LEAD SCORE CASE STUDY

ANUPAM BARASIA
ABHISHEK DAS

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%.

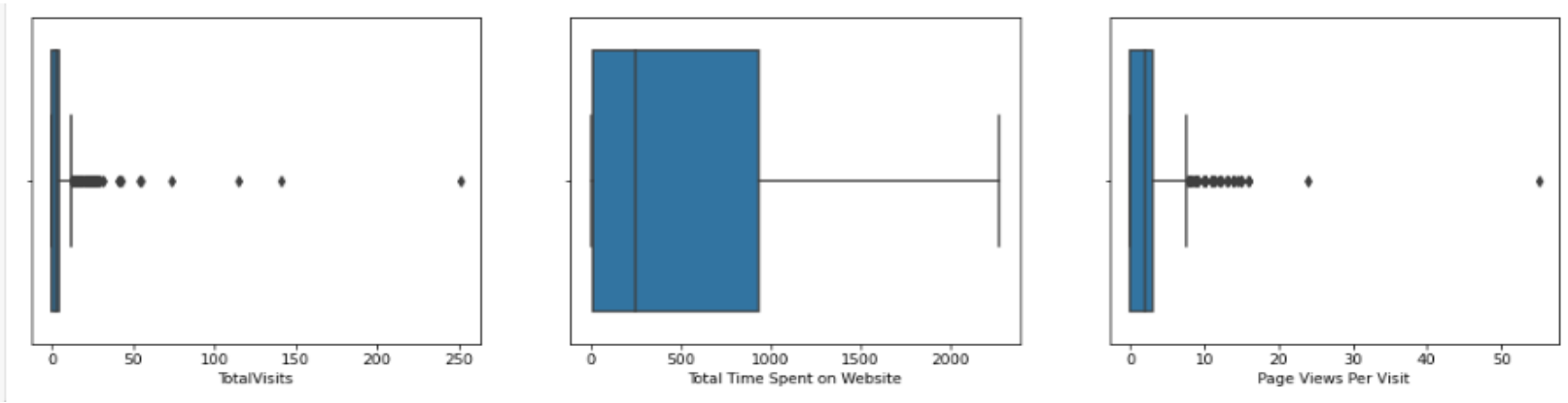
Also the model should be able to adjust if the company's requirement changes in near future.

APPROACH OF THE ANALYSIS
ANALYSIS

-Initially we cleaned the dataset by converting all the binary variables to 0 & 1 and multiple categories to dummy variables.

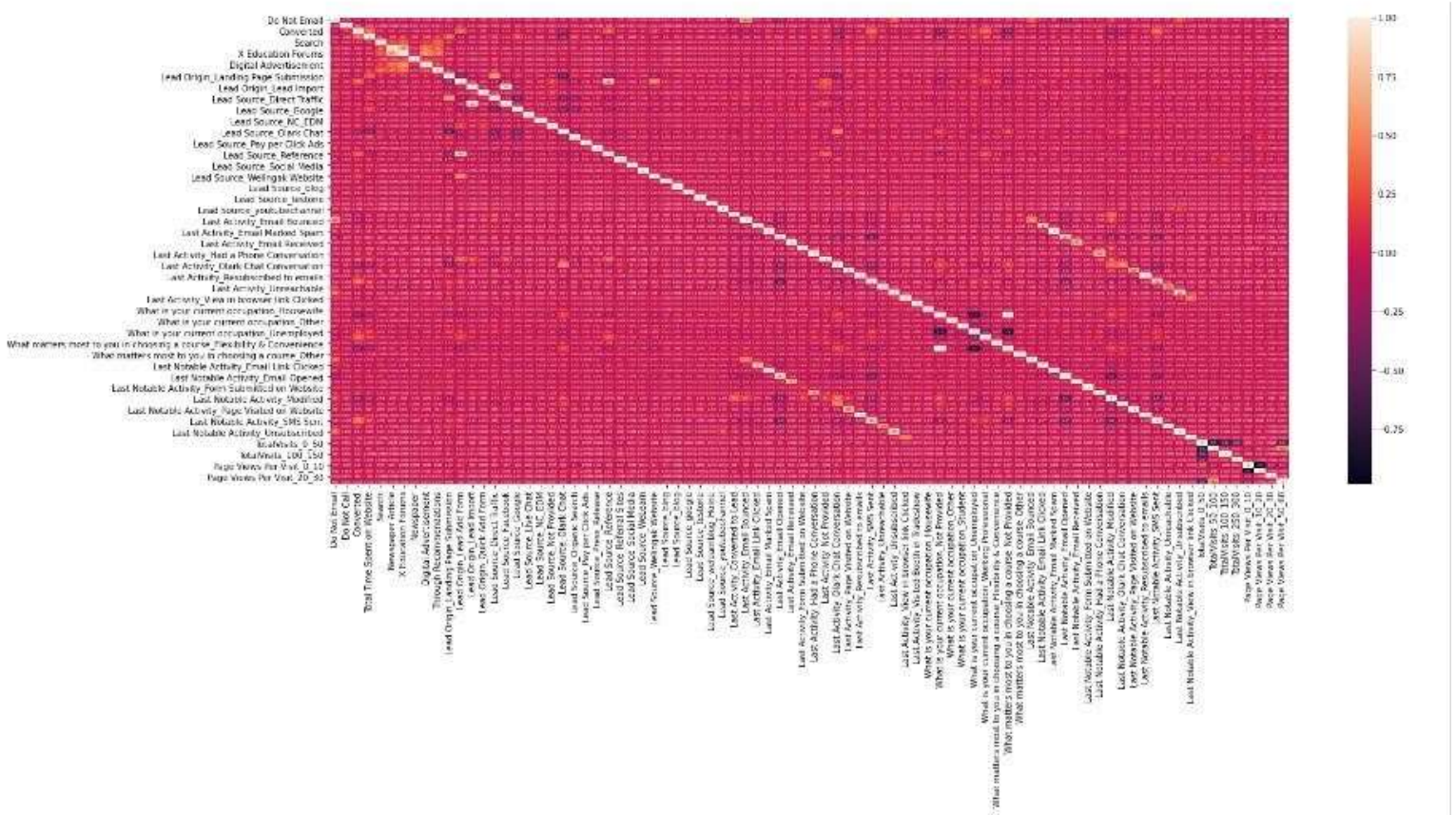
- We also checked for outliers which can be visualized from the following graphs.

-Outliers are not removed because they are necessary for bus



CORRELATION

We have plotted a heatmap to find out the highly correlated variables.



MODEL BUILDING - RFE

We built a model with all the features included and found there were many insignificant variables present in our model.

We need to drop them, but we can't do it one by one as it is time consuming and not an efficient way to do so.

Therefore, 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 went 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.

FINAL MODEL VISUALIZATION WITH VIF

	Features	VIF
11	Last Notable Activity_Modified	1.53
1	Total Time Spent on Website	1.48
9	What matters most to you in choosing a course_...	1.46
3	Lead Origin_Lead Add Form	1.40
7	Last Activity_SMS Sent	1.38
4	Lead Source_Olark Chat	1.34
5	Lead Source_Welingak Website	1.24
8	What is your current occupation_Working Profes...	1.18
0	Do Not Email	1.10
12	Last Notable Activity_Olark Chat Conversation	1.07
2	Newspaper	1.00
6	Last Activity_Resubscribed to emails	1.00
10	Last Notable Activity_Had a Phone Conversation	1.00
13	Last Notable Activity_Unreachable	1.00
14	TotalVisits_250_300	1.00

EVALUATING THE MODEL

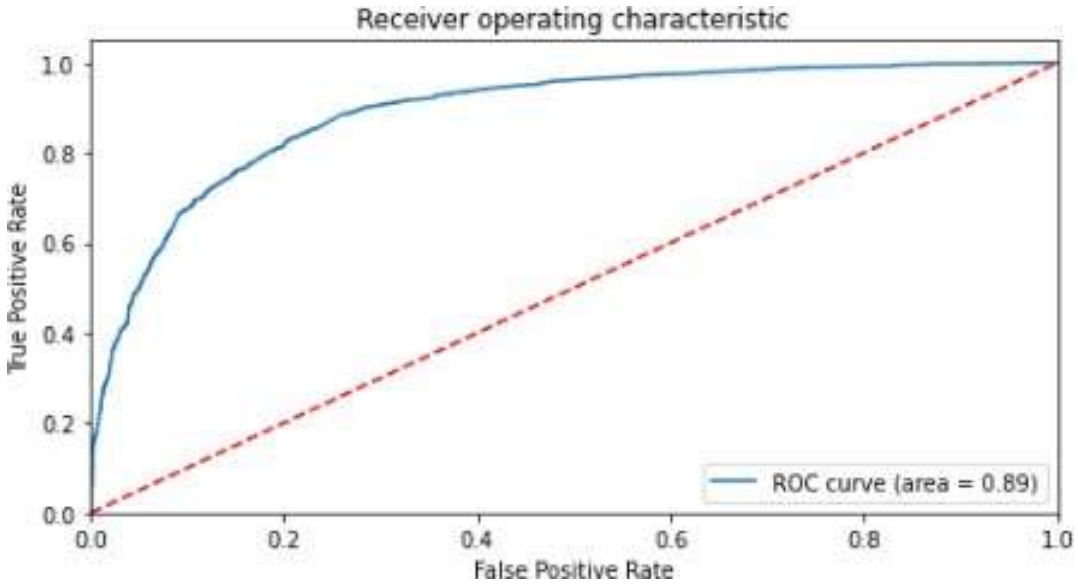
After building the final model making prediction on it(on train set),

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 is 0.89 which is a great score

We also see that our graph is leaning towards the left side of the curve which means that this is an accurate model.

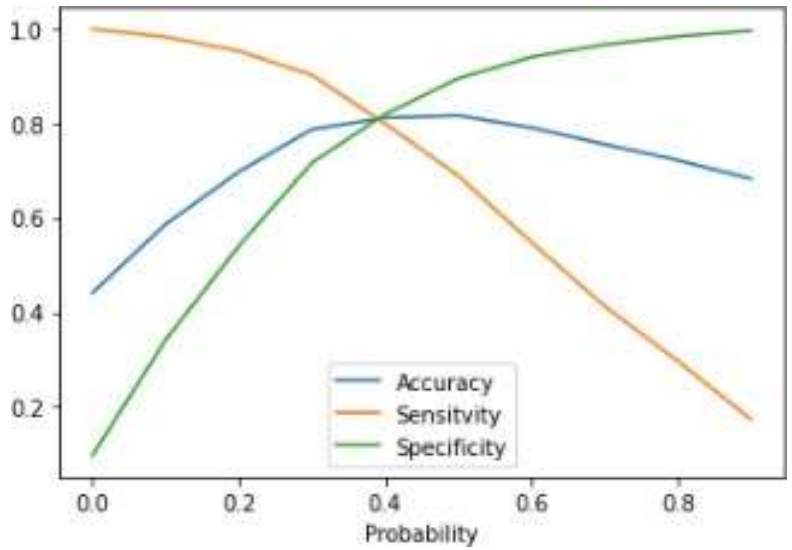


FINDING THE OPTIMAL CUTOFF POINT

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 hence we choose 0.4 as our optimal probability cutoff.



PRECISION AND RECALL

We used this cutoff point to create a new column in our final dataset for predicting the outcomes.

After this we did another type of evaluation which is by checking Precision and Recall.

Hence, we evaluated the precision and recall for this model and found the score.

as 0.74 for precision and 0.79 for recall.

Now, recall our business objective - the recall percentage I will consider more valuable because it is okay if our precision is little low which means less hot lead.

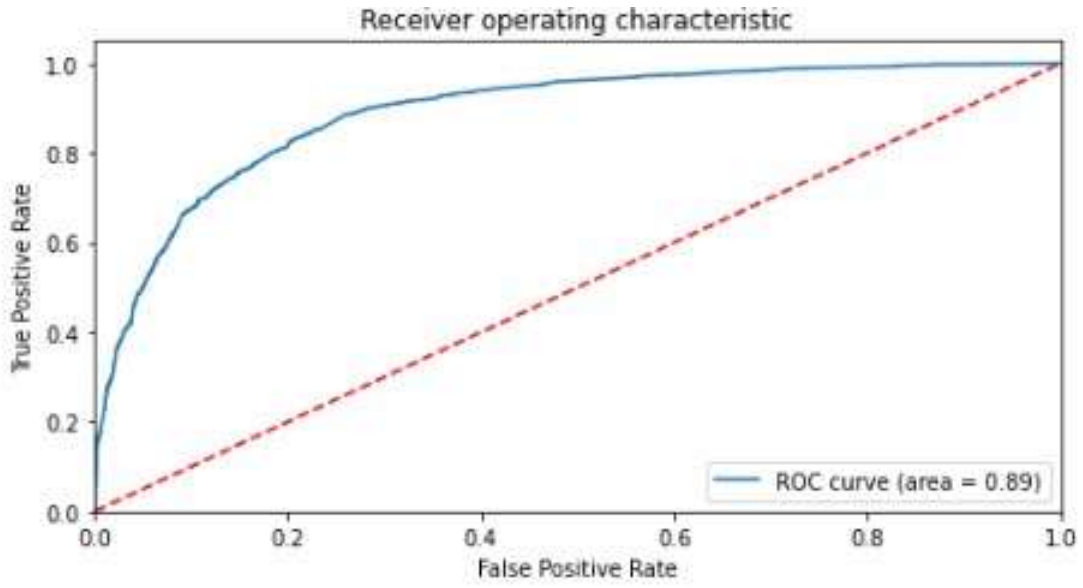
customers but we don't want to left out any hot leads which
are willing to get converted hence our focus on this will be more on Recall than Precision.

We get more relevant results - as many as hot lead customers from our model.

RFE 1 VS RFE 2

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 graph plotted for the RFE 2 on the right which has the same auc ie 0.89.



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 we found was 0.81, precision 0.74 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.

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

Insights -

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 variables from the model:

Last Notable Activity_Had a Phone Conversation

Lead Origin_Lead Add Form and

What is your current occupation_Working Professional

THANK YOU