LEAD SCORING CASE STUDY

In this case study, we created logistic regression model to help them select the most promising leads.

FAIZA AHMAD I SWAPNIL RANJAN

DS C34 31ST JULY 2021

BUSINESS OBJECTIVES

- Building a model to help the company select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

- 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.

APPROACH

01

In application new dataset we have 37 columns and 9240 rows.

02

Columns having missing values more than 45% have been dropped as it is not useful for creating insights.

03

There are days columns which is having NAN values, so we are converting that to Not Specified.

04

Performed Attribute Analysis to find the converted data..

APPROACH

05

Capping the outliers to 95% value for analysis

06

Outlier detection, analysis and correlation are performed on the dataset.

07

Feature Scaling, Model Building, Evaluation and Prediction performed on the dataset.

08

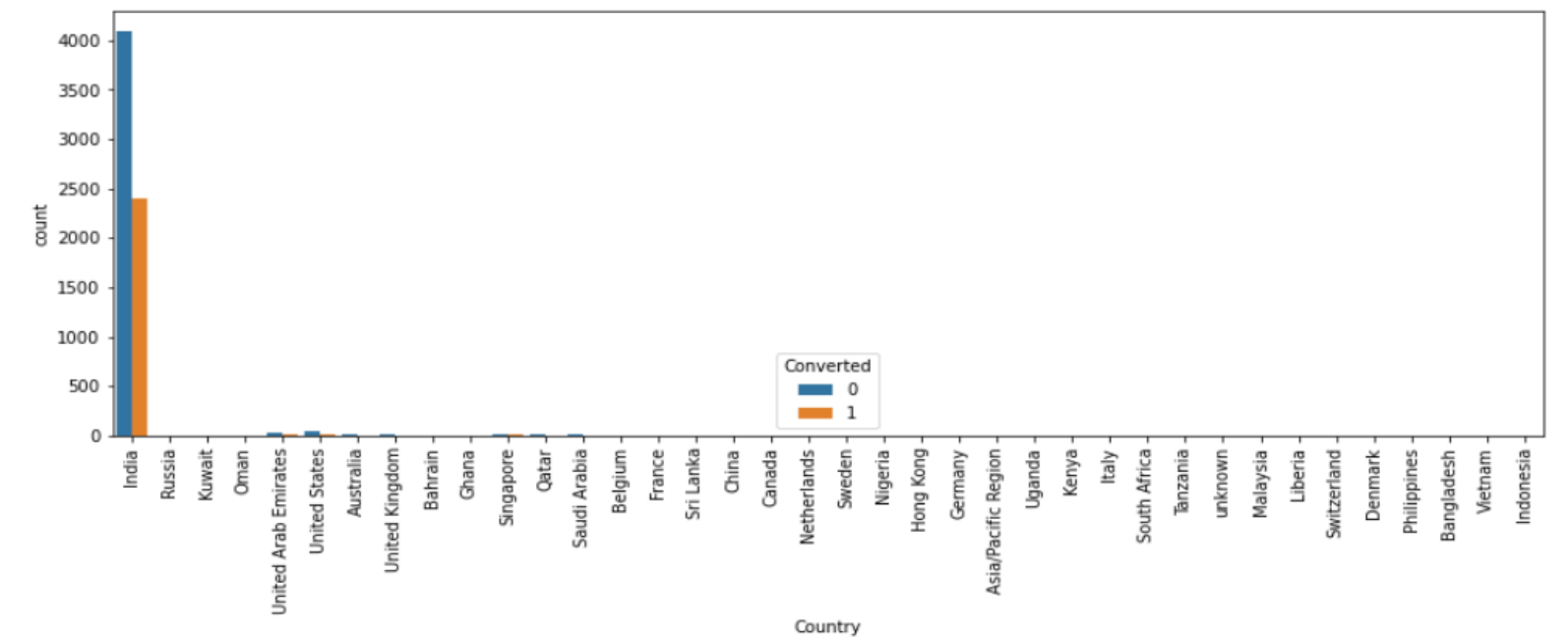
Conclusion and insights are provided at the end of the notebook.

ANALYZING Lead Scoring DATA

Exploratory Data Analysis

Country vs Converted

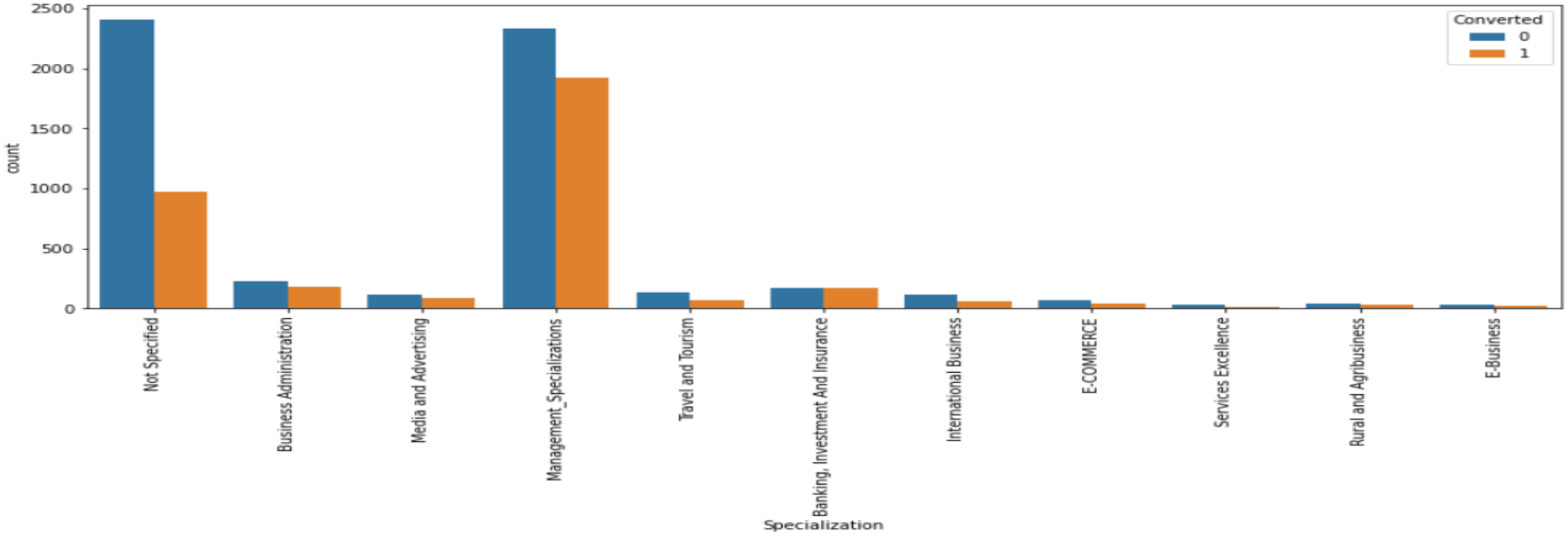
Country vs Converted: As we can see the Number of Values for India are quite high, this column can be dropped.



Exploratory Data Analysis

Specialization vs Converted

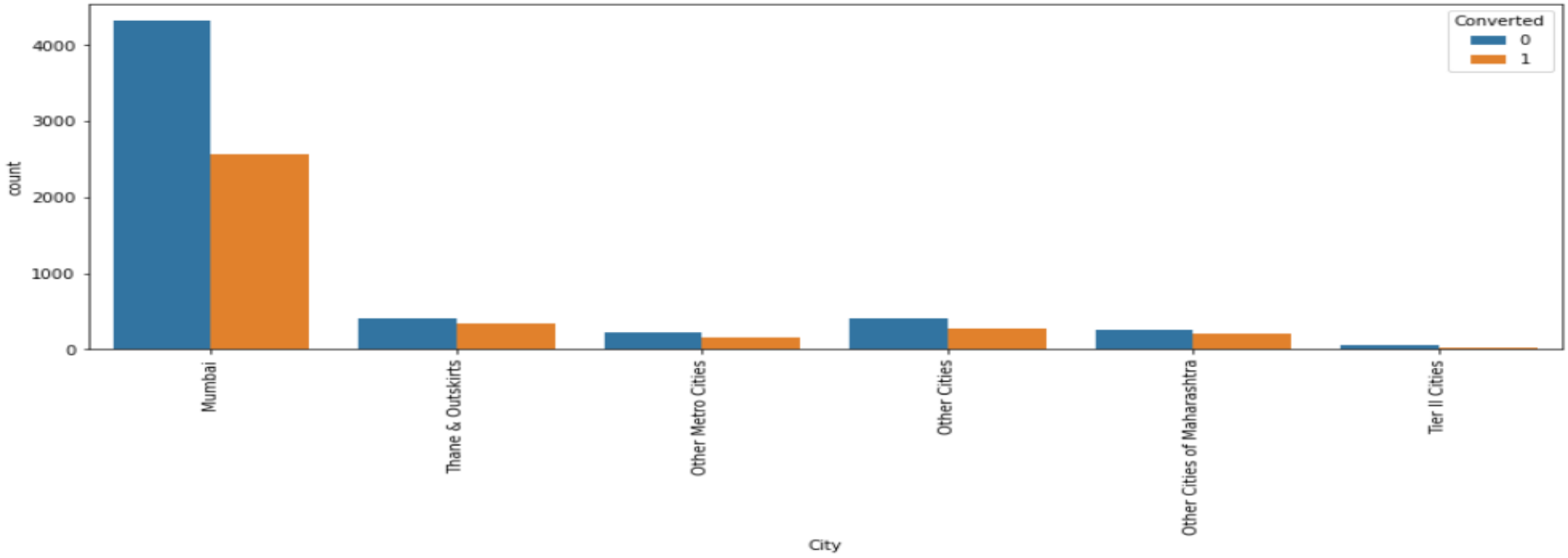
Specialization vs Converted: We see that specialization with Management have higher number of leads as well as leads converted. So, this is a significant variable and will be used further for analysis



Exploratory Data Analysis

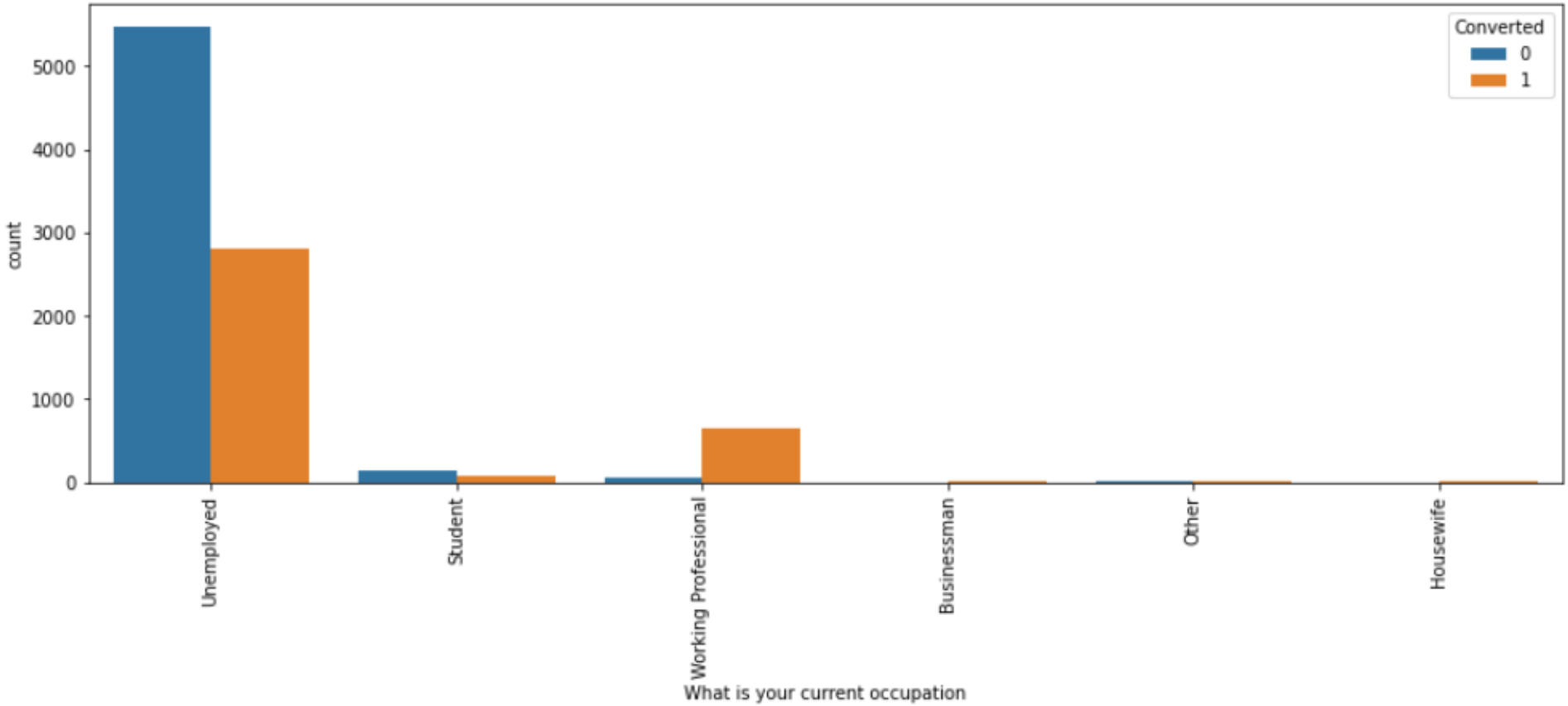
City vs Converted

City vs Converted: We see that Mumbai is having the highest count, that means having most number of leads.



Exploratory Data Analysis

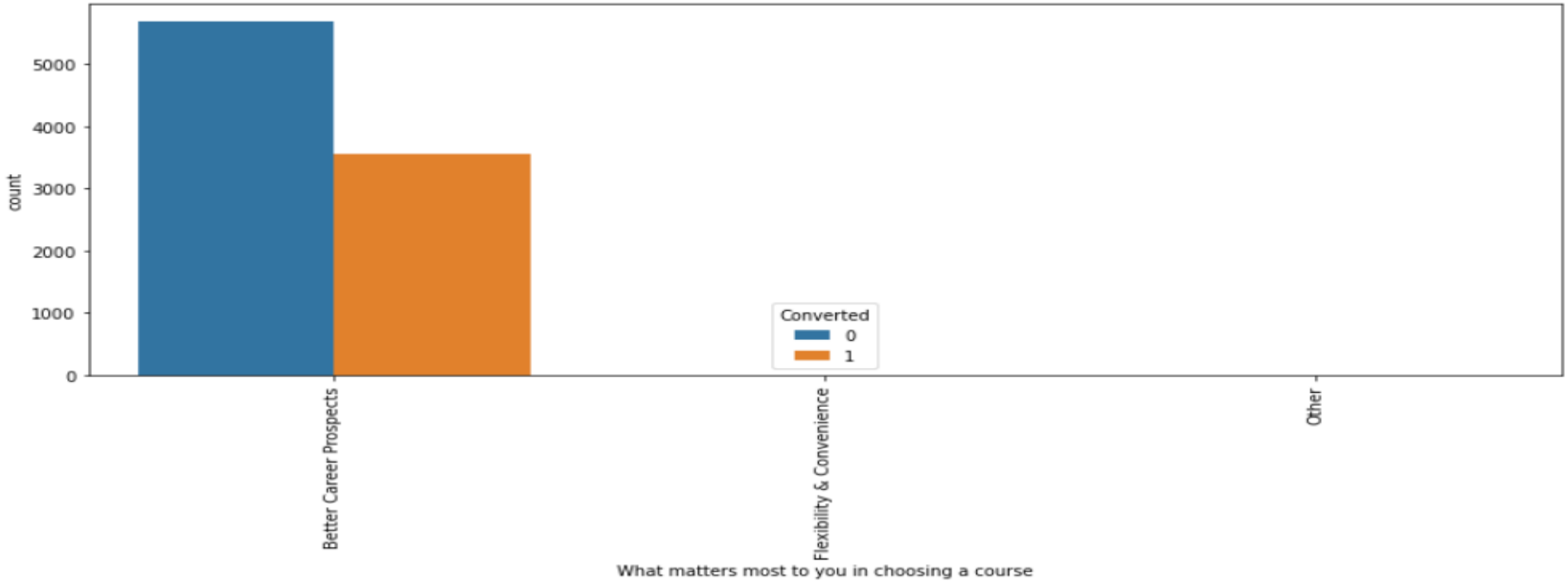
What is your current occupation vs Converted



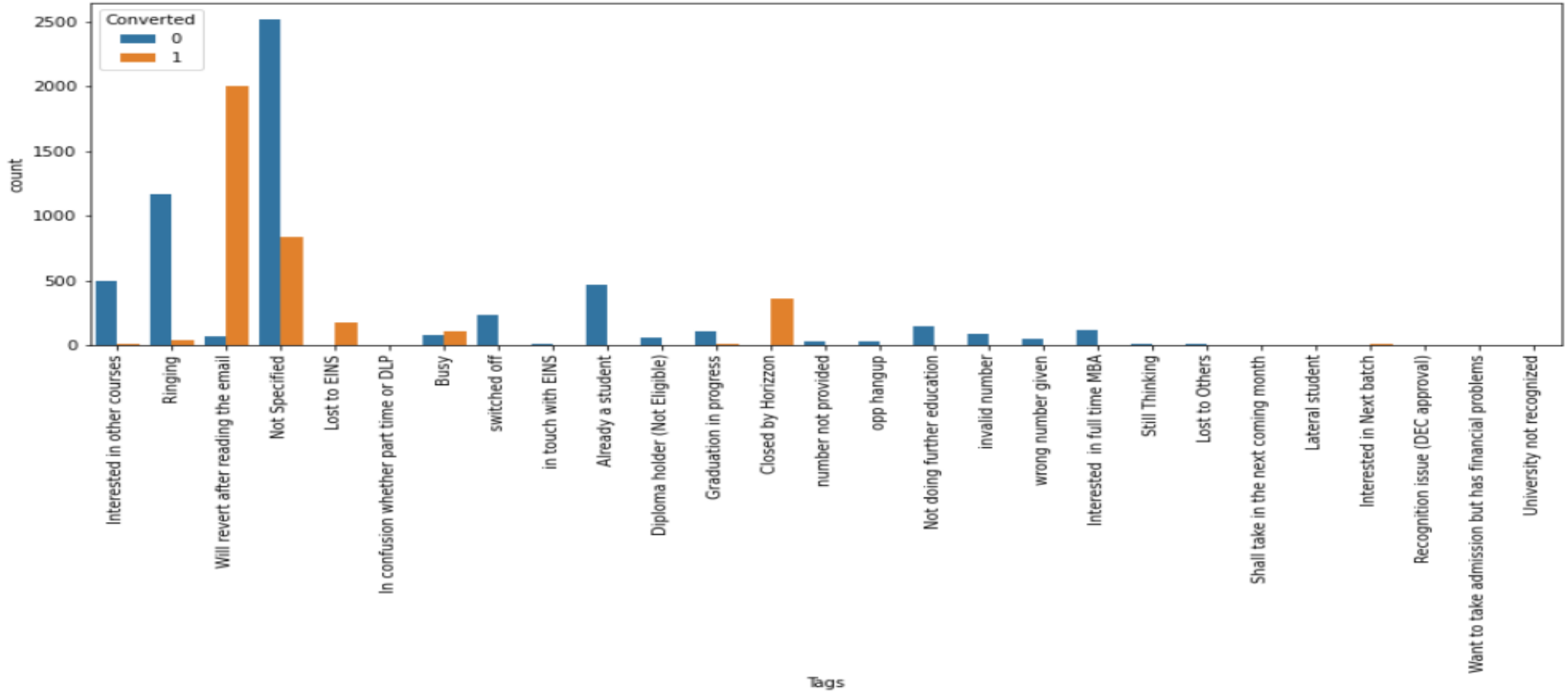
Exploratory Data Analysis

What matters most to you in choosing a course vs Converted

What matters most to you in choosing a course vs Converted: As we can see the Number of Values are quite high, this column can be dropped.

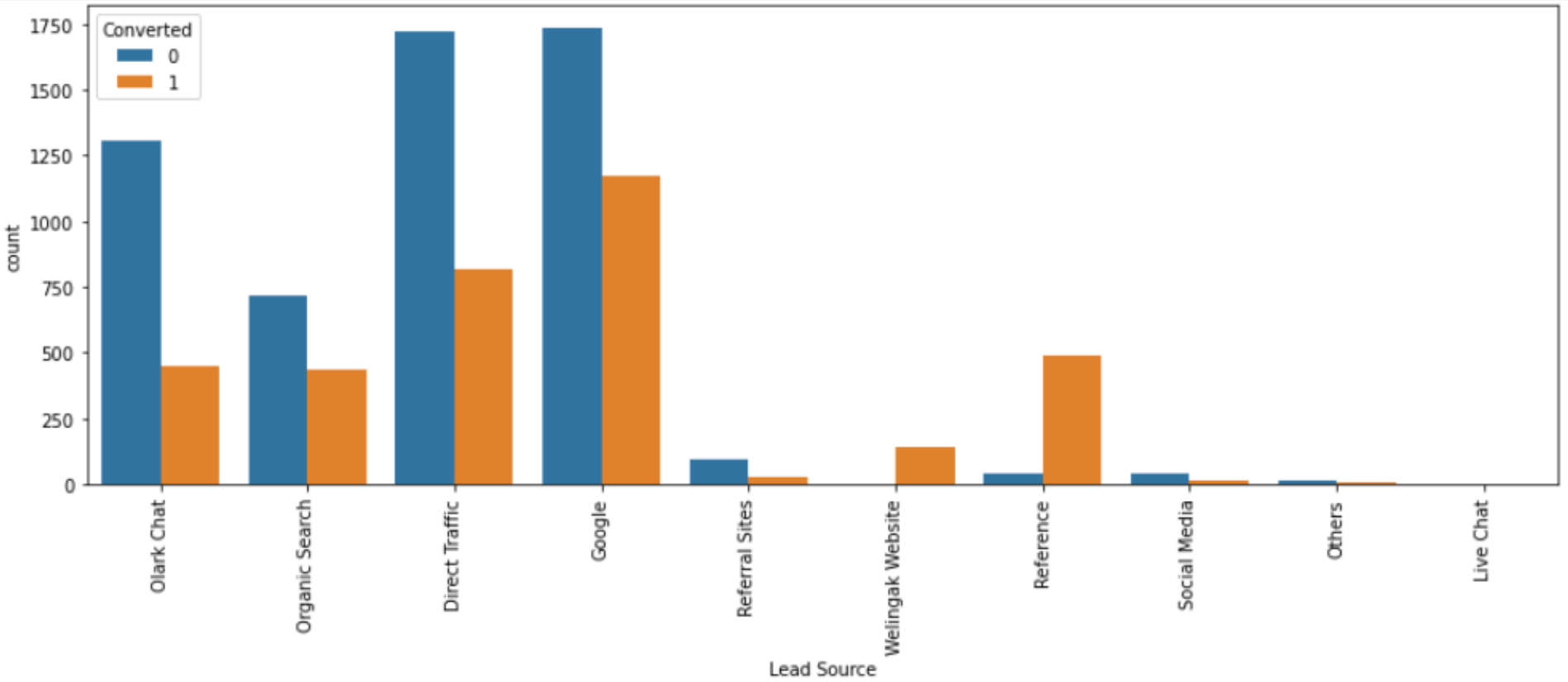


Exploratory Data Analysis
Tags vs Converted



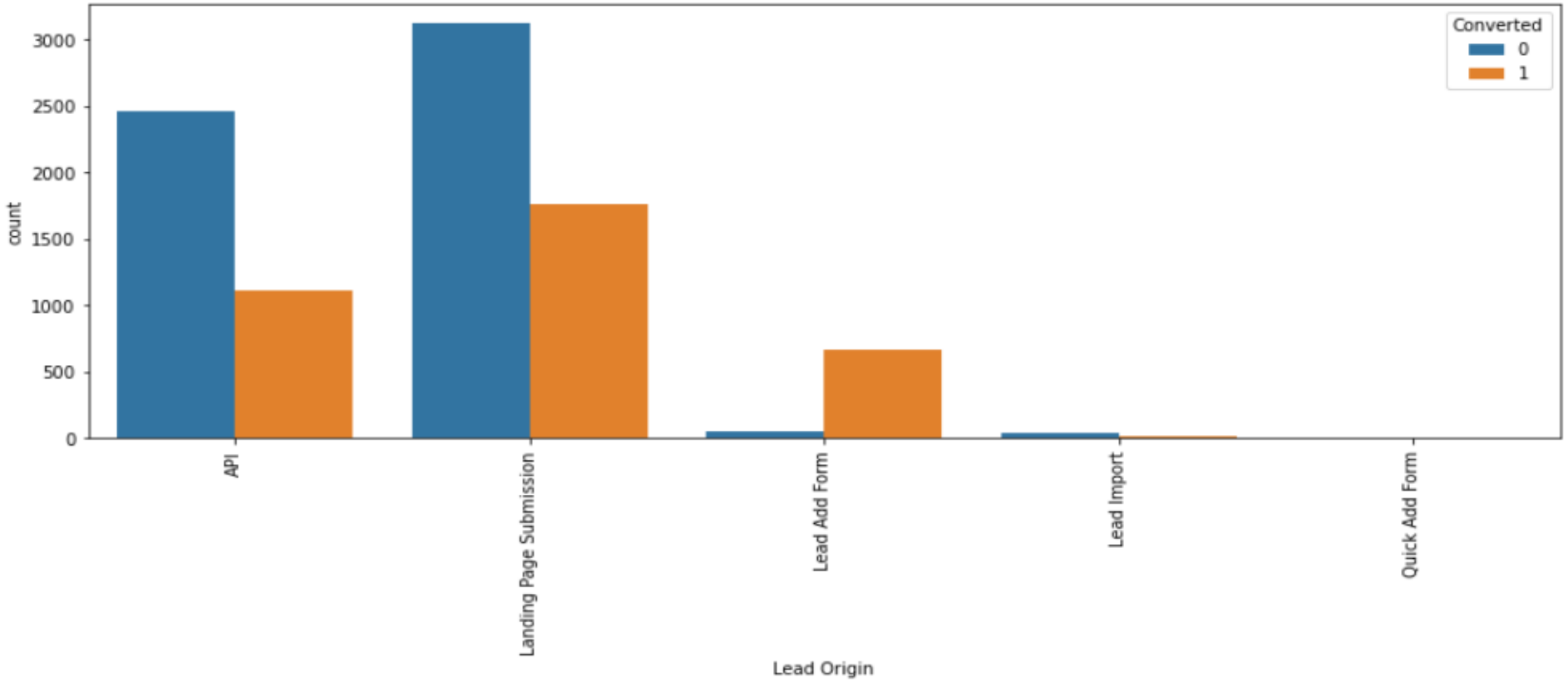
Exploratory Data Analysis

Lead Source vs Converted



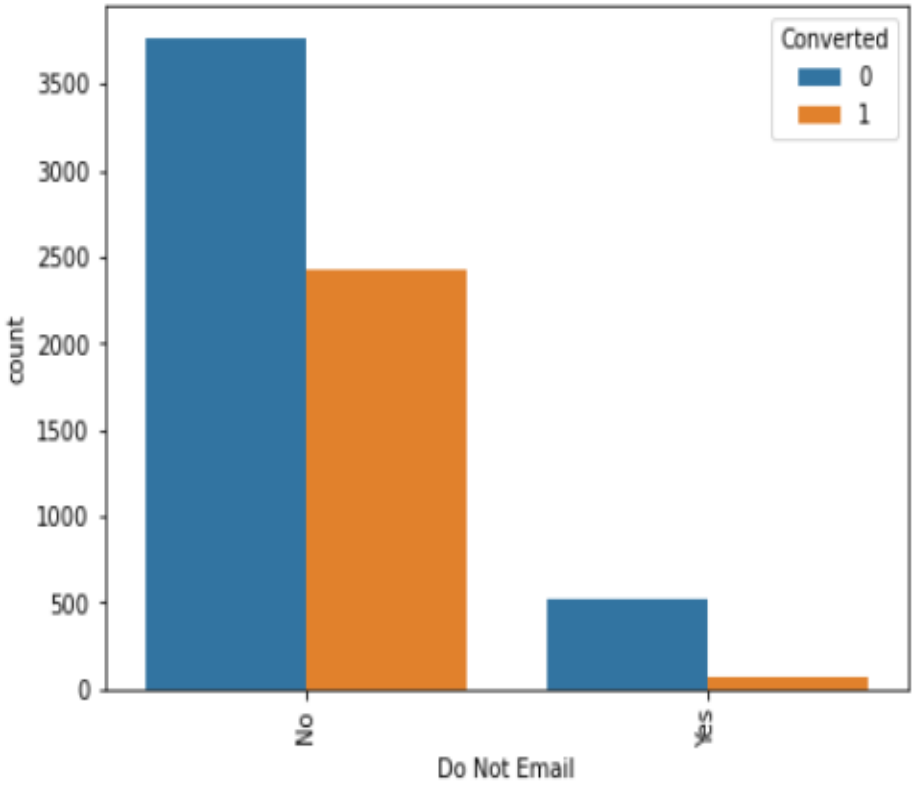
Exploratory Data Analysis

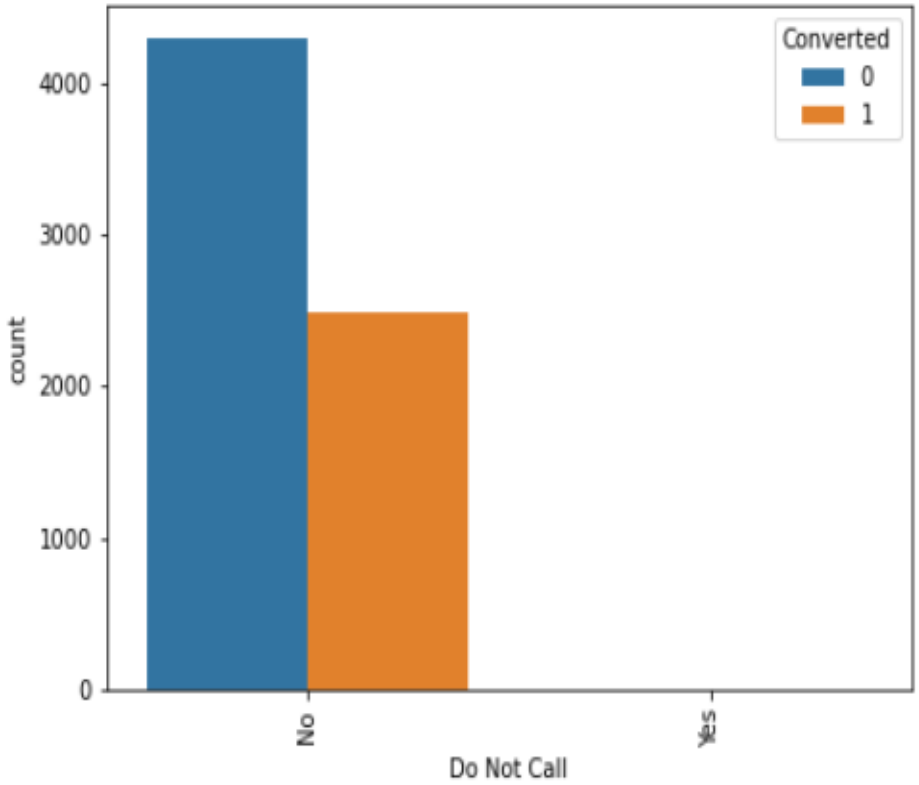
Lead Origin vs Converted



Exploratory Data Analysis

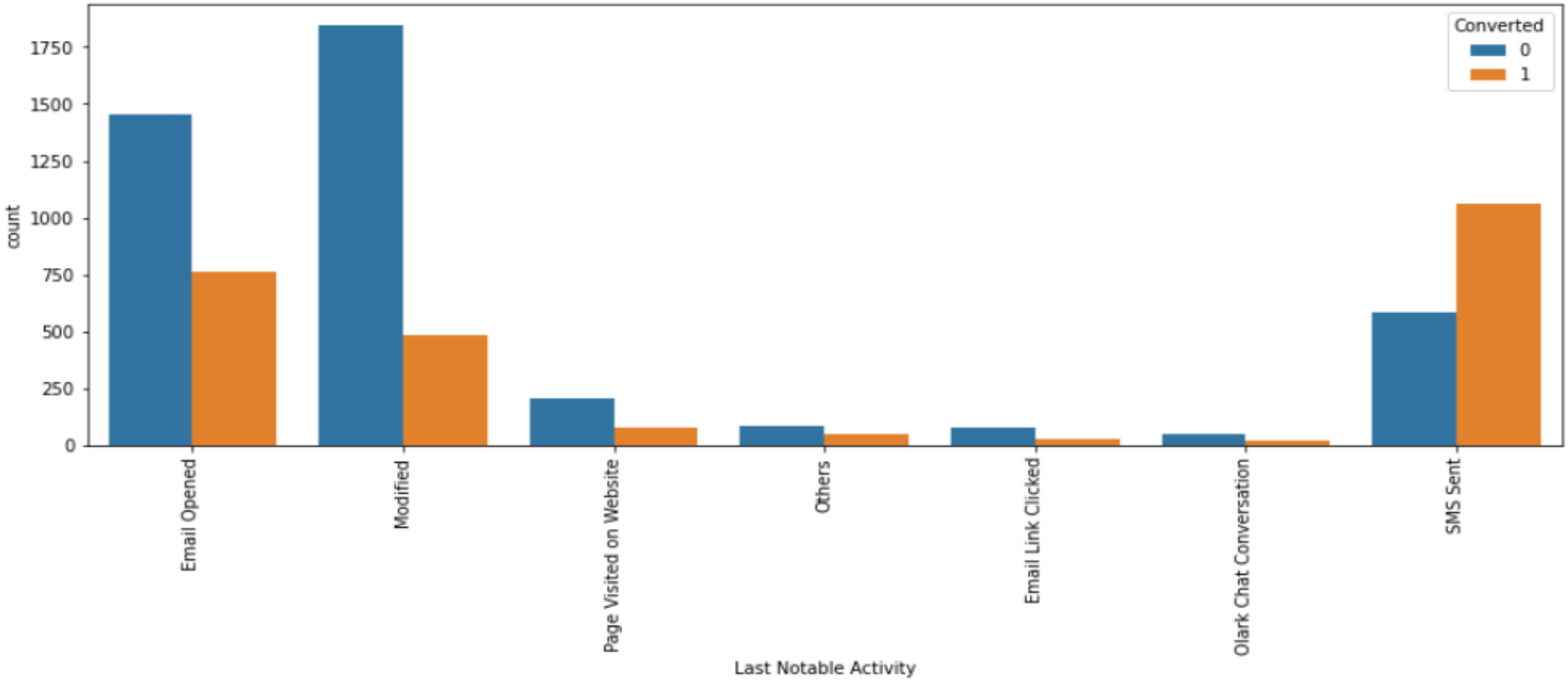
Do Not Email and Do Not Call vs Converted





Exploratory Data Analysis

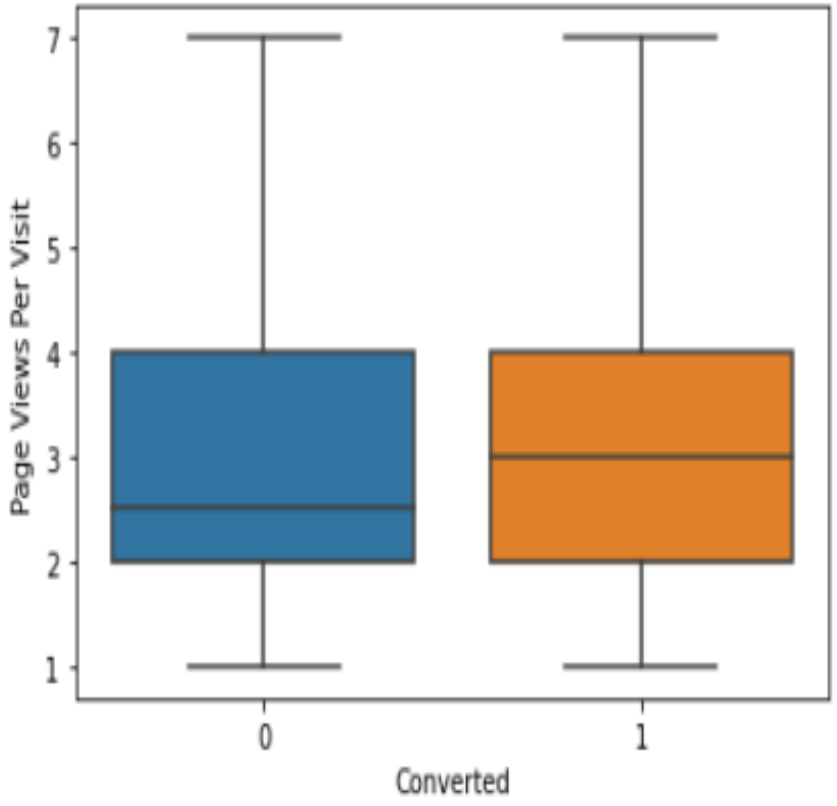
Last Notable Activity vs Converted



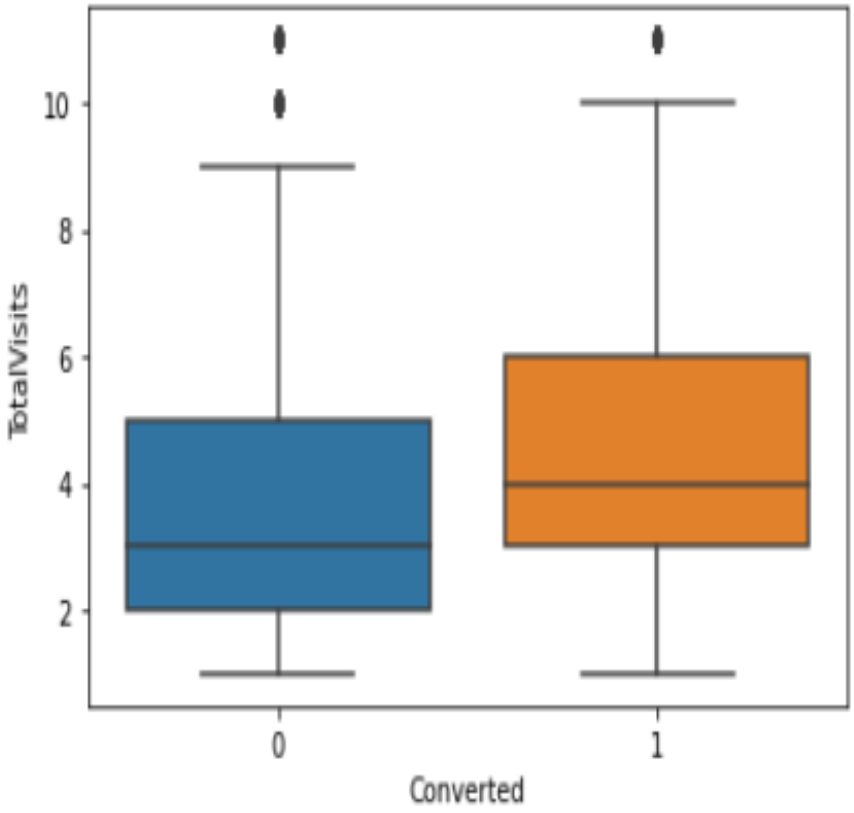
CHECKING CORRELATION Of Numeric Values

Exploratory Data Analysis

Page Views Per Visit vs Converted

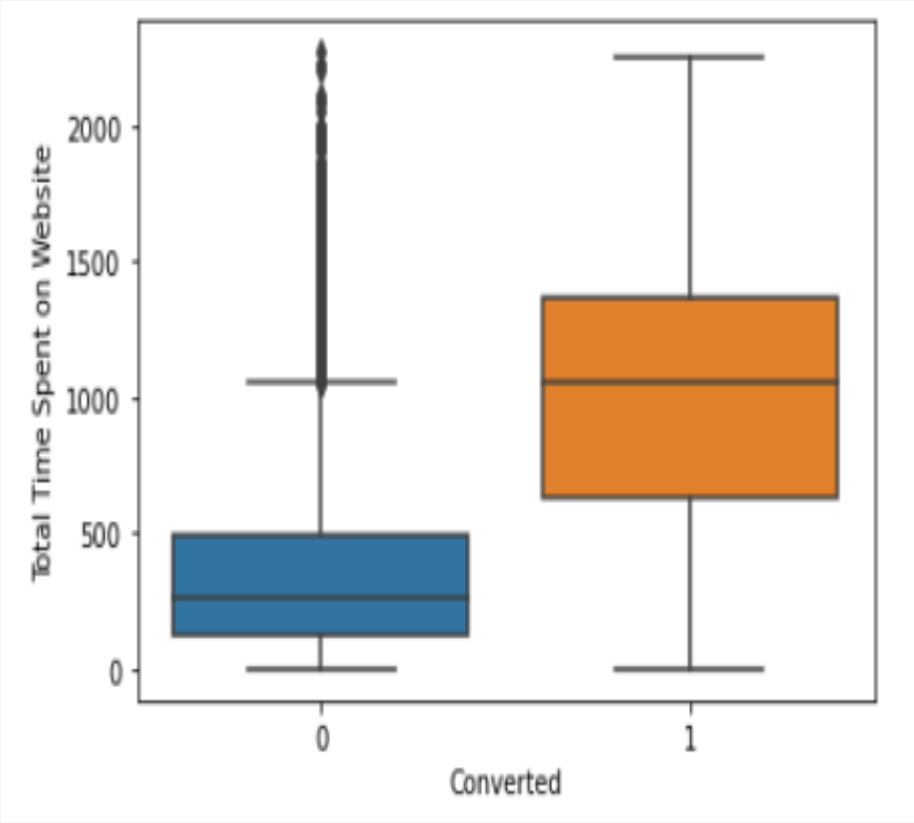


TotalVisits vs Converted

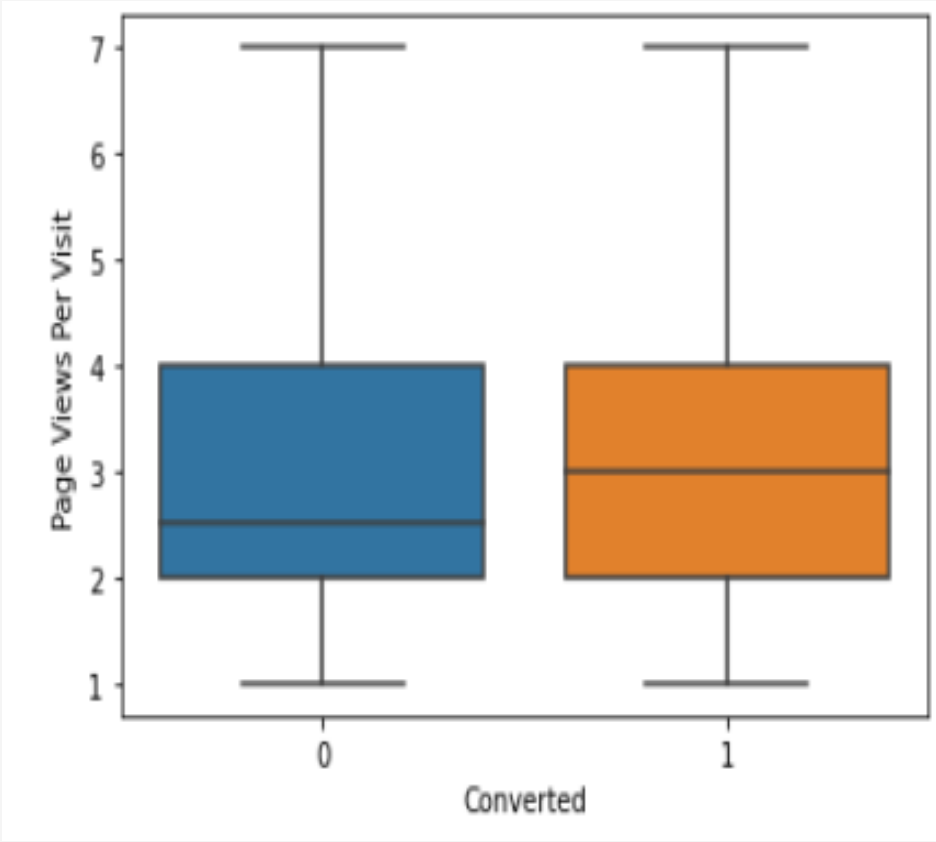


Exploratory Data Analysis

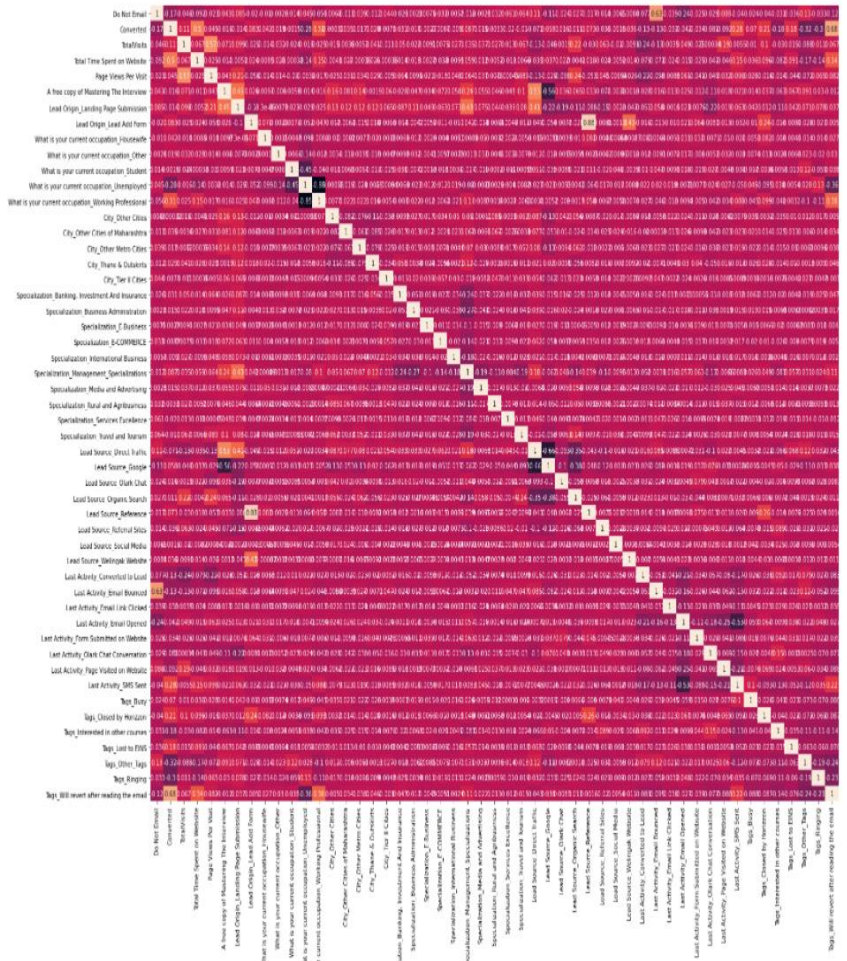
Total Time Spent on Website vs Converted



Page Views Per Visit vs Converted



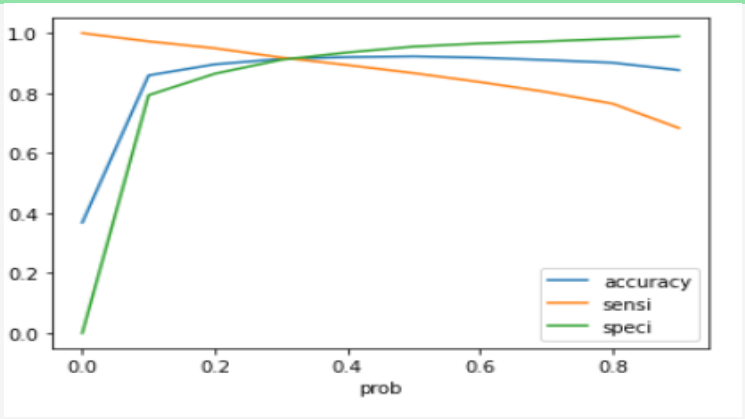
CHECKING CORRELATION Metrix



MODEL BUILDING

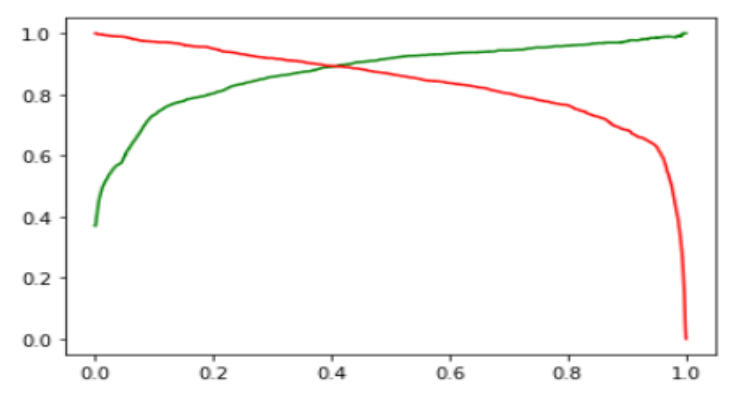
- Splitting into Train and Test Data
- Scale variables in the Train Set
- Build the Mode
- Using stats and RFE to eliminate less relevant variables
- Eliminate variables on high p-value
- Check VIF value for all the existing columns
- Predict using Train set
- Evaluate accuracy and other relevant metrics
- Predict using Test set
- Precision and Recall analysis on Test predictions.

MODEL EVALUATION (TRAIN)



Accuracy. Sensitivity and Specificity

- Accuracy : 91.40%
- Sensitivity : 91.90%
- Specificity : 91.12%



Precision and Recall

- Precision : 85.80%
- Recall : 91.87%

MODEL EVALUATION (TEST)

Accuracy. Sensitivity and Specificity

- Accuracy : 91.50%
- Sensitivity : 91.8%
- Specificity : 91.3%

Precision and Recall

- Precision : 85.78%
- Recall : 91.72%

CONCLUSIONS (APPLICATION DATA) :-

1. Data in the 'Country' column was highly skewed and thus, was dropped from the model.
2. Most of the values in the Specialization columns were missing.
3. The company seems to be doing pretty good in metropolitan areas
4. While we have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.
5. Accuracy, Sensitivity and Specificity values of test set are around 91%, 91% and 91% which are approximately closer to the respective values calculated using trained set.

THANK YOU!

FAIZA AHMAD I SWAPNIL RANJAN

DS C34 31ST JULY 2021