Micro-CreditDefaulter

A MicroFinance Institution is an organization that offers financial services to low income population.We are working with a client that it is in telecome industry. The data was given to us to predict the term of probability of each loan transaction whether the customer will pay back in 5 days of insurance of loans.

In this case they have provided the data of Defaulter and NonDefaulter according to the data we have.The data has unrealistic values which they provided.The consumer belived to be defaulter if he drifts of paying back with in the time of duration of 5 days.

As they provided the data I have perform the steps and builded the model .The first I load the data from the Jupiter notebook and performed the appropriate steps.

I loaded the libraries of pandas ,numpy and visiulization .Then imported warnings library to don't get interrupted between the process.Then load the data from the folder.Then checked the shape of the data ,columns,information,and null values.The data had 209593 rows and 37 columns.The data had 3 object columns and all are discrete data and continuous data. The data had no missing values in the data set. Then check the Exploratory Data Variable of the data which I found mobile number and date,which were not providing any information of consumer loan taken or not.

Then checked the unique value of target variable that is label which provided the number 0 and 1,which indicates 0 as a defaulter and 1 has Non defaulter. Then I done the summary statistic which I got that some value are unrealistic like number of days last recharge of main account cant be 998650 days and number of days last reacharge of data account999171.The Next steps was to get correlation of target variable which I drop the columns and plot the bar to see correlation between target variable that is label.

The number of times main account got reacharge in last in last 30 days and the number of times main account got reacharge in last 90 days has highest positive correlation with target variable. Which shows that the consumer are paying the loans regularly.

Then I perform visualization to get clear idea of data.i plot the various columns and in label i found 183431 has paid the loan on time and 26162 has not paid the loan.Then I plot violin plot its shows that the daily amount spent main account ,average over last 90 days the probability is high.There is no limit of transaction.Maximum amount loans taken by user in last 30 days according to violin plot the probability is low.After the visiulization I found some unnecessary values are present in data so I drop them .

I drop unnamed,pcircle,pdate,msisdn,last_reach_date_ma,last_reach_date_da,payback30,payback90, cnt_loans90,amnt_loans90, maxamnt_loans30,medianamnt_loans30,cnt_da_rech90 ,'fr_da_rech30', and'cnt_loans30' which some data was unrealistic affecting my model by not providing unnecessary information.

The next step I plotted the outliers which model found many outliers are present in dataset.After that I applied zscore to remove outliers but the approximately 18% data was losing ,so don't wanted loose the data more 8 to 9% as data was valueable.So I have not removed the outliers and gone for next step.

Then I checked the skewness of the data which the data found skewness present in the data set. So for removing skewness I done the feature selection and transform the data with the algorithm power transform.

After head to next step to normalize the data with the help of standardscaler and proceded to Machine Building Process.By appling train_test_split ,split the data for training and testing .For training the model x_train 28499 rows and 21 columns and 4930 and 21 columns gone for testing.I applied four model library and three Boosting libraries.Which I found RandomForestClassifier has a best model with the score 0.69 of accuracy and 0.90 cross validation score,Also roc_curve,auc I found true positive rate high.