Lead Scoring Case Study

Kamal Tamang and Siddesh Sawant

Problem Statement

- An education company named X Education sells online courses to industry professionals. It gets a lot of leads through various search engine e.g. - google, but the lead conversion rate is very poor.
- 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

- X Education want to know the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- To build a model which make this process more efficient and identify the most potential leads, also known as ‘Hot Leads’.
- Deployment of the model for the future references.

Solution

- Data cleaning and manipulation
    1. Remove duplicate data.
    2. Check and handle NA and missing values.
    3. Drop columns, if it contains large amount of missing values which are not useful for analysis.
    4. Imputation of values, if required.
    5. Check outliner in data.

- EDA
    1 Univariate data analysis: Value count, Distribution of variables etc.
    2 Bivariate data analysis: Correlation coefficients and pattern between the variables etc.

- Feature scaling & dummy variables and encoding of the data

- Classification technique: Logistic regression used for the model making and prediction.

- Validation of the model

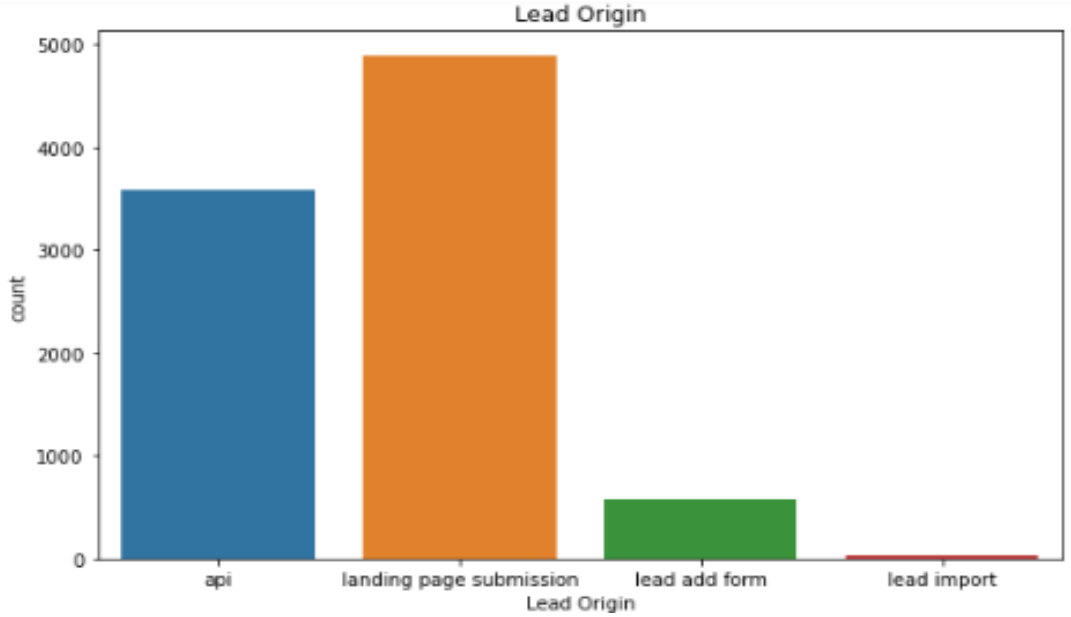
- Model presentation

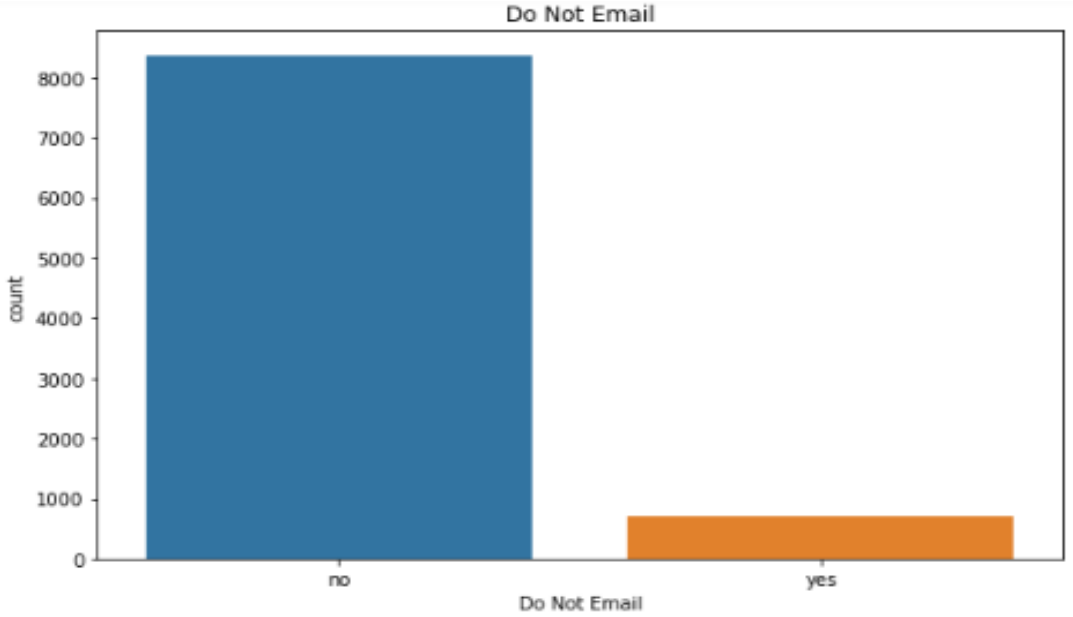
- Conclusion and recommendations

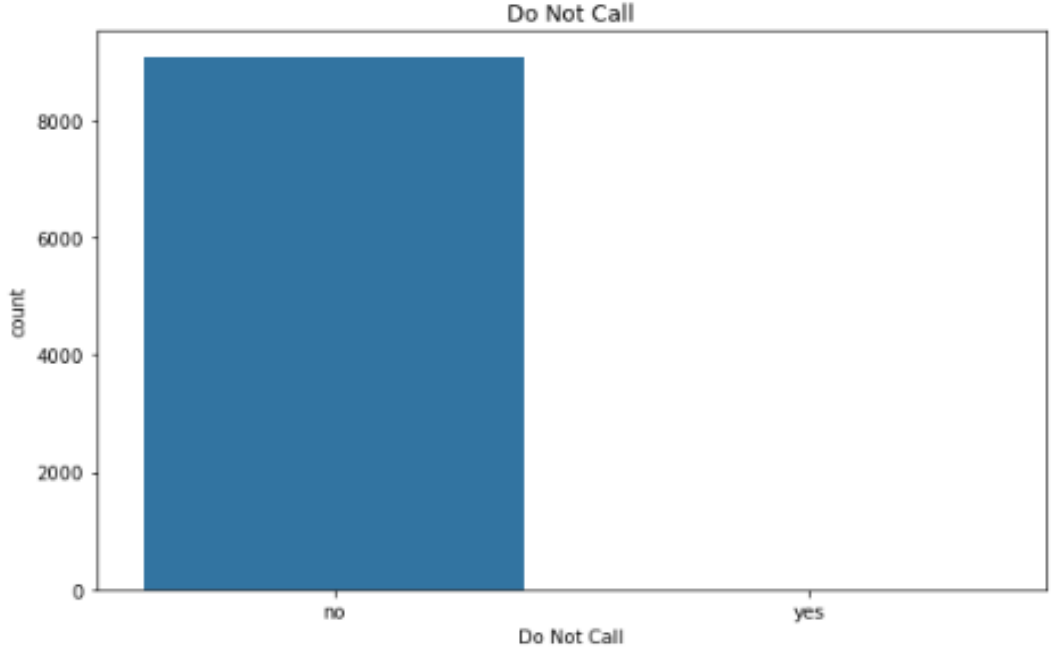
Data Manipulation

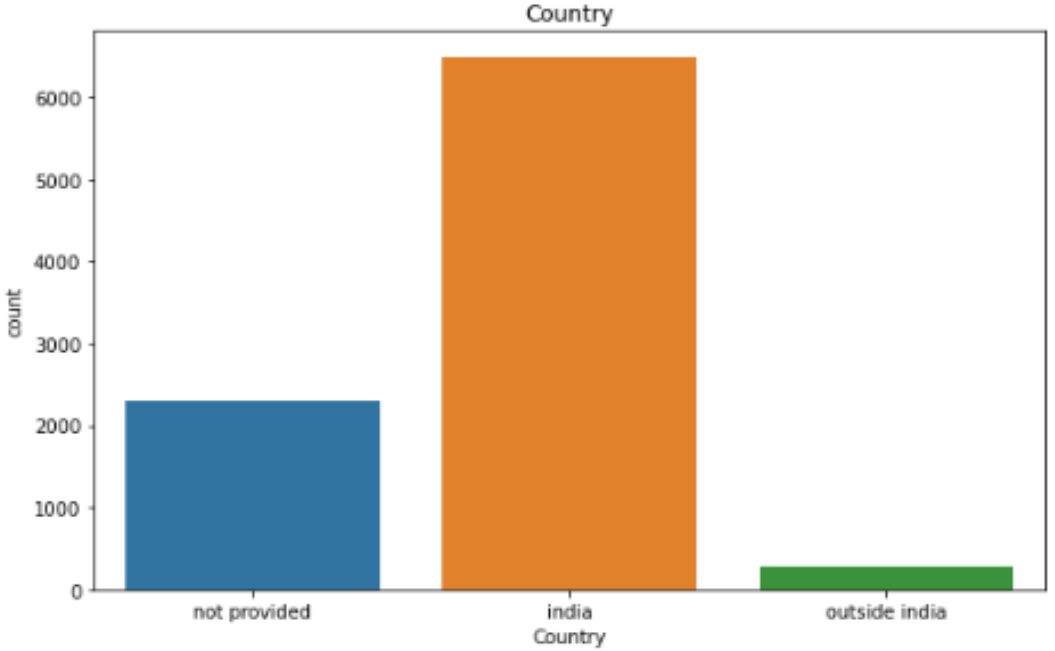
- Total number of row = 37, Total number of columns = 9240
- Single value features like ‘Magazine’, ‘Received more updated about our course’, ‘Update me on supply‘, ‘Chain content’, ‘Get updates on DM content', ‘I agree to pay the amount through cheque' etc. have been dropped.
- Removing the ‘Prospect ID' and 'Lead number‘ which is not necessary for analysis.
- After checking the value counts for some of the objective type variables, we find some of the feature which didn't have enough variance are ‘Do not call', ‘What matters most to you in choosing course’, ‘Search', ‘Newspapers article', ‘X Education forums’, ‘Newspaper’ , ‘Digital Advertising’ etc. are dropped.
- Dropping the column having more than 35% as missing values such as ‘How did you hear about X education' and ‘Lead profile’.

Exploratory Data Analysis

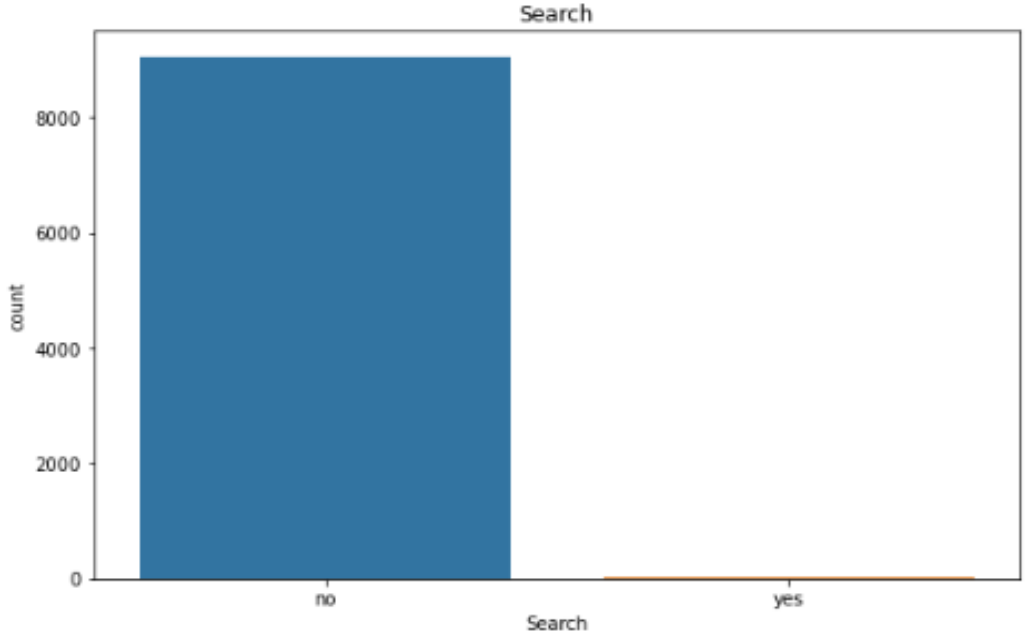


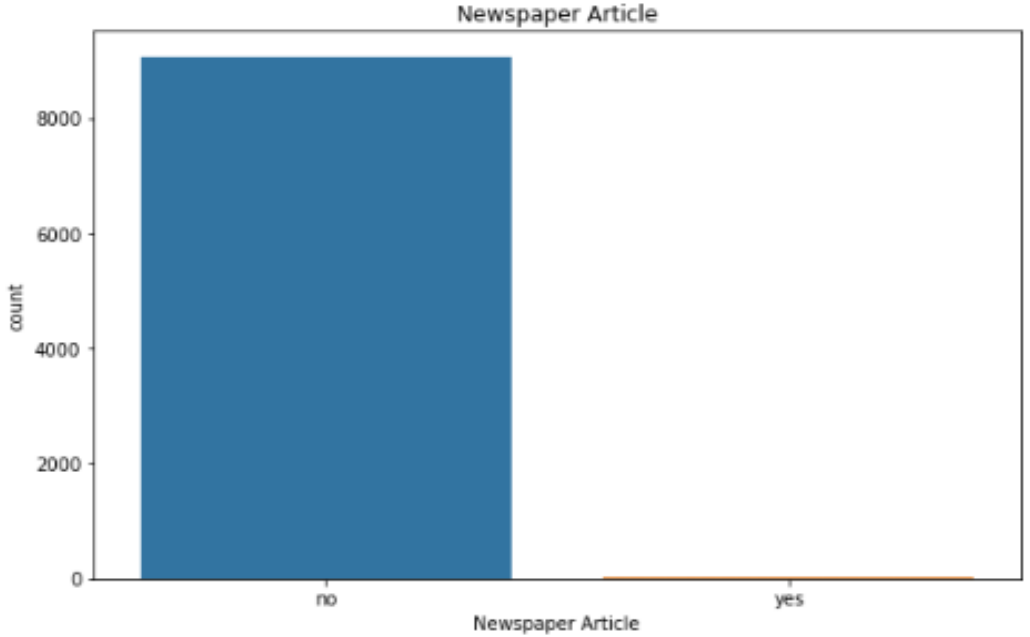


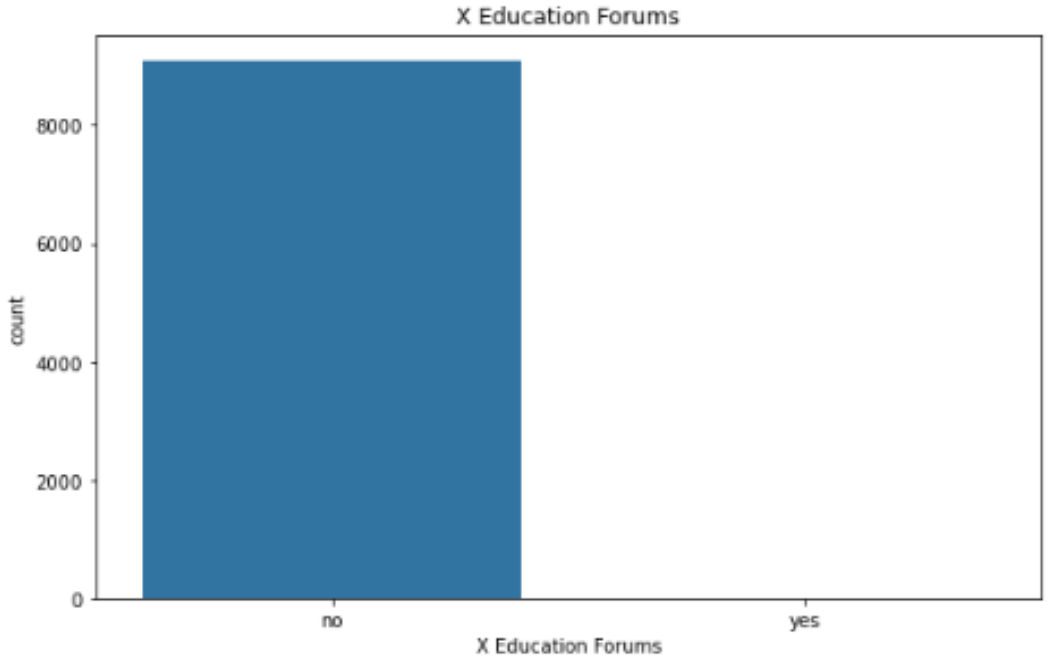


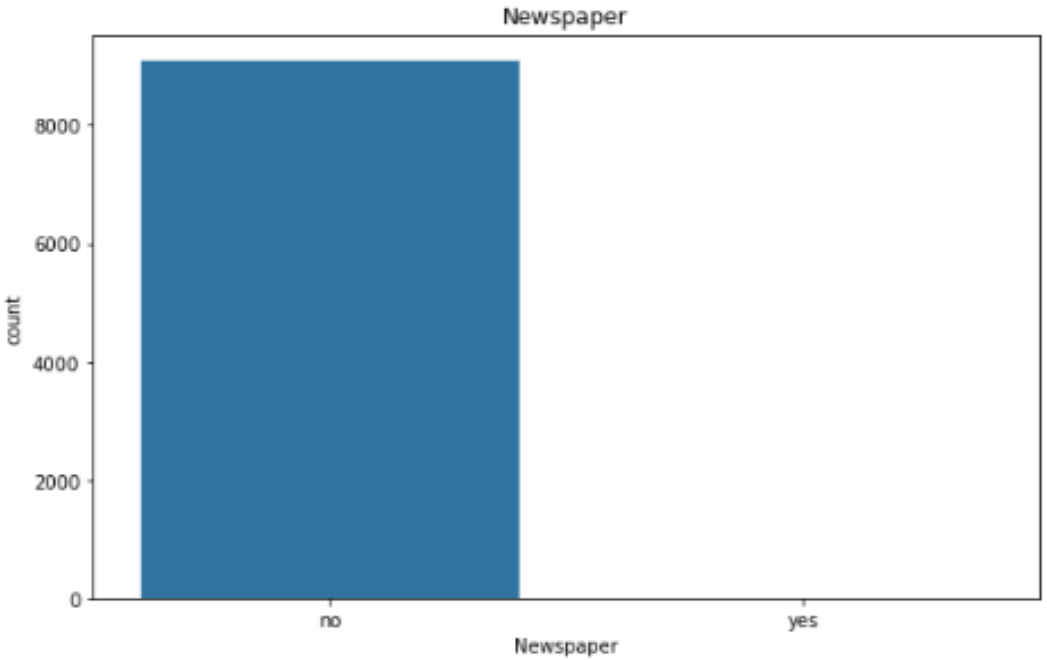


Exploratory Data Analysis

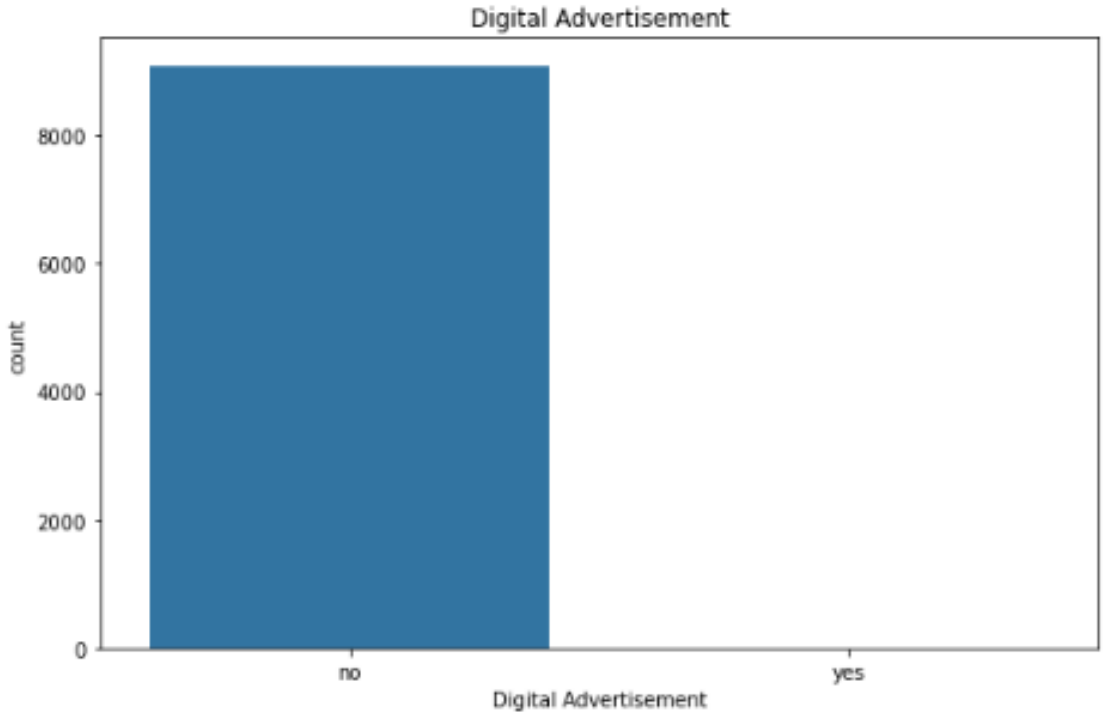


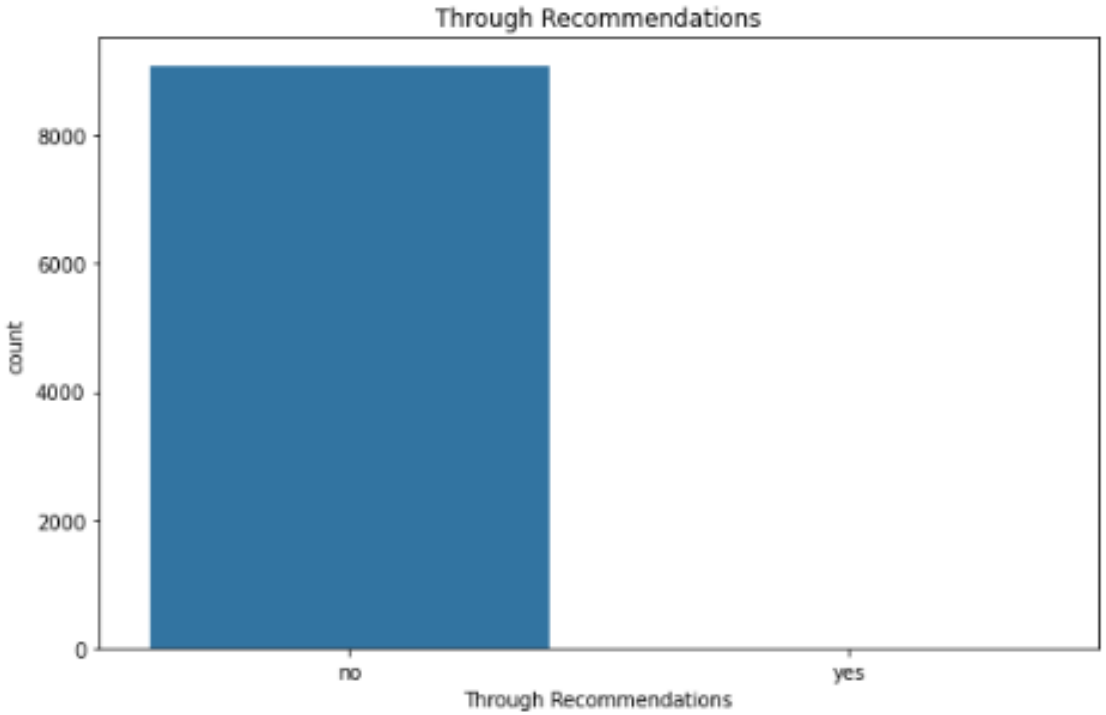




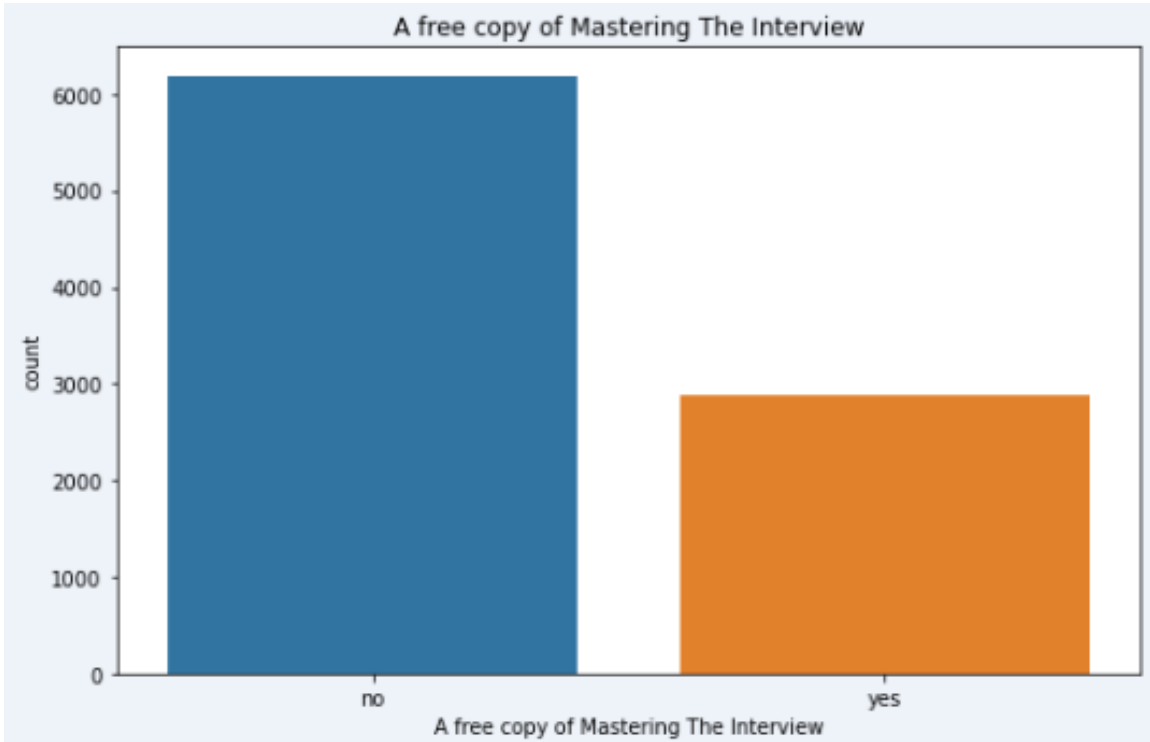


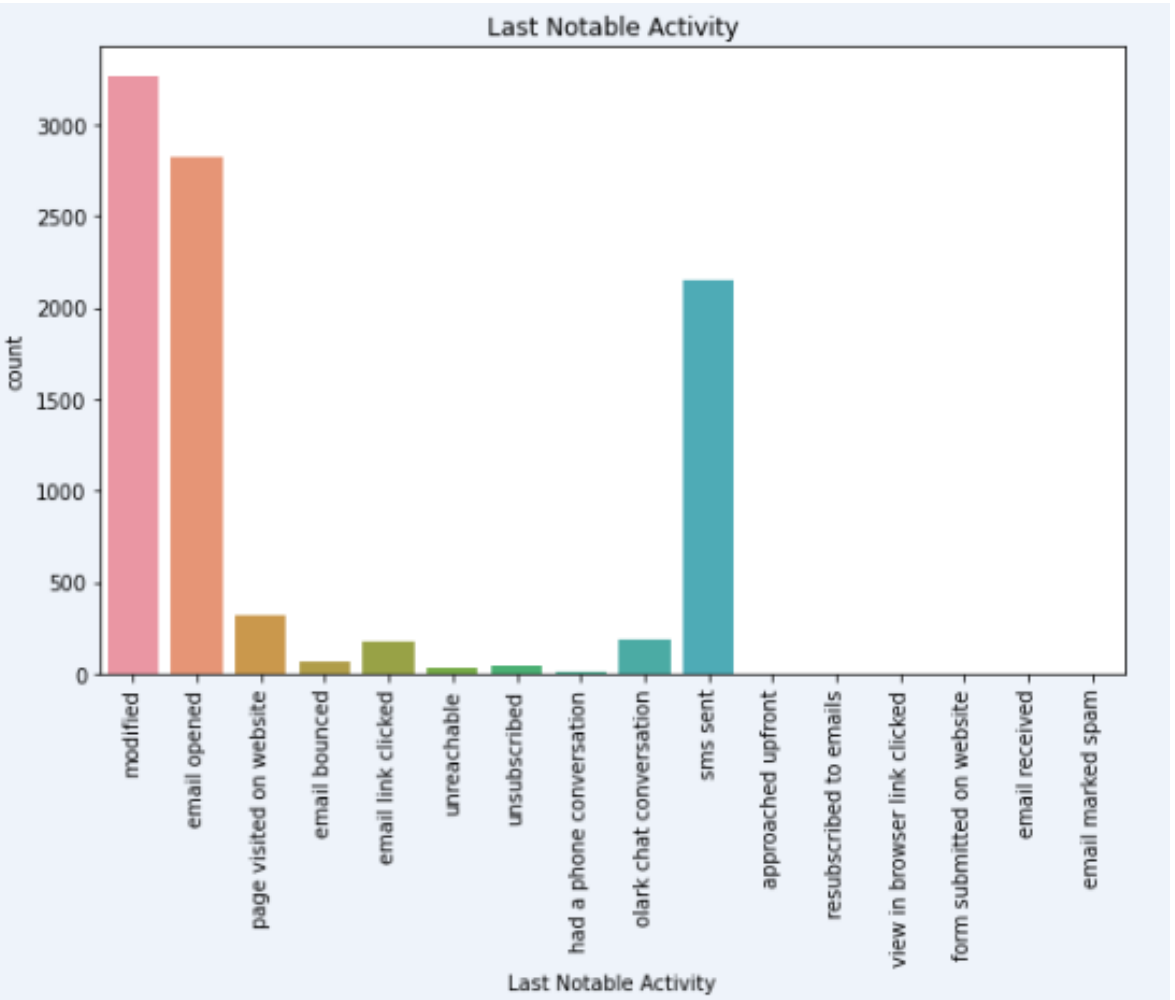
Exploratory Data Analysis



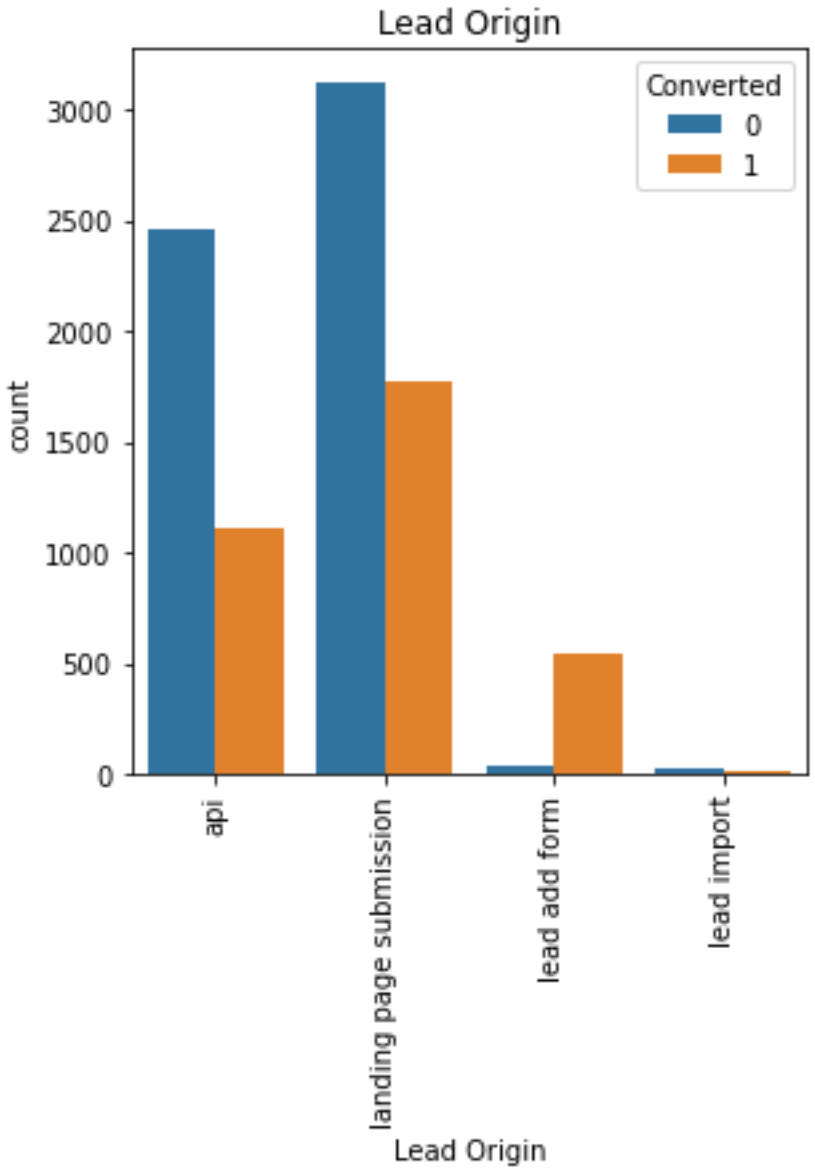


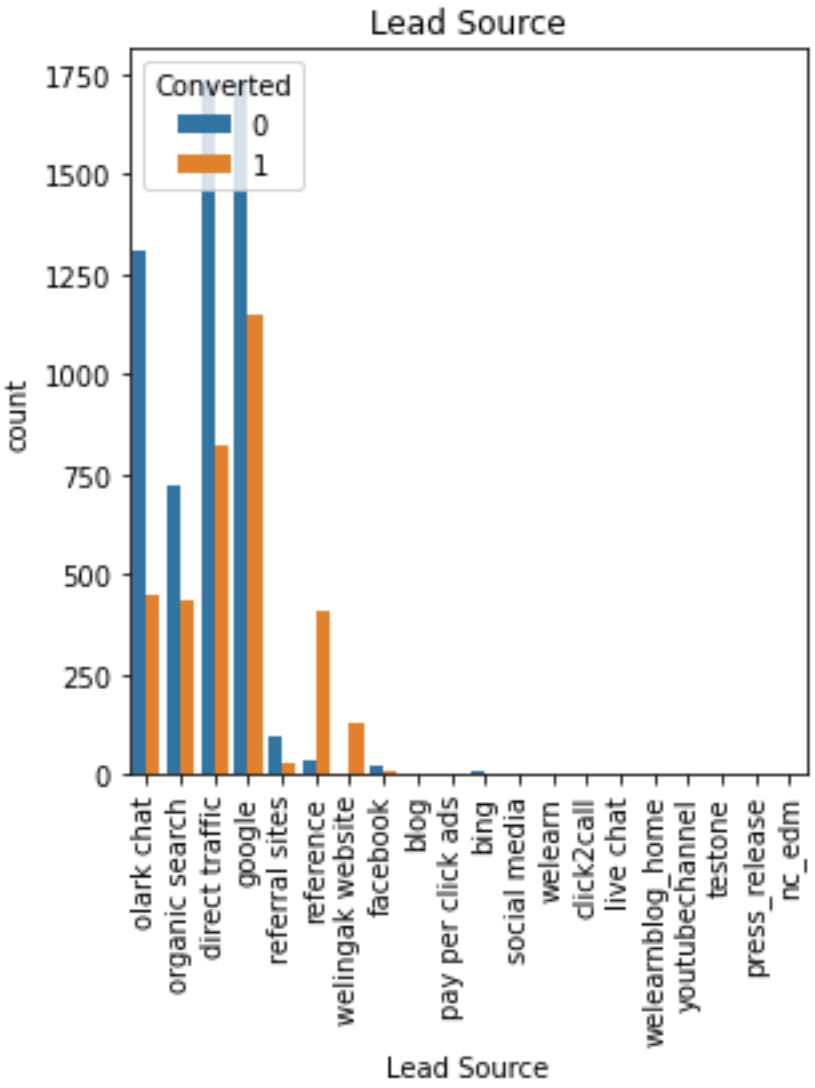
Exploratory Data Analysis



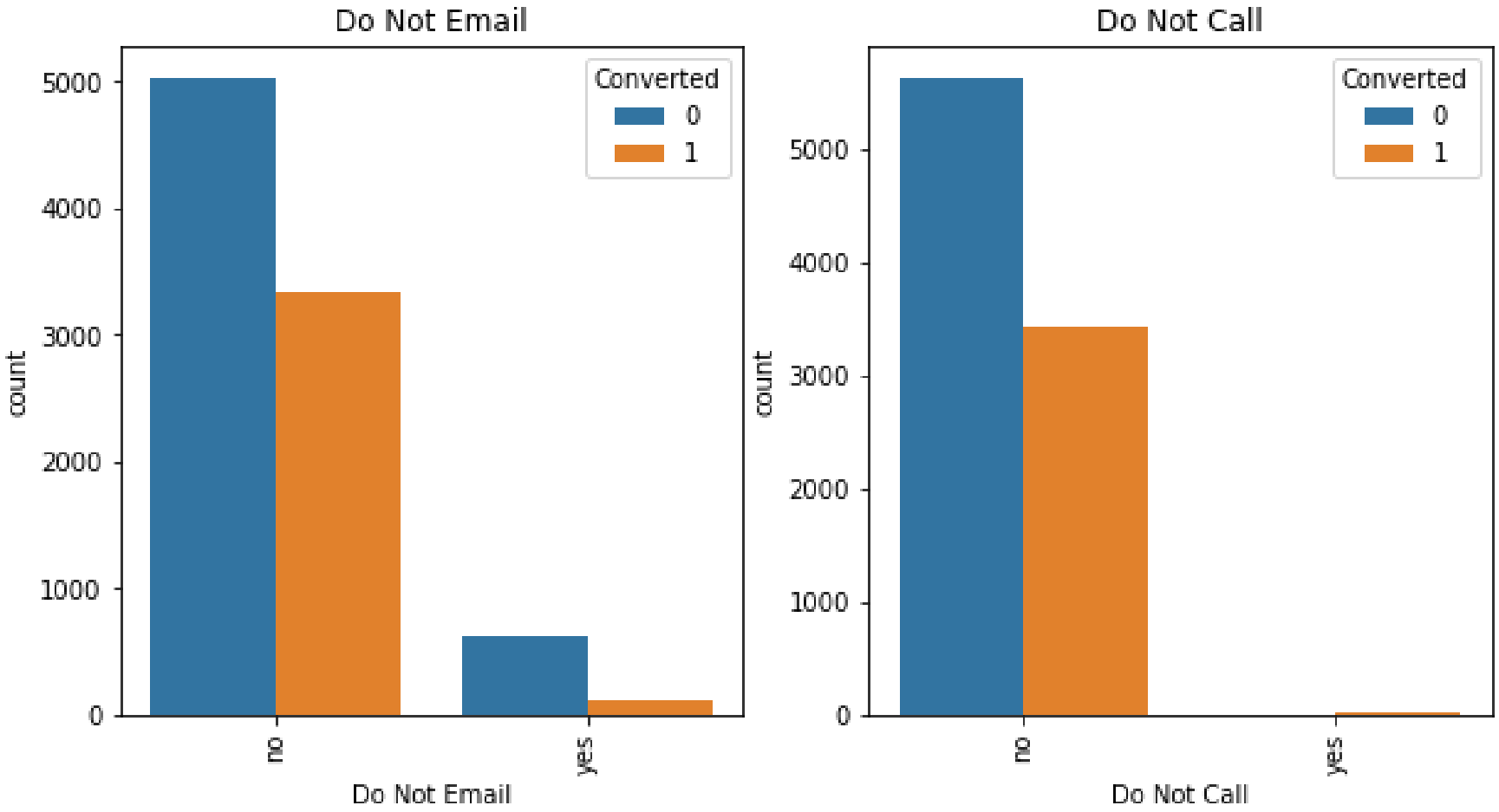


Categorical Variable Relation

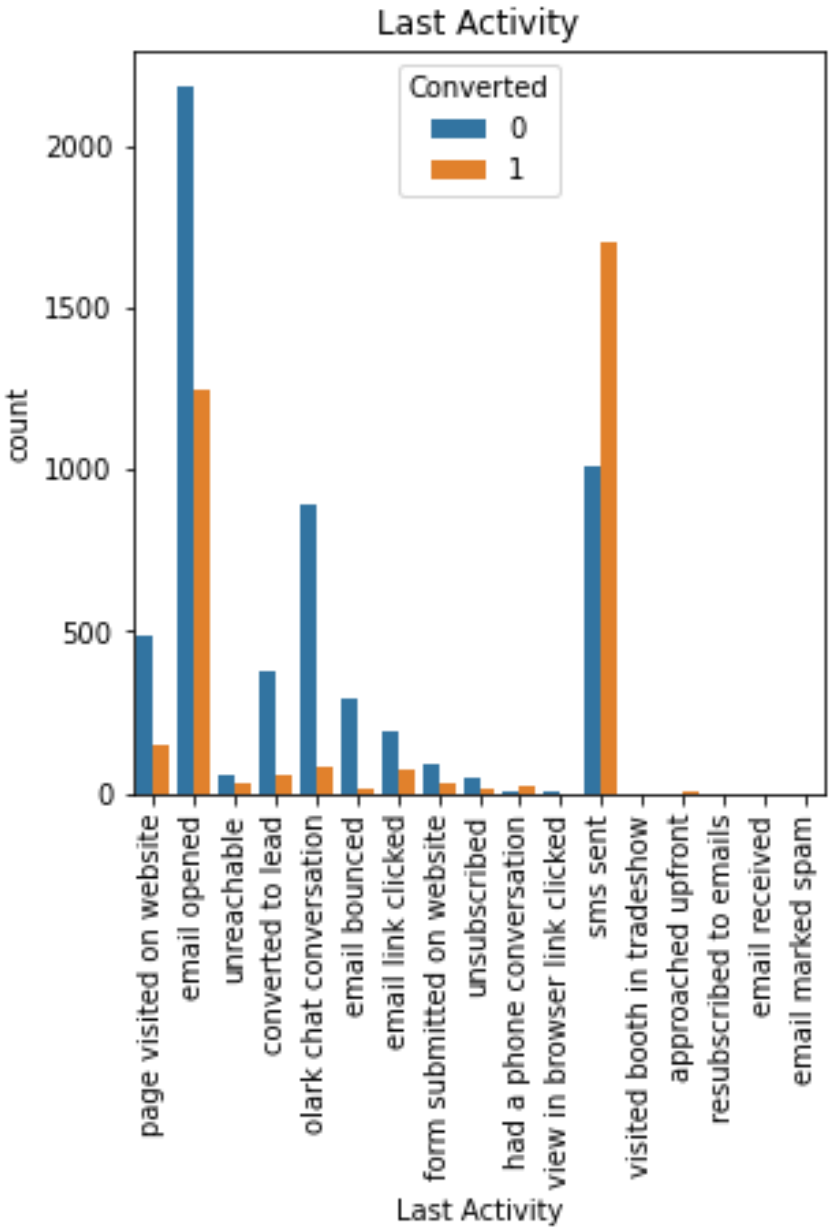


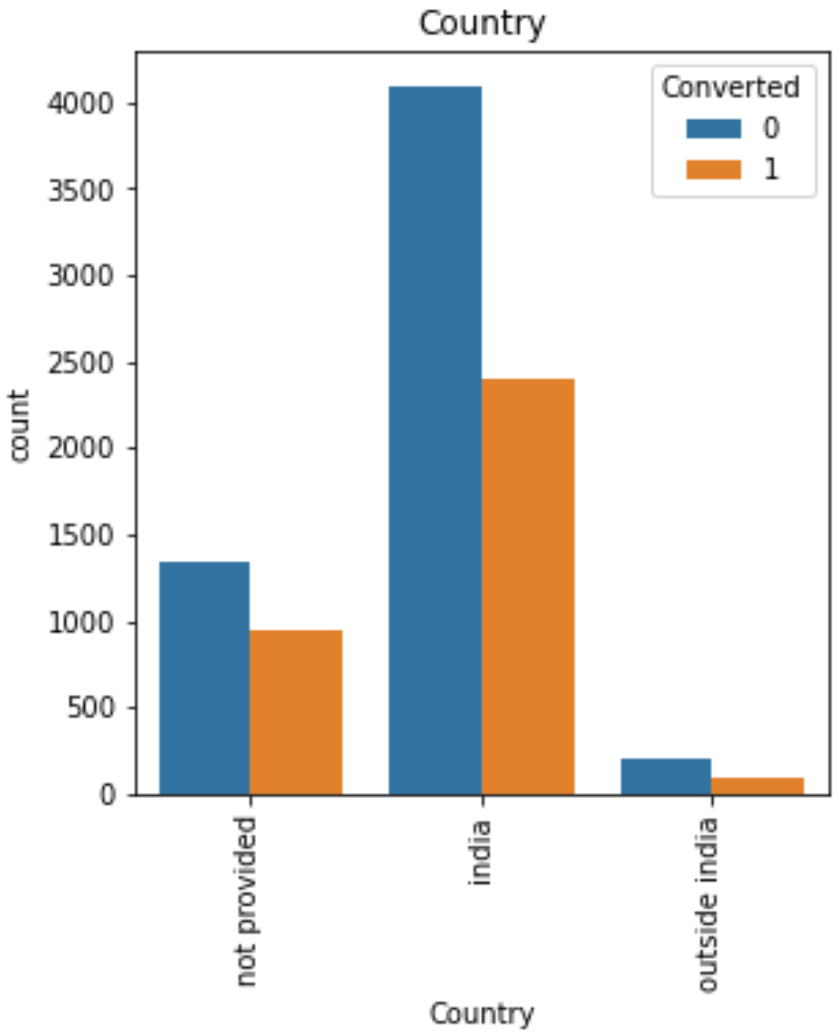


Categorical Variable Relation

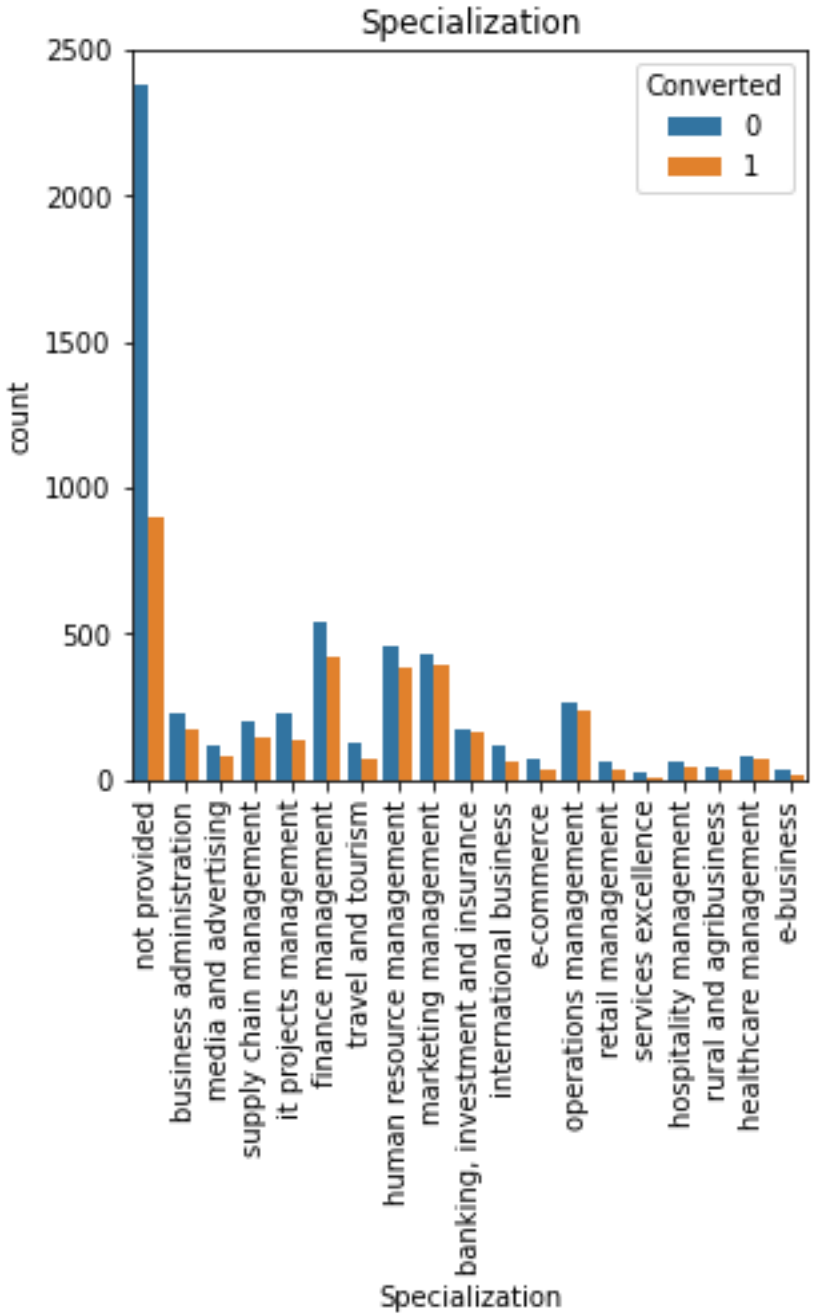


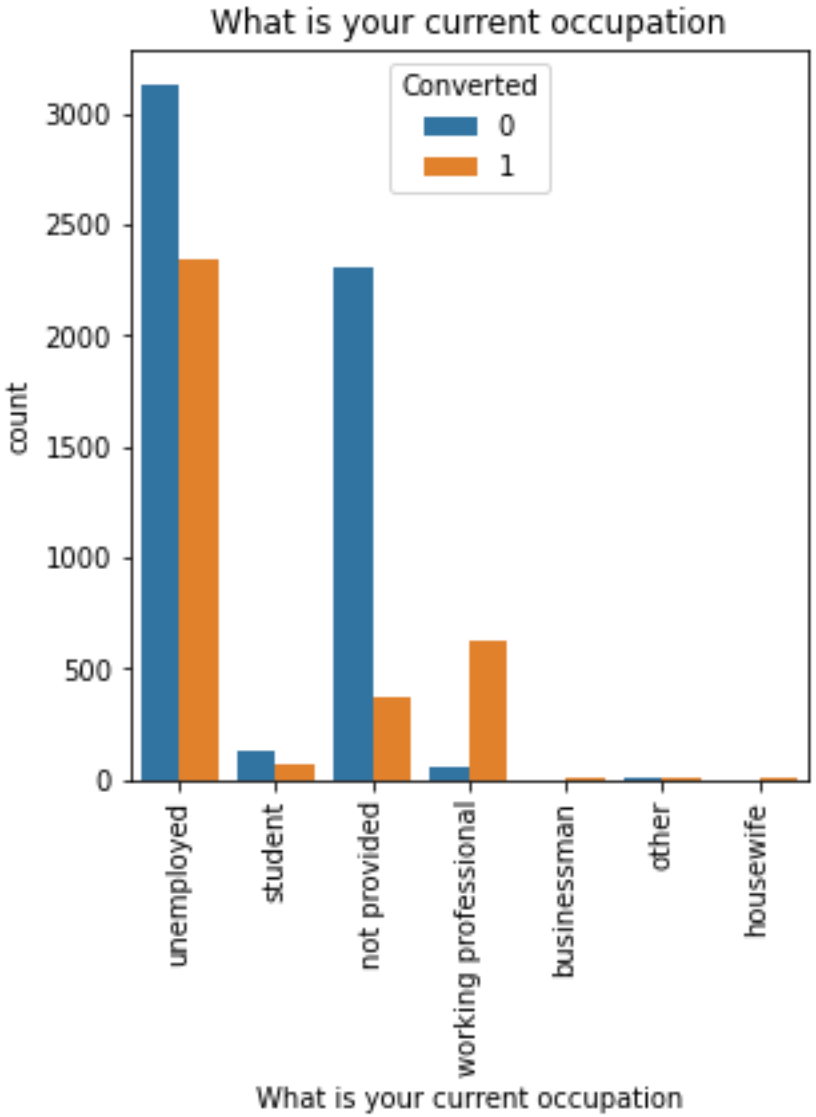
Categorical Variable Relation



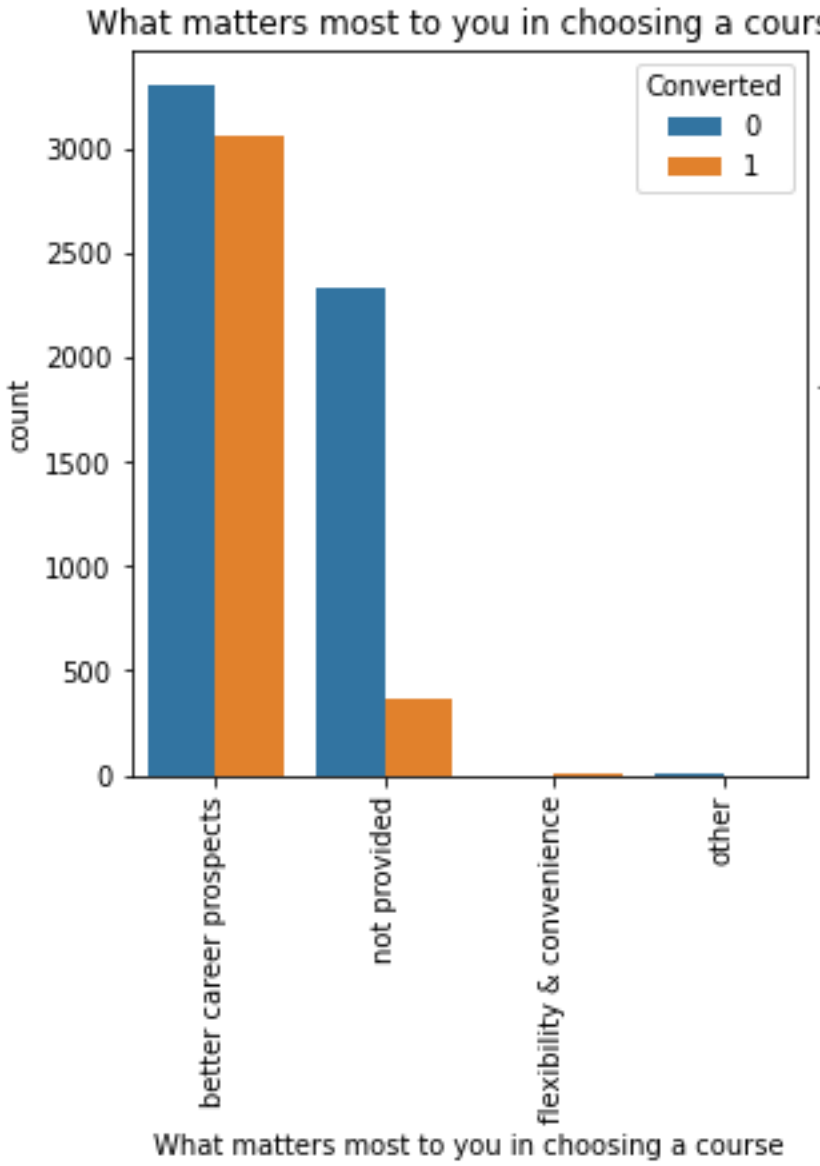


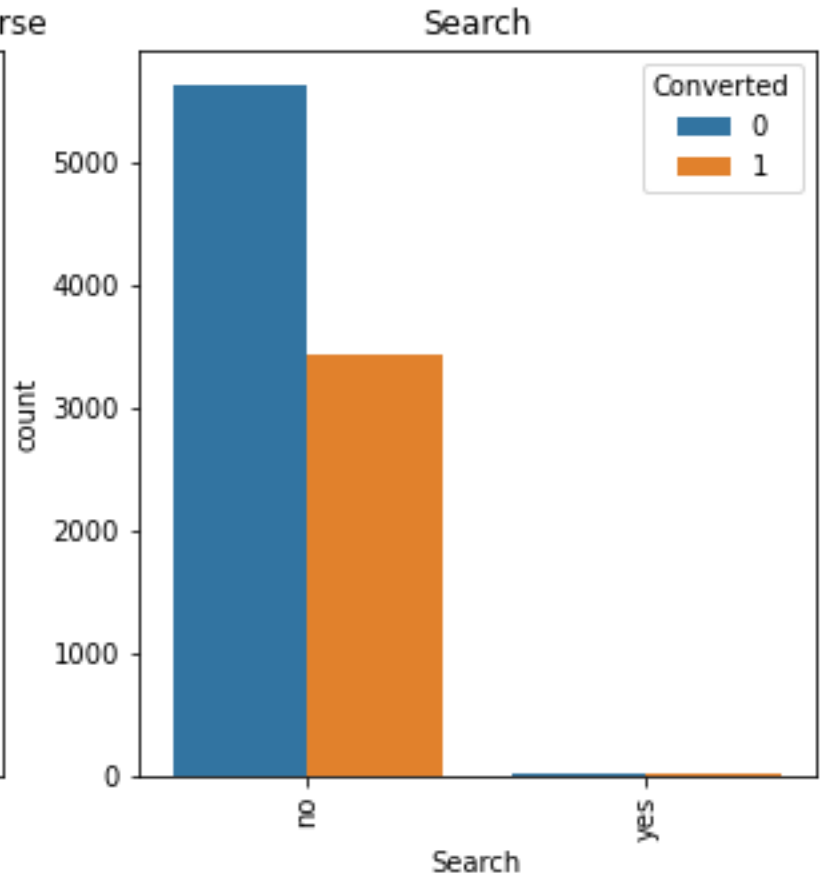
Categorical Variable Relation



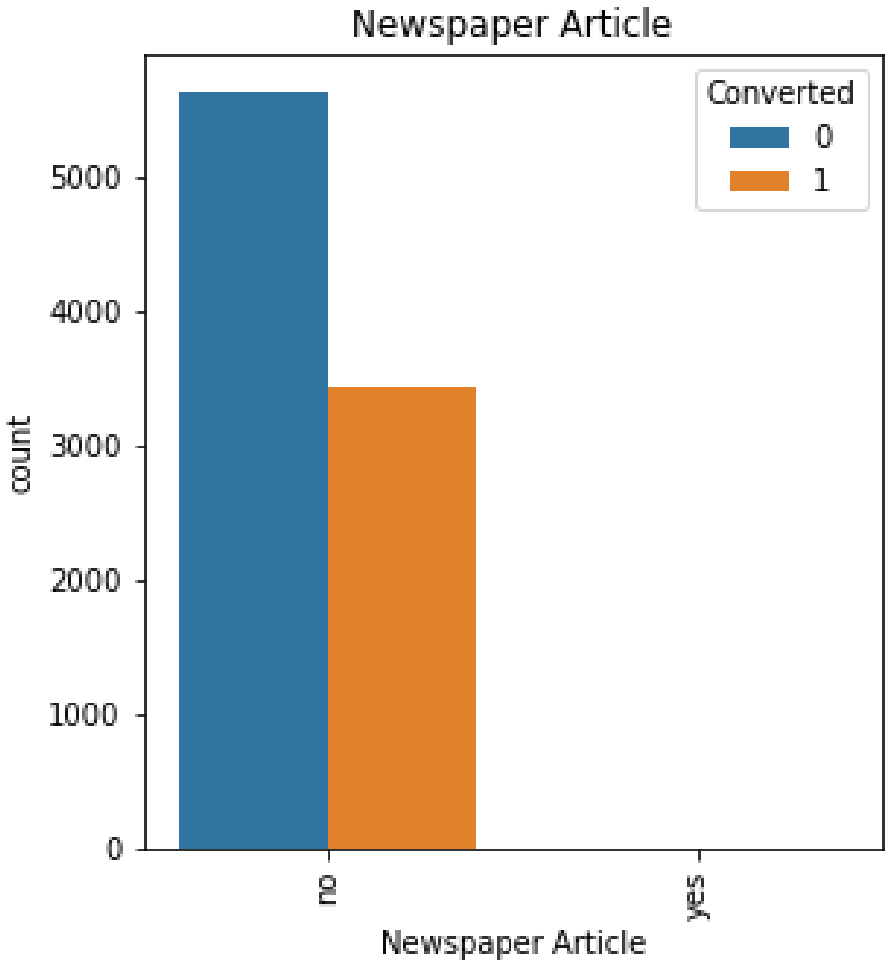


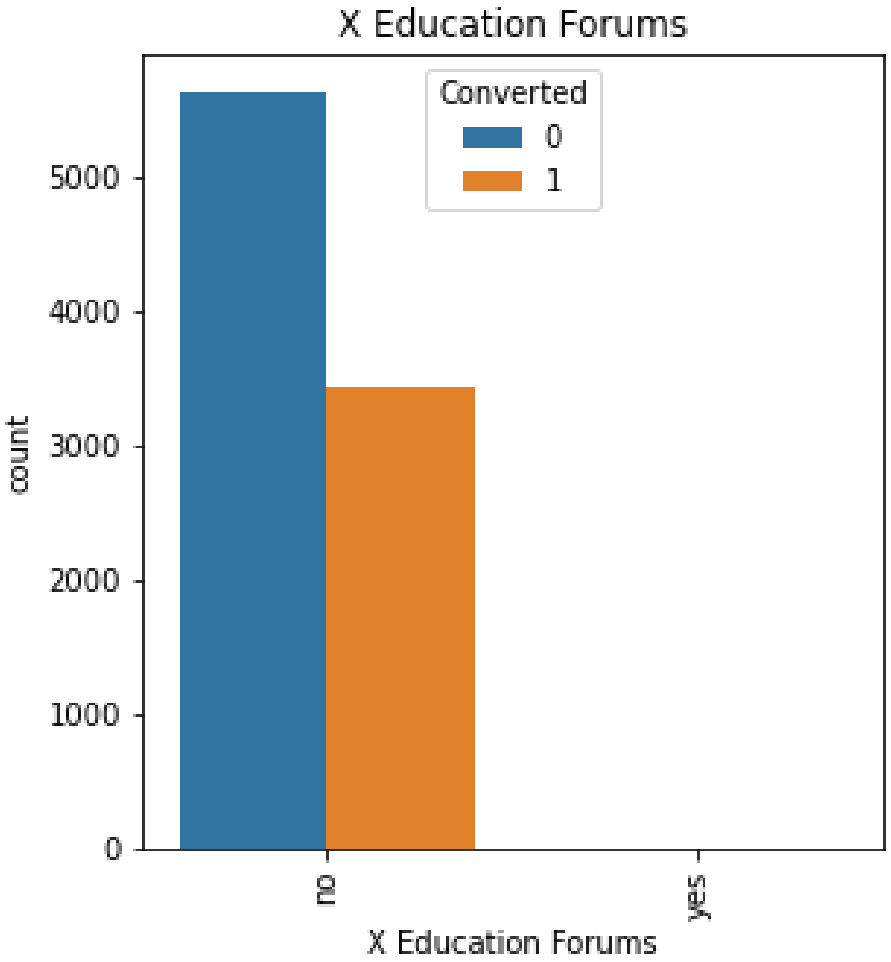
Categorical Variable Relation



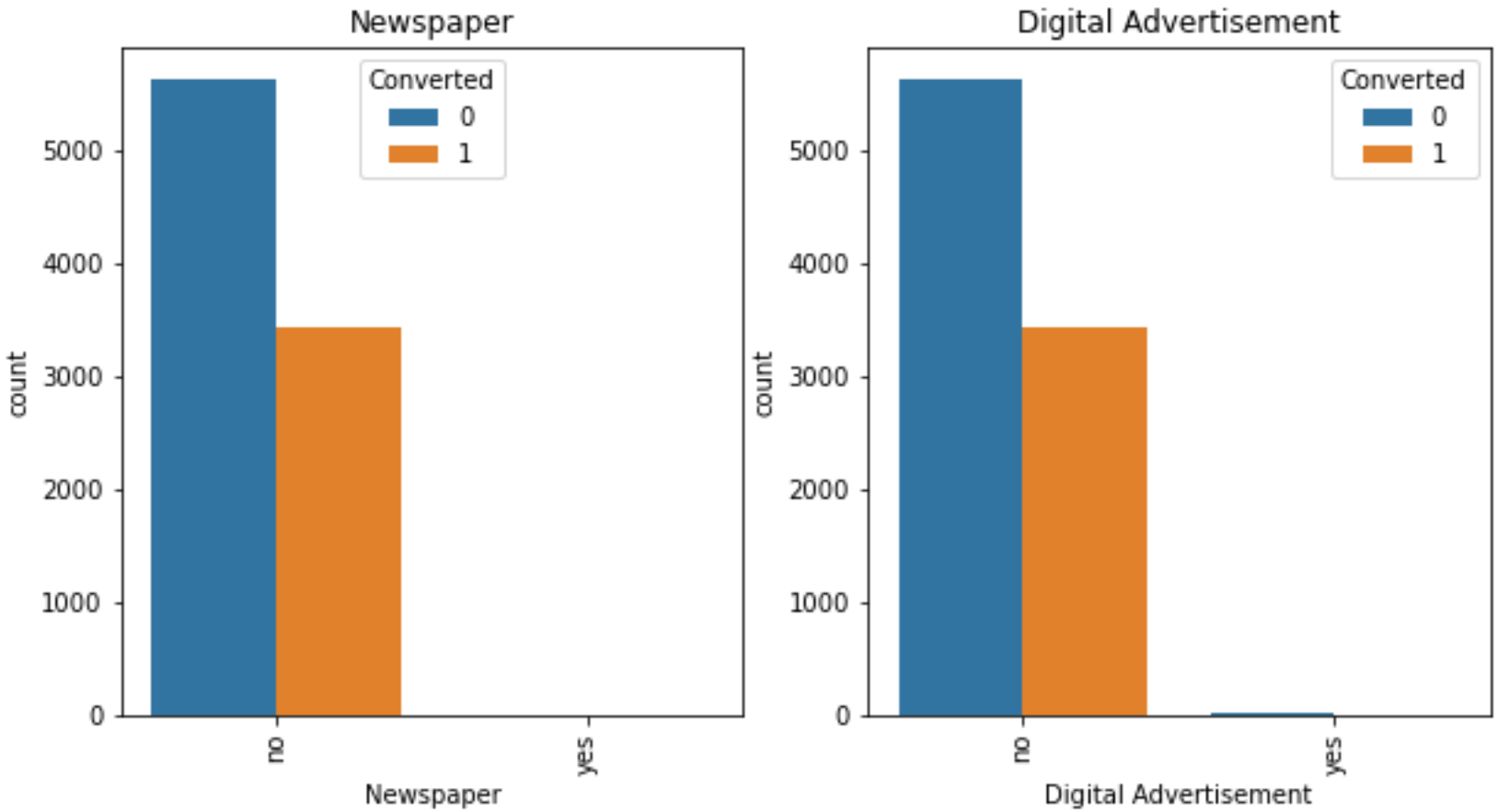


Categorical Variable Relation

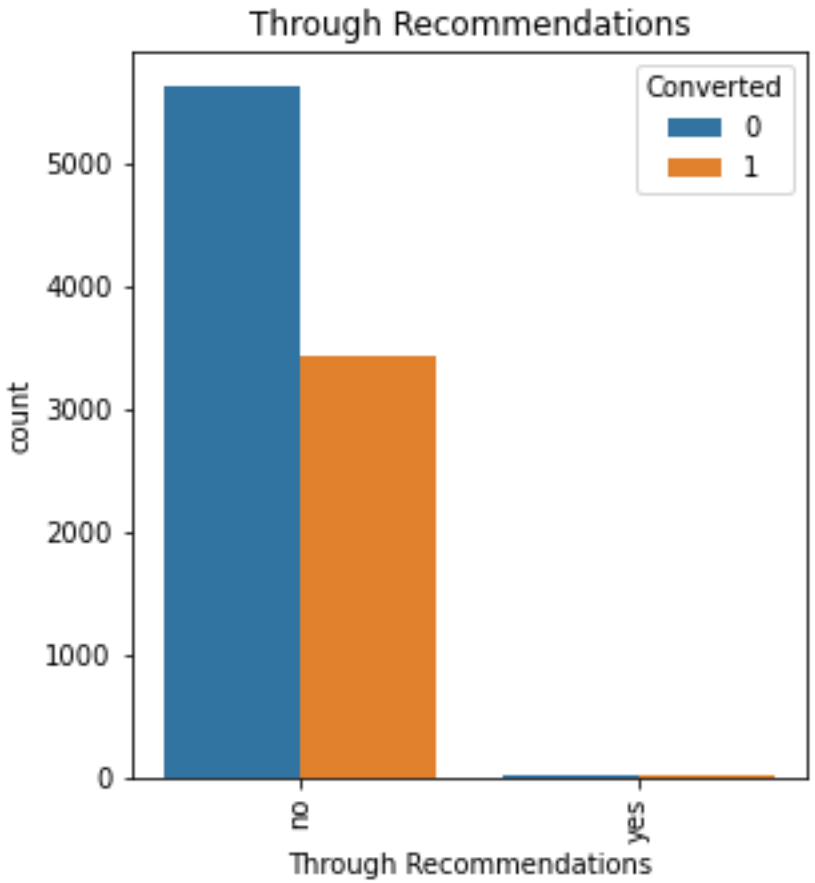


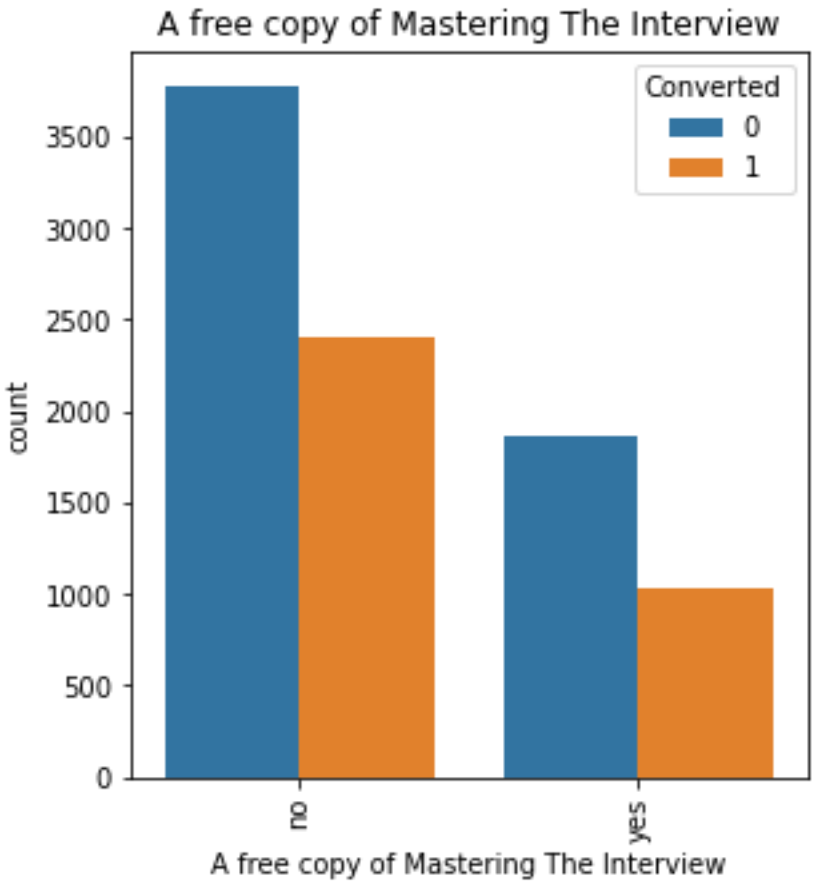


Categorical Variable Relation

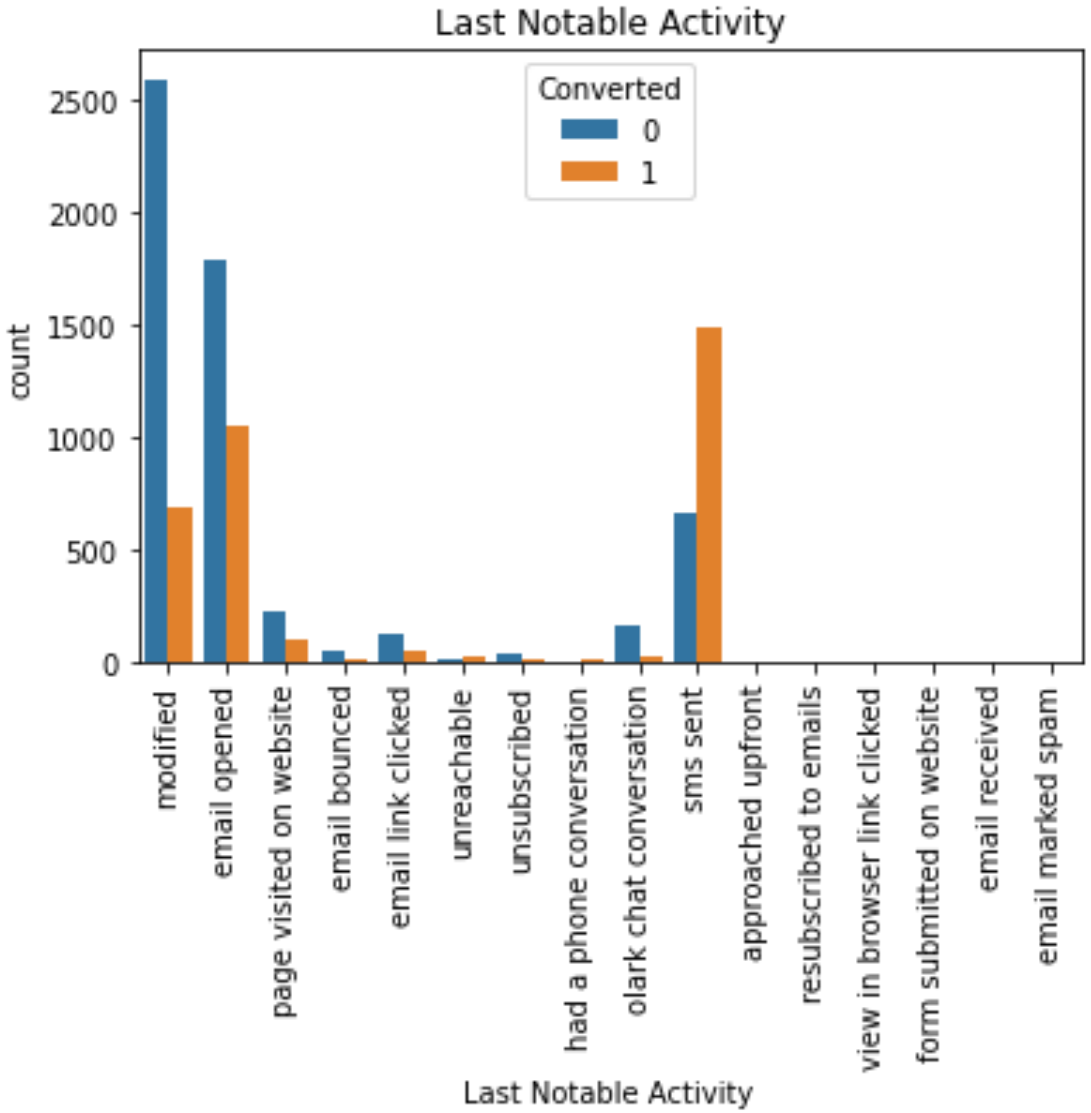


Categorical Variable Relation

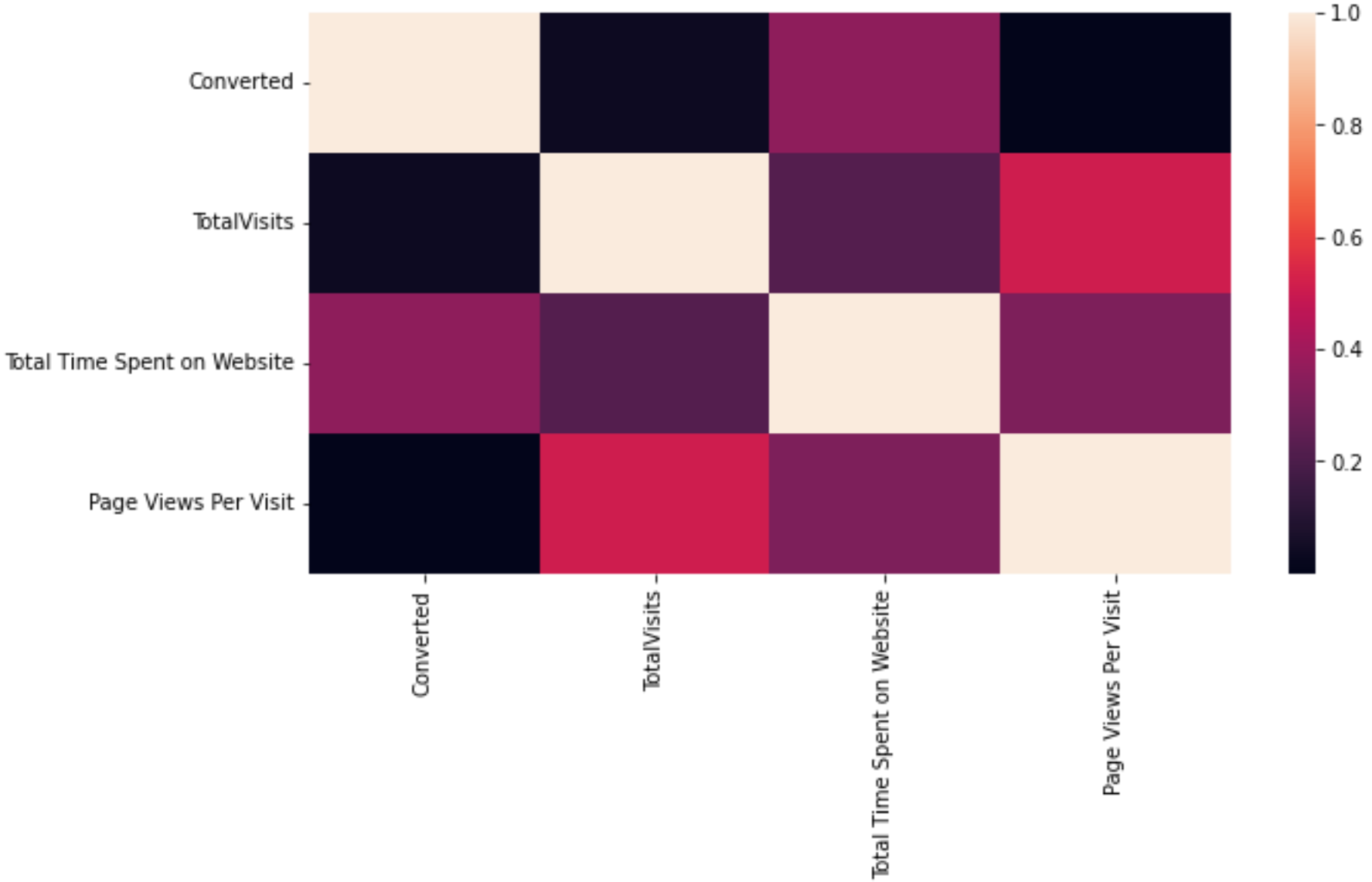




Categorical Variable Relation



Categorical Variable Relation

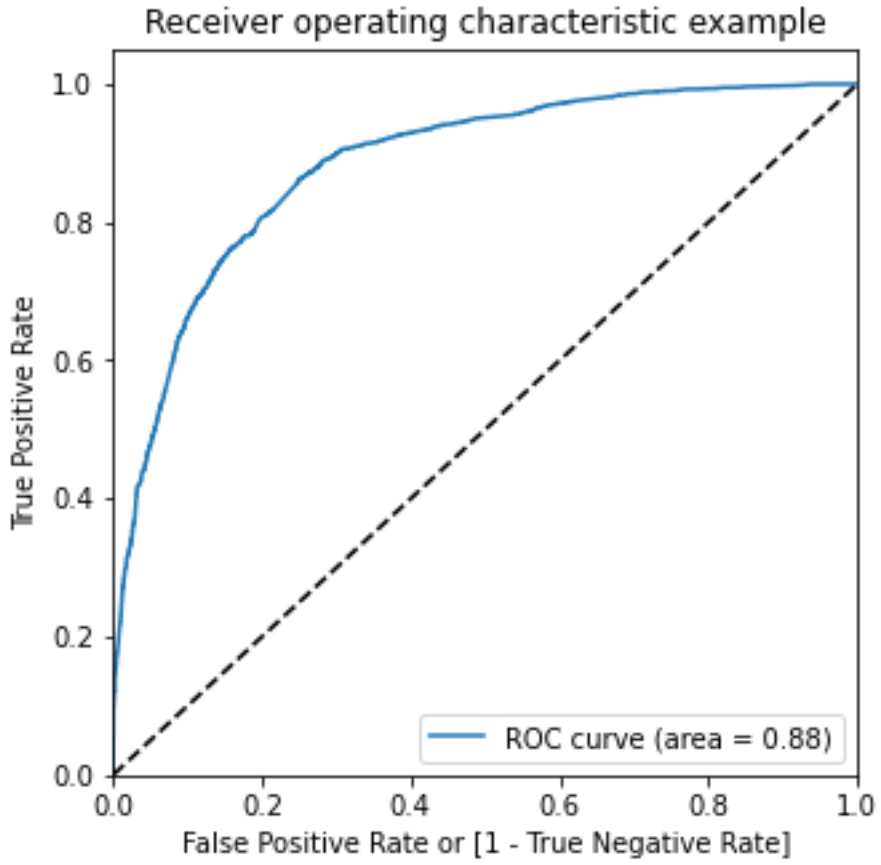


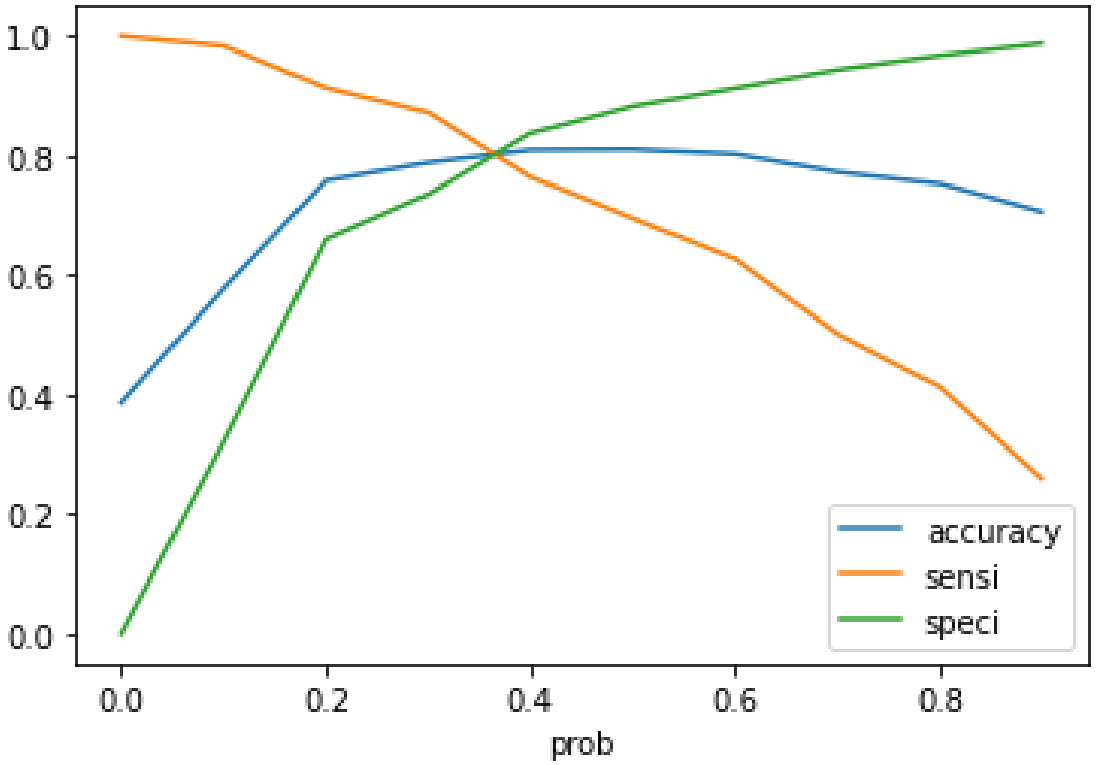
We can analysis that there are many elements that have very little data and so will be of less relevance

Model Building

- Spitted the data into training and testing sets.
- Basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Used REF for feature selection.
- Validating REF wit 15 variable as output .
- Building model by dropping the variable whose p-value is greater than 0.05 and VIF value greater than 5
- Prediction on test data set
- Overall accuracy 81%

ROC Curve





- Finding Optimal cut off point

1. Optimal cut off probability is that probability where we get balance sensitivity and specificity.

2. In Second graph it's visible that the optimal cut off id at 0.35

Conclusion

- It was found that the variables that mattered the most in the potential enroller are (In descending order) :

    1. The total time spend on the Website.
    2. Total number of visits.
    3. When the lead source was:
        a. Google
        b. Direct traffic
        c. Organic search
        d. Welingak website
    4. When the last activity was:
        a. SMS
        b. Olark chat conversation
    5. When the lead origin is Lead add format.
    6. When their current occupation is as a working professional.

Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential enroller to change their mind and enroll their courses.

Thank You