Lead Scoring Case study

Presented by

Sachin Sen

Anil Kumar

Pawan Kumar Thakur

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. A typical lead conversion process can be represented using the following funnel:

As you can see, there are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom. In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.

X Education has appointed you to help them 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 a higher lead score have a higher conversion chance and the customers with a 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%.

Goals of the Case Study

There are quite a few goals for this case study:

1.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.

2.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. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

Approach

➢ Source the data for Analysis
➢ Reading and understanding the data and problem statement
➢ Data cleaning
➢ EDA
➢ Feature scaling
➢ Feature selection
➢ Splitting data into train and test
➢ Prepare the data for Modeling
➢ RFE, VIP and Pvalue test
➢ Model Building
➢ ROC and AUC
➢ Model Evaluation
➢ Making prediction on test data
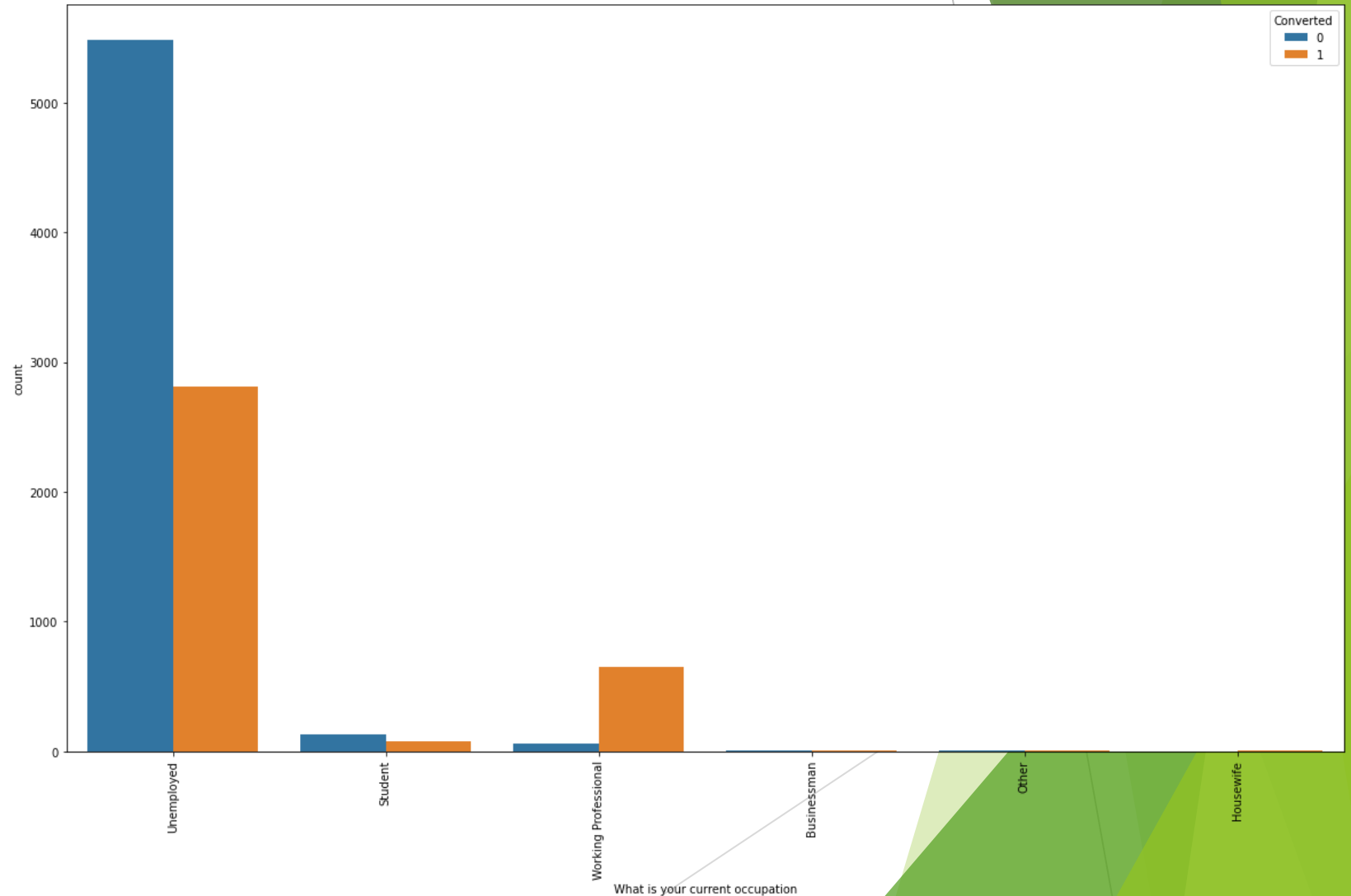
Univariate Analysis of Specialization column
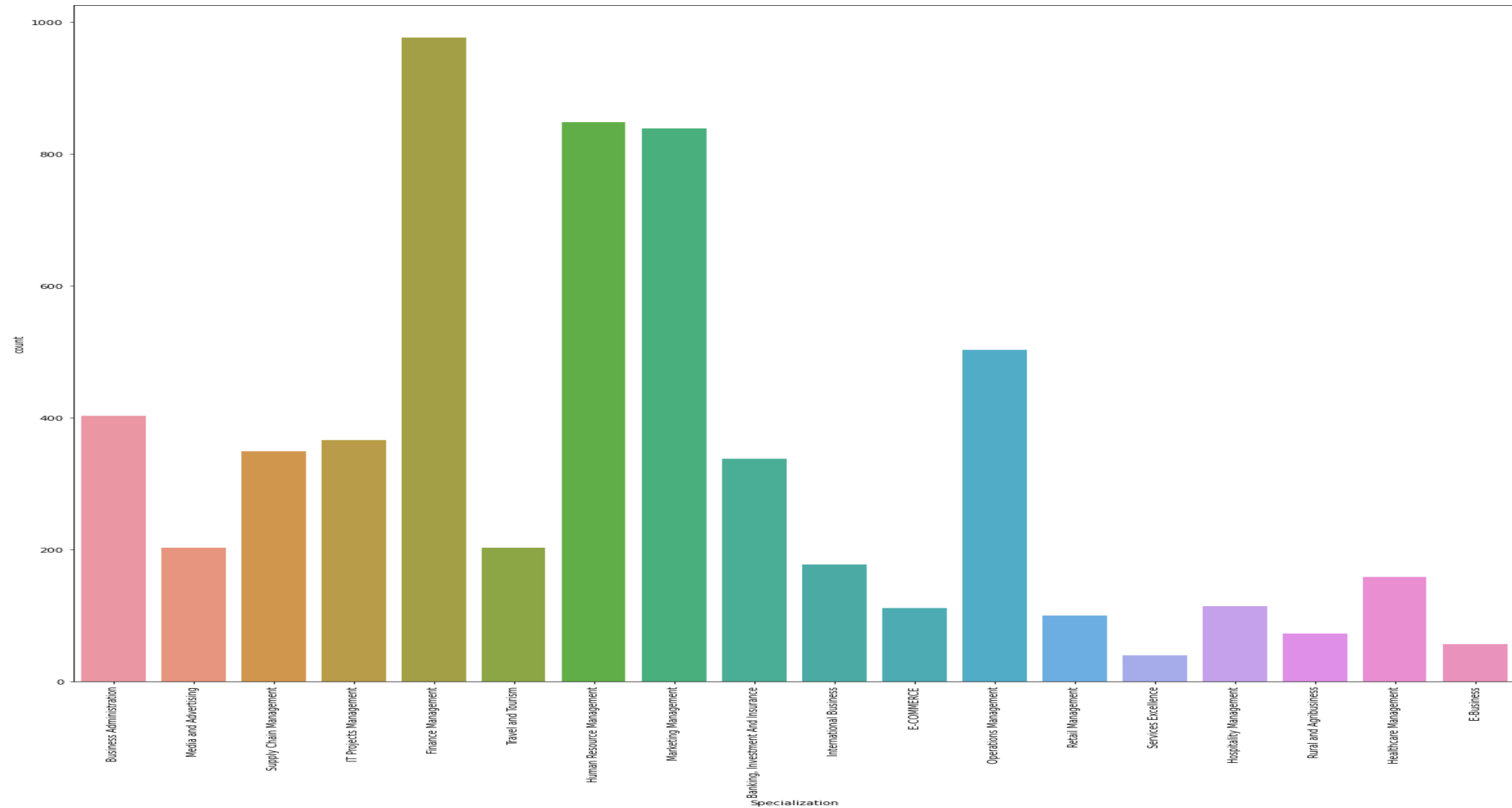
Bi variate Analysis of Specialization column with convert rate
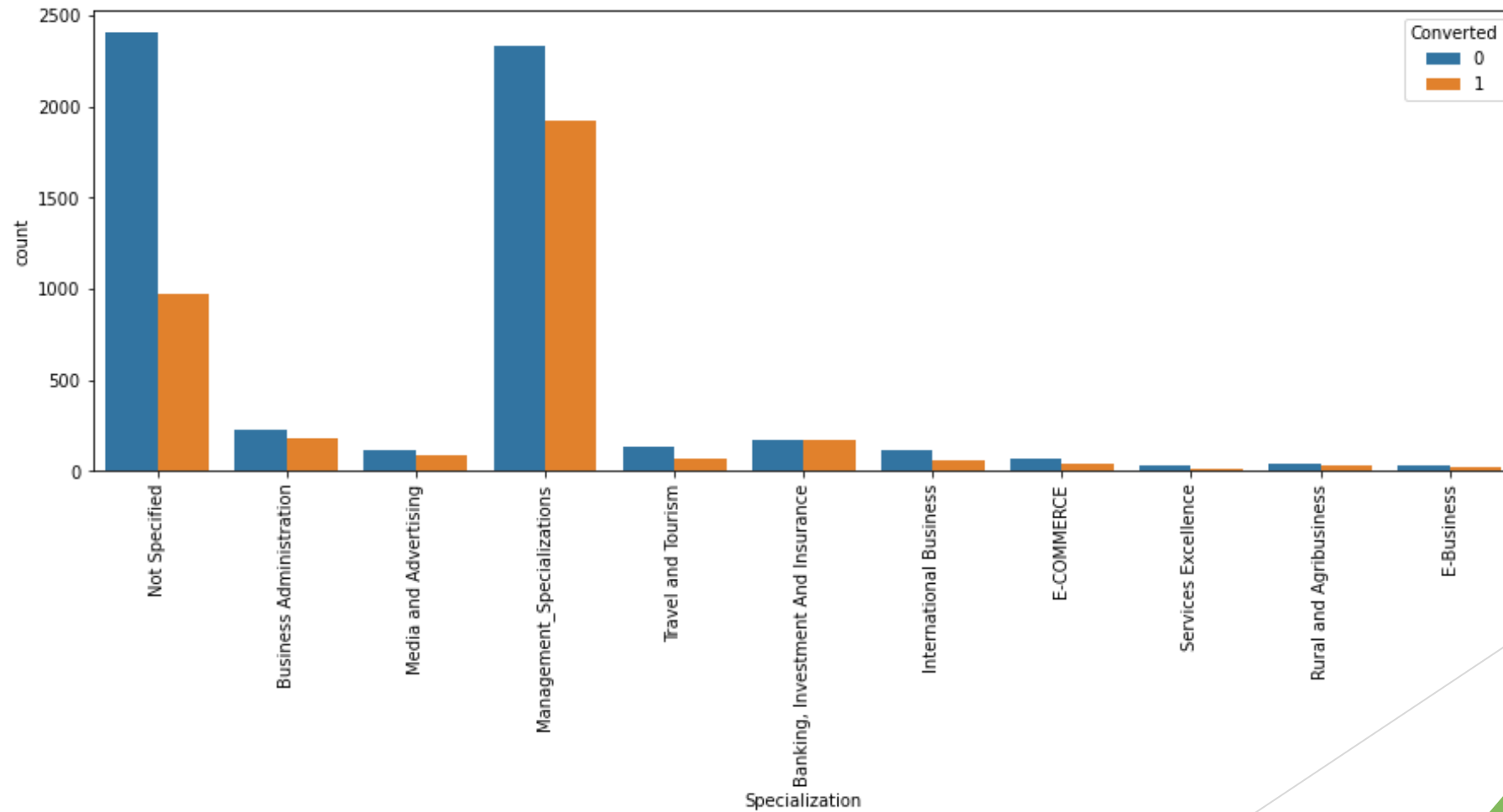
Bi variate Analysis of "What is your current occupation" column with convert rate

Observation
1. working professional are highly joining the course
2. the unemployed are high in number

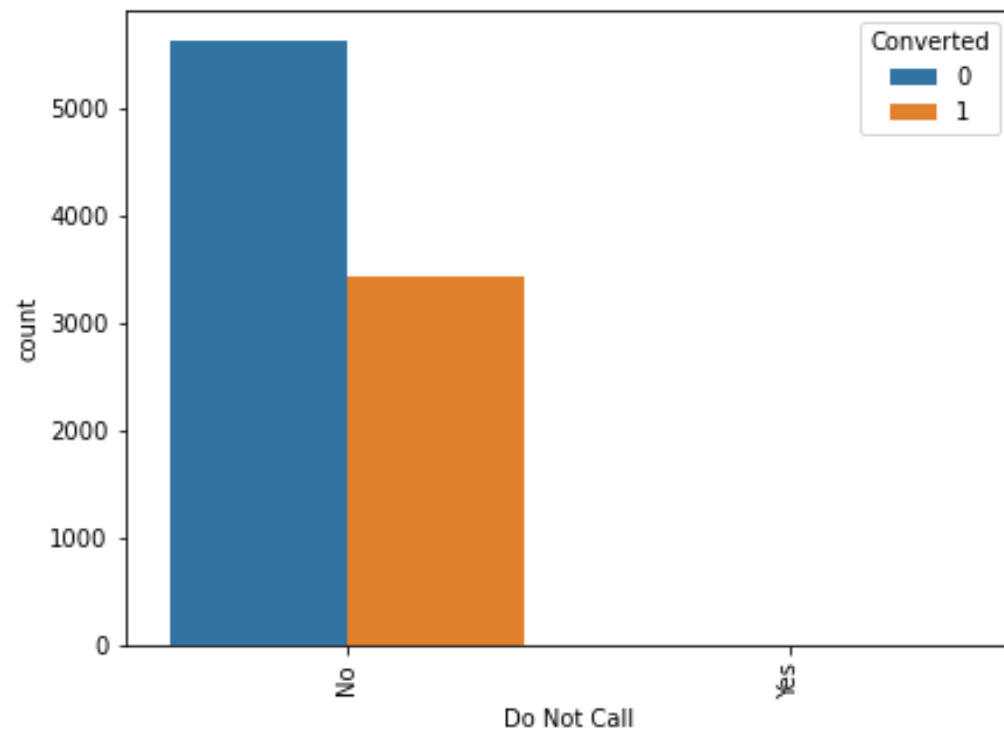Bi variate Analysis of "Lead Origin" column with convert rate

Observation

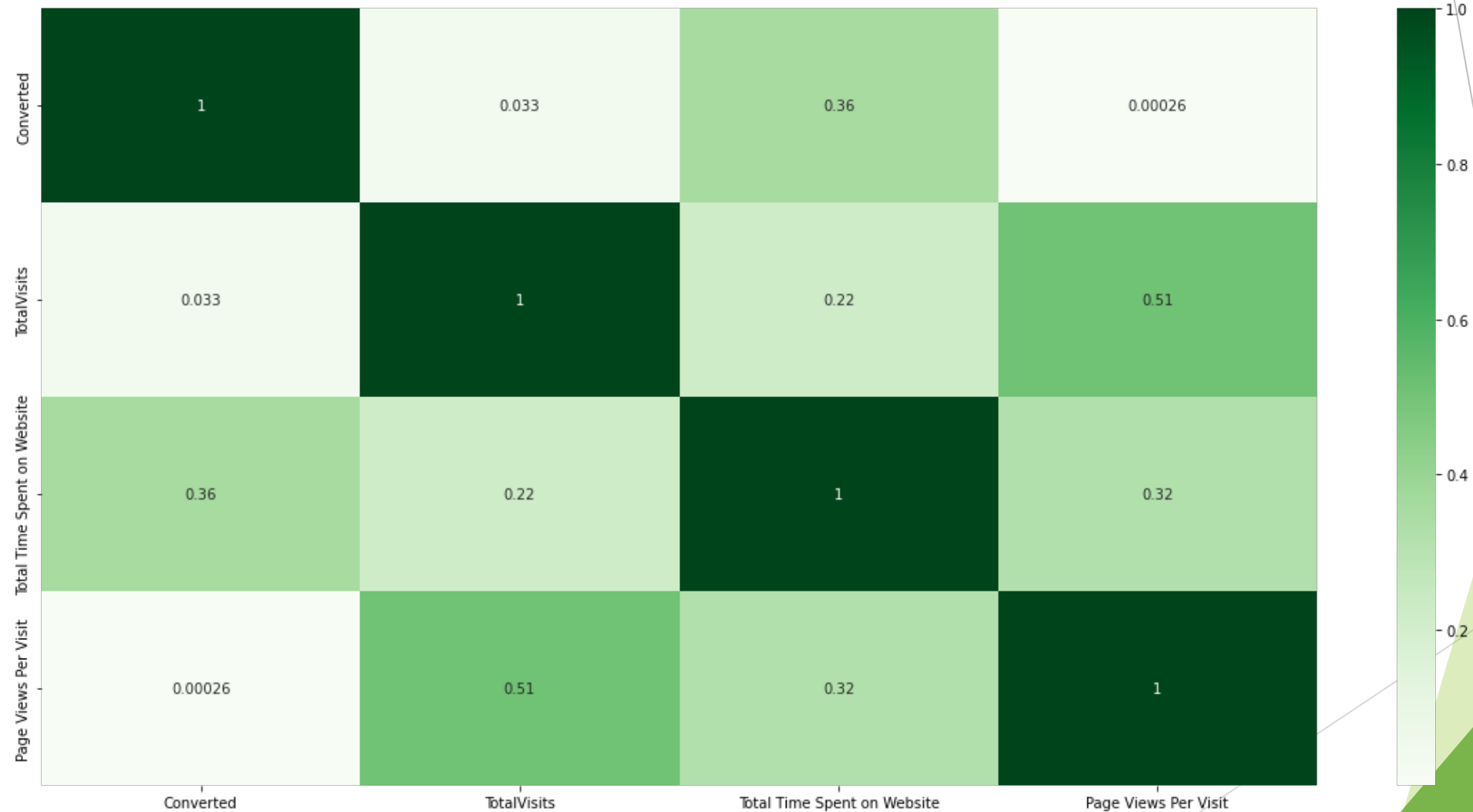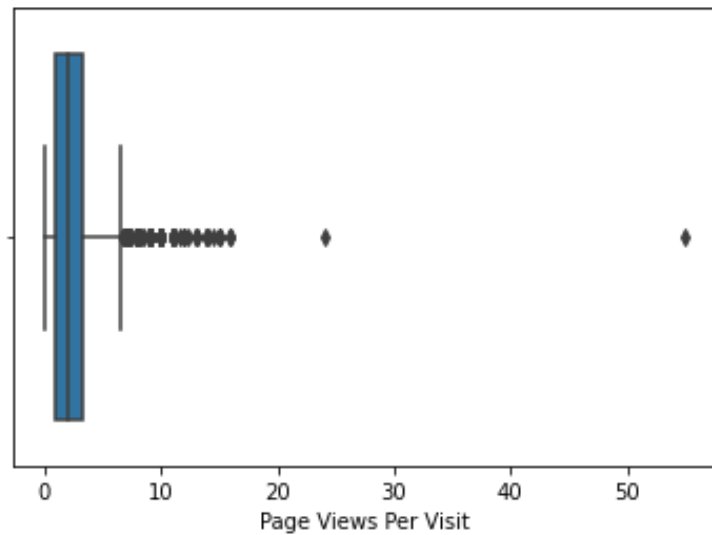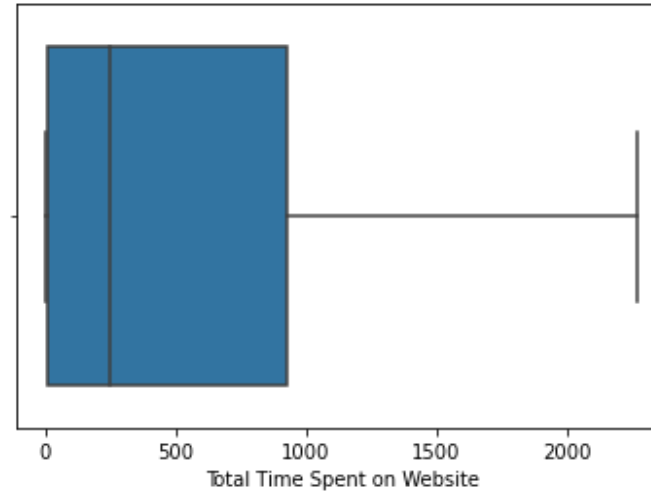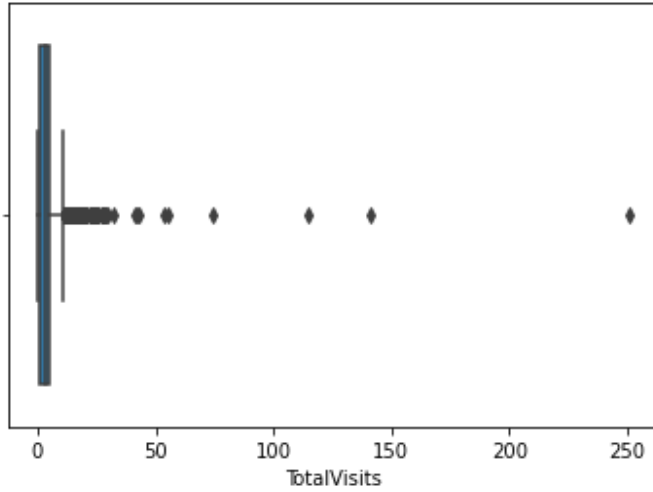1. maximum leads are coming from landing page and API and conversion rate of landing page is also high
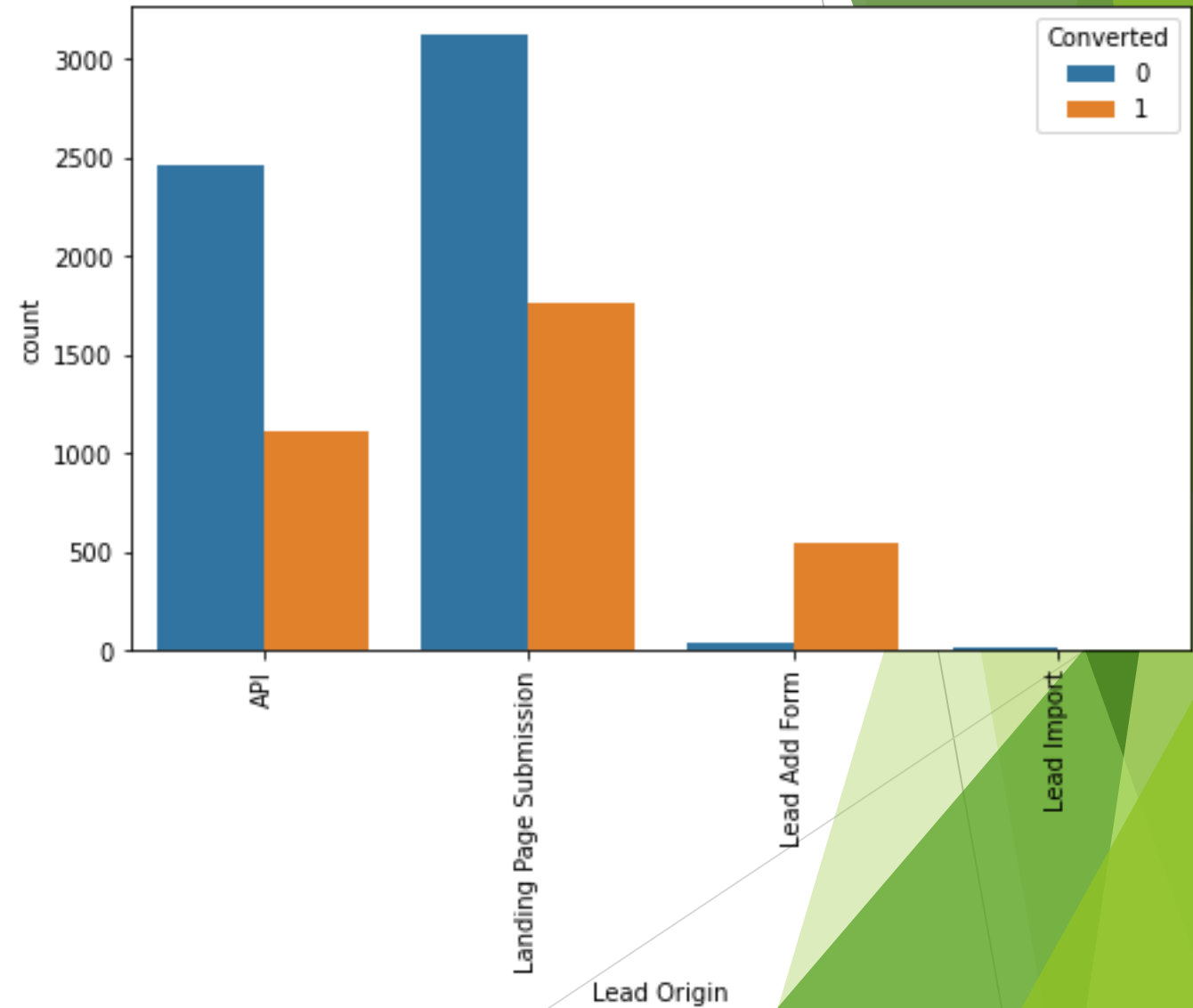
Comparison between "Do not Call" and "Do not Email"
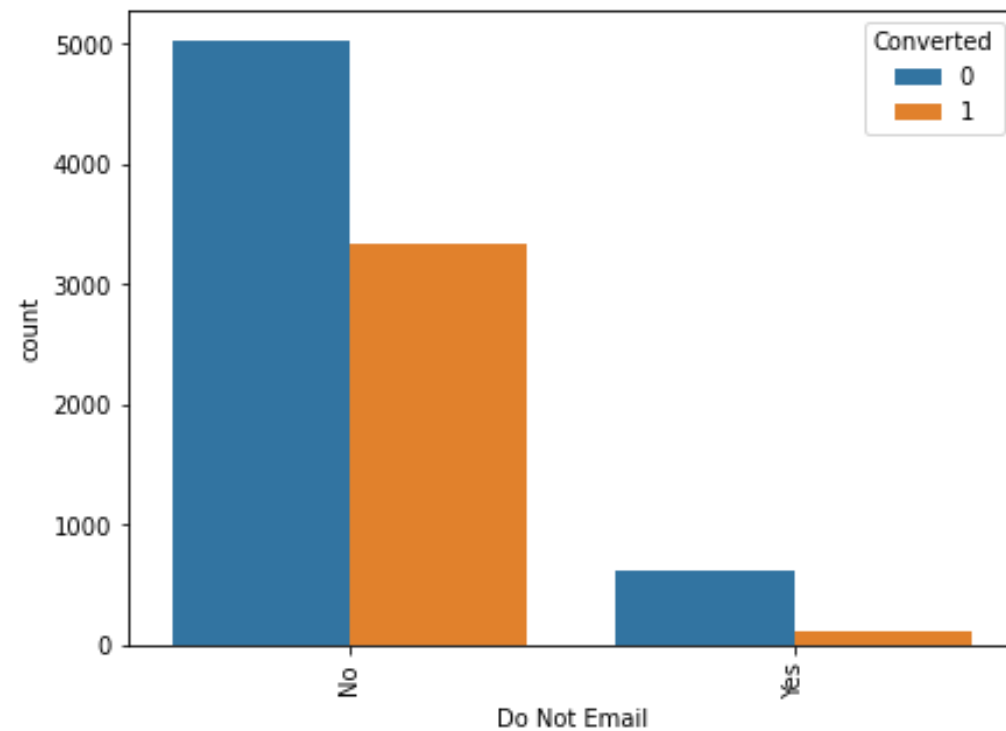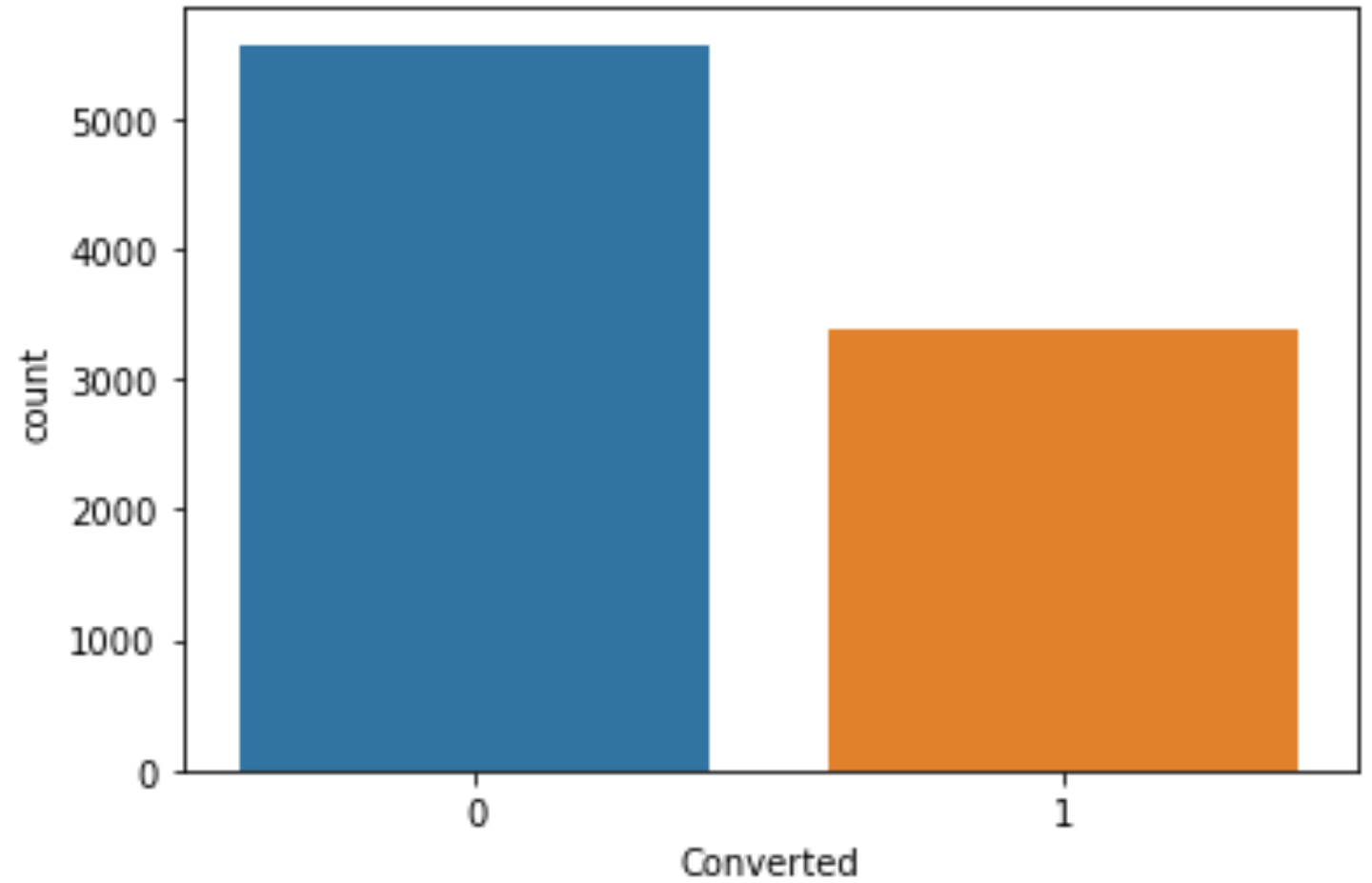
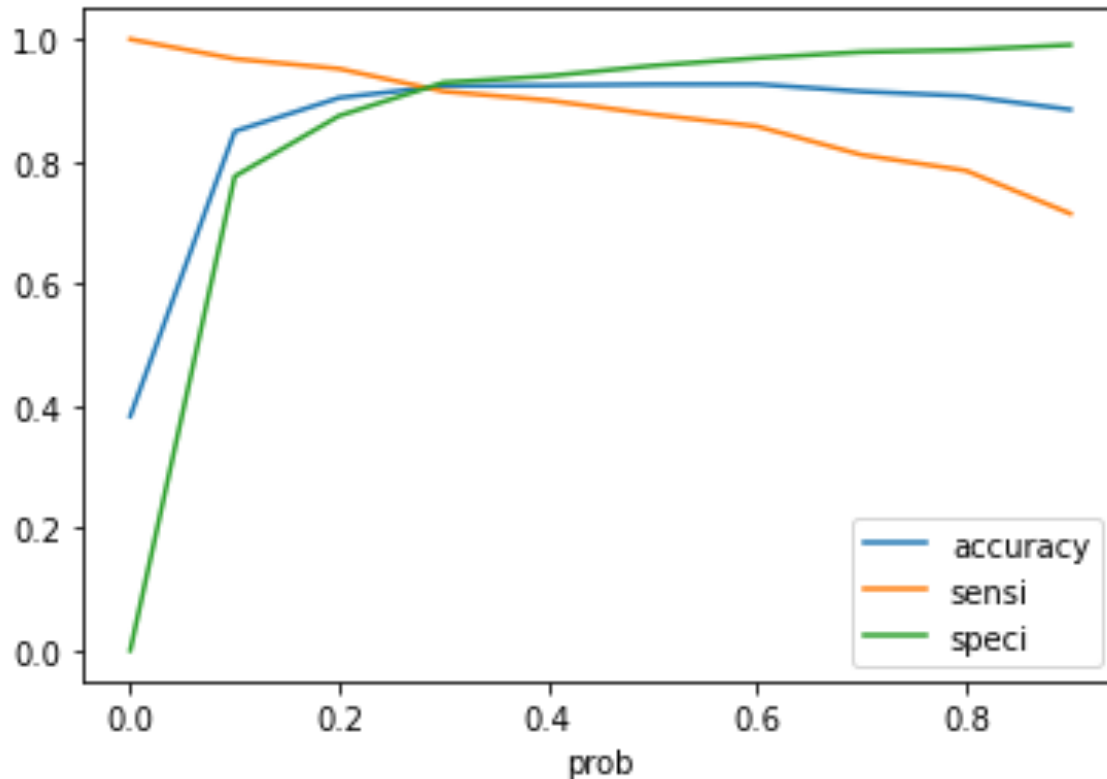# Correlation Metrics – Multivariate Analysis

Outliers

Observation
1. Total visit is having outlier
2. total time spent on website is ok
3. page view per visit is having outlier
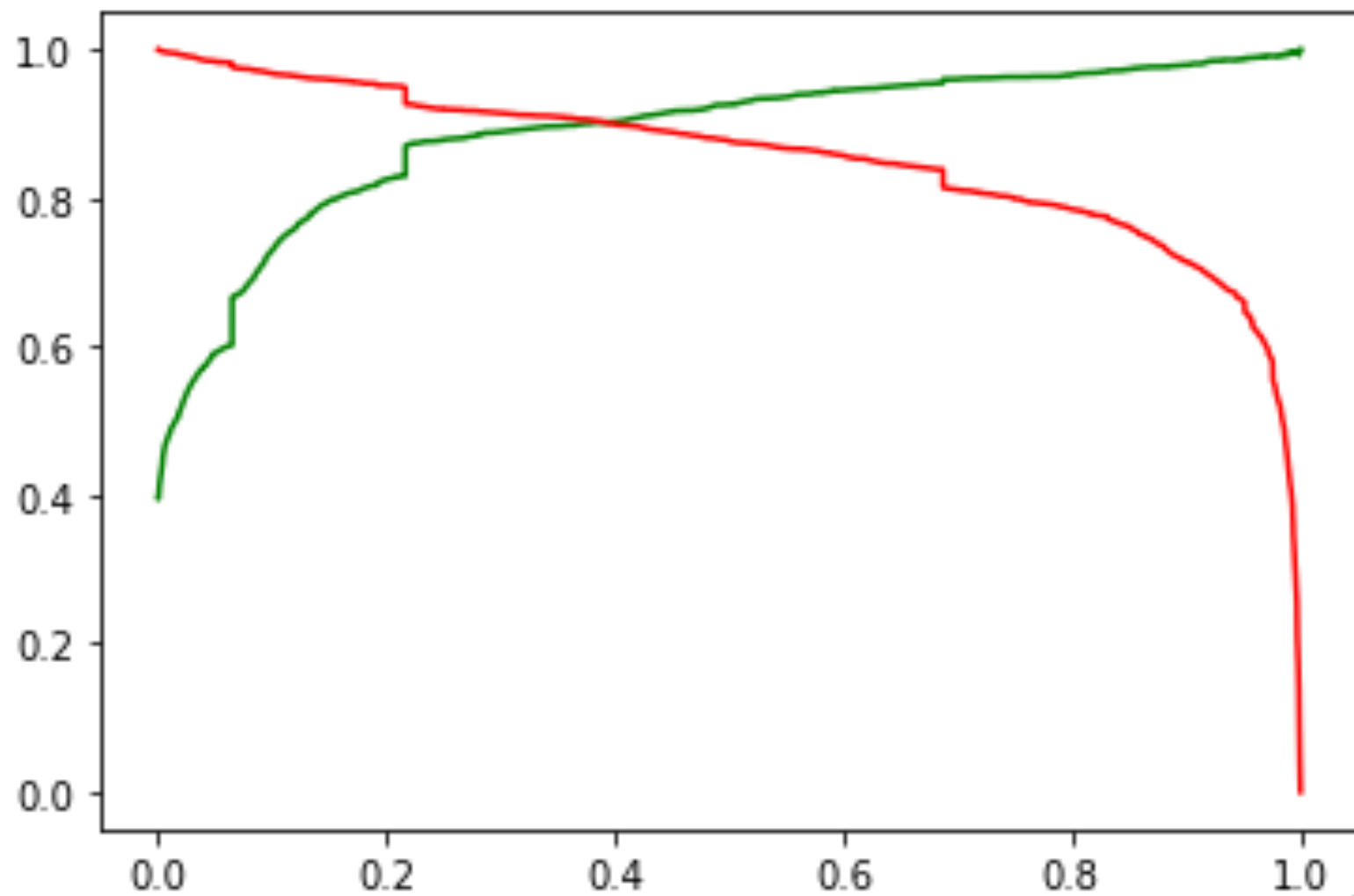
Conversion Rate is Low

Model Evaluation with Accuracy , Sensitivity and Sepcificity

Observation

1. Accuracy score : 92%

2. sensitivity : 91 %

3. specificity : 92%

Precision - 88%
Recall - 91%

Model Evaluation on Test data with Accuracy , Sensitivity and Sepcificity

Observation On test set

1. accuracy score : 93%
2. sensitivity score : 92%
3. specificity score : 93%