LEAD SCORING CASE STUDY

PROBLEM STATEMENT

- An education company named X Education sells online courses to industry professionals
- The company markets its course on several and search engine like Google.
- When any customer fill up form and providing their details. Those customers classified as leads. Then Sales teams reach out to leads via calls, emails etc.
- Although company generated lot of leads only a few of them converted into their paying user.
- The company has 30% conversion rate through the whole process of turning leads into customers by approaching those leads which are to be found having interest in taking the course. The implementation process of lead generating attributes are not efficient in helping conversation.
- The CEO aims for an 80% lead conversion rate. A dataset with 9000 data points, including attributes like Lead Source, Total Time Spent on Website, and Last Activity, is provided.
- The target variable is 'Converted' (1 for converted, 0 for not converted). Handling 'Select' levels in categorical variables is also crucial.

BUSINESS OBJECTIVE

- The company requires a model to be built for selecting most promising leads.
- A higher lead score indicates a 'hot' lead with a higher likelihood of conversion, while a lower score signifies a 'cold' lead less likely to convert.
- Additionally, the model should exhibit adaptability to address future changes in the company's requirements and challenges.
- These potential adjustments are documented separately and will be incorporated into the logistic regression model's recommendations, ensuring ongoing effectiveness and relevance in lead conversion strategies.

STEPS OF ANALYSIS

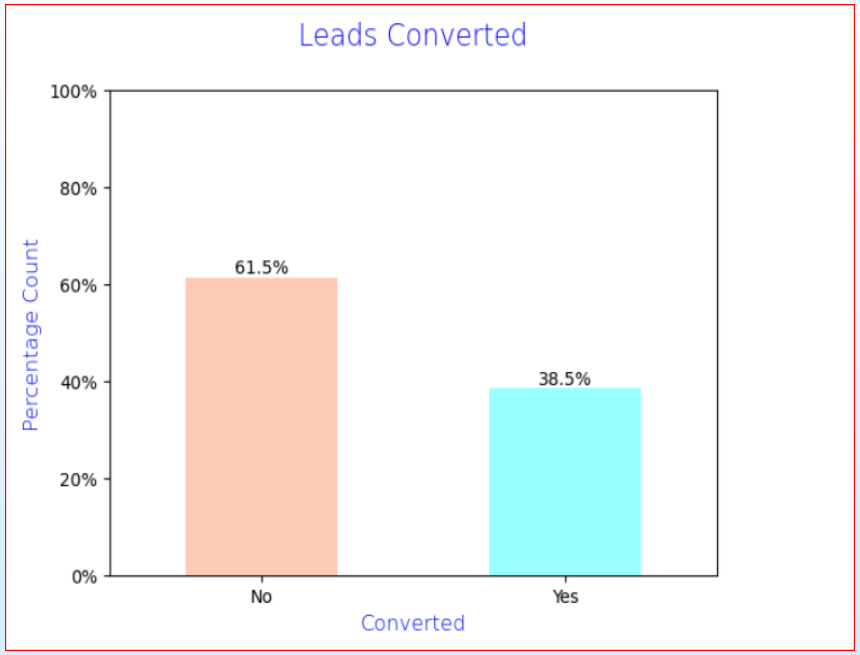
- DATA IMPORTING AND CLEANING
- EXPLORATORY DATA ANALYSIS
- DATA PREPARATION
- MODEL BUILDING AND EVALUATION
- MAKING PREDICTIONS ON TEST DATASET

DATA CLEANING

- Columns with over 40% null values were dropped
- Missing categorical values were addressed using value counts and careful consideration. Numerical data was imputed with mode after checking distribution
- “Select” value represent as null values for some categorical variables, as customers did not choose any option
- Imputation was used for some columns
- Columns with no use for modeling (Prospect ID, Lead Number) or only one category of response were dropped
- Skewed category columns were checked and dropped to avoid bias in logistic regression models.
- Other cleaning activities were performed to ensure data quality and accuracy. Fixed Invalid values & Standardizing Data in columns by checking casing styles, etc. (lead source has Google, google

EXPLORATORY DATA ANALYSIS

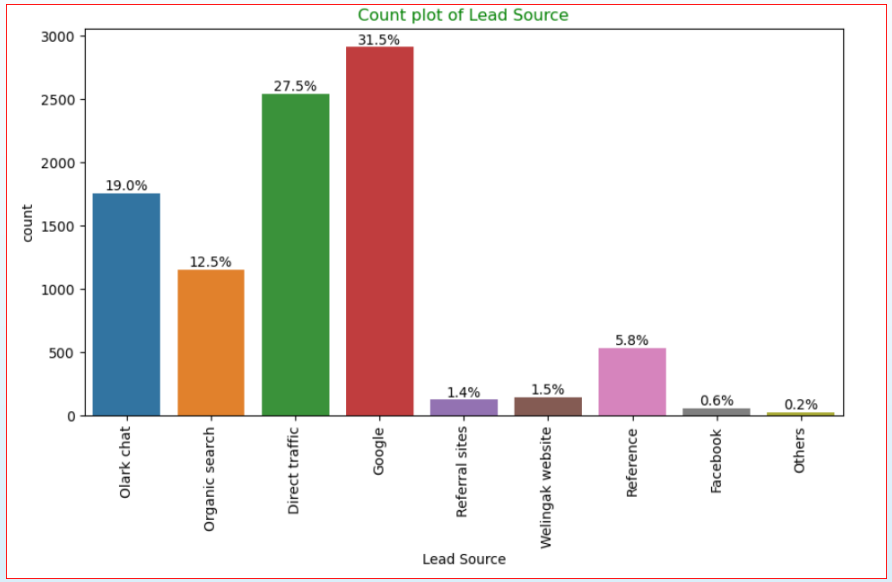
- Data is imbalanced while analyzing target variable



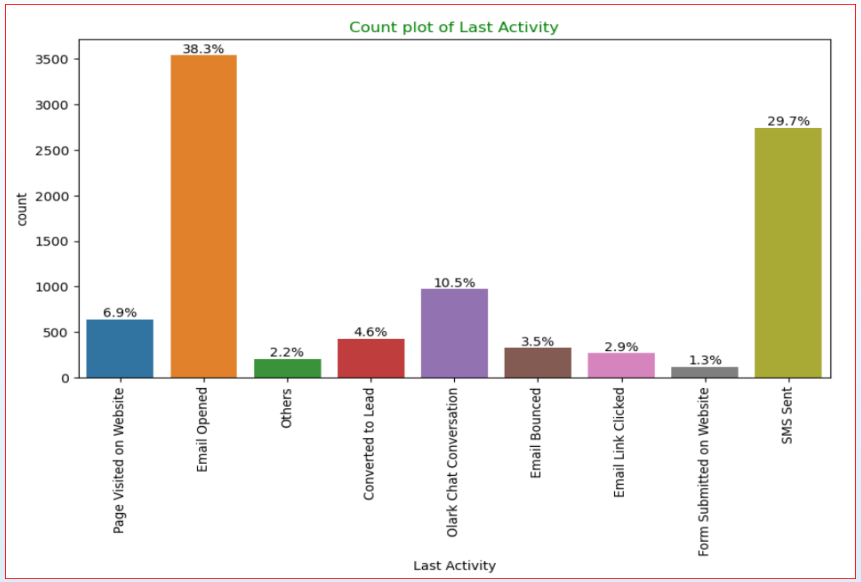
- Conversion rate is 38.5%. Its means only the 38.5% of people converted into leads which is minority
- While 61.5% of people didn't converted into leads

EXPLORATORY DATA ANALYSIS

Univariate Analysis (Categorical Variables)



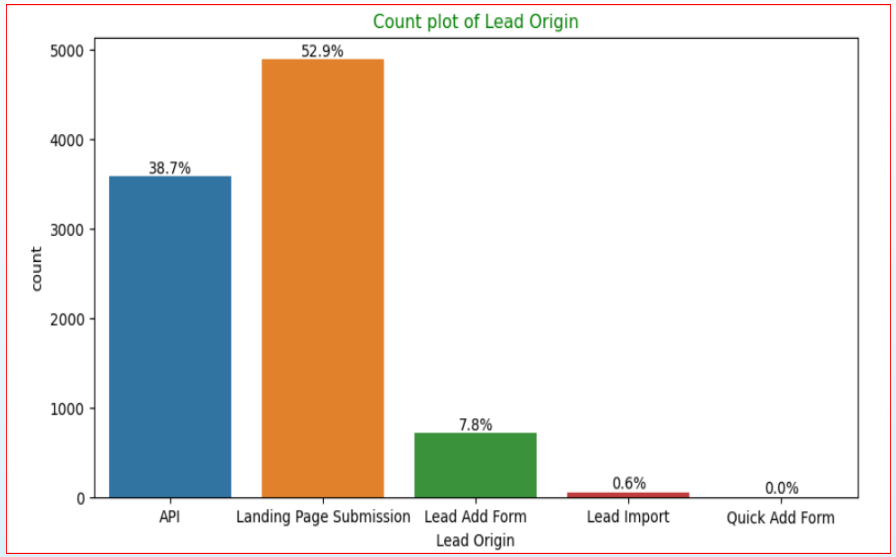
Lead Source: 78% Lead source is from Olark chat, Google & Direct Traffic combined

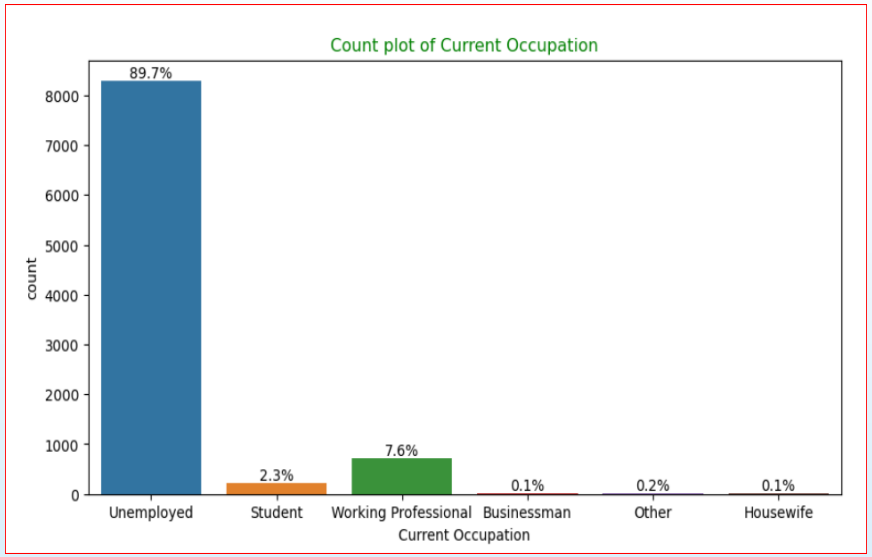


Last Activity: 68% of customer contribution in SMS Sent & Email Opened activities

EXPLORATORY DATA ANALYSIS

Univariate Analysis (Categorical Variables)



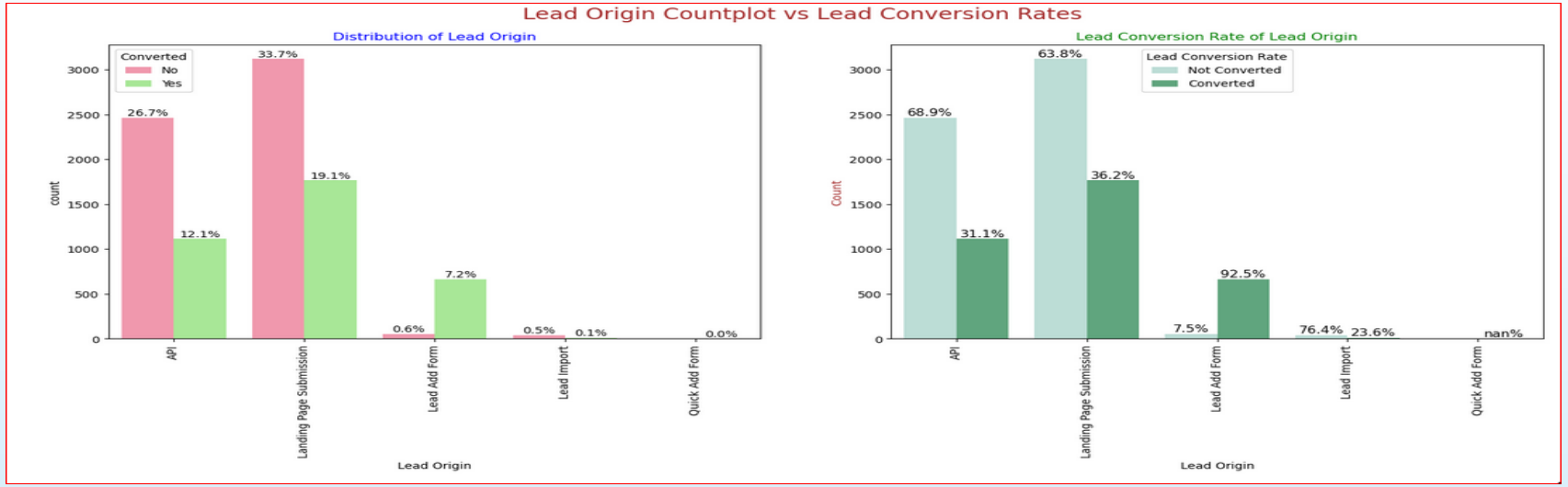


Lead Origin: "Landing Page Submission" identified 53% of customers,
"API" identified 39%

Current Occupation : It has 90% of the customers as Unemployed

EXPLORATORY DATA ANALYSIS

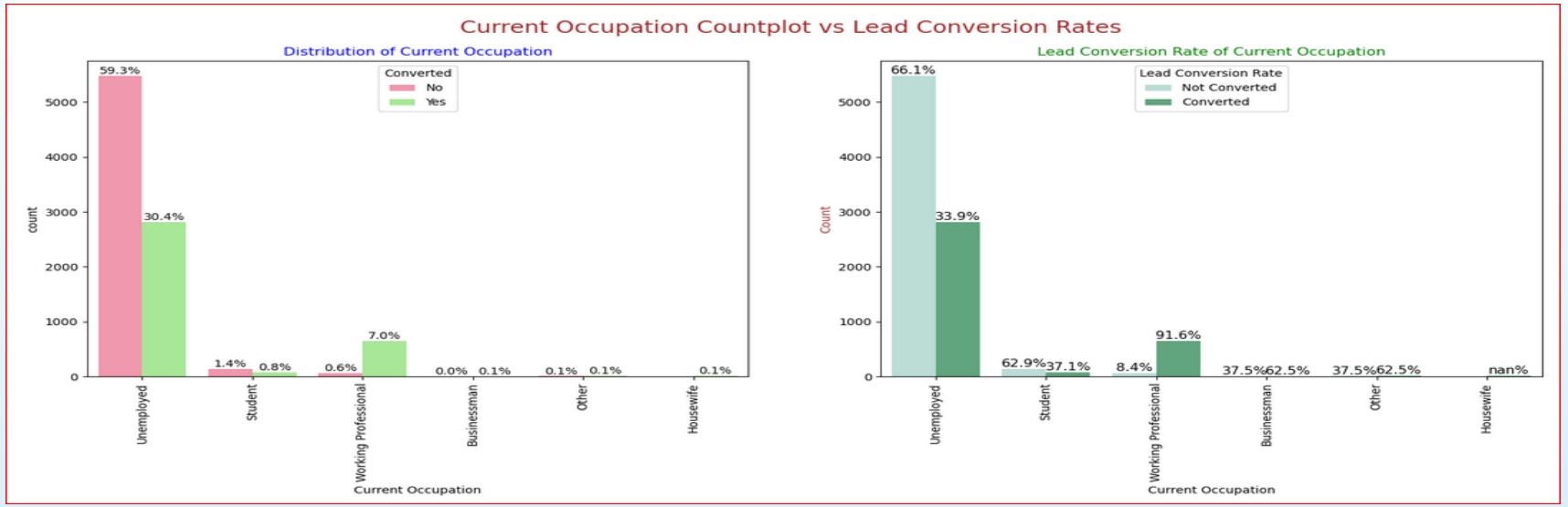
BIVARIATE ANALYSIS – CATEGORICAL VARIABLES



Lead Origin: Around 52% of all leads originated from "Landing page Submission" with a lead conversion rate of 36%
The "API" identified approximately 39% of customers with a lead conversion rate of 31%

EXPLORATORY DATA ANALYSIS

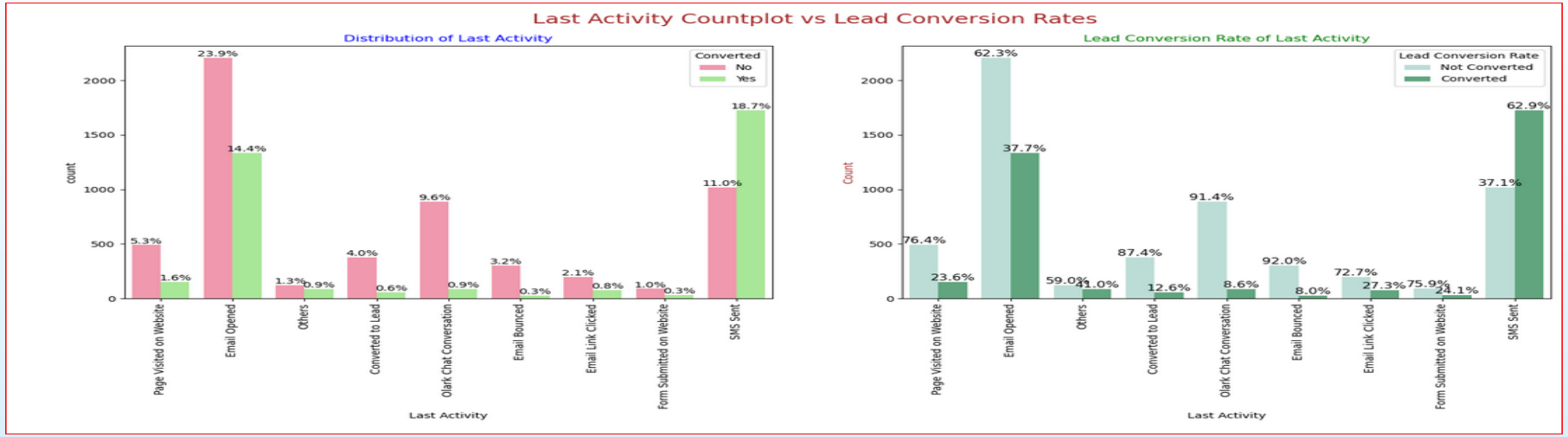
BIVARIATE ANALYSIS – CATEGORICAL VARIABLES



Current Occupation : Around 90% of the customers are Unemployed, with lead conversion rate 34%. While Working Professional contribute only 7.6% of total customers with almost 92% Lead Conversion rate.

EXPLORATORY DATA ANALYSIS

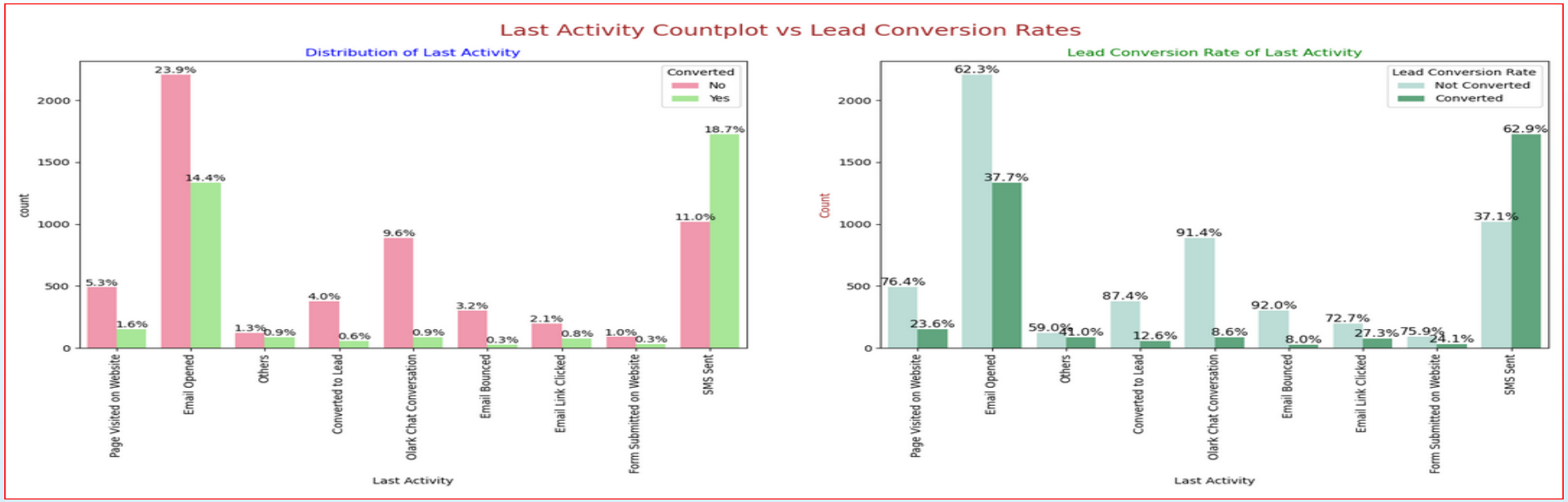
BIVARIATE ANALYSIS – CATEGORICAL VARIABLES



Last Activity: 'SMS Sent' has high lead conversion rate of 63% with 30% contribution

EXPLORATORY DATA ANALYSIS

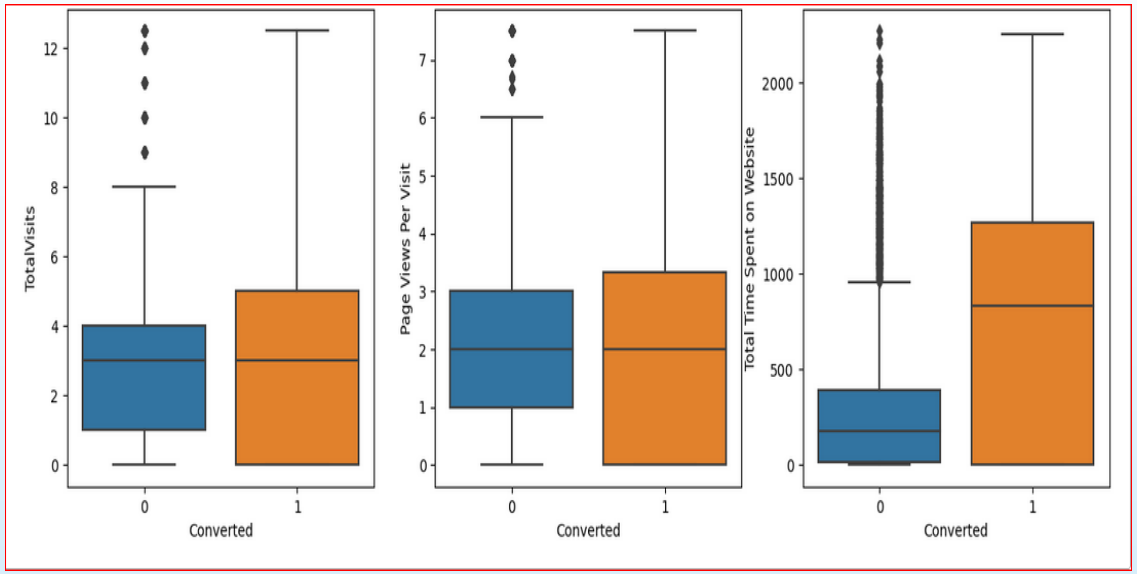
BIVARIATE ANALYSIS – CATEGORICAL VARIABLES



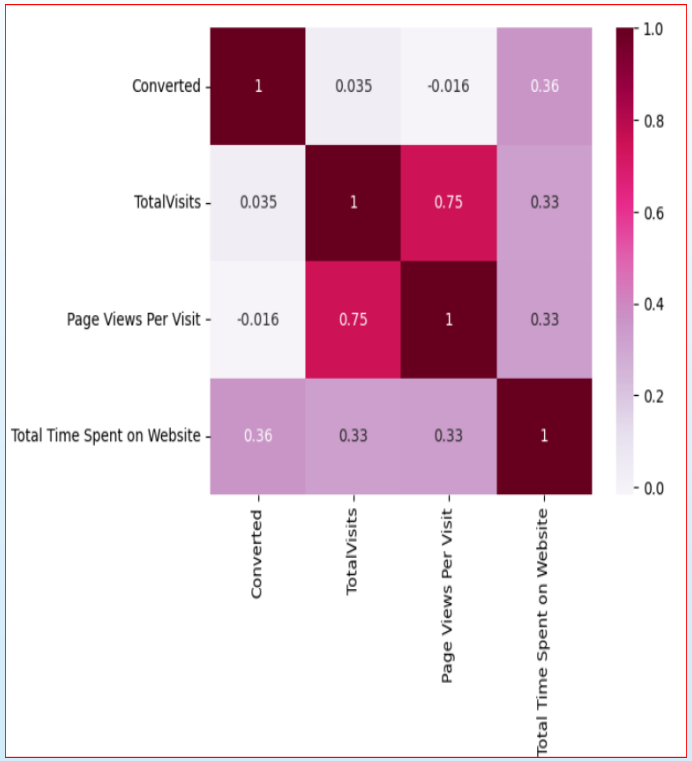
Specialization: Marketing Management, HR Management, Finance management shows good contribution in Leads conversion than other specialization.

EXPLORATORY DATA ANALYSIS

BIVARIATE ANALYSIS – NUMERICAL VARIABLES



Past Leads who **spends more time on the websites** have a higher chance of getting successfully converted than those who spends less time as seen in the box-plot



DATA PREPARATION FOR MODEL BUILDING

- Binary categorical columns were already mapped to 1/0
- Created dummy features (one-hot encoded) for categorical variables – Lead Origin, Lead Source, Last Activity, Specialization, Current occupation
- Splitting Train & Test Sets - 70:30 % ratio was chosen for the split
- Feature scaling - Standardization method was used to scale the features
- Checking the correlations - Predictor variables which were highly correlated with each other were dropped (Lead Origin Lead Import and Lead Origin Lead Add Form)

MODEL BUILDING

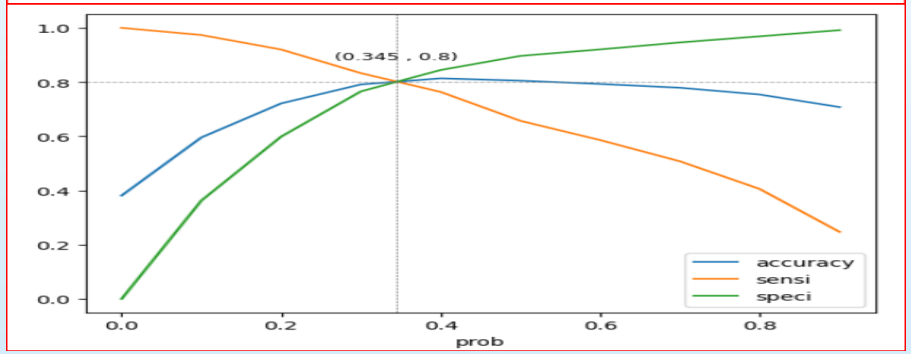
Feature Selection

- The data set has lots of dimension and large number of features.
- Hence it is important to perform Recursive Feature Elimination (RFE) and to select only the important columns. Then we can manually fine tune the model.
- RFE outcome - Pre RFE – 48 columns & Post RFE – 15 columns
- Manual Feature Reduction process was used to build models by dropping variables with p – value greater than 0.05.
- Model 4 looks stable after four iteration with: significant p-values within the threshold (p-values < 0.05) and No sign of multicollinearity with VIFs less than 5
- Model 4 looks stable after four iteration with: significant p-values within the threshold (p-values < 0.05) and No sign of multicollinearity with VIFs less than 5

MODEL EVALUATION

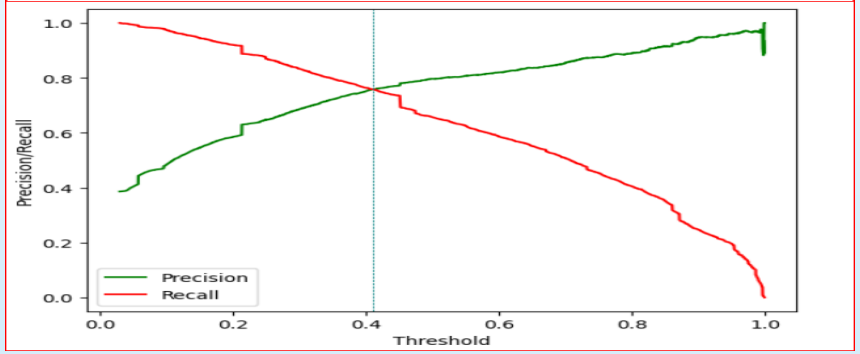
TRAIN DATA SET - It was decided to go ahead with 0.345 as cutoff after checking evaluation metrics coming from both plots

```
Confusion Matrix
[[3230  772]
 [ 492 1974]]

**********************************************************

True Negative                          :   3230
True Positive                          :   1974
False Negative                         :   492
False Positve                          :   772
Model Accuracy                         :   0.8046
Model Sensitivity                      :   0.8005
Model Specificity                      :   0.8071
Model Precision                        :   0.7189
Model Recall                           :   0.8005
Model True Positive Rate (TPR)         :   0.8005
Model False Positive Rate (FPR)        :   0.1929
```



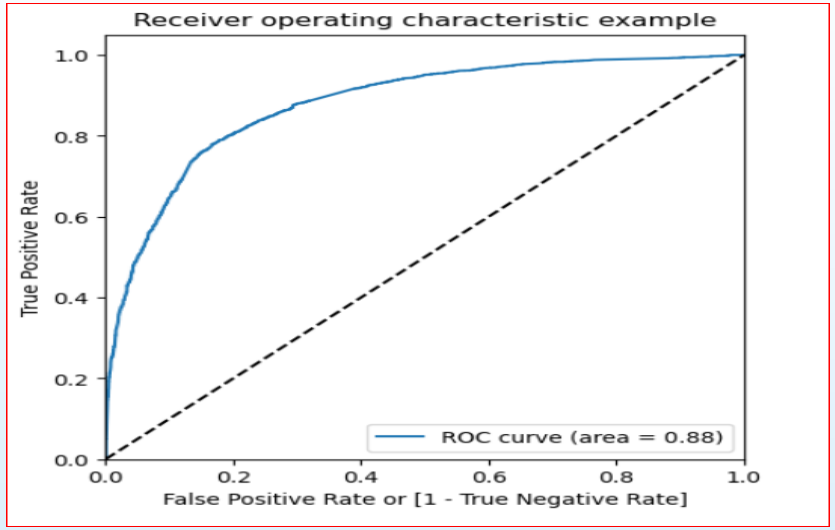
Confusion Matrix & Evaluation Metrics with 0.345 as cutoff

```
Confusion Matrix
[[3406  596]
 [ 596 1870]]

**********************************************************

True Negative                          :   3406
True Positive                          :   1870
False Negative                         :   596
False Positve                          :   596
Model Accuracy                         :   0.8157
Model Sensitivity                      :   0.7583
Model Specificity                      :   0.8511
Model Precision                        :   0.7583
Model Recall                           :   0.7583
Model True Positive Rate (TPR)         :   0.7583
Model False Positive Rate (FPR)        :   0.1489
```

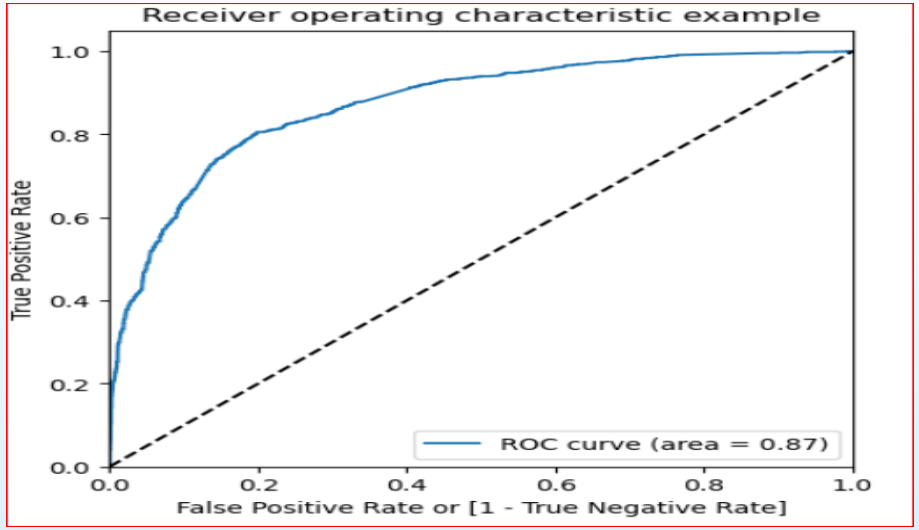


Confusion Matrix & Evaluation Metrics with 0.41 as cutoff

MODEL EVALUATION



- Area under ROC curve is 0.88 out of 1 which Indicates a good predictive model.
- The curve is as close to the top left corner of the plot, which represents a model that has a high true positive rate and a low false positive rate at all threshold values.



- Area under ROC curve is 0.87 out of 1 which indicates a good predictive model
- The curve is so close to the top left corner of the plot, which represents a model that has a high true positive rate and a low false positive rate and a low false positive rate at all threshold values.

Recommendation

- Prioritize leads originating from Lead Add Forms for targeting.
- Concentrate efforts on leads tagged as "Will revert after reading the email.“
- Give preference to leads with the last activity recorded as "SMS Sent.“
- Explore the option of using a lower cut-off threshold to aggressively target and capture more "Hot Leads.“
- Engage working professionals with personalized messaging.
- Offer incentives/discounts for successful referrals to encourage more references

THANK YOU