Micro Credit Defaulter

Submitted by:

Yatika Taneja

ACKNOWLEDGMENT

.

I am using this opportunity to express my gratitude to everyone who supported me throughout the course of this group project. I am thankful for their aspiring guidance, invaluably constructive criticism and friendly advice during the project work. I am sincerely grateful to them for sharing their truthful and illuminating views on a number of issues related to the project.
I wish to thank, all the faculties in data trained academy as this project utilized knowledge gained from every course that formed the Data science program.

INTRODUCTION

1.1> Objective of the study

The objective of our project is to build a 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 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.

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.

1.2> Business Model

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 primarily focus 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).

1.3 > Literature Survey

"Microfinance" is often seen as financial services for poor and low-income clients(In practice, the term is often used more narrowly to refer to loans and other services from providers that identify themselves as "microfinance institutions" (MFIs) [Consultative Group to Assist the Poor (CGAP) 2010]. Microfinance can also be described as a setup of a number of different operators focusing on the financially under-served people with the aim of satisfying their need for poverty alleviation, social promotion, emancipation, and inclusion. Microfinance institutions reach and serve their target market in very innovative ways Microfinance operations differ in principle, from the standard disciplines of general and entrepreneurial finance. This difference can be attributed to the fact that the size of the loans granted with microcredit is typically too small to finance growth-oriented business projects

The CGAP (2010) identifies some unique features of microfinance as follows:

a. Delivery of very small loans to unsalaried workers

b. Little or no collateral requirements o Group lending and liability
c. Pre-loan savings requirement o Gradually increasing loan sizes

Default in Microfinance

Default in microfinance is the failure of a client to repay a loan. The default could be in terms of the amount to be paid or the timing of the payment. MFIs can sustain and increase deployment of loans to stimulate the poverty reduction goal if repayment rates are high and consistent). MFIs are able to reduce interest rates and processing fees if repayment rates are high, thus increasing patronage of loans. A high repayment rate is a catalyst for increasing the volume of loan disbursements to various sectors of the economy

Several Researchers have developed several models for prediction of default). These techniques can be broadly categorized as follows: (1) Statistical models: linear discriminant analysis, logistic regression, probit regression, k nearest neighbour, classification tree, etc. (2) Mathematical programming methods: linear programming, quadratic programming, integer programming, etc. (3) Artificial intelligence techniques: artificial neural networks, support vector machines, genetic algorithm and genetic programming, rough set, etc. (4) Hybrid approaches: artificial neural network and fuzzy system, rough set and artificial neural network, fuzzy system and support vector machines etc. (5) Ensemble or combined methods: neural network ensemble, support vector machine ensemble, hybrid ensemble etc.

Analytical Problem Framing

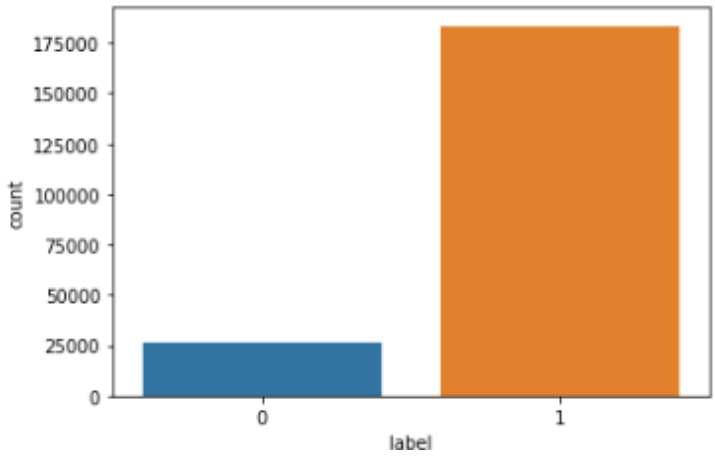
2.1> Data Sources

The sample data is provided to us from our client database. The dataset has 209593 observations and 36 features. 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 dataset is imbalanced. Label '1' has approximately 87.5% records, while, label '0' has approximately 12.5% records.

```
sns.countplot(df["label"])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x21f760442c8>
```



2.2> Data Description

VARIABLE	DESCRIPTION
label	Flag indicating whether the user paid back the credit amount within 5 days of issuing the loan(1:success, 0:failure)
msisdn	mobile number of user
aaon	Age on cellular network in days
daily_decr30	Daily amount spent from main account, averaged over last 30 days (in Indonesian Rupiah)
daily_decr90	Daily amount spent from main account, averaged over last 90 days(in Indonesian Rupiah)
rental30	Average main account balance over last 30 days
rental90	Average main account balance over last 90 days
last_rech_date_ma	Number of days till last recharge of main account
last_rech_date_da	Number of days till last recharge of data account
last_reach_amt_ma	Amount of last recharge of main account(in Indonesian Rupiah)

cnt_ma_reach30	Number of times main account got recharged in last 30 days
fr_ma_reach30	Frequency of main account recharged in last 30 days
sumamt_ma_reach30	Total amount of recharge in main account over last 30 days(in Indonesian Rupiah)
medianamnt_ma_reach30	Median of amount of recharges done in main account over last 30 days at user level(in Indonesian Rupiah)
medianmarechprebal30	Median of main account balance just before recharge in last 30 days at user level(in Indonesian Rupiah)
cnt_ma_reach90	Number of times main account got recharged in last 90 days
fr_ma_reach90	Frequency of main account recharged in last 90 days
sumamt_ma_reach90	Total amount of recharge in main account over last 90 days (in Indonesian Rupiah)
medianamnt_ma_reach90	Median of amount of recharges done in main account over last 90 days at user level(in Indonesian Rupiah)
meaduanmareachrebal90	Median of main account balance just before recharge in last 90 days at user level(in Indonesian Rupiah)
cnt_da_reach30	Number of times main account got recharged in last 30 days
fr_da_reach30	Frequency of main account recharged in last 30 days
cnt_da_reach90	Number of times data account got recharged in last 90 days
fr_da_reach90	Frequency of data account recharged in last 90 days
cnt_loans30	Number of loans taken by user in last 30 days
amnt_loans30	Total amount of loans taken by user in last 30 days
maxamnt_loans30	Maximum amount of loans taken by user in last 30 days
medianamnt_loans30	Median of amounts of loan taken by user in last 30 days
cnt_loans90	Number of loans taken by user in last 90 days
amnt_loans90	Total amount of loans taken by user in last 90 days
maxamnt_loans90	Maximum amount of loans taken by user in last 90 days
medianamnt_loans90	Median of amount of loans taken by user in last 90 days
payback30	Average payback time in days over last 30 days
payback90	Average payback time in days over last 90 days
pcircle	Telecom circle
pdate	date

2.3> Data Preprocessing

It consists into approximately 209593 observations and 37 features granted by the company, with the full set of informations about the borrower, the history of payments and the outcome of the loan. The dataset is quite clean and the figures can be considered as ground truth, but lots of columns are either irrelevant, very sparse or non informative. Moreover, the dataset is very unbalanced, Label '1' has approximately 87.5% records, while, label '0' has approximately 12.5% records. 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.

Excluding features for which the information is incomplete, or uninformative, we get a total of 19 features.

There are no missing values in the dataset and date feature is converted from object to date data type.

```
# changing date feature from object type to date type
df['pdate']=pd.to_datetime(df['pdate'])
df['Month']=df['pdate'].apply(lambda x:x.month)
df['Day']=df['pdate'].apply(lambda x:x.day)
```

```
df.drop(['pcircle','msisdn','medianamnt_loans90','last_rech_date_da','daily_decr30','rental30','cnt_ma_rech30','fr_ma_rech30','su
```

```
df.head()
```

Unnamed: 0	label	aon	daily_decr90	rental90	last_rech_date_ma	last_rech_amt_ma	medianamnt_ma_rech30	cnt_ma_rech90	fr_ma_rech90	sumamnt_ma_rec
1	0	272.0	3065.150000	260.13	2.0	1539	1539.0	2	21	3
2	1	712.0	12124.750000	3691.26	20.0	5787	5787.0	1	0	5
3	1	535.0	1398.000000	900.13	3.0	1539	1539.0	1	0	1
4	1	241.0	21.228000	159.42	41.0	947	0.0	1	0	
5	1	947.0	150.619333	1098.90	4.0	2309	2309.0	8	2	23

```
df.drop(['medianamnt_ma_rech30'],axis=1,inplace=True)
```

```
df.shape
```

```
(209593, 19)
```

Statistical Summary of dataset

```
df.describe()
```

	label	aon	daily_decr30	daily_decr90	rental30	rental90	last_rech_date_ma	last_rech_date_da	last_rech_amt_ma	cr
count	209593.000000	209593.000000	209593.000000	209593.000000	209593.000000	209593.000000	209593.000000	209593.000000	209593.000000	2
mean	0.875177	8112.343445	5381.402289	6082.515068	2692.581910	3483.406534	3755.847800	3712.202921	2064.452797	
std	0.330519	75696.082531	9220.623400	10918.812767	4308.586781	5770.461279	53905.892230	53374.833430	2370.786034	
min	0.000000	-48.000000	-93.012667	-93.012667	-23737.140000	-24720.580000	-29.000000	-29.000000	0.000000	
25%	1.000000	246.000000	42.440000	42.692000	280.420000	300.260000	1.000000	0.000000	770.000000	
50%	1.000000	527.000000	1469.175667	1500.000000	1083.570000	1334.000000	3.000000	0.000000	1539.000000	
75%	1.000000	982.000000	7244.000000	7802.790000	3356.940000	4201.790000	7.000000	0.000000	2309.000000	
max	1.000000	999860.755168	265926.000000	320630.000000	198926.110000	200148.110000	998650.377733	999171.809410	55000.000000	

8 rows × 33 columns

2.4> Tools

Python 3.7.2,
Jupyter Notebook
Numpy,
Pandas
Matplotlib
Seaborn
Scikit-learn
Scipy

2.5> Data Visualization

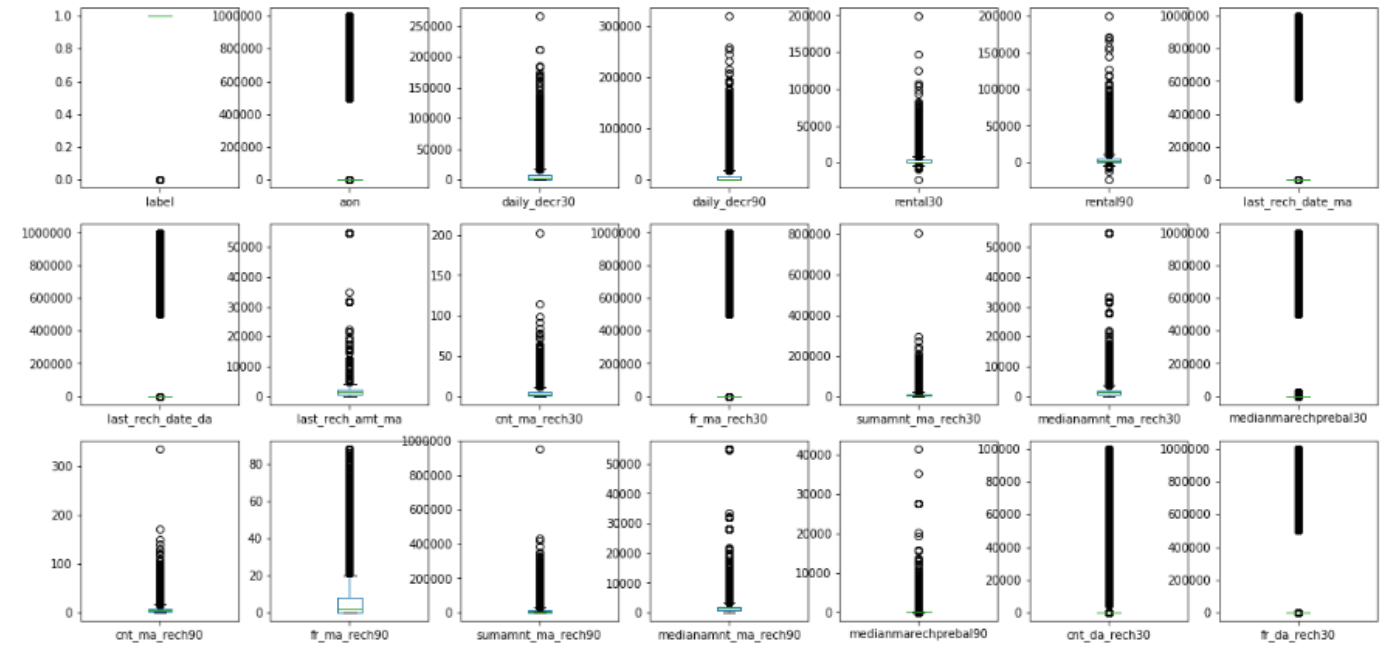
➢ Univariate Analysis

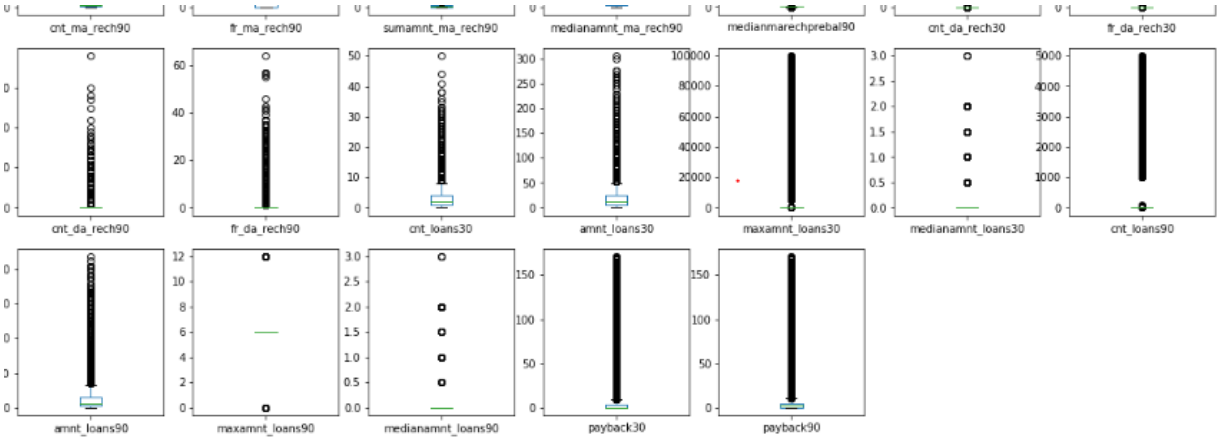
We start with some univariate plots, that is, plots of each individual variable.

a. Box Plots

It gives much clearer idea of distribution of features and detect if outliers are present or not

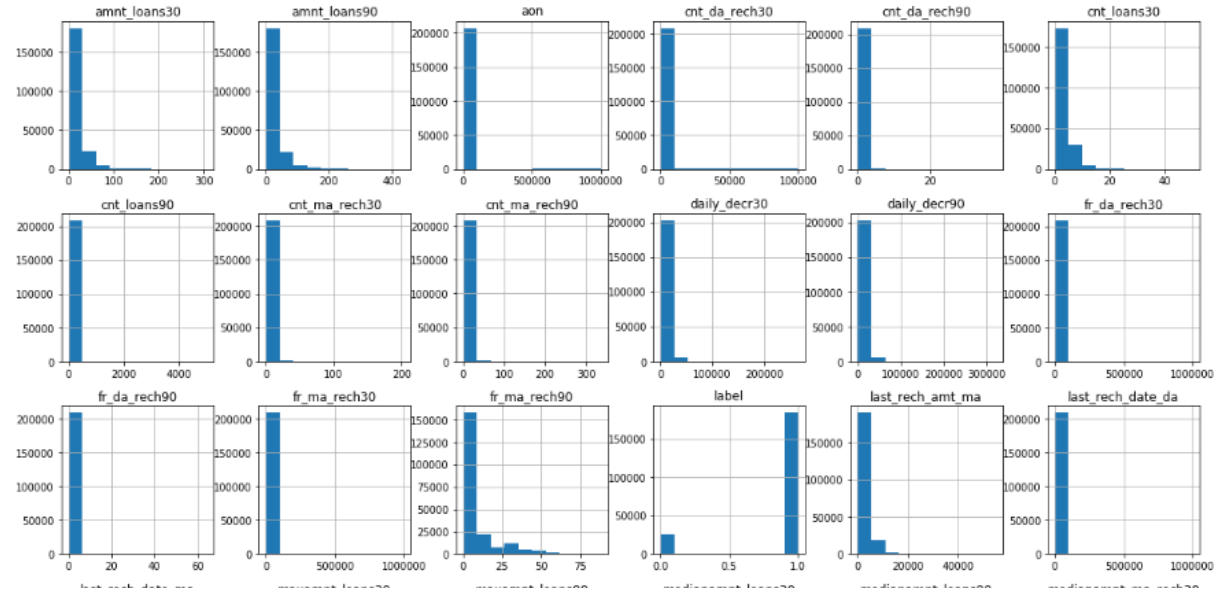
```
df.plot(kind='box', subplots=True, layout=(6,7),figsize=(20,20),sharex=False, sharey=False)
plt.show()
```

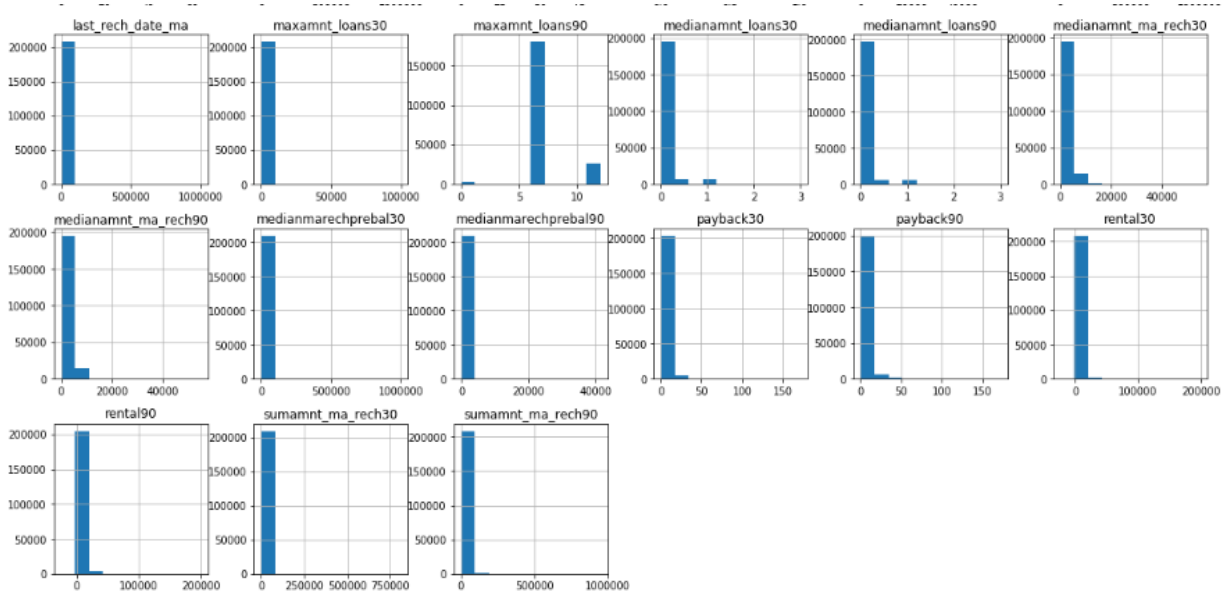




b. Histograms

```
df.hist(figsize=(20,20))
plt.show()
```





➢ Multivariate analysis

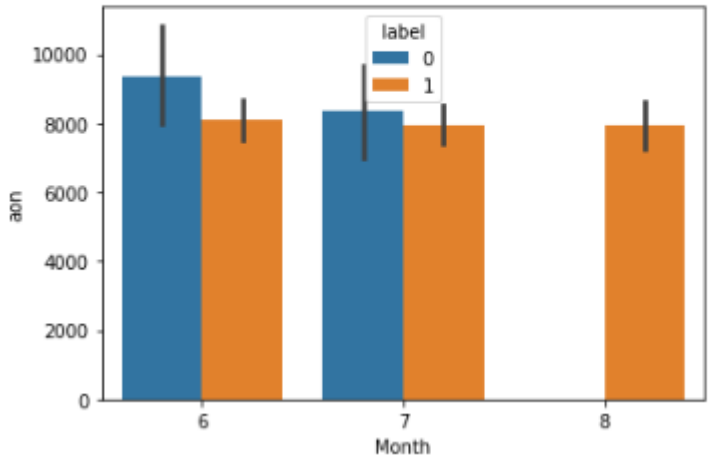
It gives the relationship between multiple features

1. Barplot

   a. Relationship between month , aon and label

```
sns.barplot(x=df["Month"],y=df["aon"],hue=df["label"])
```

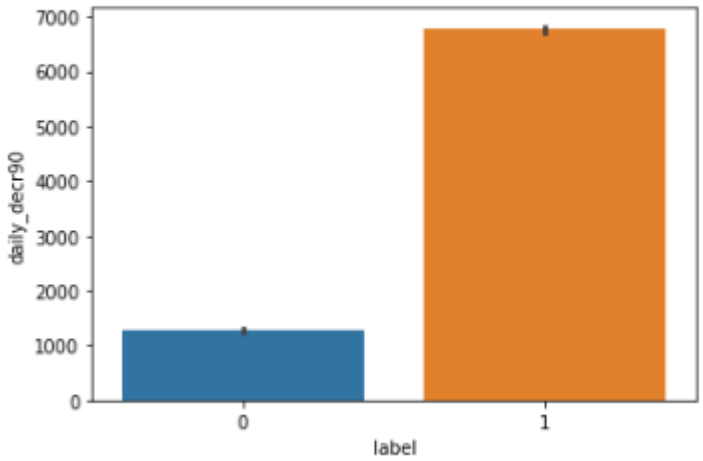
```
<matplotlib.axes._subplots.AxesSubplot at 0x211d6a52248>
```



   b. Relationship between label and daily_decr90

```
sns.barplot(x=df["label"],y=df["daily_decr90"])
```

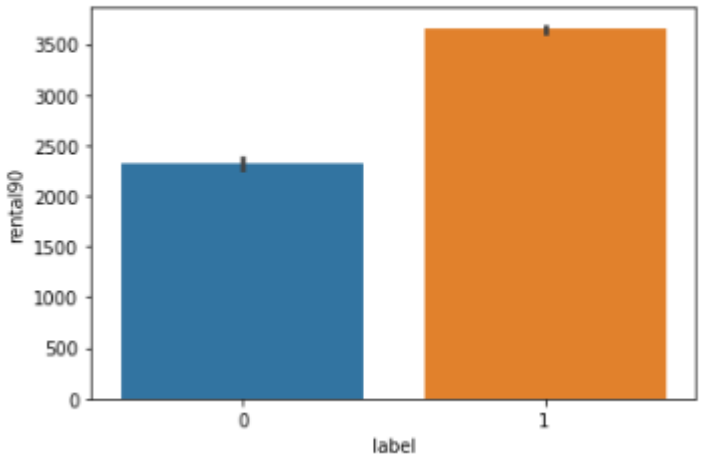
```
<matplotlib.axes._subplots.AxesSubplot at 0x211d7290ec8>
```



c. Relationship between label and rental90

```
sns.barplot(x=df["label"],y=df["rental90"])
```

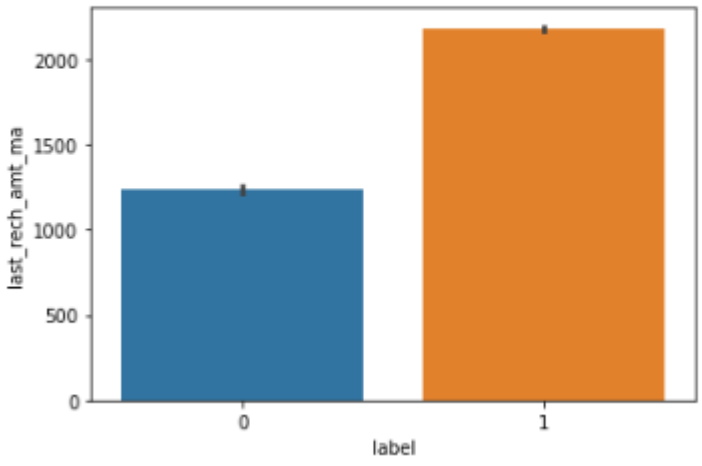
```
<matplotlib.axes._subplots.AxesSubplot at 0x211d72b28c8>
```



d. Relationship between label and last_rech_amt_ma

```
sns.barplot(x=df["label"],y=df["last_rech_amt_ma"])
```

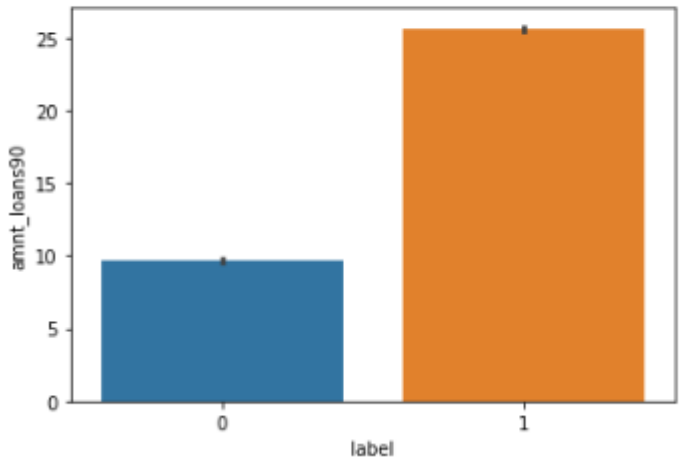
```
<matplotlib.axes._subplots.AxesSubplot at 0x211d510de48>
```



e. Relationship between label and amnt_loans90

```
sns.barplot(x=df["label"],y=df["amnt_loans90"])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x211d5136888>
```



Model/s Development and Evaluation

3.1> Models Applied

Logistic Regression:

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes. In simple words, the dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no).
Mathematically, a logistic regression model predicts P(Y=1) as a function of X. It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, Diabetes prediction, cancer detection etc.

Linear Discriminant Analysis:

Linear discriminant analysis (**LDA**), **normal discriminant analysis** (**NDA**), or **discriminant function analysis** is a generalization of **Fisher's linear discriminant**, a method used in statistics and other fields, to find a linear combination of features that characterizes or separates two or more classes of objects or events.

K-nearest Neighbors:

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.

Decision Tree Classifier:

The classification technique is a systematic approach to build classification models from an input data set. For example, decision tree classifiers, rule-based classifiers, neural networks, support vector machines, and naive Bayes classifiers are different technique to solve a classification problem. Each technique adopts a learning algorithm to identify a model that best fits the relationship between the attribute set and class label of the input data. Therefore, a key objective of the learning algorithm is to build predictive model that accurately predict the class labels of previously unknown records.

Gaussian Naïve Bayes:

In statistics, **Naive Bayes classifiers** are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naïve) independence assumptions between the features. They are among the simplest Bayesian network models,[1] but coupled **with** Kernel density estimation, they can achieve higher accuracy levels

3.2> Model Evaluation

We will use stratified 10-fold cross validation to estimate model accuracy.
This will split our dataset into 10 parts, train on 9 and test on 1 and repeat for all combinations of train-test splits.
Stratified means that each fold or split of the dataset will aim to have the same distribution of example by class as exist in the whole training dataset.

.k-Fold Cross-Validation

Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample.

The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation. When a specific value for k is chosen, it may be used in place of k in the reference to the model, such as k=10 becoming 10-fold cross-validation.

Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. That is, to use a limited sample

in order to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model.

Evaluation Metric: Accuracy

This is a ratio of the number of correctly predicted instances divided by the total number of instances in the dataset multiplied by 100 to give a percentage (e.g. 95% accurate). We will be using the scoring variable when we run build and evaluate each model next.

We now have 6 models and accuracy estimations for each. We need to compare the models to each other and select the most accurate

Accuracy Result:

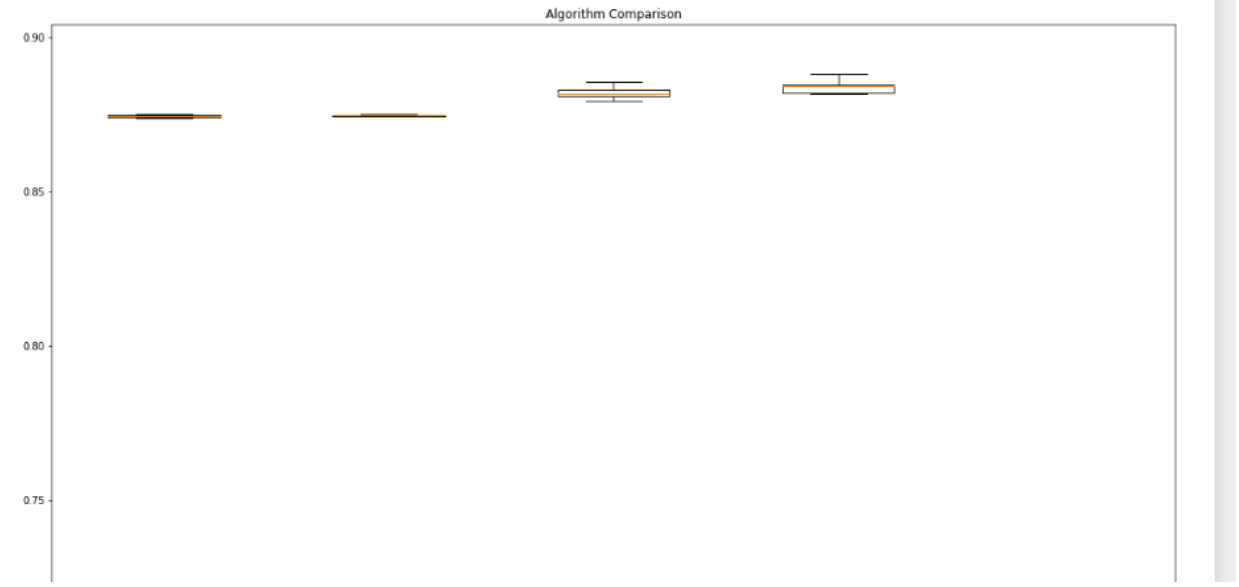
LR: 0.874524 (0.000393)
LDA: 0.874709 (0.000182)
KNN: 0.882063 (0.001791)
CART: 0.884091 (0.002200)
NB: 0.575128 (0.006369)

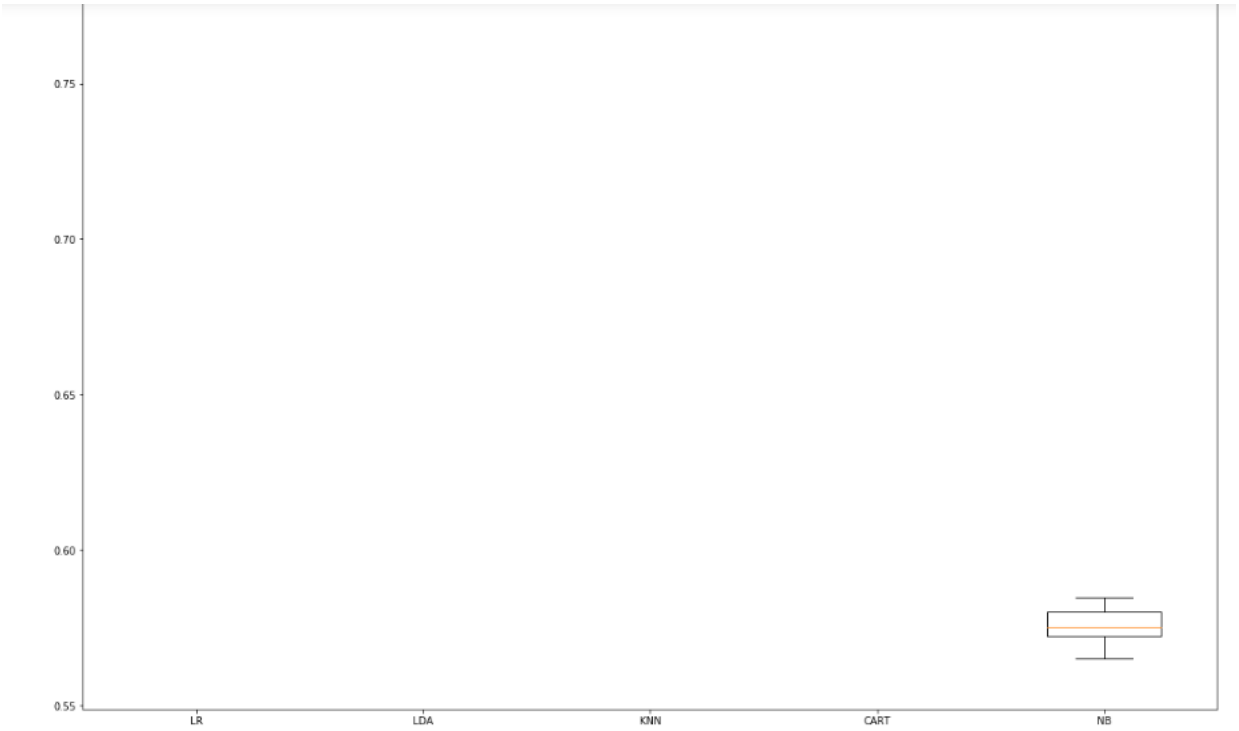
3.3> Select Best Model

In this case, we can see that it looks like Decision Tree Classifier(Cart) has the largest estimated accuracy score at about 0.884 or 88.4%.

We can also create a plot of the model evaluation results and compare the spread and the mean accuracy of each model. There is a population of accuracy measures for each algorithm because each algorithm was evaluated 10 times (via 10 fold-cross validation).

```
# Compare Algorithms
plt.figure(figsize=(20,20))
plt.boxplot(results, labels=names)
plt.title('Algorithm Comparison')
plt.show()
```





3.4> Evaluate Prediction

We can evaluate the predictions by comparing them to the expected results in the validation set, then calculate classification accuracy, as well as a confusion matrix and a classification report.

```
model = DecisionTreeClassifier()
model.fit(X_train, Y_train)
predictions = model.predict(X_validation)
```

```
print(accuracy_score(Y_validation, predictions))
print(confusion_matrix(Y_validation, predictions))
print(classification_report(Y_validation, predictions))
```

```
0.8865192394856748
[[ 2894  2286]
 [ 2471 34268]]
              precision    recall  f1-score   support

           0       0.54      0.56      0.55      5180
           1       0.94      0.93      0.94     36739

    accuracy                           0.89     41919
   macro avg       0.74      0.75      0.74     41919
weighted avg       0.89      0.89      0.89     41919
```

Key Features:

- The cross validation score suggests that Decision Tree Classifier is the best model
- The accuracy is 0.886 or about 88.6% on the hold out dataset.
- The confusion matrix provides an indication of the errors made.
- Finally, the classification report provides a breakdown of each class by precision, recall, f1-score and support showing excellent results

CONCLUSION

Microfinance has been globally accepted as the preferred medium to reach out to the rural and productive poor with banking services which includes micro credit to help alleviate poverty which is one of the United Nations millennium challenge goals. Micro credit default has been identified to be one of the major drawbacks of this laudable initiative as it depletes these revolving funds and reduces investors' confidence. Therefore, it is important to understand the factors that influence a loan beneficiary to default so that appropriate countermeasures can be developed to prevent and reduce the incidents of default. In this study, decision tree classifier was applied to identify the factors associated with the occurrence of micro credit default. The analysis showed that month,aon,label,daily_decr90, rental90, amts_loan90 and other deductions were important determinants of default.