DEVELOPING PREDICTION MODEL OF LOAN RISK IN BANKS USING MACHINE LEARNING.

[1]Anshuman rai, [2]Shubham Batra, [3]Riyazuddin Ansari, [4]Sanjeev kumar
[1]Research Student, [2]Research Student, [3]Research Student, [4]Associate Professor
Department of Computer Science and Engineering
ABES Institute of Technology, Ghaziabad, India

Abstract: This project is insight into the design and implementation of Developing Prediction Model of Loan Risk in Banks. Extending credit to individuals is essential for markets and society to act efficiently. Estimating the probability that an individual would default on their loan, is useful for bank to make decision whether to approve a loan to the individuals or not. In our project we find the accuracy for three models in python language and evaluate it to establish the finest model to forecast the finance status for an organization. We did the experiment one time for each of the three models on the same data set and find the experimental result that show the Logistic Regression Model is the best model for forecasting the finance for customers.

***IndexTerms* - Credit Risk, Classification, k-nearest neighbor, Decision Tree, logistic regression, Prediction.**

I. INTRODUCTION

Credit Risk assessment is a crucial issue faced by Banks nowadays which helps them to evaluate if a loan applicant can be defaulter at a later stage so that they can go ahead and grant the loan or not. This help the bank to minimize the possible losses and can increase the volume of credits. The Prediction of borrower status i.e. in future borrower will be defaulter or non-defaulter is a challenging task for any organization or bank. Basically, the loan defaulter prediction is a binary classification problem. Loan amount, customer's history governs his creditability for receiving loan. This problem is to classify borrower as defaulter or non-defaulter.

There are many risks related to bank loans, for the bank and those who get the loans. The analysis of risk in bank loans need understanding what is the meaning of risk. Risk denotes existing negative threat for trying to achieve a current monetary operation [5]. Till recently all the activities of banks were regulated and hence operational environment was not conductive of risk taking. Better insight, sharp intuition and longer experience were adequate to manage the limited risks. Business is the art of extracting money from other's pocket, sans resorting the violence. But profiting in business without exposing to risk is like trying to live without being born. everyone knows that risk taking is failure prone as otherwise it would be treated as sure taking. Hence risk is inherent in any walk of life in general and in financial sectors in particular. Of late, banks have grown from being a financial intermediation, the gap of which becomes thinner and thinner, banks are exposed serve competition and hence are compelled to encounter various types of financial and non-financial risks. Risk and uncertainties from an integral part of banking which by nature entail taking risks. Business grow mainly by taking risk, higher the profit and hence the business unit must strike a tradeoff between the two. The essential functions of the bank. While Non-Performing Assets are the legacy of the past in the present, Risk Management system is the pro-active in the present for the future. Managing risk is nothing but the managing the change before the risk manages. While new avenues for the bank has opened up they have brought with them new risks as well, which the banks will have to handle and overcome.

II. LITERATURE SURVEY

a. RELATED WORK

Many researches have been conducted based on data mining in the field of financial and banking sector. This section presents briefly some of these techniques which are used in loans risk management and their finding Sudhakar et al focused on specifying the data mining applications usefulness, these applications are using several data mining techniques such as decision trees and Radial Basis Neural Networks. This study came with in which way to apply these applications in a credit-risk assessment field.

McLeod presents Neural networks properties and their fitness for the credit- granting process.

Barney et al made a comparison of the performance of regression analyses and neural networks to identify the farmers who will default on the loans of their Home Administration and those farmers who return back the loans 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. [1]

Glorfeld and Hardgrave (2001) proposed a complete and useful systematic way to produce an optimal design of a high-performance model for neural network estimating of the Credit value related to applications of commercial loan. The neural network constructed using their design was able to classifying 75% of loan applicants correctly. [6]

Tessmer checked credits that offered to small Belgian businesses and he used a decision tree- based learning way. His study focused on the Type I credit errors impact (he means by type 1 taking good loans as bad loans), and Type II credit errors which mean regarding bad loans as good loans, on the accuracy, conceptual validity and constancy of the learning operation. This study has built on a previous research that compare the efficiency of several data mining tools in several credit risk estimation field.

Efficacy of neural networks and traditional techniques analyzed by Desai et al to build rating models for loans unions. he used consistent sample of data consists of 18 variables belong to three credit unions and his study proved that neural networks were more useful in bad loans detection, whereas logistic regression useful in discovering bad and good loans. [2]

Jagielska et al investigated the abilities of neural networks in classifying loans risk, uncertain logic genetic algorithms, rule stimulation software, he concluded that the genetic approach is more favorably than the neuron fuzzy and rough set methods [2].
A study by A.J.Feelders and A.J.F.le Loux about conducted a case study for personal loan evaluation by using data mining techniques. the study carried out in Netherlands in ABN AMRO bank. Data mining capabilities were applied to assess the personal loans. Historical data of clients and their return-back activities and behaviors are used to predict whether a customer will default or not [1].

Emile J. Salame focused on objectives that provide powerful tool to help in decreasing number unauthorized borrowers whose impact on the financial institution is positive. In addition, the unauthorized borrowers whose impact on the financial institution is positive. In addition, the paper gave insight in agricultural loan data to help decision-makers and to increase their ability to manage the operation of lending farmers which decreased the time and cost of checking of loan ambiguity and help loan officers to a crucial diction toward the customers [3].

Michael D. Johnson, Anders Gustafsson studied broadly the usages of data mining techniques in banking sectors and its related impacts on several operations. [4]

b. PROPOSED METHODOLOGY

The steps involved in this model building methodology are represented as below:

1. Data Selection
2. Data Pre-Processing
3. Features Selection
4. Building Classification Model
5. Predicting Class Labels of Test Dataset
6. Evaluating Predictions

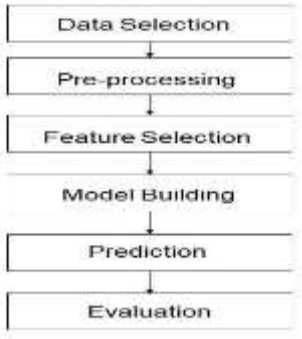


Figure 1

Figure 1 describes the methodology. Firstly, the information is collected from costumers then the data filtering is taken place where missing value are removed. In the third step, the feature importance is carried out. It makes the model accurate and more efficient. In the fourth step, the machine learning approaches, (refer Table 5) were trained and tested on the default parameters. Finally, evaluation is done.

i. Logistic Regression

MODEL: Logistic Regression

The classification methods can be classified into parametric and non-parametric problems. In fact, parametric methods are based upon the assumptions of normally distributed population and estimate the parameters of the distributions to solve the problem.

Dataset: Lending Club Time Period: 2007 – 2011

Original Dataset: 42538 samples with 111 features

Cleaned Dataset: 39685 samples with 37 features Model

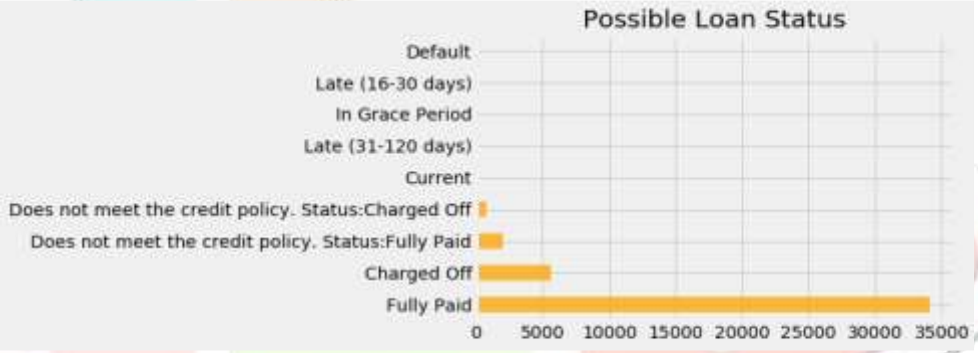


Figure 2: Possible loan outcomes

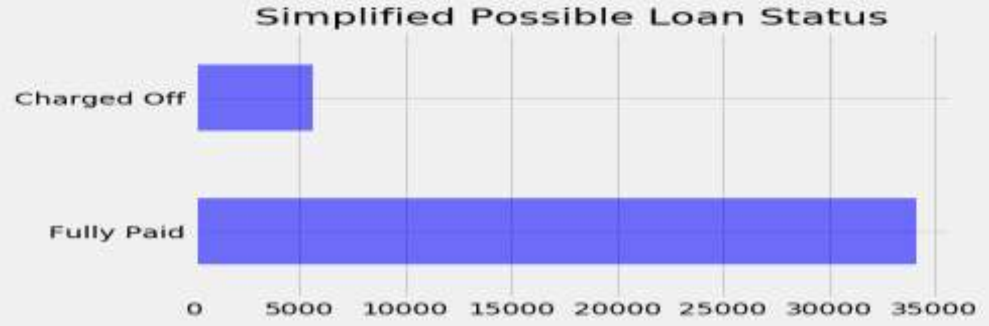


Figure 3: Simplified loan outcomes

False positives: Loans that were actually risky but were predicted to be safe

True positives: Loans that were actually safe and also predicted to be safe

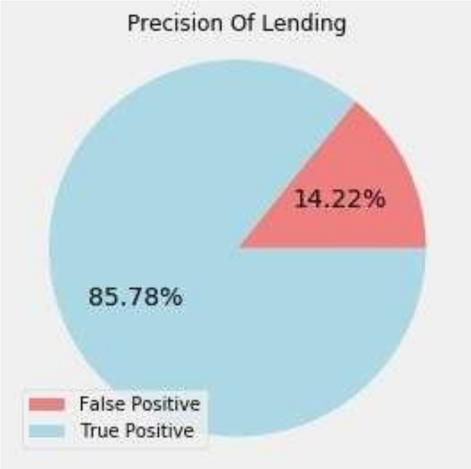


Figure 4:Precision of Lending

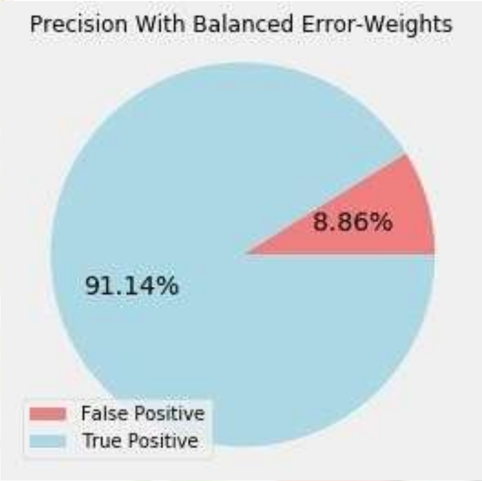


Figure 5: Precision of Lending with Logistic Regression

Accuracy Achieved: 91.14 %

ii. Decision Tree algorithm:

A decision tree model is one of the most frequent data mining models. It is popular because it is easy to understand. Decision trees are one of the useful algorithms that are used for regression and classification. They are also known as glass -box model. When the model once found the template in the data then we can see what the decision will be made for that data which we want to predict.

MODEL: Decision Tree

Dataset: Lending Club Time Period: 2007 –2011

Original:

Number of safe loan samples: 99457

Number of risky loan samples: 23150

Modified: (for training)

Number of safe loans samples: 23150

Number of risky loan samples: 23150

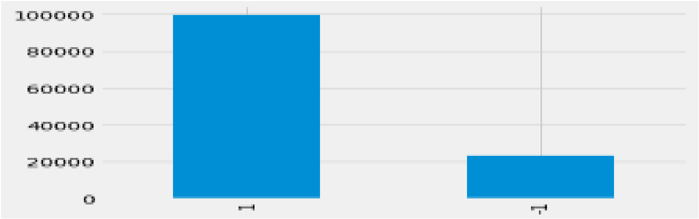


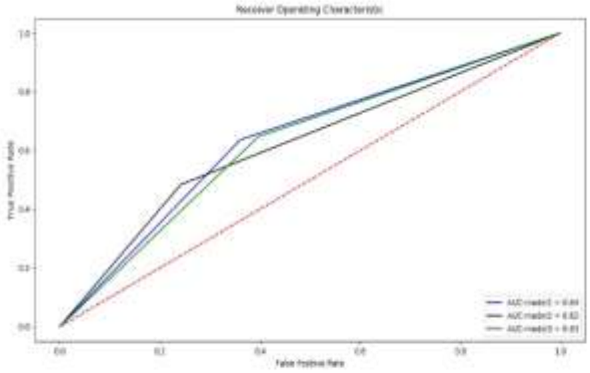
Figure 6:Number of safe and risky loan sample in dataset

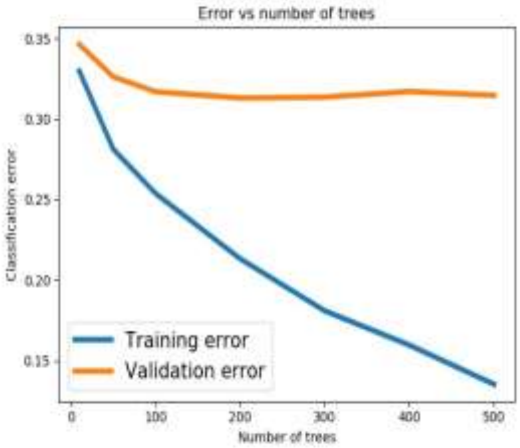
+1 : Safe loans

-1 : Risky loans

Accuracy on training data = 0.64

Accuracy on validation data = 0.63





Accuracy Achieved: ~70

iii. K-Nearest Neighbor

The k-Nearest Neighbor algorithm belongs to the family of instance-based, competitive learning and lazy learning algorithms. Instance-based algorithms are those algorithms that model the problem using data instances (or rows) in order to make predictive decisions. The k- nearest neighbor algorithm is an extreme form of instance-based methods because all training observations are retained as part of the model. It is a competitive learning algorithm, because it internally uses competition between model elements (data instances) in order to make a predictive decision.

MODEL: k-Nearest Neighbor

Dataset: LendingClub

Time Period: 2007 – 2011

Original Dataset: 42538 samples with 111 features

Cleaned Dataset: 39685 samples with 37 features

Accuracy Achieved: 85.8%

III. CONCLUSION

In this paper, we find the accuracy of three Model – (Logistic Regression, Decision tree, k- Nearest Neighbor) in python language and evaluate it to establish the best model to predict the finance status for an organization. We did the experiment for one time for each model on the same data set and the accuracy achieved is shown in the table below. The experimental results that show the k-Nearest Neighbor is the best model for forecasting the loan for customer.

Logistic Regression	91.14 %
Decision Tree	70%
k-Nearest Neighbor	85.8%

Table 1:Accuracy of models

REFERENCES

[1] Zurada, Jozef, and Martin Zurada. "How Secure Are “Good Loans”: Validating Loan-Granting Decisions and Predicting Default Rates on Consumer Loans."Review of Business Information Systems (RBIS) 6.3 (2011): 65-84.

[2] İkizler, Nazlı, and H. Altay Guvenir. "Mining interesting rules in bank loans data." Proceedings of the Tenth Turkish Symposium on Artificial Intelligence and Neural Networks. 2001.

[3] Abhijit A. Sawant and P. M. Chawan, "Comparison of Data Mining Techniques used for Financial Data Analysis," International Journal of Emerging Technology and Advanced Engineering, June 2013.

[4] R.Mahammad Shafi, A Tool for Enhancing Business Process in Banking Sector, 3rd ed.:
International Journal of Scientific & Engineering Research, 2012

[5] Strahan, Philip E. "Borrower risk and the price and non-price terms of bank loans." FRB of New York Staff Report 90 (1999).

[6] r.R.Mahammad Shafi, A Tool for Enhancing Business Process in Banking Sector, 3rd ed.: International Journal of Scientific & Engineering Research, 2012.

[7] K. Bache and M. Lichman, UCI machine learning repository.

[8] Galindo J., Tamayo P., Credit Risk Assessment Using Statistical and Machine Learning: Basic Methodology and Risk Modeling Applications (Computational Economics #15, 107-143, 2000.)

[9] Training Dataset: https://www.lendingclub.com/info/download-data.action.

[10] https://www.ibm.com/support/knowledgecenter/en/SSLVMB_22.0.0/com.ibm.spss.stistics.cs/spss/tutorials/trees_credit_intro1 html.