LOGISTIC REGRESSION – CASE STUDY

SUMA DONOJU

Dataset

- It has 9240 rows and 37 columns.
- It has missing values in few columns and the columns having more than 30% missing values are dropping.
- There are select values in the columns are replaced with null values.
- Imputation of missing values of columns having less than 30% of missing values. Replaced with “Unknown”.
- Replacing the missing values of less than 5% with mean/mode value.
- Dropping the columns the which have constant features.
- Outliers are removed and hence the data is left with (8719,19)

EDA

- The percentage of converted is more compared to the percentage of non-converted.
- In case of “Lead Origin”, Lead Add Form has more percentage of converted to other categories. So, we need to focus on Lead Add Form customers to improve the business.
- In case of Lead Source, we must focus on “Referral Sites”, “Google”, ”Facebook” and “bing”. On an overall, we need to focus on social media.
- “Do Not Email” has more percentage of non-converted in both the cases (Yes and No)

EDA

- “Do Not Call” also have more percentage of non-converted in No and almost no data in Yes.
- In the “Last Activity”, we can focus on sms sent.
- In case of “what is your current occupation” we can focus on working professionals
- The “search” shows similar results as “Do Not call”.
- “Newspaper Article” almost has no data in yes category and dropping the column. Similarly dropping, “X Education Forums” and “Newspaper”.

- “Through Recommendations” also show similar results as “Do Not Call”.
- "A free copy of Mastering The Interview“ has more of non-converted in both yes and no.
- “Last Notable Activity”, we can focus on SMS sent.
- In case of numerical features, the conversion rates are high in all the cases.
- Created dummy variables for categorical variables.
- Now the shape has become (9029,72)

- Splitting the data into train and test sets.
- Scaling the numerical variables.
- Model building
- We have 39% conversion data.

Model – 1

- P value is zero for all the features and the few variables have negative coefficients.
- We need to try RFE to choose the variables and build the model.

Model - 2

- P value is high for “Last Activity approached upfront”,”what is current occupation Housewife”. However, they have very high coefficients.
- The residual value compared to previous model has slightly increased.

Model -3

- After dropping “Last Activity approached upfront”, the residual value has drastically decreased.
- The feature “What is your current occupation_Housewife”, has high p-value.
- So, dropping this column for the next model.

Model - 4

- The features “Search”, “Lead Source_Reference”, “Last Activity_Had a Phone Conversation” and “Last Notable Activity_Had a Phone Conversation”. has pvalues in the range 0.1-0.2.
- The highest among the 4 is “Last Activity_Had a Phone Conversation” and dropping this column for next model.

Model - 5

- Lead Source_Reference and Search has high p values.
- Dropping Lead Source_Reference for the next model

Model -6

- Search feature has 0.106 and dropping this for the next model.

Model -7

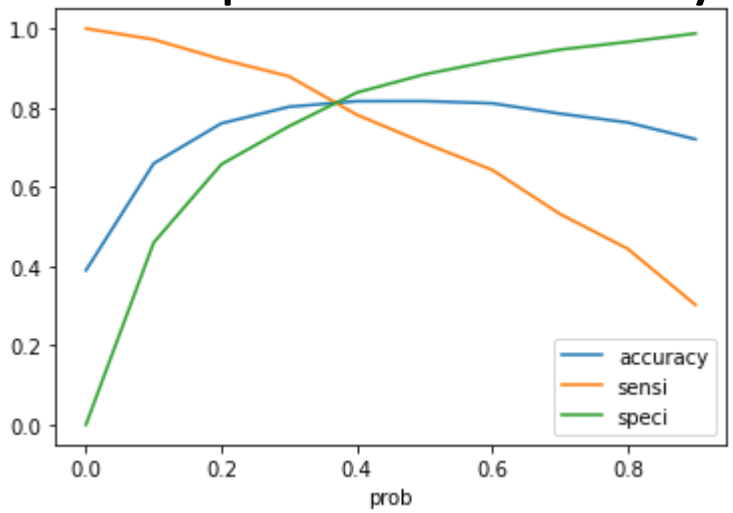
- The residual value has increased slightly.
- This model showed significant p-values and we can go with this model for further analysis.
- Converting the probability > 0.5 to 1.
- The overall accuracy of the model is 81.6%.
- Checking VIFs of the features.
- The VIF values of all the features have good values and there is no need to drop any variables.

Metrics

- Checking the metrics of the model – sensitivity, specificity, false positive rate, positive predictive value and negative predictive value.
- Sensitivity – 0.71
- Specificity – 0.88
- False Positivity Rate – 0.115
- Positive Predictive Value – 0.7969
- Negative Predictive Value – 0.827
- All the metrics are showing in good range.

Optimum cut off

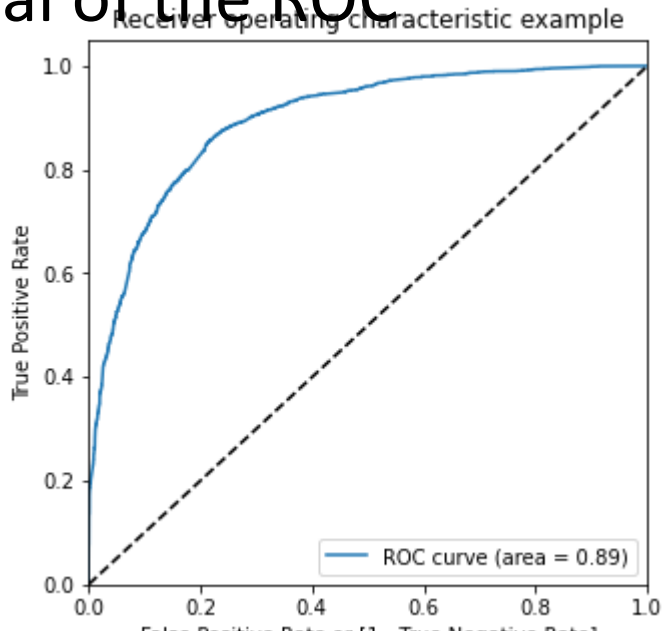
- The optimal cut off is 0.37 as per the accuracy.



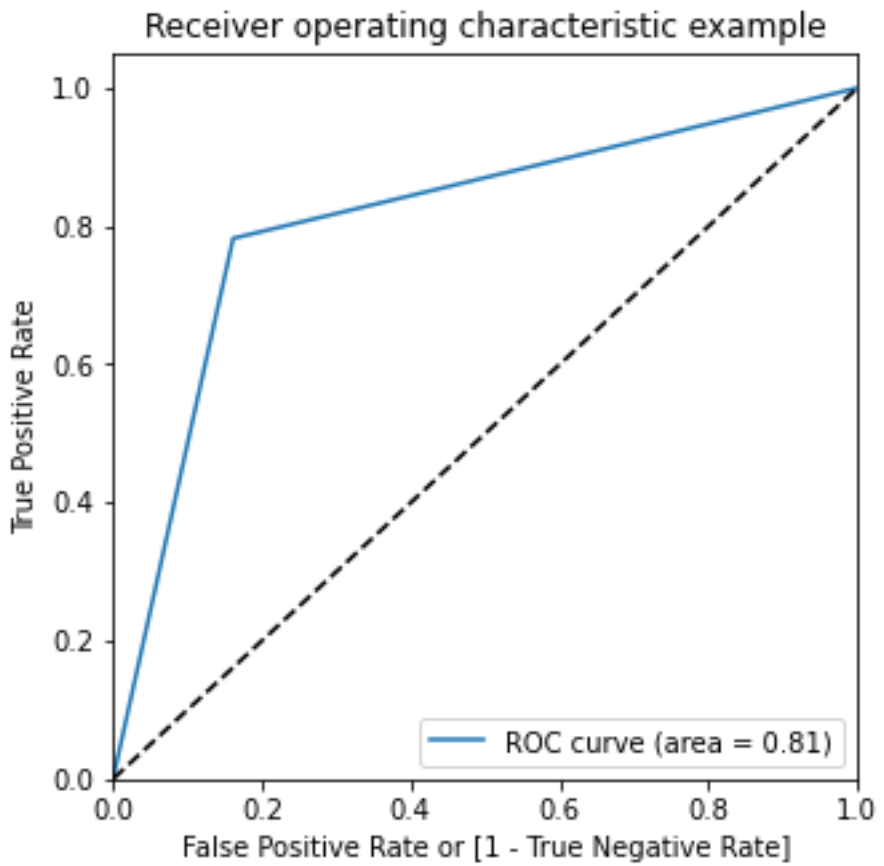
- The overall accuracy is found same as with 0.5 cut-off.

ROC Curve

- It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
- The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.



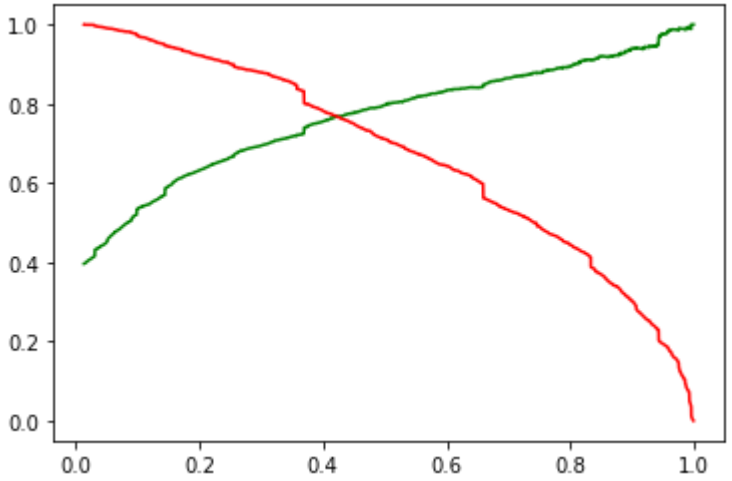
ROC Curve with optimal cut off



- Hence we can see that the final prediction of conversions have a target of 80% (79.8%) conversion as per the X Educations CEO's requirement . Hence this is a good model.
- Metrics for final model-
- Sensitivity – 0.78
- Specificity – 0.835
- False Positive Rate – 0.16
- Positive predictive value – 0.755
- Negative predictive value – 0.857

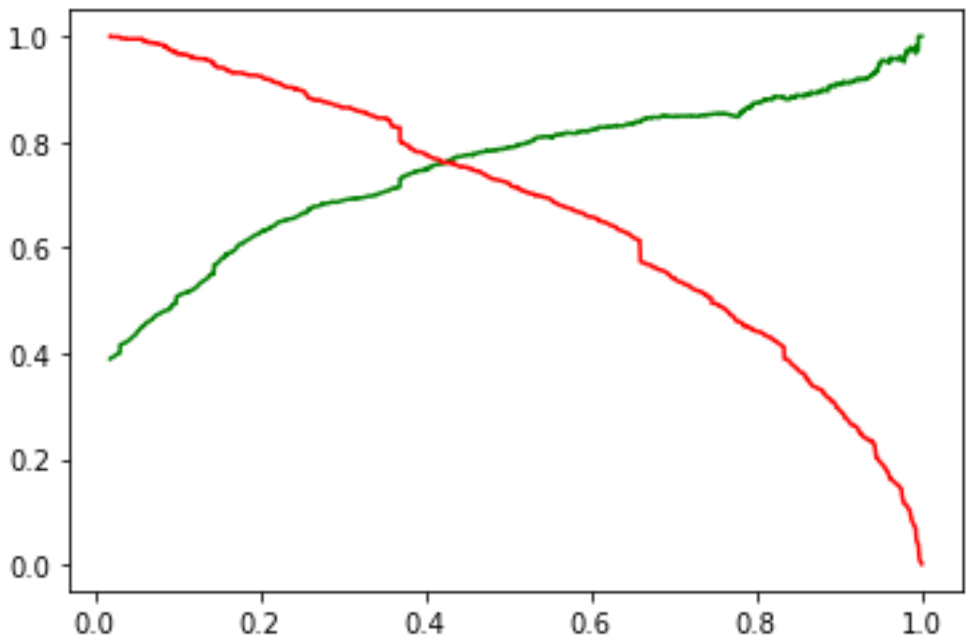
Precision and Recall

- PRECISION – 0.755
- RECALL – 0.7822
- PRECISION-RECALL TRADEOFF



Making predictions on test data.

- The final prediction of conversions have a target rate of 79% (78.5%) (Around 1 % short of the predictions made on training data set)
- Let’s check the metrics
- Accuracy – 0.81
- Sensitivity – 0.8
- Specificity – 0.822
- Precision - 0.792
- Recall – 0.8


1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

CONCLUSION

- 1.we have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the
- optimal cut off based on Sensitivity and Specificity for calculating the final prediction.
- -2.Accuracy, Sensitivity and Specificity values of test set are around 81%, 79% and 82% which are approximately closer to
- the respective values calculated using trained set.
- 3.Also the lead score calculated in the trained set of data shows the conversion rate on the final predicted model is
- around 80%
- 4.Hence overall this model seems to be good.