

MICRO CREDIT DEFAULTER

Case Study

Submitted by:

Arpita Rai

Int 33

**ACKNOWLEDGMENT**

I would like to thank Flip Robo Technologies, for giving me this opportunity to work on this project. I got to learn more from this project about Data Scraping, and practical implementations of using machine learning modules.

I take this opportunity to express my gratitude and regards to my mentor Mr. Shwetank Mishra for his guidance, monitoring and constant encouragement by giving new projects. The help and guidance given by him time to time shall carry me a long way in the journey of life on which I am about to embark.

Lastly, I thank almighty, my parents, brother, sister and friends for their constant encouragement without which this assignment would not be possible.

**INTRODUCTION**

* Business Problem Framing

We are now working with a client in the telecom industry. They are a provider of fixed wireless telecommunications networks. They've released a number of products and built their business and organisation around the budget operator model, which entails providing better products at lower prices to all value-conscious clients via a disruptive innovation strategy that focuses on the subscriber. They recognise the value of communication and how it influences a person's life, thus they focus on giving low-income families and impoverished consumers with services and products that can assist them in their time of need. They've teamed up with a microfinance institution to offer micro-credit on mobile balances that must be paid back in five days. If the Consumer deviates from the course of repaying the loaned amount within the time period of 5 days, he is considered a defaulter. The payback amount for a loan of 5 (in Indonesian Rupiah) should be 6 (in Indonesian Rupiah), whereas the payback amount for a loan of 10 (in Indonesian Rupiah) should be 12 (in Indonesian Rupiah). Here, we need to create a model that can be used to predict if a client would pay back the lent amount within 5 days of loan insurance in terms of probability for each loan transaction. Label '1' shows that the loan has been paid, indicating that it is a non-defaulter, whereas Label '0' indicates that the loan has not been paid, indicating that it is a defaulter.

* Conceptual Background of the Domain Problem
* The client from Telecom Industry 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 and make the prediction that where the people will pay the loan within 5 days or not. Based on this prediction Micro Finance Company will work based on the prediction. To make the better prediction we have clearly understand each feature. As the data is huge, we need to do the properly cleaning of data**.**
* Review of Literature

An attempt has been made in this report to review the available literature in the area of microfinance. Approaches to microfinance, issues related to measuring social impact versus profitability of MFIs, issue of sustainability, variables impacting sustainability, effect of regulations of profitability and impact assessment of MFIs have been summarized in the below report. We hope that the below report of literature will provide a platform for further research and help the industry to combine theory and practice to take microfinance forward and contribute to alleviating the poor from poverty.

* Motivation for the Problem Undertaken

We have to build a model with available independent variable data set by thorough analysis of data.

This micro credit model will help the Finance Company to decide who is defaulter and who is non-defaulter.

Who will return the loan amount within 5 days?

So, they can focus on the area which will yield in high return.

The relationship between the prediction and economy is important; this will drive a motivation in understanding the problem and providing the solution to it.

**Analytical Problem Framing**

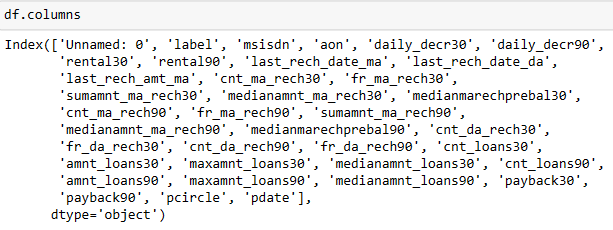
* Data Sources and their formats

The dataset has been given by Flip Robo Technologies in excel format.

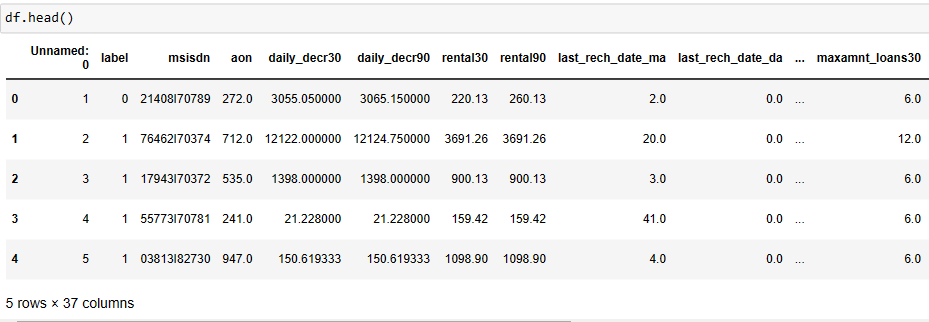
The dataset has 209593 rows and 37 columns.

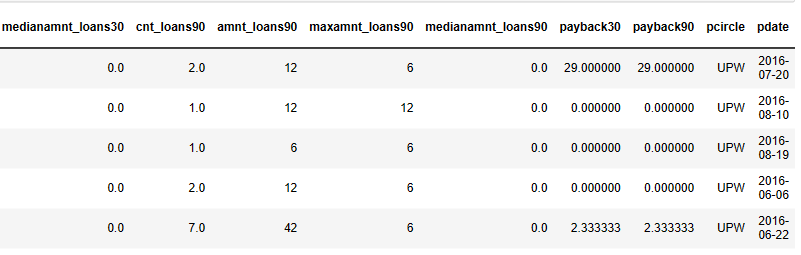
There are 36 features and 1 target in the dataset.

These are as follows:

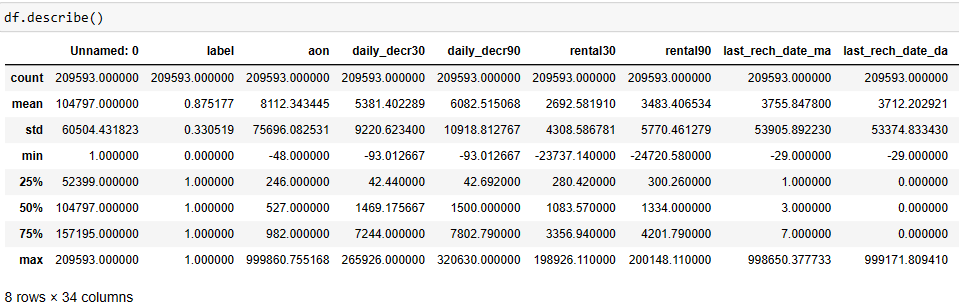


As mentioned, ‘label’ is out target variable. Label ‘1’ indicates that the loan has been paid i.e. Non- defaulter, while, Label ‘0’ indicates that the loan has not been paid i.e. defaulter. Since the target has been categorized into defaulter and non-defaulter, this becomes a project of Classification Problem. Hence we must use classification algorithms to build machine learning models.





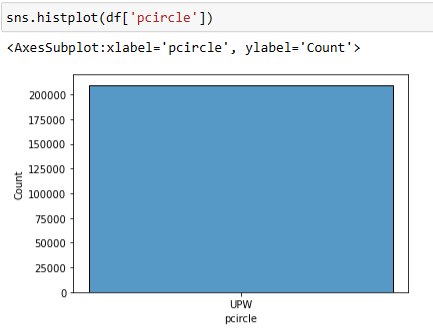
Using the describe(), we get the statistical summary of the dataset and through which we interpret that few features are defined over 30 days’ period as well as 90 days’ period and such features contains almost similar values, which leads to creating multi-collinearity of the features. Hence in order to treat this multi-collinearity, we might have to drop few features defined over different time periods, having the least correlation with the target.



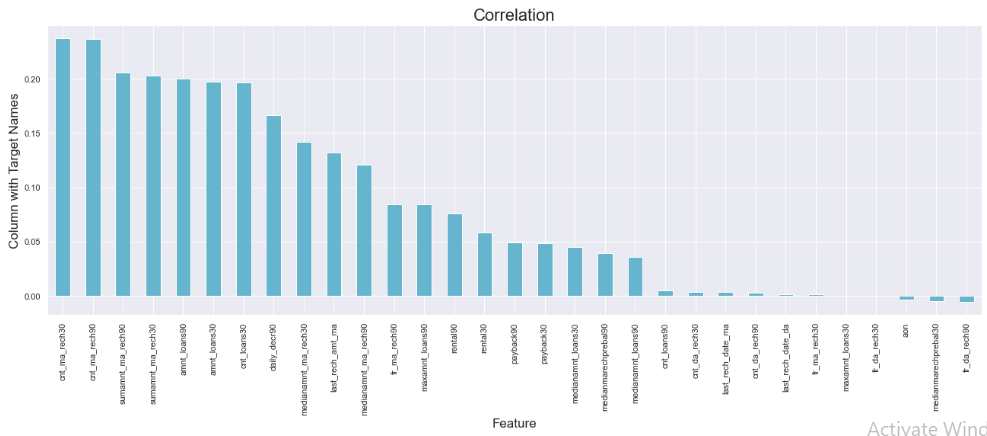
* Mathematical/ Analytical Modeling of the Problem

Using dataset.info(), we get all the information about the data like the name of features along with their data types, their non-null value count.

We got to know that there are 21 columns with float data type, 13 columns with integer data typ e and 3 columns with object data type which are ‘msisdn’, ‘pcircle’ and ‘pdat e’. We have observed that mobile number and date have very wide range of data and which does not add much in prediction of ‘Label’ , also the column ‘pcircle’ has only one value which does not contribute in ‘Label’ prediction as well.

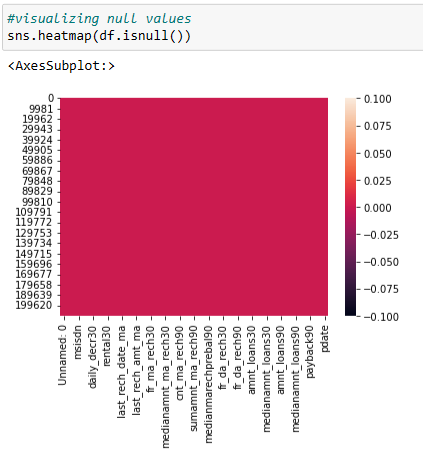


While checking the correlation using Corr(), we found out that, two features have negligible correlation with the target, which we can happily drop to make our model more effective. These features are : ‘maxamnt\_loans30’ (maximum amount of loan taken by the user in last 30 days) and ‘fr\_da\_rech30’ (Frequency of data account recharged in last 30 days).



* Data Preprocessing Done

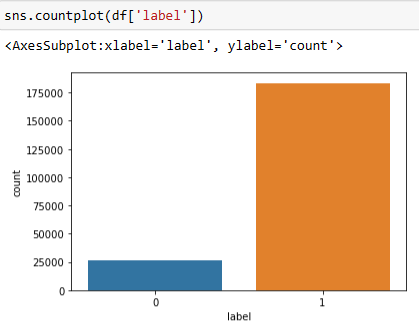
There are missing values in the given dataset which we check by using isnull().sum() then we also plot heatmap to visualize the null values in the dataset.

****

There are so many outliers in the dataset and if we remove all the outliers, we will lose almost 18% of the data and we can’t afford to lose that much of data, hence we will remove the outliers with zscore value more than 5 and in this way, we will lose only 7.5% of the data which is acceptable.

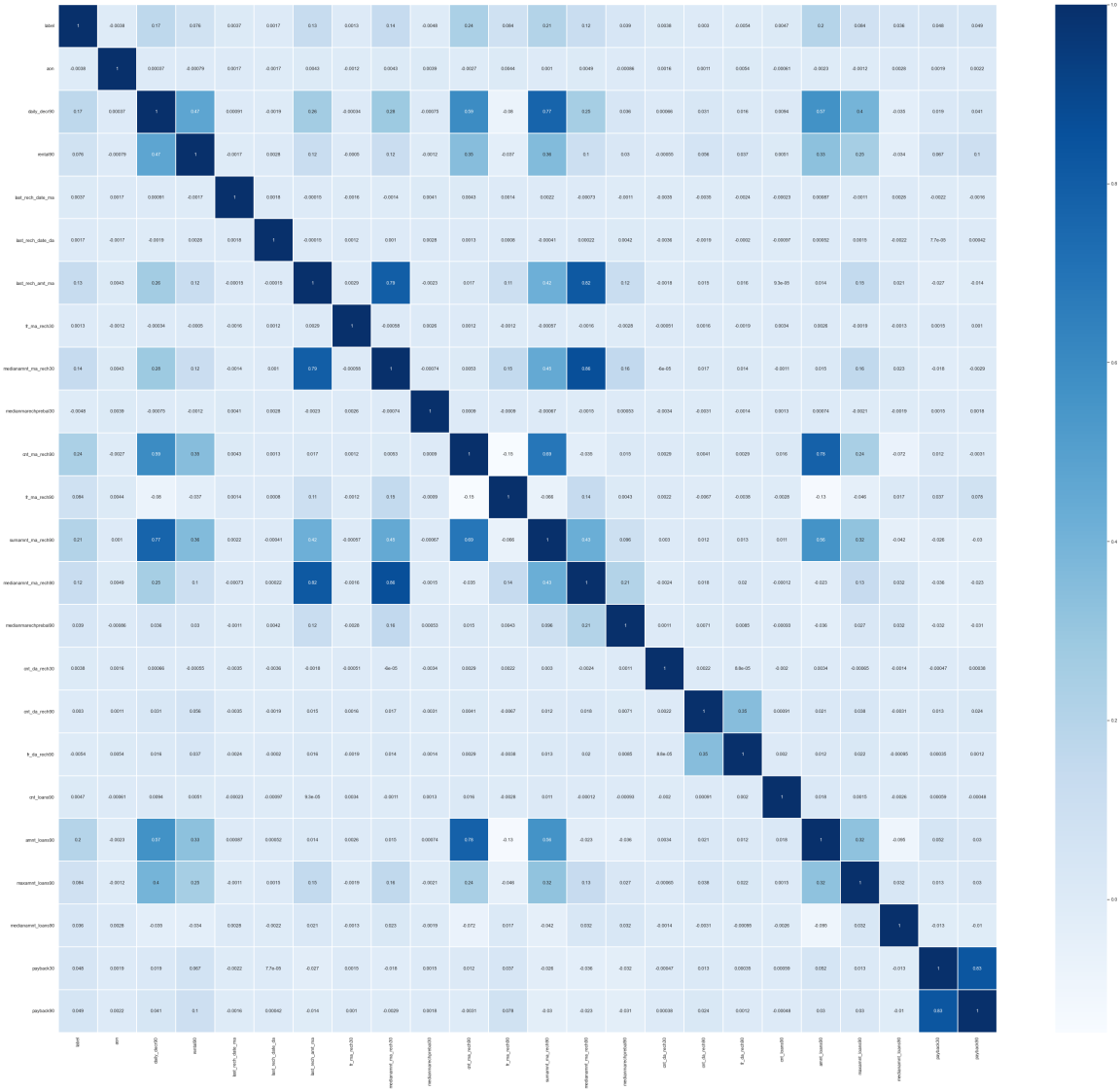
We now will check for skewness using skew(). There is skewness in three col umns which are ‘cnt\_da\_rech30’, ‘fr\_da\_rech90’ and ‘cnt\_loans90’. To treat the skewness we will apply power transformation technique. We then scale the features using StandardScalar().

There is a lot of class imbalance in the target, ‘Label’ and we will treat this imbalance using Over\_Sampling technique.



* Data Inputs- Logic- Output Relationships

Since there are 36 features in the dataset and by various statistical and graphical analysis we found out that only 24 features are actually adding much of value in predicting the ‘Label’. Hence we first drop the unwanted 12 columns. The correlation of value-adding 24 features is as follows:



From the above heatmap we conclude that ‘cnt\_ma\_rech90’ has the maximum positive correlation with the target. As the value of ‘cnt\_ma\_rech90’ increases, there are more chances of a customer being a non-defaulter. ‘sumamnt\_ma\_rech90’, ‘amnt\_loans90’ and ‘cnt\_loans30’ have same positive correlation with the target and it’s second highest. Rest all the features have comparatively less correlation with the target. Moreover, ‘aon’ , ‘fr\_da\_rech90’ and ‘medianmarechprebal30’ are negatively correlated to the data at a very minimum number.

* Hardware and Software Requirements and Tools Used

This project has been coded in Python language using Anaconda 3 Jupyter Notebook package. The libraries that we have installed and used are as follows:

1. **Pandas:** Pandas is a free data manipulation and analysis library created for the Python programming language. . For data science projects, simplified data representation facilitates better findings. Pandas aids in the speeding up of data processing. Pandas has a large feature set that you may use on our data to adapt, change, and pivot it according to our own preferences. This makes it easier to get the most out of our data.
2. **Numpy:** Numerical Python is what it stands for. NumPy aids in the creation of arrays. It is the most important Python package for scientific computing. Here we have used numpy to calculate absolute value for detecting outliers using zscore and for converting the list elements into array object.
3. **Matplotlib**: Matplotlib is a Python data visualisation package. It primarily assists us in the plotting of two-dimensional graphs. It is based on the Numpy and Scipy Python frameworks.
4. **Seaborn**: Seaborn is a Python package that helps to visualise data and make it more understandable to the user. We may plot our data and create a graphical representation of it with the help of the library.
5. **Sklearn**: Scikit-learn is an open-source machine learning library. Scikit-learn offers a large number of machine learning methods to users. The framework library focuses on data modelling rather than data loading, summarization, or manipulation.
6. **Scipy**: SciPy is a Python-based open-source library for mathematics, scientific computing, engineering, and technical computing. SciPy has a number of sub-packages that aid with the most prevalent problems in Scientific Computation.
7. **Imblearn**: The Imblearn library was created to deal with datasets that are unbalanced. To handle and remove the imbalance from the dataset, it offers several approaches such as undersampling and oversampling.
8. **Pickle**: For serialising and de-serializing a Python object structure, the Python pickle package is used. Pickling an object in Python allows it to be saved on disc.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

Firstly we try to understand the dataset. In the dataset we see that the target is already mentioned, which makes it a problem of Supervised Machine Learning.

Based on the type of target, we have decided what algorithms to be used. Here the target is already classified hence we will use Classification algorithm on the dataset.

Before applying algorithms we have cleaned and pre-process the data. After that we have split the dataset into training model and testing model.

Steps we have followed are:

1. 1. Data collection and analysis.
2. Data cleaning and pre-processing
3. Building the model and selecting the best one.
4. Conclusion and saving the model.

Testing of Identified Approaches (Algorithms)

Since the dataset is quite large in size and it’s a classification problem, we have applied 5 models. These are:-

1. Logistic Regression

2. Stochastic Gradient Descent Classifier

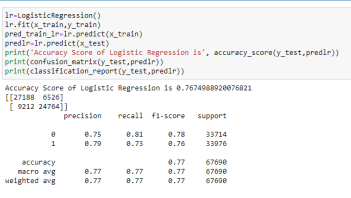
3. Decision Tree Classifier

4. Gradient Boosting Classifier

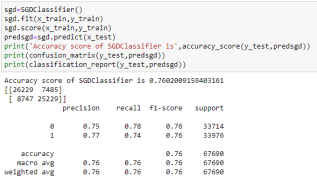
5. Random Forest Classifier

* Run and Evaluate selected models

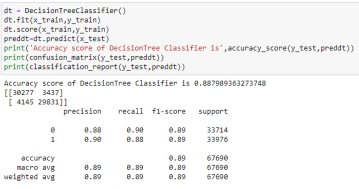
1. **Logistic Regression**: One of the most useful machine learning techniques in the world of statistics is logistic regression. It can be used to solve problems with binary and multi-class categorization. Because it can generate probabilities and classify new data using both continuous and discrete datasets, logistic regression is a key machine learning approach.



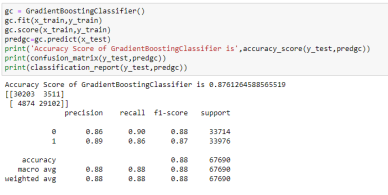
1. **Stochastic Gradient Descent Classifier**: Gradient descent is one of the most widely used optimization methods, and it is by far the most frequent method for optimising neural networks. The gradient is computed using stochastic gradient descent (SGD) with a single sample.



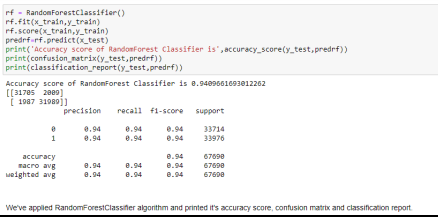
1. **Decision Tree Classifier**: Because it can tackle a wide range of issues, Decision Tree is regarded as one of the most useful Machine Learning algorithms. This algorithm checks the values of the root attribute with the values of the record (actual dataset) attribute and then follows the branch and jumps to the next node based on the comparison.



1. **Gradient Boosting Classifier:** Models of this type are popular because of their ability to accurately classify datasets. The aim behind "gradient boosting" is to take a weak hypothesis or learning algorithm and make a series of modifications to improve the hypotheses/strength. Learner’s when doing gradient boosting, decision trees are commonly employed.



1. **Random Forest Classifier**: The random forest classifier is a supervised learning method that can be used to solve problems involving regression and classification. Due to its high flexibility and ease of implementation, it is one of the most common machine learning algorithms.



* Key Metrics for success in solving problem under consideration

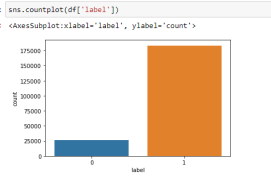
The following metrics used in the project.

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

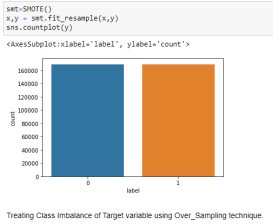
* **Recall** is used as a measure of quantity and high recall means that an algorithm returns most of the relevant results.
* **Accuracy** score is used when the True Positives and True negatives are more important. Accuracy can be used when the class distribution is similar.
* **F1**-**score** is used when the False Negatives and False Positives are crucial. While F1-score is a better metric when there are imbalanced classes.
* **Cross** **validation** **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.
* **AUC**-**ROC** Curve: A ROC curve is a graph that shows how well a classification model performs across all categorization levels. The True Positive Rate is on the y-axis, while the False Positive Rate is on the x-axis, and the plot shows the TPR and FPR values as the threshold is changed.
* Visualizations

We have drawn countplot, histogram, distribution plot, kde plot, pie charts, strip plot,box plot, bar graph and heatmap.

1. Countplot: A countplot counts the categories and returns the number of times they occur. It's one of the seaborn library's more straightforward plots.

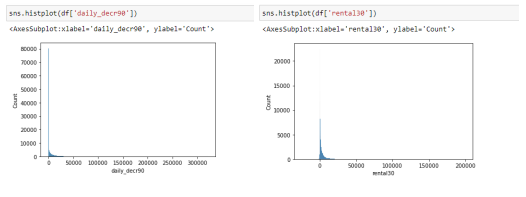


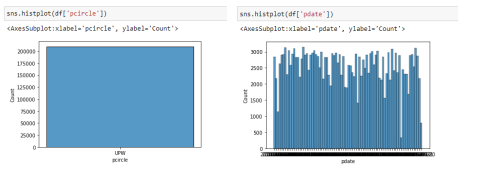
**This is a countplot of ‘Label’, before treating class imbalance**.



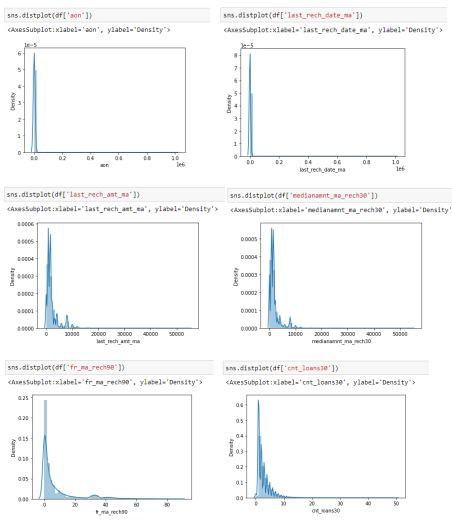
**This is a countplot of ‘Label’, after treating class imbalance.**

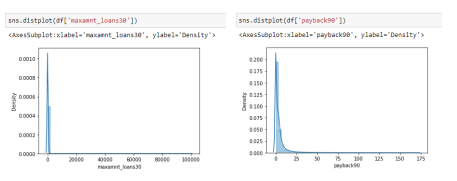
1. **Histogram**: A histogram is a graphical representation of continuous data in a categorical format. Unlike a bar graph, there are no gaps between the bars in a histogram.



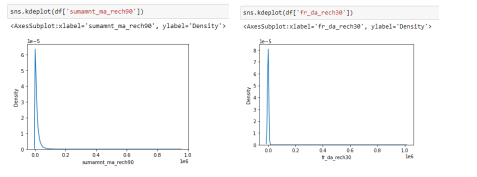


1. **Distribution Plot:** It is mostly used for univariant sets of observations and visualises them using a histogram, i.e. just one observation is used, and hence one column of the dataset is chosen.

****

****

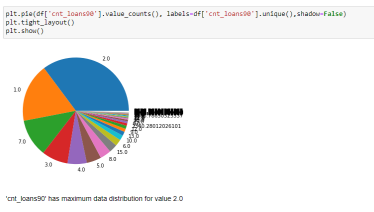
1. **Kernel Density Estimate Plot**: A kernel density estimate (KDE) plot, similar to a histogram, is a method for showing the distribution of observations in a dataset. KDE uses a continuous probability density curve in one or more dimensions to represent the data.

****

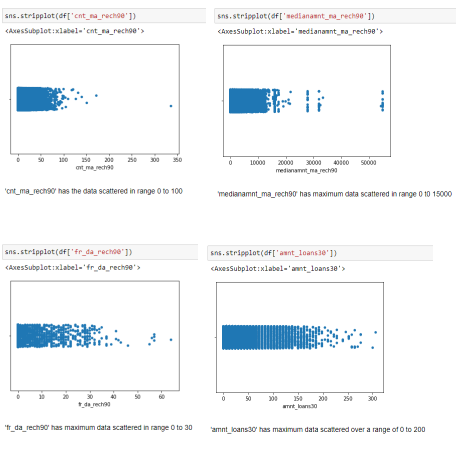
1. **Pie Charts**: A Pie Chart is a circular statistical layout that can only show one set of data at a time. The overall percentage of the provided data is represented by the chart's area. The proportion of sections of the data is represented by the area of the pie slices.

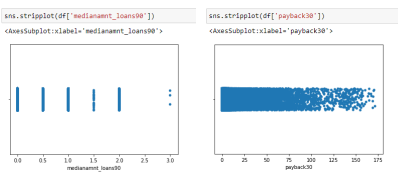
****

****

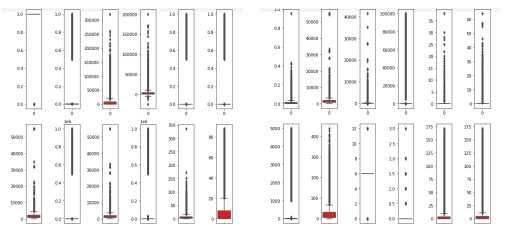
****

1. **Strip Plot:** A strip plot is a plot that is drawn on its own. In circumstances when all data are given together with some representation of the underlying distribution, it is a nice complement to a boxplot or violinplot. It's used to generate a scatter plot based on the selected category.

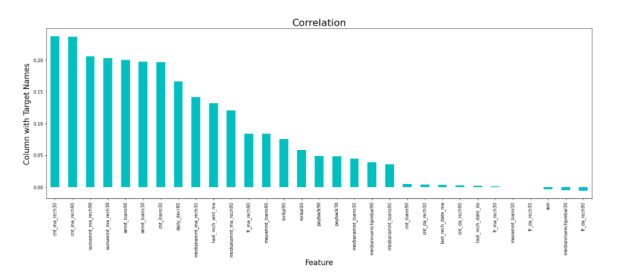
****

****

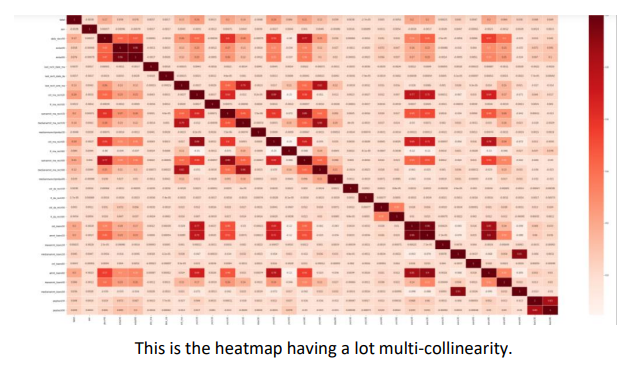
1. **Box Plot:** A single box plot can be used to illustrate multiple statistics from a vast quantity of data. It uses a number line to show the range and distribution of data. Box plots provide some insight into the symmetry and skewness of the data. Outliers are also visible in box plots.

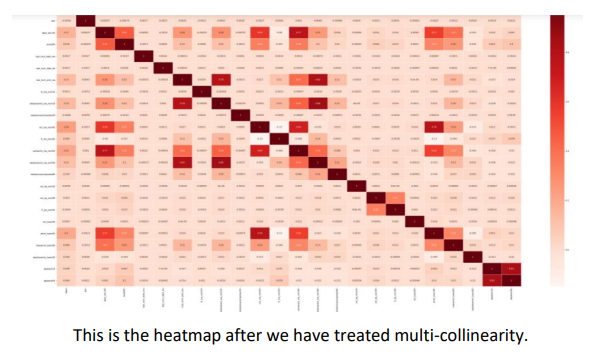
****

1. **Bar Graph**: The bar graph makes it simple to compare various sets of data among different groups. It depicts the connection using two axes, with discrete values on one axis and categories on the other. The graph depicts the most significant changes in data over time. A bar graph is a chart that graphically depicts the comparison of several data types. It uses parallel rectangular bars of identical width but varied length to show grouped data. Each rectangular block represents a distinct category, and the length of the bars is determined by the values they represent.

****

1. **Heatmap**: A heatmap is a two-dimensional graphical representation of data that uses colours to represent the individual values in a matrix. In a color-coding scheme, each data value is represented by a colour. The bigger the amount, the deeper the shade; the higher the value, the narrower the dispersion, and so on.

****

****

* Interpretation of the Results

From the visualization we have interpreted that the target has class imbalance which needs to be treated. There are many features with same same data distribution and same data value, just defined over different time periods, which are causing multi-collinearity. Hence we will have to drop few features to treat this multi-collinearity. Also, there are so many outliers in the dataset but if we remove all, we will be left with very less dataset. In order to not to lose a large part of dataset, we will have to treat with zscore value greater than 5. Also the few features have skew-ness and we will have to apply power transformation to treat such skewness. The features are not scaled properly. Hence we will also have to scale them. While reading the heatmap we also find that there are two features which are not related to target at all, hence we will also drop that. Date and Pcirlce does not add much value to the target prediction, hence we will remove those features as well.

After applying the algorithms, we find out that accuracy, precision, cross validation and F1 score of Random Forest Classifier is quite appreciable, hence we will apply hyper parameter tuning over Random forest classifier and use the hyper parameter tuned model as our final model for ‘Label’ prediction. This hyper parameter tuned final Random Forest Classifier gives us 94.23% accuracy.

**CONCLUSION**

* Key Findings and Conclusions of the Study

Mobile Money is defined as a service that allows unbanked and low income people to access financial services such as payments for goods, services, and bills through mobile devices.

The spread of Information and Communication Technologies has been shown to be a complicated process that can also be observed at the country level. Economy, culture, technology, and politics are among the country-level factors. The regulatory environment, existing alternatives, agents' behaviours, the cellular market landscape, and service providers' market share were the primary elements that influenced mobile money diffusion.

To summarise, this study's theoretical framework and synthetic analysis are concerned with mobile financial services in various economic and social viewpoints and circumstances.

* Learning Outcomes of the Study in respect of Data Science

Identifying faulty components (the features that are repeating themselves and just adding baggage to the dataset), developing solution plans (cleaning and pre-processing the data to efficiently build models), and putting the necessary changes in place without jeopardising the present system's functionality. To arrive at a realistic solution, the individual problem must be abstracted and decomposed. Identifying components that may help to solve problems. Large-scale informatics efforts require analysis. The well the dataset has been analysed, the well will be the accuracy of the model. Identifying jobs that can be automated, comprehending underlying corporate processes and assessing business wants.

* Limitations of this work and Scope for Future Work

We recommend learning more about the distinctions between mobile banking and mobile money services.

It's crucial to figure out why one service is more effective in specific nations than the other. It would be fascinating to compare and contrast developed and developing countries.

Another option is to look at the elements that influence the spread of mobile financial services in developing nations, such as country culture, financial literacy, technological progress, and the financial system.

It would be worthwhile to conduct a thorough investigation of the circumstances that contributed to their failure.

This work has a lot of scope of improvement.