IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-6, June 2018 E-ISSN: 2347-2693

Predictive Modelling for Credit Risk Detection using Ensemble Method

Anand Motwani[1*], Goldi Bajaj[2] ,Sushila Mohane[2]

[1] Department of CSE, Sagar Institute of Science, Technology& Research, (SISTec-R), Bhopal, India
[2] Department of CSE, S.V.Polytechnic College,Bhopal,India
[3] Department of CSE, SagarInstitute of Science, Technology& Research, (SISTec-R), Bhopal, India

Available online at: www.ijcseonline.org

Accepted: 22/Jun/2018, Published: 30/Jun/2018

Abstract- With the expansion of credit business, the prediction models for taking decision of credit permissions with least risk are becoming more and more admired by banking sectors. The use of Machine Learning (ML) based models has confirmed to be of practical value in resolving a range of banking risk prediction problems. The model for Credit risk prediction seeks to predict feature factors, whether an individual is bad or good applicant for loan or not. Such problems can be better solved using ML. Also, Ensemble classifiers in ML play a key role in prediction problems. The use of Ensemble Methods (EMs) for classification is among the recent areas of research in ML. Many recent researches specify that EMs lead to a major improvement in classification performance by choosing suitable class. For this work, several ML techniques are explored and evaluated on real credit card datasets. Most ML methods have achieved an accuracy of less than 80 percent. Predictive model for Credit Risk Detection based on ensemble technique is proposed. The proposed model is evaluated on basis of various performance metrics and comparison is done with base classifier (learner) resulted in 81 percent prediction accuracy and better correlation coefficient.

Keywords-Predictive Modelling, Machine Ensemble Method, Credit Risk, Data Mining

I. INTRODUCTION

In case of world-wide economic crisis banking sector is available to react against it. It entails that some financial organization specialized to retail consumer loans and mortgages etc. is better support point when the crisis breaks out. Financial organizations are more reliant on loan interests obtained from business segment that are facing crisis. Such organizations are facing problems in sanctioning the loans and with current defaulters of loans. In fact, banks have to take suitable actions to lessen credit risks to decrease costs as much as possible. Banks are having customer's portfolios that are likely to go through current crisis without much difficulty. On the basis of customer's portfolio it is possible to find the default status of payment or credit score.

Prediction model based on ML algorithms are assumed good for classifying data that is never seen before into their various categories. The predictive models work by predicting the most suitable category to which a data point belongs to by "learning" from labelled observations. Therefore such models are extensively used in various sectors including financial sectors. Some of the well-known classification algorithms used in this paper is briefly discussed in this paper.

This work starts with an overview of Ensemble Learning Methods to score credit risk. Ensemble Methods are discussed in Section II. Literature review relevant to work is presented in Section III. The proposed framework for Prediction of credit risk detection is discussed in Section IV. Finally experiments using proposed predictive model is presented in Section V. The result analysis and comparison of the ensemble method based learner and base learners is also shown in this section. This paper compares the performance in terms of Accuracy, Mean Absolute Error (MSE) and Correlation Coefficient.

There are many classifiers based on various algorithms are available for product classification. Sometimes, a classifier might override the others in classification performance for a specific set of data and sub-methods involved. In general, it is can't be said that one method always outperforms all the other methods for every possible situation. The work presented in this paper suggests the way to predict the credit risk to deal with losses. The key part of this paper is the proposed predictive model for credit risk detection which detects the potential problem clients.

.

II. ENSEMBLE LEARNING METHODS FOR CREDIT SCORING

A. Overview

Ensemble learning is a ML paradigm where several learners are trained to resolve the same problem [1]. The ordinary ML techniques try to build a model from the training data on one hypothesis, whereas EMs tries to construct a set of rules or hypotheses to use [2]. Learners composed of an ensemble are usually called base learners [3]. A combination of classifiers has been proposed in the field of ML to improve the learning models. So, an ensemble is a set of multiple classifiers, where individual classification results are pooled to get better accuracy

B. Bagging

Bagging [4, 11] is a well know ensemble method which is also called bootstrap aggregating. It is one of the most primitive ensemble learning algorithms with good performance features. It is also one of the most perceptive and easiest to implement methods. In it different training sets are obtained by drawing subsets that are randomly drawn—with replacement—from the entire training data. Each training data subset is used to train a different base learner of the same type. The strategy can lessen the variance when pooled with the basic class learner generation strategies. The method lessens the over-fitting problem and is more effective on unstable learning algorithms as it does not depend on a single classifier. The prediction result of all the classifiers is combined by voting method to obtain final prediction.

The algorithm is shown below:

Input:

Dataset: D

Base Learning Algorithm: L

Number of runs of Learning Algorithms: N

Process:

For i = 1.....N

Di = Bootstrap(D) // Generate a bag of sample.

// Train a base learner hi from the Bootstrap sample

hi = L(Di)

end.

Output:

$$\mathbf{H(x) = \arg\max_{y \epsilon Y} \sum_{i=1}^{N} 1(y = h_i(x))}$$

The value of 1(a) is 1 if 'a' is true and otherwise 0.

C. Boosting

Unlike Bagging, Boosting [5] method generates different base learners by sequentially reweighting the instance in the training set. Every misclassified instance will get a larger weight in the next run of training. The basic idea of Boosting is to repeatedly apply a base learner to modified versions of the training dataset, thereby producing a sequence of base learners for a predefined number of iterations. In it weights are adjusted for correctly and incorrectly classified instances. Finally a linear combination of base learners is obtained using boosting and the classifiers are weighted by their own performance. AdaBoost [5] is the most widely used boosting technique.

D. Stacking

Stacking is another accepted EM and it is applied to base learners constructed using different ML algorithms [6]. Unlike Bagging and Boosting, Stacking is not normally used to combine base learners of the same type.

*Corresponding Author:
C. T. Lin
e-mail: ct.lin@hotmail.com , Tel.: +00-12345-54321

III. LITERATURE REVIEW

The authors [7] applied 15 different ML algorithms to choose the best fit ML algorithm to apply on bank credit card dataset. The experiment showed that, the algorithms except Naive Bayes (Gaussian) and the Nearest Centroid performed plausibly well in terms of performance evaluation metrics. The algorithms achieved accuracy between 76% to over 80%. The 5 main features that affect the credit worthiness of customers are also determined. These features are used to detect performance through selected algorithms and it is found that there is no significant difference in their metrics when compared without feature selection i.e. using all 23 features. Also, a predictive model is formulated using linear regression for predicting customer's credit worthiness. The proposed model used 03 features.

In [8], it was stated that besides good credit and high purchasing power of customer, a certain amount of credit risk is linked with these credit groups. The paper aims at assessing the risk linked with such portfolios and finally presents a predictive model which highlights the key attributes and depicts the grouping of those attributes that categorize a client under defaulter category or non-defaulter category. The analysis is done on dataset comprising of luxurious vehicle credit range characterized by relevant attributes. The study uses conventional statistical approaches and subsequently presents ML techniques using three different decision tree classifiers: J48, Decision Tree and Random Tree.

Evaluation and Prediction of customer's Credit score is a key for preventing losses for the banking sector. The study presented in [9] analyzes the accuracy of the EMs in classifying risk group into good or bad. Authors conducted experiments using three EMs namely Bagging, AdaBoost and Random Forest combined with three ML algorithms. For selecting important features from dataset, Feature selection method is applied.

Gang Wang et al. [10] performed a performance comparison of 3 popular EMs, i.e., Boosting, Bagging, and Stacking, based on 4 basic learners, i.e., Logistic Regression, Decision Tree (DT), ANN and SVM. Experimental results shown that these 3 EMs can considerably improve individual learners. Bagging, in particular, performed better than Boosting for most credit datasets. Bagging and Stacking with DT experiments performed well in terms of performance (accuracy).

Authors [3] explored the performance of a range of systems based on ensemble of classifiers for credit scoring and bankruptcy prediction. The obtained results are better in than the individual classifiers. Multi-layer perceptron (MLP) neural net found to be the best method tested in this work.

The research is carried over 03 financial datasets including German Credit dataset. The authors concluded the paper with a note that ensemble of classifiers may be used for boosting the performance of ‘‘stand-alone” classifier. It is shown that the Random Subspace (RS) ensemble method performed better than other EMs. The maximum accuracy obtained using MLP with RS is 0.7917, i.e. 79.17%.

IV. PROPOSED PREDICTIVE MODEL FOR CREDIT RISK DETECTION USING ENSEMBLE METHOD

A. *Proposed Model*

Figure 1, depicts the proposed framework having three phases: First phase comprise of Loading and Pre-Processing the data. In Second phase model building process using training data is shown. The best model for classifying the data is chosen in this phase by iteratively applying various models and evaluating the model performance metrics. In the second phase we applied Bagging EM with REP Tree Model. In the last phase, the predictive model is deployed and used as a tool on new unclassified data. REP Tree is briefly discussed here.

B. *REP Tree*

This algorithm builds a decision/regression tree using information gain/variance and prunes it using reduced-error pruning (REP) (with back fitting). Reduced error pruning is one of the simplest forms of pruning. Starting at the leaves, each node is replaced with its most popular class. If the classification accuracy is not affected then the change is kept, else a different class is chosen. Reduced error pruning has the advantage of simplicity and speed.

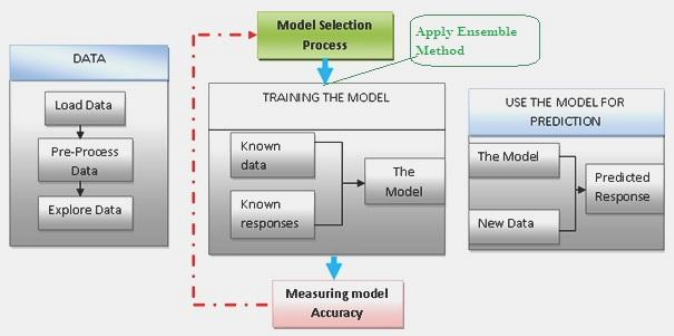

Figure 1. Proposed Predictive Model

V. EXPERIMENT SETUP AND RESULT ANALYSIS

A. *Tool used*

Weka is Data Mining (DM) tool. Its Main Features includes: data preprocessing tools and feature selection capabilities, classification, clustering, regression and association rules algorithms. It is open source, platform-independent and freely available. It is easy to use by DM specialists and academicians as well. It has kept up-to-date, with new algorithms being added as they appear in the research literature. For this work Weka version 3.8.1 is used.

B. *Dataset Used*

The dataset available is taken for experimentation purpose. This dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005. In all there are 25 variables in datasets, out of which first represents instance id and last represents the class or category of particular instance. So, there are 23 features within the bank credit defaulters’ dataset. However, not all the 23 features have considerable impact in finding the ability of a given customer in paying his/her loan or not.

C. *Methodology*

The methodology of proposed work is explained with the help of Figure 2. Figure 3 showing screenshot of real experiment performed on Weka DM tool. Weka allows configuring several simulation parameters.

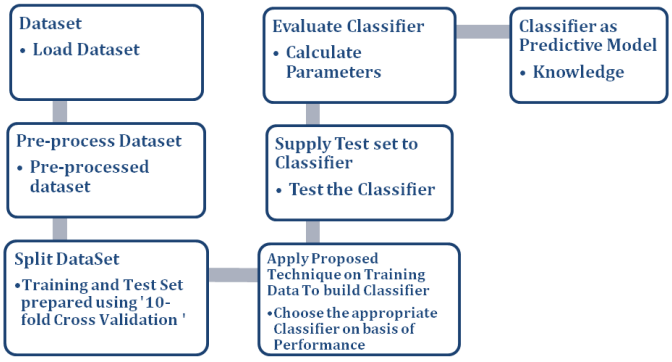

Figure 2.Methodology of Proposed Work

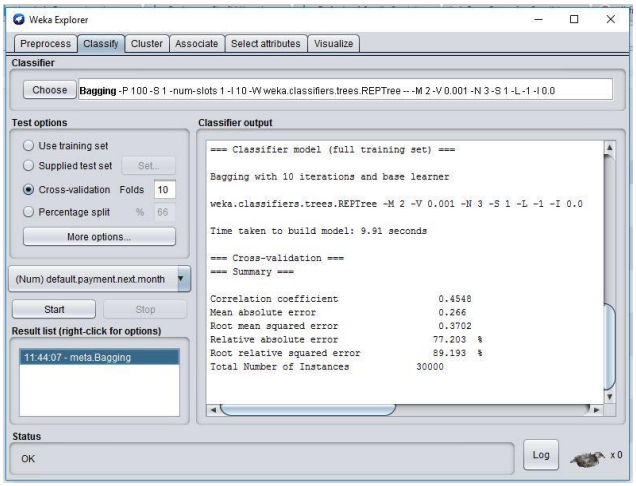

Figure 3. Screenshot of Experiment in Weka

D. Performance Parameters

Accuracy: The Accuracy of a prediction model on a given test set is the percentage of instances that are correctly classified by the model.

Correlation Coefficient: tells that relation between the value of interest (x) and estimated value (x') using some algorithm. The Correlation coefficient tells how much x and x' are related. In general model with bigger correlation and smaller error estimates are accepted.

E. Result and Analysis

For evaluation the results of Base Learners and Proposed model are compared (refer Table 1 and 2). The Correlation coefficient and MAE are better in proposed model. The Accuracy is also improved in comparison to base learners.

Table 1 Result Base Learner

Parameters	**Base Learner**		
	Linear Regression	***Decision Stump***	***REP Tree***
Accuracy	79.82%	80.9%	81.12%
Mean Absolute Error	0.307	0.292	0.27
Correlation coefficient	0.35	0.389	0.4482

Table 2 Result Proposed Work

Parameters	**Proposed Work (EM with Base Learner)**		
	Linear Regression	***Decision Stump***	***REP Tree***
Accuracy	79.85%	81.1%	81.66%
Mean Absolute Error	0.311	0.294	0.266
Correlation coefficient	0.326	0.39	0.454

VI. CONCLUSION AND FUTURE WORK

Banks and financial organizations are facing the challenge of identifying risk factors, which should be considered while advancing the loans/credit to customers. The dataset have several features/attributes of the customers, but most of these descriptions have slight predictive effect on the credit worthiness of the customer.

The learners are compared on basis of various parameters and the proposed model found to be better in terms of Accuracy and Correlation. The findings of paper have a lot of implications. In future, the proposed predictive model would definitely can be applied to find out credit worthiness of customer before granting loan. Furthermore, the result showed that an Ensemble ML algorithm is suitable for studying bank credit dataset. In future we intend to build up a ML system risk automated system over cloud for financial organizations that will incorporate key features to determine credit worthiness of customers.

REFERENCES

[1] Polikar, R. (2006). Ensemble based systems in decision making. IEEE Circuits and Systems Magazine, 6(3), 21–45.

[2] Zhou, Z. H. (2009). Ensemble. In L. Liu & T. Özsu (Eds.), Encyclopedia of database systems. Berlin: Springer.

[3] Loris Nanni , Alessandra Lumini, "An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring", Elsevier, Expert Systems with Applications 36 (2009) 3028–3033.

[4] David Opitz and Richard Maclin, "Popular Ensemble Methods: An Empirical Study", Journal of artificial intelligence research 169-198, 1999.

[5] Freund, Y., & Schapire, R. (1996). Experiments with a new boosting algorithm. In Proceedings of the thirteenth international conference on machine learning, Bari, Italy (pp. 148–156).

[6] Wolpert, D. H. (1992). Stacked generalization. Neural Networks, 5(2), 241–259.

[7] Regina Esi Turkson, Edward Yeallakuor Baagyere, Gideon Evans Wenya, "A Machine Learning Approach for Predicting Bank Credit Worthiness", ISBN: 978-1-4673-9187-0, IEEE 2016

[8] [8] U Bhuvaneswari, P. James Daniel Paul, Siddhant Sahu, "Financial Risk Modelling in Vehicle Credit Portfolio", 978-1-4799-4674-7/14/$31.00, IEEE 2014

[9] C.R.Durga devi, Dr.R.Manicka chezian, "A Relative Evaluation of the Performance of Ensemble Learning in Credit Scoring", IEEE International Conference on Advances in Computer Applications (ICACA), 978-1-5090-3770-4/16, 2016

[10] Gang Wang, Jinxing Hao, Jian Mab, Hongbing Jiang, "A comparative assessment of ensemble learning for credit scoring", Expert Systems with Applications 38 (2011) 223–230

[11] Breiman, L. (1996). Bagging predictors. Machine Learning, 123–140.

