COMPREHENSIVE RESEARCH IN EMERGING TECHNOLOGIES
https://choicemade.in/cret/ Volume 1 issue 1

A Prediction model for BankruptcyUsing Machine learning Techniques

Vardan Kumar[1], Vaibhav Patel [2] , Anurag Shrivastava[3]

[1]*M.Tech Scholar, Department of Computer Science, NIIST, Bhopal*
[2]*Assistant professor, Department of Computer Science, NIIST, Bhopal*
[3]*Associate professor, Department of Computer Science, NIIST, Bhopal*

Abstract - Financial depression and then the resultant failure of a business is usually an extremely costly and disrupting event for any company and organization. Statistical predictions of financial depression models try to predict whether a business will experience financial failure in the future. Discriminate analysis and logistic regression have been the most popular approaches which is used, but there is also a large number of data mining and machine learning techniques that can be used for this purposed. In this research work a classification method are proposed for bankruptcy prediction. This prediction used naïve bayes machine learning algorithms. This model uses the preprocessing and feature selection technique for improving classification performance. The result obtained shows that the predictive model has improved prediction model performance.

Keywords: Bankruptcy, Bankruptcy Prediction Model, Machine Learning, Classification, Classification framework

1. INTRODUCTION

Bankruptcy prediction has become a most popular research interestsover the past decade since it can have significant impact onfinancial, corporate, government and banking decision-making process. Accuracy is oneof crucial performance due to its significant economic impactnumerous statistical techniques have been used for improving the performance of bankruptcy prediction models. Bankruptcy prediction has become most important topic forresearchers and companies by using various models. The artificial neuralnetwork, support vector machine and many machine learning algorithmsare used for this purpose.Basically, there are two approaches to predict the bankruptcy: univariate analysis andmultivariate analysis. univariate analysis used to predict financial distress which is the distribution offinancial variables for companies that experiencing financial distress is different from companies thatdon't have financial distress. Deficiency of this model is contradiction between the predictedvariables. To solve this problem, multivariate models was developed. The independent variables inthis model are the financial ratios that expected to affect bankruptcy, while the dependent variable isthe prediction results. But till now, little theoretical discussion only that leads to bankruptcy research,e.g. in the selection of variables that are considered relevant. With at least the theory, bankruptprediction is more directed to the search for variables that are considered relevant to the trial and errormethods [25]. Financial fraud is a growing concern with far reaching consequences in thegovernment, corporate organizations, finance industry, In Today's world highdependency on internet technology has enjoyed increased credit cardtransactions but credit card fraud had also accelerated as online and offlinetransaction.[26]

2. LITERATURE REVIEW

An author [1] applies anomaly detection algorithms to the bankruptcy predictionproblem in an attempt to suggest a new stable model taking the data distribution into consideration. The efficacy of anomalydetection techniques is tested on bankruptcy prediction datasetsPolish banks. Author's empirical evaluation shows that IsolationForest outperforms multivariate Gaussian distribution, oneclassSVM, and other classification estimators in terms of theROC curve.

According to the authors [2] Selection of dataset for training the prediction model,the Machine Learning tool used for prediction and various otherfactors are essential in building an efficient prediction model. Thedataset includes financial ratios as attributes that are derivedfrom the financial statements of various companies. The most
Influencing ratios that are required for predicting bankruptcyare selected on the basis of the Genetic Algorithm which filtersout the most important ones from different existing bankruptcymodels. These ratios of different companies are fed as an input totrain the model being implemented in R. The predictionalgorithm used is Random Forest, which will enable us todifferentiate between bankrupt and non-bankrupt companies.

The proposed [3] algorithm is successfully applied in the bankruptcyprediction problem, where experiment data sets are originally from the UCI Machine Learning Repository. The simulation resultsshow the superiority of proposed algorithm over the traditional SVM-based methods combined with genetic algorithm (GA) orthe particle swarm optimization (PSO) algorithm alone.

Authors [4] reduce the imbalance nature of data and then train a deep neural network with the balanced data. The dataset used is that of polish companies which consists of five years of data corresponding to five years of five different tasks. Authors reduce class imbalance using an oversampling method known as SMOTE. Author's model significantly outperforms the previous neural network models and weak learners trained this dataset in terms of AUC.

In this research, authors [5] propose the implementation of Jordan Recurrent NeuralNetworks (JRNN) to classify and predict corporate bankruptcy based on financial ratios.Feedback interconnection in JRNN enables to make the network keep important informationwell allowing the network to work more effectively. The result analysis showed that JRNNworks very well in bankruptcy prediction with average success rate of 81.3785%.Neural Networks can process a tremendous amount of attribute factors; it results in overfitting frequently whenmore statistics is taken in. By using K-Nearest Neighbor and Random Forest, authors [6] obtain better results from different perspectives. Research [6] testifiesthe optimal algorithm for bankruptcy calculation by comparing the results ofthe two methods.

2.1 Dataset

The dataset was imported from UCI Machine Learning Repository [33]. The dataset consists of 64 calculated ratios which are obtained from the companies' financial annual report, including profit and loss statement and income statement. The target value is categorical with 1 means "bankrupt" and 0 for "non-bankrupt". The data was also collected for surviving companies. The size of the files is different, as well as the percentage of the bankruptcy instances. For Example, year 1 consists of 5910 instances while bankruptcy makes only 6.9% of the data.

Dataset	No. of Features	Total Instances	No. of Instances Bankrupt	No. of Instances non-bankrupt
bankruptcy data	64	5910	410	5500

Table 1: Details of Dataset

2.2 Preprocessing

Raw data is highly susceptible to noise, missing values,and inconsistency. The quality of data affects the data mining results. In order to help improve the quality of the data and consequently, of the mining result raw data is pre-processing is one of the most critical steps in a data process which deals with the preparation and transformation of the initial dataset.

2.3 Feature selection

Feature selection is an essential step to create an accuratepredictive model. There are four types of features:predictive, interacting, redundant and irrelevant [28].Predictive features provide useful information to predict thetarget. Interacting features are useful only when combinedwith other features but not by themselves. RedundantFeatures are features that have a strong correlation with otherfeatures. Irrelevant features are useless and don't provide anyinformation to predict the target value. Removing irrelevant andredundant features improve the prediction models byfocusing only on the features that are correlated to the targetvalue. In this study, finding themost important features has an economic importance becausecompanies can evaluate their performance by focusing onthose features.There are different methods to identify the key features.Each method has its pros and cons, but we observed thateach method identifies different features to be the mostimportant. In this study, we tested three techniques andcompared them based on the results of the prediction models.New naïve bayes algorithms used log probabilities. A log probability is simply the logarithm of a probability. The use of log probabilities means representing probabilities in logarithmic space, instead of the standard [0, 1] interval. In most machine learning tasks we actually formulate some probability p which should be maximized, here we would optimize the log probability log(p) instead of the probability for class θ. The use of log probabilities is widespread in several fields of computer science such as information theory and natural language processing etc.

3. Proposed framework

The framework proposed in this work is depicted in Figure 2. The proposed framework for prediction works for each transaction and separates the transaction with high or low risk using the method proposed. The proposed predictive model can be further used to generate alerts for transaction with high risks. Investigators check these alerts and provide a feedback for each alert, i.e.

true positive or false positive. The proposed model uses suitable pre-processing, attributes selection techniques along with proposed new naïve bayes machine learning algorithm.

4. EXPERIMENTAL SETUP, METHODOLOGY

4.1 Experimental Setup

Weka 3.8.1 is used as DM tool for simulation purpose. Weka is installed over Windows 10 Operating System. For this research a state of art research dataset from UCI Machine Learning Repository [29] is used. Dataset description is presented in Table 1.The experiment methodology starts with preprocessing process that remove redundancy, missing values, and inconsistency of used raw dataset. In this experiment using "all filters" from weka preprocess window after that select best feature. In second step feature selection steps applying the "CfsSubsetEval" evaluator and best first search from supervised attribute inweka preprocess window. Proposed algorithm uses in classification. Two different classifiers like naïve bayes and J48 are used for compare result with spilt 60% data. In last evaluate the result on the basis of accuracy of the proposed model and error rate.

4.2 Result Analysis

The performance analysis is done on the basis of following metrics: Accuracy and Error rate. In this experiment new naïve bayes algorithm give 93.43% accuracy and 7.3296% of error rate. Two different classifier first naïve bayes gives 90.23% accuracy and 9.85% of error rate. Another one J48 gives 92.0051% of accuracy and 7.99% of error rate.

5. CONCLUSION & FUTURE WORK

Bankruptcy Prediction is becoming the most important issue now days. This opens new confronts in the field of bankruptcy detection and prevention, but prevention is of course better than detection.It is helpful for financial organization to make decision sopredictive models are of prime importance for banks and financial organization to prevetion of bankcruptcy. In this research paper propose apredictive model which is based on classification model based on ML techniques for improving the performance of prediction model. The proposed work is compared on basis of two functional parameters: accuracy and error rate proved to be better. The efforts shown that,proposed methods are more suitable for detecting frauds. In future, more efforts methods will be worked out to improve the Fraud Catching Rate.

6. REFERENCES

[1] Shuoshuo Fan, Guohua Liu, Zhao Chen "Anomaly detection methods for bankruptcy prediction" The 2017 4th International Conference on Systems and Informatics (ICSAI 2017) IEEE

[2] Shreya Joshi1, Rachana Ramesh 2, Shagufta Tahsildar3 "A Bankruptcy Prediction Model UsingRandom Forest" (ICICCS 2018)IEEE Xplore

[3] Shanmukha Vellamcheti, Pradeep Singh "Class Imbalance Deep Learning for Bankruptcy Prediction"978-1-7281-4997-4/20/$31.00 ©2020 IEEE

[4] John Garcia "Bankruptcy prediction using synthetic sampling" https://doi.org/10.1016/j.mlwa.2022.100343
Received 3 November 2021; Received in revised form 19 April 2022; Published by Elsevier Ltd.

[5] M. Krivko, "A hybrid model for plastic card fraud detection systems," Expert Systems with Applications, vol. 37, no. 8, pp. 6070–6076, Aug. 2010.

[6] Benson Edwin Raj, A. Annie Portia, "Analysis on Credit Card Fraud Detection Methods", IEEE International Conference on Computer, Communication and Electrical Technology – ICCCET2011, 978-1-4244-9394-4/11, 2011 IEEE.

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

[8] L. Breiman, "Bagging predictors," Machine Learning, vol. 24, no. 2, pp. 123–140, 1996.

[9] 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).

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

[11]Jiwani, N., & Gupta, K. (2018). Exploring Business intelligence capabilities for supply chain:a systematic review. Transactions on Latest Trends in IoT, 1(1), 1-10. Retrieved from https://www.ijsdcs.com/index.php/TLIoT/article/view/136-

[12] Masoumeh Zareapoor, Pourya Shamsolmolia, "Application of Credit Card Fraud Detection: Based on Bagging Ensemble Classifier", International Conference on Intelligent Computing, Communication & Convergence, (ICCC 2015), Elsevier, Procedia Computer Science 48 (2015) 679 – 685.

[13]https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4180893/

[14] http://www.cs.waikato.ac.nz/~ml/weka/index.html

[15] Weka, University of Waikato, Hamilton,New Zealand.

[16] V.Mareeswari, Dr G. Gunasekaran, "Prevention of Credit Card Fraud Detection based on HSVM", IEEE, International Conference On Information Communication And Embedded System (ICICES 2016), 978-1-5090-2552-7.
[17] Alejandro Correa Bahnsen, Djamila Aouada, Aleksandar Stojanovic and Bj¨orn Ottersten, "Detecting Credit Card Fraud using Periodic Features", IEEE 14th International Conference on Machine Learning and Applications, 978-1-5090-0287-0/15, 2015 IEEE.

[18] European Central Bank, "Third report on card fraud," European Central Bank, Tech. Rep., 2014.

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[20] Benson Edwin Raj, A. Annie Portia, "Analysis on Credit Card Fraud Detection Methods", IEEE International Conference on Computer, Communication and Electrical Technology – ICCCET2011, 978-1-4244-9394-4/11, 2011 IEEE.

[21] Alejandro Correa Bahnsen, Aleksandar Stojanovic, Djamila Aouada and Bj¨orn Ottersten, "Cost Sensitive Credit Card Fraud Detection using Bayes Minimum Risk", 12th International Conference on Machine Learning and Applications 2013, 978-0-7695-5144-9/13, 2013 IEEE.

[22] Jiwani, N., & Gupta, K. (2019). Comparison of Various Tools and Techniques used for Project Risk Management. International Journal of Machine Learning for Sustainable Development, 1(1), 51-58. Retrieved from https://ijsdcs.com/index.php/IJMLSD/article/view/119-

[23] Marwan Fahmi, Abeer Hamdy, Khaled Nagati, "Data Mining Techniques for Credit Card Fraud Detection: Empirical Study", Sustainable Vital Technologies in Engineering & Informatics 2016, Published by Elsevier Ltd.

[24] Wen-Fang YU, Na Wang, "Research on Credit Card Fraud Detection Model Based on Distance Sum", International Joint Conference on Artificial Intelligence 2009, 978-0-7695-3615-6/09, 2009 IEEE.

[25] T. G. Dietterich, "Machine-learning research: four current directions," AI Magazine, vol. 18, no. 4, pp. 97–136, 1997.

[26] R. O. Duda, P. H. Hart, and D. G. Stork, Pattern Classification, Wiley-Interscience, New York, NY, USA, 2000.

[27] R. Bryll, R. Gutierrez-Osuna, and F. Quek, "Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets," Pattern Recognition, vol. 36, no. 6, pp. 1291–1302, 2003.

[28] K. Tumer and N. C. Oza, "Decimated input ensembles for improved generalization," in Proceedings of the International Joint Conference on Neural Networks (IJCNN '99), pp. 3069–3074, Washington, DC, USA, July 1999.

[29] T. Hastie and R. Tibshirani, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2009.

[30] Kalyan Nagaraj and Amulyashree Sridhar "A predictive system for detection ofbankruptcy using machine learningtechniques"International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015

[31] Lingga Hardinata1 , Budi Warsito1, Suparti1 Bankruptcy prediction based on financial ratios using JordanRecurrent Neural Networks: a case study in Polish companiesIOP Conf. Series: Journal of Physics: Conf. Series 1025 (2018) 012098 doi :10.1088/1742-6596/1025/1/012098

[32] Wenhao Zhang Machine Learning Approaches to Predicting Company Bankruptcy Journal of Financial Risk Management,2017,6,364-http://www.scirp.org/journal/jfrm ISSN Online: 2167-9541 ISSN Print: 2167-9533

[33] Björn mattsson & olof steinert corporate bankruptcy prediction using machine learning techniques department of economics university of gothenburg school of business economics and law,2017

[34] Duaa Alrasheed1 , Dongsheng Che1 Improving Bankruptcy Prediction Using Oversampling and Feature Selection Techniques Int'l Conf. Artificial Intelligence | ICAI'18 |

[35]Available:https://archive.ics.uci.edu/ml/datasets/Polish+companies+bankruptcy+data

[36]Jacky C. K. Chow "analysis of financial credit risk using machine learning"Aston UniversityBirmingham, United Kingdom April 2017

[37] Kalyan Nagaraj and Amulyashree Sridhar"a predictive system for detection of bankruptcy using machine learning techniques" International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015