LEAD SCORING CASE STUDY

Submitted By:

- Akash Verma
- Dhopte Shivkumar Vishwambhar

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 develop a model to select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a 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. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Goals for this case study:

- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.
- There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

STEPS FOLLOWED

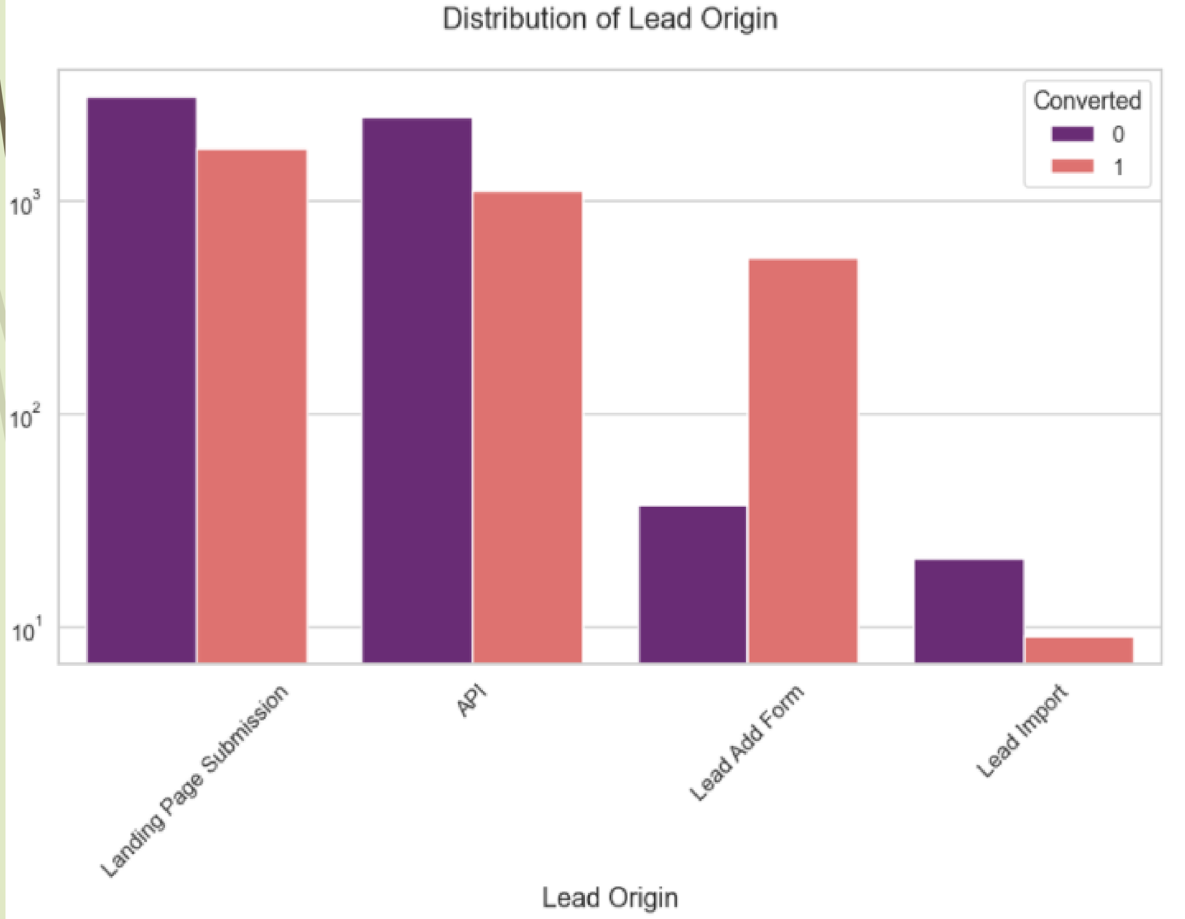
- Reading and Understanding Data.
- Data cleaning and data manipulation.
  - Check and handle duplicate data.
  - Check and handle NA values and missing values.
  - Drop columns, if it contains large amount of missing values and not useful for the analysis.
  - Imputation of the values, if necessary.
  - 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.
- Creating Dummy Variables
- Test Train Split
- Feature Rescaling
- Feature selection using RFE
- Plotting the ROC Curve
- Finding the Optimal Cutoff Point
- Computing the Precision and Recall metrics
- Conclusions and recommendations.

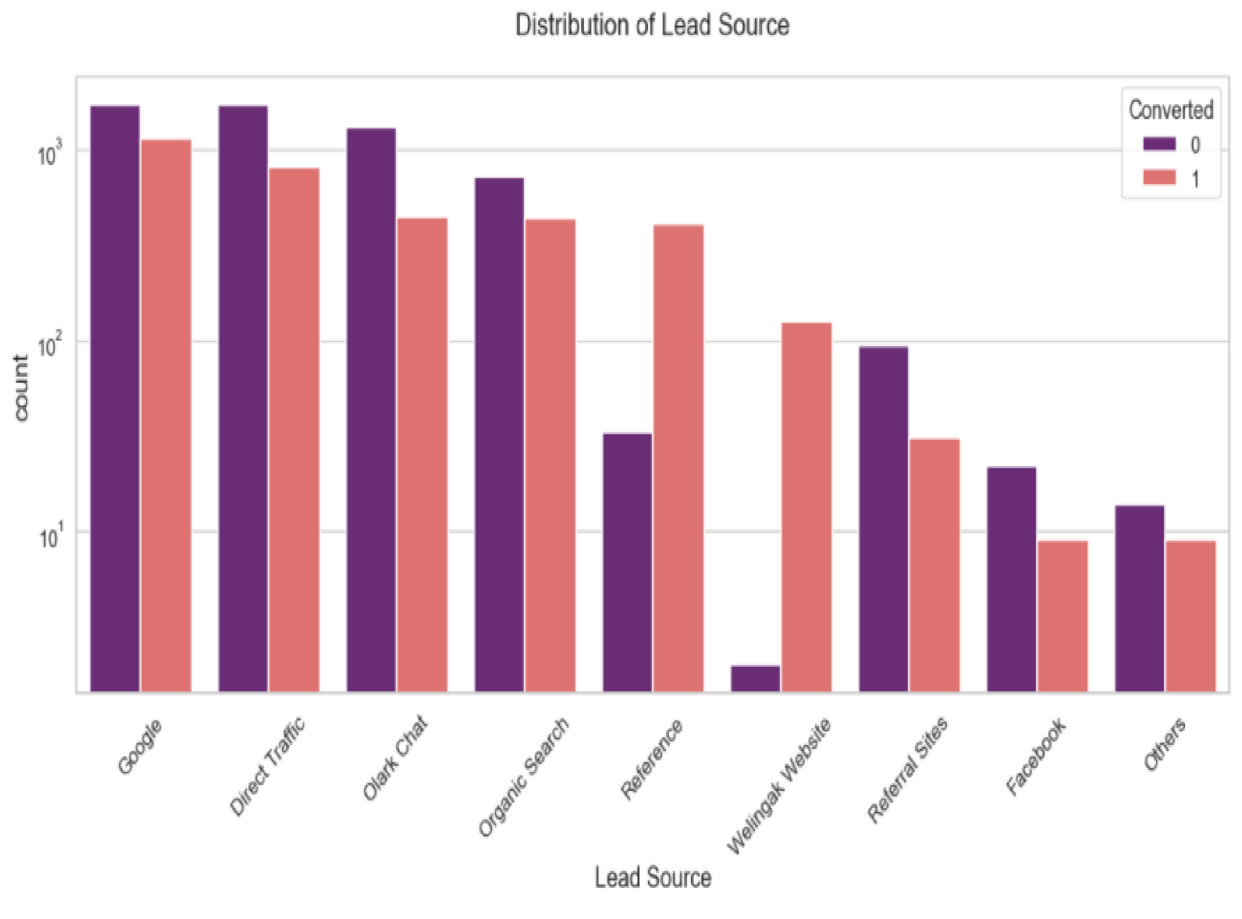
DATA MANIPULATION

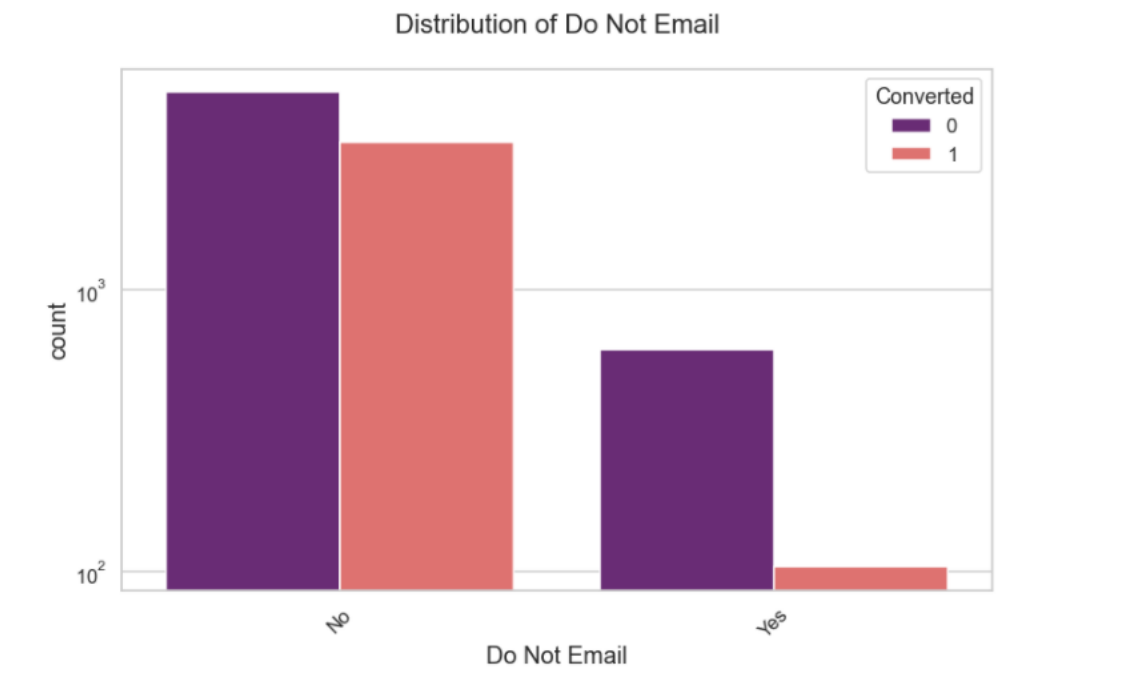
- ❑ Total Number of Rows = 37, Total Number of Columns = 9240.
- ❑ Single value features like “Magazine”, “Receive More Updates About Our Courses”, “Update me on Supply”
- ❑ Chain Content”, “Get updates on DM Content”, “I agree to pay the amount through cheque” etc. have been dropped.
- ❑ Removing the “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’.

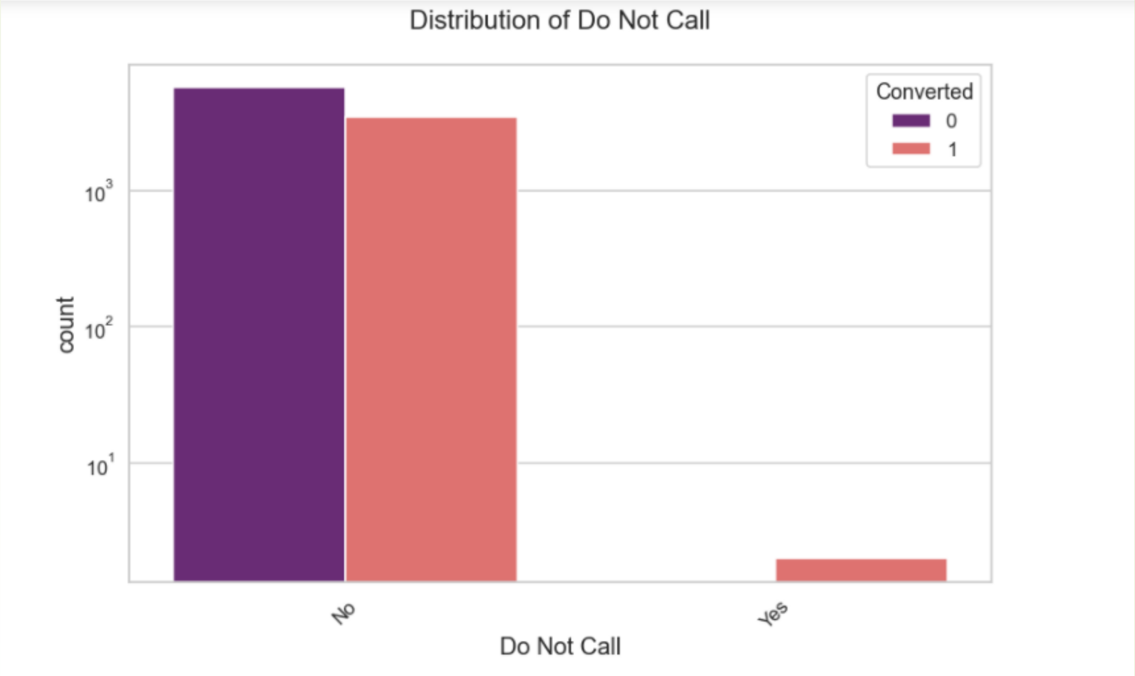
EXPLORATORY DATA ANALYSIS (EDA)

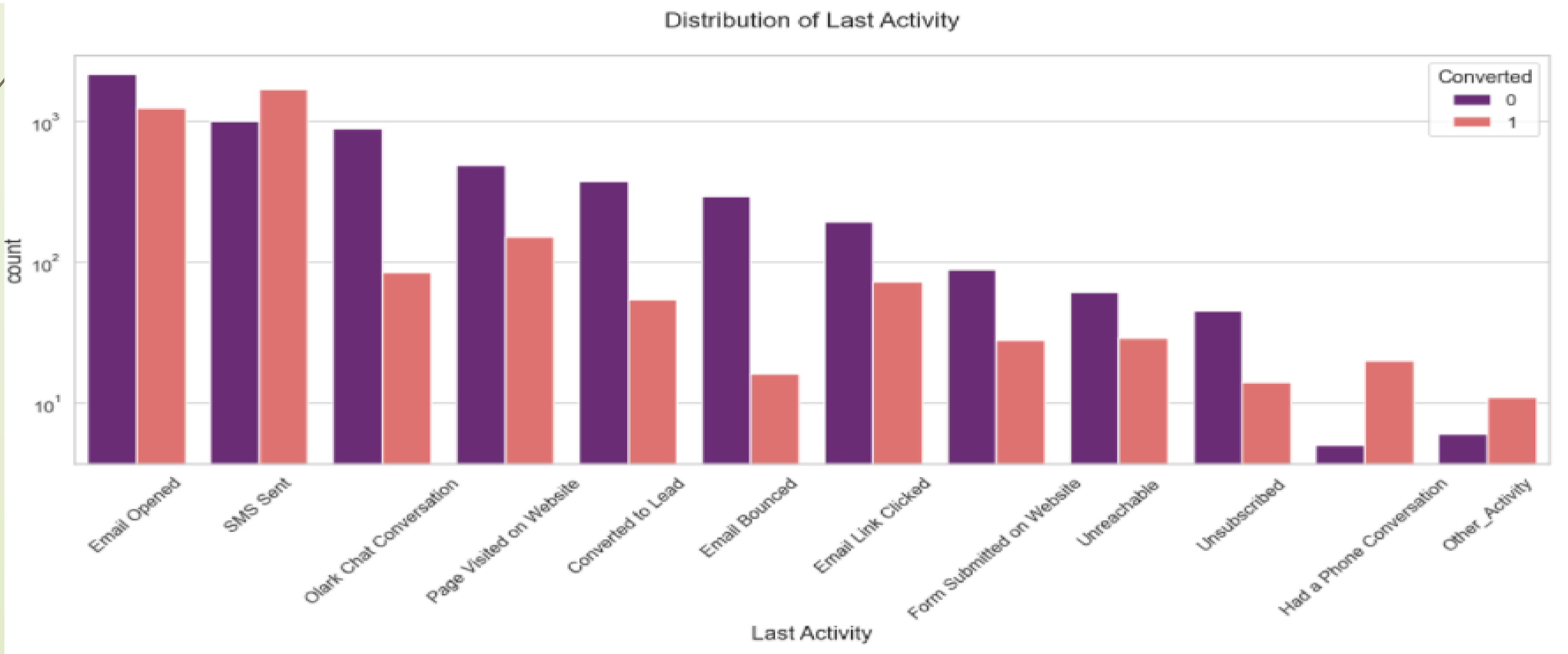
UNIVARIATE ANALYSIS – CATEGORICAL COLUMNS

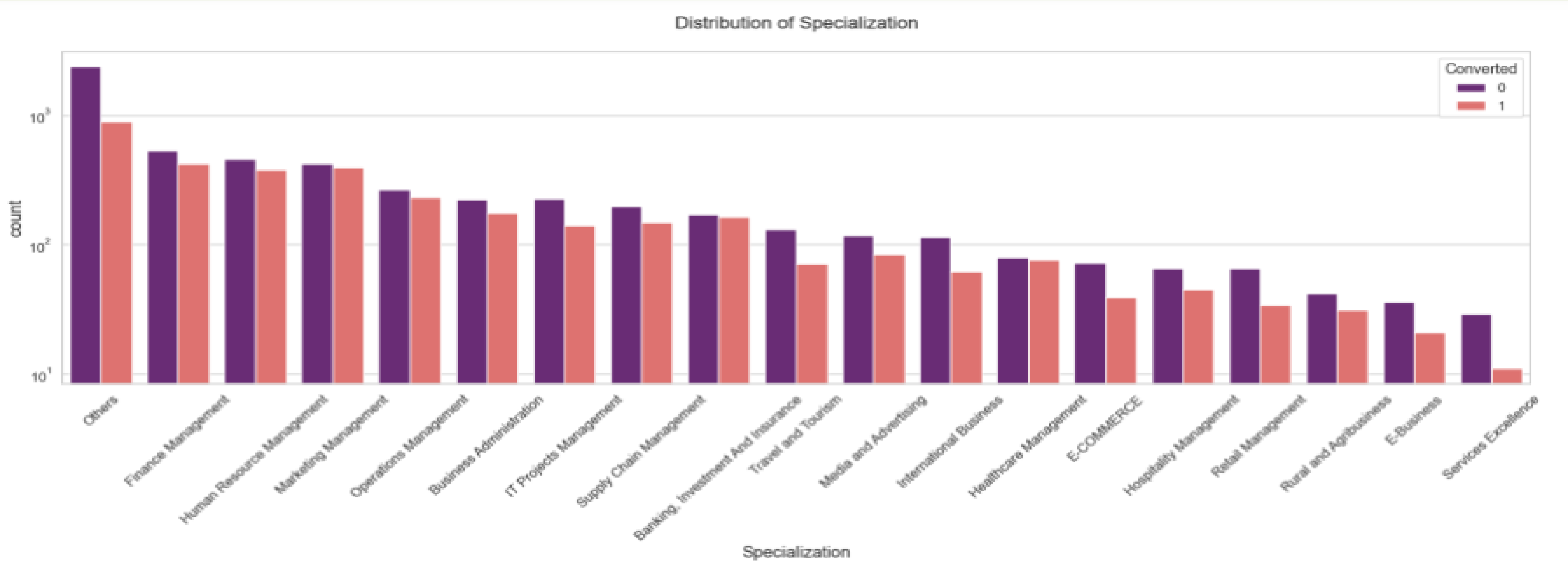


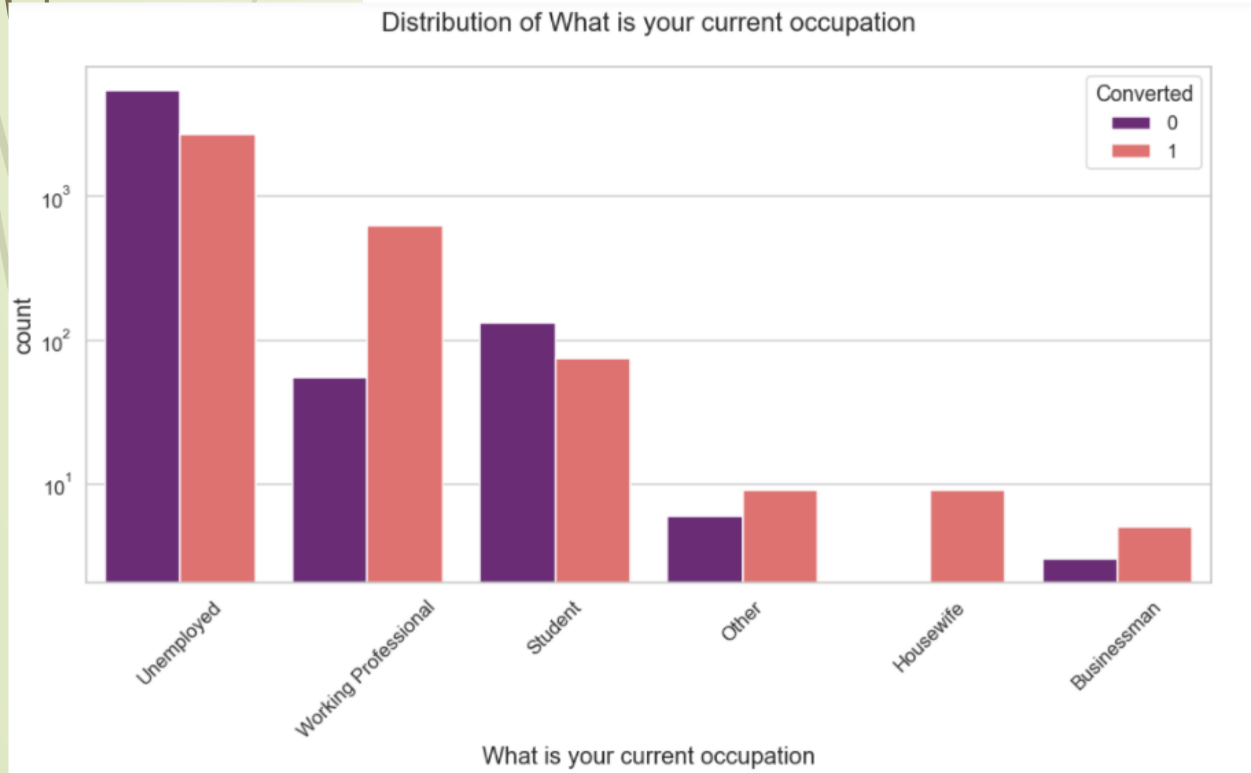


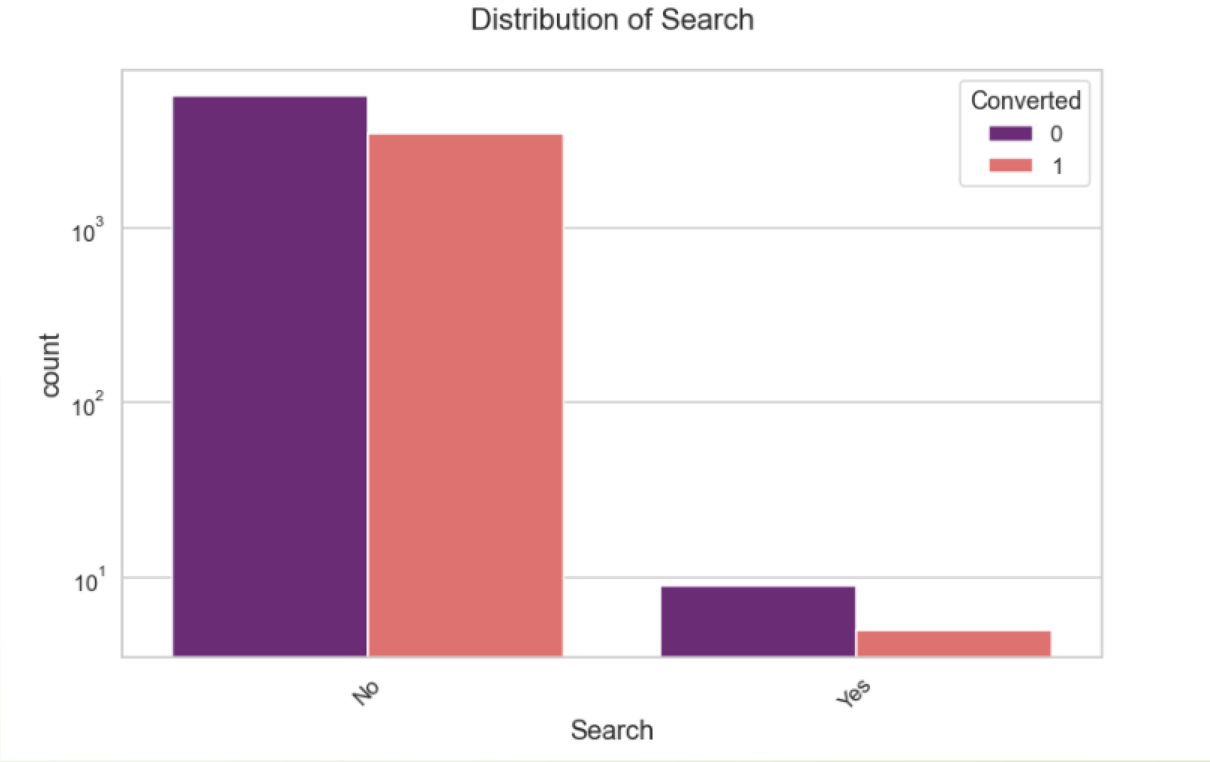

Distribution of Do Not Email
Converted
0
1
count
10^3
10^2
No
Yes
Do Not Email

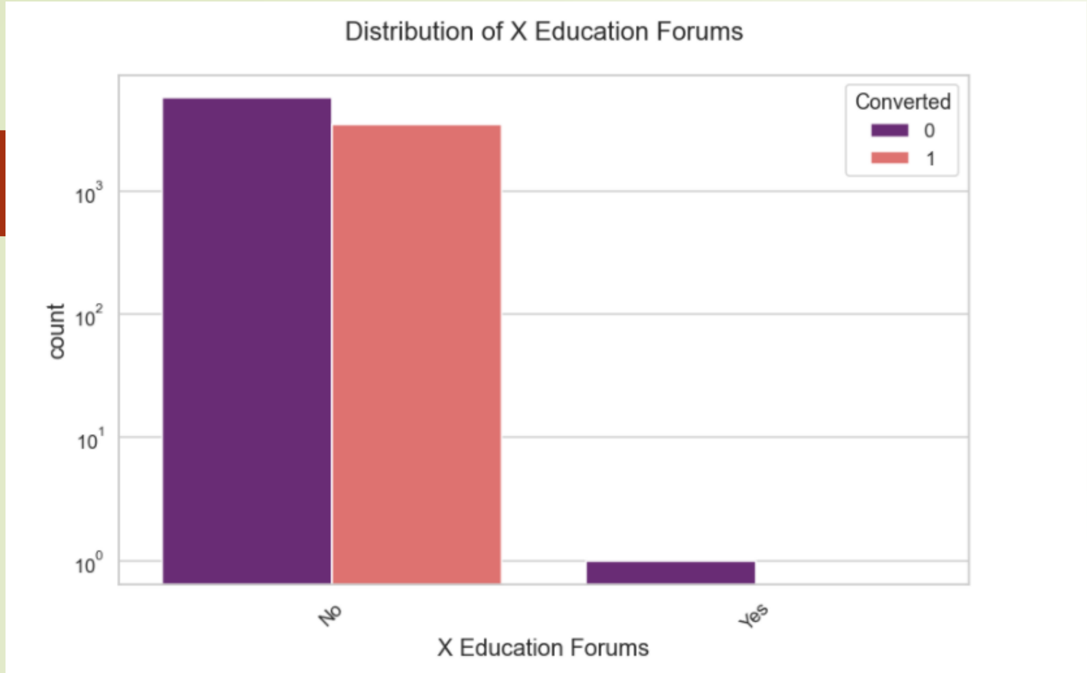

Distribution of Do Not Call
Converted
0
1
count
10^3
10^2
10^1
No
Yes
Do Not Call

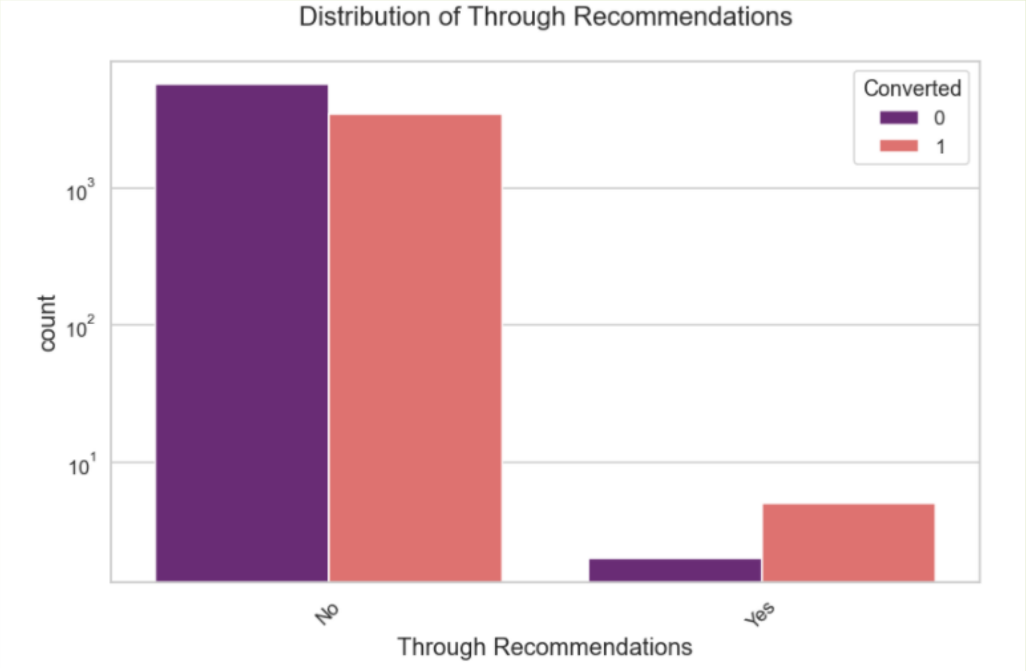

Distribution of Last Activity
Converted
0
1
count
10^3
10^2
10^1
Email Opened
SMS Sent
Olark Chat Conversation
Page Visited on Website
Converted to Lead
Email Bounced
Email Link Clicked
Form Submitted on Website
Unreachable
Unsubscribed
Had a Phone Conversation
Other_Activity
Last Activity

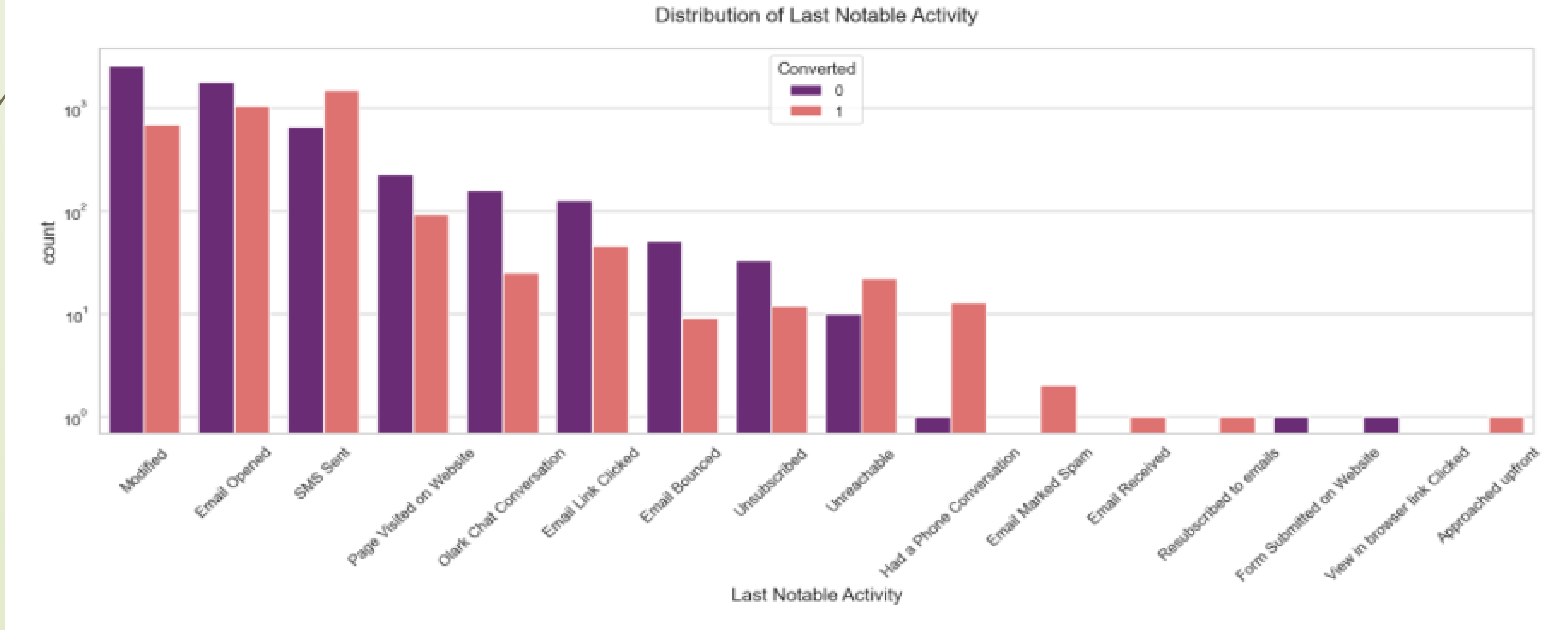

Distribution of Specialization
count
10^3
10^2
10^1
Converted
0
1
Others
Finance Management
Human Resource Management
Marketing Management
Operations Management
Business Administration
IT Projects Management
Supply Chain Management
Banking, Investment And Insurance
Travel and Tourism
Media and Advertising
International Business
Healthcare Management
E-COMMERCE
Hospitality Management
Retail Management
Rural and Agribusiness
E-Business
Services Excellence
Specialization


Distribution of What is your current occupation
count
10^3
10^2
10^1
Converted
0
1
Unemployed
Working Professional
Student
Other
Housewife
Businessman
What is your current occupation


Distribution of Search
count
10^3
10^2
10^1
Converted
0
1
No
Yes
Search


Distribution of X Education Forums
count
10^0
10^1
10^2
10^3
Converted
0
1
No
Yes
X Education Forums


Distribution of Through Recommendations
count
10^1
10^2
10^3
Converted
0
1
No
Yes
Through Recommendations


Distribution of Last Notable Activity
count
10^0
10^1
10^2
10^3
Converted
0
1
Modified
Email Opened
SMS Sent
Page Visited on Website
Olark Chat Conversation
Email Link Clicked
Email Bounced
Unsubscribed
Unreachable
Had a Phone Conversation
Email Marked Spam
Email Received
Resubscribed to emails
Form Submitted on Website
View in browser link Clicked
Approached upfront
Last Notable Activity

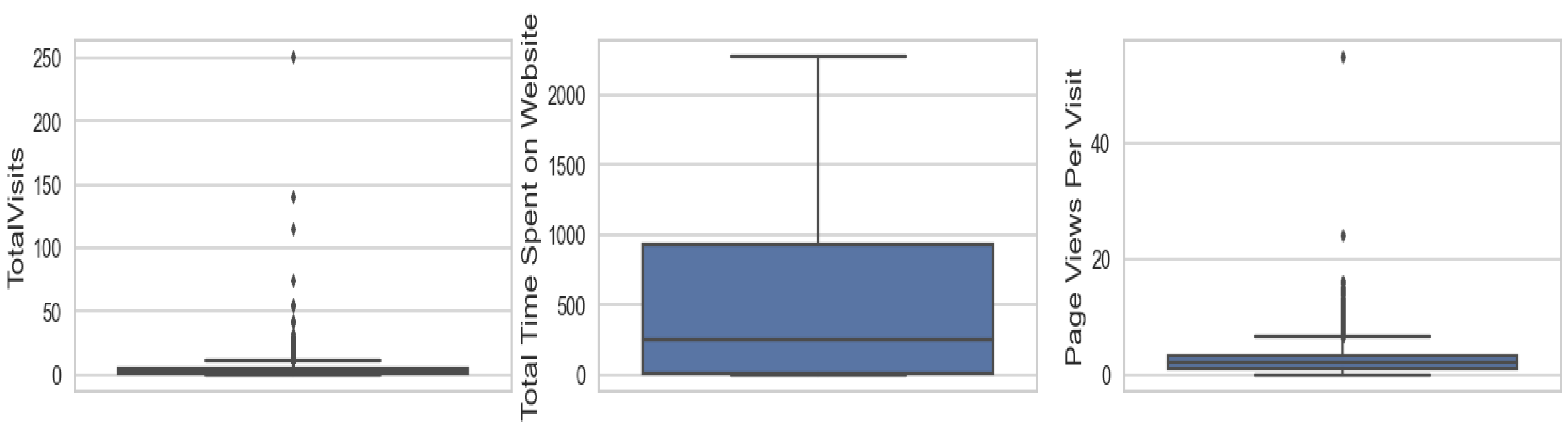
INFERENCE DRAWN FROM UNIVARIATE ANALYSIS ON CATEGORICAL COLUMNS

- **Distribution of Lead Origin:**
- Landing page submission is comparatively high than the rest of the categories in lead origin.
- lead import is the least category which is quantified in lead origin.
- Landing page submission and APO helps in lead conversion than the rest of the categories.
- Leads add form has high certainty in lead conversion #### Distribution of Lead Source.
- Google is the best lead source among all other categories in the lead source.
- Direct Traffic, Olark Chat and Organic Search are some of the best entities in lead source.
- The best category for lead conversion is Reference and Welingak Website.
- To improve overall lead conversion rate, focus should be on improving lead conversion of olark chat, organic search, direct traffic, and google leads and generate more leads from reference and welingak website.

- **Distribution of Do Not Email:**
- The customers who opted out of email communication is high.
- lead conversion through email has less certainty unlike other categories.
- **Distribution of Do Not Call:**
- The customers who opted out of call communication is high.
- lead conversion through call has less certainty unlike other categories.
- **Distribution of last Activity:**
- Most of the lead have their Email opened as their last activity.
- Conversion rate for leads with last activity as SMS Sent is almost 62%.
- **Distribution of Specialization:**
- Focus should be more on the Specialization with high conversion rate like Supply Chain, Human Resource and Finance.
- **Distribution of Occupation:**
- Working Professionals going for the course have high chances of joining it.
  - Unemployed leads are the most in numbers but has around 30-35% conversion rate.

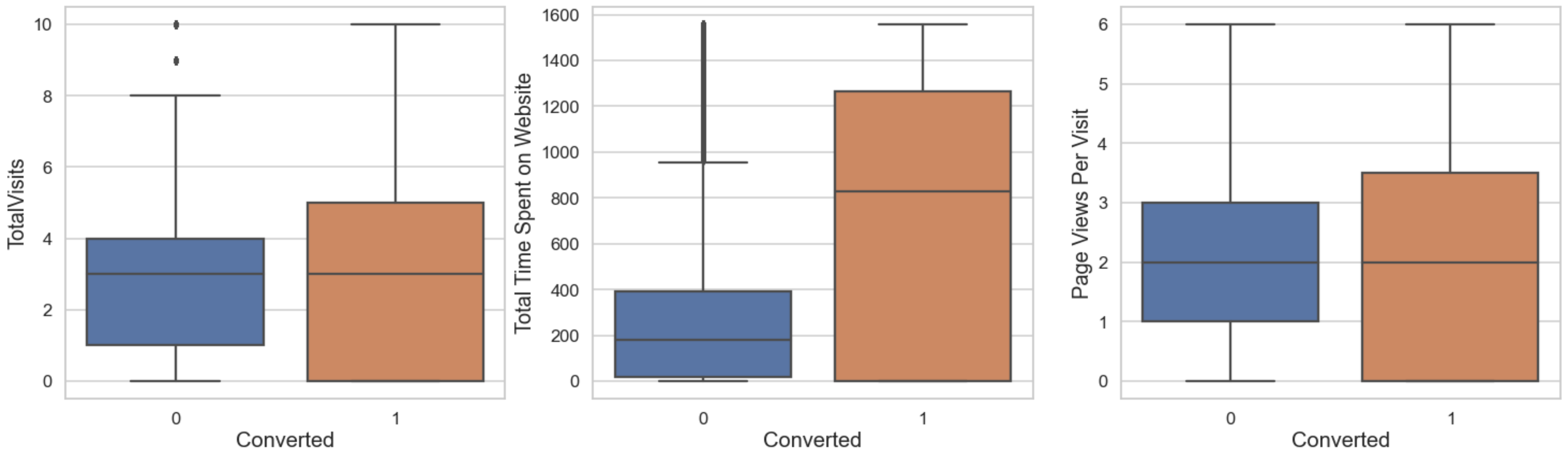
- **Distribution of Search:**
- Most entries are 'No'. No inference can be drawn with this parameter.`
- **Similar to Distribution of Search, No Inference can be drawn from the following features:**
- 'Newspaper Article', 'X Education Forums', 'Newspaper', 'Digital Advertisement', 'Through Recommendations'
- **Distribution of City:**
- Most leads are from Mumbai with around 30% conversion rate.`
- **Distribution of Last notable Activity:**
- 'SMS Sent' is strong entity for positive lead.
- **Results from Univariate Analysis for Categorical Features:**
- Based on the univariate analysis we have seen that many columns are not adding any information to the model, hence we can drop them for further analysis.

UNIVARIATE ANALYSIS - NUMERICAL COLUMNS



- We can clearly spot outliers in the features such as TotalVisits and Page Views Per Visit. Total time spent on Website doesn't have any outliers. We can see that there are as many as 250 visits recorded for total visits by possible leads. As high as this number of visits to a website seems to be not like a correct capture and hence we can remove these outliers. Similarly, for the Page Views Per Visit, as many as 20+ page views in a single visit seems to be not correct. We can remove these as well.

DISTRIBUTION OF NUMERICAL FEATURES AFTER OUTLIER TREATMENT



Inferences:

Total Visits:

- Median for converted and not converted leads are the same.
- Nothing conclusive on the basis of Total Visits.

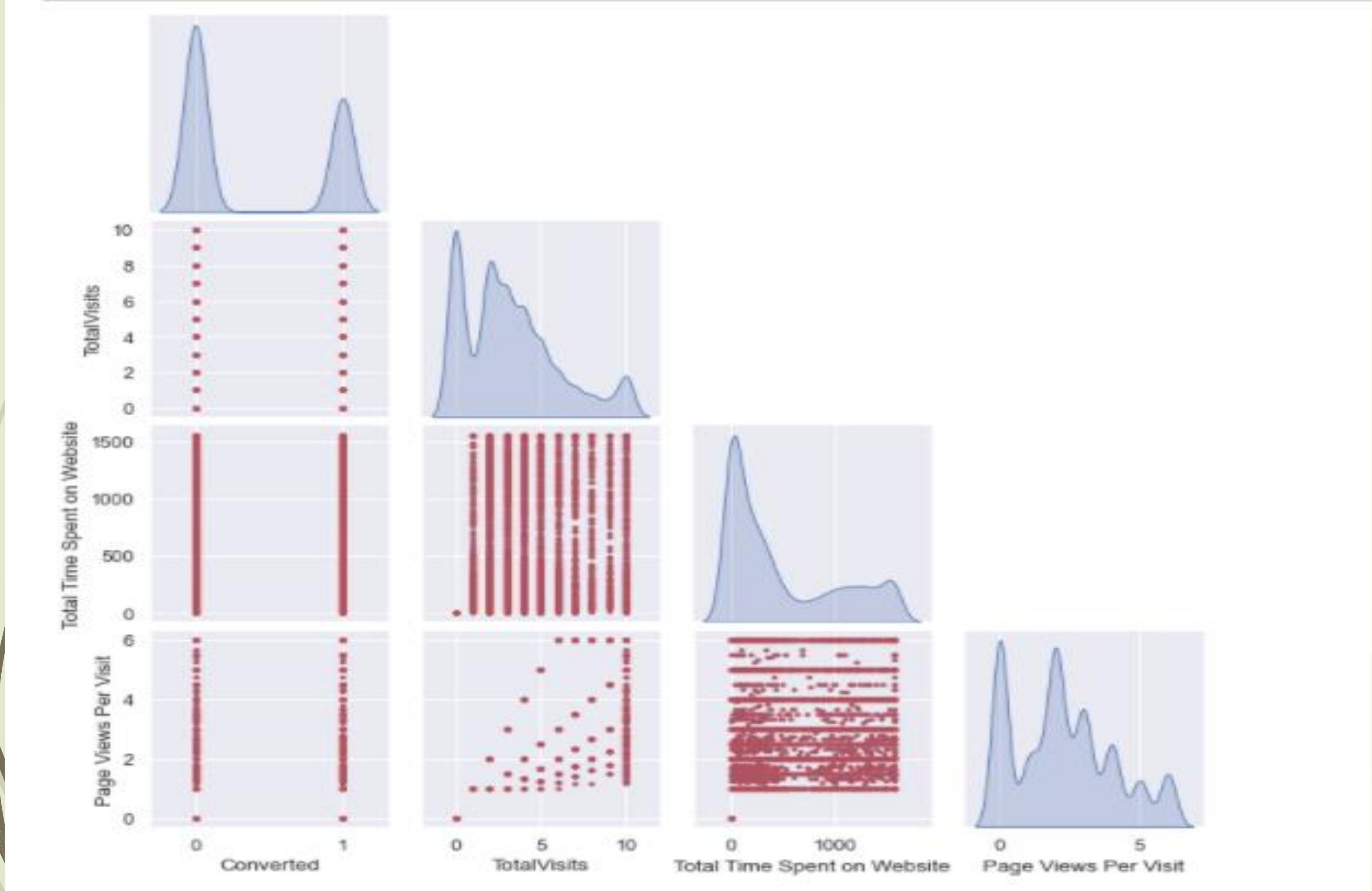
Total Time Spent on Website:

- Leads spending more time on the website are more likely to be converted.
- Websites should be made more engaging to make leads spend more time.

Page Views Per Visit:

- Nothing can be said specifically for lead conversion from Page Views Per Visit.

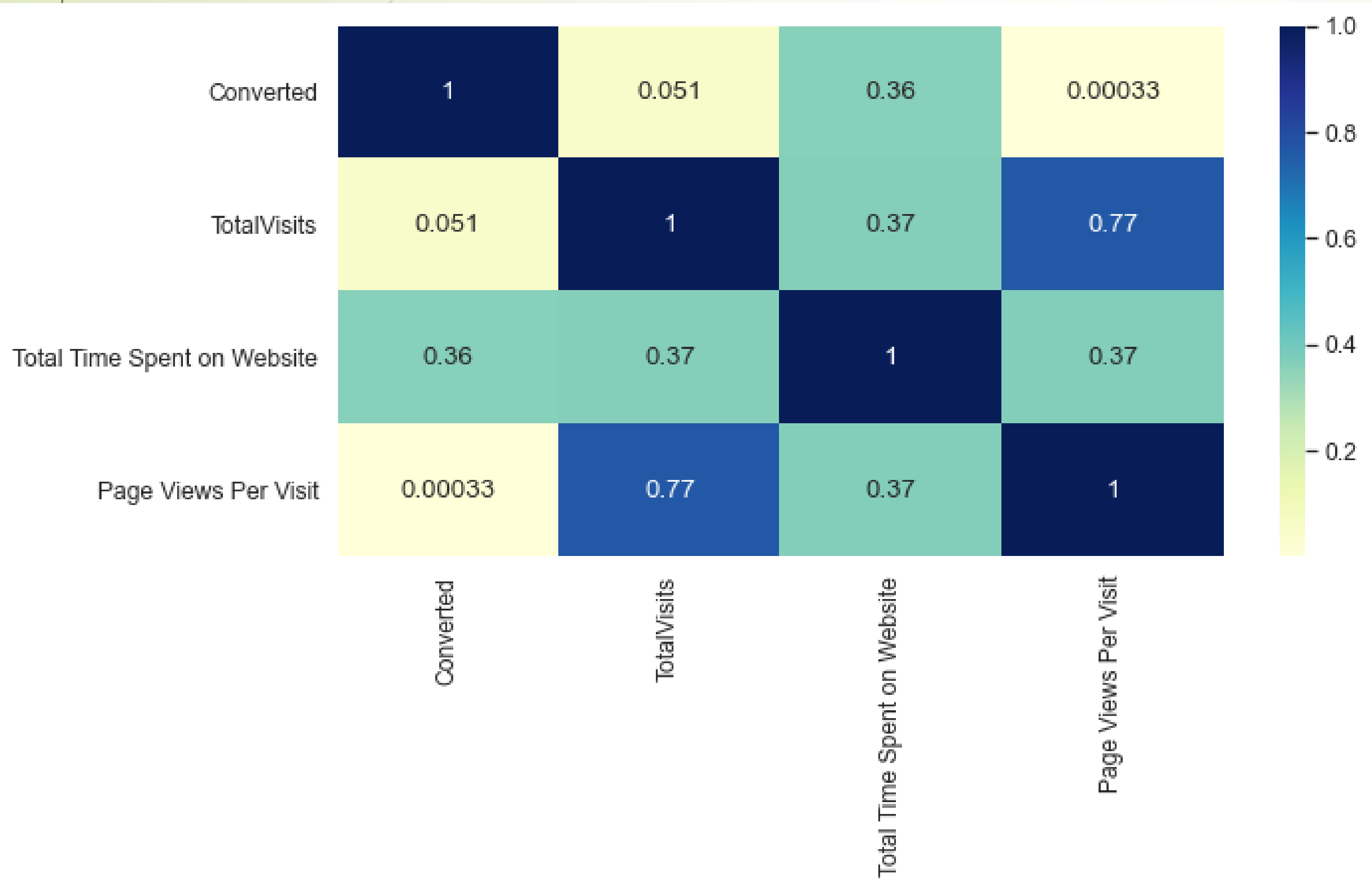
BIVARIATE ANALYSIS



Inferences:

- The data is skewed and we could witness a lot of noise in the data.

Correlation Plot to Check Multi-Collinearity



	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit
Converted	1	0.051	0.36	0.00033
TotalVisits	0.051	1	0.37	0.77
Total Time Spent on Website	0.36	0.37	1	0.37
Page Views Per Visit	0.00033	0.77	0.37	1

Inference:

- Total Visits and Page Views Per Visit has high correlation than other features.
- Total Visits and Converted has very low correlation results which means that based on Total Visits we can derive meaningful lead scoring.
- Total Visits and Total Time Spent on Website have a reasonable correlation result.
- There is positive correlation between Total Time Spent on Website and Converted.
- There is almost no correlation in Page Views Per Visit and Total Visits with Converted.

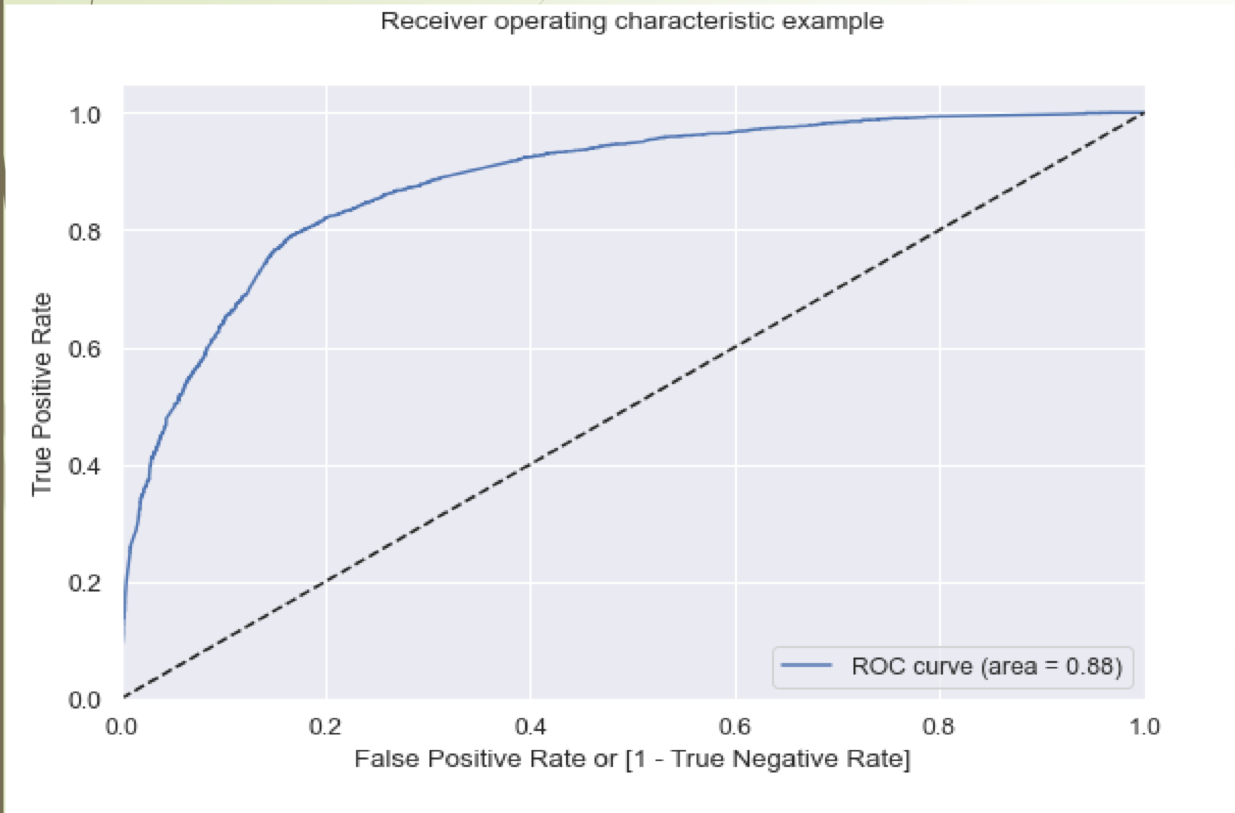
DATA CONVERSION

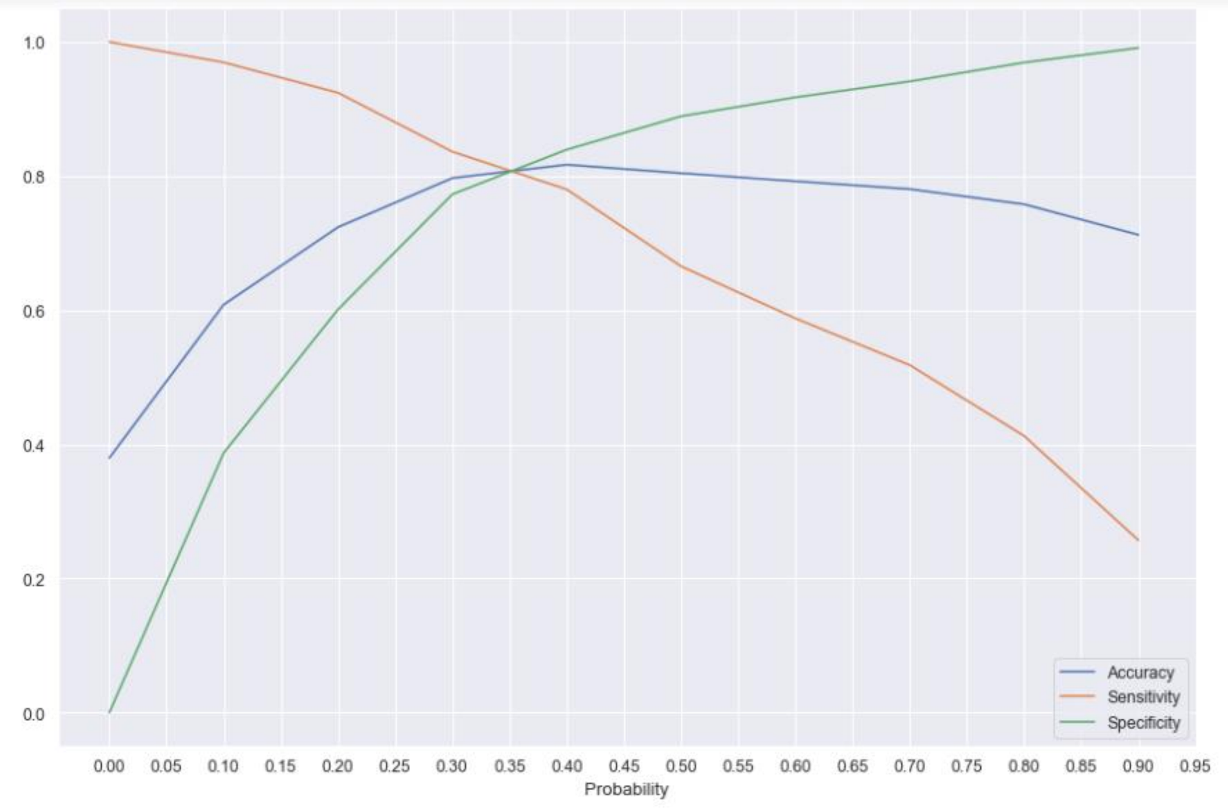
- ❑ Numerical Variables are Normalised
- ❑ Dummy Variables are created for object type variables
- ❑ Total Rows for Analysis: 9074
- ❑ Total Columns for Analysis: 54

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 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 80.4%

ROC CURVE

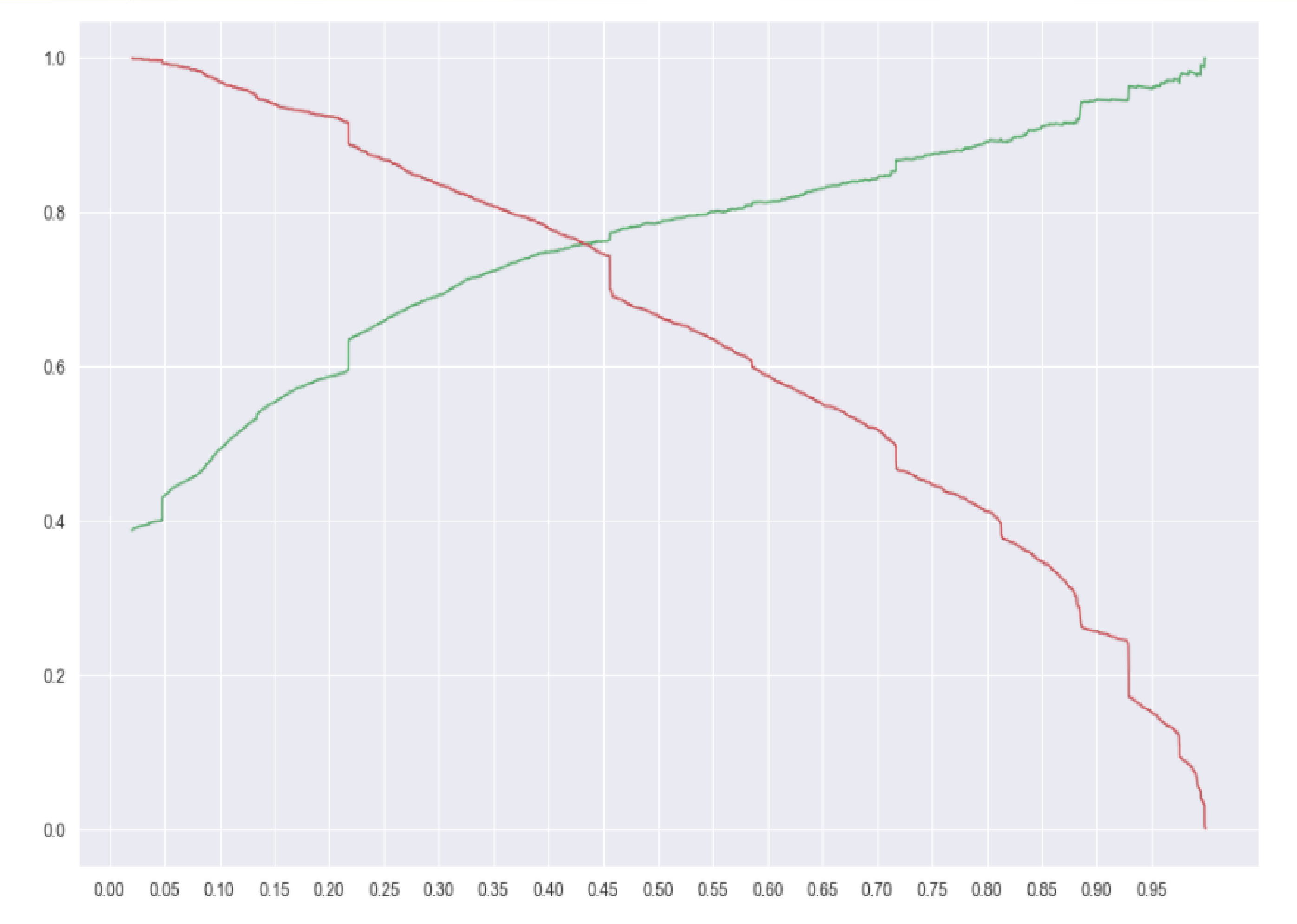




Finding the optimal cutoff point

- Optimal cut off probability is that probability where we get balanced sensitivity and specificity.
- From the second graph it is visible that the optimal cut off is at 0.35.

PRECISION AND RECALL TRADE OFF



Inference:

- As per Precision-Recall Tradeoff, the cutoff is around 0.425 (between 0.4 and 0.45).
- We can choose the cut-off as 0.45 and use the Precision-Recall-Accuracy metrics to evaluate the model.

FINAL OBSERVATIONS

For Train Data

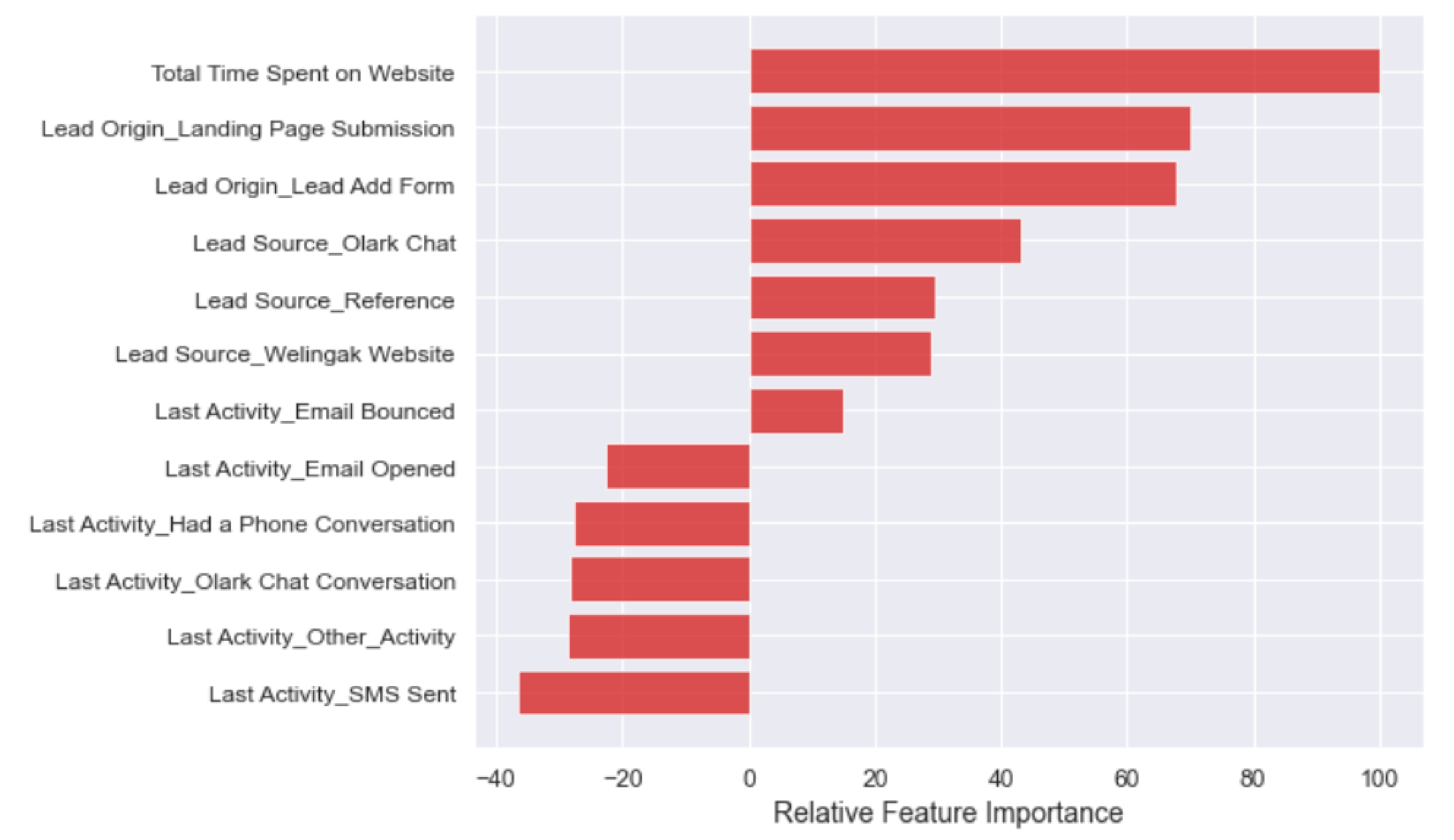
Train Data Accuracy	:81.0%
Train Data Sensitivity	:80.77%
Train Data Specificity	:81.15%
Train Data F1 Score	:0.72

For Test Data

Test Data Accuracy	:81.43%
Test Data Sensitivity	:80.77%
Test Data Specificity	:81.15%
Test Data F1 Score	:0.74

Inference: The sensitivity value on Test Data is 80.77 vs 80.77 on Train Data. The accuracy value for both Test & Train Data is around 81%. It shows that our model is performing well in test data set also and is not over-trained.

PLOT FOR RELATIVE FEATURE IMPORTANCE



RECOMMENDATIONS

- ❑ Increase user engagement on Welingak website since this helps in higher conversion.
- ❑ Focus on Working Professional which has high conversion certainly.
- ❑ Get Total Time Spent on Website increased by advertising and user experience which makes the customer engaging in the website. Since this helps in higher conversion.
- ❑ Improve the Olark Chat service since this is affecting the conversion negatively.
- ❑ Improving Lead add form also improves the lead conversion with high certainty.

THANK YOU