Lead Score Case Study Assignment

GROUP MEMBERS :

PRACHI FAYE

MADHURA PATHAK

Problem Statement

- X education sales online courses to industry professionals.
- They get lot of leads but very few gets converted (say 30 %).
- Company wishes to identify hot leads.
- The ballpark of the target lead conversion rate given by CEO is around 80%.

Business Objective

- To build a model to find most potential leads, also known as ‘Hot Leads.
- Build a model to identify hot leads.
- Deployment of model for future use.

Solution Methodology

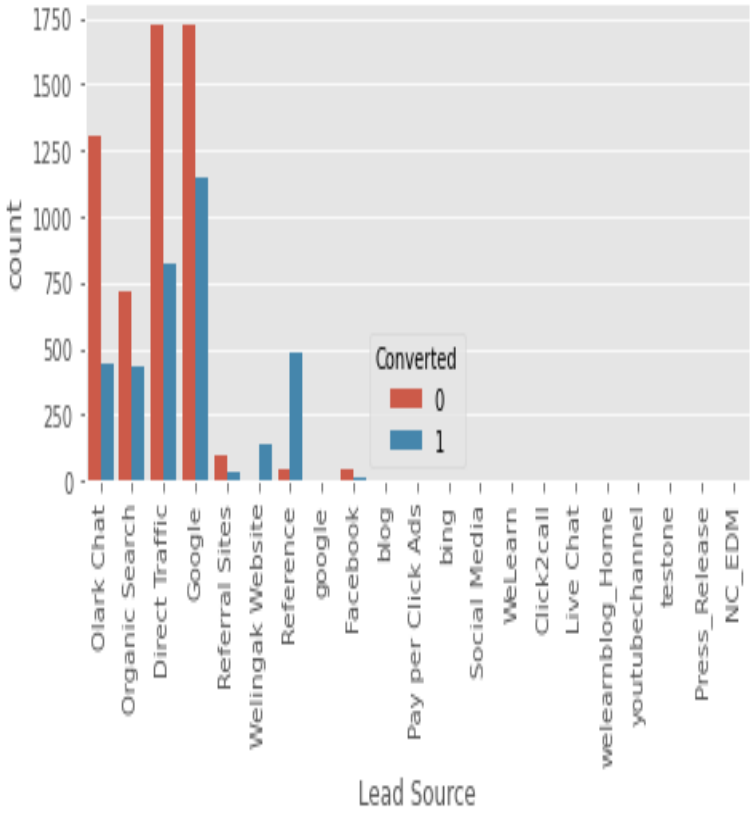
- Understanding the Data & Data Cleaning
- Check for missing values and duplicate data and treat it.
- Drop columns that have a high no of missing values and are not useful for analysis.
- Input values in other missing values.
- Check and handle outliers.
- Data manipulation
- Modify the ‘Dtype’ if necessary. “BIN’ the data for better visualization, group the data for ease of visualization

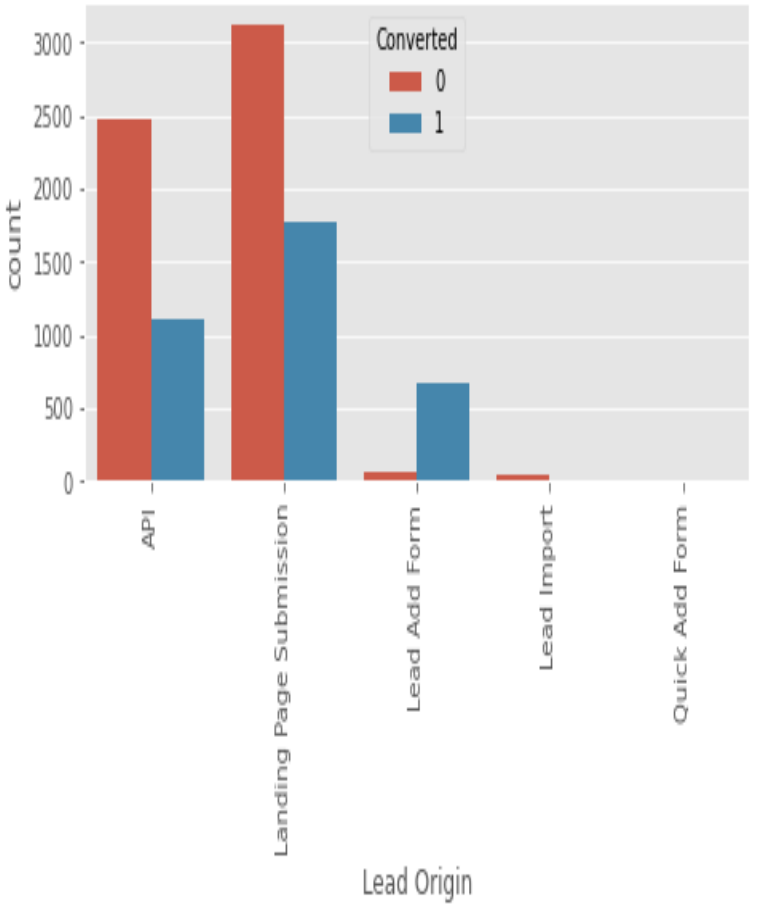
Solution Methodology

- Exploratory Data Analysis
- Univariate Analysis - Value count, distribution of the variable.
- Bivariate Analysis - Correlation Coefficient and pattern between variables
- Creating Dummy variables of categorical variables to facilitate model building.
- Features scaling (if required)
- Model making using Logistic regression
- Model Validation
- Model presentation
- Conclusion
- Recommendations

Data Cleaning and Manipulation

- There are total 9240 rows and 37 columns.
- Columns having high missing values and not relevant for analysis removed e.g. Lead Quality.
- Checked if any Lead is recorded twice. There were no Duplicates.
- Converted variable labels from Yes & No to 1& 0.
- Variable having unique value e.g Magazine dropped as they are not useful for analysis.
- Categorical variables: Missing values imputed with Mode
- Numerical Variables : Outliers were checked there were not considerable amount of outliers to treat.
- Checked if the data is imbalanced or not.



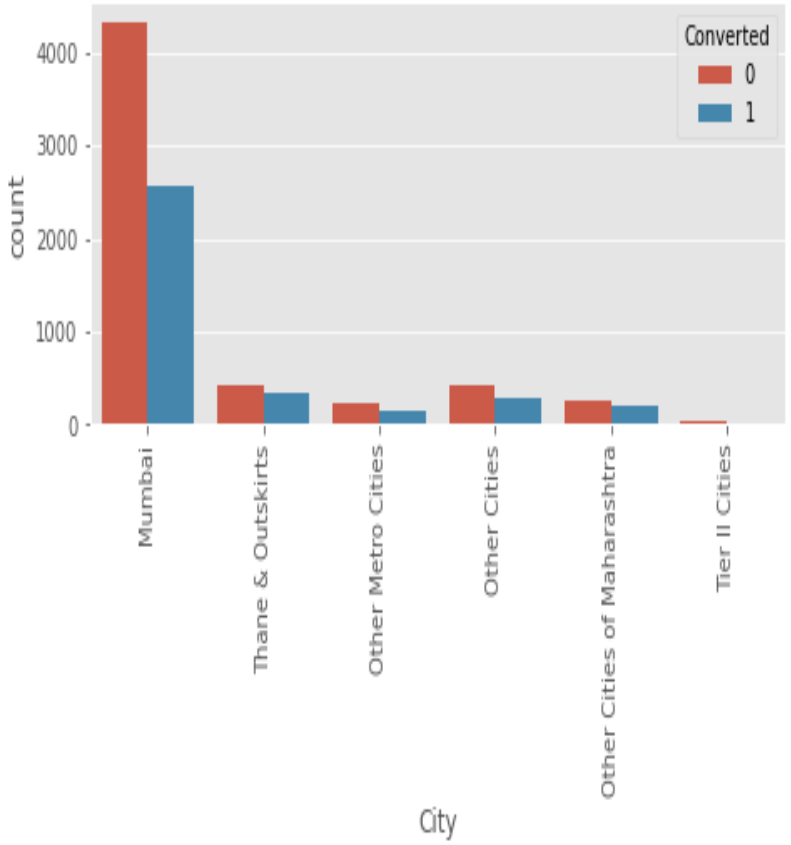


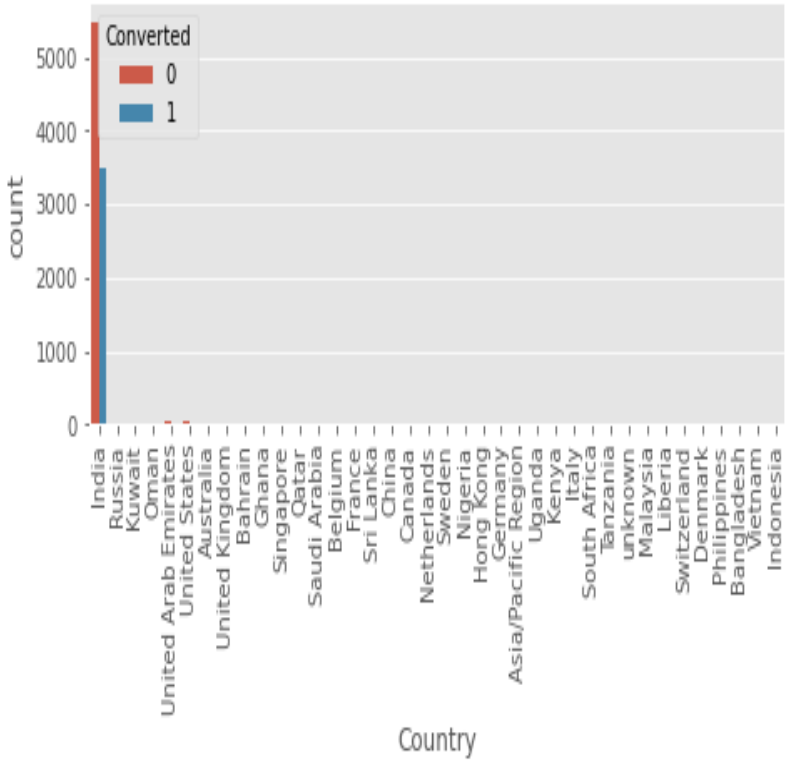
Bivariate Analysis: Categorical

Lead origin: API and Landing Page Submission have less conversion rate but counts of the leads from them are high in numbers so let's not ignore them.

Lead Source: Google is appearing twice in 'Lead Source'.

Maximum leads are from Google followed by Direct Traffic.





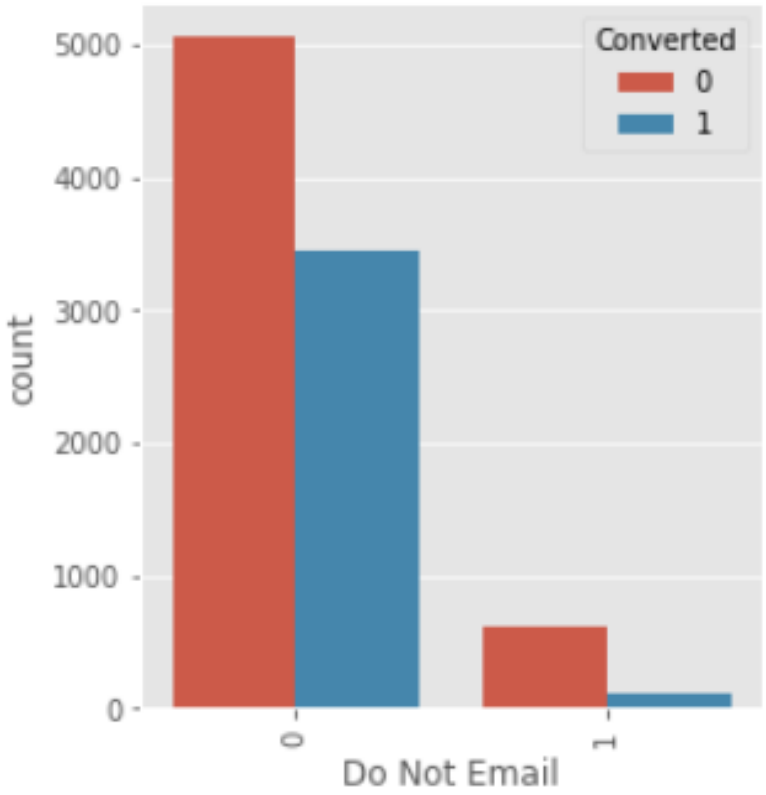
Exploratory Data Analysis

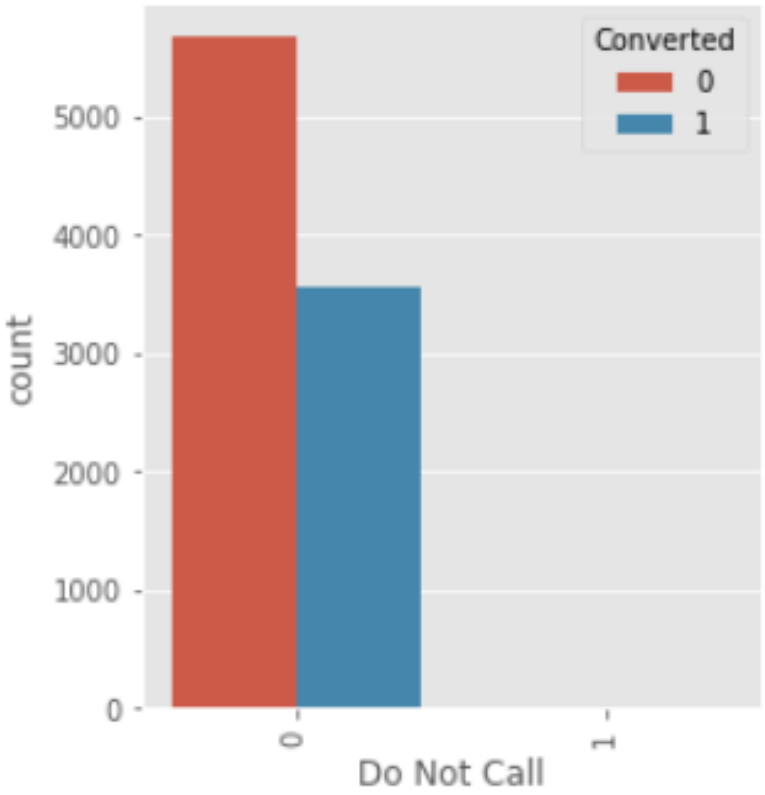
City: Mostly the reach out has been made to people from Mumbai and they have very high conversion ratio.

Tier II cities shows very less ration as compared to other cities.

Country: Mostly reach out have been made to people from Country India and few to UAE and United states.

In all, converted leads are more than total leads.





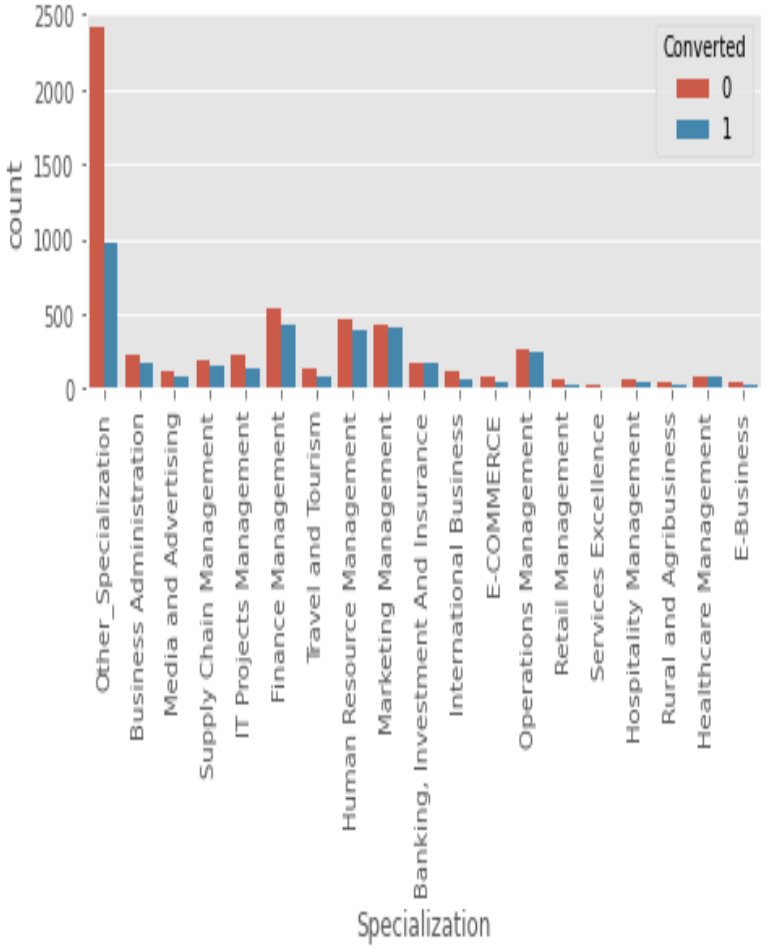
Exploratory Data Analysis

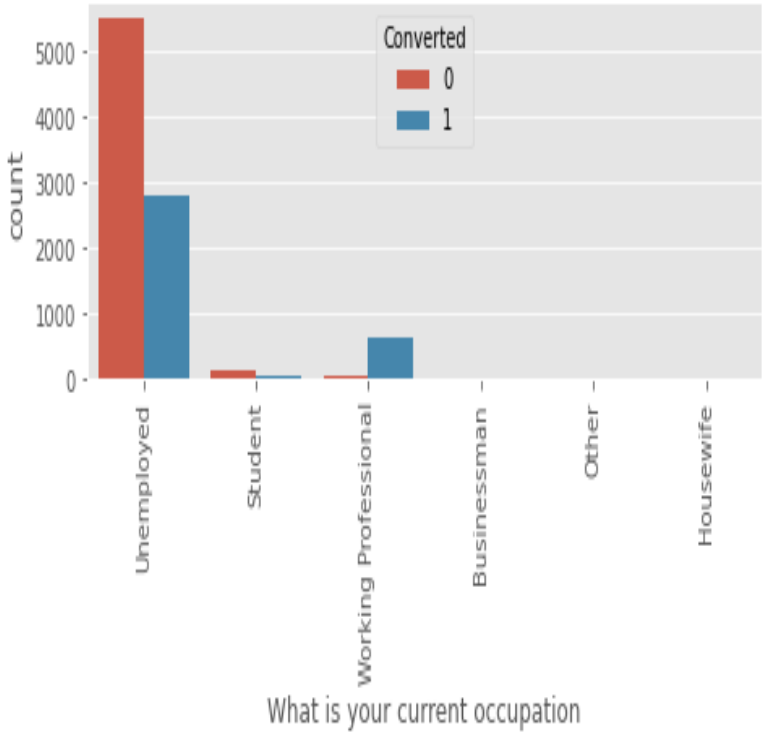
For Do Not Email

We see people advising to not to email have very low conversion ratio. It shows that mostly people are reached out by email and the conversion ratio is high.

For Do Not Call

We can see from the graph that mostly people are getting contacted by phone and they have high conversion ratio too. Mostly people prefer to get calls.





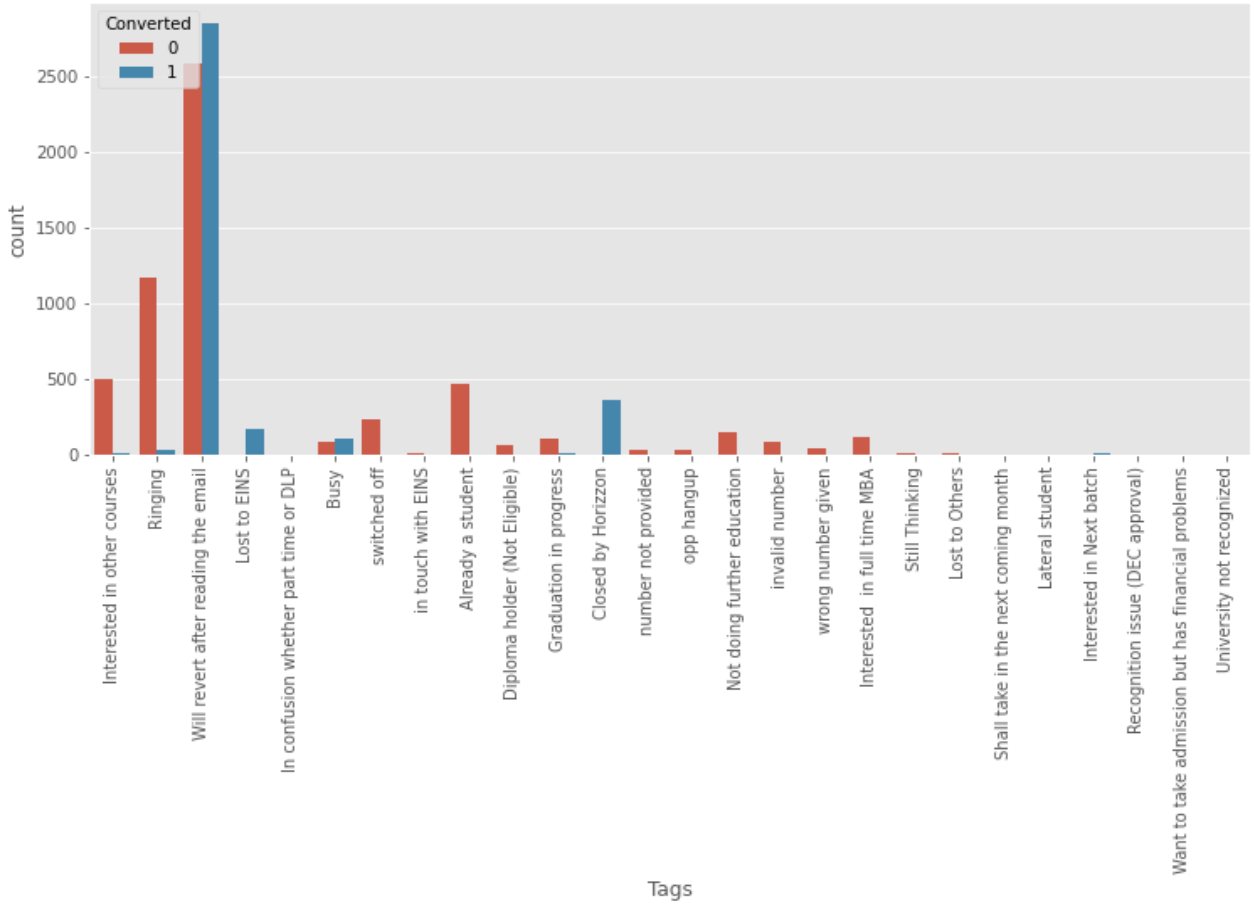
Exploratory Data Analysis

Specialization

We combined Management Specializations because they show similar trends.

What is your current occupation

Highest conversion is of working professionals.

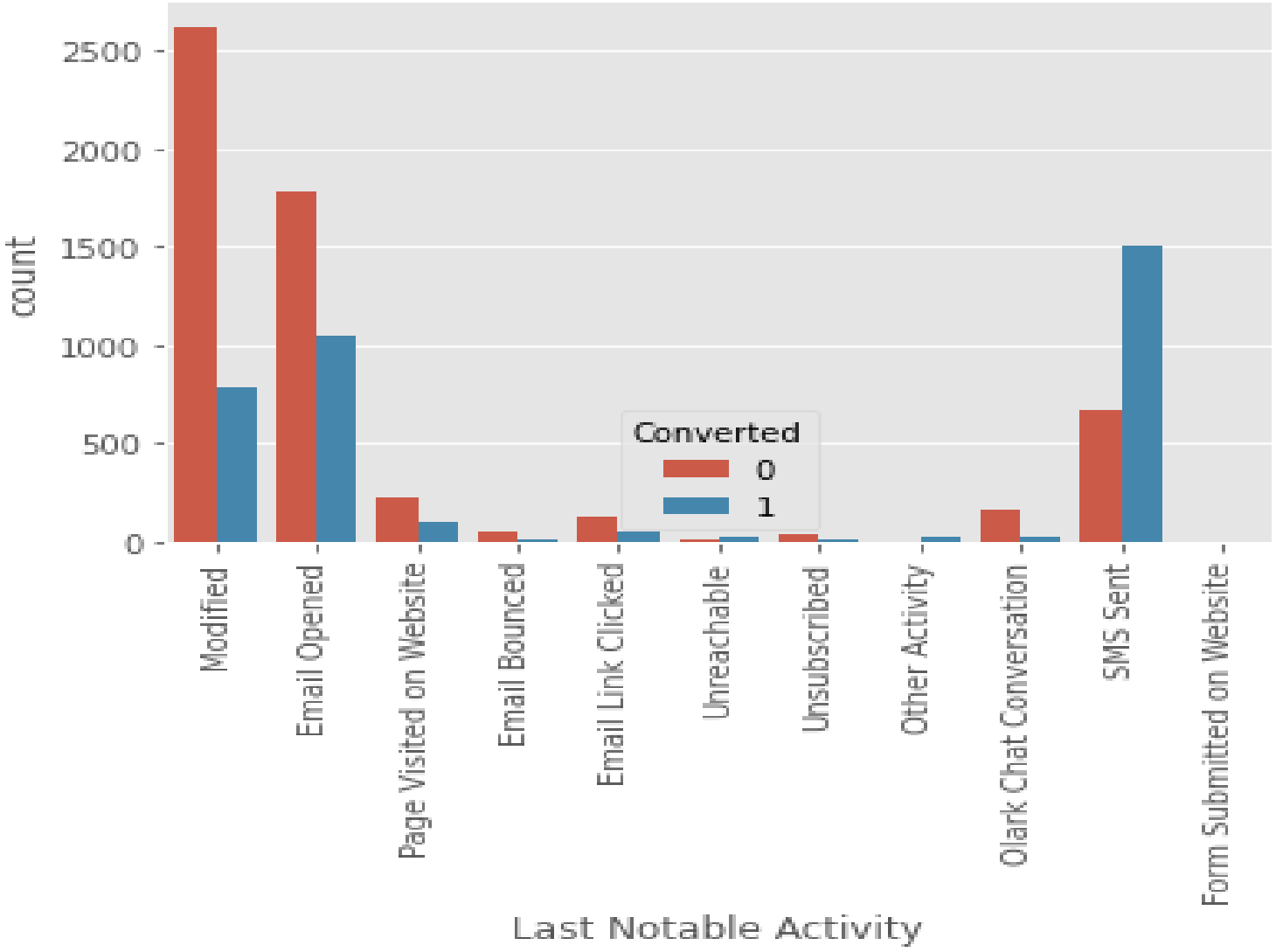


Exploratory Data Analysis

Tags

Will revert after reading the email' and 'Closed by Horizon' have a high conversion rate.

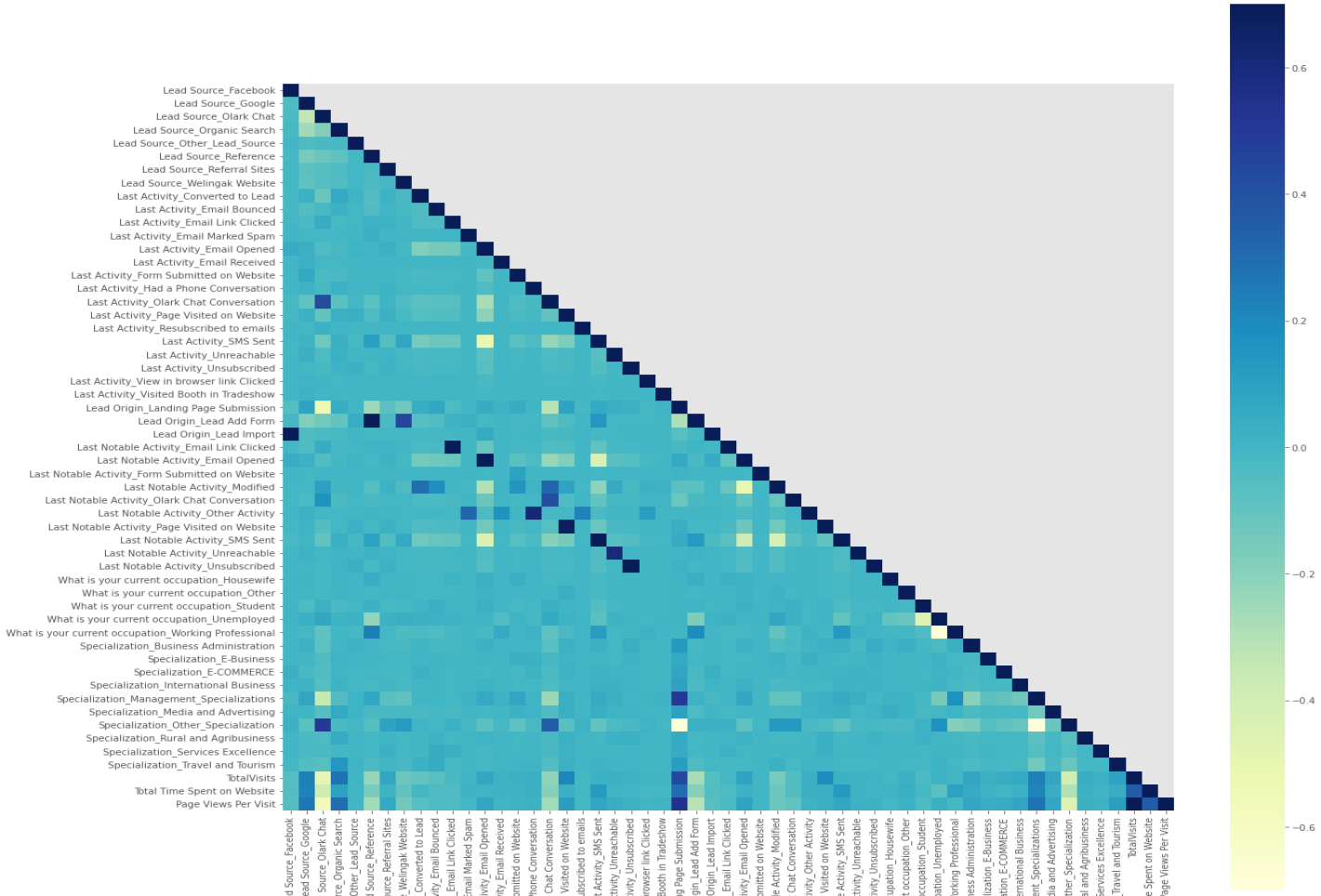
'Lost to EINS’ & ‘closed by Horizon’ have a very good conversion rate.



Exploratory Data Analysis

Last Notable Activity

We should focus on increasing the conversion rate of those having last activity as Email Opened, Modified to those leads, and also try to increase the count of the ones having last activity as SMS sent.



Checking Correlation

There are group of columns which are positively correlated with each other:

1. Lead source Reference
2. Lead_Origin_Lead_Add_Form
3. Lead Source Facebook
4. Total Visits

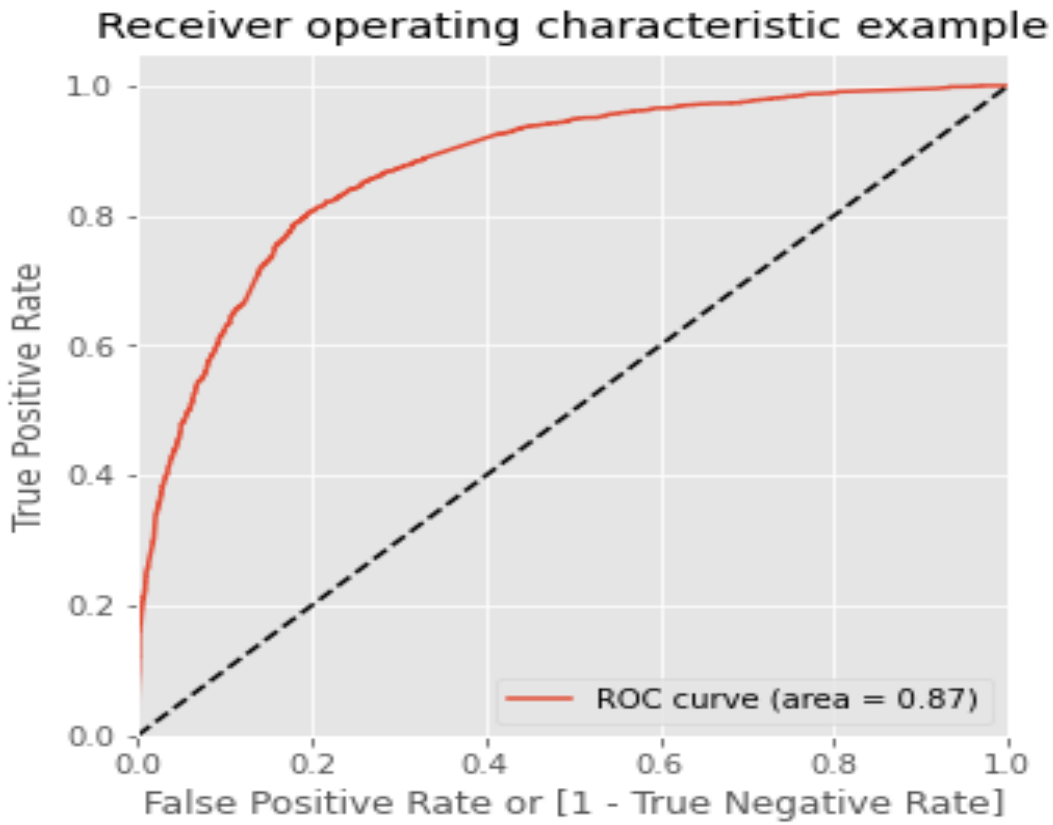
Data conversion

- Numerical Variable are normalized.
- Dummy variables are created for object type variables
- Total Rows for Analysis : 8892
- Total Columns for Analysis : 56

Model Building

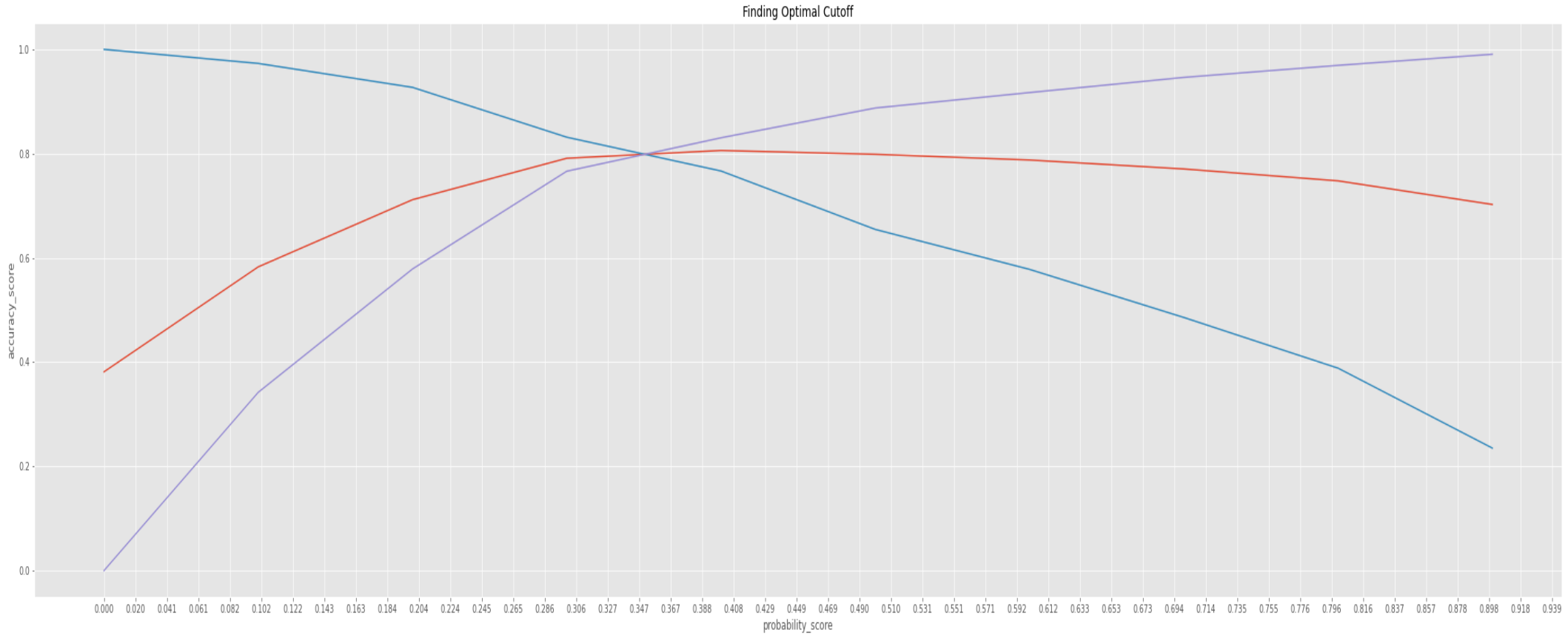
- Splitting the Data into Train and Test Sets.
- Train Test split here we have chosen 70:30 ratio.
- Used RFE for Feature selection
- Running RFE with 15 Variables as output.
- Building model by removing variable whose P value is greater than 0.05 and VIF is greater than 5.
- Prediction on Test Data.
- Overall Accuracy around 80%

The ROC Curve

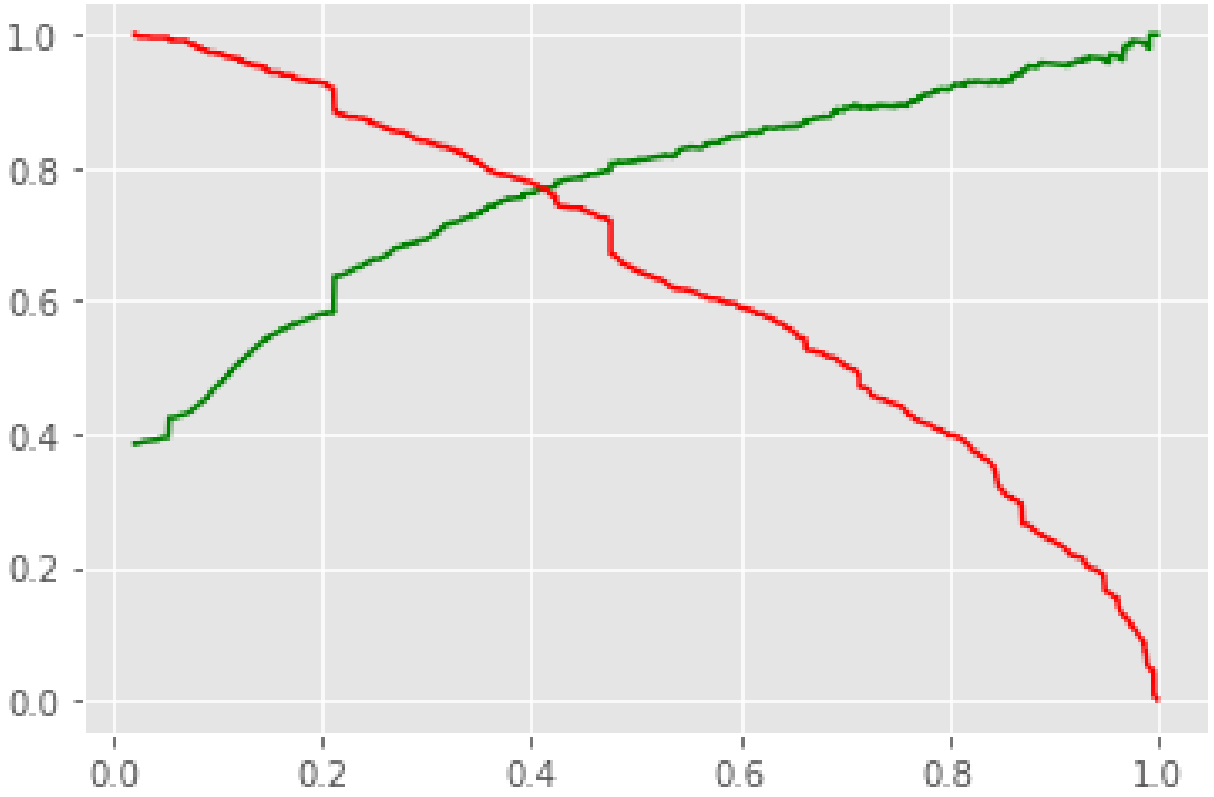


- The ROC curve for our model has a value 0.87 which is good

ROC Curve Optimal Cutoff Point is 0.34



Precision and Recall Tradeoff



- Ideal cutoff of 0.43 is observed from recall and precision plot.

Conclusion

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

- Total Time Spent on website
- Lead Origin Lead Add form
- Lead Source_Wlingak Website
- Lead Source_Olark Chat
- Leas Activity_Email Bounced
- Last Activity_Had a Phone Conversation
- TotalVisits
- Specialization_Other_Specialization

Overall the model looks good on all the parameters.

Recommendations

- By calculating the lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.
- When their current occupation is as a working professional.
- 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.
- This model will help to identify the hot leads which would enhance speed and response rate.
- So by successfully identified only hot 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.