

**Micro Credit Project**

Submitted by:

Sourav Chaudhury

**ACKNOWLEDGMENT**

I would like to take this opportunity to show my gratitude to the resources mentioned below for helping me complete this project successfully.

* Flip Robo Technologies
* DataTrained Education
* TowardsDataScience
* AnalyticsVidya
* StackOverflow
* GitHub

**INTRODUCTION**

* **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. They understand the importance of communication and how it affects a person’s life, thus, focusing on providing their services and products to low income families and poor customers that can help them in the need of hour. They are collaborating with an MFI to provide micro-credit on mobile balances to be paid back in 5 days.

The Consumer is believed to be 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).

* **Conceptual Background of the Domain Problem**

The rapid enhancement in technology, new players and huge investments coming up in the financial technology, mobile phone penetration are remodelling the financial services vista. With enhanced penetration of mobile devices globally, the opportunity to gain more understanding about consumer behaviour, comprising how they spend their money and time, has also enhanced. They can now target a distinct group of people using such knowledge. Data analytics market also relies on several mobile apps as their engine of information is stored in a data warehouse. Not just the mobile industry; Big Data influences other markets as well, such as micro finance institutions (MFIs). By incorporating digitisation.

Data Analytics plays an innovative role in the unique ecosystem. It helps to fulfil the evolving customer demands & expectations. The microfinance industry has seen rapid growth. This growth has become a reality as, a result of investors continued support to MFIs with equity infusion of around Rs 9,443 cr which is an increase of 55% from Q1 FY19-20. It can help in the development of application scorecards for making an unbiased selection of customers for funding. Customer behaviour patterns and critical information derived from industry data integration into such scorecards help to derive quick credit decisions. In a few mins, at the front end, such credit decisions are attainable through digital channels of customer service officer's handheld device. Thus, this helps to save the time spent in verification, reference checks and validation by fulfilment executive.

Adding up to, this data analytics drives customer loan approval, request processing and disbursement of the loan amount. With the help of Data Analytics, the Microfinance Industry will reach a position where they will have the ability to provide products that customers exactly need. Based on the customer needs visible through analytics such as price modelling, customer segmentation, and data modelling; MFIs will be able to service customers with just-in-time financial needs, with the right loan ticket size and insurance schemes. Microfinance companies will be able to predict portfolio behaviours at various geographies and hence, rightly analyse defaults or credit losses more precisely. Accordingly, companies can select the right portfolio for their financial products.

* **Review of Literature**

MFIs need to map their customer base to their products and this is possible only if demographic data is available to the Branch Manager of a MFI, of a particular area or region. This will enable him to target potential customers residing in that area or region.  Geospatial analytics can support this requirement. The many applications of geospatial analysis include human population forecasting by filtering out relevant data and applying it to provide accurate trend analysis, modelling and predictions.  By using this analytical technique, the Customer Sourcing Executives can expand their customer base by targeting those who have not been tapped yet.

Predictive Analytics plays a key role in coming up with behavioural patterns to determine whether a Customer is likely to default. Every MFI has to submit the KYC (Know Your Customer) data to an external agency which collates the data from all MFIs operating in the country. They also collate data of all loan defaulters. Analytics performed on this database can come up with patterns with respect to Customer behaviour so that the MFIs can target only those, who have a clean track record. This will reduce delinquencies to a great extent.

Profitability based Analytics: Most of the MFIs are upgrading to small Banks to cater more or less to their existing customer base. If the MFIs used analytics to identify the profitable customers, they could also become target customers for the small Bank as well. Customer Lifetime Value or CLTV is a measure used for deriving the value of a customer relationship, based on the NPV or Net Present Value of the projected future cash flows from the customer relationship. Customers with a better NPV are the profitable ones. With the support of Analytics, the MFI can track such customers to take them to the next level. Those with negative NPV should be ignored in this respect while effort should be made to improve recoveries from such Customers.

Analytics can also help derive a Customer’s SOW or Share of the Wallet, which is the amount of the customer’s total spending on the products and services that the MFI offers. SOW can be increased by cross-selling other products or services thereby creating a loyalty factor. MFIs have revolutionized the way “financial inclusion” has been adopted in developing countries. India, for example, has perhaps the largest “unbanked” population in the world and the statistics is really alarming. The unbanked adult population stands at roughly 40% (Urban) and 60% (Rural) while only 14% (Urban) and 10% (Rural) households have loan accounts. The MFIs catering to this large unbanked population, have indirectly created self-employment opportunities especially for women, who have become independent or sole bread winners in the family.  MFIs offering loans to Micro and Small & Medium Enterprises (MSMEs) indirectly support creation of employment opportunities thereby extending the benefit to a larger community.

The growth of MFIs indirectly support economic growth and importantly more power to women. To support growth of MFIs, Analytics has to be optimally leveraged and this is what the white paper is all about. MFIs need to map their customer base to their products and this is possible only if demographic data is available to the Branch Manager of a MFI, of a particular area or region. This will enable him to target potential customers residing in that area or region.  Geospatial analyticscan support this requirement.

The many applications of geospatial analysis include human population forecasting by filtering out relevant data and applying it to provide accurate trend analysis, modelling and predictions.  By using this analytical technique, the Customer Sourcing Executives can expand their customer base by targeting those who have not been tapped yet. Predictive Analytics plays a key role in coming up with behavioural patterns to determine whether a Customer is likely to default. Every MFI has to submit the KYC (Know Your Customer) data to an external agency which collates the data from all MFIs operating in the country. They also collate data of all loan defaulters. Analytics performed on this database can come up with patterns with respect to Customer behaviour so that the MFIs can target only those, who have a clean track record. This will reduce delinquencies to a great extent.

Profitability based Analytics: Most of the MFIs are upgrading to small Banks to cater more or less to their existing customer base. If the MFIs used analytics to identify the profitable customers, they could also become target customers for the small Bank as well. Customer Lifetime Value or CLTV is a measure used for deriving the value of a customer relationship, based on the NPV or Net Present Value of the projected future cash flows from the customer relationship. Customers with a better NPV are the profitable ones. With the support of Analytics, the MFI can track such customers to take them to the next level. Those with negative NPV should be ignored in this respect while effort should be made to improve recoveries from such Customers. Analytics can also help derive a Customer’s SOW or Share of the Wallet, which is the amount of the customer’s total spending on the products and services that the MFI offers. SOW can be increased by cross-selling other products or services thereby creating a loyalty factor. Dashboards offer a bird’s eye of KPIs (Key Performance Indicators) relevant to a particular business process through effective use of visual charts or graphs.

Analytical Reporting is perhaps the most critical part presenting the required information in different forms to the top management. This tool will enable multi-dimensional analysis, what if analysis and data drill down analysis. Score carding helps in measuring and monitoring the KPIs against the defined strategic goals and performance milestones, through effective use of traffic light indicators. The focus of MFIs is to track their Customers over a period of time, wherein they move from a lower loan category to higher loan category based on repayment history. Thus MFIs need to measure Customer retention, as one of their key KPIs. Customer retention ensures increase in Interest income as customers move to higher loan category while newly referenced Customers are added over a period of time. Some of the other key KPIs are related to Sourcing, Disbursement, Outstanding loan portfolio and Collection. MFIs would like to measure these KPIs against the goals or targets set at the start of the year. Measuring these KPIs at regular intervals, will help MFIs monitor the performance of their Branches, Regions or Zones. Using forecasting models, MFIs can ensure proactive performance monitoring. This will help the Organization, well in advance, to take preventive steps and avoid negative scenarios.

* **Motivation for the Problem Undertaken**

Microfinance is widely accepted as a poverty-reduction tool, representing $70 billion in outstanding loans and a global outreach of 200 million clients. The study here pertains client that is in Telecom Industry. They are a fixed wireless telecommunications network provider. They have launched various products and have developed its business and organization based on the budget operator model, offering better products at Lower Prices to all value conscious customers through a strategy of disruptive innovation that focuses on the subscriber.

They understand the importance of tele 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).

**Analytical Problem Framing**

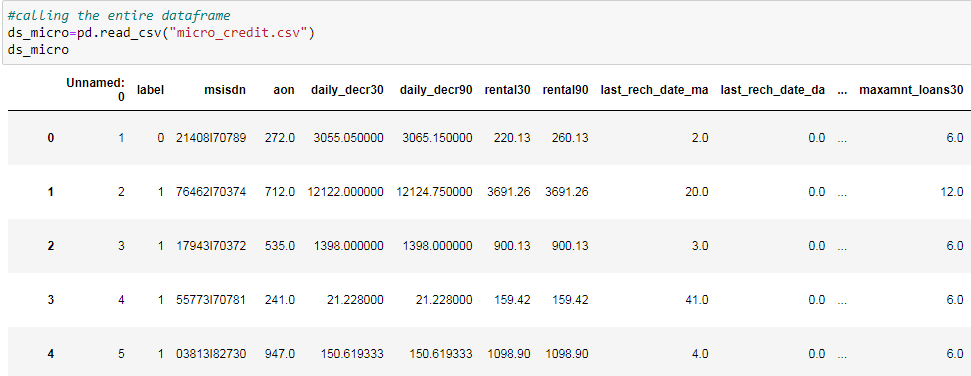
* **Mathematical/ Analytical Modeling of the Problem**

We have received the data from the client and we needed to find out the people that may default the loan payback. An extensive EDA (Exploratory Data Analysis) is performed to show the counts of different variables in the dataset and also their relationship with the target variable. There were no missing values in the dataset. I have used descriptive statistics in order to find the statistical elements of the data provided by the client. The descriptive statistics gave me certain insights and also helped me to pre conceptualize ways to deal with the problem.

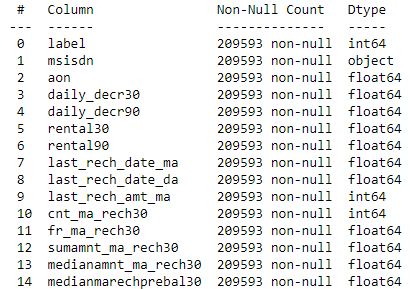
There were around 35 columns of input variable and one target variable that the loan payers and the defaulters. The data went through the pre processing steps which included replacing the given values with the absolute values in the dataset, removing outliers, fixing skewness and finally quantifying the categorical values in order to feed them into the model. Later several algorithms were used to check which works best with the given dataset in given circumstances.

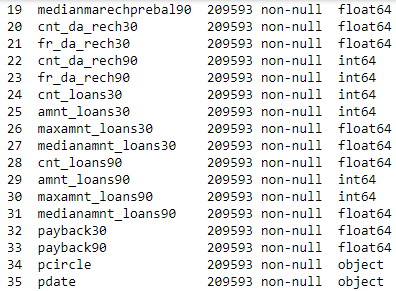
* **Data Sources and their formats**

There was only one file provided for the training and testing data. The dataset was given in a csv format. There was a separate xls file provided for the columnar description of the variables. Below is the snapshot of the dataset. All the variables in the dataset are numeric in nature.



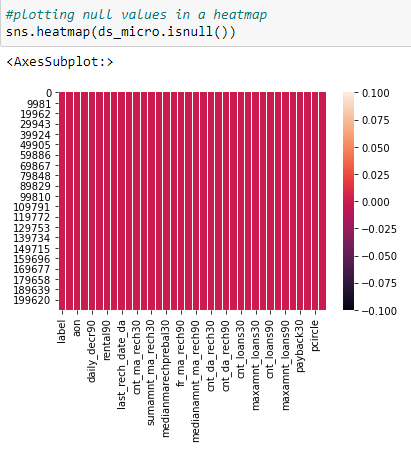
Below is the description of information of variables in the dataset





**Checking for Null Values**

We have already checked that the dataset has no missing values and now we will plot them in a heatmap for a better visual understanding.

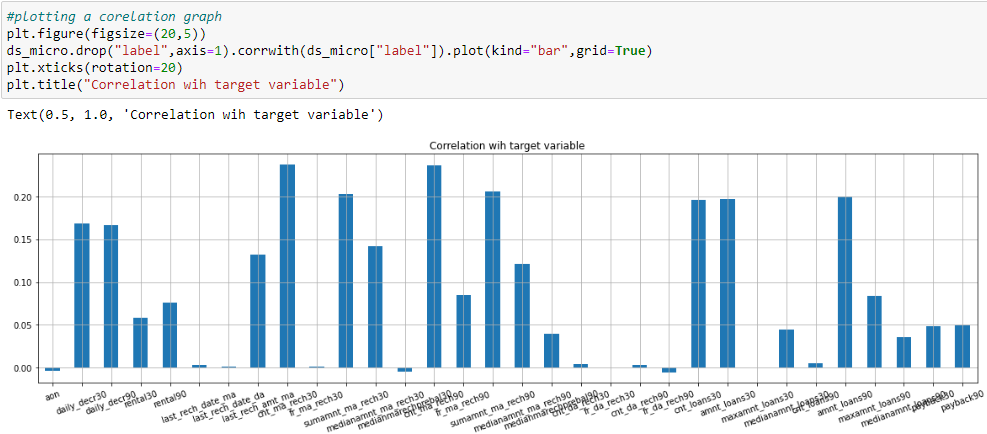


* **Data Pre-processing Done**

Data Preprocessing is one of the most important part of making any machine learning model. It involves processing the raw data in a way that it is easily understandable by the model for higher efficiency and efficacy. Below are some of the pre processing done to this dataset.

**Dropping Columns**

Correlation graph is plotted considering the target variable i.e. the labels of the loan payer or the defaulter. Any variable that shows a negative or lesser correlation is with it will be dropped. Below is the snapshot of the correlation graph and the codes to drop the variables showing a negative or lesser correlation with the target variable.

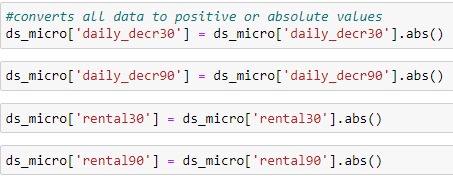




**Converting to Absolute Values**

There are some variables in the dataset with negative values. Below are the variables with negative values which are to be converted to absolute values. We will use the below code to convert them to absolute values.

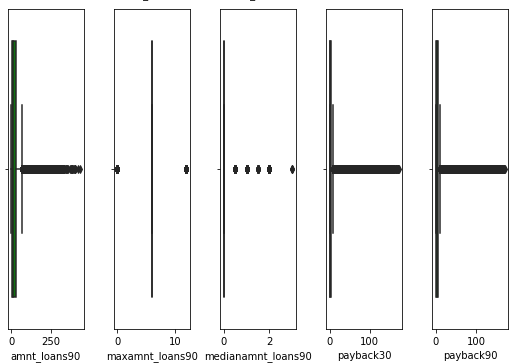
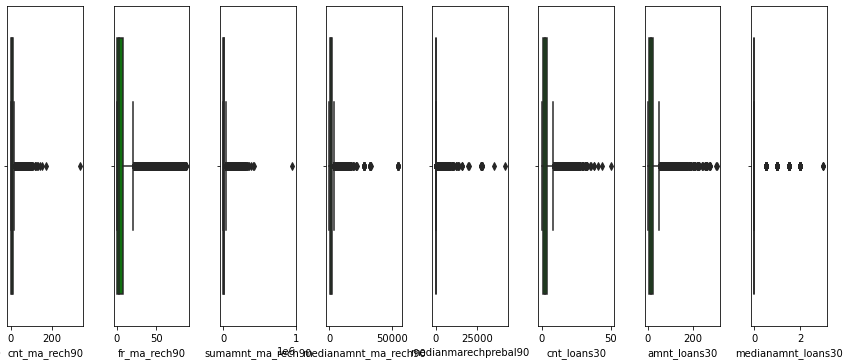
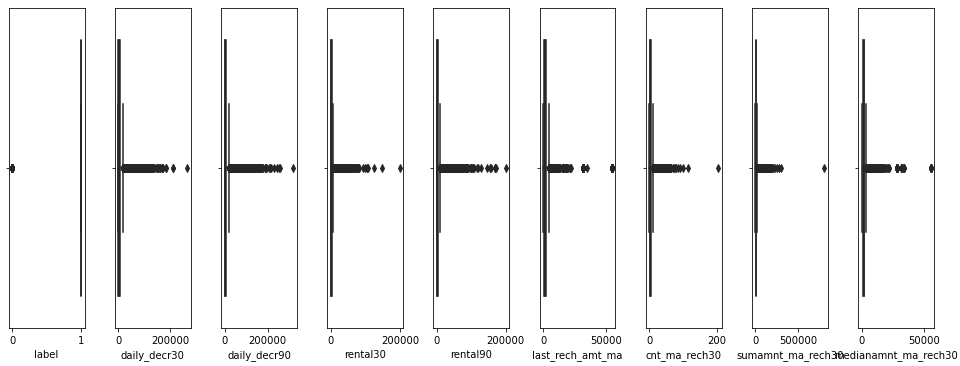
* Daily amount spent from main account, averaged over last 30 days (in Indonesian Rupiah)
* Daily amount spent from main account, averaged over last 90 days (in Indonesian Rupiah)
* Average main account balance over last 30 days
* Average main account balance over last 90 days



**Plotting Outliers**

Plotting outliers are also an essential part of pre processing and making an efficient machine learning model. Present of outliers will have affect on variance, and standard deviation of a data distribution. In a data distribution, with extreme outliers, the distribution is skewed in the direction of the outliers which makes it difficult to analyze the data and will result in biased insights.

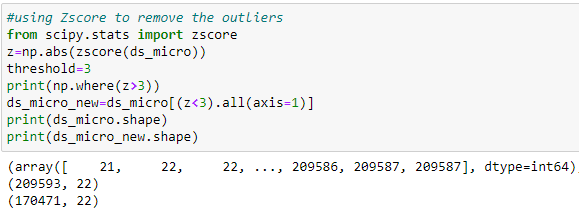
Below are the data points of the variables plotted in a boxplot. We can clearly see the outliers from the below visualization.



**Removing Outliers**

I have used Z score to remove the outliers. A Z-score is a numerical measurement that describes a value's relationship to the mean of a group of values. Z-score is measured in terms of standard deviations from the mean. If a Z-score is 0, it indicates that the data point's score is identical to the mean score. Any points that will be above Z-score 3 will be considered as outliers.

Below is the code to remove the outliers from the dataset.

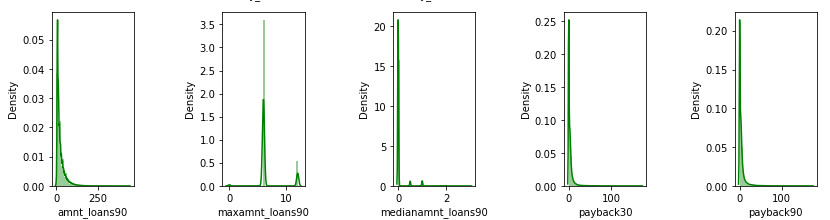
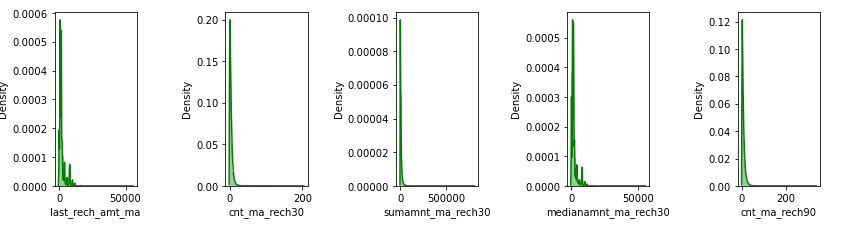
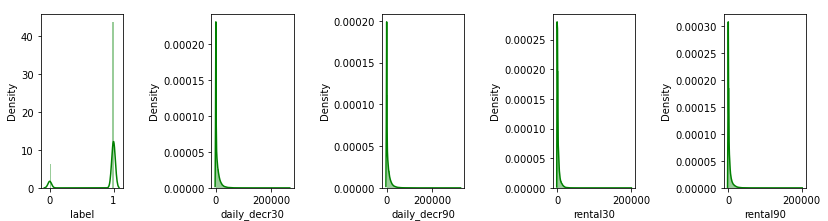


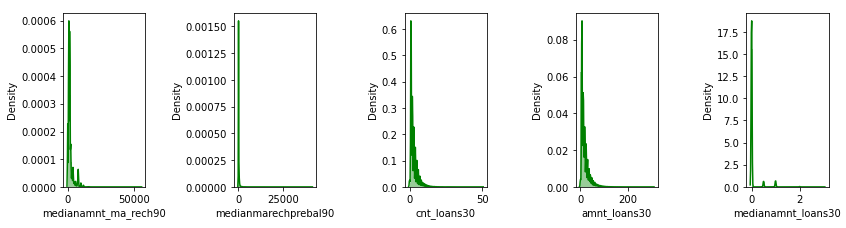
The above shape shows there were 209593 data points before removing the outliers and after removing it came down to 170471

**Checking Skewness**

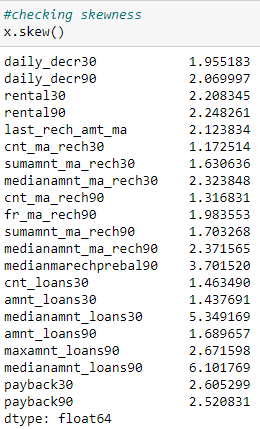
Skewness refers to a distortion or asymmetry that deviates from the symmetrical bell curve, or normal distribution, in a set of data. If the curve is shifted to the left or to the right, it is said to be skewed. A data can be right skewed or left skewed. A easy graphical representation of checking skewness of data is by plotting them in a distribution graph and also listing down there skew values.

Below are some of the data distribution of the columns in the dataset.

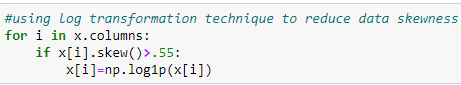




Below is the code to describe skewness of data.



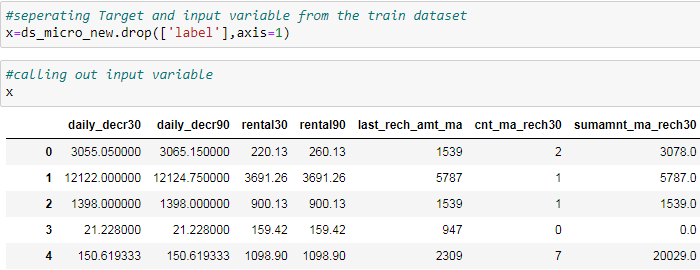
Any columns throwing a value greater than + or - .55 is highly skewed and needs to be treated. Below are the codes to treat skewness of the dataset.

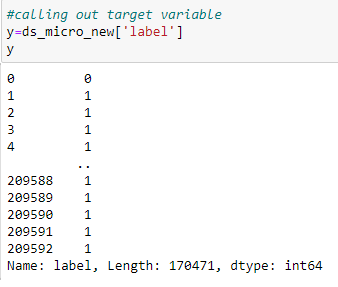


The above code log transforms any columns that shows a skewness greater than + or - .55 and does not manipulates the ones with values lesser than the threshold.

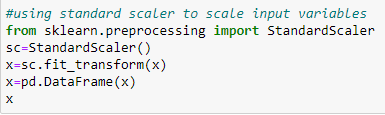
**Separating Input and Target Variable**

The target and input variables are separated from the training dataset. Below is the code to perform the same.

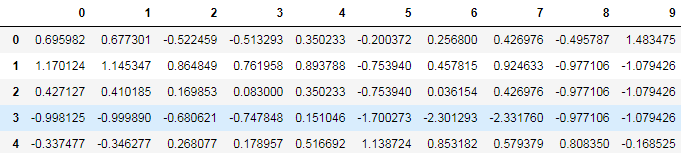
****

****

**Scaling Input Variable**

****

Below are the scaled input variables



**Removing Lesser or Negatively Correlated Data**

A correlation graph is plotted which shows relation between the target variables with the input variables. Any variable which might show a negative correlation or a lesser correlation with the target variable will be dropped from the dataset as including them might reduce the efficiency of the model and might result in the model to overfit. Below are some of the columns which shows negative or lesser correlation and is dropped from the dataset.



**Hardware and Software Requirements and Tools Used**

This project is done using Python 3.0, Ms Excel and Ms Word. The GUI (Graphical User Interface) used in this project is Jupyter Notebook. The coding part of the project is done in Jupyter Notebook while the training and testing data is received in CSV format and the project report is written in Ms Word.

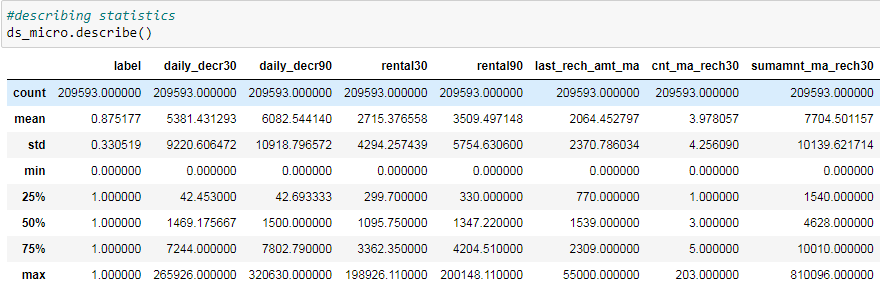
**Model/s Development and Evaluation**

* **Identification of possible problem-solving approaches (methods)**
* **Descriptive Statistics**

A descriptive statistic is a [summary statistic](https://en.wikipedia.org/wiki/Summary_statistic) that quantitatively describes or summarizes features from a collection of [information](https://en.wikipedia.org/wiki/Information). It is the process of using and analysing those statistics. Descriptive statistics is distinguished from [inferential statistics](https://en.wikipedia.org/wiki/Statistical_inference) by its aim to summarize a [sample](https://en.wikipedia.org/wiki/Sample_(statistics)), rather than use the data to learn about the [population](https://en.wikipedia.org/wiki/Statistical_population) that the sample of data is thought to represent.

Various statistical measures that were laid down to understand the data in depth. There was a combination of categorical and numerical data in the dataset. Descriptive statistics were implemented in the numeric dataset to find out various insights. Measures such as count, mean, standard deviation, minimum, maximum and inter quartile ranges are found for each of the numeric variables.

Below is the table that describes descriptive statistics.

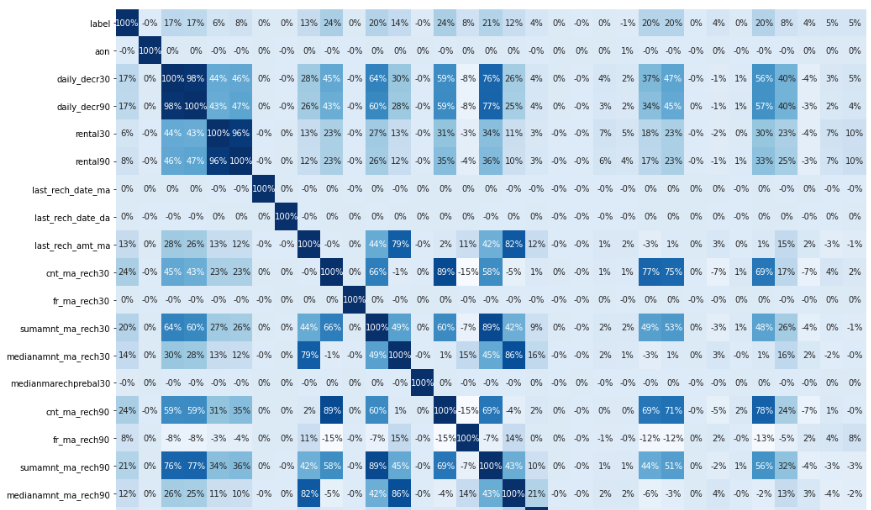


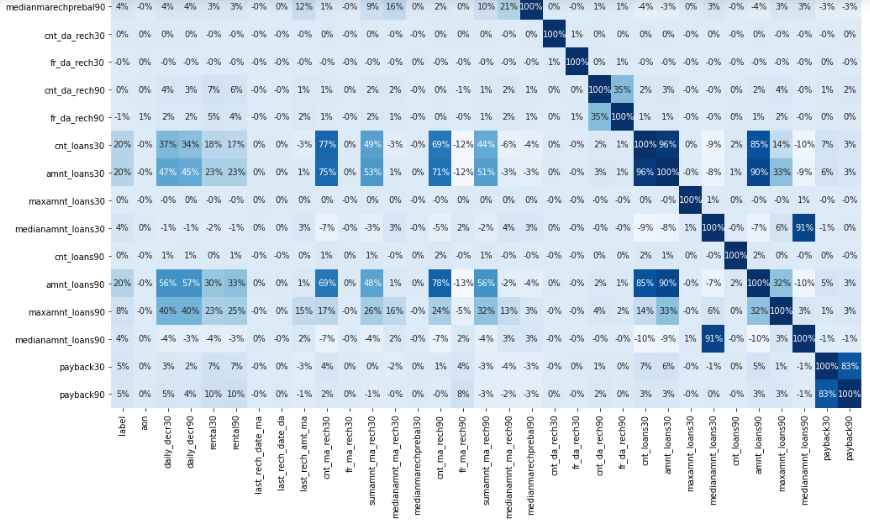
Mean column describes the average value of the column. Std (Standard Deviation) describes how far the points are from the mean. Min and Max columns describes the minimum and maximum values in the column. The 25%, 50% and 75% are the inter quartile ranges. For the columns that have mean is greater than the 50% quartile range indicates that the data in the column is right skewed. For the columns that have mean is lesser than the 50% quartile range indicates that the data in the column is left skewed. Columns which have huge differences in between the maximum and 75% of the quartile ranges means they have outliers in them.

* **Correlation Matrix, Heatmap and Graph**

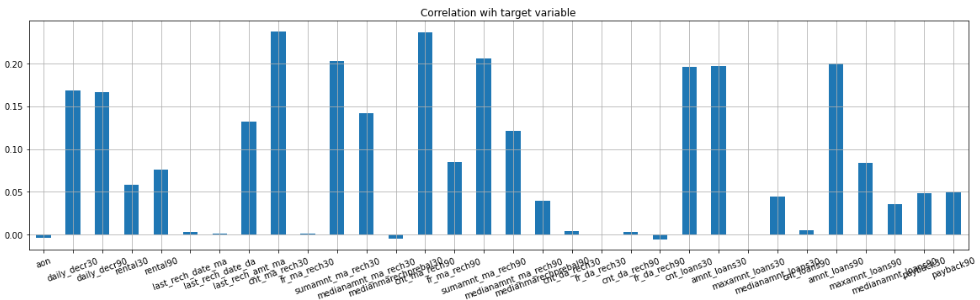
A Correlation matrix is also plotted where we can see the variables and their co-linearity with other variables and most importantly with the target variable. Below is a partial snapshot of the correlation matrix in a heatmap showing linear relationship of variables with each other. The co relation percentage is also mentioned where it shows either the variable is positively correlated or negatively correlated with each other. The co linearity of one variable with the other is more, then the matrix gets darker in shade and vice-versa.

Below is the correlation matrix for the dataset.





Other than the correlation matrix, the other thing that is very useful in understanding co-linearity among variables is to plot a co-relation graph. The below co-relation graph shows the relation of all the variables of the dataset with the target variable.



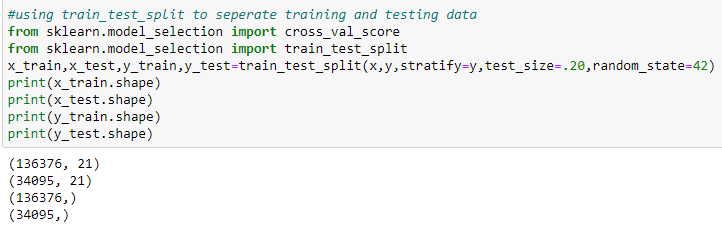
The Y axis shows the percentage of co-linearity with the target variable label and X axis shows the name of the numeric variables in the dataset. The variables which shows lesser co-linearity will be not considered in making the model as considering them might reduce the efficiency of the model. The above diagram is for the training dataset. All the variables that shows negative correlation will be removed.

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

Below are the steps and the algorithms used for training and testing.

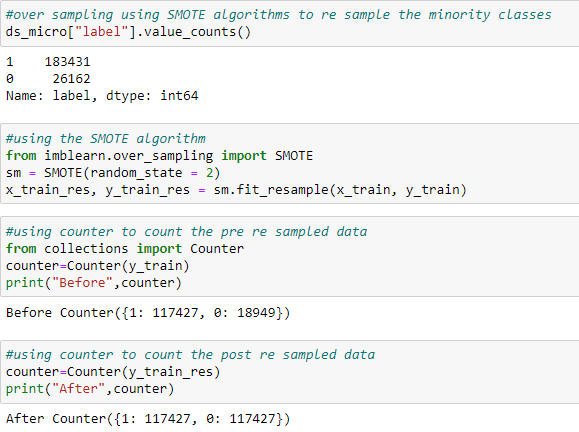
**Separating training and Testing dataset**

The entire training dataset is splitted into training and testing data. The ratio used for this study id 80-20%. 80% of the data is used to train the model the rest 20% is used for testing the model. Below is the code to do the same.



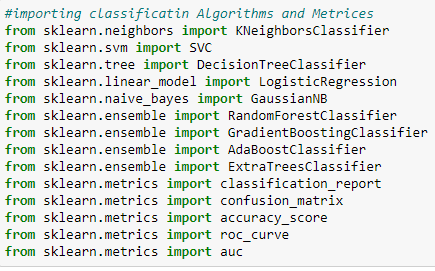
**Over Sampling using SMOTE**

One approach to addressing imbalanced datasets is to oversample the minority class. The simplest approach involves duplicating examples in the minority class, although these examples don’t add any new information to the model. Instead, new examples can be synthesized from the existing examples. This is a type of [data augmentation](https://machinelearningmastery.com/how-to-configure-image-data-augmentation-when-training-deep-learning-neural-networks/) for the minority class and is referred to as the Synthetic Minority Oversampling Technique, or SMOTE for short. Below are the codes to do the same.



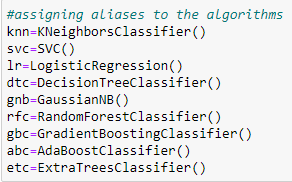
**Importing Classification Algorithms and Metrics**

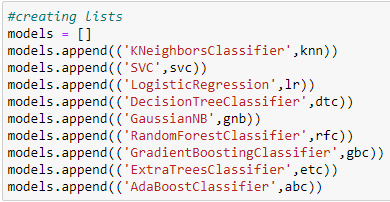
Below are the codes to import the algorithms and metrics that were used in the following study. All the below algorithms and metrics are precisely for the classification study.



**Assigning aliases and Creating Lists**

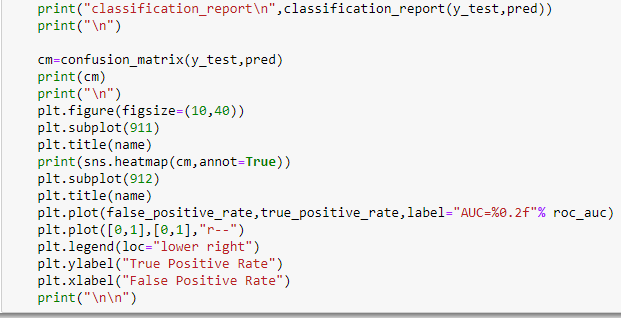
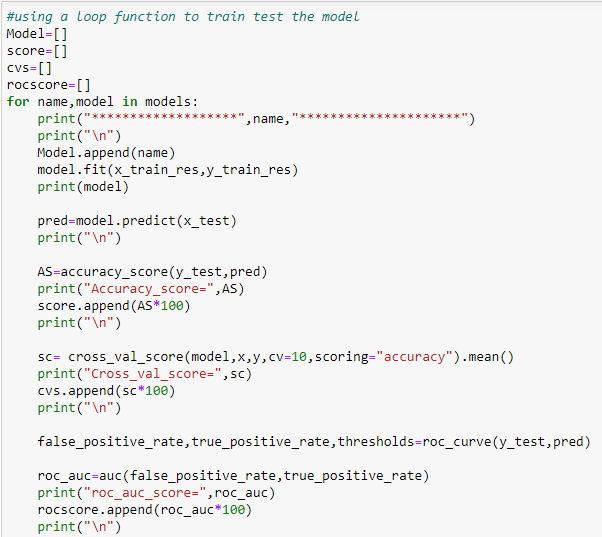
The below steps are for assigning aliases and creating lists





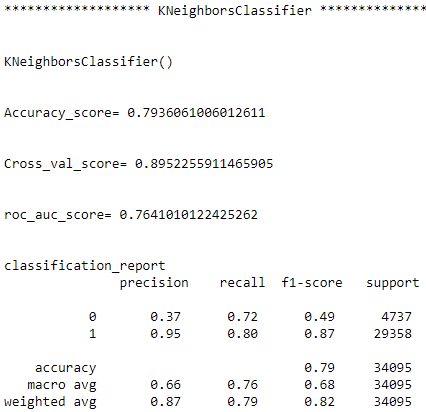
**Machine learning**

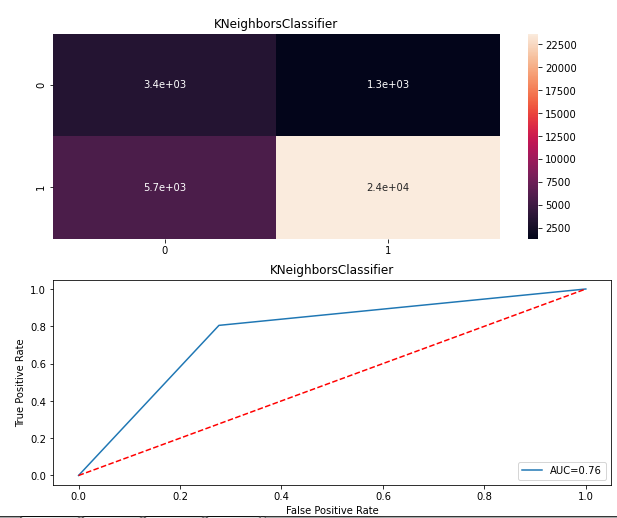
Now the most important part of the entire study, running the training dataset through the different algorithms and using metrics to evaluate the models efficiency. I am using a loop function to test the dataset. Below is the code.

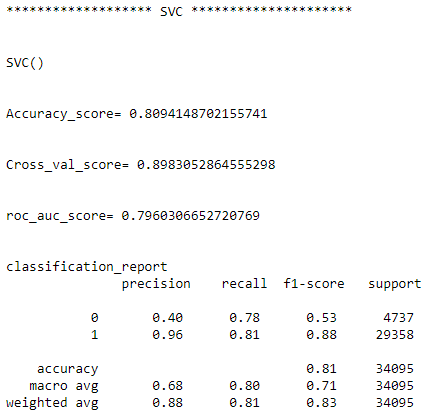


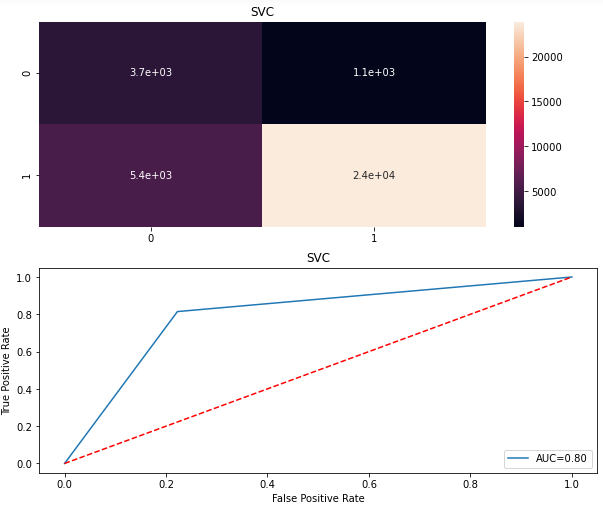
* **Run and Evaluate selected models**

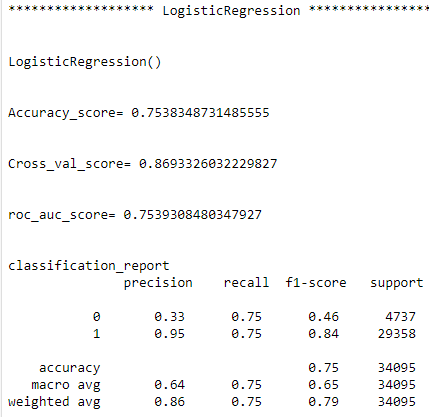
Below are the parameters and score based on different metrics used to evaluate the model.

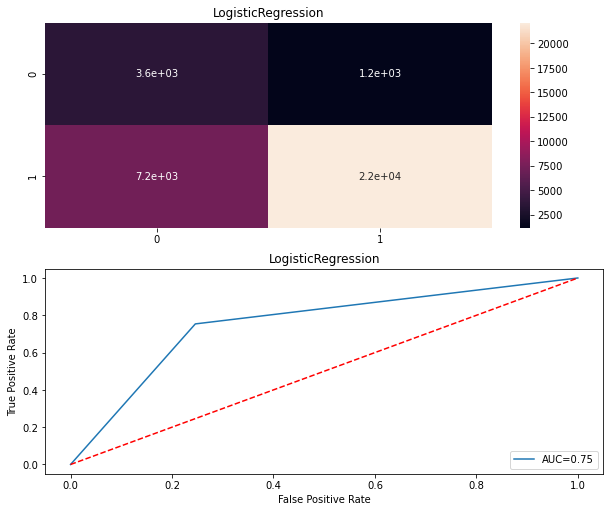


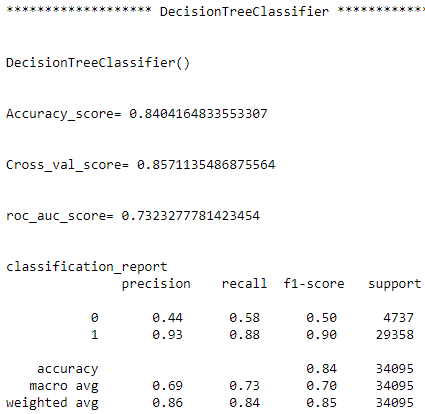


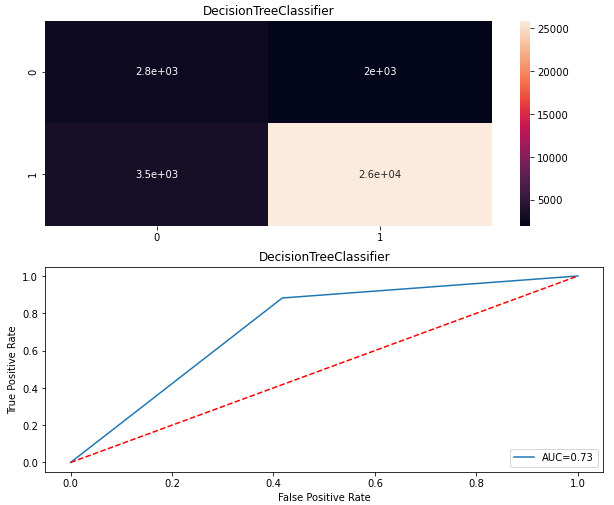


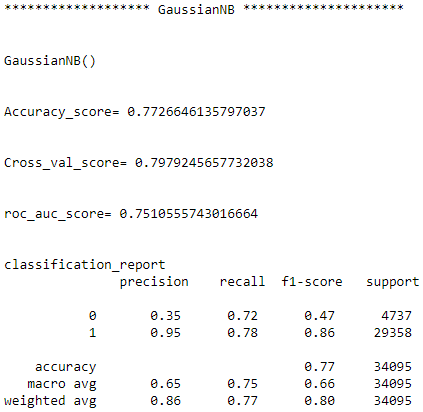


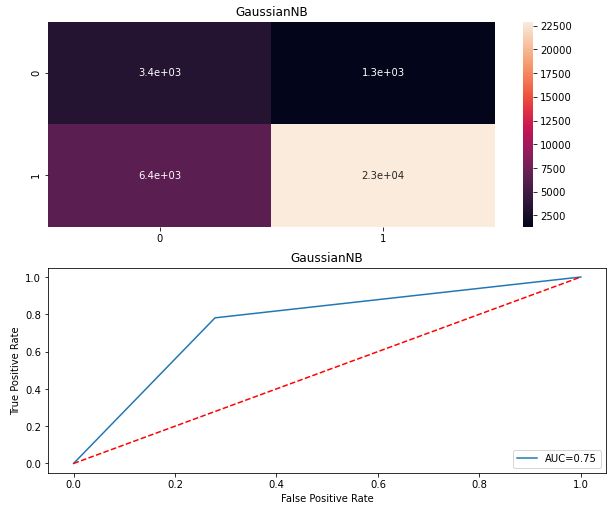


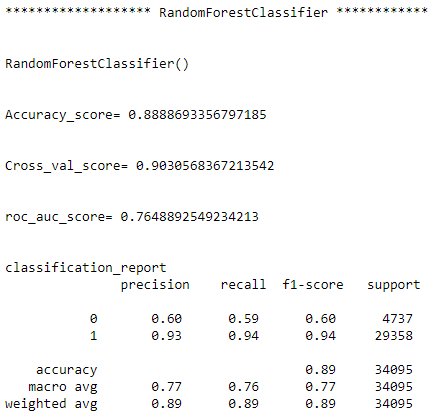


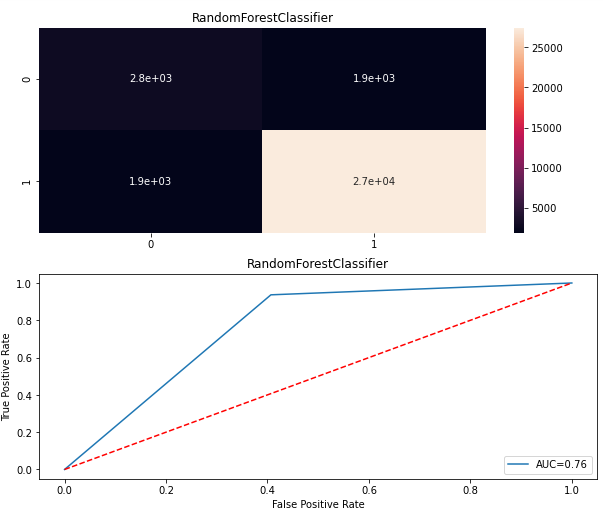


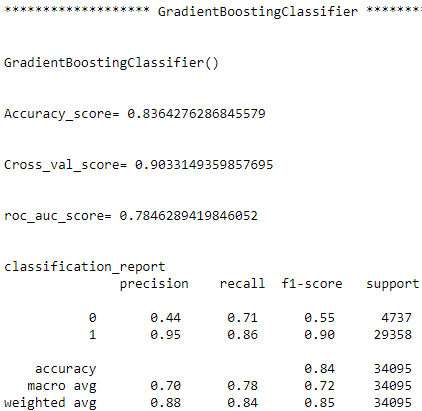


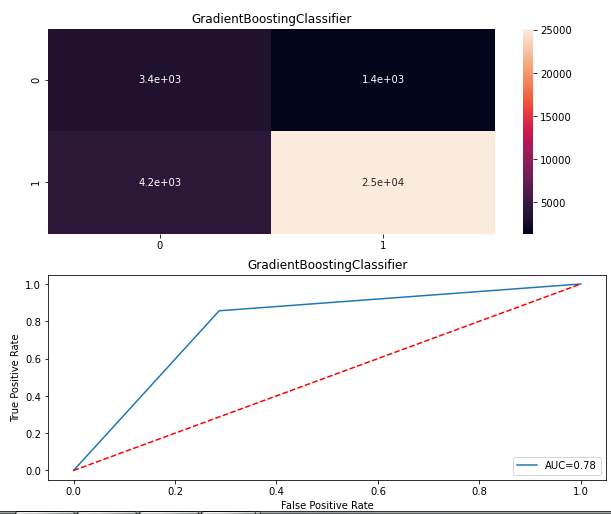


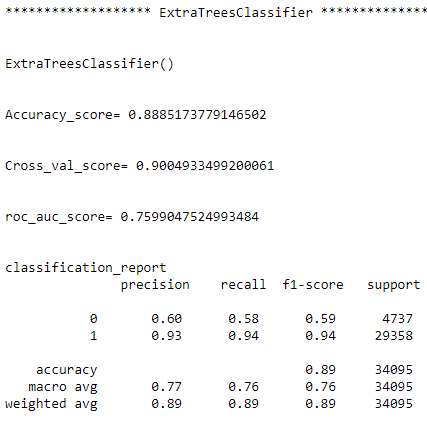


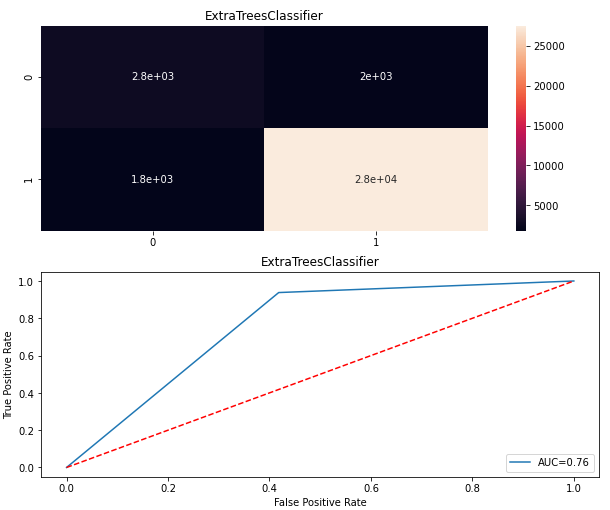


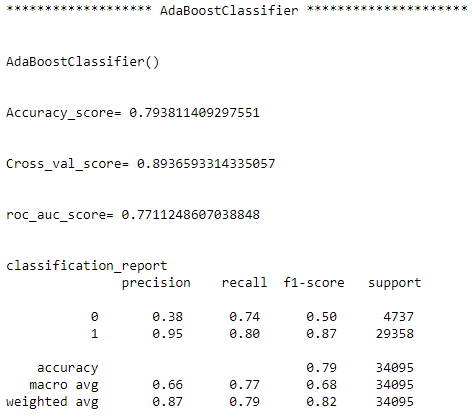


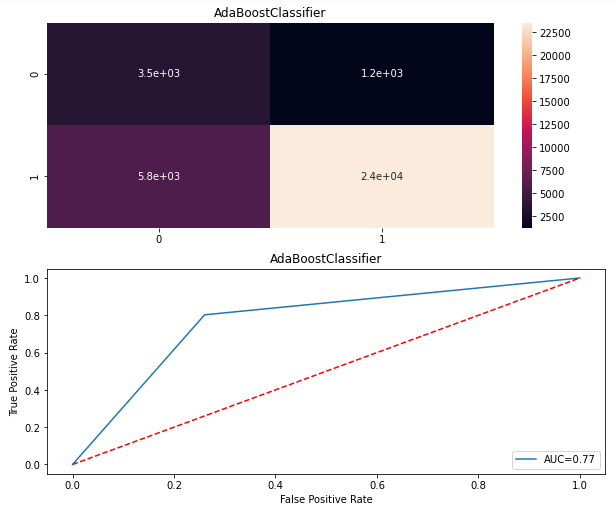




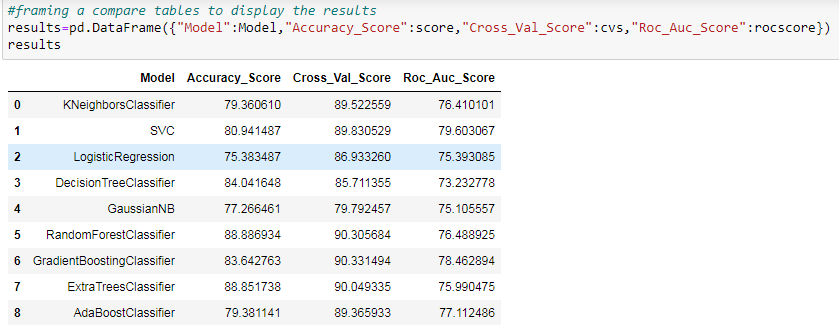




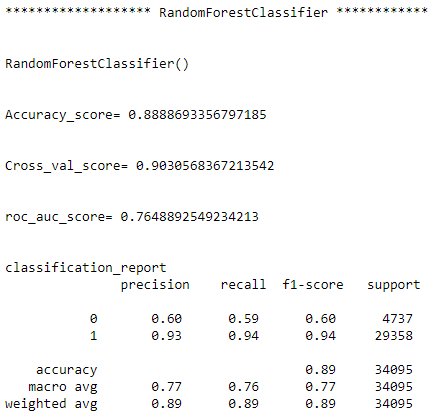




Below is the code to frame a compare table to cross check performance of all the algorithms at once



From the above table we can see that Random Forest Classifier scored the maximum in terms of all the scores. Also if considered the classification report we can see that the above model has done better considering the Precision ,Recall and the F1 Score. Below is the classification report for the Random Forest Classifier. The metric that we should be concerned about based on our problem statement is the precision for the 0s. We can see that the precision of the 0s for Random Forest Classifier is around 60% which is quite higher compared to other models.



* **Key Metrics for success in solving problem under consideration**

Metrics are various parameters that are used to judge the efficiency and efficacy of a model. The metrics used to judge different parameters of this model are below

* Accuracy Score
* Cross Validation Score
* ROC\_AUC Score
* Recall
* Precision
* F1 Score
* Confusion Matrix

**Accuracy Score**- Classification Accuracy is what we usually mean, when we use the term accuracy. It is the ratio of number of correct predictions to the total number of input samples. It works well only if there are equal number of samples belonging to each class.

**Cross Val Score**- A key challenge with overfitting, and with machine learning in general, is that we can’t know how well our model will perform on new data until we actually test it. To address this, we can split our initial dataset into separate training and test subsets. There are different types of Cross Validation Techniques but the overall concept remains the same, to partition the data into a number of subsets, hold out a set at a time and train the model on remaining set and test model on hold out set

**ROC\_AUC Score**- The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve. The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes.

**Recall**- It is the number of correct positive results divided by the number of all relevant samples (all samples that should have been identified as positive).

**Precision**-  It is the number of correct positive results divided by the number of positive results predicted by the classifier.

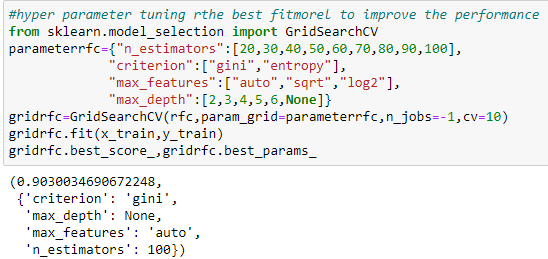
**F1 Score**- F1 Score is the Harmonic Mean between precision and recall. The range for F1 Score is [0, 1]. It tells you how precise your classifier is (how many instances it classifies correctly), as well as how robust it is (it does not miss a significant number of instances). High precision but lower recall, gives you an extremely accurate, but it then misses a large number of instances that are difficult to classify. The greater the F1 Score, the better is the performance of our model.

**Confusion Matrix** - In the field of [machine learning](https://en.wikipedia.org/wiki/Machine_learning) and specifically the problem of [statistical classification](https://en.wikipedia.org/wiki/Statistical_classification), a confusion matrix, also known as an error matrix, is a specific table layout that allows visualization of the performance of an algorithm, typically a [supervised learning](https://en.wikipedia.org/wiki/Supervised_learning) one (in [unsupervised learning](https://en.wikipedia.org/wiki/Unsupervised_learning) it is usually called a matching matrix). Each row of the [matrix](https://en.wikipedia.org/wiki/Matrix_(mathematics)) represents the instances in an actual class while each column represents the instances in a predicted class, or vice versa.

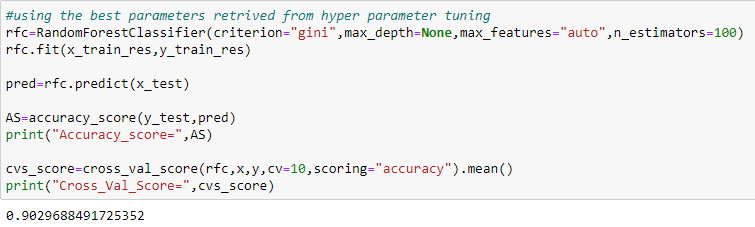
**Hyper Parameter Tuning**

In  a hyper parameter tuning the [parameter](https://en.wikipedia.org/wiki/Parameter)s are tuned to control the learning process of the model. Different datasets with specific algorithms requires different learning rate for the model to work more accurately. So it forms different combinations to get the one parameter with best result. We will be tuning the Ridge algorithm as it has the highest score.

Below is the code to perform hyper parameter tuning.



From the above output we can see that criterion: gini, max\_depth: none, max\_features: auto and n\_estimators: 100 are the best parameter that can be used for Random Forest Classifier for the best result. We will be using the above parameters for training the model. Below is the code to do the same.



* **Visualizations**

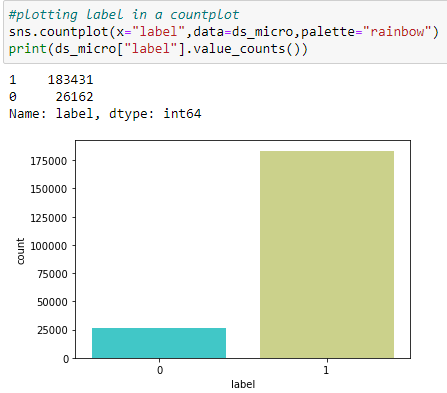
Two kinds of analysis are done in order to visualize the findings of this study, they are Univariate Analysis and Bivariate Analysis. Univariate analysis is to describe the data in order to find out the patterns in the data. This is done by looking at the mean, mode, median, standard deviation, dispersion, etc. Bivariate analysis is used to find out if there is a relationship between two sets of values. It usually involves the [variables](https://www.statisticshowto.com/probability-and-statistics/types-of-variables/)X and Y.

Below are the kinds of plots and graphs used to visualize the findings from the study.

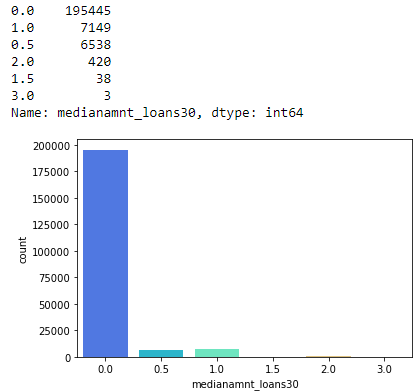
* Bar Plot
* Count Plot
* Scatter Plot

**Univariate Analysis**

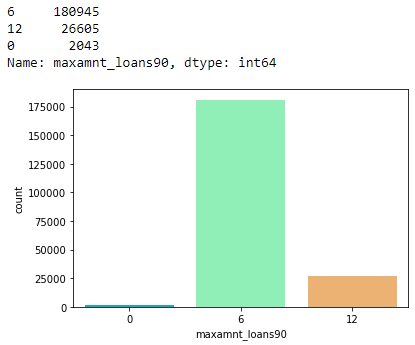
Below are some of the univariate data visualizations of the variables from the dataset.



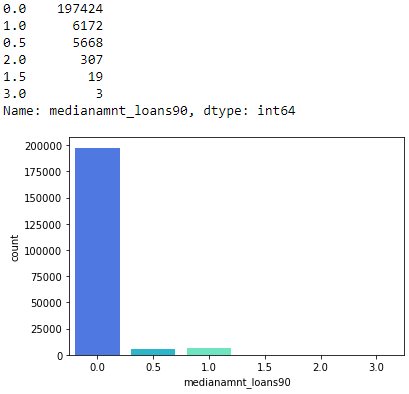
The above countplot shows instances of the target variable(labels). As we have an imbalanced dataset, we have way lesser 0s (loan defaulters) than 1s (non loan defaulters).



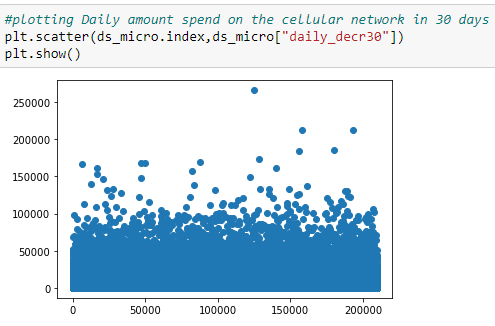
The above countplot shows Median of amounts of loan taken by the user in last 30 days. For most of the instances the median amount is 0 followed by 1 and 3 being the lowest.



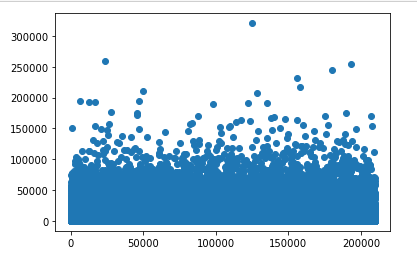
The above countplot shows maximum amount of loan taken by the user in last 90 days. For most of the instances the user have taken 6 loans followed by 12 and 0 being the lowest.



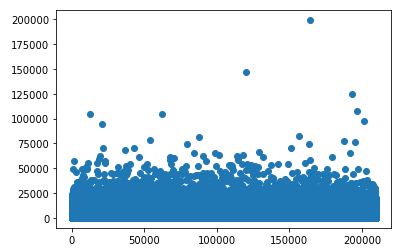
The above countplot shows Median of amounts of loan taken by the user in last 90 days. For most of the instances the median amount is 0 followed by 1.



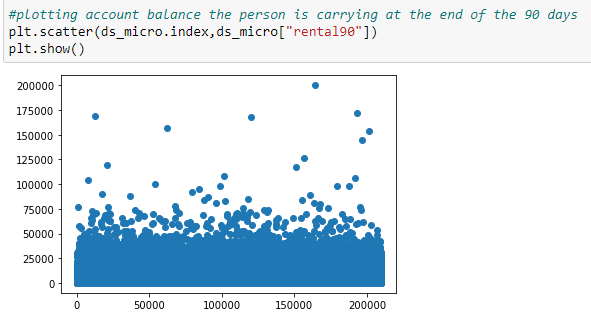
The above scatter plot shows daily amount spent on cellular network in last 30 days.



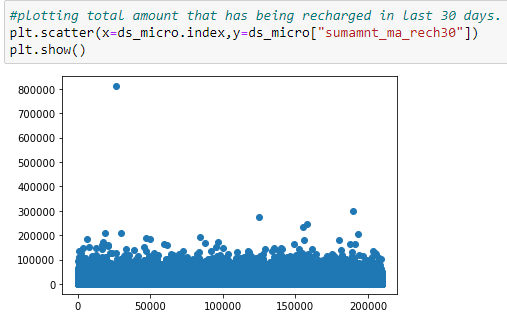
The above scatter plot shows daily amount spent on cellular network in last 90 days.



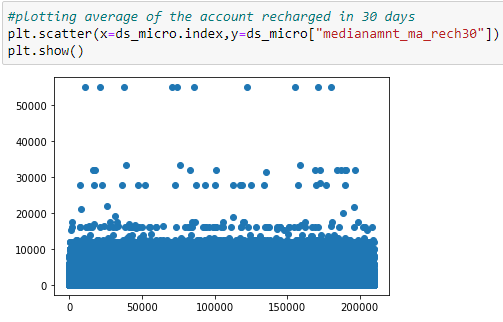
The above scatter plot shows what is the account balance the person is carrying at the end of the 30 days.



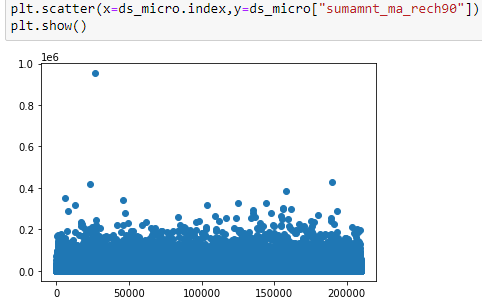
The above scatter plot shows what is the account balance the person is carrying at the end of the 90 days.



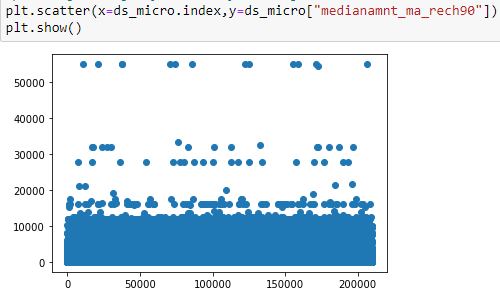
The above scatter plot shows total amount that has being recharged in last 30 days.



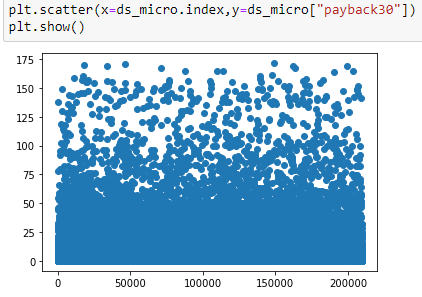
The above scatter plot shows average of the account recharged in 30 days.



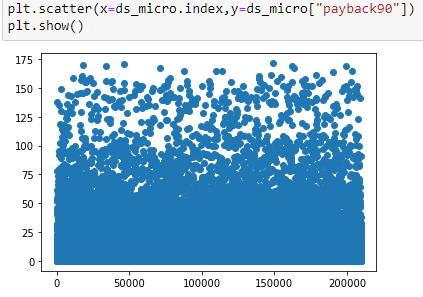
The above scatter plot shows Total amount of recharge in main account over last 90 days in Indonesian Rupiah.



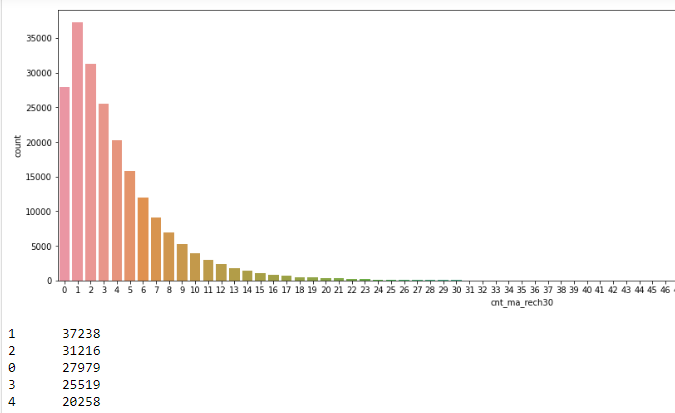
The above scatter plot shows average of amount of recharges done in main account over last 90 days at user level in Indonesian Rupiah.



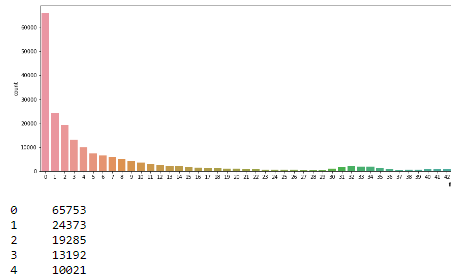
The above scatter plot shows time taken to pay back the loan in 30 days.



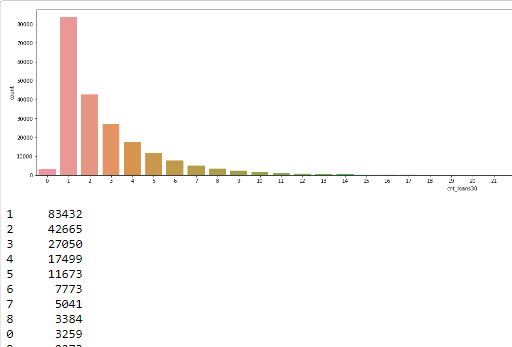
The above scatter plot shows time taken to pay back the loan in 90 days.



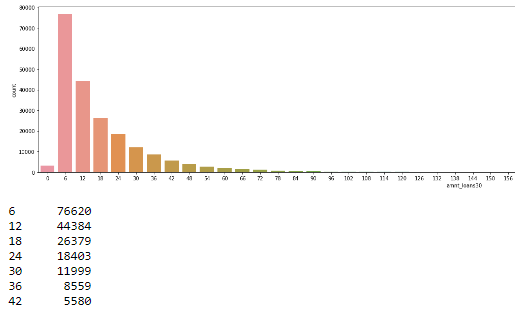
The above countplot shows frequency of the main account has being recharged in last 30 days.



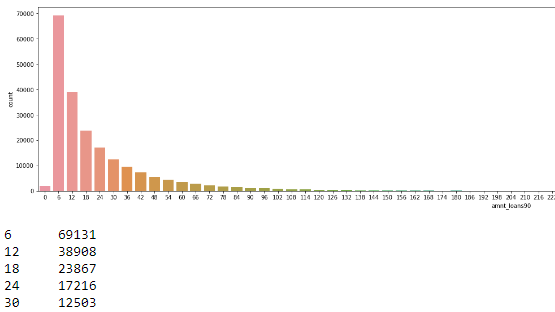
The above count plot shows number of times main account got recharged in last 90 days (total amount recharged /number of times recharged).



The above count plot shows number of loans taken by user in last 30 days.



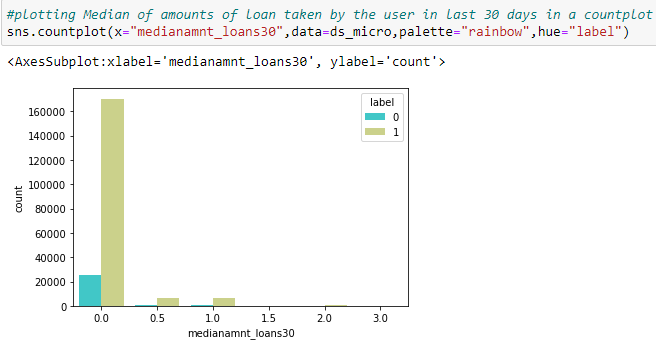
The above count plot shows total number of loans taken by user in last 30 days.



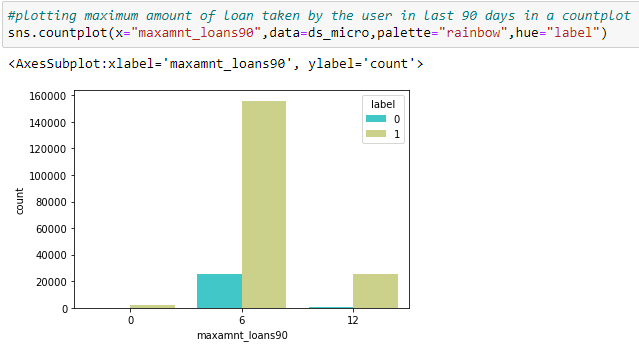
The above diagram shows total amount of loans taken by user in last 90 days.

**Bivariate Analysis**

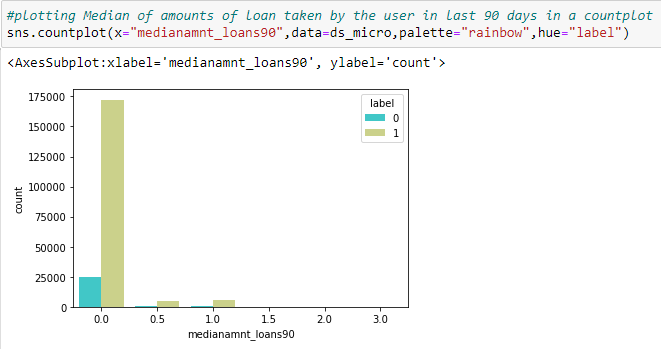
Below are some of the bivariate analysis. For most of the analysis causal relationships are shown with the target variable that is the label (0s,1s).



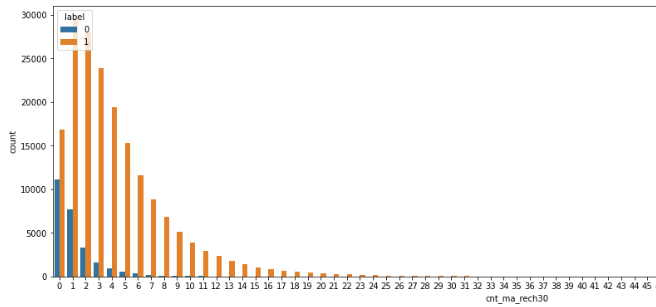
The above countplot shows median of amounts of loan taken by the user in last 30 days for both the loan payers (1s) and loan defaulters (0s).



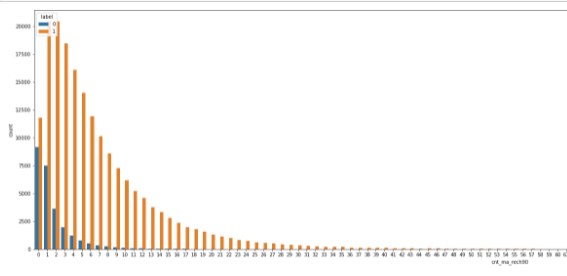
The above countplot shows maximum amount of loan taken by the user in last 90 day for both the loan payers (1s) and loan defaulters (0s).



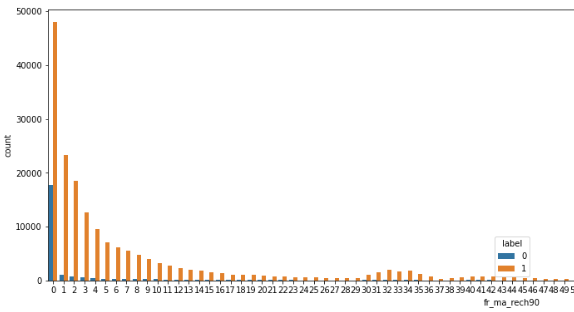
The above countplot shows median of amounts of loan taken by the user in last 90 days for both the loan payers (1s) and loan defaulters (0s).



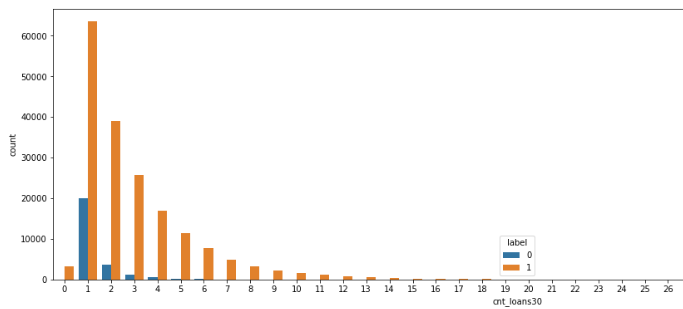
The above countplot shows frequency of the main account has being recharged in last 30 days for both the loan payers (1s) and loan defaulters (0s).



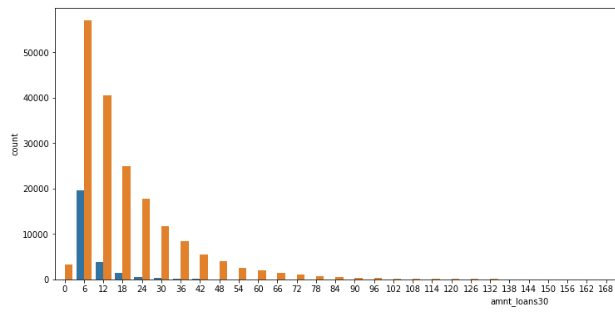
The above countplot shows number of times main account got recharged in last 90 days for both the loan payers (1s) and loan defaulters (0s).



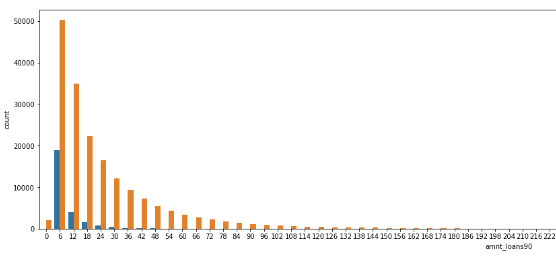
The above countplot shows Number of times main account got recharged in last 90 days(total amount recharged /number of times recharged for both the loan payers (1s) and loan defaulters (0s).



The above countplot shows number of loans taken by user in last 30 day for both the loan payers (1s) and loan defaulters (0s).



The above countplot shows number of loans taken by user in last 30 days for both the loan payers (1s) and loan defaulters (0s).



The above countplot shows number of loans taken by user in last 90 days for both the loan payers (1s) and loan defaulters (0s).

* **Interpretation of the Results**

**Interpretation from Visualizations**

The visualizations done with the variables from the dataset explains and gives insights about the problem statement. The univariate visualizations mostly talks about the dataset in particular. Most of the variables are plotted in countplot to get the counts based on the index. In the bivariate visualizations the target variable is plotted against the other variables. It gave a deeper understanding about the different labels that is 0s and 1s and their insights. We have an imbalanced dataset for which we have received a bias results and insights from the visualizations.

**Interpretation from Pre-processing**

The pre-processing started with finding the Null values. There were no missing values in the dataset, so we were good to go in that aspect. A correlation graph was made which showed the correlation co efficient of all the numeric variables with the target variable. Variables which showed lesser correlation with the target variable are then dropped from the dataset. Total of 14 columns that has shown lesser or negative correlation had been dropped from the dataset. There were 4 columns with negative values and digging deeper to the problem statement we understood that those columns should have absolute values. The 4 columns are later converted to absolute values using strategies. Then the variables were plotted in a box plot to find outliers and then z score is used to remove any data points that is beyond the threshold. The threshold here is 3. Variable which were skewed are fixed by using log transformation. The threshold for skewness is kept as .55. Then after separating the target and input variable, the input variables are scales using standard scaler.

**Interpretation from Modelling**

The modelling started with separating the training and testing data. As we have an imbalanced dataset as we used an over sampling technique called SMOTE to oversample the minority class. Then all the classification algorithm and metrics are imported. Finally the dataset is passed through all the algorithms to find out the one with the highest accuracy. The best model is selected and for the particular model the hyper parameters are tuned to find the best parameter.

**CONCLUSION**

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

The finding from the study is mostly generated from doing an extensive EDA. Using analysis like univariate analysis and bivariate analysis made a huge difference in terms of finding out the insights. Univariate analysis gave an insight about the general dataset and the way the primary research was conducted. It gave the count of different variables those were considered while collecting the data while bivariate analysis gave a relation of the variables with the problem statement and gave the variables those are more important in determining the loan payers and defaulters.

* **Learning Outcomes of the Study in respect of Data Science**

The power of visualization is something that came out pretty strongly while doing the project. Performing a univariate and bivariate analysis of the variables gave a better insight in understanding the problem statement. Also when it comes to the machine learning part, various algorithms works differently considering the dataset. In this case various algorithms has performed differently considering the different metrics and parameters. For instance various algorithms has shown efficiency considering different parameters. We have chosen the one with the highest level of efficiency considering the current business problem of the project.