# LEAD SCORE CASE STUDY

Abhinay Choday
Abhijit Ingale
Abhishek Gandham

# AGENDA

- LEAD SCORE – Problem Statement
- Approach
- Recommendations

# PROBLEM STATEMENT

- 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.
- Now, although X Education gets a lot of leads, its lead conversion rate is very poor at 30%. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'.
- X Education wants to select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company wants to build a model which assigns a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a 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%.

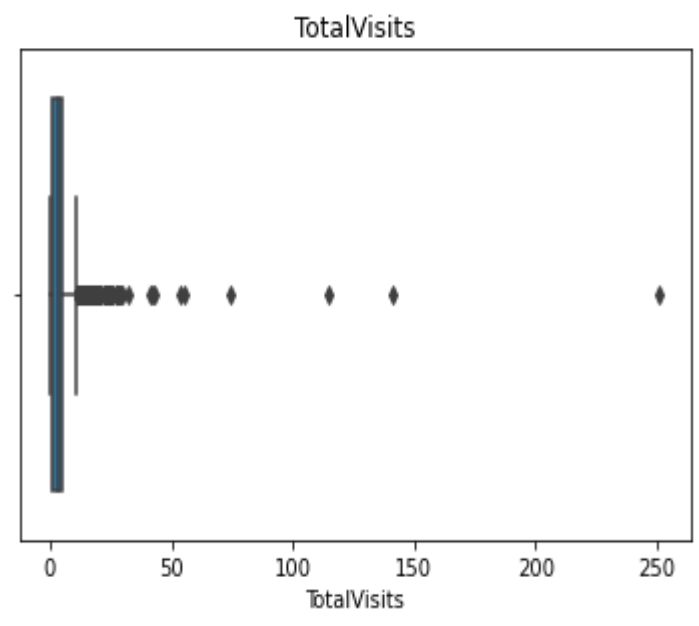
X-Education has data related to leads, which were collected from various sources. It also contains one attribute "Converted" which suggests whether a particular lead is converted or not (0-No, 1-Yes), along with various features. X-Education wants a model to be developed such that it will provide a set of hot leads which have a higher probability of the lead getting converted and crucial insights benefiting the business.

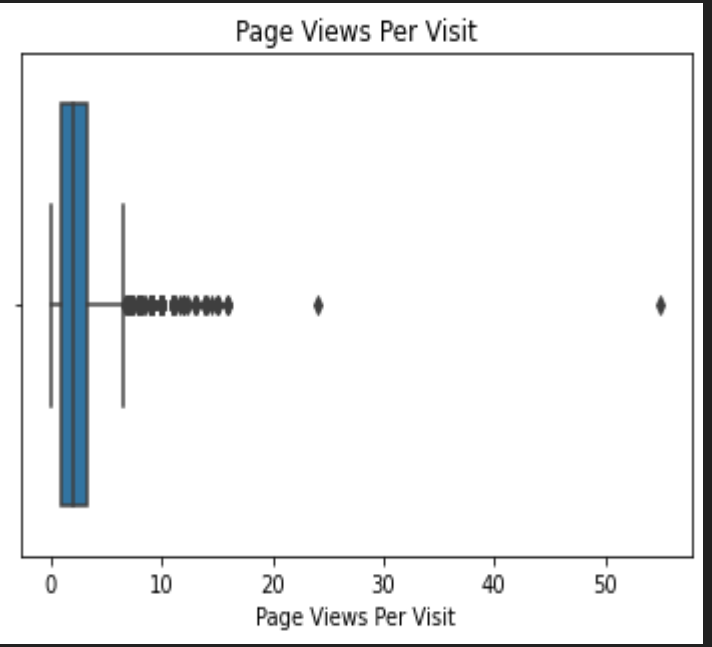
Finally stating, the main task of this case study is to find out insights that will help to improve the business of X-Education.

# APPROACH

- Understanding business scenario and problem statement
- Importing and inspecting the data
- Data cleaning and analysis
- Model Building
- Prediction
- Conclusion

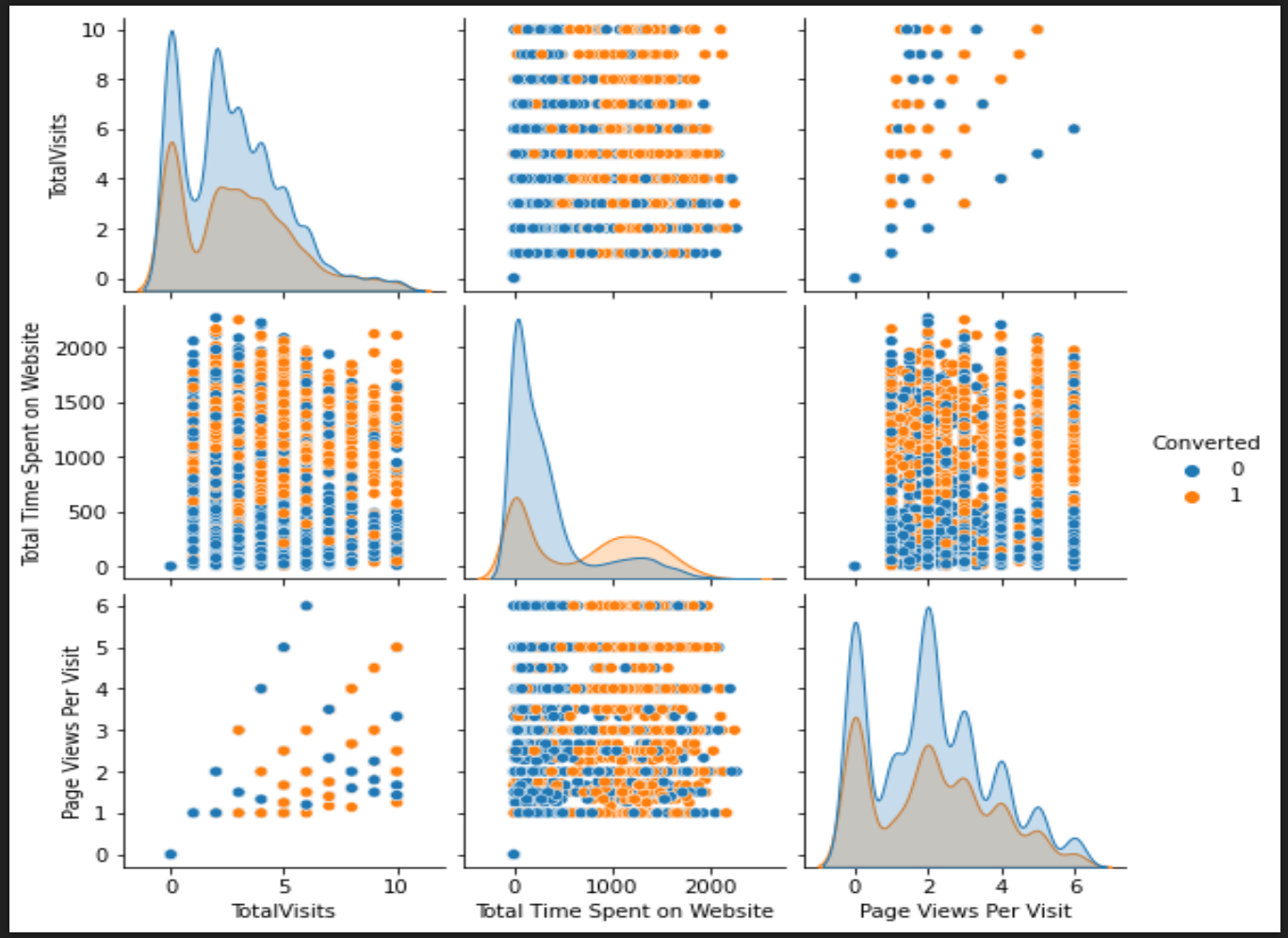
- We utilized Python language for carrying out EDA and building a Logistic Regression Model. Previous Leads conversion data containing 9240 data points was provided by the company.
- All columns having more than 40% missing values were dropped.
- Remaining feature having missing values were dealt with by imputing 'Unknown' for categorical variables.
- Rows were dropped for features having less than 2% missing values. This was followed by Outlier treatment of numerical features, this was done by dropping rows containing outliers.
- We were left with approximately 91% of data points remaining after data cleaning.



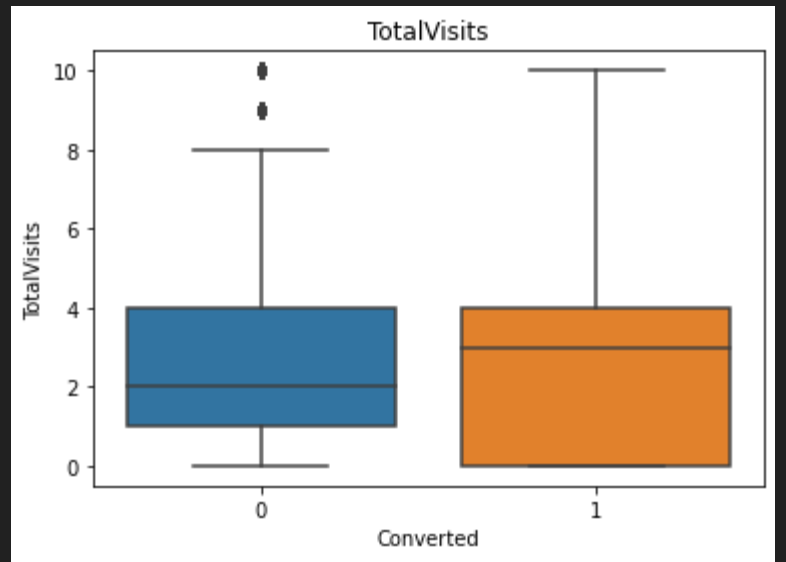
# Bivariate Analysis

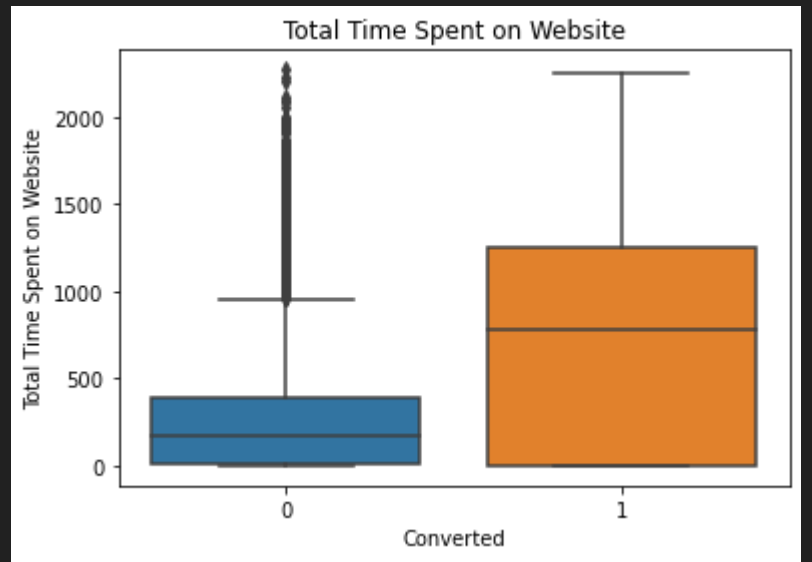
- As TotalVisits increases the Conversion rate increases. Highest conversion appears above 7 Visits to website.
- There is higher number of Converts when time spent on website is more than approximately 800. So making the visitors of website to spend more time on teh website could increase conversion rate.
- TotalVisits and Page Views Per Visit have a strong linear relationship

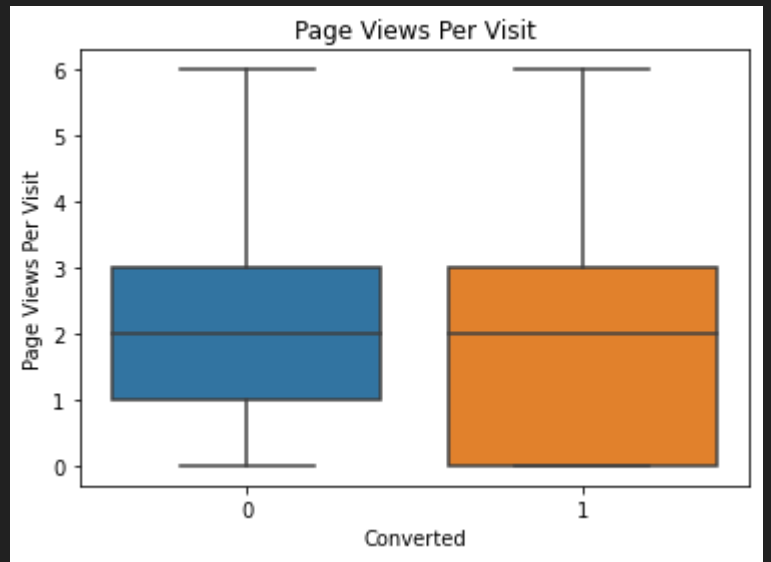


# Bivariate Analysis

- TotalVisits: The median value for Converted Leads is higher than non converted leads
- Total Time Spent On Website: The median and IQR between 25 percentyile to 75 Percentile of Total time spent on website is significantly higher. Time spent on website could be a significant indicator for conversion.
- Page Views Per Visit: The distribution appears to be identical for converted and non converted leads for Page Views per visit. It is an insignificant variable. Page Views Per Visit can be dropped since it is also having collinearity with Total Visits.

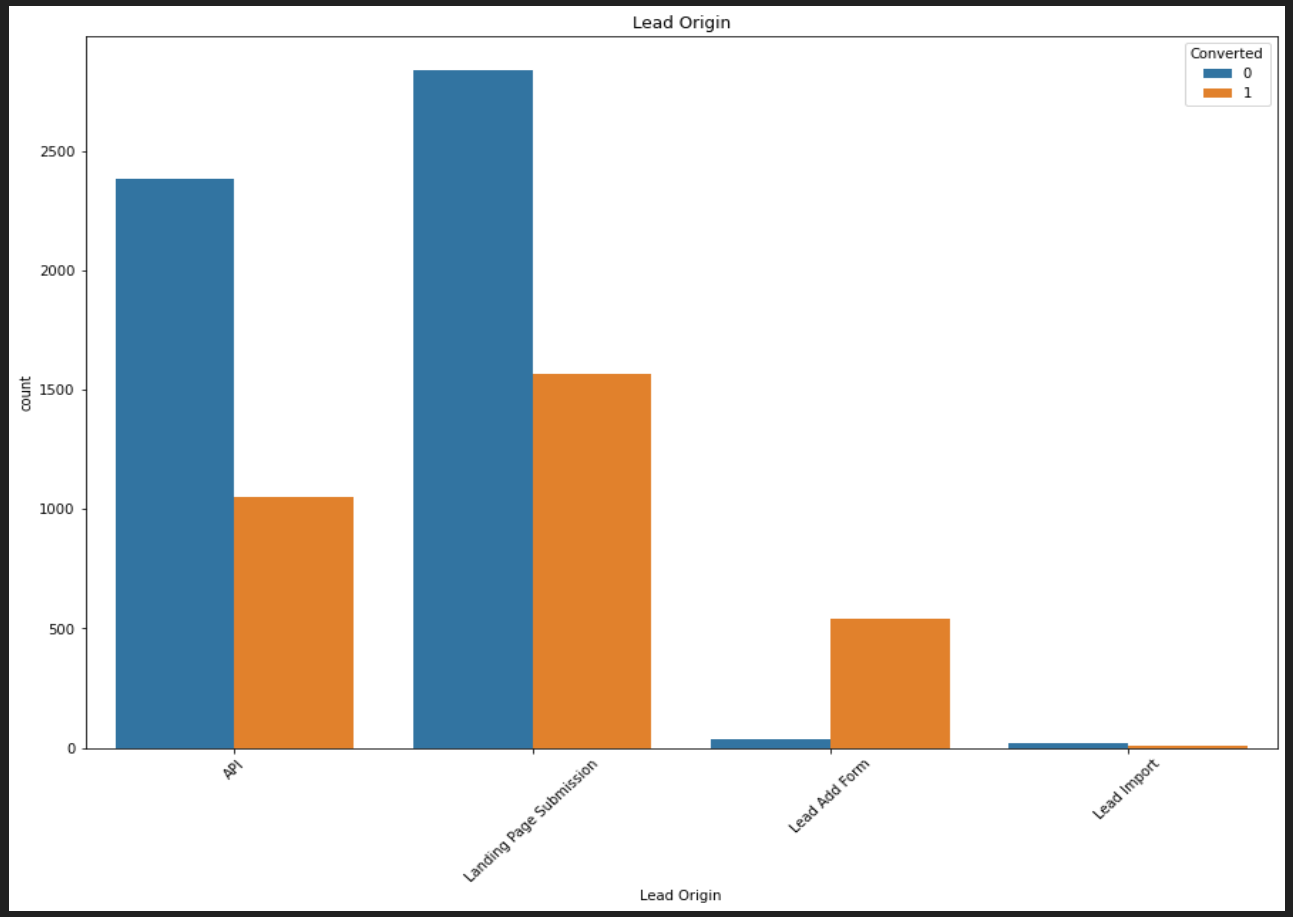


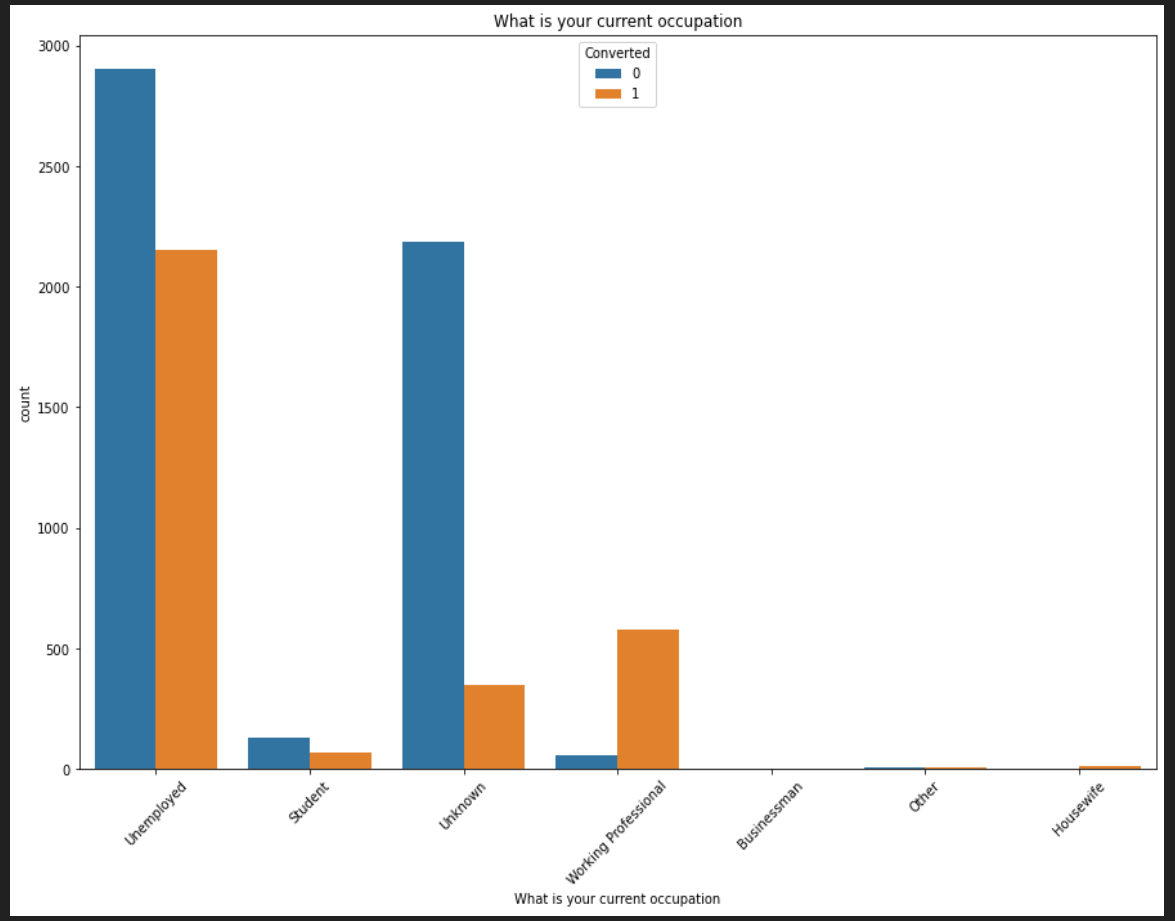



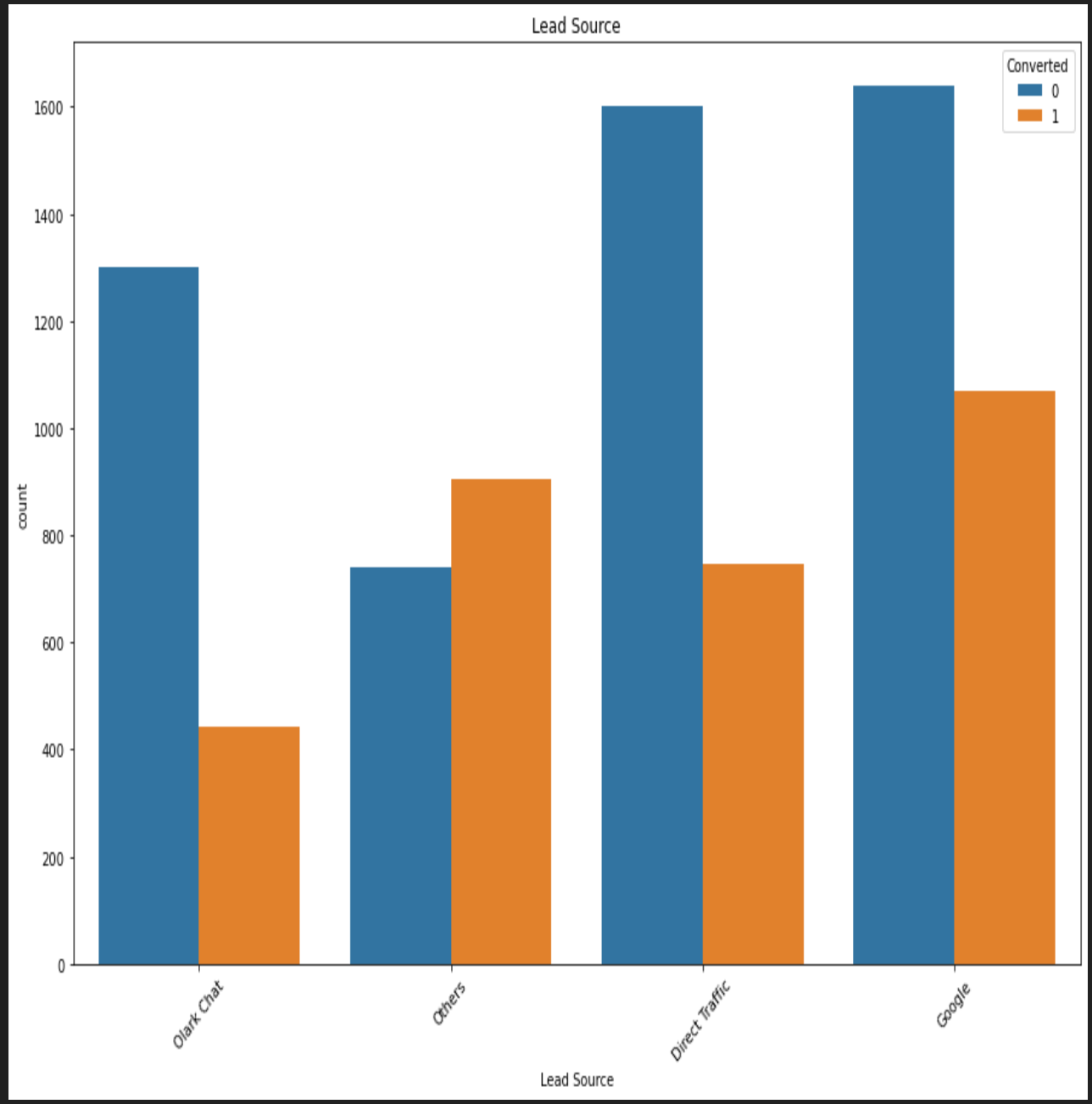



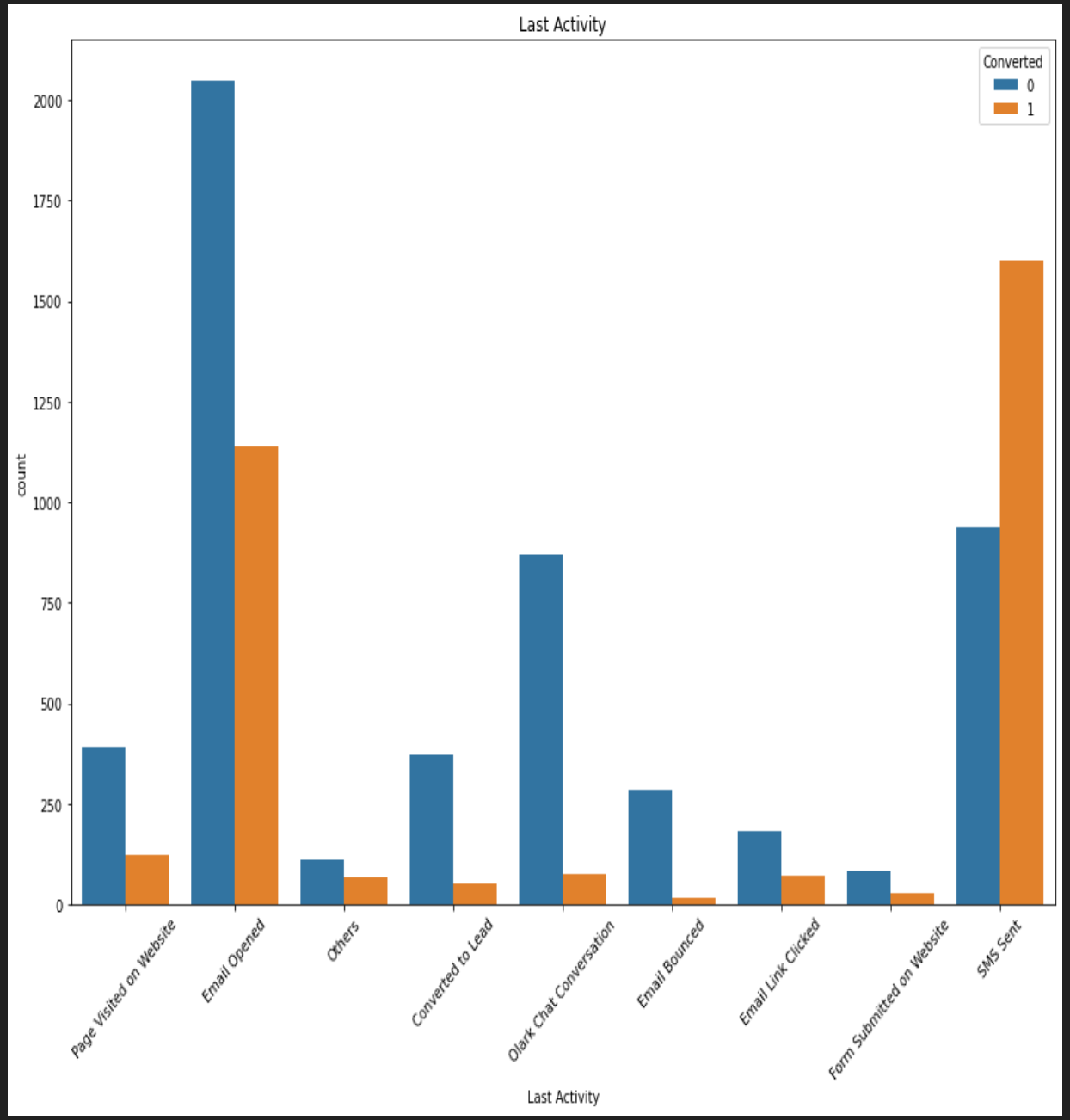
# Univariate Analysis

Notable visuals from univariate analysis



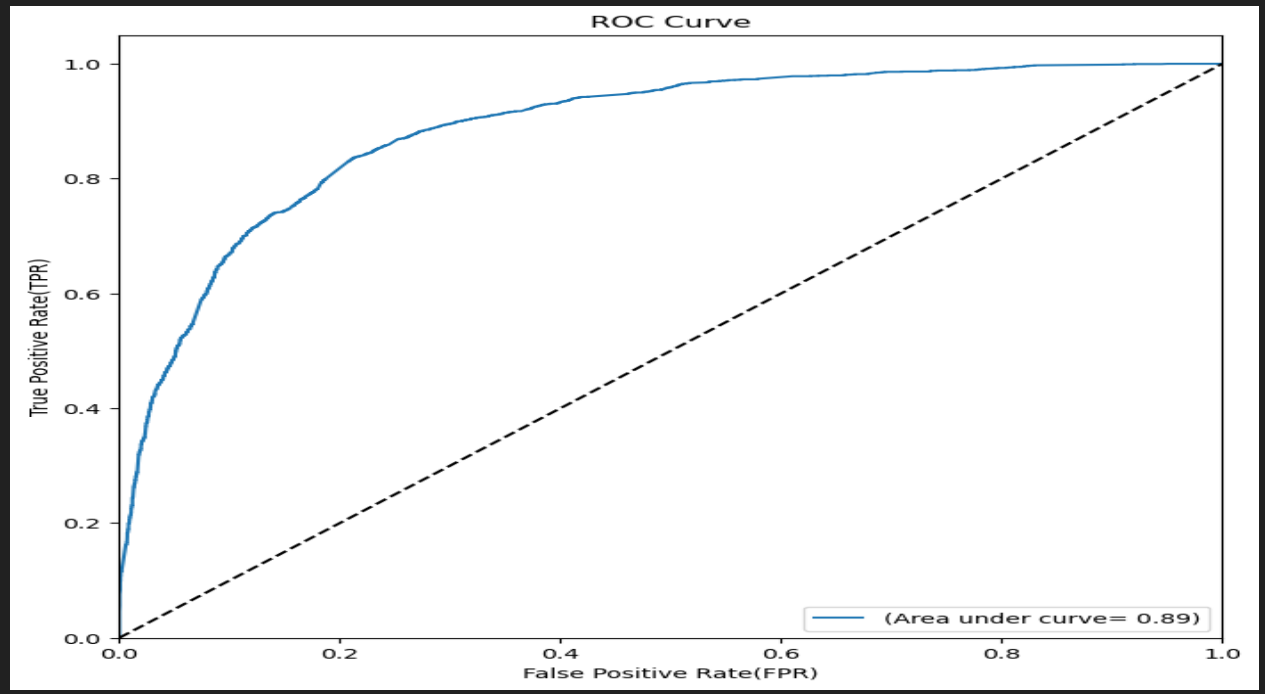




Lead Source
Converted
0
1
count
0
200
400
600
800
1000
1200
1400
1600
Olark Chat
Others
Direct Traffic
Google
Lead Source

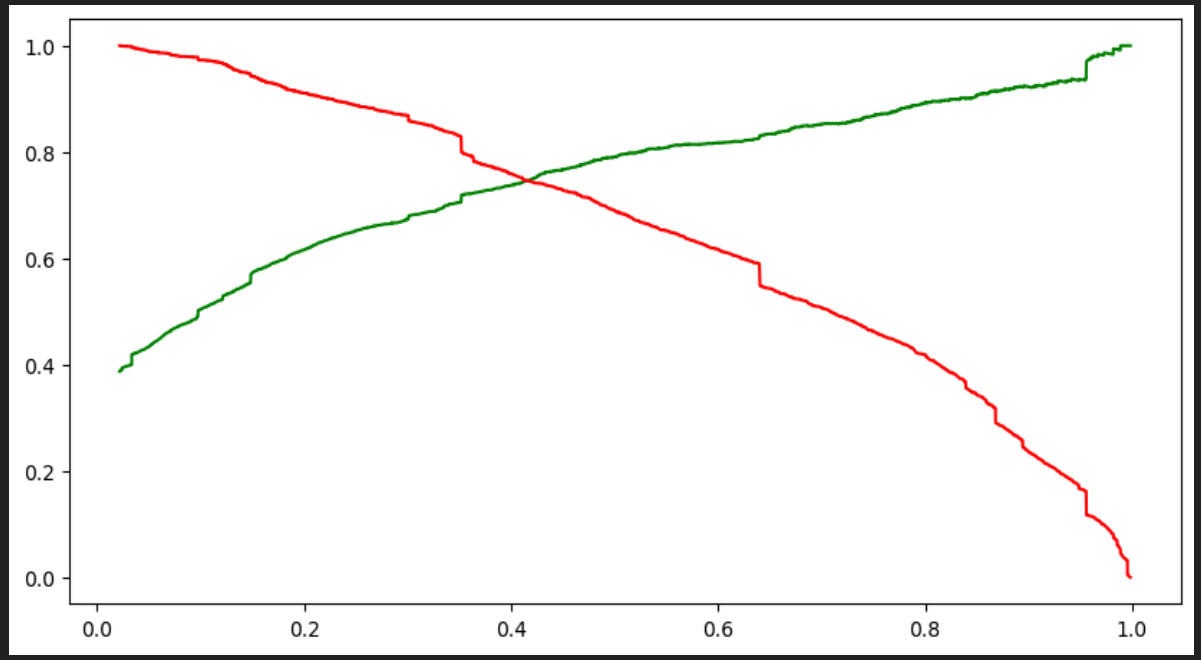

Last Activity
Converted
0
1
count
0
250
500
750
1000
1250
1500
1750
2000
Page Visited on Website
Email Opened
Others
Converted to Lead
Olark Chat Conversation
Email Bounced
Email Link Clicked
Form Submitted on Website
SMS Sent
Last Activity

# Model Building

- Logistic Regression model from Sklearn module of python was used to build Logistic Regression model
- RFE is used for initial selection of 20 features. By eliminating features based on p-values and VIF a stable model was built.
- On building a ROC curve we obtained 0.89 which is a good score indicating that the model is close to ideal.



- Now we need to set a threshold value for deciding Hot Leads based on the probabilities. Since CEO wanted a conversion rate of 80% we used Precision vs Recall view to find the optimum threshold.
- Based on the Precision vs Recall graph we need a Precision of 0.8, so threshold came close to 0.52.
- With a threshold of 0.52 we get a Precision of 0.8 and False Positive Rate was only 0.1 or 10%.
- So we decided that a threshold of 0.52 is ideal for our requirements of reaching a target of 80% conversion.



# RECOMMENDATIONS

- Based on the Logistic Regression Model used, Total Time Spent on Website, Lead Originated through Add Forms and Occupation Working Professionals have higher probability to drive conversion.

- So the company should work towards making website's UI intuitive, interesting and adding engaging content. The aim of the calling executive during initial calls should be to make the lead engage and spend time on website
- Try to contact working professionals after working hours, always take appointment for calling, have a friendly conversation by asking about their work, recommend and explain about courses available on X education which would improve their career. This needs to be done in multiple calls

# RECOMMENDATIONS

- Try to contact working professionals after working hours, always take appointment for calling, have a friendly conversation by asking about their work, recommend and explain about courses available on X education which would improve their career. This needs to be done in multiple calls and emails.
- Ensure most leads fill out the Add Forms for getting more details from them.
- Better Career prospects is a major driving factor for people to do up skilling courses. So X Education need to work on providing assistance in placements.