Lead Scoring Case study

Insights, Inferences and conclusion

Presenters - Kaushik , Bhavani and Deepa

Problem Statement

- This problem statement is about lead conversion on a educational institution whose potential lead in order to get high lead conversion.

- This case study is for building a model wherein we need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.

- The Company CEO wants a ballpark of the target lead conversion rate to be around 80%.

Analysis approach Highlights

- Importing necessary libraries
- Importing and Inspecting the Dataset
- Data Cleaning : Handling anomalies and outliers in the data
- Exploratory Data Analysis using univariate, bivariate and multivariate analysis
- Splitting the data into Train and Test Data
- Scaling the data
- Building the model
- Evaluating the model
- Making the predictions on the test dataset
- Generating Lead Scores for the sales team for the full dataset

Detailed approach - Data Cleaning

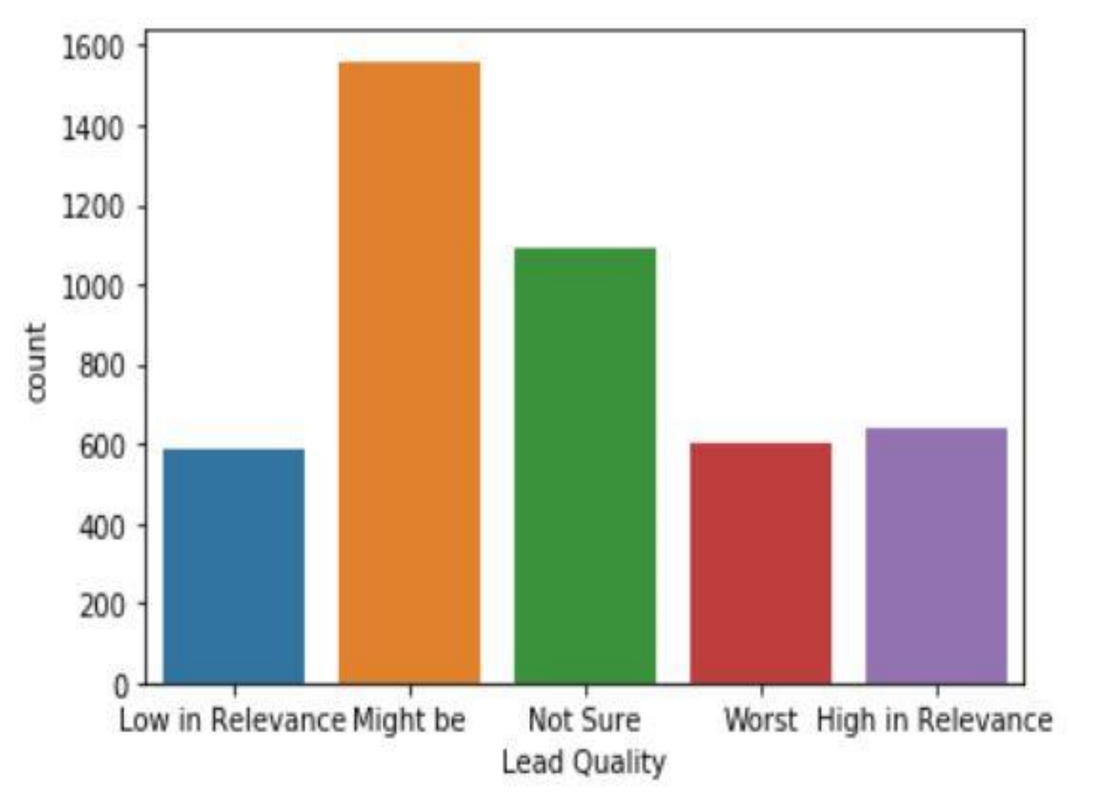
- Importing of Leads.csv data file. Understand the column values referring to Data Dictionary File.
- Analysis and Removal of columns having same data like Prospect Prospect ID and Lead Number .
- Converting “Select” values in all columns to Null.
- Checking the percentage of Null values and dropping the columns having more than 70%.
- Imputing of Null values for Lead Quality, City, Specialization, Career prospectus, Current Occupation, Most frequent, Country column.
- Plot count plots and bar plots for checking data distribution.

Detailed approach - EDA, Preparation, and Model building

- Univariate Analysis on Converted column, Lead Source, Do Not Email, Do Not Call, Total Visits
- Handle Outliers in Total Visits, Total Time spent on Website, Last Activity, column.
- Further dropping and imputing other columns like Country, Search, Magazine etc.
- Creating Dummy variables, Assigning categorical values and encoding of Data.
- Creating models using Logistic regression approach
- Comparison of Actual Lead conversion and Predicted Lead conversion by creating dataframes.
- Evaluation of model using confusion matrix.
- Plotting ROC Curve
- Getting predicted values on the Train Set.
- Feature selection using RFE.
- Model Presentations.
- Assigning Lead Score to test dataset.

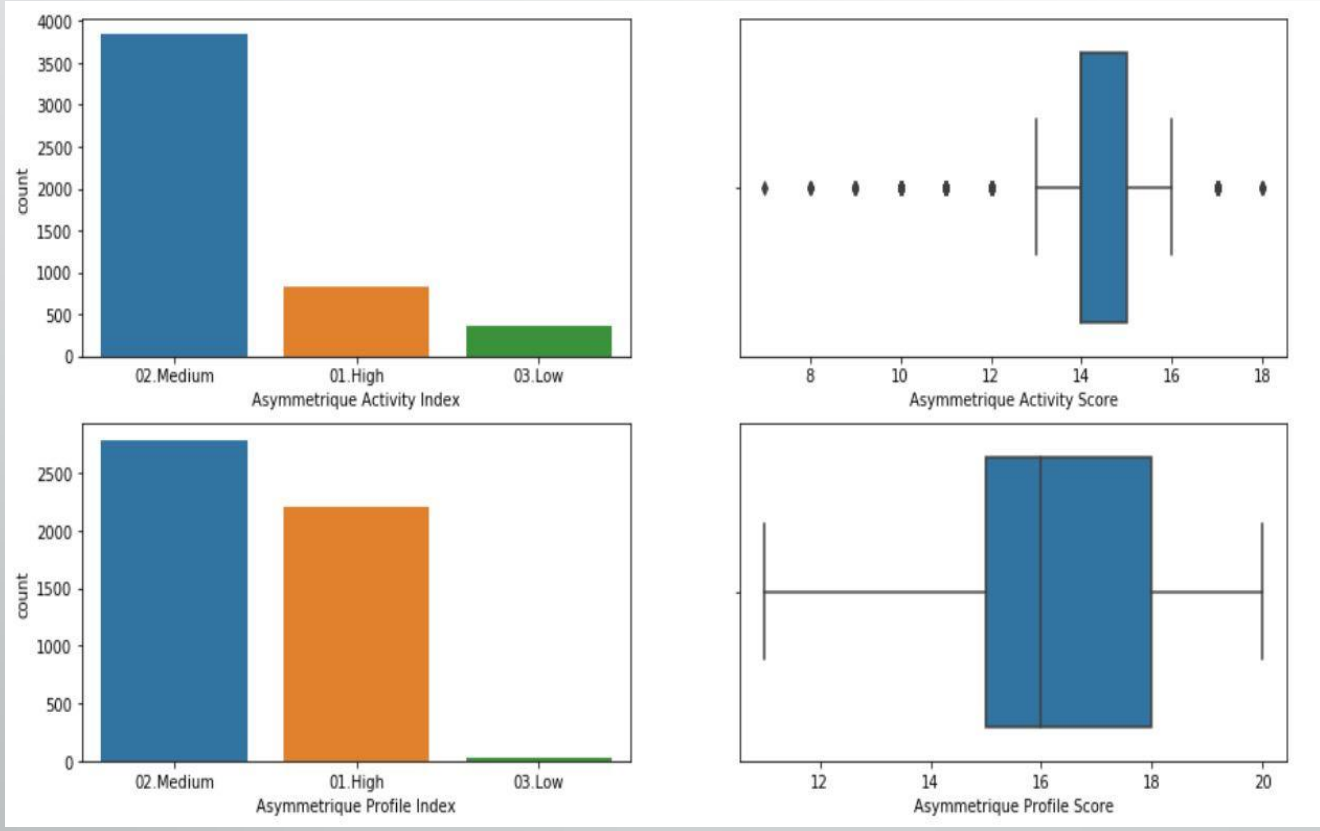
Analysis on some columns (Cleaning)

Lead Quality:



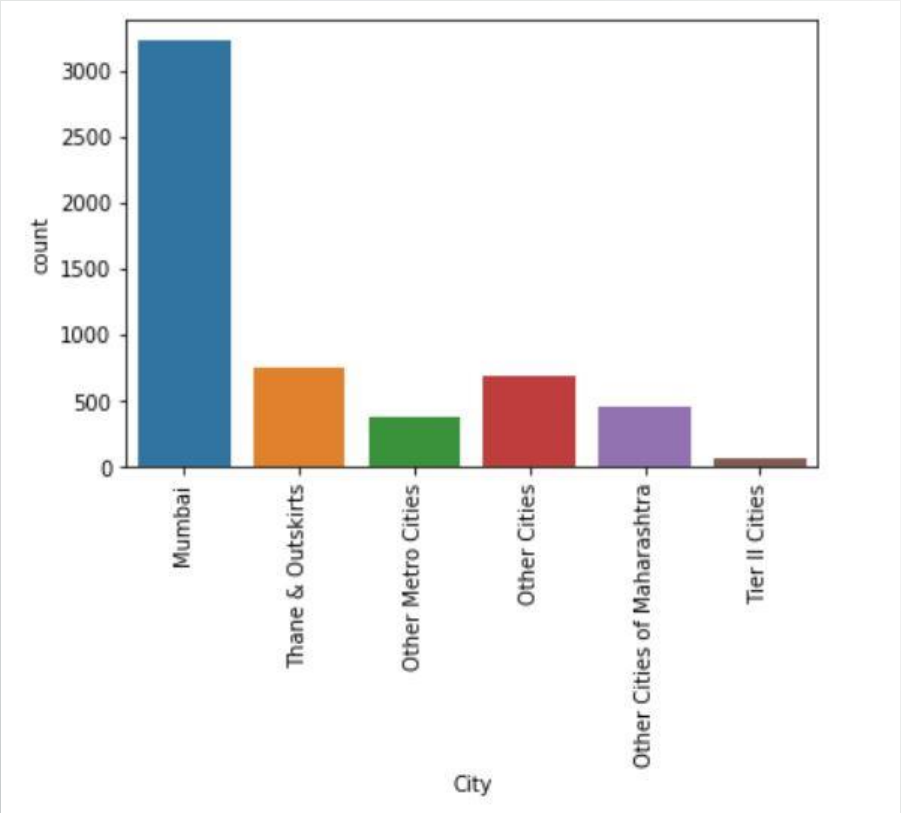
- Dropped columns which contains more than 70% of null values.
- Lead Quality column contains more than 51 % of Null value.
- Converted all the null values with 'Not Sure'.

Columns: Asymmetrique Activity Index, Asymmetrique Profile Index, Asymmetrique Activity Score & Asymmetrique Profile Score

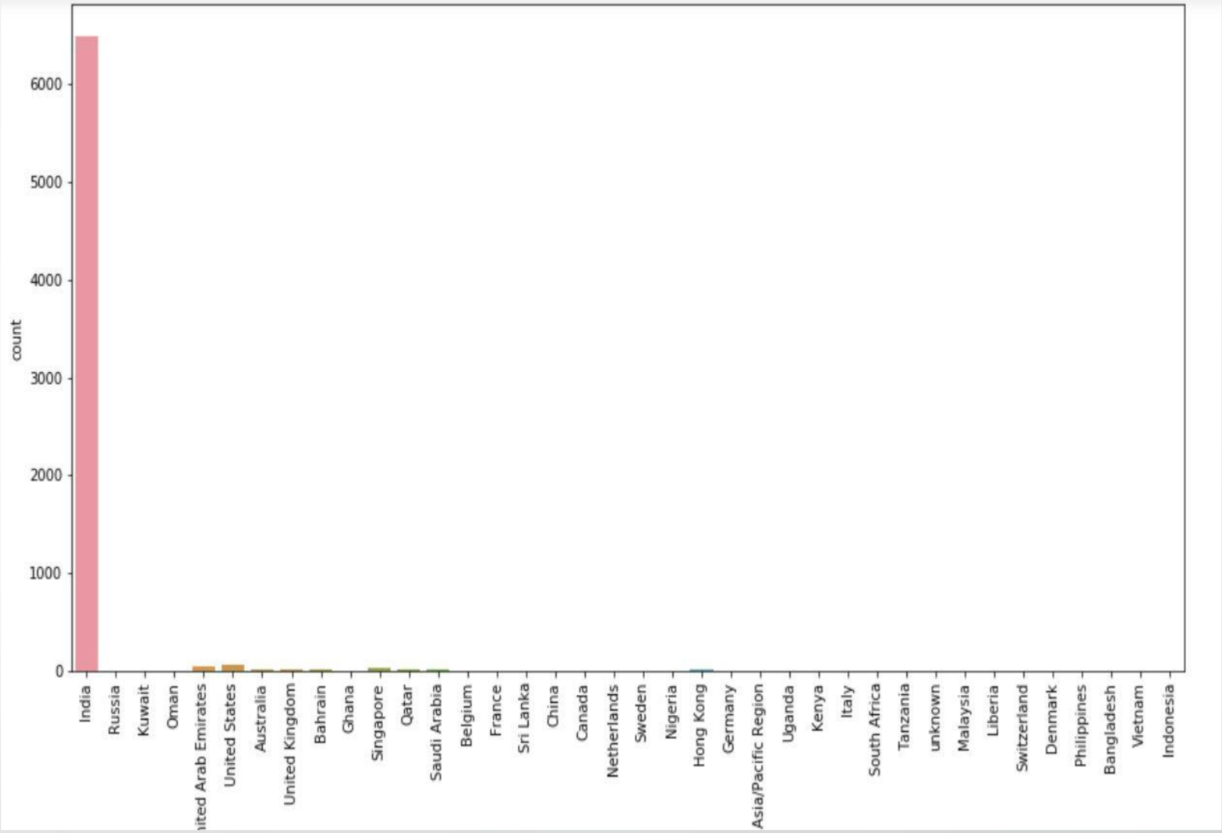


- The following column's attributes are widely speared.
- The columns contain more than 45% of Null Values.
- The columns do not contain significance with respect to the problem statement.
- Decided to drop the columns.

Columns: City & Country

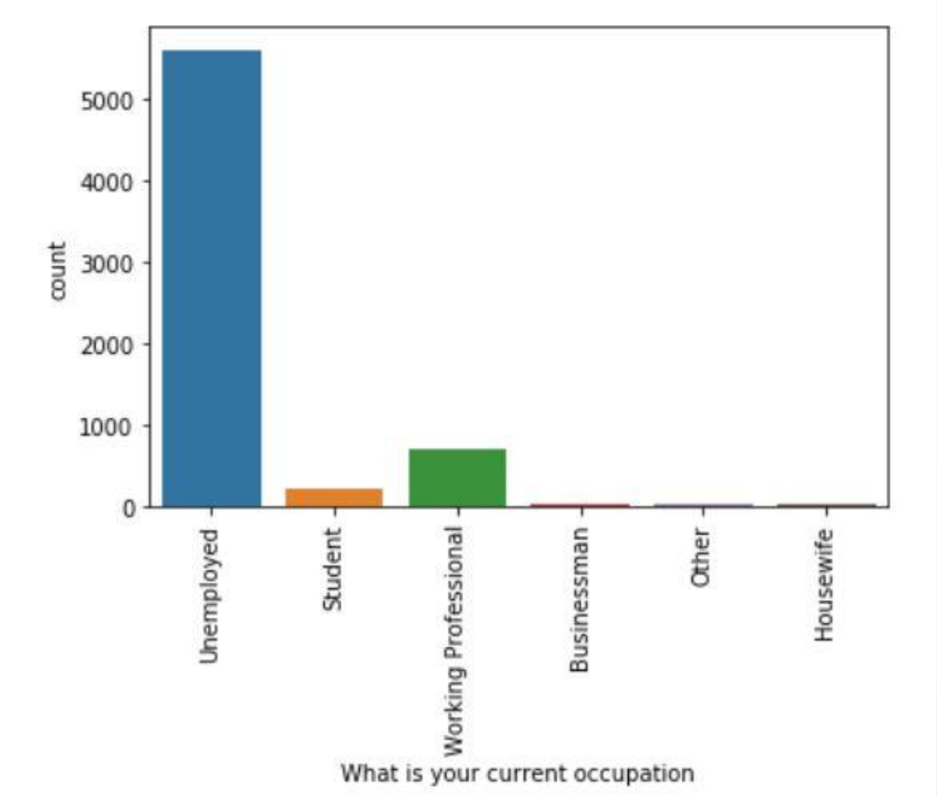


- In City column, 39% null values are present.
- Converted nulls as other city for analysis.

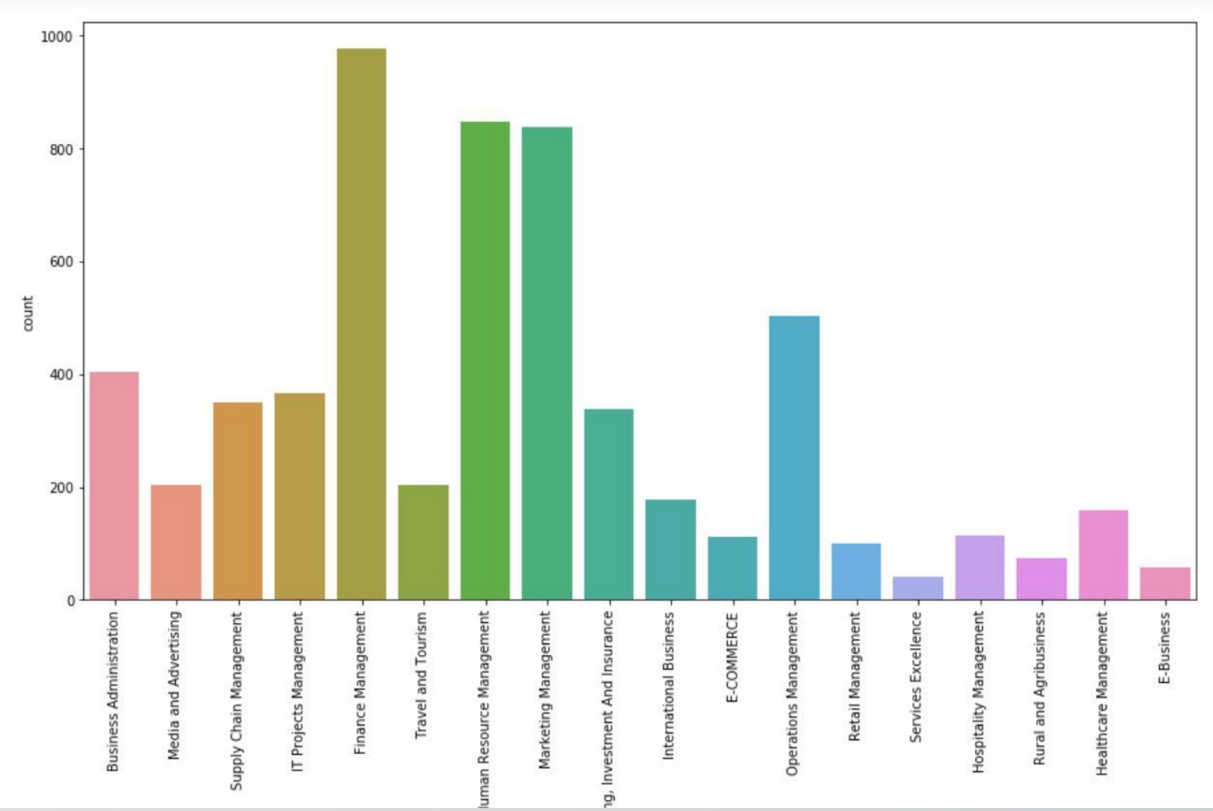


- In Country columns, 26% null values are present.
- India is the highest attribute.
- Imputed null values as most frequent values.

Columns: Current Occupation and Specialization

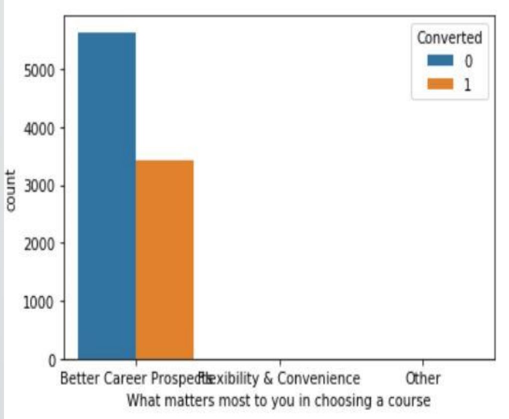


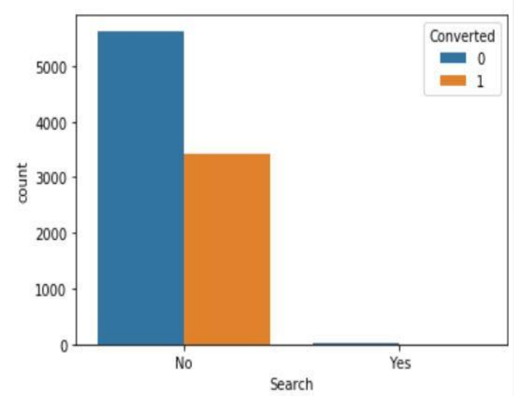
- Current Occupation contains about 29% of null values.
- Imputed null values with most frequent values

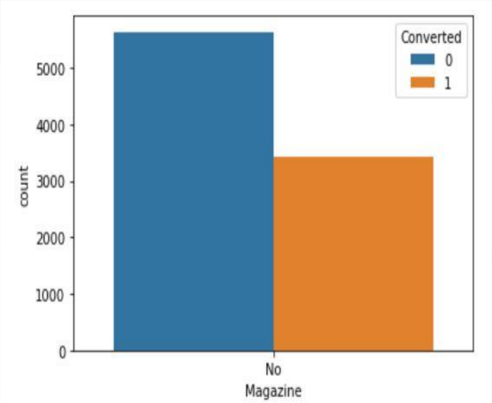


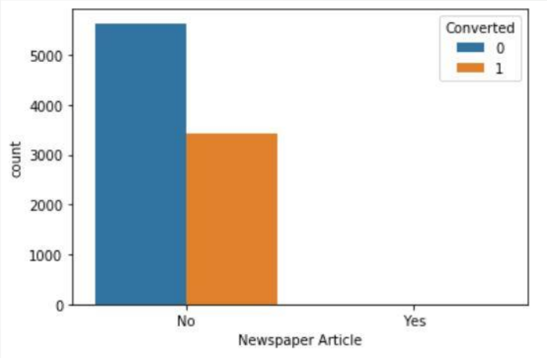
- Specialization contains more than 36 % of null values.
- Imputes them with Other Specialization

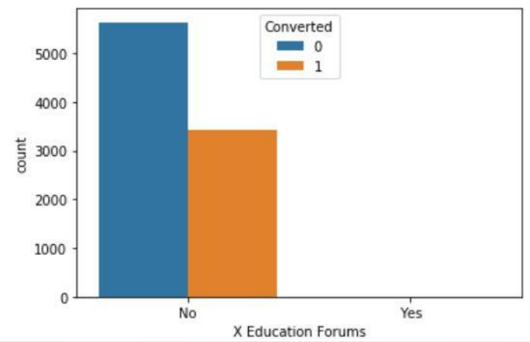
Columns with high single sided responses are dropped as inferences can not be drawn :

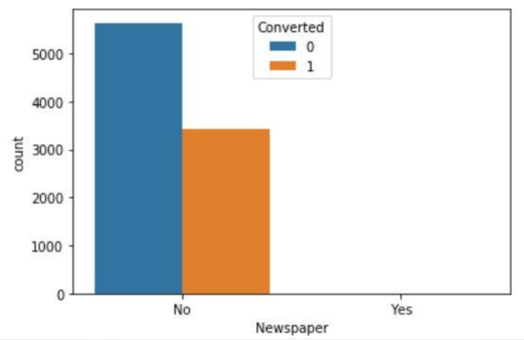


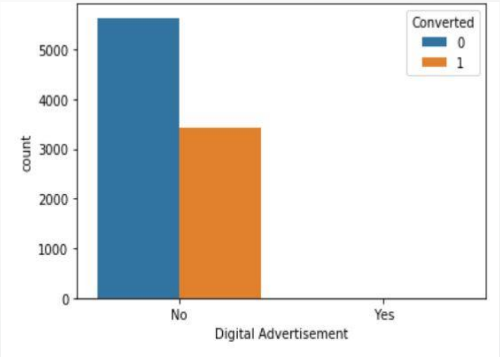


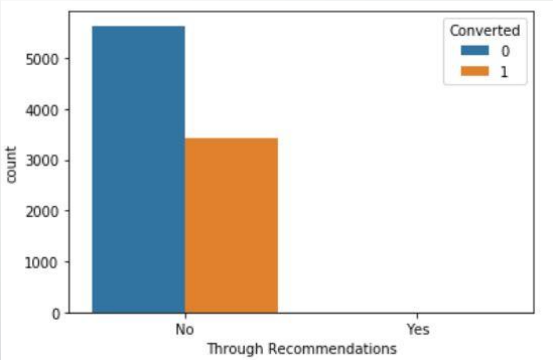


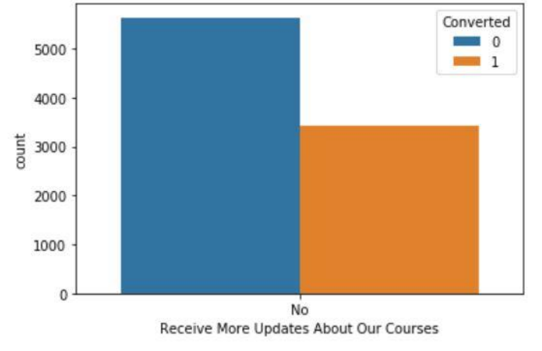


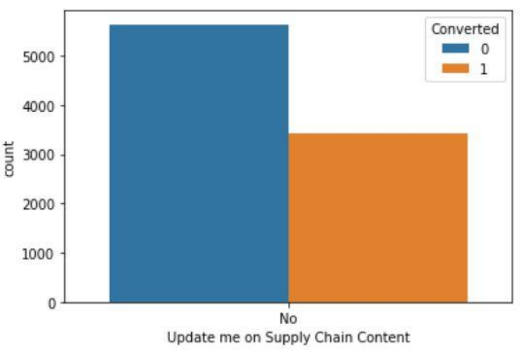


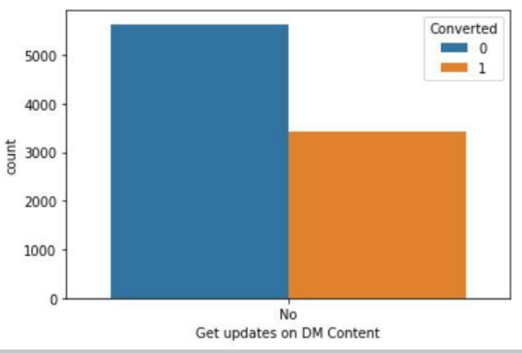






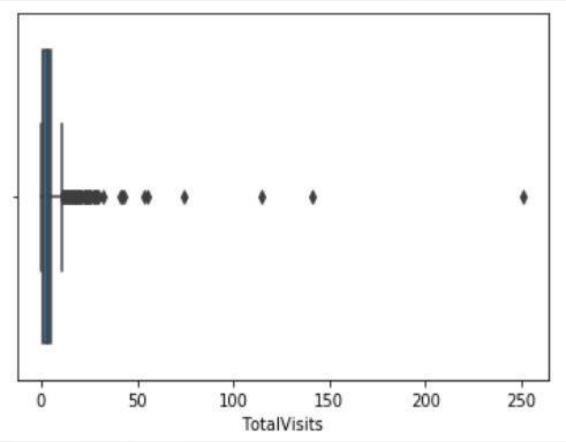




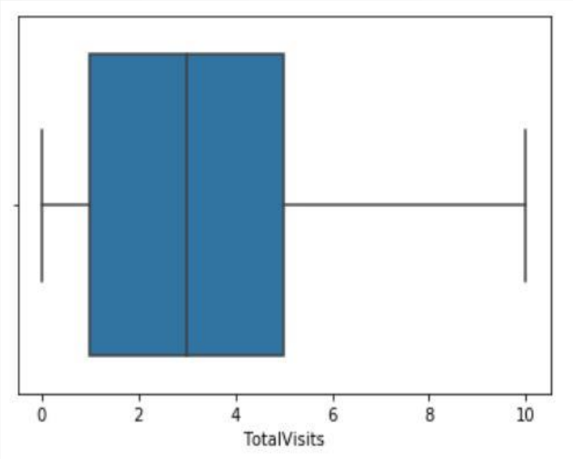


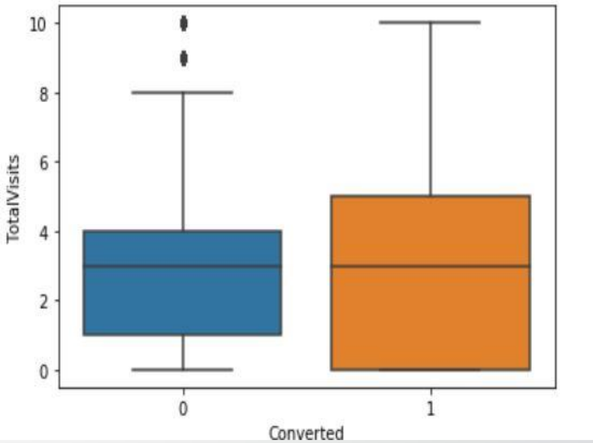
Handling outliers

Total Visits:

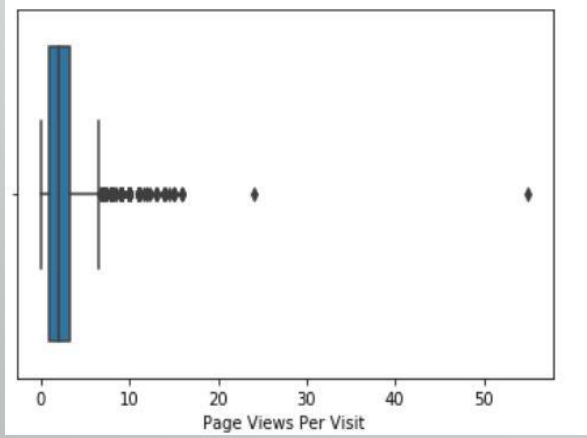


Outliers handled

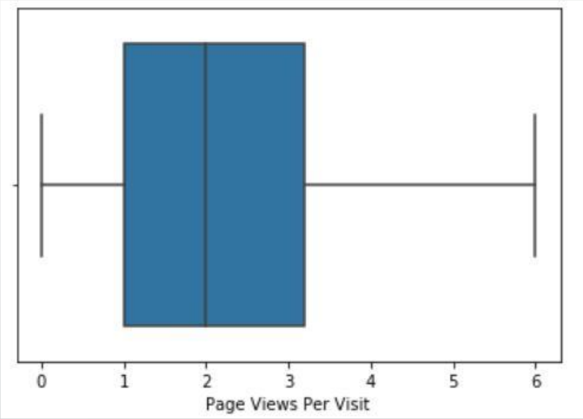


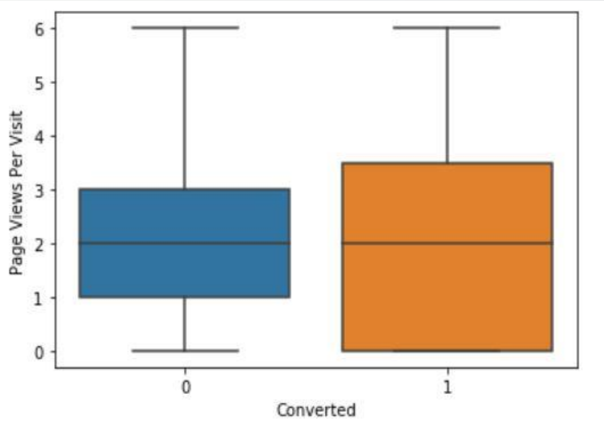


Page views per visits:

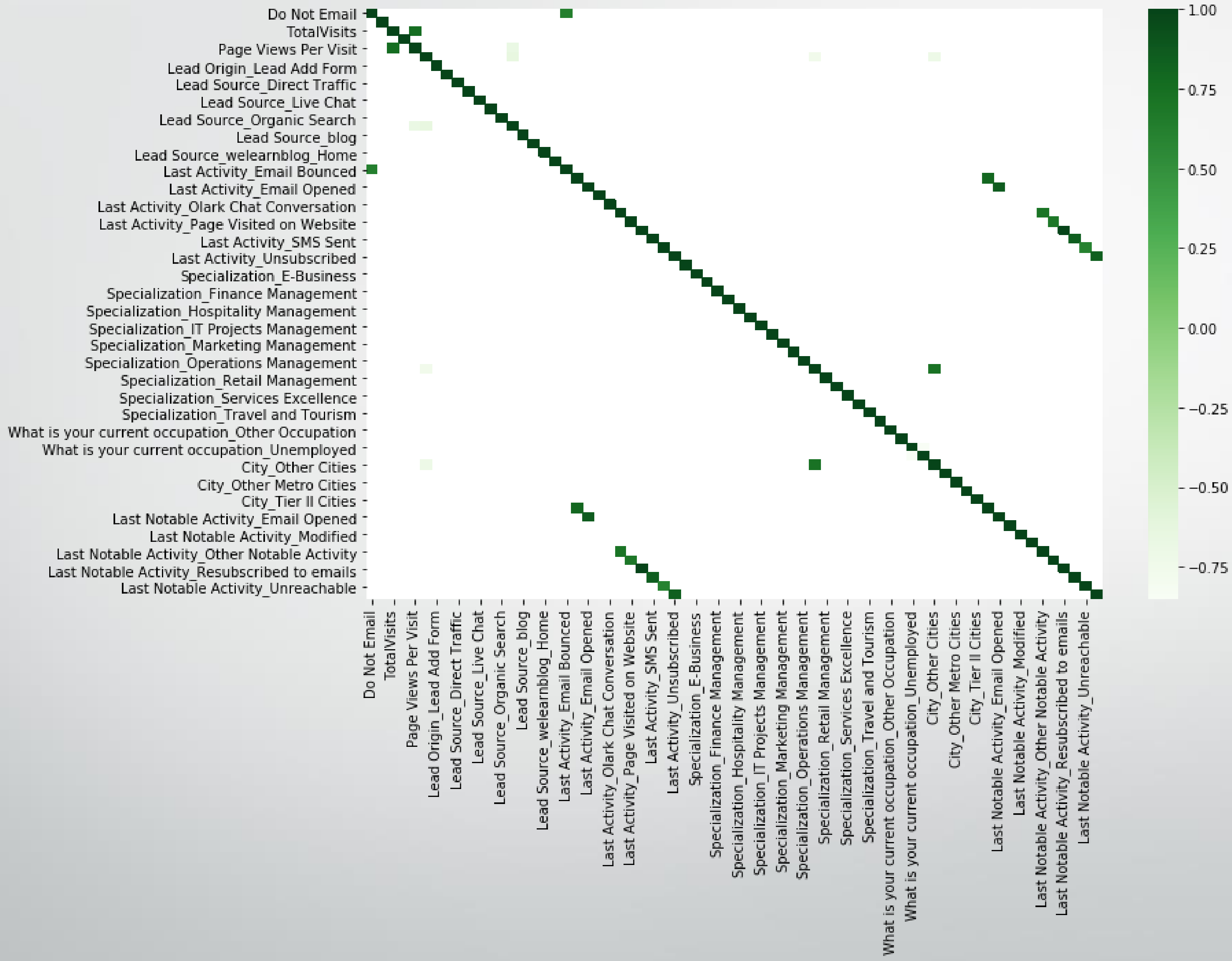


Outliers handled





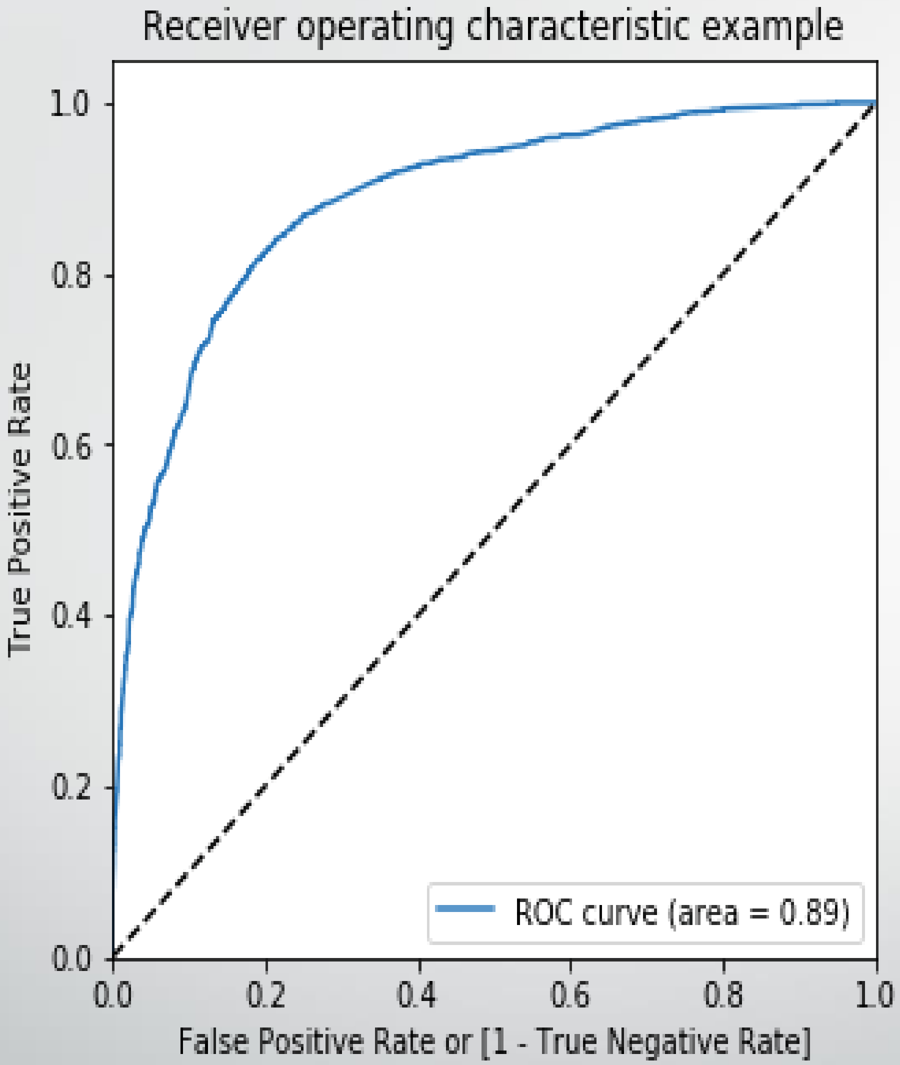
Correlation check



Checking of highly co-related columns

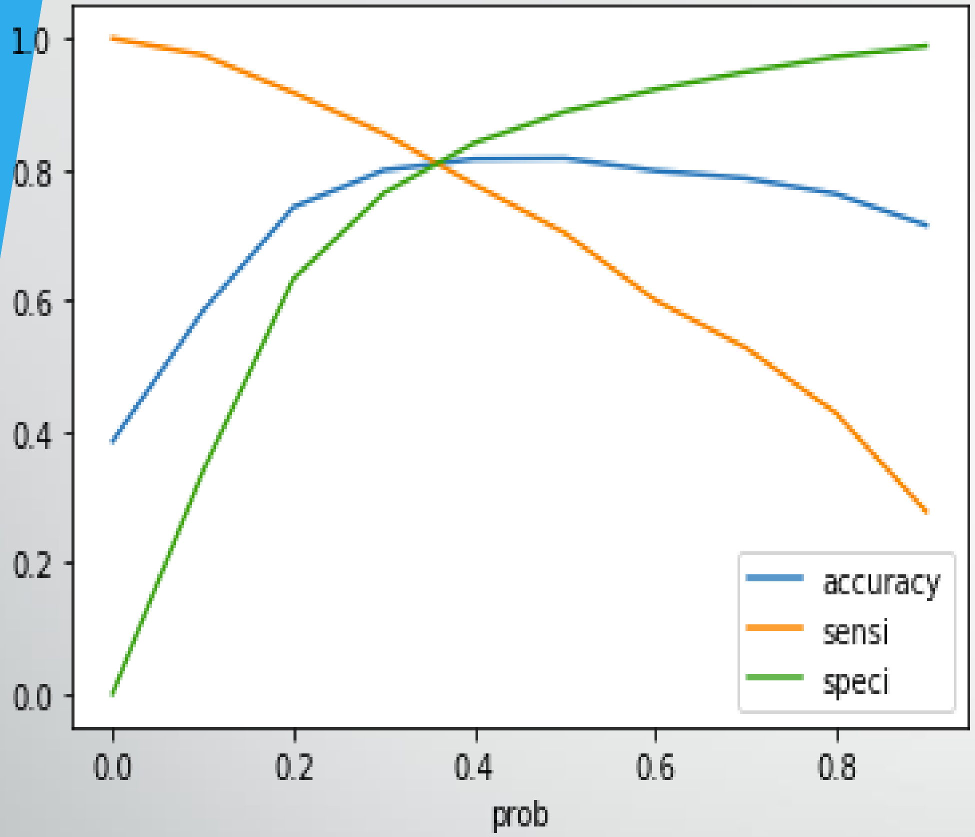
Dropped them as required for the model

ROC Curve



- After modelling, plotted ROC curve whether TPR & FPR are plotted in Y and X axis.
- The curve is more in left-upper as the AUC is high.
- It is good sign for the model.

Finding of Optimal Cutoff point



- We found model accuracy, sensitivity & specificity of the model as 81%, 70% and 89% with a default cut-off value of 0.5.

- For optimal cut-off we plotted accuracy, sensitivity and specificity as pic shown. We found optimal cut-off near about 0.38.

- With optimal cut-off, we rebuilt the model and found the accuracy, sensitivity and specificity as 81%, 82% and 81%.

- **With optimal cut-off, the model overall accuracy has increased**.

Testing of model on test data set, assignment of lead score and Feature importance:

- Tested the model on testing data set and found accuracy, sensitivity and specificity as 81%, 80% and 81% which signifies that the model is accurately built and successfully tested on testing data set results overall good accuracy on testing data.

- Based on model evaluation, the lead scores have been assigned for the Sales representative of X Education company.

- We found three highly features after finding features importance which are highly important for converting leads:
  - Total time spent on web sites
  - Lead Origin
  - What is your current occupation

Observations

- X Education to look into the persons/clients who basically searched for the X-Education website more. There may be the highly chances to convert them.

- They definitely have to look into the interested people who currently working professionals. Leads can be converted easily.

- Education company should focus on Lead Origin specially the origin from lead add forms. They can be easily converted.

THANK YOU