# Lead Scoring Case Study

Anurag Kumar Pal

Prasanna Hanumanthu

Malarvizhi Paulraj

# Content

- Problem Statement
- Business Objective
- Dataset Provided
- Approach Followed
- EDA Results
- Conclusions
- Recommendations

# Problem Statement

- An education company named X Education 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%.
- 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 Objective

- 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.

# Dataset Provided

- **Assignment Subjective Questions -** contains four subjective question of Lead Scoring analysis and looking for its outcome.
- **Leads.csv** - contains 37 columns and 9240 rows of lead scoring data.
- **Leads Data Dictionary -** contains Leads data variables and its descriptive part.

# Approach Followed

- Importing Libraries
- Data Reading and understanding
- Data Cleaning
- Visualize the data and apply EDA
- Model Building
- Splitting the data for training and testing
- Model Evaluation
- ROC Curve and Checking Accuracy
- Calculate Precision and Recall
- Assigning Lead Score

37.85541106458012

# EDA Results

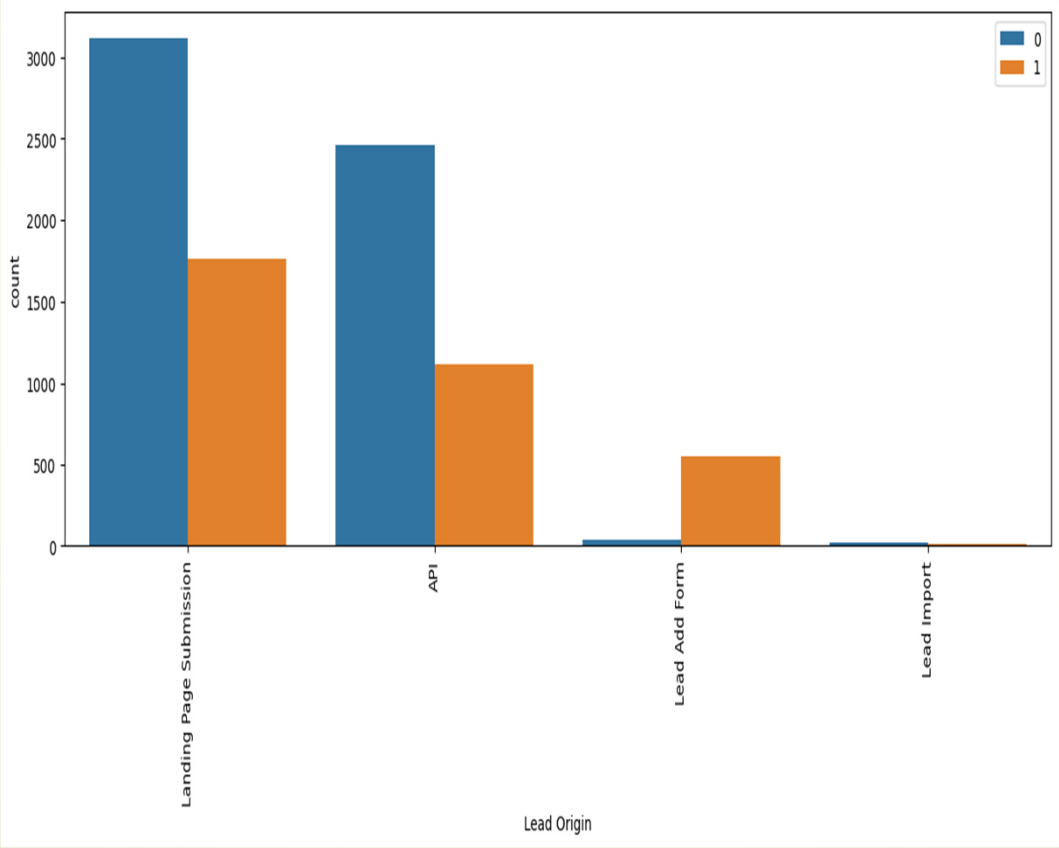
- Calculating the conversion rate:

ConversionRate=(sum(df_leads['Converted'])/len(df_leads['Converted'].index))*100

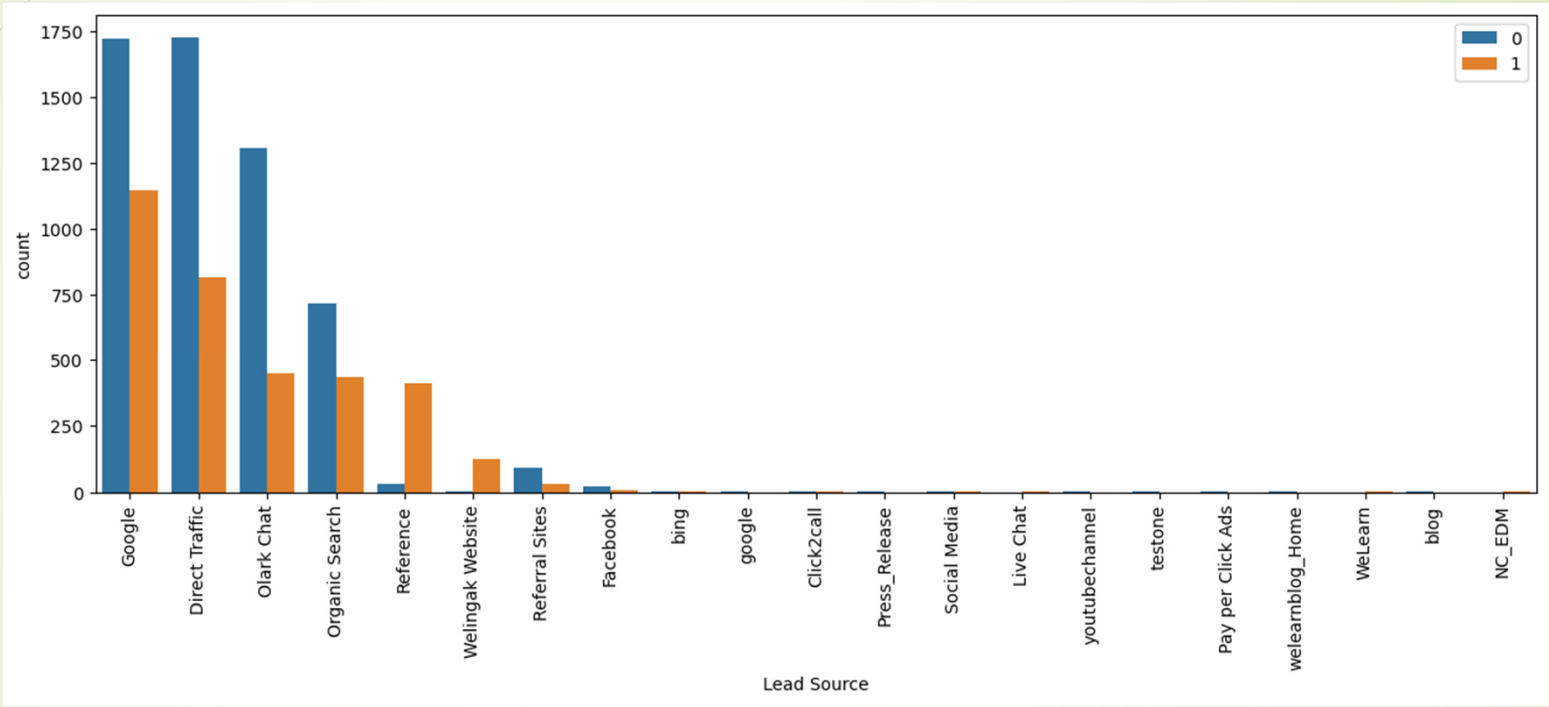
ConversionRate

Output:37.85541106458012

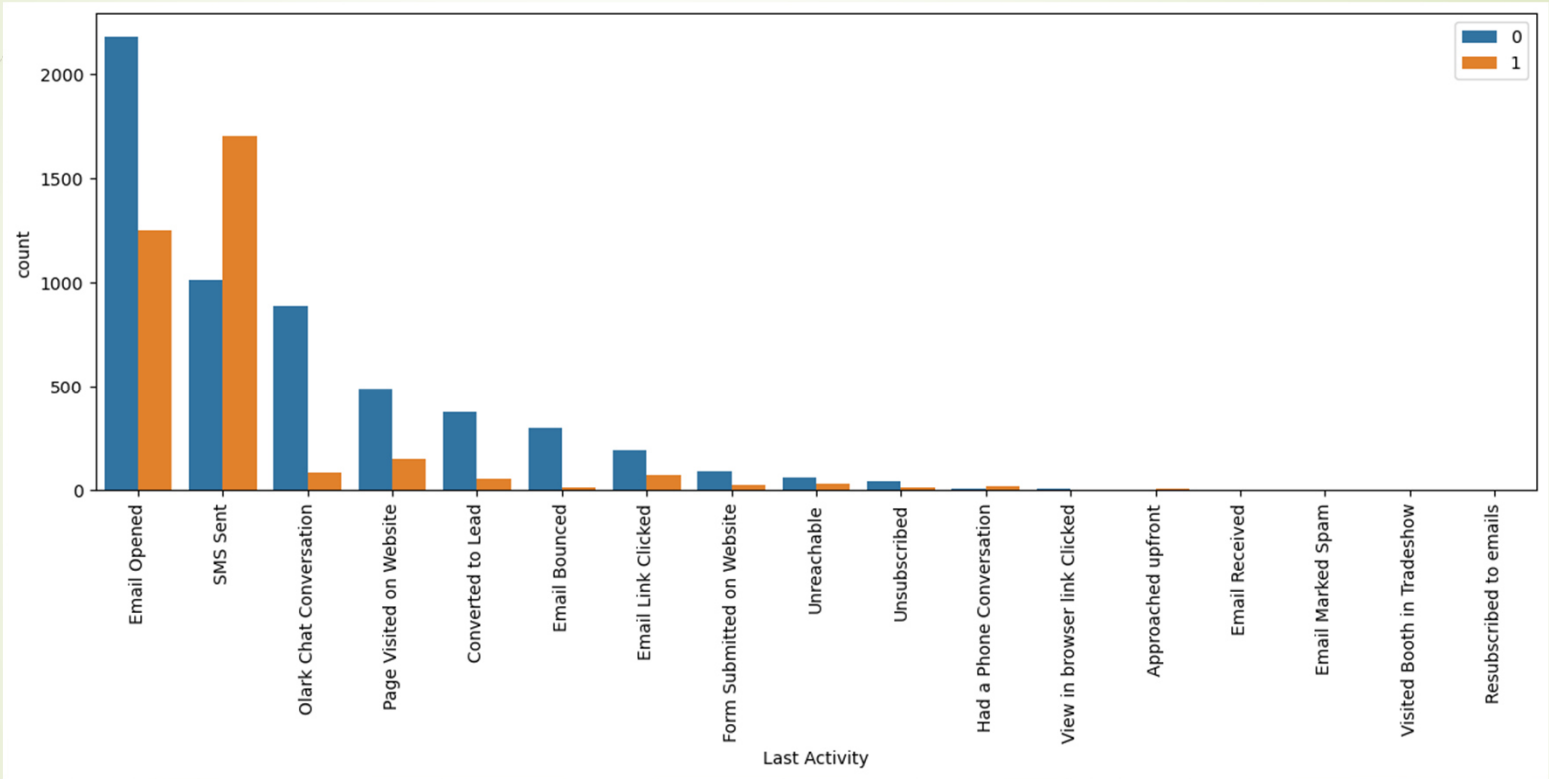
# Univariate Analysis of Lead Origin



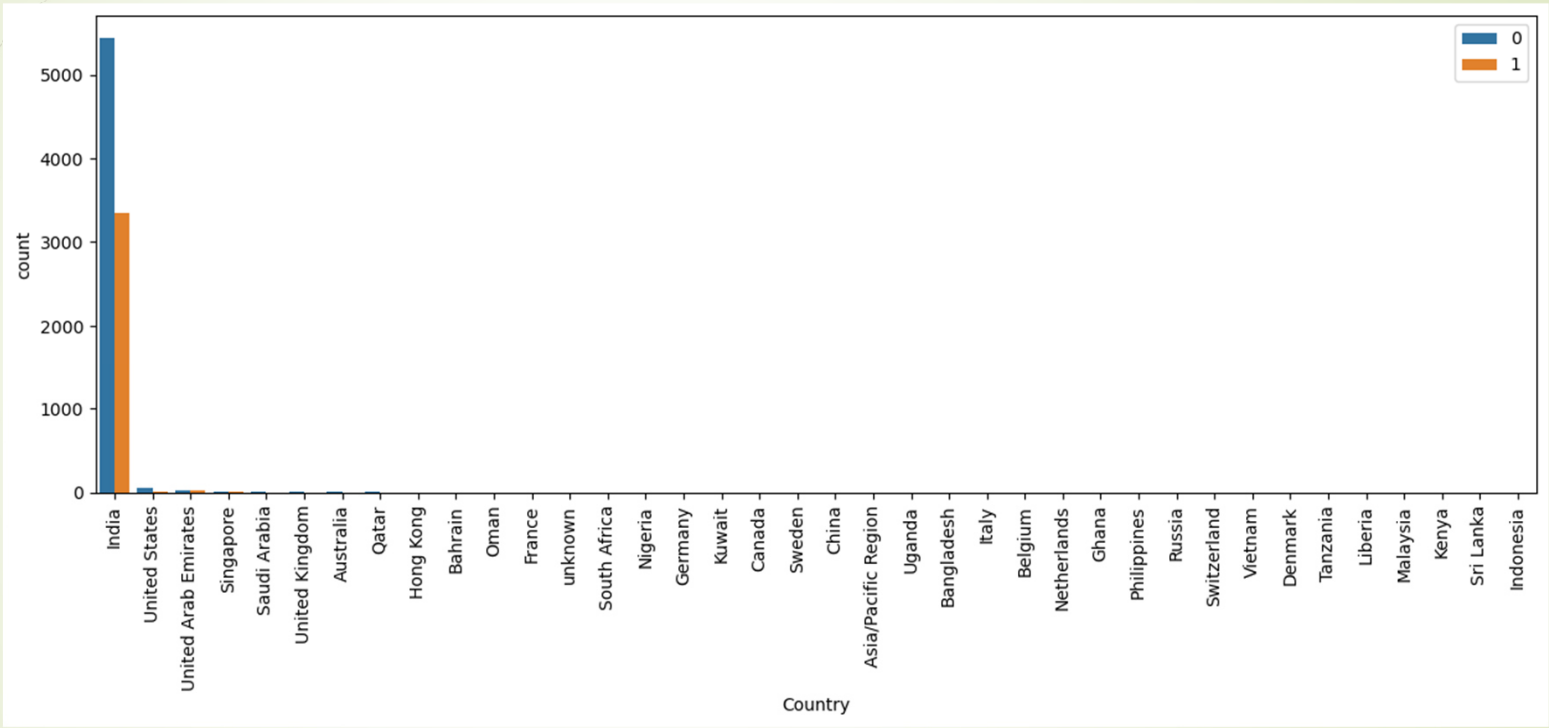
# Univariate Analysis of Lead Source



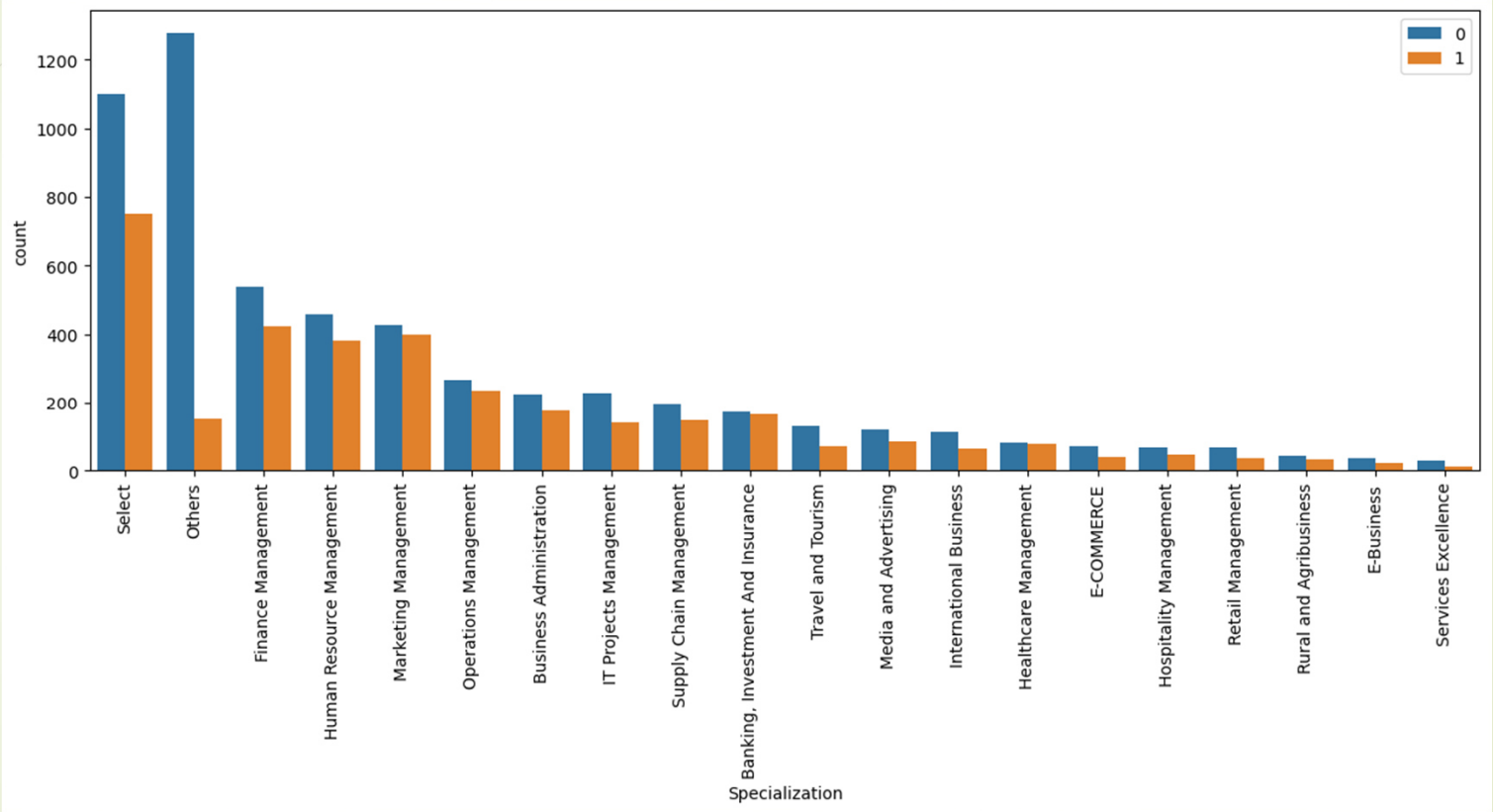
# Univariate Analysis of Last Activity



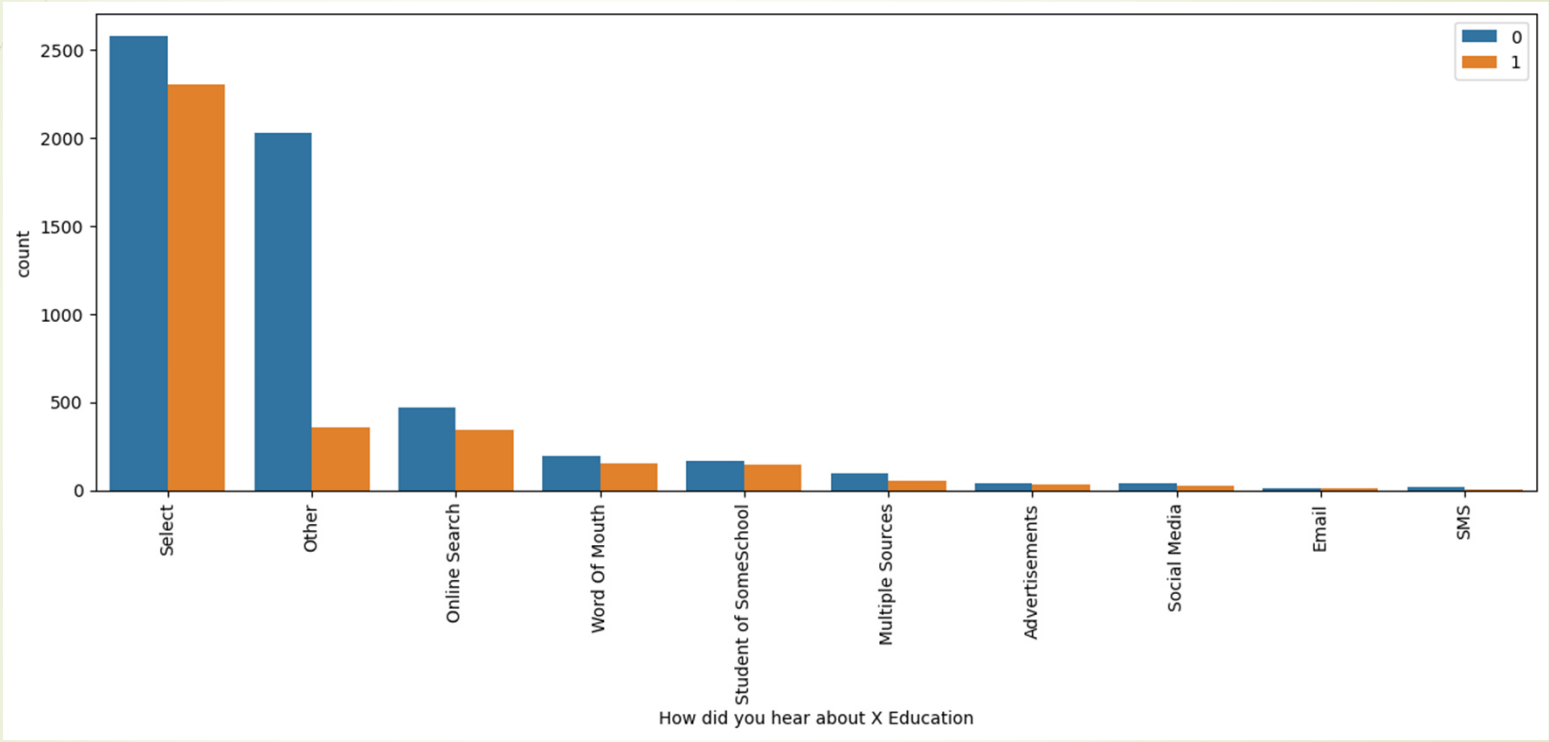
# Univariate Analysis of Country



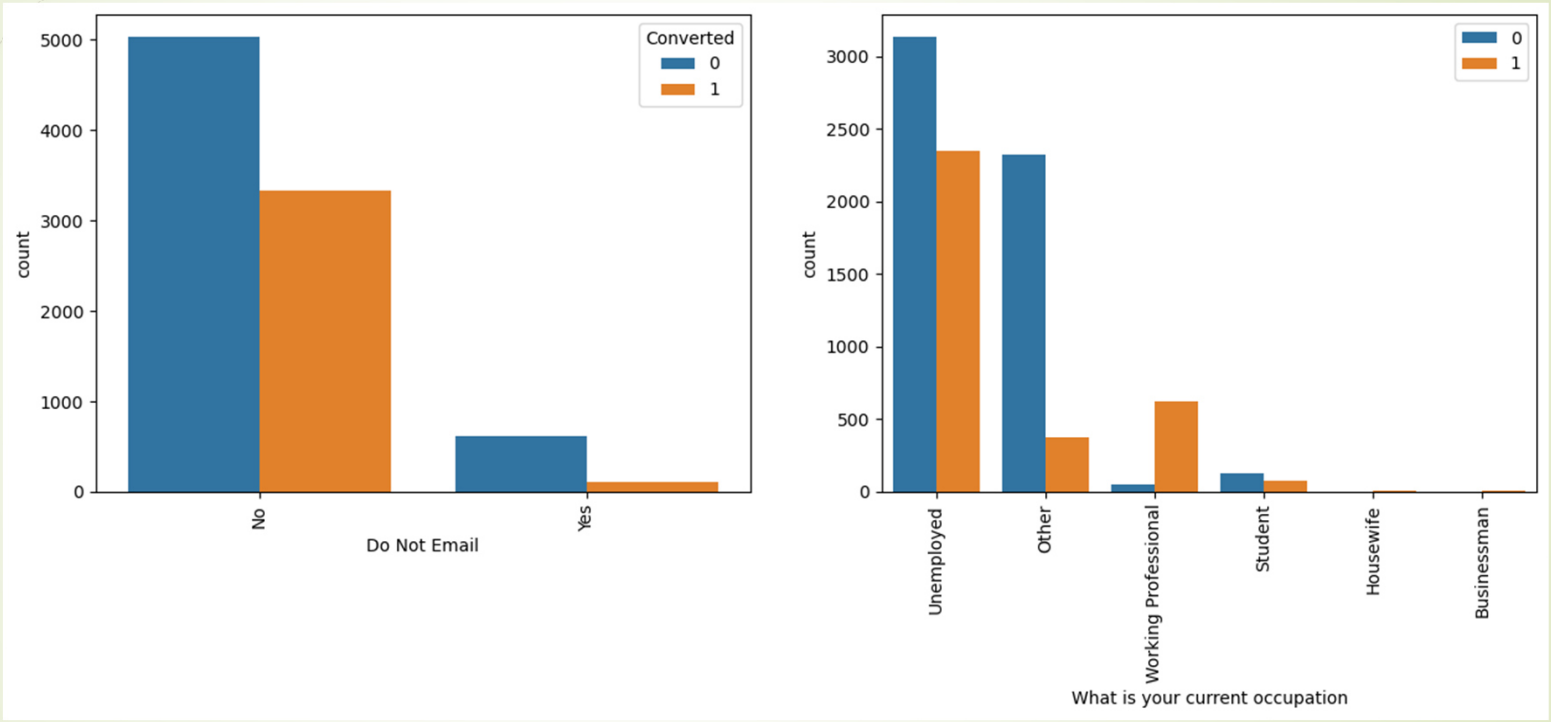
# Univariate Analysis of Specialization



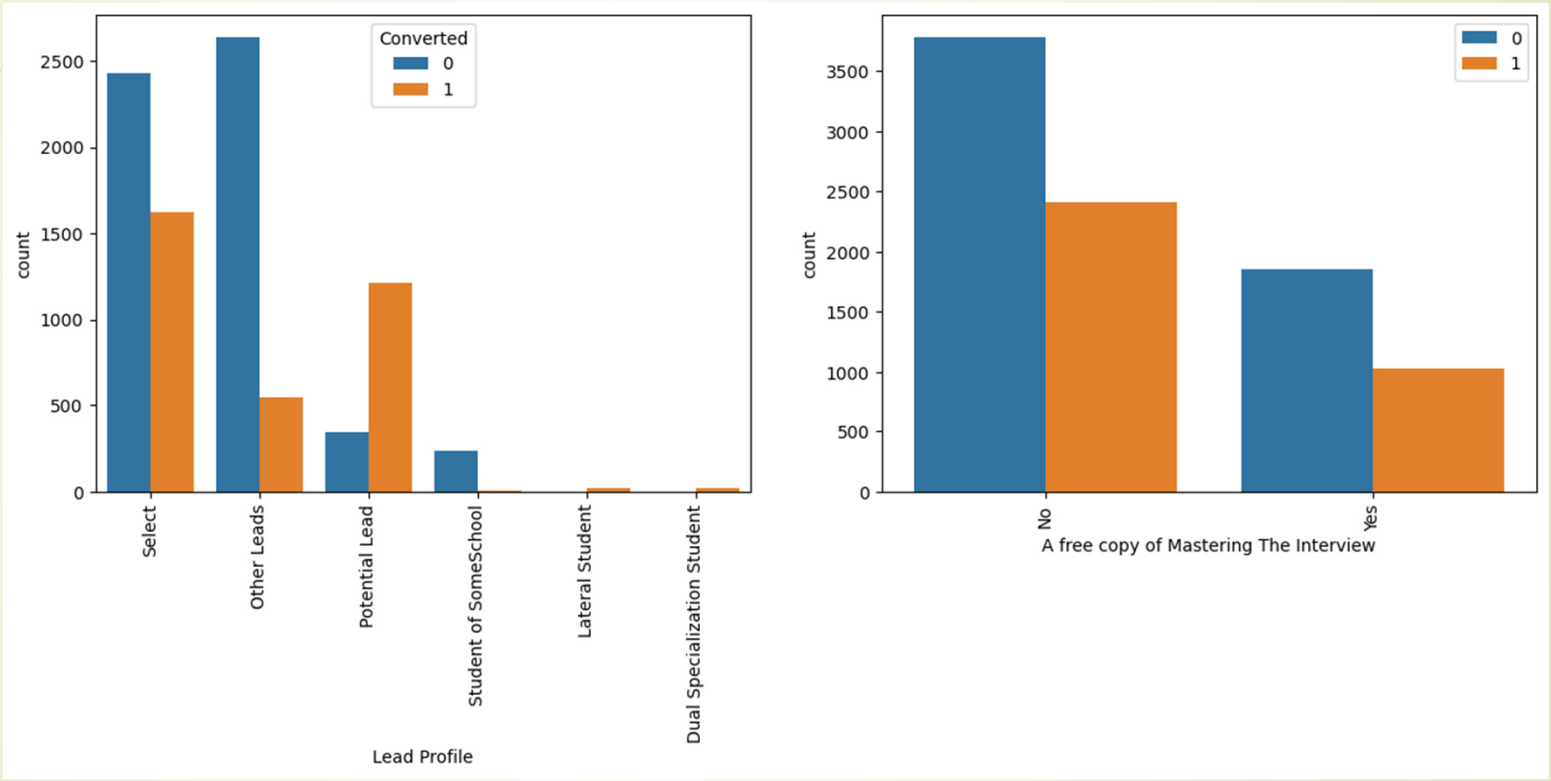
# Univariate Analysis of ‘How did you hear about X Education’



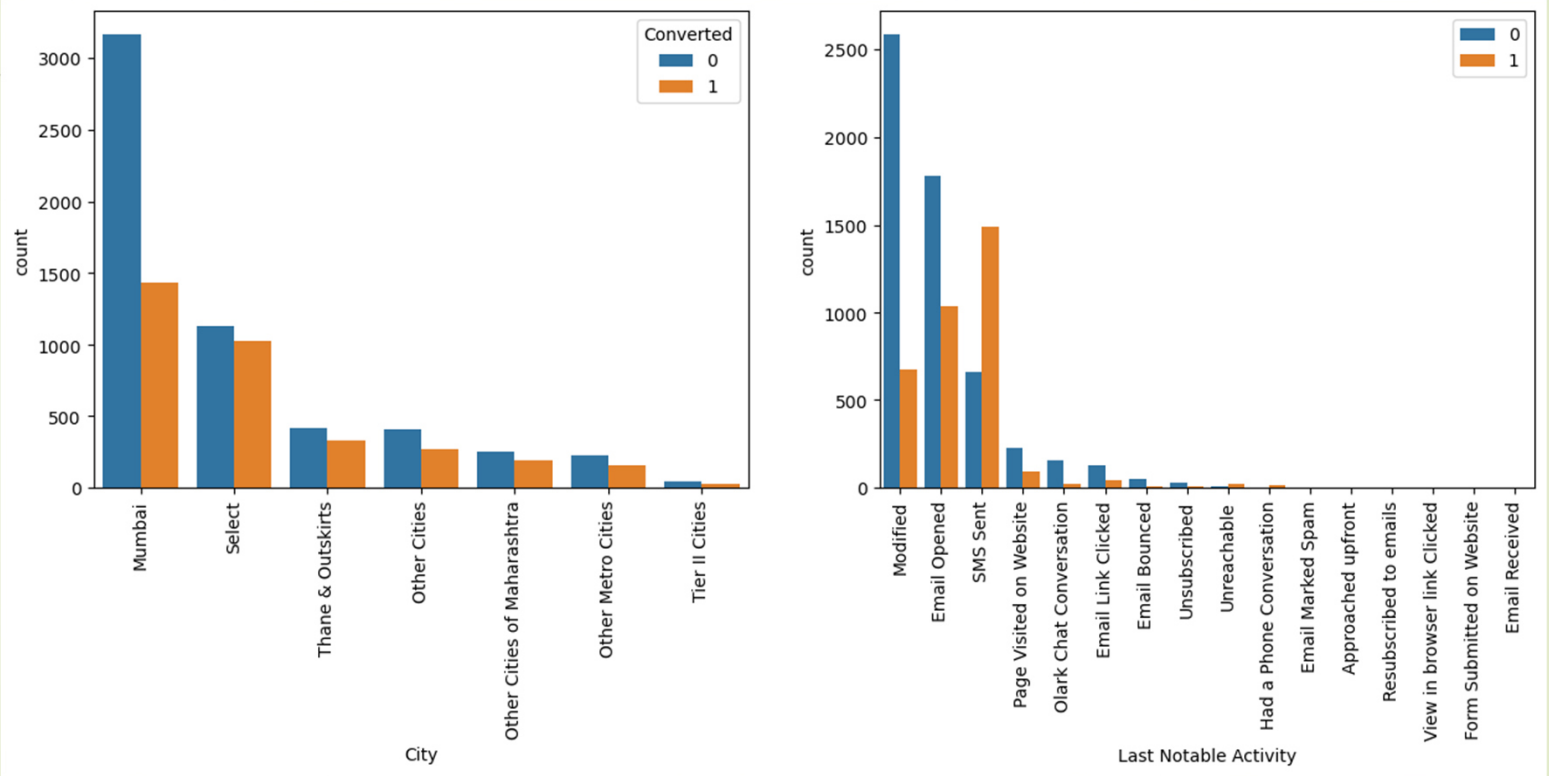
# Multivariate Analysis



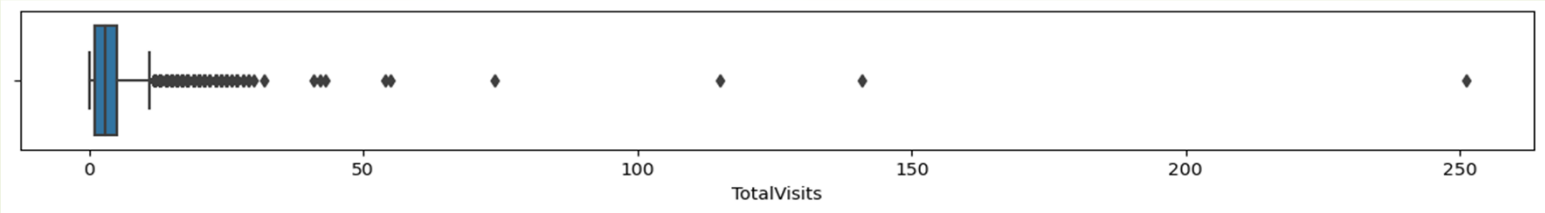
# Multivariate Analysis



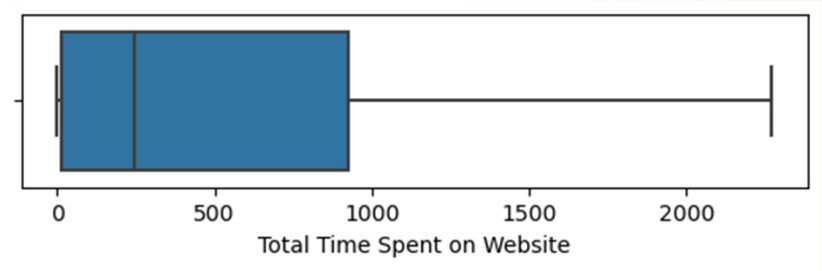
# Multivariate Analysis



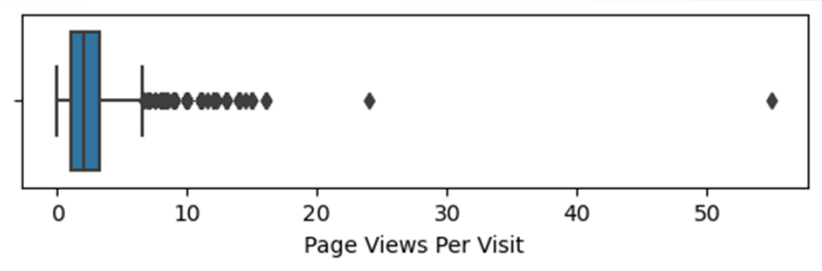
# Boxplot for TotalVisits before Outliers



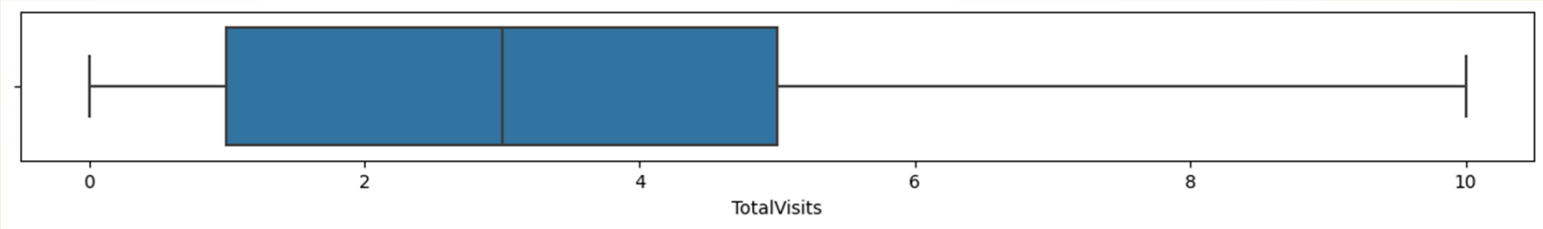
# Boxplot for Total time spent on website before Outliers



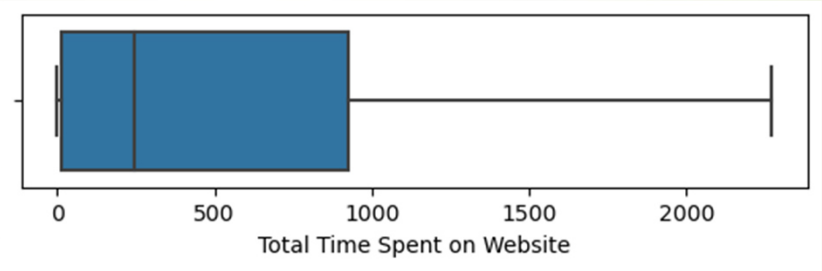
# Boxplot for page views per visit before Outliers



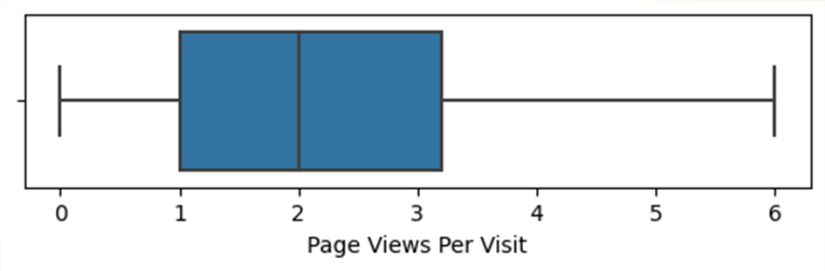
# Boxplot for TotalVisits after Outliers



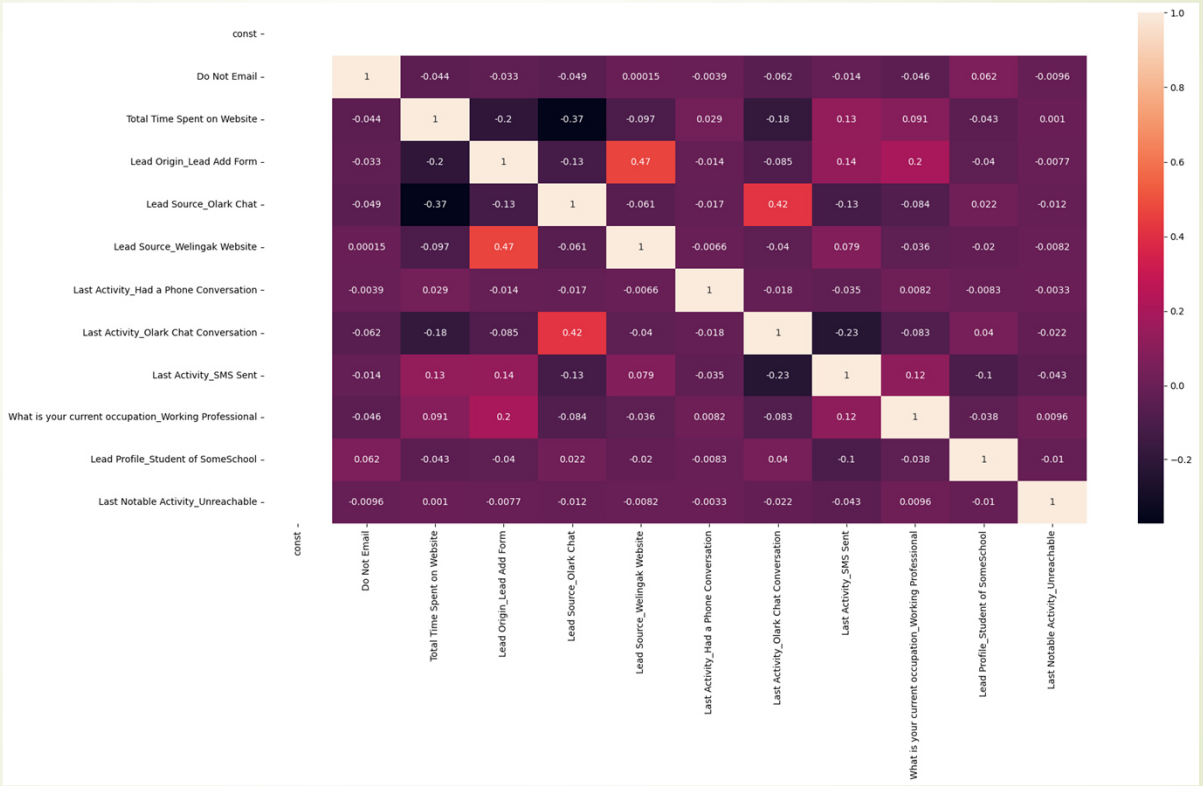
# Boxplot for Total time spent on website after Outliers



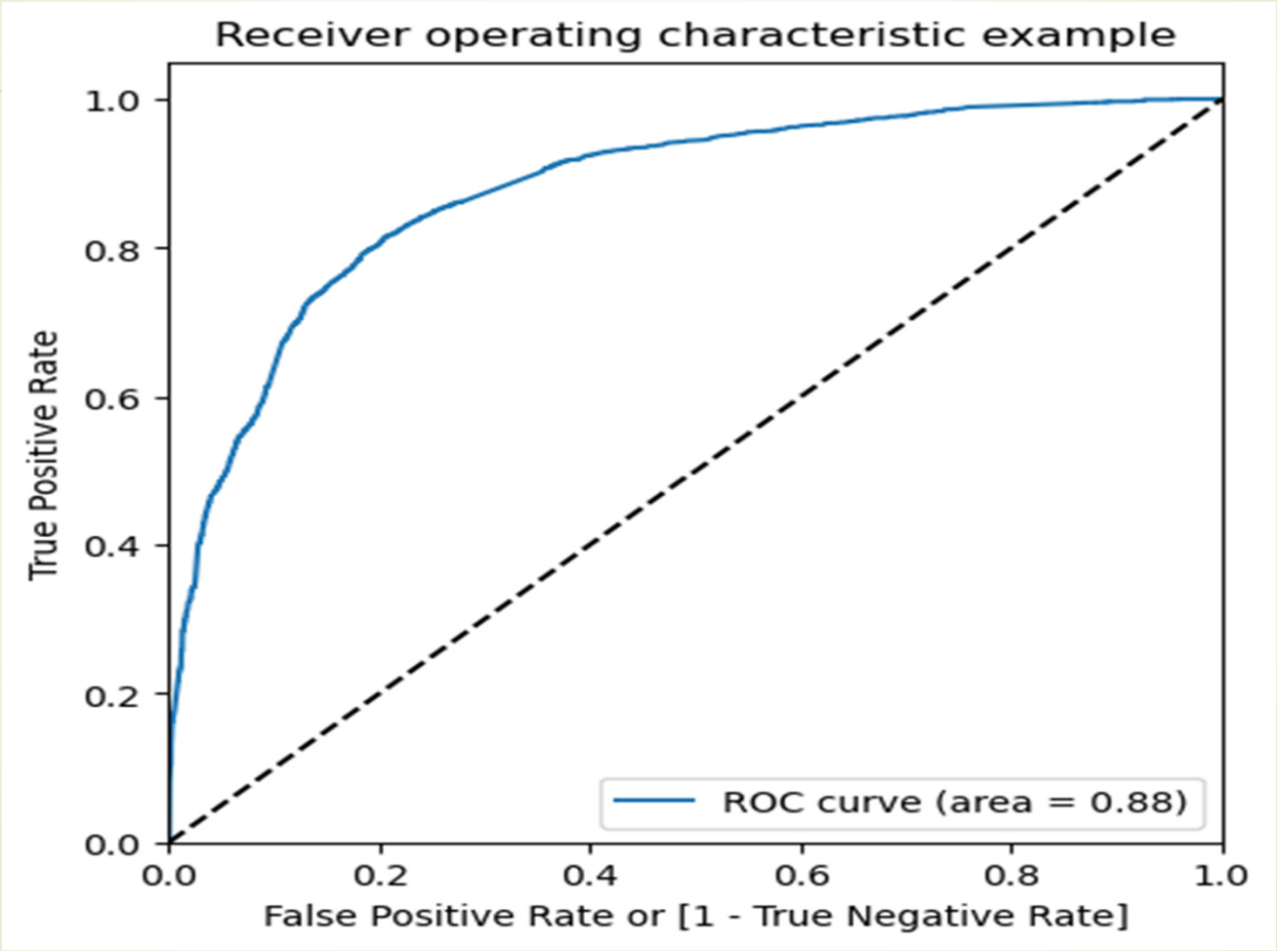
# Boxplot for page views per visit  after Outliers



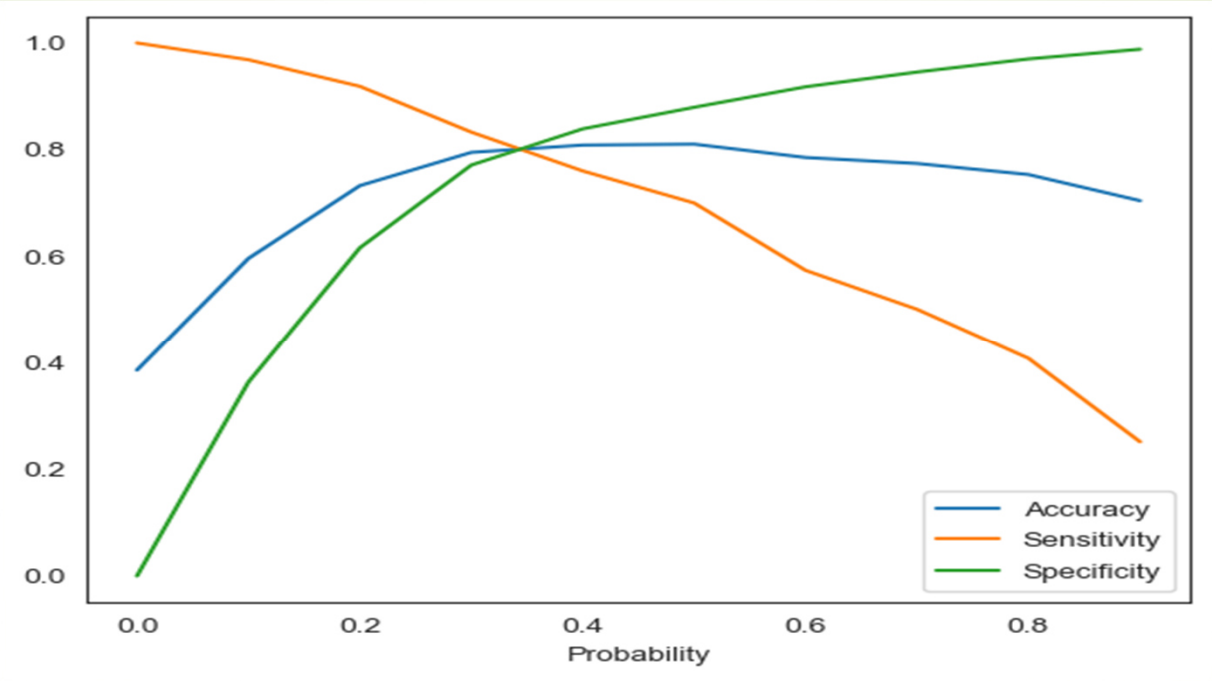
# Correlations



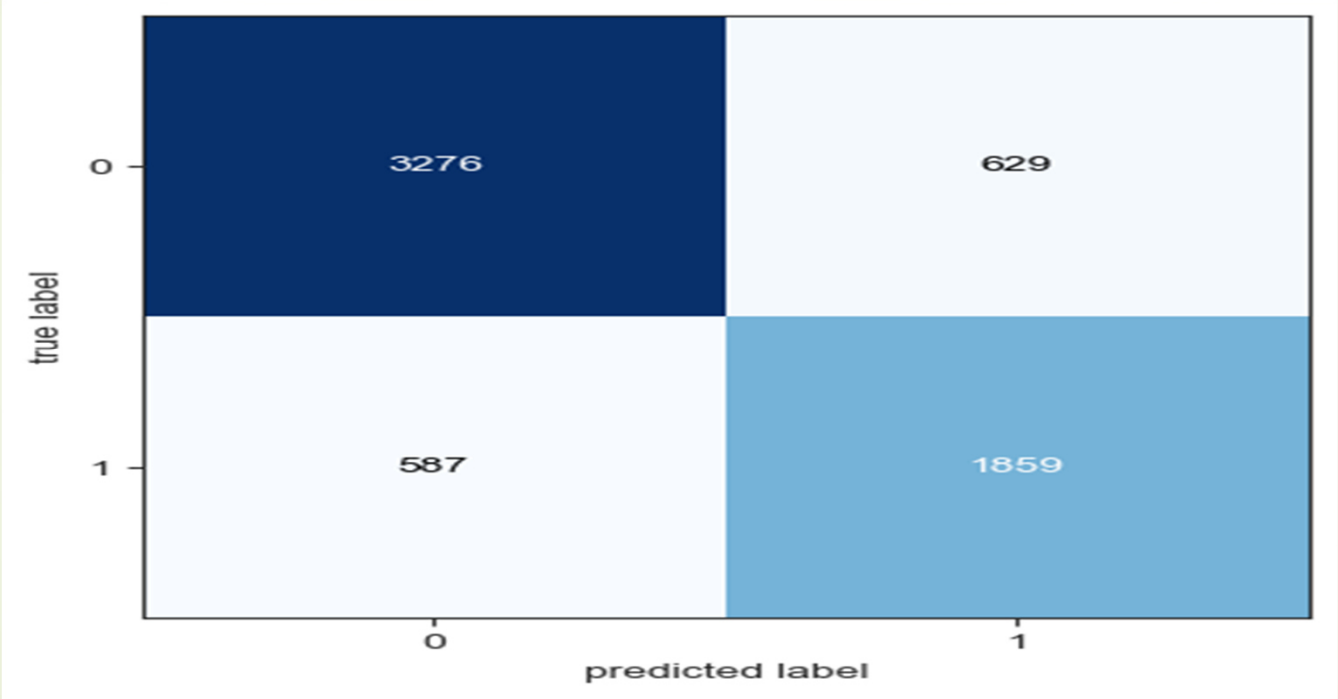
# Plotting ROC curve



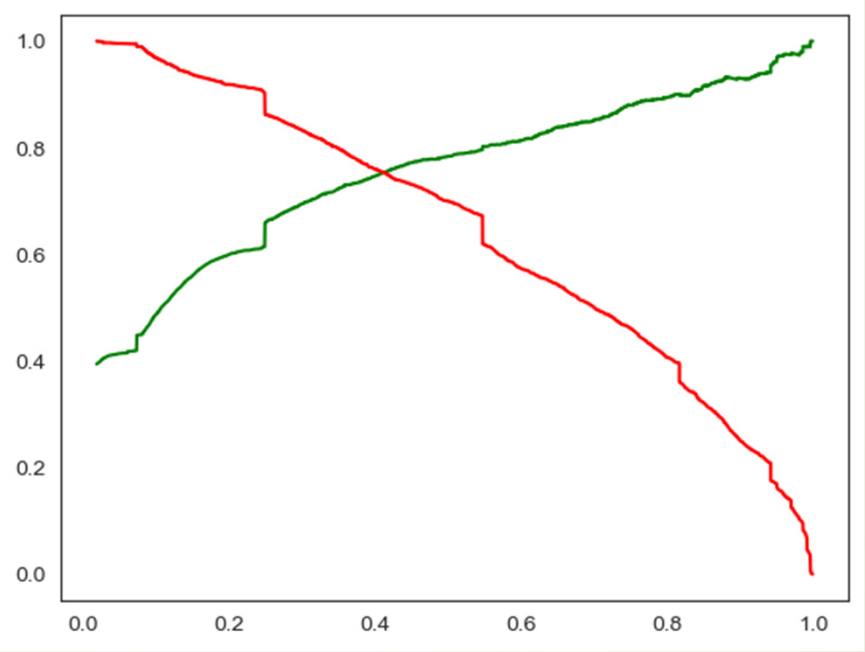
# Plotting Accuracy,Sensitivity,Specificity



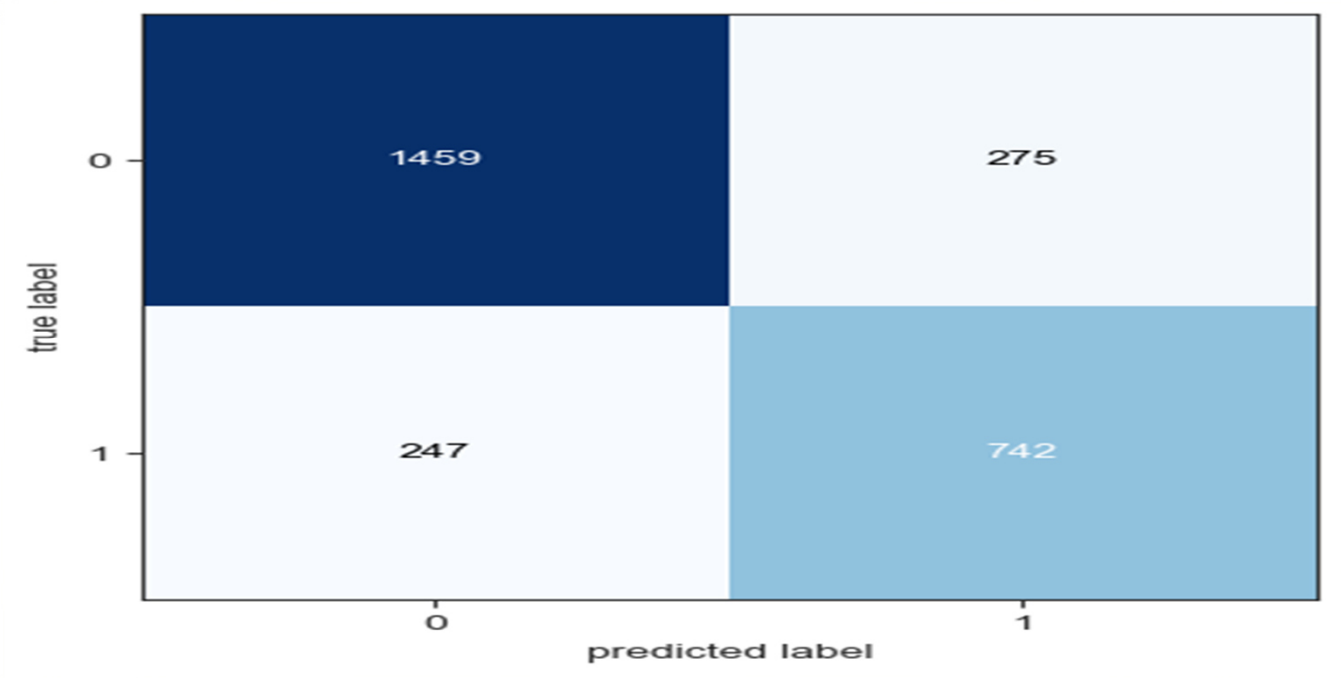
# Plotting Confusion Matrix of Training Data



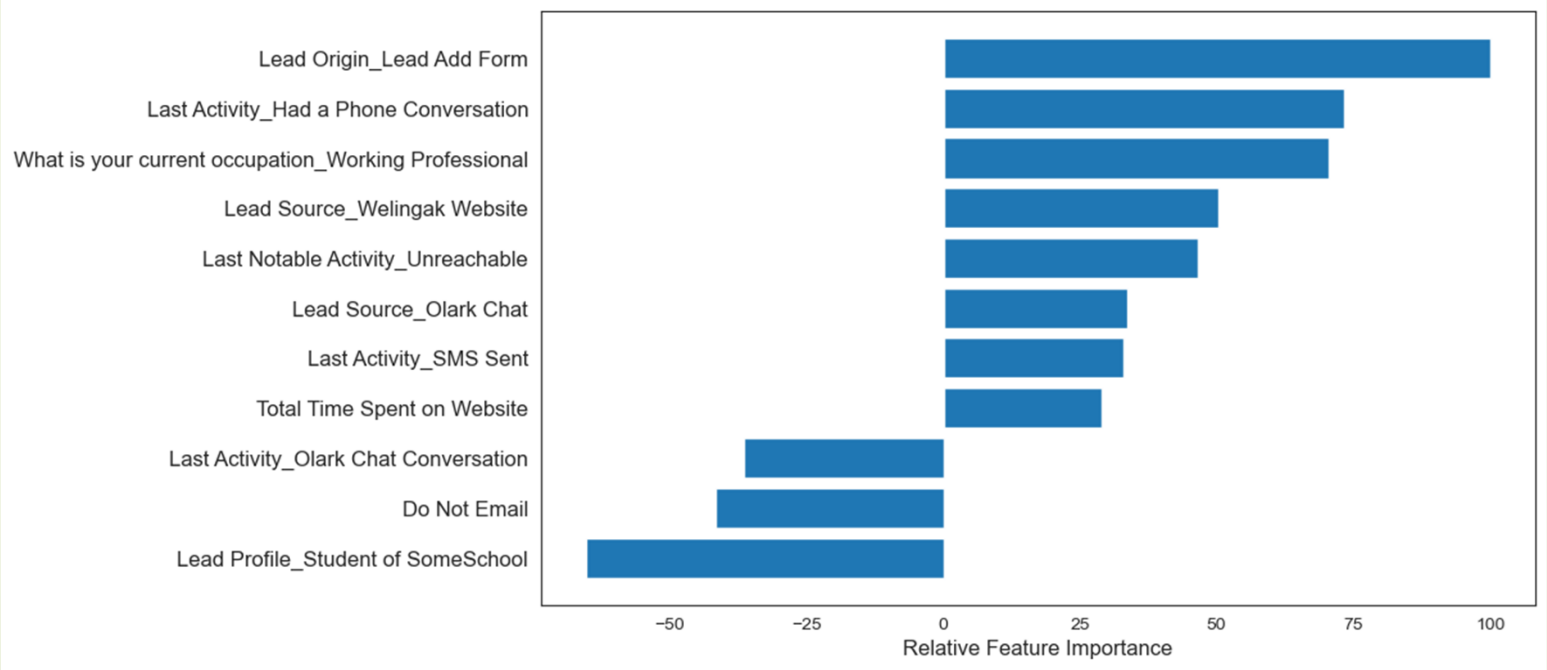
# Precision Vs Recall



# Plotting Confusion Matrix of Testing Data



# Relative Feature Importance



# Conclusions

## Evaluation Metrics for the Train Dataset:

- Accuracy: 0.81
- Sensitivity: 0.70
- Specificity: 0.88
- Precision: 0.78
- Recall: 0.70
- Positive predictive value: 0.78
- Negative predictive value: 0.82

## Evaluation Metrics for the Test Dataset:

- Accuracy: 0.81
- Sensitivity: 0.75
- Specificity: 0.84
- Precision: 0.73
- Recall: 0.75
- Positive predictive value: 0.73
- Negative predictive value: 0.86

# Recommendations

- There are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom. In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.
- First, sort out the best prospects from the leads you have generated. 'TotalVisits' , 'Total Time Spent on Website' , 'Page Views Per Visit' which contribute most towards the probability of a lead getting converted.
- Then, You must keep a list of leads handy so that you can inform them about new courses, services, job offers and future higher studies. Monitor each lead carefully so that you can tailor the information you send to them. Carefully provide job offerings, information or courses that suits best according to the interest of the leads. A proper plan to chart the needs of each lead will go a long way to capture the leads as prospects.

Thank You