

# PROJECT REPORT

# ON

# MICRO CREDIT DEFAULTER -PREDICATION & ANALYSIS PROJECT

Submitted by:

**Ms. Yashshree Baviskar**

**FLIPROBO SME:**

**Mr. Mohd Kashif**

**ACKNOWLEDGMENT**

I would like to express my special gratitude to “Flip Robo” team, who has given me this opportunity to deal with a beautiful dataset and it has helped me to improve my analyzation skills. And I want to express my huge gratitude to Mr. Mohd Kashif (SME Flip Robo), he is the person who has helped me to get out of all the difficulties I faced while doing the project.

A huge thanks to “Data trained” who are the reason behind my Internship at Fliprobo. Last but not least my parents who have been my backbone in every step of my life.

References use in this project:

1. Does Microfinance Affect Poverty Reduction and Inequality in Indonesia?

<https://www.ijstr.org/final-print/apr2019/Does-Microfinance-Affect-Poverty-Reduction-And-Inequality-In-Indonesia.pdf>

1. Transformative Technology in Microfinance: Delivering Hope Electronically?

<https://www.researchgate.net/publication/305875278_Transformative_Technology_in_Microfinance_Delivering_Hope_Electronically>

1. A Brief Overview of Outlier Detection Techniques.

<https://towardsdatascience.com/a-brief-overview-of-outlier-detection-techniques-1e0b2c19e561>

1. How to remove outliers for machine learning:

<https://machinelearningmastery.com/how-to-use-statistics-to-identify-outliers-in-data/>

1. Data: How to handle Imbalanced Classification Problems

<https://www.analyticsvidhya.com/blog/2017/03/imbalanced-data-classification/>

1. Top 10 Data Visualization Techniques, Concepts & Methods In Business (datapine.com)

<https://www.datapine.com/blog/data-visualization-techniques-concepts-and-methods/>

1. Performance Metrics for Classification Machine Learning Problems

<https://towardsdatascience.com/performance-metrics-for-classification-machine-learning-problems-97e7e774a007>

# Accelerating Indonesian microfinance with high tech and high touch

<https://asianbankingandfinance.net/banking-technology/commentary/accelerating-indonesian-microfinance-high-tech-and-high-touch>

**Chap. 1 INTRODUCTION**

This study was launched to explore how microfinance institutions (MFIs) treat Micro finance clients who are unable to repay their loans. It was motivated by a dearth of information on the client-facing actions MFIs take when a borrower moves into default.

The microfinance industry’s progress on prioritizing client protection, there is a surprising lack of shared information on what happens when the contractual relationship breaks down between the lender and the borrower. While there are many studies on the causes of over-indebtedness, there are far fewer on the consequences. A multitude of decisions occur between the time when the MFI first observes a delinquency and when it makes the determination that the lender-borrower relationship is irreparably broken and terminates contact with the client.

This analysis attempts to understand the practices, and whether and how they are influenced by the market context, regulation or incentives which influence MFIs. The timing of the study coincides with an increased awareness of client protection as fundamental to microfinance. During the early years of microfinance, providers did not openly discuss sensitive internal practices like collections or what happened to clients who defaulted. Now, there is more willingness to discuss default management, due in large measure to the emergence of industry codes of conduct, various high-profile repayment

crises, and the MFIs’ real desire to learn how their peers are handling the issue. Nonetheless, the research team is deeply grateful to the participants in this survey for their openness to discussing their practices, and their trust in us to preserve their anonymity.

**What Are the Responsibilities of MFIs?**

The intent of this study is to inspire MFIs to reconsider their treatment of clients in default, and specifically, to define acceptable parameters for humane treatment of clients during the default process. Clients who are facing default may very well be in the most precarious financial position of their lives, which is often accompanied by an equally fragile mental state. Therefore, the obligation falls squarely on MFIs to handle these clients with care.

* 1. ***BUSINESS PROBLEM FRAMING***

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 donors are 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 MFS has 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.

* 1. ***CONCEPTUAL BACKGROUND OF THE DOMAIN PROBLEM***

Telecom Industries 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 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 be 6 (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. We have to build a model which can be used to predict in terms of a probability for each loan transaction, whether the customer will be paying back the loaned amount within 5 days of insurance of loan. In this case, Label ‘1’ indicates that the loan has been paid i.e., non-defaulter, while, Label ‘0’ indicates that the loan has not been paid i.e., defaulter.

***REVIEW OF LITERATURE***

**What is Microfinance?**

“Microfinance” is often seen as financial services for poor and lowincome clients (Ayayi, 2012; Mensah, 2013; Tang, 2002). In practice, the term is often used more narrowly to refer to loans and other services from providers that identify themselves as “microfinance institutions” (MFIs) [Consultative Group to Assist the Poor (CGAP) 2010]. Microfinance can also be described as a setup of a number of different operators focusing on the financially under-served people with the aim of satisfying their need for poverty alleviation, social promotion, emancipation, and inclusion. Microfinance institutions reach and serve their target market in very innovative ways (Milana 2012). The CGAP (2010) identifies some unique features of microfinance as follows: ➢ Delivery of very small loans to unsalaried workers ➢ Little or no collateral requirements ➢ Group lending and liability ➢ Pre-loan savings requirement ➢ Gradually increasing loan sizes Implicit guarantee of ready access to future loans if present loans are repaid fully and promptly Microfinance is seen as a catalyst for poverty alleviation, delivered in innovative and sustainable ways to assist the underserved poor, especially in developing countries (Dixon, Ritchie, & Siwale, 2007; Spiegel, 2012). Economic development may be achieved by helping the underserved poor to engage in income-generating/poverty reduction activities through entrepreneurship (Milana 2012). On December 18, 1997, the United Nations (UN) passed a microcredit resolution, also known as the Grameen Dialogue of 1998 at its General Assembly. The resolution was adopted because of the importance of microcredit programs in poverty reduction (Elahi & Demopoulos 2004). The UN later declared the year 2005 as International Year of Micro Credit. Globally, Microfinance has become an important sector. It is estimated that more than 3,500 institutions are meeting the demands of 205 million clients with a volume that is still uncertain but substantial (Maes and Reed 2012). Default in Microfinance Default in microfinance is the failure of a client to repay a loan. The default could be in terms of the amount to be paid or the timing of the payment. MFIs can sustain and increase deployment of loans to stimulate the poverty reduction goal if repayment rates are high and consistent (Wongnaa 2013). Machine Learning Techniques for microfinance & finance Pollio and Obuobie [ ] applied logistic regression on four factors and concluded that the probability of default increases with the number of dependents, whether the proceeds are used to acquire fixed assets, the frequency of monitoring, decreases with the availability of non-business income, years in business, the number of guarantors, whether the proceeds were used for working capital purposes and whether the client is a first-time borrower. In Addo et al. (2018) the authors examined credit risk scoring by employing various machine and deep learning techniques. The authors used binary classifiers in modelling loan default probability (DP) estimations by incorporating ten key features to test the classifiers’ stability by evaluating performance on separate data. Their results indicated that the models such as the logistic regression, random forest, and gradient boosting modelling generated more accurate results than the models based on the neural network approach incorporating various technicalities. Machine learning-based systems are growing in popularity in research applications in most disciplines. Considerable decision-making knowledge from data has been acquired in the broad area of machine learning, in which decision-making tree-based ensemble techniques are recognized for supervised classification problems. Classification is an essential form of data analysis in data mining that formulates models while describing significant data classes (Rastogi and Shim 2000). Accordingly, such models estimate categorical class labels, which can provide users with an enhanced understanding of the data at large Han et al. (2012) resulted in significant advancements in classification accuracy.

***MOTIVATION FOR THE PROBLEM UNDERTAKEN:***

The project was the first provided to me by Flip Robo Technologies as a part of the internship programme. The exposure to real world data and the opportunity to deploy my skillset in solving a real time problem has been the primary motivation. Further diving into the dataset, the motive is to help the poor or low-income band to have continuous access to their mobile accounts, and to make emergency calls even when they do not have account balance making use of the loan facility

This project was highly motivated project as it includes the real time problem for Microfinance Institution (MFI), and to the poor families in remote areas with low income, and it is related to financial sectors, as I believe that with growing technologies and Idea can make a difference, there are so much in the financial market to explore and analyze and with Data Science the financial world becomes more interesting.

The objective of the project is to prepare a model based on the sample dataset that classifies all loan defaulters and help our client in further investment and improvement in selection of customers.

**Chap 2. Analytical Problem Framing**

**MATHEMATICAL/ ANALYTICAL MODELING OF THE PROBLEM**

The dataset is a csv file with 37 attributes (36 features and 1 target). The target variable is either 1 or 0 which means non defaulter and defaulter respectively. The other key attributes are the account balances, days since last recharge, age on network, median recharge balance for 30 and 90 days and many more. The similar attributes for 30 and 90 days are highly correlated and conveys the same. Hence for the purpose of the project, highly correlated attributes needs to be removed

The statistical figure I get to know by the .describe() so many information the min max standard deviation the 25 percentile the 50th percentile the 75 percentile. Then by the help of correlation function I get to know the correlation of each feature with each other. From the heatmap I can visualized to see them clearly that they are positive correlated or the negative correlated the dark side is show the negative correlation among each other the lighter side represent the positive correlation among the each other

From an initial statistical overview of the dataset, we infer that some data features are binary or ordinal, whereas other features are continuous. Further, the minimum is negative which is not even possible for most of the features notably daily recharge, main account balance, aon, and last recharge which can't be negative and maximum values for some features, notably for aon, maxamnt\_loans30, medianmarechprebal90, medianmarechprebal30 are unrealistic. Most the features have mode is greater than median this suggests the presence of outliers in the data and All Features are not Normally Distributed (Theatrically if feature is normally distributed, Mean = Median = Mode) like weight and height are right and left skewed.

The Dataset we are having consists of some features giving information about the user for the time span of 30 days and 90 days. According to me if we have data of large number of days for a particular user then we could interpret User's behavior more precisely because many users have the tendency of repeating the same things. Thus, the features having the data with a time span of 90 days gives more information about the user as compared to the features with a time span of 30 days. All the categories that is being made to make the visualizations easy are solemnly based on the Description i.e. statistical summary of the data plotted above for instance low comes under (0-25%), average comes under (25-75%) and high comes over 75% of the data values in a given feature

Using MS EXCEL, we have found the maximum values a feature can have, beyond these values the values are unimaginable. For an example beyond the value [2500], the very next value in "aon" feature comes out to be around 2379 years, which means a user is using the telephone services from 359 BCE which is clearly not possible.

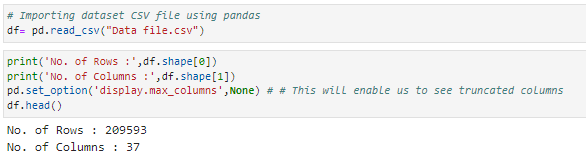
we checked the correlation of the independent and dependent features and from the correlation table it is also clear that the features with time span of 30 and 90 days almost have the same correlation thus we can drop one for the same information.

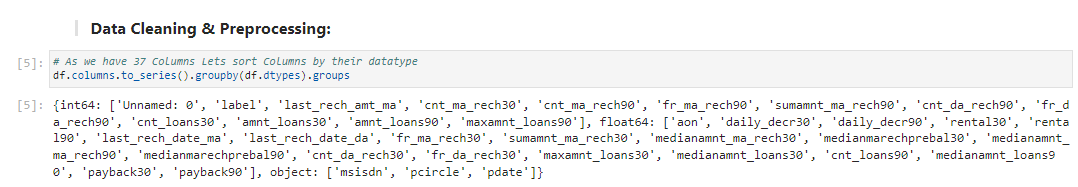
**DATA SOURCES AND THEIR FORMATS**

The given data was in CSV (Comma Separated Values) format. In the dataset there were 209593 rows and 37 features.

The data descriptions are as follow: -

|  |  |
| --- | --- |
| Variable | Definition |
| 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 Indonesian Rupiah) |
| medianamnt\_ma\_rech90 | Median of amount of recharges done in main account over last 90 days at user level (in Indonesian Rupiah) |
| medianmarechprebal90 | Median of main account balance just before recharge in last 90 days at user level (in Indonesian 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 |
| payback90 | Average payback time in days over last 90 days |
| pcircle | telecom circle |
| pdate | date |





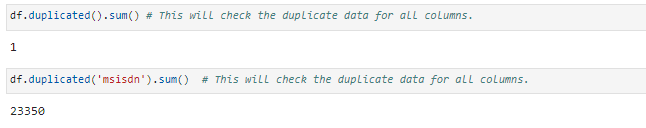
The different datatypes of these features are as shown in above figure. Out of all features only three features with object datatypes and rest are int64. We can note here ‘pdate’ has datatype of object instead of datetime datatype.

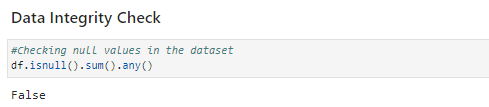
**3. DATA PRE-PROCESSING**

The dataset is large and it may contain some data error. In order to reach clean, error free data some pre-processing is done on data. At first integrity check is perform on data for presence of missing values, whitespaces. After that statistical matrix is plotted using df.describe () command to gain more insight about data.

• **Missing value check – Data contain no missing value**

**• Data integrity check –**



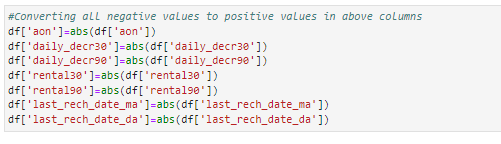




• **Statistical Matrix –** From df.describe () command we got some key observation about data. One of it was that some features contain negative values and another observation few features contain extreme maximum value indicating possible outliers or invalid data.

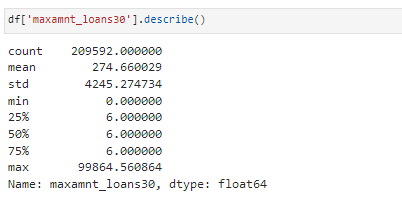
• Strategy to handle data error in min and max column

**Assumption**- All negative values are typing error happen accidentally by type - in front of original value (except feature depicting median). Corrective approach - Negative values are converted into absolute value to correct negative typing error whenever applicable except feature depicting median.



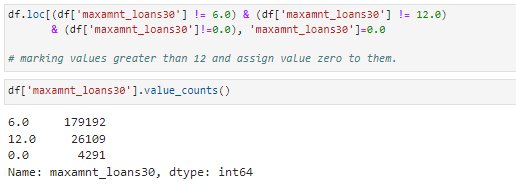
Upper limit of these features handles by outlier removal.

• Data error and correction in maxamnt\_loans30 column (maxamnt\_loans30: maximum amount of loan taken by the user in last 30 days)



The maximum value in maxamnt\_loans30 is not reliable. We already know maximum loan amount taken by customers can be 0,5,10 and which can be repay with amount of 0,6,12.

**Assumption** - The maximum value in maxamnt\_loans30 is 12. We gone replace values greater than 12 into category of zero.



• **Feature Engineering on 'pdate' column**

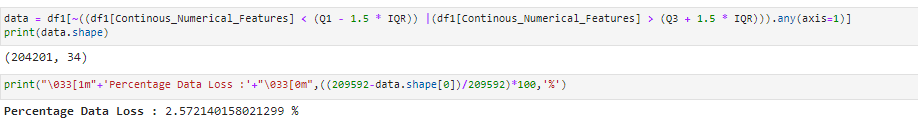
Simple feature engineering operation perform on ‘pdate’ to extract day, month and year column. At last Unnamed :0, PCircle , msisdn columns are drop as they are unnecessary for further investigation.



* **Outliers Detection and removal –**

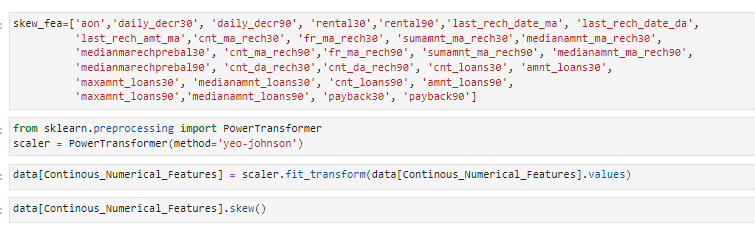
Outliers detected in boxplot. In order to remove outliers Z-score method employ but it results in huge data loss of 23.42 %, which we cannot afford. We got observation from boxplot that outliers do not exist in lower bound but outliers exist in upper bound of features. Based on this observation we decided to employ quantile-based flooring- capping method. Flooring is performed at 0th percentile for lower bound and capping perform at 99th percentile for upper bound.





* **Skewness in features & it’s transformation**

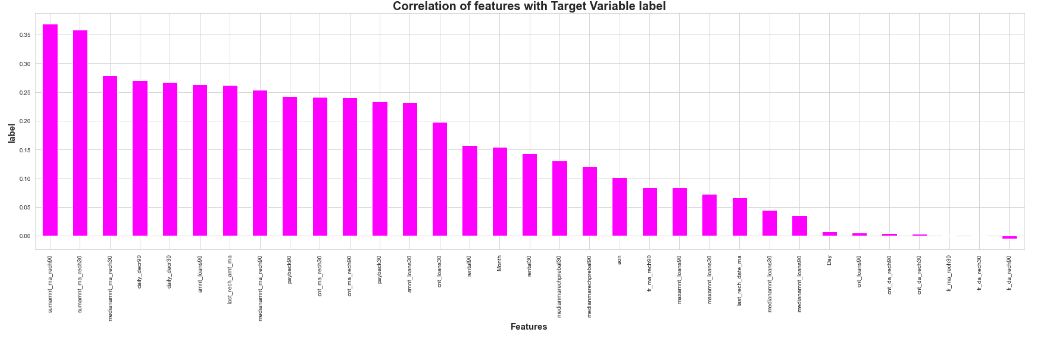
Considerable amount of skewness found in most features by skew () function. Power transformer from sklearn.preprocessing library used to transform skewness in features.



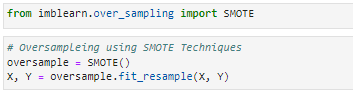
For most of feature’s skewness is reduce within permissible limit except few ones.

**4.Data Inputs- Logic- Output Relationships**

Correlation heatmap is plotted to gain understanding of relationship between target features & independent features. We can see that lot of features are highly correlated with target variable Label. To gain insights about relationship between Input & output different types of visualization are plotted which we will see in EDA section of this report.



We can see that most of independent features are poorly or moderately correlated with target variable label. After that data is split into X and Y and data is scaled using standard scalar. The target variable label is imbalanced in nature, in order to resolved it SMOTE is applied to oversample minority label class.



We have successfully resolved the class imbalanced problem and now all the categories have same data ensuring that the ML model does not get biased towards one category. The multicollinearity between features checked using variance inflation factor. Few findings are as below:

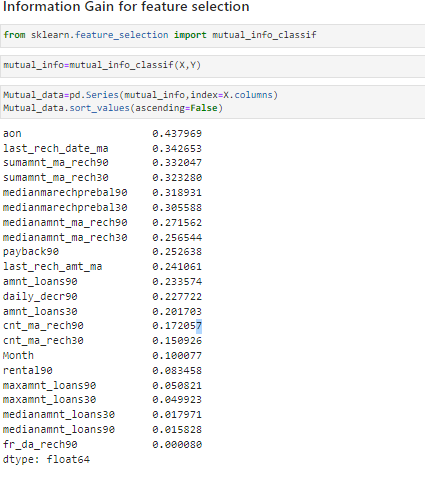
➢ daily\_decr30 and daily\_decr90 are highly correlated with each other.

➢ cnt\_loans90 and amnt\_loans90 are highly correlated with each other.

➢ cnt\_loans30 and amnt\_loans30 are highly correlated with each other.

➢ cnt\_ma\_rech30 and sumamnt\_ma\_rech30 are highly correlated with each other

Here we have done the feature selection by using mutual\_info\_classif



**5.Hardware & Software Requirements with Tool Used**

**Hardware Used –**

1. Processor — Intel i7 processor with 2.4GHZ

2. RAM — 4 GB

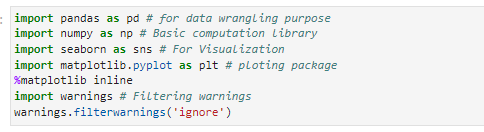
3. GPU — 2GB N Series Graphics card

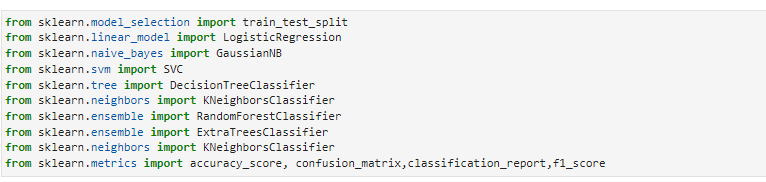
**Software utilized –**

1.Anaconda – Jupiter Notebook/Jupiter Lab

**Libraries Used –**

Different libraries are used while building ML model and Visualization of data.





**Chap. 3 Models Development & Evaluation**

1. **Identification of Possible Problem**

The target variable label has two classes i.e., label ’1’ indicates nondefaulter & label ‘0’ indicates defaulter. Our objective is to predict whether customer is defaulter or not. This becomes binary classification problem which can be solved using various classification algorithms. In order to gain high accuracy of model we will train model with different classification model and select final model among them. To enhance performance of best model will employ hyper parameter tuning over it. At end we will save our final model using joblib

1. **Testing of Identified Approaches (Algorithms)**

The different classification algorithm used in this project to build ML model are as below:

❖ Logistics Regression

❖ Decision Tree Classifier

❖ Random Forest Classifier

❖ Extra Tree Classifier

1. **KEY METRICS FOR SUCCESS IN SOLVING PROBLEM UNDER CONSIDERATION**

**Precision:** can be seen as a measure of quality, higher precision means that an algorithm returns more relevant results than irrelevant ones.

**Recall** is used as a measure of quantity and high recall means that an algorithm returns most of the relevant results.

**Accuracy score** is used when the True Positives and True negatives are more important. Accuracy can be used when the class distribution is similar.

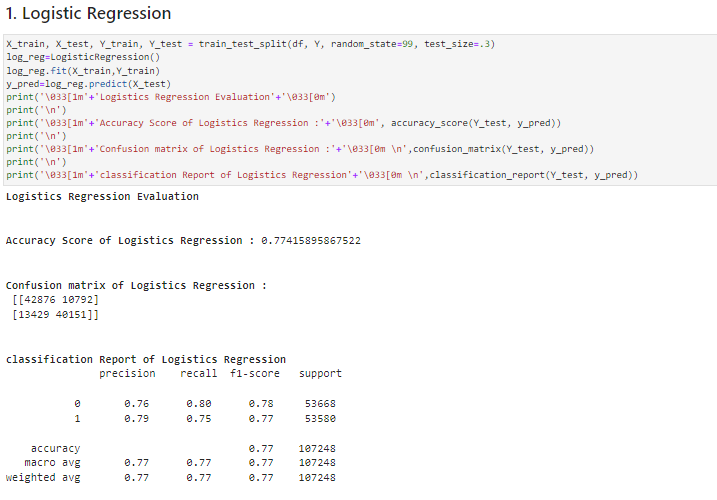
**F1-score** is used when the False Negatives and False Positives are crucial. While F1-score is a better metric when there are imbalanced classes.

**Cross\_val\_score**: To run cross-validation on multiple metrics and also to return train scores, fit times and score times. Get predictions from each split of cross-validation for diagnostic purposes. Make a scorer from a performance metric or loss function.

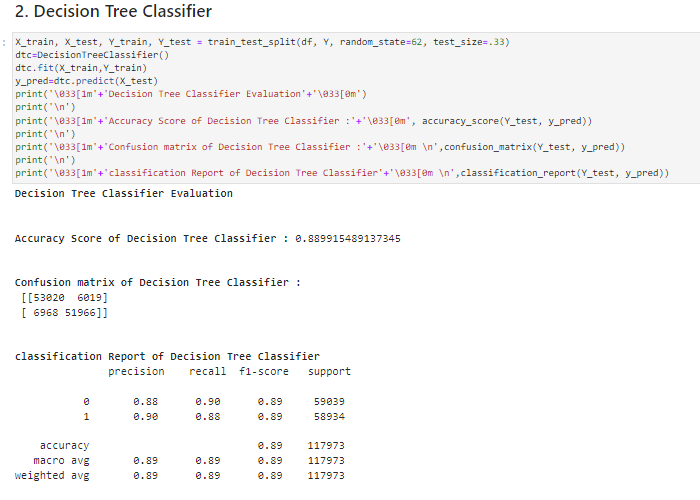
**ROC\_AUC\_SCORE:**  ROC curve. It is a plot of the false positive rate (x-axis) versus the true positive rate (y-axis) for a number of different candidate threshold values between 0.0 and 1.0

1. **RUN AND EVALUATE SELECTED MODELS**

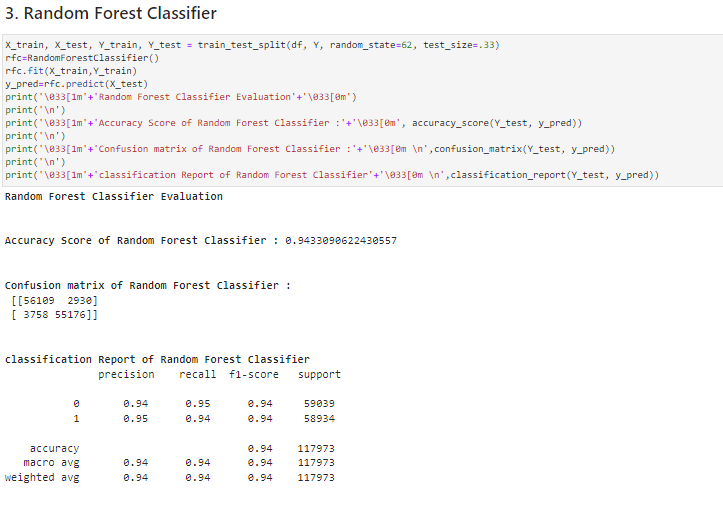
**1.Logistic Regression Model:**



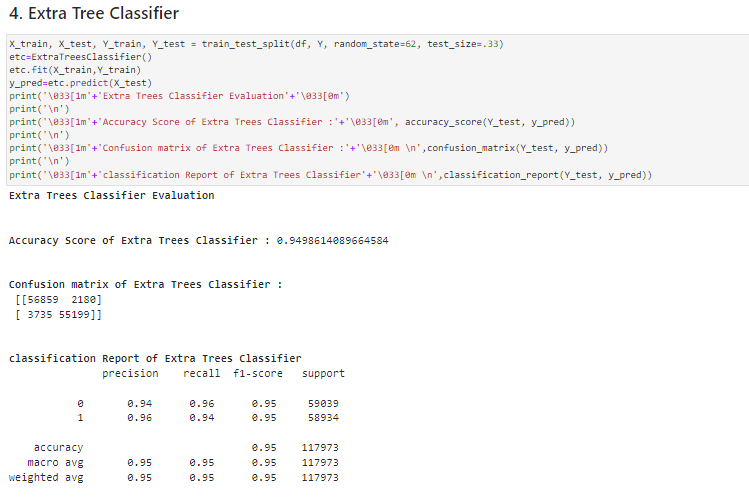
**2. Decision Tree Classifier:**



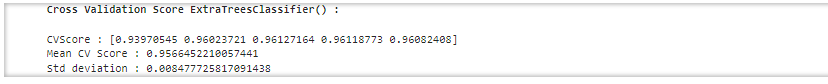
**3. Random Forest Classifier:**



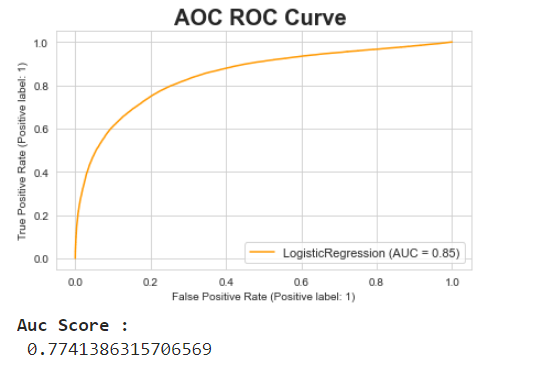
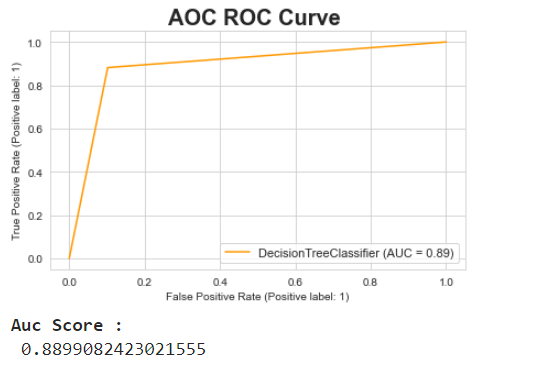
**4. Extra Trees Classifier:**

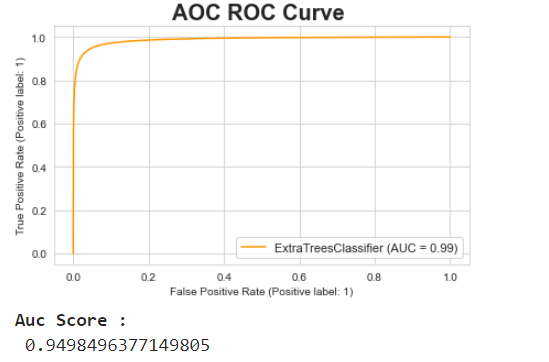
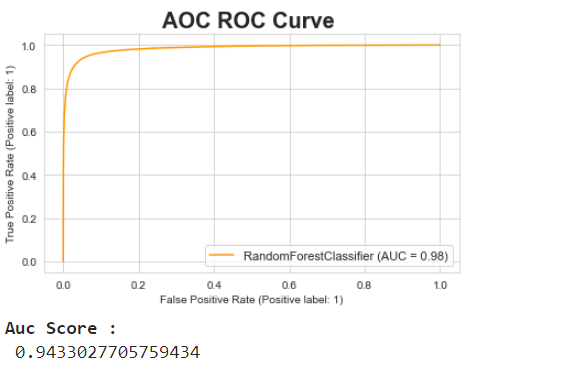


**k-5 fold cross validation is perform**

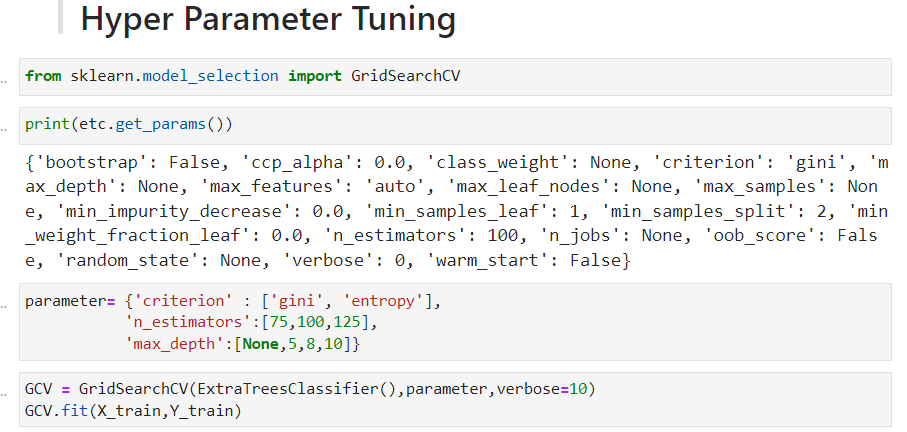


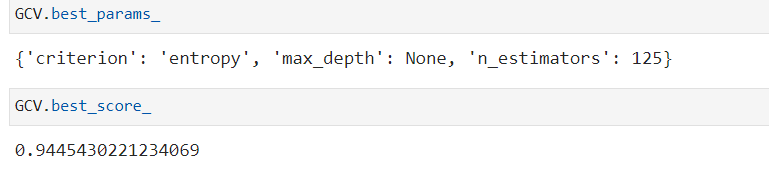
**AOC -ROC CURVE OF DIFFERENT ML MODELS**

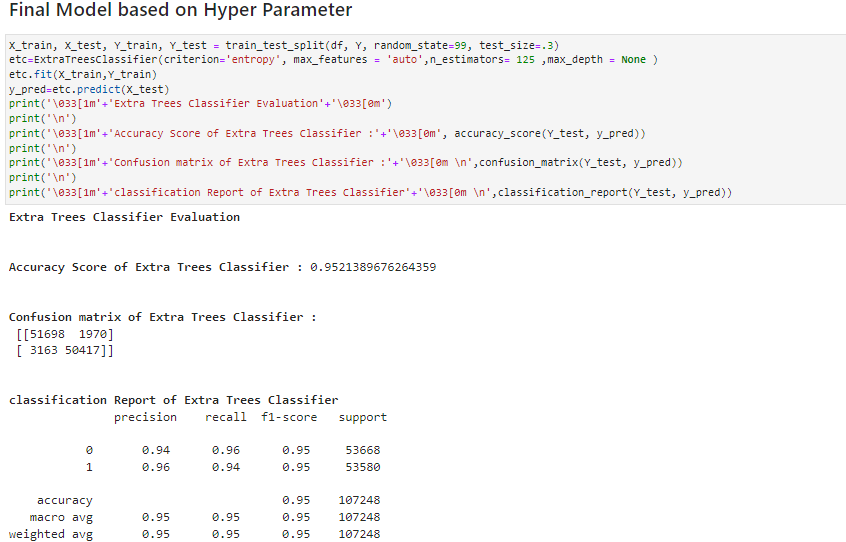


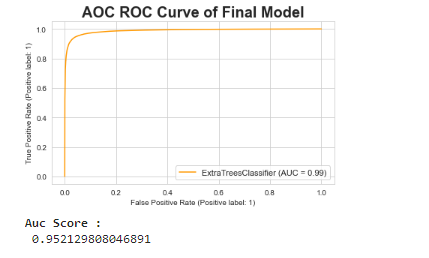
 

We can see Extra Tree Classifier gives maximum AUC. It also gives us highest accuracy score and cross validation score. Hyper parameter tuning performs on this model to enhance accuracy of model.

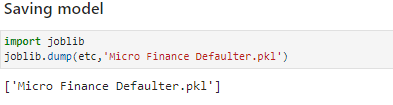






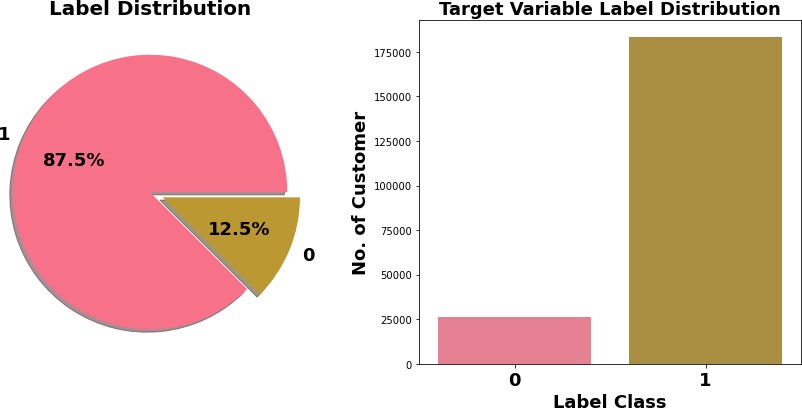


Final model is saved using joblib library.

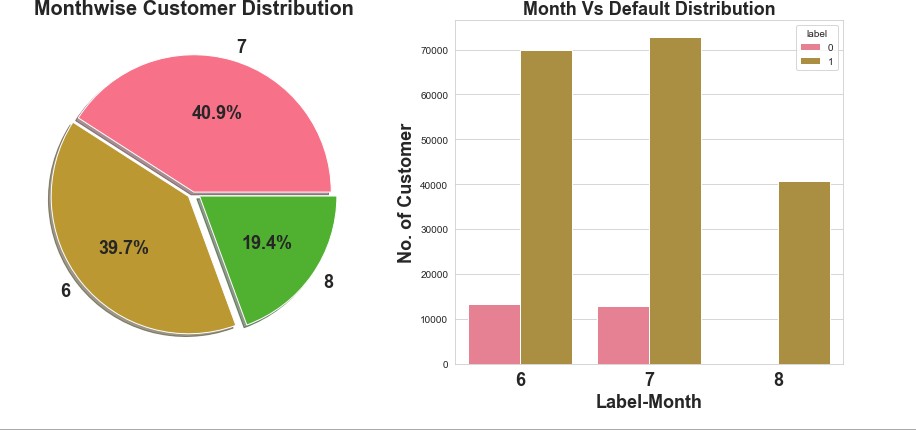


**VISUALIZATIONS**

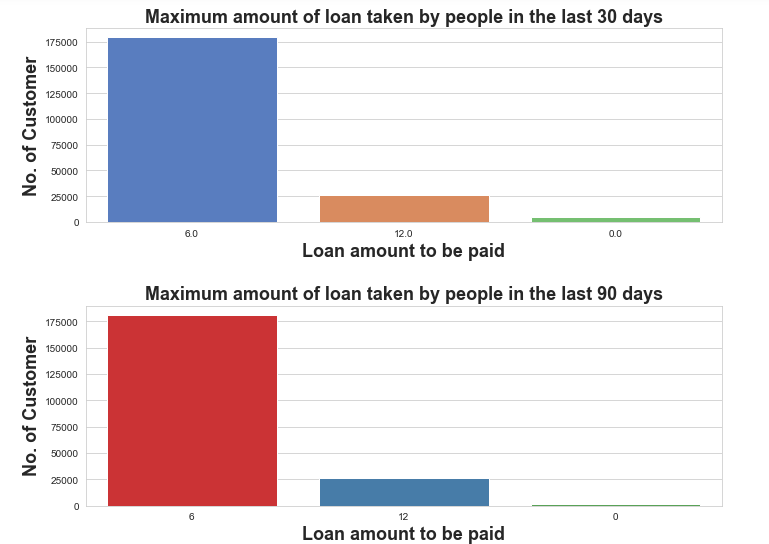
Let’s see target variable distribution before balancing data



Here target variable Label class 1 represent non-defaulter while Label class 0 represent defaulter i.e., Loan not paid. We can see Most of customers are non-defaulter while very few are defaulter. From ML model building point of view target variable is imbalanced which need to balance using balancing techniques.

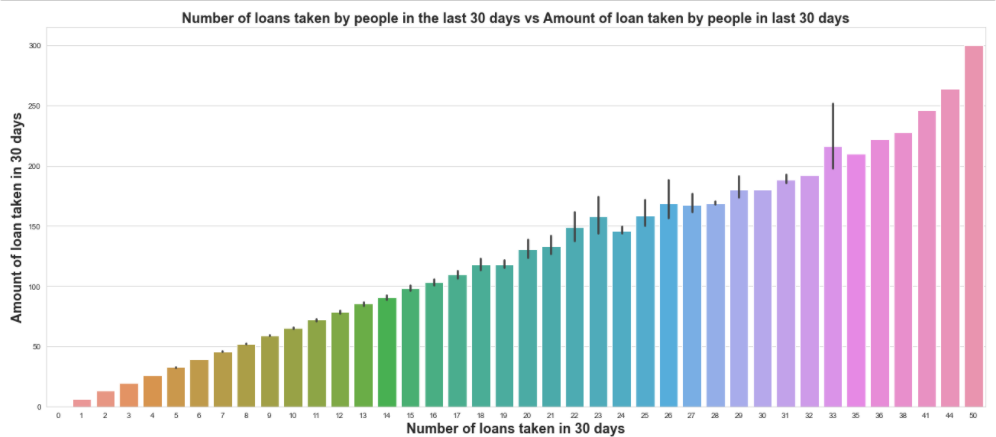


Most of data belong to month 6 and 7, followed my month 8. We can see very few defaulters in month 8.

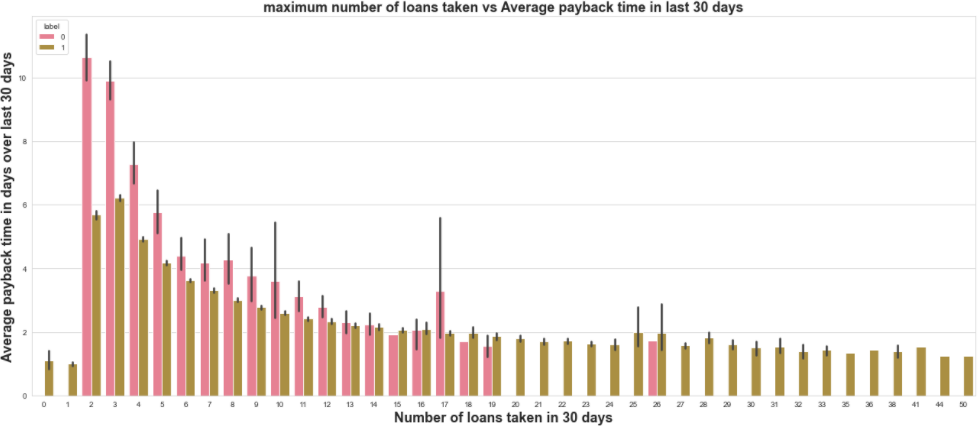


Observations:

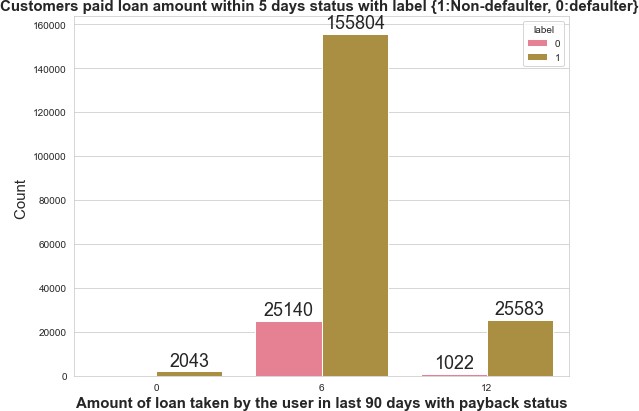
1. In 30 days, maximum number of people had taken 6Rs as the loan amount and the number of people is 179192 whereas the number of people had not taken loan and their number is 4291.
2. In 90 days, maximum number of people had taken 6Rs as the loan amount and the number of people is 180944 whereas the number of people had not taken loan and their number is 2043.
3. Maximum number of people had taken 12Rs as the loan amount within 90 days and their number is 26605 whereas for 30 days the number of people who had taken 12Rs is 26109 respectively.



Maximum number of loans taken by the people is 50 and the Average loan amount is equivalent to 300. Minimum number of loans taken by the people is 0.



We can observe that the Average payback time over last 30 days is higher for people who had taken loan 2 times.



Very few defaulters in case of customers who have taken loan in amount of 12.

## **6. Interpretation of the Results**

* As this dataset belongs from the year 2016, the data are recorded in the month of June, July and August. From the visualization, we can say that the most loan amount taken is rupiah 6 and most of the users are paying the loan within the time frame of 5 days, but many early users failed to do so. They usually take almost 7 to 8 days to pay the loan amount and even the valuable customers some time fails to pay the amount within the time frame.
* One more thing I noticed that, the smaller number of loans taken by the people are more defaulters and the frequently loan taking customers are less defaulters.
* Most importantly, the people are paying the amount early or lately and sometimes they might fail to pay within the time frame, but I observed that almost 80% of users are paying the amount within 7-8 days. It is recommended that to extent loan repayment time frame from 5 days to 7 days.
* The collected data is only for one Telecom circle area as per Dataset Documentation so that we had dropped that column.
* Customer who takes a greater number of loans are non-defaulters (i.e., 98% of the category) as they repay the loan within the given time i.e., 5 days

# Chap 4. Conclusion

**1. Key Findings and Conclusions of the Study**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Algorithm | Accuracy Score | Recall | Precision | F1 Score | CV Score | AUC  Score |
| Logistics Regression | 0.7746 | 0.77 | 0.77 | 0.77 | 0.7733 | 0.7746 |
| Decision Tree Classifier | 0.8886 | 0.89 | 0.89 | 0.89 | 0.8932 | 0.8886 |
| Random Forest Classifier | 0.9423 | 0.94 | 0.94 | 0.94 | 0.9439 | 0.9423 |
| Extra Tree Classifier (ETC) | 0.9502 | 0.95 | 0.95 | 0.95 | 0.9566 | 0.9502 |
| Final Model (ETC- Tuned) | 0.9521 | 0.95 | 0.95 | 0.95 | 0.9507 | 0.9521 |

1. Extra Tree Classifier Hyper parameter tuned gives maximum accuracy score of 0.9521 with cross validation score of 0.9507. It also gives us maximum AUC score.
2. 12.5 % customers are defaulters out of whole dataset.
3. Tendency to pay loan within 5 days is high among customer who take loan many times within month compare to those who take loan 1-2 times.
4. It is recommended that to extent loan repayment time frame from 5 days to 7 days.

## **2. Learning Outcomes of the Study in respect of Data Science**

1. First time I handle such huge dataset.
2. First time any project I worked on ever need such data clean operation. I paid attention realistic & unrealistic data, considering it corrective measure taken as per need. This was beyond normal missing value imputation for me.
3. As data was huge require high computational capacity, it made me switch to Google Colab for running model and for hyperparameter Tuning. I Hyper Tuned Final model with Google Colab GPU.
4. I run Hyper parameter tuning 2-3 times with serval parameter. It was taking lot of times so at end I reduce Hyperparameter search parameter and still it was taken 6-7 hr for finding best parameter.

## **3. Limitations of this work and Scope for Future Work**

1. Limited computational resources put limitation on optimization through hyper parameter tuning. Accuracy of model can increase with hyperparameter tuning with several different parameter. Here we use only two parameters for tuning.
2. Data is imbalanced, we utilized SMOTE for it but if get label data which at least in ratio of 70:30, It can give us much more realistic model.