Lead Score Case Study

By Sandeep Suman Pradhan

Problem Statement

- X education sells online courses to industry professionals
- 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
- X education gets a lot of leads however the lead conversion is poor only about 30% of the leads get 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 and the sales team then would focus on converting the leads rather than focus on everyone

Business Objective:

- X education wants to find out more promising lead such that the lead conversion is good (‘Hot Leads’)
- For this purpose a model should be built that identifies ‘Hot Leads’
- Relieve work effort of the sales team by deploying the model for future use to avoid calls to everyone

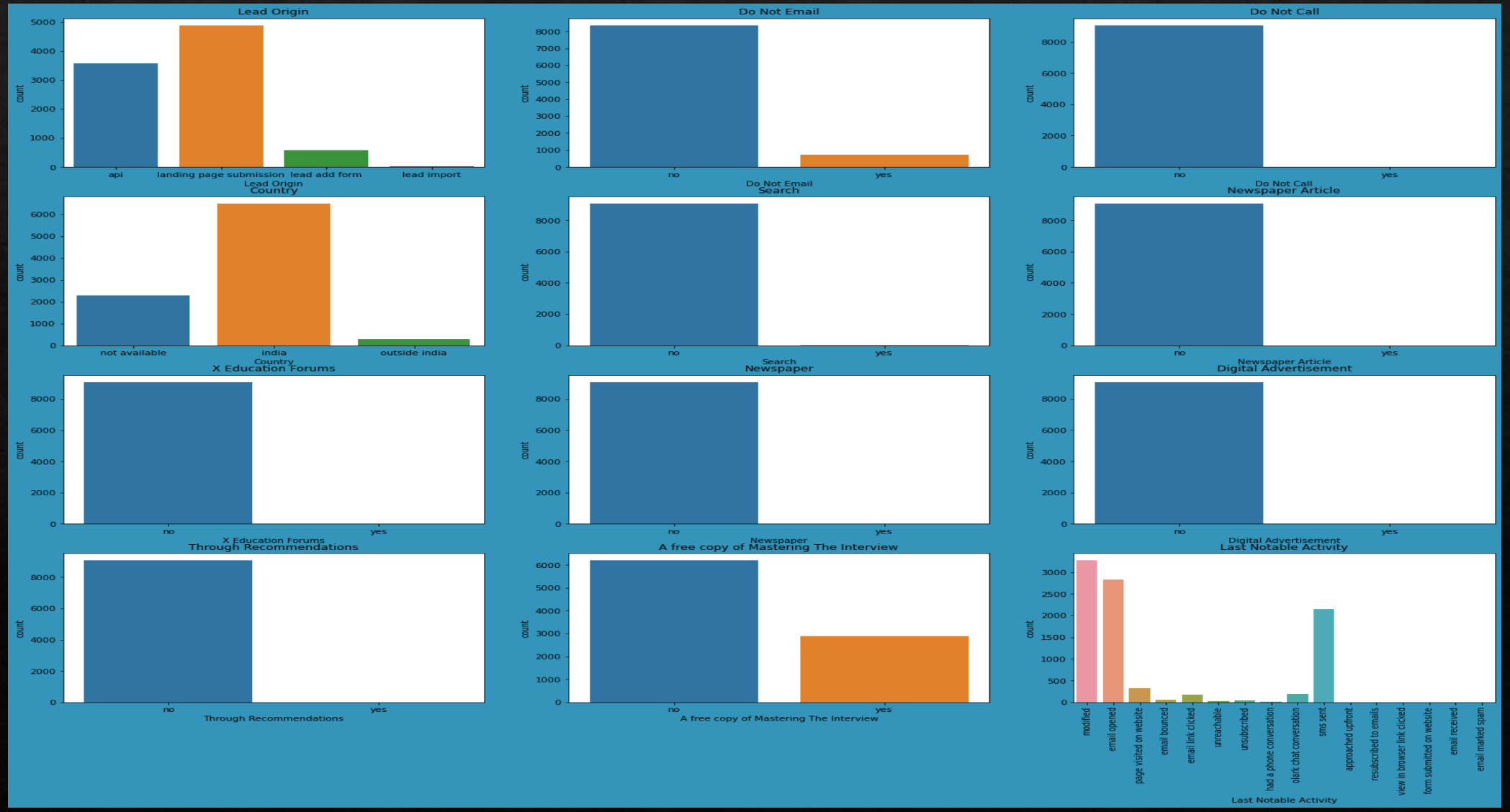
Solution Process

- **Data cleaning , understanding and manipulation**
  - Handle duplicate data
  - Check and handle missing and nan values
  - Drop columns if they contain large amount of missing values which will not be useful for analysis
  - Impute values if required
  - Handle outliers if any
- **Exploratory Data Analysis(EDA)**
  - Univariate Data Analysis
  - Bivariate Data analysis
- **Feature scaling and dummy variable encoding**
- **Logistic regression used for model building and prediction**
- **Validation of the model**
- **Model presentation**
- **Conclusion and Recommendation**

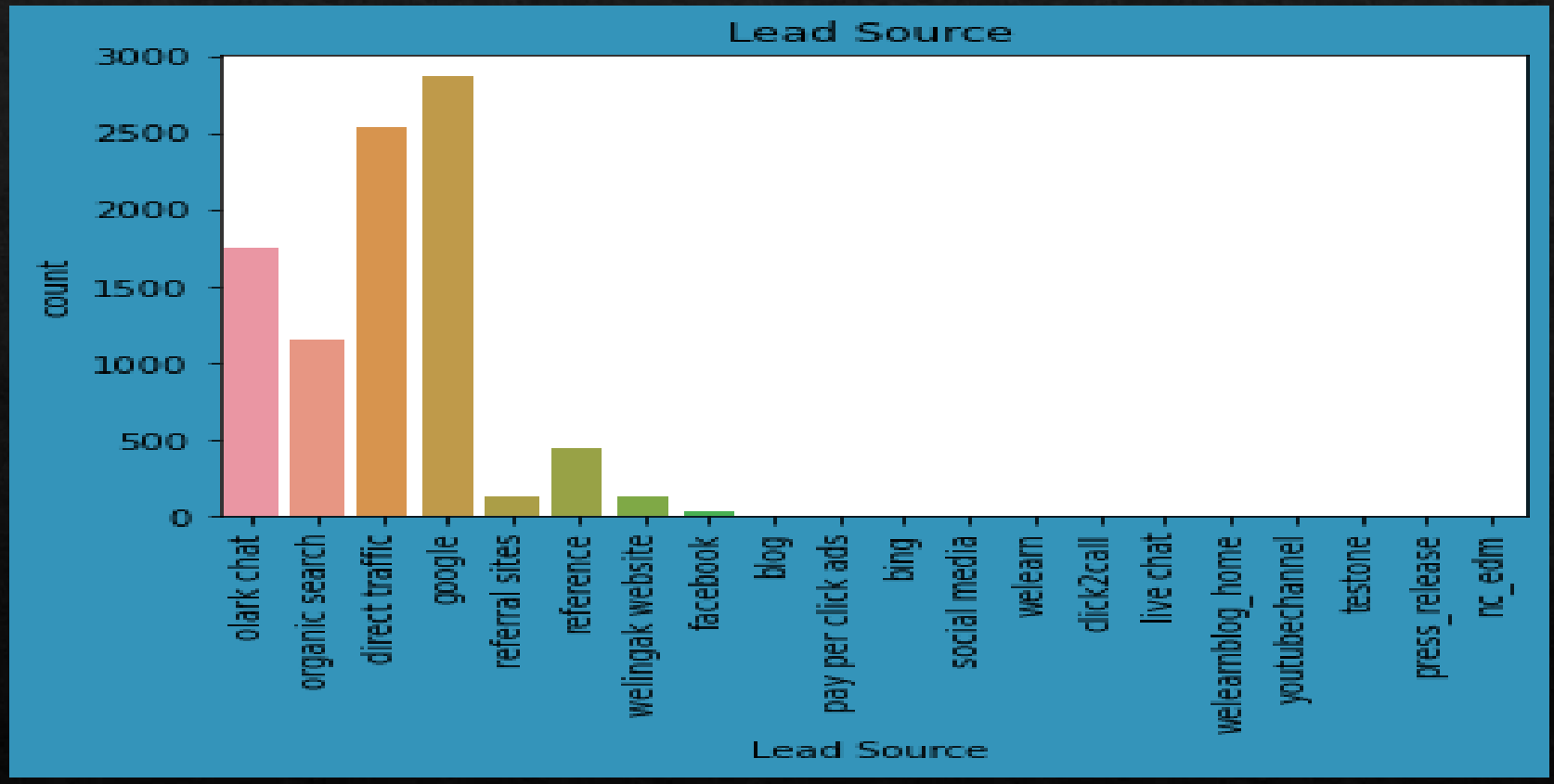
Data Manipulation

- Total number of rows=37, Total number of columns=9240
- Columns with single unique values such as ‘Magazine',’ Receive More Updates About Our Courses’, 'Update me on Supply Chain Content’, 'Get updates on DM Content’, 'I agree to pay the amount through cheque’ were dropped as they add no value to the analysis
- Removed ‘Prospect ID’ and ‘Lead number’ columns as they added no value to the analysis
- Checked the value counts for some object type variables and decided to drop features with not enough variance such as ‘Do Not Call’ , ‘What matters most to you in choosing Course’, ‘Search’ , ‘Newspaper Article’, ‘X education Forums’, etc
- Dropped columns with more than 35% missing values

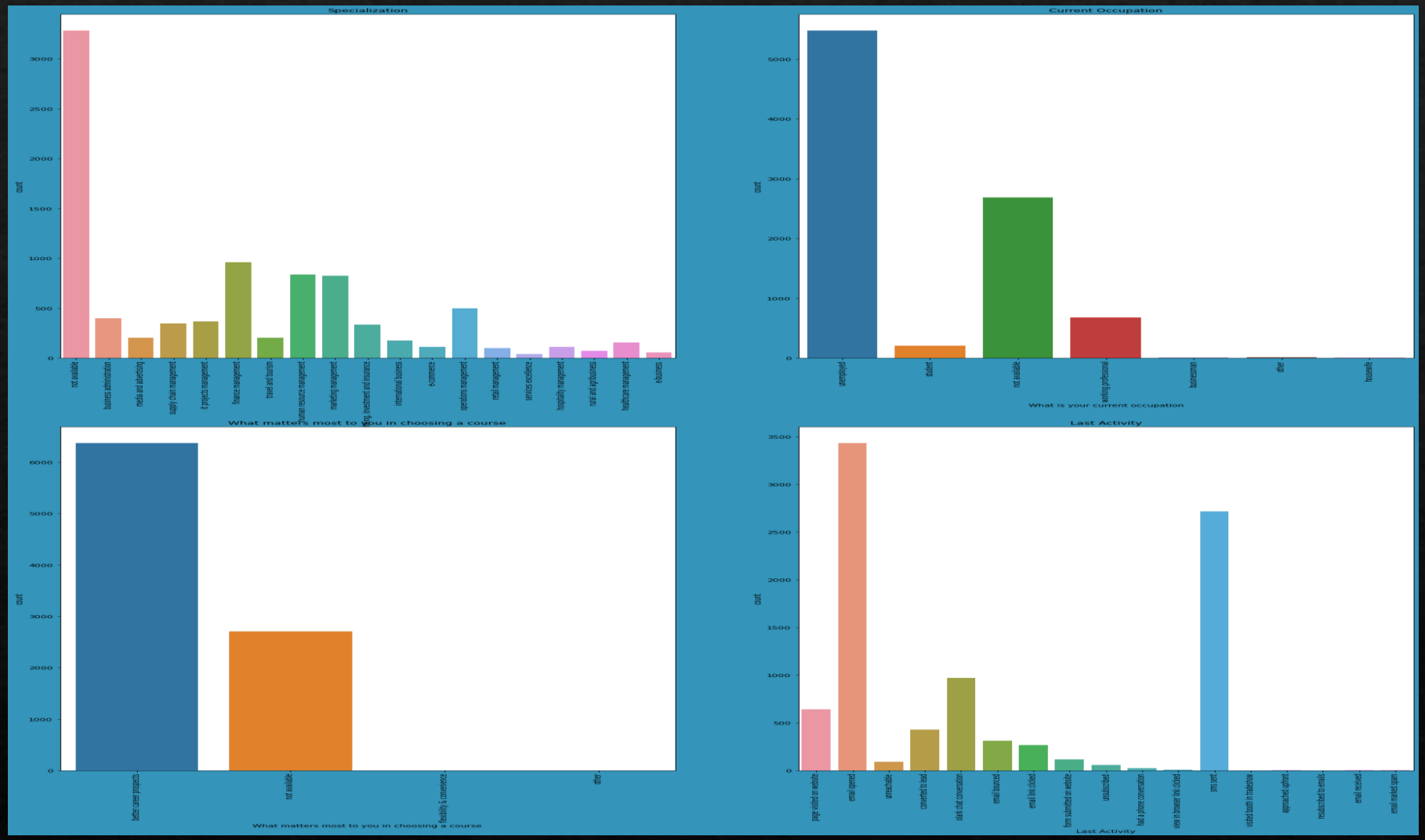
Exploratory Data Analysis(EDA)



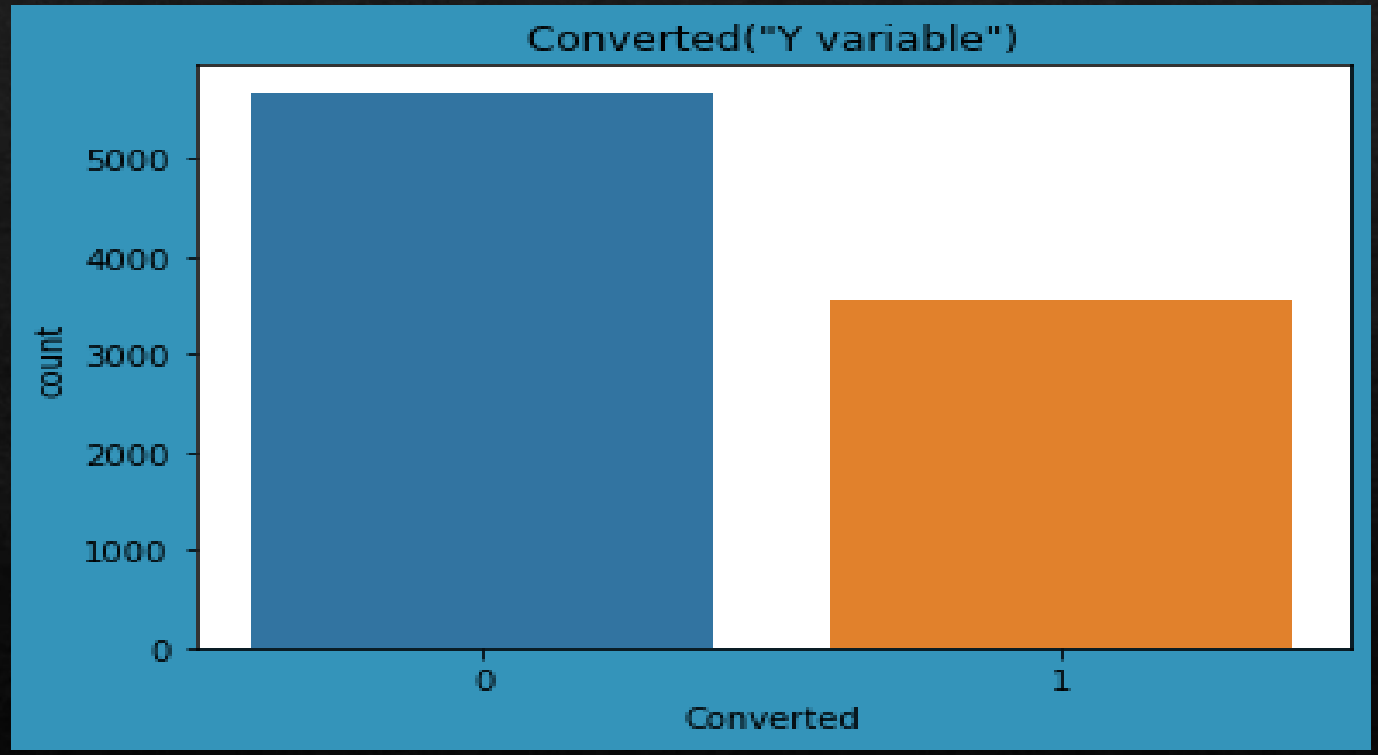
Exploratory Data Analysis(EDA)



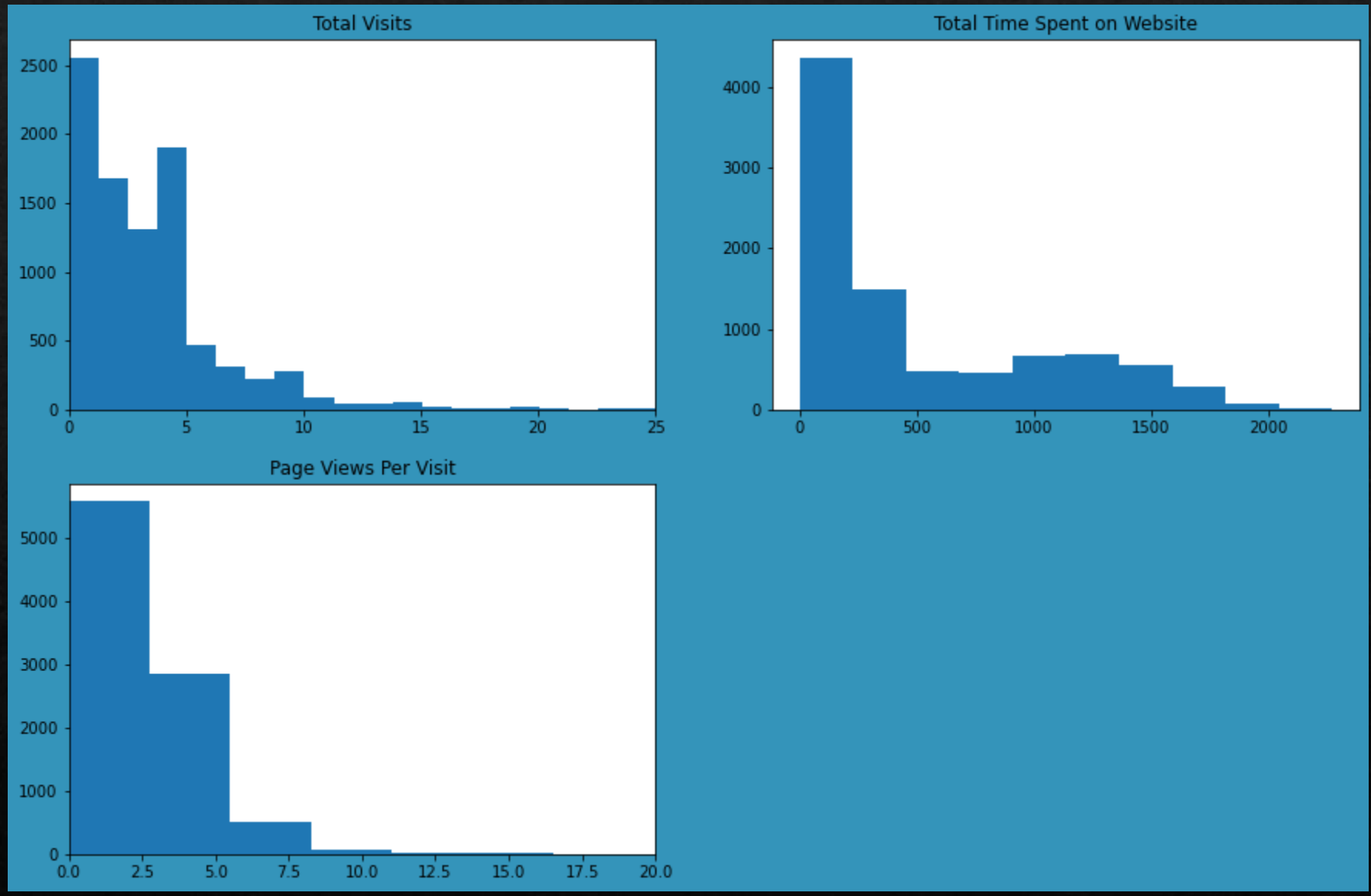
Exploratory Data Analysis(EDA)



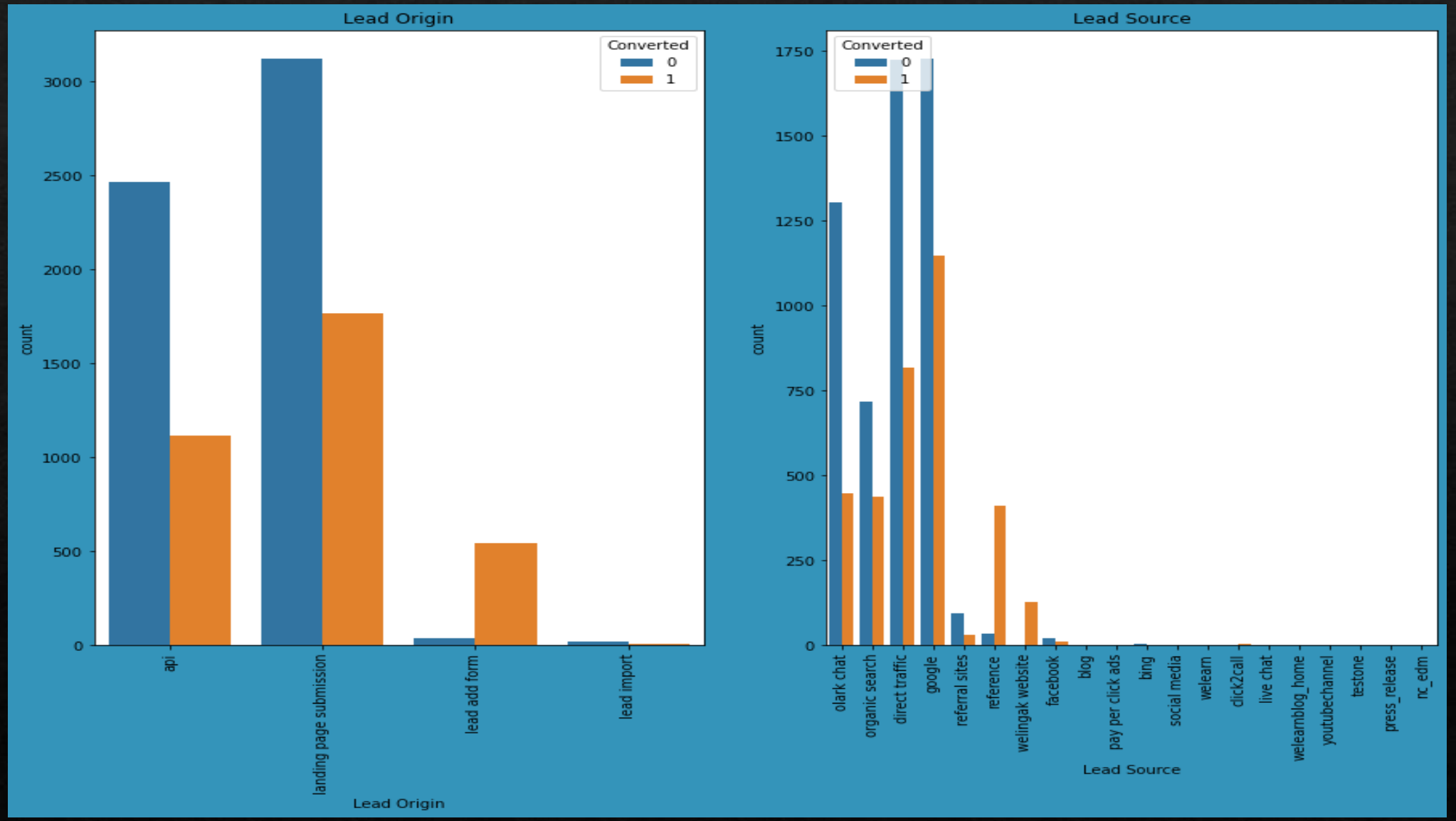
Exploratory Data Analysis(EDA)



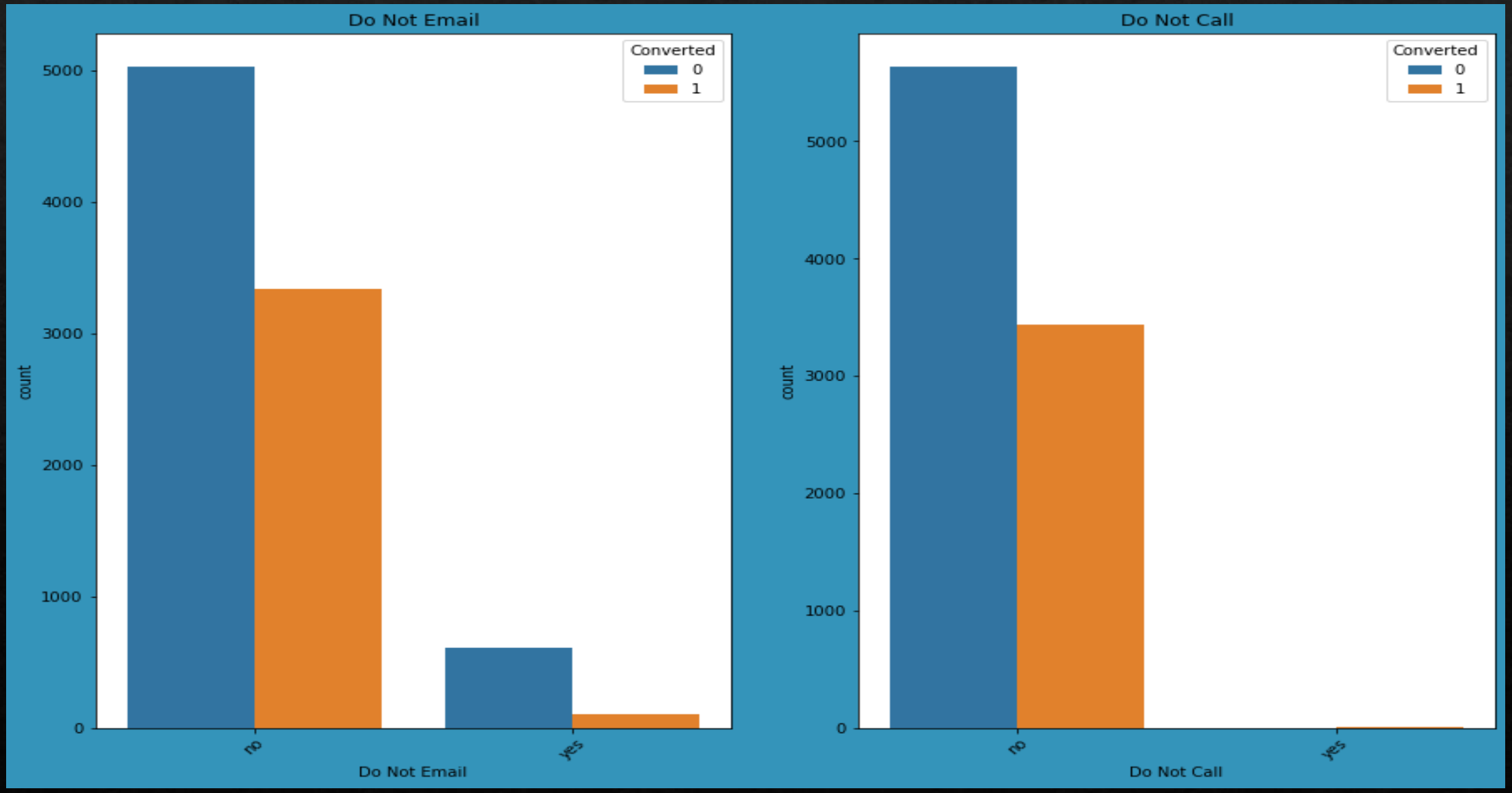
Exploratory Data Analysis(EDA)



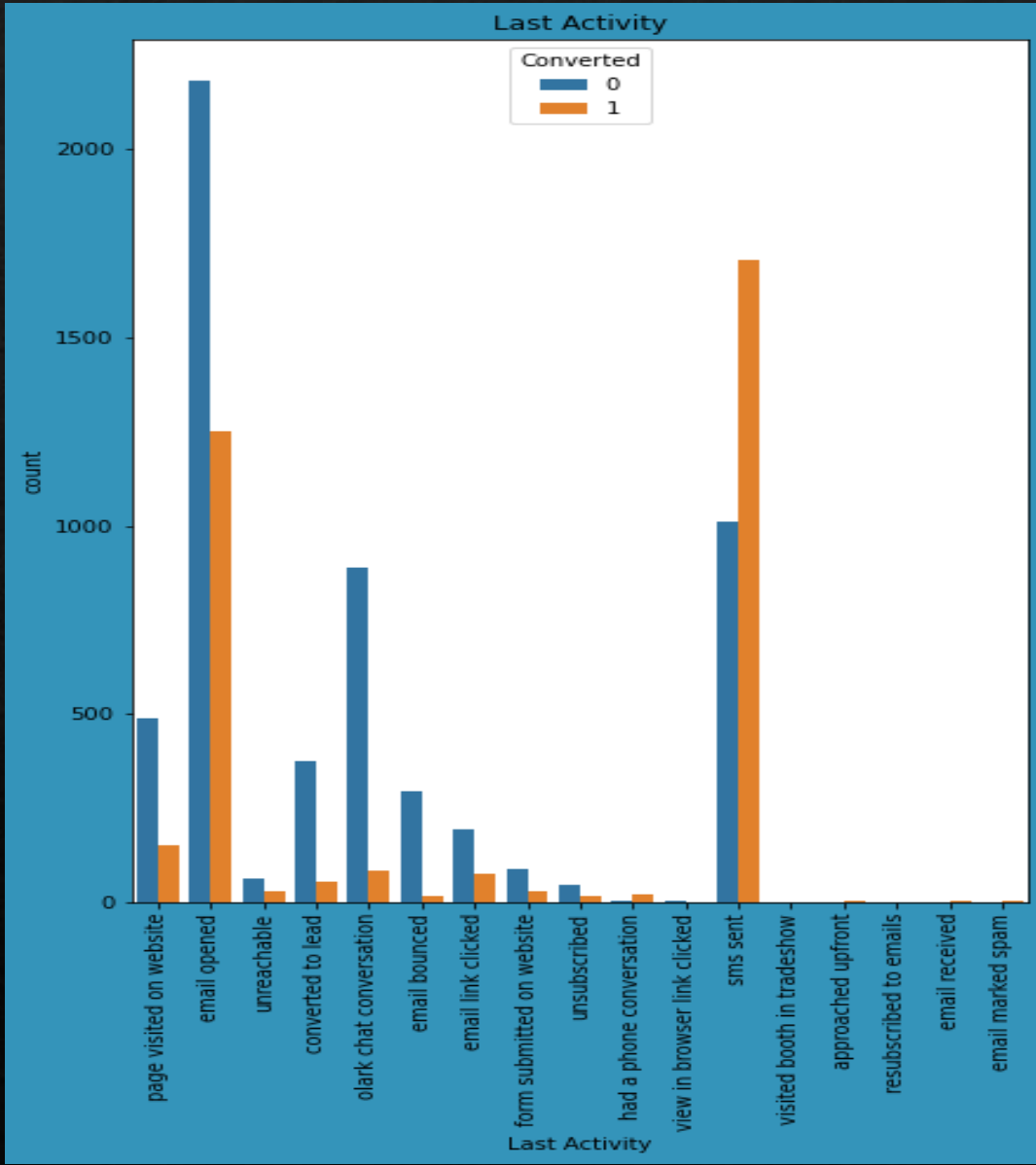
Exploratory Data Analysis(EDA)

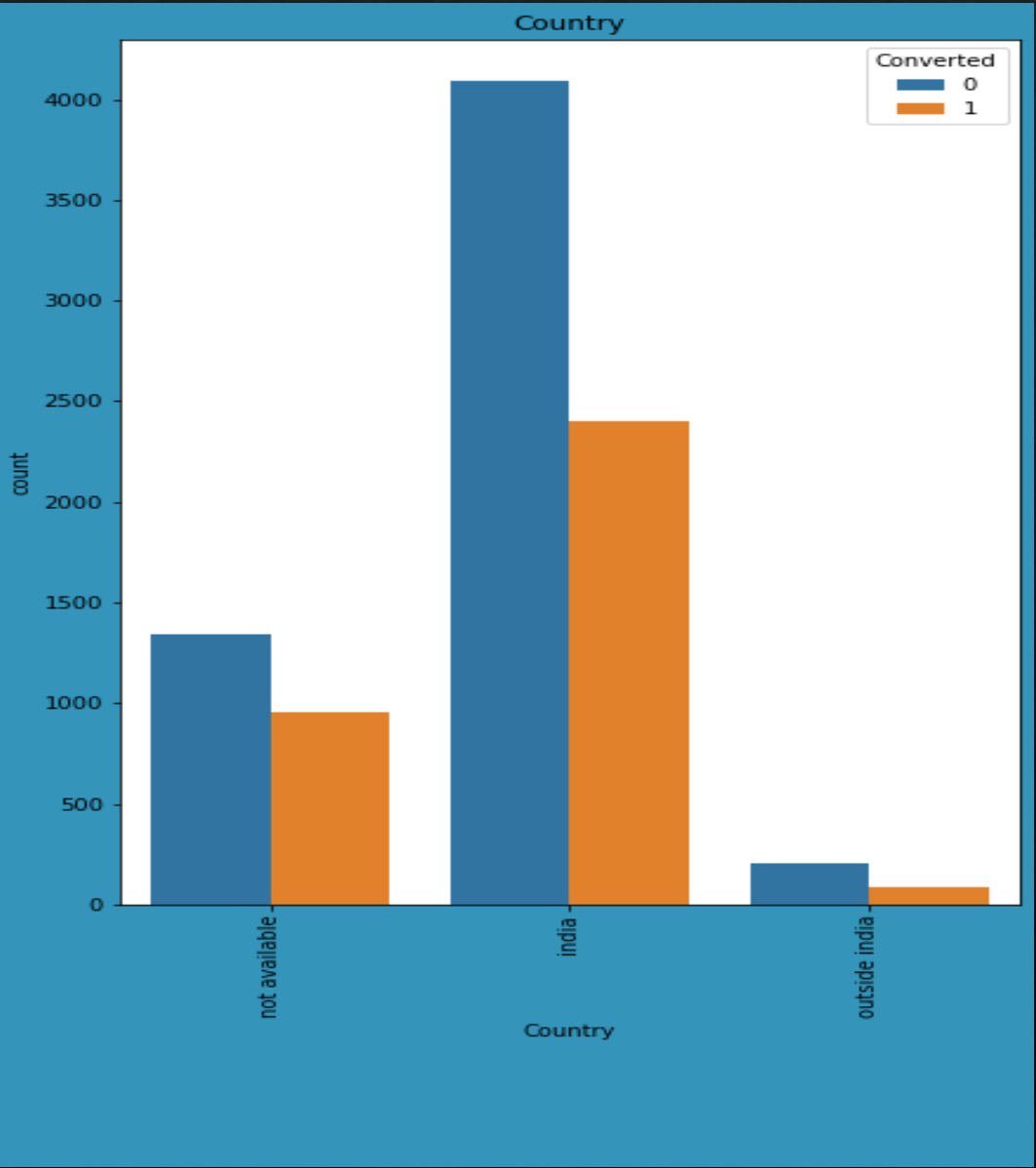


Exploratory Data Analysis(EDA)

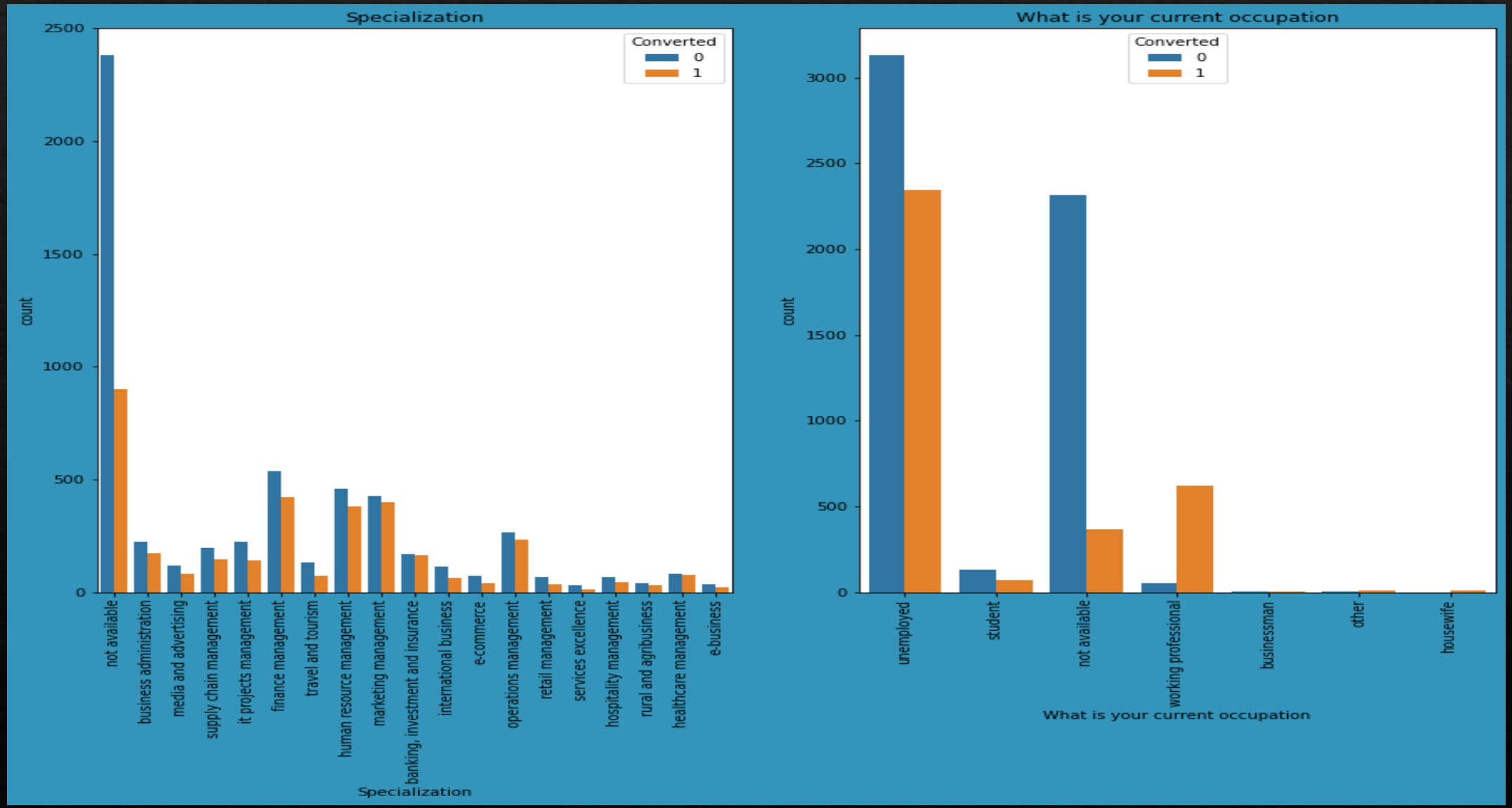


Exploratory Data Analysis(EDA)

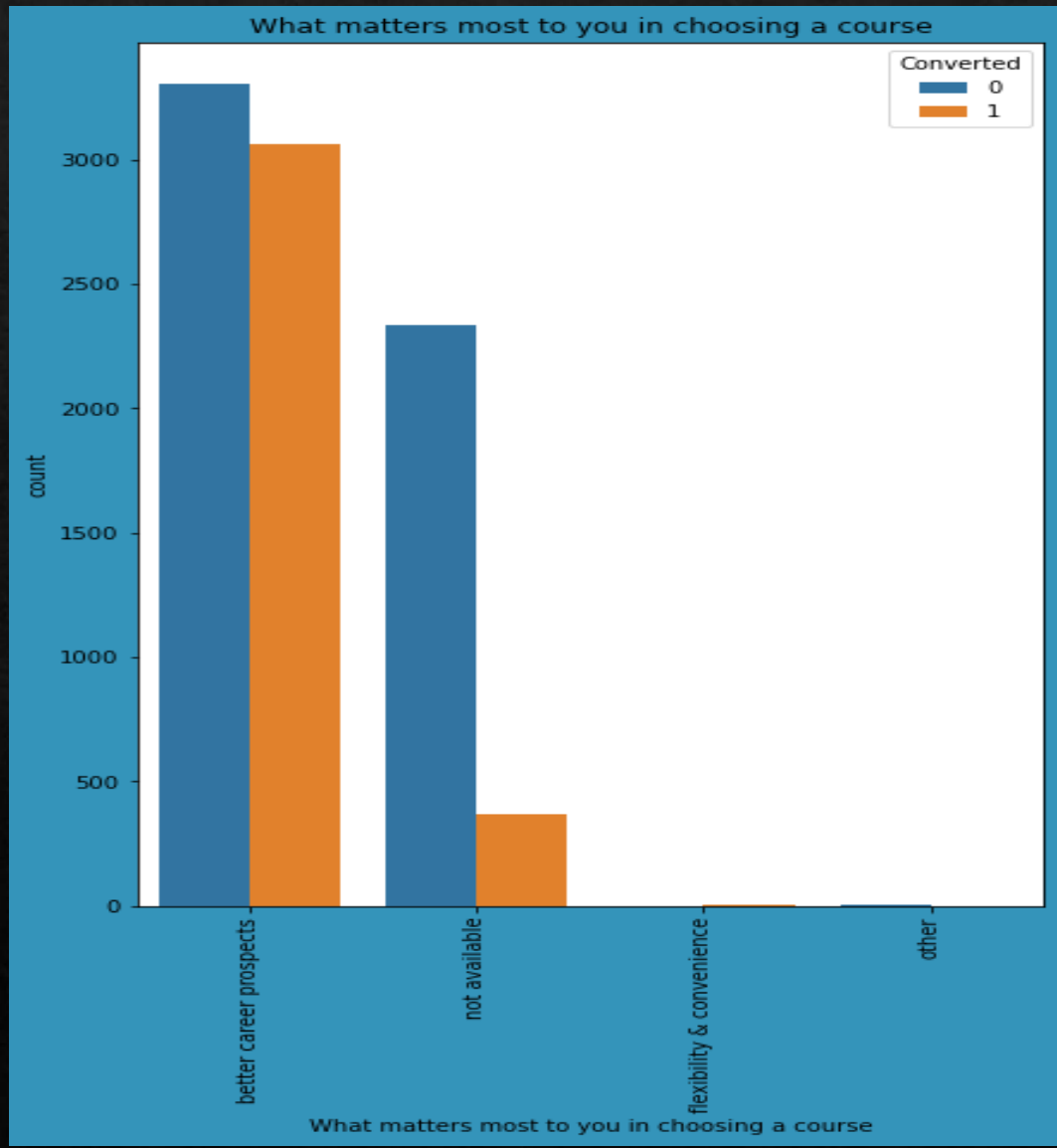


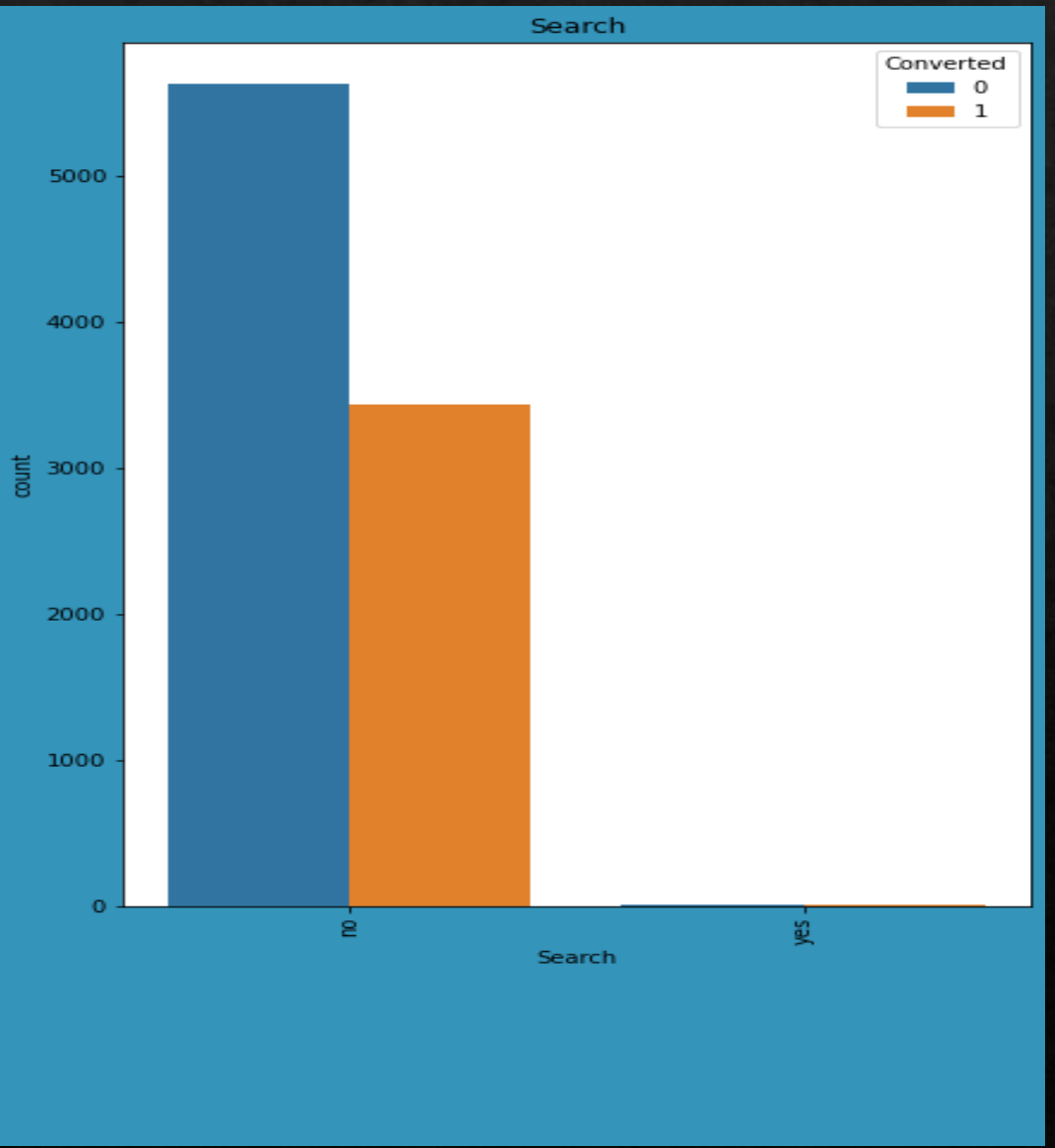


Exploratory Data Analysis(EDA)

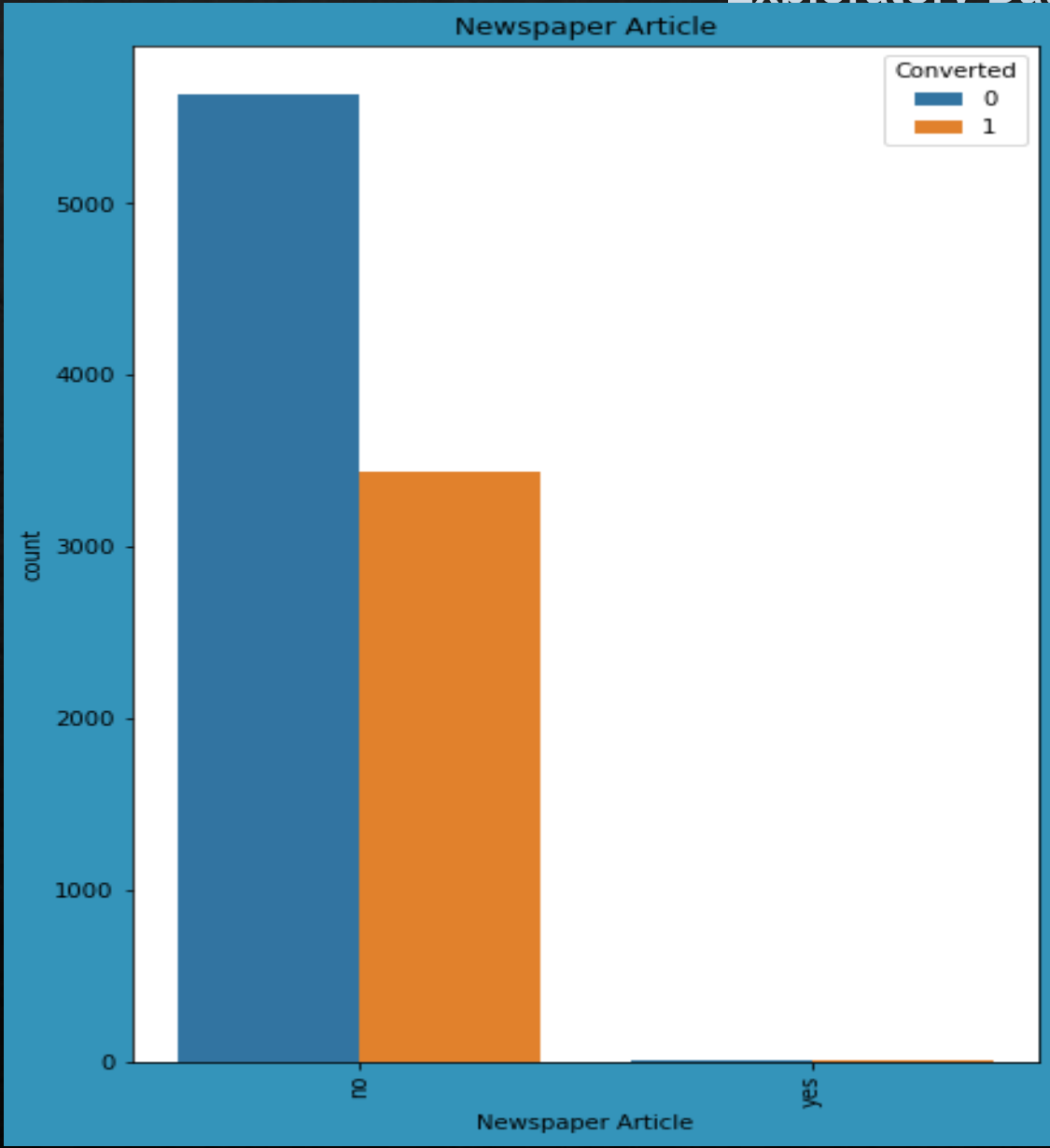


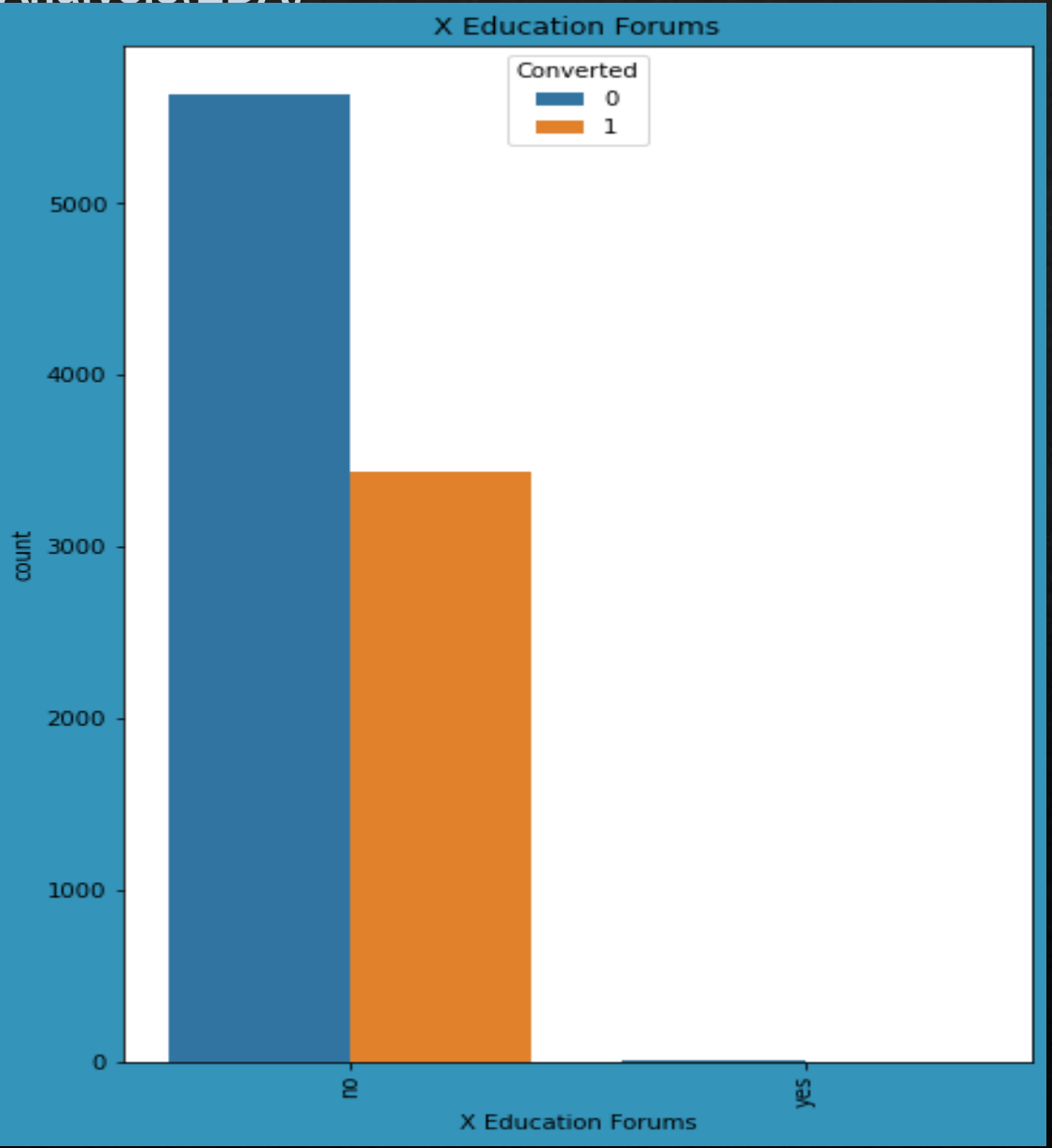
Exploratory Data Analysis(EDA)



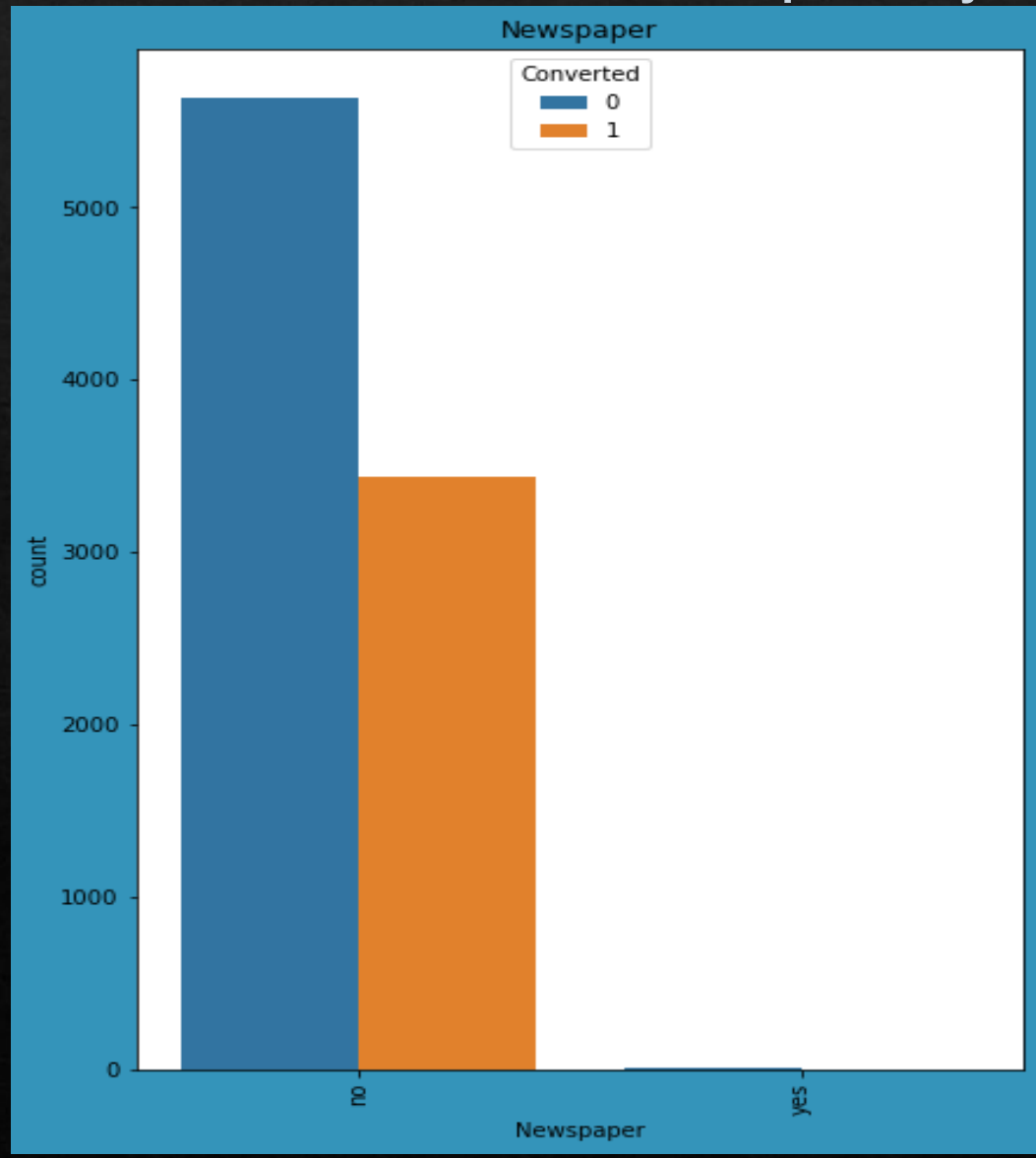


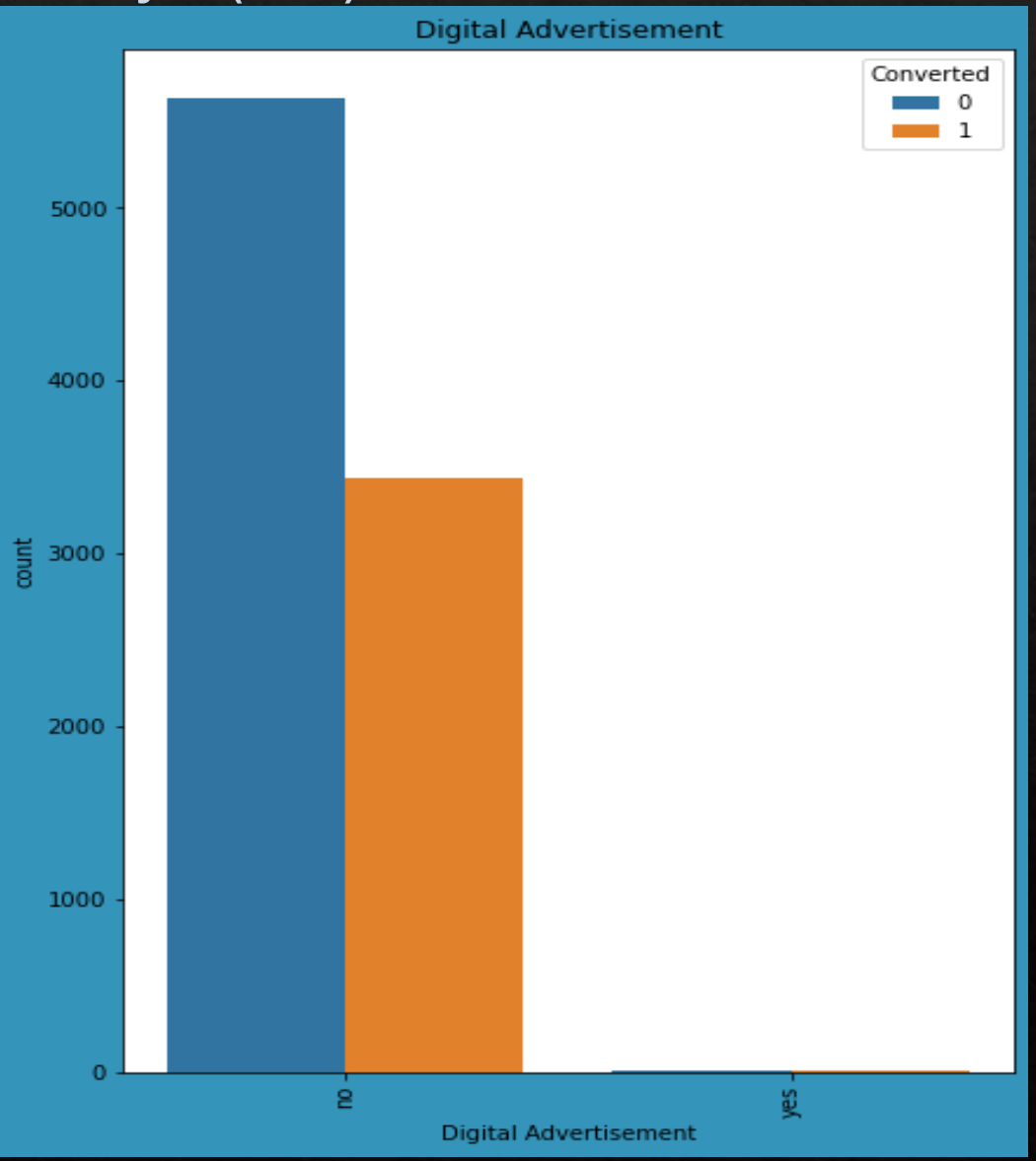
Exploratory Data Analysis(EDA)



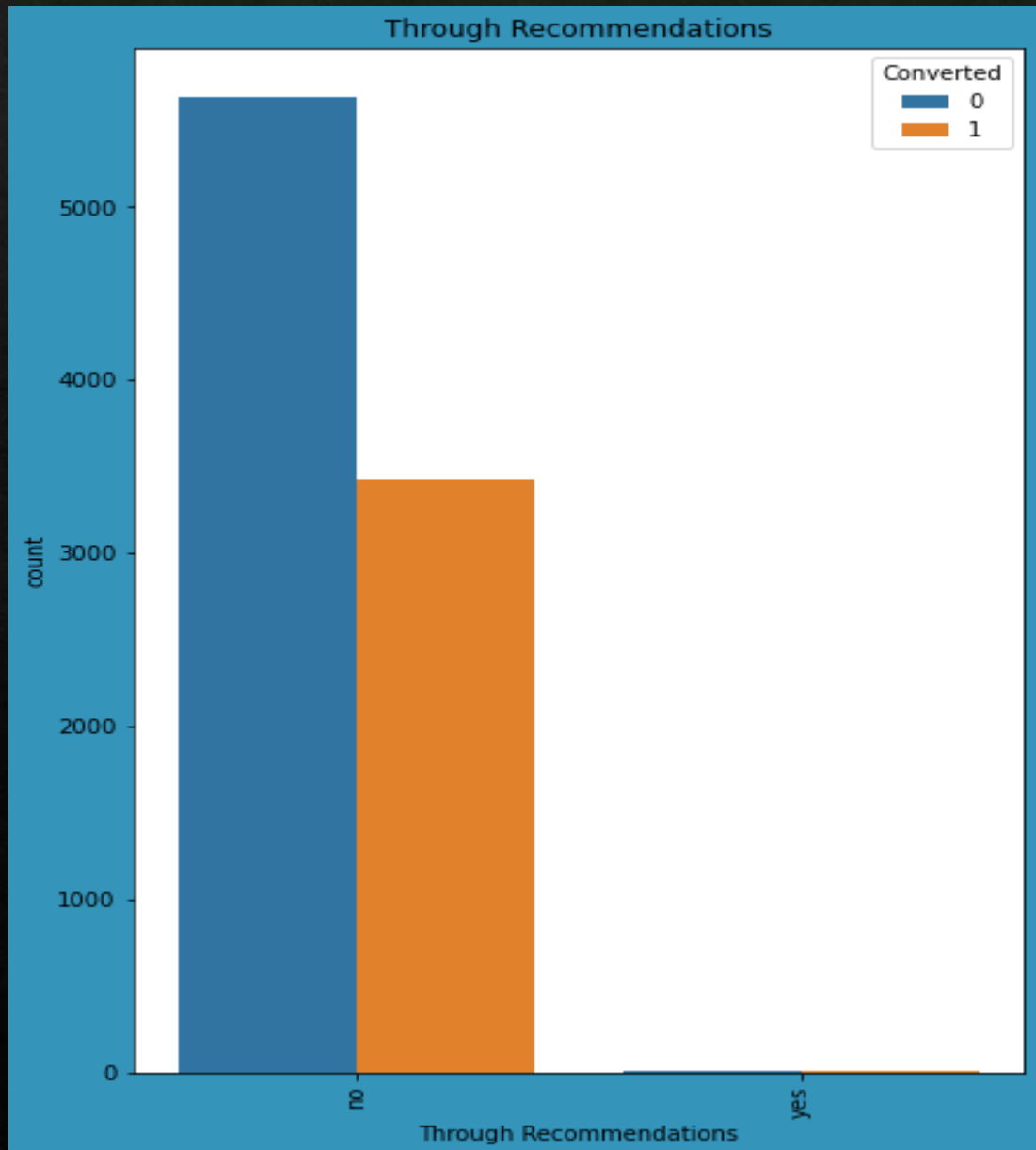


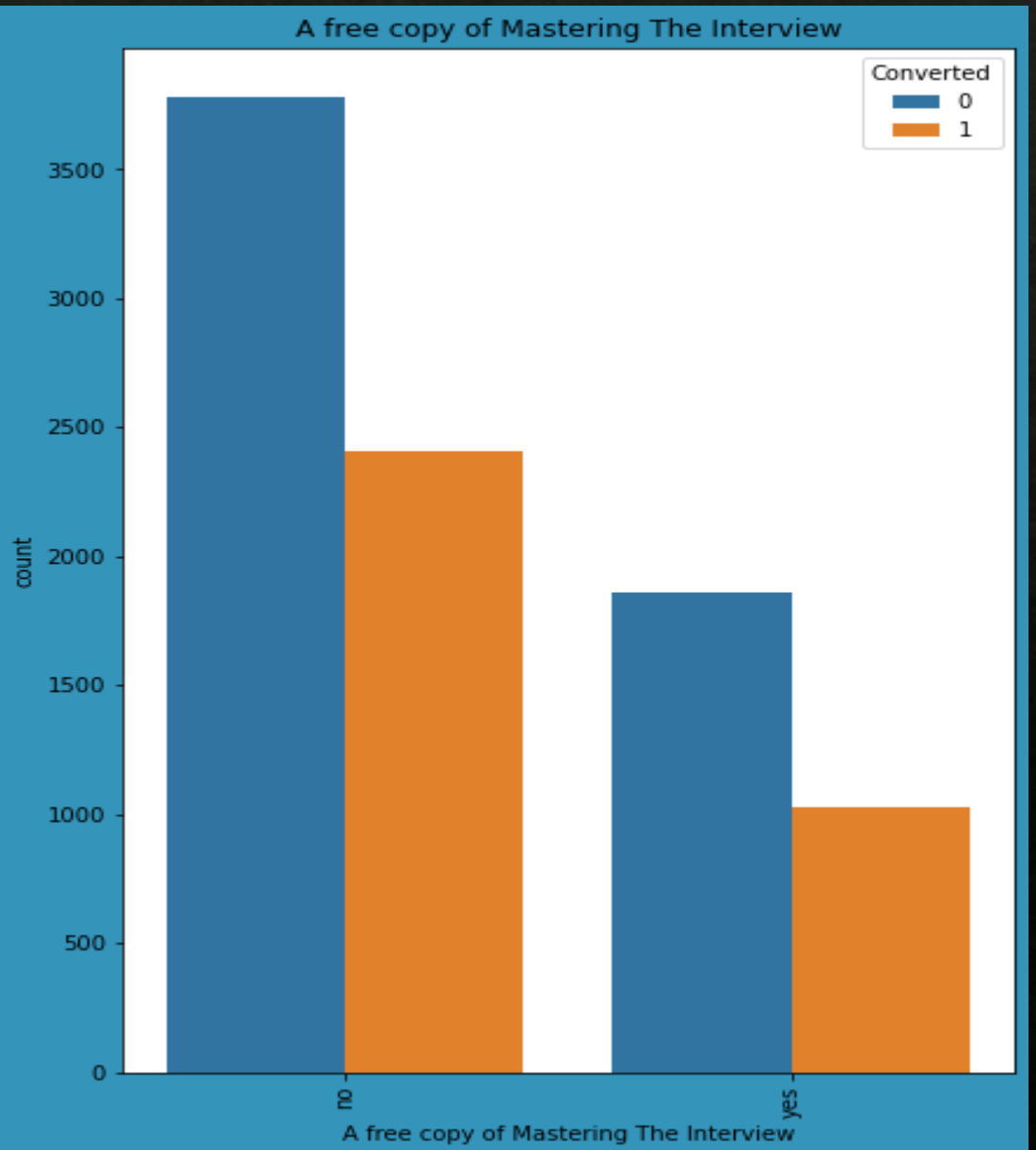
Exploratory Data Analysis(EDA)



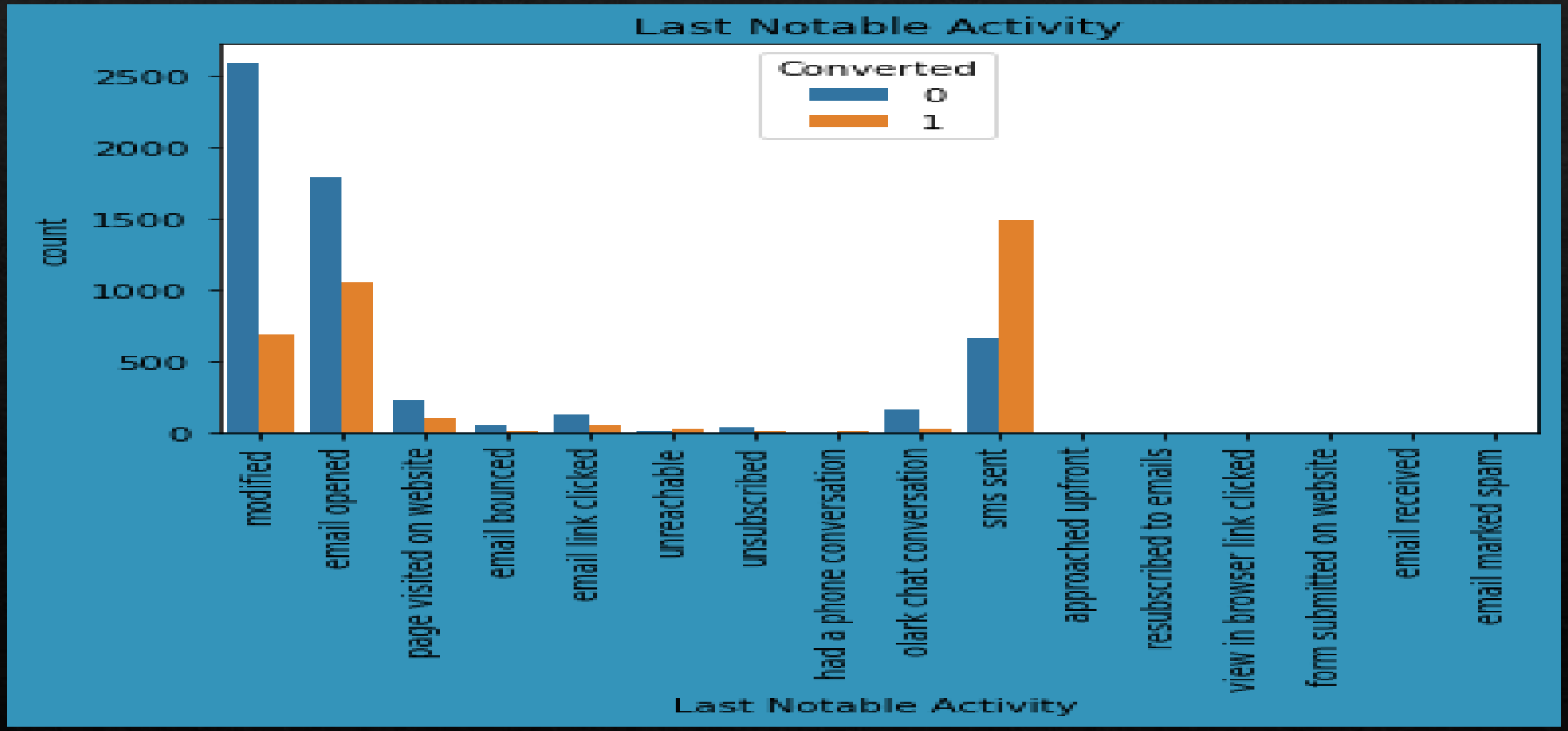


Exploratory Data Analysis(EDA)

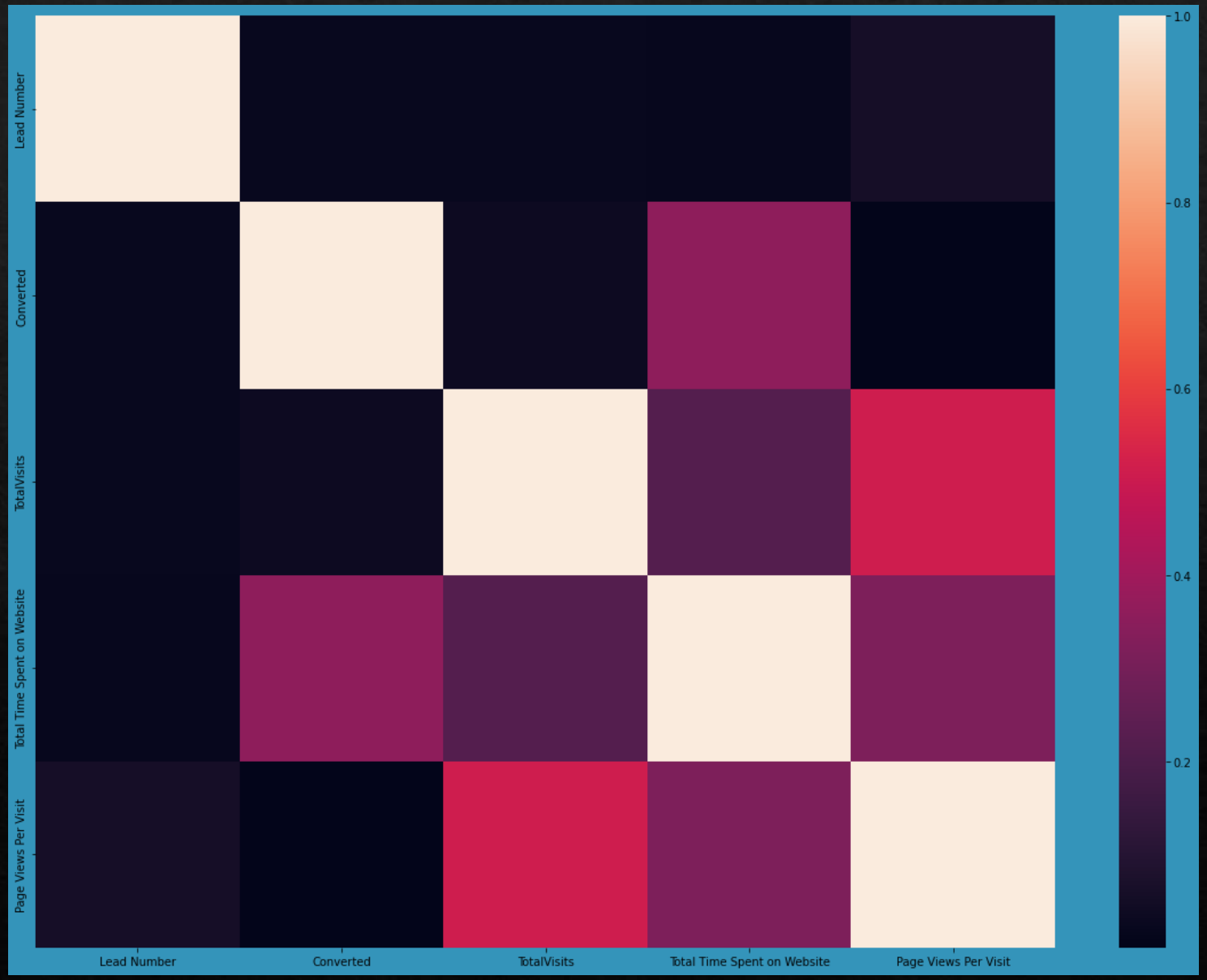




Exploratory Data Analysis(EDA)



Exploratory Data Analysis(EDA)



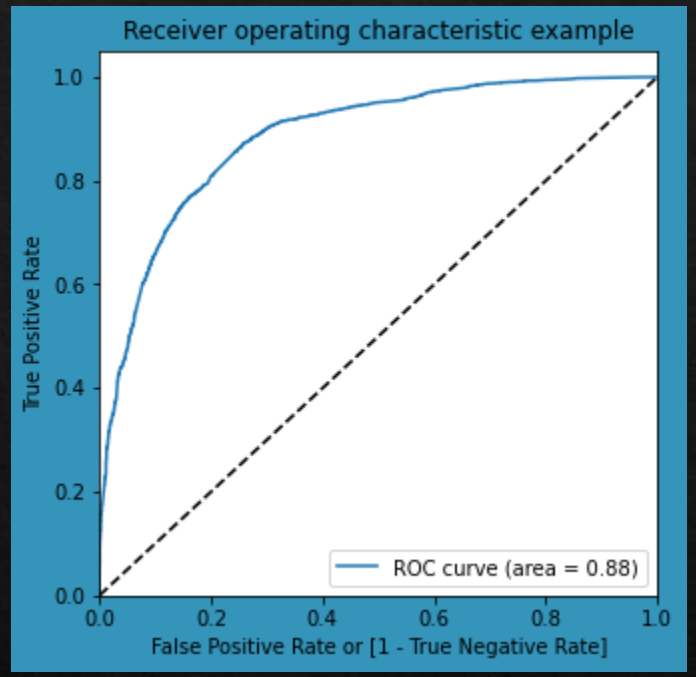
Data Conversion

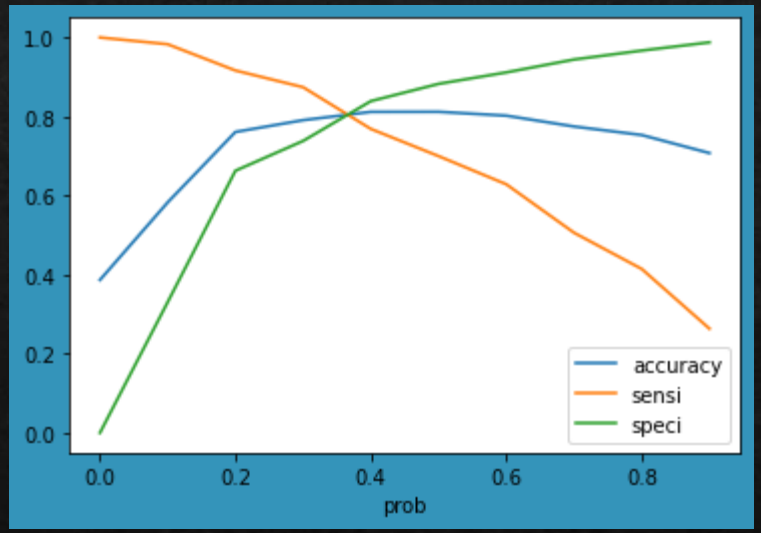
- Numerical variables were scaled and normalized
- Dummy Variables were created for Object type variables
- Total Rows for Analysis: 8792
- Total columns for analysis: 43

Model Building

- The data was split 70, 30 into train and test sets
- The train and test split was performed using linear regression in a 70-30 ratio
- Used RFE for feature selection
- We ranked the features and chose most relevant 15 features through RFE
- We built the models by removing features whose p-value was greater than 0.05 and VIF value was greater than 5
- We did predictions on the Test data set
- Overall Accuracy was 81%

ROC Curve





- We found the Optimal cut off point
- Optimal cut off probability gives the point at which sensitivity and specificity intersect and are balanced
- From the second graph it was found that the optimal cut off point was 0.38

Conclusion

- The variables which mattered most to potential buyers were:
  - Total time spent on the website
  - Total number of visits
- When the lead source was:
  - Google
  - Direct Traffic
  - Organic Search
  - Welingak wensite
- When the last activity was:
  - SMS
  - Olark chat conversion
- When the lead origin is lead add format
- When they were working professional

X education should focus on these points in order to increase their lead conversion rate and spend less time on calling prospective buyers