

Micro Credit Project

Submitted by:

RAVI VERMA

**ACKNOWLEDGMENT**

For the following project I used certain references which are mentioned below:

1. towardsdatascience.com
2. medium.com
3. stackoverflow.com
4. bing.com

**INTRODUCTION**

* Business Problem Framing

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).

* Conceptual Background of the Domain Problem

A Microfinance Institution (MFI) is an organization that offers financial services to low income populations. Microfinance services (MFS) becomes very useful when targeting especially the unbanked poor families living in remote areas with not much sources of income. The MFS provided by MFI are different type of Loans.

Basically, here a one telecom industry provides the 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.

We know that telecom sector is very much competitive so this data is very helpful in understanding the problem for the lower-class people specially by providing them the facility of network and the credit amount provided by the help of MFI and MFS. From this data we get to know that what the criteria to become defaulters and successor are. And the useful information from the data to know how much amount people spend on data recharge or on the main balance recharge.

* Review of Literature

From dataset we know that it is a classification problem and there are two categories which is successor and the other is defaulters.

* Motivation for the Problem Undertaken

As we can see there are different kind of information every recharge done by the user on which kind of recharge user is using mostly and the data service or the main balance the frequency of recharge in 30 day or 90 days.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

The statistical figure I get to know by the df.describe() there are so many information the min max standard deviation 25 percentile the 50th percentile the 75 percentile.

With the help of correlation function, we get to know the correlation of each columns with each other.

From the heatmap we can visualize to see them clearly that they are positive correlated or negative correlated.

The dark side is showing the negative correlation among each other. Lighter side shows the positive correlation among the each other.

**The z-score** function computes the relative **Z-score** of the input data, relative to the sample mean and standard deviation.

* Data Sources and their formats

The dataset provided by Flip Robo is Data file.csv which included 209595 rows and 36 columns.



* Data Pre-processing Done

No-null values are present in dataset but we found some outliers which has to be removed from the dataset, we found 48128 outliers which we removed from the dataset. After that categorical are change to integer & float using **LabelEncoder**.

Then we split the dataset into “X” and “Y” using Standard Scalar. Which transforms the dataset in a distribution which will have a mean value 0 and standard deviation of 1.

* Hardware and Software Requirements and Tools Used

**Hardware:** A good configuration Laptop

**Software:** Python (Jupyter Notebook)

**Library Used:** numpy, pandas, seaborn, matplotlib, warnings, LabelEncoder, StandardScaler, zscore, PCA, KNeighborsClassifier, LogisticRegression, DecisionTreeClassifier, GaussianNB

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

Descriptive statistics are used to describe the basic features of the data in a study which are mean count max standard deviations 25%, 50%, 70% it all help me to understand the data in terms of statistically for the problem solving

* Testing of Identified Approaches (Algorithms)

Below are the Algorithms used for training and testing.

1. KNN=KNeighborsClassifier(n\_neighbors=6)
2. LR=LogisticRegression()

1. DT=DecisionTreeClassifier(random\_state=6)
2. GNB=GaussianNB()

* Run and Evaluate selected models

Describe all the algorithms used along with the snapshot of their code and what were the results observed over different evaluation metrics.

* Key Metrics for success in solving problem under consideration

Model : KNeighborsClassifier

KNeighborsClassifier(n\_neighbors=6)

Accuracy\_Score = 0.967382328654005

Cross\_Val\_Score = 0.9690273546726893

roc\_auc\_score = 0.8961728986896772

Classification\_report

precision recall f1-score support

0 0.96 0.80 0.87 6720

1 0.97 0.99 0.98 41720

accuracy 0.97 48440

macro avg 0.96 0.90 0.93 48440

weighted avg 0.97 0.97 0.97 48440

[[ 5360 1360]

[ 220 41500]]

AxesSubplot(0.125,0.808774;0.62x0.0712264)

Model : LogisticRegression

LogisticRegression()

Accuracy\_Score = 0.9401940545004129

Cross\_Val\_Score = 0.9401294464717204

roc\_auc\_score = 0.8275192753275806

Classification\_report

precision recall f1-score support

0 0.87 0.67 0.76 6720

1 0.95 0.98 0.97 41720

accuracy 0.94 48440

macro avg 0.91 0.83 0.86 48440

weighted avg 0.94 0.94 0.94 48440

[[ 4513 2207]

[ 690 41030]]

AxesSubplot(0.125,0.808774;0.62x0.0712264)

Model : DecisionTreeClassifier

DecisionTreeClassifier(random\_state=6)

Accuracy\_Score = 0.9562964492155244

Cross\_Val\_Score = 0.9578732397522265

roc\_auc\_score = 0.9093994686800896

Classification\_report

precision recall f1-score support

0 0.84 0.84 0.84 6720

1 0.97 0.97 0.97 41720

accuracy 0.96 48440

macro avg 0.91 0.91 0.91 48440

weighted avg 0.96 0.96 0.96 48440

[[ 5675 1045]

[ 1072 40648]]

AxesSubplot(0.125,0.808774;0.62x0.0712264)

Model : GaussianNB

GaussianNB()

Accuracy\_Score = 0.829335260115607

Cross\_Val\_Score = 0.8320069802453123

roc\_auc\_score = 0.8022990572067752

Classification\_report

precision recall f1-score support

0 0.43 0.76 0.55 6720

1 0.96 0.84 0.89 41720

accuracy 0.83 48440

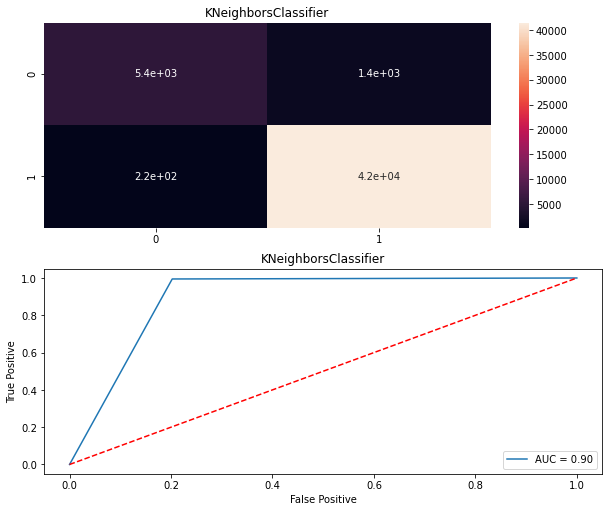
macro avg 0.70 0.80 0.72 48440

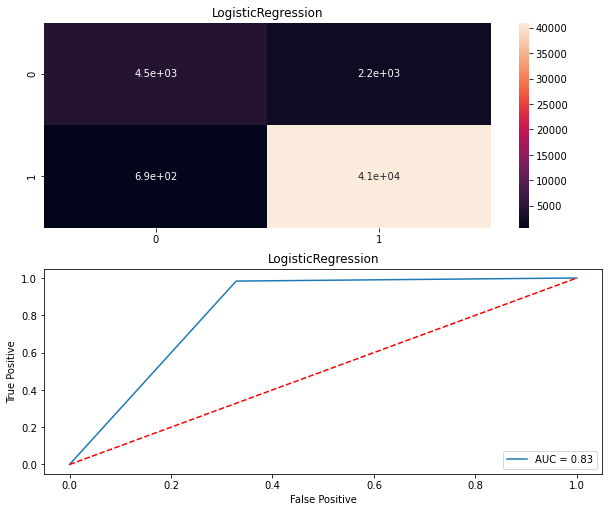
weighted avg 0.88 0.83 0.85 48440

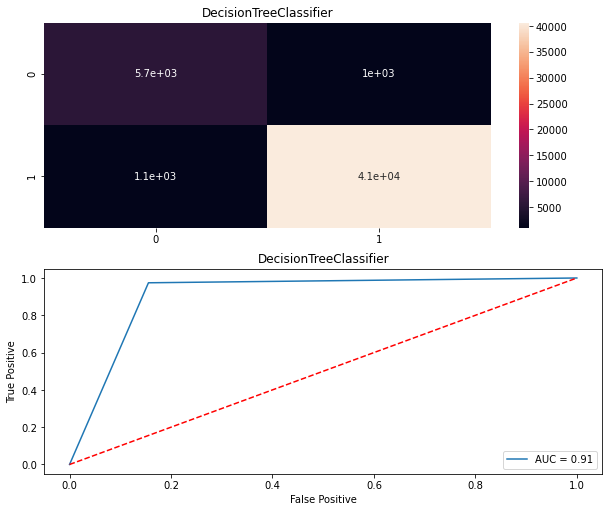
[[ 5140 1580]

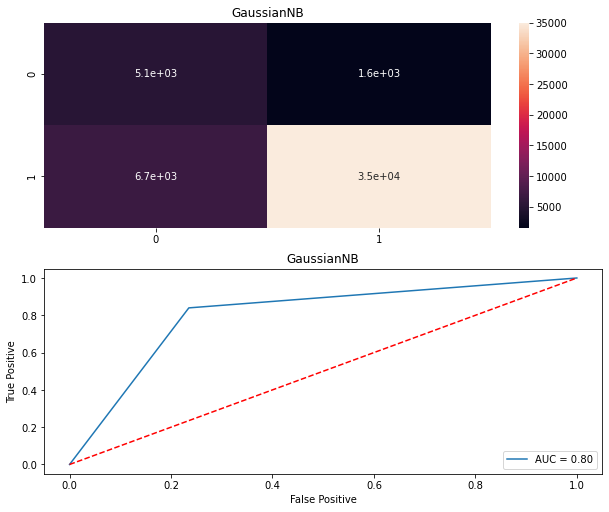
[ 6687 35033]]

AxesSubplot(0.125,0.808774;0.62x0.0712264)

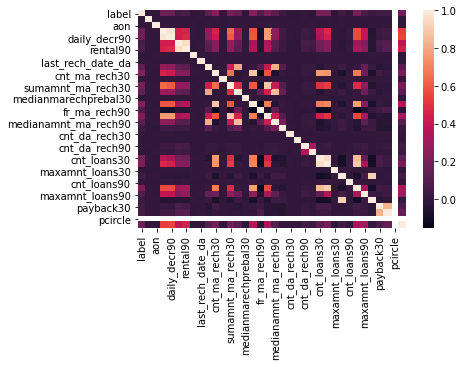




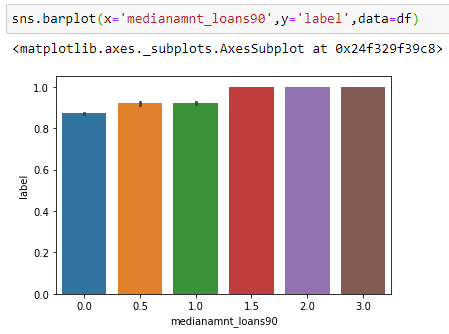
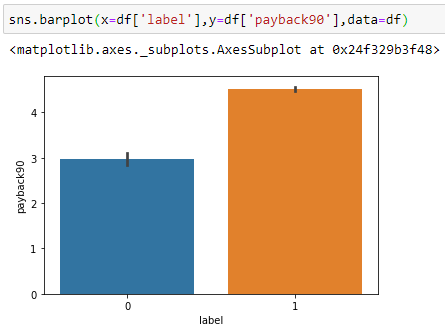
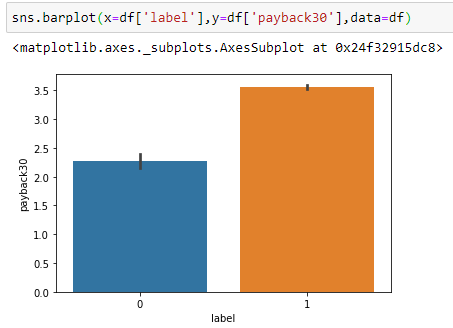
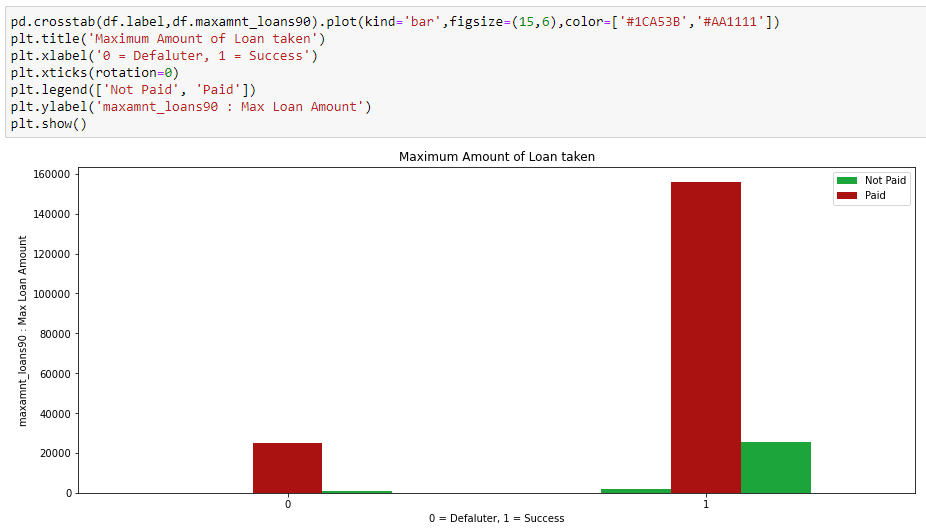
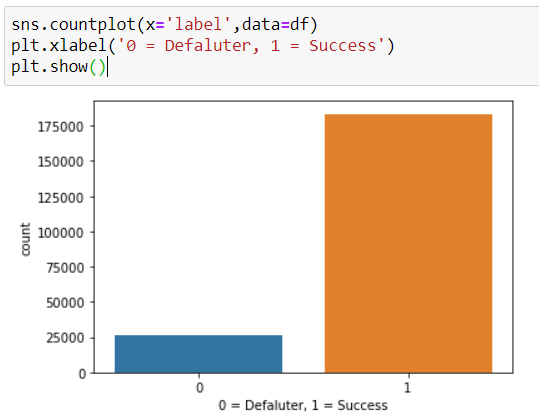
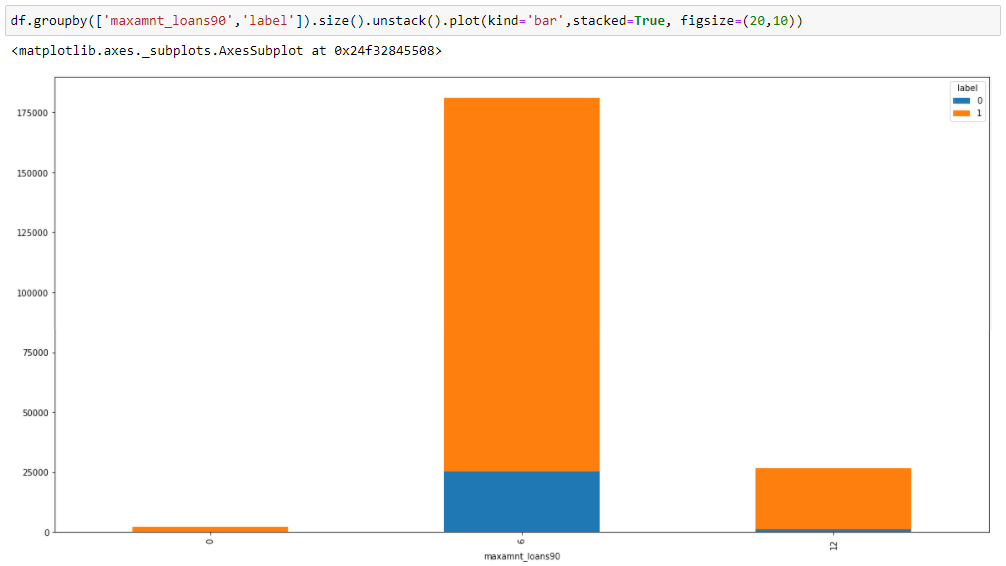




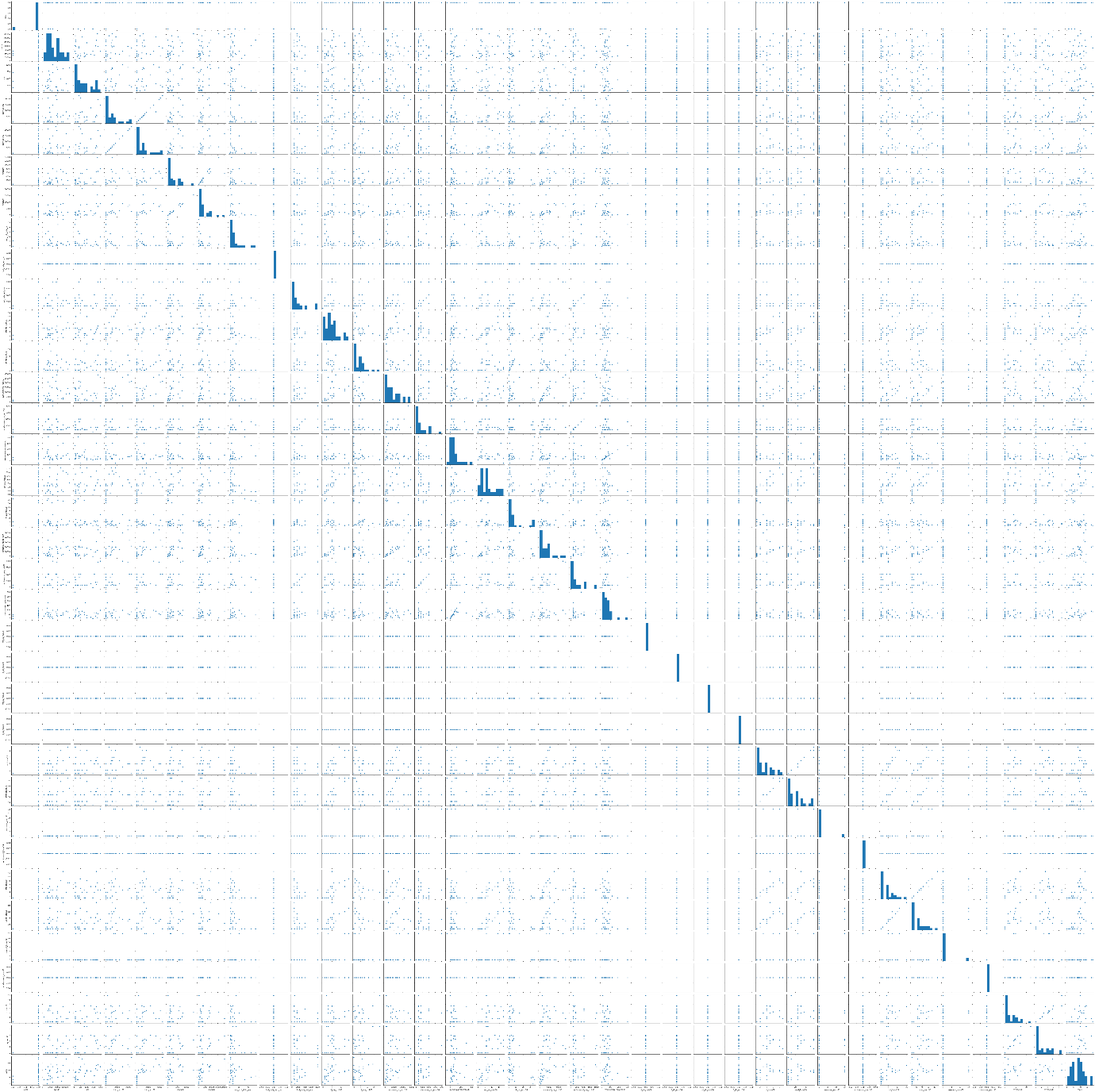
* Visualizations



Using **sns.heatmap(dfcor)** We get to know that the darker side represents the negative correlation and the higher side represent the positive correlation.



**Using Code:** sns.pairplot(df\_new.sample(30))



**CONCLUSION**

* Key Findings and Conclusions of the Study

In the above dataset every features of the dataset plays an important role to understand the data also in visualization and applying models and algorithms.

* Learning Outcomes of the Study in respect of Data Science

Visualization is very helpful and it plays a vital role to understand the data into graphical form. So, that we can understand what the data is trying to say.

Data cleaning is also import part it helps me to remove null, nan or missing values by using mean, median, mode or by simply using 0.

I used four types of algorithm like KNeighborsClassifier, LogisticRegression, DecisionTreeClassifier and GaussianNB. In this dataset the best model works is KNeighborsClassifier.

The problem I faced while on this dataset is using SVC Algorithm and pair plot because of huge rows and columns it takes more than one hour for output. I overcome this situation with the help of towardsdatascience.com by running the code using small chunk of dataset.