INTERNATIONAL RESEARCH JOURNAL OF ENGINEERING & APPLIED SCIENCES
ISSN: 2322-0821(O) VOLUME 7 ISSUE 1
ISSN: 2394-9910(P) Jan 2019 - Mar 2019
www.irjeas.org

IMPROVED BANKRUPTCY PREDICTION SYSTEM USING NEW NAÏVE BAYES ALGORITHM

Surabhi Girare[1], Vaibhav Patel[2], Anurag Shrivastava[3]

[1] *M.Tech Scholar, CSE, NIIST, Bhopal*
[2] *Assistant Professor, CSE, NIIST, Bhopal*
[3] *Associate Professor, CSE, NIIST, Bhopal*

ABSTRACT - Prediction of Bankruptcy is the most important task. The techniques to assist fraud investigators, for banks and other financial organizations, rely on machine learning algorithms. Proposing a predictive model for Fraud determination is however mainly exigent due to the highly distributed data and the availability of only few transactions labeled as fraud in overall transactions. To seek out whether the transaction is fraud on E-commerce websites, is role of prediction models. To find out such transaction can be treated as a sort of machine learning (ML) problem. Many researches use machine learning techniques to improve performance of prediction and classification accuracy. In this paper find classification methods that are helpful in building model for bankruptcy prediction. Further in this paper, a predictive model for improved bankruptcy based on new naïve bayes machine learning is proposed. New naïve bayes classifier is used for better accuracy. The dataset imported from UCI Machine Learning Repository. The results obtained shows that the predictive model has potential in determining fraud and minimizing the risk in e-commerce transactions. The paper directs about the future research in the field.

Keywords - Bankruptcy, Bankruptcy Prediction, E-commerce, Machine Learning, Financial Organization, Classification

1. INTRODUCTION

Bankruptcy prediction has been an important and widely topic in accounting and finance because it's significant impact on management, employees, stockholders, and nation. Accuracy is one of crucial performance due to its significant economic impact numerous statistical techniques have been used for improving the performance of bankruptcy prediction models. Bankruptcy prediction has become most important topic for researchers and companies by using various models. The artificial neural network, support vector machine and many machine learning algorithms are used for this purpose. Basically, there are two approaches to predict the companies bankruptcy: univariate analysis and multivariat analysis. Univariate analysis used to predict financial distress which is the distribution of financial variables for companies that experiencing financial distress is different from companies that don't have financial distress. Deficiency of this model is contradiction between the predicted variables. To solve this problem, multivariate models was developed. The independent variables in this model are the financial ratios that expected to affect bankruptcy, while the dependent variable is the 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, bankrupt prediction is more directed to the search for variables that are considered relevant to the trial and error methods [25]. Financial fraud is a growing concern with far reaching consequences in the government, corporate organizations, finance industry, In Today's world high dependency on internet technology has enjoyed increased credit card transactions but credit card fraud had also accelerated as online and offline transaction.

In this paper, different machine learning techniques are employed to predict bankruptcy. The support system can be utilized by stock holders and investors to predict the performance of a company based on the nature of risk associated.

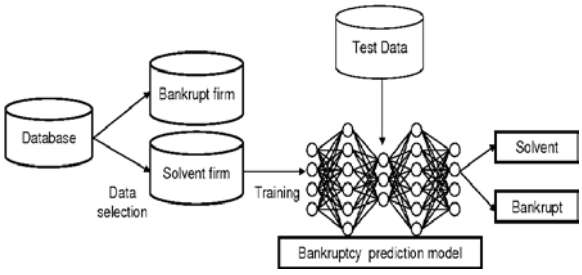


Figure 1: Bankruptcy prediction model

2. LITERATURE REVIEW

There is a continuous attention in finding better methods to predict bankruptcy because many financial decisions can be made based on the result of such methods. The machine learning techniques are increasingly being developed to improve the prediction of bankruptcy. Author [28] tested six major ML algorithms for predicting bankruptcy which are Neural Networks, Decision Trees, Random Forests, Support Vector Machine, K-Nearest Neighbor and Logistic Regression. Random Forest, Decision Tree, and KNN were found to be the best techniques for such problem as they produced higher prediction accuracy. In [25], authors propose the implementation of Jordan Recurrent Neural Networks (JRNN) to classify and predict corporate bankruptcy based on financial ratios. Feedback inters connection in JRNN enable to make the network keep important information well allowing the network to work more effectively. The result analysis showed that JRNN works very well in bankruptcy prediction with average success rate of 81.3785%. Authors [26] said Machine Learning is important assistance, and many companies would use Neural Network, a model in bankruptcy prediction, as their guide to prevent potential failure. However, although Neural Networks can process a tremendous amount of attribute factors, it results in over fitting frequently when more statistics is taken in by using K-Nearest Neighbor and Random Forest; Authors obtain better results from different perspectives. Author [26] testifies the optimal algorithm for bankruptcy calculation by comparing the results of the two methods.

3. PROPOSED WORK

3.1 Dataset

The dataset was imported from UCI Machine Learning Repository [29]. 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

3.2 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.

3.3 Feature selection

Feature selection is an essential step to create an accurate predictive model. There are four types of features: predictive, interacting, redundant and irrelevant [28]. Predictive features provide useful information to predict the target. Interacting features are useful only when combined with other features but not by themselves. Redundant Features are features that have a strong correlation with other features. Irrelevant features are useless and don't provide any information to predict the target value. Thus, we try to identify those features in order to find the best subset that gives the best prediction results. Removing irrelevant and redundant features improve the prediction models by focusing only on the features that are correlated to the target value. This also leads to avoiding over fitting which makes the model limited to predict the testing set only but not instances that are new to the model. In this study, finding the most important features has an economic importance because companies can evaluate their performance by focusing on those features. There are different methods to identify the key features. Each method has its pros and cons, but we observed that each method identifies different features to be the most important. In this study, we tested three techniques and compared them based on the results of the prediction models.

3.4 New Naïve bayes Algorithms: 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.

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

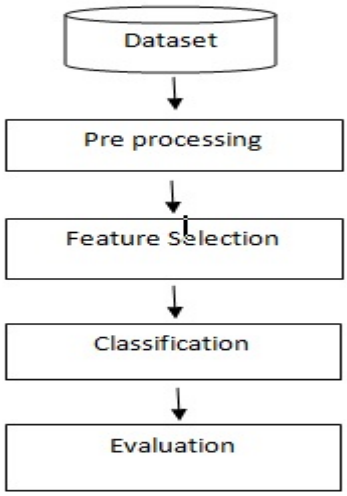


Figure 2: Proposed Model for Credit Card Fraud Detection

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.

4.2 Methodology

The experiment methodology involves following steps:

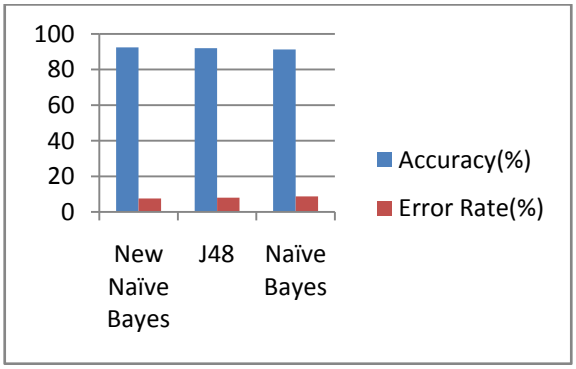
1. **Preprocessing of Dataset:** In preprocessing step remove redundancy, missing values, and inconsistency of used raw dataset. In this experiment using "all filters" from weka preprocess window after that select best feature.
2. **Applying Feature selection:** In feature selection steps applying the "CfsSubsetEval" evaluator and best first search from supervised attribute inweka preprocess window.
3. **Applying new naïve bayes classifier :** In this classification phase used new naïve bayes classifer for classification. Two different classifiers like naïve bayes and J48 are used for compare result with spilt 60% data.
4. **Evaluate result:** After that evaluate the result on the basis of accuracy of the proposed model and error rate.

4.3 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 92.4704% accuracy and 7.5296% of error rate. Two different classifier first naïve bayes gives 91.2437% accuracy and 8.75% of error rate. Another one J48 gives 92.0051% of accuracy and 7.99% of error rate.

Classifier	Accuracy(%)	Error Rate(%)
New Naïve Bayes	92.4704	7.52
J48	92.0051	7.99
Naïve Bayes	91.2437	8.75

Table 2: Comparison Result



Result Comparison Graph

5. CONCLUSION & FUTURE WORK

Prediction of bankruptcy is most important now days. This opens new confronts in the field of fraud detection and prevention, but prevention is of course better than detection. The simple techniques like database comparison and pattern matching are not enough for detecting such frauds because fraudulent transactions are rare within huge number of genuine transactions. So, Predictive models are of prime importance for banks to detect fraud. The proposed predictive model is compared with two other models. The proposed work is compared on basis of two functional metrics: 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. At the same time proposed predictive model would be integrated with live stream to find the online fraudulent transaction instantly. In future we intend to build up a cloud based ML application for detecting frauds in financial transactions done with cards.

6. REFERENCES

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

[2] 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.

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

[4] L. Breiman, “Bagging predictors,” Machine Learning, vol. 24, no. 2, pp. 123–140, 1996.

[5] 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).

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

[7] 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.

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

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

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

[11] 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.

[12] 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.

[13] European Central Bank, “Third report on card fraud,” European Central Bank, Tech. Rep., 2014.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

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

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

[21] 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.

[22] 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.

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

[24] 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

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

[26] 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

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

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

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