FLIP ROBO

Micro-Credit Defaulter Mode

Submitted by:

Abdul Khan

ACKNOWLEDGMENT

I would like to express my gratitude to my SME for guidance and insightful comment and observation throughout the duration of this project. During the project I took references from books like Think stats, OReilly Python for Data Analysis, Building Machine Learning Systems with Python, Data Science and Productivity Analytics, Coding for Python, Data Analysis with IBM SPSS Statistics, Mandal J. Algorithms in Machine Learning Paradigms etc . and paper to complete my project.

INTRODUCTION

The microfinance movement has changed perceptions towards helping the poor in both, South Asia and Latin America. In many countries, microfinance has been used as a tool to increase financial depth in rural areas and it has typically targeted very low-income group. The potential role financial institutions and intermediation play in economic growth and development, when it comes with Telecom industries it will be game-changer for dealing with low-income families and poor families Because Tele communication industries are one of the widest industries in the world. According to the International Telecom Union survey, south Asia is on Top to telecom user. In 2001 alone South Asia having 1345.5 million users, among them around 72% are from Rural Area and this figure goes up rampantly, which contribute approximate 128.4 billion Dollar in GDP.

Microfinance is a credit methodology, which employs effective collateral substitute for short-term and working capital loans to micro-entrepreneurs. The level of a country's poverty has long been linked with measures of its economic development. The economies with a positive growth rate of Gross National Product (GNP) were measured by their poverty mitigation. This gratitude emphasized the achievement of wealth and technology as a path for development and assumed that improved lives for all would be the natural consequence.

Microfinance is not a new development. Some developed countries as well as developing countries particularly in Asia have a long history of microfinance. During the eighteenth and nineteenth centuries, in the number of European countries, microfinance evolved as a type of informal banking for the poor. Informal finance and self-help have been at the foundation of microfinance in Europe. Microfinance has a huge impact on the lives of millions of poor people particularly women. Numerous scholars and NGOs have been working to take microfinance within the reach of poor people, who are still not benefited by the conventional financial system. It was believed that microfinance is not important for all people but most groups can benefit from this idea.

Micro-credit:

It is a component of microfinance and is the extension of small loans to entrepreneurs, who are too poor to qualify for traditional bank loans. Especially in developing countries, micro-credit enables very poor people to engage in self-employment projects that generate income, thus allowing them to improve the standard of living for themselves and their families.

Microcredit with Telecom Industries

Telecom industries are widely used in all over the globe. Around 72% of users came from Rural Area, There are different mode where microcredit is used in these industries. Mode of Credit solution provided by operator and service provider with the ability to extend their service to their user through small and short term facility like Emergency credit service or Seamless Distribution Services.

Analytical Problem Framing

Mathematical/ Analytical Modeling of the Problem

In alone Indonesia, there are more than 60,000 MFI's are present and they able to connect 50 million people. In our sample datasets, there is around 209593 customers information is available through which 12.5% is a defaulter. Finding out from EDA -

- Average cellular Age in the whole dataset is above 300 but if we compare with respect to Defaulter it gradually comes down from 300.

- Total Average amount user is lying under 1000 and the main account got average 2 times recharged with a frequency more than 3500.

- In terms of Defaulter, most of the defaulter is recharge only one time but others are more than that.

- Around 80% of the users took loan by 6 Indonesian Rupees and rests of them are 12 Rupees.

- There is not any single strong correlation between two columns.

Data Pre-processing –

1-There are not a single null values in datasets but it having strings in cellular age columns and it was treated by `df['msisdn']=df['msisdn'].str.replace('I','1').astype(int)`

2- There are more unrealistic values present in data set especially in '**cnt_da_rech90**'in order to filter these value, a threshold is given.

3- Treating outliers with z score.

Model/s Development and Evaluation

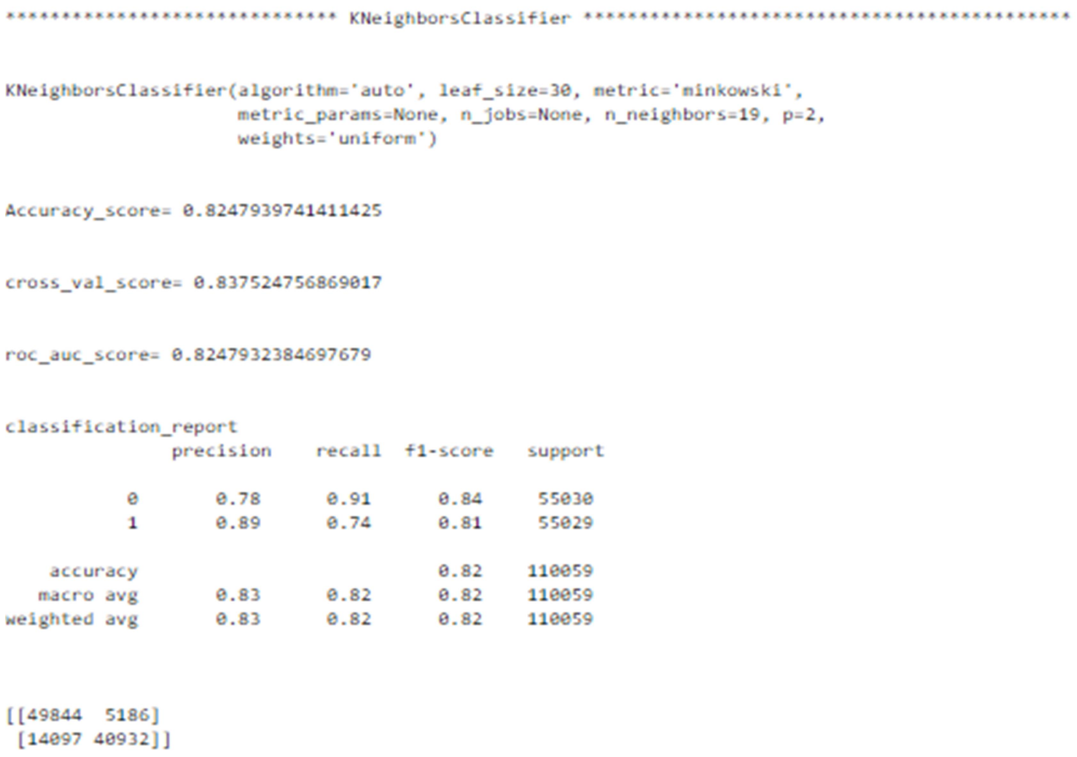
Testing of Identified Approaches (Algorithms)-

More than 5 models where used to solve the problem-

- `KNeighborsClassifier`
- `LogisticRegression`
- `DecisionTreeClassifier`
- `GaussianNB`
- `RandomForestClassifier`
- `AdaBoostClassifier`
- `GradientBoostingClassifier`

1- KNeighborsClassifier-

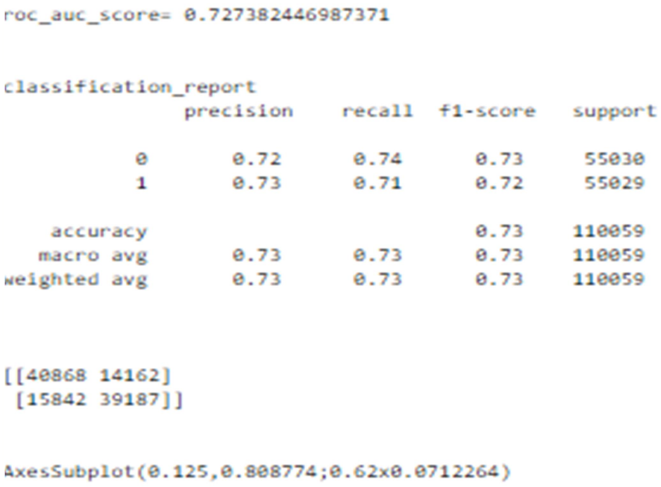
K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a simi measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition already beginning of 1970's as a non-parametric technique.



```
****************************** KNeighborsClassifier ********************************************


KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',
                     metric_params=None, n_jobs=None, n_neighbors=19, p=2,
                     weights='uniform')


Accuracy_score= 0.8247939741411425


cross_val_score= 0.837524756869017


roc_auc_score= 0.8247932384697679


classification_report
              precision    recall  f1-score   support

           0       0.78      0.91      0.84     55030
           1       0.89      0.74      0.81     55029

    accuracy                           0.82    110059
   macro avg       0.83      0.82      0.82    110059
weighted avg       0.83      0.82      0.82    110059



[[49844  5186]
 [14097 40932]]
```

- 2- LogisticRegression-

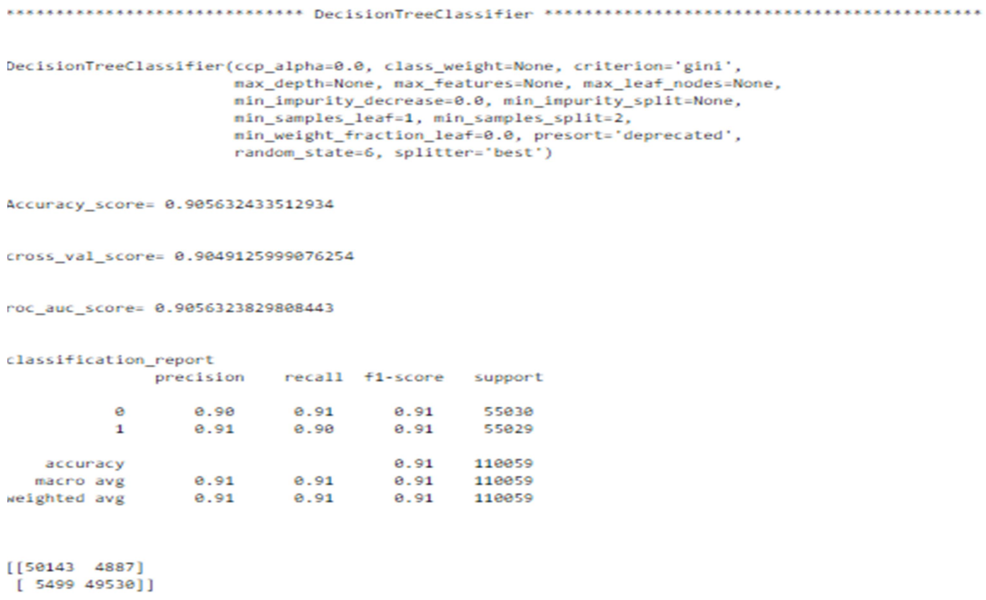
Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist.



```
roc_auc_score= 0.727382446987371


classification_report
              precision    recall  f1-score   support

           0       0.72      0.74      0.73     55030
           1       0.73      0.71      0.72     55029

    accuracy                           0.73    110059
   macro avg       0.73      0.73      0.73    110059
weighted avg       0.73      0.73      0.73    110059



[[40868 14162]
 [15842 39187]]


AxesSubplot(0.125,0.808774;0.62x0.0712264)
```

2- DecisionTreeClassifier

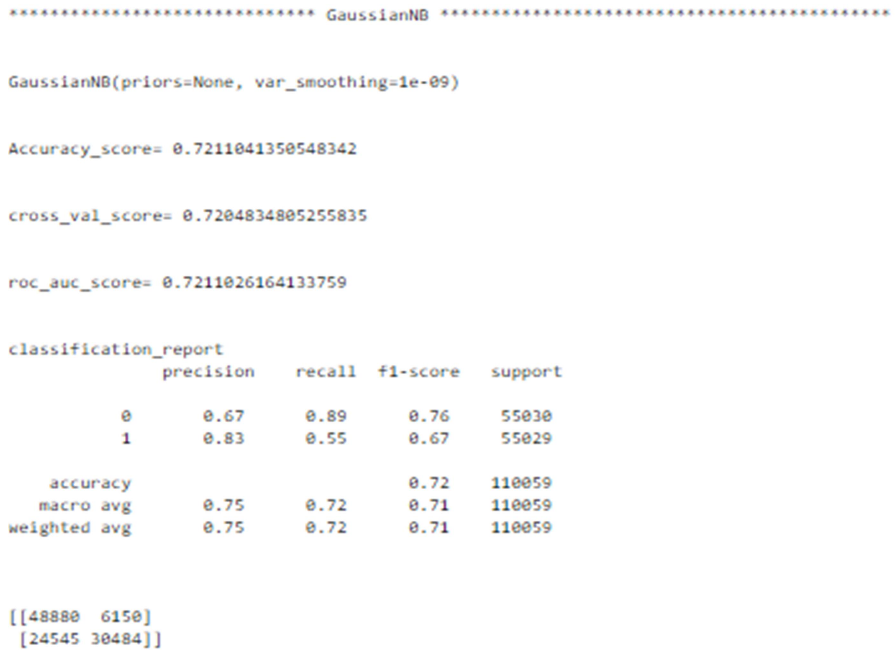
The decision tree classifier creates the classification model by building a decision tree. Each node in the tree specifies a test on an attribute, each branch descending from that node corresponds to one of the possible values for that attribute.



```
****************************** DecisionTreeClassifier ********************************************

DecisionTreeClassifier(ccp_alpha=0.0, class_weight=None, criterion='gini',
                       max_depth=None, max_features=None, max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, presort='deprecated',
                       random_state=6, splitter='best')

Accuracy_score= 0.905632433512934

cross_val_score= 0.9049125999076254

roc_auc_score= 0.9056323829808443

classification_report
              precision    recall  f1-score   support

           0       0.90      0.91      0.91     55030
           1       0.91      0.90      0.91     55029

    accuracy                           0.91    110059
   macro avg       0.91      0.91      0.91    110059
weighted avg       0.91      0.91      0.91    110059


[[50143  4887]
 [ 5499 49530]]
```

4- GaussianNB-

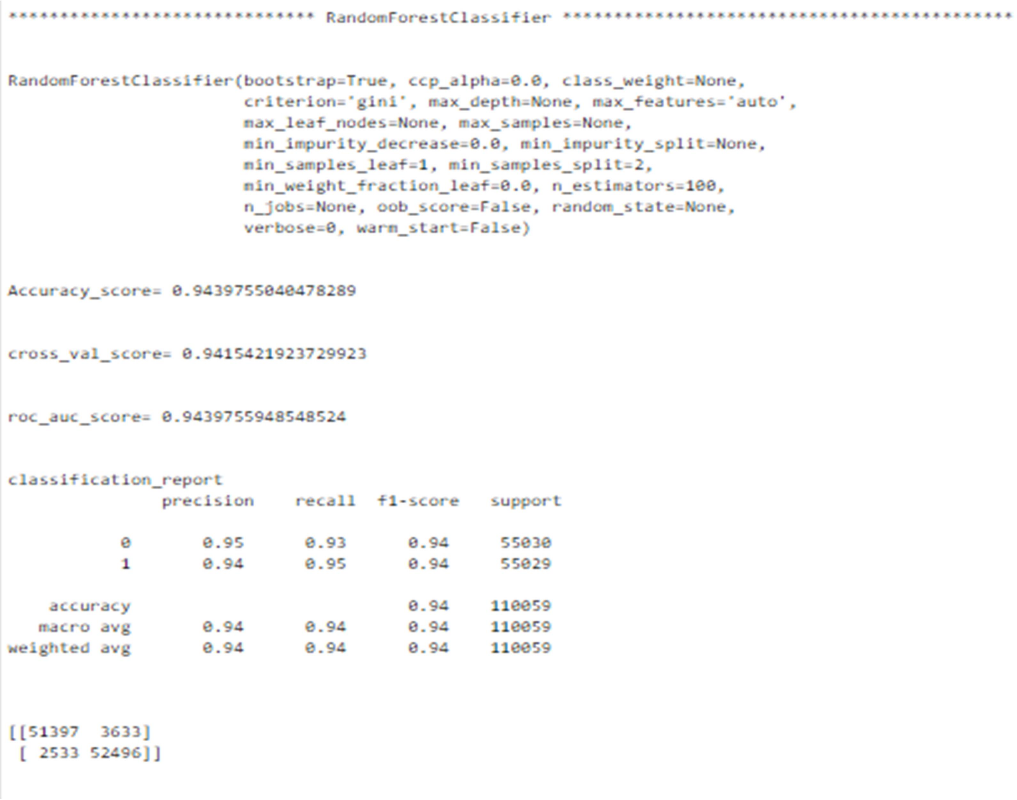
Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem**.** It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.



```
****************************** GaussianNB *********************************************


GaussianNB(priors=None, var_smoothing=1e-09)


Accuracy_score= 0.7211041350548342


cross_val_score= 0.7204834805255835


roc_auc_score= 0.7211026164133759


classification_report
              precision    recall  f1-score   support

           0       0.67      0.89      0.76     55030
           1       0.83      0.55      0.67     55029

    accuracy                           0.72    110059
   macro avg       0.75      0.72      0.71    110059
weighted avg       0.75      0.72      0.71    110059



[[48880  6150]
 [24545 30484]]
```

5- RandomForestClassifier-

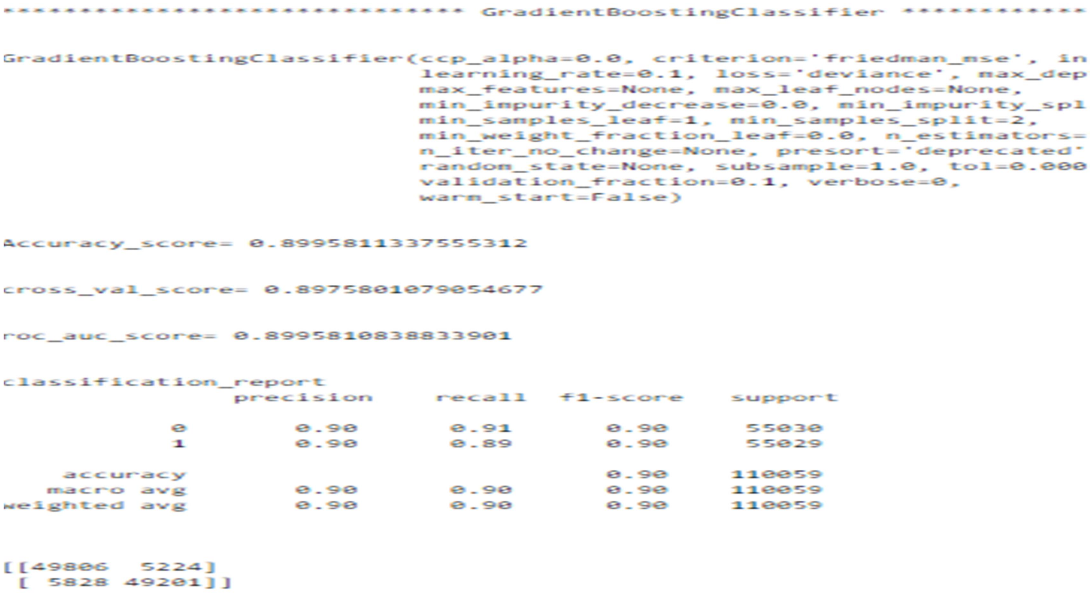
It is an ensemble tree-based learning algorithm. The Random Forest Classifier is a set of decision trees from randomly selected subset of training set. It aggregates the votes from different decision trees to decide the final class of the test object.



```
****************************** RandomForestClassifier ********************************************


RandomForestClassifier(bootstrap=True, ccp_alpha=0.0, class_weight=None,
                       criterion='gini', max_depth=None, max_features='auto',
                       max_leaf_nodes=None, max_samples=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, n_estimators=100,
                       n_jobs=None, oob_score=False, random_state=None,
                       verbose=0, warm_start=False)


Accuracy_score= 0.9439755040478289


cross_val_score= 0.9415421923729923


roc_auc_score= 0.9439755948548524


classification_report
               precision    recall  f1-score   support

           0       0.95      0.93      0.94     55030
           1       0.94      0.95      0.94     55029

    accuracy                           0.94    110059
   macro avg       0.94      0.94      0.94    110059
weighted avg       0.94      0.94      0.94    110059



[[51397  3633]
 [ 2533 52496]]
```

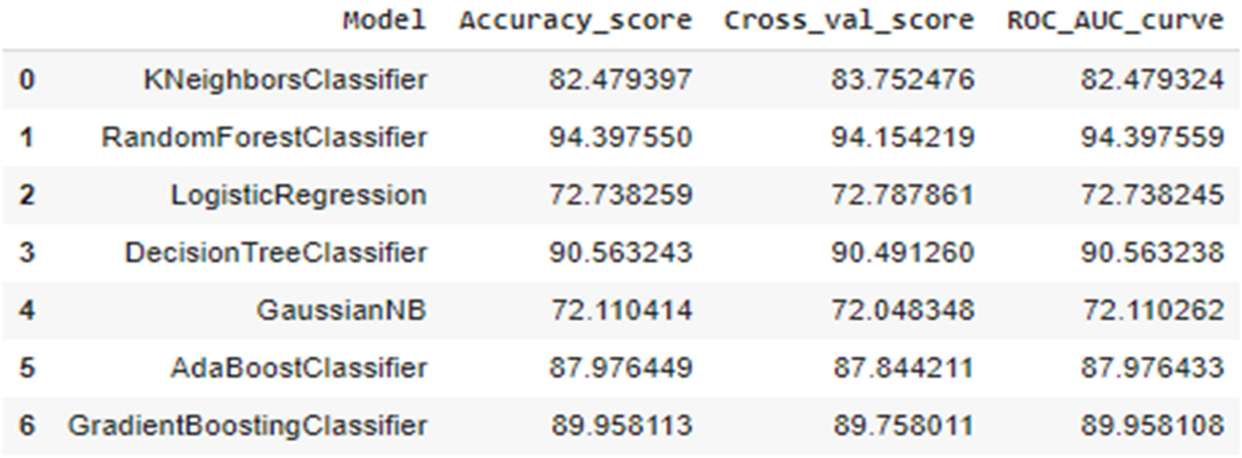
6-GradientBoostingClassifier

Gradient boosting classifiers are a group of machine learning algorithms that combine many weak learning models together to create a strong predictive model.



```
****************************** GradientBoostingClassifier ************

GradientBoostingClassifier(ccp_alpha=0.0, criterion='friedman_mse', in
                           learning_rate=0.1, loss='deviance', max_dep
                           max_features=None, max_leaf_nodes=None,
                           min_impurity_decrease=0.0, min_impurity_spl
                           min_samples_leaf=1, min_samples_split=2,
                           min_weight_fraction_leaf=0.0, n_estimators=
                           n_iter_no_change=None, presort='deprecated'
                           random_state=None, subsample=1.0, tol=0.000
                           validation_fraction=0.1, verbose=0,
                           warm_start=False)

Accuracy_score= 0.8995811337555312

cross_val_score= 0.8975801079054677

roc_auc_score= 0.8995810838833901

classification_report
               precision    recall  f1-score   support

           0       0.90      0.91      0.90     55030
           1       0.90      0.89      0.90     55029

    accuracy                           0.90    110059
   macro avg       0.90      0.90      0.90    110059
weighted avg       0.90      0.90      0.90    110059


[[49806  5224]
 [ 5828 49201]]
```

Among all these Models Random Forest classifier perform good with accuracy 94% and roc 0.94.



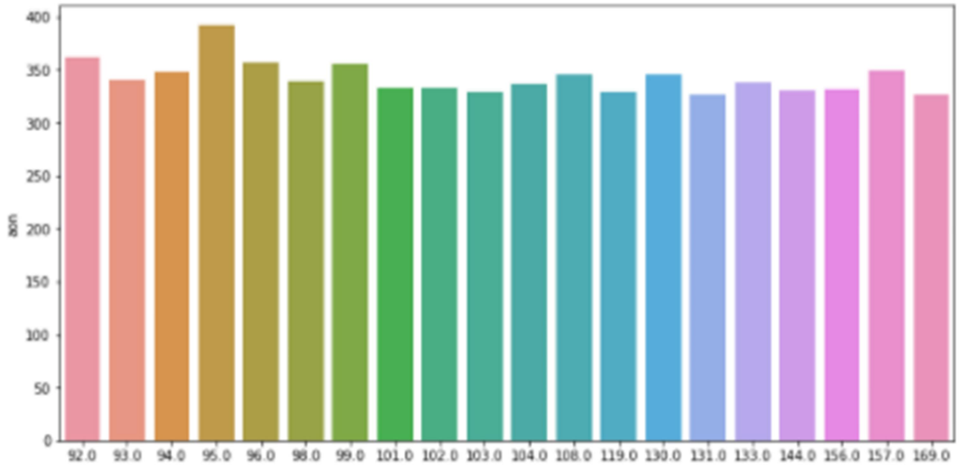
	Model	Accuracy_score	Cross_val_score	ROC_AUC_curve
0	KNeighborsClassifier	82.479397	83.752476	82.479324
1	RandomForestClassifier	94.397550	94.154219	94.397559
2	LogisticRegression	72.738259	72.787861	72.738245
3	DecisionTreeClassifier	90.563243	90.491260	90.563238
4	GaussianNB	72.110414	72.048348	72.110262
5	AdaBoostClassifier	87.976449	87.844211	87.976433
6	GradientBoostingClassifier	89.958113	89.758011	89.958108

Visualizations for Datasets-

Cellular Age of the Users-

Top cellular Age counts.

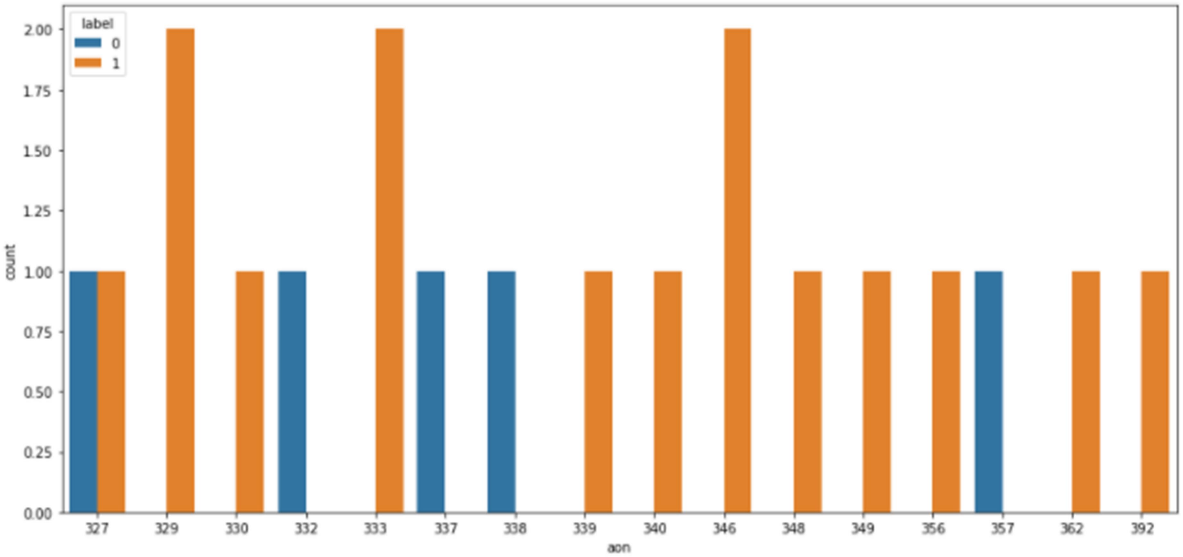
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f415b8abb70>
```



Most of the cellular age in the data set is above 300 days and if we compare to Defaulter then this value goes down which is below 300.It’s very difficult to verify from age of cell network.

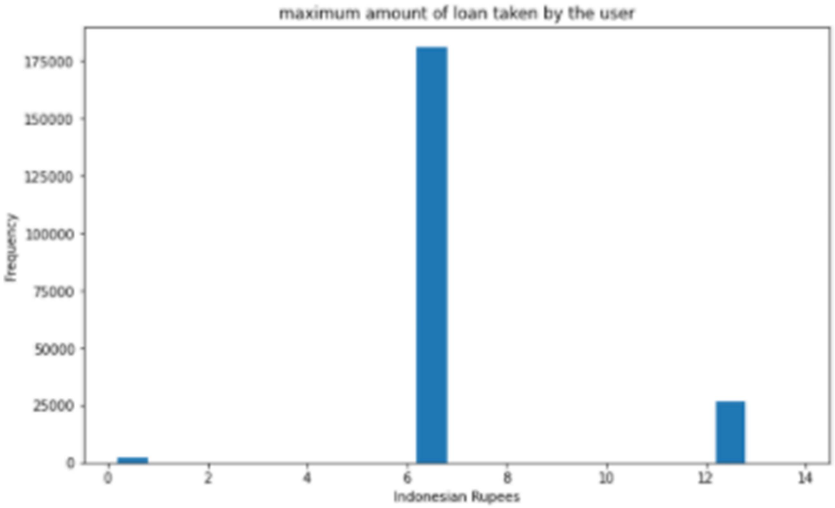
Cellular Age with respect to Defaulter

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f414b96d860>
```



Maximum loan counts

In whole datasets only two loan amount is credited to the user which is 6 & 12 . If we use count method to maximum loan then we observe that the maximum loan was 6 .The frequency of mini loan of Rupees

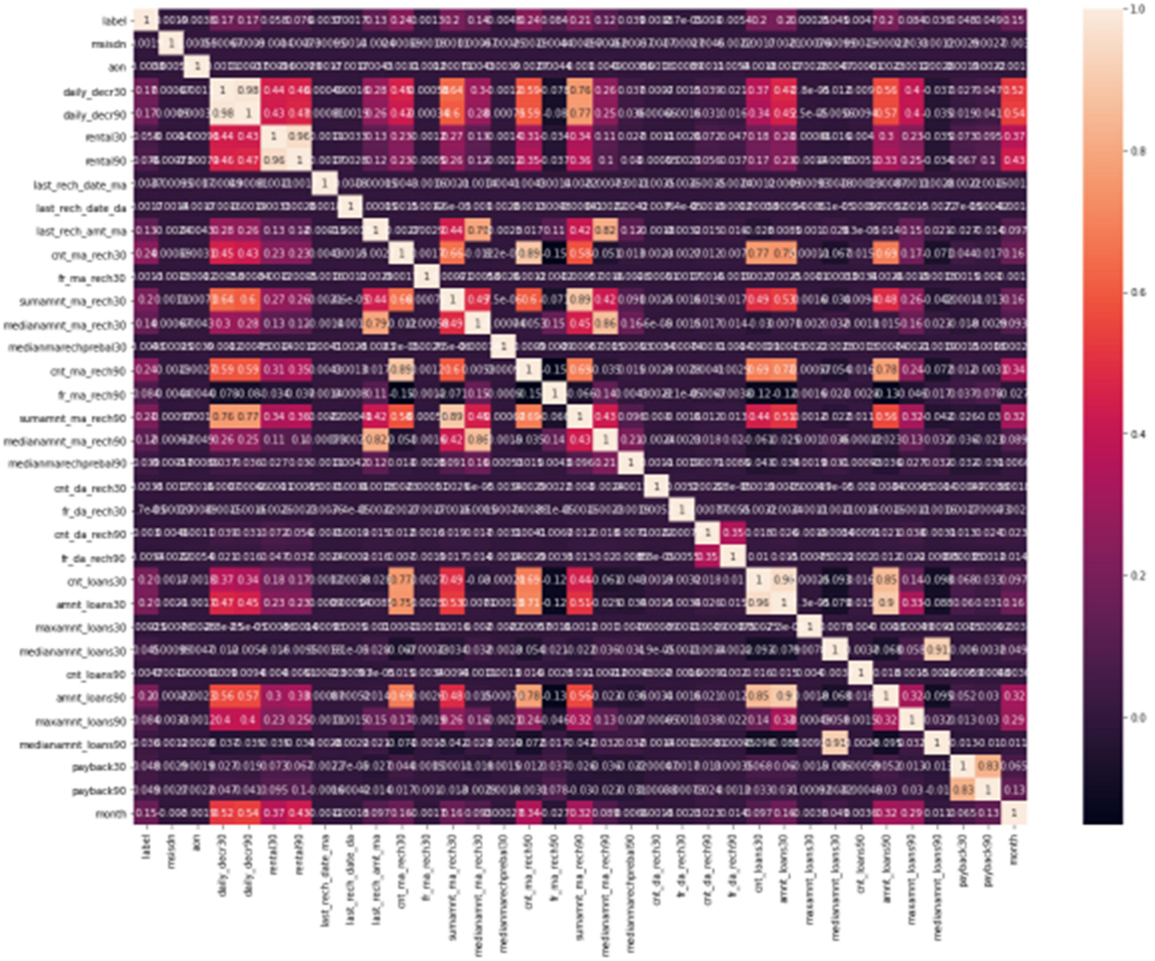


6 is much more higher than 12 Rupees.

If we go with maximum amount with respect to defaulter then also we find the number of defaulter comes in 6 Rupees loan group.

Correlation Between columns –

When we come to the correlation we didn’t find any strong correlation between two columns



but we found some columns which is correlate each other like medianamnt_loans30 and medianamnt_loans90 etc. As we see these 30 days and 90 days data is very similar to each other.

CONCLUSION

Normally the datasets contain lots of feature but after analyzing the whole data, there is some pattern in Defaulter list. The Defaulter doesn’t having old cellular connection because if we see the age of defaulter cellular , it always less than 300 days that mean most of people who having new connection is in defaulter list. The other thing is Defaulter comes only in 6th and 7th months with maximum amount of loan is Rupees 6. Apart from that the tendency of main account is also very less as compare to Non defaulter.

Limitations of this work and Scope for Future Work-

There are some difficulty come when working with large dataset. Normally our computer is not made for tough and hectic work so several times it got stuck especially during Modeling. In order to solve this problem I switch from Jupyter Notebook to Google colab and I does not involved in much depth like hyper tuning and other so that my computer not stuck too much.