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.
- A confusion matrix was created, and overall accuracy was checked which came out to be 80.91%.

4. Model Evaluation

- Sensitivity – Specificity

If we go with Sensitivity- Specificity Evaluation. We will get :

- On Training Data

  - The optimum cut off value was found using ROC curve. The area under ROC curve was 0.88.
  - After Plotting we found that optimum cutoff was 0.35 which gave

Accuracy 80.91%

Sensitivity 79.94%

Specificity 81.50%.

- Prediction on Test Data

  - We get

Accuracy 80.02%

Sensitivity 79.23%

Specificity 80.50%

- Precision – Recall:

If we go with Precision – Recall Evaluation

 On Training Data

o With the cutoff of 0.35 we get the Precision & Recall of 79.29% & 70.22% respectively.

o So to increase the above percentage we need to change the cut off value. After plotting we found the optimum cut off value of 0.44 which gave

Accuracy 81.80%

Precision 75.71%

Recall 76.32%

 Prediction on Test Data

o We get

Accuracy 80.57%

Precision 74.87%

Recall 73.26%

5. So if we go with Sensitivity-Specificity Evaluation the optimal cut off value would be 0.35

&

If we go with Precision – Recall Evaluation the optimal cut off value would be 0.44

CONCLUSION

TOP VARIABLE CONTRIBUTING TO CONVERSION:

• LEAD SOURCE:

o Total Visits

o Total Time Spent on Website

• Lead Origin:

o Lead Add Form

- Lead source:

o Direct traffic

o Google

o Welingak website

o Organic search

o Referral Sites

Last Activity:

- Do Not Email_Yes
- Last Activity_Email Bounced
- 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.

Target leads that spend a lot of time on X-Education site (Total Time Spent on Website)

Target leads that repeatedly visit the site (Page Views Per Visit). However they might be repeatedly visiting to compare courses from the other sites, as the number of visits might be for that reason. So the interns should be a bit more agressive and should ensure competitive points where X-Education is better, are stringly highlighted.

Target leads that have come through References as they have a higher probability of converting

Students can be approached, but they will have a lower probability of converting due to the course being industry based. However, this can also be a motivating factor to ensure industry readiness by the time they complete their education