Summary Report

Lead Scoring Case Study

STEP 1: Reading the Dataset

Importing and reading the dataset.

STEP 2: inspecting the Dataset

Checking dimensions. Their variables types, null values percentage in each of the variables. Along with some checks on the continuous data regarding their distribution.

STEP 3: Cleaning the Dataset

Removing all the columns with null value content <40%. Imputing the rest with the mode for some categorical variables wherever possible or can create a new sub category. In case of continuous variables all the missing values rows were removed as they were very low in count.

STEP 4: Data Analysis

Plotting all the categorical and continuous variables against the target variable and viewing their distribution. Looking for any underlying patterns. Also, to look out for the outliers in the continuous variables present remove them. Several categorical columns were also removed as they have imbalance in their data.

STEP 5: Data Preparation

Converting Yes/No categorical variable to binary values 0/1. Creating dummy variables for categorical variables with multiple categories.

STEP 6: Test-Train split

As per the general rule splitting the data into train and train in the ratio of 70:30.

STEP 7: Feature Scaling

All the continuous variables were scaled. Dividing the data of the variables with 0 mean and 1 as their variance. Looking for any imbalance in the target variable and checking the correlation of all the variables.

STEP 8: Looking for Correlation

Looking for any imbalance in the target variable and checking the correlation of all the variables and removing ones with high correlation between them.

STEP 9: Model Building

Building the first model using logistic regression and observing the initial P-values.

STEP 10: Feature selection using RFE

Using recursive feature elimination we select the 20 best features. Repeating the process iteratively until we get all the P-values below the threshold levels. After that checking all the other metrics like VIF and accuracy, sensitivity and specificity.

After building 9 models P-values and VIF were found within threshold with 13 features. Initially selecting 0.5 as the optimal cut-off accuracy of 78% was observed with sensitivity at 63% and specificity at 88%

STEP 11: Plotting the ROC Curve

ROC curve was plotted between Sensitivity and Specificity AUC was found 0.84 which is decent enough to proceed Further

STEP 12: Finding Optimal Cut-Off Point

Optimal cutoff point was found by plotting the curve between sensitivity, specificity and Accuracy. The intersection of all three values are found out to be at 0.3.

By using optimal cut-off point with accuracy of 79% sensitivity of 76% and specificity of 81% was observed on the predicted values in the training dataset.

STEP 12: Precision and Recall Tradeoff

Precision and Recall values are calculated as 76% and 63% respectively. After plotting the Precision-Recall Tradeoff cut-off value of 0.38.

STEP 12: Making Prediction on the test dataset.

Using the train set parameters prediction were made on the test data set. Final Accuracy of 79%, Sensitivity of 76% and specificity of 81% was observed.

Prepared By-

Gunjan Bhardwaj

Neha B

Rohit Singh