IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 3, March 2016

Loan Prediction Using Ensemble Technique

Anchal Goyal[1], Ranpreet Kaur[2]

Research Scholar, Computer Science, RIMT, Gobindgarh, India[1]

Assistant Professor, Computer Science, RIMT, Gobindgarh, India[2]

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. We introduce an effective prediction technique that helps the banker to predict the credit risk for customers who have applied for loan. A prototype is described in the paper which can be used by the organizations for making the correct or right decision for approve or reject the request for loan of the customers. The paper uses three different models (SVM Model, Random Forest Network and Tree Model for Genetic Algorithm) and the Ensemble Model, which combines these three models and analyses the credit risk for optimum results.

Keywords: Ensemble, Prediction, Loan, Random forest.

I. INTRODUCTION

Now a day's bank plays a vital role in market economy. The success or failure of organization 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.

Loan amount, costumer's history governs his creditability for receiving loan. The problem is to classify borrower as defaulter or non defaulter. However developing such a model is a very challenging task due to increasing in demands for loans. A prototype of the model is described in the paper which can be used by the organizations for making the correct or right decision for approve or reject the request for loan of the customers. This work includes the construction of an ensemble model by combining three different machine learning models.

II. RELATED WORK

Amira Kamil Ibrahim Hassan, Ajith Abraham (2008) uses a prediction model which is constructed using three different training algorithms to train a supervised two-layer feed-forward network. The results show that the training algorithm improves the design of loan default prediction model.

Angelini (2008) used a neural network with standard topology and a feed-forward neural network with ad hoc connections. Neural network can be used for prediction model. This paper shows that the above two models give optimum results with less error.

Ngai (2009) uses the classification model for predicting the future behaviour of costumers in CRM. In CRM domain, the mostly used model is neural network. He recognized eighty seven articles associated to data mining applications and techniques between 2000 and 2006.

Dr. A. Chitra and S. Uma (2010) introduced a ensemble learning method for prediction of time series based on Radial Basis Function networks (RBF), K - Nearest Neighbor (KNN) and Self Organizing Map (SOM). They proposed a model namely PAPEM which perform better than individual model.

Akkoç (2012) used a model namely hybrid Adaptive Neuro-Fuzzy Inference model, grouping of statistics and Neuro-Fuzzy network. A 10-fold cross validation is used for better results and a comparison with other models.

Sarwesh Site, Dr. Sadhna K. Mishra (2013) proposed a method in which two or more classifiers are combined together to produce an ensemble model for the better prediction. They used the bagging and boosting techniques and then used random forest technique.

Maher Alaraj, Maysam Abbod, and Ziad Hunaiti (2014) 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.

AlarajM , AbbodM (2015) introduced a 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.

III. PROPERTIES AND TECHNIQUES

A. Data set and its properties

Company wants to automate the loan eligibility process (real time) based on customer detail provided while filling online application form.

To automate this process, they have given a problem to Identify the customers segments; those are eligible for loan amount so that they can specifically target these customers. Here they have provided a partial data set.

DOI 10.17148/IJARCCE.2016.53128

The data set include 13 attributes such as Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and others. The data sets contain 615 records which is unfiltered data. The filtered Train Data set file contains 479 records. Table 1 gives information about the data set. The correlation between each feature is shown in Table 3.

TABLE1: DESCRIPTION OF THE FEATURES

Feature FID	Features	Information
FID 1	Loan ID	Unique Loan FID
FID 2	Gender	Male/ Female
FID 3	Married	Applicant married (Y/N)
FID 4	Dependents	Number of dependents
FID 5	Education	Applicant Education (Graduate/ Under Graduate)
FID 6	Self-employed	Self employed (Y/N)
FID 7	Applicant Income	Applicant income
FID 8	Co applicant Income	Co applicant income
FID 9	Loan Amount	Loan amount in thousands
FID10	Loan Amount Term	Term of loan in months
FID11	Credit History	credit history meets guidelines
FID12	Property Area	Urban/ Semi Urban/ Rural
FID13	Loan Status	Loan approved (Y/N)

B. Methodology

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 on H, Gini, AUC, AUCH, Accuracy etc.

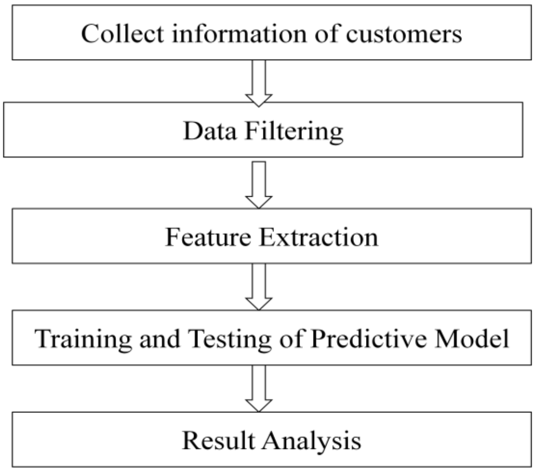
Collect information of customers
Data Filtering
Feature Extraction
Training and Testing of Predictive Model
Result Analysis

Figure 1: Methodology used

TABLE 2: SAMPLE DATASET

FID1	FID2	FID3	FID4	FID5	FID6	FID7	FID8	FID9	FID10	FID11	FID12	FID1 3
LP001003	Male	Yes	1	Graduate	No	4583	1508	128	360	1	Rural	0
LP001005	Male	Yes	0	Graduate	Yes	3000	0	66	360	1	Urban	1
LP001006	Male	Yes	0	Not Graduate	No	2583	2358	120	360	1	Urban	1
LP001008	Male	No	0	Graduate	No	6000	0	141	360	1	Urban	1
LP001011	Male	Yes	2	Graduate	Yes	5417	4196	267	360	1	Urban	1

TABLE3: CORRELATION BETWEEN EACH FEATURE

	FID2	FID3	FID4	FID5	FID6	FID7	FID8	FID9	FID10	FID11	FID12
FID2	1.00	0.33	0.20	-0.04	-0.01	0.02	0.16	0.05	-0.12	-0.05	0.07
FID3	0.33	1.00	0.36	0.01	0.05	0.06	0.04	0.15	-0.09	0.01	0.06
FID4	0.20	0.36	1.00	0.02	0.03	0.10	-0.03	0.11	-0.04	-0.05	0.01
FID5	-0.04	0.01	0.02	1.00	0.01	0.15	0.09	0.19	0.10	0.13	0.06
FID6	-0.01	0.05	0.03	0.01	1.00	0.18	-0.01	0.14	0.03	-0.04	-0.05
FID7	0.02	0.06	0.10	0.15	0.18	1.00	-0.10	0.48	0.02	-0.07	-0.07
FID8	0.16	0.04	-0.03	0.09	-0.01	-0.10	1.00	0.13	0.03	0.00	0.04
FID9	0.05	0.15	0.11	0.19	0.14	0.48	0.13	1.00	0.10	-0.01	-0.07
FID10	-0.12	-0.09	-0.04	0.10	0.03	0.02	0.03	0.10	1.00	0.06	-0.05
FID11	-0.05	0.01	-0.05	0.13	-0.04	-0.07	0.00	-0.01	0.06	1.00	-0.01
FID12	0.07	0.06	0.01	0.06	-0.05	-0.07	0.04	-0.07	-0.05	-0.01	1.00

TABLE 4: RCGA FEATURE IMPORTANCE

No. of executions	FID11	FID7	FID9	FID10	FID3	FID2	FID8	FID4	FID5	FID12	FID6
1	0.98	0.20	0.14	0.12	0.11	0.08	0.08	0.06	0.04	0.02	0.00
2	0.96	0.22	0.16	0.14	0.13	0.10	0.07	0.05	0.06	0.05	0.02
3	0.95	0.18	0.12	0.10	0.09	0.06	0.09	0.02	0.06	0.02	0.04
4	0.92	0.21	0.15	0.13	0.12	0.09	0.05	0.07	0.02	0.01	0.01
5	0.90	0.20	0.18	0.15	0.10	0.07	0.04	0.02	0.05	0.04	0.05
Average	0.94	0.20	0.15	0.13	0.11	0.08	0.07	0.04	0.05	0.03	0.02
Ranking	1	2	3	4	5	6	7	8	9	10	11

C. Feature Importance

The importance of each features calculated using RCGA. The objective function is used to assign weight to each feature. The coding is executed five times and weight thus obtained is shown in Table4.

IV. ANALYSIS OF MACHINE LEARNING TECHNIQUES

A. Machine learning techniques

Various machine learning techniques have been applied for prediction as explained below:

TABLE5: TECHNIQUES USED

Models	Method Used	Packages	Tuning Parameters
Bagged CART	bagging	Iperd	None
Random Forest	Random Forestb	Random Forest	mtry
Tree Model For Genetic Algorithm	evtree	Evtree	alpha
Decision Trees	rpart	Rpart	Min split, max depth, min bucket
Linear Model	multinom	Car, nnet	size
Neural Network	nnet	Nnet	Size, decay
SVM	ksvm	Kernlab	C
Extreme Learning Machine	elmtrain	elmNN	NhId, actfun
Multivariate Adaptive Regression spline	earth	Earth	degree
Bayesian Generalized Linear Model	bayesglm	Arm	None
Model Tree	tree	Tree	None

B. Model Analysis

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

DOI 10.17148/IJARCCE.2016.53128

1) 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:

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

2) 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.

3) 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.

$$\text{Gini}=2\text{AUC}-1 \qquad (2)$$

4) ROC Curve

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

5) 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.

6) MER

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

7) MWL

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

8) K-Fold Cross Validation

Cross-validation is a technique to evaluate predictive models by partitioning the data into a training set to train the model, and a test set to evaluate it. In k-fold cross-validation, the original sample is arbitrarily divided into k equal size subsamples .The advantage of this method is that it matters less how the data gets divided.

V. 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.

The results are presented in two parts. First part of this section displays accuracy rate for each of model (Table 6). Second part of this section shows the cross validation of predicted values of top three individual models having high accuracy and the ensemble model (Figure 2). Optimum results show that the ensemble model provides optimum results.

VI. 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.

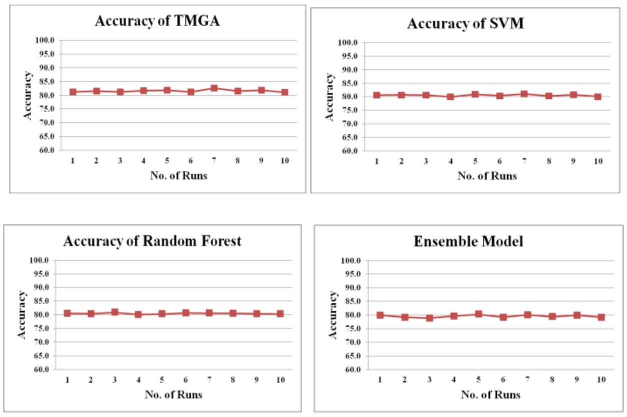

Figure 2: Cross Validation graph for Ensemble Model

TABLE 6: TRAINING DATA SET 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**

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.

[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.

[17]. M. Yaghini , T. Zhiyan , and M. Fallahi, "A Prediction Model for Recognition of Bad Credit Customers in Saman Bank Using Neural Networks", Int'l Conf. Data Mining , 2011.

[18]. Hussain Ali Bekhet , Shorouq Fathi Kamel Eletter , "Credit risk assessment model for Jordanian commercial banks: Neural scoring approach" , Review of Development Finance,Vol.4, Issue 1, pp. 20-28, April2014.

DOI 10.17148/IJARCCE.2016.53128