Summary Report for the Lead Scoring Case Study

Here the data consists of 37 columns and 9240 rows, and this data includes both numerical and categorical data, it consists of many missing values. This dataset consists ‘select’ keyword in some of the columns which can be considered as null values, we change all the select values to null values and then we check for percentage of null values in each column.

We remove columns with more than 70% missing values, we perform the missing value treatment for the columns with less missing values, and we remove the data with less than 2% missing values. After that we check for outliers on the numeric columns as the outliers are there in some columns we perform outlier treatment on the dataset. Along with that remove some of the columns with varying values and whose data doesn’t affect the dataset much.

After performing all the above mentioned tasks, for the columns with only yes or no type values we’ll do binary mapping with Yes as 1 and No as 0. For the rest of the categorical variables we’ll perform dummy variable creation, and then we’ll drop the 1st column among them.

Then we merge all the dummy variables with the original dataset and we remove the columns for which we had already created the dummy variables,

We split the data into train and test datasets and for the X_train dataset we scale the data by using standard scalar method. Then we first train the data and check the summary. Then we perform the RFE with 15 variables as output on the X_train dataset. We consider only the variable which comes under top list, and on that dataset we perform the training and check the summary. After removing the values with more p value, we perform the VFE on the data and check the colinearity between the variables. After training the data we check the predicted score, after that we’ll put a threshold on the score as the one which is greater than 0.5 is considered as 1 and the rest as 0. We check for the confusion matrix. WRT the values in the confusion matrix we calculate the accuracy, sensitivity, specificity, false positive rate (FPR), Negative predicted value and Positive predicted value.

Now we change the threshold from 0.1 to 0.9 and depending on the value we calculate the accuracy, sensitivity and specificity for various probabilities. Then we plot them on a graph to find the optimum threshold. From the above plot we can say that 0.2 is the optimum threshold.

We change the threshold to 0.2, hence we get a different prediction, the obtained data we again check the sensitivity, Specificity, FPR, Precision and Recall.

After performing the training we test the data on the test dataset and predict the value for by considering 0.2 as the threshold, after obtaining the final data we again create the confusion matrix and check the accuracy, sensitivity and specificity of the data.