Lead Scoring Model for X Education

"Optimizing Lead Conversion Rates using Logistic Regression"

Sauradip Pradhan

Date - 04-05-2024

- Agenda:

- - Introduction to X Education
- - Explanation of the lead conversion process at X Education
- - Data Overview
- - EDA Analysis
- - Model Development
- - Model Evaluation
- - Recommendations
- - Conclusion

- Brief introduction to X Education

- - Mission and Vision of the company
  - X Education sells online courses to professionals. Many visit their website daily but only a few become leads by filling out forms. The conversion rate is typically 30%. They want to boost this by identifying 'Hot Leads' to focus their efforts on. We need to build a model to assign a lead score to each lead. The goal is to prioritize leads with higher scores, as they are more likely to convert. The CEO aims for an 80% lead conversion rate.
- - Importance of lead conversion for business growth
  - X Education has a high volume of leads but struggles with a low conversion rate; for instance, out of 100 leads, only around 30 get converted. To improve efficiency, the company aims to pinpoint potential leads, or 'Hot Leads'. By doing so, they anticipate a higher conversion rate as the sales team can prioritize communication with these promising leads, rather than spreading efforts across all leads.

- Explanation of the lead conversion process at X Education

- - Challenges faced by the company in lead conversion
  - We've received a leads dataset with about 9000 data points, containing various attributes like Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. The target variable is 'Converted', where 1 indicates conversion and 0 indicates non-conversion. We should investigate and handle the 'Select' level present in categorical variables, as it's akin to a null value.
- - Objective of the case study
  - 1. Develop a logistic regression model to assign lead scores ranging from 0 to 100, helping the company target potential leads. Higher scores indicate hotter leads with a higher likelihood of conversion, while lower scores signify colder leads less likely to convert.
  - 2. Address additional problems presented by the company, ensuring the model can adapt to future changes in requirements. These issues will be documented based on the logistic regression model's findings and included in the final presentation for recommendations.

- Overview of the dataset provided

- The name of the csv file is Lead.csv
- The dataset has 9240 rows and 37 columns
- Three columns are integer datatype, four columns are float data type and the rests are object datatype

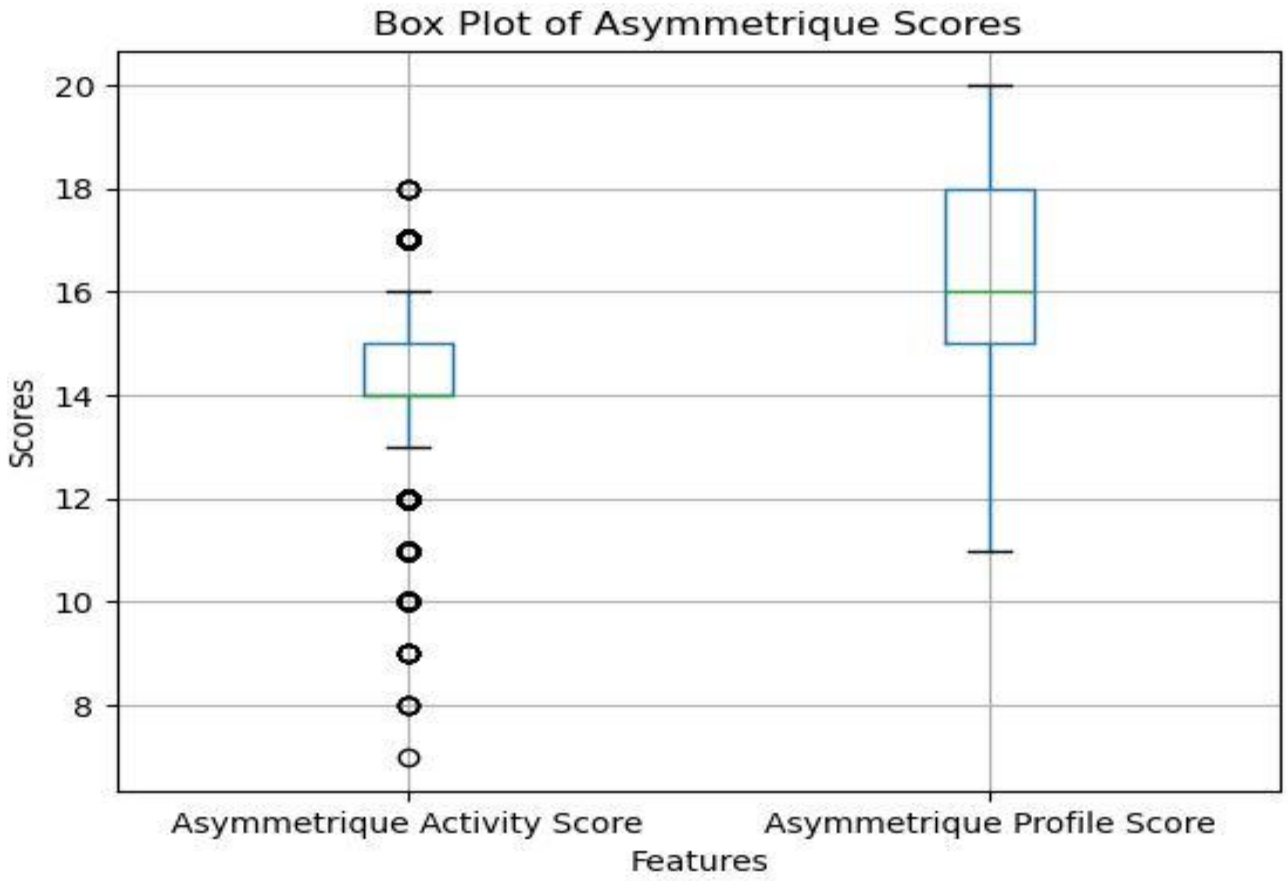
- EDA Analysis

- Data Cleaning
  - In the data set, few columns are contain the 'Select' Feature. First we convert the 'Select' Feature as Null Value.
  - There are lots of missing values in the dataset. We drop the columns which have more than 70% missing values.
  - For the remaining categorical columns, we replace null values with the most frequent feature within each respective column.
  - For numerical columns, we replace null values with the mean of the respective column.
  - For columns with very few missing values, we can drop the corresponding rows.(less than 2% missing values)

- EDA Analysis

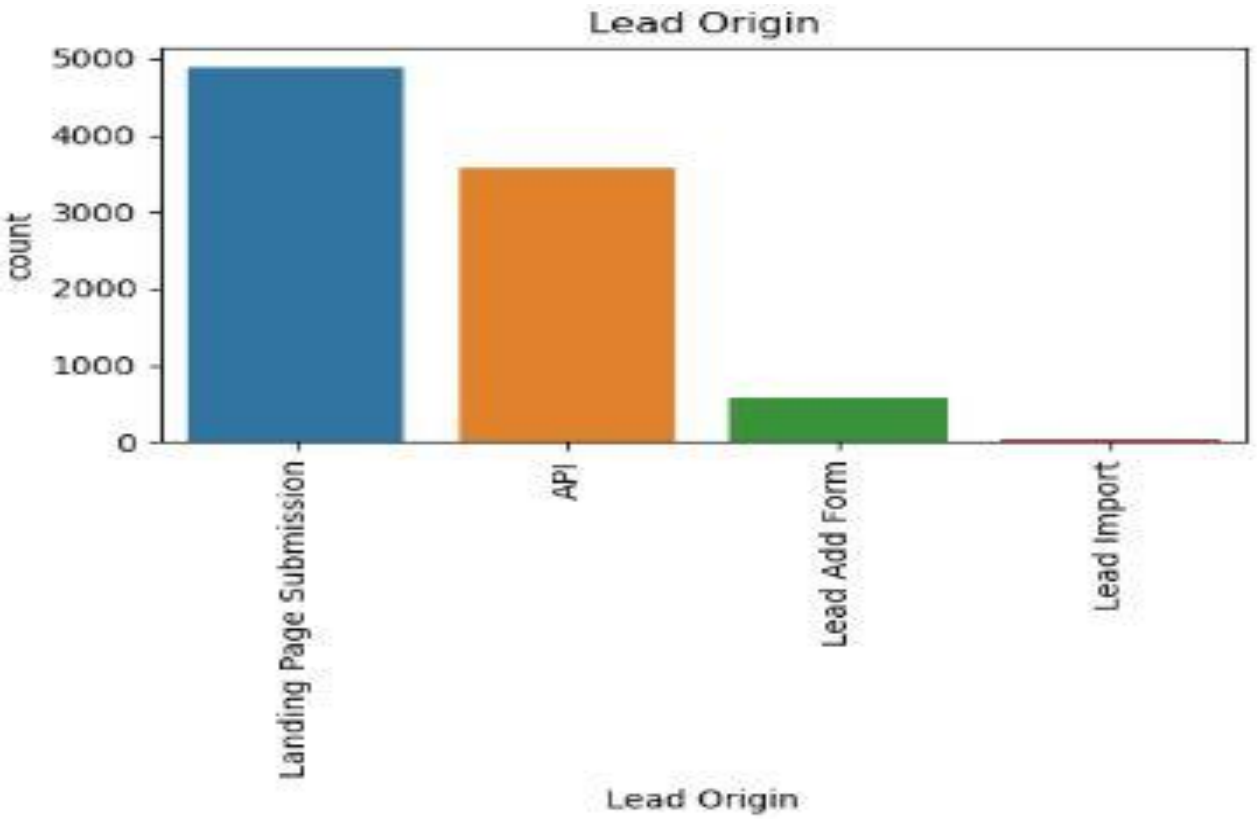
- Outlier Treatment

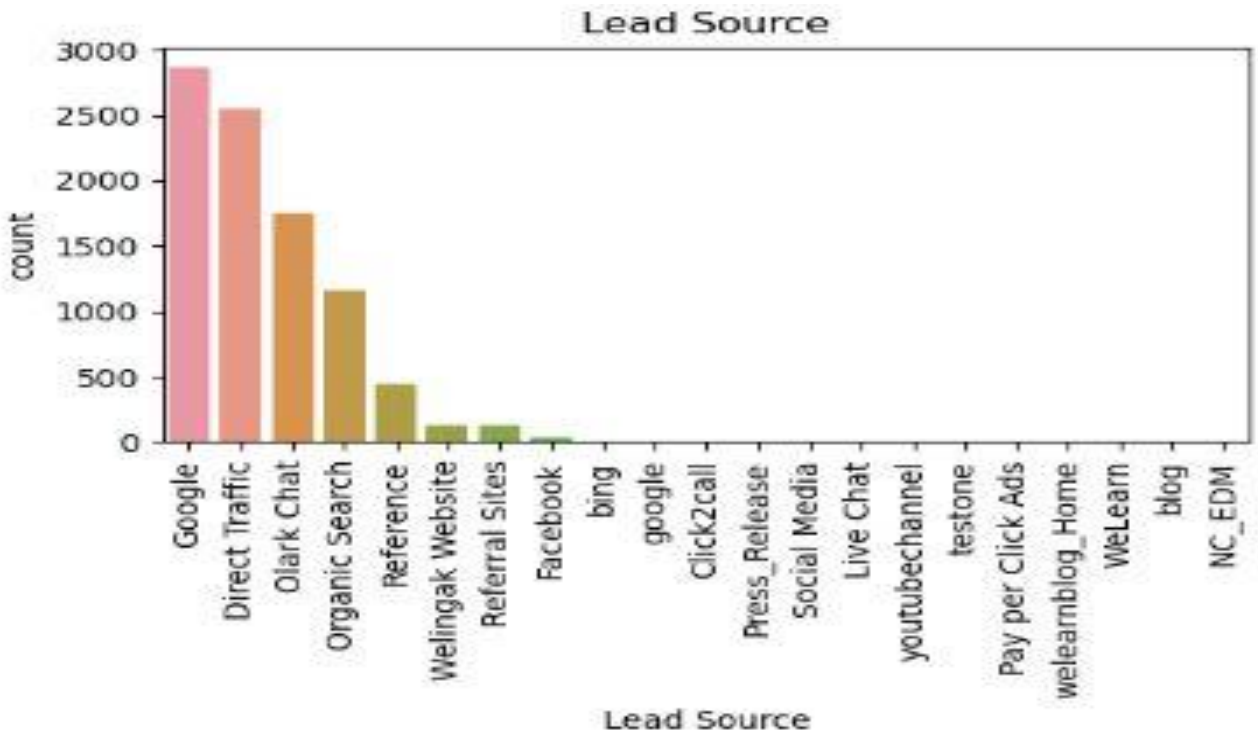
As there is so many outliers we drop the Asymmetrique Activity score



- EDA Analysis

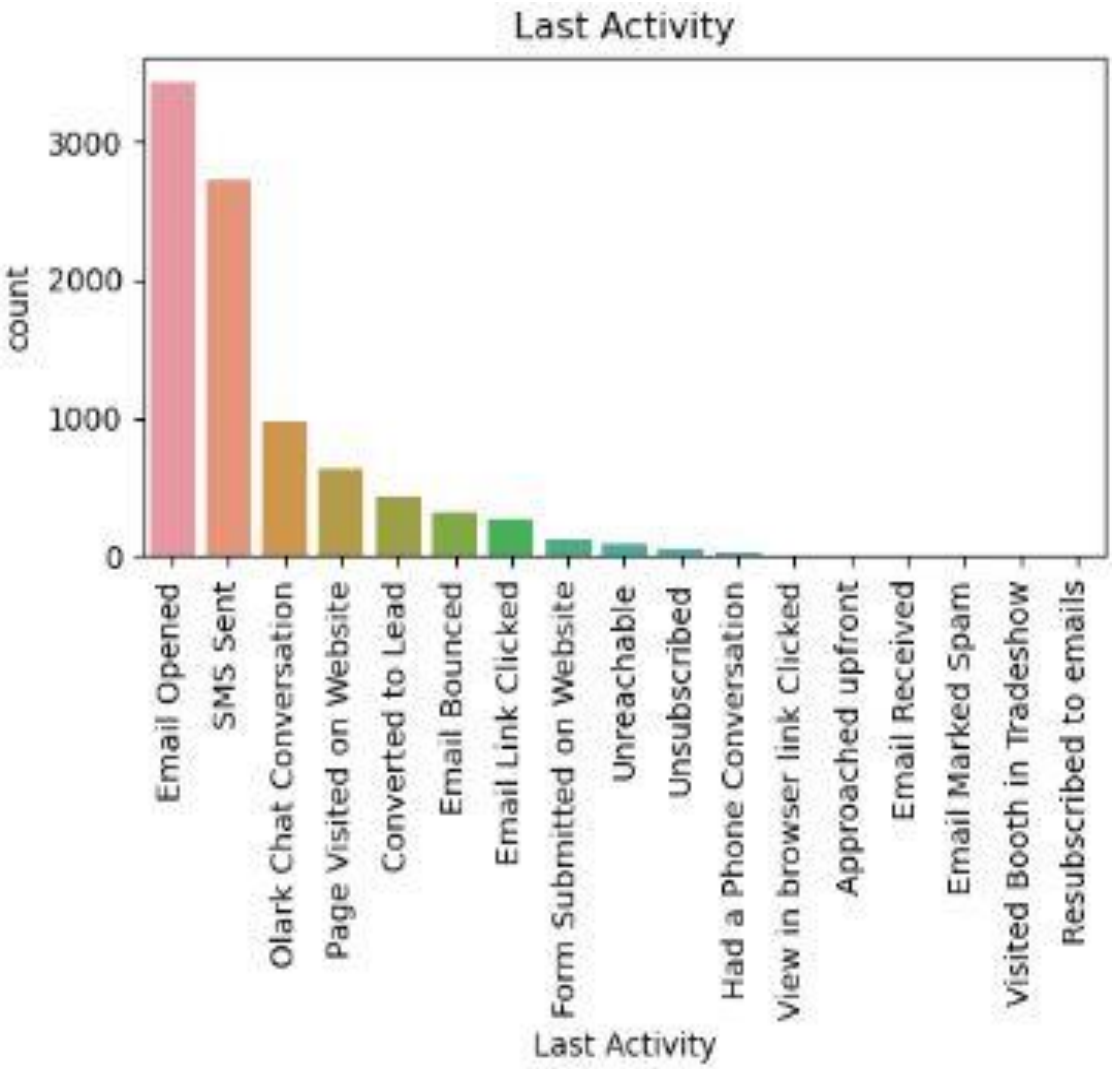
- Univariate Data Analysis (Visualization of categorical columns)

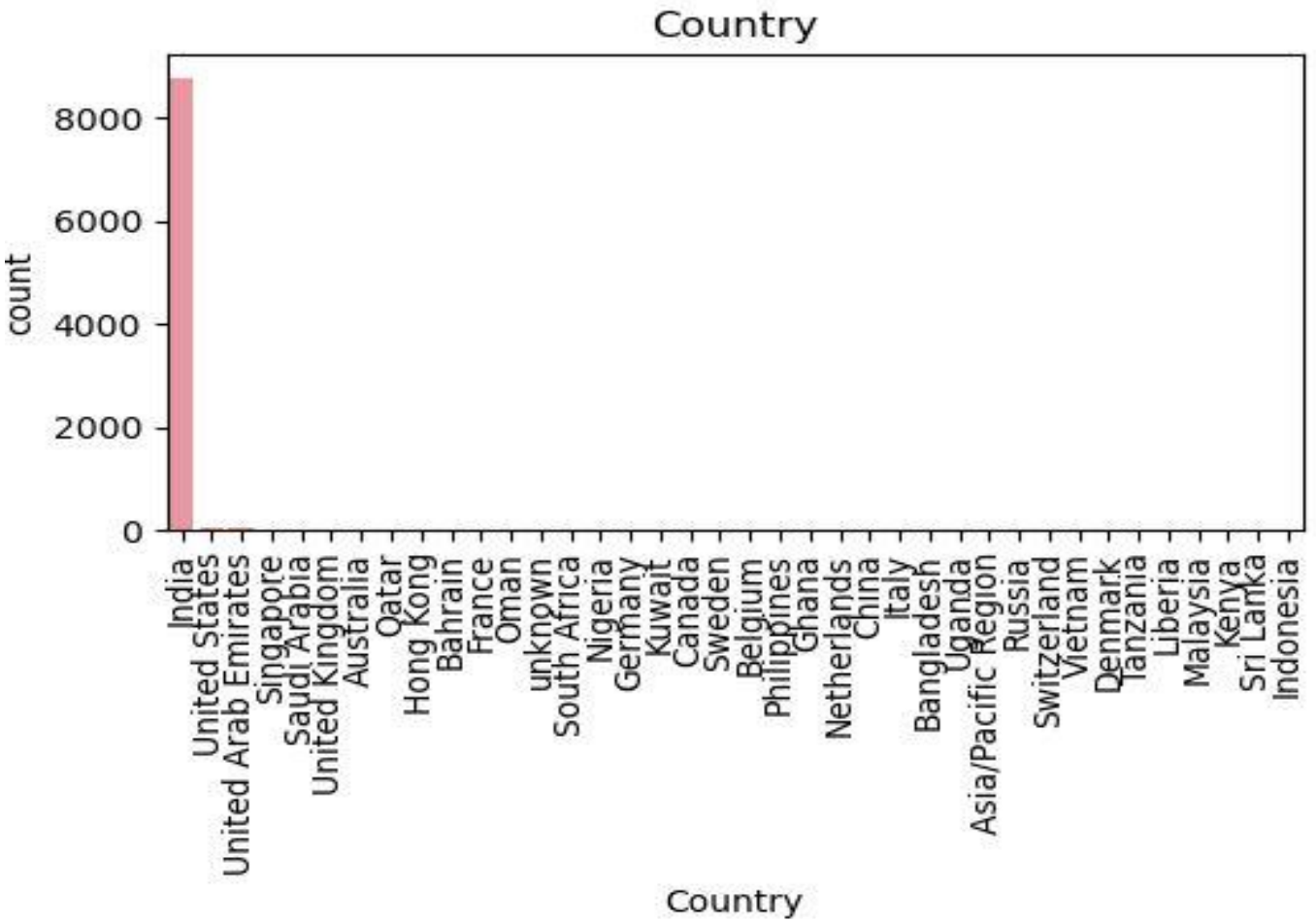




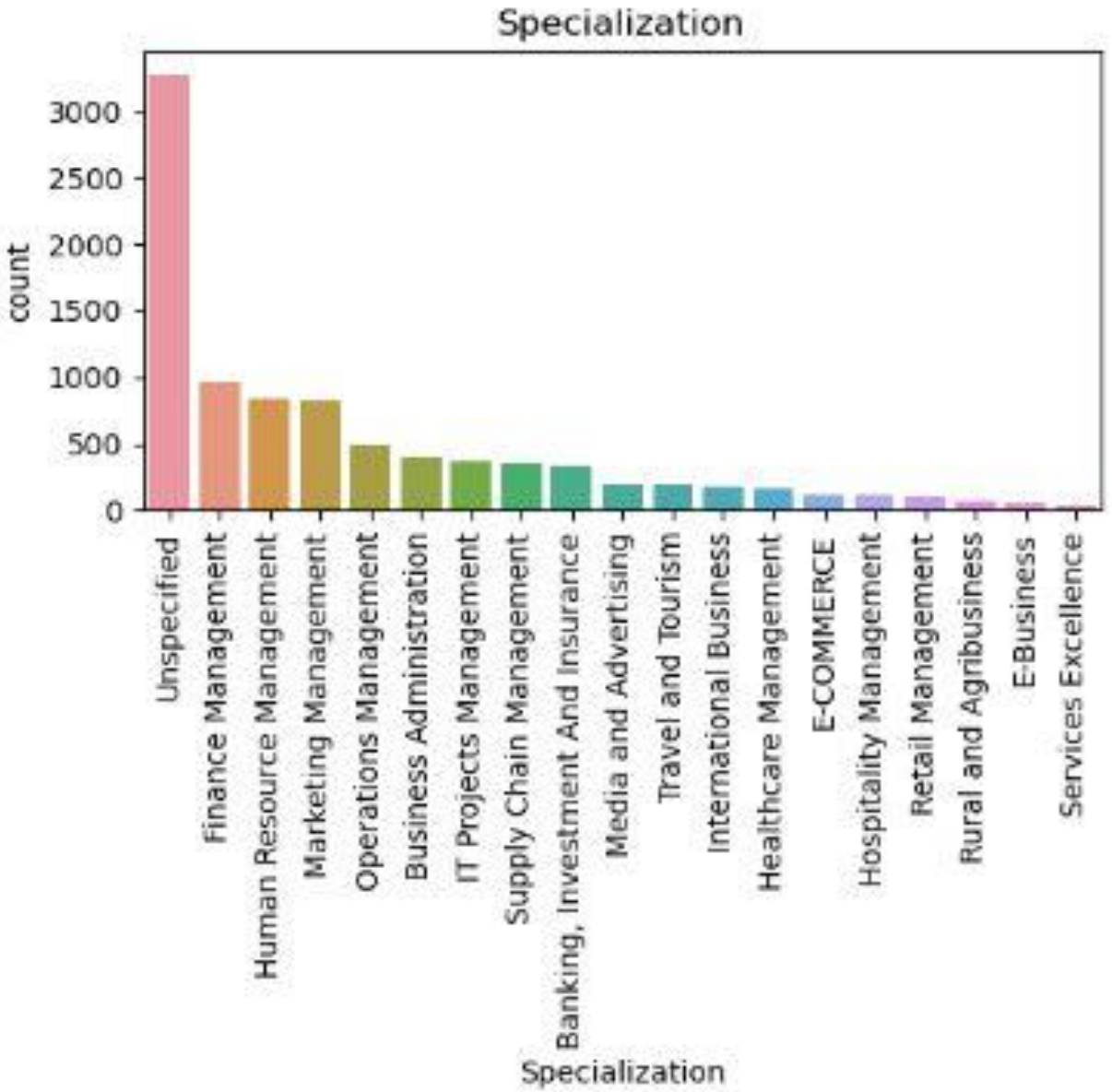
Important factors:

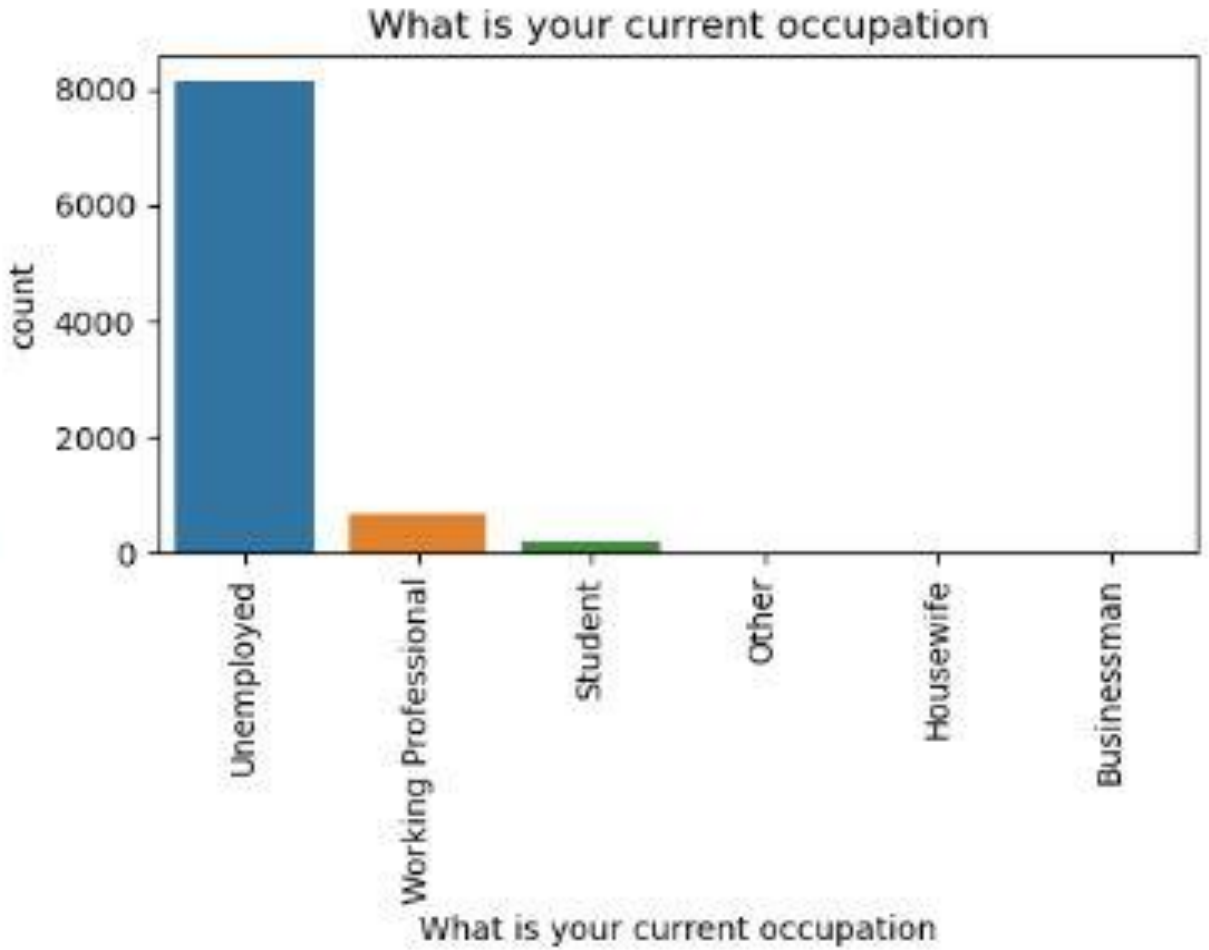
- Landing page submission is higher than API, Lead Add Form and Lead Import in Lead Origin graph
- Most of the people Know about the course from Google



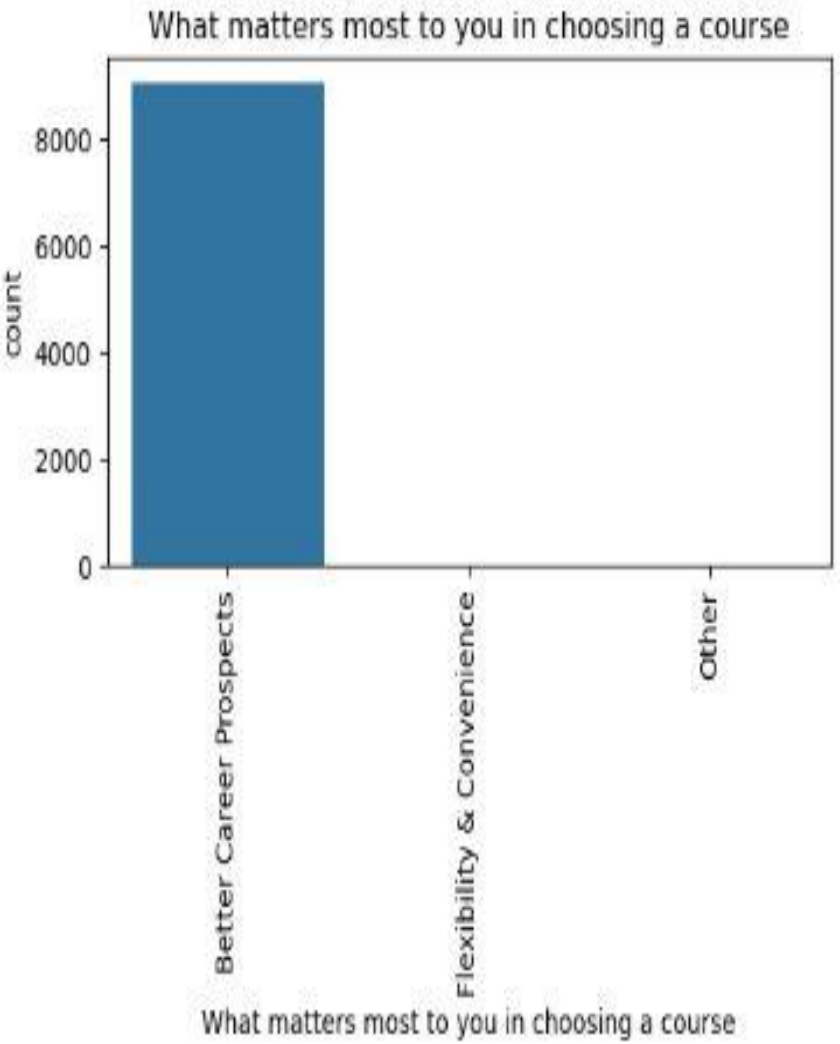


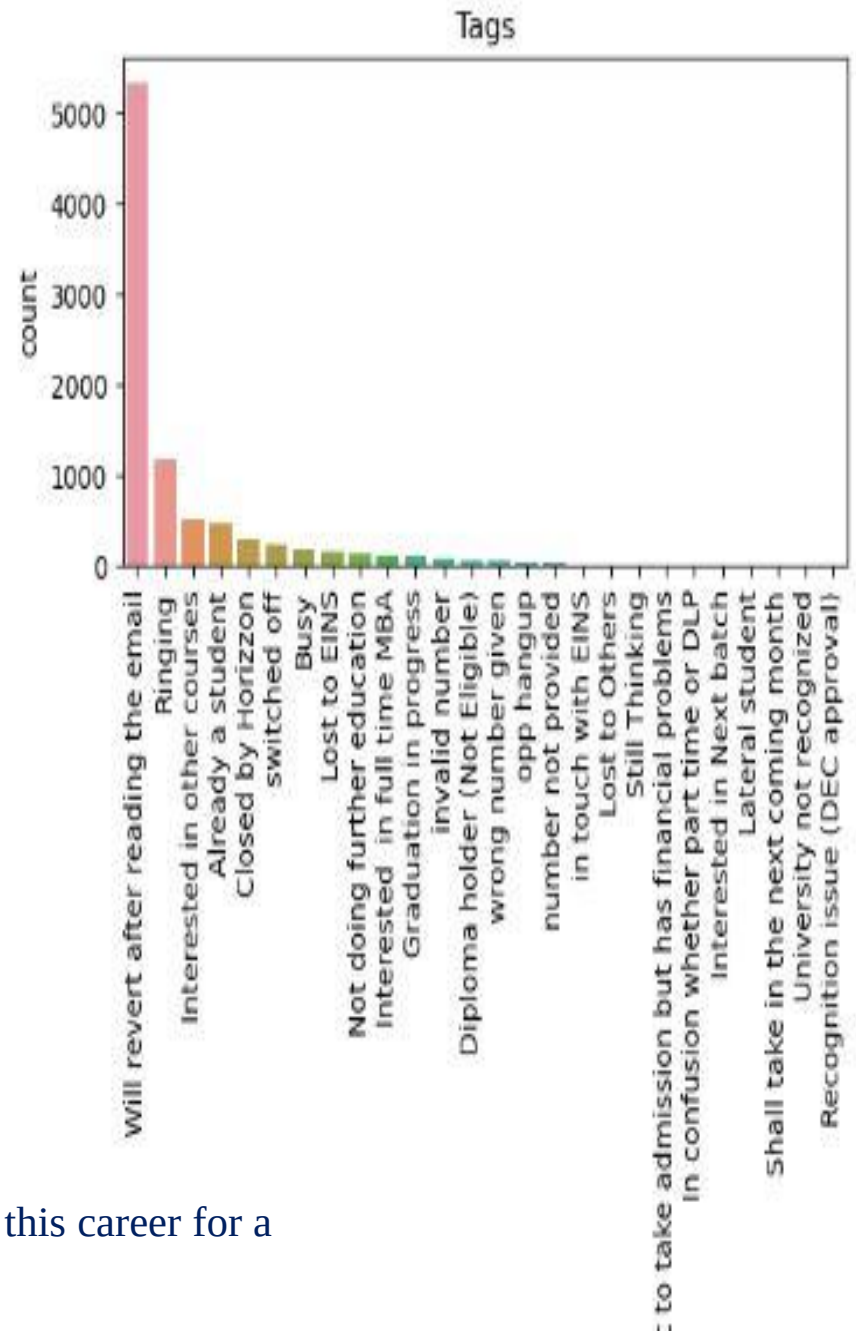
Important factors: Here we observe that most of the peoples are from India

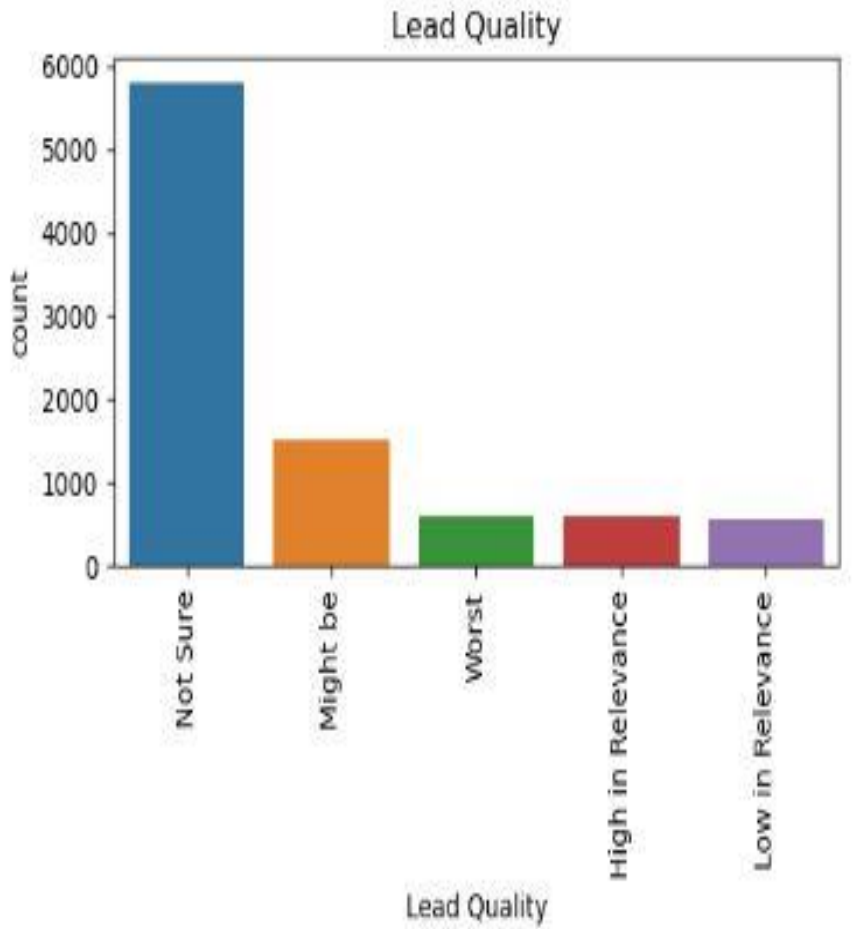




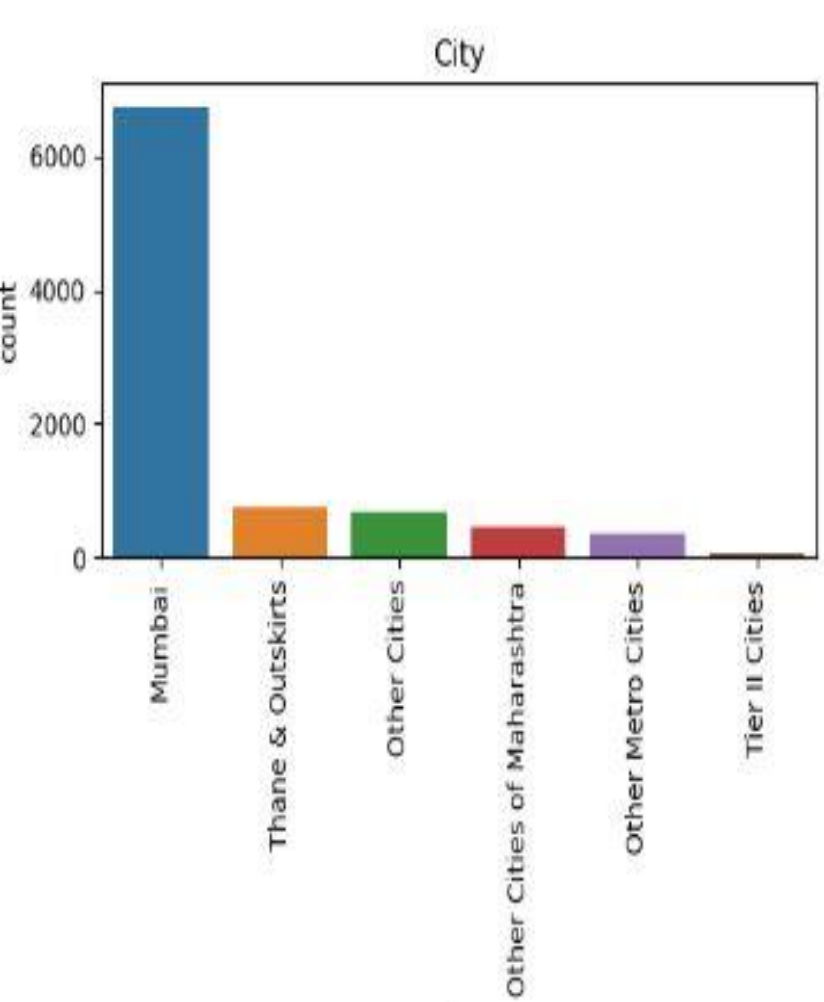
Important factors: Most of the people are Freshers because they most of them are not specified about their skills. And also they are Unemployed.

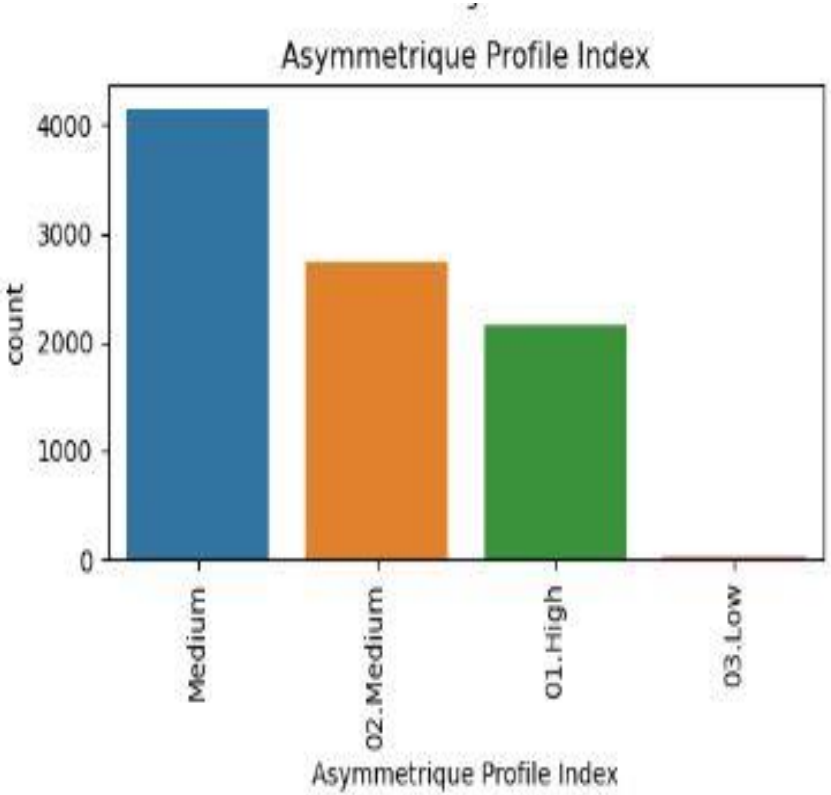


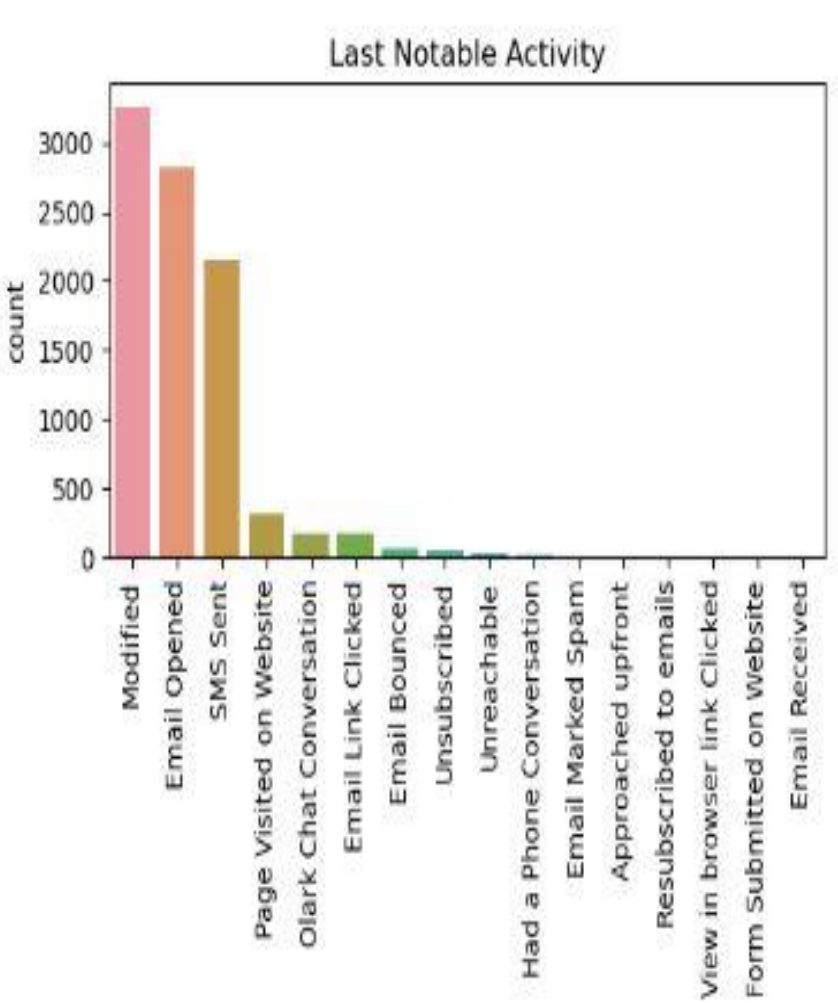




Important factors: Most of the people choosing this career for a better career

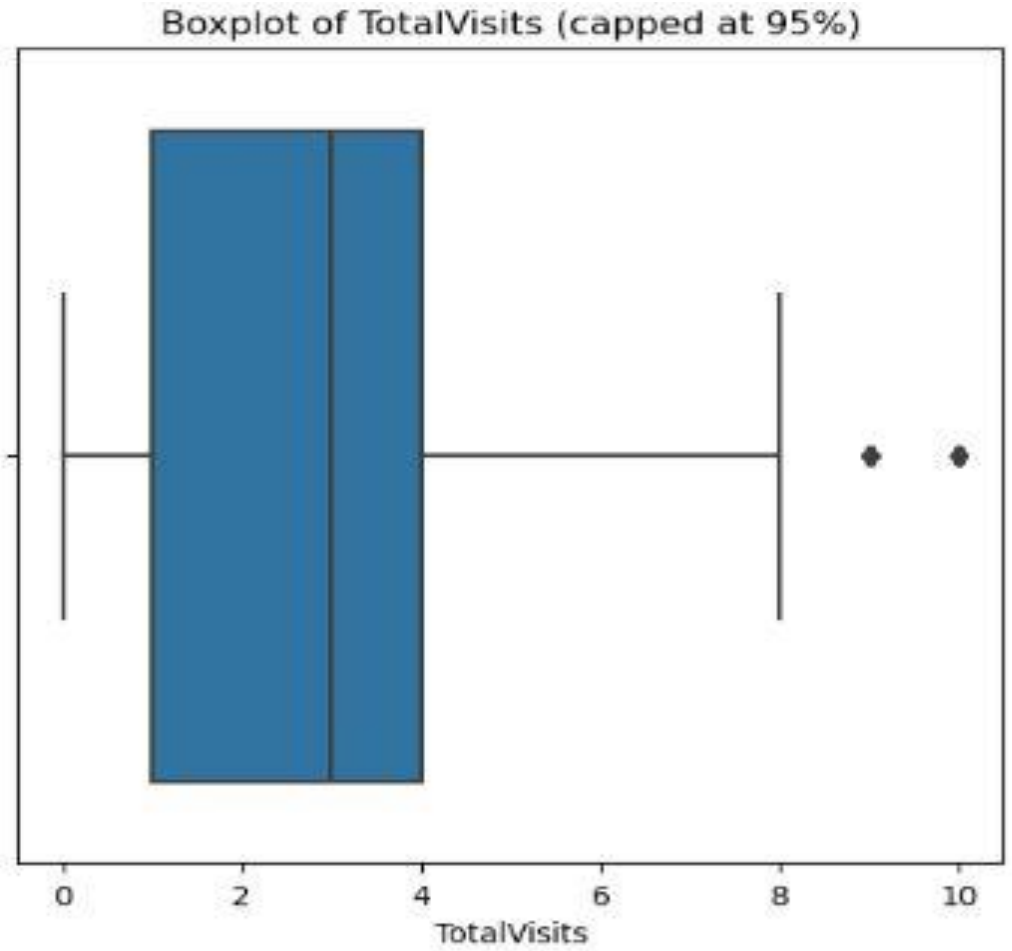


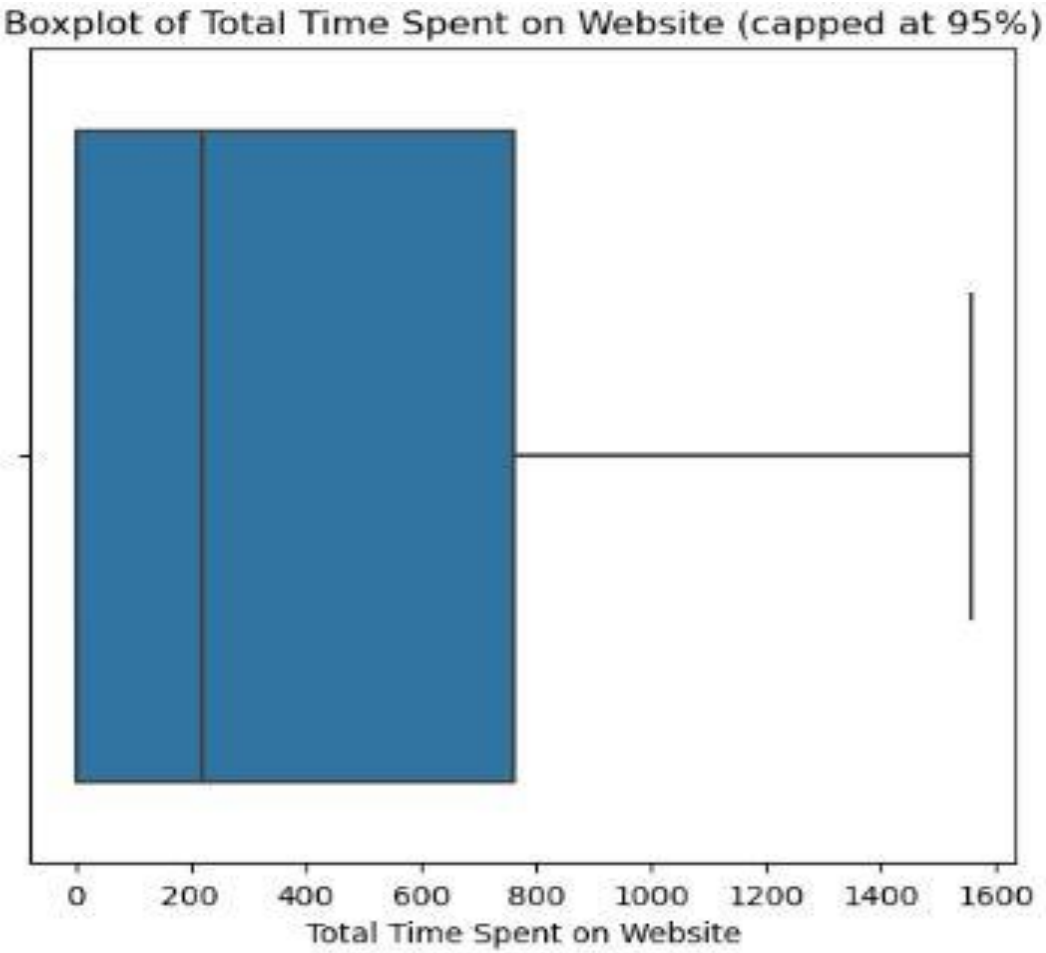


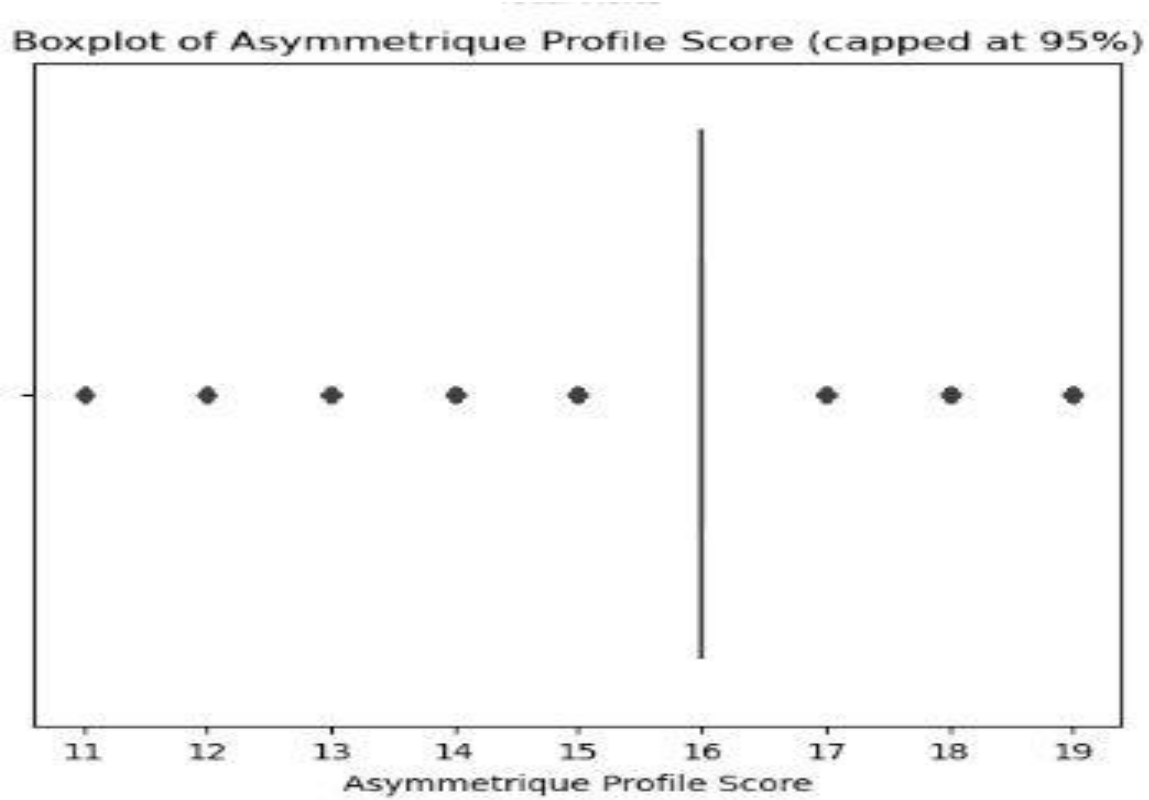


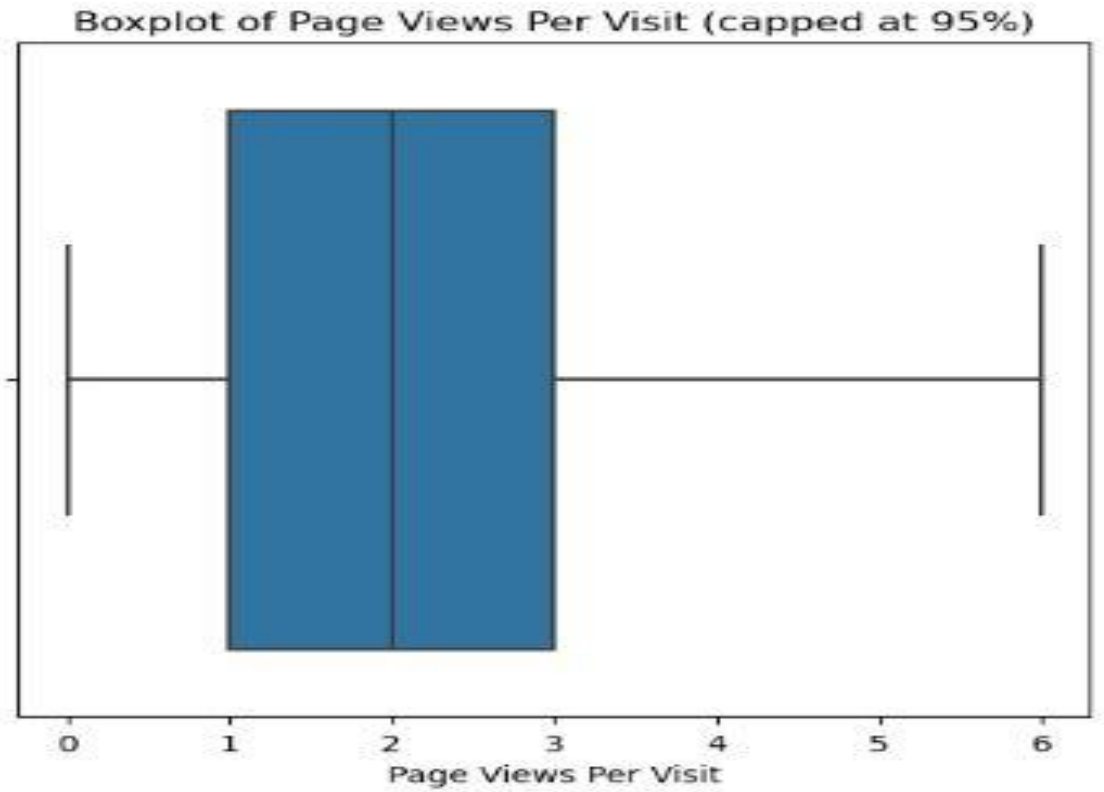
Important factors: High number of people who are interested for this course are from Mumbai

Univariate Data Analysis (Visualization of Numerical columns)



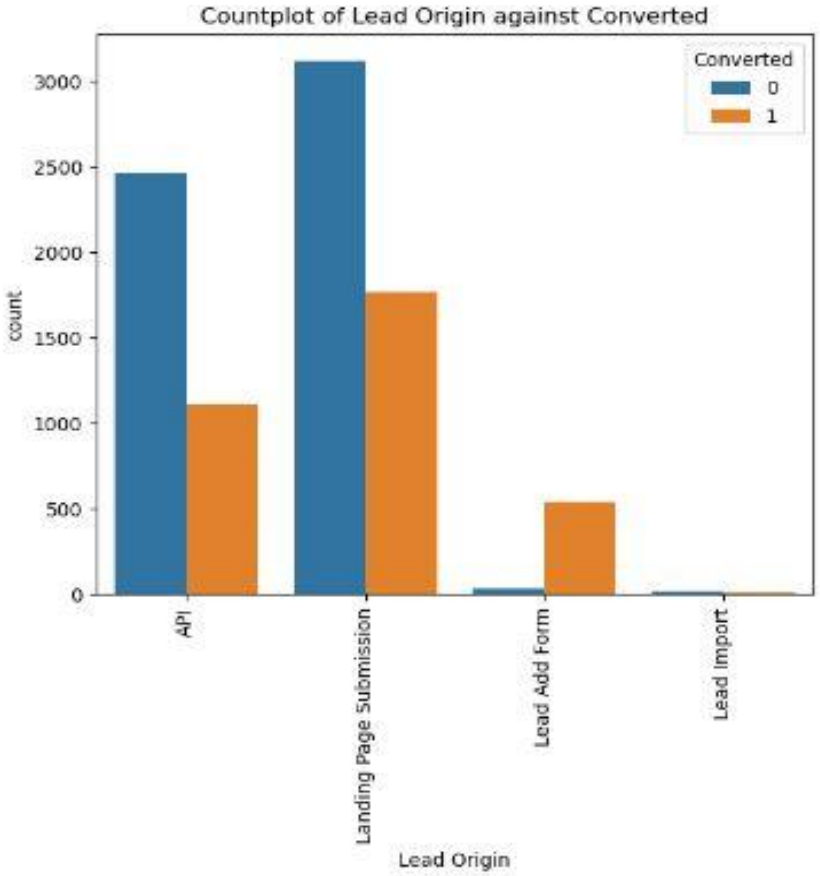


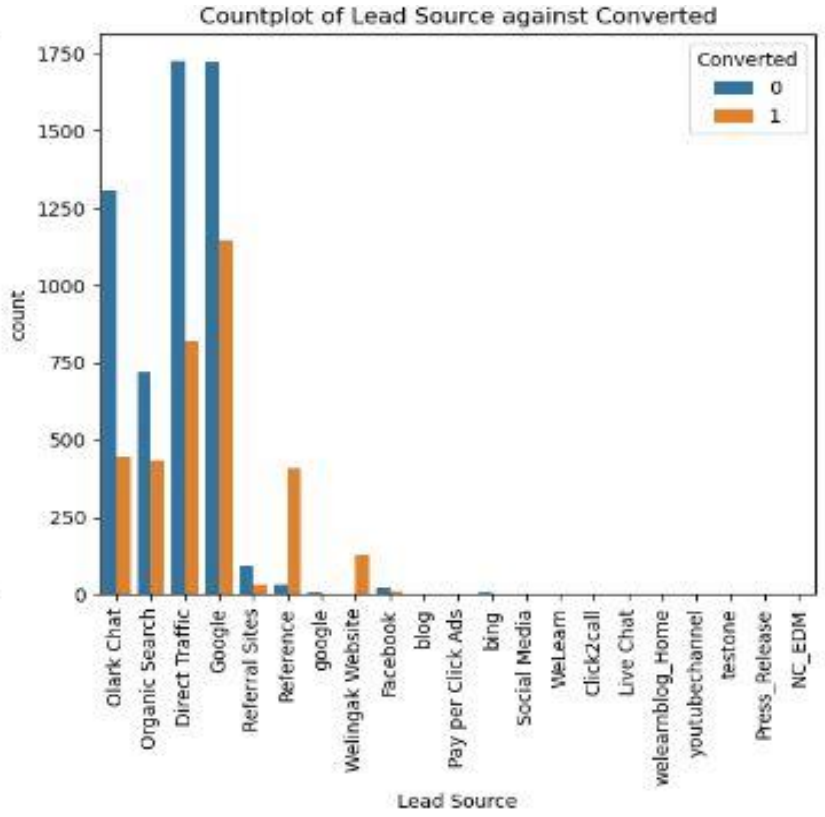


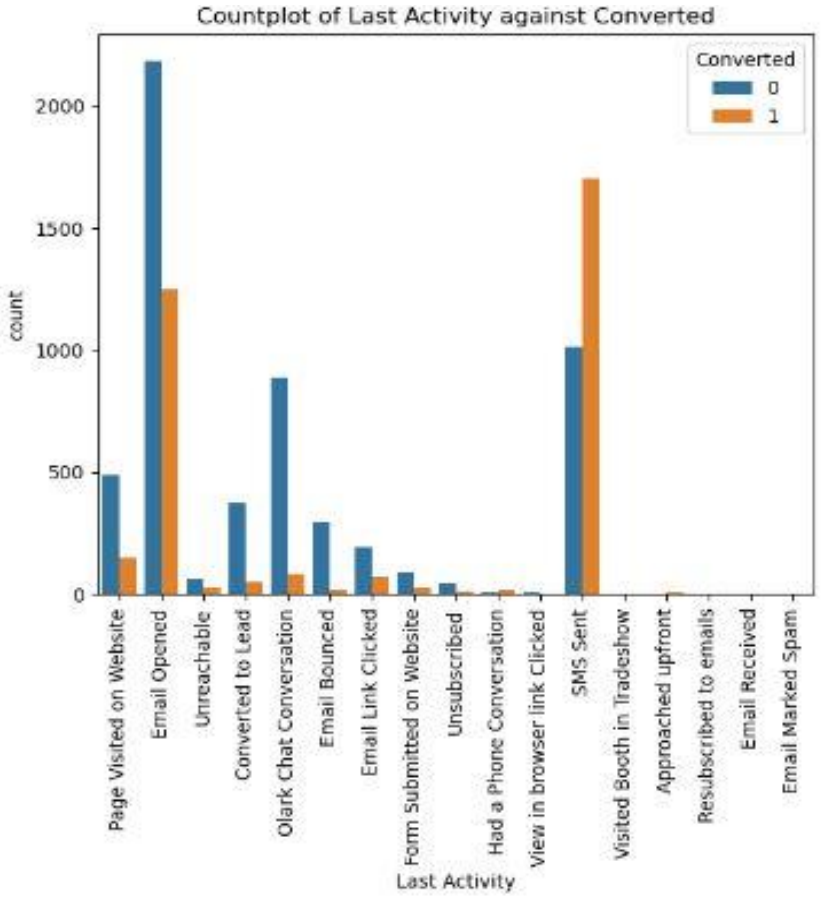


Importanat factors: Here we see that there is Outliers in Assymmetrique profile score. So we Drop that column

Multi variate Data Analysis (Visualization of categorical columns)

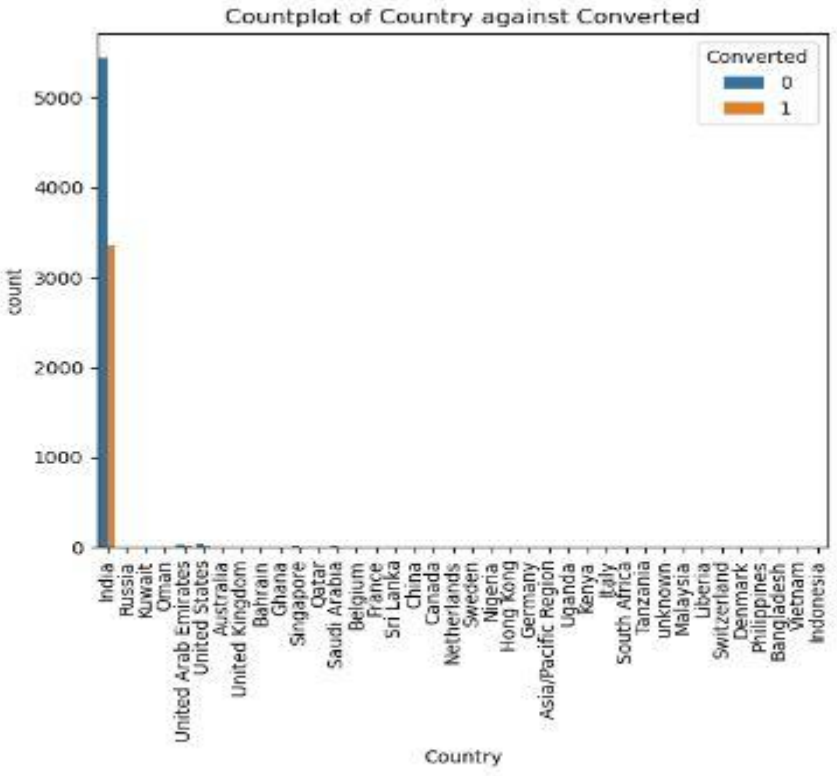


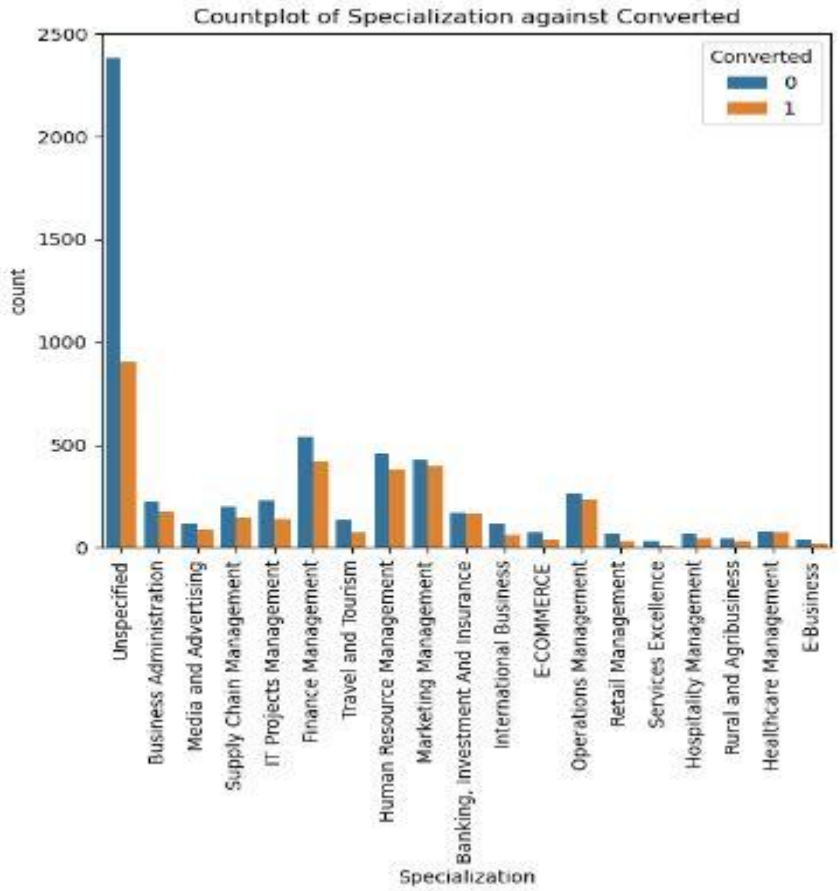


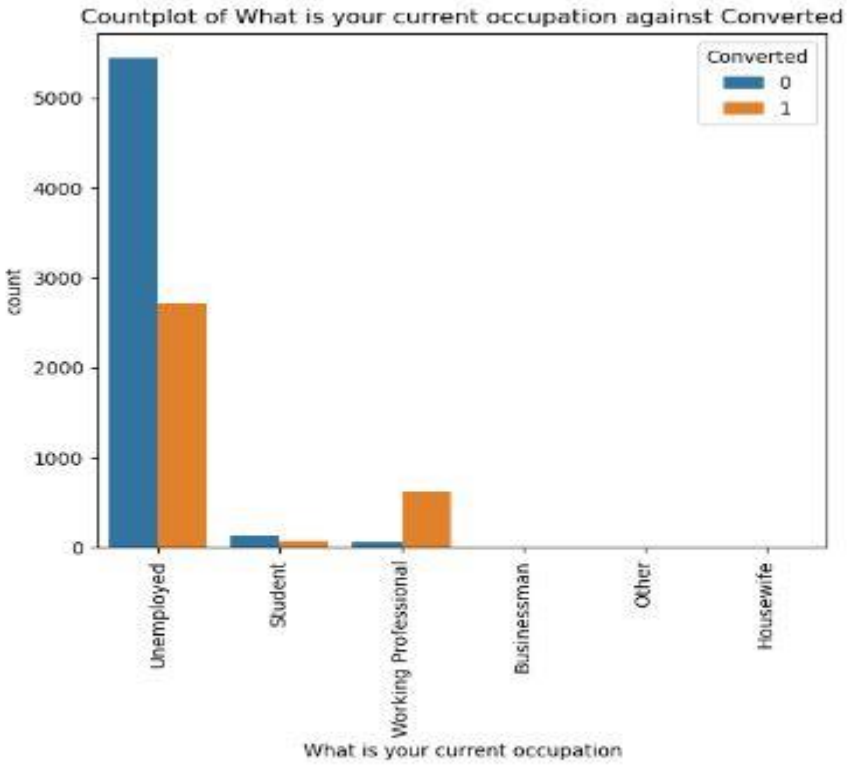


Important Factors:

- The conversion rate of the 'Lead Add Form' category is much higher than its non-conversion rate.
- Additionally, the conversion rate of the 'Reference' category is higher than its non-conversion rate. Furthermore, people whose last activity is 'SMS sent' are mostly converted.



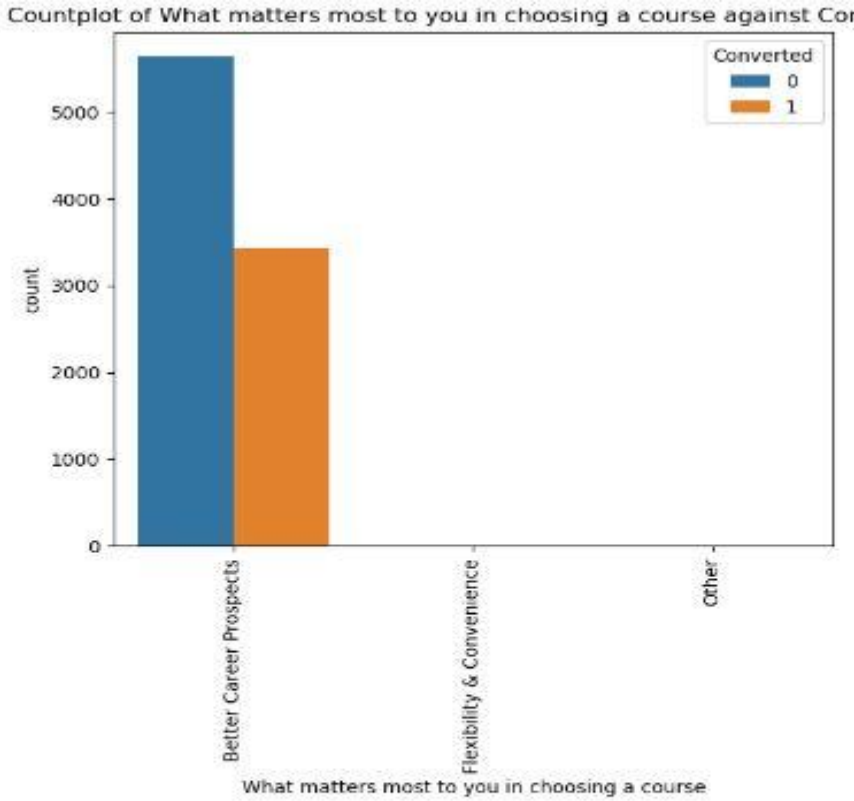


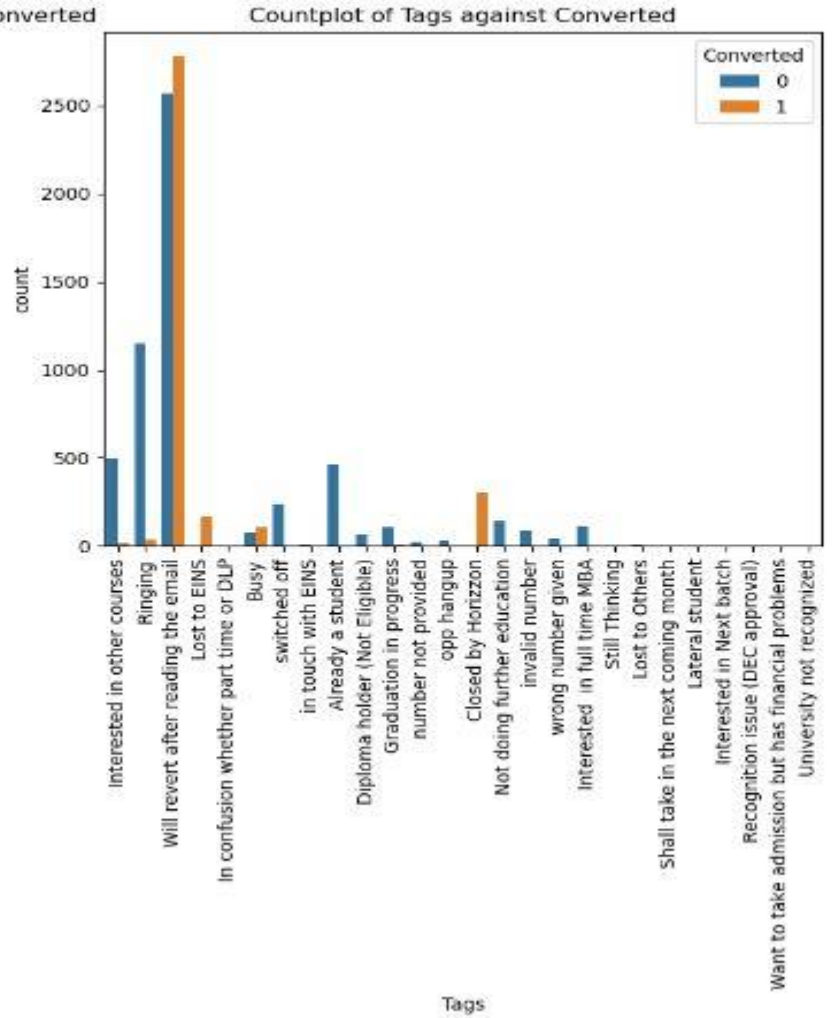


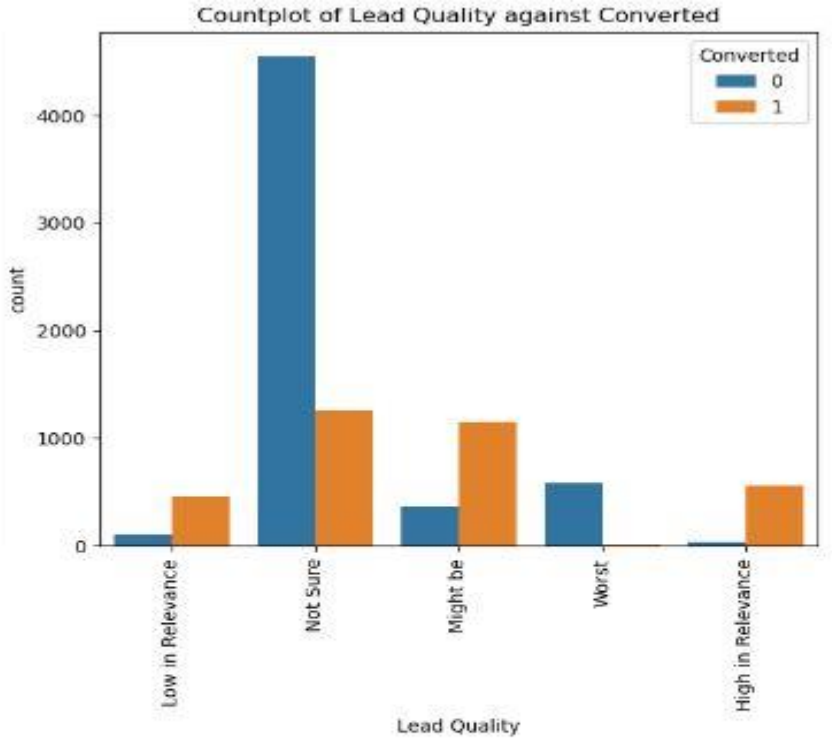
Important Factors:

- We clearly see that the 'County' column has no impact on the target variable because it mostly contains 'India' features. Therefore, we can simply drop this column.
- Both the conversion and non-conversion rates of the categories 'Unspecified' and 'Unemployed' are high. However, the non-conversion rate is higher than the conversion rate.

Important Factors: We clearly see that the column 'What Matters most to you in choosing the course' has no impact on t

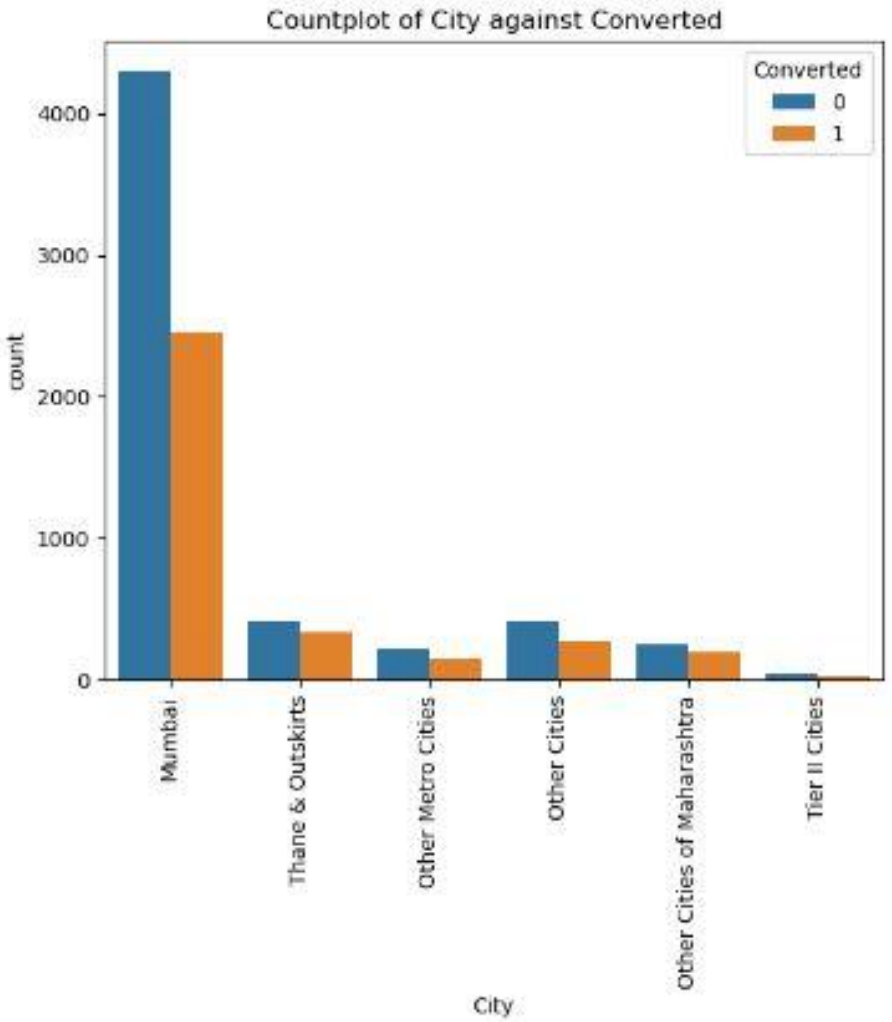


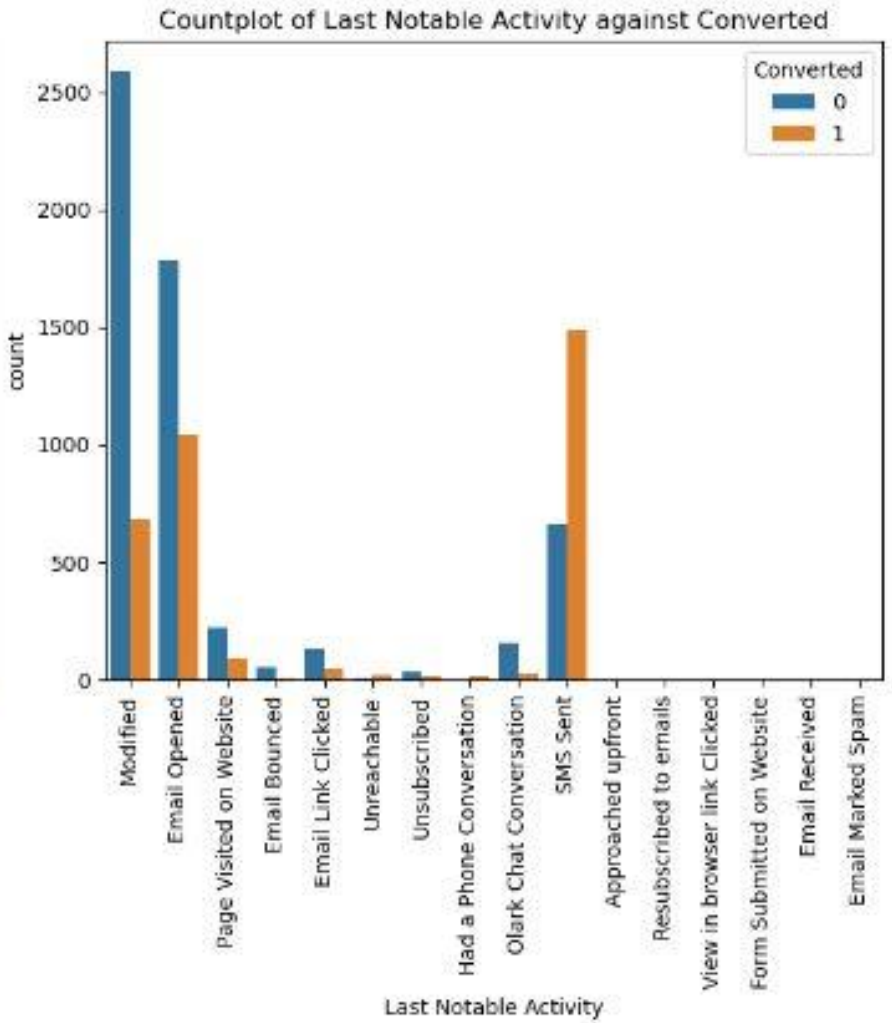




Important Factors:

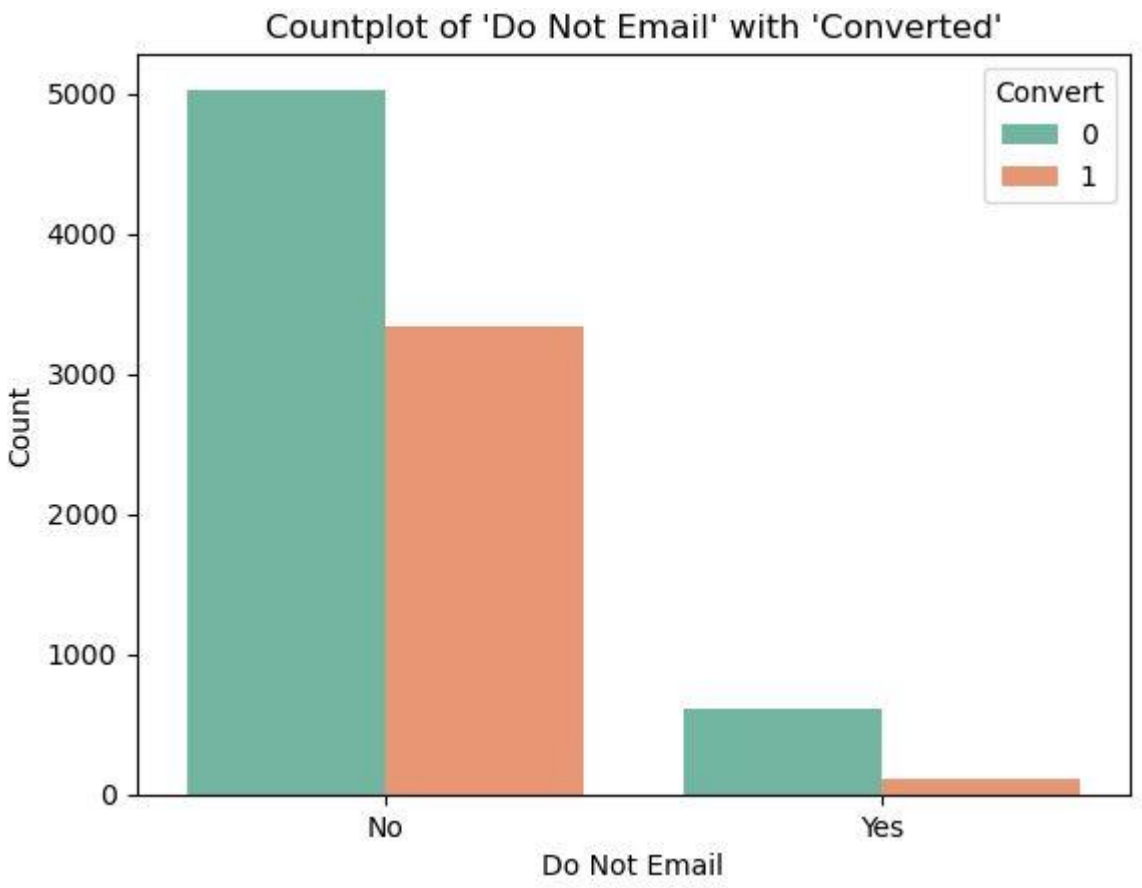
- we clearly see the Column 'What Matters most to you in choosing the course' has no impact on the target Variable. 'what matters most to you in choosing the course' is contain the 'Better Career prospect' feature mostly. So simply we drop this column.
- ‘Will revert after reading the mail’ category has higher converted rate.
- Most of the People who are not sure about the Lead Quality are not converted.

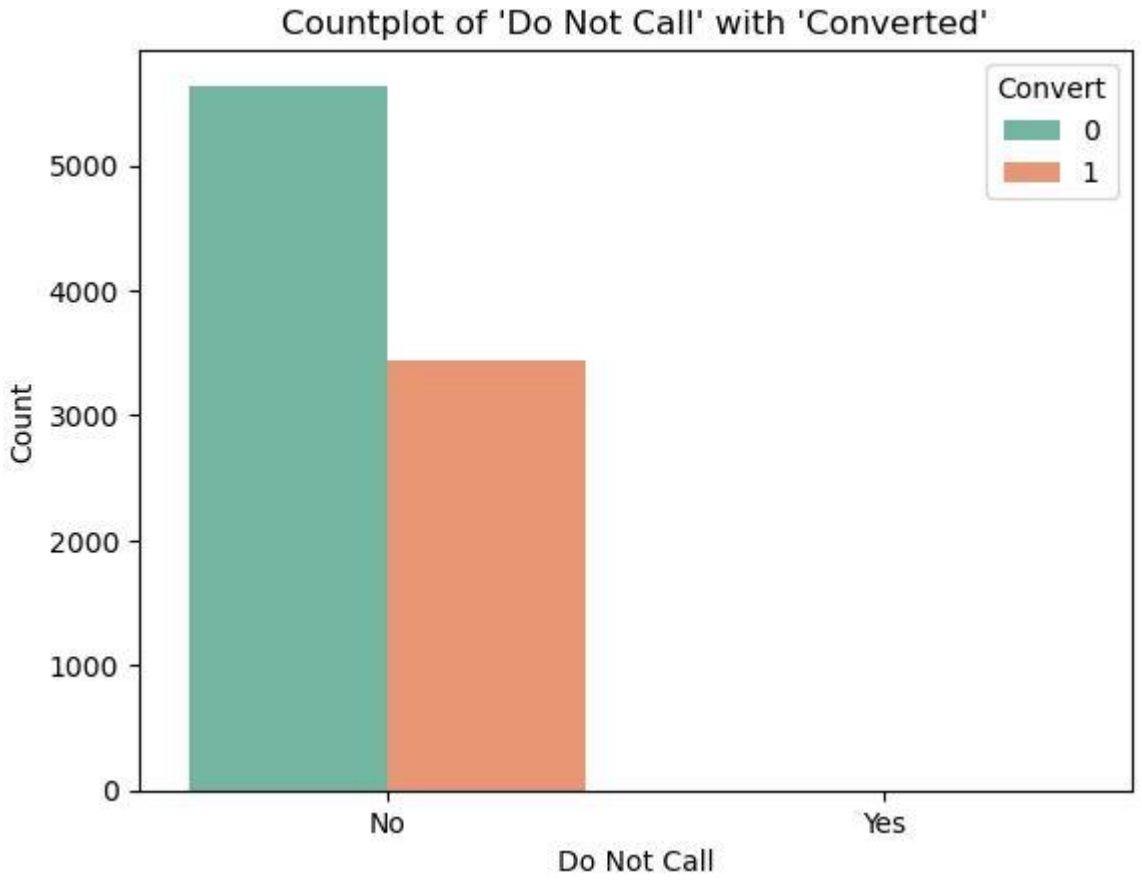




Important Factors:

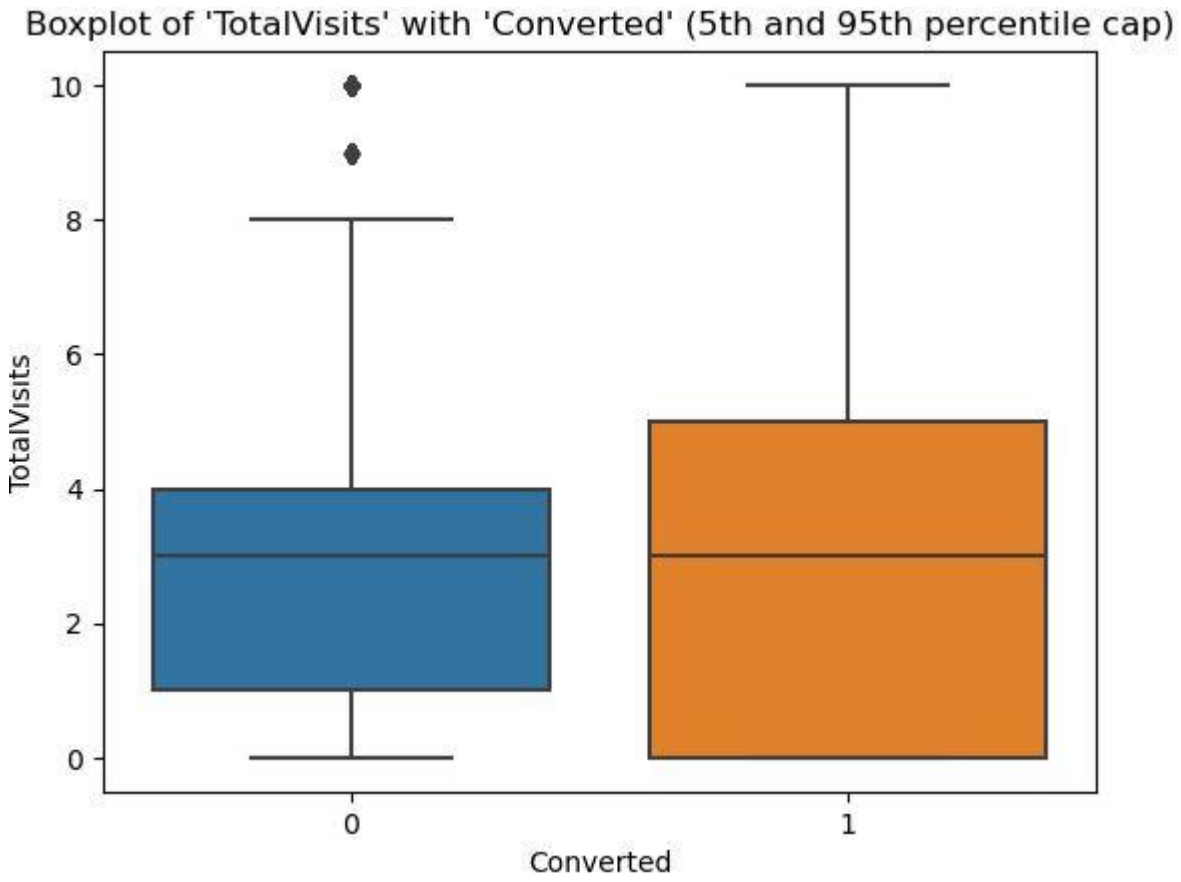
- The conversion and non-conversion rates of Mumbai are high, but the non-conversion rate is even higher.
- The category 'SMS Sent' has a higher conversion rate.

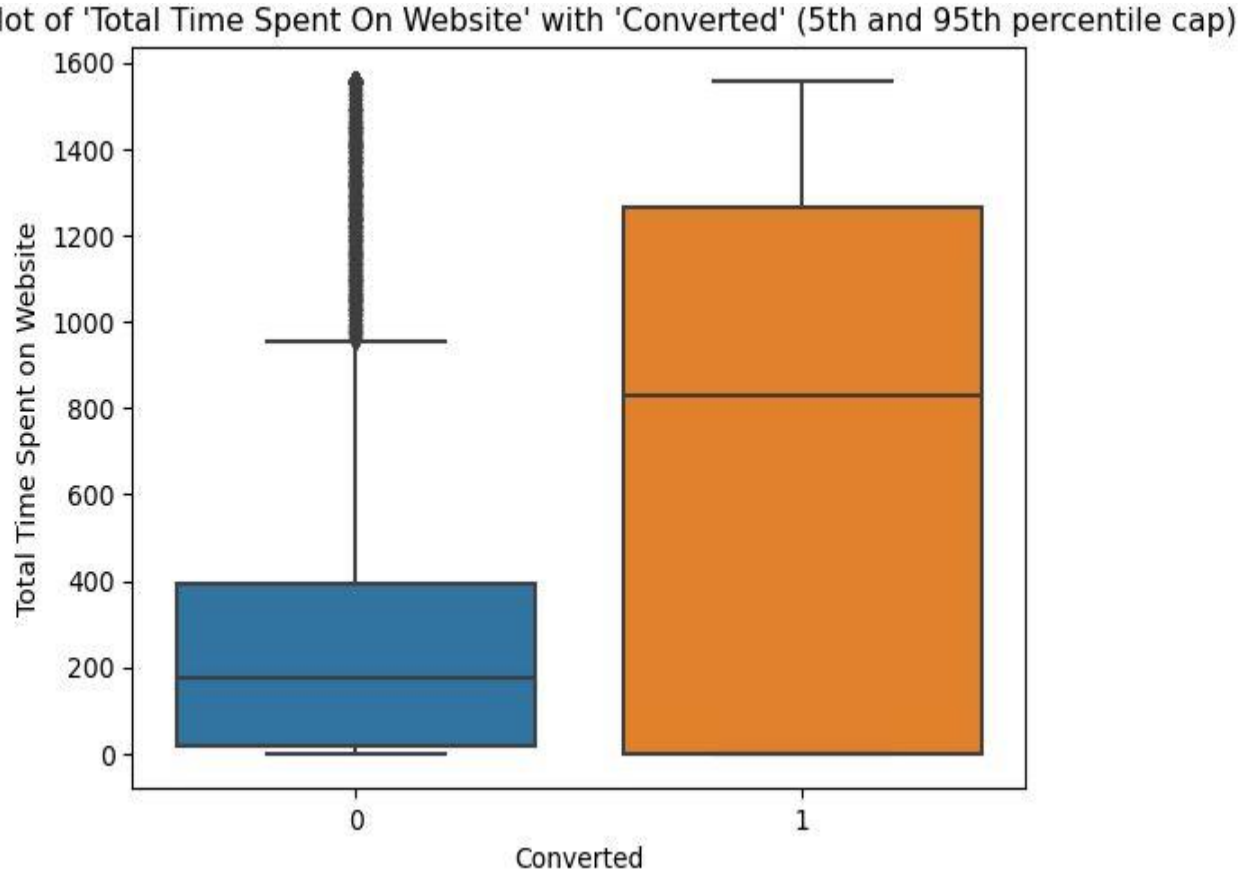




The analysis of the two graphs indicates that the "Do Not Email" column significantly influences the target variable. Instances with a "No" designation have a higher conversion rate, while non-conversion instances are also prevalent within the "No" category. Conversely, the "Do Not Call" column shows minimal impact on the target variable.All Instances are in the "NO" designation. So We drop this column

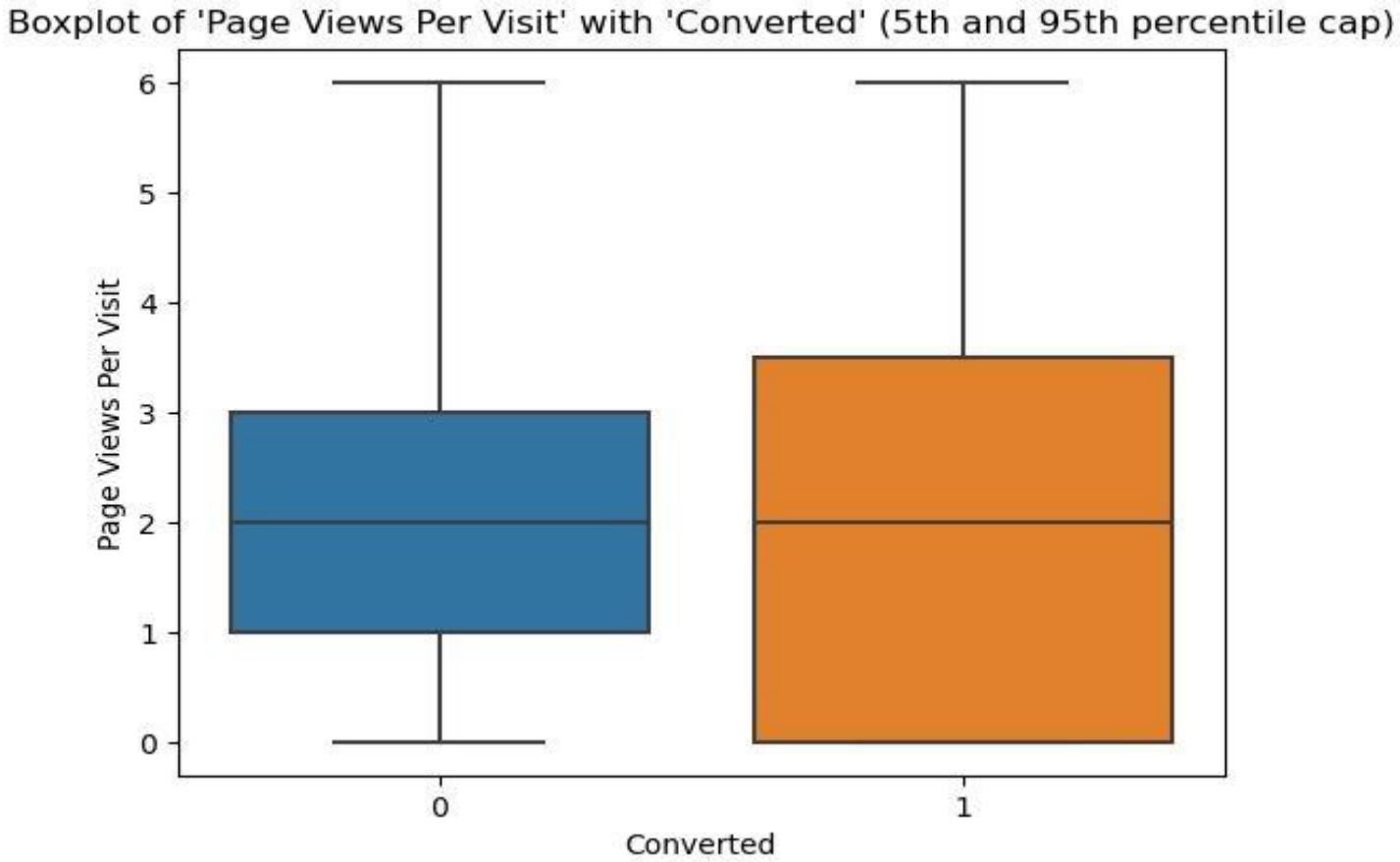
Multi variate Data Analysis (Visualization of Numerical columns)





For Total Visit, we observe that both the non-conversion and conversion features have identical medians. Consequently, it is challenging to determine which factor more significantly affects the target variable

The graph reveals a notable correlation between the conversion rate and the "Total time spent on Website" column. Higher time spent on the website corresponds to a heightened likelihood of course enrollment, indicating that individuals who invest more time on the website are more inclined to enroll in the course.



In this observation, we note that the medians for both conversion and non-conversion instances remain consistent concerning the "Page Views per Visit" column. Consequently, drawing definitive conclusions regarding the influence of this column on the target variable proves challenging.

Model Development

- Data Preperation
  - Create Binary Dummy Variable for the columns which have two features
  - Create Dummy Variable for multiple features
  - Checking the correlation
  - Drop the columns which have higher correlation(>.8)

- - Explanation of logistic regression model for lead scoring
  - - Training and testing of the model (Create test train dataset using sklearn model)
  - - Feature Scaling (use the minmax scaling method)
  - -Feature selection process (Select 15 features for the model using RFE)
  - - Build a Logistic Regression Model and Run it
  - Check the p-values and drop the columns with higher p-value and again run the model
  - Check the VIF value and drop the columns which have higher VIF (>5) and again check it
  - Get the predicted value on train data set

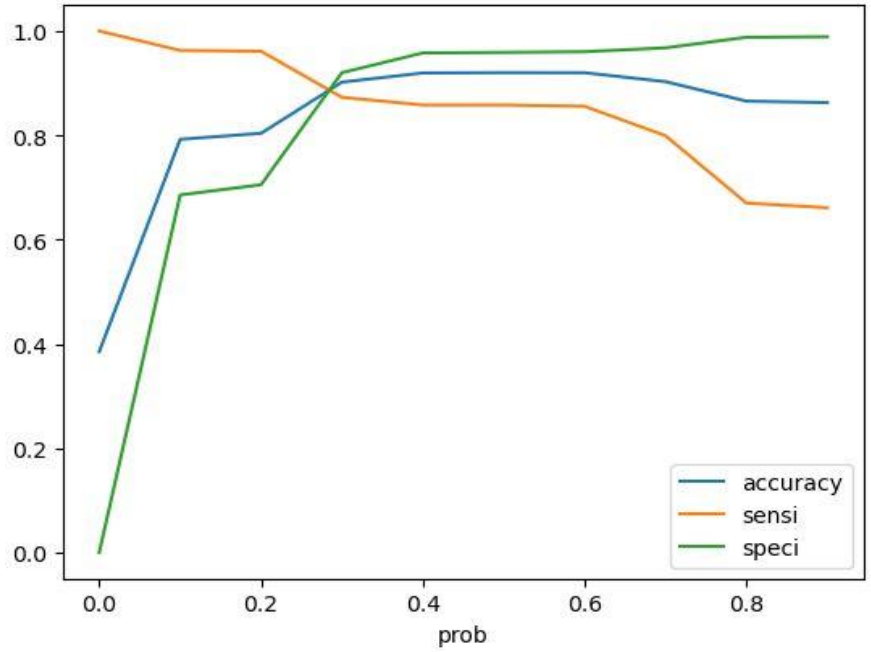
Model Development

- Create a data frame with actual converted and converted probability
- Probability with 50% above are consider as converted predicted lead
- Calculate confusion matrix

Pricision: 0.9291722000885347

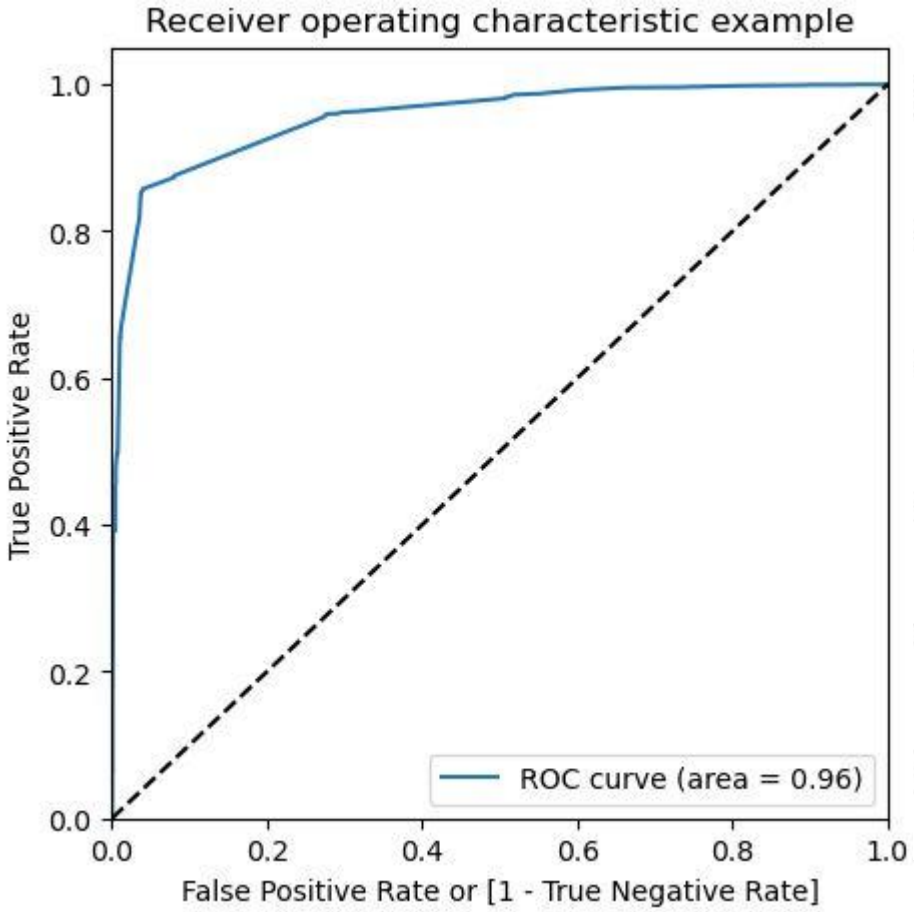
Model Evaluation

- - Evaluation metrics used: accuracy, precision, recall, F1-score
  - Train Data:-
    - Accuracy: 90.19%-
    - Sensitivity: 87.28%-
    - Specificity: 92.01%
    - Precision – 92.91%
    - Recall – 85.81%
  - Test Data:-
    - Accuracy: 89.60%-
    - Sensitivity: 85.23%-
    - Specificity: 92.09%
    - Precision – 86.02%
    - Recall – 85.23%
  - Optimal cutoff =0.3

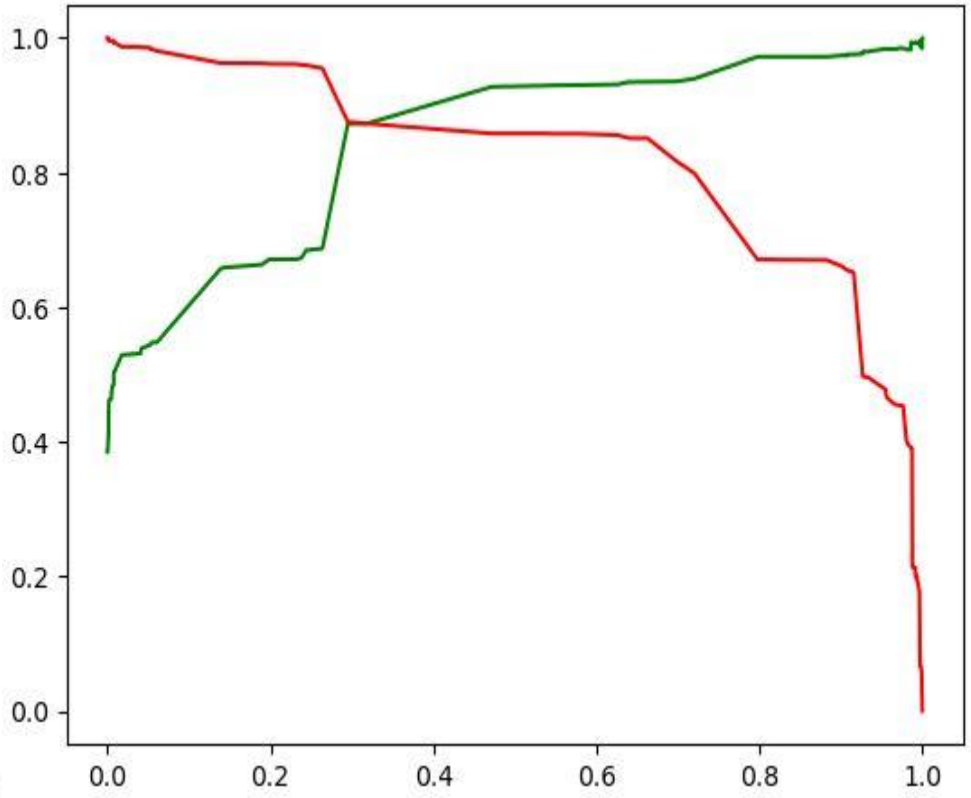


Model Evaluation

- - ROC curve and AUC score(Area= 0.96)



Precision vs Recall Plot



Recommendations

- Focus on catering to freshers, as they constitute a significant portion of potential candidates, especially those who are unspecified about their skills and unemployed.
- Emphasize the career advancement opportunities associated with the course to attract individuals seeking better career prospects.
- Direct marketing efforts towards the Mumbai region, considering the high number of interested individuals from there.
- Prioritize strategies that capitalize on the effectiveness of lead generation forms, as they exhibit a significantly higher conversion rate.
- Allocate resources towards promoting references from existing participants, as they tend to yield a higher conversion rate.
- Implement targeted follow-up strategies for individuals indicating an intention to revert after reading emails, as they exhibit a higher conversion rate.
- Address concerns regarding lead quality assessment, as individuals unsure about lead quality are less likely to convert.
- Enhance the website's user experience and content to encourage prolonged engagement, as higher time spent on the website correlates positively with conversion rates.
- The top three dummy variables in the model are Tags_lost to EINS, Tags_Ringing, and Tags_Closed by Horizzon. Therefore, it is crucial to focus on these variables when analyzing the data and making decisions.

Conclusion

- In conclusion, the analysis highlights several key insights that can guide the company's marketing and enrollment strategies effectively. By focusing on targeting freshers, emphasizing career advancement opportunities, prioritizing regions like Mumbai, optimizing lead generation forms, leveraging references and email follow-ups, addressing lead quality concerns, and enhancing website engagement, the company can maximize its conversion rates and drive course enrollments. Thank you for the opportunity to provide recommendations based on the data analysis.

Thank You...