

MICRO CREDIT DEFAULTER

Submitted by:

VEENA CHAWLA

**ACKNOWLEDGMENT**

Excerpts have been taken from below websites and the same have been citated as footnote wherever used.

<https://github.com/sonarsushant/Loan-Defaulter-Prediction/blob/master/Model_Training_and_Evaluation.ipynb>

1. [https://www.convergences.org/en/119115/#:~:text=In%20ten%20years%2C%20microfinance%20institutions,over%20the%20past%20five%20years](https://www.convergences.org/en/119115/" \l ":~:text=In%20ten%20years%2C%20microfinance%20institutions,over%20the%20past%20five%20years).

1. <https://en.wikipedia.org/wiki/Microfinance>
2. <https://www.bpcbt.com/smartvista-solutions/microfinance-services?gclid=CjwKCAjwgOGCBhAlEiwA7FUXkgzH0zeWOAkh1VDE7yCVA7zNpJiYSxRz4Bb-xYKziuzGE1Z_qK60AxoCyawQAvD_BwE>
3. <http://www.gbgindonesia.com/en/finance/article/2013/an_outlook_on_indonesia_s_microfinance_sector.php#:~:text=Indonesia%20is%20renowned%20for%20its,50%20million%20people%20(CGAP).&text=Indonesia%20has%20more%20than%2050,growth%20in%202012%20(OECD)>.
4. <http://www.gbgindonesia.com/en/finance/article/2013/an_outlook_on_indonesia_s_microfinance_sector.php#:~:text=Indonesia%20is%20renowned%20for%20its,of%202013%20(MIX%20Market)>.

<https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-python/>

# INTRODUCTION

In today’s world lending has become an essential business requirement. The definition of lending is changing from conventional application through bank branches and stringent background checks, towards document-less and automated processes. The issue with conventional lending is that its time consuming and an expensive affair. It becomes difficult for consumers falling in low to medium income category to avail for any funds the traditional way. Here non-financial institutions and alternate financing companies comes to the rescue.

Institutions funding these categories of companies are called Micro Finance Institutions. Over the last decade MFI’s have lent hundreds of billions of dollars. Annual growth rate of 11.5% was observed over the last 5years in MFI’s. In the last 10 years, MFIs have also improved their efficiency. Despite a decade marked by a sharp increase in the cost per borrower, from an average of $68.4 in 2009 to $106.7 in 2018 (+56%), the operating expense ratio decreased by 2.7 points over the period. Between 2009 and 2018, MFIs also recorded an increase in their returns on assets (+1.3 points) and equity (+2.9 points).[[1]](#footnote-1)

Nevertheless, there was a slight deterioration in the quality of the portfolio over the entire period, with the portfolio at risk (PAR) over 30 days having risen from 6.4% in 2009 to 7% in 2018. After a decline in the PAR > 30 days between 2010 and 2012, it rose again and stabilised between 2016 and 2018 at around 7%.[[2]](#footnote-2)

Initially microfinance had limited definition which was, providing loans to poor entrepreneurs and small businesses lacking credit history. But over time microfinance has emerged as a larger movement whose objective now is "a world in which everyone, especially the poor and socially marginalized people and households have access to a wide range of affordable, high quality financial products and services, including not just credit but also savings, insurance, payment services, and [fund transfers](https://en.wikipedia.org/wiki/Electronic_funds_transfer)." [[3]](#footnote-3)

In many parts of the world, government goals for financial inclusion and poverty reduction are driving an increase in microfinance[[4]](#footnote-4). Though this has not been proven yet, but much research is being undertaken on this subject.

Micro-credit comes under microfinance, it is a form of microfinance which involves extremely small loan amounts. The purpose of this form of credit is to assist borrowers from extremely low-income categories for either start-up or expansion of existing business. Other term used for micro credit is microlending or microloan.

Indonesia is renowned for its large-scale microfinance sector, with a range of commercial banks and over 60,000 MFIs reaching more than 50 million people (CGAP). There are about $11.2 billion USD loans disbursed, 722,249 borrowers and $13.1billion USD in deposits as of 2013 (MIX Market). Indonesia has more than 50 million MSMEs, representing some 97% of all enterprises and contributing no less than 30% of GDP growth in 2012 (OECD). Currently, however, many of these do not have adequate access to the bank financing they need to grow their businesses, particularly in rural areas. To that end, Bank Indonesia issued a rule that requires banks to have at least 20% of their loan portfolio dedicated to micro loans by 2018 opening new opportunities to further grow the sector.[[5]](#footnote-5)

In Indonesia, there are 260 million mobile subscribers, with 143 million unique mobile subscribers enabling the poorest people to have access to reliable financial transactions.[[6]](#footnote-6)

Our client here owns a Telecommunication company, who provides fixed wireless telecommunications network. There are various other products also launched by them and they have developed their 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.

The company understands the importance of communication and how it affects a person’s life, thus, they are focusing on providing their services and products to low-income families and poor customers that can help them in the need of hour. For this they are collaborating with an MFI to provide micro-credit on mobile balances that needs to be paid back in 5days of issuance of credit. The consumer is believed to be a defaulter the loaned amount is not repaid within the stipulated time-period.

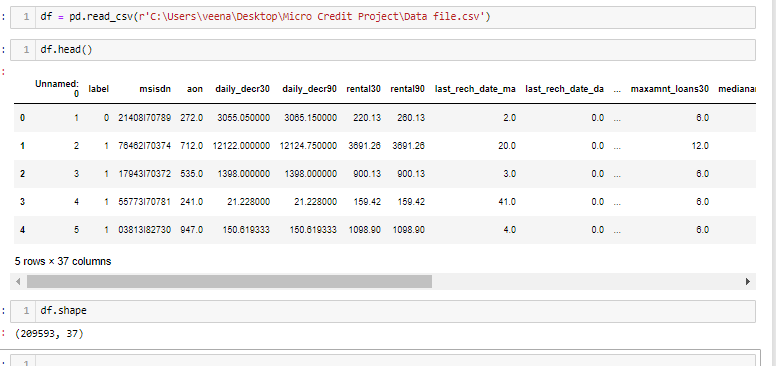
The purpose of this project is to analyse the information from the data provided by the client, to predict the repayment behaviour of the customers which can help if future investments & better return on this investment.

# Data Sources and their formats

The data that was provided by the client, consisted of 37 columns and 209593 rows; description of the columns are as mentioned below:

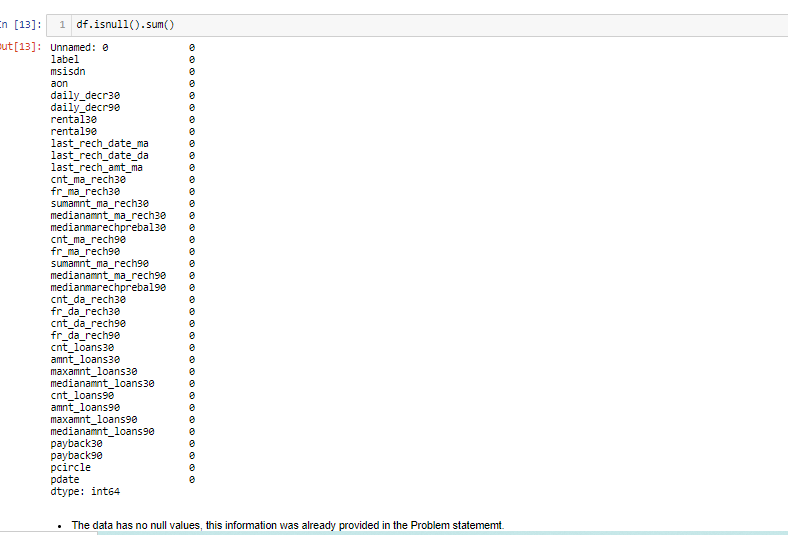
|  |  |
| --- | --- |
| Variable | Definition |
| label | Flag indicating whether the user paid back the credit amount within 5 days of issuing the loan{1:success, 0:failure} |
| msisdn | mobile number of user |
| aon | age on cellular network in days |
| daily\_decr30 | Daily amount spent from main 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 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 |
| payback90 | Average payback time in days over last 90 days |
| pcircle | telecom circle |
| pdate | date |

The snapshot of 5 head column along with the shape is pasted below:



# Data Pre-processing Done

The dataset had no null values present in it, this information was already provided and were verified as per below output:

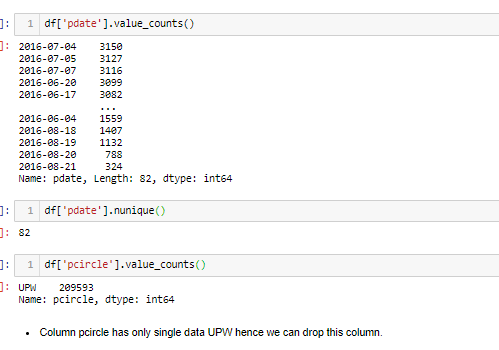


Under pre-processing process the data was checked for unique values in msisdn column which had – 186243 unique values, pdate which had 82 unique values & pcircle which had single unique value in it.

Since the objective of the data analysis is to identify the attributes which can assist in predicting the default probabilities, certain attributes were considered irrelevant to this analysis.

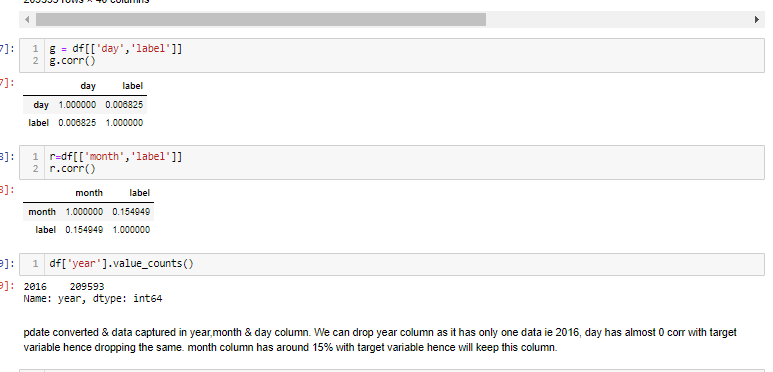
The irrelevant attributes which were eliminated from the analysis are

* Unnamed: 0, 🡪 index column containing numbers of each row, unique to each row.
* msisdn 🡪 contained mobile number of the customers which is an identifier of an individual and not an attribute of a customer.
* pcircle 🡪 had unique data for each row
* pdate 🡪 was object type date column



For better understanding of the data the following steps were undertaken

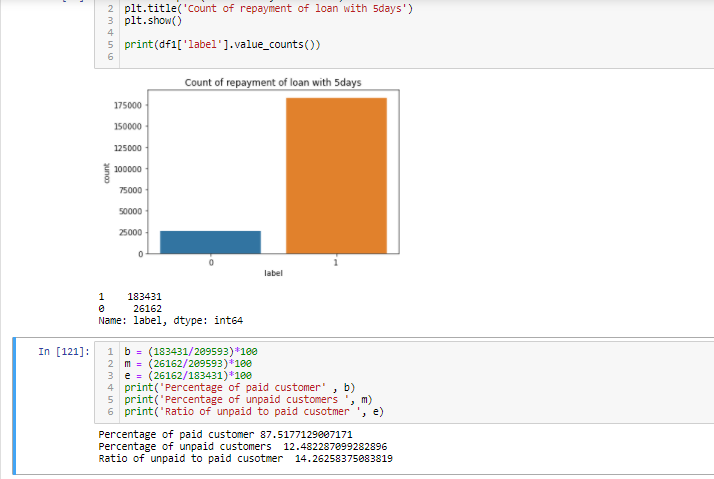
1. Initially attribute “pdate” was converted to datetime format & then was split into day, month & year columns.
2. Subsequently the correlation of these new columns with target variable was checked to understand the relevancy of the columns and to understand which column can be dropped as the dataset now had 40 columns.



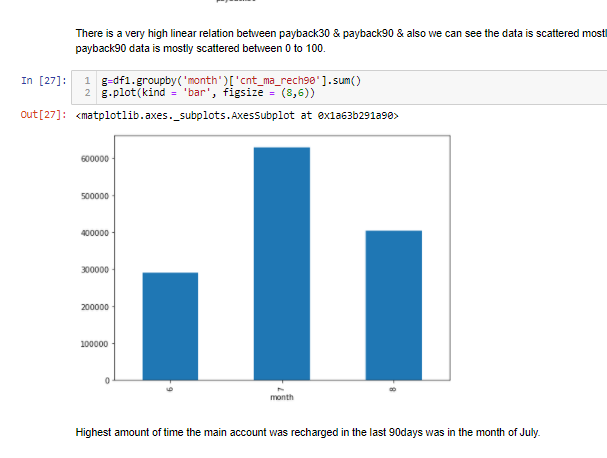
1. From the output as depicted above, day had <1% corr with target variable, month approximately 15% & year was a unique column with record of only 2016.
2. As a conclusion the columns that were removed (dropped) from further analysis were Unnamed: 0, msisdn, pcircle, pdate, day & year column from the dataset.
3. The truncated dataset was then processed further for visualisation.

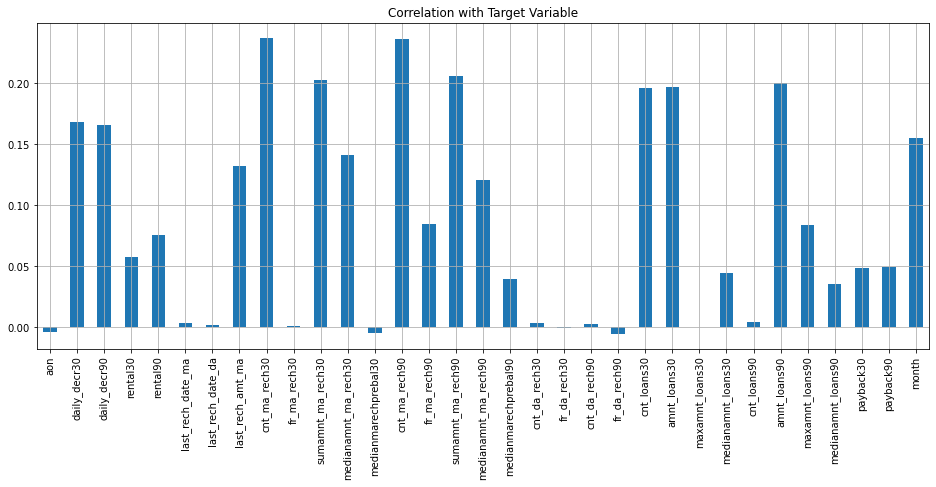
# Data Inputs- Logic- Output Relationships

Univariate Analysis of label column confirmed that the data was highly in favour of paid customers, with unpaid records of around 12.5% & paid at around 87.5%. The ratio of unpaid to paid is at 14.26%. This indicates that for every 100 paid customers there were 14 unpaid customers.



Another aspect of visualisation was to see the distribution of recharges across various months. For this, the attribute that was plotted was “recharge in 90 days” against the month. July month stood out at the month which depicted the highest number of “recharge in 90 days” while, the preceding month and the next month depicted an identical value, it can be argued that the “recharge in 90 days” might follow a cyclic method every alternate month with peaks and troughs, however since the data available is only for a limited time period of 3 months, this assumption cannot be accurately verified.

: 



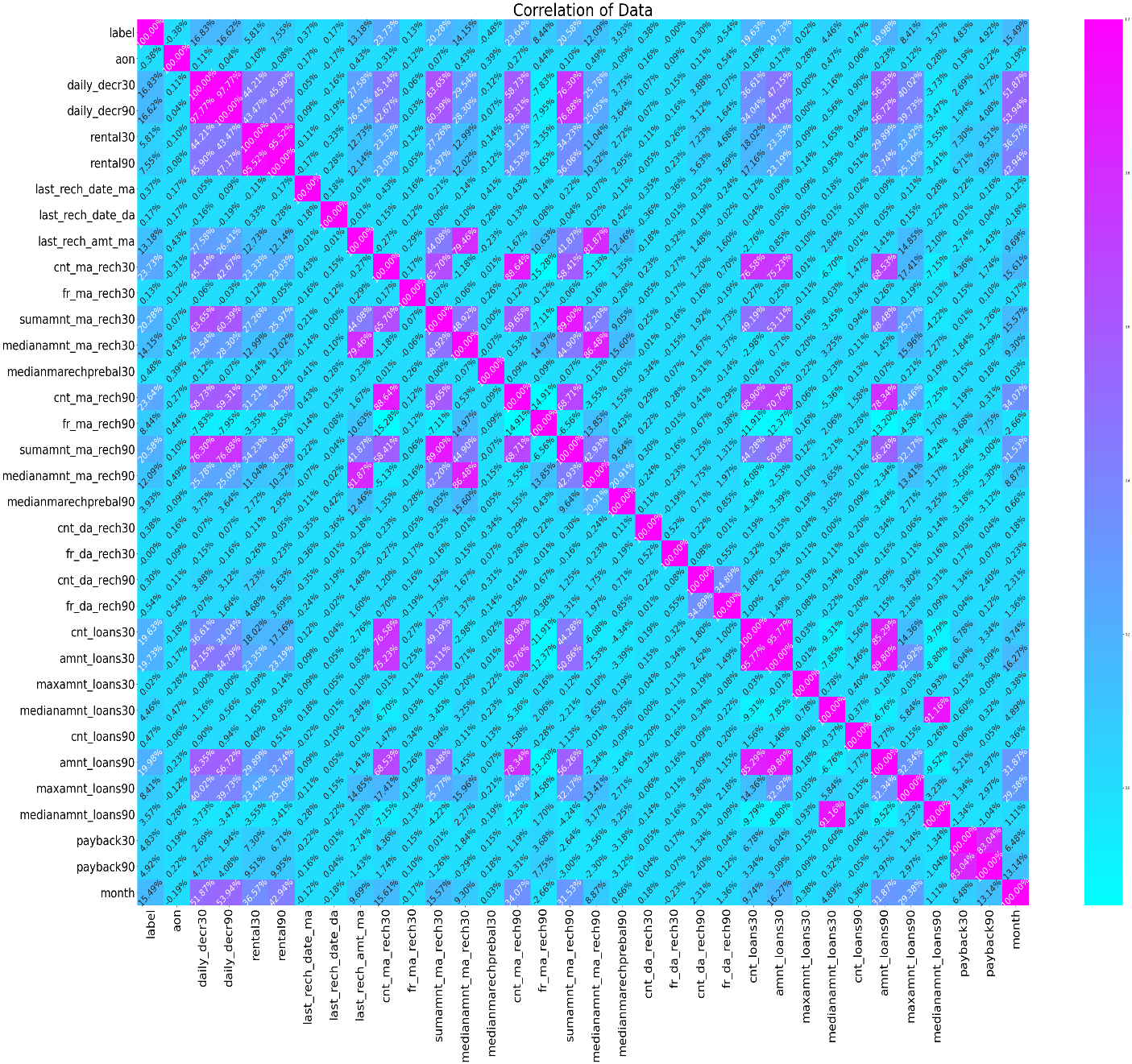
To determine the correlation of all attributes with the target variable, visualisation as shown in the above picture was generated. The observation from this graph were:

**Observations:**

* All the attributes except aon, medianmarechprebal30, fr\_da\_rech90 have negative corelation with the target variable.
* last\_rech\_date\_ma, last\_rech\_date\_da, fr\_ma\_rech30, cnt\_da\_rech30, cnt\_da\_rech90, cnt\_loans90 all these attributes have very low correlation with the target variable (a further check for skewness and outliers in these columns can assist in deciding on the elimination of any of these attributes at a later stage.)
* fr\_da\_rech30 & maxamnt\_loans30 show correlation closer to 0% with targe variable.
* medianmarechprebal90 shows a correlation of around 4% with target variable (this attribute was eliminated from the dataset for further analysis).
* cnt\_ma\_rech30, cnt\_ma\_rech90 are highly positively correlated with the target variable.
* amnt\_loans 30 & amnt\_loans90 show a correlation of around 20% with target variable.
* Attributes - daily\_decr30 & daily\_decr90 show high correlation with each other but also show high +ve corr with target variable to approx. 16% hence these attributes are retained at this stage for further analysis.
* Attributes- rental30 & rental90, are highly correlated with each other at over 90% & rental30 has a correlation of around 6% with target variable. (a further check for skewness and outliers in these columns can assist in deciding on the elimination of any of these attributes at a later stage.)
* cnt\_loans90 & amnt\_loans90 have high correlation with each other & cnt\_loans90 have a very low +ve corr with taret variable ie around 1%.(a further check for skewness and outliers in these columns can assist in deciding on the elimination of any of these attributes at a later stage.)
* medianamnt\_loans30 & medianamnt\_loans90 are highly correlated with each other at 91.16% & with the target variable show a correlation of around 4.5 & 3.6 hence shall be eliminated from the dataset for further analysis.

# State the set of assumptions (if any) related to the problem under consideration

The heatmap for correlation of the whole dataset indicate that some attributes have high correlation with each other, the assumption taken at this stage was that these attributes were duplicate & post checking the outliers and skewness of these attributes one of the these would be eliminated from further analysis. Snip of heatmap is pasted below followed by a detailed observation of the heatmap.



**Observations:**

\*Columns daily\_decr30 & daily\_decr90 have a corr of 97.77% & with label both have a corr of 16.83% & 16.62%. We can drop one of columns from these as they have almost similar corr with target variable & are highly corr with each other.

\* Same is the case with rental30 & rental90, highly correlated with each other @ 95.52% & with target variable-label has a corr of 5.81% & 7.55%. We can drop one of these columns.

\* cnt\_loans30 & amnt\_loans30 has cor of 95.77% wih each other so we can drop one of these columns.

\* medianamnt\_loans30 & medianamnt\_loans90 have a corr of 91.16% with each other.

**From these observations, 3 assumptions were derived,**

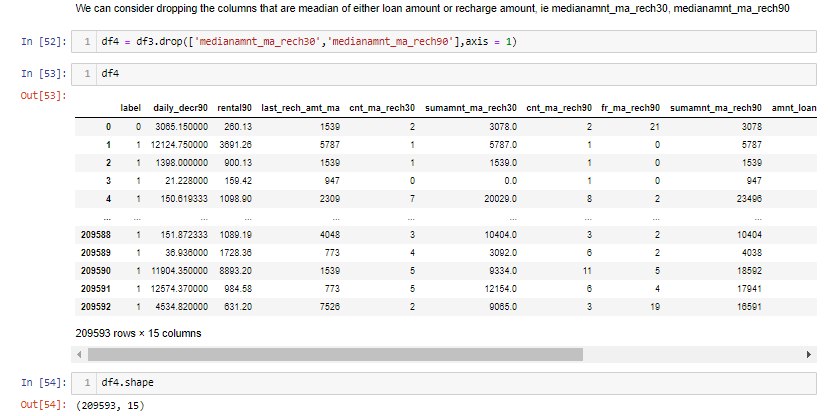
1. 1st was to drop the columns which were negative to <4% correlated with the target variable
2. 2nd was to check for duplicates in columns with correlation of >90% with each other.
3. The data provided had records of past 90 days & past 30days, 3rd assumption was to keep the records of 30 days or drop the same.

**To verify the above assumptions below steps taken:**

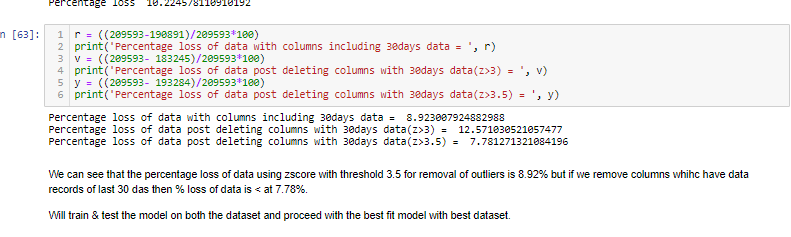
1. Create a separate dataframe for each of the assumptions.
2. Check outliers & skewness for these columns and drop the ones with too many outliers and high skewness.
3. model was tested on 2 different datasets for elimination or retention of records of 30days. (Information on the same is provided under model building part of this report).

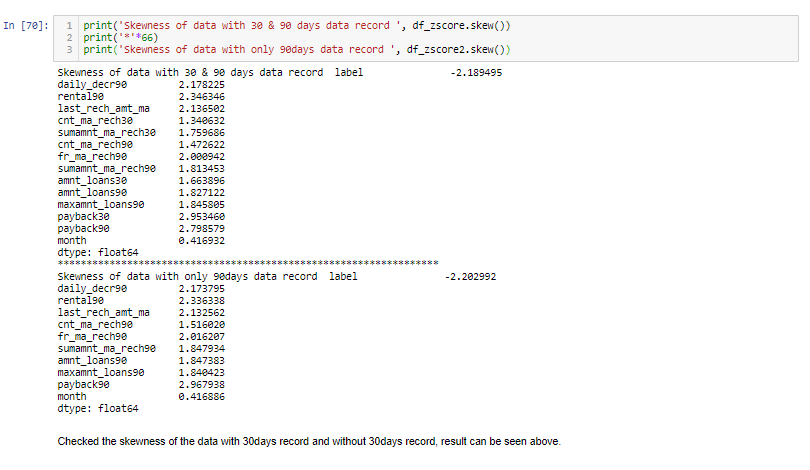
**Process and Output for verifying Assumptions**

1. Correlation of the attributes <4% with target variable, were dropped, as they depicted high skewness & too many outliers. Similarly, from the attributes with >90% correlation with each other, single columns which depicted high skewness and many outliers were dropped.



1. The attributes that had median values present in them were eliminated from the final for further analysis.
2. The final dataset had 15 attributes in it, which were further bifurcated into 2 datasets. 1st had records of both 30 days & 90days, 2nd had records for only 30 days.
3. Outliers for both the dataset were treated using z-score method & threshold of 3.5.
4. The percentage loss for the dataset with 15 attributes was 8.92%, while for the 2nd dataset which had records of only 90days (11 attributes) loss of data was 7.78%



1. The skewness for both the dataset ranged in between 0.40 to 3, below image shows this data. This skewness was treated using power\_transfor(‘yeo-johnson’) method. 
2. StandardScaler was imported from sklearn library & both the datasets were standardized for further analysis.
3. Both the dataset were ready for model building and testing.

# Hardware and Software Requirements and Tools Used

1. The analysis of this project was done using python.
2. The libraries used were pandas, numpy for statistical analysis.
3. For visualization used matplotlib & seaborn. Imported zscore from scipy.stats & from sklearn.preprocessing imported power\_transform, & StandardScaler to treat outliers,skewness & scaling the data.Algorithms were imported from sklearn.
4. Featured data was saved in pikle format under joblib library.

# Model/s Development and Evaluation

The problem statement for this dataset was to predict the probability of a customer repaying the loan in stipulated timeframe. Since the outcome was binary probabilities, classification algorithms were used for machine learning.

Algorithms used:

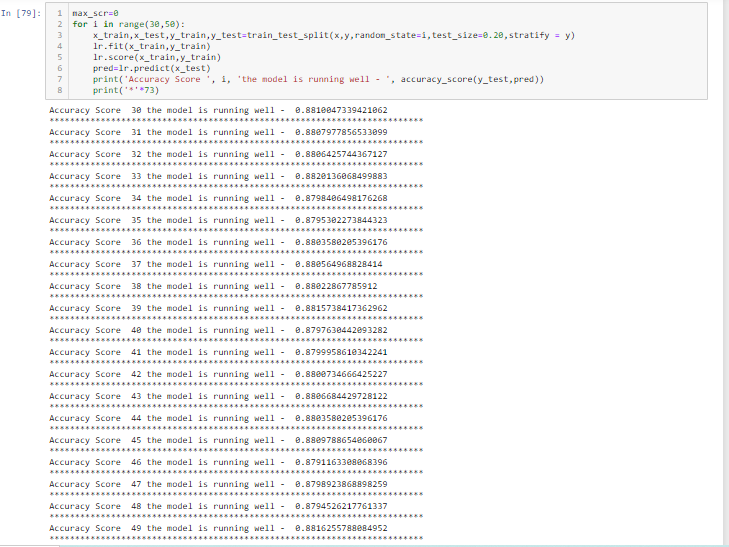
1. Logistic Regression
2. GaussianNB
3. RandomForestClassifier.

To check the efficiency and precision of the predicted score, confusion matrix & classification report method was adopted

**Model Development**

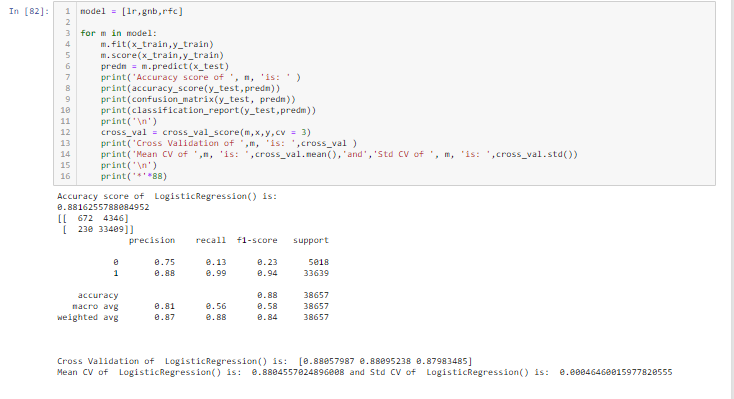
Both the dataset were split into train\_test\_split, a ratio od 80:20 was assumed where 80 meant 80% of the data will be trained and on 20% the model will be tested resulting in accuracy score.

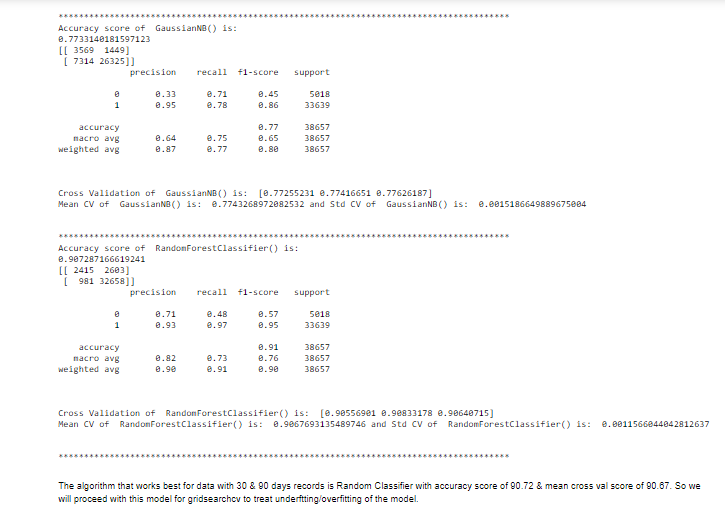
Value of random\_state was derived using logistic regression algorithm, which confirmed the value as 49, below snip confirms this:



Each dataset were trained & tested separately to understand which gave the best output with least accuracy score and least data loss. A detailed conclusion for same is presented at a later stage of this report.

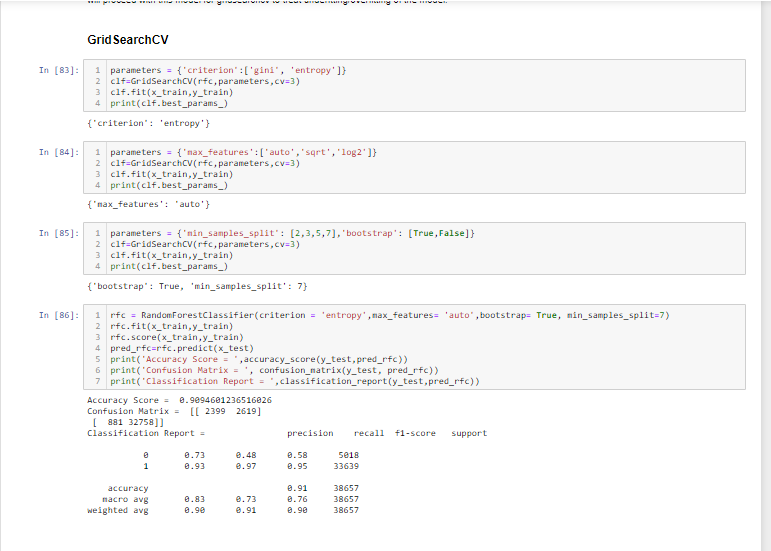
Once with random\_state value , data was further trained & tested on the above mentioned 3 algorithms, & to check the efficiency of the model – confusion matrix along with classification report was used. For better accuracy score cross validation method was adopted. Below snip shows the codes used and the output of the code





Conclusion that RandomForestClassifier gave the best accuracy score was depicted from the above images.

The model was further fine-tuned using GridSearchCV to treat any underfitting or overfitting of the mode, below image depicts the code with the output



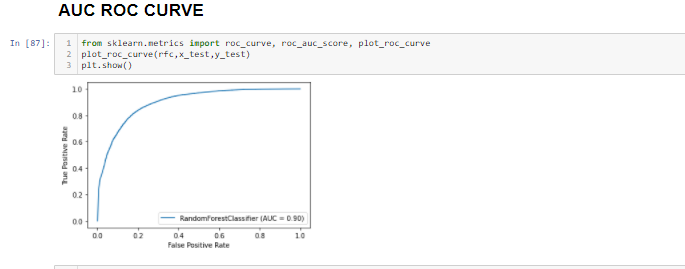
Model was trained and tested on different parameters of RandomForestClassifier, best parameters were (criterion = 'entropy',max\_features= 'auto',bootstrap= True, min\_samples\_split=7). With these new parameters the score of the model improved slightly, from 90.72% it went up to 90.94%.

Visuals to check the accuracy of the model were made adopting AUC ROC curve. AUC means Area Under the Curve and ROC means Receiver Operator Characteristics. ROC plots false positive rate versus false negatives, it derives that the model predicted positive is same as actual positive. Below are is the mathematical calculation of ROC & AUC[[7]](#footnote-7)

True Positive Rate (TPR) = True Positives / (True Positives + False Negatives), also called as Sensitivity

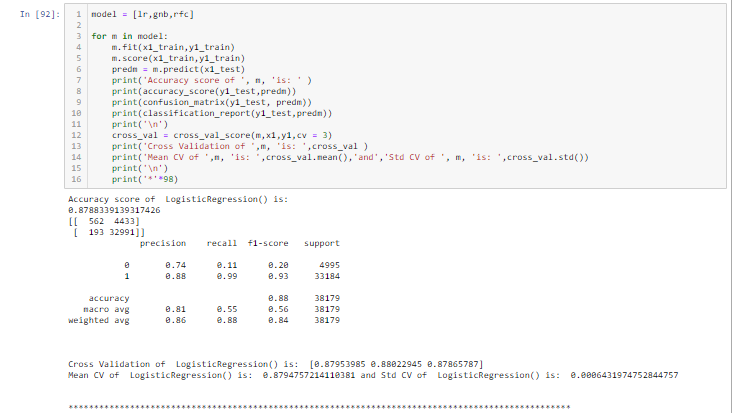
False Positive Rate (FPR) = False Positives / (False Positives + True Negatives), also called as Specificity.

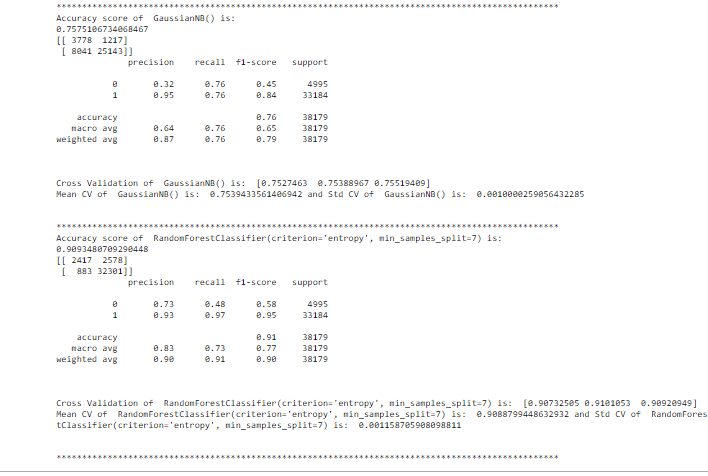
The AUC is used as a summary of the model built, below image throws light on this concept.



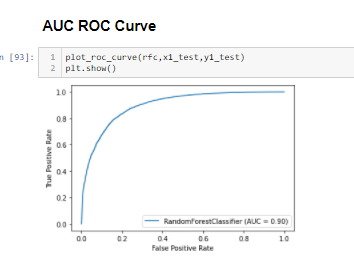
The curve above confirmed the performance of the model, which was at 90% which means that the predicted positive is 90% accurate to be actual positive.

Performance of the model for 2nd dataset (30days& 90days record) was tested with the same steps used to build the model for above dataset. Below are the snapshots of the code used and the outputs for same.





The above images conclude that RandomForestClassifier gave the best output for this dataset too @ 90.90%.



The model predicted score was 90% accurate to actual score and the same can be confirmed from above image.

# CONCLUSION

The conclusion from the model building for this dataset was made that RandomForestClassier performed well with accuracy score of 90%. The dataset of using records of only 90days was determined by the percentage loss of data which was <8%.

The dataset and the model was saved in pickle file using joblib library.

Below image shows the dataframe recording the predicted and original value depicting the performance of the saved model



The study gave a sneak-peek into the micro finance industry of Indonesia. An insightful learning was made for performance of micro-credit and micro-finance over the last decade in Indonesia. It also enlightened the growth and the business opportunities provided in the field of micro-credit in Indonesia and globally. A

Even though the data gave an accuracy score of 90%, the prediction that 90% of population will not default cannot be predicted due to the imbalance in target variable. Unpaid to paid ratio was 14% which is not an adequate ratio for predicting the default rate. Require more data with wider spread for a better prediction of this problem.

1. <https://www.convergences.org/en/119115/#:~:text=In%20ten%20years%2C%20microfinance%20institutions,over%20the%20past%20five%20years>. [↑](#footnote-ref-1)
2. <https://www.convergences.org/en/119115/#:~:text=In%20ten%20years%2C%20microfinance%20institutions,over%20the%20past%20five%20years>. [↑](#footnote-ref-2)
3. <https://en.wikipedia.org/wiki/Microfinance> [↑](#footnote-ref-3)
4. <http://www.gbgindonesia.com/en/finance/article/2013/an_outlook_on_indonesia_s_microfinance_sector.php#:~:text=Indonesia%20is%20renowned%20for%20its,50%20million%20people%20(CGAP).&text=Indonesia%20has%20more%20than%2050,growth%20in%202012%20(OECD)>. [↑](#footnote-ref-4)
5. <http://www.gbgindonesia.com/en/finance/article/2013/an_outlook_on_indonesia_s_microfinance_sector.php#:~:text=Indonesia%20is%20renowned%20for%20its,of%202013%20(MIX%20Market)>. [↑](#footnote-ref-5)
6. <http://www.gbgindonesia.com/en/finance/article/2013/an_outlook_on_indonesia_s_microfinance_sector.php#:~:text=Indonesia%20is%20renowned%20for%20its,of%202013%20(MIX%20Market)> [↑](#footnote-ref-6)
7. <https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-python/> [↑](#footnote-ref-7)