

**Micro-Credit Defaulter**

Submitted by:

Akchay Pradhan

**ACKNOWLEDGMENT**

I would like to sincerely thank our SME Mr. Shubham Yadav for providing the Dataset and helping us to solve the problem by giving help as to when was needed.

**INTRODUCTION**

* Business Problem Framing

With the advancement of technology, there is no doubt the banking sector has advanced with its methodology of providing loans to the people.one such monumental sector is microfinance institutions (MFI). these institutions are providing micro finance services such as Group Loans, Agricultural Loans, Individual Business Loans and so on to the poor people and to the small entrepreneurs so that they don’t fall back into the poverty. This segment also focuses on small scale industry, animal husbandry and helping the farmers through loans needed to procure seeds and industrial equipment for farming. This institution has provided a methodology of providing mobile financial services 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

* Conceptual Background of the Domain Problem

The role of microfinance institutions in poverty reduction and wellbeing improvement has attracted the policymakers' attention in the developing countries across the globe. The underlying logic is that by providing microfinance services, poor will be able to participate in the economic market through forming their small businesses. Consequently, they will be able to generate income and improve their households, self-esteem and efficacy. Despite the popularity and apparent success of microfinance, there is no clear evidence that asserts the positive impact of microfinance scheme.

Review of Literature

Micro finance can be considered a new field in the banking sector and it’s a emerging field. Let’s look into the definition of what micro finance actually is- Asian Development Bank (ADB) defines microfinance as “the provision of a broad range of financial services such as deposits, loans, money transfers, and insurance to small enterprise and households.”

We need to understand its role in the banking world and how it has changed the banking system in the present world era. Poverty has been a topic of discussion both for the academia and government as poor have been completely or partially excluded from the formal financial system for decades. Lack of access of formal financial services has prompted the formal sector organizations, which includes government, non-government, and private, to meet the needs of the poor. Microfinance has been the term which refers to the formal and informal arrangements of providing financial services to the poor for the upliftment from poverty. It is microfinance which over the past decades has changed the perception of poor from non-bankable to bankable and recommending various methodologies to provide financial services. Also, Microfinance over the years has not only tried to alleviate poverty across the world but also shown glimpses of sustaining themselves from profit earned in the process. Hence as suggested by Brau and Woller (2004), they are “doing well by doing good” which has brought Microfinance to a global stage.

* Motivation for the Problem Undertaken

When I first got this project there wasn’t a lot I knew about what micro finance is all about.as I read articles about these new branch of banking sector it intrigued me further more.it is very important for the government and the policy makers to include the poorest of the poor people in the society and give them a chance to come out of that vicious circle of poverty.it is very important to understand that poor people may not have the knowledge of these kind of services let alone know about the mobile micro finances services. I believe the change has to come from the grassroot level and the efforts has to be made by the government to educate people about what all services are at their disposal and how they can make the best use of it and come out of the poverty loophole.

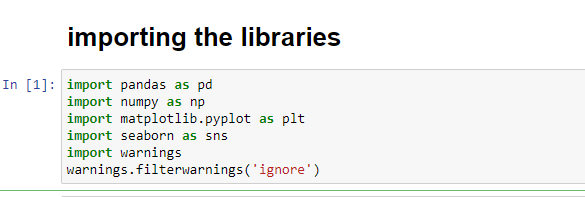
**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

In this model we are predicting micro credit defaulter weather the customer will be paying back the loaned amount within 5 days of insurance of loan. In this case, Label ‘1’ indicates that the loan has been paid i.e. Non- defaulter, while, Label ‘0’ indicates that the loan has not been paid i.e. defaulter. The data for micro credit defaulter is provided in the csv files where we will take these data and start our analysis in python and try to predict the best possible outcomes for the banking sector to provide loan or not.

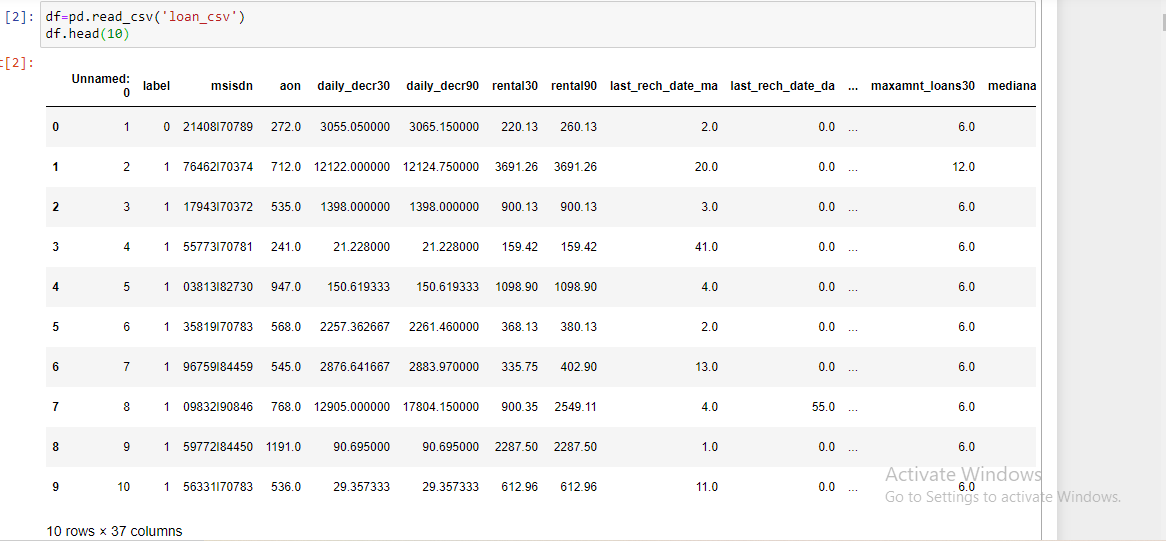
* Data Sources and their formats

The data for micro credit defaulter is provide in a csv files. We are going to load this csv files in python but first it is very important to import all the necessary libraries.



We have imported libraries such as pandas, NumPy and for visualization we have imported matplotlib and seaborn.

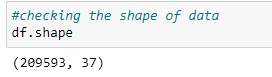
The next step is loading the csv files in python



After loading the data we can move ahead for data pre processing.

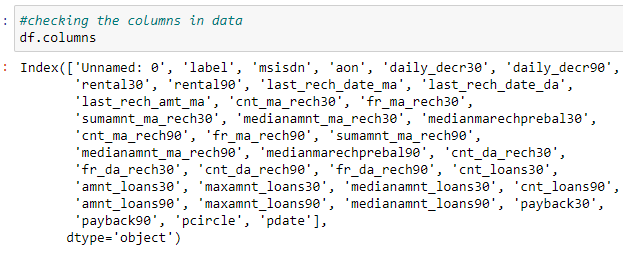
* Data Preprocessing Done

the first and the foremost step is to check the shape of the data to understand a what volume af data we are dealing with

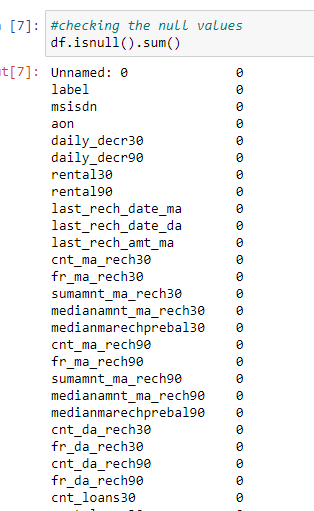
****

so there is 37 columns and 209593 rows in these data.

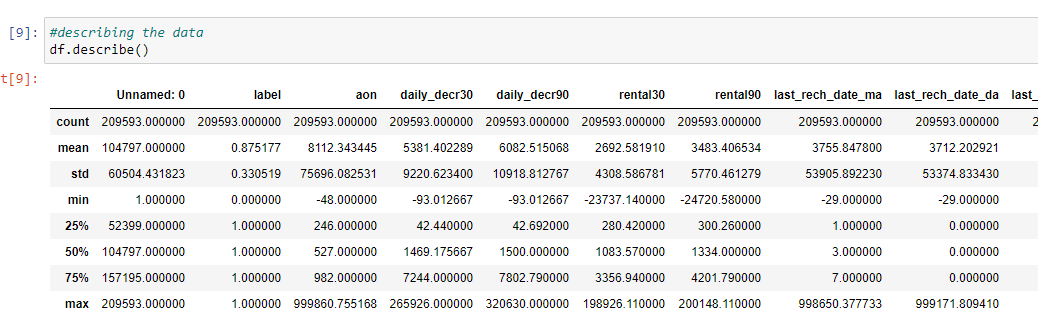
Checking all the columns and their names:



Checking null values in the dataset

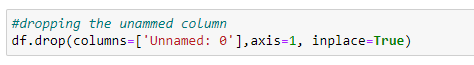


There is no null values in these dataset we won’t be treating it.

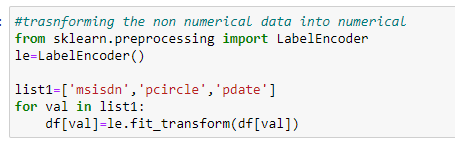
Statistical summary of the dataset 

We can see the mean, median, standard deviation 25th percentile, 50th percentile etc.

Drooping the columns



Changing all the object variable by using label encoder



* Data Inputs- Logic- Output Relationships

We can use correlations algorithm to check the relationship with the input and the output variables. Here in this data set our output variables will be labels which is 1 (non defaulter) and 0 (defaulter).we have used heatmap for the graphical representation of the correlations.

The darker colour represents higher correlations and the lighter colour represents lower correlations. our target variable is label which has strongest correlations with amnt\_loans90.,amont\_loans 30,cnt\_loans30,daily\_decr90.

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

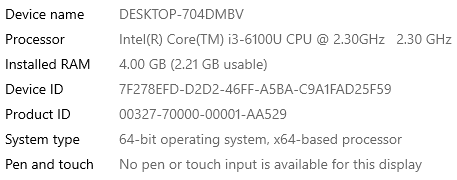
Here, in this micro credit defaulter data there is no NaN values and we are not required treat it. Although the data is huge and there are many columns there aren’t any columns which is heavily correlated with our output variable. Since the target variable has two outcomes 1 and 0 which is non defaulter and defaulter we are going to use matrix such as precision, recall,f1 score, classification report etc.

* Hardware and Software Requirements and Tools Used

Hardware

The first thing is to ensure we have the right hardware for data analytics. There isn’t much we can do if our hardware doesn’t have what we need. Since laptops are the most mainstream device for computing nowadays I have done the same.

The laptop I am using has the following hardware-



Software

There is no doubt that there is a growing demand for the data analytics in the market and have generated many openings worldwide.  It becomes slightly tough to shortlist the top data analytics tools as the open-source tools are more popular, user-friendly and performance oriented than the paid version. There are many open-source tools which doesn’t require much/any coding and manages to deliver better results than paid versions e.g. – R programming in data mining and Tableau public, Python in data visualization.

I have performed my data analysis in pyhton as it is object-oriented scripting language which is easy to read, write, maintain and is a free.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

In this micro credit defaulter data the problem statement is to find the defaulter weather the customer is going to pay the loan back or not. Since there is only two possible outcomes from this dataset we will be considering it as a binary logistics regression. Logistic Regression is used to solve the classification problems, so it’s called as Classification Algorithm that models the probability of output class.

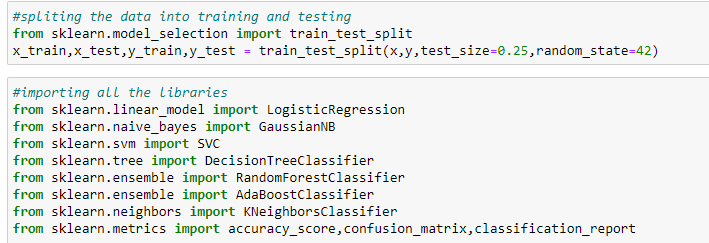
Consider a scenario where we need to classify whether a patient has diabetes or not. If we use linear regression for this problem, there is a need for setting up a threshold based on which classification can be done. Say if the actual class is malignant, predicted continuous value 0.3 and the threshold value is 0.6, the data point will be classified as not malignant which can lead to serious consequence in real time.

From this example, it can be inferred that linear regression is not suitable for classification problem. Linear regression is unbounded, and this brings logistic regression into picture. Their value strictly ranges from 0 to 1.

Here in this data set the output is either 1 or 0 so it is only best that we use logistic regression approach in this dataset to get the best possible outcomes.

* Testing of Identified Approaches (Algorithms)

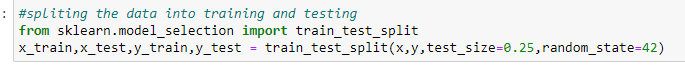
Since we are taking the logistic regression approach I have impetrated all the libraries needed for the problem solving



I have imported all the libraries from scikit learn.Scikit-learn is a free software machine learning library for the Python programming language.

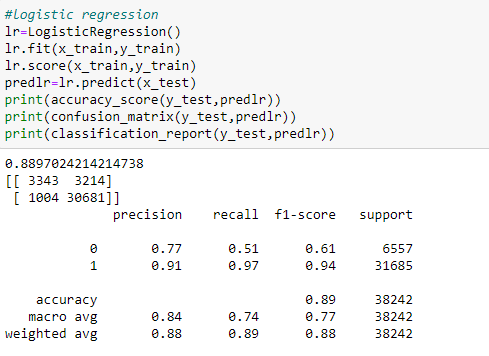
* Run and Evaluate selected models

First and foremost, I have split the data into training and testing and have also given the test size and random state where all these algorithms will be working.



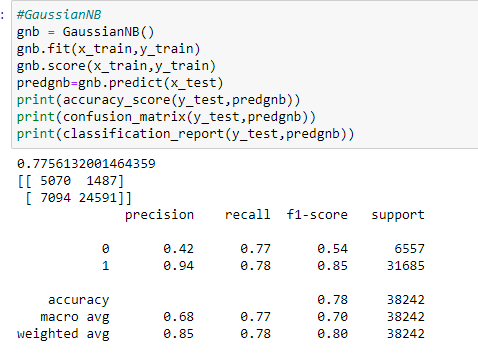
After splitting the data I have used various algorithms for model building

Logistic regression:

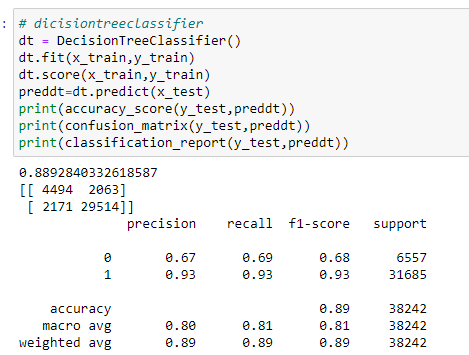


Its giving us the accuracy of 89%

GuassiaNB:

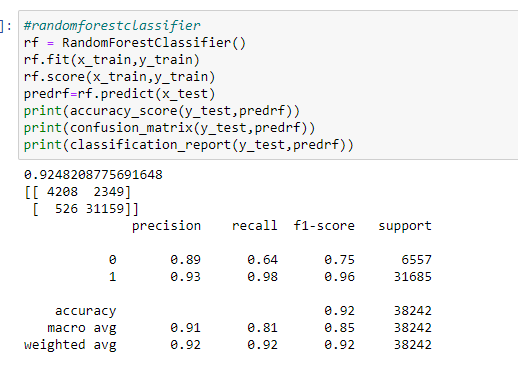


It giving us the accuracy of 77%

Decision tree:

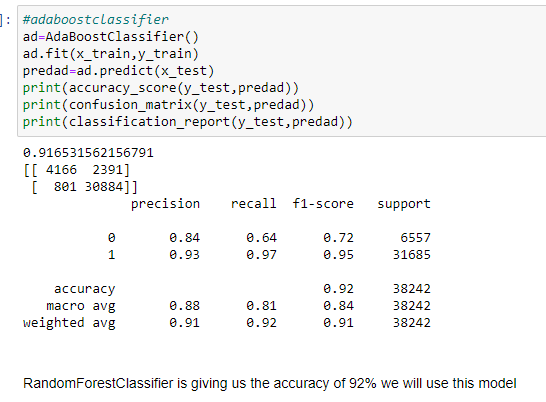
It’s giving the accuracy of 88%

Random Forest:



It giving the accuracy of 92%

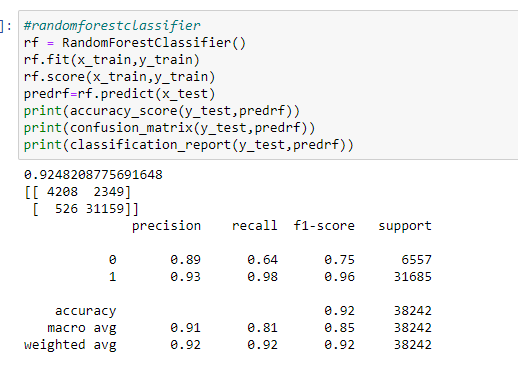
Ada boost:



It giving us the accuracy of 91 % so after using multiple algorithms it is evident that random forest is giving the result with highest accuracy so we will be using this model for predictions for credit defaulter.

* Key Metrics for success in solving problem under consideration

I have used metrics such as accuracy score, confusion matrix , classification report, auc-roc curve

I will be taking random forest as it is the highest performing model to explain all the metrics

Classification accuracy: Classification accuracy is the number of correct predictions made as a ratio of all predictions made. Our classification accuracy is 92%

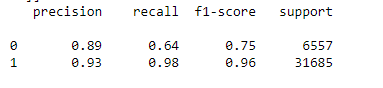


Confusion matrix: The confusion metrics is a handy presentation of the accuracy of a model with two or more classes.

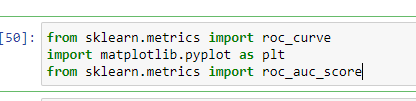


The table presents predictions on the x-axis and accuracy outcomes on the y-axis. The cells of the table are the number of predictions made by a machine learning algorithm.

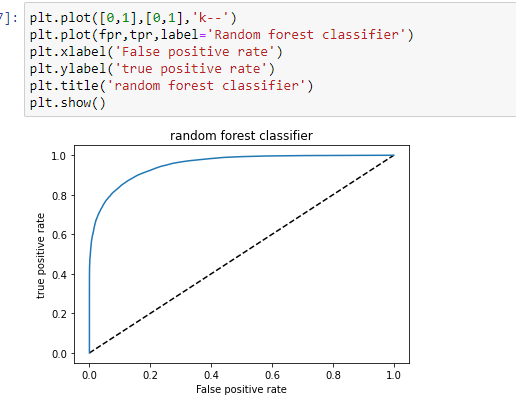
Classification report: Scikit-learn does provide a convenience report when working on classification problems to give you a quick idea of the accuracy of a model using a number of measures.



The classification report function displays the precision, recall, f1-score and support for each class.

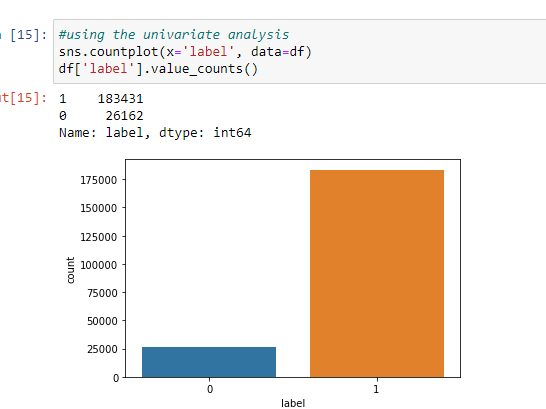
Auc-roc curve: Area Under ROC Curve (or ROC AUC for short) is a performance metric for binary classification problems. 

The AUC represents a model’s ability to discriminate between positive and negative classes. An area of 1.0 represents a model that made all predictions perfectly. An area of 0.5 represents a model as good as random.

A ROC Curve is a plot of the true positive rate and the false positive rate for a given set of probability predictions at different thresholds used to map the probabilities to class labels. The area under the curve is then the approximate integral under the ROC Curve.

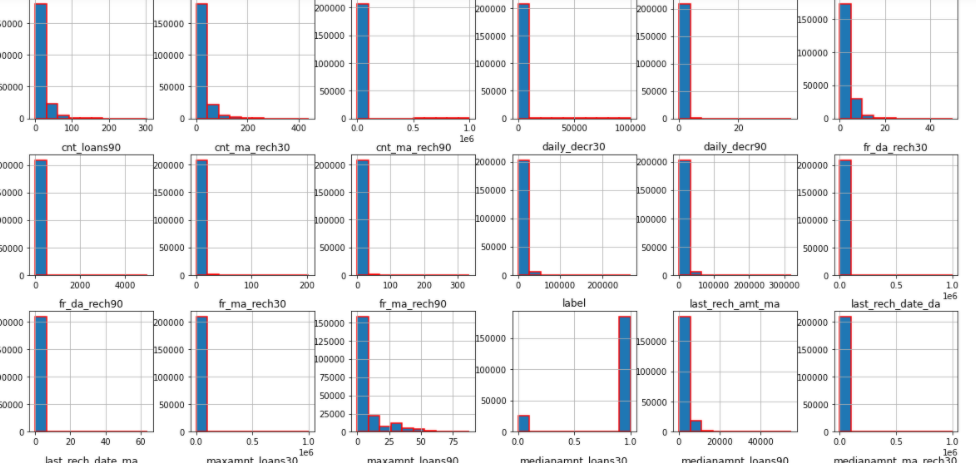
Visualizations

Univariate analysis:



We have used countplot and we can see the counts of target variable label 1 and 0.

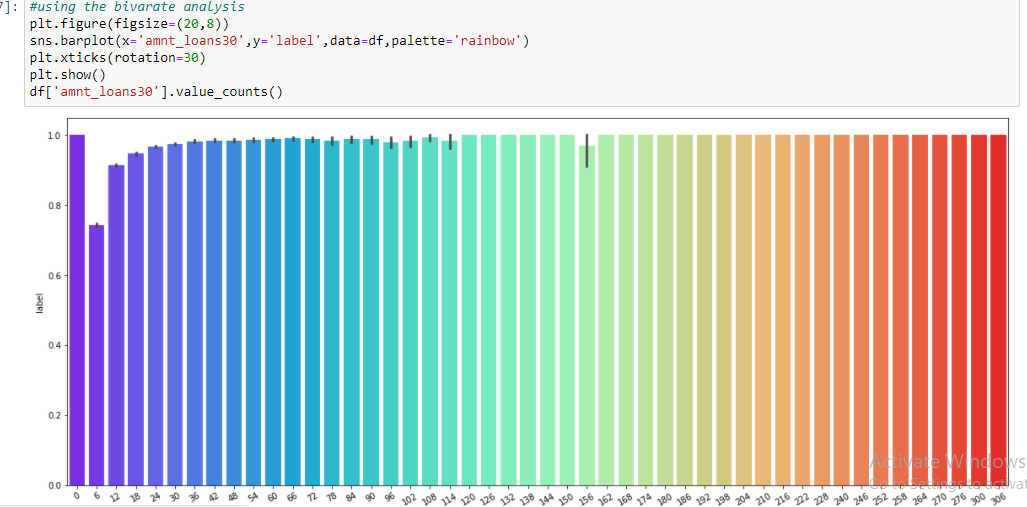
Using histplot:



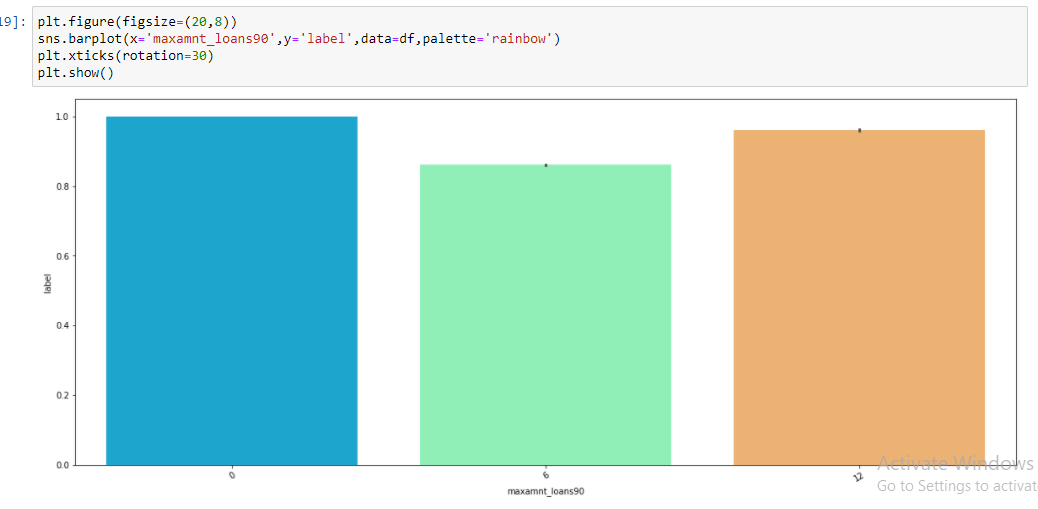
We have used histplot to show the distribution of multiple varaibles by counting the number of observations that falls within bins.

Bivariate analysis:

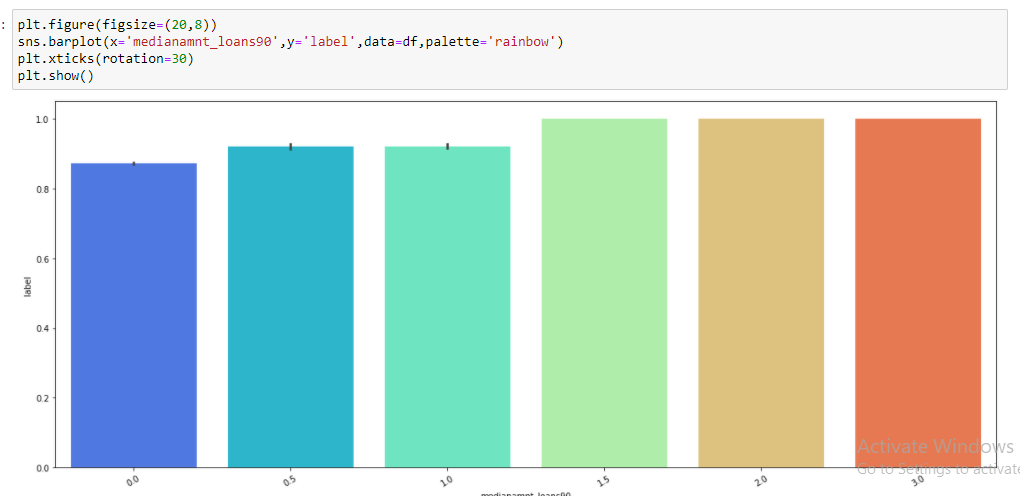
Using barplot: A bar plot represents an estimate of central tendency for a numeric variable with the height of each rectangle and provides some indication of the uncertainty around that estimate using error bars.



In the graph we can see the bar with 6 has lowest relations with label and 156 has highest outliers.

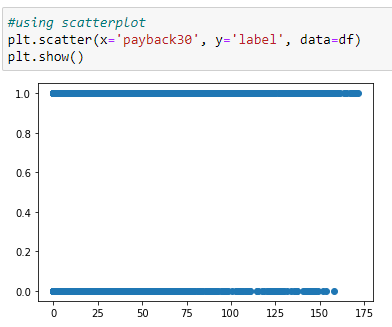


here the bar with 0 is the highest and 6 is the lowest.

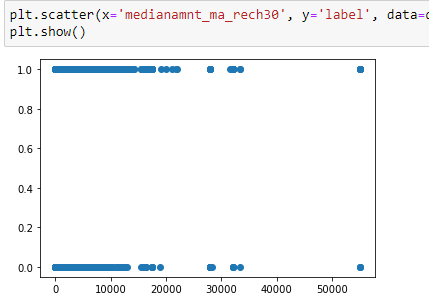


Here the bar with 3.0 is the highest and 0.0 is the lowest

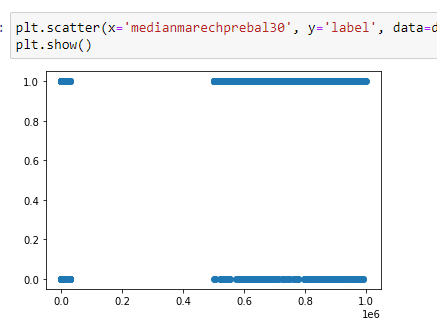
Using scatterplot: Scatter plots are the graphs that present the relationship between two variables in a data-set. It represents data points on a two-dimensional plane or on a Cartesian system. The independent variable or attribute is plotted on the X-axis, while the dependent variable is plotted on the Y-axis. These plots are often called scatter graphs or scatter diagrams.



Here the data x is scattered heavily from 0-152 and y from 0-175



Here the x axis is scattered heavily around 0-120000 and y axis around 0-200000



Here the x axis is scattered from 0.5-1.0 and y axis is scattered from 0.5-1.0

**CONCLUSION**

* Key Findings and Conclusions of the Study

We have explored the data, cleaned the data, visualized the data, and have built various models for predictions. There is no doubt the data was huge but luckily there no NaN values so we didn’t have to treat them. There were 37 columns in this dataset and dropping even one column would have resulted in huge data loss so I only dropped one column ‘unnamed’ which was completely not related and useless, however there were some columns which was object types so I have to use label encoder to convert them into numeric. Since the data was huge, I was expecting high correlations with target variable with at least few columns but it was not the case but any correlations is a good correlations so I had to proceed with what I had. The next step was visualization of the data since there were many columns I had to use for loops for multiple occasion. After exploring the data through visualizations, I found that all the columns have outliers and the data is skewed. I treated the outliers using IQR method since z score method was reducing the data to 0 and for removing skewness, I used power transform method. Next was splitting the data and model building. I used various machine learning algorithms such as logistic regression, Decision tree classifier, AdaBoost classifier etc and I found the best result was given by Random forest Classifier so I saved the model for prediction. I also used various metrics such as classification report, accuracy score, Auc-Roc curve to check how the model is performing.