INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 5, Issue 2)
Available online at: www.ijariit.com

An adaptive approach to prognosticate an individual's capability for emolument through Machine Learning

Jujjuri Goutham
jujjuri.9194@gmail.com
Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, Andhra Pradesh

T. Anitha
anitha.cse@anits.edu.in
Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, Andhra Pradesh

S. Joshua Johnson
joshua.cse@anits.edu.in
Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, Andhra Pradesh

Routhu Dhanunjay
routhudhanunjay@gmail.com
Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, Andhra Pradesh

Vemuri Susmitha
vsusmitha1@gmail.com
Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, Andhra Pradesh

Nagavarapu Sravani
nsravani1205@gmail.com
Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, Andhra Pradesh

ABSTRACT

One of the major integrant to be considered while granting a loan is the customer's ability to pay back the amount to the bank as per the banks provided a schedule. Our work focuses on the analysis of all the attributes that might affect the customer's ability to pay the loan. It is basically a credit scoring mechanism used by the bank to make sure a customer's intentions to apply for a loan are legit using Ensemble Algorithms. Our work gives a probabilistic predictive model or a scorecard to estimate the probability of defaulters in the current global scenario. Our work is due diligence fulfilled by the investors involved with the bank. Our aim is to prognosticate correct credit worth which will cause a significant increment in the profits of commercial institutions.

***Keywords**— Random forest classifier, Defaulters, Mathew's correlation coefficient, Credit scoring*

1. INTRODUCTION

The aim of our project is to reform our current banking system to control and diminish the consistent rise of frauds. This paper addresses the problem of Bank Indessa which has not done well in the last 3 quarters, NPA (Non-Performing Assets) are quite high which is caused by loan defaulters. It is our responsibility to identify the capability/probability of the a person to pay back the loan in other words we predict the chance of fraud that might happen basing on the credentials of a singular person for example, consider a person having a few public derogatory records and is associated with a few crimes, such a person has a greater chance of committing fraud than a person who has no public derogatory records and is not associated with any crime. Consuming goods on credit led to a rather unstable economic situation. To attend to this issue we propose a reliable model to reduce operating risk and cost. Even a small improvement in identifying the correct credit worth would largely improve the gain of financial institutions. It identifies the risks associated with the loan requests with respect to their characteristics such as income, age and occupation. In most of the researches, only a single classification model is built for loan default prediction. So, it is much better to compare the performance of several ensemble classification models for loan defaulting on similar criteria. The aim of the study is to compare the performance of a wide range of a classification technique in the credit scoring each applicant of the loan.

Machine Learning (ML) is a category of an algorithm that allows software applications to become more accurate in predicting outcomes without being explicitly programmed. The basic premise of machine learning is to build algorithms that can receive input data and use statistical analysis to predict an output while updating outputs as new data becomes available.

Many people are familiar with machine learning from shopping on the internet and being served ads related to their purchase. This happens because recommendation engines use machine learning to personalize online ad delivery in almost real time. Beyond personalized marketing, other common machine learning use cases include fraud detection, spam filtering, network security threat detection, predictive maintenance and building news feeds. For this problem, Random Forest classifier is the best-used classification because it uses many decision tree models to predict the result. It uses the Bagging method for predicting the outputs. The predictions it builds is an ensemble of decision trees.

1.1 Random Forest Classifier

Random Forest is a flexible and easy to use machine learning algorithm and also one of the best widely used Classifier which uses the Bagging method where a set of base classifiers are combined to fit the data. The random forest can be used for both classification and regression problems which is the main advantage in it. In bagging the train data is divided into subparts in which each part is trained with one base classifier (decision tree). It is a combination of learning methods which increases the entire result. The Random forest classifier adds the randomness as additional while growing the trees. While splitting a node instead of searching for the most important feature, it searches for the best feature among a random subset of features.

2. RELATED WORK

Over the past few years, several classification techniques have been developed for credit rating and loan default prediction; relating to this, we studied various latest improvements in the loan applicants' classification. In 2015, Lessmann has introduced a benchmarking study with 41 classifiers on eight credit scoring datasets and various ensemble selection techniques. The accuracy was measured by several indicators.

A literature survey considering the theories and applications of many binary classification methods have been discussed. The result shows the importance and use of these methods. An ensemble classification technique is proposed and used on the supervised mechanism with the K-NN approach. The proposed method improves the accuracy of the previously proposed model. An adaptive approach is performed on the base classifiers for the different ensemble methods to improve the classification task has been carried out.

3. EXPERIMENTAL SETUP AND DATA SETS

Loan default prediction is a group of decision models and their included techniques that act as a helping hand to the lenders when providing a loan to customers. In order to prognosticate an individual's capability of emolument, we have performed ensembles giving out the probability of an individual by which they can pay back a loan.

4. IMPLEMENTATION

The implementation of the several machine learning algorithms is categorized into two types namely individual and ensemble. Figure 1 shows the implementation of individual and ensemble-based classifiers. 10-fold cross validation is used in all the classifications where the original dataset is divided into 10 folds where 9 out of 10 folds are combined and are used as a training set; the remaining one fold is used as the testing set. Later, every classifier is applied separately to these data sets. A probability is given by the model at the end, greater the probability of loan default.

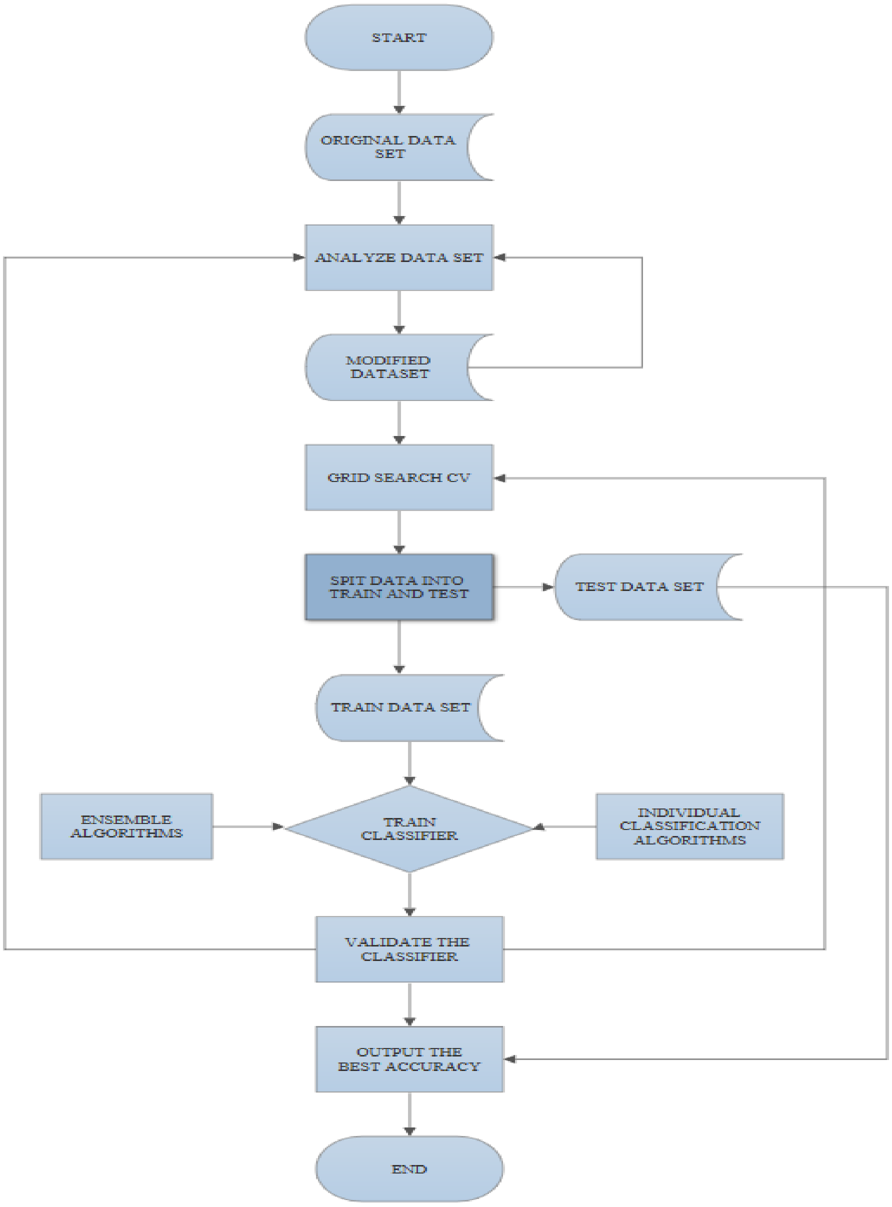

Fig. 1: System architecture

4.1 Individual Classifiers

In individual classification, all the various categories of classifiers are used for evaluation. Various categories are mainly from Logistic regression, KNN algorithm. We use individual classifiers to provide a stark contrast between the accuracy acquired utilizing individual and ensemble classifiers.

4.2 Ensemble classifiers

They use multiple learning algorithms to obtain better predictive performance than individual learning algorithms. Ensembles perform better when there is large diversity in the models. Ensemble algorithms are a divide-and-conquer approach which is used for the betterment of performance. The most significant feature in the stream of research in supervised learning is to study the techniques to build better ensemble classifiers. The main idea is that ensemble method are generally much more exact than the individual classifiers of which they are built. The principle of ensemble classifiers stated as 'a group of "weak learners" that can merge together and can form a "strong learner". Singularly, each classifier is considered a "weak learner," where all the combined classifiers are considered "strong learners".

The ensemble used for the creation of the model is: Bagging with a base of fast decision tree classifiers i.e., random forest.

4.3 Dataset characteristics

Our adaptive approach includes lending club credit dataset. The data set is obtained from the Kaggle online community owned by Google. The summary for each data set is provided in Table***, data sets are binary class and slightly lean towards the "good" class i.e. 'the customers who are not defaulting'. Our study includes the classification of the original data sets.

4.4 Performance indicators

A variety of indicators to measure predictive accuracy are available. We consider following accuracy indicators in our study: the percentage correctly classified (ACC), the area under a receiver-operating-characteristics (ROC) curve, the TP rate, FP rate Precision, Recall and F-measure. We chose these indicators because they assess the predictive performance of a scorecard from different angles. Table shows the configuration of a confusion matrix and the formulations of other accuracy parameters that were used to assess the classifiers used in our study.

$$\textbf{Precision} = tp / (tp+fp) \qquad (1)$$

The precision is in the equation (1) where tp = true positive and fp = false positive. Precision focuses on the class cooperation of the data labels with the positive labels given by the classifier

$$\textbf{Recall} = tp / (tp+fn) \qquad (2)$$

Where tp=true positive and fn=false negatives. Recall evaluated the effectiveness of a classifier to identify positive labels.

$$\textbf{Fmajor} = 2*(precision*recall)/(precision+recall) \qquad (3)$$

F1 score can be attained as a weighted average of precision and recall.

$$\textbf{Accuracy} = (tp+tn) / (tp+tn+fp+fn) \qquad (4)$$

Accuracy shows the overall effectiveness of a classifier.

AUC provides the classifier's ability to avoid false classification. Table shows all machine learning algorithms and their abbreviations, used in this study.

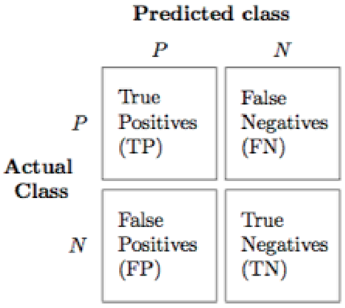

Fig. 2: Predicted class and actual class

5. RESULT AND DISCUSSIONS

The Random forest credit scoring model was successful in classifying default and non-default loans. Hence, commercial lenders can reduce the risk of investment failure by selecting profitable borrowers after processing loan applications through the model. This model correctly classified the default and no-default is 89% in the test data set.

Table 1: Classification Result

S.no.	Percentage split	Accuracy	Error rate
1.	80%:20%	0.89	0.0903
2.	70%:30%	0.85	0.0913
3.	60%:40%	0.83	0.0916
4.	50%:50%	0.87	0.0911

Figure 3 shows the minimum average error rate. The graph is drawn with the iteration level on the x-axis and the percentage of error rate on the y-axis.

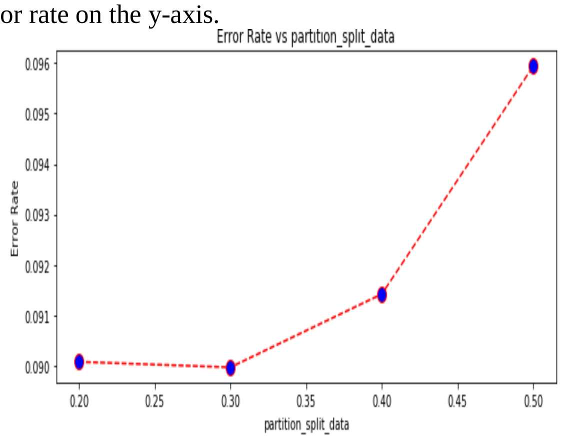

Fig. 3: Minimum average rate graph

The applied Random forest credit scoring model successfully demonstrates the applicability of Random forest classifier in credit scoring for classification and prediction of the loan status. Machine Learning techniques are used to develop the credit scoring system. The Bagging method of Random forest classifier provides higher accuracy than other classifiers. Fig 3 shows the minimum error rate of training data. If the partition split data is increased then the error rate is gradually decreased. This minimum error rate based credit scoring model provides significant results for predicting the loan status in the commercial banks. This proposed model presented in this study can be effectively used by commercial loan lenders to predict the loan applicant. Lenders can use this model to predict the loan status of the loan applicant.

6. REFERENCES

[1] The general introduction to the classification and related theory can be found in Khashman (2010), Elizondo (2006) or in a great book focused on statistical learning Hastie Et al. (2013).

[2] Arutjothi, G., & Senthamarai, C. (2017) Prediction of loan status in the commercial bank using machine learning classifier. 2017 International Conference on Intelligent Sustainable Systems (ICISS) doi:10.1109/ iss1.2017.8389442

[3] Zhang, C. (2011). Whether education has an effect on loan defaults: A theoretical and empirical study. 2011 IEEE 18th International Conference on Industrial Engineering and Engineering Management. doi:10.1109/ icieem.2011.6035516.

[4] Vaidya, A. (2017). A predictive and probabilistic approach using logistic regression: Application to the prediction of loan approval. 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT). doi:10.1109/ icccnt.2017.8203946

[5] Sutrisno, H., & Halim, S. (2017). Credit Scoring Refinement Using Optimized Logistic Regression. 2017 International Conference on Soft Computing, Intelligent System and Information Technology

[6] Singh, P. (2017). Comparative study of Individual and ensemble methods of Classification for credit scoring. 2017 International Conference on Inventive Computing and Informatics (ICICI).