A Model to Predict Loan Defaulters using Machine Learning

R. B. Saroo Raj [1], Gurpartap Singh [2], Balaji S [3], K. H. Ajit Baskar [4]
[1] Asistant Professor, Computer Science Department, SRM Institute of Science and Technology, Chennai,India
[2, 3, 4] Computer Science Department, SRM Institute of Science and Technology, Chennai,India

Abstract – With the upgrade in the managing an account part bunches of individuals are applying for bank loans however the bank has its restricted resources which it needs to give to constrained individuals just, so discovering to whom the loan can be allowed which will be a more secure alternative for the bank is a regular procedure. So, in this paper we endeavour to lessen this hazard factor behind choosing the protected individual to spare heaps of bank endeavours and resources. This is finished by mining the Data of the past records of the general population to whom the loan was allowed previously and based on these records/encounters the machine was prepared utilizing the machine learning model which give the most precise outcome. The primary target of this paper is to anticipate in the case of allotting the loan to specific individual will be protected or not. This paper is separated into four segments (i)Data Collection (ii) Comparison of machine learning models on gathered data (iii) Training of framework on most encouraging model (iv) Testing.

Index Terms – loan, machine learning, data set, prediction.

1. INTRODUCTION

Conveyance of the loans is the core business part of every banks. The principle partitions of bank's profit is straightforwardly originated from the benefit earned from the loans distributed by the banks. The prime goal in managing an account domain is to put their benefits in safe hands where it is. Today numerous banks/monetary organizations favour loan after a relapse procedure of confirmation and approval yet at the same time there is no surety whether the picked candidate is the meriting right application out everything being equal. Through this framework we can anticipate whether that specific candidate is protected or not and the entire procedure of approval of highlights is computerized by machine learning method. The burden of this model is that it underscores diverse weights to each factor yet in genuine at some point loan can be affirmed based on single solid factor just, which isn't conceivable through this framework. Loan Prediction is extremely useful for representative of banks and also for the candidate moreover. The point of this Paper is to give brisk, quick and simple approach to pick the meriting candidates. It can give extraordinary preferences to the bank. The Loan Prediction System can consequently figure the heaviness of every element participating in loan handling and on new test data same highlights are prepared as for their related weight. A period breaking point can be set for the candidate to check whether his/her loan can be authorized or not. Loan Prediction System enables hopping to particular application with the goal that it very well may be keep an eye on need premise. This Paper is solely for the overseeing expert of Bank/fund organization, entire procedure of expectation is done secretly no partners would have the capacity to modify the handling. Result against specific Loan Id can be send to different division of banks so they can make proper move on application. This encourages all others office to did different customs

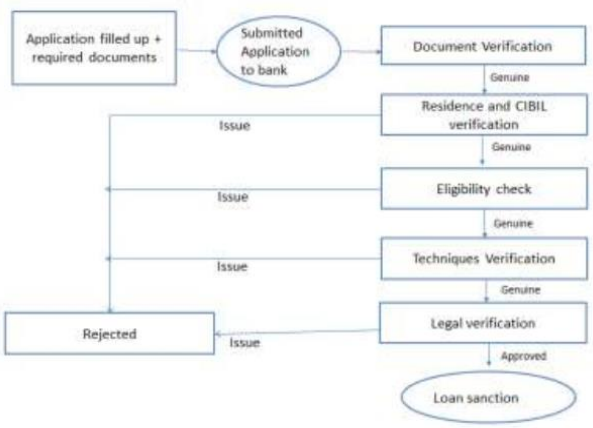

Figure 1 Process of Loan Sanction

2. RELATED WORK

Many researches have been conducted based on data mining and data analysing in the field of financial and banking sector. This section presents briefly some of these techniques which are used in loans management and their finding Sudhakar et al focused on specifying the data mining applications usefulness, these applications are using several machine learning algorithms such as decision trees and Radial Basis Neural Networks. This study came with in which way to apply these applications in a loan approval assessment field. McLeod presents Neural networks properties and their fitness for the credit granting process. [1]

Barney et al made a comparison of the performance of prediction algorithms to identify the farmers who will default on the loans of their Home Administration and those farmers who return back the credits as in the appointment. By using an

unstable data, this study proofed that neural networks regarding better logistic regression to classify farmers into two groups, those who pay back on time and those who default to return their loans [2].

3. PORPOSED MODELLING

We mean by loan assessment process, the grouping of steps that are taken to decide about giving a loan to the client or not. At the point when the client applies for a loan conceding application, the bank officer must research about what called 5 C's which are Character (or Credit History), Cash Flow (or Capacity), Collateral, Capitalization and Conditions. It is useful for assessment loan application and it viewed as a supportive system for gauge the credit hazard identified with a plausible bank.

4. DATA SET

The raw data set contains 75 fields for each loan begun. In any case, not the majority of the fields are helpful for our learning models, for example, the loan ID and the month in which last instalment was received, and in this way we evacuated such fields. We additionally evacuated fields for which more noteworthy than 10% of the loans were missing data for. Clear cut highlights, for example, address state (for instance, California), were ventured into Boolean sections, one segment for each particular esteem that the highlight could take. At long last, we evacuated any loans that were missing data for any field (around 3% of the loans in our dataset). To name the dataset, we characterized any loan that defaulted, were charged off, or were late on instalments was ordered as negative precedents, while we arranged any loan that was completely paid or current was named positive precedents.

No	The attribute	Description	Data type
1	Credit_history	Previous history of customer credit	Nominal
2	Purpose	The loan purpose	Nominal
3	Gender	Male or female	Nominal
4	Credit_amount	The amount of credit	Numeric
5	Age	Customer Age	Numeric
6	Housing	Rent, own or for free	Nominal
7	Job	Is the customer has a job	Nominal
8	Class	The class of loan good/bad	Nominal

Table 1 Data set Description

5. MODEL IMPLEMENTATION

The process of classification crowds the data set into groups of classes according to their similarity. There are several classification algorithm or classifiers like Naïve Bayes Classifier, Neural Network Classifier, decision Tree Classifier. There are several algorithms in each of this technique which used to produce a model to predict the class of unknown class tables. The major goal of this algorithm is the provision by a model for predicting the class of unknown records

Every classifier algorithm consists of these few steps:

- Prepare the training set, a record that are already have known class label
- Build model by applying one of learning algorithm to learning data set
- Apply model to unknown test data set
- Evaluate the accuracy of model

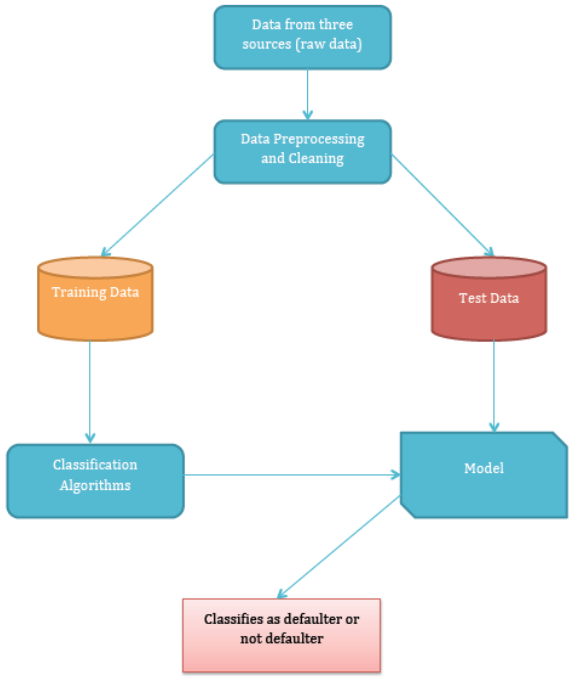

Figure 2 Workflow Diagram

In our research we use three different classification algorithms to build two different models. These algorithms are j48 decision tree, and naïve Bayes. Original data set has been divided into two parts of 20% and 80%. These are used for testing and training data respectively.

a. *J48 Classification Algorithm* -J48 is the upgraded version of C4.5 algorithm or can be seen as C4.5 implementation. J48 takes as an info the arrangement of tables and produce a decision tree as a yield. The created decision tree is indistinguishable to the structure of tree. It comprises of root, halfway and leaf hub. The hubs in the created tree contain a decision which manual for the outcome. It split the information data set into totally unrelated sets, each set with a name. Part measure is connected to figure out which credit prompt the ideal part like data gain foundation [3]

b. *Naïve Bayes*- The Bayesian is a supervised learning method. It portrayed with it is style,

straightforwardness, and robustness. Therefore, it is turned out to be broadly utilized in prediction or classification purposes. It guesses that properties of a class are self-determining in real life. [4]

6. RESULTS AND DISCUSSIONS

Naive Bayes, we acquired the outcomes from the two analyses in table 2 and in the wake of looking at the effectively grouped occurrence percent we find that the best algorithm for loan arrangement is j48 algorithm. J48 algorithm is best since it has high exactness and low mean supreme mistake as appeared in the outcome. Additionally, it is able to characterize the cases effectively than alternate strategies. Perplexity grid of the two algorithms demonstrated that the j48 algorithm is the best one. The tests have been completed a few times and, in each time, the preparing and test sets estimate have been changed (80% preparing 20% test set,60% preparing 40% test and 70% preparing 30% test) and we acquire a similar outcome which is J48 algorithm is best in arranging loans to great and terrible loan. this model help bank administrator to acknowledge or dismiss loan applications by anticipating that if the exchange will lead bank to chance or not and bolster decision producer to settle on a compose decisions.

Techniques	**Correctly classified instance percent**
J48	78.3784%
Naïve Bayes	73.8739%

Table 2 Results from algorithms

5. CONCLUSION

In this paper, two algorithms - j48, and naive Bayes algorithms were used to build a predictive model that can be used to predict and classify the applications of loans that introduced by the customers to good or bad loan by investigating customer behaviors and previous pay back credit. The model has been implemented by using python and machine learning. After applying classification's data mining techniques algorithms which are j48, and naive Bayes, we find that the best algorithm for loan classification is j48 algorithm. J48 algorithm is best because it has high accuracy and low mean absolute error

ACKNOWLEDGEMENT

This work was supported by computer science department of SRM Institute of science and technology, Chennai. All faculties of department helped us in making this paper. We are greatly thankful to all of them. We would like to thank Mr. R. B. Saroo Raj for guiding us in this project.

REFERENCES

[1] Ogawa, Ms Sumiko, et al. Financial Interconnectedness and Financial Sector Reforms in the Caribbean. No. 13-175. International Monetary Fund, 2013.

[2] Strahan, Philip E. "Borrower risk and the price and nonprice terms of bank loans." FRB of New York Staff Report 90 (1999).Tomar, Divya, and Sonali Agarwal. "A survey on Data Mining pproaches for Healthcare." International Journal of Bio-Science and Bio-Technology 5.5 (2013): 241-266

[3] AboobydaJafar Hamid and Tarig Mohammed Ahmed, "DEVELOPING PREDICTION MODEL OF LOAN RISK IN BANKS USING DATA INING", Machine Learning and Applications: An International Journal (MLAIJ) Vol.3, No.1, March 2016.

[4] vTomar, Divya, and Sonali Agarwal. "A survey on Data Mining pproaches for Healthcare." International Journal of Bio-Science and Bio-Technology 5.5 (2013): 241-266.

[5] Sharma, Poonam, and Gudla Balakrishna. "PrefixSpan: Mining Sequential Patterns by Prefix-Projected Pattern." International Journal of Computer Science and Engineering Survey 2.4 (2011):111.

[6] Chitra, K., and B. Subashini. "Data Mining Techniques and its Applications in Banking Sector."International Journal of Emerging Technology and Advanced Engineering 3.8 (2013): 219-226.