LEAD SCORE CASE STUDY

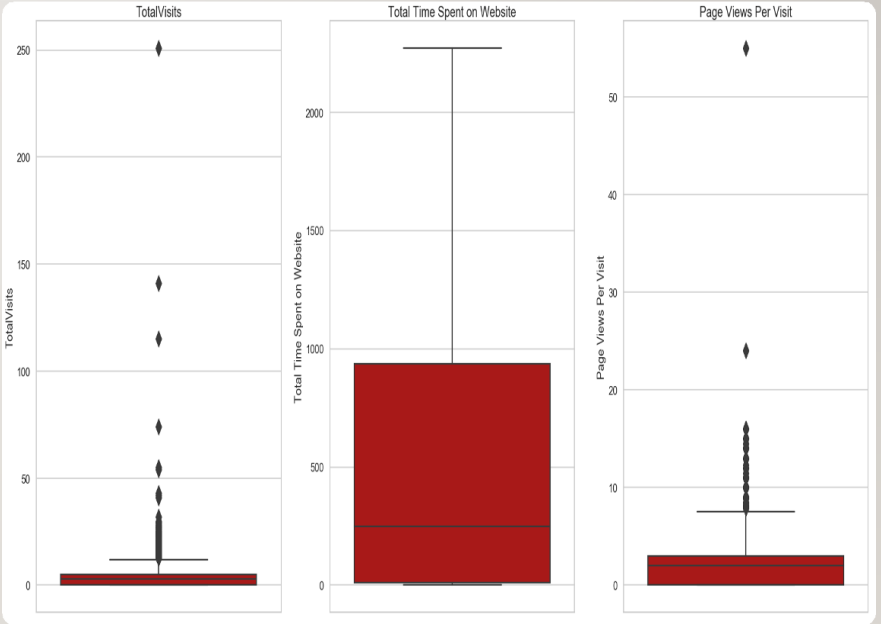
T SAI VISHWANATH

PROBLEM STATEMENT

- Make a model so that consumers with high lead scores have a higher likelihood of converting, while customers with low lead scores have a lesser chance of converting. The target lead conversion rate is in the neighbourhood of 80%.
- Also, the model should be flexible enough to evolve as the needs of the business do in the near future.

APPROACH OF THE ANALYSIS

- With our cleaned dataset, we began our analysis by changing all of the binary variables to "0" and "1" and all of the multiple categories into dummy variables.
- The dataset's outliers were then examined. On the graph that is attached, on the right side, we can see how such outliers are visualised.
- Because outliers in a logistic model are so delicate, we must deal with them carefully to avoid losing important data. By making bins, this can be accomplished. So we did it.



CORRELATION

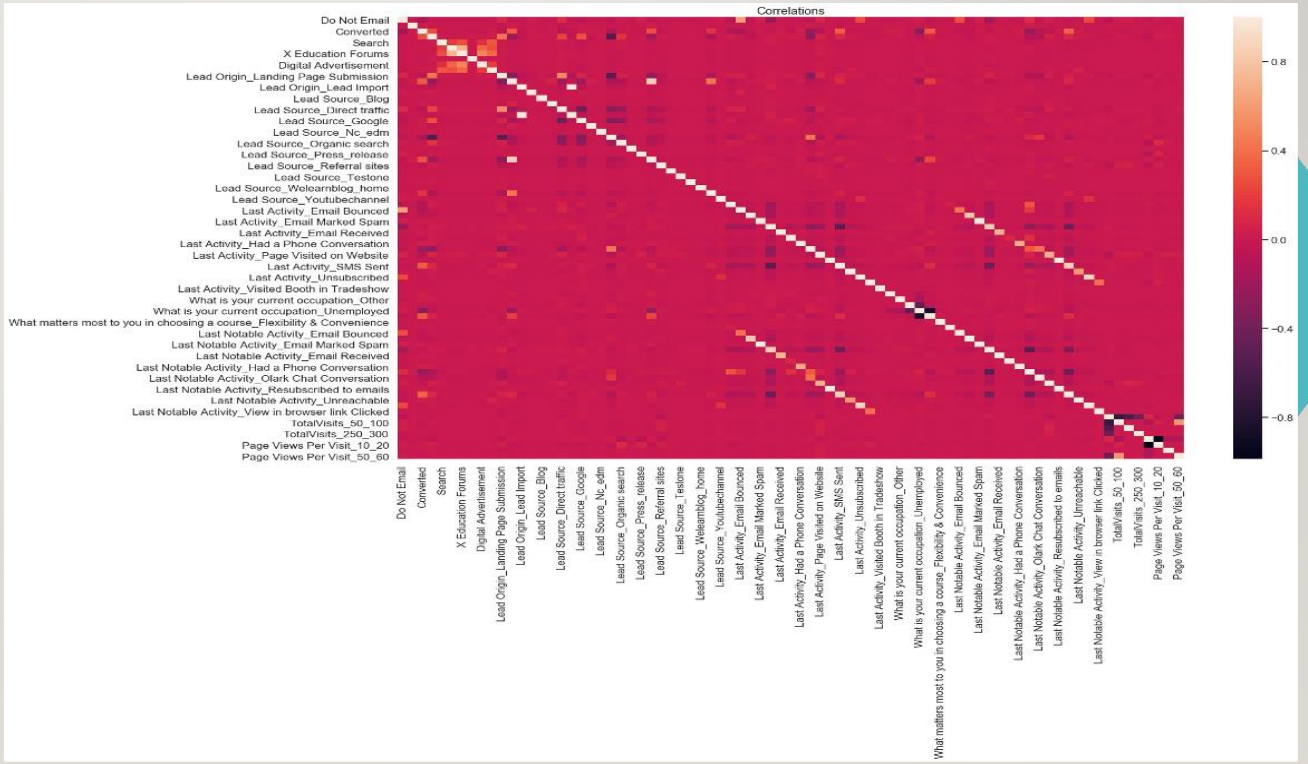
After removing the outliers and creating the dummy, we go on to the next step of analysis, data preparation.

a) We standardise the features and divide the dataset into a train and test set.

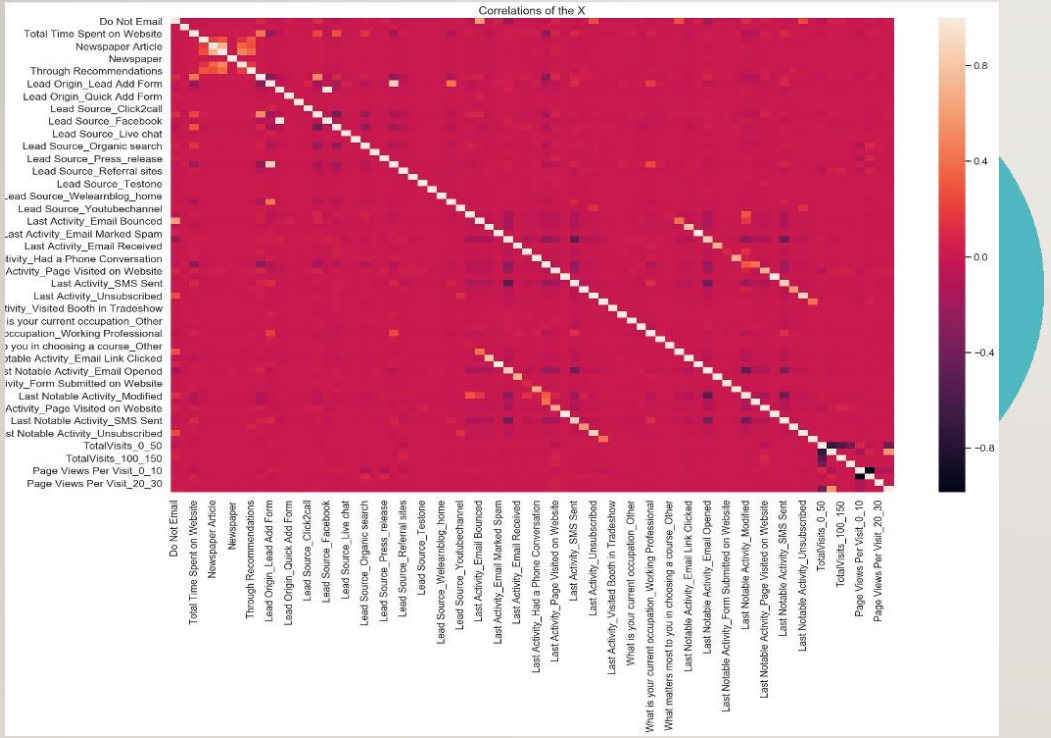
b) To keep all the variables on the same scale and make computations more effective, standardisation is necessary.

c) Examined the dataset's connection. The heatmap that is attached displays the correlation between each attribute in the dataset.

d)There are several high correlations in the heatmap which we dropped



CORRELATION



- We produced a heatmap again after eliminating those features with high correlations to confirm, and it was found that the highly associated variables had been eliminated.
- There are still a small number of them, but we will check them after building our model to see how much of an impact they have because it is difficult to tell from the plot on the right which variable has a large correlation.

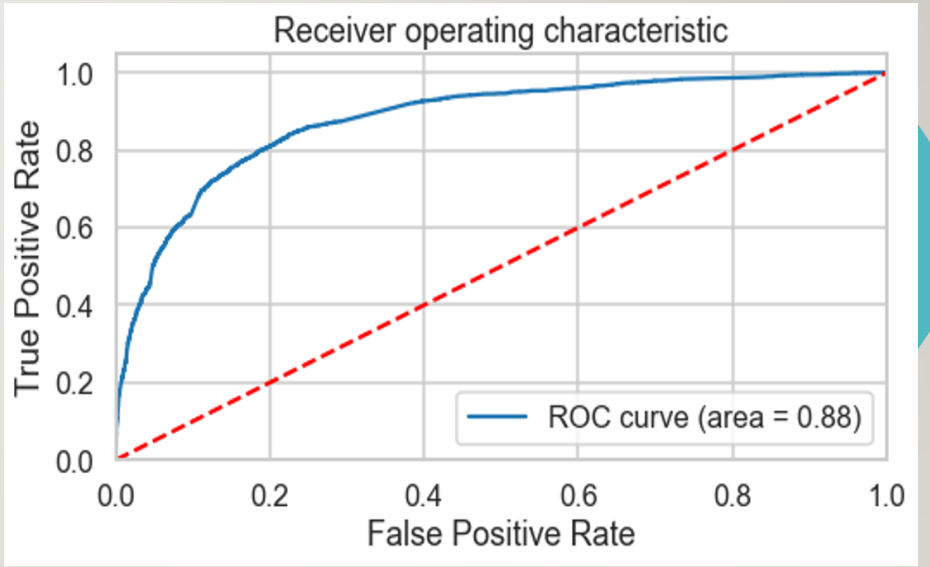
BUILDING A MODEL – RFE 1

- We created a model using all the features given and discovered that our model contained a large number of irrelevant variables.
- We must get rid of them, but doing so one by one would take too much time and be ineffective.
- Hence, we began by subtracting those irrelevant variables using the RFE approach. 19 and 15 RFEs are used to make our decision.
- We conducted two rfe counts in order to determine the stability of our final model.
- We began building our model with rfe count 19 and gradually eliminated variables until the model contained only meaningful variables and had low VIF values.
- Now we tested our model by making a prediction. We produced a fresh dataset using
- Prediction values and the original converted values.

EVALUATING THE MODEL

- As we can see from the graph plotted on the right side, the area score is 0.88, which is a fantastic score. After creating the final model and performing predictions on it (on the train set), we produced ROC curve to find the model stability using auc score (area under the curve).

- Also, the fact that our graph leans to the left of the border indicates that our accuracy is high.

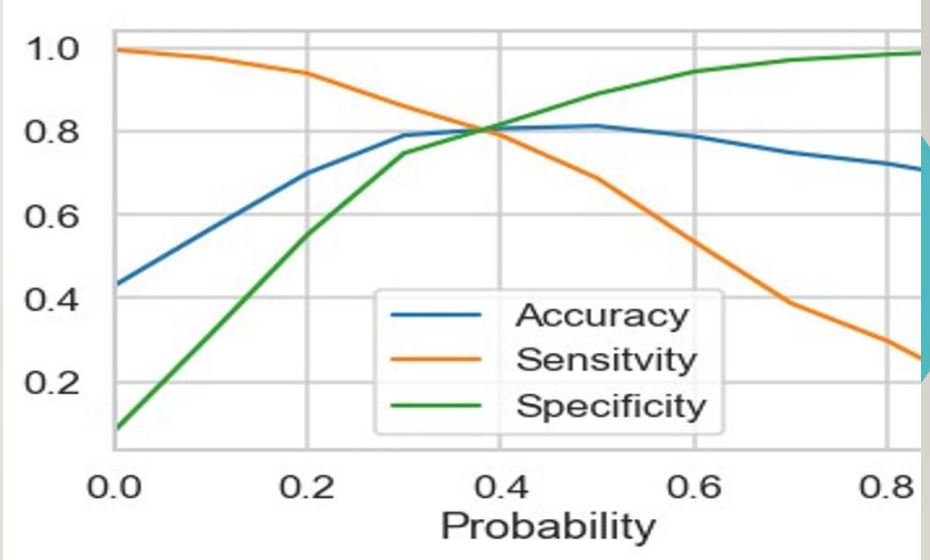


FINDING THE OPTIMAL CUTOFF POINT

- We have now established a range of points for which we will test each point's accuracy, sensitivity, and specificity before deciding which point to use as the 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.

- We put this in a graph to validate our conclusion. The line plot is on the right. We stand corrected that the meeting point is near to 0.4, thus we select 0.4 as our ideal probability cutoff.

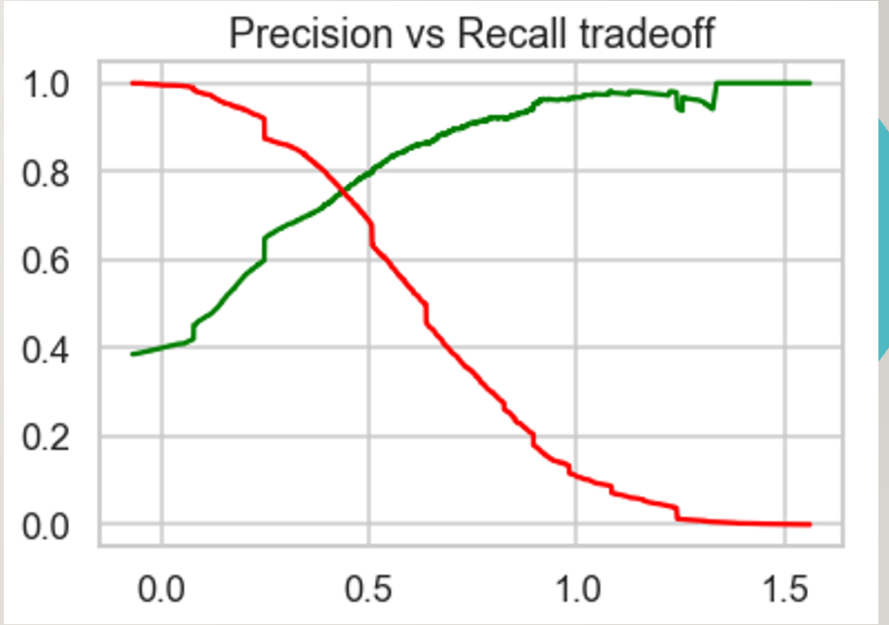


PRECISION AND RECALL

- In our final dataset for the predictions, we added a new column using this cutoff point.
- Following this, we conducted another form of examination by examining Precision and Recall.
- As everyone is aware, Precision and Recall play a crucial role in making our model more commercially focused and in describing how our model works.
- As a result, we assessed the model's precision and recall and discovered that the scores were 0.73 for precision and 0.79 for recall.
- Now, think back to our business goal, the recall %. As we don't want to miss any hot leads who are willing to convert, our attention will be more on recall than precision. It is okay if our precision is a little poor, which results in fewer hot lead customers.
- i.e. Our methodology generates more hot lead customers and more relevant outcomes .

PRECISION AND RECALL TRADEOFF

- Using a graph, we were able to visualise the trade-off between recall and precision.
- The meeting point between Precision and Recall, according to our research, is roughly at 0.5.

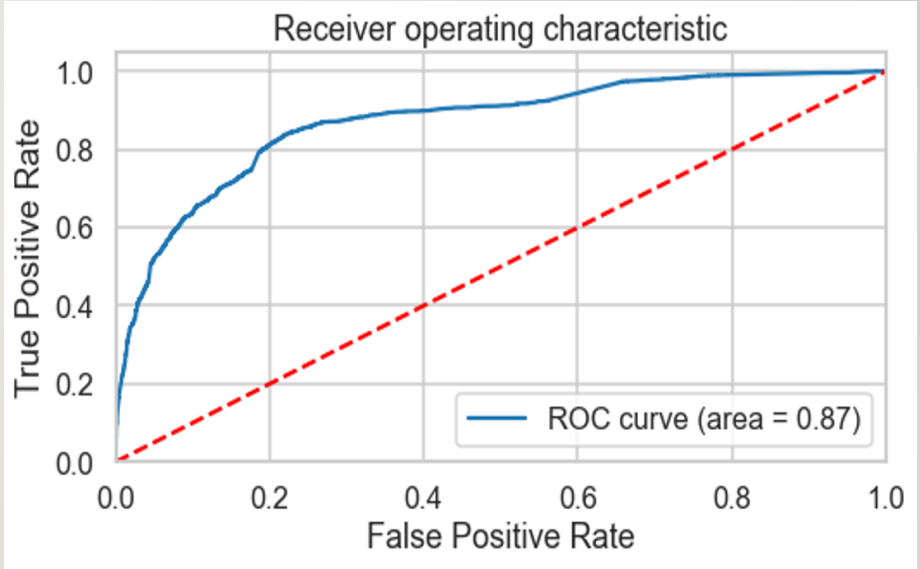


WITH RFE 2

- We continued with our second rfe technique with count 15 after finishing our model evaluation from rfe 1.
- We used the same procedures as in rfe 1: we created a model, checked it for insignificant values and VIFs, eliminated them, and then ran the model once again until it had no unimportant variables and had a low VIF.
- Finally, a model with all significant values and a low VIF was discovered.
- In the train set, we predicted the final model, and we generated a new dataset containing both the original conversion values and the forecast values.
- After that, you should determine which of the two final models the one developed with 19 variables or the one created with 15 variables is the best.

RFE 1 VS RFE 2

- In order to select our final model for test dataset prediction, we plotted the RFE 2 model's ROC curve and compared these two graphs.
- Attached graph plotted for the RFE 2 on the right.
- We discovered that the area under the curve (auc score) generated by rfe 2 was 0.87, which was lower than the auc score generated by rfe 1.
- We discovered that the final model produced by RFE 1 is more accurate and stable than the final model produced by RFE 2, despite the fact that we are all aware that the auc score indicates the model accuracy and stability.



PREDICTION ON TEST SET

- We must standardise the test set and ensure that the exact identical columns are present in our final train dataset before making predictions on the test set.
- We began predicting the test set after completing the above step, and the new predictions values were saved in a new dataframe.
- After this we did model evaluation i.e. finding the accuracy, precision and recall.
- The accuracy score 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 in an acceptable range.
- This also shows that our model is stable with good accuracy and recall/sensitivity.
- For a test dataset, a lead score is developed to detect hot leads; the higher the lead score, the higher the likelihood that the lead will convert, and the lower the lead score, the less likely it will be to convert.

CONCLUSION

- Valuable Insights -
- After examining the same in train set evaluation stages, the Accuracy, Precision, and Recall/Sensitivity are exhibiting promising scores in the test set, which is expected. This indicates that the recall has a high score value compared to the precision, which is suitable for business demands.
- In terms of business, this model offers the flexibility to change to meet the needs of the company in the 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) Last Notable Activity_Had a Phone Conversation

b) Lead Origin_Lead Add Form and

C) What is your current occupation_Working Professional

THANK YOU