Lead Score Case Study

Problem Statement:

Here we are provided with a dataset of X education Company which sells online courses to industry professionals. The person who is willing to do a course lands on their website and enquires to a particular course. The advertise of course to spread it worldwide is done through several website's and search engines like Google, Olark chats , Bing etc. Once interested people lands on their website the look for course academy, will try to gather more information about the course which they are interested in. Once they fill the form for enquire they provide with the email address as well as the mobile number of the person interested. Once they provide the phone number or email id they are classified as a lead to the X Education Company. The company gets many leads through pass referrals also. Once the interested person is classified as a Lead the sales team starts making calls, writing emails etc is done. Through in this process not all leads get converted some do and some don't. The typical conversion ratio here is 30 % gets converted. The main goal of this case study is to identify the "Hot Leads" which are actually going to be converted into potential leads. If this leads are classified correctly the lead conversion rate will increase as the sales team will more focus on the potential leads and will focus more to communicate with them rather than calling everyone. The main problem given here is to identify the leads that are most likely to convert into paying customers.

While performing data inspection step we found that there are some outliers in the numeric columns i.e. Total Visits , Total time spend on website and Page views per visit
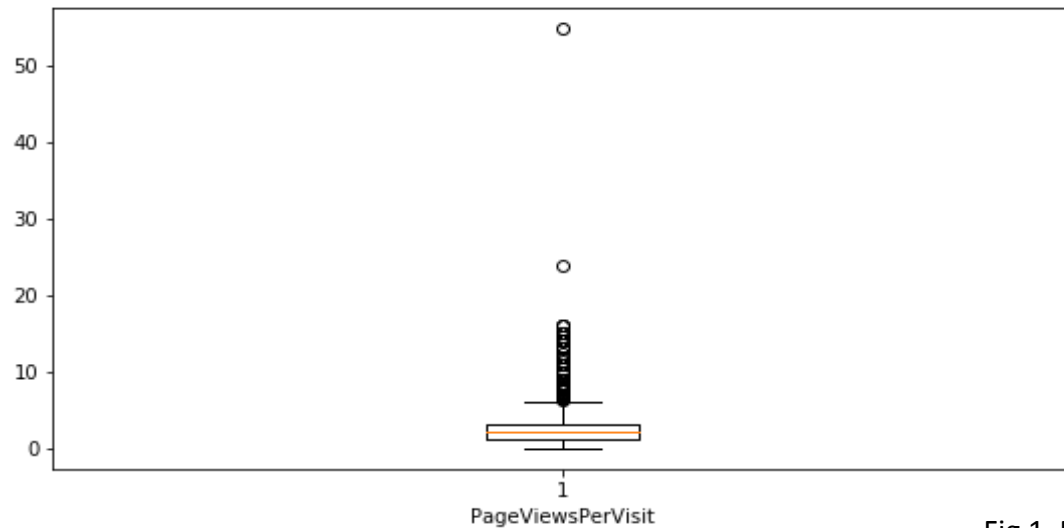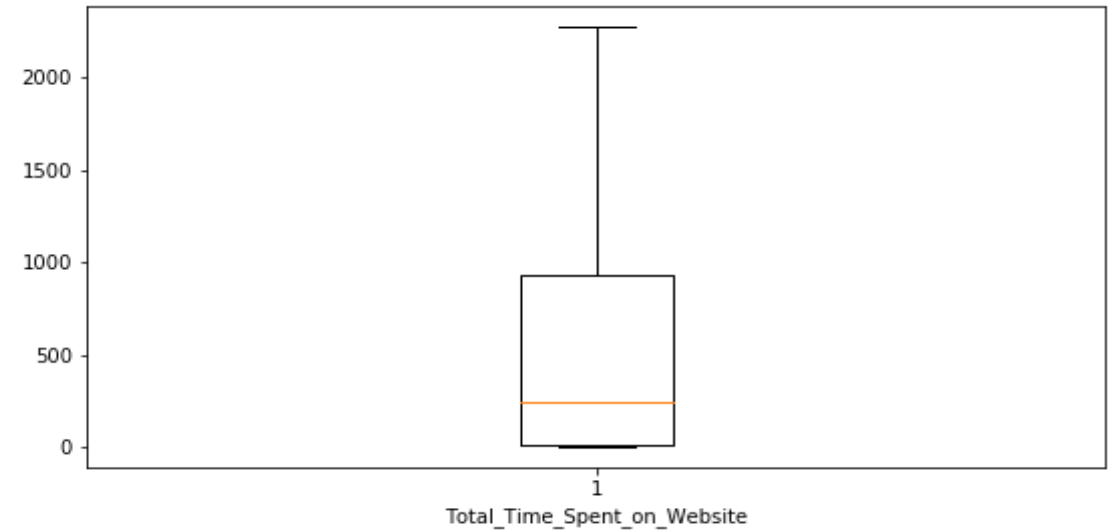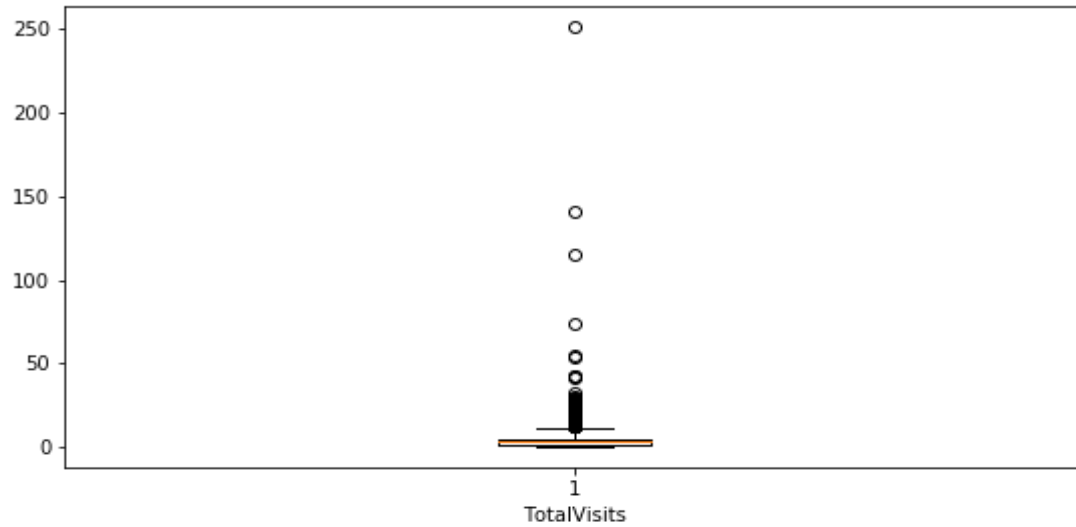

Fig 1. Data before Outlier

After Outlier Detections, we have performed outlier treatment on data using IQR Rule and dropped the outliers from the data .
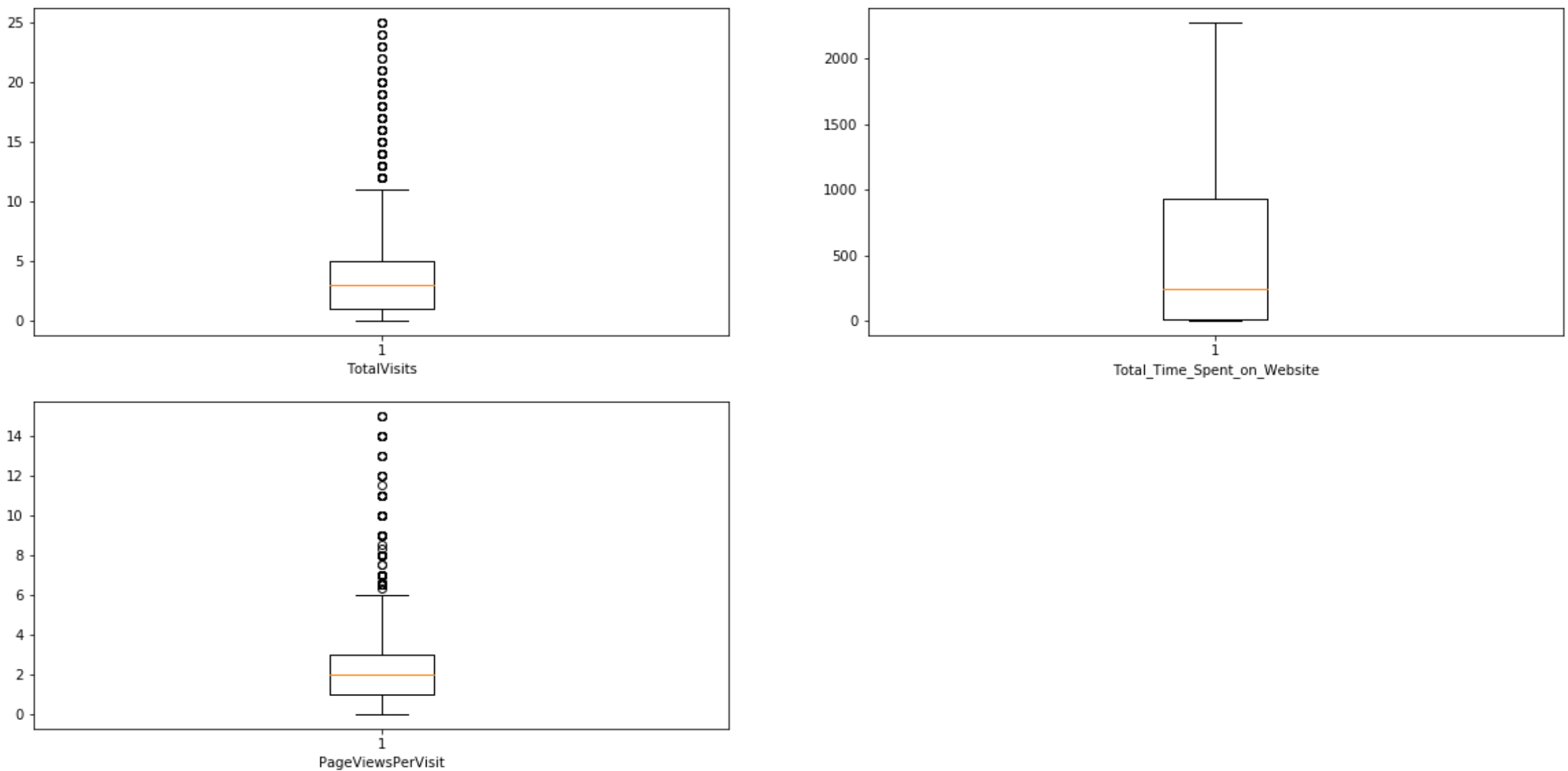

Fig 2. After Outlier Treatment

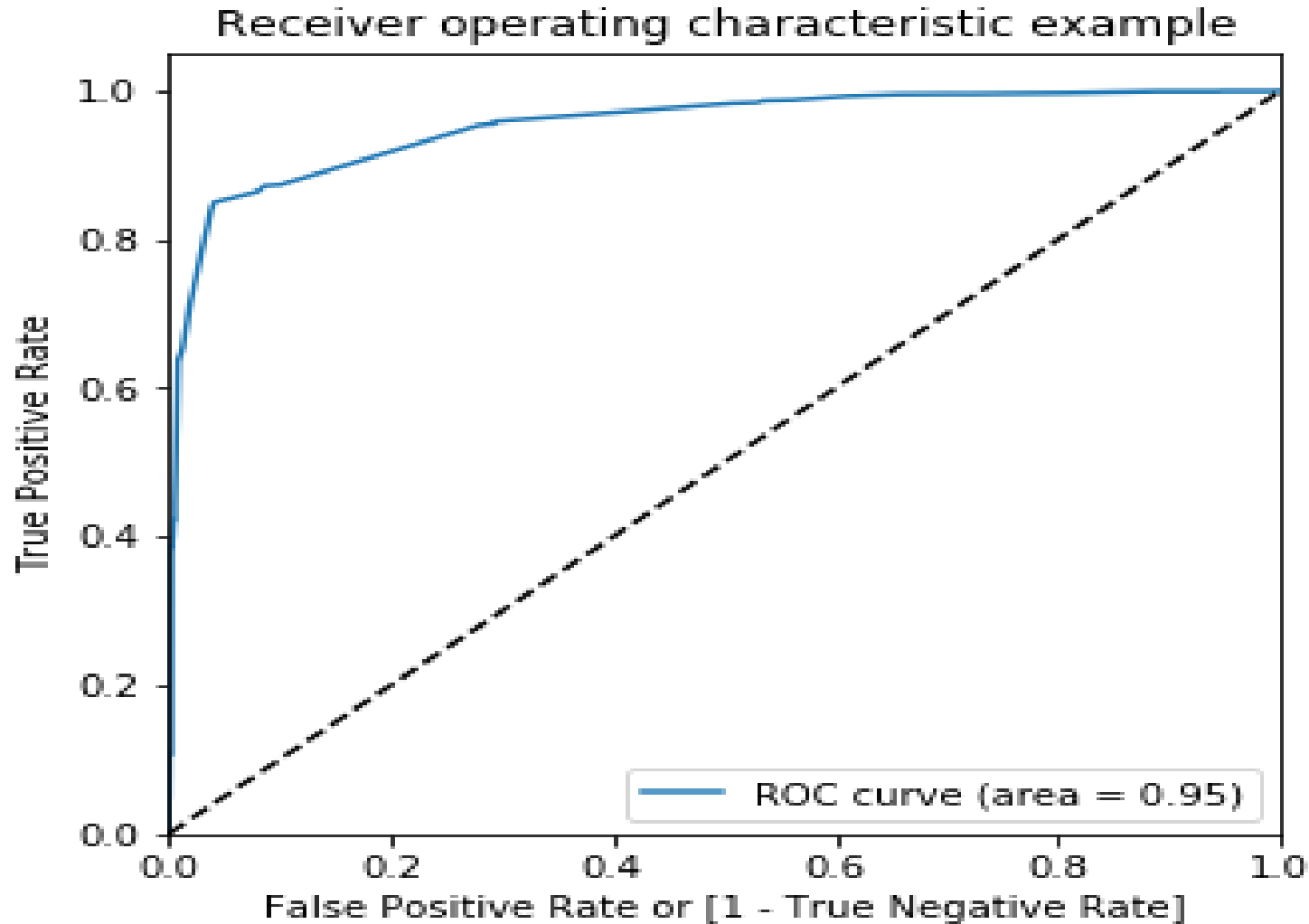# ROC Curve : Receiver Operating Characteristics

Fig 3. ROC Curve

ROC Curve :

- ROC Curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.
- Higher the AUC means that the model is capable of distinguishing between different classes.
- Here we get 0.95 AUC (Area under the Curve), higher the AUC value better the model.
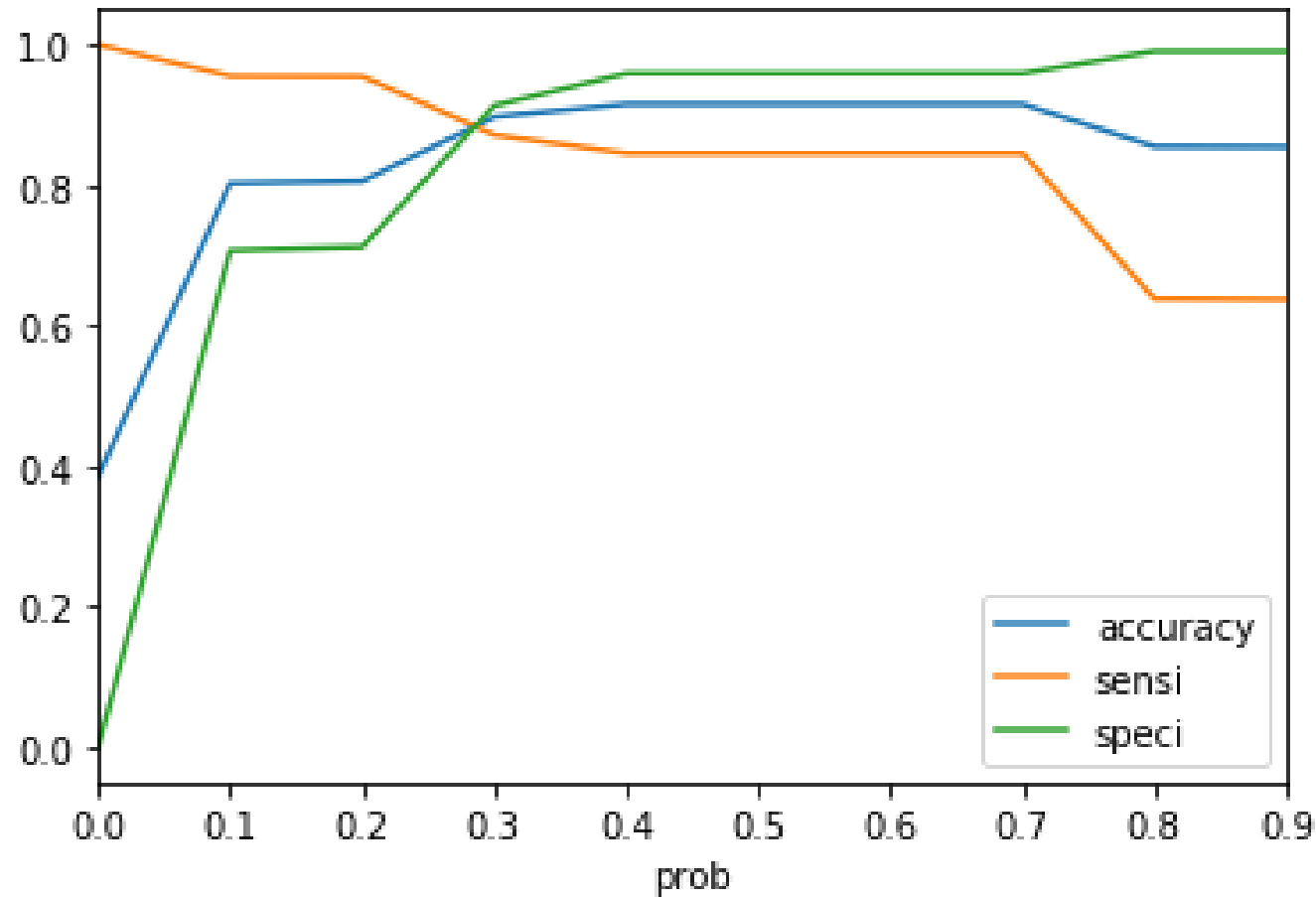
Optimal Cut off

Fig 4. Finding Optimal Cutoff

Finding Optimal Cut Off :

- We plot the Sensitivity , Specificity and Accuracy to identify the cut off value.
- We identify cut off as where the three curve intersect each other.
- Here we can see that the three curve intersect at 0.3 probability value hence we take 0.3 as a cut off
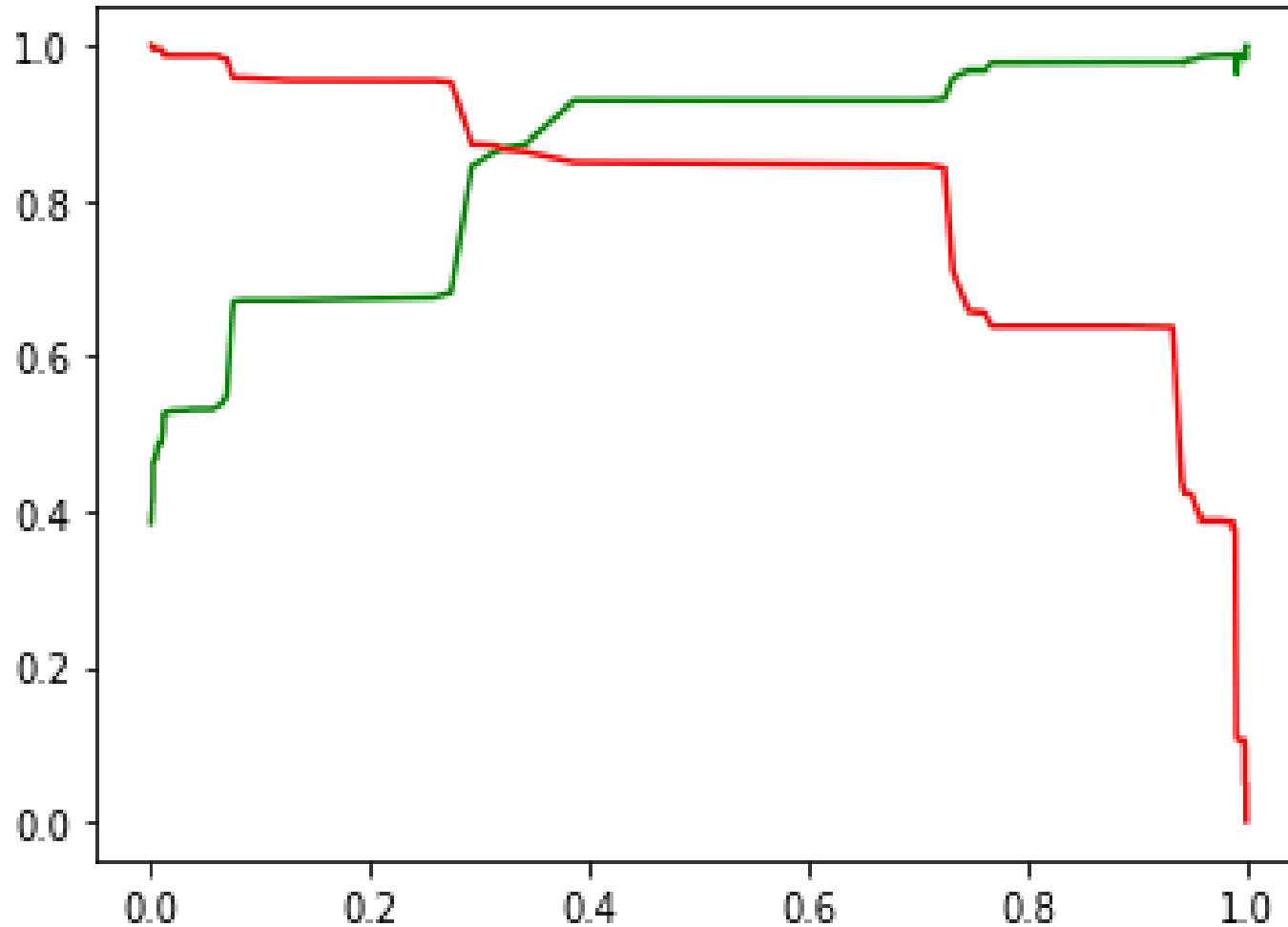
Precision & Recall Trade off

Fig 5. Precision & Recall Trade off

Precision & Recall Trade off :

- While recall expresses the ability to find all relevant instances in a dataset, precision expresses the proportion of the data points our model says was relevant actually were relevant.
- When we increase the **recall**, we decrease the **precision**

RECOMMENDATIONS:

- As we can see that the model is performing good on test dataset s the accuracy of the train dataset and test dataset is almost similar.

- Here are the most 3 significant features are Last Activity – SMS Sent, Tags – Ringing and Last Notable Activity Modified can contribute more to Lead conversion.