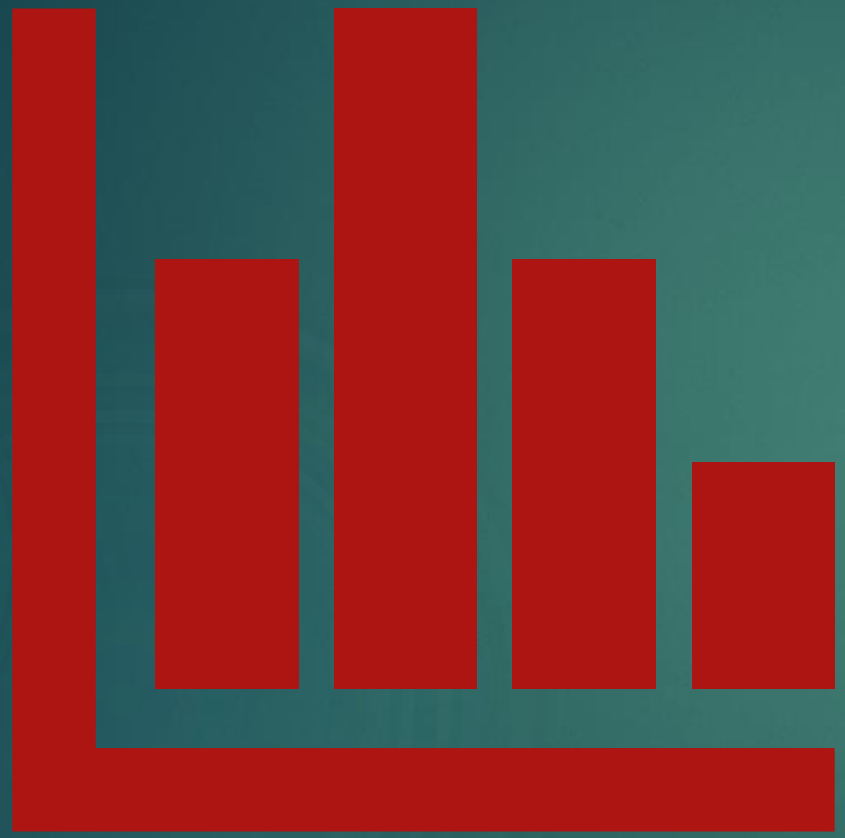


Leads Scoring Case Study

Yash Kumar Singh

Problemstatement

Model should be created in such a way that the customers with high lead score have higher conversion chance and low lead score have lower conversion chance. The ballpark of the target lead conversion rate is around 80%.

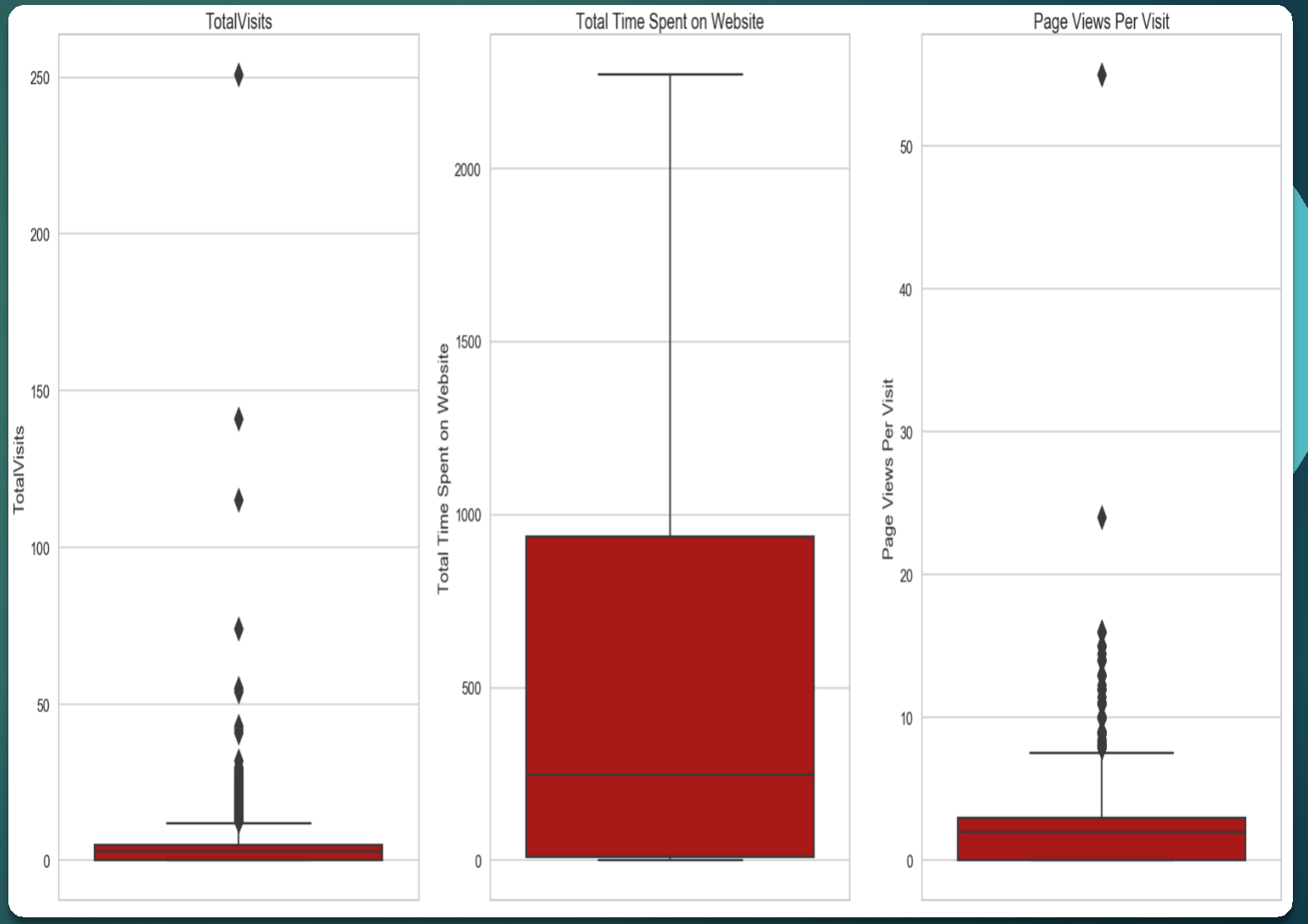
Also the model should be able to adapt if the company’s target and requirements change

Approach of the analysis

I. We started our analysis with our cleaned dataset by converting all the binary variables to '0' and '1' and multiple categories into dummy variables.

II. Next, we have checked the outliers of the dataset. The visualization of those outliers we can see on the graph attached .

III. Outliers in logistic regression model is very sensitive hence we need to deal with it so This can be achieved by creating bins.



Correlation

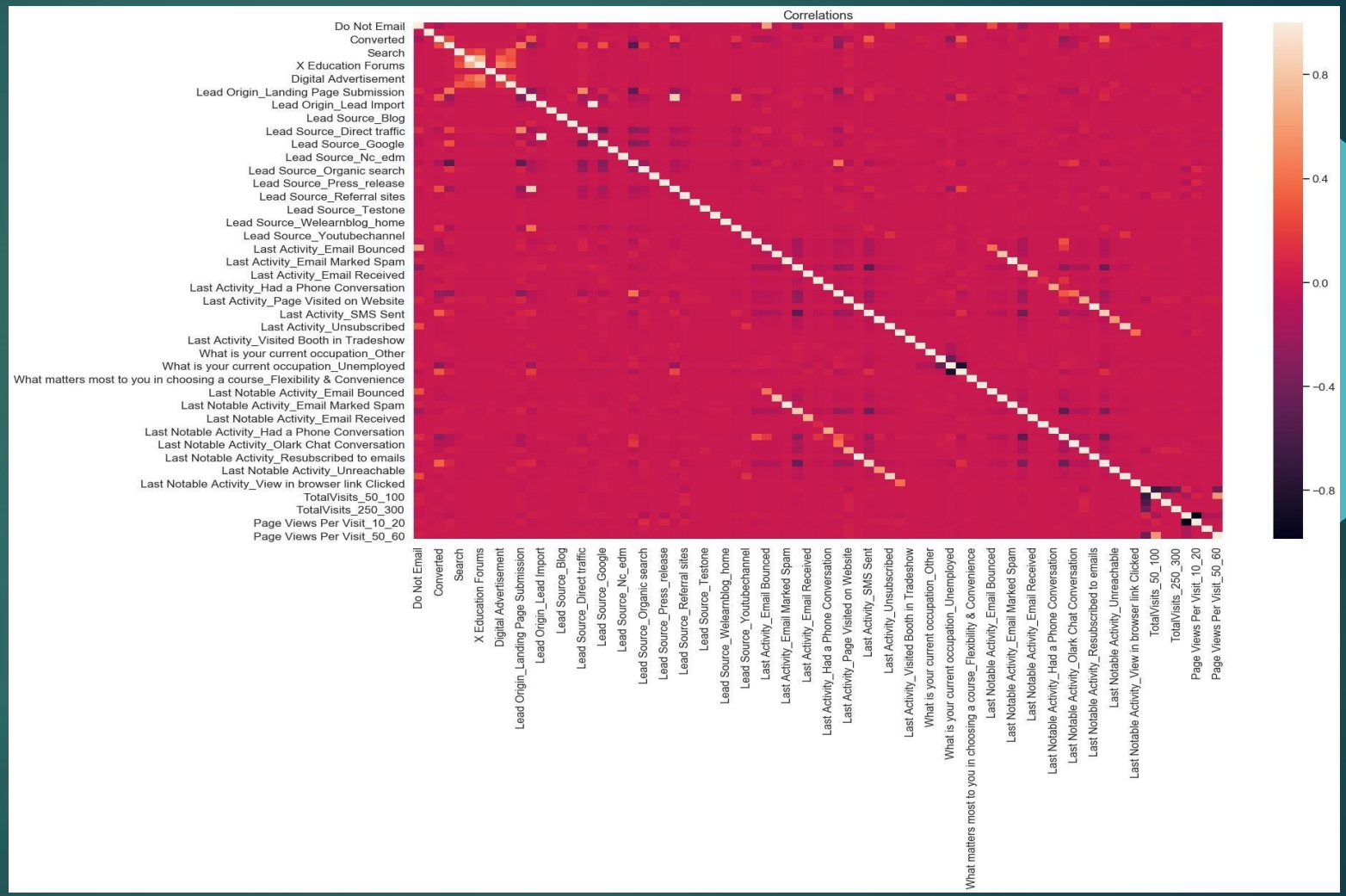
After fixing the outliers we are now doing the data preparation.

a) We have split the dataset into train and test set and do standardization on the features.

b) Standardization is done in order to keep all the variables in same scale ,

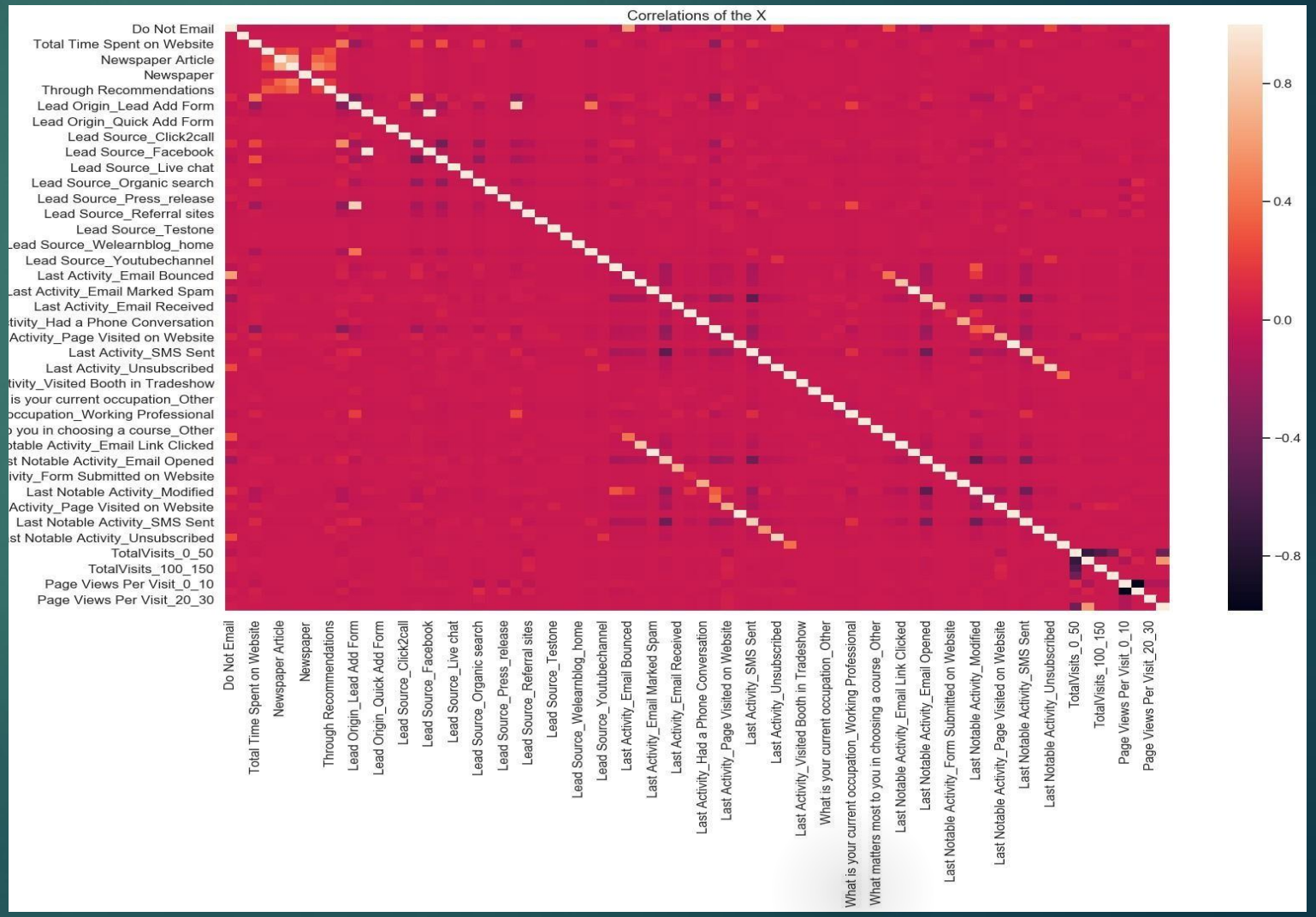
c) Checked the correlation of the dataset. The Attached heatmap is showing the correlation of all features present in the dataset.

d) We observed that There are some high correlations in the heatmap which we dropped.



Correlation

- We have plotted this again to ensure whether those were dropped or not.
  As from the plot we cannot clearly understand anything so we will check them after
- creating a model.



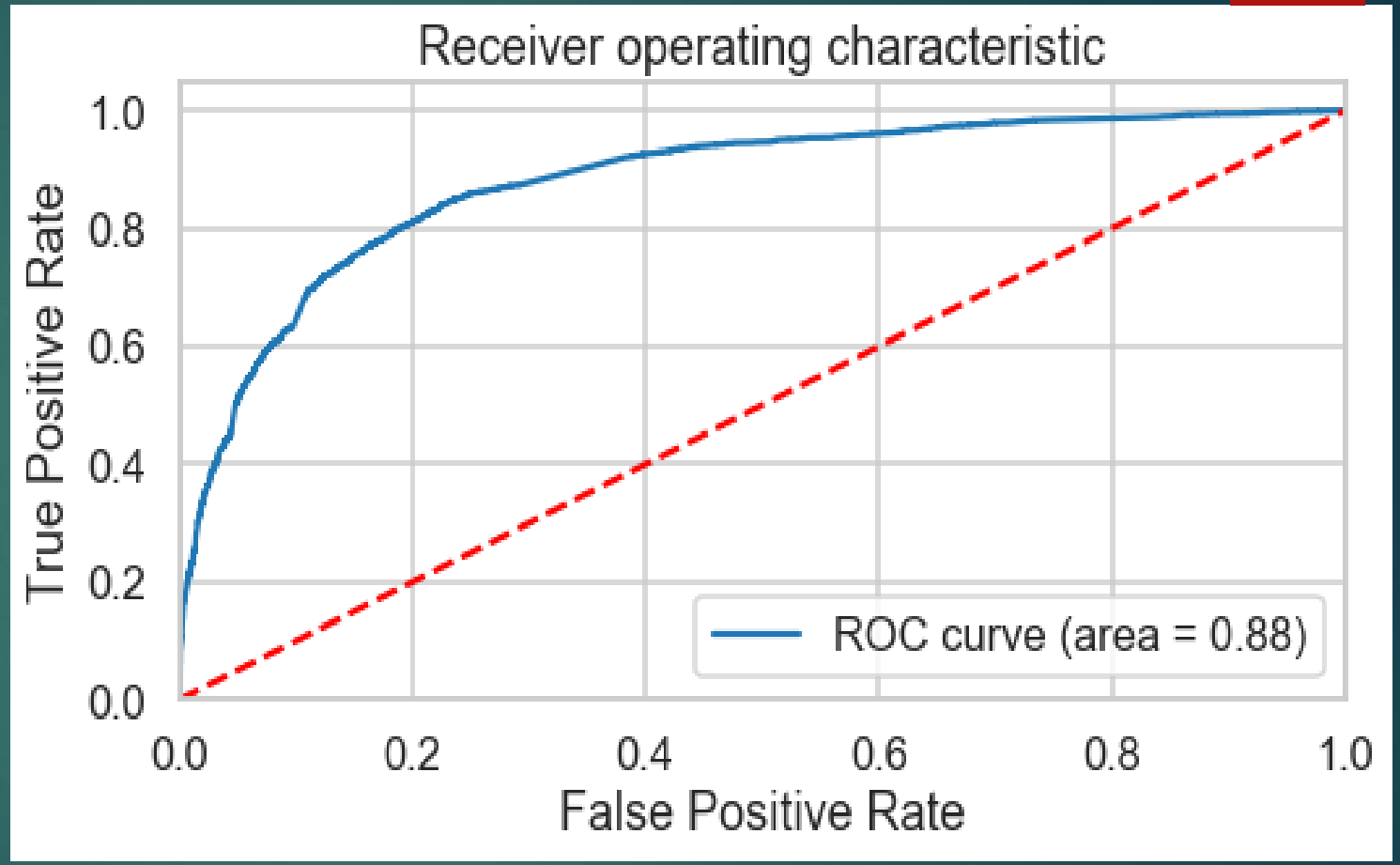
Building a Model – RFE 1

- We built a model which has all the features included and found there were many insignificant variables present in our model.
- We should drop them, but we can't do it one by one as it is time consuming and not an efficient way to do so.
- Hence, we started with RFE method to deduct those insignificant variables. We choose with RFE count 19 and 15.
- We did two rfe count because we want to find out our final model stability.
- We started creating our model with rfe count 19 and started dropping variables one by one until we reach the point where the model is having all significant variables and low VIF values.
- Now we evaluated our model by first predicting it. We created new dataset with original converted values and the prediction values.

Evaluating the model

- After building the final model making prediction on it(on train set), we created ROC curve to find the model stability with auc score(area under the curve) As we can see from the graph plotted on the right side, the area score we got is 0.88 which
- is a considered a great score.

  And our graph is leaning towards the left side of the border which means we have good accuracy.

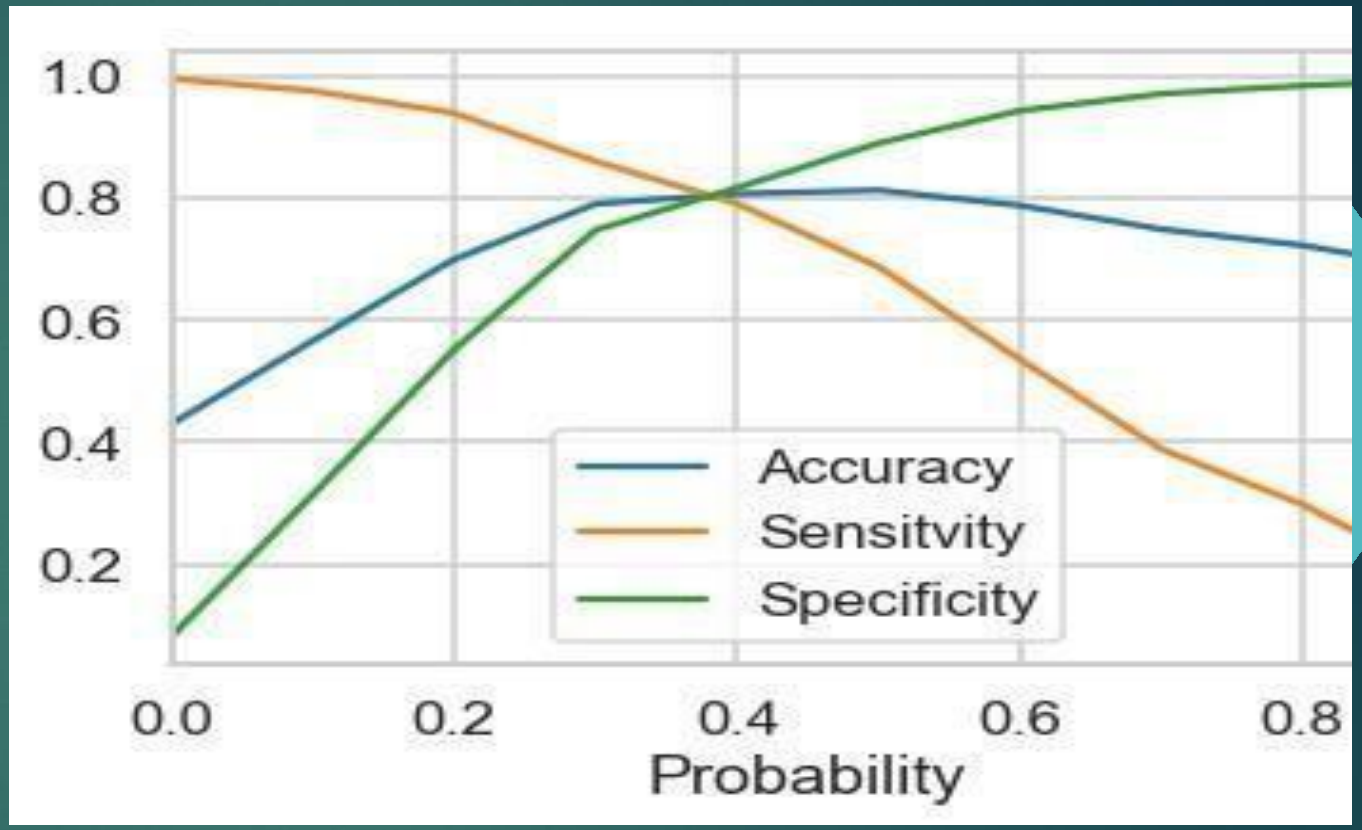


Finding the optimal cutoff point

Now, we have created range of points for which we will find the accuracy, sensitivity and specificity for each points and analyze which point to chose for probability cutoff.

We found that on 0.4 point all the score of accuracy, sensitivity and specificity are in a close range which is the ideal point to select and hence it was selected.

To verify our answer we plotted this in a graph – line plot which is on the right side and we stand corrected that the meeting point is close to 0.4 and therefore we choose 0.4 as our optimal probability cutoff.

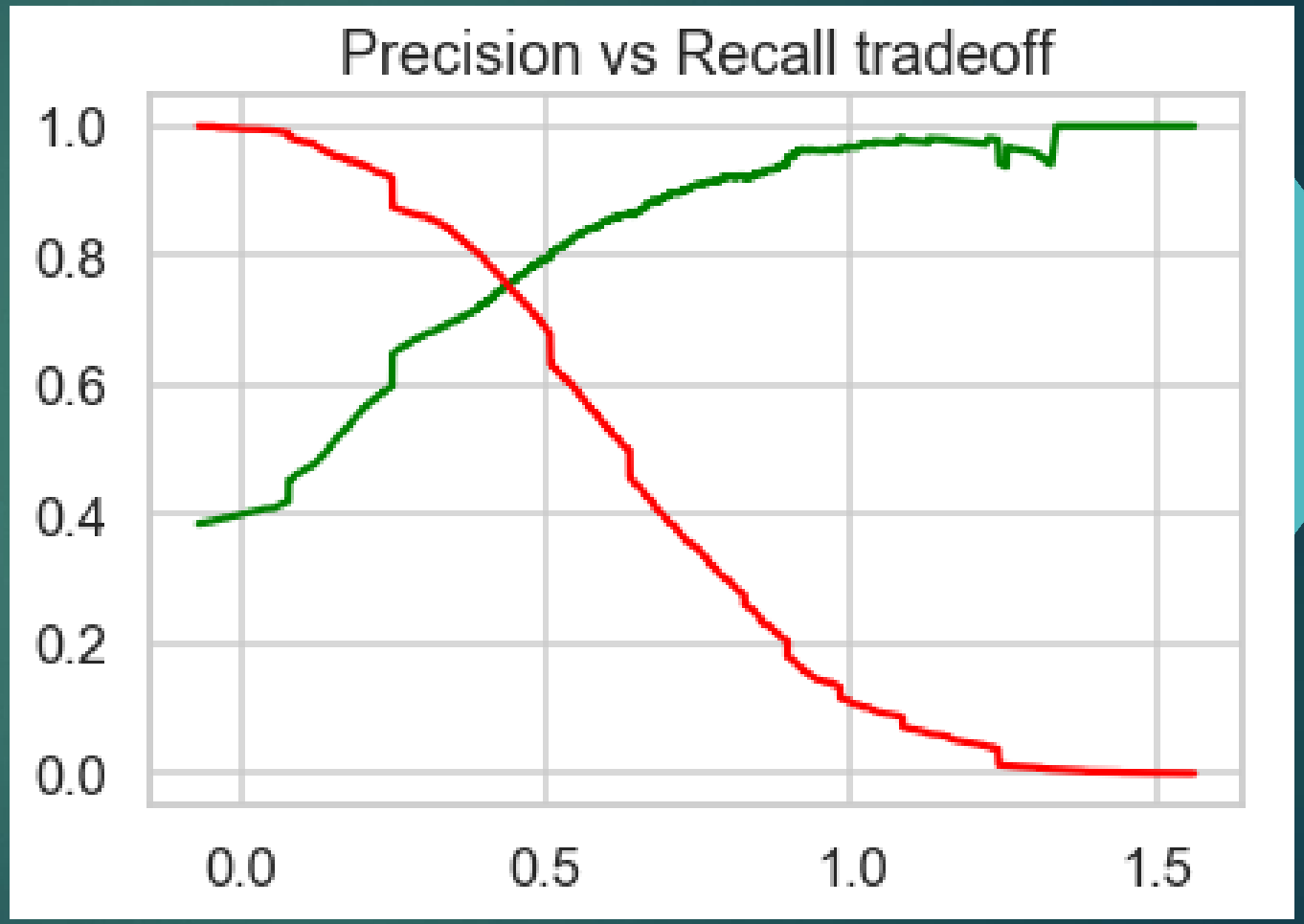


Precision and Recall

- We have used this cutoff point to create a new column in our final dataset for predicting the outcome
- After this we did different type of evaluation which is by checking Precision and Recall
- As we all know, Precision and Recall plays very important role in build our model more business ofriendly and it also tells how our model behaves.
- Hence, we evaluated the precision and recall for this model and found the score as 0.73 for precision and 0.79 for recall.
- Now, recall our business objective - the recall percentage we will consider more valuable because it is okay if our precision is little low which means less hot lead customers but we don't want to left out any hot leads which are willing to get converted hence our focus on this will be more on Recall than Precision.

Precision and Recall tradeoff

- We have created a graph which will show us the tradeoff between Precision and recall.
- We found that there is a trade off between Precision and Recal and the meeting point is approximately at 0.5.



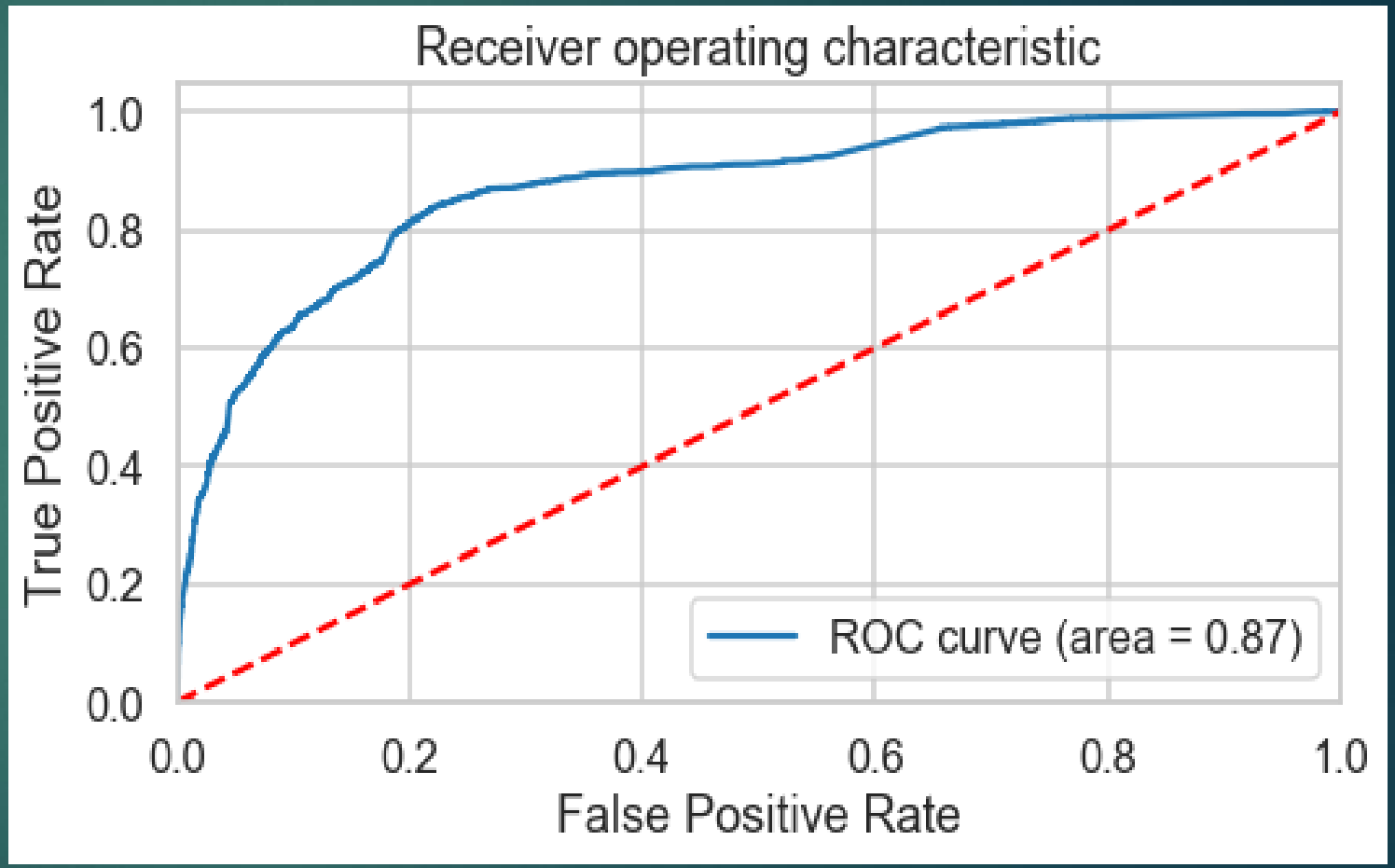
With RFE 2

- After completing our last model evaluation from rfe 1, we proceeded with our second rfe method with count 15.
- We did that same steps like we did in rfe 1, like creating a model and checking the insignificant values and VIFs and dropping those and running again until we reach our model with no insignificant variables and low VIFs.
- Ultimately, we found out last final model with all significant values and low VIFs.
- We predicted the final model in train set and created a new dataset with original converted values and prediction values.
- After this want to verify which final model is the best – one that was created with 19 variables or the one created with 15 variables.

RFE1vsRFE2

- Now We want to choose our final
model for test dataset prediction and in order to do that we plotted ROC curve for the RFE 2 model and compared these two graphs
- Attached is the graph plotted for
the RFE 2 on the right.
- What we found was the auc score(area under the curve) in rfe 2 was 0.87 which was less than auc score generated in rfe 1.

  As we all know that the auc score shows the model accuracy and stability, we established that the final model created by RFE 1 is more stable and accurate than the final



Prediction on test set

- Before predicting on test set, we need to standardize the test set and need to have exact same columns present in our final train dataset.
- After doing the above step, we started predicting the test set and the new predictions values were saved in new dataframe.
- After this we did model evaluation i.e. finding the accuracy, precision and recall.
- The accuracy score that we found was 0.82, precision 0.76 and recall 0.79 approximately.
- This shows that our test prediction is having accuracy , precision and recall score Of an acceptable range.
- This also shows that our model is stable with good accuracy and recall/sensitivity.
- Lead score is created on test dataset to identify hot leads – high the lead score higher the chance of converted, low the lead score lower the chance of getting converted.

Conclusion

Valuable Insights -

- The Accuracy, Precision and Recall/Sensitivity are showing very promising scores in test set which is as expected after looking the same in train set evaluation steps. So we established that recall is having high score value than precision which is acceptable for business needs.
- In business terms, this model has an ability to go along with the company's requirements in coming future.
- This concludes that our model is stable.

Important features responsible for good conversion rate are :

a) **Last Notable Activity_Had a Phone Conversation**
b) **Lead Origin_Lead Add Form and**
c) **What is your current occupation_Working Professional**

Thank You