A survey on Ensemble Model for Loan Prediction

Anchal Goyal[1], Ranpreet Kaur[2]
Research Scholar[1], Assistant Professor[2]
Department of Computer Science,
RIMT –IET (PTU),Mandi Gobindgarh,
Punjab, India.

ABSTRACT

Extending credit to individuals is necessary for markets and society to function smoothly. Estimating the probability that an individual would default on their loan, is useful for banks to decide whether to sanction a loan to the individual or not. In this paper we discuss the ensemble model that is combination of two or more algorithms and give better results as compared to stand alone models. The performance is also enhanced through the ensemble model.

Keywords: *Ensemble, Boosting, Bagging, Stacking.*

1. INTRODUCTION

1.1 Introduction to Machine Learning

Machine learning is a field of computer science that involves the learning of pattern identification and computational learning theory in artificial intelligence. Machine learning generally refers to the changes in systems that carry out tasks linked with artificial intelligence (AI). Such tasks include recognition, analysis, planning, robot control, forecasting, etc. 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.

Machine learning can be broken into two categories:

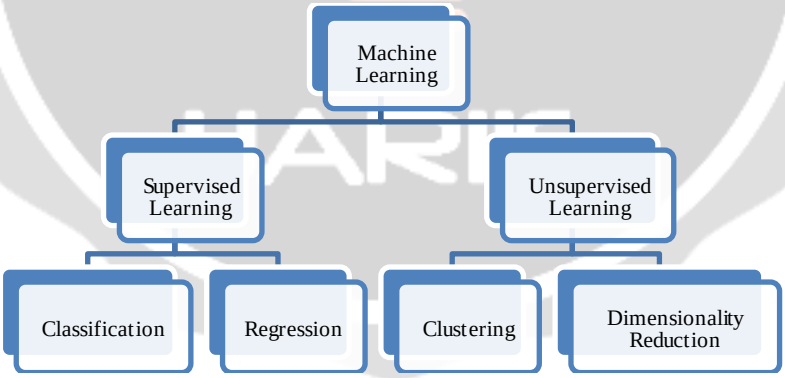

Fig-1 Categories of Machine Learning

a) **In** Supervised Learning, a data set includes both *features* and *labels*. The task is to build an estimator which is able to forecast the label of an object with the set of features. Supervised learning is further broken down into two parts: *classification* and *regression*.

Classification is the task of forecasting the value of a categorical variable given some input variables.

Regression is the task of forecasting the value of a continuously changeable variable (e.g. a price, a temperature) given some input variables.

***b)* In Unsupervised Learning**, a data set has no label and we find similarities among the objects. We can use this technique to display the best arrangement of data. It includes dimensionality reduction, clustering.

Dimensionality reduction is the task of derive a set of new features that is smaller than the original feature set while keep most of the variation of the original data.

Clustering is the method of gathering samples into groups of analogous samples according to some predefined similar or dissimilar measure.

1.2 Ensemble Model

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.

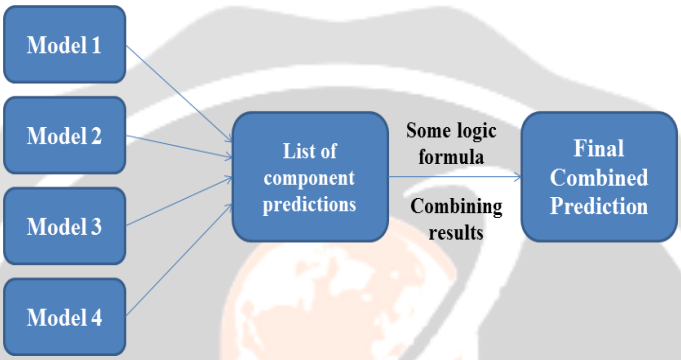

Fig-2 Ensemble Model

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

1.2.1 Ensemble Learning Algorithms

Bagging: Bagging stands for bootstrap aggregating. It is the one of the earliest model. It is the simplest ensemble based algorithm with a good performance. In the bagging algorithm each model has an equal weight in the ensemble vote. The class which have maximum votes are considered as final result for the given classification problem. In order to show the inconsistency, bagging trains each model with a randomly drawn subset of training dataset. Example: Random forest algorithm combine random decision tree with bagging to achieve the higher accuracy.

Firstly we have to create the random samples of the training data set and then build classifier for each sample. These several classifiers generate the final results using average voting system. This algorithm helps to reduce to variance error.

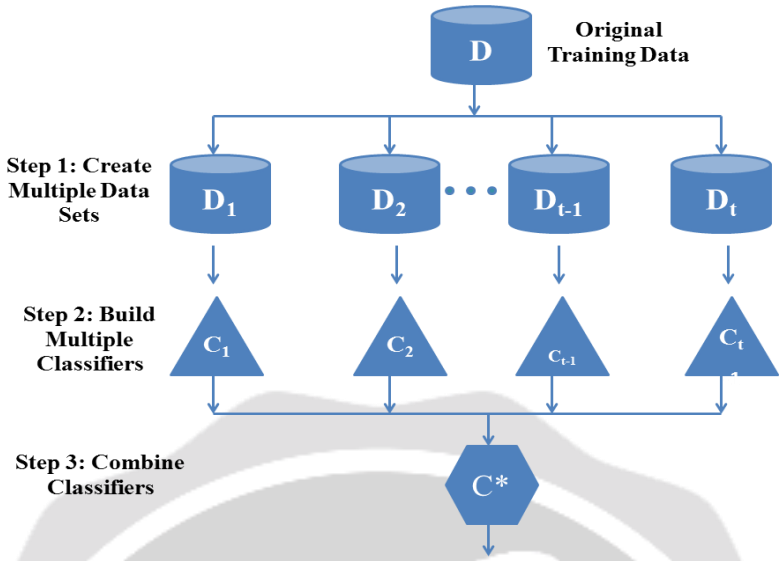

Fig-3 Bagging

Boosting: Boosting is another ensemble model. The term "Boosting" refers to a family of algorithm that converts weak learner to a strong learner. As similar to bagging, Boosting also generates ensembles of classifiers on different data that are then integrated by votes of majority. In the boosting algorithm each training set has some weight and the weight of each set is updated on each iteration. In some cases, boosting algorithm shows better results and accuracy as compared to bagging algorithm but the drawback of this algorithm is over fitting of the training data.

Stacking: Stacking is also known as stacked generalization. This techniques used on both type of learning either it is supervised learning or it is unsupervised learning. It includes a learning algorithm that combines the predictions of several other machine learning algorithms. Firstly, all other algorithms are trained over the given data set and then a combiner algorithm is used to make final prediction. The performance of this algorithm is better than the individual trained model. Basically it works on two phases: Firstly use the multiple base classifiers to forecast the class .Secondly, a new learner is used to combine their predictions and the aim is to reduce the error.

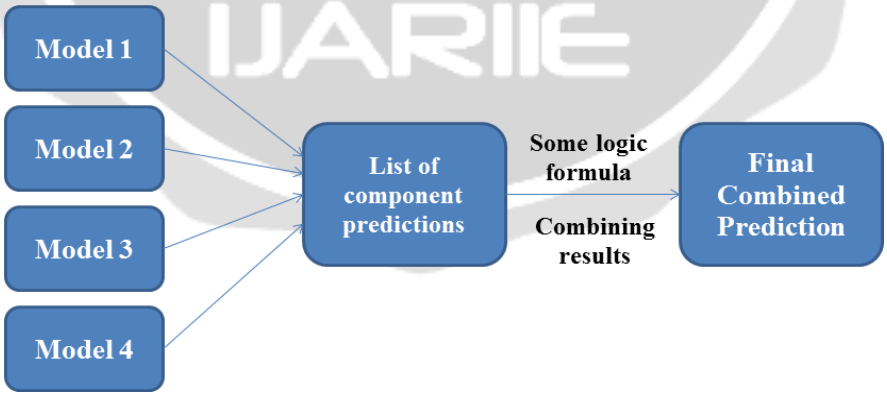

Fig4 -Stacking

Adaboost: Adaboost is very successful machine learning method. It is a type of ensemble learning where several learners are engaged to construct a stronger learning algorithm. It works on the basis on which a base algorithm is choose and then iteratively improving it for the given classified training dataset. In this algorithm, equal weights are assigned to all the training data and then choose a base algorithm. At each stage, base algorithm is applied to data set and increases the weights of incorrectly classified data. If it is not classified correctly then it doesn't change its weights. It is the algorithm that minimizes the error during learning.

Bucket of models: It is an ensemble method in which a model selection algorithm is used to select the best model for every problem. This model didn't generate better results if it is tested on one problem. If the number of problems is huge, then this model gives better results. The most general method used for model selection is cross validation selection. Cross validation is the method of analyzing and comparing machine learning algorithms by dividing the data into two segments. One segment is used to train a model and other segment is used to validate the model. The vital form of cross validation is K-fold cross validation in which data is first partitioned into k equally sized segments. Then k iterations are applied for training and validation of the data. Within each iteration, a part of data is held out for validation and remaining k-1 part are used for learning.

2. LITERATURE REVIEW

Sarwesh Site, Dr. Sadhna K. Mishra[1] proposed a method in which two or more classifiers are combined together to produce a ensemble model for the better prediction. They used the bagging and boosting techniques and then used random forest technique. The process of classifiers is to improve the performance of the data and it gives better efficiency. In this work, the authors describe various ensemble techniques for binary classification and also for multi class classification. The new technique that is described by the authors for ensemble is COB which gives effective performance of classification but it also compromised with noise and outlier data of classification. Finally they concluded that the ensemble based algorithm improves the results for training data set.

Amira Kamil Ibrahim Hassan, Ajith Abraham [2] 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.

A.R.Ghatge, P.P.Halkarnikar[3] develops the artificial neural network model for predict the credit risk of a bank. The Feed- forward back propagation neural network is used to forecast the credit default. They also compare the results with the manual calculations of the bank conducted in year 2004, 2005 and 2006. The results give the better and higher performance over manual calculations of bank.

Suresh Ramakrishnan, Maryam Mirzaei and Mahmoud Bekri[5] explores Adaboost ensemble method and makes an empirical comparison. The main goal is to compare ensemble classifiers. This study explores Ada Boost and bagging ensemble for default prediction to contrast with several classifiers including learning Logistic Regression (LR), Decision Tree (DT), artificial Neural Networks (NN) and support vector machine (SVM) as base learner.

Dr. A. Chitra and S. Uma[4] 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 that uses Mackey dataset, Sunspots dataset and Stock Price dataset as dataset and shows the proposed model performs better than the individuals. 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.

Alaraj, M. , Abbod, M.[6] 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.

Maher Alaraj, Maysam Abbod, and Ziad Hunaiti[7] 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.

Marc Claesen, Frank De Smet, Johan A.K. Suykens, Bart De Moor[8] proposed a model based on support vector machine that reduced the complexity of training data and predicts the model with high accuracy. This model is used to avoid duplicate storage of data.

Gang Wang, Jian Ma[9] 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.

M. Yaghini , T. Zhiyan , and M. Fallahi [11] presents a model that is based on feed forward neural networks to identify the bad costumers in the bank. They use three different strategies such as quick, dynamic and multiple strategies. To prevent the model from over fitting, cross validation is done on the model. To evaluate the proposed model, the result of neural network is compared with some common predictions methods namely decision tree and logistic regression. The results state that the three layer neural network based on the back propagation learning algorithm with quick strategy has higher accuracy.

Hussain Ali Bekhet , Shorouq Fathi Kamel Eletter[12] proposed two credit model namely logistic regression model and Radial basis function scoring model to support loan decision for Jordanian commercial banks using data mining techniques. The experimental result shows that the logistic regression model performed slightly better than the radial basis function model in terms of accuracy.

Table 1: Summary of Research Work Discussed

Title	Authors	Methods	Year
Modeling Consumer Loan Default Prediction Using Ensemble Neural Networks	Amira Kamil Ibrahim Hassan, Ajith Abraham	Neural Network	2008
An Ensemble Model of Multiple Classifiers for Time Series Prediction	Dr. A. Chitra and S. Uma	Pattern prediction Ensemble Model	2010
Study of corporate credit risk prediction based on integrating boosting and random subspace	Gang Wang, Jian Ma	boosting and random subspace	2011
A Prediction Model for Recognition of Bad Credit Customers in Saman Bank Using Neural Networks	M. Yaghini , T. Zhiyan , and M. Fallahi	Neural network	2011
Ensemble Neural Network Strategy for Predicting Credit Default Evaluation	A.R.Ghatge, P.P.Halkarnikar	Feed- forward back propagation neural network	2013
A Review of Ensemble Technique for Improving Majority Voting for Classifier	*Sarwesh Site, Dr. Sadhna K. Mishra*	Bagging, Boosting and Random Forest	2013
Credit risk assessment model for Jordanian commercial banks: Neural scoring approach	Hussain Ali Bekhet , Shorouq Fathi Kamel Eletter	Logistic regression model, Radial basis function scoring model	2014
Evaluating Consumer Loans Using Neural Networks Ensembles	Maher Alaraj, Maysam Abbod, and Ziad Hunaiti	Neural Network	2014
A Library for Ensemble Learning Using Support Vector Machines	Marc Claesen, Frank De Smet, Johan A.K. Suykens, Bart De Moor	Support Vector Machines	2014
A systematic credit scoring model based on heterogeneous classifier ensembles	Alaraj, M. , Abbod, M.	logistic artificial neural network, logistic regression and support vector machine	2015
Adaboost Ensemble Classifiers for Corporate Default Prediction	Suresh Ramakrishnan, Maryam Mirzaei and Mahmoud Bekri	Adaboost and Bagging	2015

3. CONCLUSION:

Ensemble Model gives the better prediction than the individual models. This model also enhances the performance and accuracy of the model. Through Ensemble model we compare the several models and choose the best model for our data that helps the organization to make the right decision for the loan request of the costumer.

4. REFERENCE:

[1] Sarwesh Site, Dr. Sadhna K. Mishra, " A Review of Ensemble Technique for Improving Majority Voting for Classifier", Volume 3, Issue 1, January 2013 .

[2] Amira Kamil Ibrahim Hassan, Ajith Abraham, "Modeling Consumer Loan Default Prediction Using Ensemble Neural Networks", Aug. 2013.

[3] A.R.Ghatge, P.P.Halkarnikar, "Ensemble Neural Network Strategy for Predicting Credit Default Evaluation", Volume 2, Issue 7, January 2013.

[4] Dr. A. Chitra and S. Uma, "An Ensemble Model of Multiple Classifiers for Time Series Prediction", Vol. 2, No. 3, June 2010

[5] Suresh Ramakrishnan, Maryam Mirzaei and Mahmoud Bekri, "Adaboost Ensemble Classifiers for Corporate Default Prediction", jan 2015.

[6]Alaraj, M., Abbod, M., "A systematic credit scoring model based on heterogeneous classifier ensembles",Sep2015

[7]Maher Alaraj, Maysam Abbod, and Ziad Hunaiti, "Evaluating Consumer Loans Using Neural Networks Ensembles", Jan. 8-9, 2014

[8]Marc Claesen, Frank De Smet, Johan A.K. Suykens, Bart De Moor, "A Library for Ensemble Learning Using Support Vector Machines", Jan 2014

[9]Gang Wang, Jian Ma, "Study of corporate credit risk prediction based on integrating boosting and random subspace", 2011.

[10] Wo- Chiang Lee, " Genetic Programming Decision Tree for Bankruptcy Prediction", JCIS-Oct 2006, 1951-6851.

[11] M. Yaghini , T. Zhiyan , and M. Fallahi, "A Prediction Model for Recognition of Bad Credit Customers in Saman Bank Using Neural Networks", 2011.

[12] Hussain Ali Bekhet , Shorouq Fathi Kamel Eletter , "Credit risk assessment model for Jordanian commercial banks: Neural scoring approach" ,Apr2014.

[13] Lim, T.-S., Loh, W.-Y., & Shih, Y.-S, "A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms." Machine Learning, 2000.

[14] Niculescu-Mizil, A., & Caruana, R, "Predicting good probabilities with supervised learning.", Proc. 22nd International Conference on Machine Learning 2015.

[15] H. Faris, B. Al-Shboul and N. Ghatasheh, "A Genetic Programming Based Framework for Churn Prediction in Telecommunication Industry", LNCS, vol. 8733, **2014.**

[16] D. Fantazzini and S. Figini, "Random Survival Forests Models For SME Credit Risk Measurement," Methodology and Computing in Applied Probability, vol. 11, no. 1,2009.