LEAD SCORING CASE STUDY

Group Members:

1) Mahesh Kumar Patra

2) Priyadharshan T

PROBLEM STATEMENT

- An education company named 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.

METHODS FOR SOLUTION

Data Loading and Cleaning

- Import the data
- Inspect the data
- Cleaning the data

Exploratory data analysis

- Univariate analysis
- Segmented univariate analysis
- Bivariate analysis

Dummy variables

- Convert binary variables from yes/no to 1/0
- Create dummy variables

Train test split and scaling

Model building

Model evaluation

Plot ROC curve and find optimal cut-off point

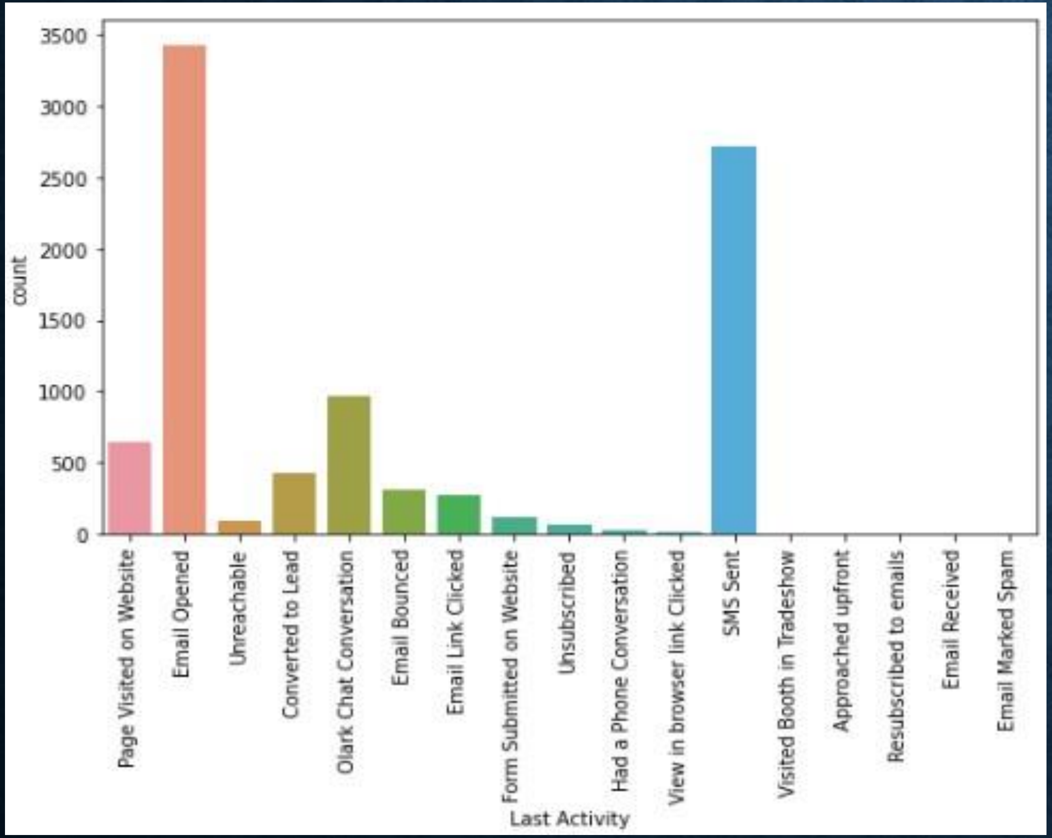
Make prediction on test data

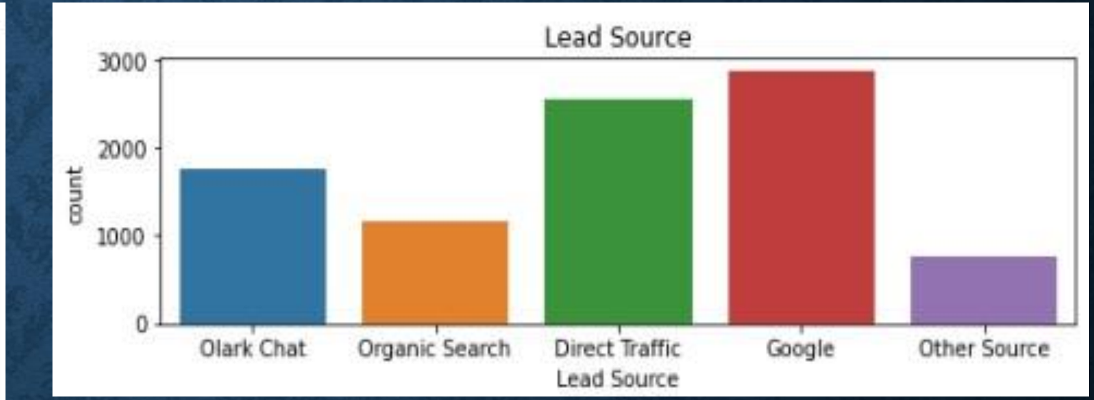
ASSUMPTIONS

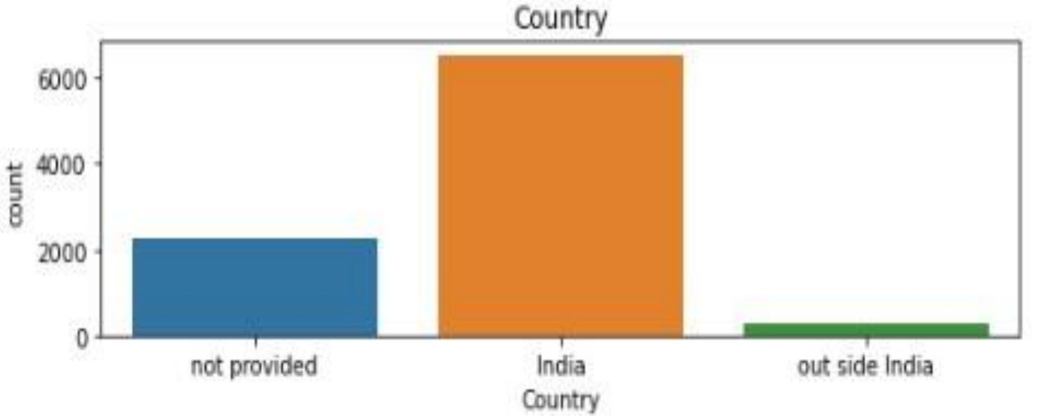
- Replace ‘select’ with null
- Drop the features having null values more than 45%
- Drop the features having only single value
- Drop the features, those have imbalance data
- Impute null values with ‘not provided’
- Combine some data into single group
- Removing the “Prospect ID” and “Lead Number” which is not necessary for the analysis.

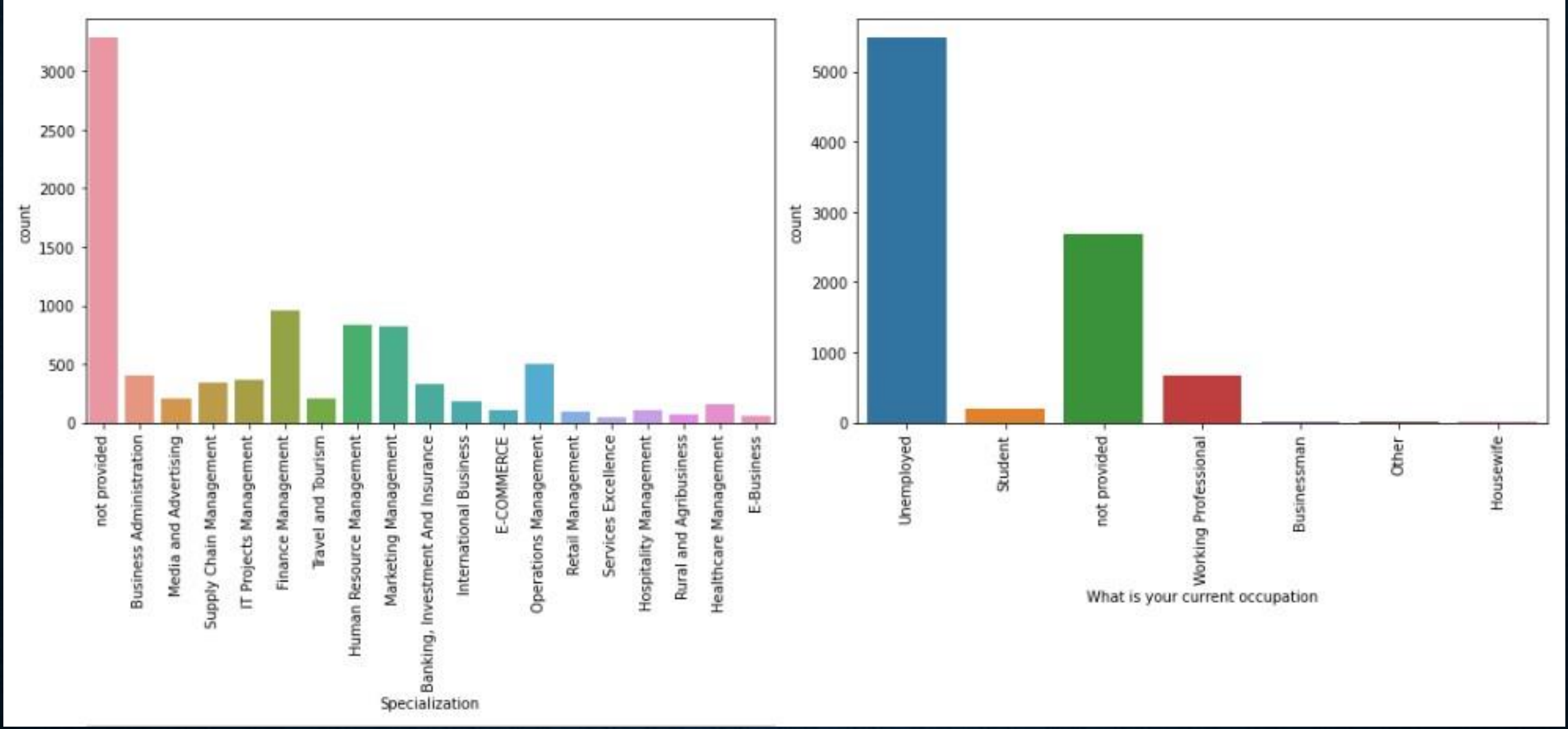
EXPLORATORY DATA ANALYSIS

- Univariate analysis on categorical varibles

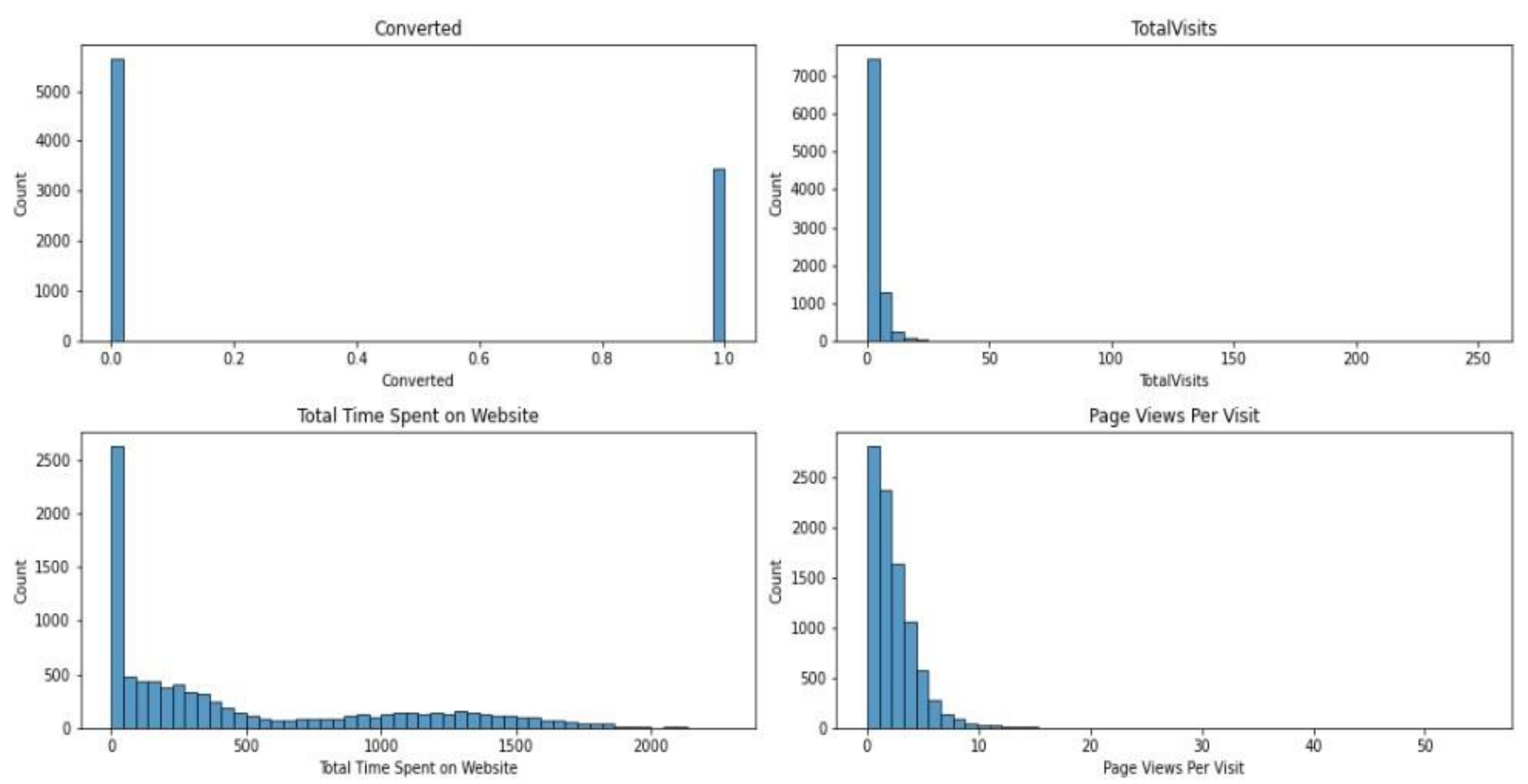




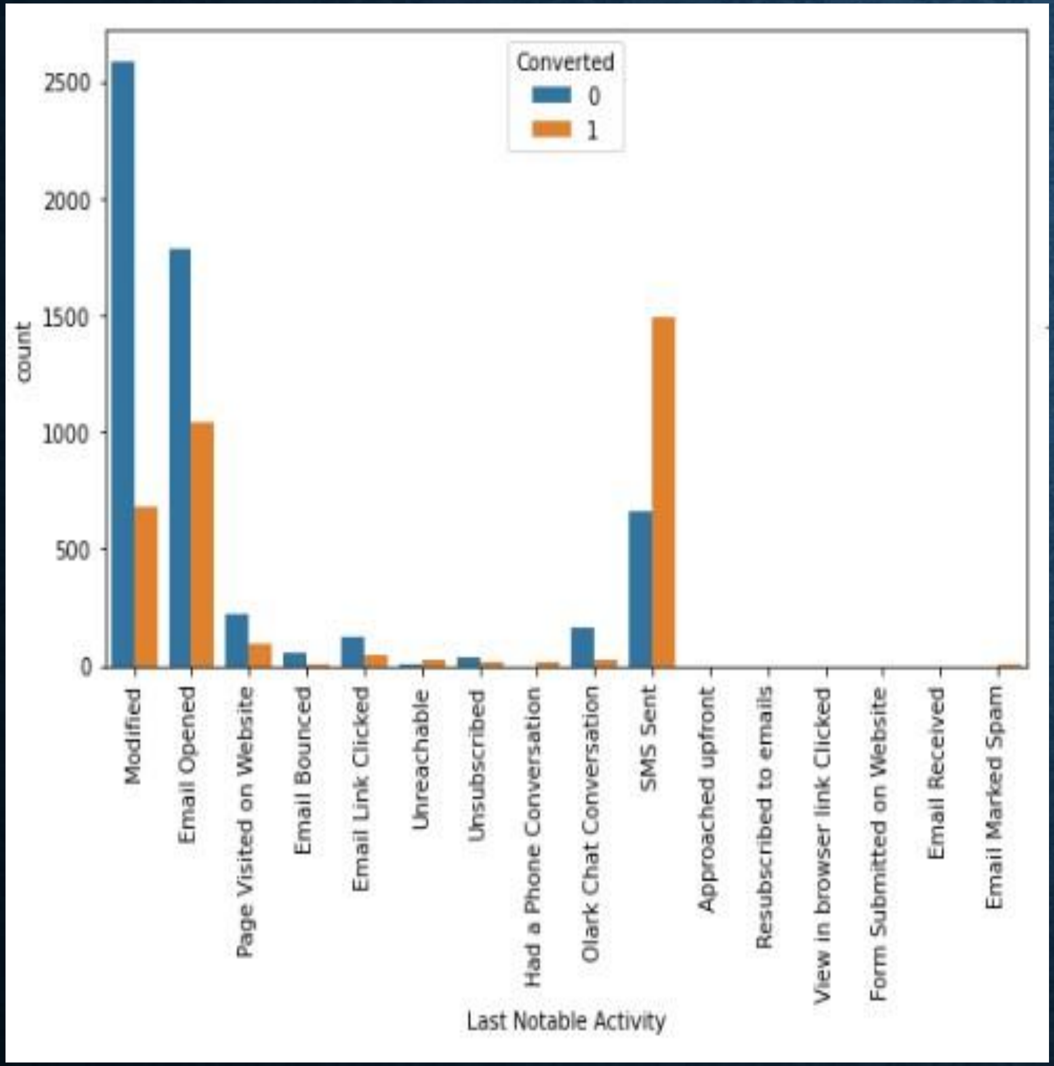


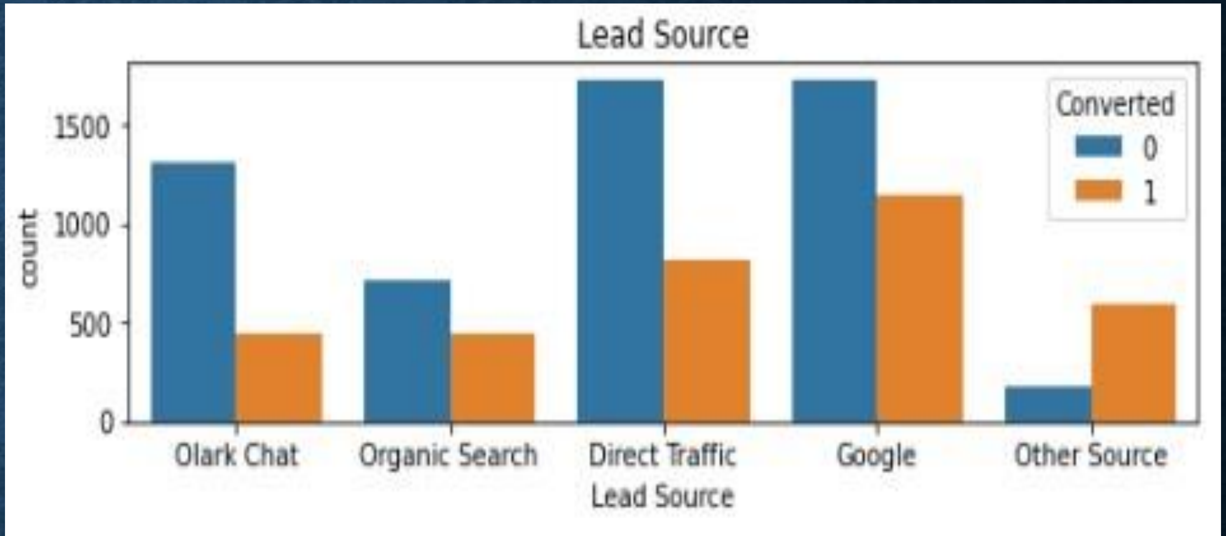

count
3000
2500
2000
1500
1000
500
0
not provided
Business Administration
Media and Advertising
Supply Chain Management
IT Projects Management
Finance Management
Travel and Tourism
Human Resource Management
Marketing Management
Banking, Investment And Insurance
International Business
E-COMMERCE
Operations Management
Retail Management
Services Excellence
Hospitality Management
Rural and Agribusiness
Healthcare Management
E-Business
Specialization
count
5000
4000
3000
2000
1000
0
Unemployed
Student
not provided
Working Professional
Businessman
Other
Housewife
What is your current occupation

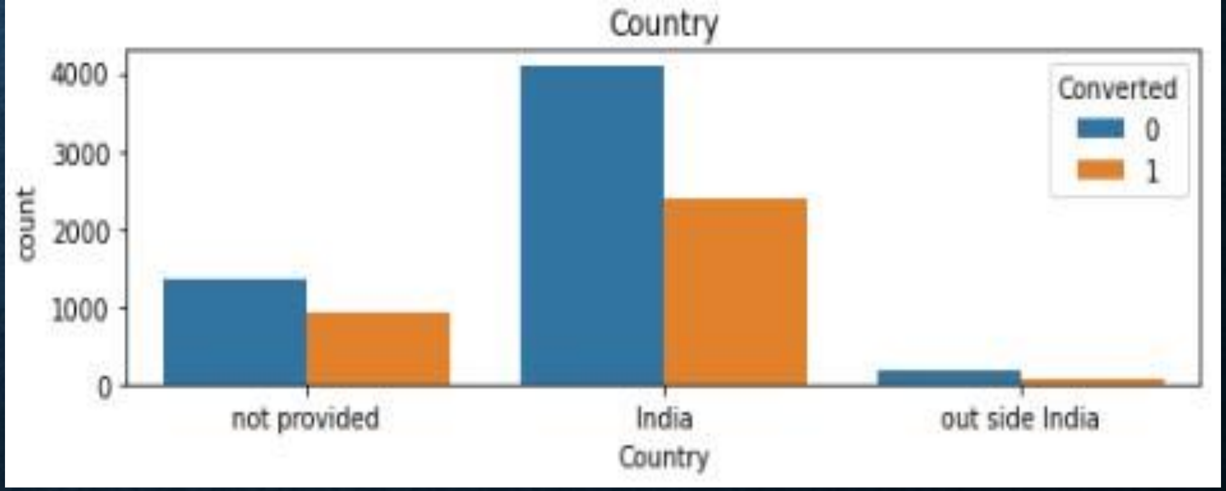
- Univariate analysis on numerical varibles

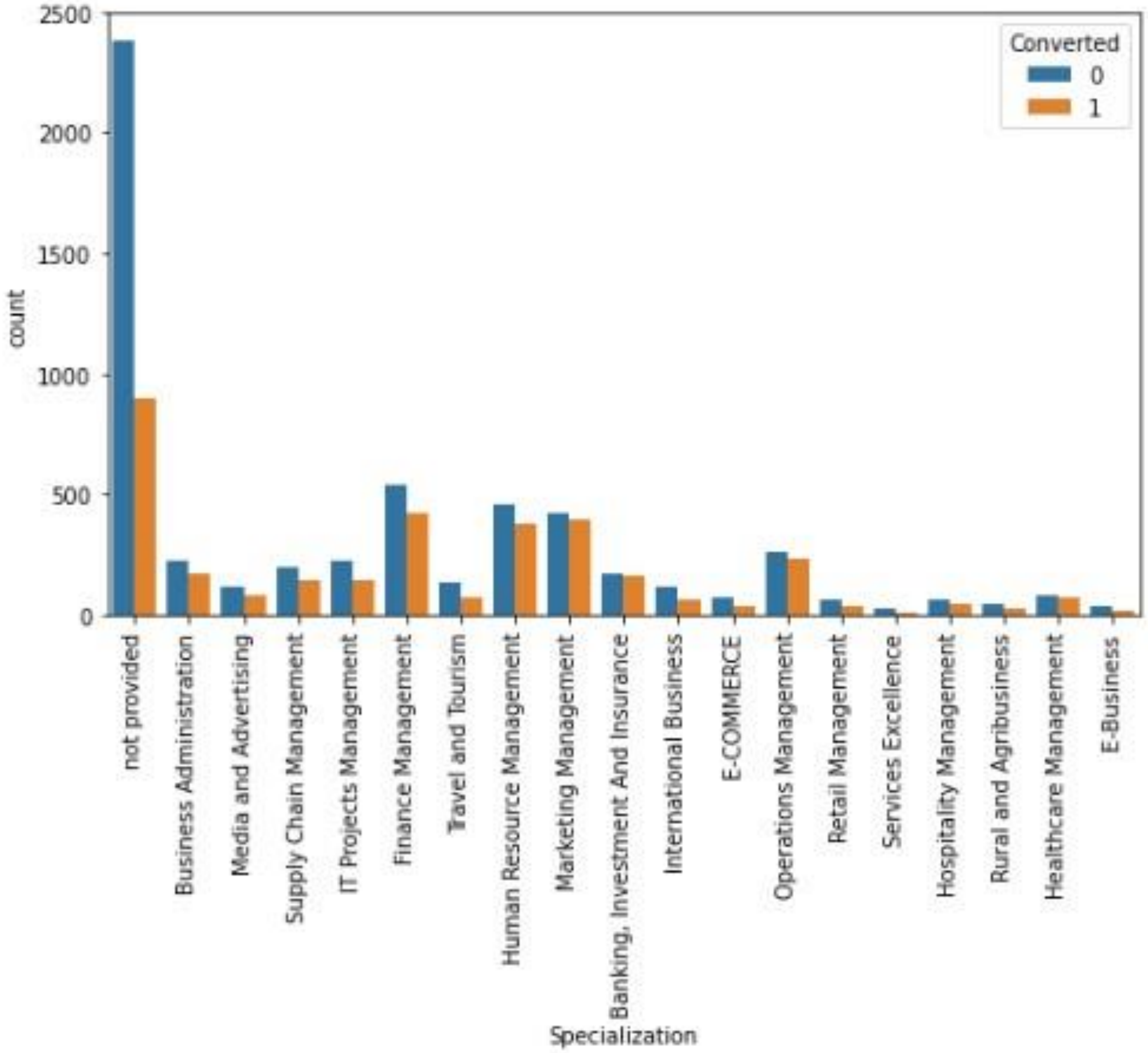


- Segmented univariate analysis on categorical variables

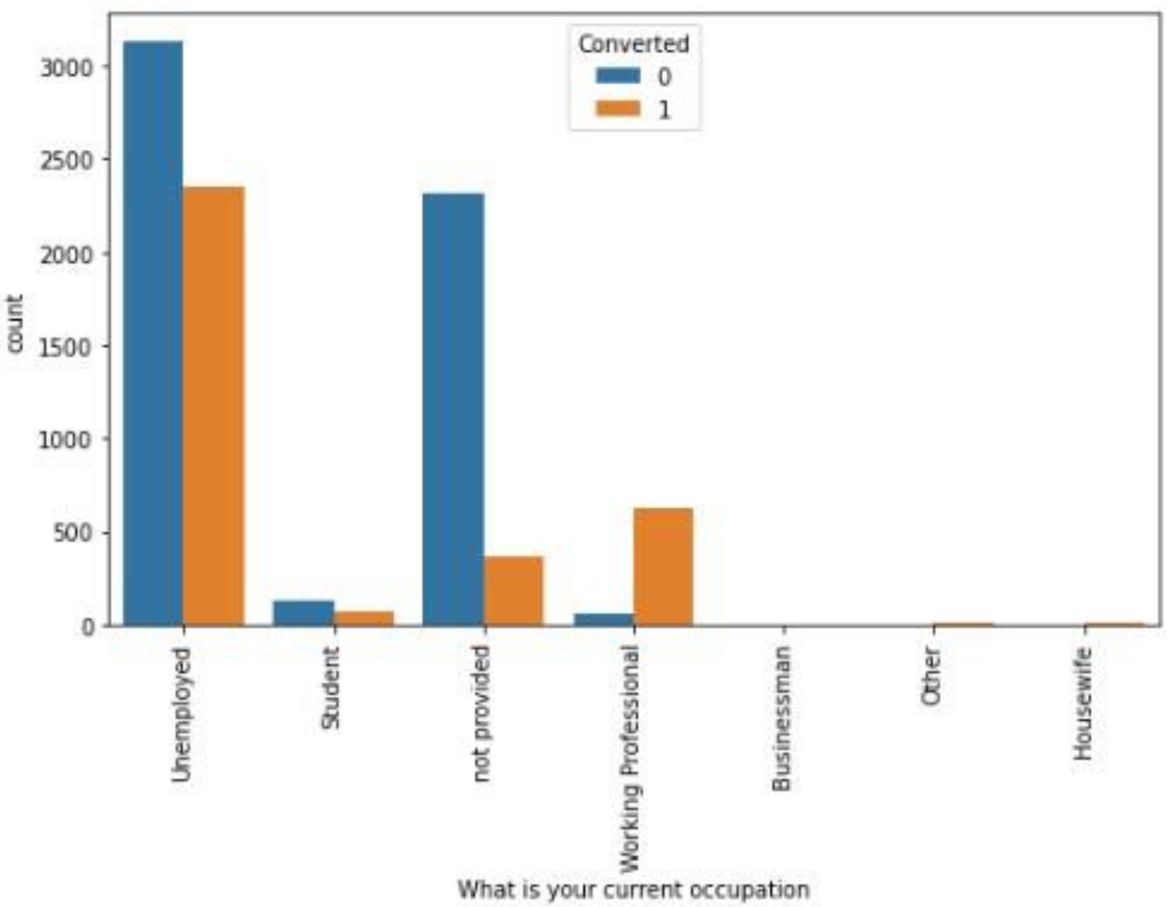








count
2500
2000
1500
1000
500
0
Converted
🟦 0
🟧 1
not provided
Business Administration
Media and Advertising
Supply Chain Management
IT Projects Management
Finance Management
Travel and Tourism
Human Resource Management
Marketing Management
Banking, Investment And Insurance
International Business
E-COMMERCE
Operations Management
Retail Management
Services Excellence
Hospitality Management
Rural and Agribusiness
Healthcare Management
E-Business
Specialization

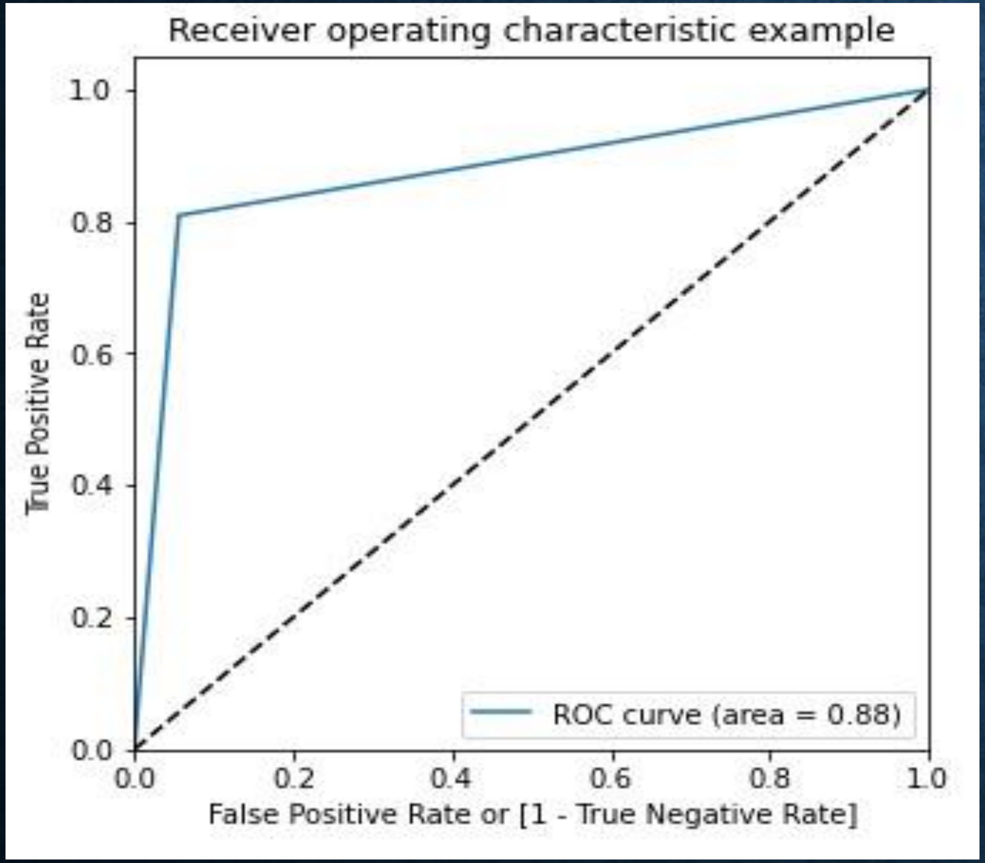


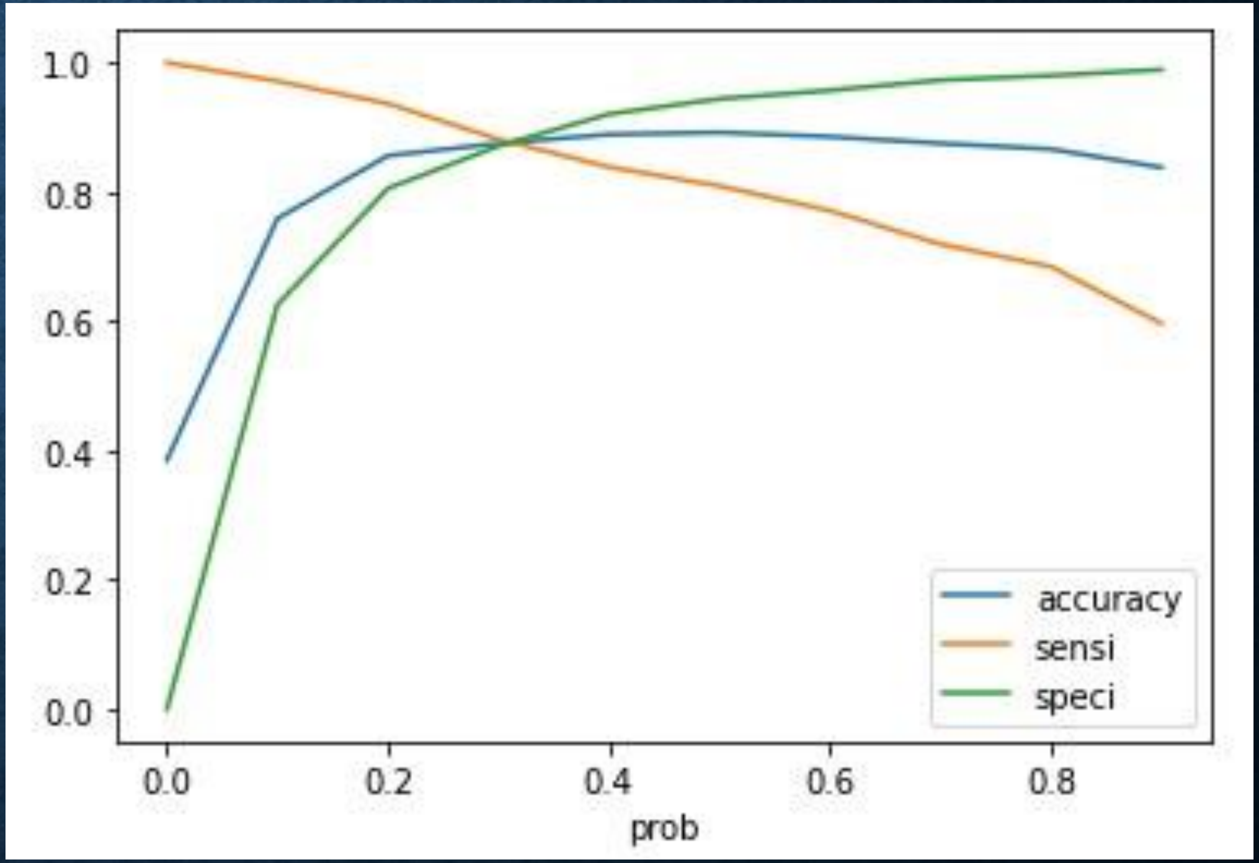
count
3000
2500
2000
1500
1000
500
0
Converted
🟦 0
🟧 1
Unemployed
Student
not provided
Working Professional
Businessman
Other
Housewife
What is your current occupation

MODEL BUILDING

- Create Dummy variables
- Split the data set into train and test sets
- Scale the numerical features for both train and test data set
- Create a logistic regression model
- Use RFE for feature selection
- Check the p-values and VIF values
- Make the prediction on train data set and create confusion matrix
- Calculate accuracy, sensitivity, specificity
- Plot ROC curve to find out the optimal cutoff point
- Calculate precision and recall
- Plot precision-recall tradeoff curve
- Make the prediction on test data set and create confusion matrix again
- Calculate precision and recall again

- <u>ROC curve</u>


Receiver operating characteristic example
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate or [1 - True Negative Rate]
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.88)


0.0
0.2
0.4
0.6
0.8
1.0
prob
0.0
0.2
0.4
0.6
0.8
accuracy
sensi
speci

CONCLUSION

- It was found that the variables that mattered the most in the potential buyers are,
    - Total time spent on website
    - Total number of visits
    - Last activity was SMS sent and Olark chat conversation
    - Lead source was Google, Direct traffic, Organic search
    - Lead origin is lead add form
    - Current occupation is working professional