MICRO CREDIT LOANS

Micro credit prediction model

Bhatia, Jasminder Singh

**Acknowledgments**

* Investopedia - <https://www.investopedia.com/terms/m/microcredit.asp>
* The balance - <https://www.thebalance.com/what-is-microfinance-and-how-does-it-work-4165939>

**What is Microcredit?**

* Microcredit is a common form of microfinance that involves an extremely small loan given to an individual to help them become self-employed or grow a small business. These borrowers tend to be low-income individuals, especially from less developed countries (LDCs). Microcredit is also known as "microlending" or "microloan.“
* The concept of microcredit was built on the idea that skilled people in underdeveloped countries, who live outside of traditional banking and monetary systems could enter an economy through the assistance of a small loan. The people to whom such microcredit is offered may live in barter systems where no actual currency is exchanged.
* Microfinance, pioneered by the Nobel-Prize winner Muhammad Yunus, helps the financially marginalized by providing them with the necessary capital to start a business and work toward financial independence.These loans are significant because they are given even though the borrower has no collateral. However, the interest rates for these microloans are often very high due to the risk of default.
* Microfinance is important because it provides resources and access to capital to the financially underserved, such as those who are unable to get checking accounts, lines of credit, or loans from traditional banks.
* Without microfinance, these groups may have to resort to using loans or payday advances with extremely high-interest rates or even borrow money from family and friends. Microfinance helps them invest in their businesses, and as a result, invest in themselves.

**Problem Statement**

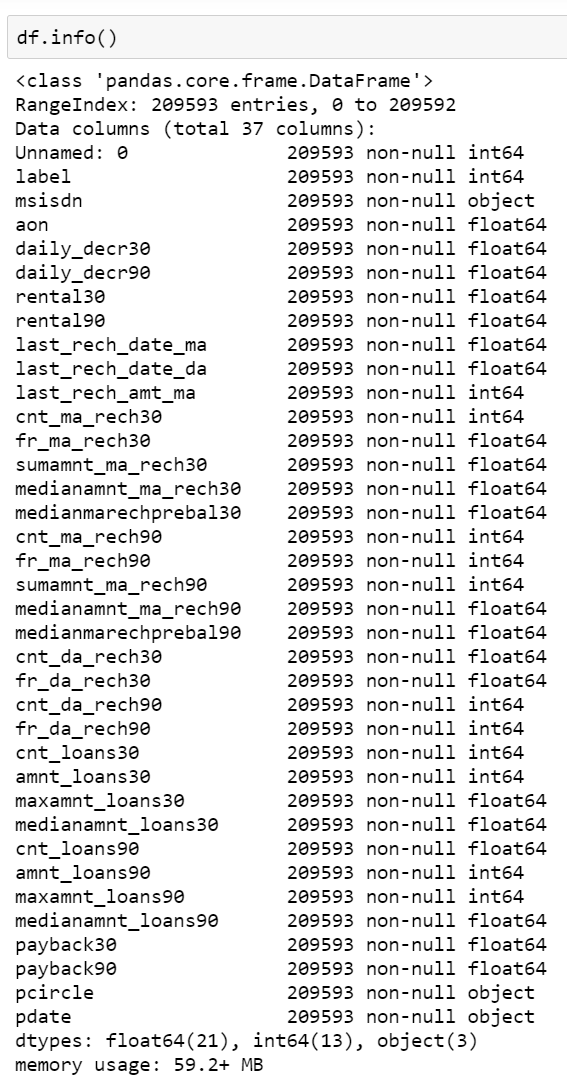
* Today, microfinance is widely accepted as a poverty-reduction tool, representing $70 billion in outstanding loans and a global outreach of 200 million clients.
* One of the most important criteria for microfinance institutions to provide credit to their clients is to be able to assess if the client would be able to payback or repay the credit amount in the agreed time.

This is the question which any lending institution would like an answer to before providing credit/loan to a client.

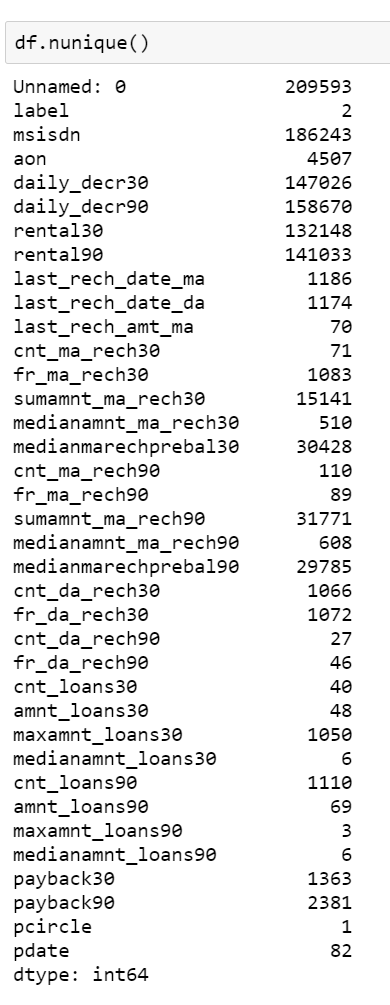
* We are working with one such client that is in Telecom Industry. They are a fixed wireless telecommunications network provider. They have launched various products and have developed its business and organization based on the budget operator model, offering better products at Lower Prices to all value 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).
* To help the Microfinance Institution in finding an answer to their problem of improving the selection of customers for the credit, we will create a predicting model which can be used to predict in terms of a probability for each loan transaction, whether 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 payed i.e., Non- defaulter, while Label ‘0’ indicates that the loan has not been payed i.e., defaulter.
* The sample data has been provided by the client.

**Analytical Problem Framing**

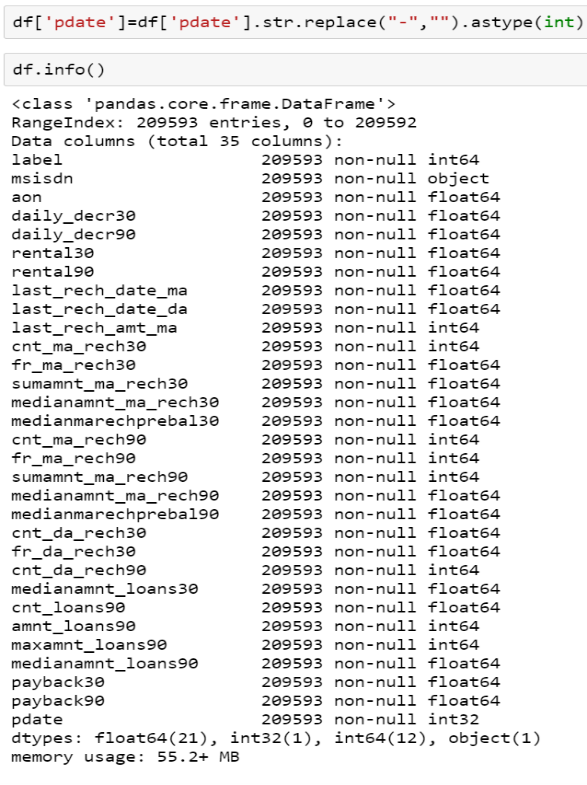
* The dataset comprises of 37 columns and 209593 rows.
* As we can see in the image below, there are no null value columns.
* The values in the columns are of following datatypes: float64 – 21 columns; int64 - 13 columns; object – 3 columns;

****

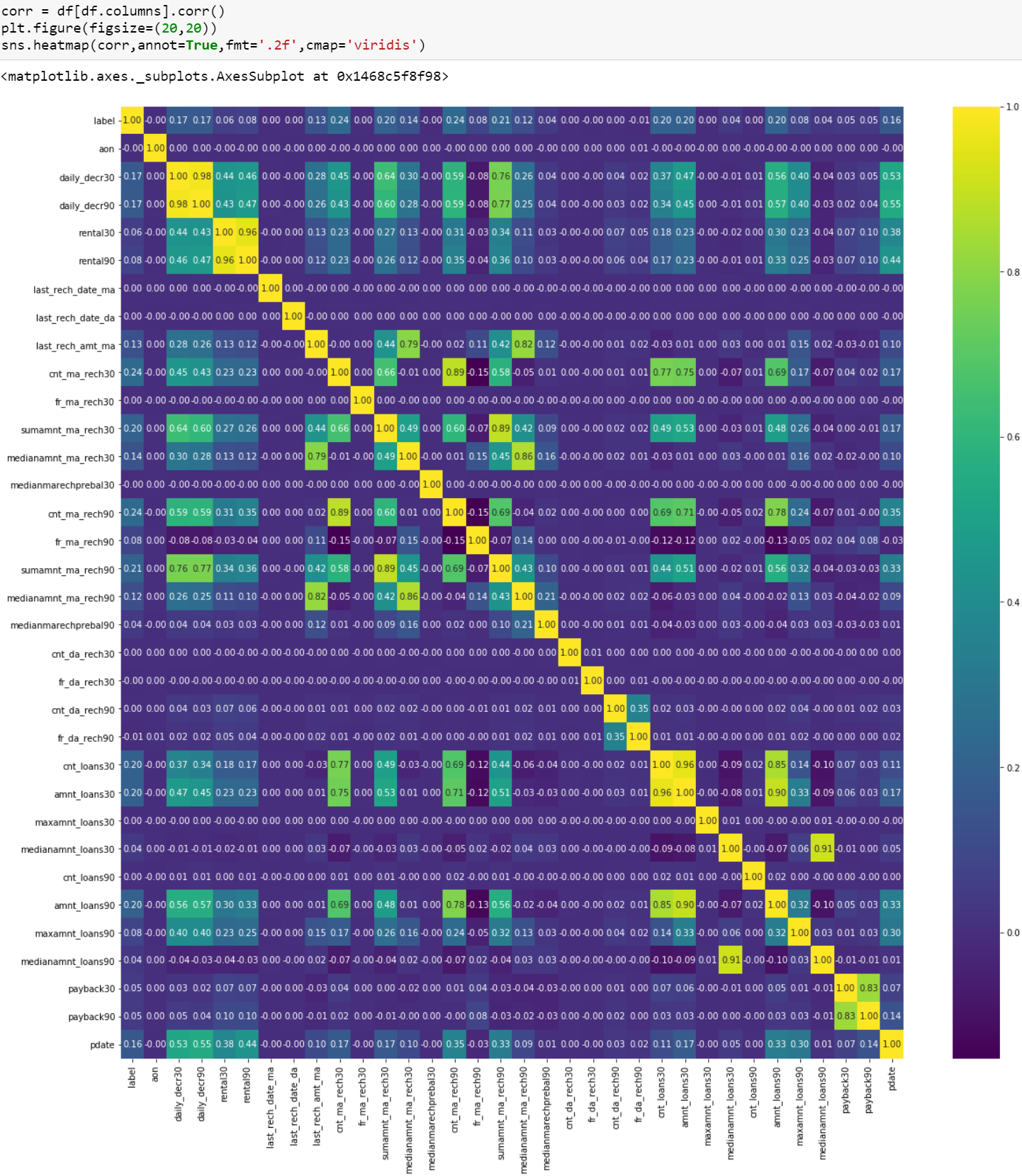
* When we look at the unique values in each column, it gives us information about the data.
* Based on the observation we can remove the Unnamed column as it is just serial numbers.
* We can also remove the pcircle column as there is only one value for all the rows, so it won't make a difference statistically or contribute to our model.



**Converting the pdate column to int datatype from object datatype for using it effectively in future stage for creating the predicting model**

****

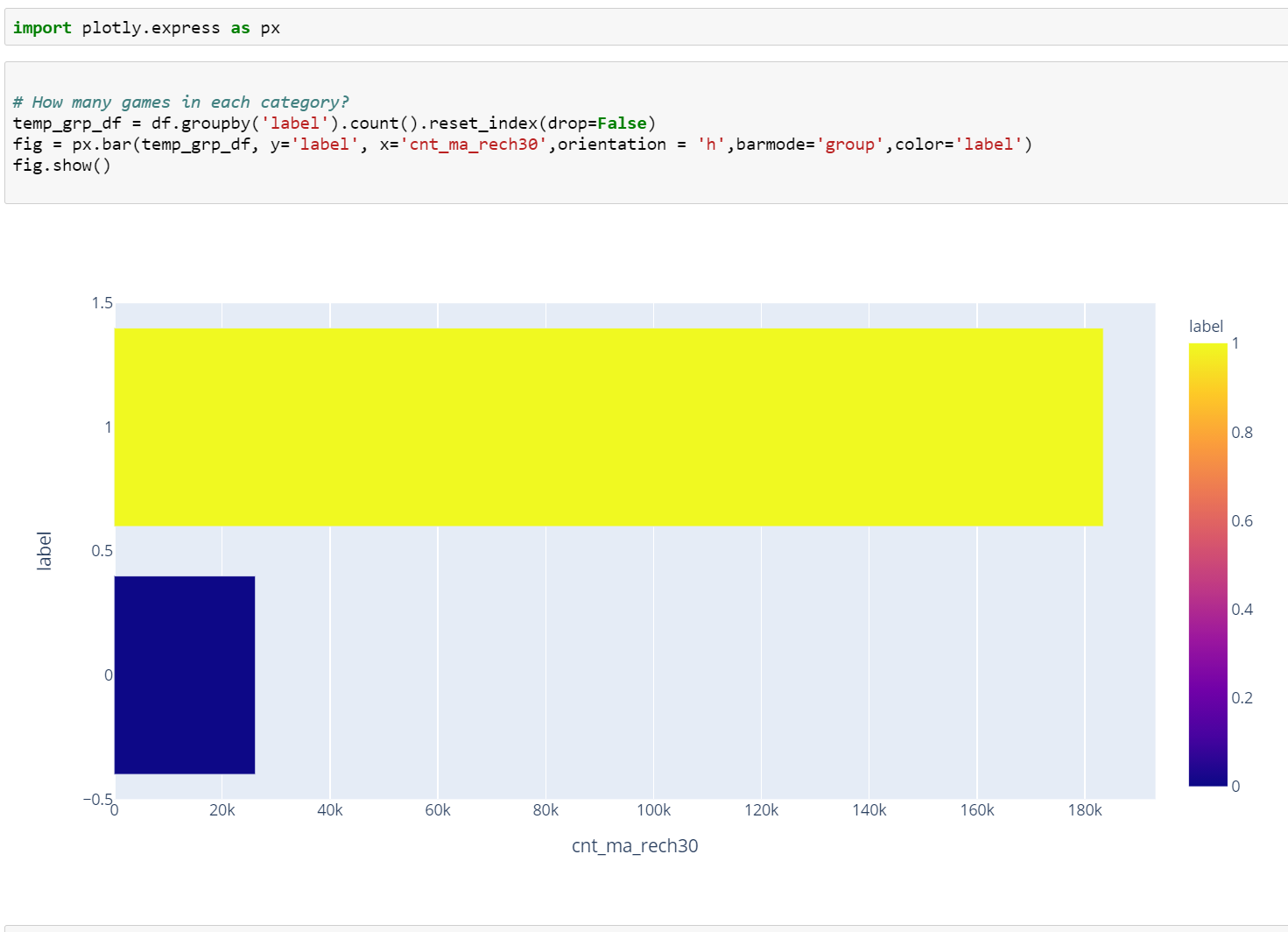
* Now we will look at the correlation between variables, mainly focusing on correlation of target variable ‘label’ with other columns to determine the relationship.
* Based on the relativity as we can see in the image, I have removed columns with no or low relativity with the target variable ‘label’ - cnt\_loans90, maxamnt\_loans30,cnt\_da\_rech90,fr\_da\_rech30,cnt\_da\_rech30,medianmarechprebal90,fr\_ma\_rech30,last\_rech\_date\_da,last\_rech\_date\_ma,aon,msisdn.



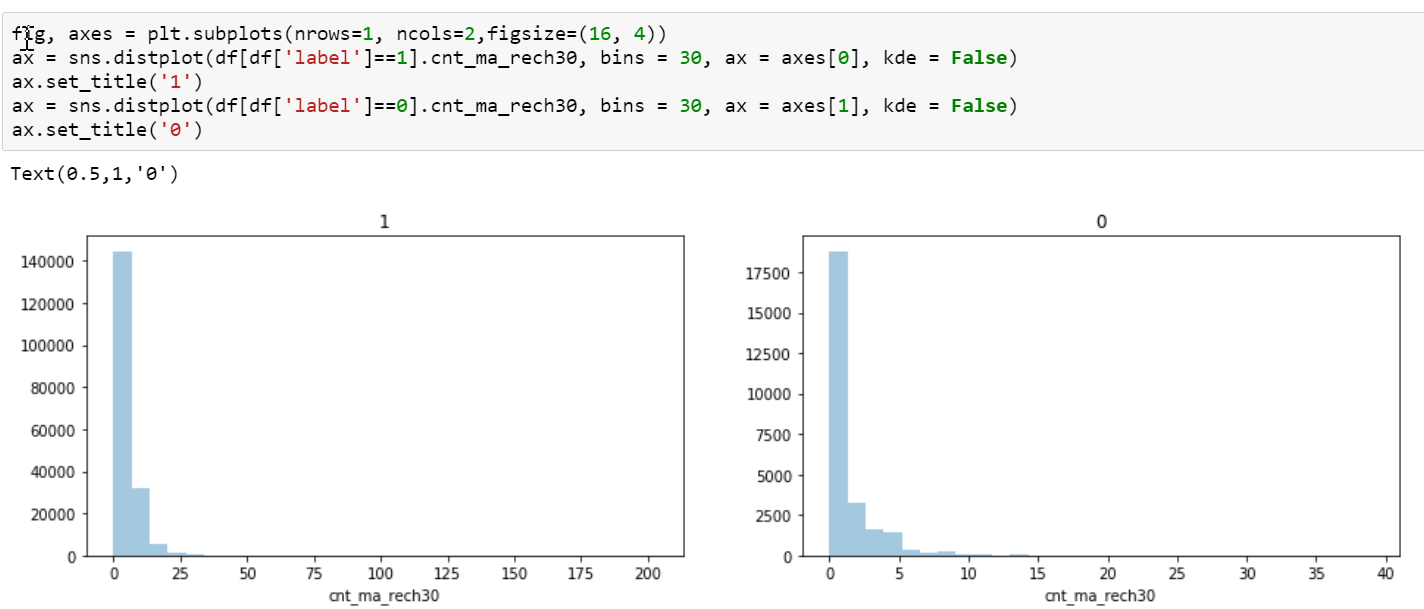
* Now we will look at the correlation between variables, mainly focusing on correlation of target variable ‘label’ with other columns to determine the relationship.
* Based on the relativity as we can see in the image, I have removed columns with no or low relativity with the target variable ‘label’ - cnt\_loans90, maxamnt\_loans30,cnt\_da\_rech90,fr\_da\_rech30,cnt\_da\_rech30,medianmarechprebal90,fr\_ma\_rech30,last\_rech\_date\_da,last\_rech\_date\_ma,aon,msisdn.

**Data Visualization**

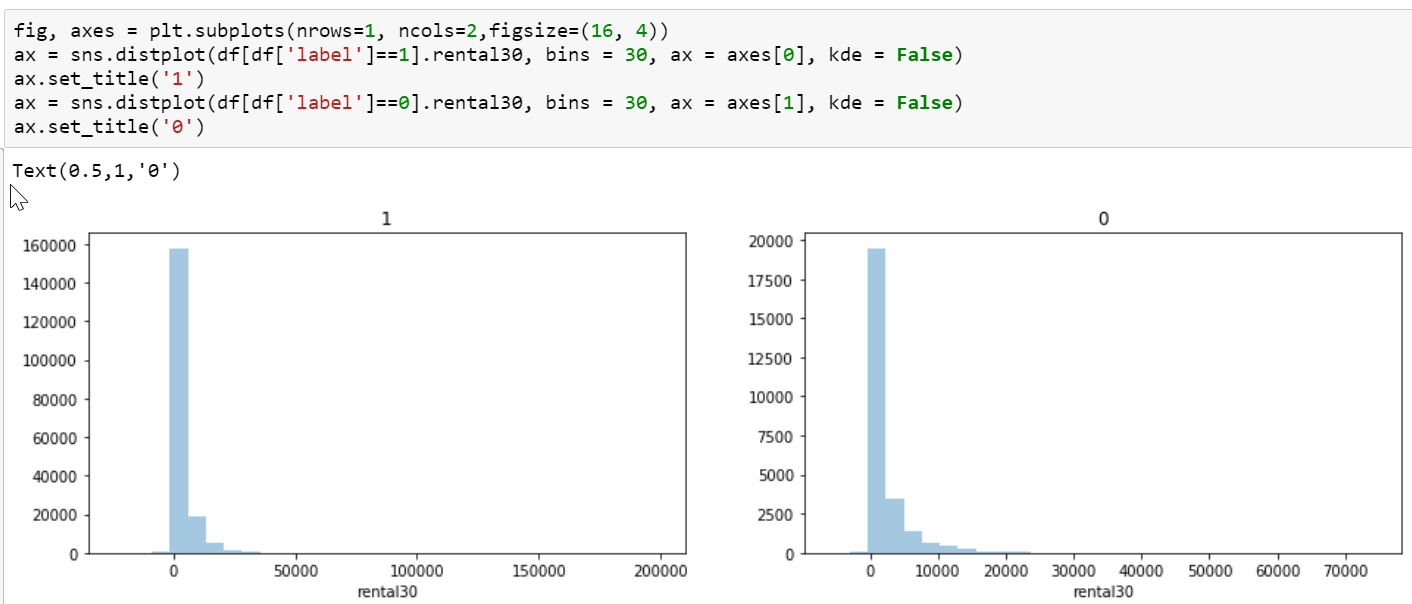
* Now we will investigate the dataset with the help of visualization to understand our target variable ‘label’ in a better way.
* This will help us understand the relationships better and provide an insight if we need to remove any irrelevant/outlier data.
* Let's start with checking the number of values for the two results in the column ‘label’ as seen in the image.



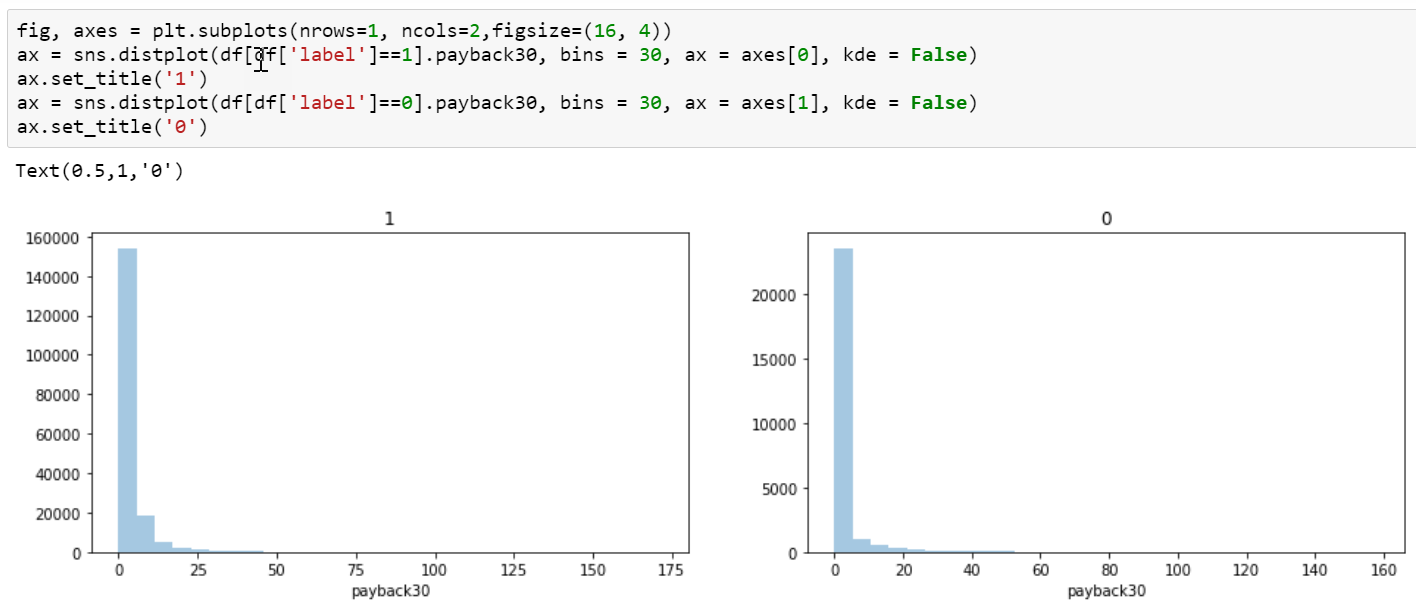
* **Plotting Cnt\_ma\_rech30 over label**

****

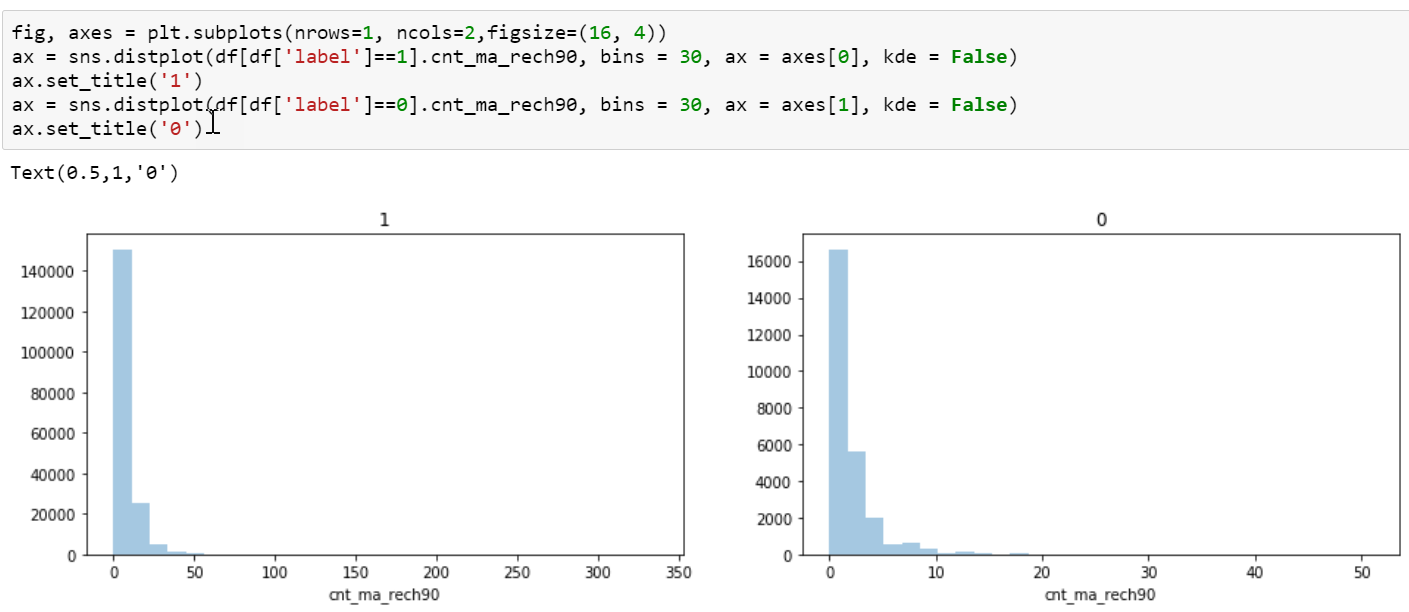
* **Plotting rental30 over label**

****

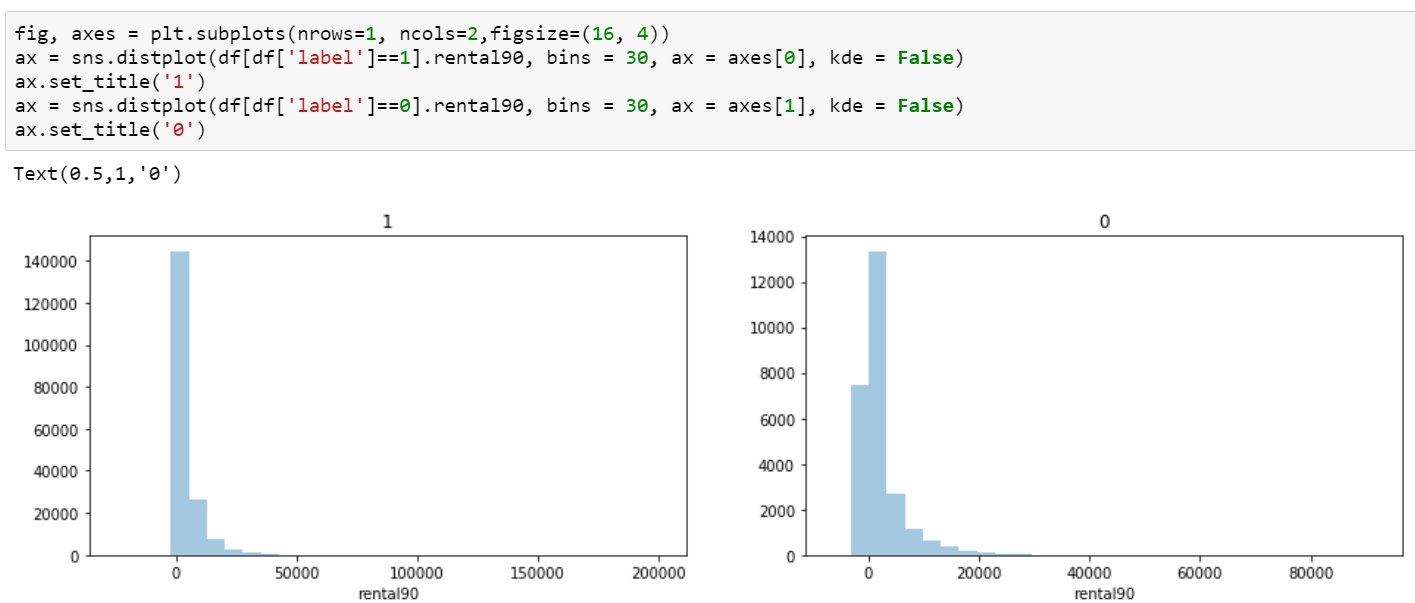
* **Plotting payback30 over label**

****

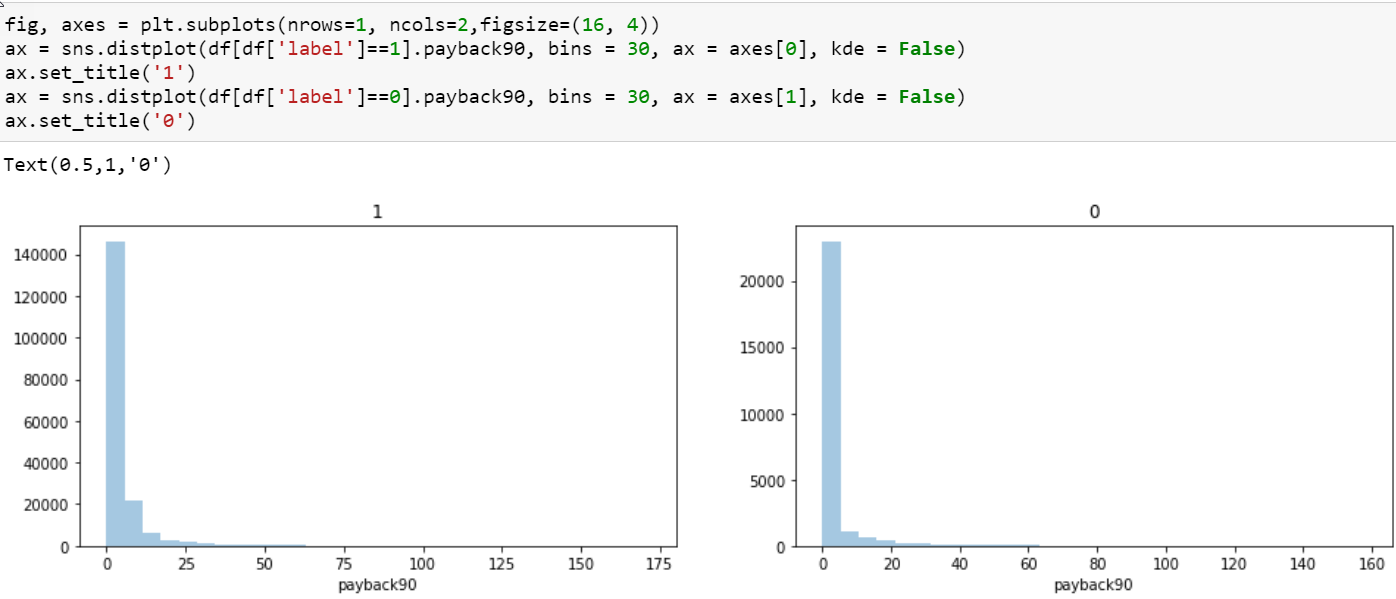
* **Plotting cnt\_ma\_rech90 over label**

****

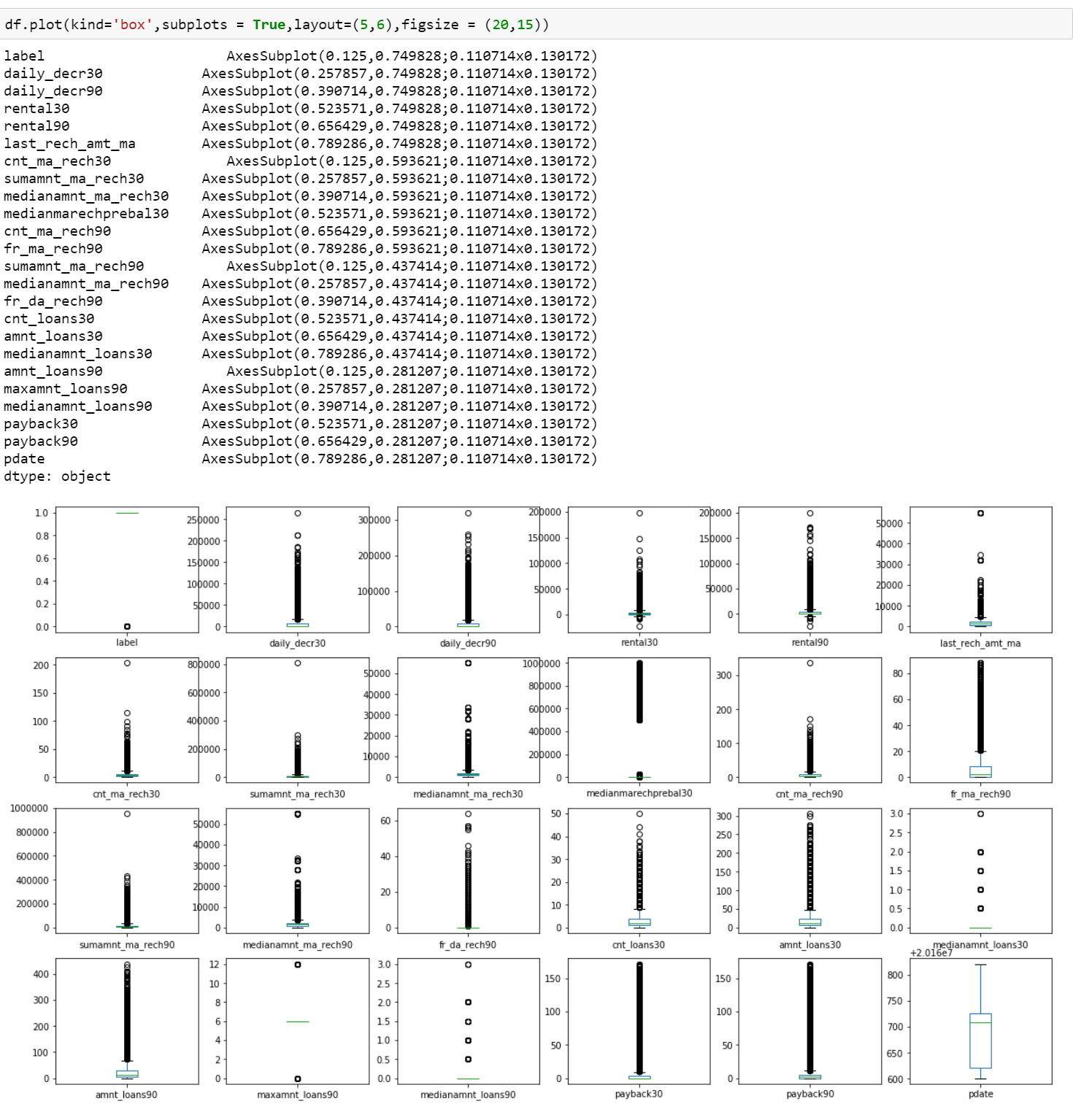
* **Plotting rental90 over label**

****

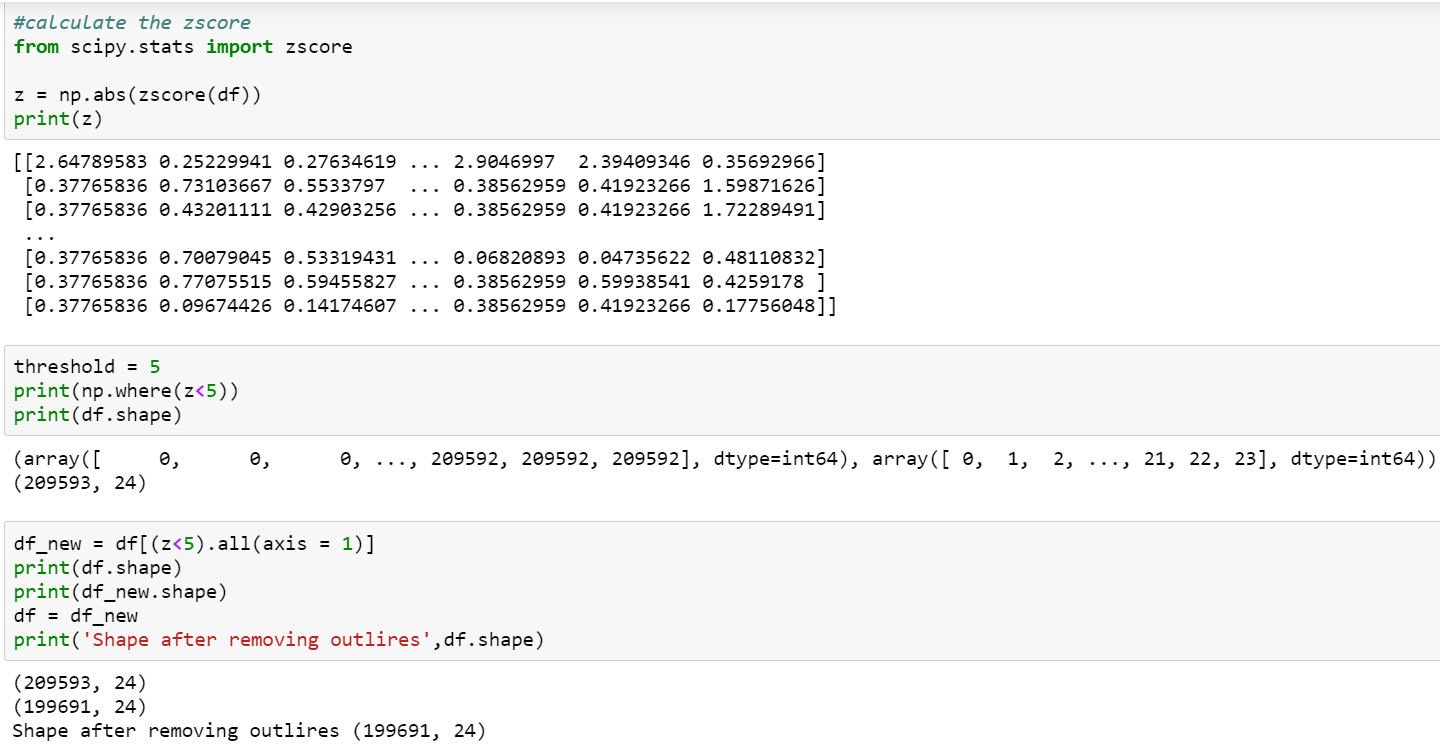
* **Plotting payback90 over label**

****

* We can see from the plots that majority of the values are under same bin when we compare it with different variables.
* Another point that we can see is that the values are a little more spread in 90-day variables.
* The plotting of data helps us to see the spread of the values for each variable per category of our target variable, ’label’, with values 0 and 1.
* Now let's plot all the variables using univariate analysis to identify if there are any outliers in the data which can be eliminated from our dataset.
* **Plotting all the columns using boxplots to understand the spread and identify outliers if any.**

****

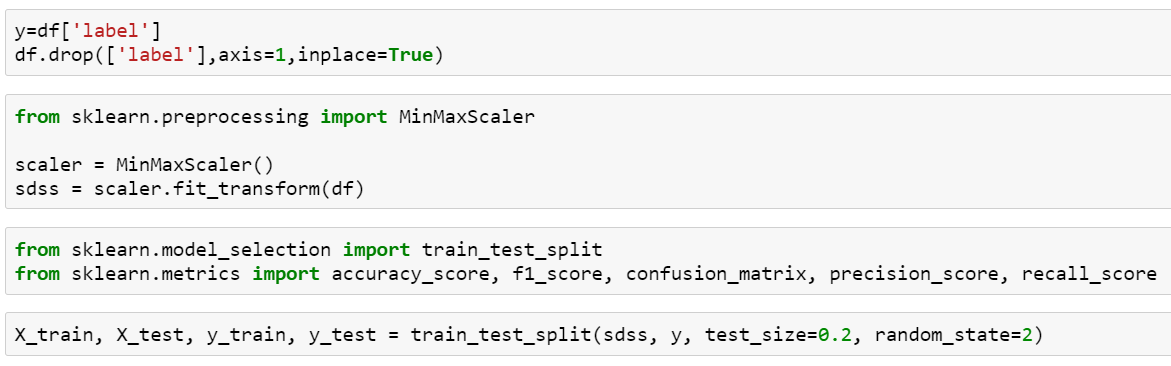
* As it can be observed in the plots, there are a few outliers for certain columns. We will work on evaluating the outliers and removing them from our dataset.
* In the first step we will calculate the Z-scores. I have set a threshold of 5, so we will eliminate the values with Z>5. Choosing the threshold as 5 so that we do not lose a lot of data but still eliminate the extreme outliers.



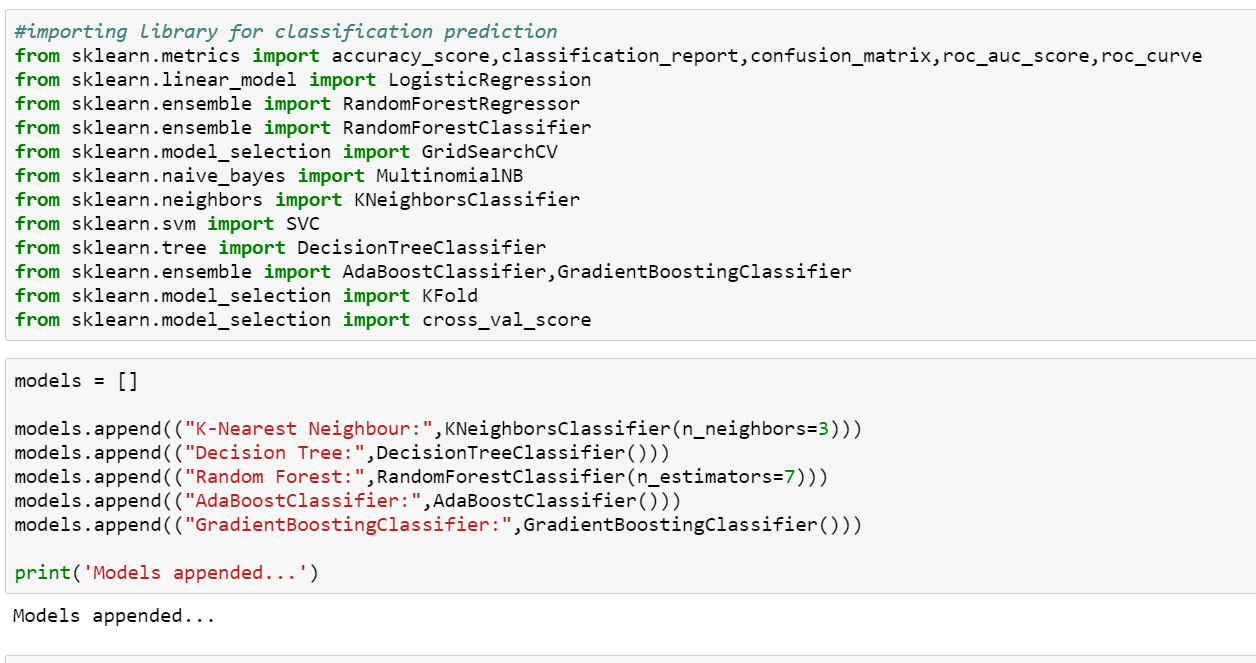
* Initially we had 209593 rows and post the removal of outliers, we now have 199691 rows in our dataset.

**Model Development and Evaluation**

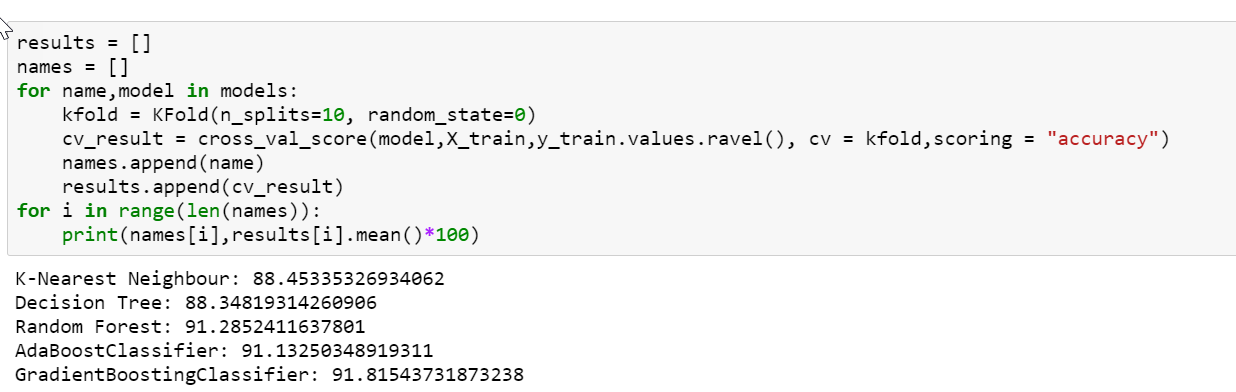
* The dataset is now ready to be trained. We will use MinMaxScaler to scale and transform the data.
* The dataset will now be split into train and test data so that once the training of the data is done, we can evaluate the model created on train data with test data.
* I have chosen 5 algorithms for training our dataset which are best for Classification prediction: K-Nearest Neighbor, Ada Boost Classifier, Random Forest Classifier, Decision Tree, Classifier and Gradient Boosting Classifier.



* We will now import several libraries for using the algorithms listed earlier and creating our model as seen in the image on right.
* In order to optimize the code, I have created a function to run all the 5 models at once and provide the accuracy score for each model so that it is easy for us to pick the model with best accuracy. This will save us the runtime and provide us accuracy scores for all the models trained at once.



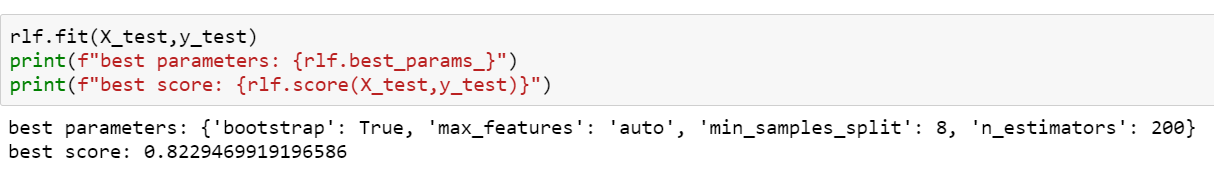
* I used K-Fold Cross validation procedure to train and create our models on the train data.
* The models have been created and the accuracy score for each model is calculated as seen in the image.
* We have an accuracy score of above 91% for 3 models, Random Forest, Ada Boost Classifier and Gradient Boosting Classifier which is pretty good.



* In order to avoid Overfitting and Underfitting of the models with our dataset, we will optimize our model using Grid Search CV. We will also find out the best parameters for the same.
* I have opted to go with Random Forest Model for applying Grid Search CV and optimizing our model.
* Best Score on applying Grid Search CV for our model is ~ 79%. We have also got best parameters to be used in achieving this score for the model.



* Now let's test our model with the test data we have created earlier from our original dataset.
* This will help us evaluate our model and give us a better picture of the accuracy of the model been created for prediction.
* ~82% accuracy, which is a pretty good score.



**Conclusion**

* We were able successfully create a predicting model to predict whether a customer would be paying back the credit amount taken in due time with an accuracy of ~80%.
* This will help the client in evaluating and understanding their customers better while also helping them in making right business decisions of choosing a customer for offering credit.
* A large dataset helped in understanding the data in a much better way to evaluate the features and find relationships between them. Using Machine Learning, we were able to develop a predicting model which would help the client in shaping up their business.
* While there are several algorithms that can be used for creating a prediction model, choosing the right model and understanding the problem at hand is the key.
* In this project, we were able to identify the type of problem, Classification, and select the algorithms best suited for solving a classification problem.
* Data cleaning also played a very important role in optimizing the original dataset as we removed certain irrelevant columns from our dataset as they have would not contribute in any way for our model creation but their existence in the final dataset would make the model slow and may even impact accuracy at time.
* With the help of Visualizations, we were able to identify there are outliers in the dataset and removed the outlier values which helped our algorithm to work in a much better and efficient way.