**Micro Credit Loan Project**

**Submitted by: Shobha**



**Introduction**

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). 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.

**Problem Statement:**

This is a classic Business problem which helps Micro Financing Institutions and other Lending companies reduce Credit risks by recognizing potential Defaulters.

**Below are the steps that we will perform to analyse the case study to draw an outcome.**

1. **Data Pre-processing**: Pre-Process the data to suit them with the analysis method.The Pre-Processing may involve cleaning up the data, transforming the data, or creating new variables that may bring useful information for further analysis.
2. **Exploratory Data Analysis (EDA):**

This step creates textual and visual summaries of the dataset that highlight some characteristics of the data.

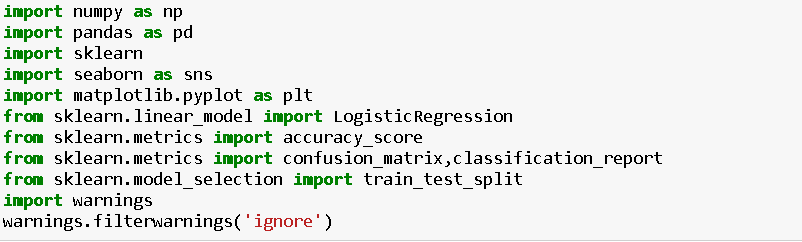
1. **Feature Engineering**
2. **Model Selection and Training**,

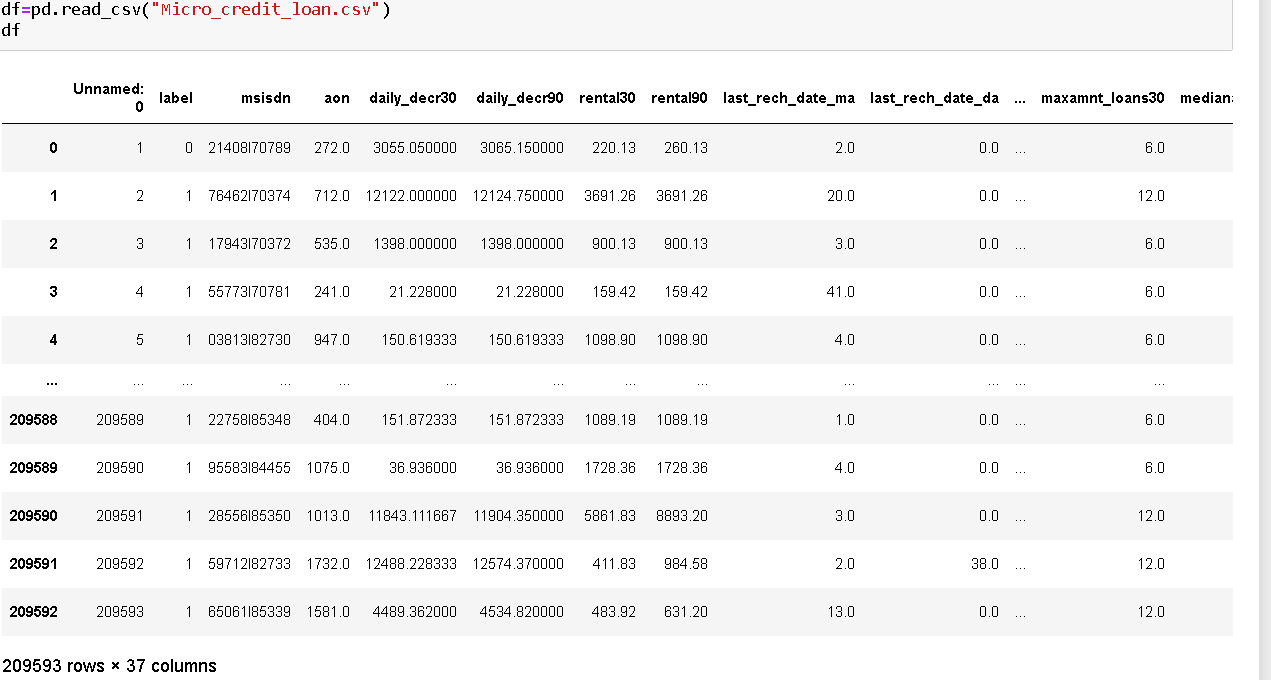
Test and Evaluate the Model: Evaluate the performance of the proposed model

**Data Pre-Processing**

Getting the system ready and loading the data

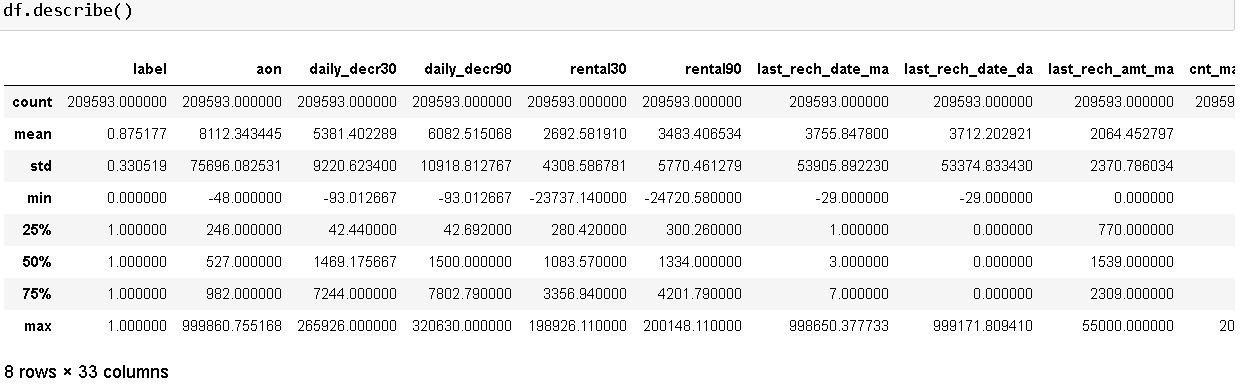
We will be using Python for this course along with the below-listed libraries.



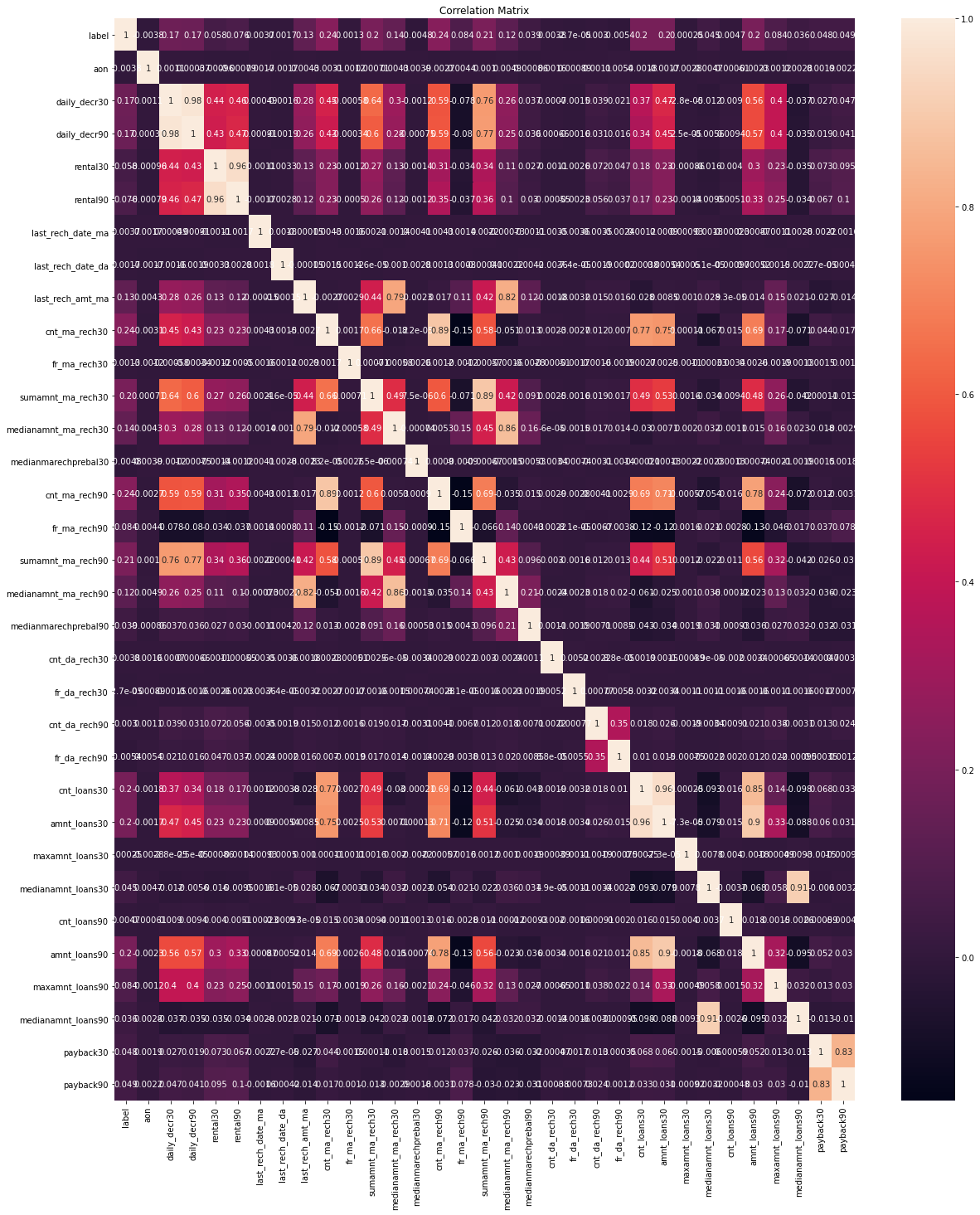


I am dropping columns Unnamed, msisdn, pcircle and pdate. columns unnamed is repetition of serial number. msisdn is customer mobile number and pdate is date both of the columns will not contribute any value while predicting the model accuracy. pcircle have same value in each row so I decided to remove it.

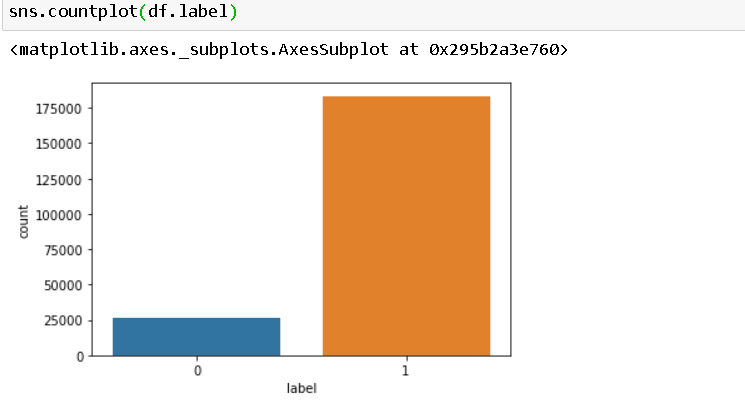
**Exploratory Data Analysis (EDA)**



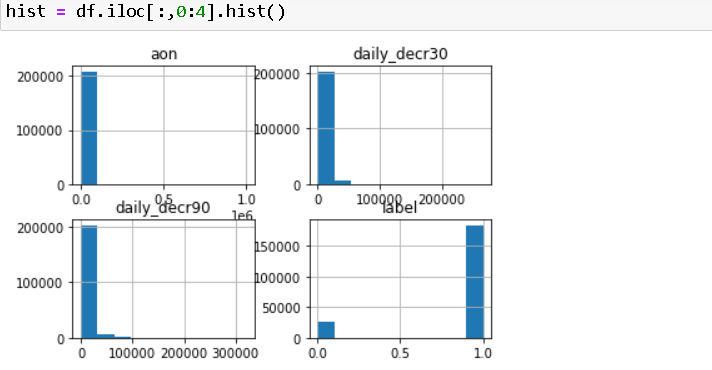
As per above statistics description there is no null value in given dataset. Most of the columns have outliers. Standard deviation is on higher side in some of the columns. In some of the columns there is difference in mean and 50% and mean and standard deviation so the data is again skewed in most of the columns.

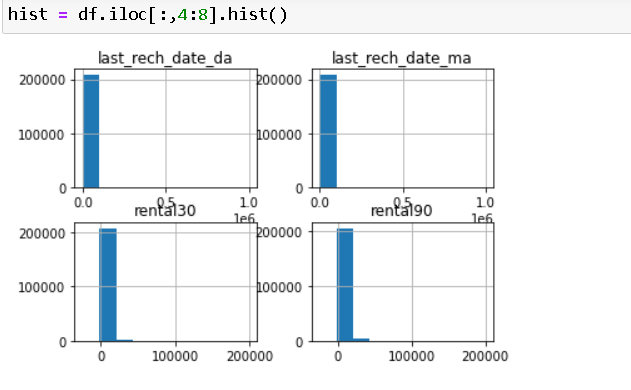


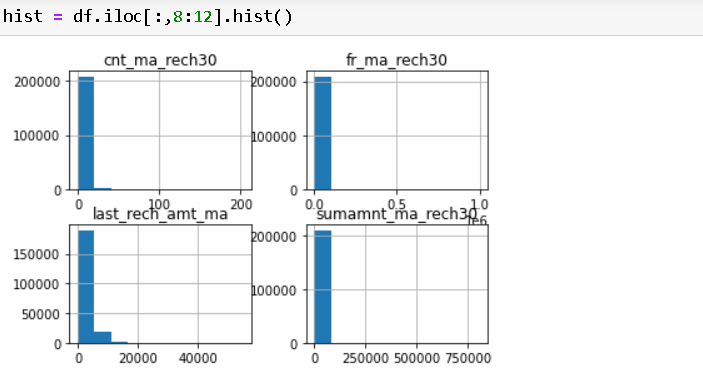
The graphs shows that the independent variables are very moderately correlated with target variable.

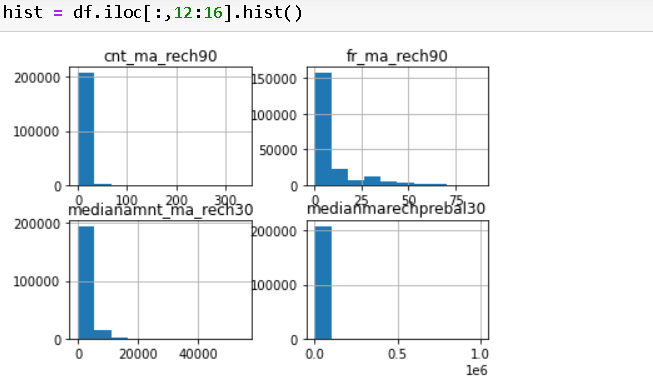


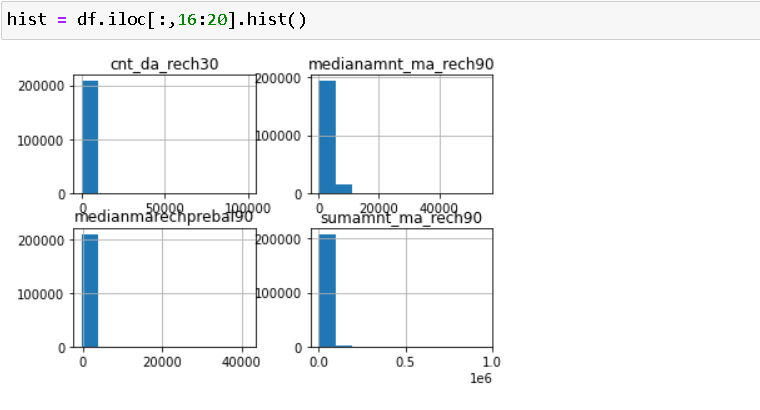
The dataset is imbalanced. Label ‘1’ has approximately 87.5% records, while, label ‘0’ has approximately 12.5% records.

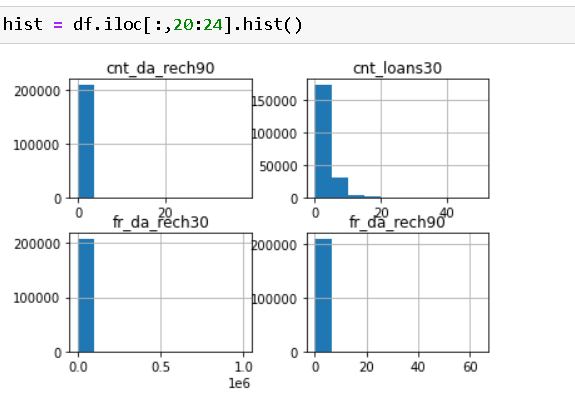


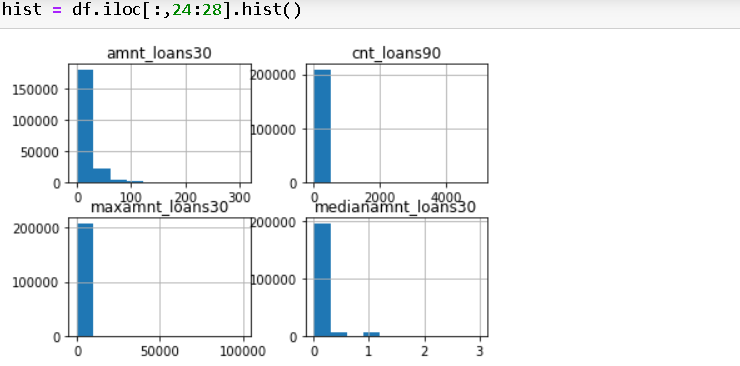


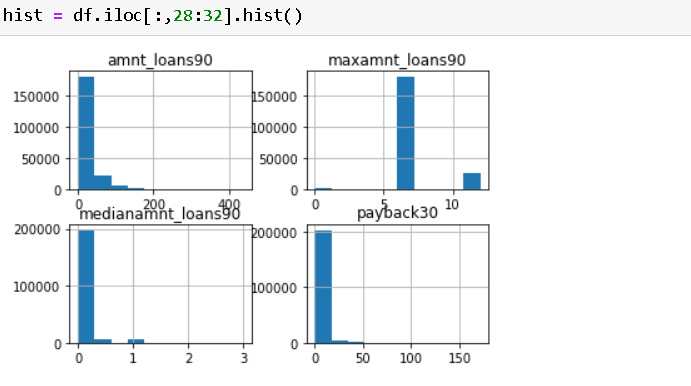


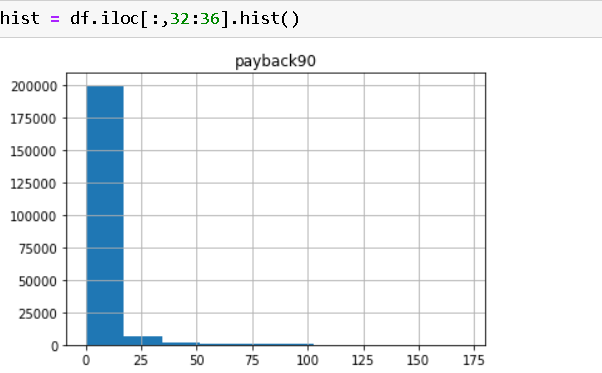


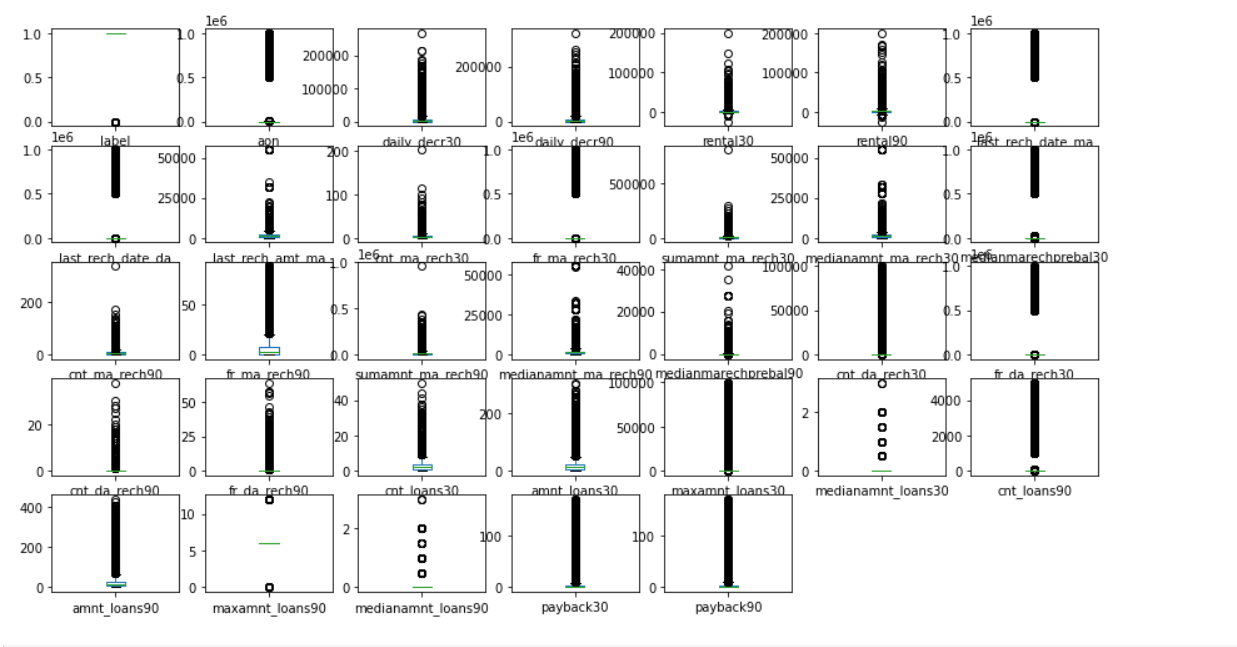




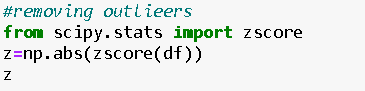


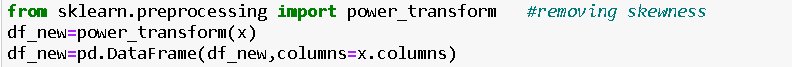


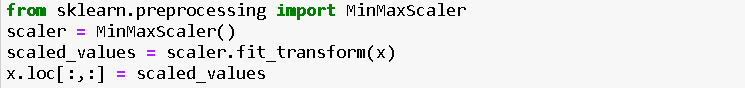




**Feature Engineering**



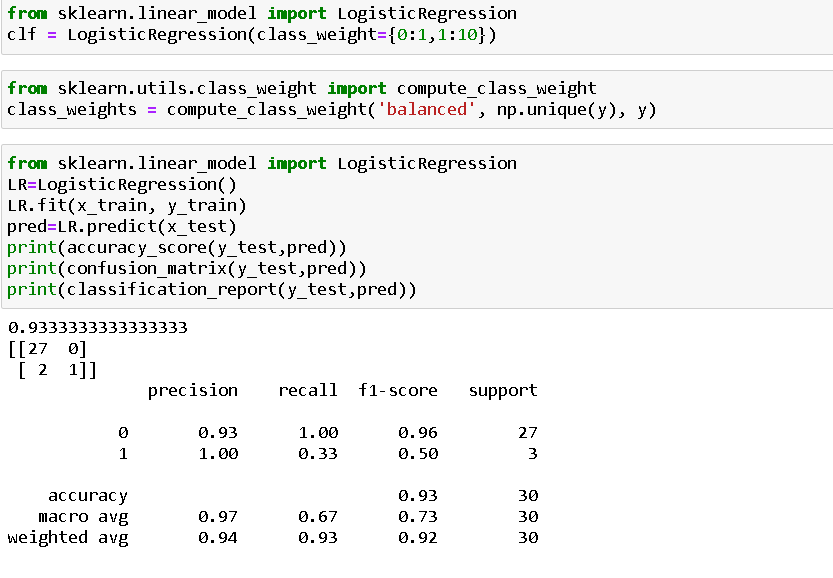


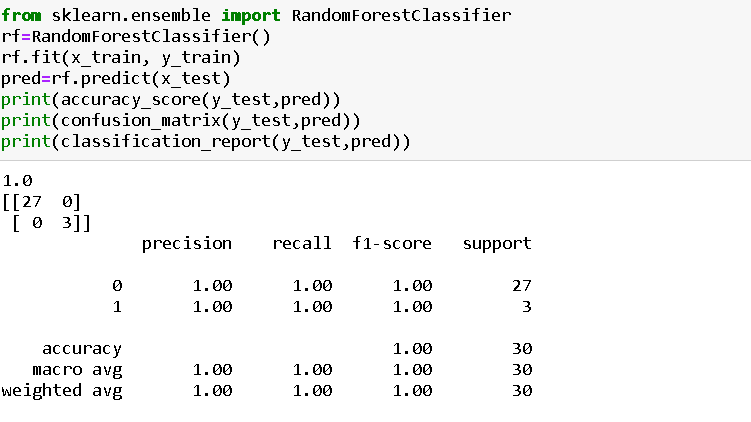


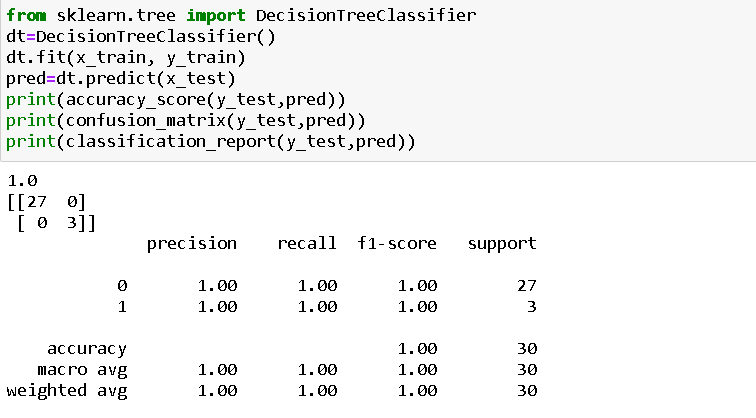
**Model Selection and Training**,

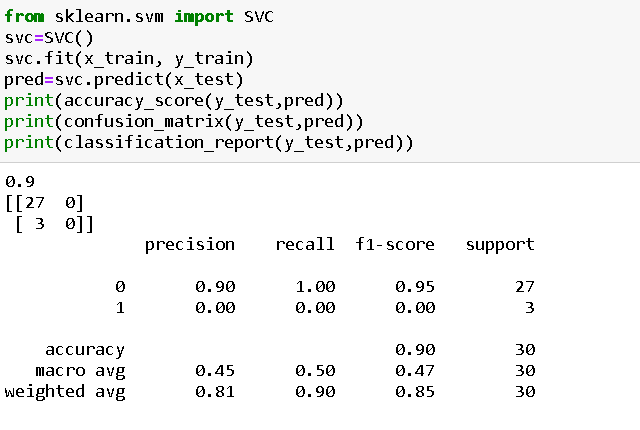
To begin, let’s split the dataset into training and test sets using 70/30 split; 70% of data will be used to train the model and the rest 30% to test the accuracy of the model. Then we can up sample the minority class, in this case the positive class.

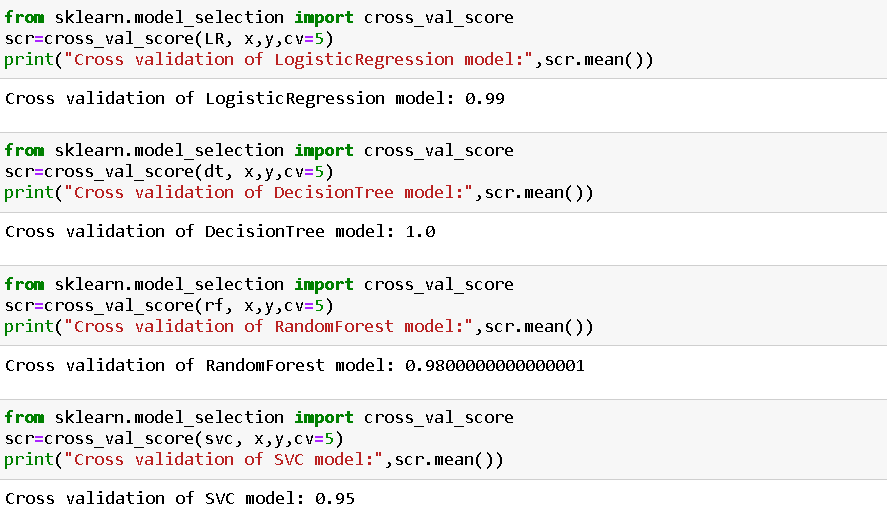
After partitioning and balancing, our data will make our model ready to be input of the machine learning models. We will train 5 different models: Logistic Regression, Random Forest, Decision Tree Classifier and SVC. In this step, we will start modifying model parameters, perform feature engineering and balancing data strategies to improve the performance of the models. Try with more trees in the Random Forest model, include new variables, penalize wrong predictions from the minority class until you beat the performance of our current best model.

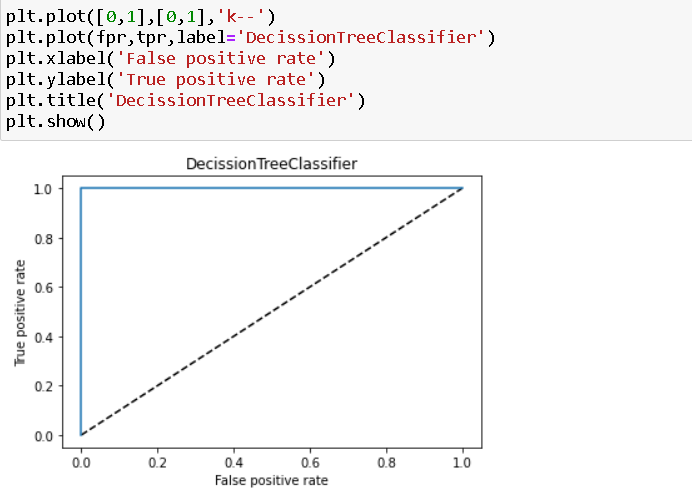


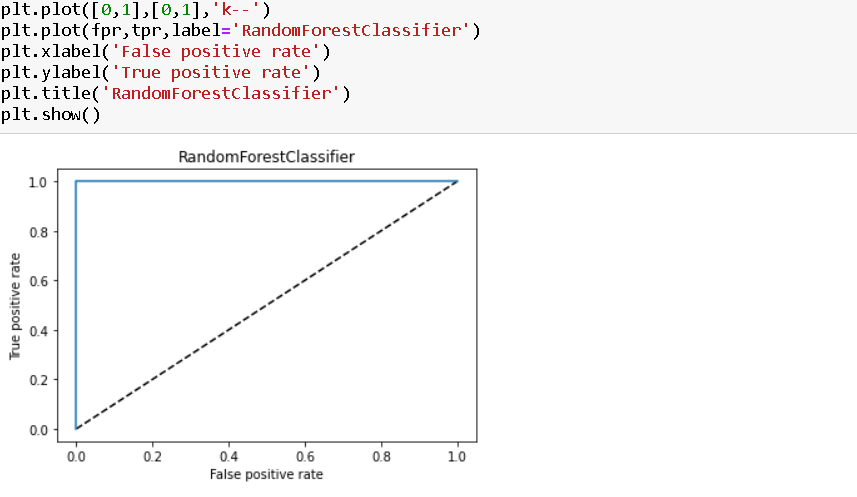


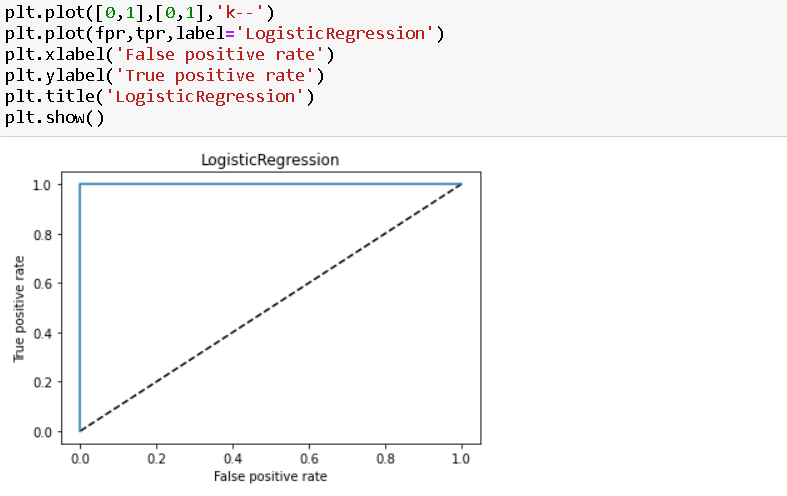


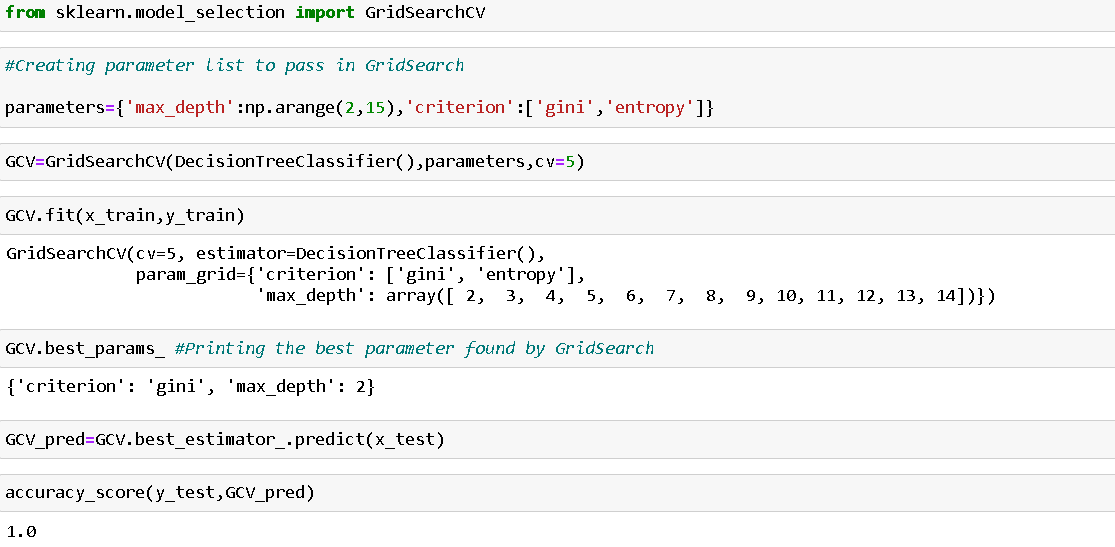












**Conclusion:**

Nowadays, the loan business becomes more and popular, and many people apply for loans for various reasons. However, there are defaulter cases which results in huge financial loss. In this study, the dataset was cleaned first, and the exploratory data analysis and feature engineering were performed. The strategies to deal with outliers, skewness and imbalanced data sets were covered. Then I propose four machine learning models to predict if the applicant could repay the loan, which are Random Forest, Logistic Regression, Support Vector Machine, and DecisionTreeClassifier. When tuning parameters, Cross Validation, AUC curve and Grid Search Cross Validation methods are applied in different situations. It is noticed that the model was found which best fits the dataset with highest F1 accuracy is the DecisionTreeClassifier model, and the model with highest AUC score is with DecisionTreeClassifier.