

NAME OF THE PROJECT

Micro Credit Project

Submitted by: Surendra Singh Devda

YOUR NAME

**ACKNOWLEDGMENT**

Medium, geeksforgeeks, Kaggle and stack over flow were some of the websites that helped in completion of this project.

**INTRODUCTION**

* Business Problem Framing

The project was related to the micro credit loans which are being provided to small businesses and farmers or shopkeepers maybe and by making such model it would help us predict whether or not a person is liable to be given loan and so that there may be no future frauds of loans.

* Conceptual Background of the Domain Problem

Manually seeing the file records of the person all together and matching different criteria for the person becomes so exhausting for any bank and individual and even we may miss some records and even we might fail in actually predicting whether or not the person would Pay a loan and we would end ourselves in some fraud so better why not use the technology for our betterment.

* Review of Literature

This is a comprehensive summary of the research done on the topic. The review should enumerate, describe, summarize, evaluate and clarify the research done.

The model need to be made on few important aspects, most important of them being cleaning the data which is most important,

We cleaned the data, but before that we use label encoder to convert the categorical data into numerical data.

We need to check for the null values as well, if any then convert them into some values by using mean function.

We will check for the outliers and remove them, plot them by using boxplot and remove them using zscore.

Skewness is other important parameter , we need to remove the skewness from the data, if any is present.

And then we need to do scaling so that whole data has a same unit.

We are going to use logistic regression initially as it is used for the discrete data.

We also do resampling as the data is more biased towards giving loan that is 1 and this can lead to false model building although the accuracy score may be good but the prediction would not be valid enough. We can get to know that by recall value.

We are going to use SMOTE function for the oversampling.

* Motivation for the Problem Undertaken

Now days there are too many frauds all going on and if country like India which is a developing country losses a huge money due to such frauds then this will keep India from becoming a developing nation more slow and even downgrade it.

So by using technology driven things to safe guard our money is a better option.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

We have used zscore to remove the outliers, zscore is basically a threshold value that is given, it shows how far is data away from the mean of the data and is therefore an outlier.

Zscore= datapoint- mean/ standard deviation

We have plotted the outliers from boxplot, we can use either matplotlib or seaborn library to plot boxplot.

We dropped pcircle column as it had no such important significance and the pcircle column showed some NAN values so we had remove it as it was showing some problem in zscore.

We have seen different bar graph to understand whether or not we require to do resampling as we saw in it was highly recommended to do oversampling as undersampling would lead to loss of data, so we used SMOTE technique to oversample.

Then we used logistic regression into model as the data has discrete values and logistic regression is best preferred for the discrete values.

* Data Sources and their formats

What are the data sources, their origins, their formats and other details that you find necessary? They can be described here. Provide a proper data description. You can also add a snapshot of the data.

* Data Preprocessing Done

Firstly there was no null data.

Then we went on to remove the outliers by using the zscore and any data having standard deviation more than + and – 3 are to removed.

We tried to remove the skewness in the data but log and sqrt and boxcox all the function showed NAN and null values so we could not skew much of a data.

* Data Inputs- Logic- Output Relationships

Describe the relationship behind the data input, its format, the logic in between and the output. Describe how the input affects the output.

There we many such columns which affected the yes and no of the output the target column which was label to give the right decision.

Like daily amount paid for last 30 days etc.

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

Assumptions were not any as such, still we could say that SMOTE analysis which was oversampling could be considered as an assumption.

* Hardware and Software Requirements and Tools Used

Hp, 8GB ram, libraries like pandas, numpy, matplotlib, scikit learn, imblearn etc were used for the model building along with many other algorithms like logistic regression, KNN, Decision Tree classifier etc.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

Describe the approaches you followed, both statistical and analytical, for solving of this problem.

* Testing of Identified Approaches (Algorithms)

Listing down all the algorithms used for the training and testing.

Logistic Regressioin

KNN

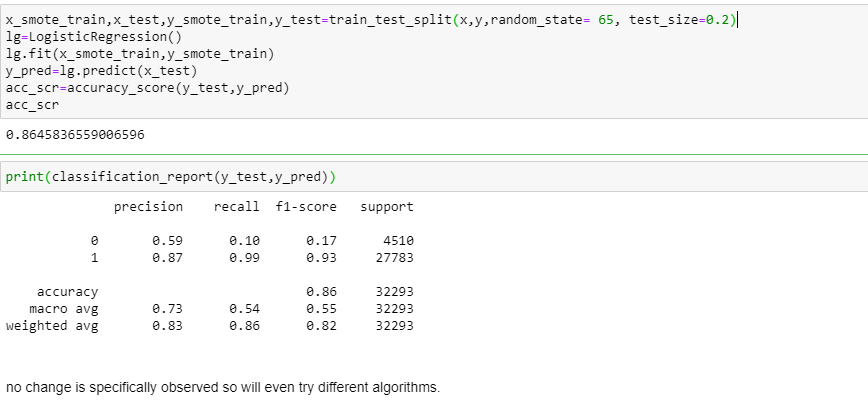
Decision Tree classifier

Gussian NB

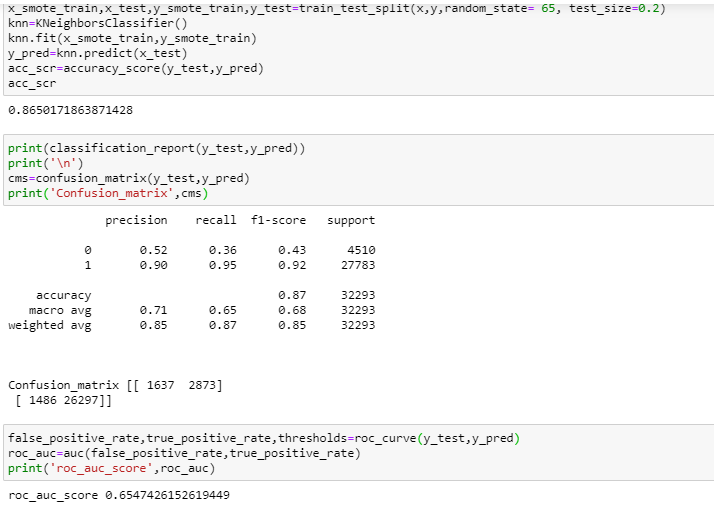
* Run and Evaluate selected models

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

Logistic Regression



KNN



* Key Metrics for success in solving problem under consideration

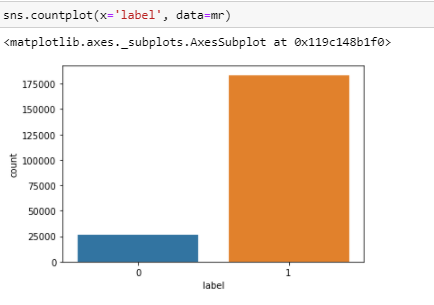
Decision Tree Classifier was a key matrix which was used as a final model for prediction as well.

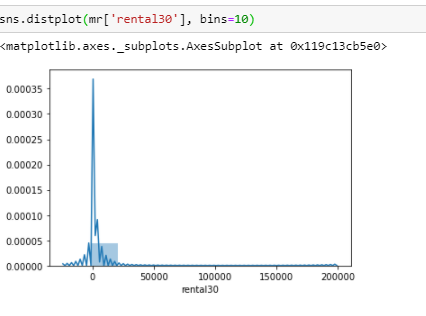
* Visualizations

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

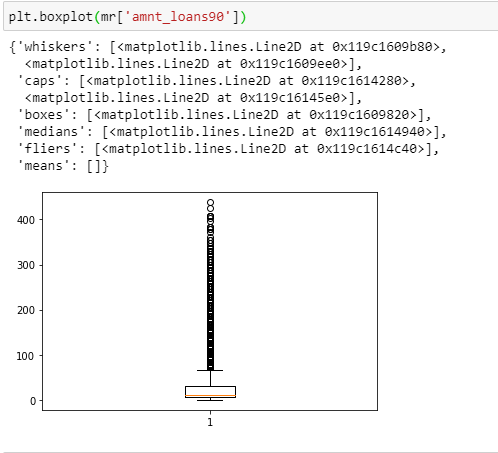
If different platforms were used, mention that as well.

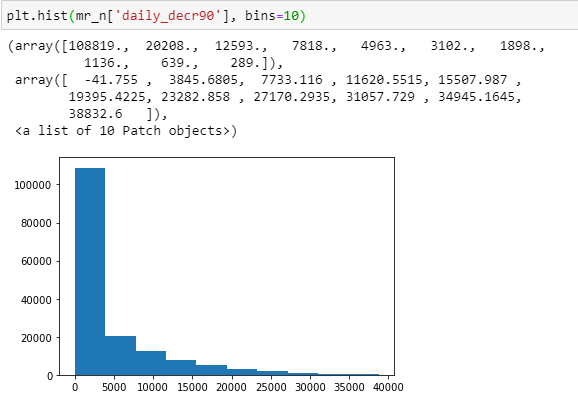
This showed us the need of the resampling so we used SMOTE technique .



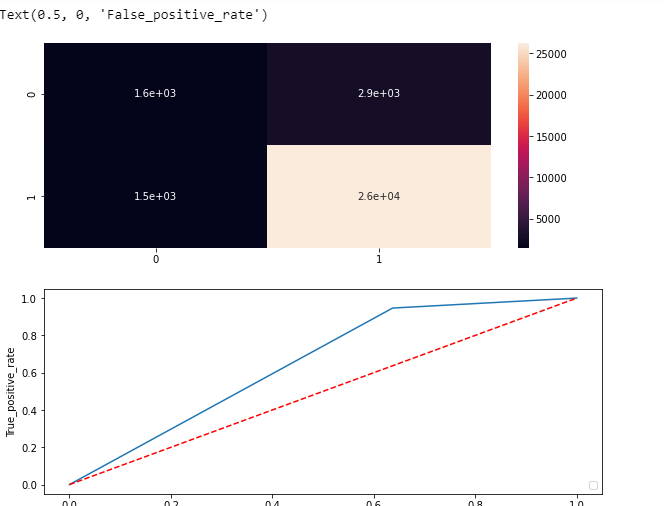


This below showed the outliers.

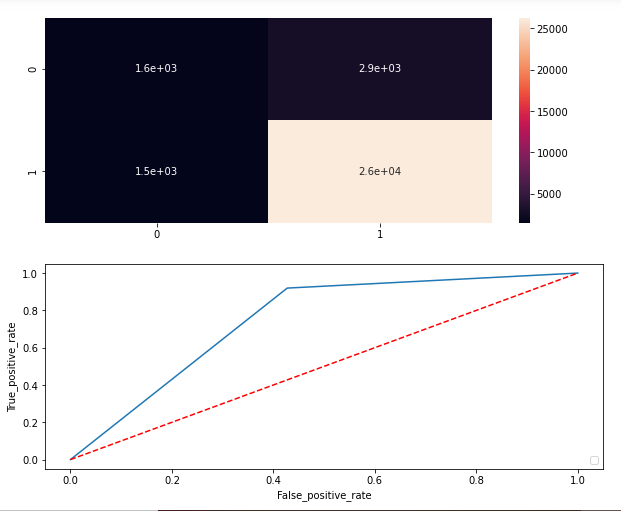




The below diagram is showing us the area under the curve for KNN.



The below diagram is showing us the area under the curve for DTC.



* Interpretation of the Results

After using SMOTE we saw there was not much of a change in recall value of 0 it was 0.1 in logistic regression then we tried different models and we ended getting best results into decision tree classifier where the accuracy score and the RUC AUC score was also better than all other algorithms used it had accuracy score 87% and RUC AUC score was 74.5%, although we even noticed Gussian NB could be better as the recall value for 0 was 83% and for 1 was 63% although accuracy score was less 63% and AUC RUC score is 73 %, so we selected DTC for better scores.

**CONCLUSION**

* Key Findings and Conclusions of the Study

we saw there was not much of a change in recall value of 0 it was 0.1 in logistic regression then we tried different models and we ended getting best results into decision tree classifier where the accuracy score and the RUC AUC score was also better than all other algorithms used it had accuracy score 87% and RUC AUC score was 74.5%, although we even noticed Gussian NB could be better as the recall value for 0 was 83% and for 1 was 63% although accuracy score was less 63% and AUC RUC score is 73 %, so we selected DTC for better scores.

* Learning Outcomes of the Study in respect of Data Science.

Most important learning was there was some continuous issue with the zscore duplicating the values and I could not figure out what was going at end I got to know it was because of pcircle which had just single value and which of because it was creating null values so had to remove it later.

* Limitations of this work and Scope for Future Work

Codes ran very slow for the first time, it never happened such, but for the first all the codes did not run, even after waiting whole night.