Ensemble Model for Loan Prediction Using Machine Learning

Anchal Goyal[1] and Ranpreet Kaur[2]

[1]*Research Scholar*
Department of Computer Science, RIMT –IET, Mandi Gobindgarh, Punjab
anchal91.goyal@gmail.com

[2]*Assistant Professor*
Department of Computer Science, RIMT –IET, Mandi Gobindgarh, Punjab
er.ranpreet@gmail.com

Abstract

Banks frequently rely on credit risk prediction models to decide whether to grant a loan request or not. For a bank, reliable model is essential so that the bank can provide credit without any threat. We introduce an efficient prediction technique based on accuracy that helps the banker to predict the credit risk for customers who have applied for loan. We apply three different models (SVM Model, Random Forest Network and Tree Model for Genetic Algorithm) and the Ensemble Model, which is combination of these three models and analyses the credit risk for optimum results. Ensemble Model gives better results as compared to stand alone model.

Keywords: *Ensemble, Prediction, Tree Model, Random forest.*

1. INTRODUCTION

A loan is the lending of money from one individual, organization or entity to another individual, organization or entity. Now a day's bank plays a crucial role in market economy. The success or failure of the banking industry largely depends on the industry's ability to evaluate credit risk. Before giving the credit loan to borrowers, bank decides whether the borrower is bad (defaulter) or good (non defaulter). 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 [14]. Credit risk is one of the most studied and researched area in banking system. The loan predicting model [4] uses analysis model and techniques and uses the current and previous data of the customer to make prediction about the customer ability to pay back the loan amount on time.

Ensemble modeling is the method of running two or more associated but different models and then combines the results into a single score to improve the accuracy of predictive data and data mining applications. In machine learning, ensemble methods use several algorithms to get better predictive performance.

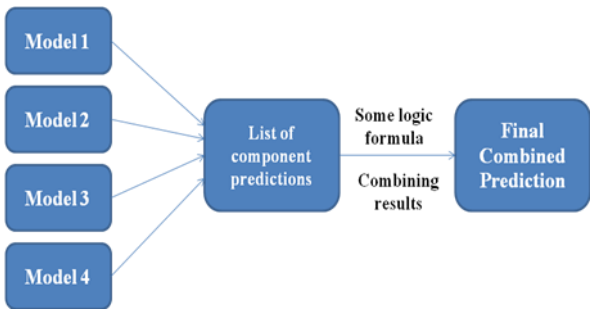

Fig-1 Ensemble Model

ISSN: 2456-1118

An ensemble is a supervised learning algorithm. Supervised Learning Algorithms are usually described as performing the task of find suitable data that will make better predictions with a specific problem. Ensembles combine multiple facts to form a better result. When several prediction models are used to try to make a forecast, the method is termed as multi-model ensemble forecasting.

This method of prediction has been shown to enhance forecasts when compared to a single model-based approach. The main benefits of Ensemble models are:

- Better Forecasting
- More Constant model
- Better results
- Reduces error

2. RELATED WORK

Amira Kamil Ibrahim Hassan, Ajith Abraham constructed a loan default predication model using three several neural network training algorithms. The aim is to test accuracy using attribute filter technique and develop a model called ensemble model by combining the results of those three algorithms. The experiment did on several parameters like training time, MSE, R, iteration for comparison. The best algorithm was Levenberg-Marquardt (LM) because it had largest R and the slowest algorithm is One Step Secant (OSS). For the accuracy purpose, the filtering function was applied on original dataset that produced two another datasets. Then for each data set different training algorithm of neural network is applied and the filtering function gave the better model among all the models.

E. Angelini, A. Roli, and G. di Tollo [5] used a feed-forward neural network with classical topology and a feed-forward neural network with ad hoc connections, justifying their use of neural network that it is one of the best methods to design a prediction model. The conclusions reached that both methods produced efficient models that can correctly predict default with low error.

Gang Wang, Jian Ma [15] proposed an ensemble approach based on boosting and random subspace and the model named as RS-Boosting for the risk prediction. It gives better performance. The results shows that the proposed approach gives best performance among seven other methods i.e., logistic regression analysis (LRA), decision tree (DT), artificial neural network (ANN), bagging, boosting and random subspace.

Ngai et al. identified eighty seven articles related to application of data mining techniques in CRM, and published between 2000 and 2006. The classification model is the most commonly applied model in CRM for predicting future customer behaviors. They also stated that neural networks were used in a wide range of CRM domains. However this study has some limitations, it surveyed articles published between 2000 and 2006.

Dr. A. Chitra and S. Uma [1] introduces a two level ensemble model for prediction of time series based on radial bias function network(RBF), k nearest neighbor (KNN) and self organizing map (SOP). The aim is to increasing the prediction accuracy. They construct a model named PAPEM i.e. Pattern prediction Ensemble Model. The Comparison of various classifiers done on root mean square, mean absolute percentage error and prediction accuracy. The results show that the PAPEM model is better than standalone classifier.

Maher Alaraj, Maysam Abbod, and Ziad Hunaiti [13] proposed a new ensemble method for classification of costumer loan. This ensemble method is based on neural network. They state that the proposed method give better results and accuracy as compared to single classifier and any other model.

ISSN: 2456-1118

Akkoç [10] used a three stage hybrid Adaptive Neuro-Fuzzy Inference model, which is combination of statistics and Neuro-Fuzzy. A 10-fold cross was used for validation and a comparison with traditional models show that the produced model is much better.

Alaraj M , Abbod M [13] introduce a credit risk model that are based on homogenous and heterogeneous classifiers. Ensemble model based on three classifiers that are logistic artificial neural network, logistic regression and support vector machine. The results show that the heterogeneous classifiers ensemble gave improved performance and accurateness as compared to homogeneous classifiers ensemble.

3. DATA SET AND FEATURES

The data set include 13 attributes such as Gender, Marital Status, Education, Income, Loan Amount, Credit History and others which are shown as:

Table1: Feature Description

Feature ID	Features	Information
F1	Loan ID	Unique Loan ID
F2	Gender	Male/ Female
F3	Married	Applicant married (Y/N)
F4	Dependents	Number of dependents
F5	Education	Applicant Education (Graduate/ Under Graduate)
F6	Self-employed	Self employed (Y/N)
F7	Applicant Income	Applicant income
F8	Co applicant Income	Co applicant income
F9	Loan Amount	Loan amount in thousands
F10	Loan Amount Term	Term of loan in months
F11	Credit History	credit history meets guidelines
F12	Property Area	Urban/ Semi Urban/ Rural
F13	Loan Status	Loan approved (Y/N)

4. MACHINE LEARNING MODELS

Machine learning is a field of computer science that involves the learning of pattern identification and computational learning theory in artificial intelligence [3]. It explores the study and construction of algorithm that can make prediction on data. Machine Learning is used to build programs with its tuning parameters that are adapted consequentially to increase their functioning by adapting to earlier data.

Various machine learning models that have been applied for the prediction of accuracy as explained below [3].

A. Decision Tree Model

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.

B. Linear model

A linear model is the one of the method for fitting a statistical model to data. It is appropriate when the target variable is numeric and persistant. This model helps to analyze the data and also helps to recognize and predict the performance of the complicated system.

ISSN: 2456-1118

C. Random Forest Model

A random forest model is basically a collection (i.e. ensemble) of tens or hundrerds of decision trees. These models are mainly used if we have large no. of input variables i.e. in hundreds and thousands and if we have very vast dataset. This model is very efficient if we have large no. of variables and it distributes the variable into different subsets.

D. Neural Network Model

This model is basically based on various layers that are connected to each other like neurons. This model combines the numbers and provides the numeric data to produce the final results throughout the network. These models are identical to biological neural network in order to perform functions parallel and collectively rather than individually.

E. Support vector machine

SVM is supervised machine learning model with learning algorithms which examine the data and uses that data for regression and classification [3]. This model uses a technique namely a kernel trick to transform the data and based on these transforms of data, it finds the best optimum results. It is not considered as better as than the other machine learning models because it works on less data set.

F. Extreme learning machines

ELM is a modification is a feed forward network with single layer which have a hidden nodes for single layer. The Weights are randomly given to hidden nodes and it never be updated. The name to this model was given by Guang-Bin Huang. Different from other traditional models, the extreme learning model not only provide the smaller training error but also better performance.

G. Multivariate Adaptive Regression Splines

This model is established by Jerome H. Friedman in 1991. This model is used for both regression and classification type problem with the purpose to predict the values. This technique has popular in data mining because it is used to find the difficult data mining problems.

H. Model tree

Model tree is a classification model that is combination of decision tree learning and logistic regression model. The package named ‘tree’ is used in implementation of this model. This model tree works on when have to predict the numeric quantities.

I. Bayesian Generalized Linear Model

BGLM is most generally used technique for creating the relationship. This model is used when have huge dataset and BGLM is used to fit the dataset into pragmatic size and remove the problem of over fitting. This model is included in package “arm” in r language.

J. Bagged Cart Model

This model is used for classification and regreession problems. This model build under the package ‘iperd’ and ‘plyr’. Bagging for classification and regression trees were suggested by Breiman in 1996.

K. Tree model form Genetic Algorithm

Genetic algorithm is a search heuristic i.e. it is an algorithm for finding and solving a problem more quickly and produces the result in reasonable time. This model is very efficient, flexible and finds optimal solutions for given problem.

ISSN: 2456-1118

Table2: Techniques Used

Models	Method Used	Packages
Bagged CART	bagging	Iperd
Random Forest	randomForestb	Random Forest
Tree Model For Genetic Algorithm	evtree	Evtree
Decision Trees	rpart	Rpart
Linear Model	multinom	Car, nnet
Neural Network	nnet	Nnet
SVM	ksvm	Kernlab
Extreme Learning Machine	elmtrain	elmNN
Multivariate Adaptive Regression spline	earth	Earth
Bayesian Generalized Linear Model	bayesglm	Arm
Model Tree	tree	Tree

5. MODEL ANALYSIS

Models are analysed for performance of prediction. The measure used for analysis is shown as follows:

A. Accuracy

Accuracy depends on how data is collected, and judged on basis of comparison of several parameters. True positive (TP) depicts amount of predictions which are positive, the actual value being positive. Similar in the case of true negative (TN). The accuracy is computed as [12]:

$$\text{Accuracy} = \frac{TP+TN}{\text{Toatal Data}} * 100 \qquad (1)$$

B. AUC

AUC or Area under Curve is a metric for binary calculation. It's a probably the second most popular parameter after Accuracy. It compute the area under the curve of a given performance measure. Its value lies between 0.5 and 1. It depicts the quality of models used for classification problems.

C. Gini Coefficient

The disparity of a distribution is calculated by using Gini coefficient and its values lies between 0 and 1.These are mainly used for model comparison.

$$Gini = 2AUC - 1 \qquad (2)$$

D. ROC Curve

A receiver operating characteristic (ROC) curve is used to classify problem of binary type. The function is included in pROC package.

E. K-S chart

K-S or Kolmogorov-Smirnov chart measures performance of classification models. More accurately, K-S is a measure of the degree of separation between the positive and negative distributions.

F. MER

MER metrics represents the Minimum Error Rate. Here threshold value act as a free parameter

G. MWL

ISSN: 2456-1118

MWL metrics represents the Minimum Cost- Weighted Error Rate. It is related to the KS statistics. Cost guides the threshold value in this measure.

6. RESULTS

In it we calculate the results of prediction of all models on the training dataset. The machine learning models may go through from over fitting. To overcome this over fitting problem, all models run on their defaulting parameters and the data is distributed among training and testing set are 70% and 30% correspondingly for all the models. The performance is calculated on basis of its Accuracy, H, Gini, AUC, AUCH, KS, MER, MWL, and ROC. Table 3 displays accuracy rate for each of model. Optimum results show that the ensemble model provides optimum results.

Table 3: Results

Models	Accuracy	H	Gini	AUC	AUCH	KS	MER	MWL	ROC
Decision Tree	78.47	0.26	0.52	0.76	0.76	0.52	0.22	0.17	0.76
Linear Model	79.86	0.30	0.60	0.80	0.80	0.60	0.18	0.12	0.80
Neural Network	79.86	0.30	0.60	0.80	0.80	0.60	0.18	0.12	0.80
Random Forest	**80.56**	**0.32**	**0.60**	**0.80**	**0.80**	**0.60**	**0.19**	**0.13**	**0.80**
SVM	**80.56**	**0.32**	**0.60**	**0.80**	**0.80**	**0.60**	**0.19**	**0.13**	**0.82**
Bagged Cart	78.47	0.26	0.52	0.76	0.76	0.52	0.22	0.17	0.76
Tree model for genetic algorithm	**81.25**	**0.35**	**0.68**	**0.84**	**0.84**	**0.68**	**0.17**	**0.09**	**0.84**
model tree	79.86	0.30	0.59	0.79	0.79	0.59	0.19	0.13	0.79
Extreme learning machine	68.75	0.27	0.49	0.66	0.59	0.48	0.16	0.11	0.64
Multivariate Adaptive Regression Spline	79.86	0.30	0.60	0.80	0.80	0.60	0.18	0.12	0.80
BGLM	79.86	0.30	0.60	0.80	0.80	0.60	0.18	0.12	0.80
ENSEMBLED MODEL (SVM + RF + TMGA)	**79.86**	**0.31**	**0.63**	**0.78**	**0.78**	**0.63**	**0.20**	**0.14**	**0.79**

7. CONCLUSION

In the proposed work, eleven machine learning models are constructed which have nine properties that are used to predict the credit risk of costumers who have applied for loan. Under different training algorithms, this paper presented an ensemble models for loan predications by using several parameters like Accuracy, Gini, Auc, Roc etc to do the comparison. The main purpose of this paper is to test the accuracy of models and develop a new model called ensemble model that combine the outputs of the three different models to predict the loan of costumers. Real Coded Genetic Algorithms is used to calculate the feature importance. These features help to predict the credit risk for costumers. K- Fold validation method is used to calculate the robustness of the predictive model.

REFERENCES

[1]. Dr. A. Chitra and S. Uma., "An Ensemble Model of Multiple Classifiers for Time Series Prediction", International Journal of Computer Theory and Engineering, Vol. 2, Issue 3, pp. 454–458, June 2010.

[2]. M. V. Jagannatha Reddy and B. Kavitha, " Extracting Prediction Rules for Loan Default Using Neural Networks through Attribute Relevance Analysis", International Journal of Computer Theory and Engineering, Vol. 2, Issue 4, pp. 596-601, August 2010.

[3]. Ms. Neethu Baby, Mrs. Priyanka L.T., " Customer Classification And Prediction Based On Data Mining Technique" , International Journal of Emerging Technology and Advanced Engineering, Vol. 2, Issue 12, pp. 314-318, December 2012.

[4]. Sivasree M S, Rekha Sunny T, "Loan Credibility Prediction System Based on Decision Tree Algorithm", International Journal of Engineering Research & Technology , Vol. 4 ,Issue 09, pp. 825-830, September 2015.

[5]. E. Angelini, A. Roli, and G. di Tollo, "A neural network approach for credit risk evaluation" elsevier, The Quarterly Review of Economics and Finance, Vol. 48, Issue 4, pp. 733–755, November 2008.

ISSN: 2456-1118

[6]. Suresh Ramakrishna, Maryam Mirzaei and Mahmoud Bekri, "Adaboost Ensemble Classifiers for Corporate Default Prediction" , 1st International Conference of Recent Trends in Information and Communication Technologies, pp. 258-269, September 2014.
[7]. Amira Kamil Ibrahim Hassan and Ajith Abraham, "Modeling Consumer Loan Default Prediction Using Ensemble Neural Networks", International Conference on Computing, Electrical and Electronics Engineering , pp. 719 – 724, August 2013.
[8]. C. F. Tsai and J. W. Wu, "Using neural network ensembles for bankruptcy prediction and credit scoring " ,Expert Systems with Applications, Vol. 34, Issue 4, pp. 2639–2649, May 2008.
[9]. Amir F. Atiya ,"Bankruptcy prediction for credit risk using neural networks: A survey and new results", IEEE TRANSACTIONS ON NEURAL NETWORKS, Vol. 12, Issue 4, pp. 929-935,July 2001.
[10]. S. Akkoç, "An empirical comparison of conventional techniques, neural networks and the three stage hybrFId Adaptive Neuro Fuzzy Inference System (ANFIS) model for credit scoring analysis : The case of Turkish credit card data", Elsevier Europezan Journal of Operational Research, Vol. 222, Issue 1, pp. 168–178, October 2012.
[11]. Sarwesh Site, Dr. Sadhna K. Mishra, " A Review of Ensemble Technique for Improving Majority Voting for Classifier", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 3, Issue 1, pp. 177- 180,January 2013.
[12]. A.R.Ghatge, P.P.Halkarnikar, "Ensemble Neural Network Strategy for Predicting Credit Default Evaluation", International Journal of Engineering and Innovative Technology, Vol. 2, Issue 7, pp. 223-225, January 2013.
[13]. Maher Ala'raj and Maysam Abbod,"A systematic credit scoring model based on heterogeneous classifier ensembles", Innovations in Intelligent Systems and Applications (INISTA), pp. 1-7, September 2015.
[14]. Marc Claesen, Frank De Smet, Johan A.K. Suykens and Bart De Moor, "A Library for Ensemble Learning Using Support Vector Machines", Journal of Machine Learning Research 15, pp. 141-145 , January 2014.
[15]. Gang Wang, Jian Ma, "Study of corporate credit risk prediction based on integrating boosting and random subspace", Elsevier Expert Systems with Applications, Vol. 38, Issue 11, pp. 13871–13878,October 2011.
[16]. Wo- Chiang Lee, "Genetic Programming Decision Tree for Bankruptcy Prediction", Joint Conference on Information Science, October 2006.

Authors

Anchal Goyal is a Research Scholar studied in Department of Computer Science & Engineering, RIMT–IET, Mandi Gobindgarh (Punjab). Her area of research is Machine Learning.

Ranpreet Kaur is working as Assistant Professor in Department of Computer Science & Engineering, RIMT–IET, Mandi Gobindgarh (Punjab). She has a teaching experience of more than 10 years. Her area of research is Digital Image Processing.

ISSN: 2456-1118