

# Credit Risk Analysis Using Machine Learning Models

<sup>1</sup>Sandip Chobhe, <sup>2</sup>Shubham Madalapure, <sup>3</sup>Sanket Chandake, <sup>4</sup>Akshay Baraskar  
<sup>1</sup>B.E(Computer Engineering), <sup>2</sup>B.E(Computer Engineering), <sup>3</sup>B.E(Computer Engineering), <sup>4</sup>B.E(Computer Engineering)  
<sup>1</sup>Pune District Education Association college Of Engineering Manjari bk, Pune,  
<sup>2</sup>Pune District Education Association college Of Engineering Manjari bk, Pune,  
<sup>3</sup>Pune District Education Association college Of Engineering Manjari bk, Pune,  
<sup>4</sup>Pune District education association Collage of engineering manjari bk

**Abstract** - Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention. credit risk evaluation is one of the critical and biggest challenge faced by banks accuracy plays very important role in classification of credit data to avoid financial loss. in Banking industry has the major activity of lending money to those who are in need of money. In order to payback the principle borrowed from the depositor bank collects the interest made by the principle borrowers. Credit risk analysis is becoming an important field in financial risk management Credit risk predictions, monitoring, model reliability and effective loan processing are key to decision-making and transparency. In this work, we build binary classifiers based on machine and deep learning models on real data in predicting loan default probability In the given paper we are study different techniques for the credit risk analysis which are used for the evaluation for the credit risk data sets.

**keywords** - Machine Learning, Credit risk; Machine learning; Bayesian classifier; Naive Bayes classifier; Decision tree; KNN; K-means clustering.

## I. INTRODUCTION

Credit risk is one of the major financial challenges that exist in the banking system. Yet, so far, many lenders have been slow to fully utilize the predictive power of digitising risk. This is despite a recent report from McKinsey showing that machine learning may reduce credit losses by up to 10 per cent, with over half of risk managers expecting credit decision times to fall by 25 to 50 per cent.

The banking system evaluates the accuracy of the datasets in order to classify the loan applicants into good and bad classes. The applicants which are in the good classes have the high probability of returning the money to the bank. The applicants which are in the bad classes have the low probability of returning of the money to the bank so, they are the defaulters of the loans.

As credit risk prediction plays an important role in the banking sector and it is a very critical and biggest challenge faced by all the banks, accuracy plays a very important role in classification of credit data to avoid the financial loss of Banks. The increase in defaulter's rate in the credit risk data set which is not reliable gives motivation towards this sector.

## II. LITERATURE SURVEY

This section of the literature survey eventually reveals some facts based on thought analysis of many authors work as follows.

[1] Pornwattana Wongchinsri, Werasak Kuratach in this survey, we propose a binary classification approach that can classify customers who apply for loans. A statistical technique called Stepwise Regression (SR) is used as a pre-process to select important features for the classifier. Artificial Neural Network (ANN) is the classification model type that has been chosen. The concept of confusion matrix combined with business rules has been used for obtaining an appropriate classification model. Based on the experimentation, we found that our method, SR-based Binary classification has the accuracy rate of 95.65%, which is higher than using ANN alone (91.30%).

[2] Shishi Dahita, N.P, Singh in this paper hybrid approach is used to enhance the classification accuracy for better credit evaluation of consumer loans. Two benchmarked of MLP Neural network Technique with FS and bagging for higher classification accuracy and then improving loan granting.in this paper ANN is are introduced with a special emphasis on MultiLayer preception architecture this is followed by a description of the method used for ensemble of classifiers 'with an emphasis on bagging method and on FS.

[3] Bhuvaneswari, This study deals with the analysis of a data set comprising of opulent vehicle credit portfolios characterized by relevant variables. It aims at assessing the risk associated with these portfolios and finally presents a predictive model which highlights the important variables and depicts the combination of those variables that classify a client under defaulter or no defaulter. The study starts with the use of conventional statistical techniques and subsequently presents machine learning approach using three different decision tree classifiers.

[4] Peter Martey Add, Dominique Guegan, and Bertrand Hassani in this paper, we will focus on the algorithms that are used to make these decisions. Algorithms are used in different domains with different objectives. For instance, they are used in enterprises to recruit persons suitable for the profile proposed. Algorithms can simplify the process, make it quicker and more fluid, etc. Nevertheless, algorithms are a set of codes with specific objectives to attain certain objectives. For instance, in the process of recruitment, it can introduce discrimination or a specific profile, and then, “format” the persons working in the enterprise.

### III. Method

Different types of method are used for the evaluation of credit datasets for the good and reliable credit risk prediction

#### [1] DECISION TREE

Decision Tree Analysis is a general, predictive modelling tool that has applications spanning a number of different areas. In general, decision trees are constructed via an algorithmic approach that identifies ways to split a data set based on different conditions. It is one of the most widely used and practical methods for supervised learning. Decision Trees are a non-parametric supervised learning method used for both classification and regression tasks.

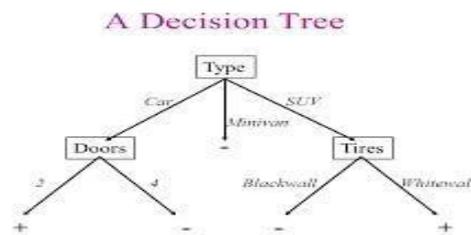


Fig.1- structure of Decision Tree.

#### [2] NAIVE-BAYES CLASSIFIER

Naive Bayes classifier is a straightforward and powerful algorithm for the classification task. Even if we are working on a data set with millions of records with some attributes, it is suggested to try Naive Bayes approach. Naive Bayes classifier gives great results when we use it for textual data analysis. Such as Natural Language Processing.

#### [3] K-MEANS

It is a type of unsupervised learning which is used when you have unlabeled data. The goal of this algorithm is to find groups in the data with the number of groups represented by the variable  $k$ . The algorithm works iteratively to assign each data point to one of  $K$  groups based on the features that are provided. The centroids of  $K$  clusters which can be used as low labeled new data and labels for training data.

#### [4] EXTREME LEARNING MACHINE

Extreme learning machines are feedforward neural networks for classification, regression, clustering, sparse approximation, compression and feature learning with a single layer or multiple layers of hidden nodes, where the parameters of hidden nodes (not just the weights connecting inputs to hidden nodes) need not be tuned.

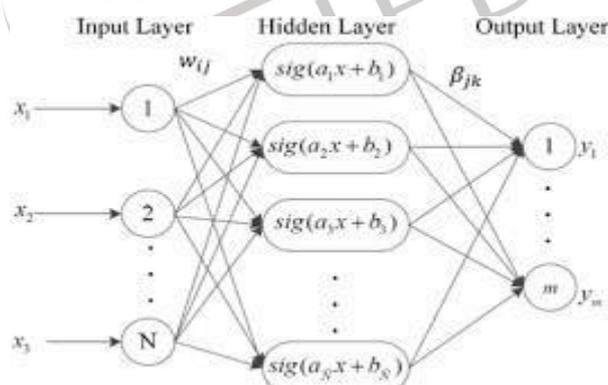


Fig.2- Extreme learning machines model.

These hidden nodes can be randomly assigned and never updated (i.e. they are random projection but with nonlinear transforms), or can be inherited from their ancestors without being changed. In most cases, the output weights of hidden nodes are usually learned in a single step, which essentially amounts to learning a linear model. The name "extreme learning machine" (ELM) was given to such models by its main inventor Guang-Bin Huang.

### IV. CONCLUSION

In this paper is to surveying on the different classifier which are used in the credit risk evaluation. In this paper different types of classifiers are discussed i.e. decision tree, naive Bayes classifier, k-means, extreme learning machine and also different types of ensemble classifiers are briefed. The dataset which are used in the classifier is discussed in the paper.

We have analyzed and compare their accuracies using different types classifiers and from comparison table we found that the ELM classifier gives better accuracies compare to other classifiers

## V. REFERENCES

- [1] Cawley, G., Talbot, N., 'Improved sparse least squares support vector machines, *Neurocomputing*, Vol. 48(1-4), pp. 1025-1031, 2002.
- [2] Lee, T-S., Chiu, C-C., Chou, Y-C., Lu, C-J., 'Mining the customer credit using classification and regression tree and multivariate adaptive regression splines', *Computational Statistics and Data Analysis*, Vol. 50, pp.1113-1130, 2006
- [3] Kruppa, J., Schwrz, A., Arminger, G. and Zeigler, A., 'Consumer credit risk: Individual probability estimate using machine learning', *Expert System with Application*, Vol. 40, pp. 5125-5131, 2013.
- [4] Yang, L., Xiaohvi, Y., Jimmy, X.H. and Aijun, A., 'Combining integrated sampling with SVM ensembles for learning from imbalanced dataset', *Journal of Information Processing and Management*, Vol. 47, pp.617-631,2011.
- [5] Dhaiya, S., Singh, N.P., 'Impact of Bagging on MLP classifier', *International Conferences on Computing for Sustainable Global Development*, pp. 3794-3799. 2016.
- [6] Akko, Soner," An empirical comparison of conventional techniques, neural networks and the three-stage hybrid Adaptive Neuro Fuzzy Inference System (ANFIS)model for credit scoring analysis: The case of Turkish credit card data." *European Journal of Operational Research* 222.1 (2012): 168-178.
- [7] Riedmiller, Martin, and Heinrich Braun.," A direct adaptive method for faster backpropagation learning: The RPROP algorithm." *Neural Networks*, 1993., IEEE International Conference on. IEEE, 1993.
- [8] Cheng-Lung Huang, Mu-Chen Chen, Chieh-Jen Wang, "Credit scoring with a data mining approach based on support vector machines", *Expert Systems with Applications*, vol. 33, pp 847-856, 2007.
- [9] Christophe Mues, Bart Baesens, Craig M. Files, Jan Vanthienen, "Decision diagrams in machine learning: an empirical study on real-life credit-risk data", *Expert Systems with Applications*, vol. 27, 257-264, 2004.
- [10] Yang, L., Xiaohvi, Y., Jimmy, X.H. and Aijun, A., 'Combining integrated sampling with SVM ensembles for learning from imbalanced dataset', *Journal of Information Processing and Management*, Vol. 47, pp.617-631,2011.

