MICRO CREDIT PROJECT

BY

MOHIT MEEL

Acknowledgment

- The data is secondary as it is provided by the company, it is not collected from anywhere.
- The company has provided all the guidelines to complete the project .
- The data is in csv format.
- The data is secondary therefore no researches been made.
- The company has provided sample documentation and description of data in order to complete the project and also to study the project.

Introduction

The problem of business is to improve the selection and also to predict the customers which are going to pay back the loan amount within the days limit for further investment. There are so many customers which take loan from the financial institutions but not able to pay or doesn't pay the loan amount so it is necessary for the institution to analyze or predict the customers while giving the loan .

In today's world , the microfinance institution helps the poor families of remote areas in providing the financial assistance. The poor families of remote areas are not able to take financial help from banking sector due to improper documents. The micro finance institutions faces the major issue in the repayment of the loaned amount due to which the institutions faces loss therefore it is necessary to select the honest and reliable customers in providing financial assistance. The customers not able to pay or doesn't pay the loan amount which makes the institutions to predict the customers that are going to pay.

The research done on the project is that the customers willing to take loan for financial help , pay the loan amount and some don't. The institution which provides financial help wants that every loan amount should be paid off within the days limit given by the customers.

The objective behind this project is to prevent losses occurred by the wrong selection of customers which don't pay the loan amount on the time and to select those customers which are honest to their payments on time. To make investment in lending the money to poor and needy families.

Analytical Problem Framing

- The dataset provided by the client database is in csv format. It is huge data as it comprises of 209593 rows and 37 columns. Most of column sin the dataset is numerical only three are of categorical . The output variable is of discrete. There are no null values present in the dataset. The categorical columns are converted into numeric by label encoder for not losing important information in respect to prediction. The statistical view of the data is that the mean of the columns is higher or equal(one or two columns) to the median of the columns. The most of the columns standard deviation is higher than its mean. The correlation between the columns is moderate, the three columns which shows very less relation with target variable so they are dropped out of the dataset. The outliers present is checked with the help of boxplot and came to know that outliers are present but after removing the outliers from the dataset, the data loss % is 22% which means that the accuracy of data is being compromised and though the dataset is becoming biased if we remove the outliers. The acceptable range of data loss is 7-8%. So the outliers present are not removed. The skewness present in the dataset is checked with the histplot and thus removed with the power transform but after removing there is one columns which shows highly negative skewness so it is been removed from the dataset. The target variable has class imbalance which is treated with the help of smote. After balancing the target variable, the data now has 366862 rows and 32 columns. At last before sending the feature and target column for training & testing and for prediction modelling, the dataset is scaled with the standard scaler for better learning of the model by the machine.

Models Development and Evaluation

The training and testing is now started after the cleaning of data, for training 80% data has been send for machine and for testing 20%. The random state is 59 at which the highest accuracy is achieved by the logistic regression model which is 76.7%. Different models used to predict the target variable but before predicting I have seen the accuracy score of such models and that model is selected which has scored highest accuracy among the predicted models.

The first model is logistic which has given the score of 76.7% at random state of 59.

The second model is decision tree classifier which has scored of 91.5% , that is a nice accuracy score.

The third model which I tried is random forest classifier, this model scores the highest among all that is 95.4% which is a good accuracy score.

There were some models too which I tried but I didn't come up with accuracy score and some shown error due to the huge dataset, those models names are kneighbors , SVC, adaboost , gradientboosting.

The roc auc score of the decision tree classifier and random forest classifier is also done and thus the roc curve pot is also implemented.

The best model for this project for predicting the target variable is random forest classifier.

Conclusion

After knowing the best model that comes up in all models is random forest classifier. The most important facts that come up while studying and analyzing the dataset was that the credit facility given to those customers only which are able to pay off the credit within the time limit plays an important role in terms of preventing losses. The target variable of the dataset which was predicted by the machine is same as the actual label given by the machine by the help of random forest classifier model in most of the rows. There are so many observations which I found that most of the columns are moderate correlated to the target variable. The outliers present if removed then 22% of data loss was there in removing the outliers which is not right for the accuracy of the dataset. One column was there which was highly negatively skewed thus it is removed.

The visualization part is done basically for the checking of the outliers , seeing the correlation and for the checking of the skewness.

The solution for the company is to predict it customers with more of the attributes too such as their income, their daily expenses, how much amount of loan they take, the financial situation, and some relevant extra information about the family, what is their work, etc. To first customers, the financial institutions should give small amount of loan so that if the loan amount is not paid then the losses of the institution is not at that level which was there with the customers.