LEAD SCORING CASE STUDY

Group members:

Sayan Debnath, Saurabh Singh, Shubhankar Sardeshmukh

PROBLEM STATEMENT

- X Education sells online courses to industry professionals.
- 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 successful, 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.

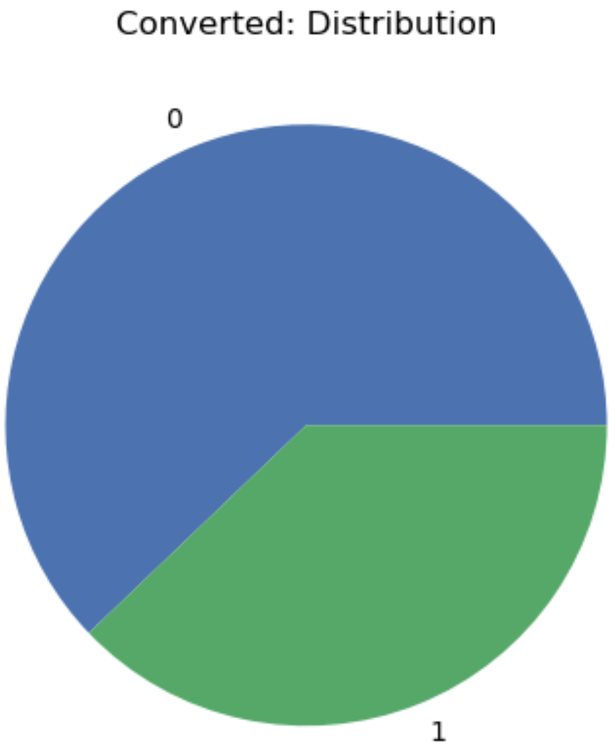
BUSINESS OBJECTIVE

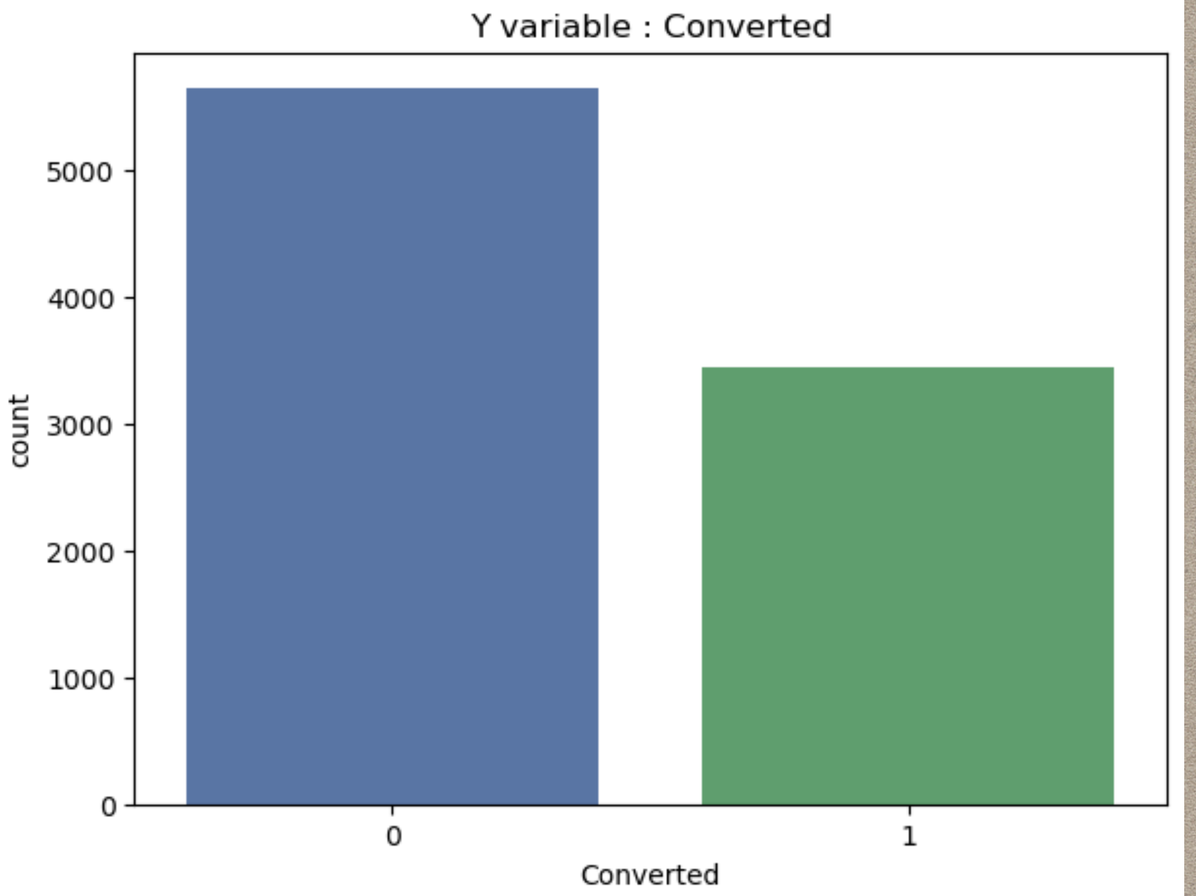
- Which leads are the most promising, 'hot leads'. They want us to give every lead a score between 0 and 100 to identify them.
- A model to identify the 'hot leads'.
- Deployment of the model for future use.
- The CEO wants to achieve a lead conversion rate of 80%

PROBLEM APPROACH:

- Importing the data.
- Inspecting the data-frame.
- Data Preparation.
- EDA(Exploratory Data Analysis).
- Dummy Variable Creation.
- Train-Test Split.
- Feature Scaling.
- Correlations.
- Model Building (RFE, R-squared, p-value and VIFs).
- Model Evaluations.
- Making predictions on the test set.

EXPLORATORY DATA ANALYSIS (EDA)





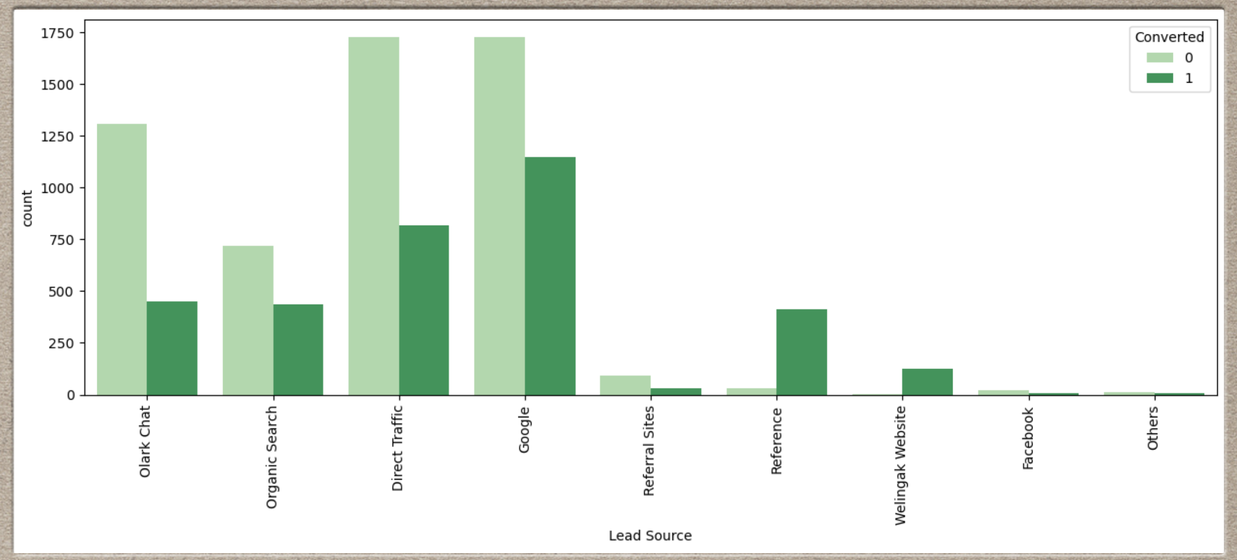
FEATURE: CONVERTED (TARGET VARIABLE)

LEAD CONVERSION RATE : 38%

LEAD SOURCE:

1. ‘Google’ and ‘Direct Traffic’ Produces the Most Number of Leads but the Conversion Rate Is Low.

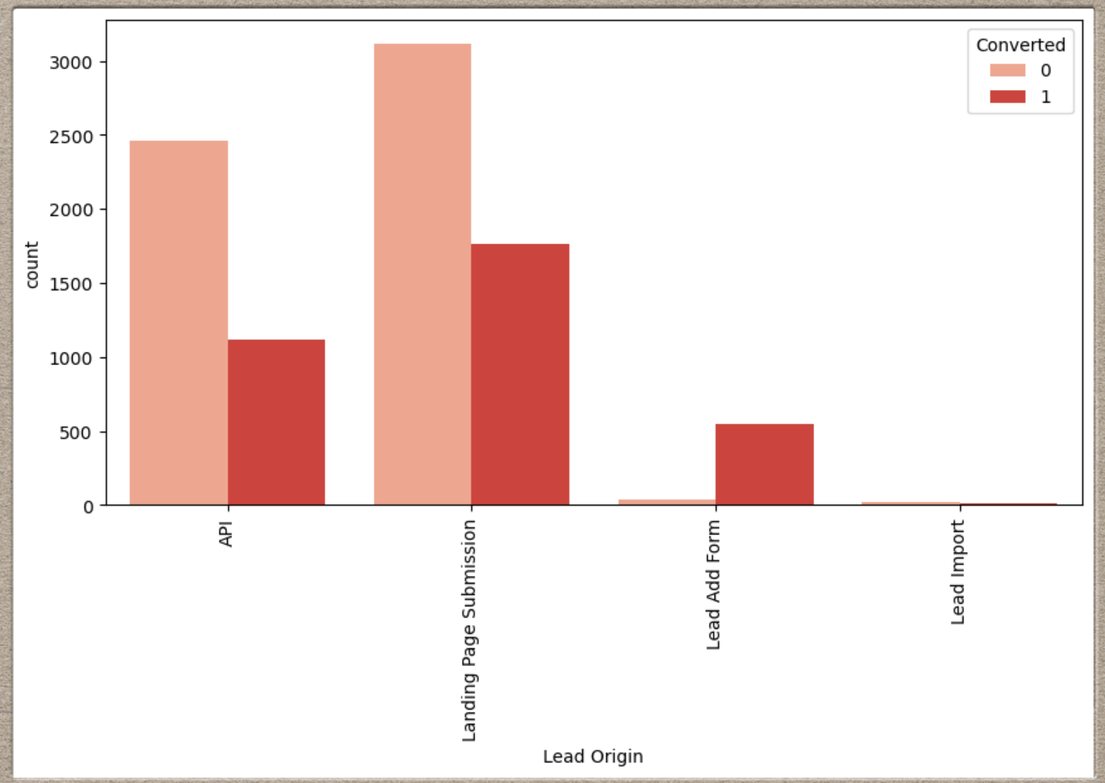
2. ‘Reference’ and ‘Welingak Website’ Have High Conversion Rates.



LEAD ORIGIN:

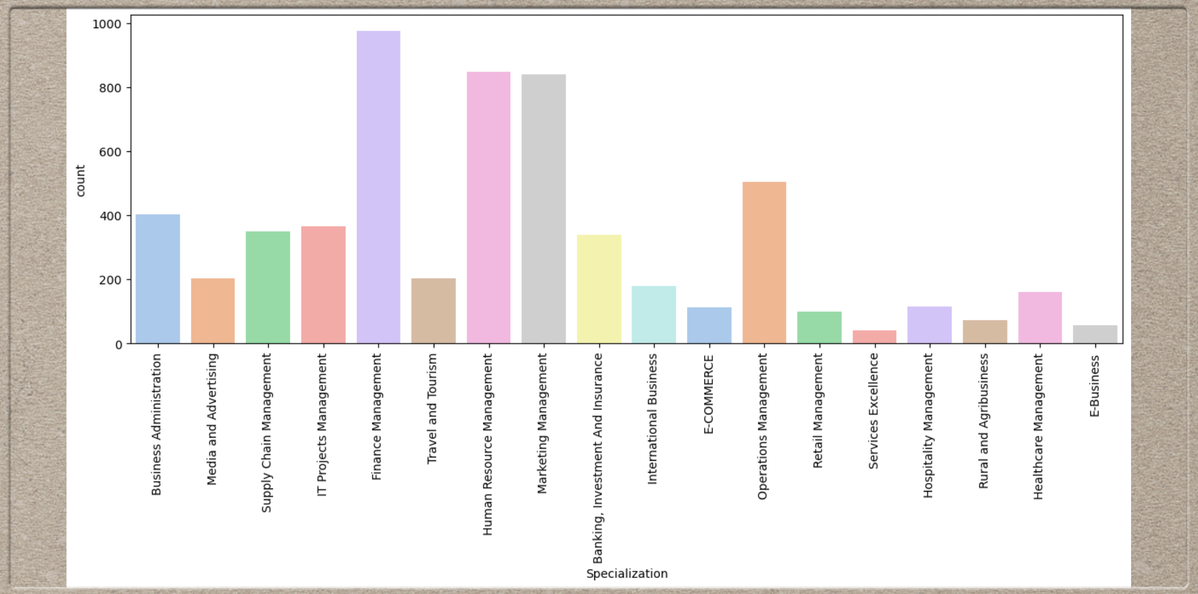
1. ‘Landing Page Submission’ and ‘API’, Are the Highest 2 Categories in Terms of Quantity.

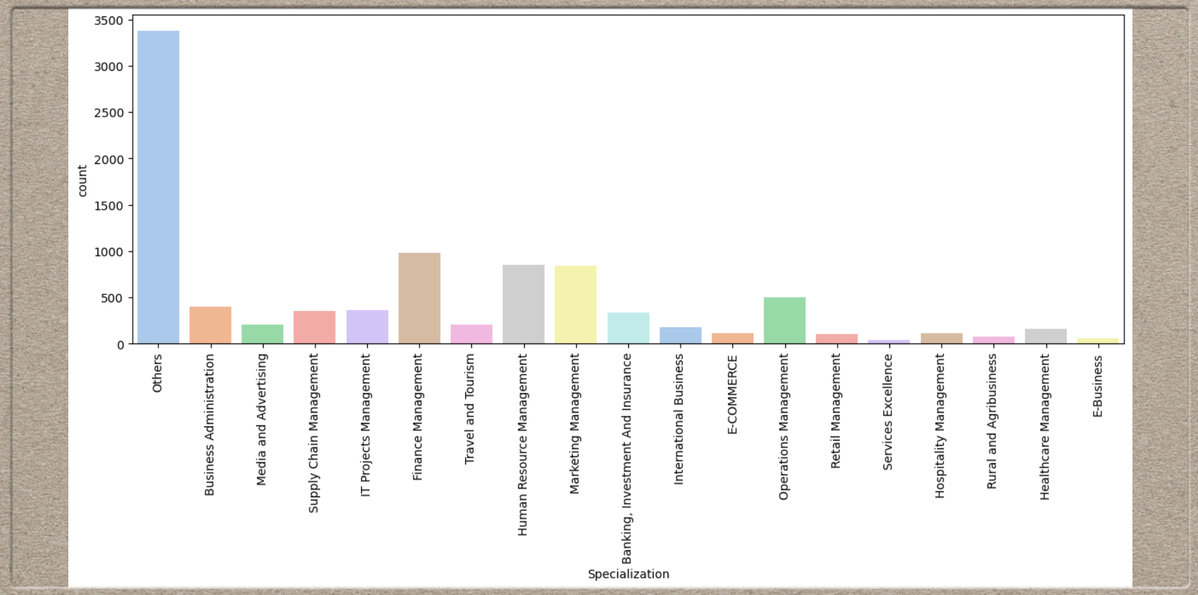
2. Leads From 'Lead Add Form' Have the Least Amount of Rejection and the Most Amount of Conversion.

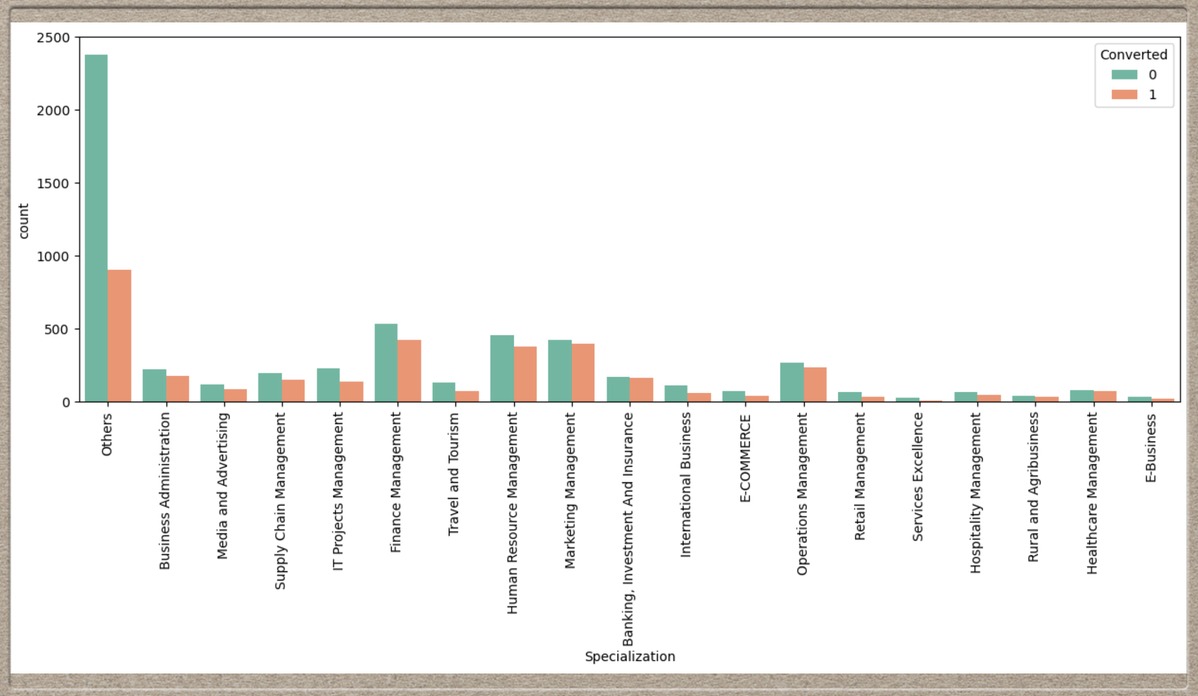


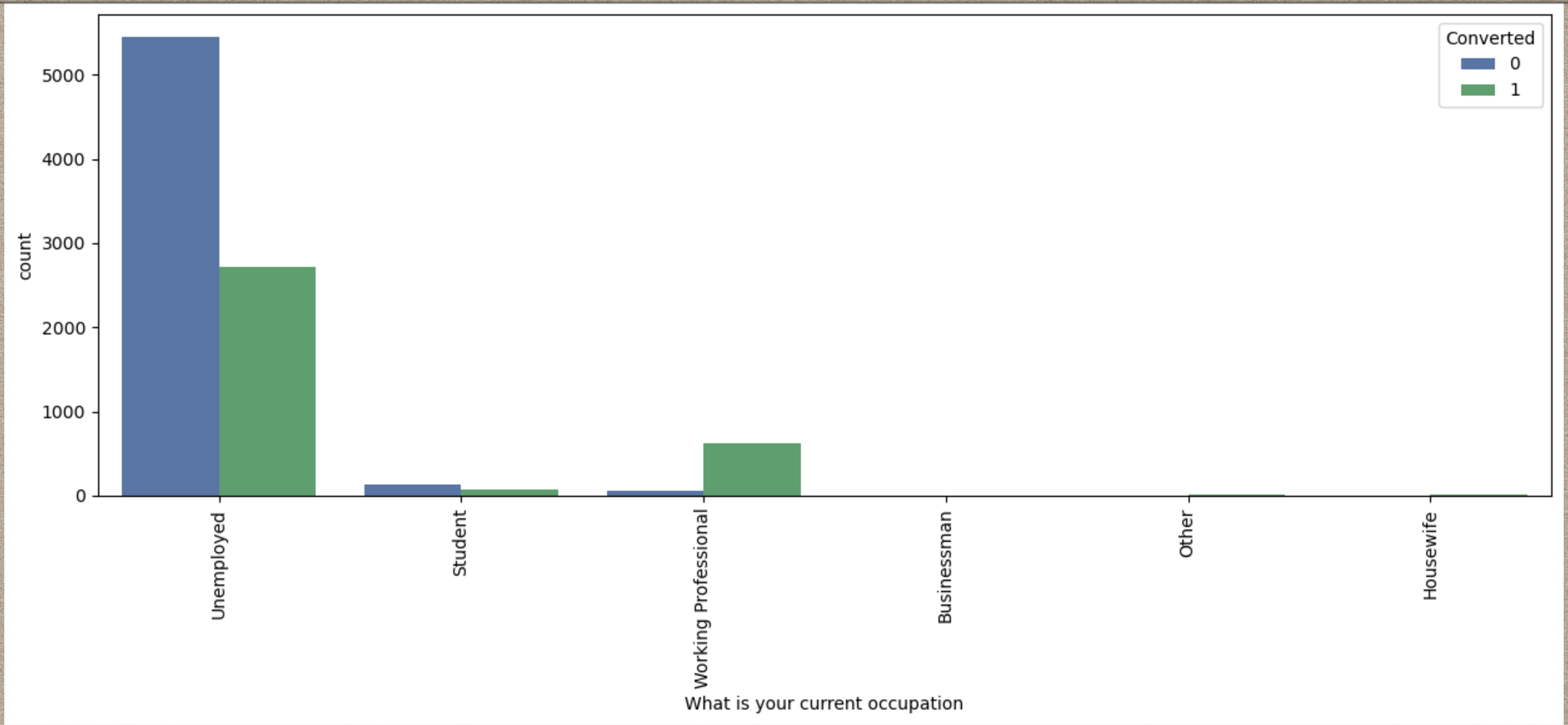
FEATURE: SPECIALIZATION

- **Top Right:** ‘Specialization’ without ‘Others’.
- **Bottom Left:** ‘Specialization’ with ‘Others’.
- **Bottom Left:** ‘Specialization’ with the target variable ‘Others’.
- Leads who do not mention their specialisation have a lower chance of conversion.



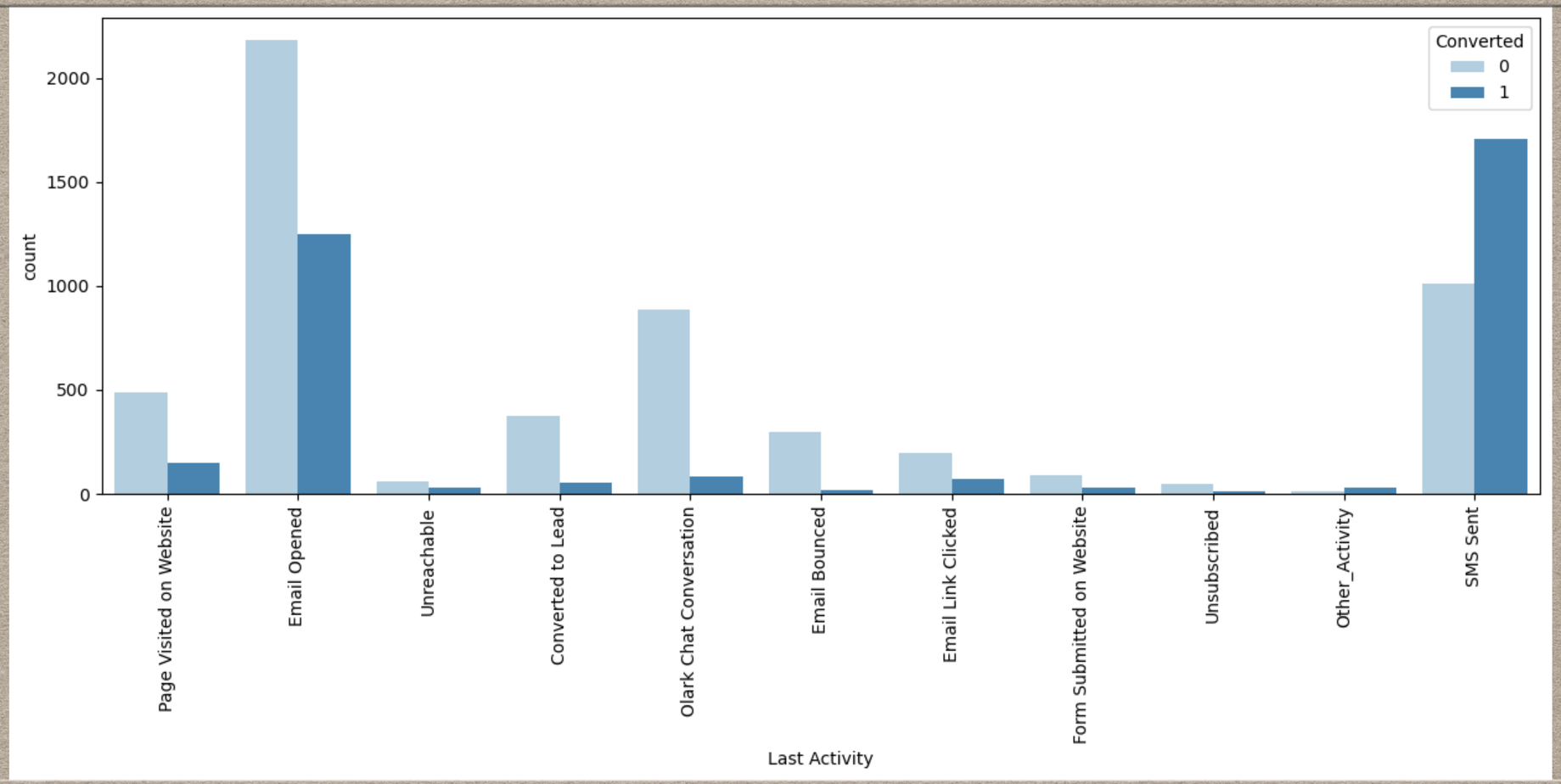






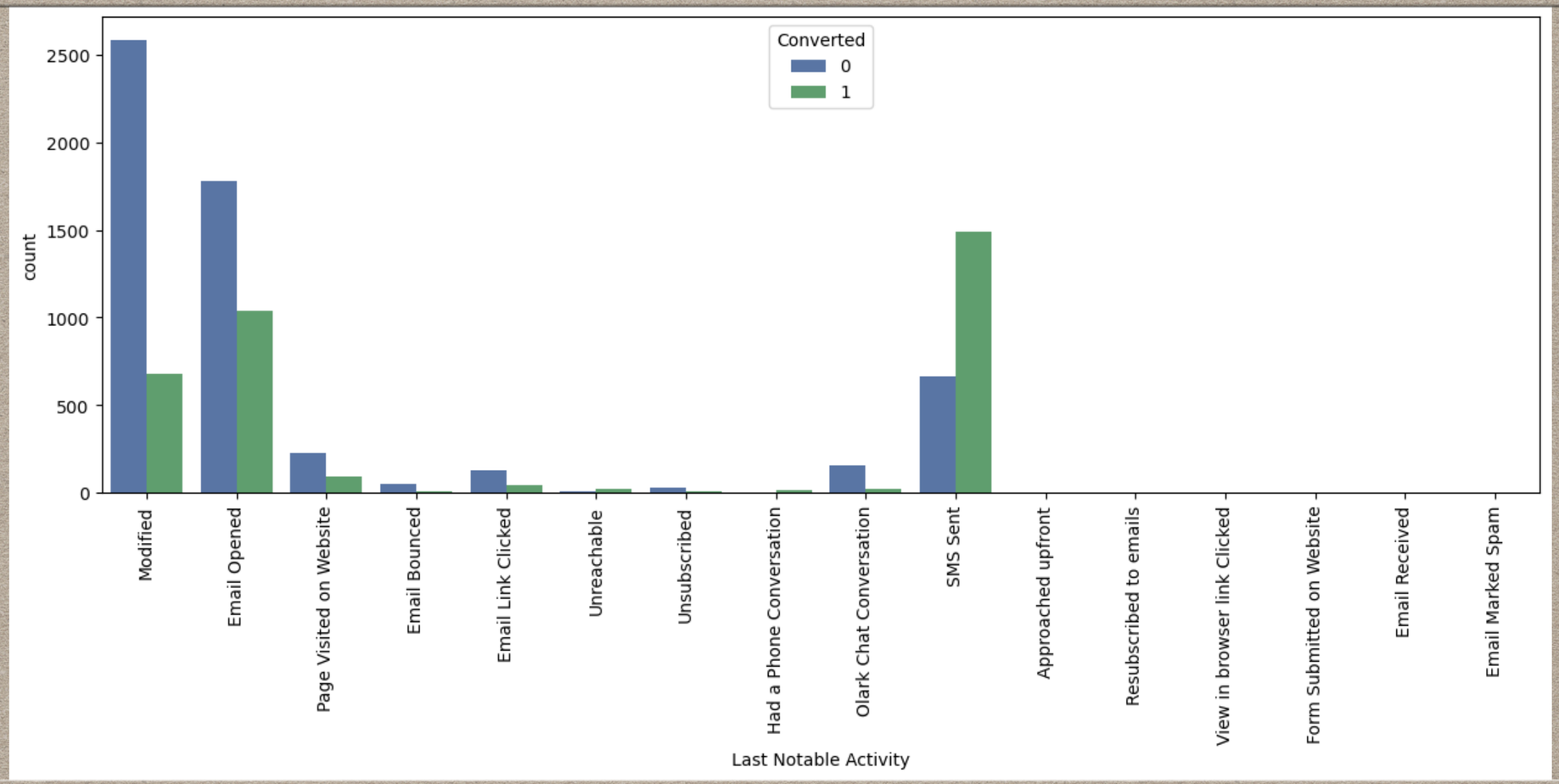
FEATURE : WHAT IS YOUR CURRENT OCCUPATION

1. 'UNEMPLOYED' LEADS ARE THE HIGHEST IN TERMS OF QUANTITY OF LEAD.
2. CONVERSION RATE HIGHEST FOR 'WORKING PROFESSIONAL'.



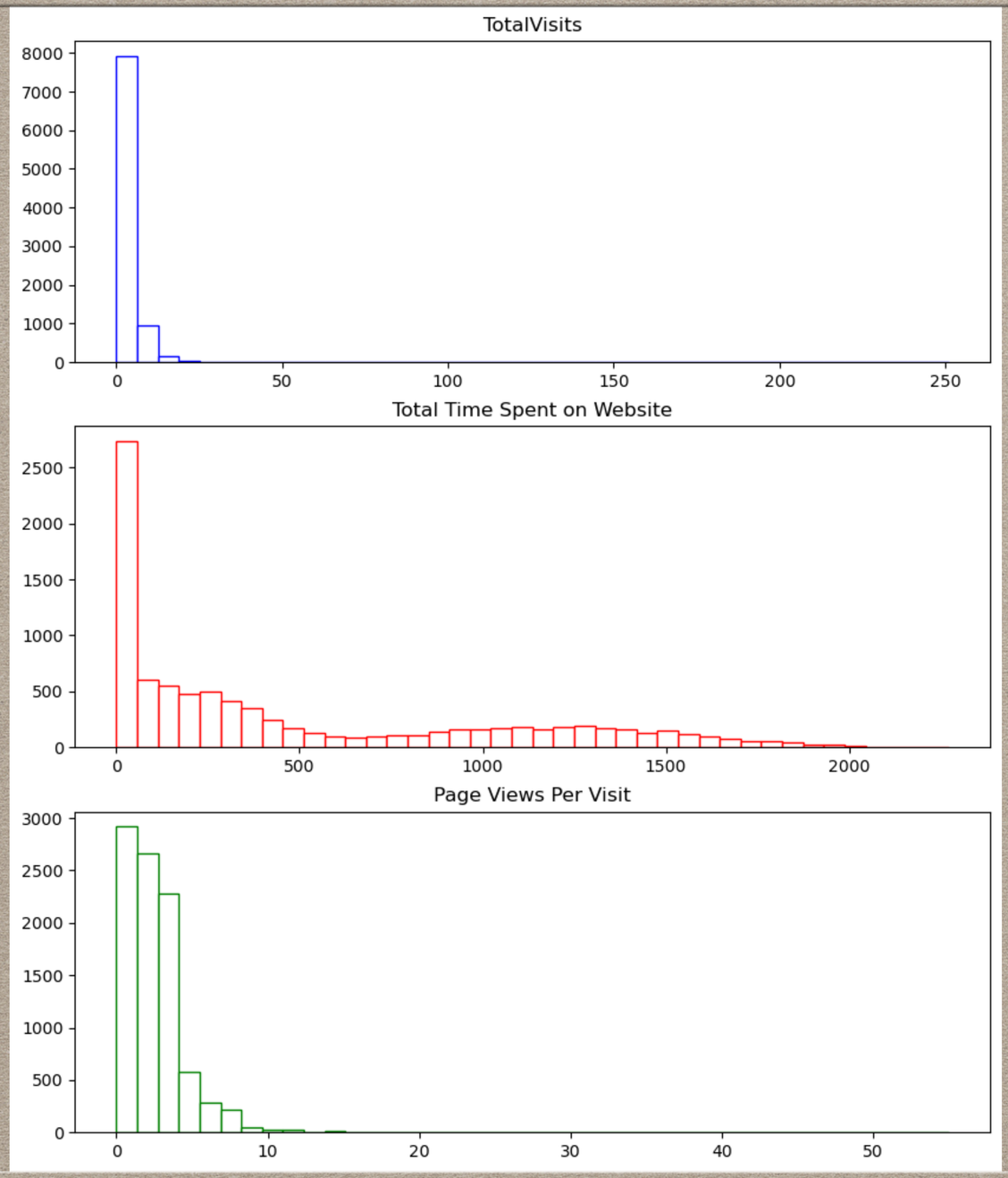
FEATURE : LAST ACTIVITY

'EMAIL OPENED' HAVE THE HIGHEST LEADS FOLLOWED BY 'OLARK CHAT CONVERSION' BUT THE RATE OF REJECTION IS HIGHER. 'SMS SENT' HAVE THE HIGHEST CONVERSION RATE.



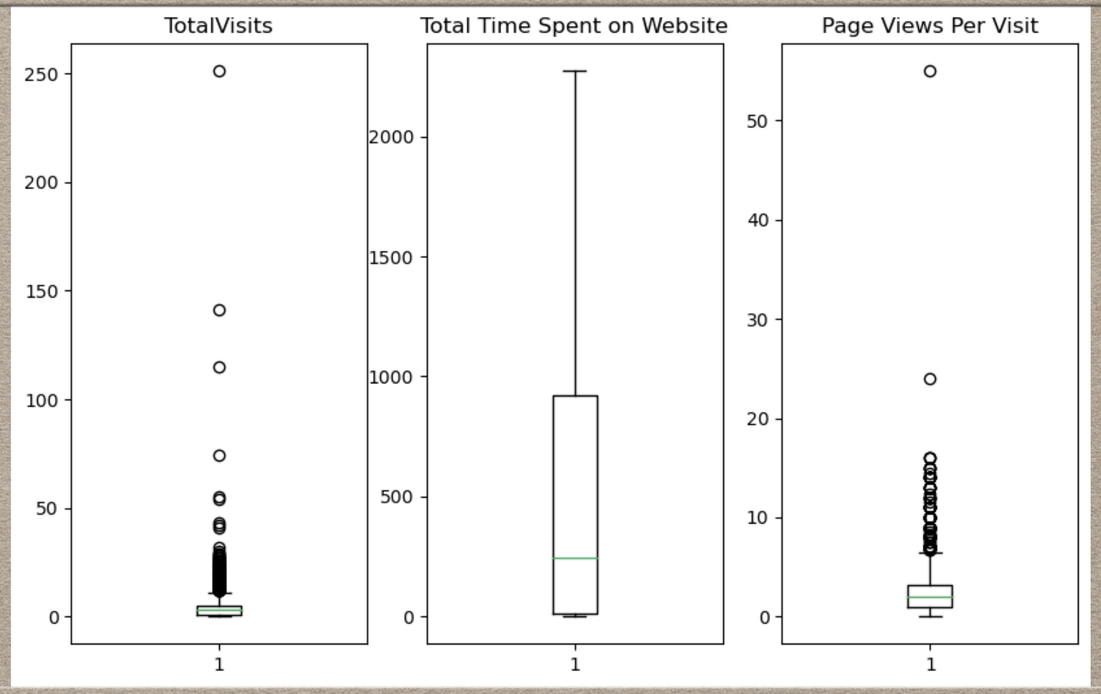
FEATURE : LAST NOTABLE ACTIVITY

LEADS WHO SELECT 'MODIFIED' AND 'EMAIL OPENED' ARE THE HIGHEST FOLLOWED BY SMS SENT. LEADS WHO SELECT 'SMS SENT' HAS THE HIGHEST CONVERSION RATE. THEY HAVE THE HIGHEST PROBABILITY OF CONVERTING.



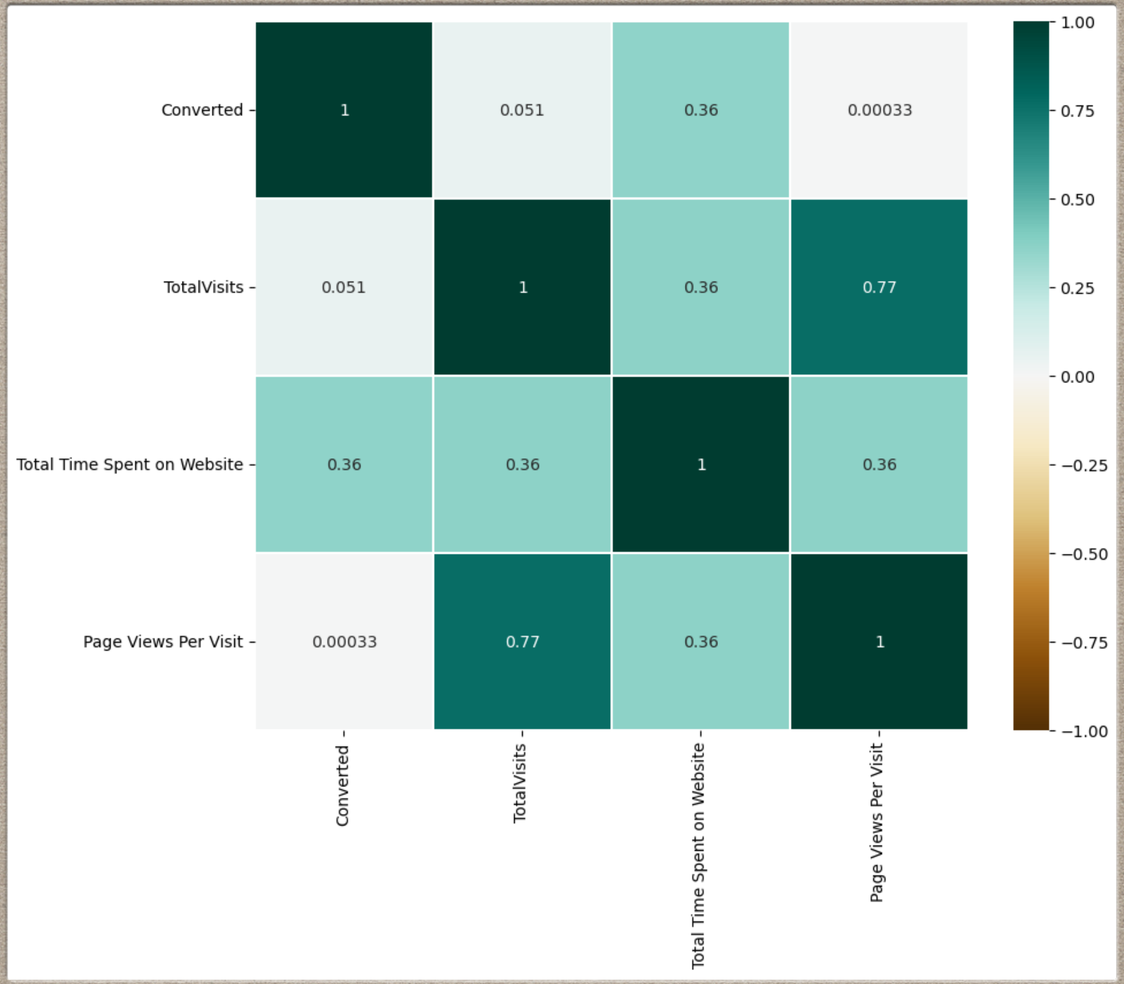
NUMERICAL VARIABLES

- 'TotalVisits' and 'Page Views Per Visit' have outliers. We can cap them to 95 percentile.
- 'Total Time Spent on Website' doesn't show any outliers.
- Leads who spend more time on the website have higher possibility of converting.



NUMERICAL VARIABLES: CORRELATION

We see the correlation between the variables: TotalVisits, Total Time Spent on Website, Page Views Per Visit and the target variable Converted.



DATA CONVERSION

- Total number of rows after data cleaning : 9074
- Total number of remaining features before creating dummy variables : 13
- Total number of features after creating dummy variables : 66
- Numerical variables are normalised.

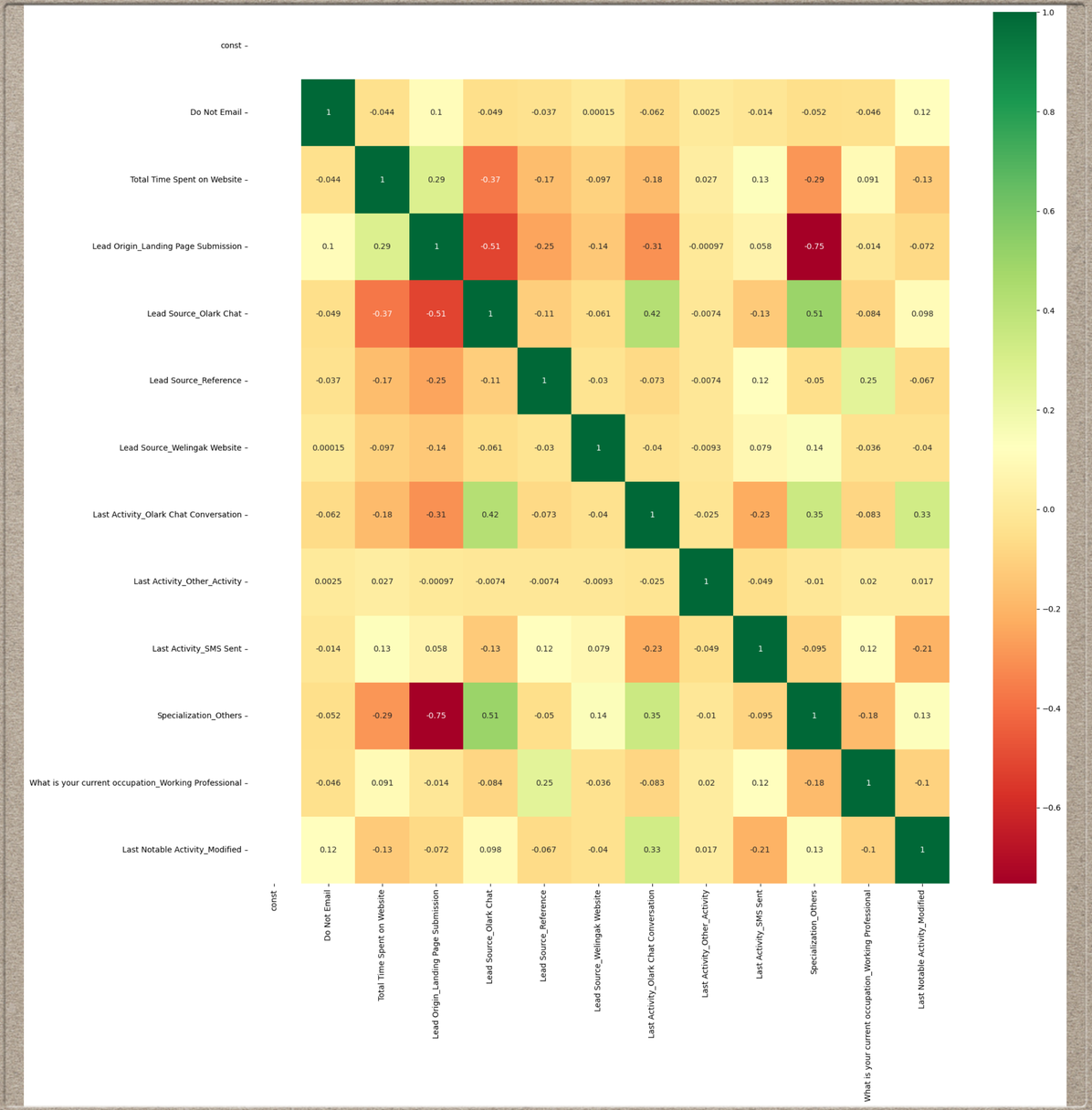
MODEL BUILDING

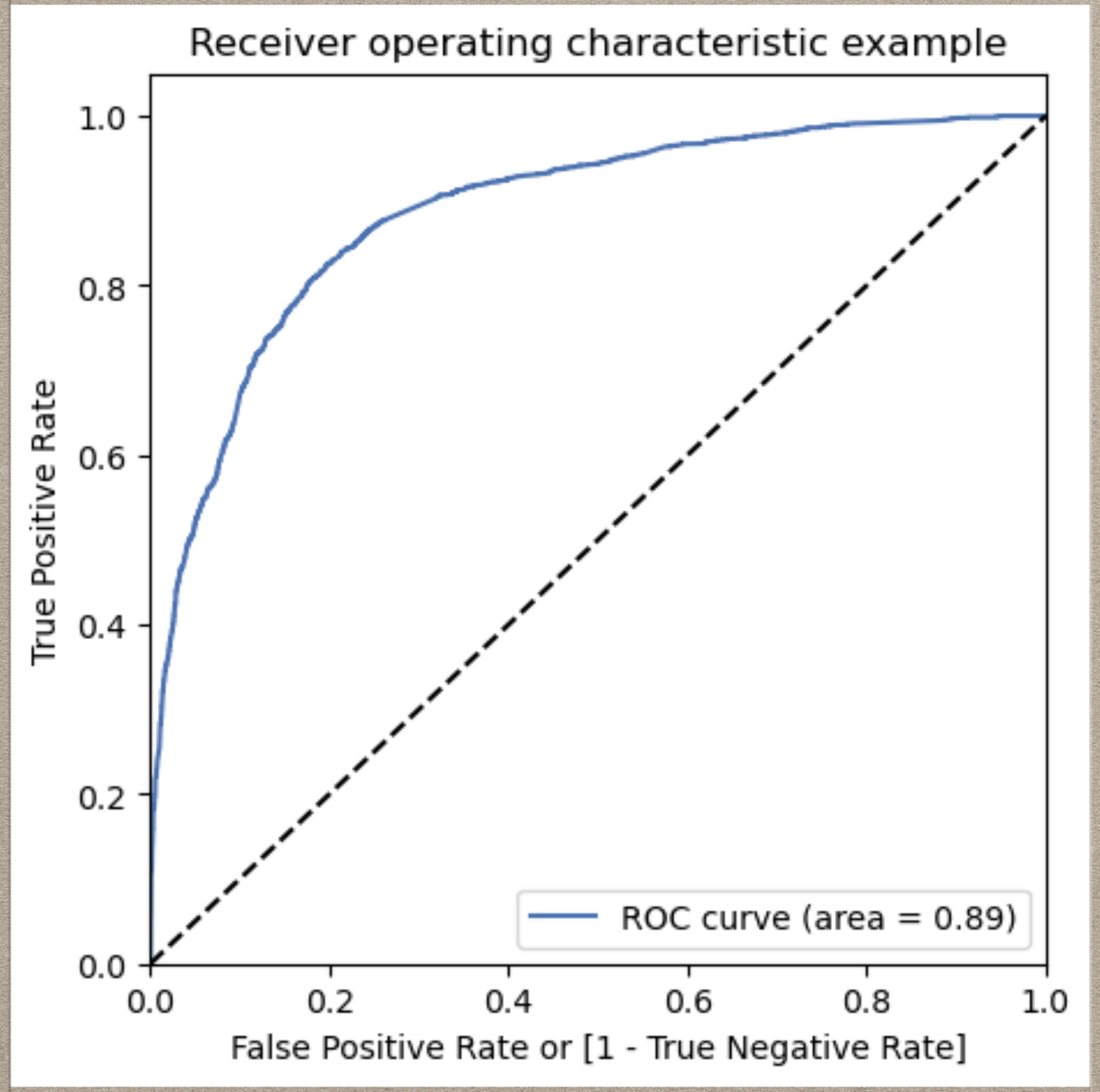
- Splitting the data into Training and Testing Sets.
- The first step for regression is performing a train-test split.
- We have selected the 70-30 ratio.
- Use RFE for feature selection.
- Running RFE with 20 variables as output.
- A total of 9 models were created.
- Final feature count : 12
- Overall Accuracy of the Model : 81%

CORRELATION HEAT MAP FOR THE FINAL 12 FEATURES

LIST OF ALL THE FINAL FEATURES ARE GIVEN BELOW FOR REFERENCE

- Specialization_Others
- Lead Source_Olark Chat
- Last Notable Activity_Modified
- Lead Origin_Landing Page Submission
- Last Activity_Olark Chat Conversation
- Last Activity_SMS Sent
- Total Time Spent On Website
- Lead Source_Reference
- What is your current occupation_Working Professional
- Do Not Email
- Lead Source_Welingak Website
- Last Activity_Other Activity

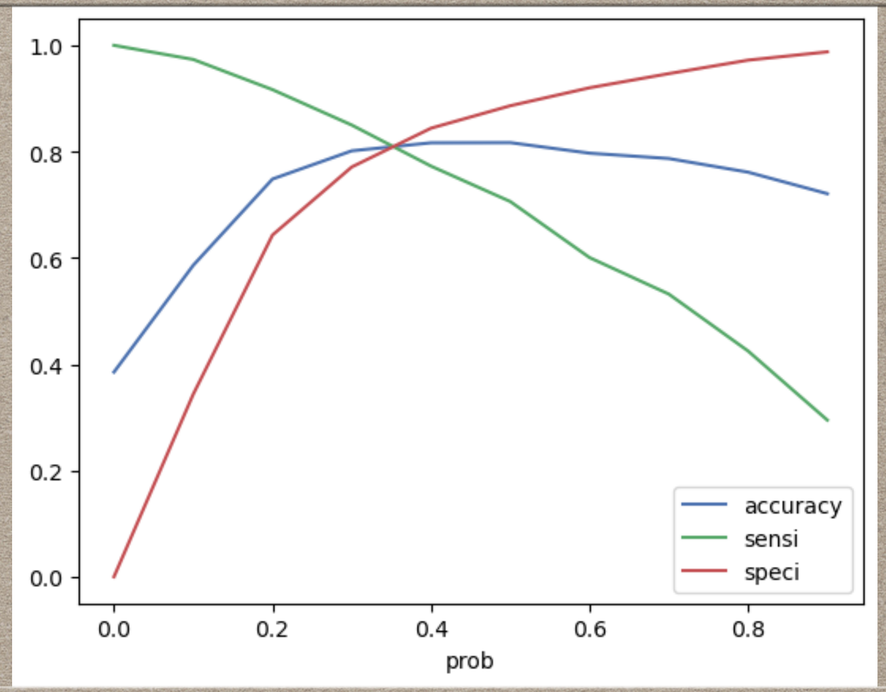




ROC CURVE

We have higher area (0.89) under the ROC curve.

We have accepted approximately 0.34 as the optimum point to take as a cut-off probability.



FINAL OBSERVATIONS

- Optimum point for cut-off is **0.34**
- **Training Data Observation:**
- Accuracy : **81.0 %**
- Sensitivity : **81.7 %** (Recall)
- Specificity : **80.6 %**
- **Testing Data Observation:**
- Accuracy : **80.4 %**
- Sensitivity : **80.4 %** (Recall)
- Specificity : **80.5 %**

CONCLUSION AND RECOMMENDATION

- Leads who do not mention their **specialization** (others,select,etc.) have a lower chance of conversion.
- Leads from sources **"Welingak Website"** and **"Reference"** are more likely to get converted.
- **"Working Professionals"** are more likely to get converted.
- Leads who spend more time on the website are more likely to get converted.
- Leads from the source **"Olark Chat"** are more likely to get converted.
- Leads with **"Last Activity"** being **"SMS Sent"** are more likely to get converted.
- Leads who chose **"Do not Email"** as **"yes"** are less likely to get converted.
- Leads who's lead origin is **"Landing page Submission"** are less likely to get converted.