LEAD SCORE CASE STUDY

PROBLEM STATEMENT

- An education company named X Education sells online courses to industry professionals.
- People searches the course in search engine and fill up the form for the course by providing email address or phone number, they are classified as lead.
- Now, although 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 OJECTIVE

- X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- The company needs a model wherein you a lead score is assigned 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%

APPROACH AND METHODOLOGY

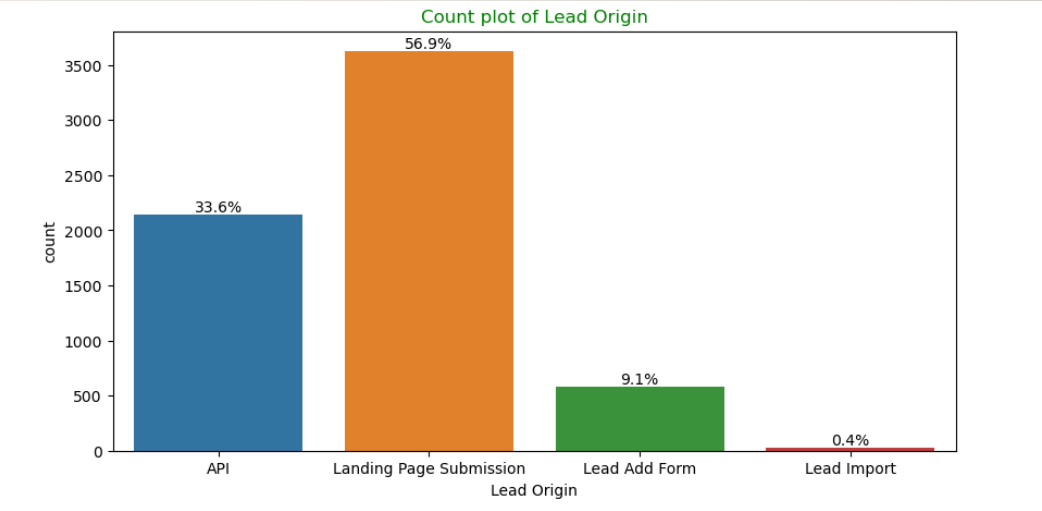
- **Data cleaning and data manipulation**
- 1. Check and handle duplicate data.
- 2. Check and handle NA values and missing values.
- 3. Drop columns, if it contains many missing values and are not useful for the analysis.
- 4. Imputation of the values, if necessary.
- 5. Check and handle outliers in data.

APPROACH AND METHODOLOGY (CONT...)

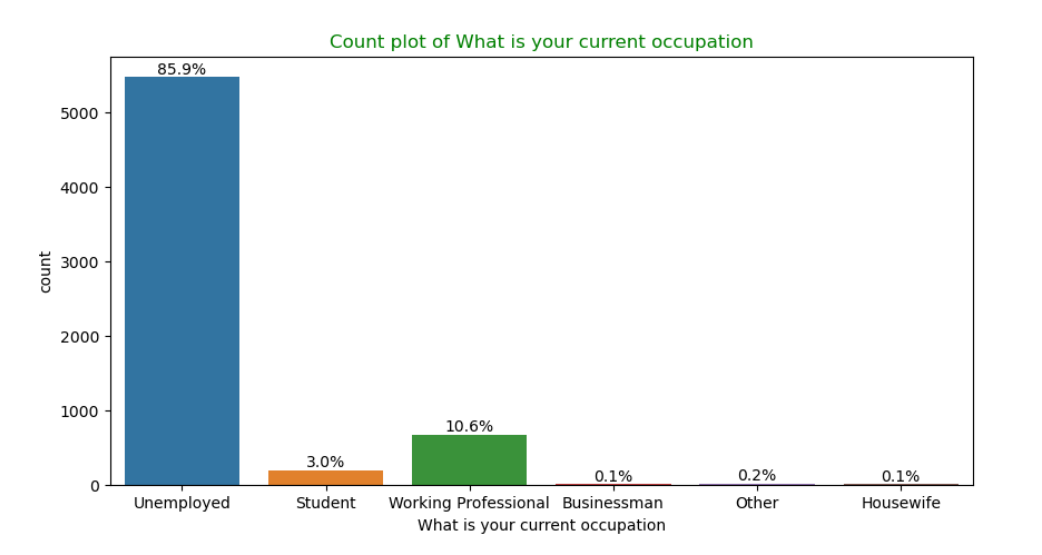
- **Exploratory Data Analysis (EDA)**
- 1. Univariate data analysis: value count, distribution of variables, etc.
- 2. Bivariate data analysis: correlation coefficients and pattern between the variables etc.
- 3. Feature Scaling & Dummy variables and encoding of the data.
- 4. Classification technique: logistic regression is used for model making and prediction.
- 5. Validation of the model.
- 6. Model presentation.
- 7. Conclusions and recommendations.

EXPLORATORY DATA ANALYSIS

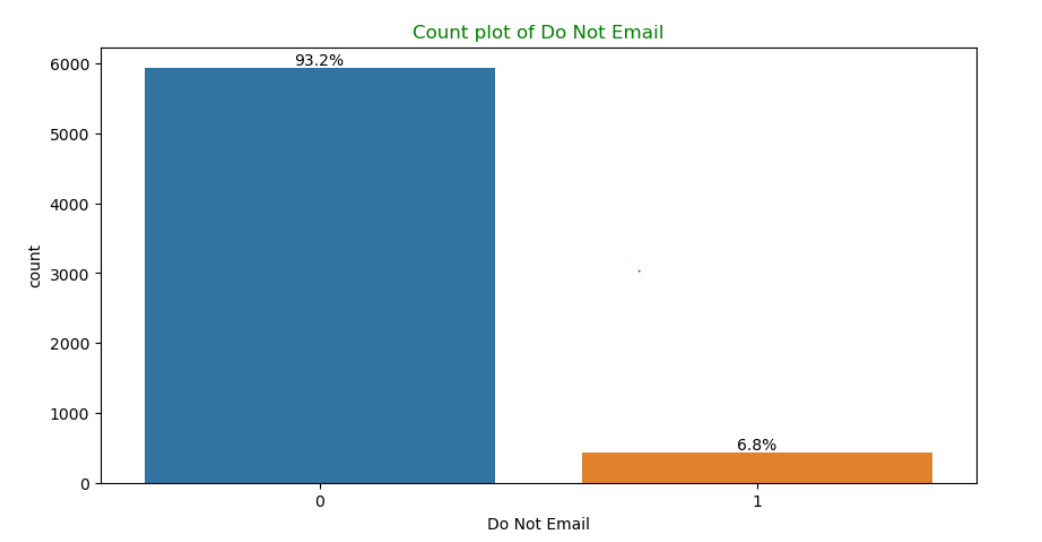
- **Univariate Analysis for Categorical Variables**
- 1. Lead Origin: "Landing Page Submission" identified 57% customers; "API" identified 34%.



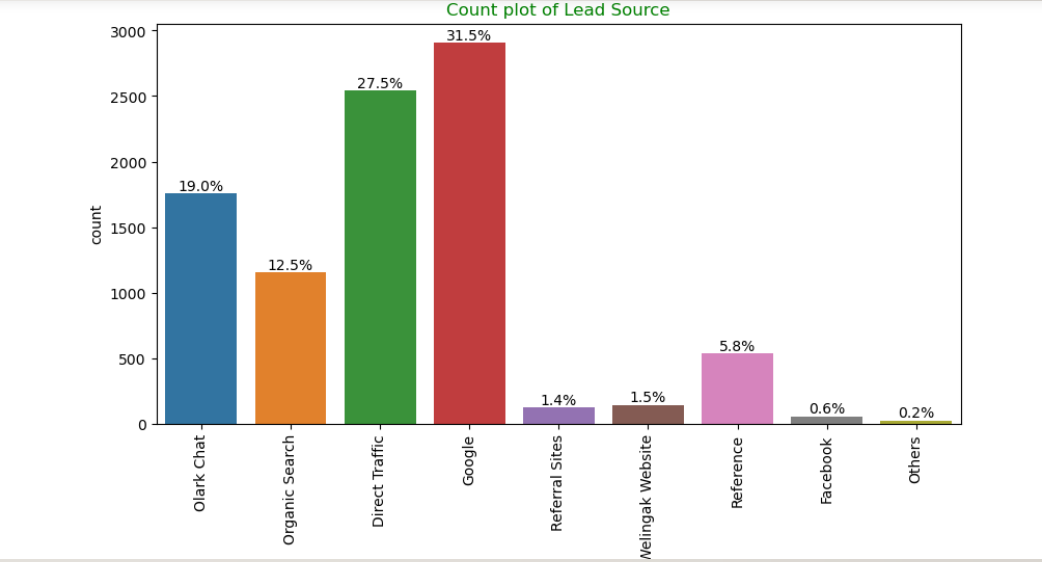
- **Univariate Analysis for Categorical Variables**
- 2. Current occupation: It has 86% of the customers as Unemployed



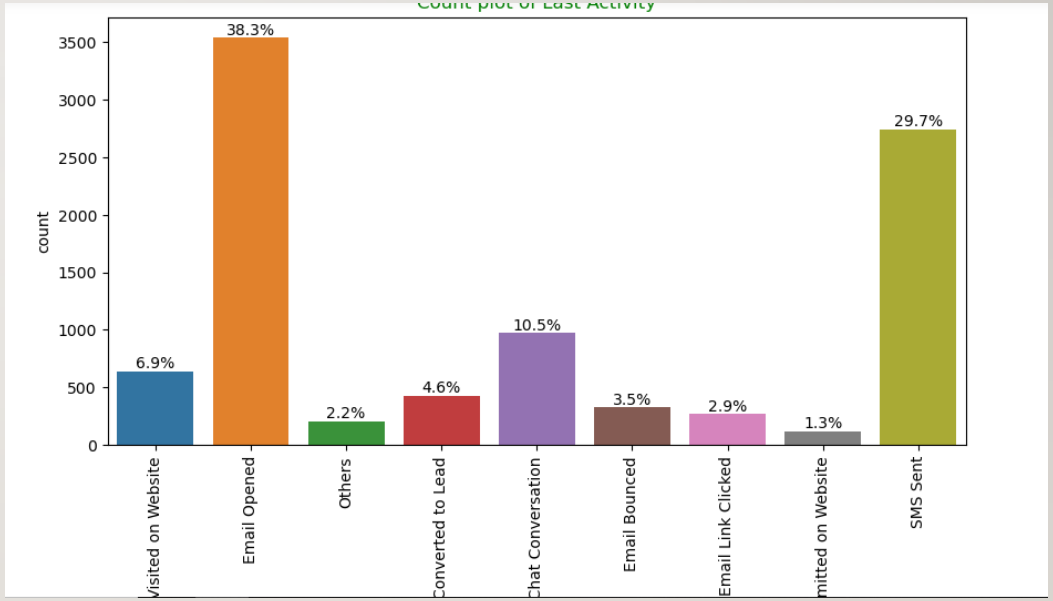
- **Univariate Analysis for Categorical Variables**
- 3. Do Not Email: 93% of the people has opted that they don't want to be emailed about the course.



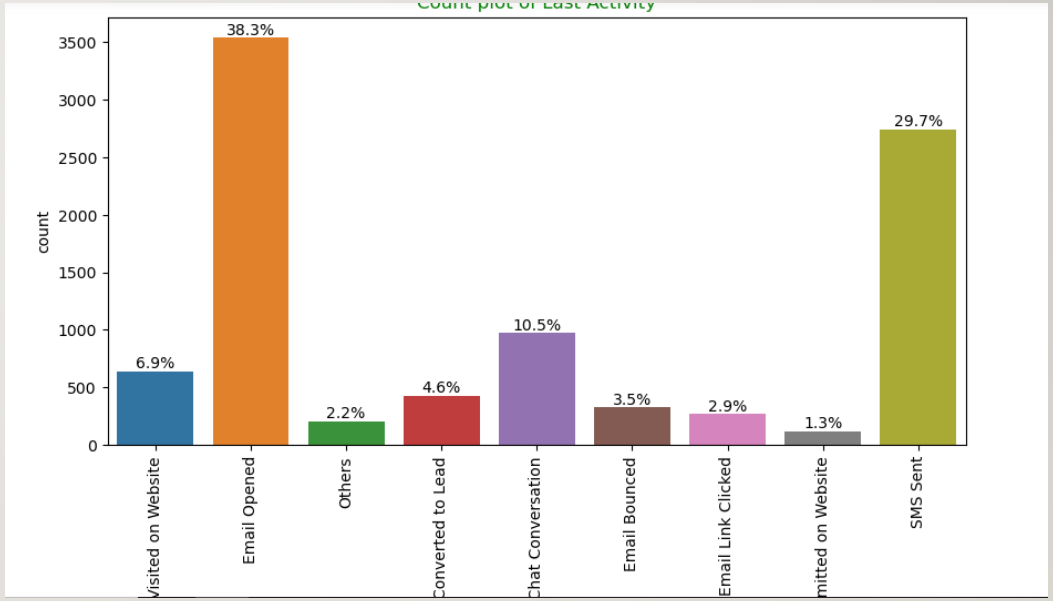
- **Univariate Analysis for Categorical Variables**
- 4. Lead Source: 60% Lead source is from Google & Direct Traffic combined



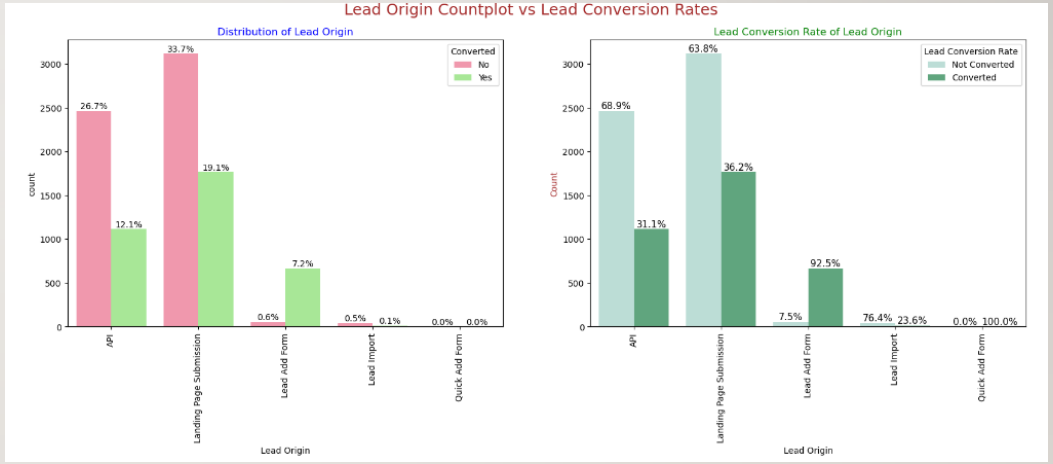
- **Univariate Analysis for Categorical Variables**
- 5.Last Activity: 67% of customers contribution in SMS Sent & Email Opened activities



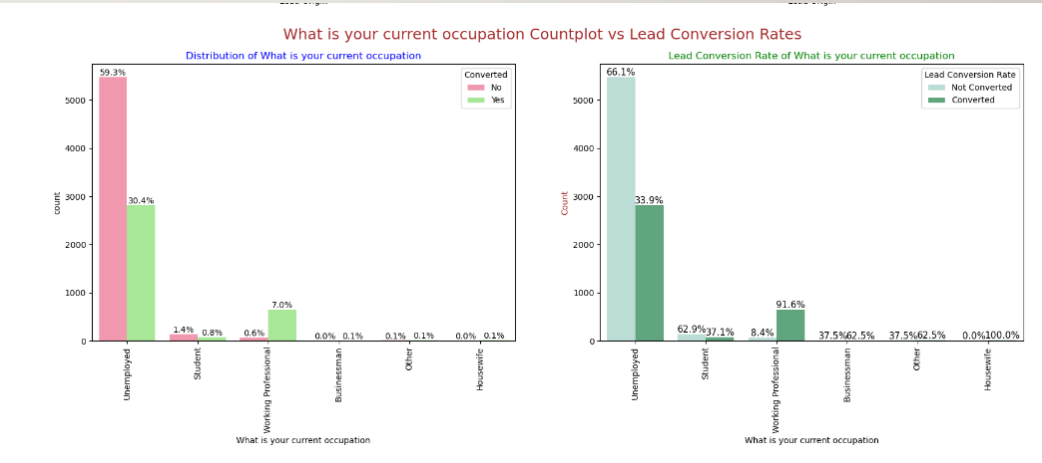
- **Univariate Analysis for Categorical Variables**
- 6.Last Activity: 67% of customers contribution in SMS Sent & Email Opened activities



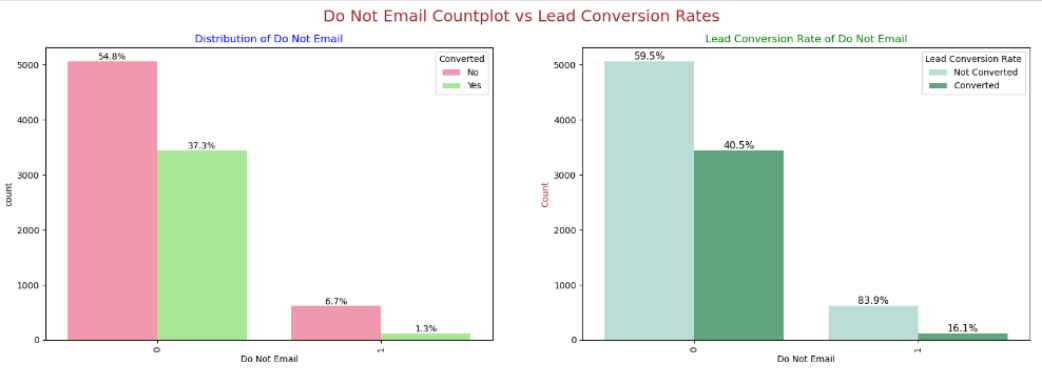
- **Bivariate Analysis for Categorical Variables**
- 1. Lead Origin: Around 52% of all leads originated from "Landing Page Submission" with a lead conversion rate (LCR) of 36%. The "API" identified approximately 39% of customers with a lead conversion rate (LCR) of 31%.



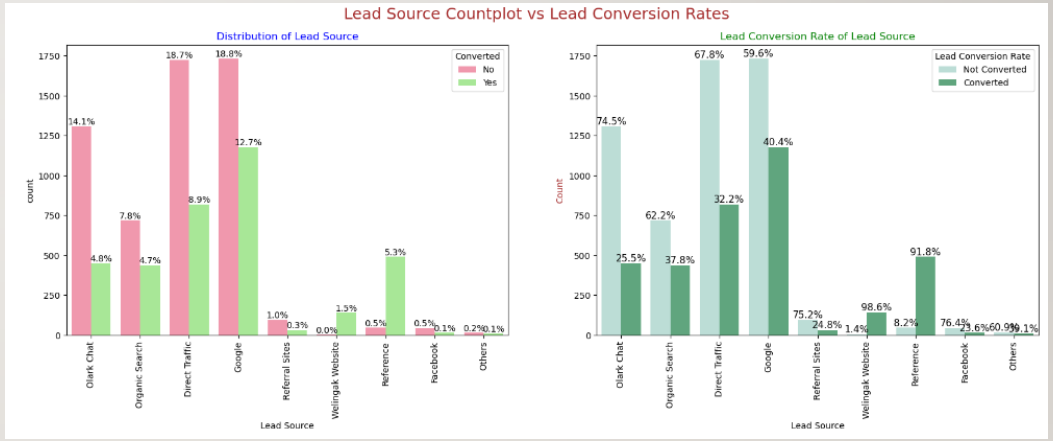
- **Bivariate Analysis for Categorical Variables**
- 2. Current Occupation: Around 90% of the customers are Unemployed with lead conversion rate (LCR) of 34%. While Working Professional contribute only 7.6% of total customers with almost 92% lead conversion rate (LCR).



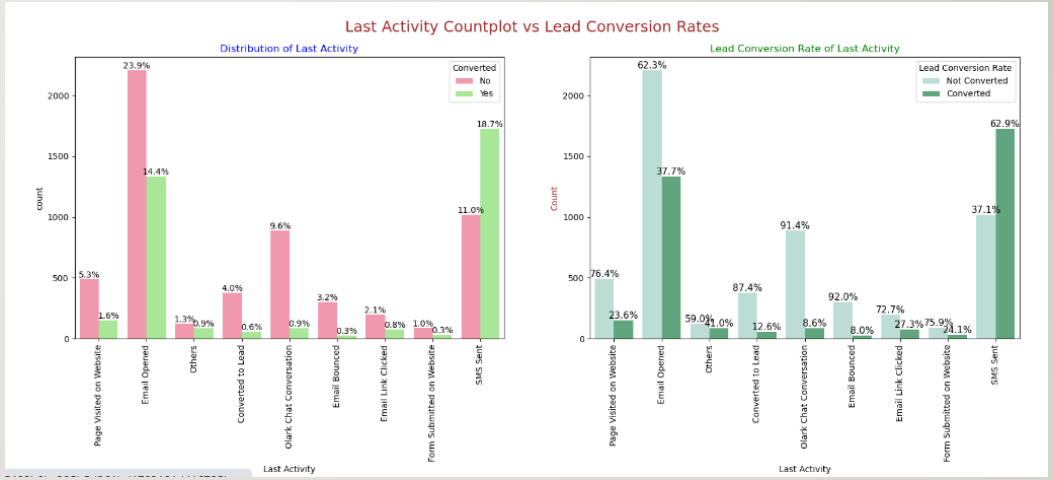
- **Bivariate Analysis for Categorical Variables**
- 3. Do Not Email: 92% of the people has opted that they don't want to be emailed about the course.



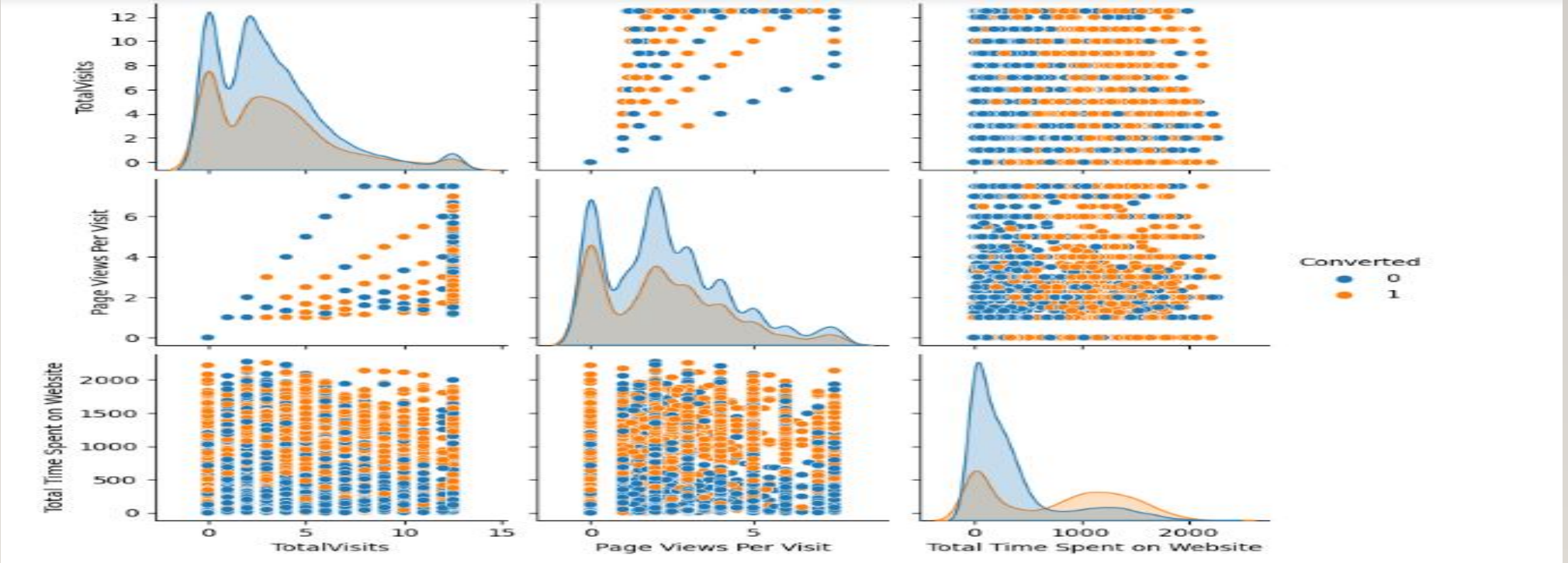
- **Bivariate Analysis for Categorical Variables**
- 4. Lead Source: Google has LCR of 40% out of 31% customers , Direct Traffic contributes 32% LCR with 27% customers which is lower than Google, Organic Search also gives 37.8% of LCR but the contribution is by only 12.5% of customers ,Reference has LCR of 91% but there are only around 6% of customers through this Lead Source.



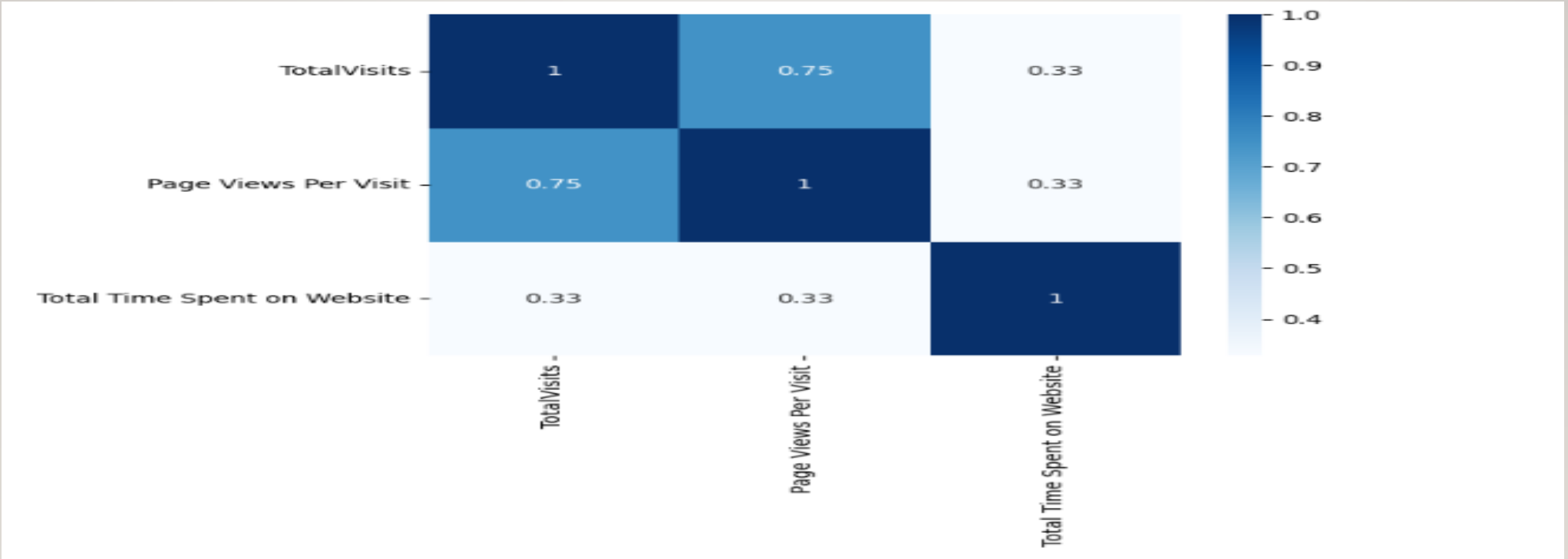
- **Bivariate Analysis for Categorical Variables**
- **5.** Last Activity: 'SMS Sent' has high lead conversion rate of 63% with 30% contribution from last activities, 'Email Opened' activity contributed 38% of last activities performed by the customers with 37% lead conversion rate..



- **Bivariate Analysis for Numerical Variables**



- **Heatmap**


TotalVisits
Page Views Per Visit
Total Time Spent on Website
1
0.75
0.33
0.75
1
0.33
0.33
0.33
1
TotalVisits
Page Views Per Visit
Total Time Spent on Website
1.0
0.9
0.8
0.7
0.6
0.5
0.4

MODEL BUILDING

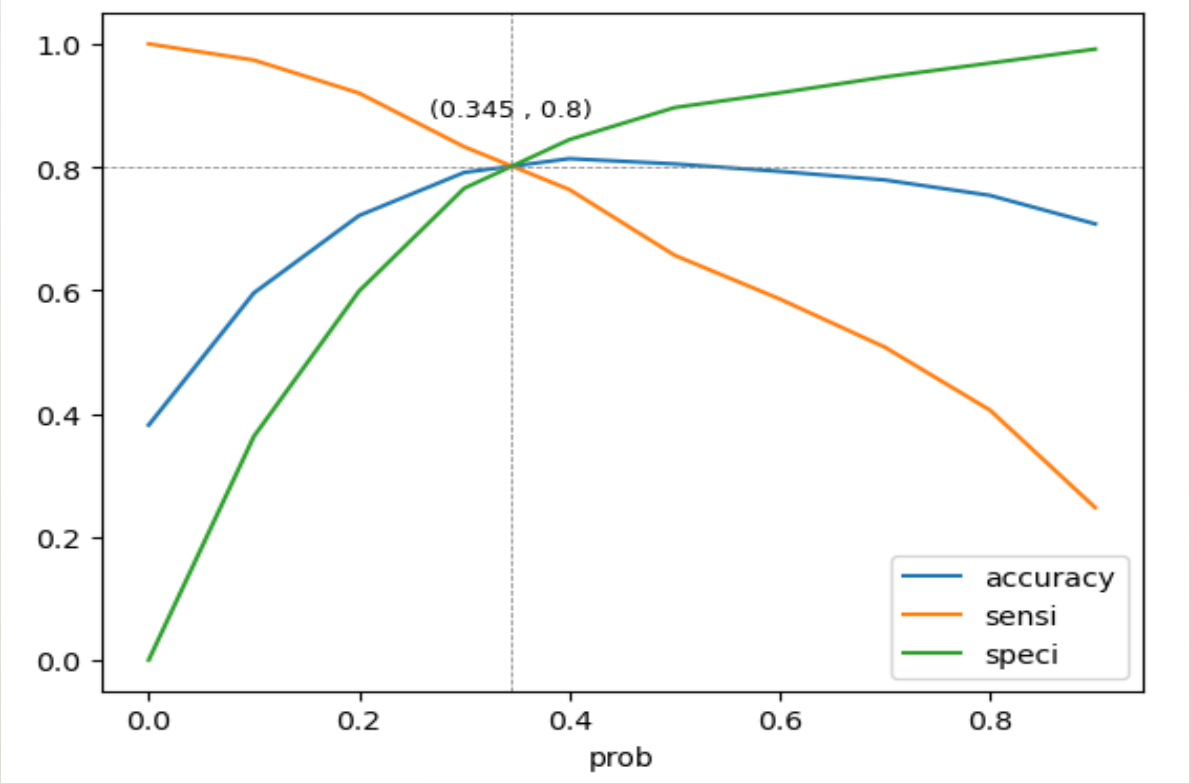
- Splitting into train and test set
- Scale variables in train set
- Build the first model
- Use RFE to eliminate less relevant variables
- Build the next model
- Eliminate variables based on high p-values
- Check VIF value for all the existing columns
- Predict using train set
- Evaluate accuracy and other metric
- Predict using test set
- Precision and recall analysis on test predictions

MODEL EVALUATION (TRAIN)

- **Accuracy Sensitivity and Specificity**
- Confusion Matrix

3230	772
492	1974

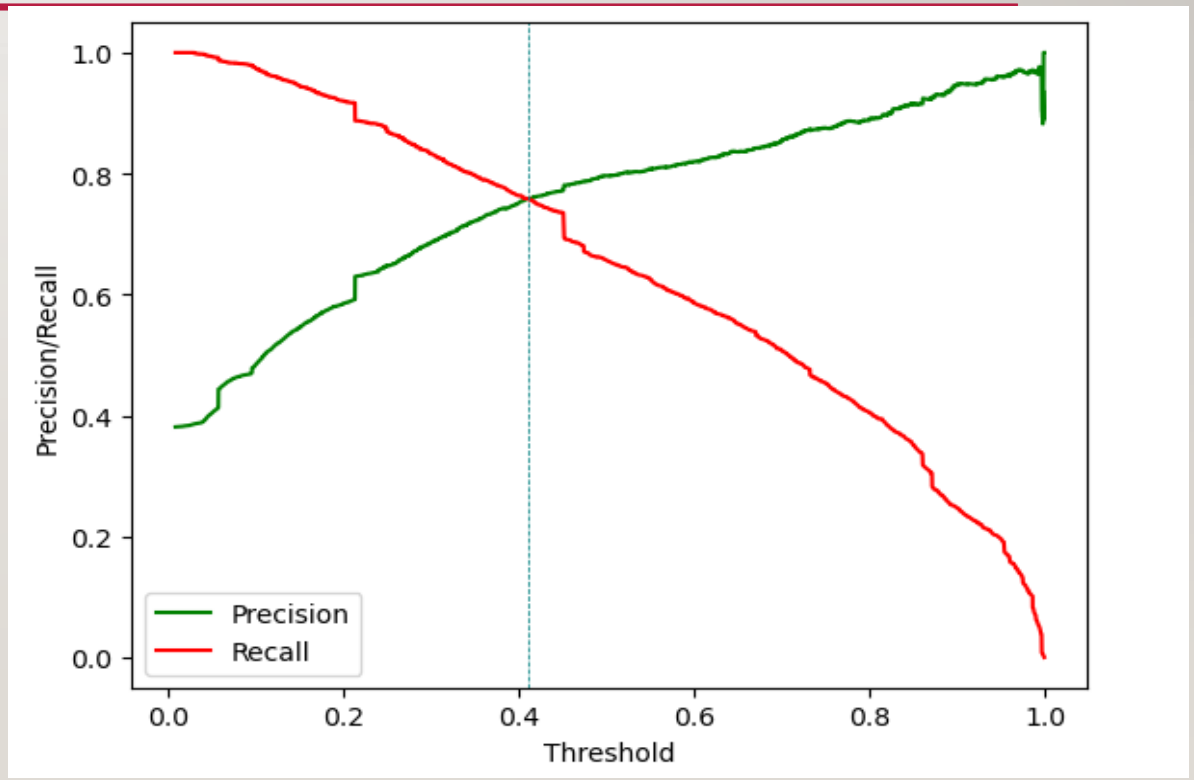
- Accuracy - 80%
- Sensitivity - 80 %
- Specificity - 80 %



- **Precision and Recall**
- Confusion Matrix

3406	596
596	1870

- Precision – 75.83%
- Recall – 75.83%

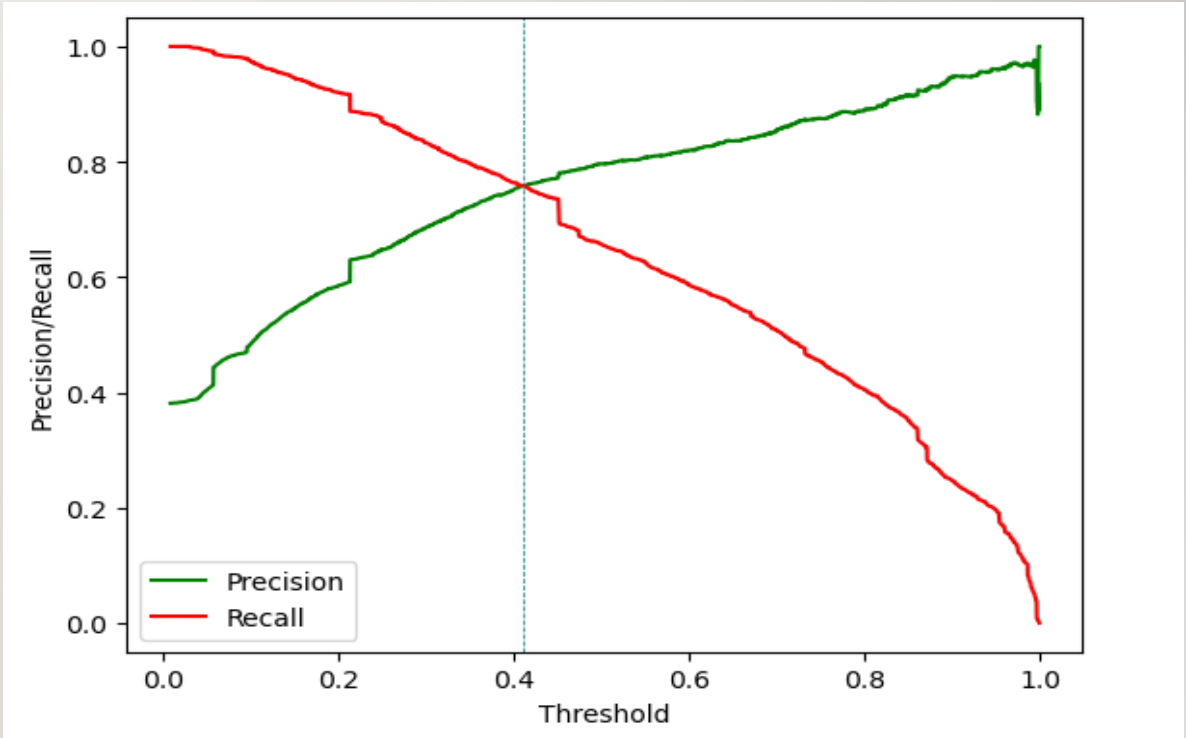


MODEL EVALUATION (TEST)

- Confusion Matrix

1353	324
221	874

- Accuracy - 80%
- Sensitivity - 79 %
- Specificity - 80 %



CONCLUSION

- The model shows high close to 80% accuracy
- The threshold has been selected from Accuracy, Specificity, specificity measures and precision, recall curves.
- The model shows 80% of sensitivity and specificity
- The model finds correct promising leads and lead that have less chances of getting converted
- Overall, this model proves to be accurate

THANK YOU

BY – RAKSHITHA KULAL

VANITA RATNANI

RAKHI RIJHWANI