

**Micro Credit Loan**

**Submitted by:**

**Ajay Sonawane**

**ACKNOWLEDGMENT**

This includes mentioning of all the references, research papers, data sources, professionals and other resources that helped you and guided you in completion of the project.

**INTRODUCTION**

**• Business Problem Framing**

We have a dataset from MFI(Microfinance Institution) which contains loan related data and we have to predict whether the customer is defaulter or non defaulter.

In real word by recognising default customers financial assistance can be provided by MFI in form of mobile recharge.

**• Conceptual Background of the Domain Problem**

It will be helpfull to have basic idea of how potential customer for financial assistance are selected. Like how banks check CIBIL score before approving any loan and the loan amount can very based on the credit score and banking history of customer.

**• Review of Literature**

MFI is an organization that offers financial help to low income people.It mainly focuses on non banking customers who are living in remote areas.MFI have better reach and are cost effective when compared to tradition financial instituions.

Till now MFI have has given $70 billion as outstanding loan to customers and have 200 million clients.

Similary financial assistance can be provided based to transaction history of customer.MFI are working with a Telecome company where they want to help financially weak customers by providing mmico credit in form of recharge.

The consumer is belieed to be defaulter if he does not pay back the credit loan within 5 days. Same data is provided to MFI for analysis so that they can detect the default customers and provide them micro credit.

**• Motivation for the Problem Undertaken**

This project will help MFI to find the needy consumers and will help them with micro credit which will help them financially.

**Analytical Problem Framing**

* **Mathematical/ Analytical Modeling of the Problem**

This is a Classification problem hence we have used regression models.Data is imbalanced so we tried to balance it using SMOTE method.Also there are many outliers which we tried to remove using Zscore.

No null values present in dataset.Few columns were of Object datatype which we have converted to int using label encoder.After all the Preprocessing and EDA done we devided data in X and Y and passed it to different models.

We have also hypertuned our best model to increase our accuracy.

**• Data Sources and their formats**

Dataset is csv format file.Dataset have total 209593 rows and 37 columns.

Three columns are object type and rest all are int datatype columns

**• Data Preprocessing Done**

1.First checked for the Null values if any.There were no null values in dataset.

2.Checked for data type of columns and found 3 columns data type as object which we converted to int using label encoder.

3.Removed outlier using zscore

4.used standard scalar to normalise the data in same range.

**• Data Inputs- Logic- Output Relationships**

Some of the important columns are cnt\_loans90,amnt\_loans90,payback30 and payback90.In these columns we can see how much loan customer has taken and what was he`s repayment time.Which gives us the idea about financial condition of customer.

If customer is taking too much loan and not paying back in time then it will affect our output which is defaulter/non defaulter

**• State the set of assumptions (if any) related to the problem under consideration**

Here, you can describe any presumptions taken by you.

**• Hardware and Software Requirements and Tools Used**

We need Python and jupyter notebook installed in system.

we will use libraries like numpy,pandas for accessing the and different operation on dataset.

seaborn and matplotlib for visualisations,

sklearn is used for all the algorithams which we used

imblearn is used for balancing the imbalanced dataset.

**Model/s Development and Evaluation**

**•Identification of possible problem-solving approaches (methods)**

There are 209593 rows and 37 columns in dataset.There is no null value present in dataset.Checked the data type of all columns and found there are 3 columns with object datatype which I

converted to int type because our model will not work on object data type columns.Checked for the correlation of columns using the heatmap and found daily\_decr30,daily\_decr90,rental30,rental 90,

last\_rech\_data\_da,cnt\_ma\_rech90,amnt\_loan90,payback90 are highly correlated where as medianmarechprebal90.maxmnt\_loan30,unnamed,msisdn are least correlated columns.

Columns like aon,

daily\_decr30,daily\_decr90,last\_reach\_data\_ma,last\_rech\_date\_da,sumamnt\_ma\_rech30,fr\_ma\_rech30,medianamnt\_ma\_rech30 have many outliers which we can remove using zscore method.

Also data was imbalaced so we used SMOTE method to balance our dataset

Once data is pre processed we gave it to our model for train and testing.

**• Testing of Identified Approaches (Algorithms)**

I used Logistic Regression,Random Forest Regressor,GaussianNB and Support vector machine regression

**• 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.

Logistic Regression :

Accuracy of 66.59%

GaussianNB :

Accuracy of 53.67%

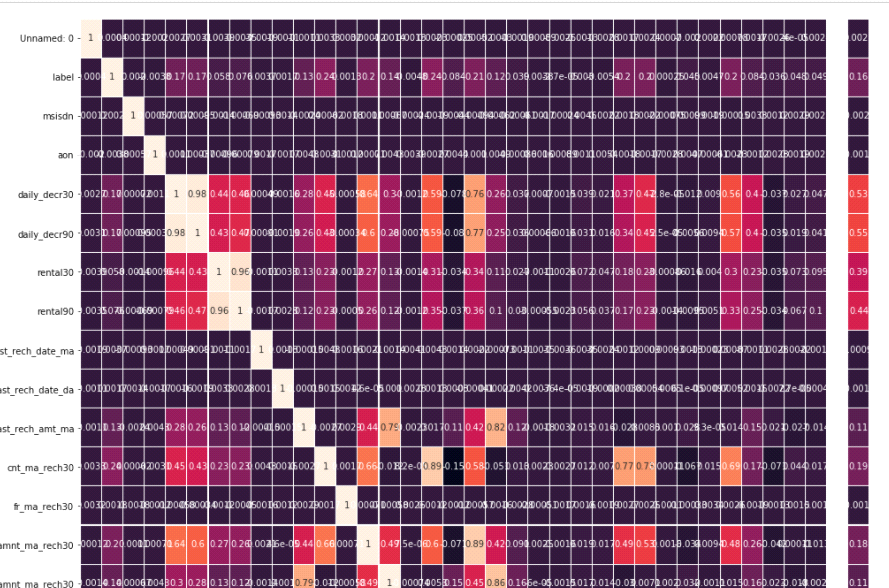
Random Forest Regressor :

Accuracy of 91.31%

**• Visualizations**

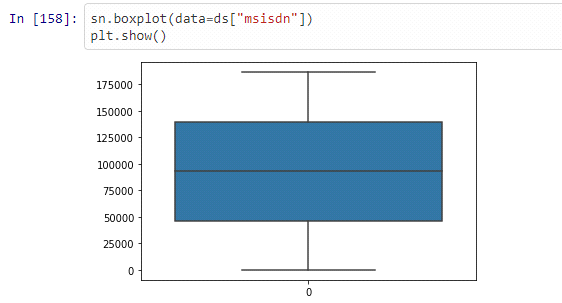
Mention all the plots made along with their pictures and what were the inferences and observations obtained from those. Describe them in detail.

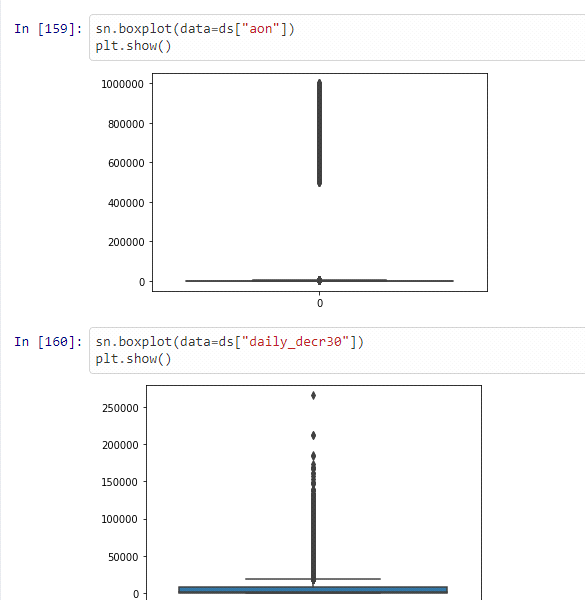
If different platforms were used, mention that as well.

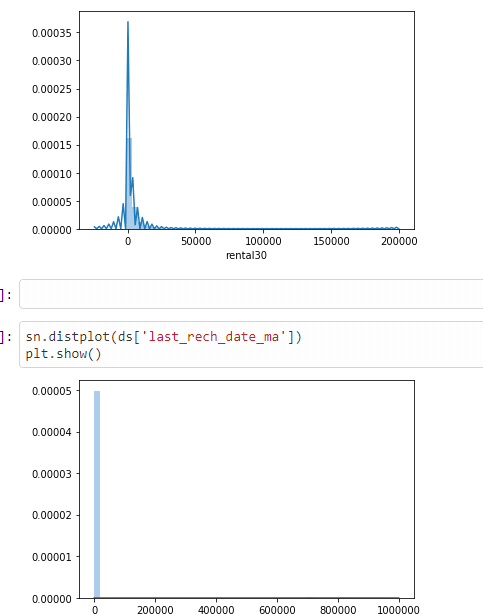


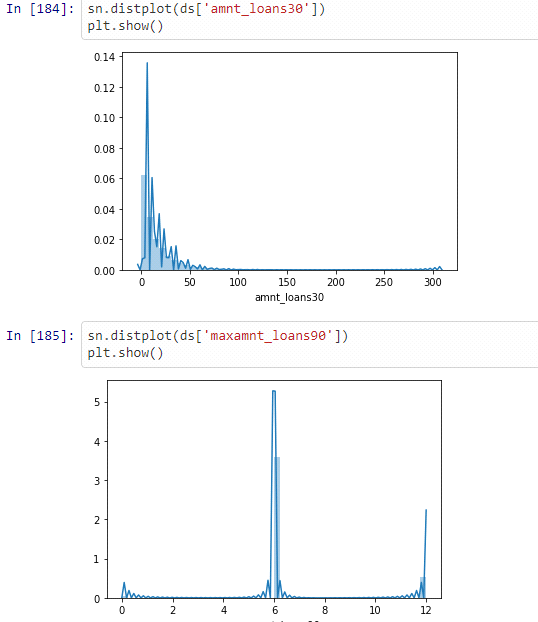
Checked for the correlation of columns using the heatmap. found daily\_decr30,daily\_decr90,rental30,rental90,last\_rech\_data\_da,cnt\_ma\_rech90,amnt\_loan90,payback90 are highly correlated where as medianmarechprebal90.maxmnt\_loan30,unnamed,msisdn are least correlated columns.





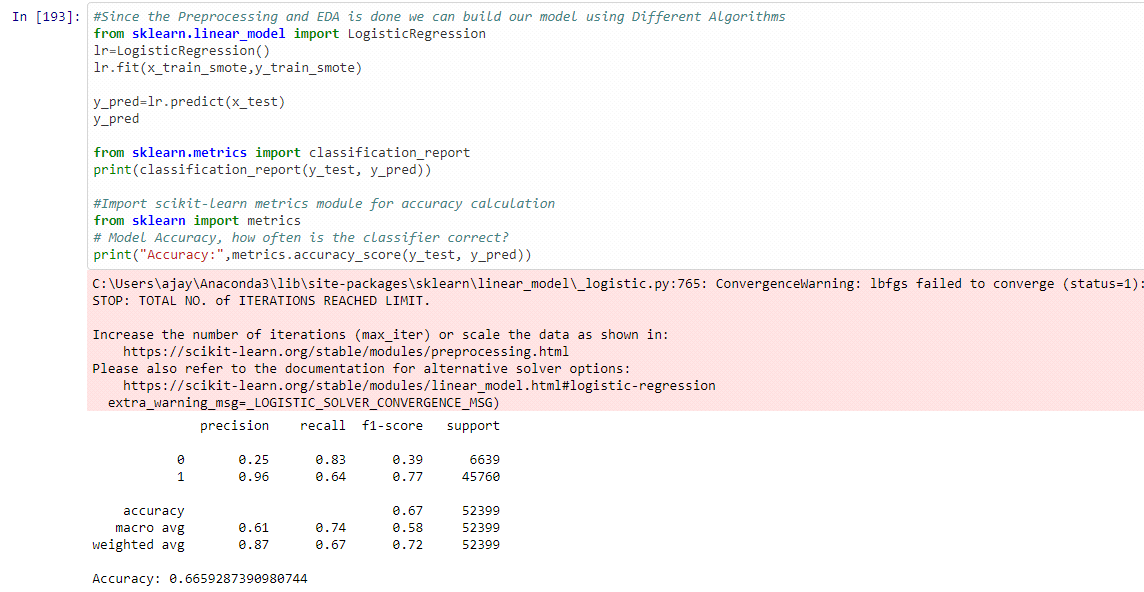


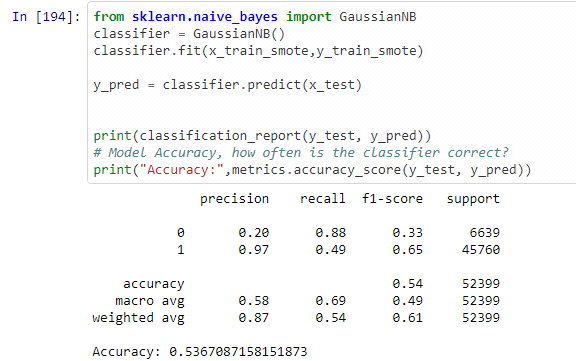


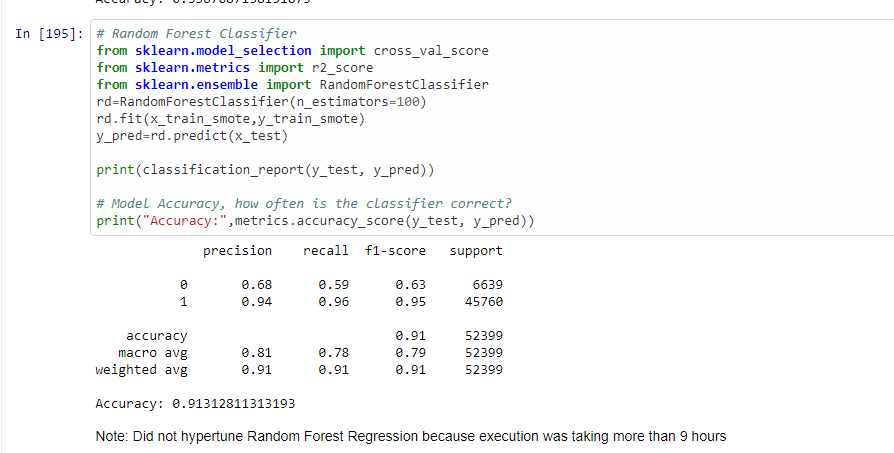


**• Interpretation of the Results**

Give a summary of what results were interpreted from the visualizations, preprocessing and modelling.







We have used LogisticRegression, GuassianNB and Random Forest Regression.

Accuracy for all the models is-

Logistic Regression Accuracy 66.59%

GaussianNB accuracy 53.67%

Random Forest Regressor accuracy 91.31%.

We got the best result with randome forest with accuracy of 91.21%

**CONCLUSION**

**• Key Findings and Conclusions of the Study**

Describe the key findings, inferences, observations from the whole problem.

Our target variable was "label" column and it was imbalanced.

total record for 1:137671 and 0: 19523. We balanced it using SMOTE.

There are outliers in columns aon,daily\_decr30,daily\_decr90,last\_reach\_data\_ma,last\_rech\_date\_da,sumamnt\_ma\_rech30,fr\_ma\_rech30,medianamnt\_ma\_rech30.

We converted object datatype columns to int type.

We used Logistic Regression,GuassianNB and Random Forest Classifier. We recieved best accuracy score with random forest model.

**• Limitations of this work and Scope for Future Work**

What are the limitations of this solution provided, the future scope? What all steps/techniques can be followed to further extend this study and improve the results.

Could not remove the outliers because all the columns were getting removed after applying zscore.