

Prediction of Credit Defaulters

Submitted by:

Amey Takle

**ACKNOWLEDGMENT**

The project takes inspiration from research published in Research Journal of Finance and Accounting.

Articles published on site like Vaultz and Vox also helped a lot in formulating the solution.

Study of research done by Subex was done to get an idea of what microcredit business looks like around the world.

**INTRODUCTION**

* Business Problem Framing

The task is to predict on some given parameters, whether a customer will pay the micro-credit availed in a particular time.

This model will help service providers to know what percent of credit they will receive back in the given time.

* Conceptual Background of the Domain Problem

This is a binary classification problem where 0 stands for the customer has failed to pay the credit and 1 stands for customer has successfully paid the credit back.

Visualizing the data WRT to target will help us understand the data better.

As binary classification logistic regression will be the first approach.

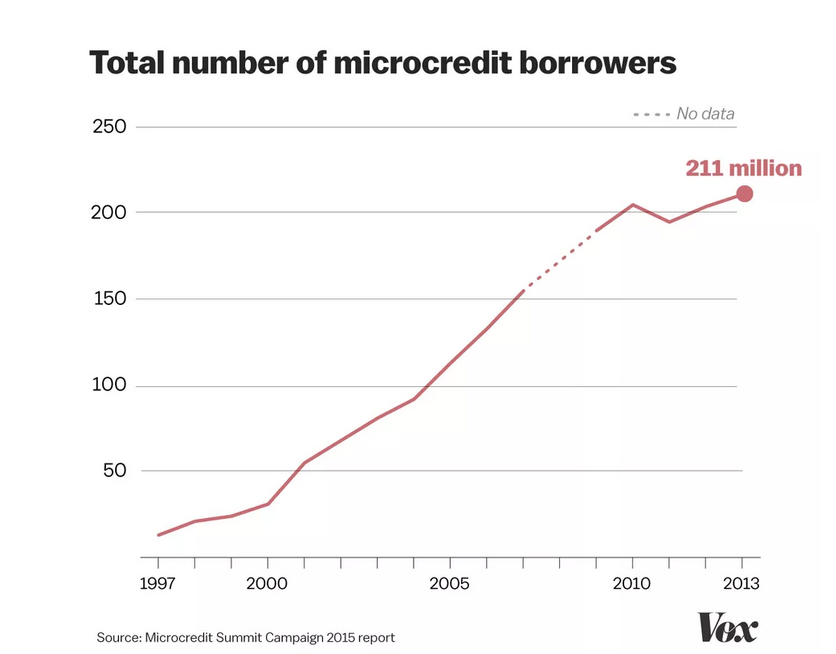
Further we will test other algorithms to check for better accuracy and metrics.

* Review of Literature

Microcredit is the amount borrowed by the customer on a repay duration of 5 days in this problem.

According to SUBEX and VOX the idea of microcredit flourished in 1980s to early 2000s.

The graph following shows the increase in the number of microcredit borrowers



* Motivation for the Problem Undertaken

Objective behind this project is to help Service Providers determine what amount of consumers will pay back the credit in a given amount of time.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

The target variable (dependent) is label which tells us whether the user has paid the credit in given period of time or now.

All the other variables are independently collected data which will help us formulate the target

* Data Preprocessing Done

PCircle and PDate were the first columns to be dropped as PCircle has only one value and PDate is the date the customer pays the microcredit.

Anyways our label is giving us same data (whether the credit was paid back or not)

Also, phone number of the user will not help us in predicting the target hence it was dropped too.

Further checked the correlation of all columns with the label and the least negatively correlated columns were dropped.

Further cleaning was done by dropping the duplicated entries.

The data description revealed that there is a large difference between the 3rd quantile and the maximum, hence it was clear that there were outliers in the data.

Outliers were dropped using the ZScore method.

* Data Inputs- Logic- Output Relationships

The label is the output variable here, rest all are the input variables.

Saw that target is imbalanced, only about 12.5% of negative values while all other are positive values.

* Hardware and Software Requirements and Tools Used

Hardware used: system memory 8GB, Processor: 5th gen core i7

Model is developed on Jupyter Notebook

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

As this is a binary classification, the first approach here was logistic regression.

Further for improved accuracy I have employed classification algorithms like Random Forest Classifier, Switched Vector Classifier.

Other classification algos like KNN Classifiers also can be used.

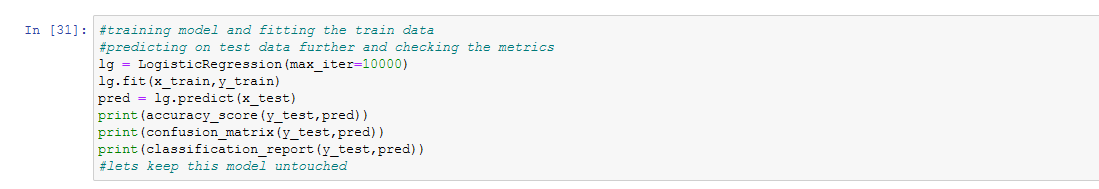
* Testing of Identified Approaches (Algorithms)

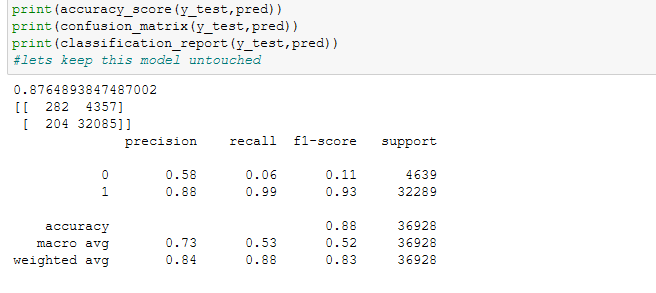
Algorithms used here were

1. LogisticRegression
2. RandomForestClassifier
3. SwitchedVectorClassifier

* Run and Evaluate selected models

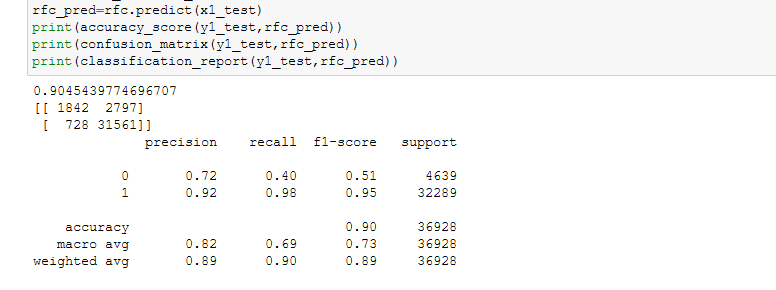
1. Logistic Regression:



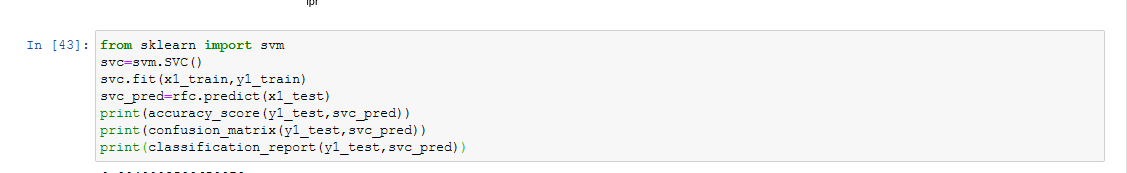


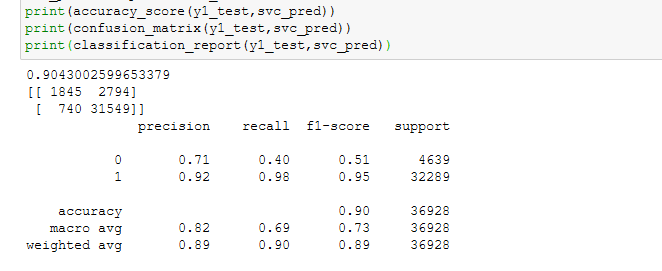
1. RandomForestClassifier





1. SwitchedVectorClassifier





* Key Metrics for success in solving problem under consideration

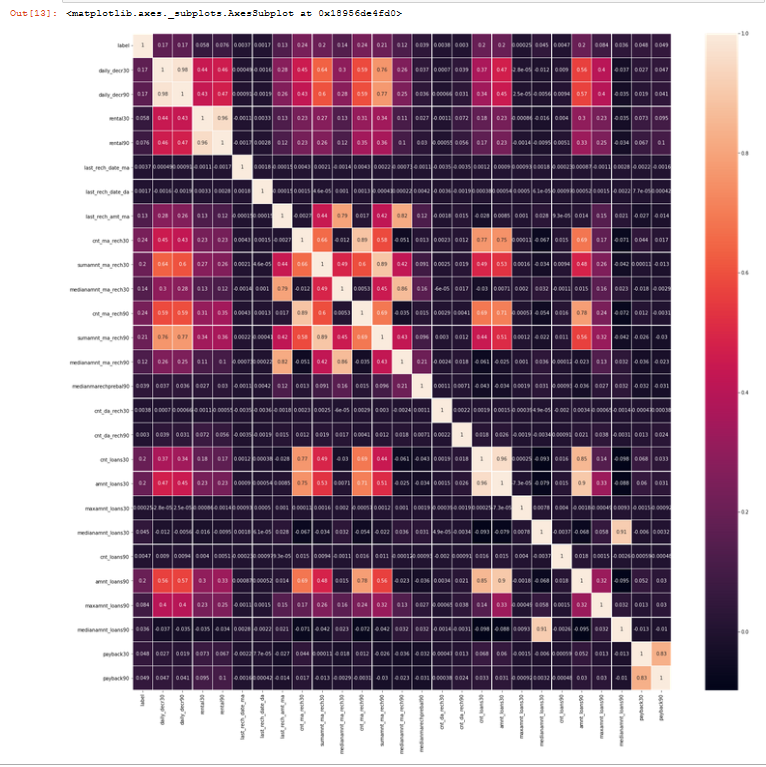
This being a classification problem, the key metrics that are employed here are Accuracy Score, Confusion Matrix.

Classification Report is also employed to get metrics like precision, recall, f1-score

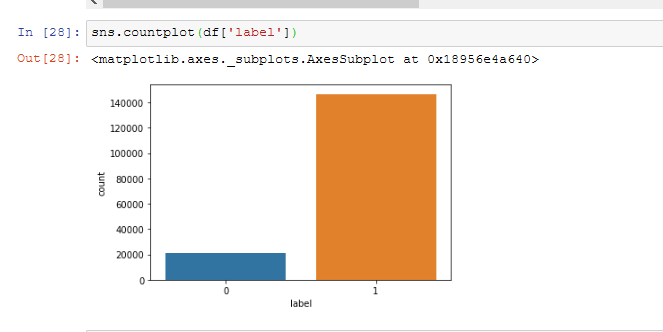
Auc-Roc Score and Auc-Roc Cusrve is also plotter to get ideas of TPR and FPR

* Visualizations

First plot that is done is a heat map showing the correlation between the variables.

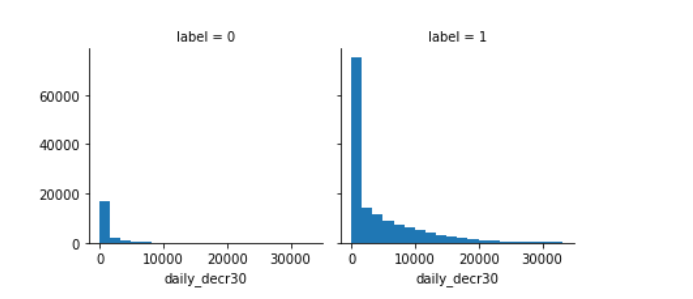


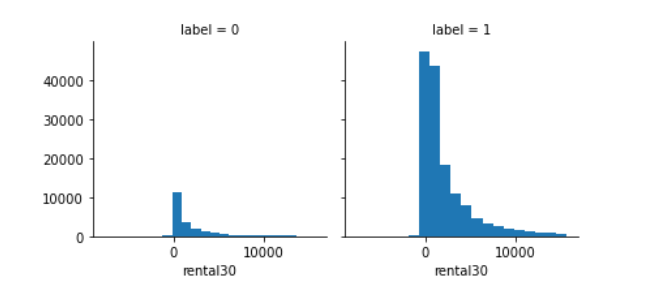
Further a countplot is made on the target to get how the label is distributed.

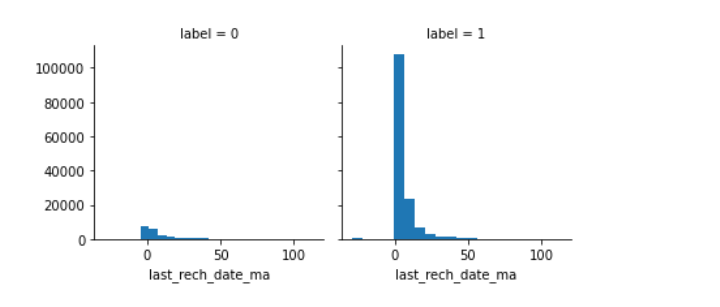


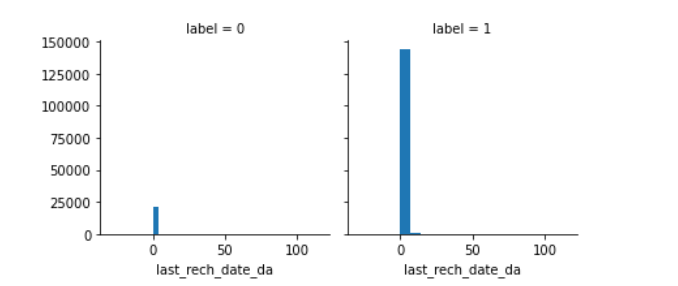
A histogram was plotted to view data in each column which revealed that the data is hugely skewed.

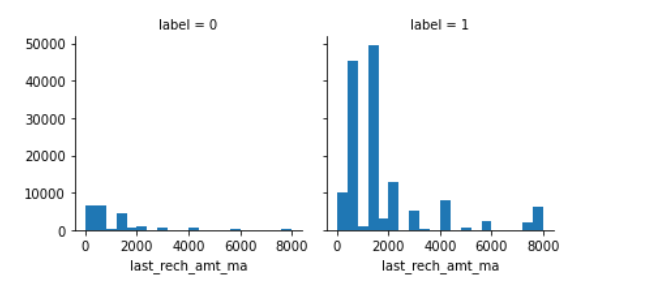
Further each column distribution is checked with respect to label:

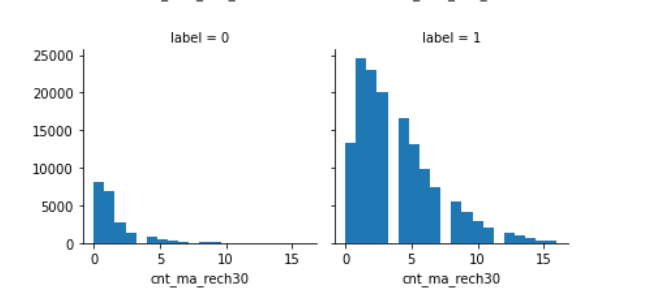


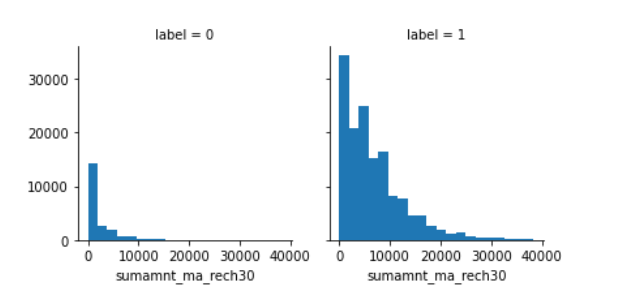


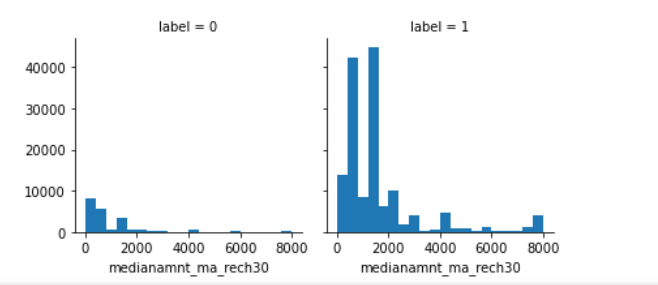


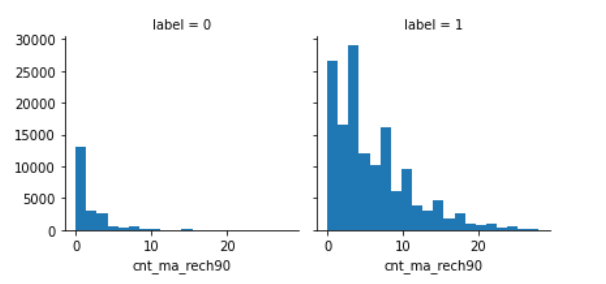


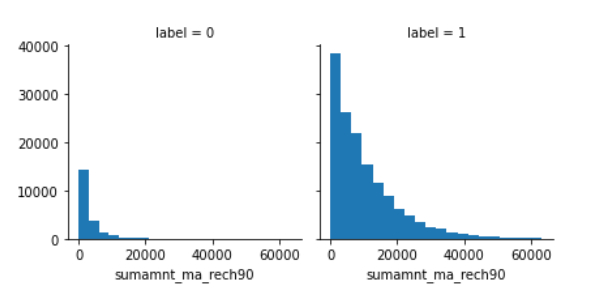


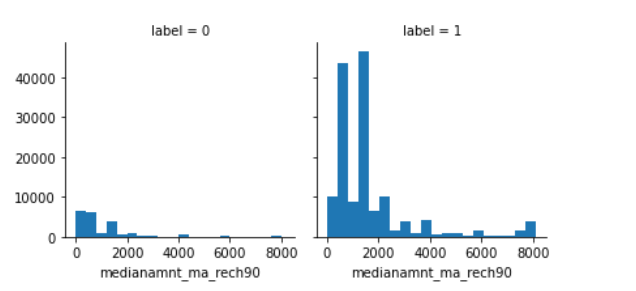


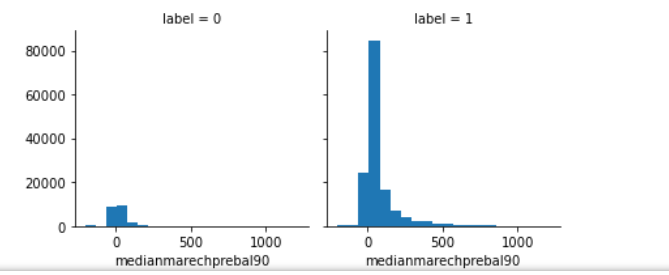


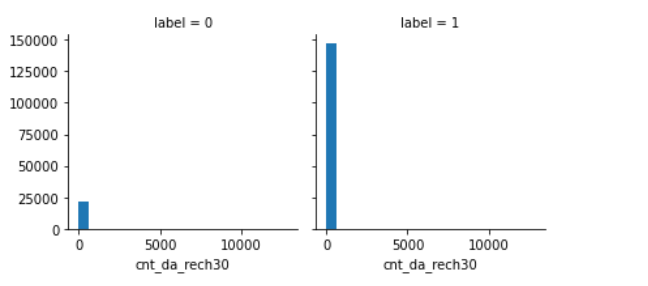


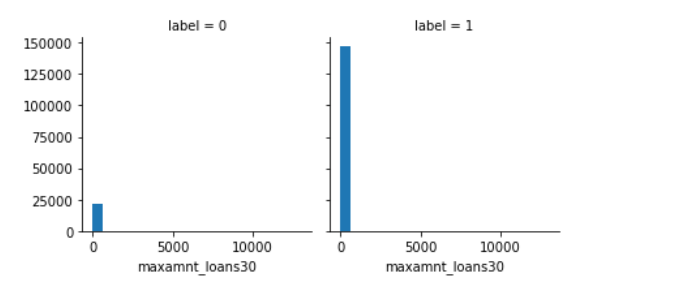
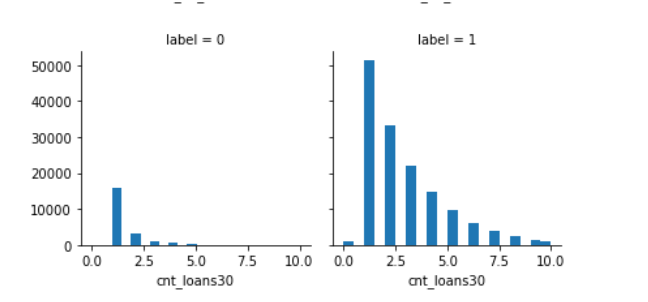












As there are every low records for 0s we can see same trend in all other columns.

* Interpretation of the Visualizations

The countplot revealed the imbalance in the label.

The correlation heatmap gave an idea of which are the features that are highly correlated.

One of these features can be dropped as both with add same value of importance to the target.

Further the distplots were plotted to get an idea of how the data is distributed and it gave an idea about the skewness of the data.

Facetplot help us with the distribution of each feature with respect to label.

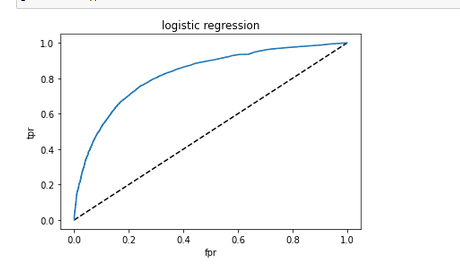
* Results

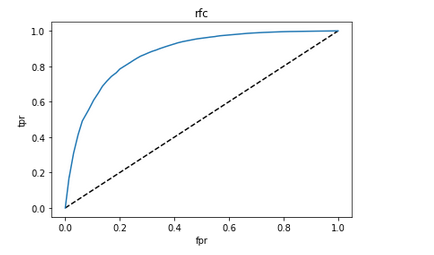
If we observe the accuracy and classification report we can estimate the results of the model.

Applying logistic regression gives us a good accuracy but prediction of 0s is very bad.

If we use Random Forest Classifier accuracy increses and ability of model to predict zeroes accurately also increases.

We need to employ GridSearchCV for hyperparameter tuning to get better results in predicting the zeroes.





Looking at the AUC-ROC curves and AUC score we can infer that the RFC is a better performing model.

**CONCLUSION**

* Key Findings and Conclusions of the Study

Data Cleaning is an important step to develop a better performing model.

Outliers in model will impact the performance of the model to great extent.

There exists people who default the microcredit availed by them.

The number is very less.

Ensembling techniques can be used to improve the performance of the model.

HyperParameter tuning techniques will result in better performance of the model.

Vizualization of data gives us a detailed view of what data actually is.

* Limitations of this work and Scope for Future Work

Limitation of this model is that no hyperparameter tuning could be done as it was taking a lot of time(even after running for about 2 hours no outcome was obtained)

Improvement on same must be done to ensure a better performing model is presented