

SUMMARY

The model building and prediction is being done for company X Education and to find ways to convert potential users. The data provided gave us a lot of information about how the potential customers visit the site, the time they spend there, how they reached the site and the conversion rate.

The following are the steps used:

1. READING & UNDERSTANDING THE DATA

- First, we read the data to get a basic idea and checked shape and datatypes.

2.CLEANING DATA & HANDLING NULL VALUES:

- 'select' option was converted to null values.
- Columns having null values greater than 35% were dropped.
- Columns having null values <35% were treated by replacing them with median & mode values respective to their column types.

3.BINARY COLUMNS AND DATA IMBALANCE:

- Dropped columns having significant data imbalance.
- Dropped columns having unique values.

5.EXPLORATORY DATA ANALYSIS:

- Plotted box-plots and histograms for numerical columns and outliers were treated by capping at 99% percentile.
- Plotted count-plot for categorical columns with respect to target variable.

5.DATA PREPARATION & DUMMRY VARIABLE:

- Converted binary valued column to 0 and 1.
- Created dummy variable for categorical columns except binary one.

6. TRAIN-TEST SPLIT:

- The split was done at 70% and 30% for train and test data respectively.
- Scaling for numerical columns.
- Correlation was plotted and dropped few columns which were highly correlated.

7. MODEL BUILDING:

- 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 (The variables with $VIF < 5$ and $p\text{-value} < 0.05$ were kept).
- 4th model was selected.
- Predicted the possibility of conversion on train dataset.

8. MODEL EVALUATION:

- A confusion matrix was made. Later on, the optimum cut off value (using ROC curve) was used to find the accuracy, sensitivity and specificity which came to be around 78-80%.
- The optimum cut off value was found using ROC curve. The area under ROC curve was 0.86.
- After Plotting we found that optimum cutoff was **0.38** which gave
Accuracy 70.01%
Sensitivity 77.96%
Specificity 79.66%.

9. PREDICTION ON TEST DATASET:

- Scaled the numerical columns.
- With cutoff as **0.38** we got;
Accuracy 77.70%
Sensitivity 77.41%
Specificity 77.89%
Precision 76.18%
Recall 67.05%

10. PRECISION-RECALL TRADEOFF:

- With cutoff as **0.45** we get;
Accuracy 79.74%
Precision 73.73%
Recall 72.85%

11. PREDICTION ON TEST DATASET AGAIN:

- Prediction on test dataset with new cut off of 0.45
- We get:
Accuracy 78.84%
Precision 73.60%
Recall 71.60%

12. CONCLUSION:

With Precision, Recall and Accuracy of 70+ on test dataset, the Model seems to predict the Conversion Rate very well.

Features that mattered the most for potential buyers are;

What is your current occupation_Unemployed,Last Activity_Email Opened,Last Activity_SMS Sent,Total Time Spent on Website,Last Notable Activity_Modified,Lead Source_Olark Chat,Last Notable Activity_Other notable activities,Last Activity_Other activities,Do Not Email,Lead Source_Reference,Last Notable Activity_Olark Chat Conversation,Lead Source_Other Social Sites.

Suggestion:

- Three groups having conversion rate of 75%+(Hot lead/High priority),40-74% and <40% can be made, and pursued by sales team as per availability of resources.