Lead Scoring
Case Study

Submitted by-

ANKITA SAHA RAY & VAIBHAV BHARGAVA

(DS-C42)

Problem statement

- **'X Education'** is an EdTech company which sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

- The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

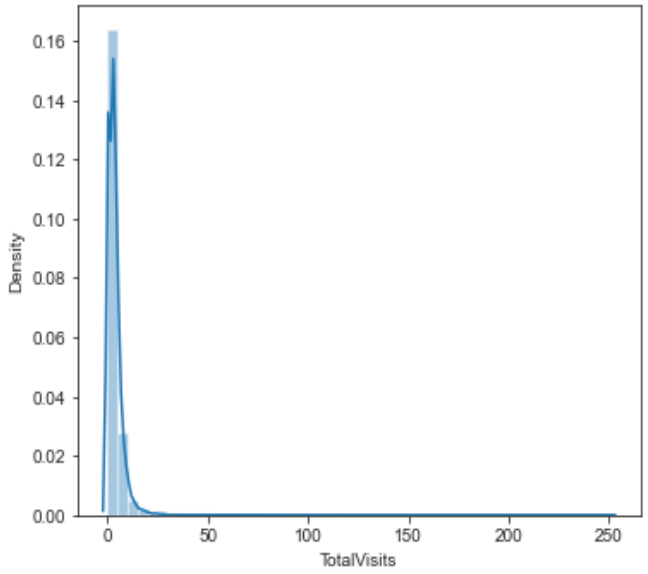
🎯 Goal of the Case Study

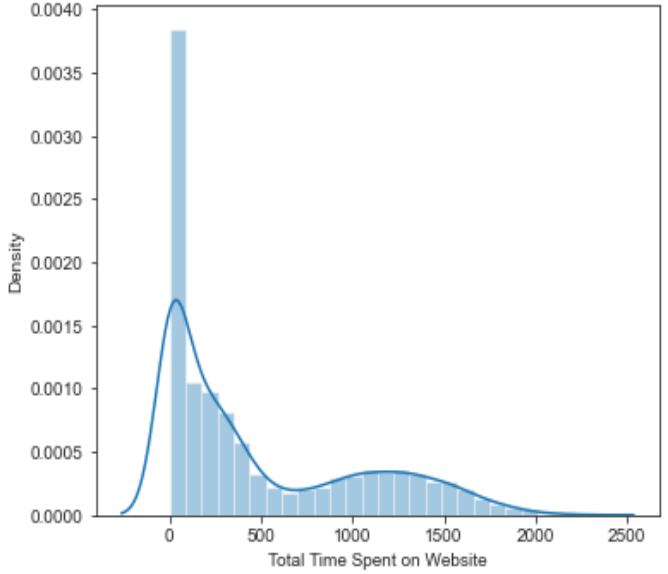
- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

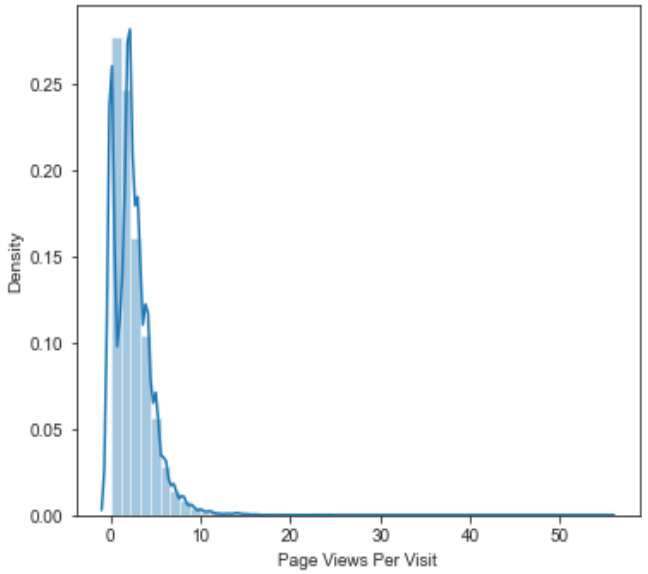
Approach for the Case Study

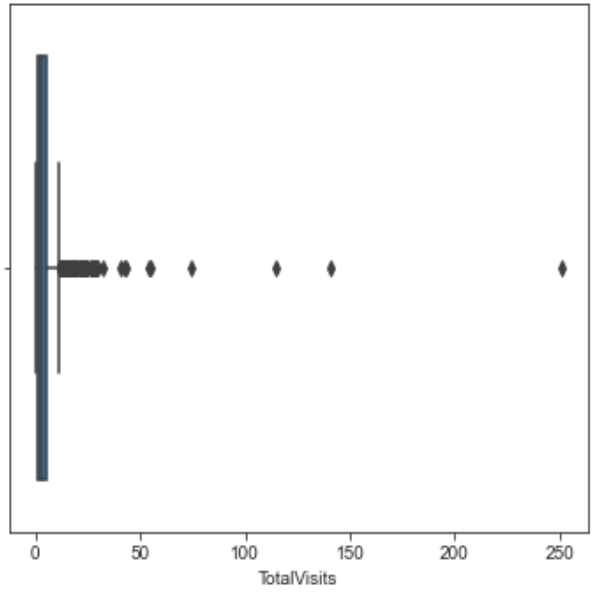
1. • **Importing** “Leads” Dataset
2. • **Understanding** basic structure of the data
3. • **Cleaning the data** from discrepancies like missing values, unique values, and ‘Select’ which is as good as a null value
4. • Performing **Exploratory Data Analysis** on numerical and categorical columns separately with respect to the target variable
5. • **Pre-processing data** for Model Building which includes conversion of Yes/No values to 1/0 and creation of dummy variables
6. • **Preparing data for modelling** by splitting into 70:30 Train and Test ratio and **rescaling** the features using **Standard Scaler**
7. • Creating a **Logistic Regression model** using a mixed approach of **RFE** and manually dropping feature variables
8. • **Predicting values of Training Dataset** using the final model with an initial cut-off of 0.5
9. • Plotting **ROC** curve to find area under curve **(AUC),** evaluating optimal cut-off point based on **‘Precision’** and **‘Recall’** trade-off
10. • Plotting **confusion matrix** to evaluate all important metrics on the **Training** data
11. • Predicting the **Test data** with previously obtained optimum cut-off
12. • Evaluating all important metrics like- **Accuracy, Sensitivity, Specificity, Precision, Recall and F1-score** of the **Test** data

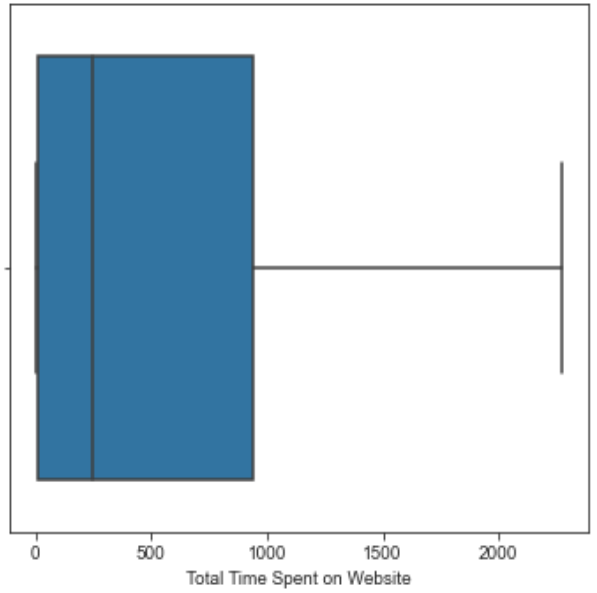
Univariate Analysis - Continuous Variables

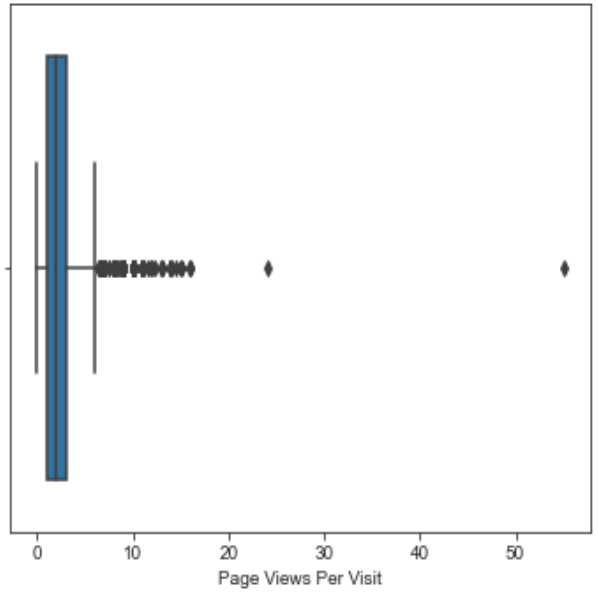












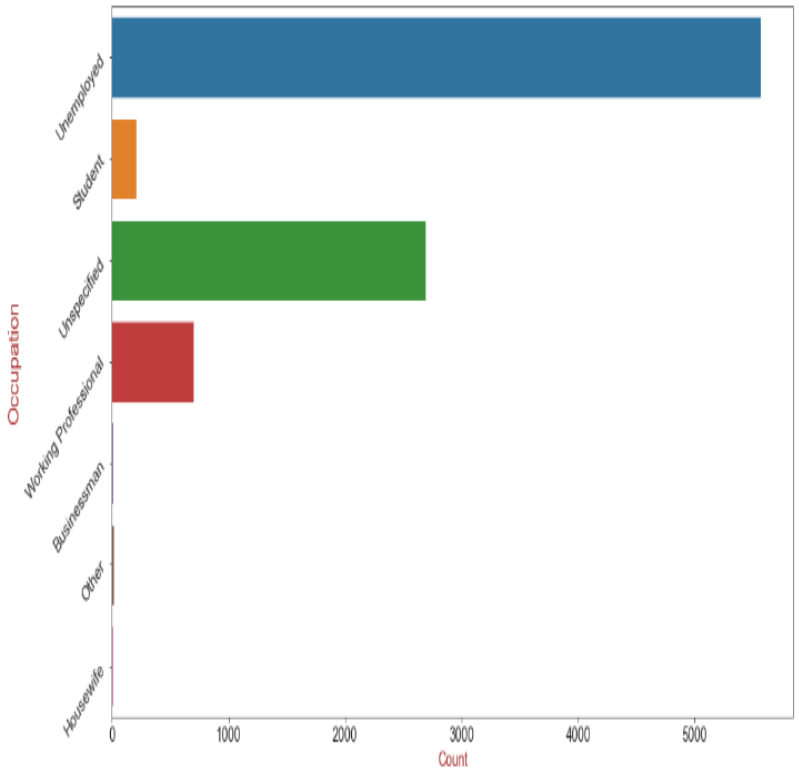
❑ **Observations:**

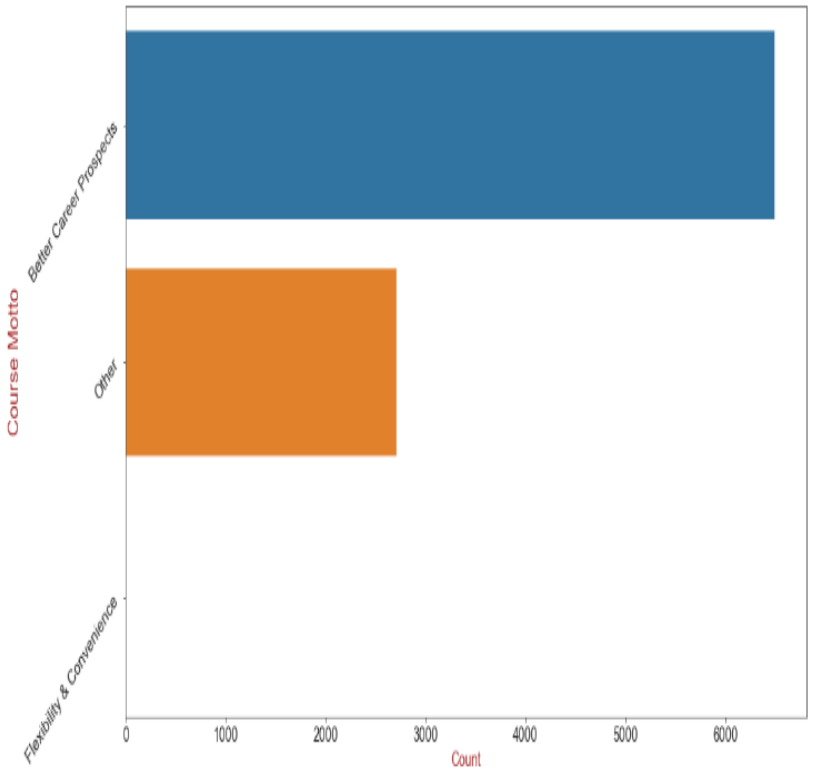
1. The total number of visits made by the customer on the website is mostly between 0-30.

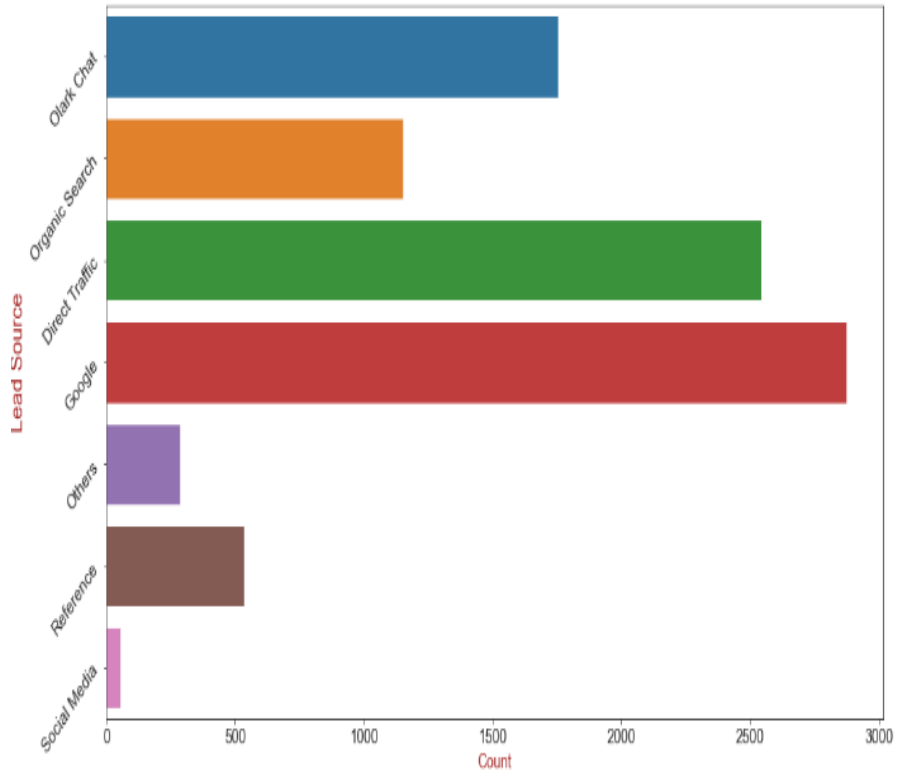
2. The total time spent by the customer on the website is high in the beginning with a gradual decrease, then again rising slightly. It is possible that customers who are likely to convert are re-visiting the website.

3. Page views per visit lies mostly in the range of 0-10.
4. Outliers are present but it is not significantly high as per distribution of data.

Univariate Analysis-Categorical Variables



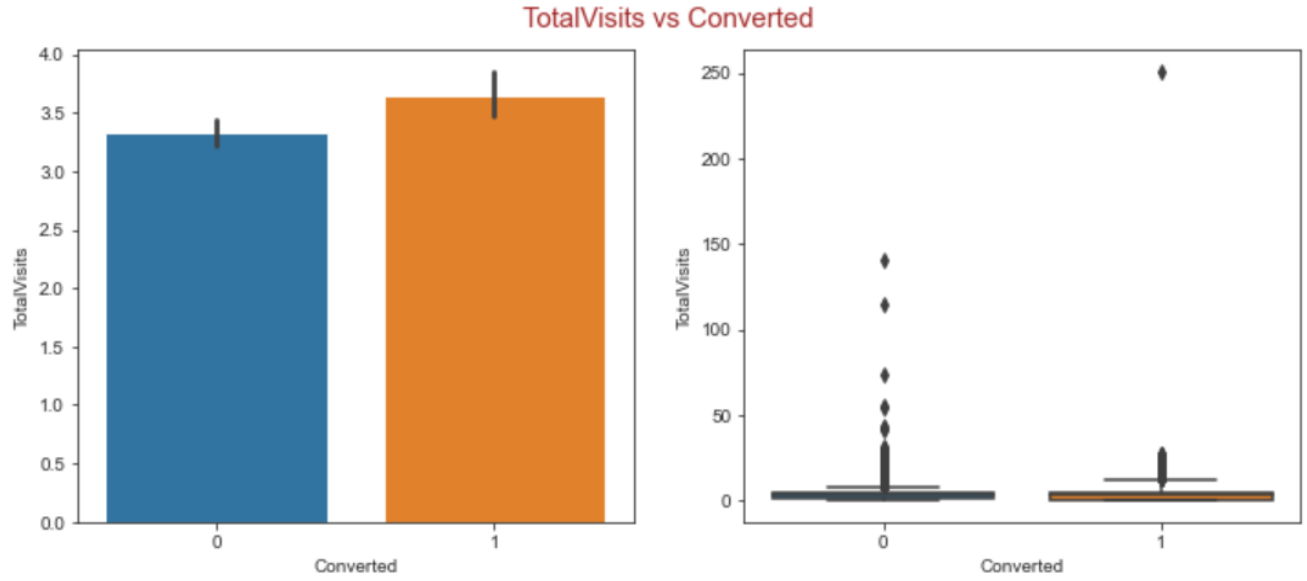


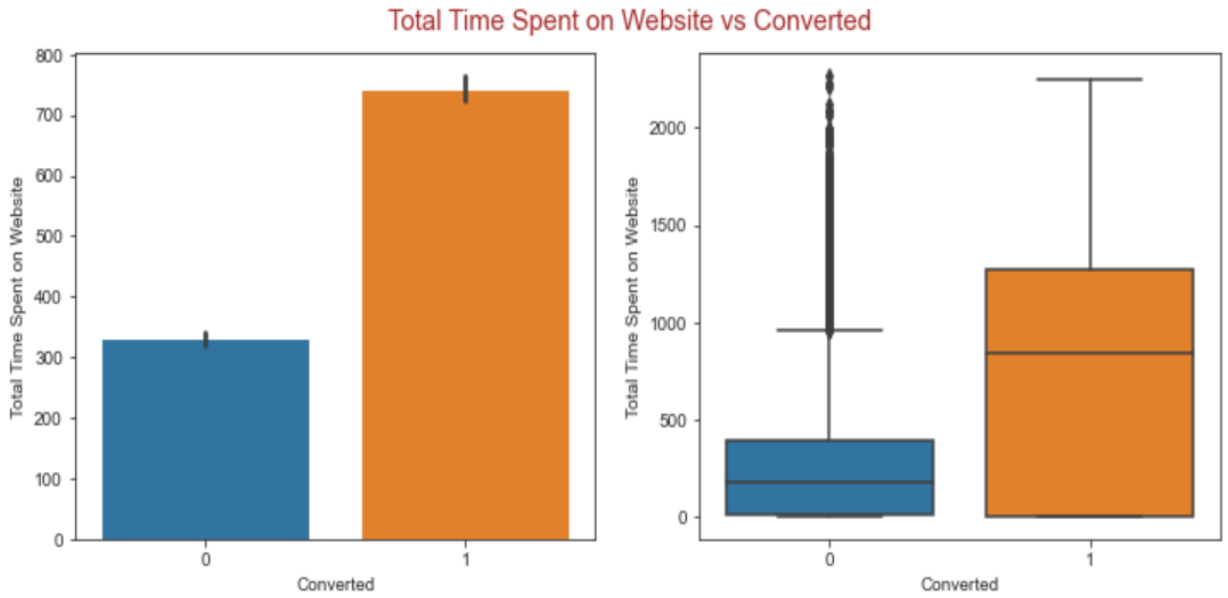


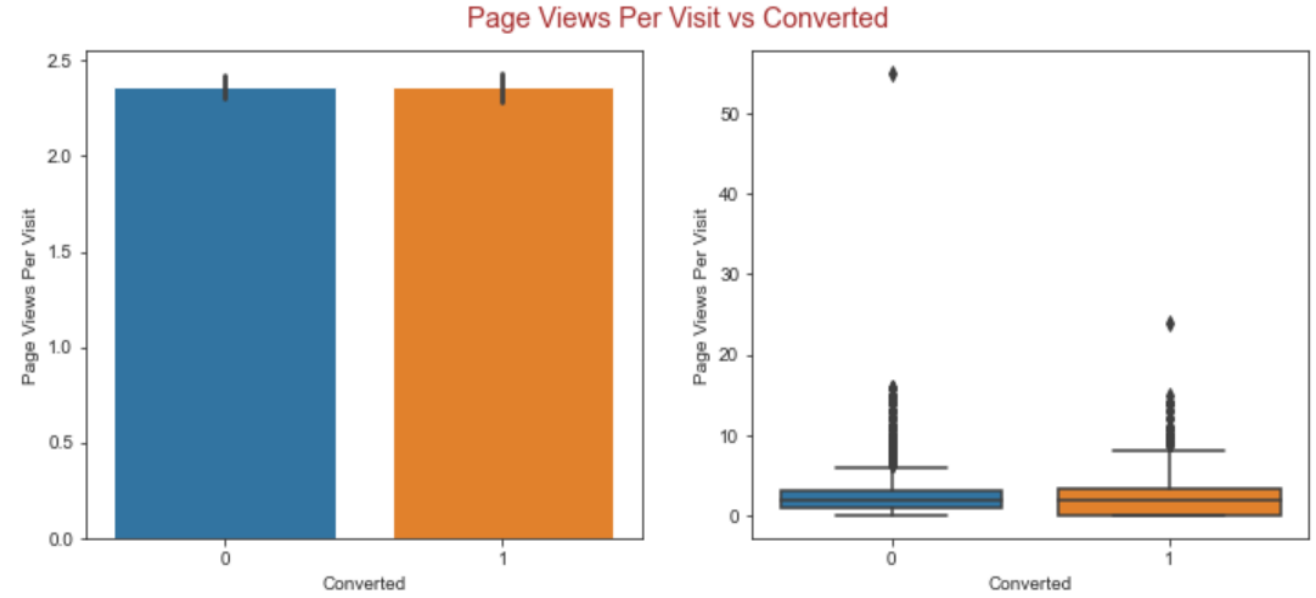
❑ **Observations:**

1. Google as the source of the lead is higher in number, followed by direct traffic.
2. Unemployed people are more in number among the pool of leads.
3. Most people look for a course for getting better career prospects.

Bivariate Analysis - Continuous Variables







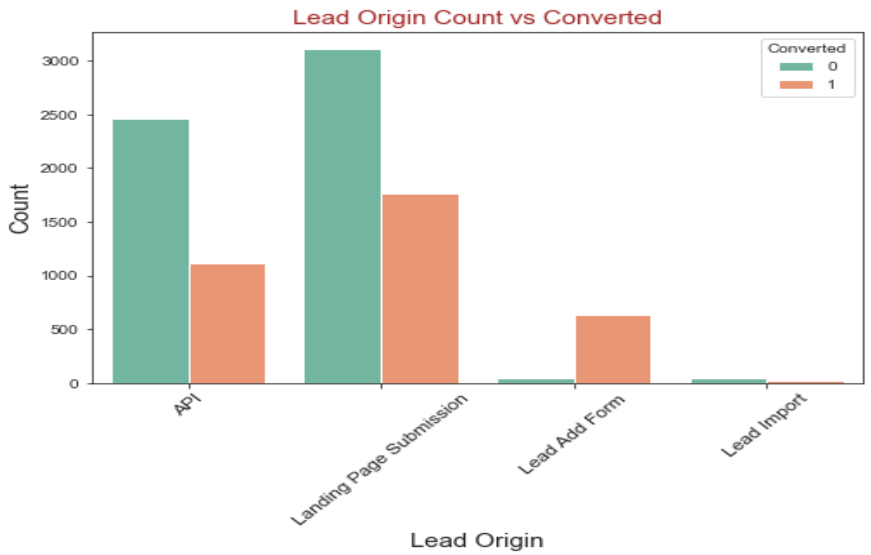
❑ **Observations:**

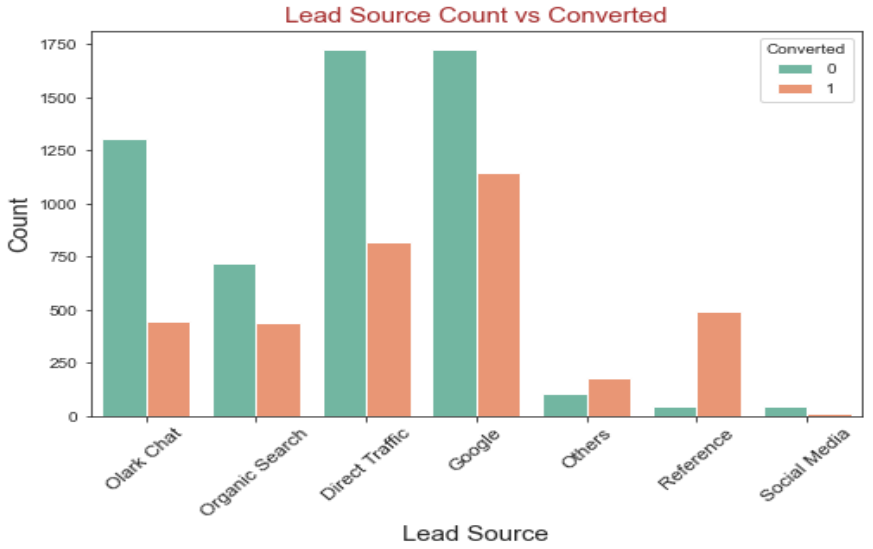
1. People with higher number of **'TotalVisits'** seem to have converted more.

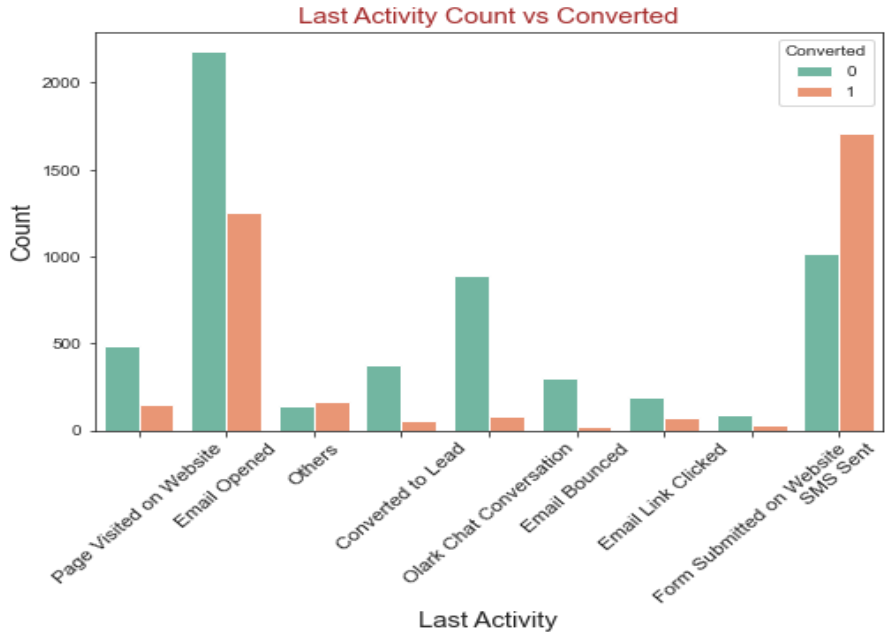
2. **'Total Time Spent on Website'** by people who have converted is significantly higher than people who have not converted.

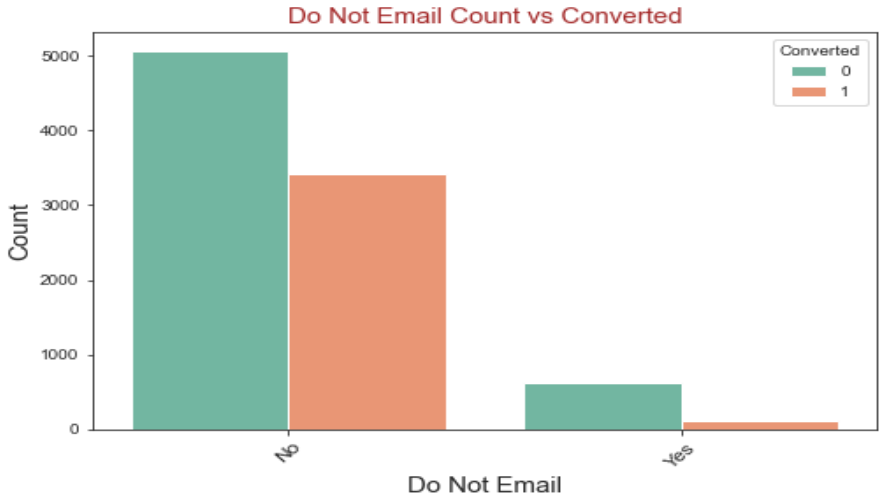
3. **'Page Views Per Visit'** does not make any difference to the conversion.

Bivariate Analysis - Categorical Variables









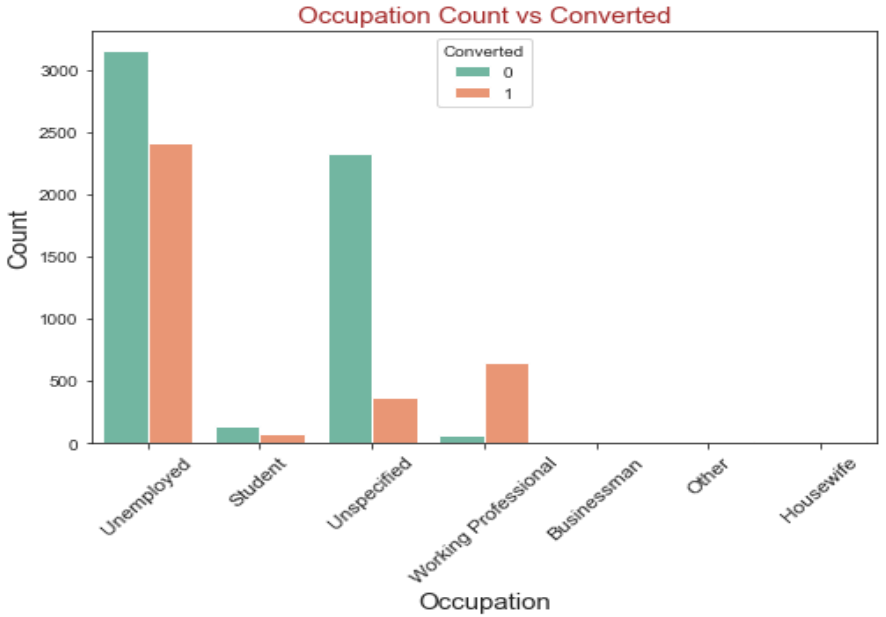
❑ Observations:

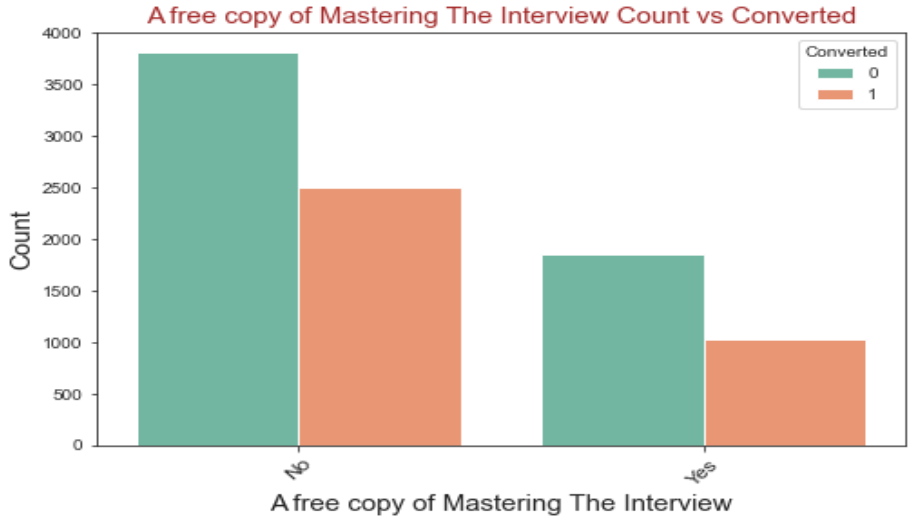
1. Most people who have converted have 'Google' as a source of lead. Google searches resulted in higher conversions, but Reference has the highest conversion rate.

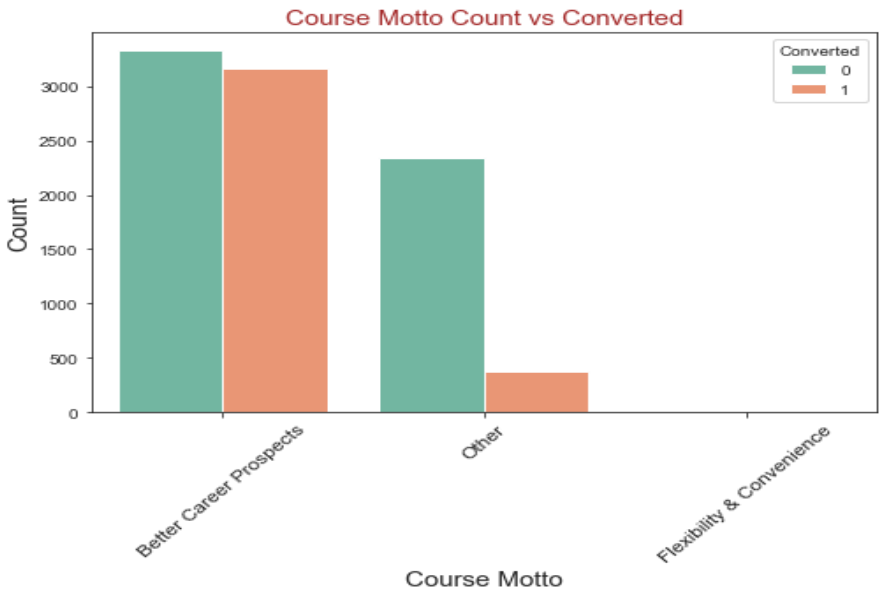
2. People who have not converted to customers have asked not to email.

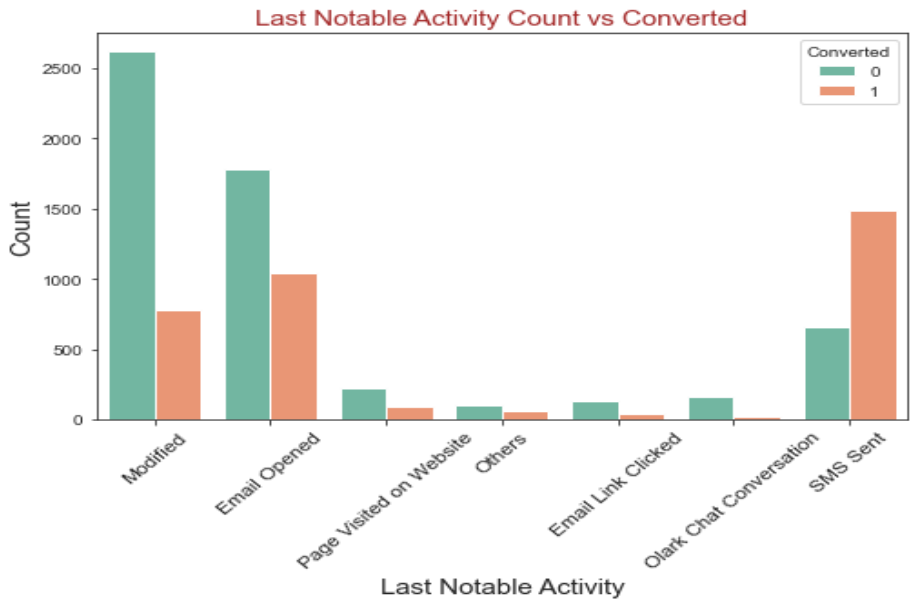
3. Most leads have usually opened their emails or sent an SMS as their last activity.

Contd. : Bivariate Analysis - Categorical Variables









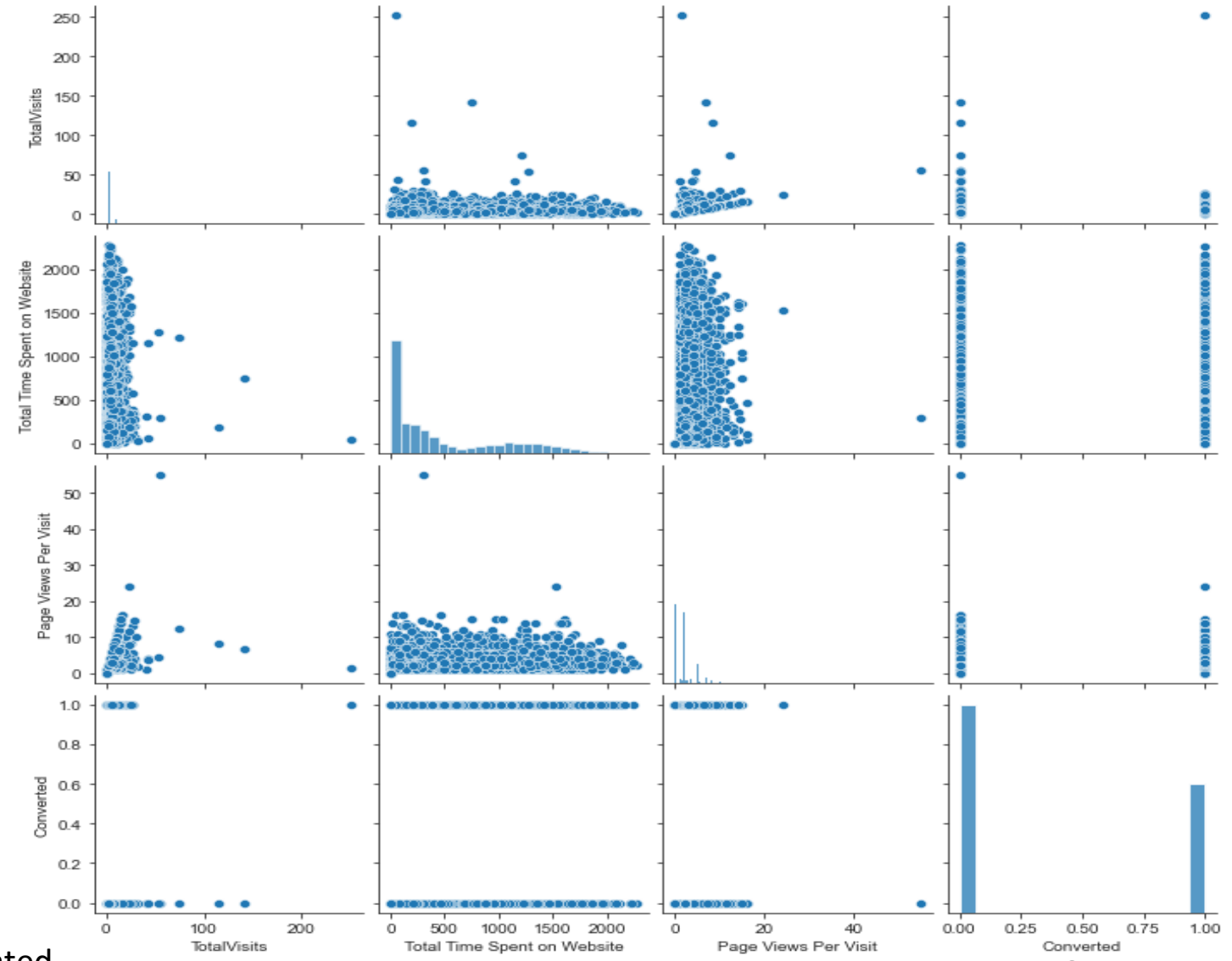
❑ **Observations:**

1. Unemployed people are higher in number for both converted and non-converted.

2. People who are looking for better career prospects are more likely to be converted as customers.

3. People who have not converted to customers have asked not to share the free copy of mastering the interview.

Multivariate Analysis - Continuous Variables

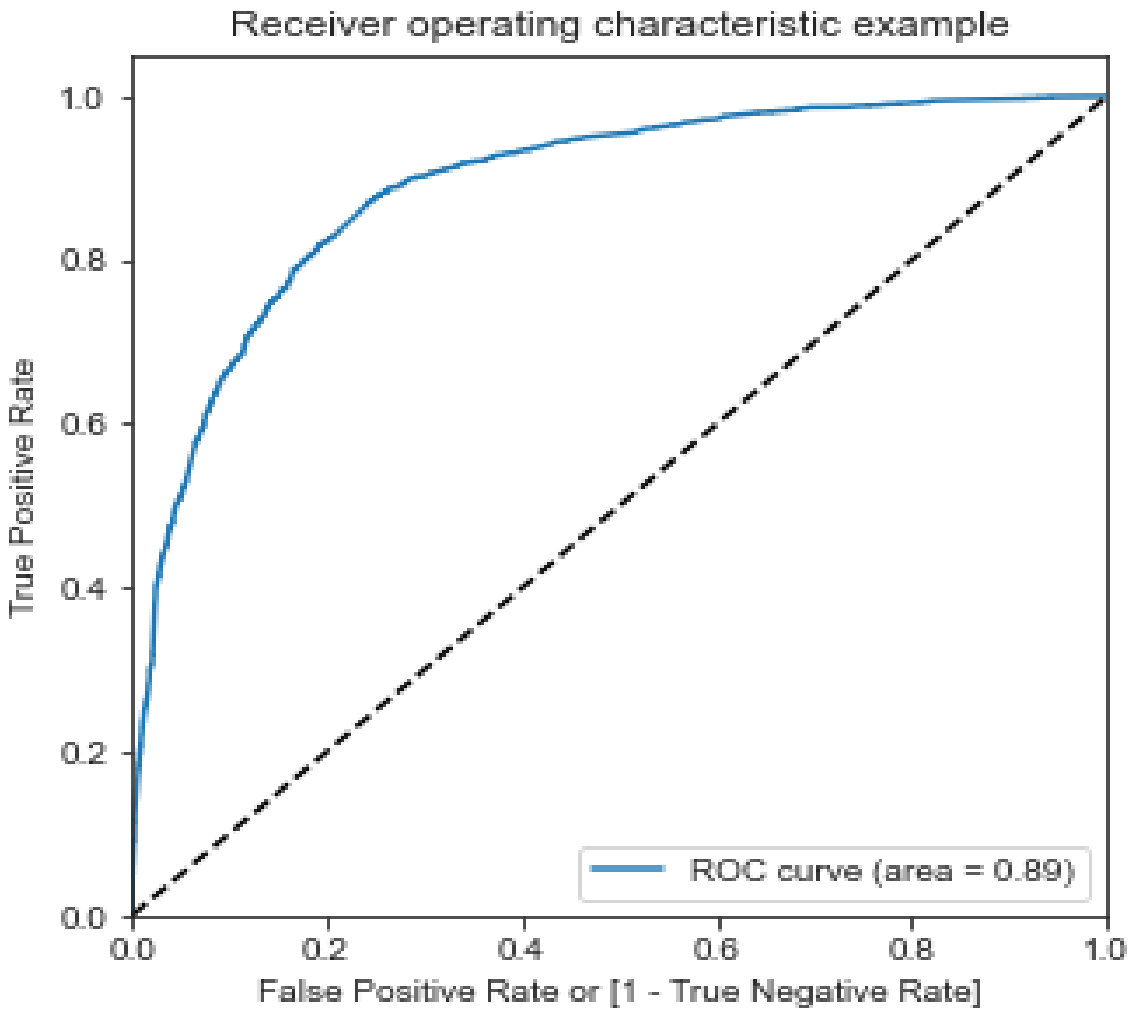


- **Observations:**

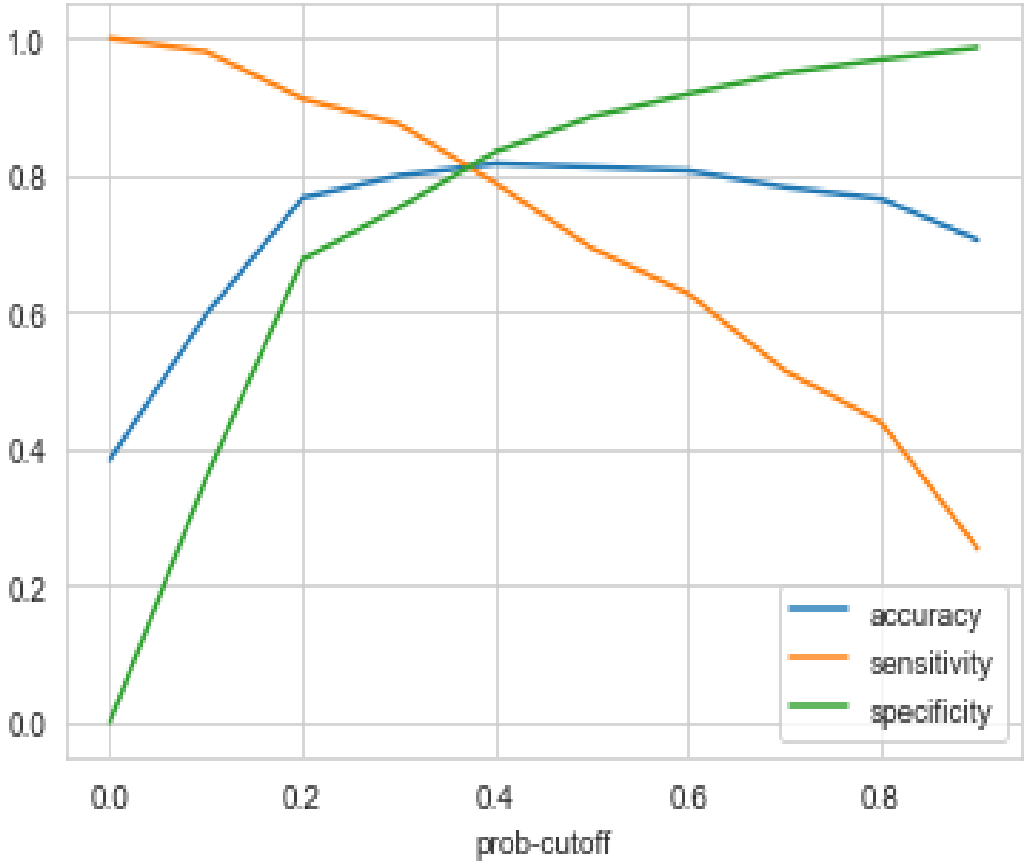
'Page Views Per Visit' and **'TotalVisits'** are moderately correlated.

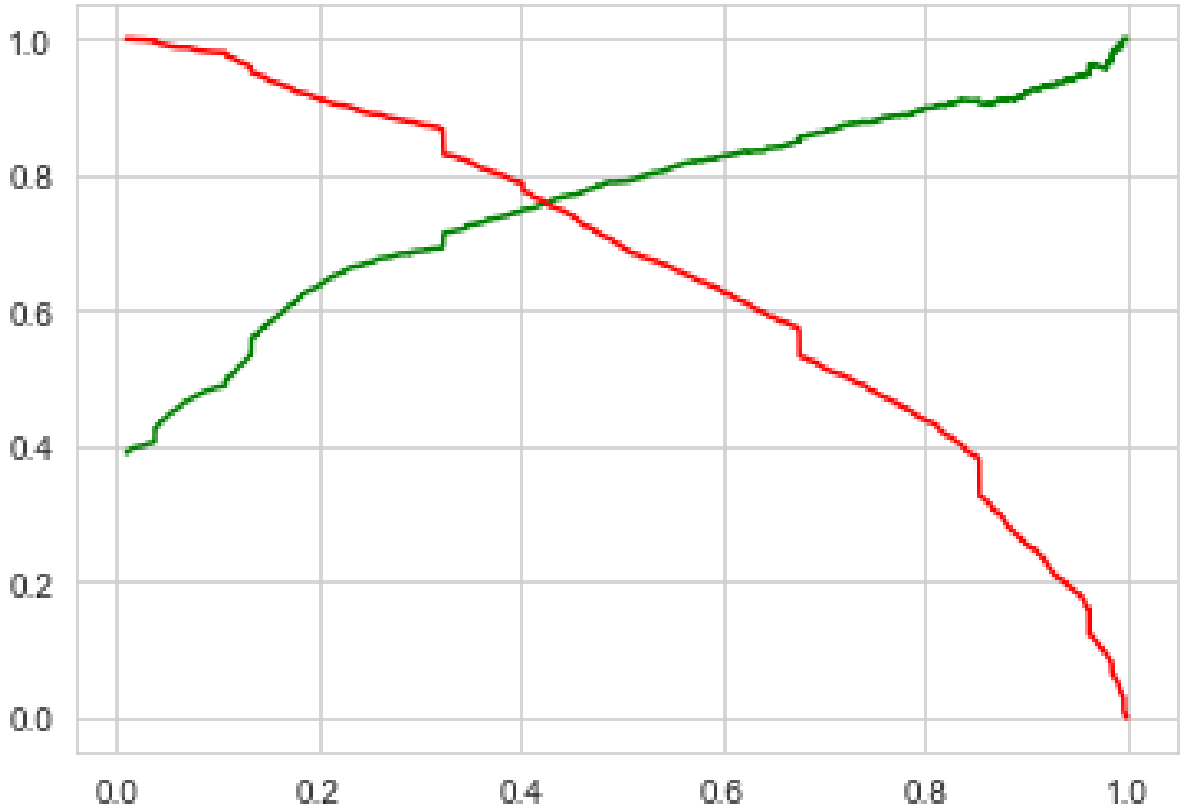
Model Building - ROC Curve

- Created a logistic regression model using **RFE (Recursive Feature Elimination)** to get 15 features initially.
- Dropped variables one by one manually on the basis of high P-value (>0.05) and high VIF (>5). Finally, we retained 12 variables in 'X_train' data.
- Settled with a stable model and predicted the training data with an initial cut-off of 0.5.
- Plotted the **ROC Curve** (Receiver Operating Characteristic) and obtained area under the curve **(AUC)** of **0.89.**



Optimal Cut-off Point



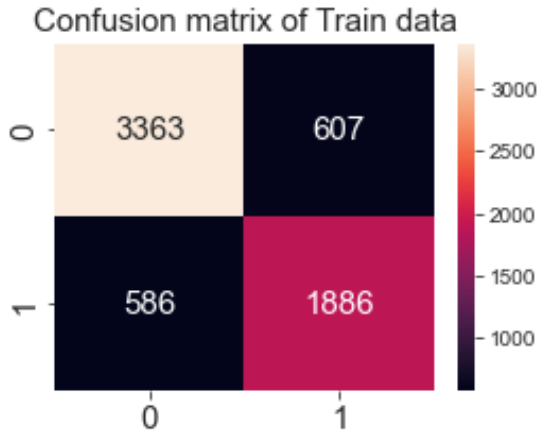


Precision – Recall trade-off

- **Optimal Cut-off Point** is that probability value where we get balanced Sensitivity and Specificity. We settled for the optimal cut-off point as **0.42** on the basis **of 'Precision' and 'Recall' trade-off**.
- Predictions on the **Test Data** are made with the optimal cut-off of **0.42.**

Model Metrics

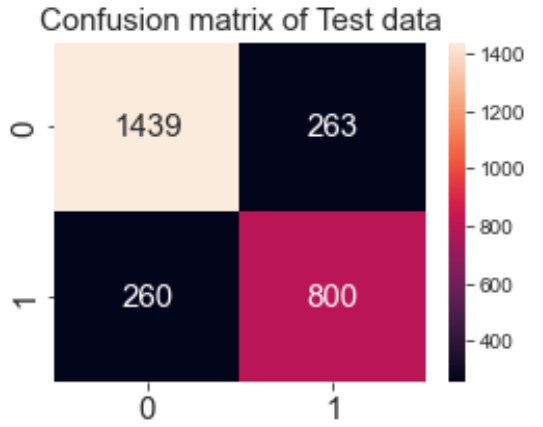
Train Data Metrics



Important metrics on Train data with optimal cut-off 0.42

Metric	Value (%)
Accuracy	81.48
Recall	76.29
Precision	75.65
F1-Score	75.97
Sensitivity	76.29
Specificity	84.71

Test Data metrics



Important metrics on Test data with optimal cut-off 0.42

Metric	Value (%)
Accuracy	81.06
Recall	75.47
Precision	75.25
F1-Score	75.36
Sensitivity	75.47
Specificity	84.54

Final Model Parameters

- **The following parameters have been obtained from the final model :**

Parameters	Coefficient
Do Not Email	-1.492
Total Time Spent on Website	1.107
Lead Origin_Lead Add Form	3.833
Lead Source_Olark Chat	1.334
Last Activity_Email Bounced	-1.257
Last Activity_Olark Chat Conversation	-1.372
Last Activity_Page Visited on Website	-0.499
Occupation_Student	1.236
Occupation_Unemployed	1.131
Occupation_Working Professional	3.514
Last Notable Activity_Others	1.825
Last Notable Activity_SMS Sent	1.474

Summary & Recommendations

- **Summary:**
  - Model shows an '**accuracy' of 81.06%** on the Test data. It is close to the value obtained for Training set (81.48%), which suggests that the model is robust enough.
  - **Top 3 features which are contributing significantly towards the probability of a lead getting converted are the following :**

    i. ***Occupation_Working Professional:*** *Indicates whether the customer is a student, unemployed or employed. It can be seen that 'Working Professionals' have the highest chances to convert to a customer.*
    ii. ***Lead Origin_Lead Add Form and Lead Source_Olark Chat:*** *Lead Origin and Lead Source also play an important role in increasing the probability of lead conversion.*
    iii. ***Last Notable Activity_SMS Sent:*** *The last notable activity where the student has sent an SMS seems to play an important role. It can be the sign of an interested customer.*
  - Increase in these factors will also increase the probability of lead conversion, so these factors should be given an equal importance.
    i. *Occupation_Student*
    ii. *Occupation_Unemployed*
    iii. *Total Time Spent on Website*

- **Recommendations:**
  - Leads who are investing more time on visiting the website or leads gathered from sources like Olark Chat, References are more likely to convert.
  - Leads which are tagged as they will revert after reading the email are having very high potential of converting. Hence these users should be regularly followed up.
  - Leads who are unemployed or working professional have higher potential to become paying customers. Having them in loop can be beneficial for the company.

Thank you!