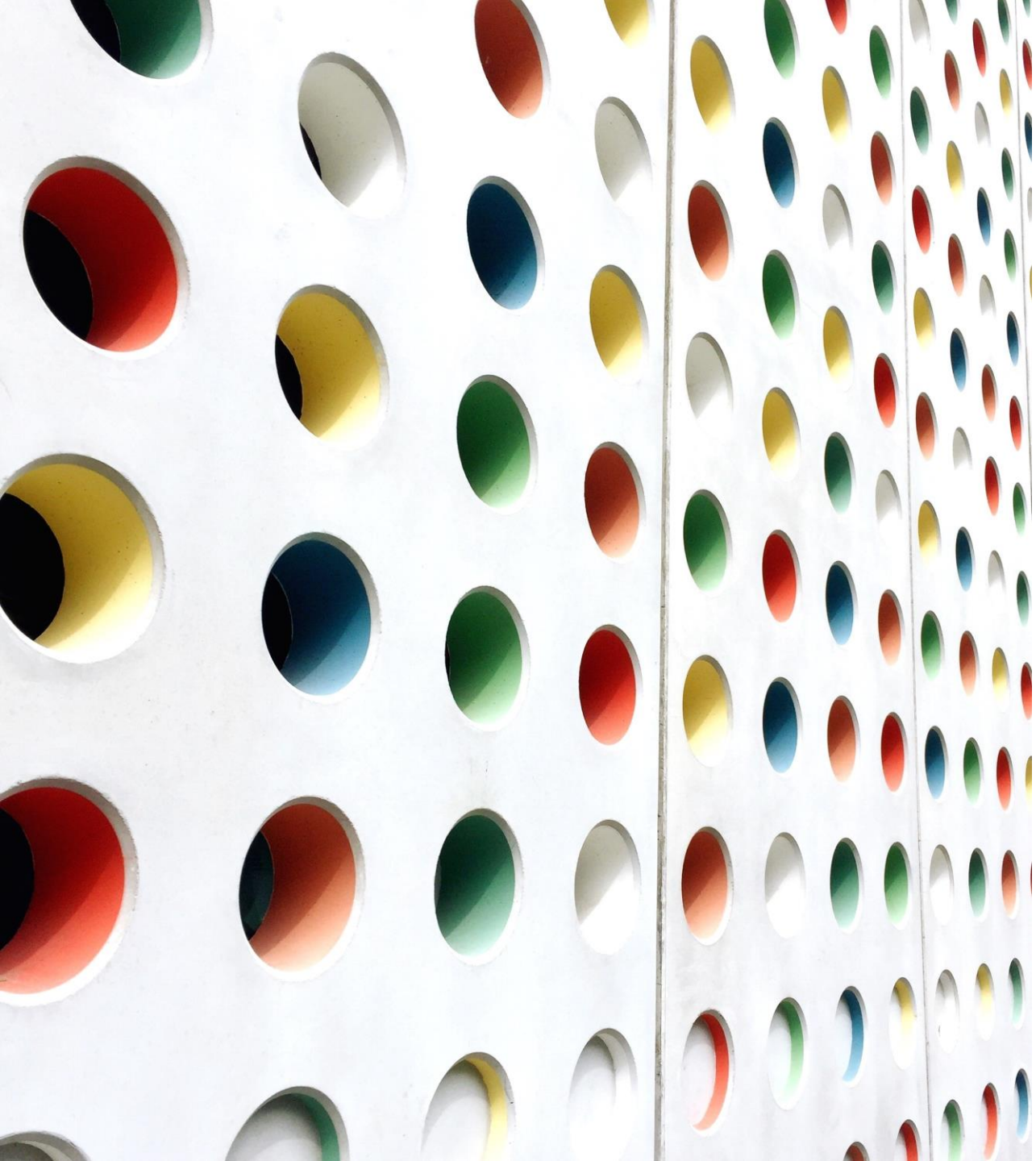


Lead Scoring Assignment

BY :- AJAY BENDALE

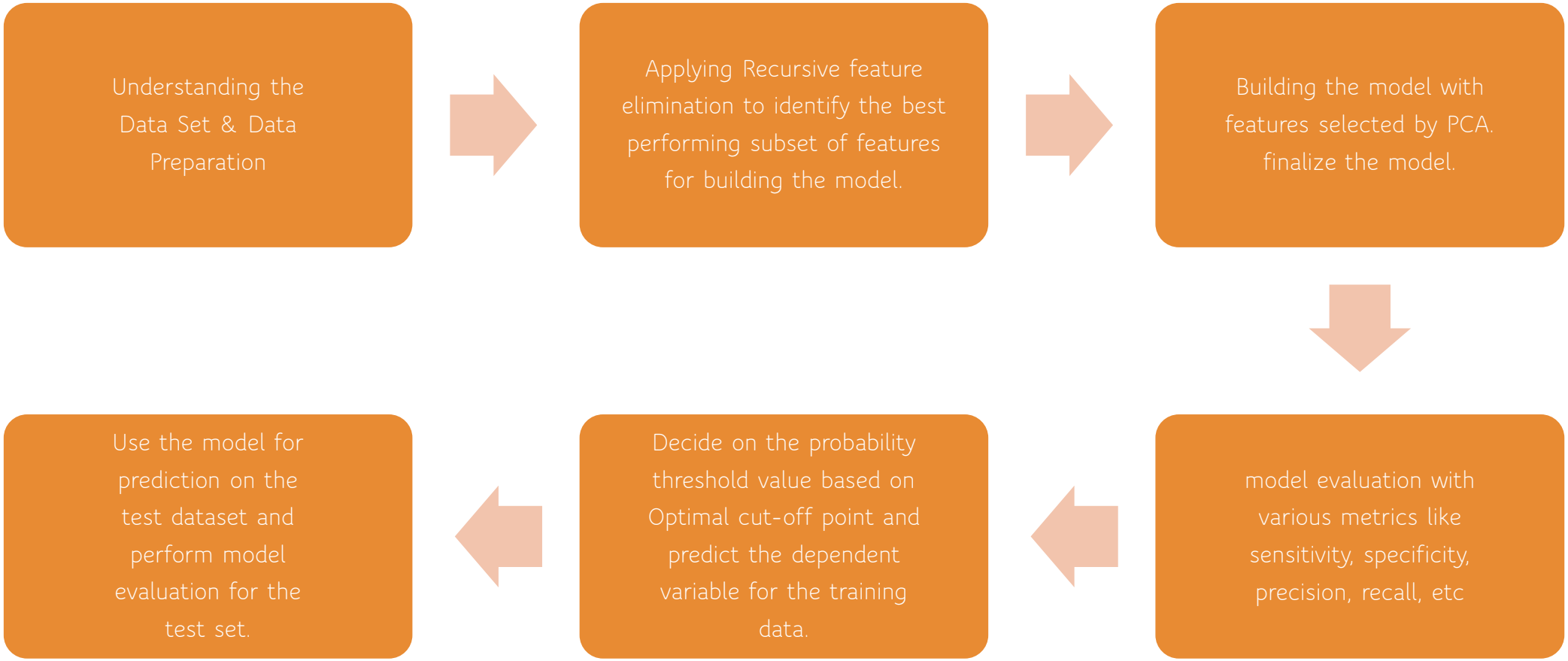
Problem statement

- Create a model 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 adjust if the company's requirement changes in near future.

Business Objective

- To help X Education to select the most promising leads(Hot Leads), i.e. the leads that are most likely to convert into paying customers.
- To build a logistic regression model to assign a lead score value between 0 and 100 to each of the leads which can be used by the company to target potential leads.

Analysis methodology

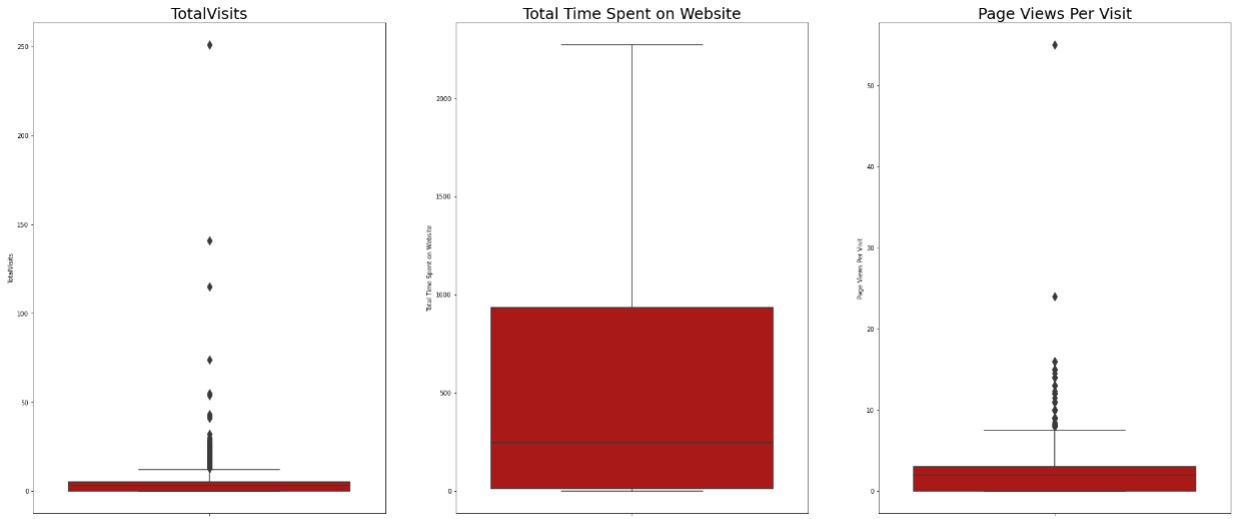


Approach of the analysis

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 checked the outliers of the dataset. The visualization of those outliers we can see on the graph attached on the right side.

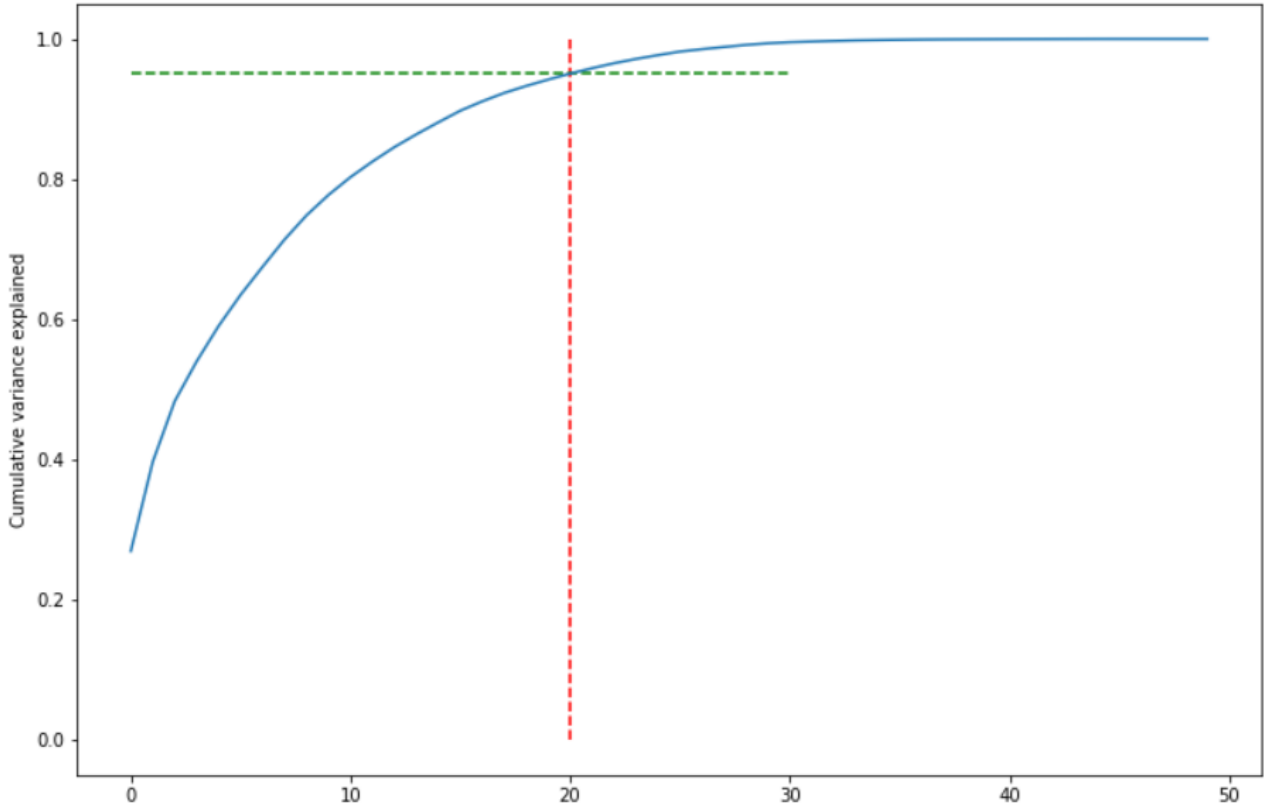
III. Outliers in logistic model is very sensitive hence we need to deal with it without losing our valuable information. This can be achieved by creating bins. Hence, we did it.



scree plot for the explained variance

Perform PCA with 20 components for

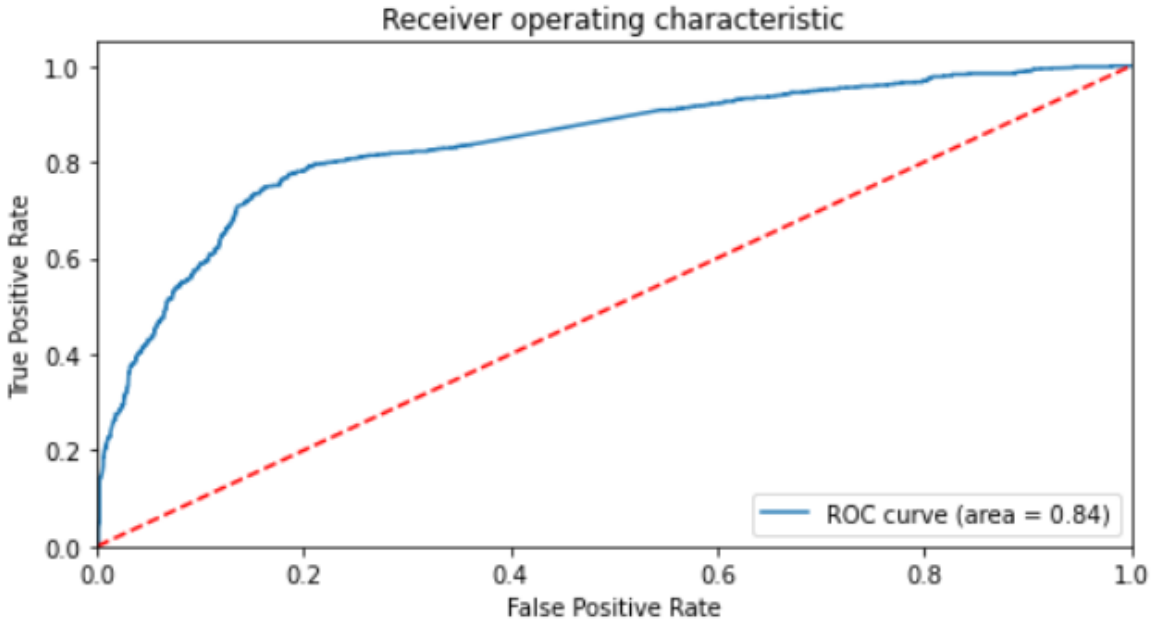
which includes 95% of data variance



Evaluating the model

 After building the 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 is 0.84 which is a great score.

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

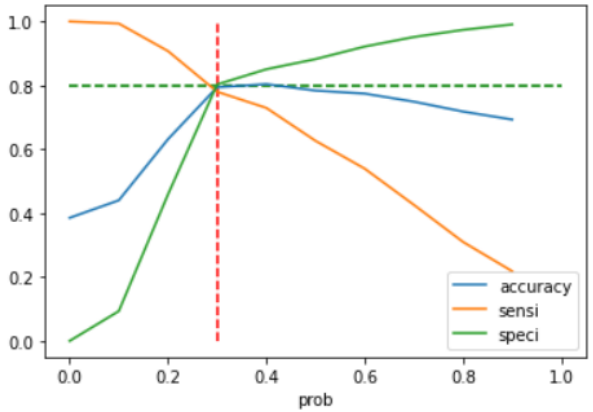


Finding the optimal cutoff point

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 0.3 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 0.3 and hence we choose 0.3 as our optimal probability cutoff.



Summary

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

- Important features responsible for good conversion rate or the ones' which contributes more towards the probability of a lead getting converted are :
    a) Total Time Spent on Website
    b) Lead Origin Lead Add Form and
    c) What is your current occupation Working Professional


THANK
YOU!