**INTRODUCTION**

A Microfinance Institution (MFI) is an organization that offers financial services to low income populations. MFS becomes very useful when targeting especially the unbanked poor families living in remote areas with not much sources of income. The Microfinance services (MFS) provided by MFI are Group Loans, Agricultural Loans, Individual Business Loans and so on.

Many microfinance institutions (MFI), experts and donorsare supporting the idea of using mobile financial services (MFS) which they feel are more convenient and efficient, and cost saving, than the traditional high-touch model used since long for the purpose of delivering microfinance services. Though, the MFI industry is primarily focusing on low income families and are very useful in such areas, the implementation of MFShas been uneven with both significant challenges and successes. Today, microfinance is widely accepted as a poverty-reduction tool, representing $70 billion in outstanding loans and a global outreach of 200 million clients.

We are working with one such client that is in Telecom Industry. They are a fixed wireless telecommunications network provider. They have launched various products and have developed its business and organization based on the budget operator model, offering better products at Lower Prices to all value conscious customers through a strategy of disruptive innovation that focuses on the subscriber.

They understand the importance of communication and how it affects a person’s life, thus, focusing on providing their services and products to low income families and poor customers that can help them in the need of hour.

They are collaborating with an MFI to provide micro-credit on mobile balances to be paid back in 5 days. The Consumer is believed to be a defaulter if he deviates from the path of paying back the loaned amount within the time duration of 5 days. For the loan amount of 5 (in Indonesian Rupiah), payback amount should be6(in Indonesian Rupiah), while, for the loan amount of 10(in Indonesian Rupiah), the payback amount should be 12(in Indonesian Rupiah).

The sample data is provided to us from our client database. It is hereby given to you for this exercise. In order to improve the selection of customers for the credit, the client wants some predictions that could help them in further investment and improvement in selection of customers.

##### **Feature Description:**

1. label :Flag indicating whether the user paid back the credit amount within 5 days of issuing the loan{1:success, 0:failure}
2. msisdn :mobile number of user
3. aon :age on cellular network in days
4. daily\_decr30 :Daily amount spent from main account, averaged over last 30 days (in Indonesian Rupiah)
5. daily\_decr90 :Daily amount spent from main account, averaged over last 90 days (in Indonesian Rupiah)
6. rental30 :Average main account balance over last 30 days
7. rental90 :Average main account balance over last 90 days
8. last\_rech\_date\_ma :Number of days till last recharge of main account
9. last\_rech\_date\_da : Number of days till last recharge of data account
10. last\_rech\_amt\_ma : Amount of last recharge of main account (in Indonesian Rupiah)
11. cnt\_ma\_rech30 : Number of times main account got recharged in last 30 days
12. fr\_ma\_rech30 : Frequency of main account recharged in last 30 days
13. sumamnt\_ma\_rech30 : Total amount of recharge in main account over last 30 days (in Indonesian Rupiah)
14. medianamnt\_ma\_rech30 : Median of amount of recharges done in main account over last 30 days at user level (in Indonesian Rupiah)
15. medianmarechprebal30 : Median of main account balance just before recharge in last 30 days at user level (in Indonesian Rupiah)
16. cnt\_ma\_rech90 : Number of times main account got recharged in last 90 days
17. fr\_ma\_rech90 : Frequency of main account recharged in last 90 days
18. sumamnt\_ma\_rech90: Total amount of recharge in main account over last 90 days (in Indian Rupee)
19. medianamnt\_ma\_rech90 :Median of amount of recharges done in main account over last 90 days at user level (in Indian Rupee)
20. medianmarechprebal90 : Median of main account balance just before recharge in last 90 days at user level (in Indian Rupee)
21. cnt\_da\_rech30 : Number of times data account got recharged in last 30 days
22. fr\_da\_rech30 : Frequency of data account recharged in last 30 days
23. cnt\_da\_rech90 : Number of times data account got recharged in last 90 days
24. fr\_da\_rech90 : Frequency of data account recharged in last 90 days
25. cnt\_loans30 : Number of loans taken by user in last 30 days
26. amnt\_loans30 : Total amount of loans taken by user in last 30 days
27. maxamnt\_loans30 : maximum amount of loan taken by the user in last 30 days
28. medianamnt\_loans30 : Median of amounts of loan taken by the user in last 30 days
29. cnt\_loans90: Number of loans taken by user in last 90 days
30. amnt\_loans90 :Total amount of loans taken by user in last 90 days
31. maxamnt\_loans90 : maximum amount of loan taken by the user in last 90 days
32. medianamnt\_loans90: Median of amounts of loan taken by the user in last 90 days
33. payback30 :Average payback time in days over last 30 days
34. payback90: Average payback time in days over last 90 days
35. pcircle: telecom circle
36. pdate :date

### **Data Preprocessing:**

### Remove columns where number of unique value is only 1.

We have looked at no. of unique values for each column.We will remove all columns where number of unique value is only 1 because that will not make any sense in the analysis.

Observation:

* Summary statistics shows all the statistics of our dataset i.e. mean, median and other calculation.
* Mean is greater than median in all the columns so aur data is right skewed.
* The difference between 75% and maximum is higher that's why outliers are removed which needs to be removed.
* The pdate column tells the date when the data is collect. It contains only three month data.
* msidn is a mobile number of user and mobile number is unique for every customers. There are only 186243 unique number out of 209593 so rest of the data is duplicates entry so we have to remove those entry.

### **Data Exploration:**

Observation:

* It contains only one circle area data. So it have not any impact in our model if we drop this feature.
* After seeing the label column which is also our target feature for this dataset it is clearly shown that 86.11% of data is label 1 and only 13.8% of data is label 0 so our dataset is implanced. So before making the ML model first we have to do sampling to get rid off imblance dataset.
* Daily\_decr30 and daily\_decr90 features are highly correlated with each other.
* Rental30 and rental90 features are highly correlated with each other.
* Cnt\_loans30 and amount\_loans30 columns are highly correlated with each other.
* Amount\_loans30 is also highly correlated with amount\_loans90 column.
* Medianamnt\_loans30 and medianamnt\_loans90 is highly correlated with each other.
* We have to drop one of the features which are highly correlated with other features. And if we dont do this then our model will face a multicollinearity problem.

### **Data Visualization:**

Observation:

* Here we see the correlation of the columns with respect to the target column that is label.
* Label 1 indicates loan has been paid i.e Non-Defaulter and label 0 indicates indicates that the loan has not been paid i.e. defaulter.
* We plot the histogram to display the shape and spread of continuous sample data.In a histogram, each bar groups numbers into ranges. Taller bars show that more data falls in that range.
* The first figure which is date vs label shows that the customers who did not pay their loans are from date 10 to 23.
* There are severals customers at June and July month who did not pay their loan.
* There are too many outliers present in our dataset.So we need to remove it. But before removing please check that only 8 to 10% of data removed.
* By looking at the daily\_decr90 which is Daily amount spent from main account, averaged over last 90 days (in Indonesian Rupiah), it seems that this feature helps to discriminate the data indeed. This feature can bring insights for company when analyzing a customers.

### **Model Training:**

We have used KNeighborsClassifier, LogisticRegression, DecisionTreeClassifier, GaussianNB and RandomForestClassifier.