# **MICRO-CREDIT-DEFAULTER-MODEL**

**Submitted by:**

**NILOTPAL MUKHERJEE**

PROBLEM STATEMENT:

A telecom company has tied up with a MFI and have decided to help customers by providing a small amount of loan as credit on the mobile balance and the payback time is 5 days. As MFIs have turned out to be a great force in the field of economics as it serves as an effective tool for the low income group. The telecom industry is of great importance especially as it can reach out to the common masses. So the telecom industry coupled with the MFI can prove to be really effective tool reach out more amount of people and in a quicker and cleaner way and provide support to the really needy. But for that to happen the company needs to be able to make better business decisions when it comes to lending money to the people. It would be helpful for the company could somehow be able to identify the customers to whom they should be giving the loans to, i.e., they want to identify the defaulters from potential customers (people paying back the loan amount). This would prove to viable to sustain the business as it would help the companied to reach out the potential people and in turn keep the wheel of economy moving. So it becomes important to come up with a solution that will help the telecom services to separate the defaulters from the non-defaulters. A way subtle way to predict form the information already available with company seems a viable approach.

The company wants to explore a way to predict out defaulters from non-defaulters for future endeavours based on the current data they have collected. So as this project deals in such a sector the telecom company already have a way to differentiate among the each and every one of its customers within their network via the phone numbers. It also has a way to differentiate among the people in the network based on who is actively using the network to make calls and who is using it less or not using it at all. Along with this a lot of other features like how many times the person recharges the account and what is the gap in between each recharge. Also which of its customers are paying back the loan amount and who are not paying back along with frequency of paybacks in a span of time like 30-90 days. So these data can help in predicting the future behaviours of the customers based on the current behaviour and this is a substantial for the telecom company in a way that is similar to the banking industry.

ANALYSIS OF THE DATASET AND REMARKS

The dataset contained 209593 rows and 37 columns.

The columns present are as follows:

Label-------------> Flag indicating whether the user paid back the credit amount within 5 days of issuing the loan{1:success, 0:failure}

msisdn-------------> mobile number of user

aon -------------> age on cellular network in days

daily\_decr30 -------------> Daily amount spent from main account, averaged over last 30 days (in Indonesian Rupiah)

daily\_decr90-------------> Daily amount spent from main account, averaged over last 90 days (in Indonesian Rupiah)

rental30-------------> Average main account balance over last 30 days

rental90-------------> Average main account balance over last 90 days

last\_rech\_date\_ma-------------> Number of days till last recharge of main account

last\_rech\_date\_da-------------> Number of days till last recharge of data account

last\_rech\_amt\_ma-------------> Amount of last recharge of main account (in Indonesian Rupiah)

cnt\_ma\_rech30 -------------> Number of times main account got recharged in last 30 days

fr\_ma\_rech30-------------> Frequency of main account recharged in last 30 days

sumamnt\_ma\_rech30-------------> Total amount of recharge in main account over last 30 days (in Indonesian Rupiah)

medianamnt\_ma\_rech30-------------> Median of amount of recharges done in main account over last 30 days at user level (in Indonesian Rupiah)

medianmarechprebal30 Median of main account balance just before recharge in last 30 days at user level (in Indonesian Rupiah)

cnt\_ma\_rech90-------------> Number of times main account got recharged in last 90 days

fr\_ma\_rech90-------------> Frequency of main account recharged in last 90 days

sumamnt\_ma\_rech90-------------> Total amount of recharge in main account over last 90 days (in Indonasian Rupiah)

medianamnt\_ma\_rech90-------------> Median of amount of recharges done in main account over last 90 days at user level (in Indonasian Rupiah)

medianmarechprebal90 Median of main account balance just before recharge in last 90 days at user level (in Indonasian Rupiah)

cnt\_da\_rech30 -------------> Number of times data account got recharged in last 30 days

fr\_da\_rech30-------------> Frequency of data account recharged in last 30 days

cnt\_da\_rech90-------------> Number of times data account got recharged in last 90 days

fr\_da\_rech90-------------> Frequency of data account recharged in last 90 days

cnt\_loans30-------------> Number of loans taken by user in last 30 days

amnt\_loans30-------------> Total amount of loans taken by user in last 30 days

maxamnt\_loans30-------------> maximum amount of loan taken by the user in last 30 days

medianamnt\_loans30-------------> Median of amounts of loan taken by the user in last 30 days

cnt\_loans90-------------> Number of loans taken by user in last 90 days

amnt\_loans90-------------> Total amount of loans taken by user in last 90 days

maxamnt\_loans90-------------> maximum amount of loan taken by the user in last 90 days

medianamnt\_loans90-------------> Median of amounts of loan taken by the user in last 90 days

payback30-------------> Average payback time in days over last 30 days

pcircle -------------> telecom circle

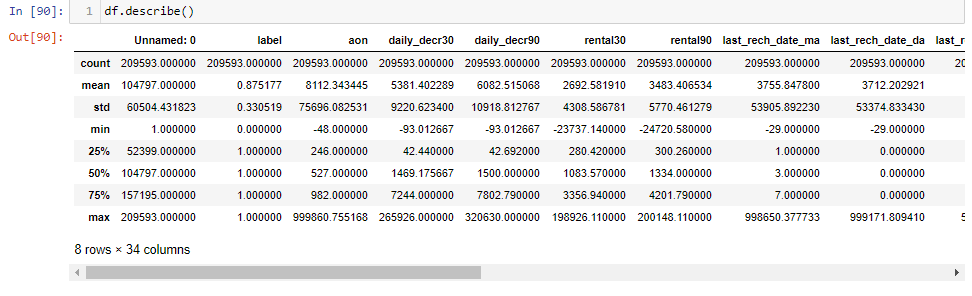
pdate-------------> date

For a grouped dataset with so much information statistical approach was done. Firstly the mean, standard deviation(std.), median was calculated for checking distribution of the data. For all the datapoints that were not in alignment with desired std. ,zscore values were calculated simply by finding the difference between each datapoint and the mean and dividing the difference by the std. and later on removing all the values more than 3 . After which correlations among each of the feature was explored and depicted with help of a heatmap that plots values based on the correlation values. The ‘label’ column proved to be the dependent feature(Y) and the multiple independent features (X) and it was clear that it was a classification problem that could be represent the relationship of the X and Y values through a sigmoid curve and for that reason logistic regression model was trained to predict the values. As the dataset contained a lot of outliers the best approach would be to classify the dataset with the help of Decision tree classifier as it breaks up into multiple nodes and leaves to reach a result so it is generally not affected by outliers. Random Tree classifier is another such algorithm that is going to give proper results as it consists of multiple decision trees so it gives better result.

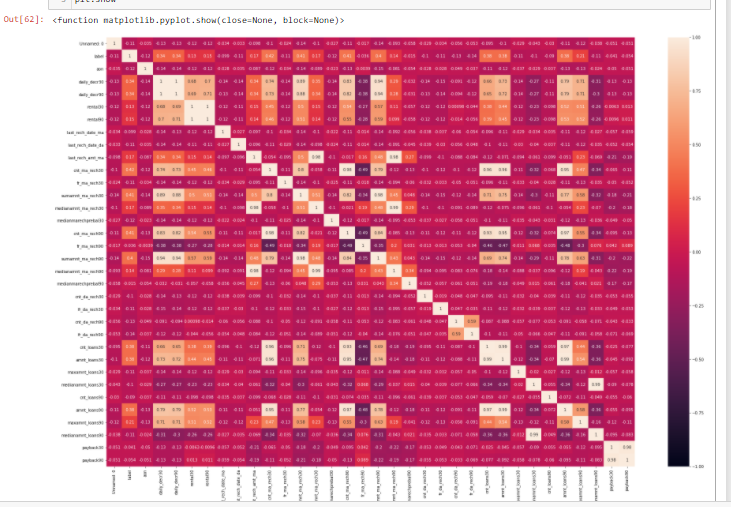
For predicting the dataset, the following libraries were used –

So it can be assumed that in a machine with better processor it might not happen. Libraries used for the project

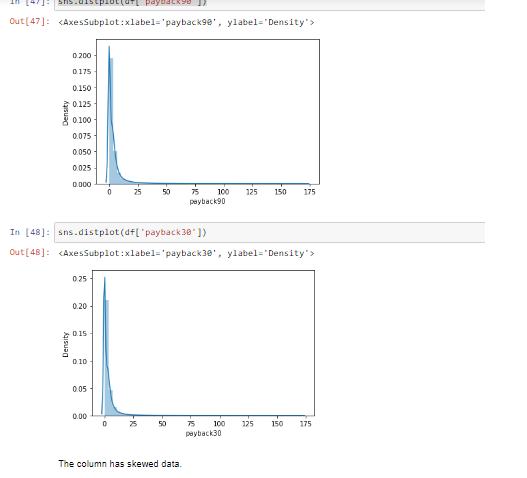
1. Pandas dataframe
2. Numpy array
3. Matplotlib.pyplot(boxplot)
4. Seaborn (countplot, heatmap)
5. Scipy.stats.zscore
6. Imblearn.over\_sampling.SMOTE
7. sklearn.preprocessing .MinMaxScaler

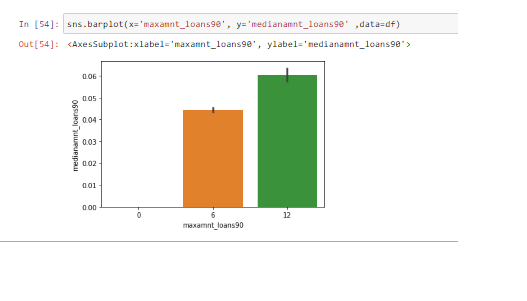


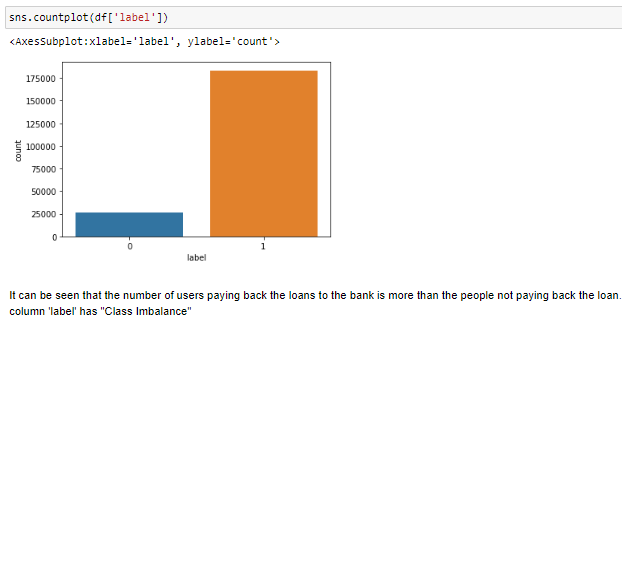
THE DATA WAS CHECKED FOR CORRELATIONS VIA PLOTING WITH A HEATMAP.



THE HEATMAP SHOWS THAT THERE ARE MULTIPLE CORELATIONS IN BETWEEN VARIOUS INDEPENDENT AND TARGET VARIABLES. THERE ARE DATAPOINTS SHOWING HIGH CORTELATIPON VALUES AND ATAPOINTS SHOWING LOW CORELATION VALUES .

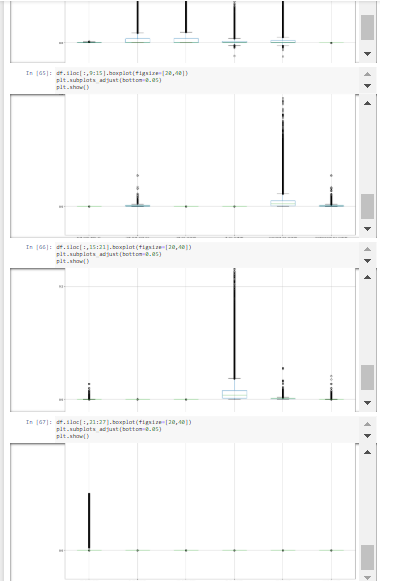






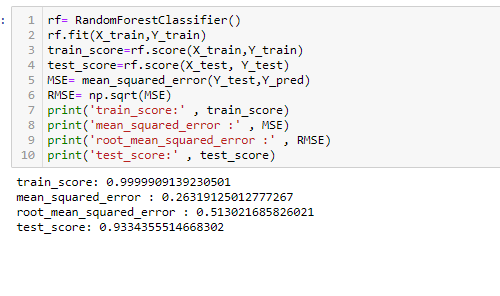
THE DISTPLOT HELPED TO IDENTIFY THAT THE NUMBER OF NONE PAYERS OF LOAN IS VERY M UCH MORE THAN THE NUMBER OF PEOPLE PAYING BACK THE LOAN. IT LSO PROVD THAT THE TARGET VARIABLE HAS A CLASS IMBALANCE.

THE BOX-PLOTS SHOWS THAT THERE ARE NUMEROUS OUTLIERS PRESENT THIS SHOWS THAT THE DATA DISTRIBUTION MIGHT BE SKEWED IN NATURE AND THAT ITS STANDARD DEVIATION WILL BE MORE THAN THE DESIRED +1/-1 RANGE.

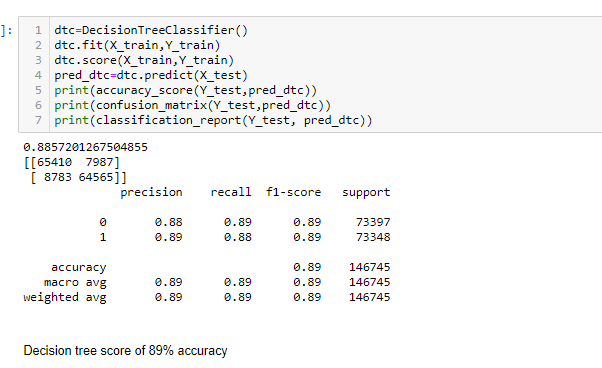


MODEL BUILDING

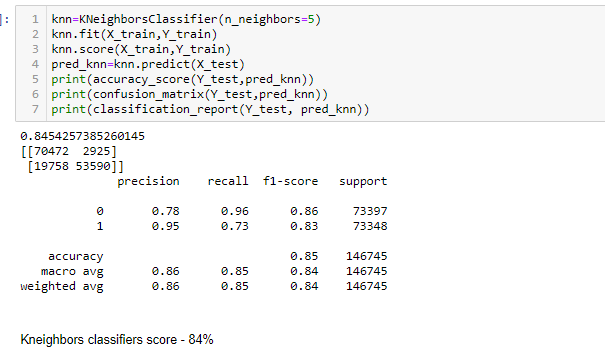
Random forest classifier - A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. It gave a score of 93% accuracy which was the highest among all the scores.



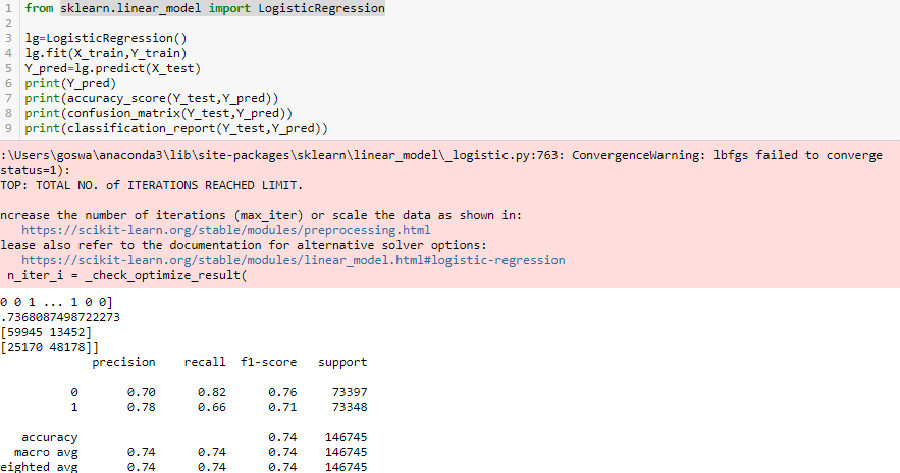
Decision Tree - non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation.



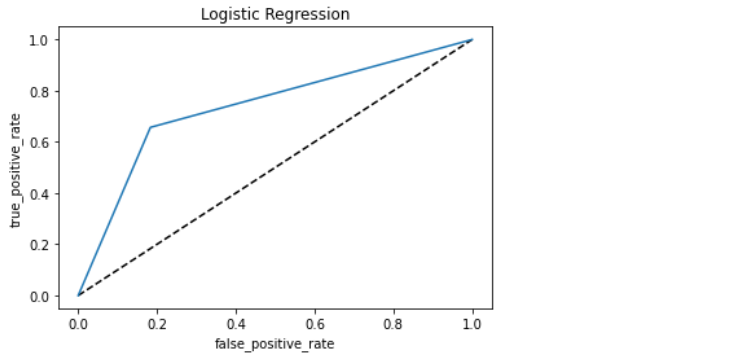
Kneighbors classifiers - implements learning based on the k nearest neighbors of each query point, where k is an integer value specified by the user



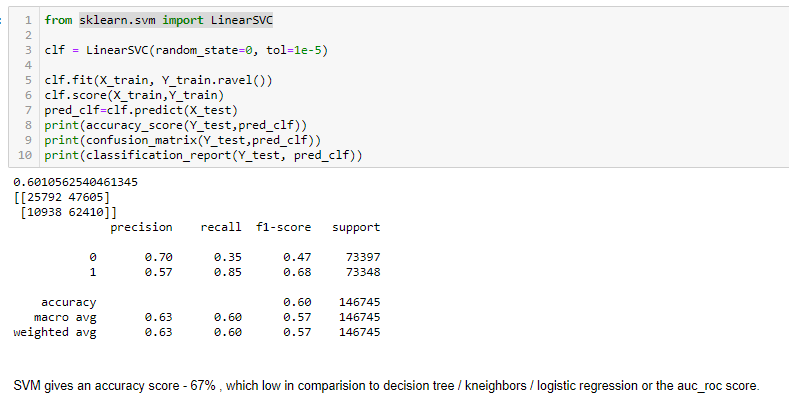
Logistic Regression- is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable.



AUC\_ROC -   is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the **roc curve**. The higher the **auc** the better the performance of the model at distinguishing between the positive and negative classes.



SUPPORT VECTOR MACHINE-A support vector machine (**SVM**) is a supervised machine learning model that uses classification algorithms for two-group classification problems.



* Key Metrics for success in solving problem under consideration

1. RANDOM FOREST- 93%
2. DECISION TREE - 89%
3. KNEIGHBORS- 84%
4. LOGISTIC REGRESSION - 74%
5. AUC\_ROC - 73%
6. SVM - 67%

Final thoughts

It was a large dataset having -

209593 rows x 37 columns

The dataset is based on real world data hence it was having a lot of outliers. Removing the outliers would have resulted in removal of 20%+ of the data that would have resulted in the formation of a biased model.

Out of 6 evaluation metrics the random-forest classifier gave the best results. The reason for it being as it creates multiple decision trees as part of its ensemble techniques it can really work with a dataset that is having a huge number of outliers.

As the dataset was containing outliers during the testing and training phase of the data it might vary during actual ground work. The Random forest though ended up with a high score in this particular case it cannot be considered as the ultimate model for prediction as the ensemble techniques for the random forest allows it work with outliers. This project has many real world applications especially in the financial fields. The banking industry is always on the lookout for potential customers to give loan to but they would not want to end up on a trade that would result in losing out money. The debt situation in an economy always helps to move the economy forward but if there are a lot defaulter then it might prove to be a financial disaster. So this projects not only provides insight into the pattern of human behaviour but also provides a way out of a financial crisis. It will be interesting to observe the relationship between each user individually can be singled out with the help of their phone numbers.