

Summary

The model building and prediction is being done for company X Education and to find ways to convert potential users. We will further understand and validate the data to reach a conclusion to target the correct group and increase conversion rate. Let us discuss steps followed:

1. EDA:

- Quick check was done on % of null value and we dropped columns with more than 45% missing values.
- We also saw that the rows with the null value would cost us a lot of data and they were important columns. So, instead we replaced the NaN values with 'not provided'.
- Since India was the most common occurrence among the non-missing values, we imputed all not provided values with India.
- Then we saw the Number of Values for India were quite high (nearly 97% of the Data), so this column was dropped.
- We also worked on numerical variable, outliers and dummy variables.

2. Train-Test split & Scaling :

- The split was done at 70% and 30% for train and test data respectively.
- We will do min-max scaling on the variables ['TotalVisits', 'Page Views Per Visit', 'Total Time Spent on Website']

3. Model Building

- RFE was used for feature selection.
- Then RFE was done to attain the top 15 relevant variables.
- Later the rest of the variables were removed manually depending on the VIF values and p-value.

4. Model Evaluation

The 82% accuracy with is a very good value.
sensitivity of around 70% and specificity of around 89%.

5. Roc curve

The area under ROC curve is 0.88 which is a very good value.

From the graph it is visible that the optimal cut off is at 0.35.

6.Precision – Recall:

With the current cut off as 0.35 we have Precision around 80% and Recall around 70%.

Prediction on Test set

Variable: TotalVisits', 'Page Views Per Visit', 'Total Time Spent on Website.

With the current cut off as 0.41 we have Precision around 75% , Recall around 79% and accuracy 82.5%.

The Model seems to predict the Conversion Rate very well and we should be able to give the CEO confidence in making good calls based on this model

The most potential buyers are:

CONCLUSION

TOP VARIABLE CONTRIBUTING TO CONVERSION:

TotalVisits

#The total time spend on the Website.

#Lead Origin_Lead Add Form

#Lead Source_Google

#Lead Source_Direct Traffic

#Lead Source_Welingak Website

#Do Not Email_Yes

#Last Activity_Olark Chat Conversation

The Model seems to predict the Conversion Rate very well and we should be able to give the Company confidence in making good calls based on this model.