Summary of Leading Score

Summary of Leading Score:

Our focus lies on constructing the model, making predictions, and optimizing strategies for company X Education. The objective is to identify effective approaches for converting prospective users. Subsequently, we will thoroughly analyze and validate the gathered data, leading us to actionable insights for targeting the right audience and enhancing the conversion rate.

Let us discuss the sequence of actions undertaken:

1. EDA:

We performed a rapid assessment of the percentage of null values and consequently eliminated columns with missing values exceeding 45%. Furthermore, we recognized that retaining rows containing null values would result in a substantial data loss, particularly given the significance of these columns. As an alternative approach, we opted to substitute the NaN values with 'not provided'.

Given that 'India' emerged as the predominant entry within the non-missing values, we proceeded to replace all instances of 'not provided' with 'India'. Subsequently, upon observing a substantial predominance of values attributed to India (approximately 97% of the dataset), we made the decision to remove this column. We also worked on numerical variables, outliers and dummy variables.

2. Train-Test split & Scaling:

The dataset was divided into a 70% training set and a 30% test set. Following this, we will apply min-max scaling to the following variables: ['TotalVisits', 'Page Views Per Visit', 'Total Time Spent on Website'].

3. Model Building:

Feature selection was conducted using Recursive Feature Elimination (RFE). Initially, RFE was employed to identify the top 15 relevant variables. Subsequently, the remaining variables were manually eliminated based on considerations such as VIF values and p-values.

Following this process, a confusion matrix was generated, and the overall accuracy was evaluated, yielding a result of 80.91%.

4. Model Evaluation:

- Sensitivity – Specificity

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

Summary of Leading Score

On Training Data:

I} The optimum cut off value was found using ROC curve. The area under ROC curve was 0.88.

2} 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:

I} We get

Accuracy 80.02%

Sensitivity 79.23%

Specificity 80.50%

- Precision – Recall:

If we go with Precision – Recall Evaluation

On Training Data:

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

2} 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%

Summary of Leading Score

Recall 76.32%

Prediction on Test Data:

We get

Accuracy 80.57%

Precision 74.87%

Recall 73.26%

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:
    - Total Visits
    - Total Time Spent on Website
- Lead Origin:
    - Lead Add Form
- Lead source:
    - Direct traffic

Summary of Leading Score

- Google
- Welingak website
- Organic search & Referral Sites

Last Activity:

- Do Not Email_Yes
- Last Activity_Email Bounced
- Olark chat conversation

The model demonstrates a strong capability to predict Conversion Rate, instilling a sense of confidence for the company to make informed decisions guided by this model's insights.