Lead scoring case study

Submitted by:

Chetan R Tippa

Madhurima Sumit Falls

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%.

Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are 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 as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Goals of the study

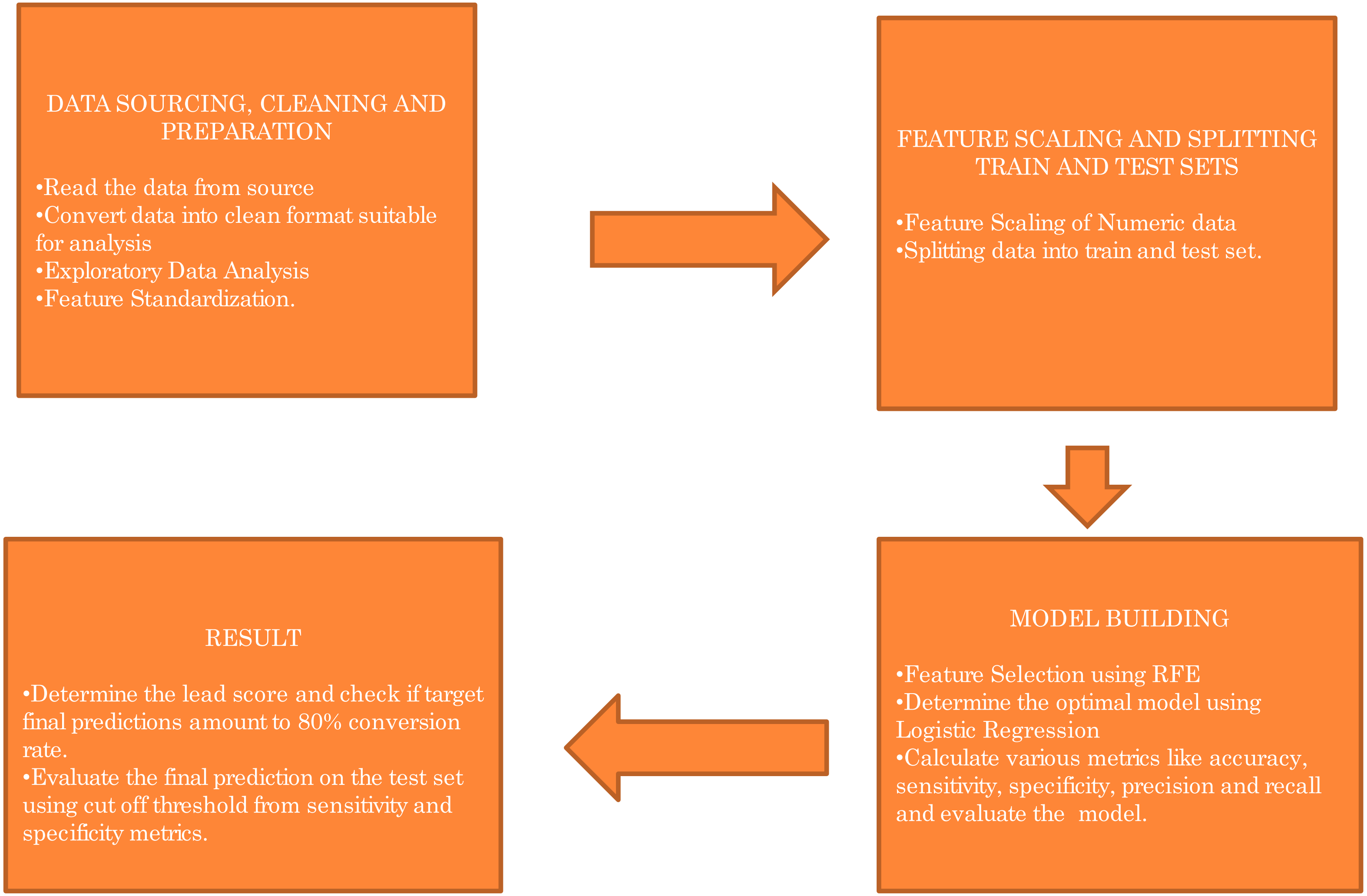
- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

- There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well.

APPROACH

- Source the data for analysis
- Reading and understanding the data
- Data Cleaning
- EDA
- Feature scaling
- Splitting the data into test and train dataset
- Preparing the data for modelling
- Model building
- Model evaluation-specificity and sensitivity or precision recall
- Making predictions on the test set

PROBLEM SOLVING METHODOLOGY



DATA SOURCING, CLEANING AND PREPARATION

- Read the data from source
- Convert data into clean format suitable for analysis
- Exploratory Data Analysis
- Feature Standardization.

FEATURE SCALING AND SPLITTING TRAIN AND TEST SETS

- Feature Scaling of Numeric data
- Splitting data into train and test set.

MODEL BUILDING

- Feature Selection using RFE
- Determine the optimal model using Logistic Regression
- Calculate various metrics like accuracy, sensitivity, specificity, precision and recall and evaluate the model.

RESULT

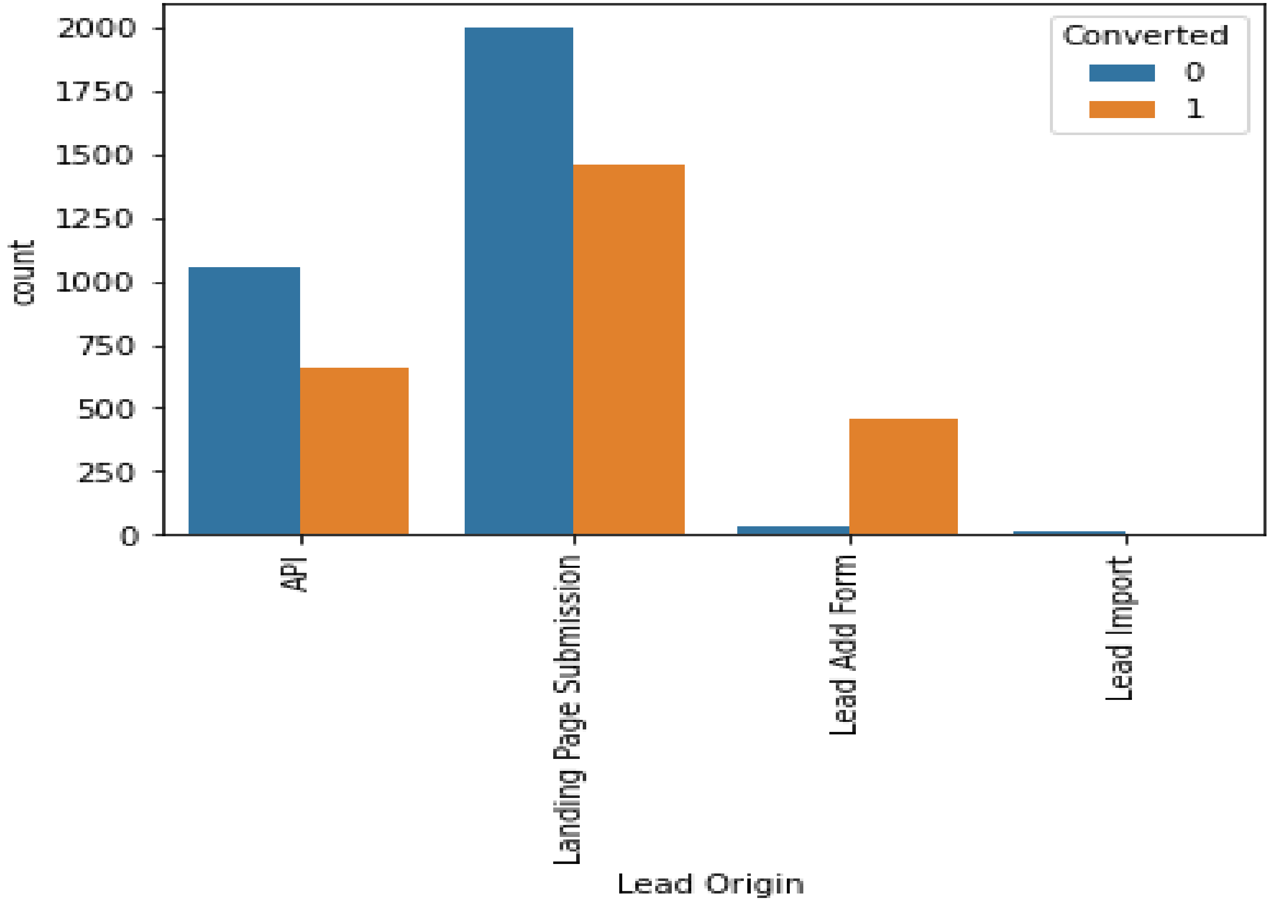
- Determine the lead score and check if target final predictions amount to 80% conversion rate.
- Evaluate the final prediction on the test set using cut off threshold from sensitivity and specificity metrics.

DATA SOURCING, CLEANING & PREPARATION

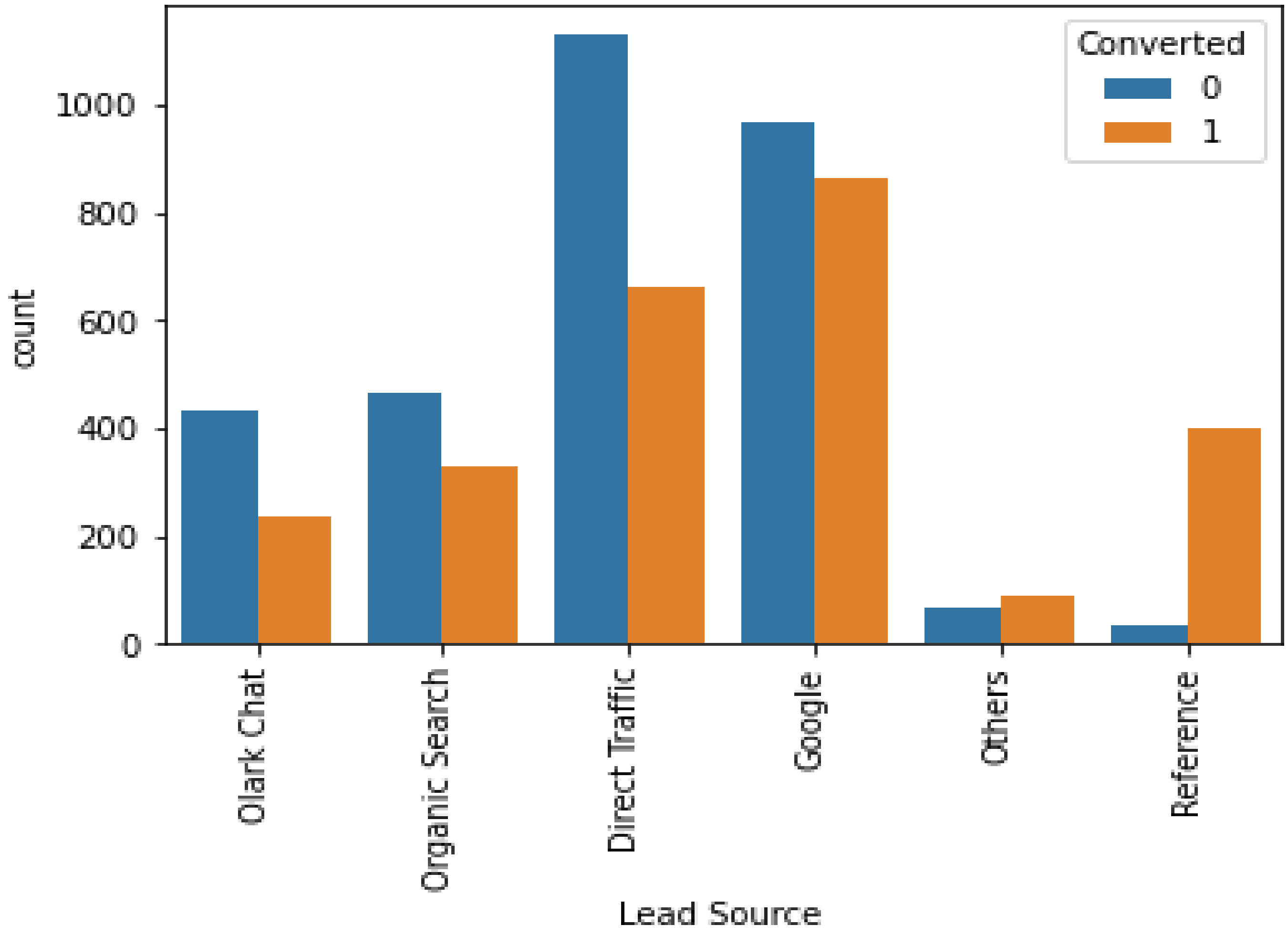
- Read the data from CSV file.
- Outliers treatment
- Data cleaning- Handling Null Values & removing higher Null values data
- Removing Redundant columns in the data
- Imputing Null values
- Exploratory data analysis
- Feature Standardisation

Data visualisation

- Lead origin :

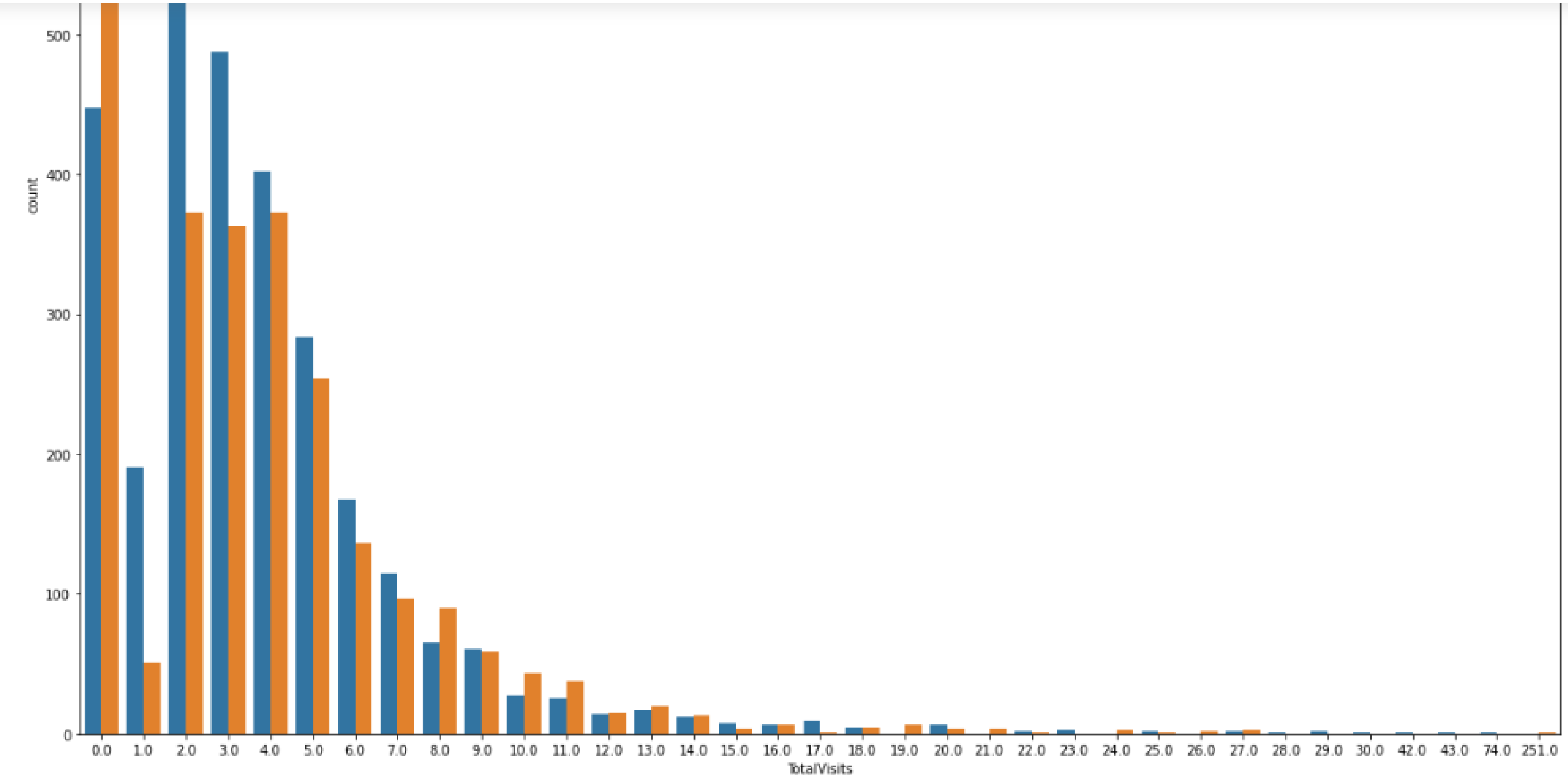


- Lead Source:

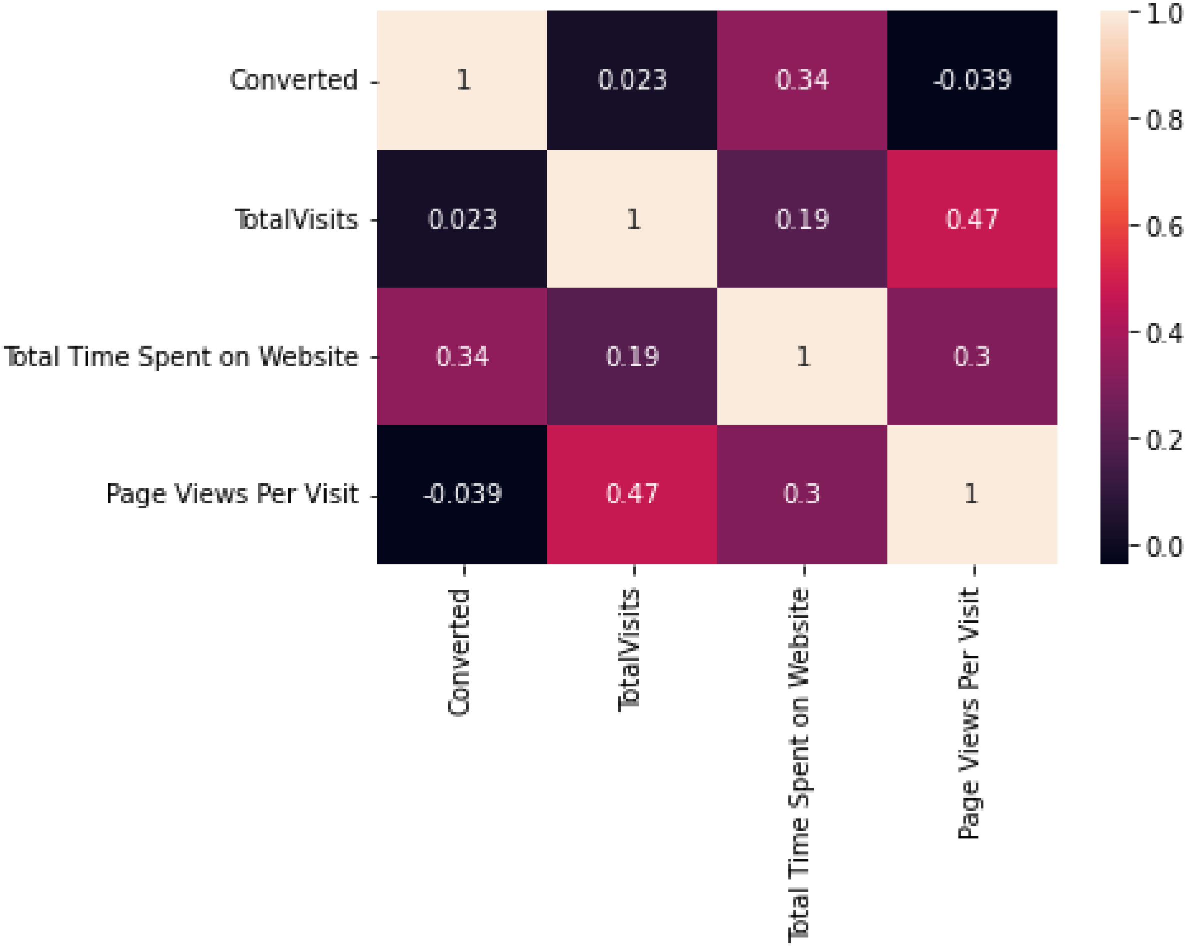


Converted
0
1
count
1000
800
600
400
200
0
Olark Chat
Organic Search
Direct Traffic
Google
Others
Reference
Lead Source

- Total Visits:

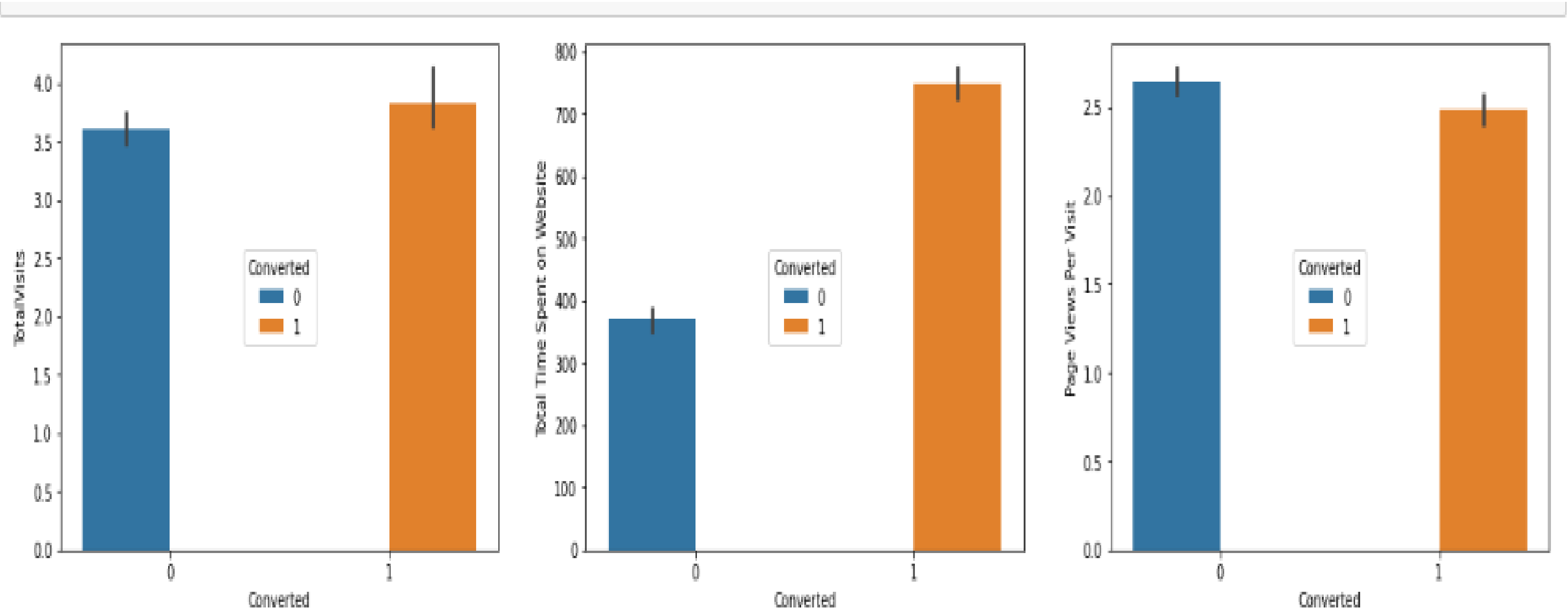


- TotalVisits, Total Time Spent on Website,Page Views Per Visit - Corelation matrix :

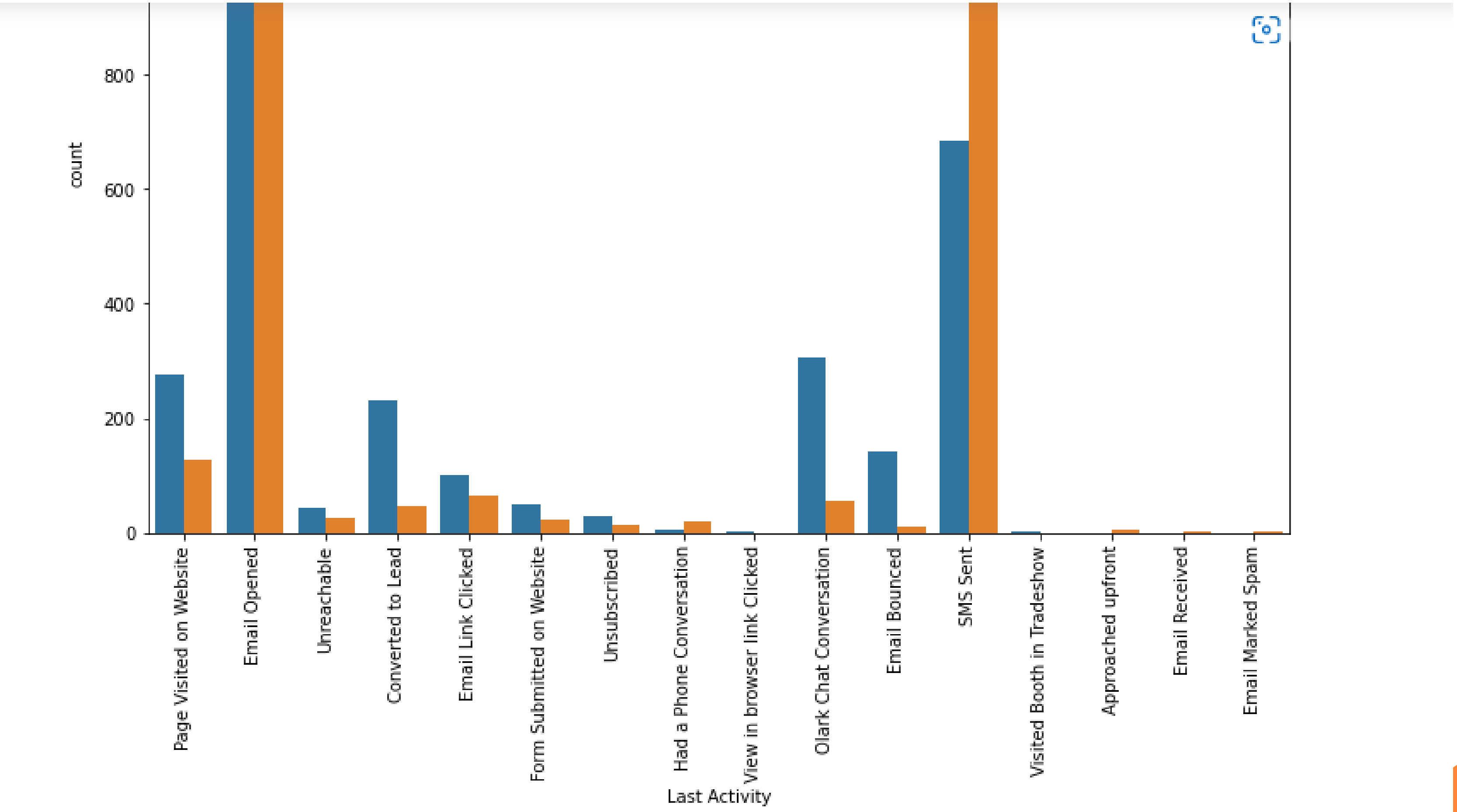


	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit
Converted	1	0.023	0.34	-0.039
TotalVisits	0.023	1	0.19	0.47
Total Time Spent on Website	0.34	0.19	1	0.3
Page Views Per Visit	-0.039	0.47	0.3	1

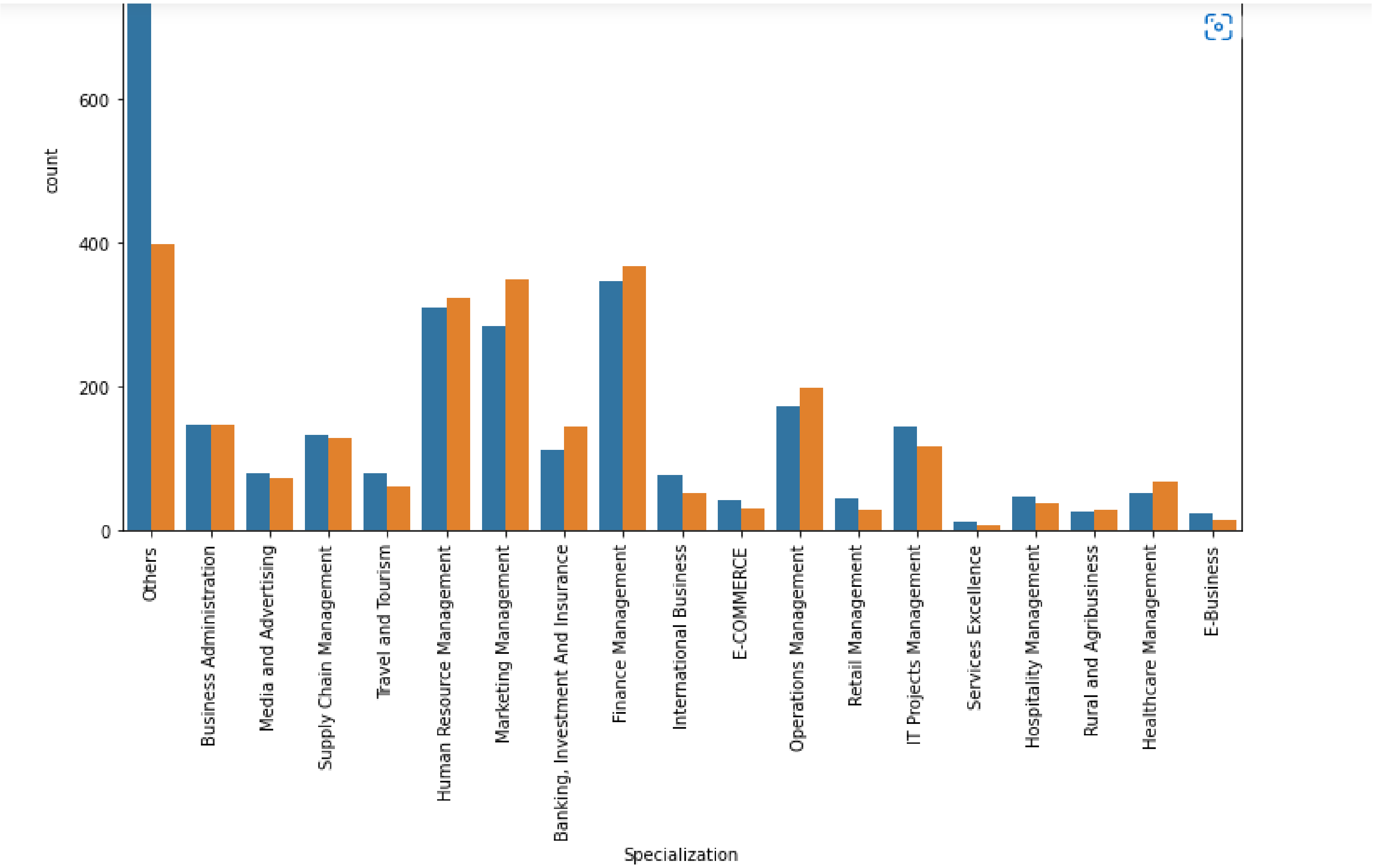
- Total Visits, Total Time Spent on Website, Page Views Per Visit



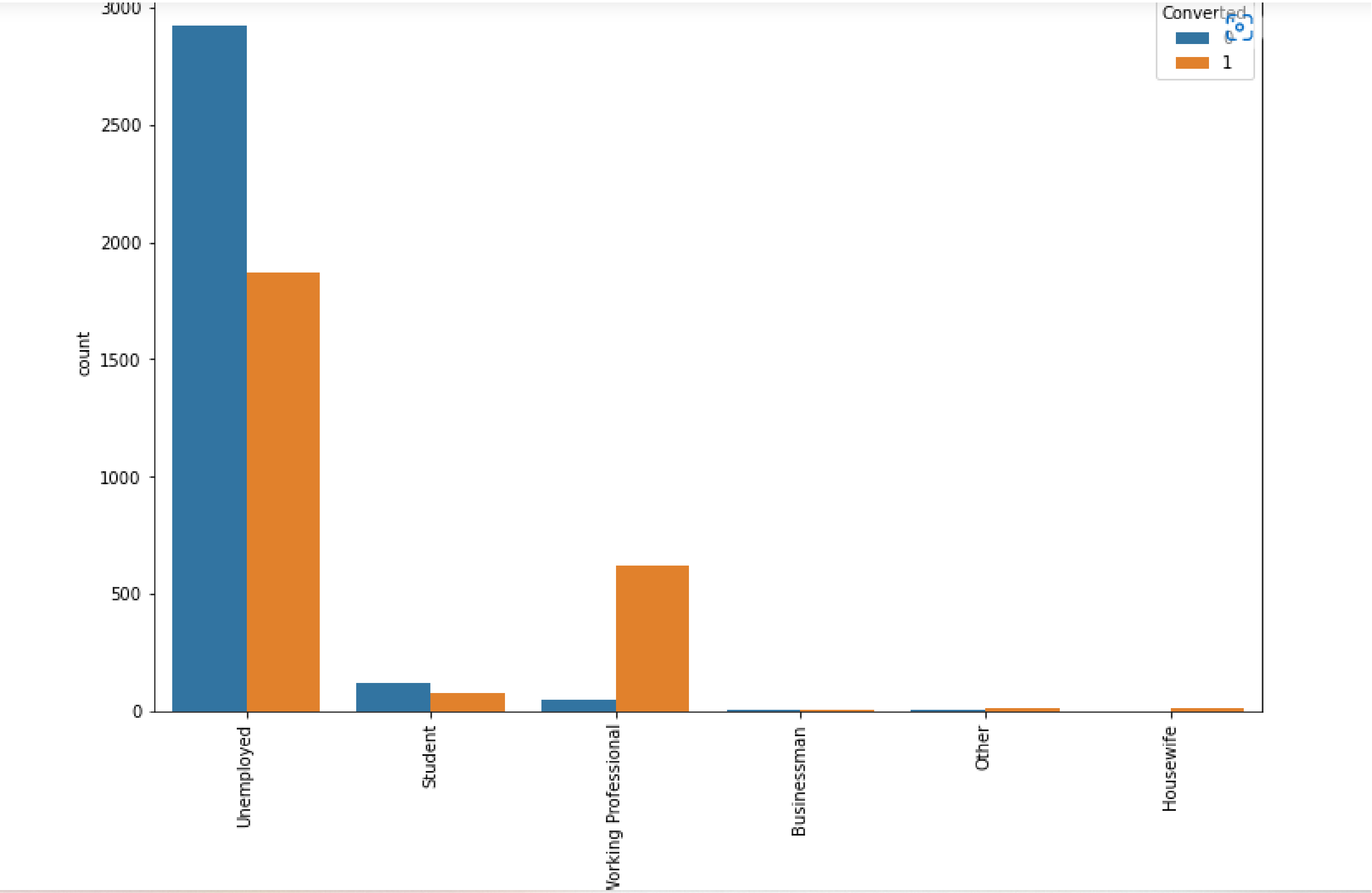
- Last Activity:



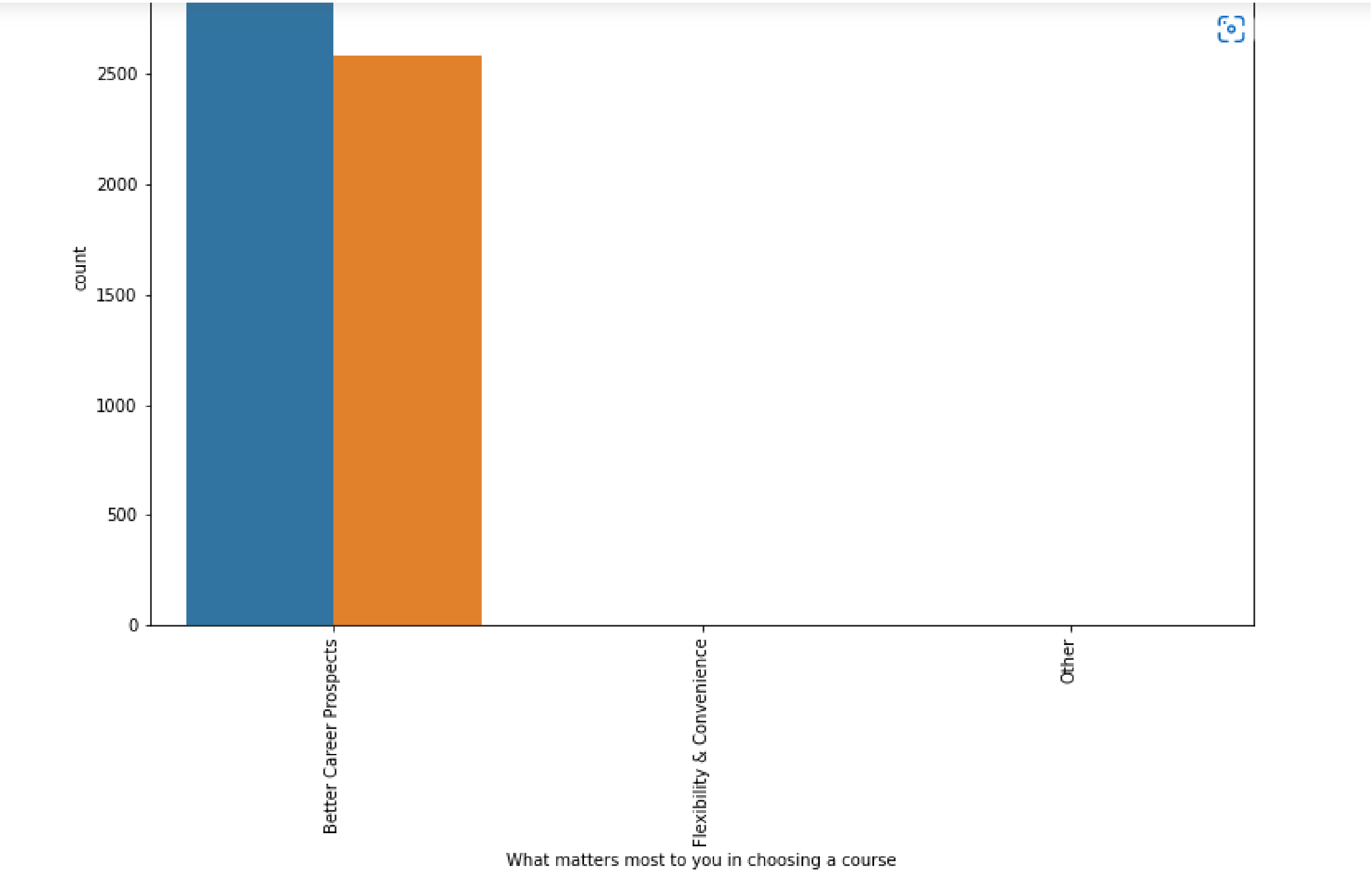
- Specialization:



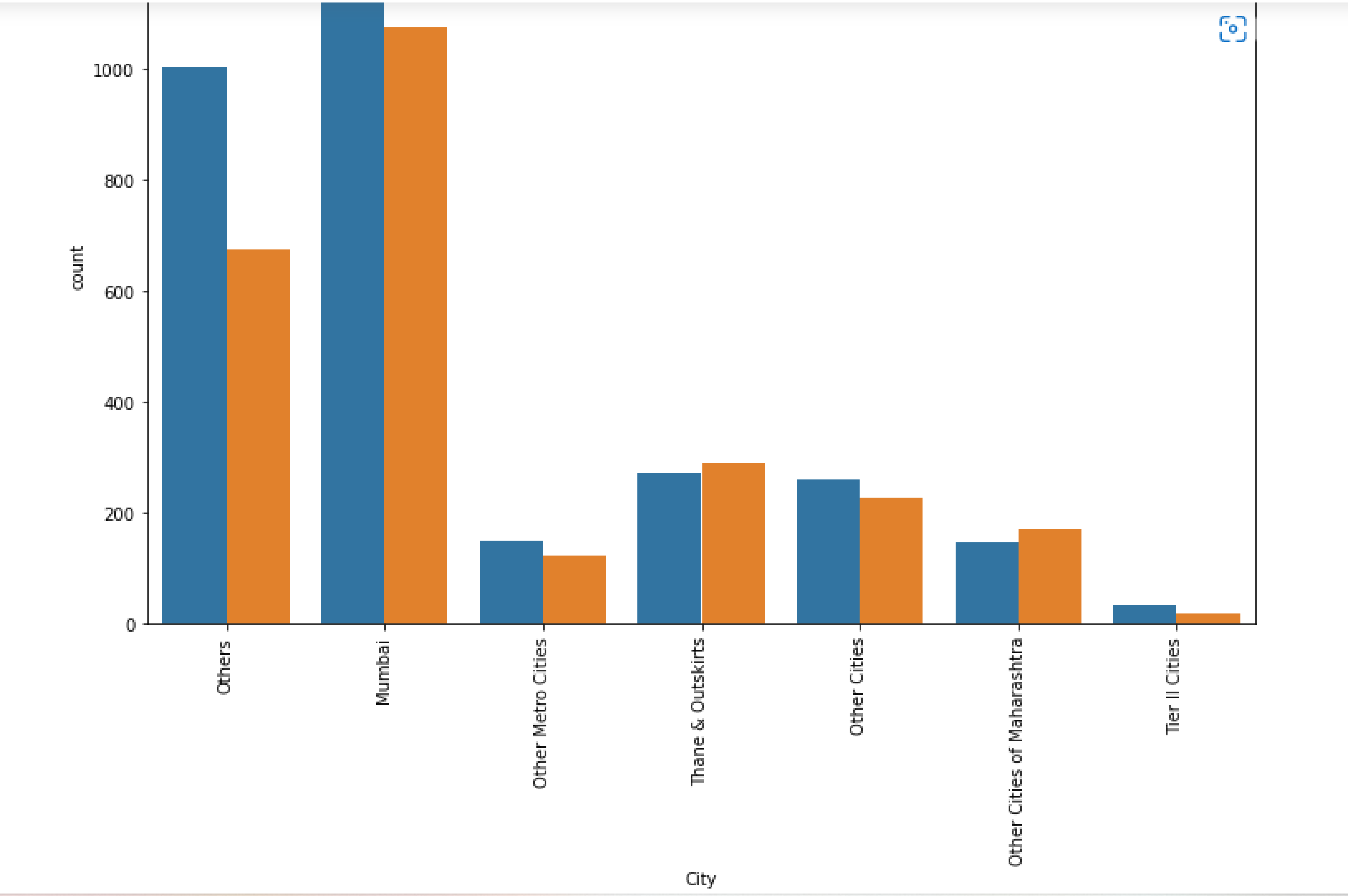
- What is your current occupation?



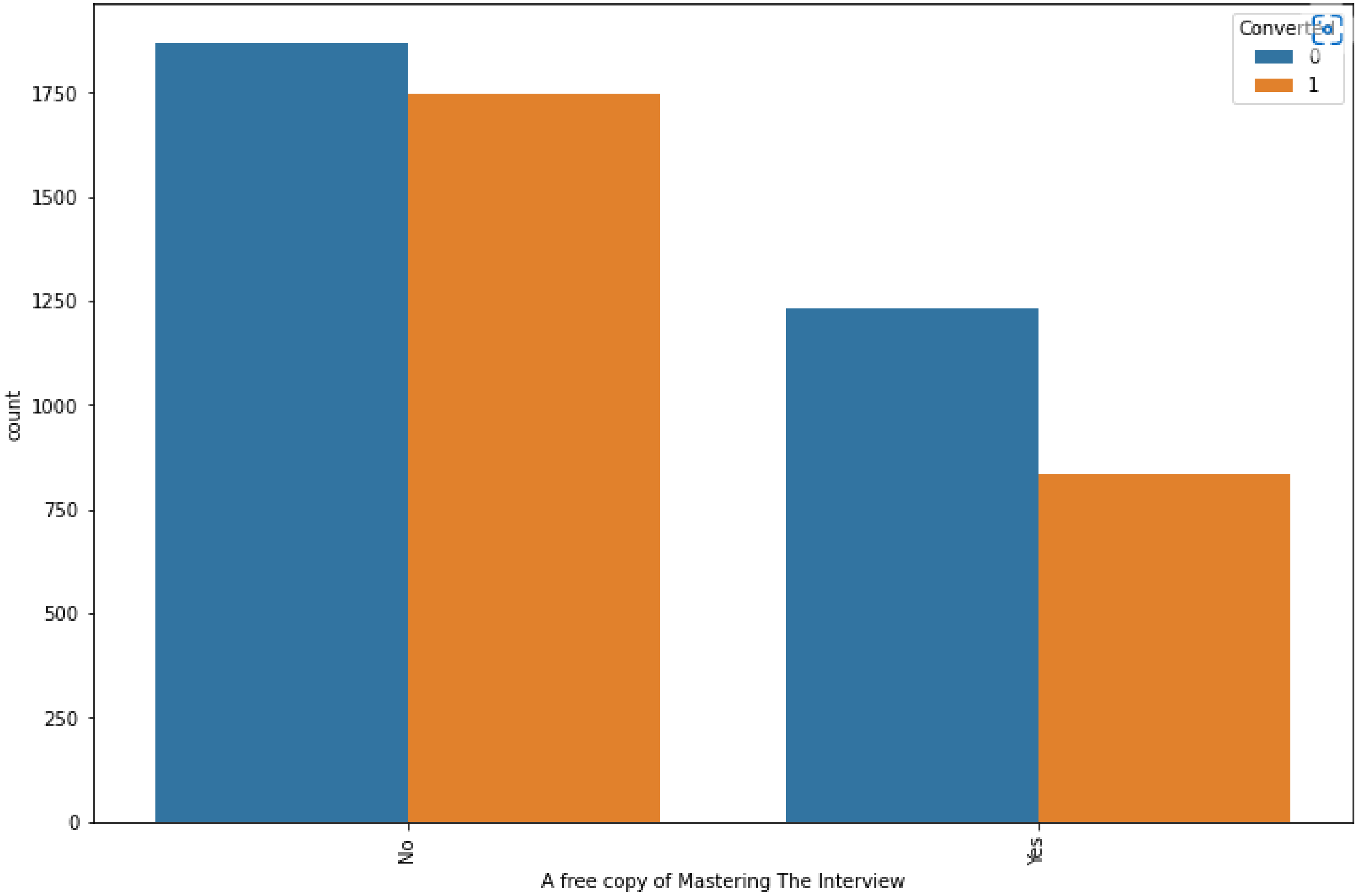
- What matters most to you in choosing a course?



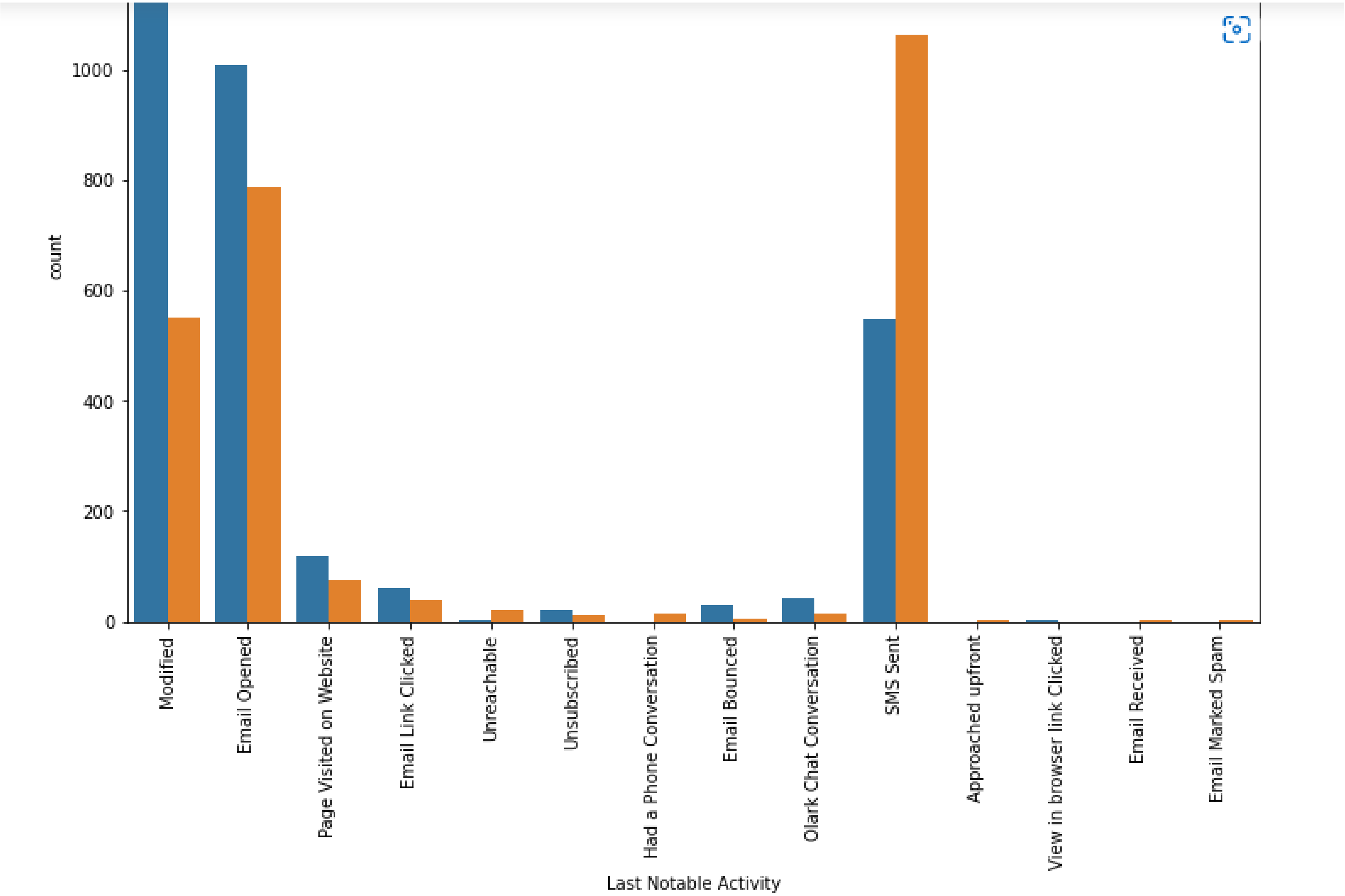
- City:


count
1000
800
600
400
200
0
Others
Mumbai
Other Metro Cities
Thane & Outskirts
Other Cities
Other Cities of Maharashtra
Tier II Cities
City

- A free copy of Mastering The Interview:



Last Notable Activity:



VARIABLES IMPACTING THE CONVERSION RATE

- Total Visits
- Total Time Spent on website
- Lead Source_Reference
- Lead Origin_Lead Add Form
- Lead Source_Welingak Website
- Do Not Email
- Lead Source _Referance….etc.

DATA PREPARATION

- Converted Binary variables into 0 & 1
- Created dummy variables for categorical variables

Feature Scaling & Splitting Train & Test Sets

- Feature Scaling of Numeric Data □
- Splitting data into Train & Test Set

MODEL BUILDING

- Feature Selection using RFE
- Determined Optimal Model using Logistic Regression
- Calculated accuracy ,sensitivity specificity, precision & Recall & evaluate model

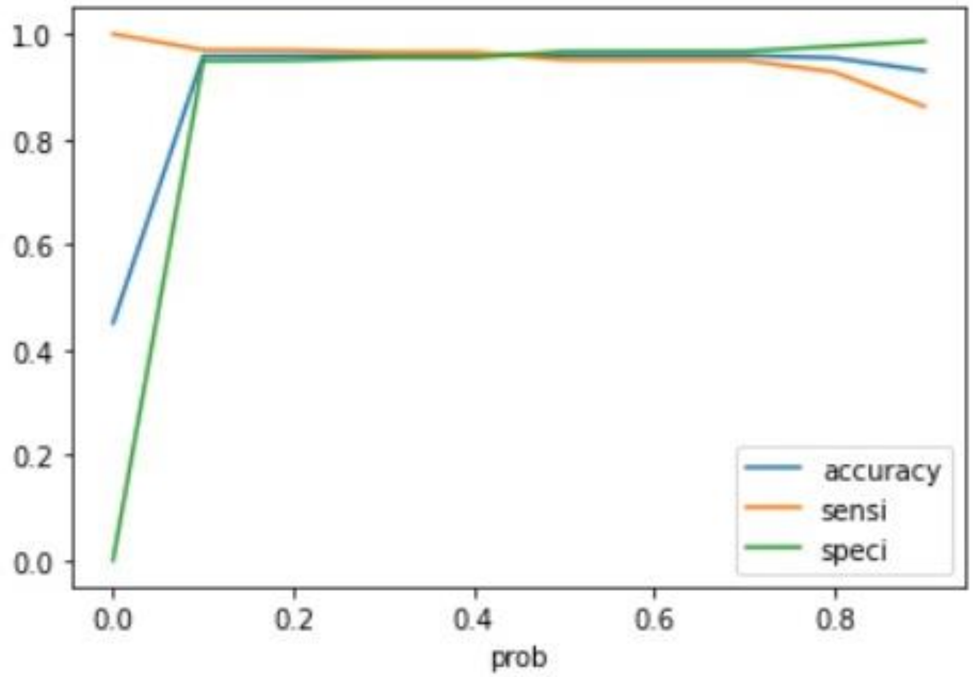
Model Evaluation-Sensitivity & Specificity on Train Data Set

Graph depicts an optimal cutoff of 0.42 bases on Accuracy,Sensitivity,Specificity.

Accuracy= 95.9%

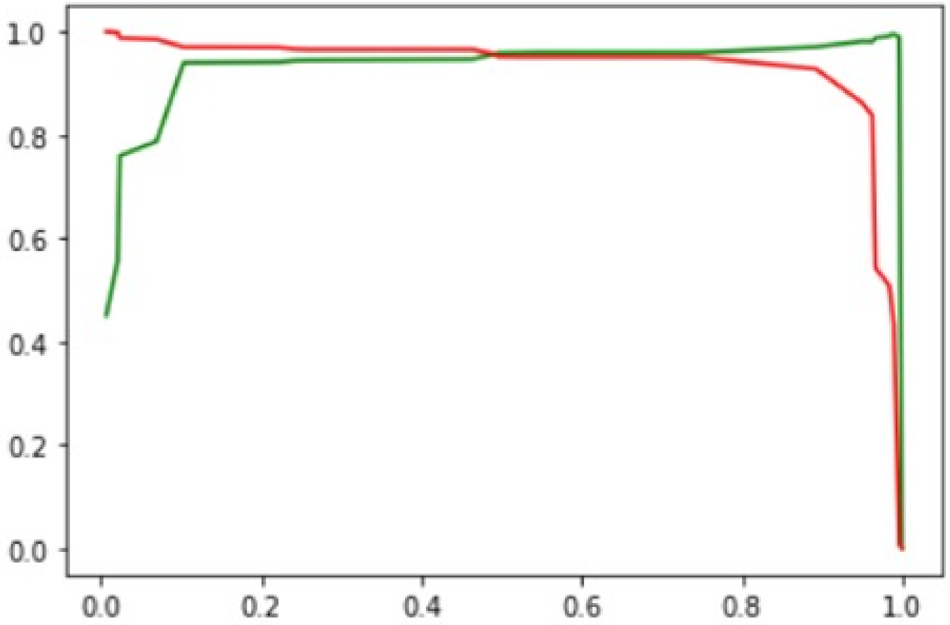
Sensitivity= 96.5%

Specificity= 95.5%



Model Evaluation Precision & Recall on Train dataset

The precision and recall have a trade-off at 0.5. Hence, 0.5 wil be used as threshold on test data



Precision= 95.9 %

Recall= 95.1%

MODEL EVALUATION

Sensitivity & Specificity on Test Dataset

Accuracy= 95.7 %

Sensitivity= 94.4%

Specificity= 96.9%

RESULT

- Accuracy, Sensitivity and Specificity values of training and test set are close to training set
- Accuracy, Sensitivity and Specificity values of training set are 95.9%,96.5%,95.5% Respectively
- Accuracy, sensitivity & Specificity values of test are 95.7%,94.4%,96.9% Respectively
- We have done the prediction on the test set using cut off threshold from sensitivity & specificity metrics

CONCLUSION

- While we have checked both sensitivity-specificity as well as Precision & recall metrics, we have considered the optimal cut off based on sensitivity & specificity for calculating the final prediction
- Accuracy, Sensitivity & specificity values of test set are around 95.7%, 94.4%, 96.9% which are approximately closer to Values calculated using Trained Data Set
- Hence, Overall Model seems to be Good