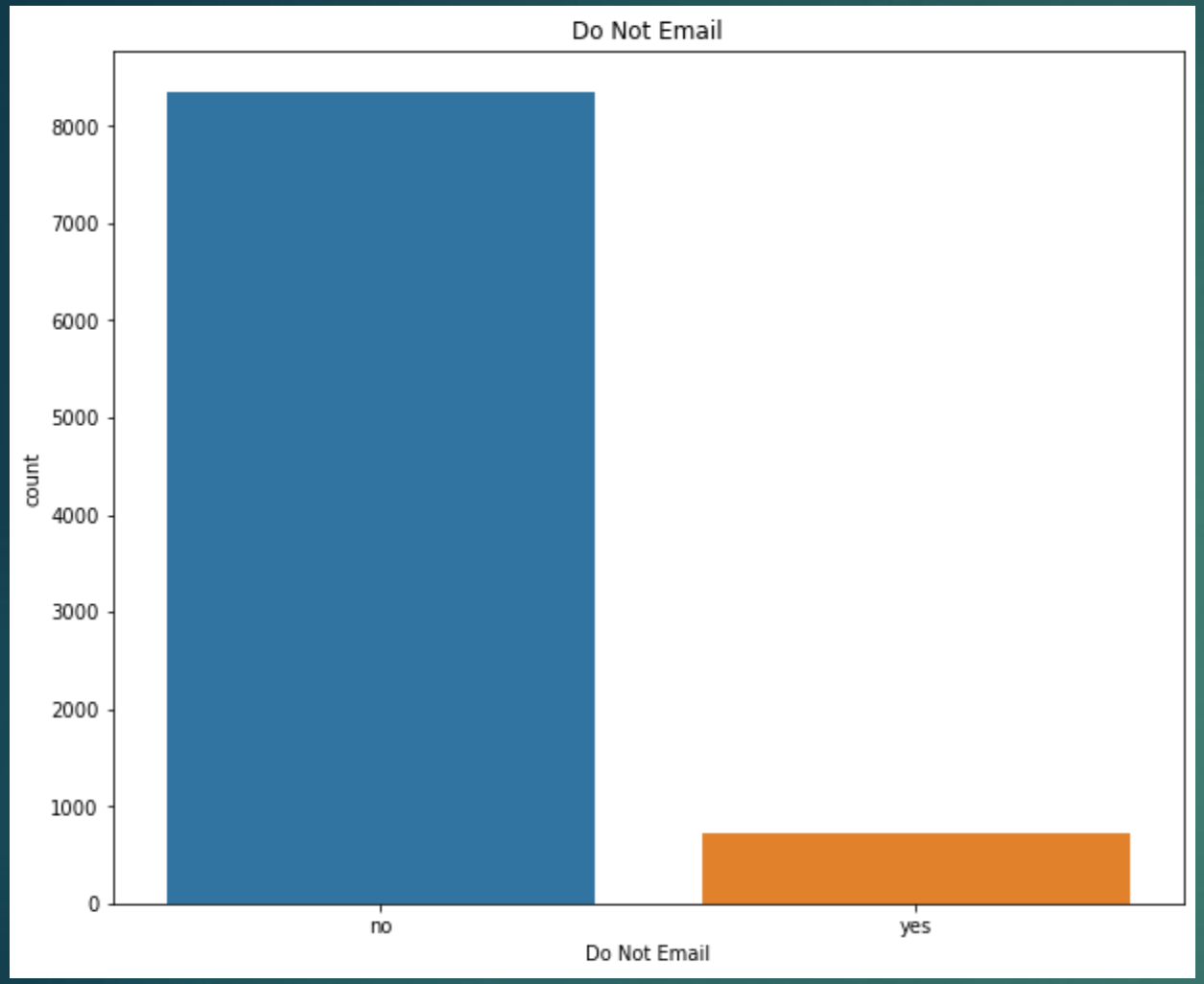
EDA

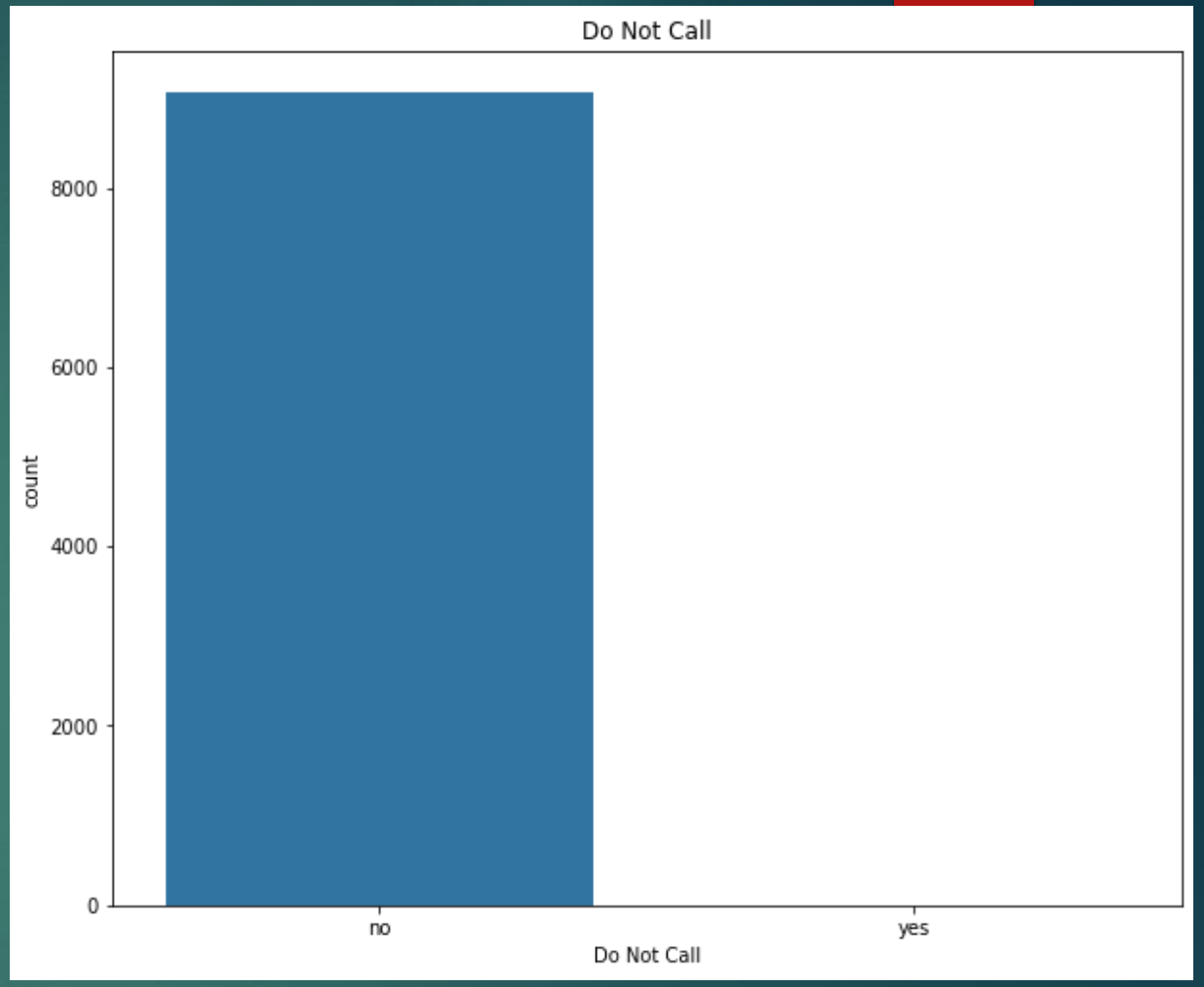
Univariate Analysis of Categorical Variables



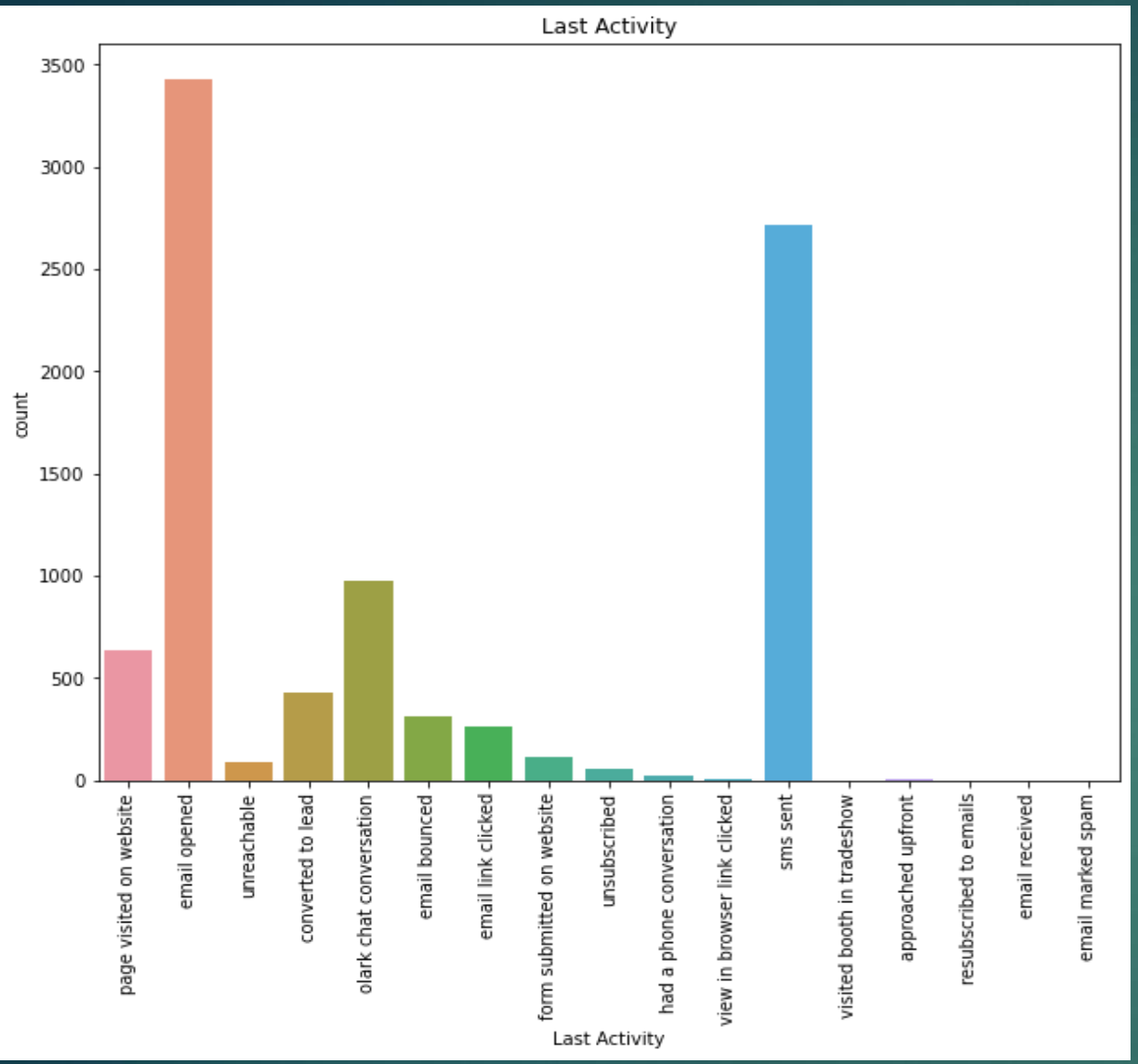


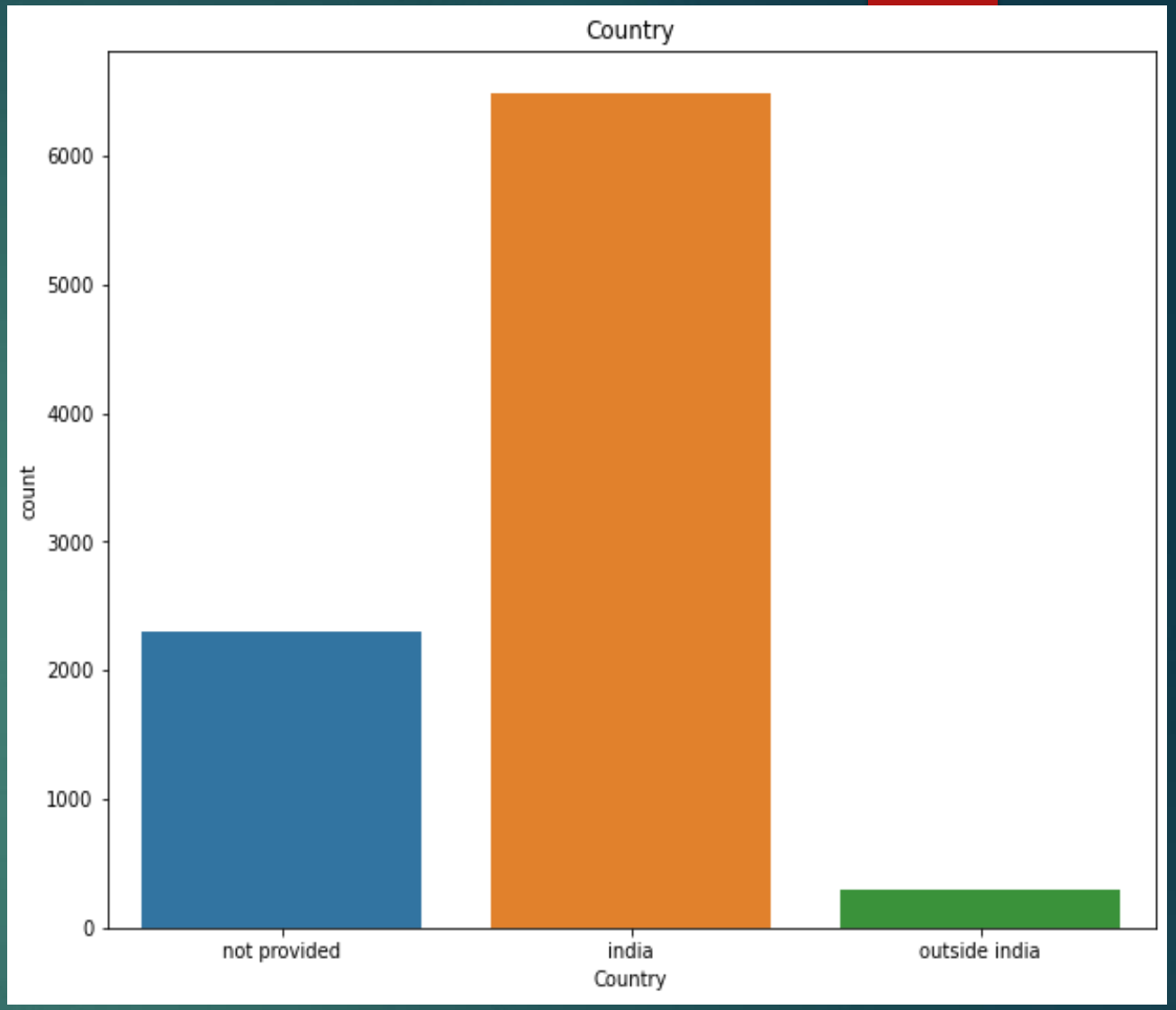
We can see from the above plot of Lead Origin that landing page submission has highest number of count than api,lead add form and lead import and also we can see from the above plot of Lead Source that google has highest number of count than others.



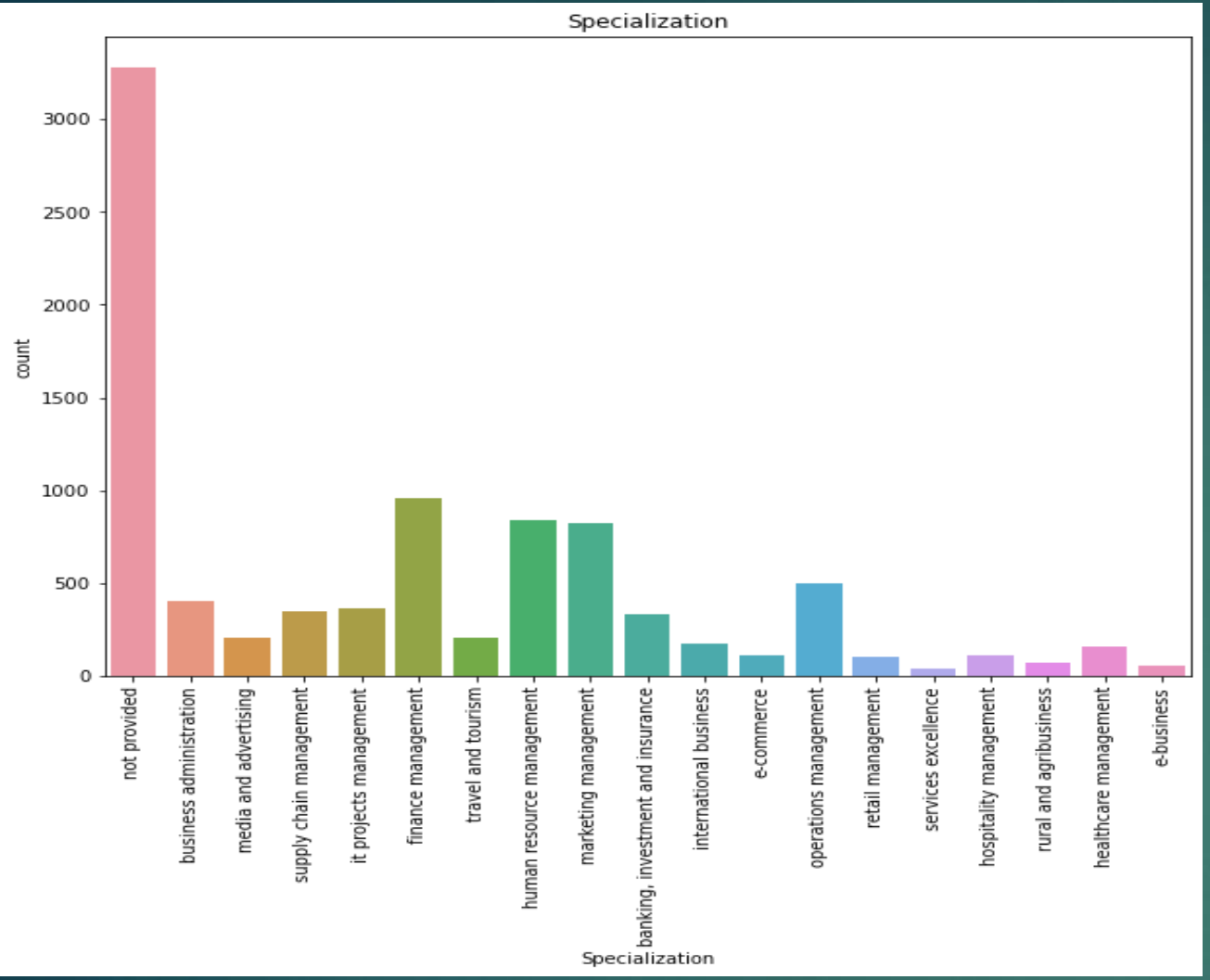


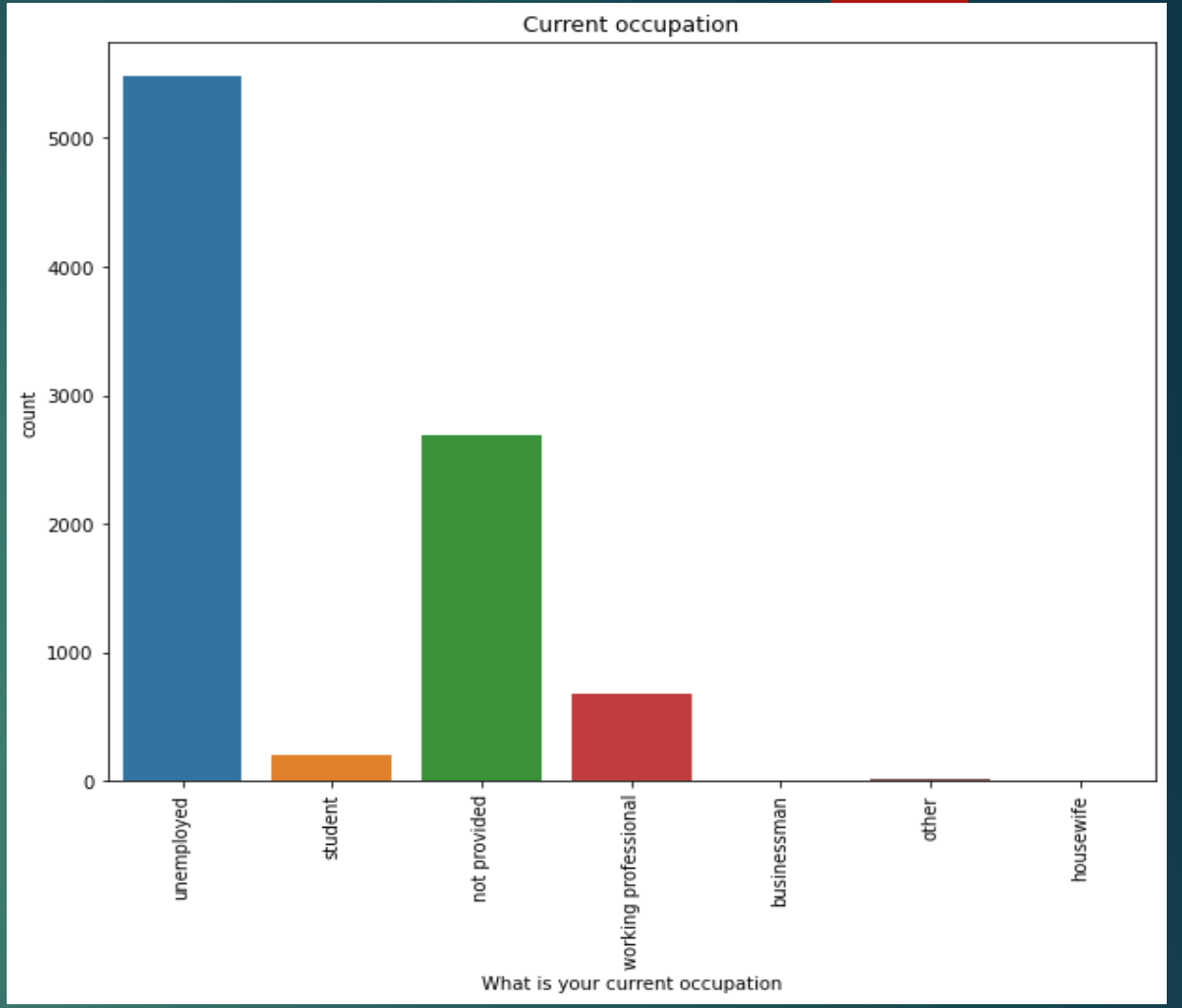
We can see from the above plot of Do Not Email that user has opted for no much higher than yes also we can see from the above plot of Do Not Call that user has opted for no than yes.



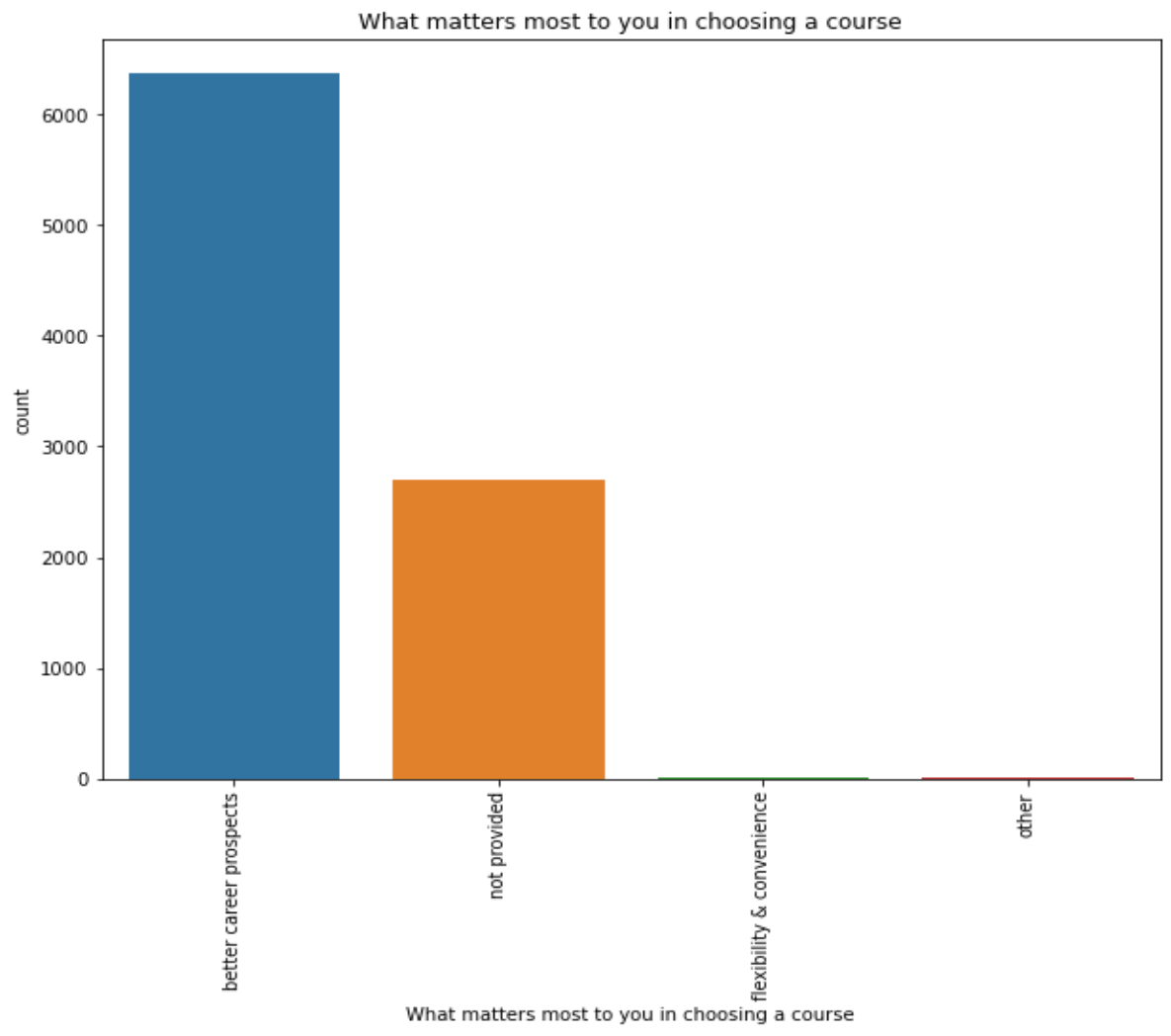


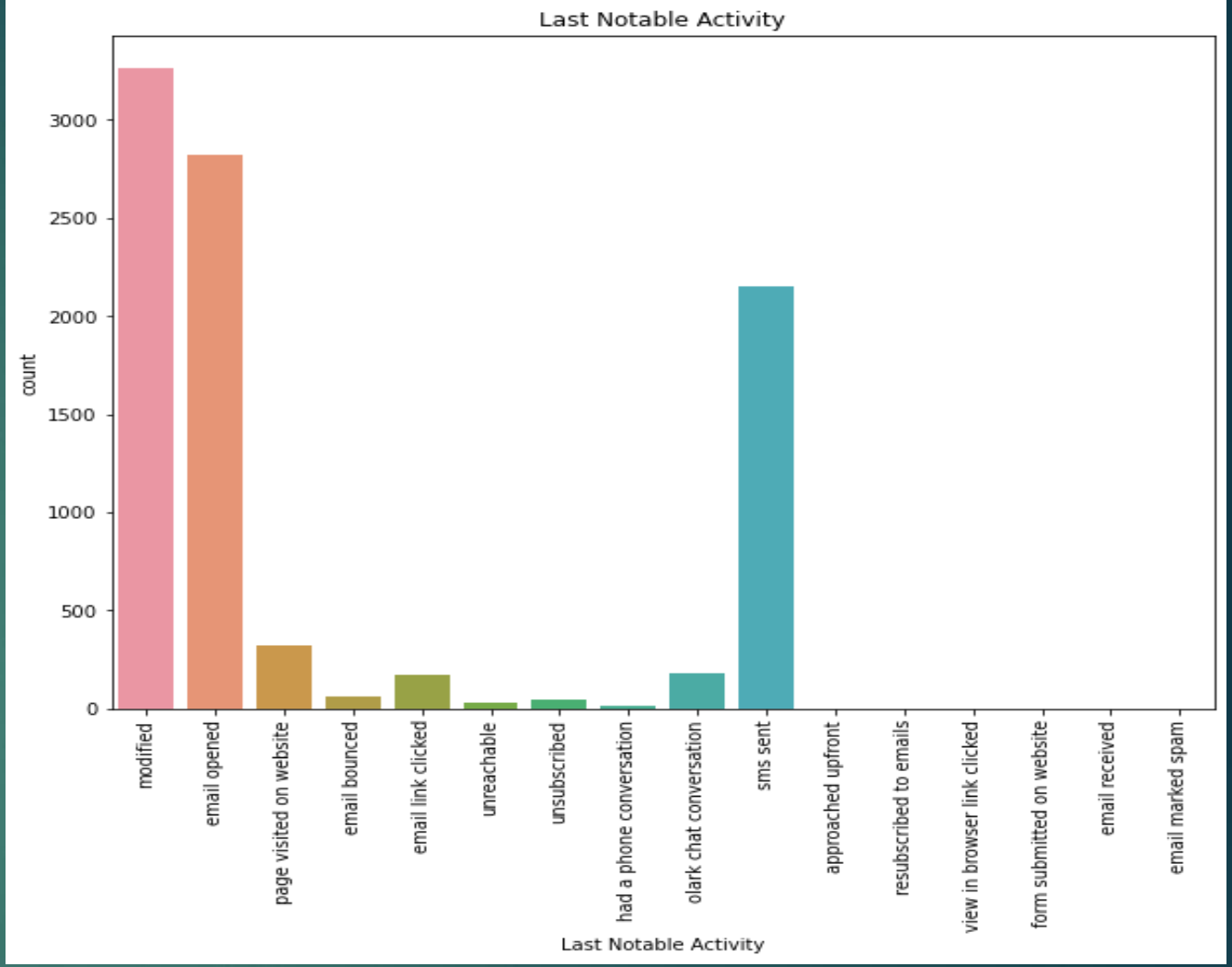
We can see from the above plot of Last Activity that the Last Activity of most of the people is opening of the email. We can see from the above plot of Country that most of the users are from india and there are many users who have not provided the country details.



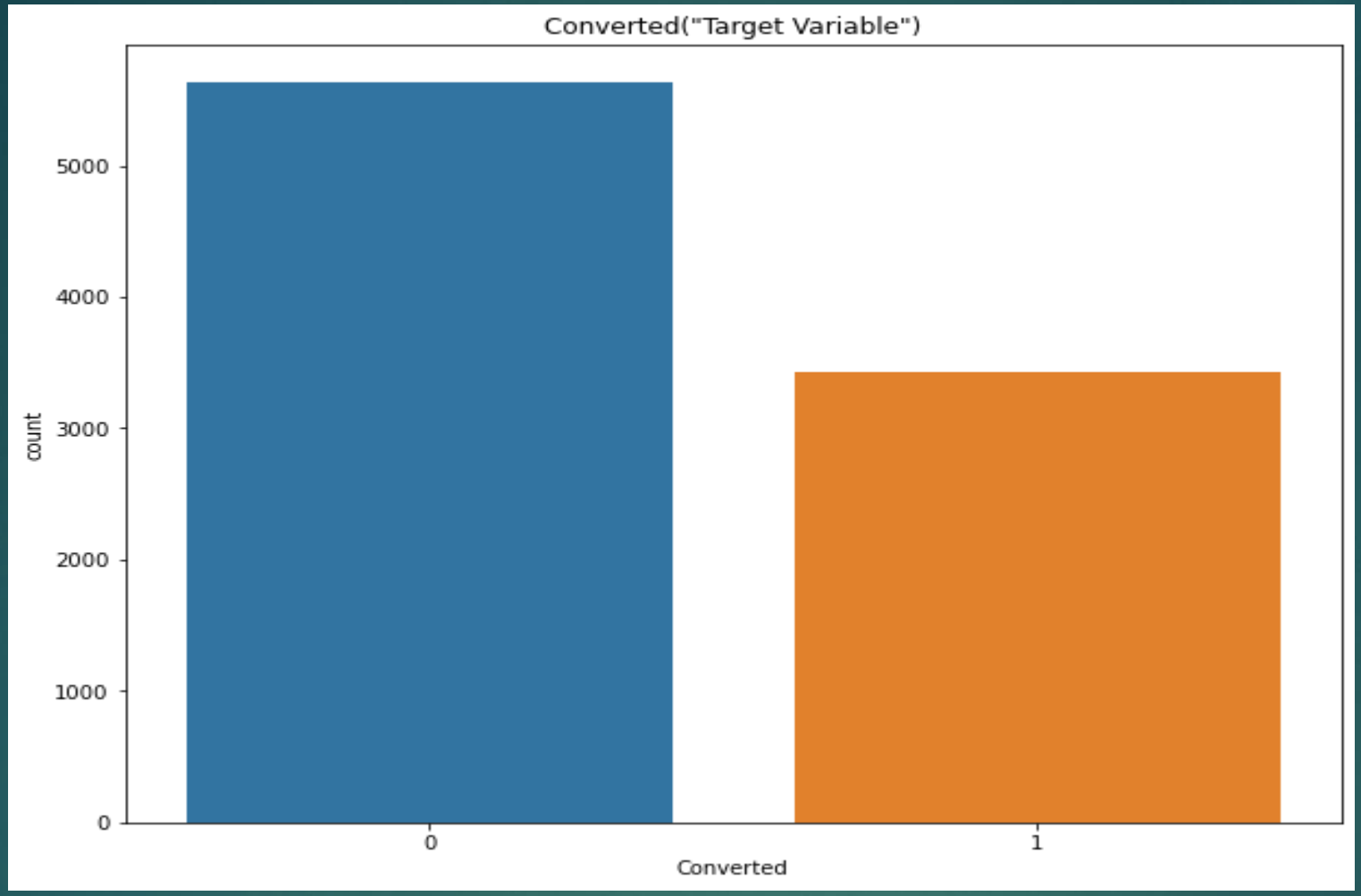


We can see from the above plot of Specializations that users have not provided the details of the specialization other than that finance management specialization is much higher than other specializations. We can see from the above plot of What is your current occupation that most of the users are unemployed and many of them have not provided the details of their current occupation.



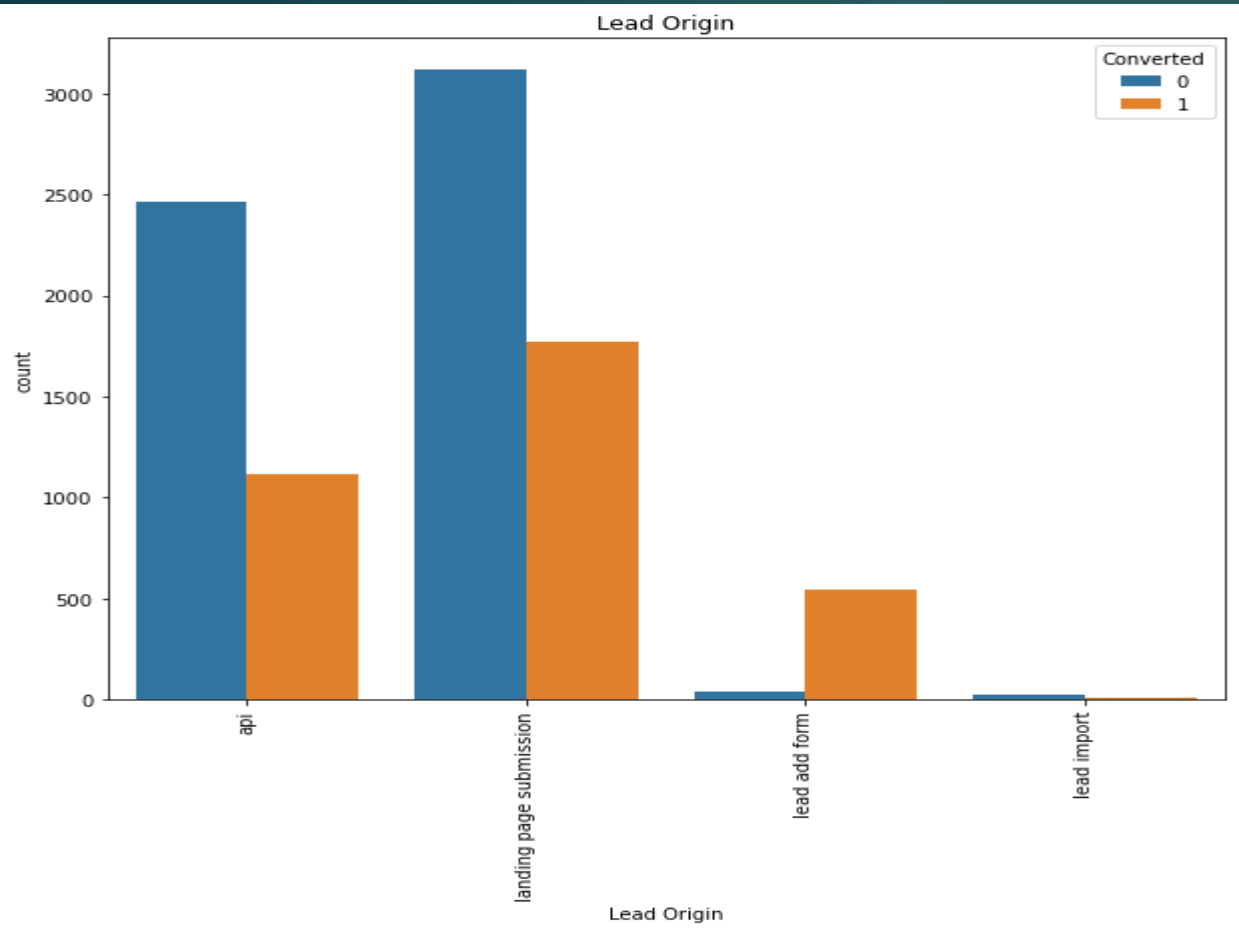


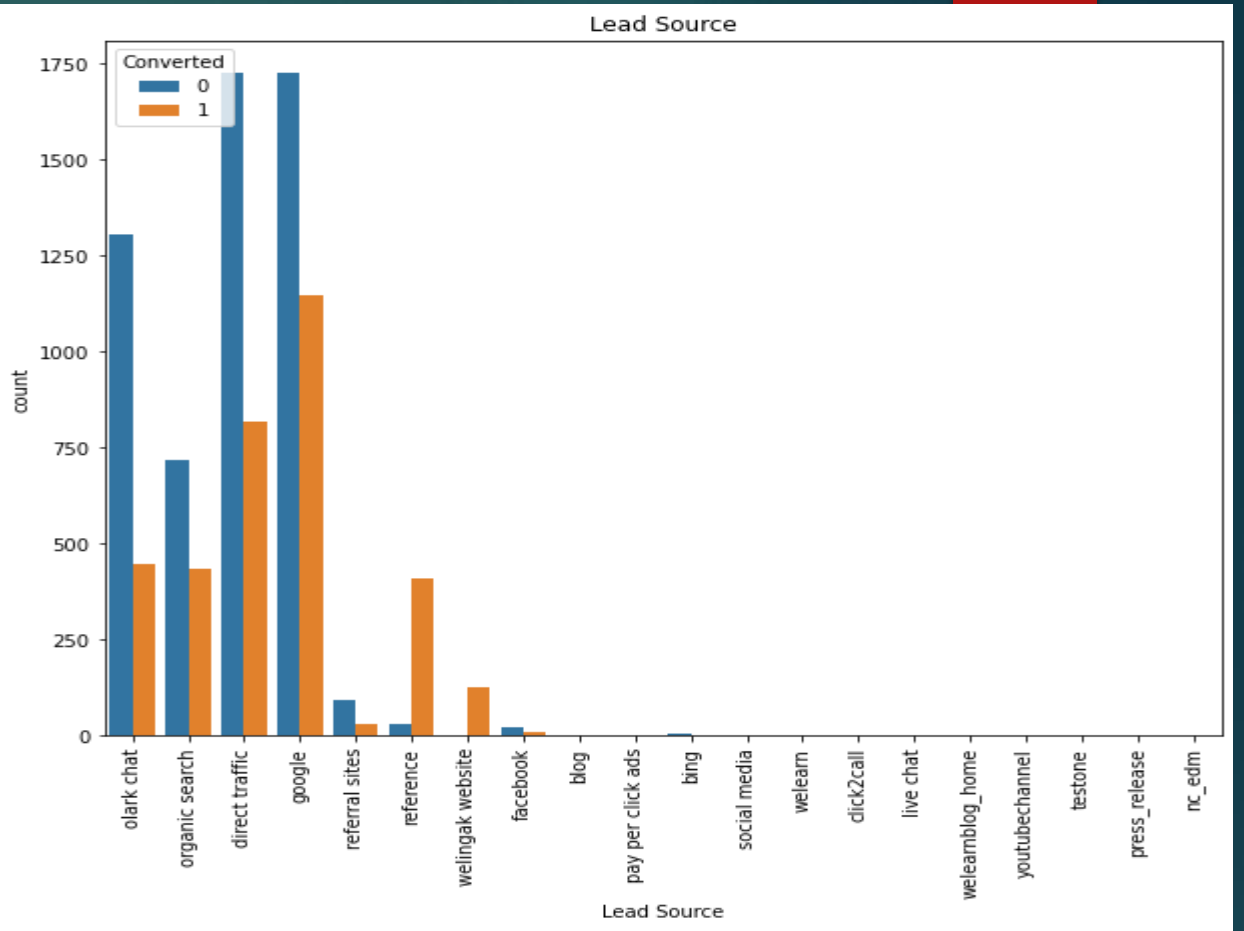
We can see from the above plot of What matters most to you in choosing a course that most of the users have opted the reason for choosing the course is better career prospects also we can see from the above plot of Last Notable Activity that modified activity is much higher than other activities.



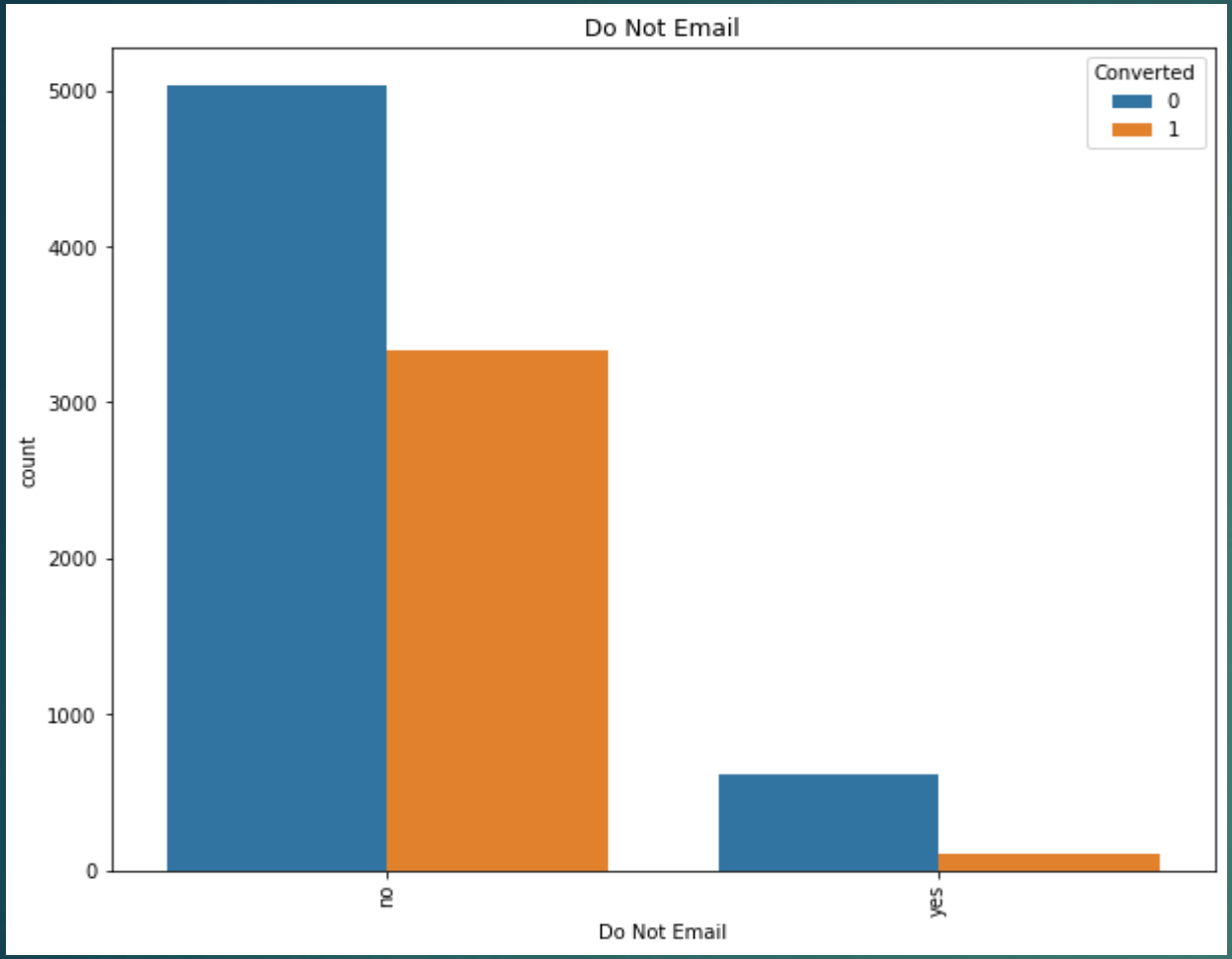
We can see from the above plot of the Target Variable Converted that leads which are not converted which is indicated by 0 are more than the leads converted which is indicated by 1.

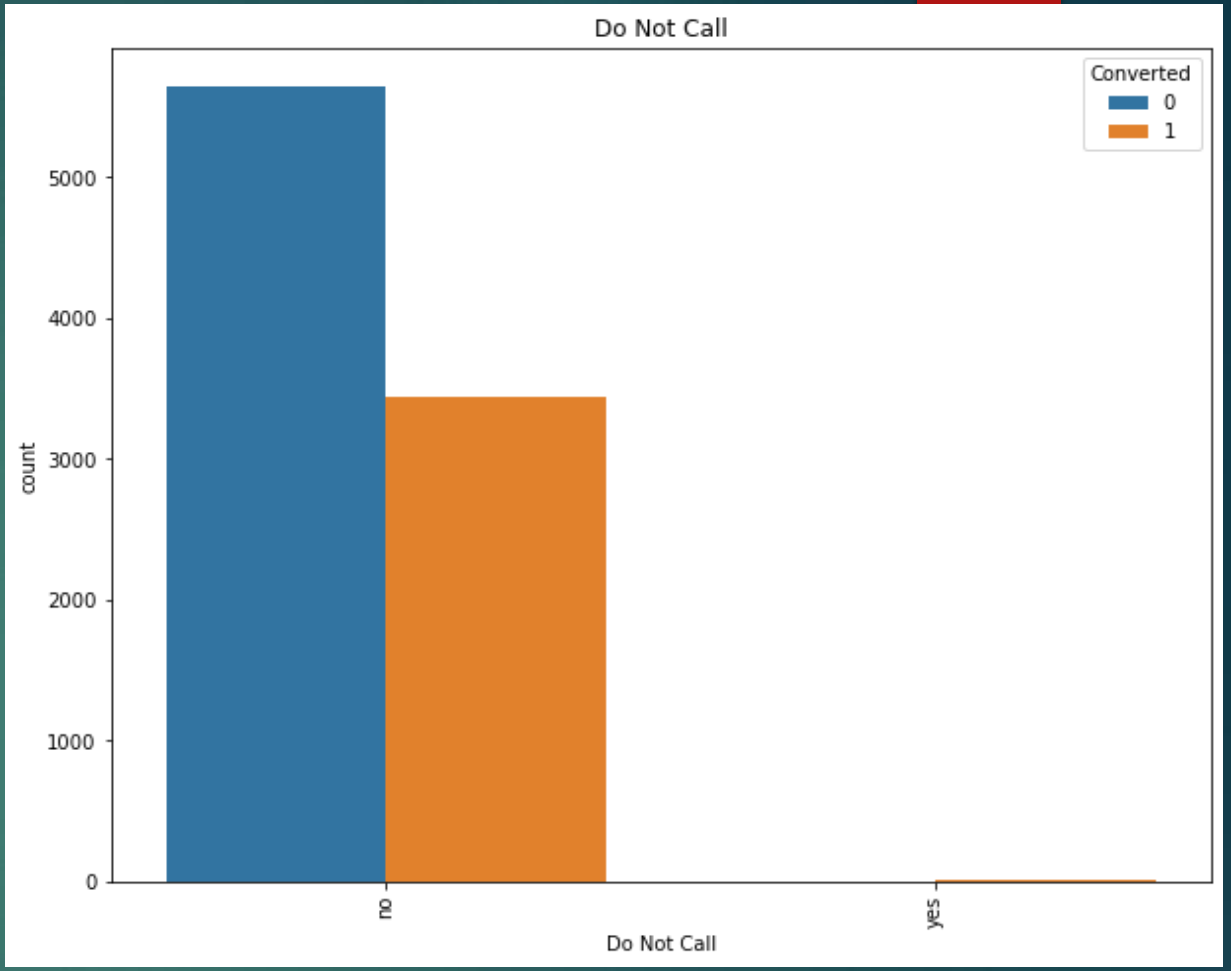
Bivariate Analysis of Categorical Variables with Target Variable



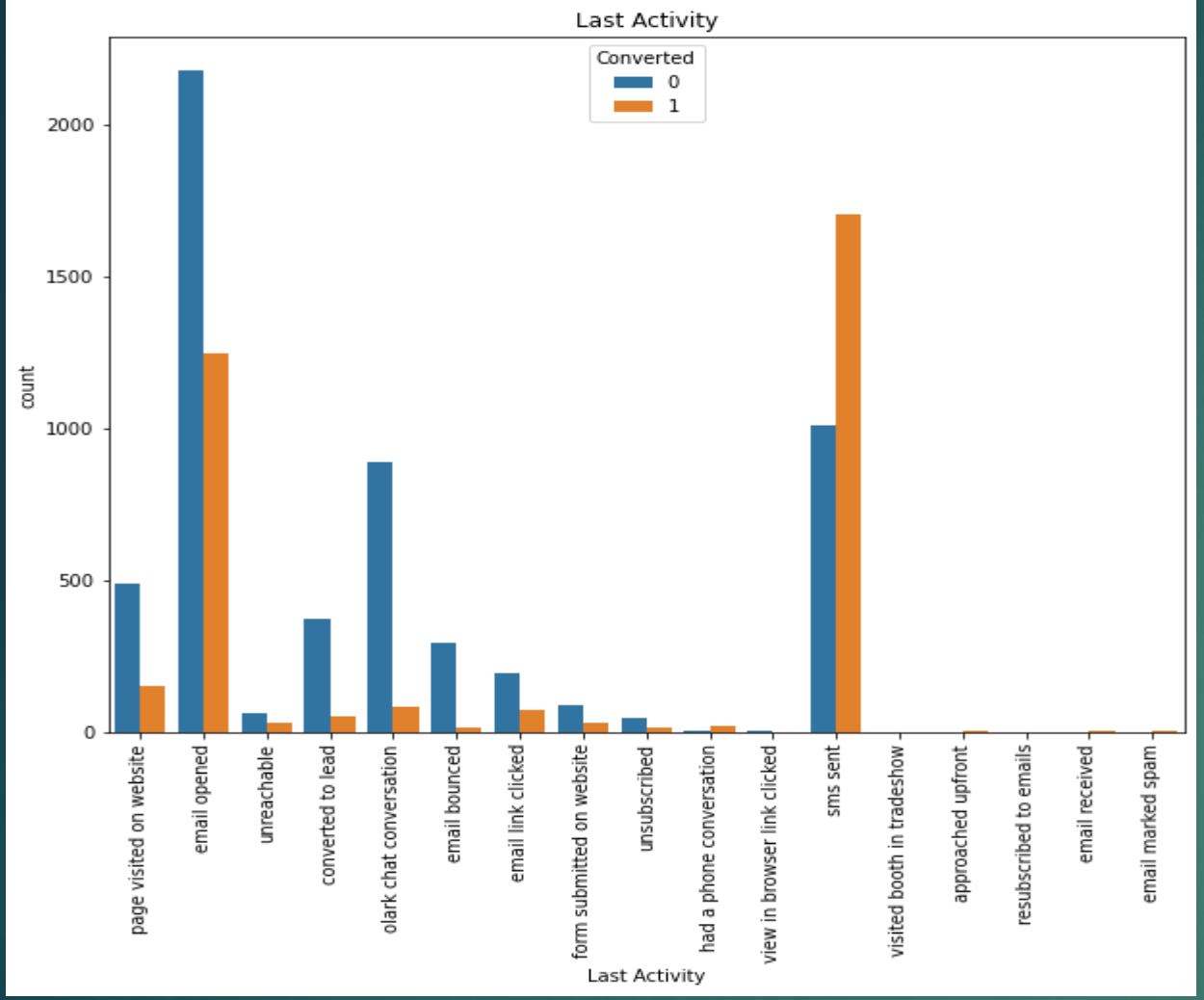


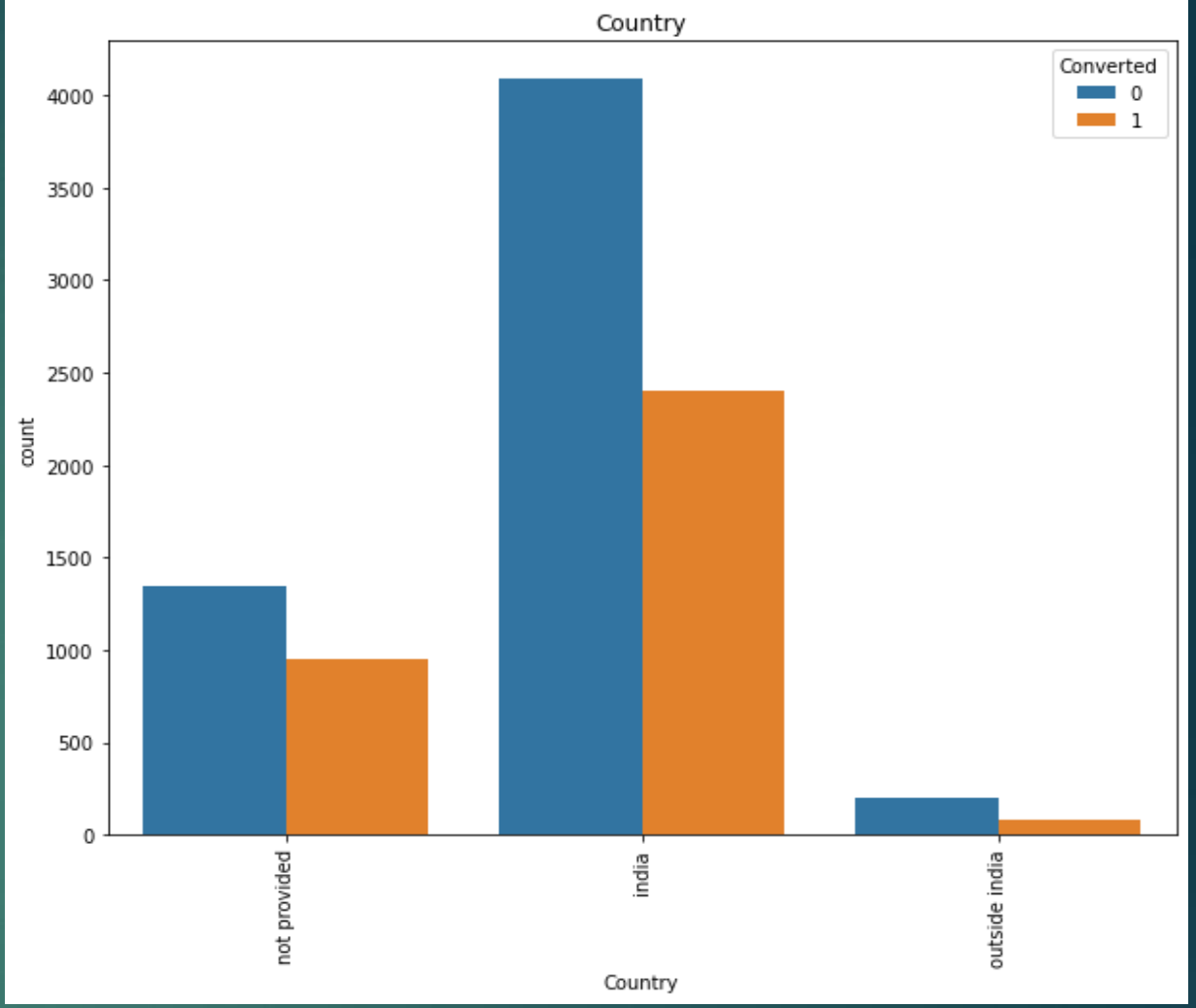
We can see from the above plot of Lead Origin that most of the leads come from landing page submission and most of the leads are converted from landing page submission and also from lead add form also we can see from the above plot of Lead Source that most of the leads are converted from the lead source which are google,references,direct traffic.



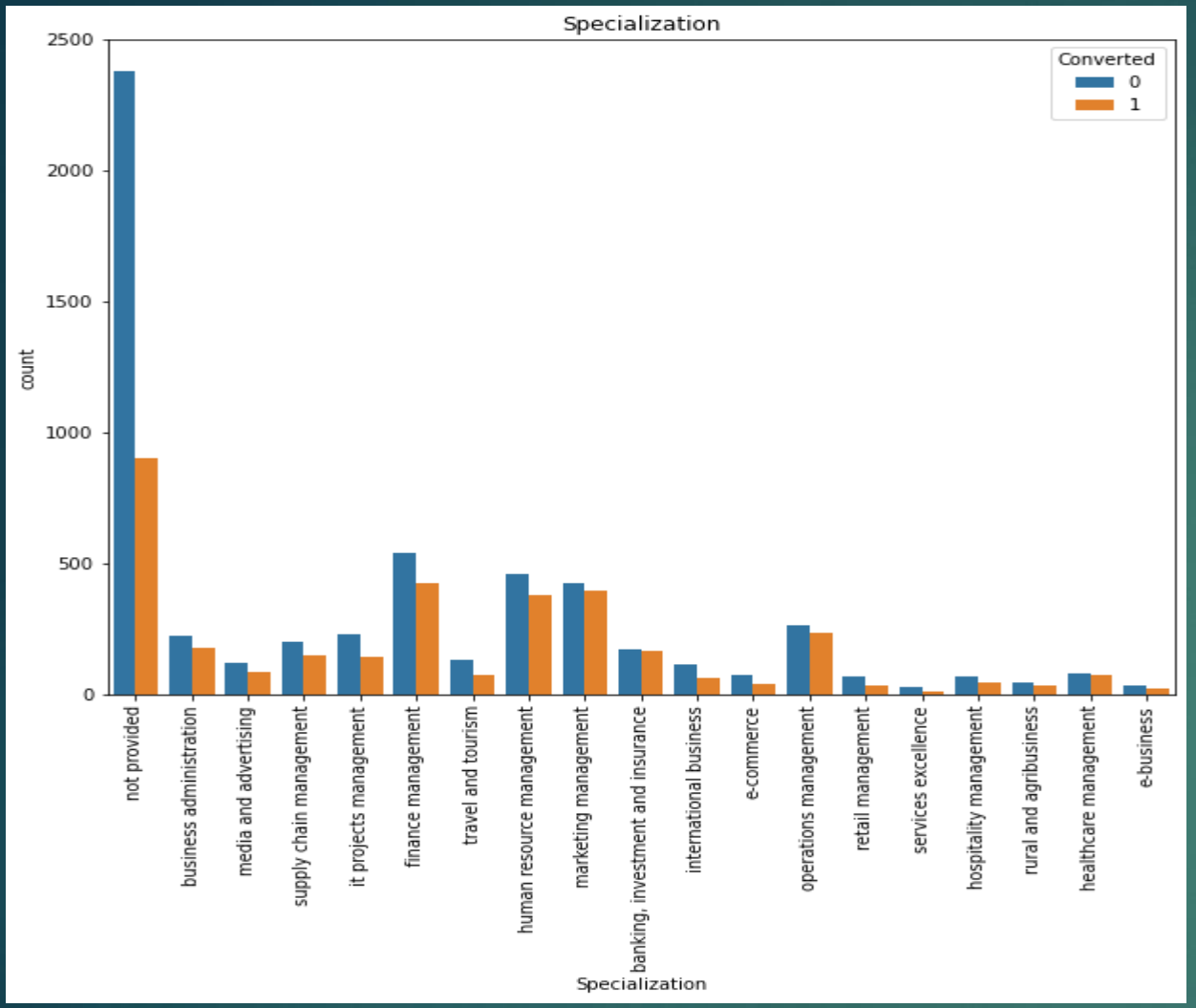


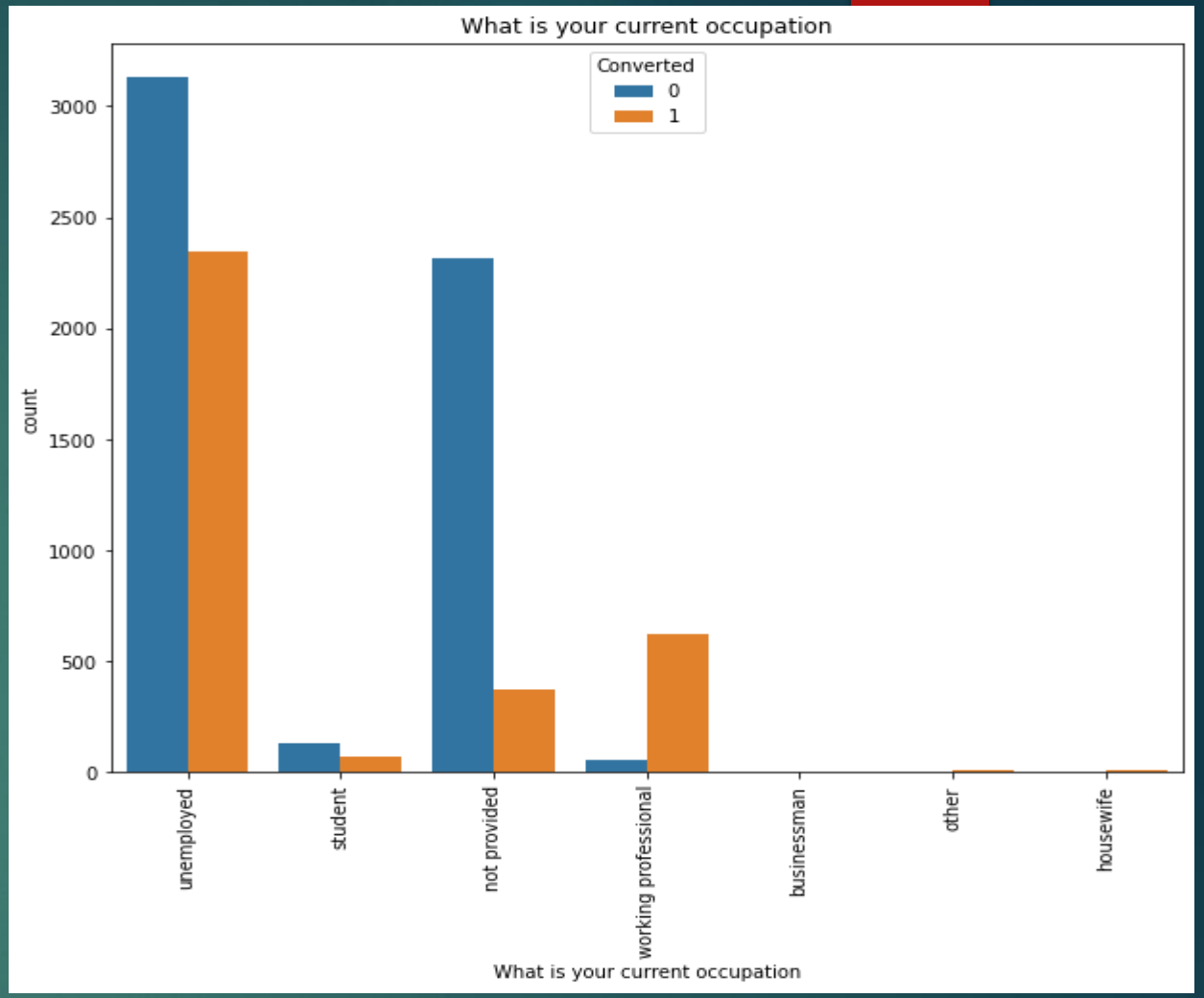
We can see from the above plot of Do Not Email leads converted when a customer selects not to email is more than the when the customer selects the option to email also we can see from the above plot of Do Not Call that the leads converted when the customer selects not to call is more than when the customer selects to call.



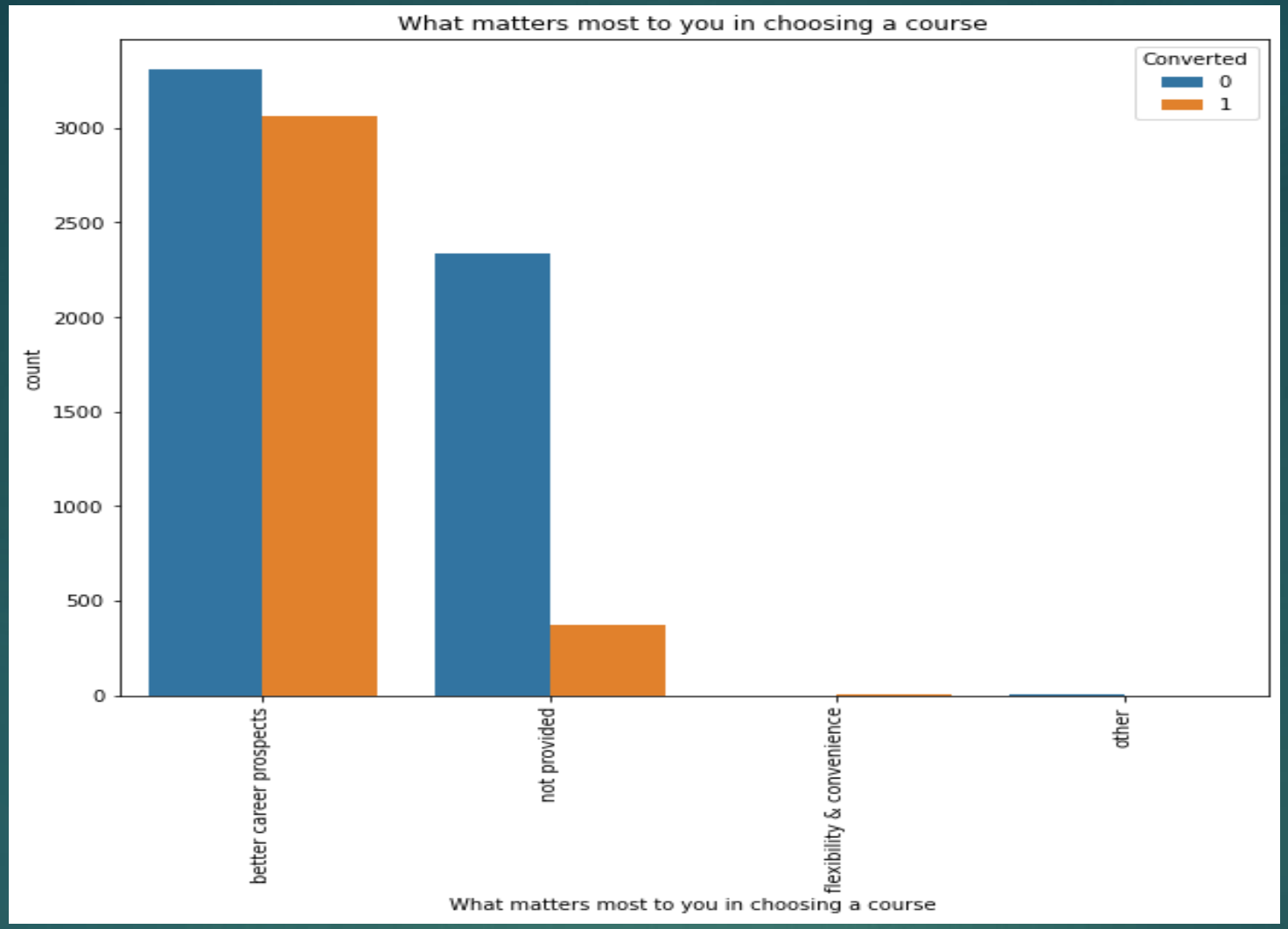


We can see that from the above plot of Last Activity most number of leads that were converted when the user's last activity is sms sent also we can see from the above plot of Country that most of the leads that were converted were from country India.





We can see from the above plot of Specialization that leads are from people who have not provided their specialization but ratio of leads conversion is best in banking,investment and insurance also we can see from the above plot of What is your current occupation that leads conveted are more in people who are working professionals.



We can see from the above plot of What matters most to you in choosing a course that leads are converted more when people choose better career prospects in What matters most to you in choosing a course.

Data Preparation

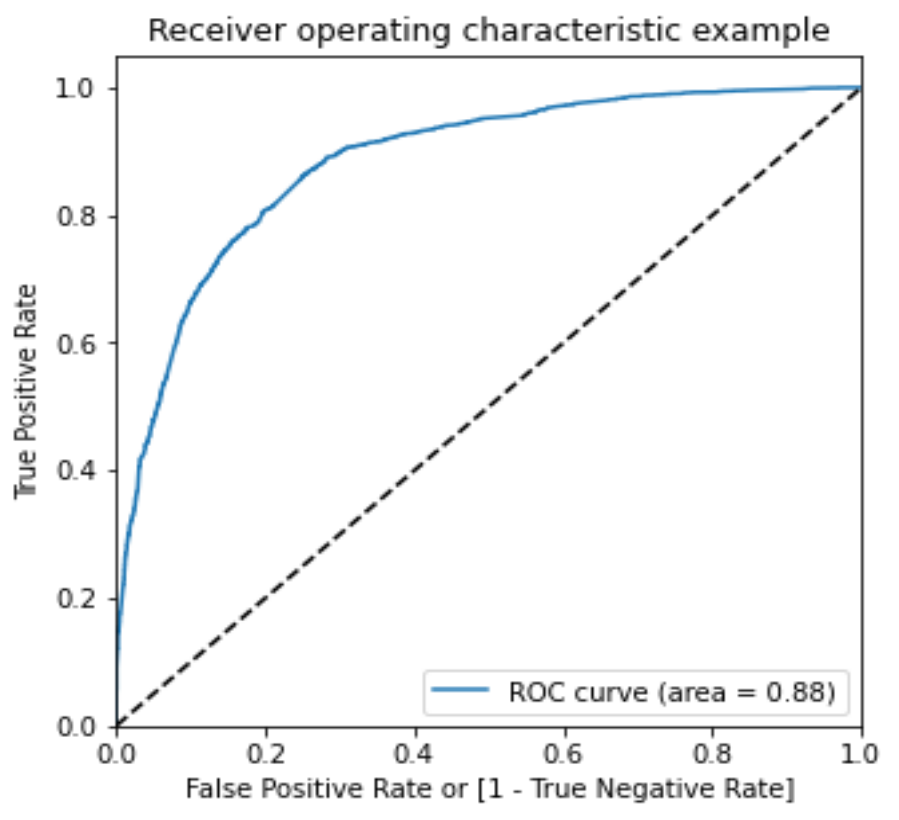
Data is Prepared by the following steps

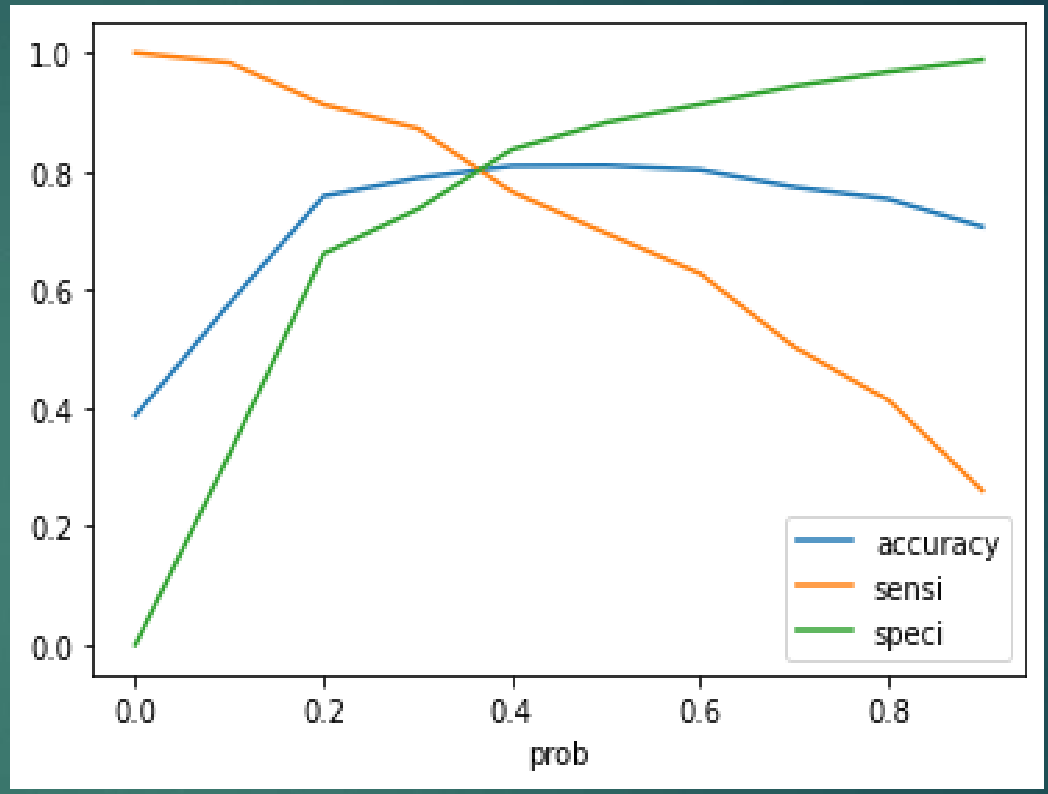
- The Numerical Variables are Normalised
- Dummy Variables are created for object type variables
- Total Rows for Analysis: 9074
- Total Columns for Analysis: 81

Model Building

- The Data Split into Training and Testing Sets
- The first basic step for logistic regression model building is performing a train-test split, we have chosen 70:30 ratio for test-train sets respectively.
- Used 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
- Making Predictions on test data set
- Overall accuracy 81%

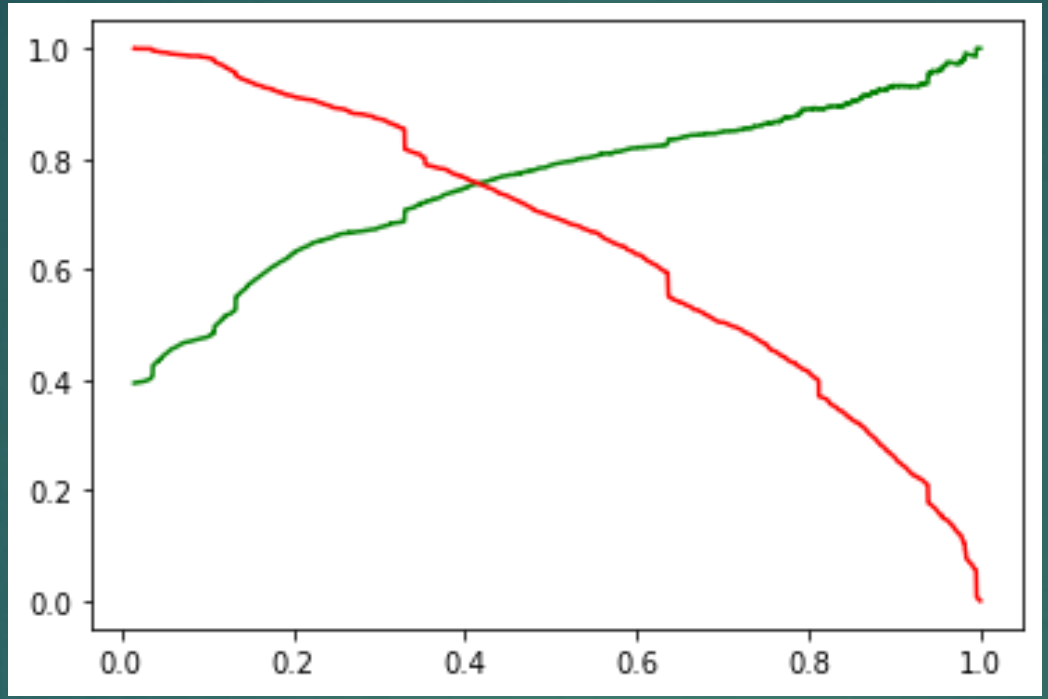
ROC Curve





By the use of ROC Curve we find optimal cut-off probability which is probability where we get balanced sensitivity and specificity and from the second plot we can see that the optimal cut-off is 0.35.

Precision Recall Trade Off



By the use of Precision Recall Trade off plot above we found that the intersection point is 0.41 and by this we found that point we found that the Precision value is 75% and Recall value is around 76%.

Conclusion

It was found that the variables that mattered the most in the potential buyers are (In descending order) :

- 1.What is your current occupation_unemployed
- 2. The total time spend on the Website.
- 3.Total Visits.
- 4.Lead Origin Add form
- 5.When the last activity was:
- a. SMS
- b.olark chat conversation
- 6.When the lead source was:
- a.Olark_chat
- b.welingak website
- 7.What is your current occupation_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.