

**Micro Credit Defaulter Project**

Submitted by:

N THARANI

**ACKNOWLEDGMENT**

Various research papers and websites are used. Research papers like Predicting Credit Default among Micro Borrowers in Ghana by [Kwame Simpe Ofori](https://www.researchgate.net/profile/Kwame-Ofori), [Eli Fianu](https://www.researchgate.net/profile/Eli-Fianu), [Osaretin Kayode Omoregie](https://www.researchgate.net/profile/Osaretin-Omoregie), [Nii Afotey Odai](https://www.researchgate.net/profile/Nii-Odai); Telecom microfinance banking versus commercial banking: a battle in the financial services sector - [Waqar Younas](https://link.springer.com/article/10.1057/s41264-020-00085-7#auth-Waqar-Younas), [K. Ramanathan Kalimuthu](https://link.springer.com/article/10.1057/s41264-020-00085-7#auth-K__Ramanathan-Kalimuthu) and papaers by Acquah & Addo, 2011 and Wongnaa 2013, etc., are used. Website like researchgate.net, geeksforgeeks, economictimes, stackoverflow.com, etc., are used as references. The data is received from the client which is their own data.

Thanking SRISHTI MAAN, my guide from FLIPROBO TECHNOLOGIES for clearing all my doubts while undergoing the project.

**INTRODUCTION**

* Business Problem

Microfinance institutions play a major role in economic development in many developing countries. However many of these microfinance institutions are faced with the problem of default because of the non-formal nature of the business and individuals they lend money to. Today, microfinance is widely accepted as a poverty-reduction tool, representing $70 billion in outstanding loans and a global outreach of 200 million clients. 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.

This problem is based on Telecom industry. They have launched various products and have developed its business and organization based on the budget operator model, offering better products at Lower Prices. They are collaborating with an MFI to provide micro-credit on mobile balances to be paid back in 5 days. The Consumer should pay back the loaned amount within the time duration of 5 days. This study seeks to find the determinants of credit default in microfinance institutions.

* Conceptual Background of the Domain Problem

Microfinance Institutions are organizations that offer financial services to low income populations and everyone who is in need. This becomes very useful when targeting the unbanked poor families living in remote areas with not much sources of income. The Microfinance services are Group Loans, Agricultural Loans, Individual Business Loans, etc., Many microfinance institutions, experts and donors are supporting the idea of using mobile financial services (MFS) which they feel are more convenient , efficient and cost saving.

* Review of Literature

This is a comprehensive summary of the research done on the topic. The review should enumerate, describe, summarize, evaluate and clarify the research done.

Certain websites and papers that helped me to take insights from are:

1. Predicting Credit Default among Micro Borrowers in Ghana - [Kwame Simpe Ofori](https://www.researchgate.net/profile/Kwame-Ofori), [Eli Fianu](https://www.researchgate.net/profile/Eli-Fianu), [Osaretin Kayode Omoregie](https://www.researchgate.net/profile/Osaretin-Omoregie), [Nii Afotey Odai](https://www.researchgate.net/profile/Nii-Odai)
2. Acquah & Addo, 2011
3. Wongnaa 2013
4. reserachgate.net
5. Telecom microfinance banking versus commercial banking: a battle in the financial services sector - [Waqar Younas](https://link.springer.com/article/10.1057/s41264-020-00085-7#auth-Waqar-Younas), [K. Ramanathan Kalimuthu](https://link.springer.com/article/10.1057/s41264-020-00085-7#auth-K__Ramanathan-Kalimuthu)

* Motivation for the Problem Undertaken

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 are able to reduce interest rates and processing fees if repayment rates are high, thus increasing patronage of loans. A high repayment rate is a catalyst for increasing the volume of loan disbursements to various sectors of the economy. It is necessary here to find the determinants of defaulters and stop providing loan on expecting for no defaulters.

Telecom companies have chosen microfinance banking as their alternative revenue stream. Micro financing is one of the needed service for telecom users. It is said that telecom industry gains from micro financing. This happens only when there is no defaulter or the repayment of loan is done at time. When repaying gets delayed or repaying is not done, both are considered as defaulter. On concerning this issue, our project has made certain determinants to be necessary for the customer to receive a loan from the company.

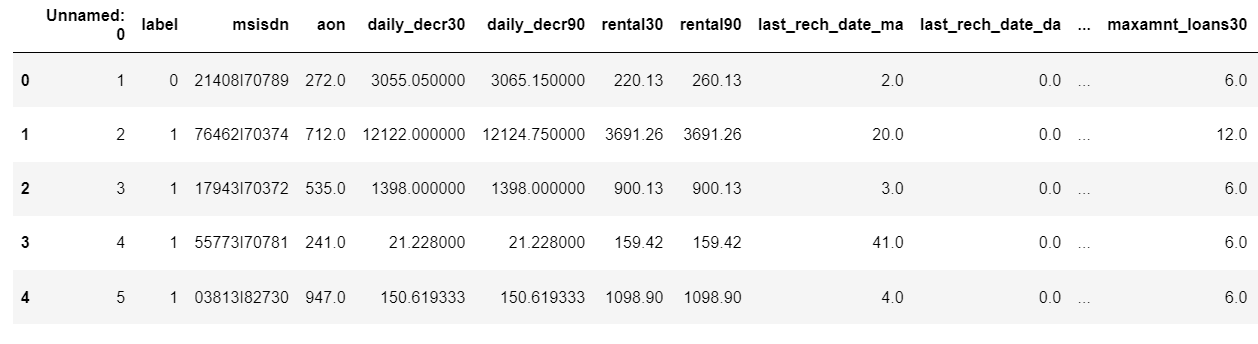
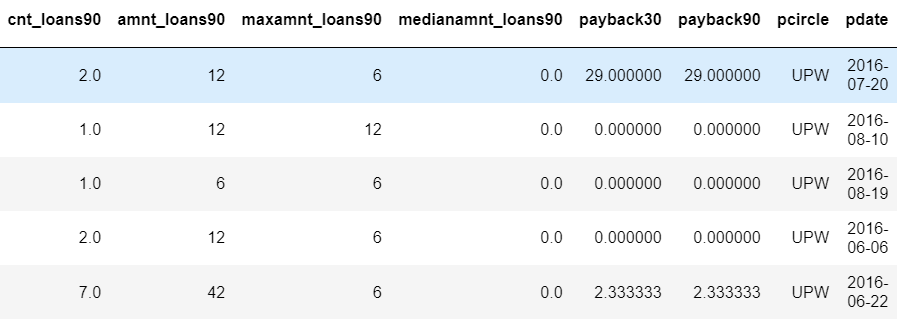
The main motivation of this project is that the Microfinance institutions could use this model to screen prospective loan applicants in order to reduce the level of default.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

Various statistical, mathematical, analytical algorithms are used. Experimental design, outliers detection using Z-score, transformation, scaling, re-sampling methods, statistical hypothesis tests, estimation statistics, data manipulation, feature selection, feature reduction, balancing the dataset, etc are done for the problem.

* Data Sources and their formats

The data is received from the client with necessary data description. The data is in excel format which can be imported using pandas from local library easily.

* Data Preprocessing Done

Certain column names are having the same meaning, unnamed:0 are not necessary. Based on the data description, features which does not help to determine the loan defaulters are removed. pcircle is a feature which has the same data for the whole dataset which will not help in determining the result. Check for null values, string data type to encode. Check for outliers and check the curve for normalization. Correlation and multi collinearity are checked and treated.

* Data Inputs- Logic- Output Relationships

There are various features used to detect the defaulter. All the features which are helping to detect the defaulter is visualised using matplotlib and seaborn. The relationship between the features are determined. After data cleaning, the relationship between input and output variables is fairly linear.

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

In telecom industry, micro loans are very normal where the defaulters are assumed to be low. The timing of repayment or repayment amount is assumed to be high. Only after checking out the model after building, it can be found.

* Hardware and Software Requirements and Tools Used

Any laptop and computer can be used as hardware. Processor used is Intel(R) Core(TM) i7-4510U CPU. System type is 64-bit operating system, x-64 based processor. RAM of the systerm is 8.00 GB. Microsoft Windows 8.1 version 6.3 is the OS used. Python 2.7.10 is used with the interface Jupyter notebook with many installed libraries.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

There are statistical and analytical approaches in problem-solving. Data cleaning, data mining, wrangling comes under statistical approach whereas data manipulation, feature selection, feature reduction, creating ML models comes under analytical approach.

**Statistical:**

* + - ***Experimental Design*** - Methods to design systematic experiments to compare the effect of independent variables on an outcome
  + Univariate - measure relies on only one variable - Statistical summary, information on variables, etc
  + Bivariate - measure relies on two variables - Correlation, multicollinearity
    - ***Outliers detection using Z-Score (z = (x-μ)/σ)*** - where x is the observed value, μ is the population mean, and σ is the population standard deviation
    - ***Tranformation*** - Transformed the data using Power Transformation to reduce the skewness.
    - ***Scaling*** - Scaled the data using Standard Scaler to get unbiased result.
    - ***Re-sampling Methods*** - Train Test Split is used to systematically split a dataset into subsets for the purposes of training and evaluating this predictive model.
    - ***Statistical Hypothesis Tests*** - Cross validation quantifies the likelihood of observing the result given an assumption or expectation about the result whether the model is overfitting/underfitting or fitting good.
    - ***Estimation Statistics*** - GridSearchCV is used to quantify the best parameter from the listed to fit in the model and give better result. It uses data analysis framework which has a combination of effect sizes, confidence intervals, precision planning, and meta-analysis to plan experiments, analyze data and interpret results.

**Analytical:**

It concerns the design and development of algorithms.

* Data Manipulation - Treat the skewness and balance the imbalanced dataset using SMOTE
* Feature selection - Check the relationship between independent and dependent variables. Select the features which gives meaning to the problem based on the subject knowledge.
* Feature reduction - Reduce the features which does not give proper solution to the problem using Principal Component Analysis.
* Balancing the dataset - Balance the dataset and normalize the data to get better result.
* Testing of Identified Approaches (Algorithms)

Various evaluation metrics can be used for this classification type of model. Some of the popular algorithms are the following:

* Logistic Regression
* k-Nearest Neighbors
* Decision Trees
* Support Vector Machine
* Random Forest Classification
* Naive Bayes
* Run and Evaluate selected models

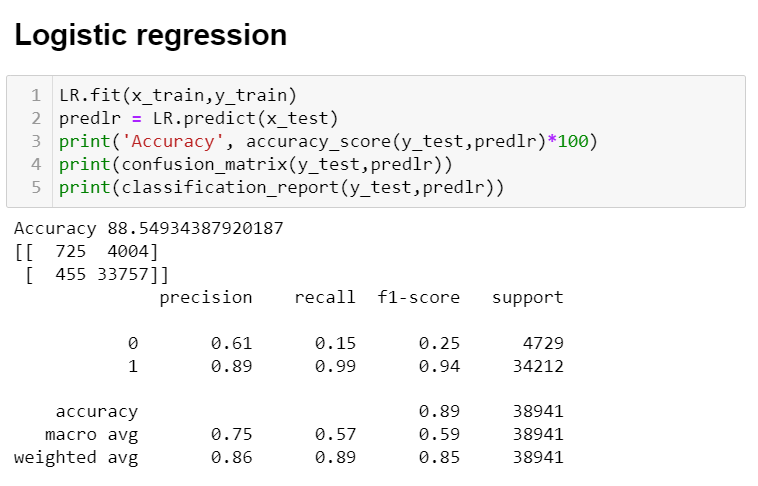
Accuracy - Percentage of correct predictions for dataset

Recall - Percentage of positive cases

Precision - Percentage of predictions are correct

F1 Score - Percentage of positive predictions are correct

***Logistic Regression:***



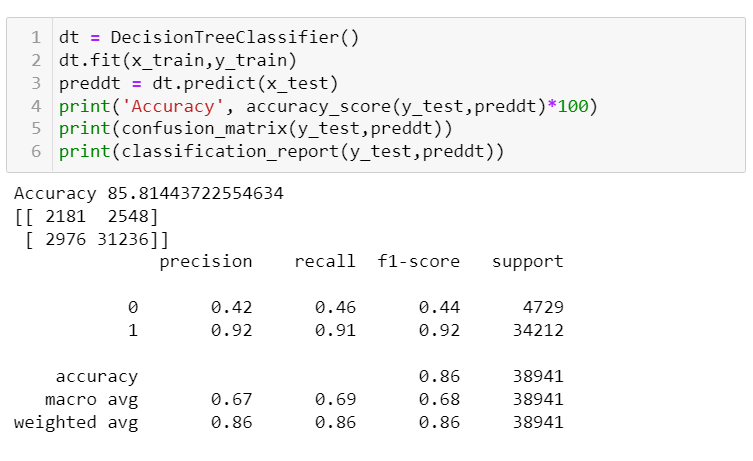
Accuracy - 88.5

Precision - Not paid - 0.61, Paid - 0.89

Recall - Not paid - 0.15, Paid - 0.99

F1 Score - Not paid - 0.25, Paid - 0.94

***Decision Tree:***



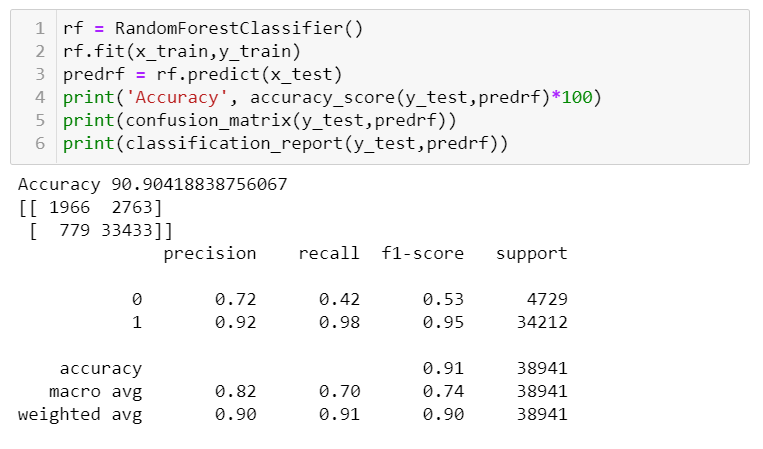
Accuracy - 85.8

Precision - Not paid - 0.42, Paid - 0.92

Recall - Not paid - 0.46, Paid - 0.91

F1 Score - Not paid - 0.44, Paid - 0.92

***Random Forest Classifier:***

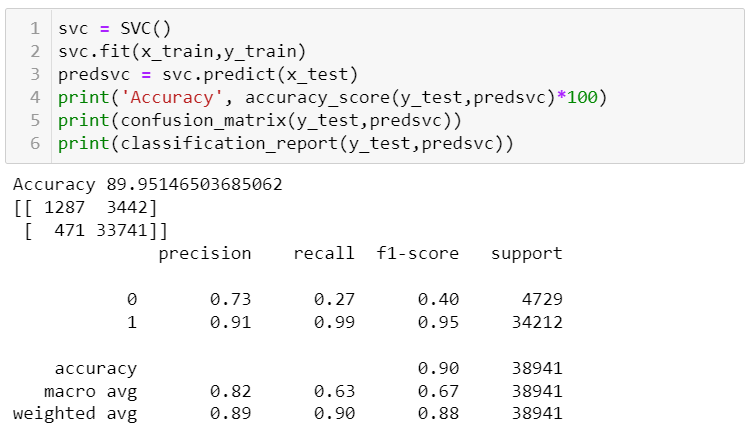
Accuracy - 90.9

Precision - Not paid - 0.72, Paid - 0.92

Recall - Not paid - 0.42, Paid - 0.98

F1 Score - Not paid - 0.53, Paid - 0.95

***Support Vector Classification:***



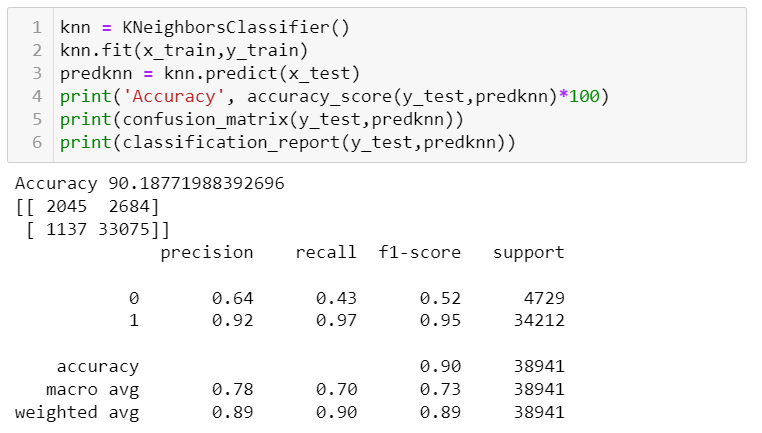
Accuracy - 89.9

Precision - Not paid - 0.73, Paid - 0.91

Recall - Not paid - 0.27, Paid - 0.99

F1 Score - Not paid - 0.40, Paid - 0.95

***KNeighbors Classifier:***



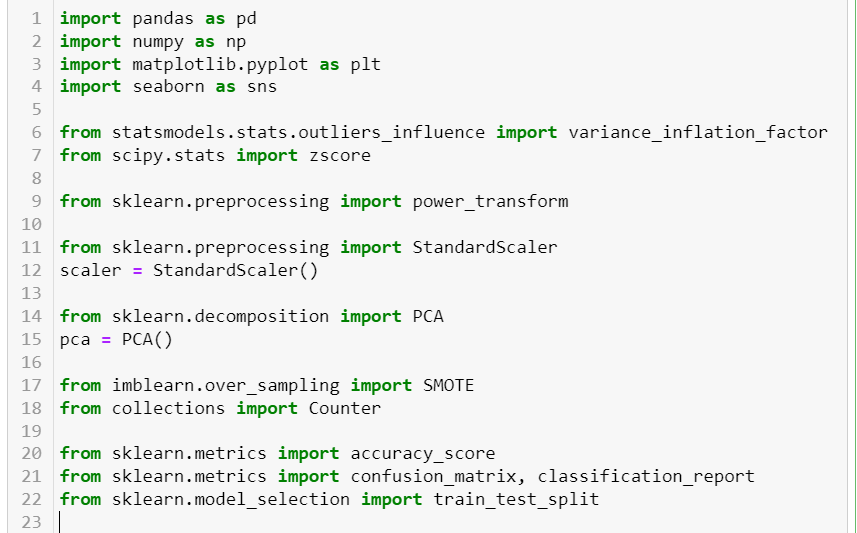
Accuracy - 90.1

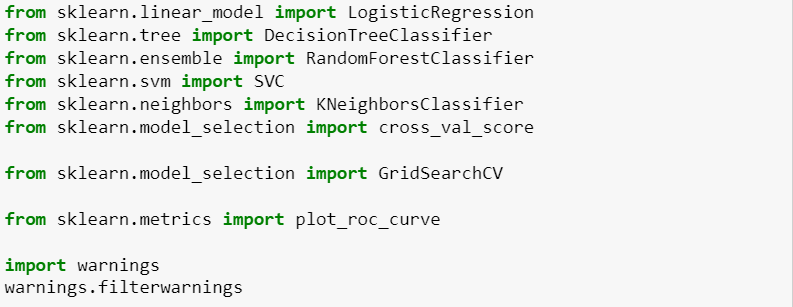
Precision - Not paid - 0.64, Paid - 0.92

Recall - Not paid - 0.43, Paid - 0.97

F1 Score - Not paid - 0.52, Paid - 0.95

* Key Metrics for success in solving problem under consideration

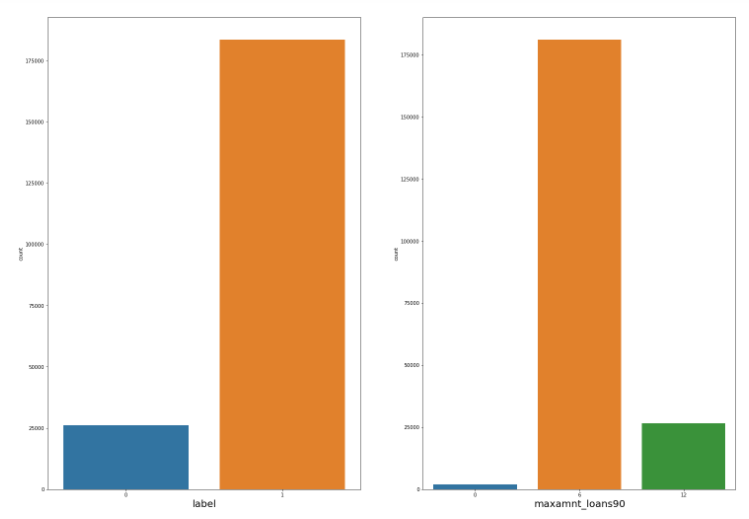




|  |  |
| --- | --- |
| **Libraries** | **Usage** |
| Pandas and numpy | Importing dataset, data cleaning, data wrangling and exploratory data analysis |
| Matplotlib and seaborn | Visualisation libraries |
| Variance\_inflation\_factor | To check the multicollinearity |
| z-score | To check and remove outliers with fixed threshold |
| Power transform | Power transformation to remove skewness |
| Standard Scaler | To standardise the data and normalise the curve |
| PCA | Feature reduction |
| SMOTE and Counter | To balance the imbalanced dataset |
| Accuracy score, confusion matrix, classification report | For concluding the results |
| Train\_test\_split | To separate the training and testing dataset |
| Logistic Regression, DecisionTreeClassifier, RandomForestClassifier, SVC, KNeighborsClassifier | All these are machine learning algorithms to find the results |
| Cross\_val\_score | To check the best fitting of the model |
| GridSearchCV | For hyper parameter tuning |
| Plot\_roc\_curve | To check whether the model is good by checking the area under the curve |

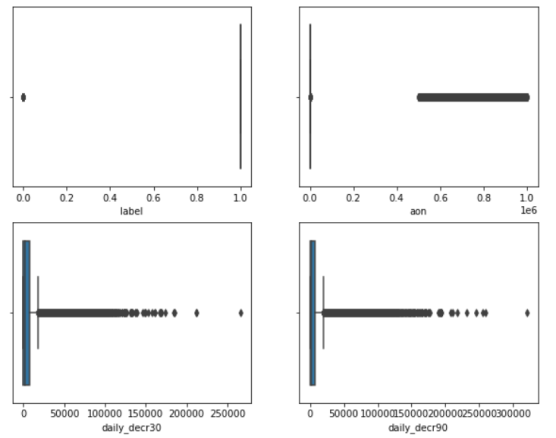
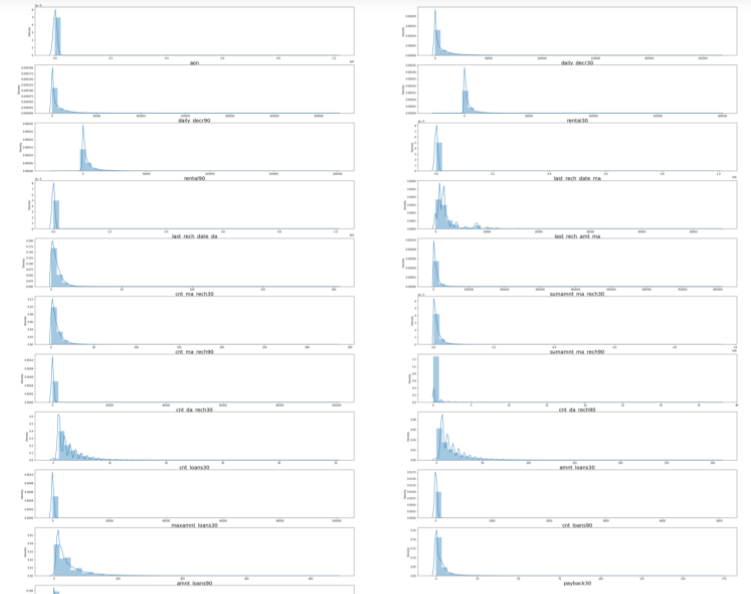
* Visualizations

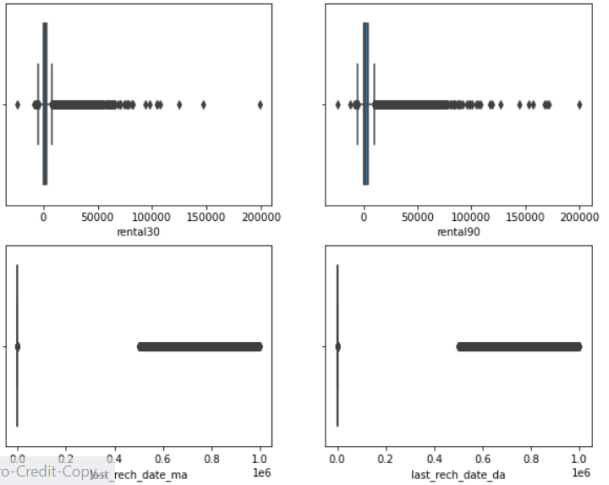
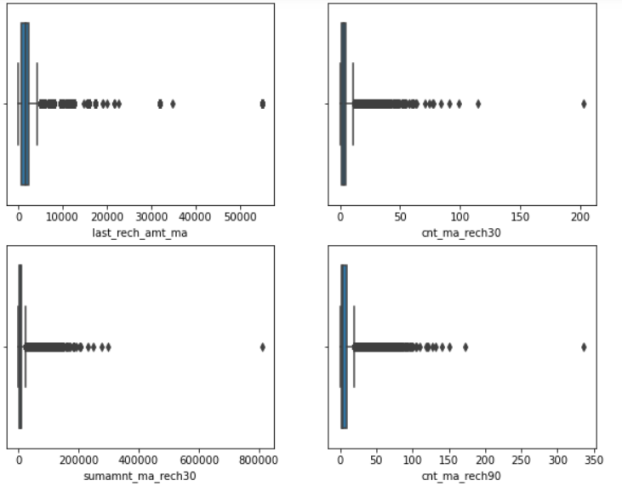
Matplotlib and seaborn is used for visualisations

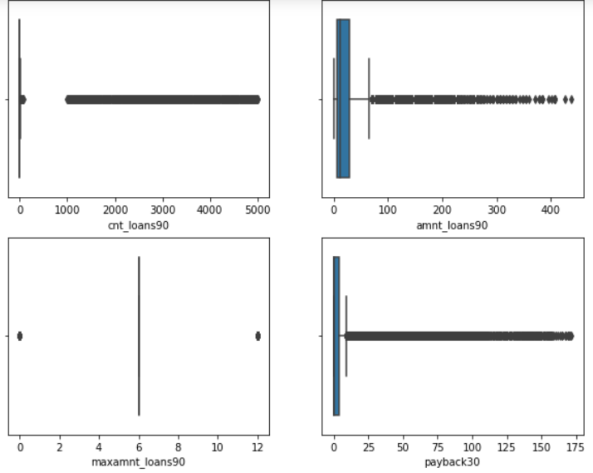
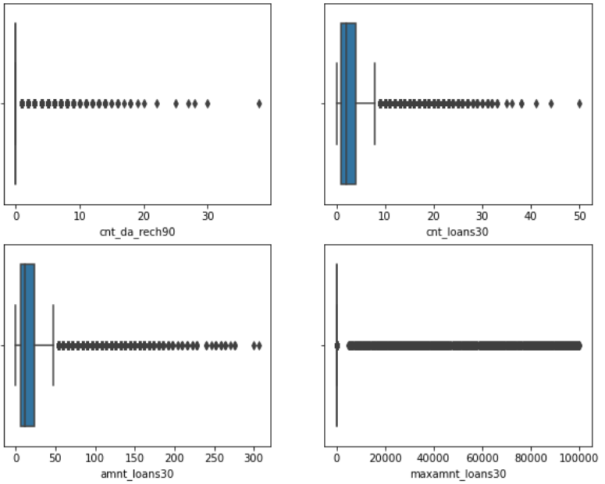


Categorical variables are visualized separately to understand the distribution where the target is imbalanced using countplot.

Continuous variables are visualized separately to understand the distribution where they are normally distributed.





To detect outliers boxplot is used which showed many outliers

* Interpretation of the Results
* Categorical variables and continuous variables are visualized separately to understand the distribution where the target is imbalanced and the other variables are normally distributed
* To detect outliers boxplot is used which showed many outliers
* The relationship between dependent and independent variables are visualised where it looks fairly linear
* In data preprocessing, duplicates are removed which helps to get unbiased result

**CONCLUSION**

* Key Findings and Conclusions of the Study

The relation between the input and output variables looks fairly linear after data cleaning. Correlation and multicollinearity are checked and treated. The features are selected and reduced to get better results. From the models, random forest classifier gives better result and also on comparing the cross validation the model fitted well.

On observing the ROC curve, the area under the curve is more and that shows that the model is good. Save the model and that can be used for predictions later since the model is trained well.

* Learning Outcomes of the Study in respect of Data Science

Visualising the categorical and continuous variable tells how the variable is distributed and the balancing of the dataset. After checking the correlation, reconfirm it using scatterplot. The data already does not have null value and the data is not needed to encode.

Logistic Regression and decision tree is good to go with the model. Random forest regression, KNN takes time for training. SVC takes very long time to train the model. Cross-validation took time to complete and checked whether the model is underfitting and overfitting the model.

Random Forest Classifier gave 91% accuracy approximately. Even after hyper parameter tuning, the model score did not increase. So, we can use Random Forest Classifier for further predictions.

* Limitations of this work and Scope for Future Work

The features can be reduced to get a very better result. Data description must be clearly explained. Time extension for the repayment of loan with increase in credit rates can be determined. Increasing loan amount with increase in credit time also can be analysed. Long-term credits can be given to avoid defaulters.