Lead Scoring Case Study

Group Members -
Anusheya M
Arti Kumari
Arindam Chaki

Problem Statement

An education company named 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. 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. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Business Objective -

- Identify the most potential leads, known as ‘Hot Leads’
- Build a machine learning model that identifies the hot and promising leads
- Deployment of the model for future usage

Solution Approach

Data cleaning & transformation -

Read the data
Data sanity checks
Missing value treatment - numerical & categorical
Converting Binary variables from Y/N to 1/0
Treating categorical variables - grouping columns to reduce number of variables, dummy encoding
Univariate & Bivariate Analysis
Outlier check and removal
Class Imbalance check
Correlation check and removal of highly correlated columns

Model Building -

Classification technique - Logistic Regression used for model building and prediction
Feature selection using RFE
Validation of the model

Interpretation and Recommendations -

Model Interpretation
Conclusion & Recommendations

Data Manipulation

Number of rows : 37, Number of columns : 9240.

Replacing ‘select’ with null value as it represents no data variable.

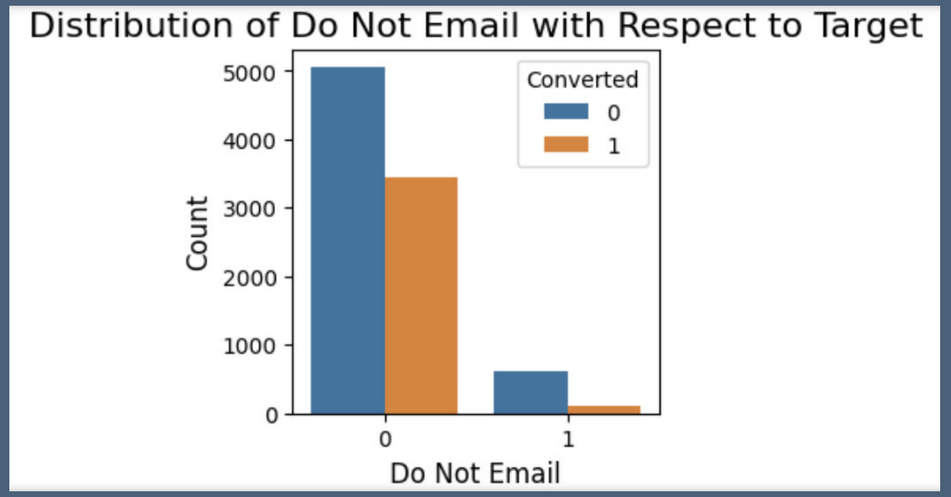
Removing the columns that has missing values more than 30% and replacing/filling missing values with their mode, if categorical variable and mean ,if numerical variable.

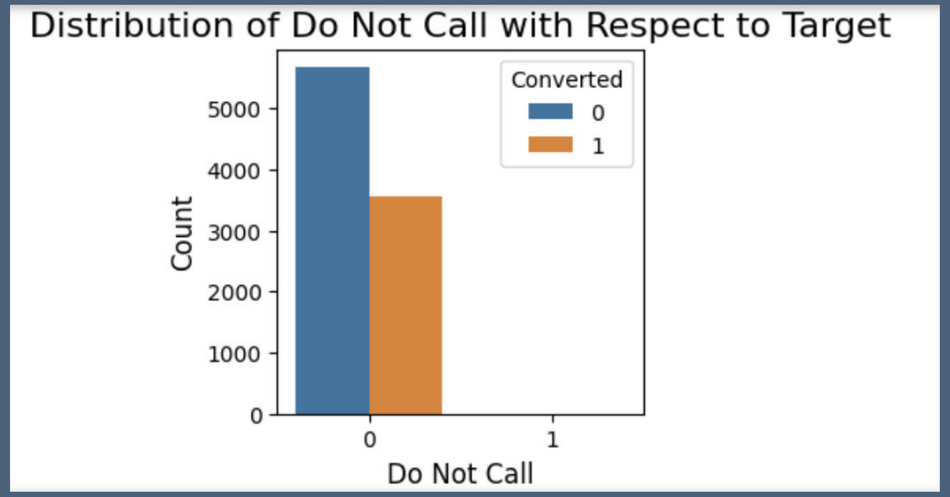
Converting binary variables like ‘Do not Email’, ‘Do Not Call’, etc from Y/N to 1/0.

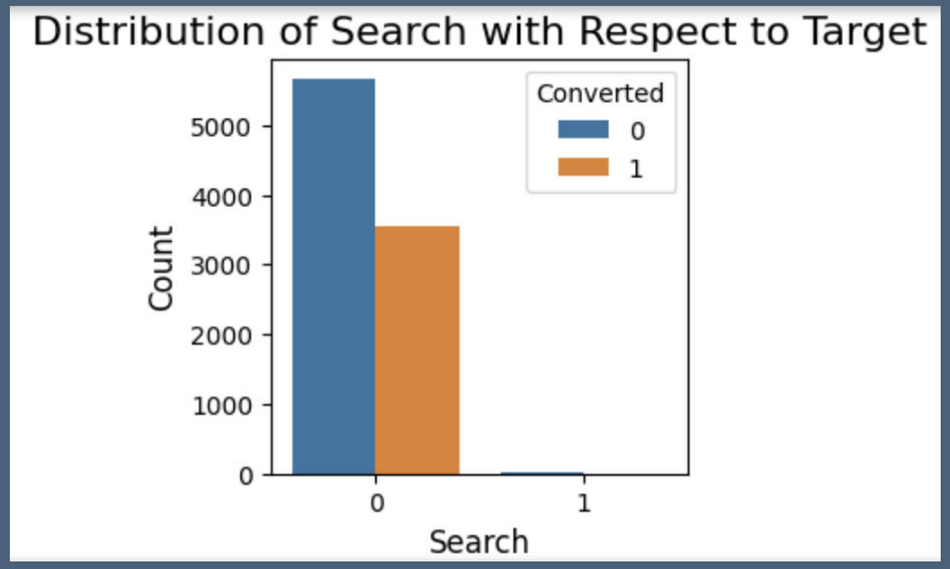
Grouping categorical variables to reduce the number of variables - creating new columns like ‘Continent’, ‘Last notable activity’, etc.

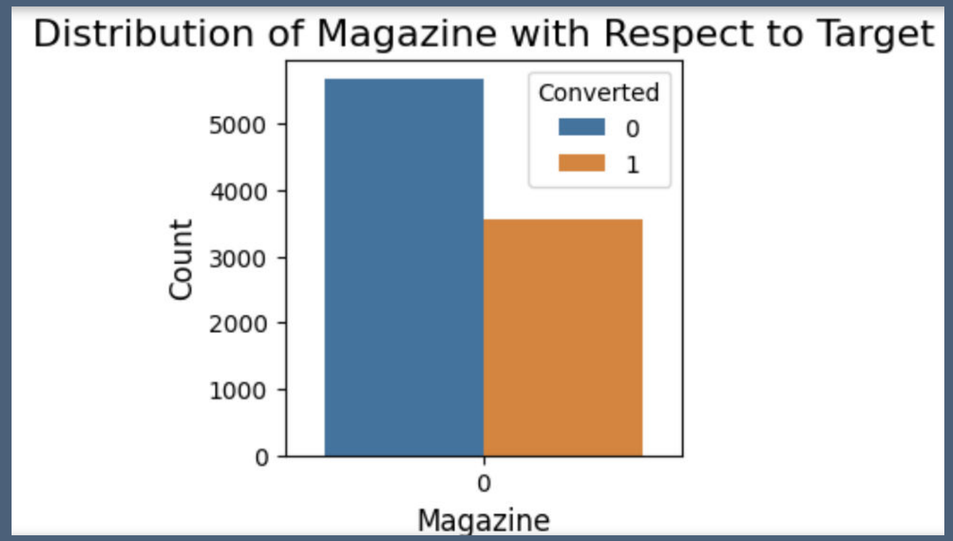
Creating a dummy variable for some of the categorical variables(‘Lead Origin’, ’Lead Source’, etc) and dropping the first occurrence of the column.

EDA - Distribution of feature w.r.t. Target

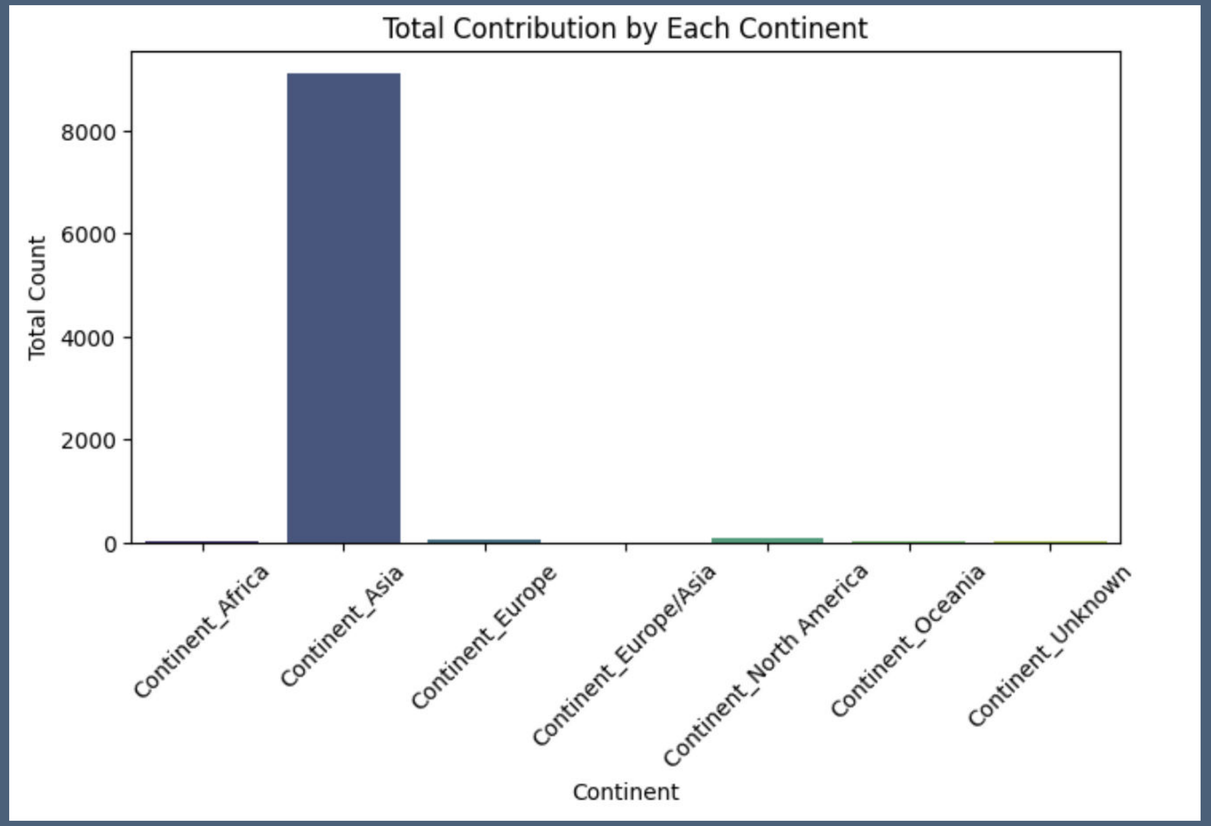


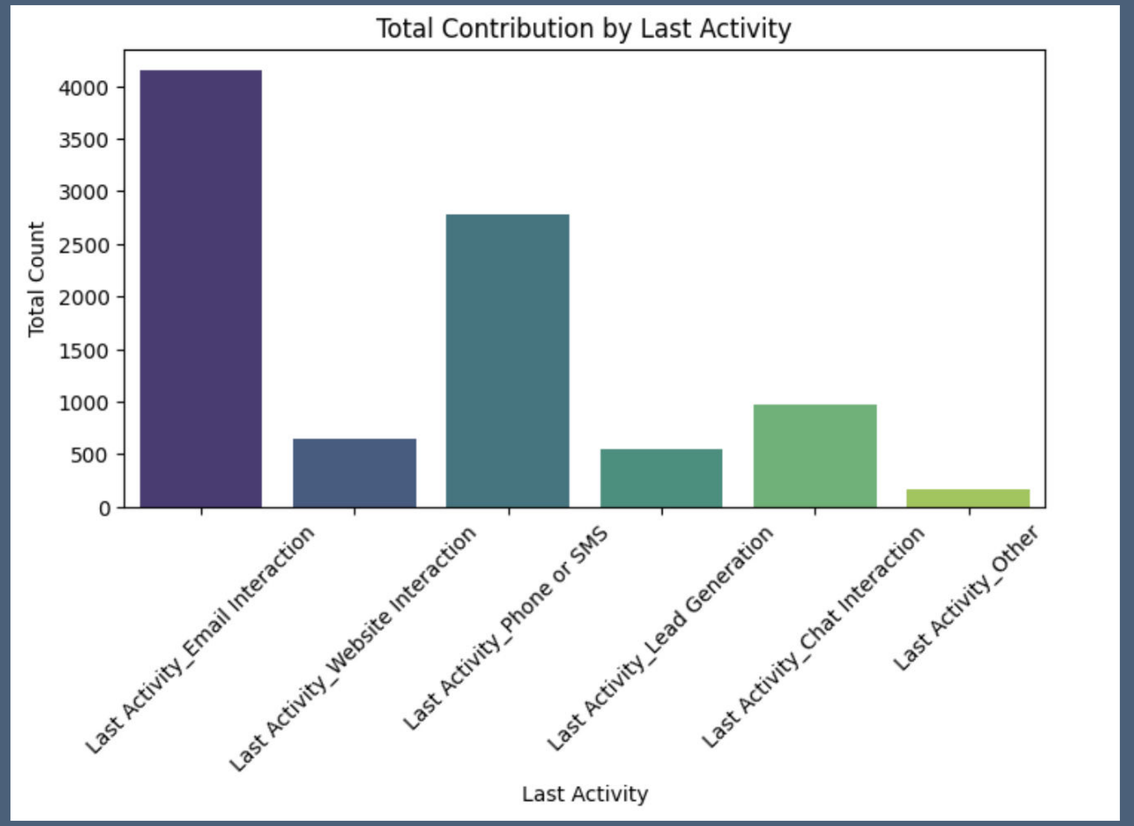




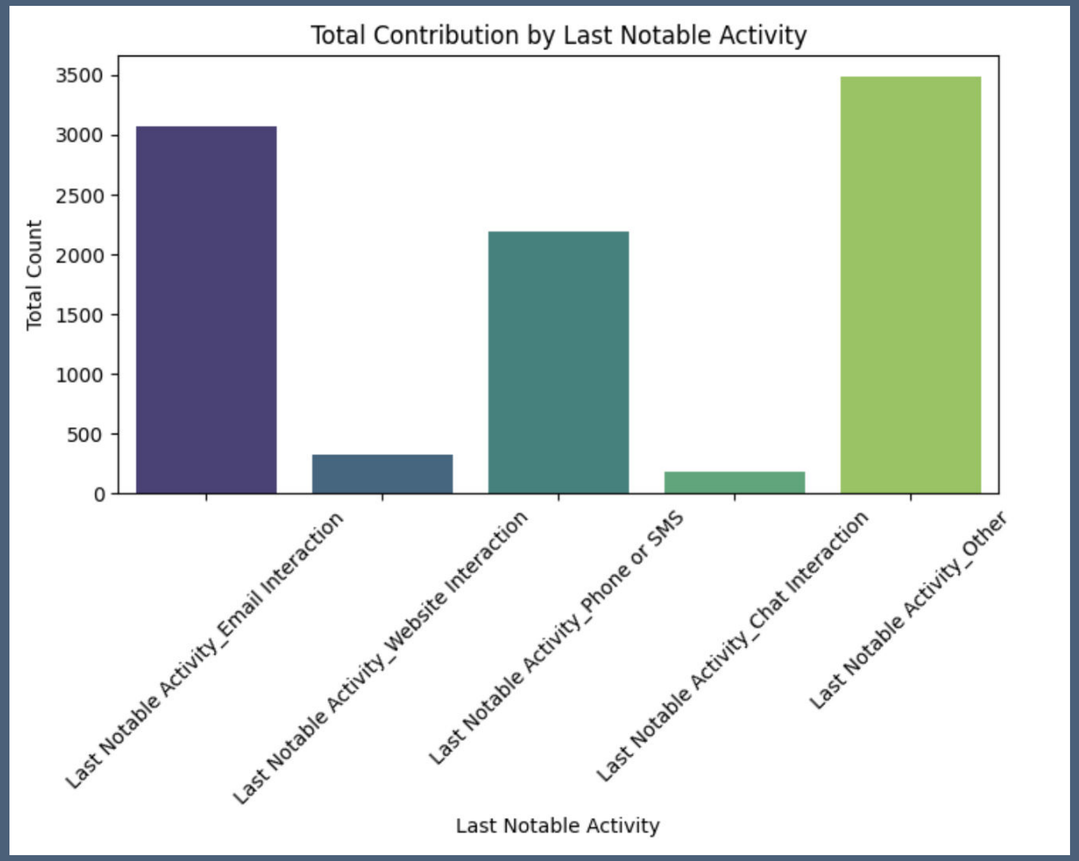


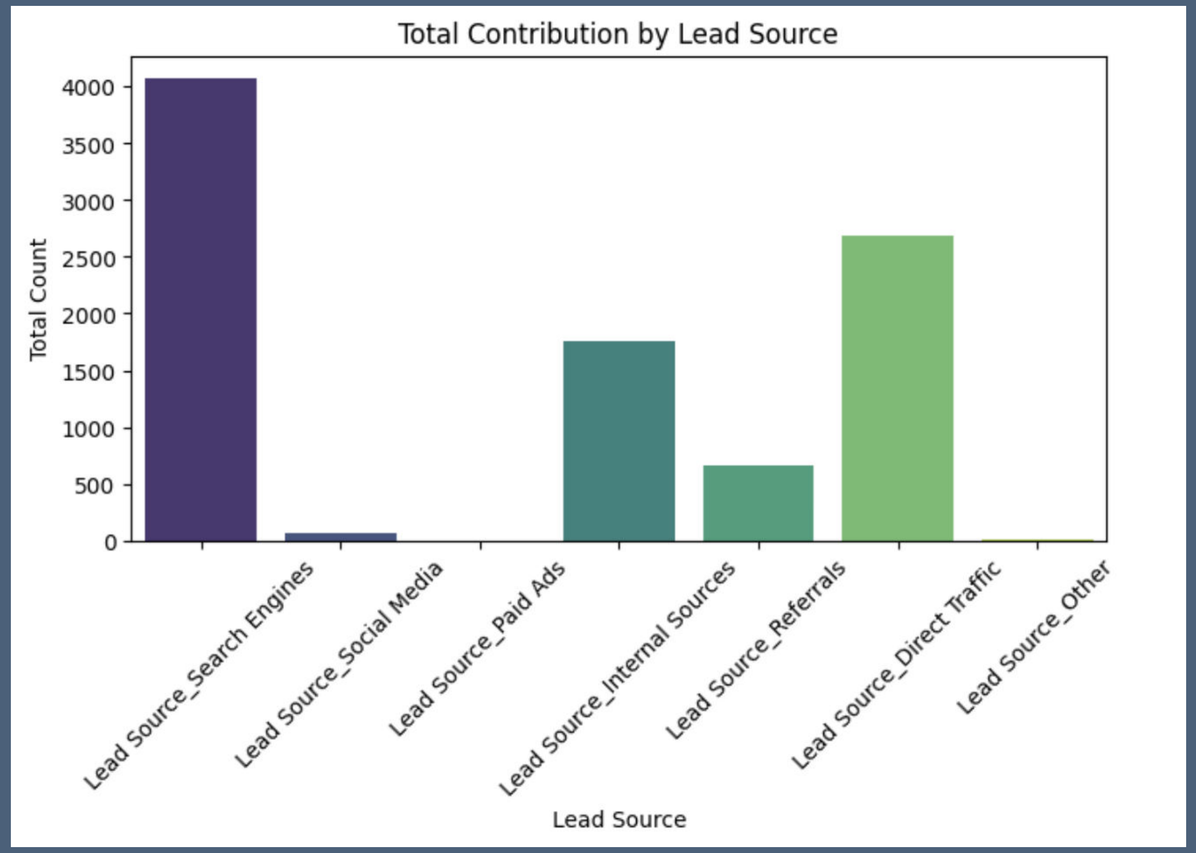
EDA - grouped/categorical variable relation



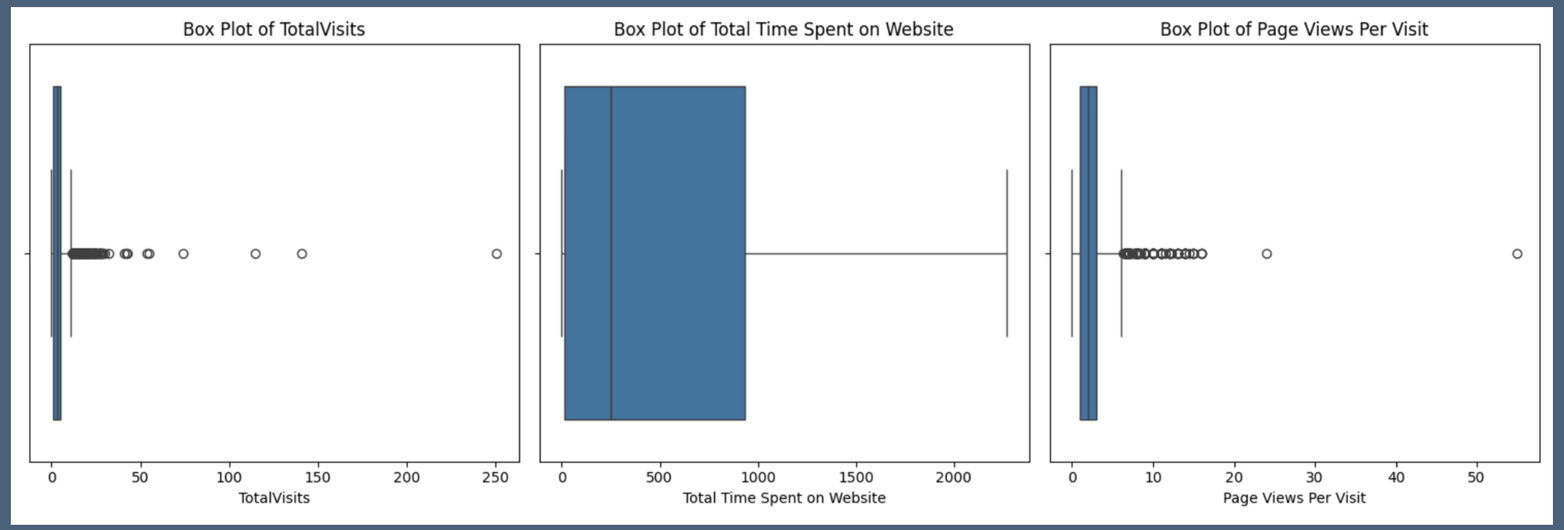


EDA - grouped/categorical variable relation



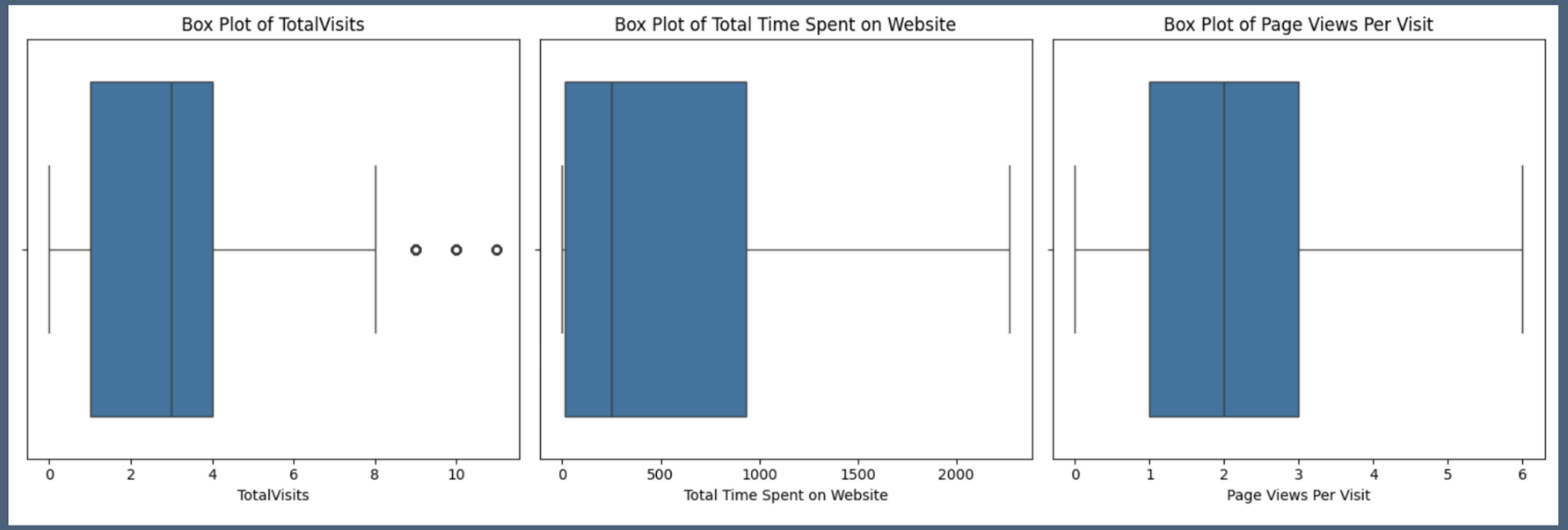


Outlier Check & Removal

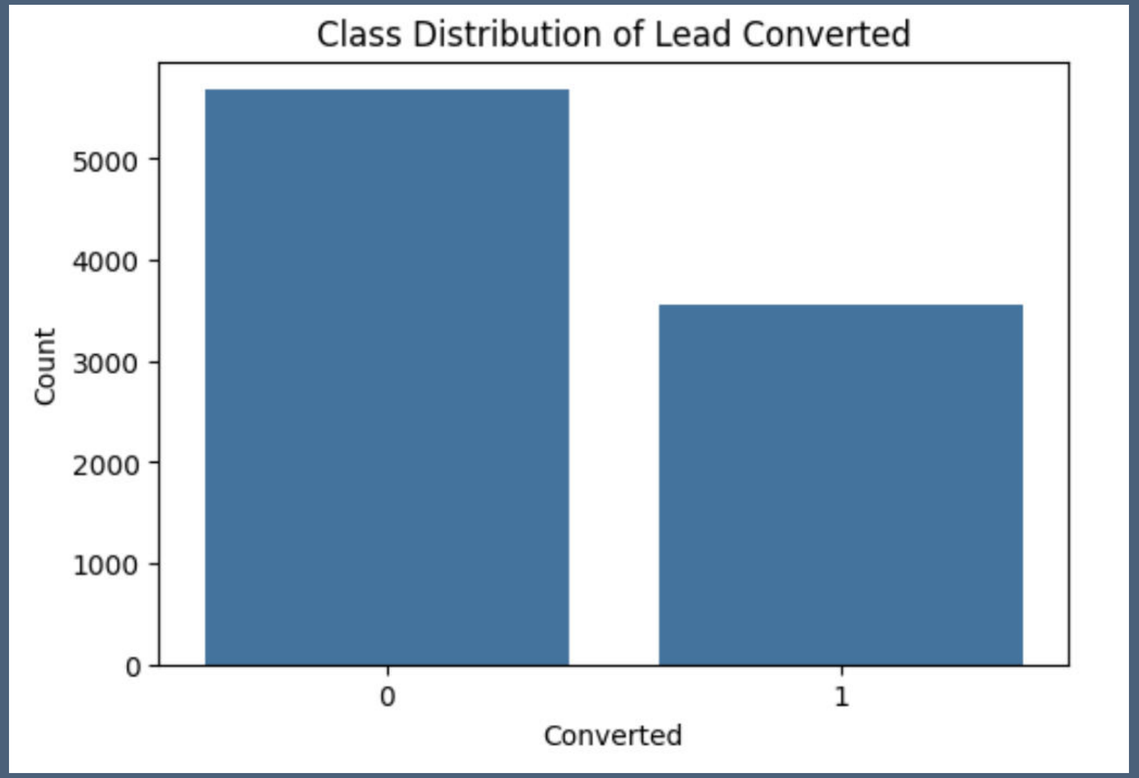


<- Before

After ->

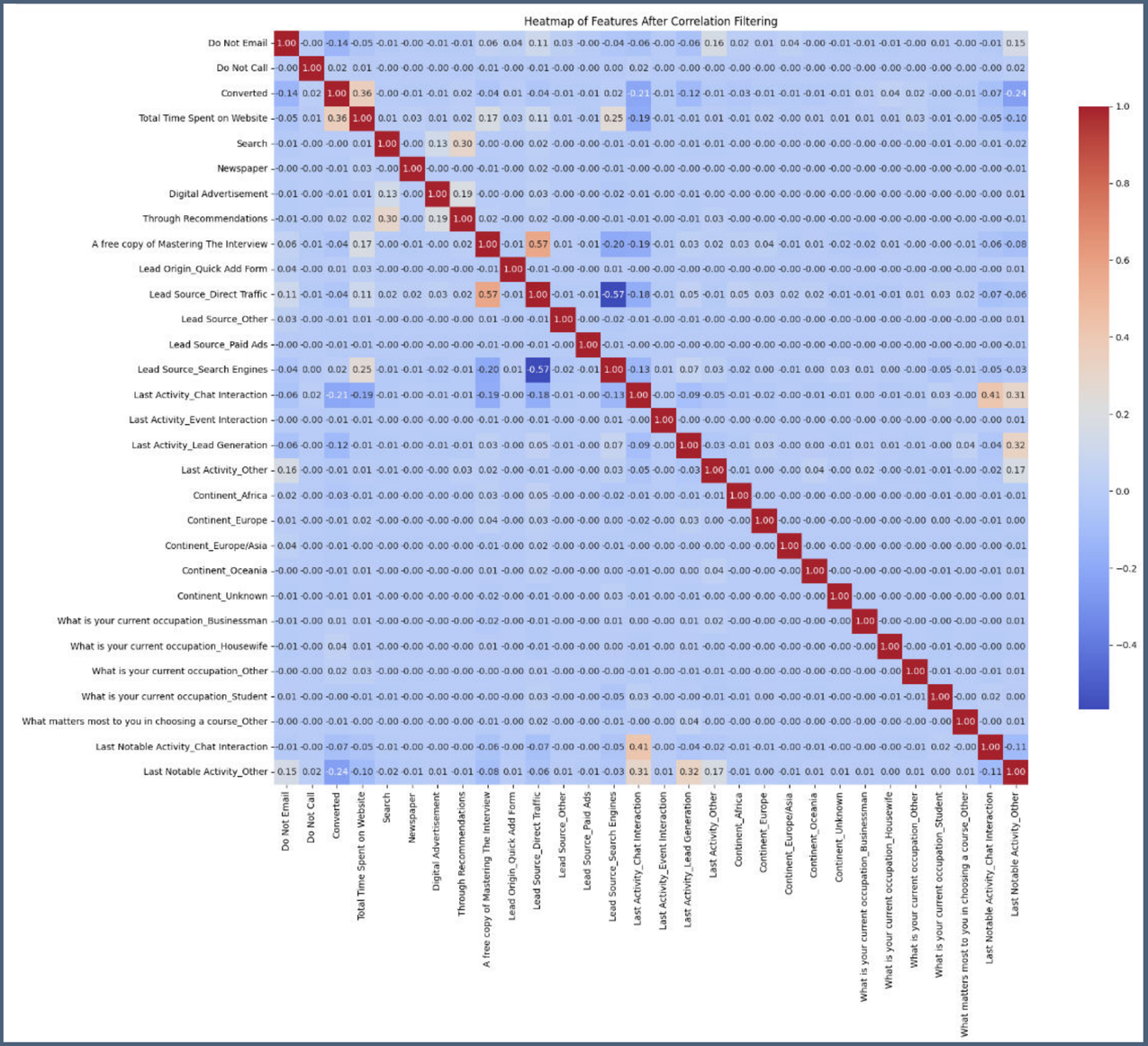


Check for Class Imbalance



Imbalance Ratio: 1.59
This means for every 1 lead that converts,
there are approximately 1.59 leads that do not convert. It's not an extreme imbalance.

Correlation Matrix



Dropped features due to high correlation: ['X Education Forums', 'Lead Source_Social Media', 'Last Notable Activity_Phone or SMS', 'Last Activity_Phone or SMS', 'Lead Origin_Lead Import', 'TotalVisits', 'Last Activity_Website Interaction', 'Lead Origin_API', 'Lead Source_Referrals', 'What is your current occupation_Unemployed', 'Continent_North America', 'Page Views Per Visit', 'What is your current occupation_Working Professional', 'What matters most to you in choosing a course_Better Career Prospects', 'Last Notable Activity_Website Interaction', 'Lead Origin_Landing Page Submission', 'Continent_Asia', 'What matters most to you in choosing a course_Flexibility & Convenience', 'Last Notable Activity_Email Interaction', 'Newspaper Article', 'Lead Source_Internal Sources', 'Last Activity_Email Interaction', 'Lead Origin_Lead Add Form']

Model Building

Splitting Data to train and test sets (70:30 ratio).

Standardize numerical columns.

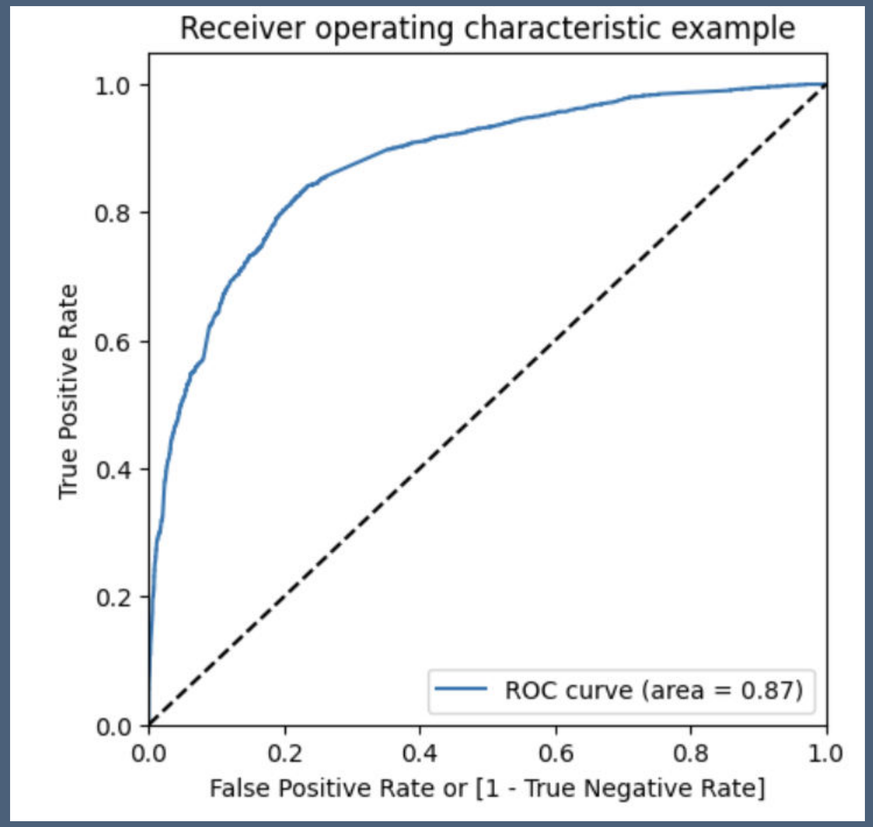
Feature selection using RFE with top 15 variables.

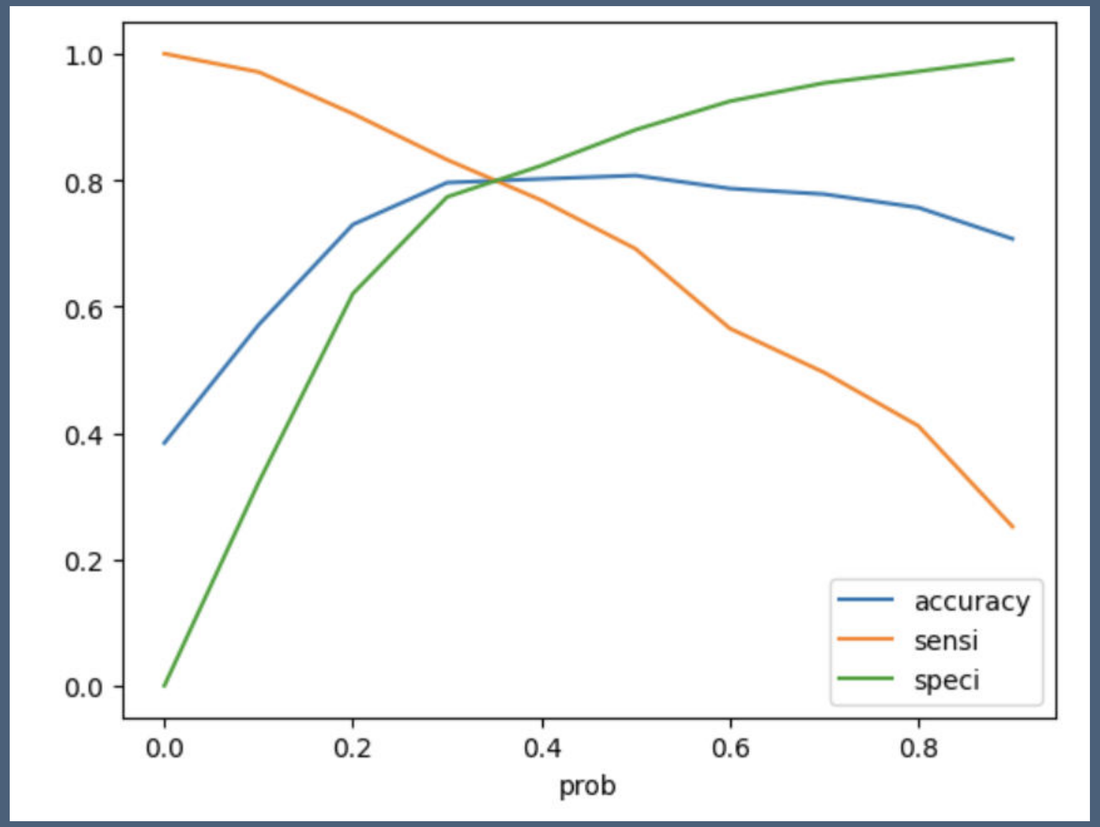
Assess the model using stats model and build the model by removing variables with p-value greater than 0.05 and VIF greater than 5.

Prediction on test data set.

Overall accuracy of the model : 81%

ROC Curve





The area under the curve of the ROC is 0.87 which is quite good. So we seem to have a good model.
Also, while checking the sensitivity specificity trade off, we get 0.375 as the optimal cut off point.

Prediction on test set

Train set Prediction-
Accuracy - 0.8007
Sensitivity - 0.8092
Specificity - 0.7954

Vs

Test set Prediction-
Accuracy - 0.8153
Sensitivity - 0.8459
Specificity - 0.7959

Conclusion & Recommendations

It was found that the features that mattered the most in the potential buyers are -

- Lead Origin
- Current Occupation
- Last Notable Activity
- When the lead source was Internal source
- When the last activity is chat interaction
- When the lead origin is API
- When their current occupation is as a working professional.

Keeping the above in mind, X Education can flourish/expand as they have a very high chance to get almost all their potential buyers/leads to change their mind and buy their courses.