Lead Case Study

This case study is for building a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potentials leads.

We started with reading the files and then inspecting it. We saw few columns had SELECT as the field entry so we replaced it with NAN value, we then observed the missing values and treated them by deleting rows for the columns which had lower missing values and deleting columns which had maximum values and no significance with the target variable and we even treated few columns of missing value by imputing them with the VALUE COUNT () function of python which impute the missing value of the column with the most highest number of value for that column.

We observed few outliers in the dataset so we treated it with IQR (Interquartile Range).

We converted YES/NO to 1/0 and created dummy variable for all categorical columns for easy modeling.

Now the dataset was divided into train and test data and scaling was done for both of the dataset.

We used RFE for feature selection in modeling, giving 40 columns number for selection.

Train data values:

Accuracy: 0.900990099009901

Sensitivity: 0.8459831335996449

Specificity: 0.934430652995143

All the steps were repeated for the test dataset too and the measures we got are:

Accuracy: 0.8931088488645262

Sensitivity: 0.8530954879328436

Specificity: 0.9169269206745784.

Finally we calculated the lead score against each lead ID.

Model was good with a Accuracy of 0.900990099009901 and 0.8931088488645262 for train and test data respectively.