Data Mining Techniques for Credit Risk Assessment Task

ADNAN DZELIHODZIC, DZENANA DONKO
International Burch University
Francuske revolucije bb, Ilidza, Sarajevo
BOSNIA AND HERZEGOVINA
adnandz@gmail.com, ddonko@etf.unsa.ba

Abstract: - This paper is review of current usage of data mining, machine learning and other algorithms for credit risk assessment. We are witnessing importance of credit risk assessment, especially after the global economic crisis on 2008.S o, it is very important to have a proper way to deal with the credit risk and provide powerful and accurate model for credit risk assessment. Many credit scoring techniques such as statistical techniques (logistic regression, discriminant analysis) or advanced techniques such as neural networks, decision trees, genetic algorithm, or support vector machines are used for credit risk assessment. Some of them are described in this article with theirs advantages/disadvantages. Even with many models and methods, it is still hard to say which model is the best or which classifier or which data mining technique is the best. Each model depends on particular data set or attributes set, so it is very important to develop flexible model which is adaptable to every dataset or attribute set.

Key words: - Credit risk assessments, credit scoring techniques, single classifiers

1 Introduction

Banks and banking activities have evolved significantly through the time and they have important role in the economy. Banks act as brokers between supply and demand of securities, and they transform short-term deposits into medium and long-term credits [1]. History of credit stretches back 5000 years and it is still one of main focus of research in financial sector [2]. Banks as all other companies have risks in their business. According to GARP banks face several types of risk. All the following are examples of the various risks banks encounter:

- Borrowers may submit payments late or fail altogether to make payments.
- Depositors may demand the return of their money at a faster rate than the bank has reserved for.
- Market interest rates may change and hurt the value of a bank's loans.
- Investments made by the bank in securities or private companies may lose value.
- Human input errors or fraud in computer systems can lead to losses.

Credit scoring is one of the most important and the most interesting area for research all around the world. There are a lot of reasons for that and one of them is that credit scoring is important for global economy and as we know it is one of the reasons for the world financial crises from 2008.

2 Credit Risk

Generally banks should focus on three main types of risk: Credit, Market and Operational [10]. Credit risk is the potential loss a bank would suffer if a bank borrower, also known as the counterpart, fails to meet its obligations—pay interest on the loan and repay the amount borrowed—in accordance with agreed terms. Credit risk is the single largest risk most banks face and arises from the possibility that loans or bonds held by a bank will not be repaid either partially or fully [7].

Credit risk is typically represented by means of three factors: default risk, loss risk and exposure risk. Credit and default risk are often synonymous. Credit risk management is a process that involves the identification of potential risks, the measurement of these risks, the appropriate treatment, and the actual implementation of risk models [1]. Credit risk assessment was the first tool developed in financial services 60 years ago. Establishing a standardized and practical assessment system for commercial banks is of positive and practical significance to comprehensively improve the bank's management level and to effectively reduce and prevent credit risks [9].

ISBN: 978-960-474-317-9

3 Credit Risk Assessments

The overall idea of credit evaluation is to compare the features or the characteristics of a customer with other previous customers, whose loans they have already paid back. So, credit scoring is often used to analyze a sample of past customers to differentiate present and future credit customers. Credit scoring can be formally defined as a mathematical model for the quantitative measurement of credit [11]. Another definition of credit scoring is: Credit scoring is the set of decision models and their underlying techniques that aid lenders in the granting of consumer credit [2].

Generally,in order to mitigate credit risk and evaluate credit application, two techniques can be used: Loan officer's subjective assessment and credit scoring [20].

The success of a judgemental process depends on the experience and the common sense of the credit analyst and it is are associated with subjectivity, inconsistency and individual preferences motivating decisions. Judgemental methods have some strengths, such as taking account of qualitative characteristics and having a good track record in evaluating past credit by utilising the wealth of the credit analyst's past experience [21].

Credit scoring was introduced for the first time by Fisher in 1936 and the only methods used were statistical discrimination and classification methods [12]. The other credit analysis methods that are using decision tree analysis, k-nearest neighbor, neural networks, Support Vector Machine (SVM), rule-based have been developed lately [13][14][16][17][18]. Common for all this approach is that they are data-driven. In recent years, there is a lot of implementation of credit scoring using different machine learning and data mining techniques.

Traditional credit scoring has been less effective in credit assessment for new and innovative loan products. For example, „the once-vaunted Fair, Isaac and Company (F ICO) credit scoring system" is now being blamed for failing to signal risky borrowers in the mortgage market [3]. Also one of the reasons of financial crises from 2008 was bad credit risk assessment. That is why credit risk assessment task is still one of the main topics in yhe banking industry.

4 Credit scoring techniques

Generally, credit scoring techniques can be divided into two main categories: traditional methods and advanced methods. Goal of all these techniques is to provide powerful, accurate and meaningful models for classification. In advance techniques, typically there is two phases: learning or training and testing phase.

4.1 Traditional statistical methods

Traditional statistical methods are statistical-based learning tools and they are first developed tools for credit scoring [12].

Even if Fisher developed Credit scoring model using discriminant analysis developed by Fisher in , this is a still valid technique used in building credit scoring models [29].

Linear regression has been used in credit scoring applications, as the two class problem can be represented using a dummy variable. Factors, such as customers‟ historical payments, guarantees, default rates in a timely manner, can be analyzed by credit analysts, with linear regression to set up a score for each factor, and then to compare it with the bank‟s cut-off score. If a new customer’s score passes the banks score, the credit will be granted.

Orgler (1971) used a regression approach for evaluating outstanding consumer loans. According to this research, information that is not included on the application form had greater predictive ability than information on application form.

Logistic regression in difference to linear regression model has dichotomous outcome variable. On theoretical grounds it might be supposed that logistic regression is a more proper statistical instrument than linear regression, given that the two classes “good” credit and “bad” credit have been described (Hand & Henley, 1997).

Logistic regression is a statistical technique where the dependent variable is a Bernoulli variable. It models the logarithm of the odds as a linear function of the independent variables, Xi. The form of the model is:

$$logiit(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

Where p_i is the probability that Y=1 and X_1, $X_{2,\ldots,}$ X_k are the independent variables (predictors) and $\beta_0, \beta_1, \ldots, \beta_k$ are known as regression coefficients.

In order to build a classification model with logistic regression, the following rule can be used:

$$class = \begin{Bmatrix} 1, & p_i \leq 0.5 \\ -1 & p_i > 0.5 \end{Bmatrix}$$

There are many logistic regression credit scoring models and it is one of the most widely used statistical techniques [26][29].

ISBN: 978-960-474-317-9

4.2 Advanced techniques

Advanced techniques such as neural networks, decision trees or support vector machines have widely implemented models of credit scoring. Reason for that is their capability of modeling extremely complex functions and getting better results, primarily in accuracy. Advanced techniques can be categorized in a couple of ways such as: single classifiers, ensemble techniques and hybrid classification techniques.

4.2.1 Single classifiers

Credit scoring models are usually developed using single classifiers and they are widespread. These single classification techniques are divided into several groups, which are supervised learning, unsupervised learning, and other techniques. Most common used supervised learning single classifiers are decision trees, support vector machines and neural networks. Examples of unsupervised learning classifiers are self-organized maps and k-means but they are rarely used. Genetic algorithm as example of the other techniques, besides neural networks and support vector machines, is one of common used techniques for credit scoring.

a) Neural Networks

Neural networks that are used to solve the problem of credit evaluation can be regarded as a statistical method, which transform linear combination variables with a non-linear manner and then recycling the process.

A neural network is most accurate in bank failure prediction, followed by linear discriminant analysis, logistic regression, decision trees, and k-nearest neighbor [32]. In comparison with other techniques, they concluded that neural network models are more accurate, adaptive and robust. There are many examples of ANN classifiers for credit scoring [33][34][35][38].

b) Genetic Algorithm

Genetic Algorithms (GAs) is a stochastic global search method that mimics the metaphor of the natural biological evolution. Basis of these algorithms and their development was given by in 1975 and since then GAs can be viewed as general-purpose optimization method, but also as algorithms for difficult search problems and other machine learning problems. Recently GAs is more frequently used in business, scientific and engineering applications. Genetic programming is one of the most recent techniques that has been applied in the field of credit scoring.

In a genetic algorithm, a population of strings (called chromosomes), which encode candidate solutions (called individuals, members, or phenotypes) to an optimization problem, evolves toward better solutions. Traditionally, solutions are represented in the binary form as strings of 0s and 1s. The evolution usually starts from a population of randomly generated individuals and happens in generations. In each generation, the fitness of every individual in the population is evaluated, multiple individuals are stochastically selected from the current population (based on their fitness), and modified (recombined and possibly randomly mutated) to form a new population. The new population is then used in the next iteration of the algorithm. Commonly, the algorithm terminates when either a maximum number of generations has been produced, or a satisfactory fitness level has been reached for the population [32]. There are three essential operations in GAs: Selection, Crossover, and Mutation. Implementation of GAs for credit scoring we can find in many research papers and also there is a widespread number of implementations of credit scoring models where genetic algorithms are used for optimization or preprocessing for classifiers such as SVM [32] [36]

c) Support Vector Machines

The Support Vector Machine (SVM) was first developed by Vapnik (1995) for binary classification. To achieve this, the algorithm attempts to find the optimal separating hyperplane between classes by maximizing the class margin. If we have a training dataset $\{x_i, y_i\}$ (i=1...N) where x_i are input and y_i are corresponding observed binary variable (output or class). Decision boundary is given by $\boldsymbol{\omega}\, \boldsymbol{x} + b = 0$, and it is necessary to find maximum margin. Margin is maximum distance of decision boundary from the data of both classes. Support Vectors are those data points that the margin pushes up against.

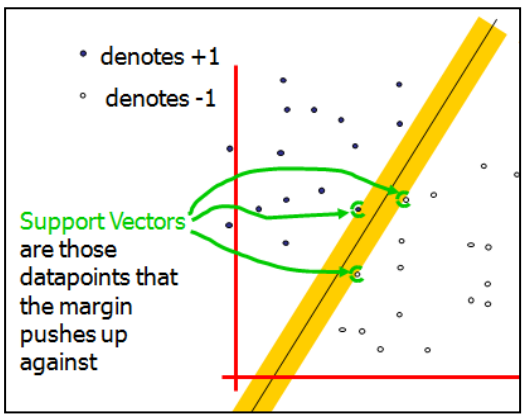

Fig.1 . SVM method

ISBN: 978-960-474-317-9

This is an example of linear SVM which is the simplest SVM. But, in real- world applications it is very hard to classify data with this approach. So we have to map data in high-dimensional feature space through a mapping function Φ(·) and find optimal separating hyper plane.

The separate hyper plane can be represented as follows:

$$z(x) = \omega^T \phi(x) + b = 0 \quad (1)$$

where $\boldsymbol{\omega}$ is the normal vector of the hyper plane and b is the bias that is a scalar.

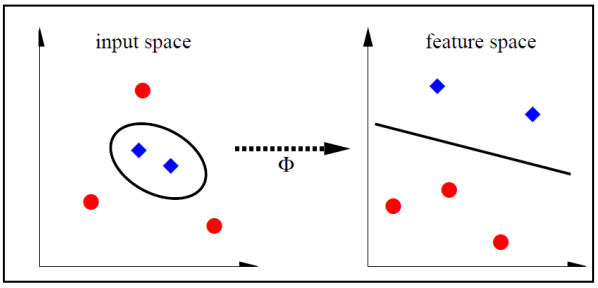

Fig. 2. Mapping function Φ(·)

Suppose that ϕ (·) is a nonlinear function that maps the input space into a higher dimensional feature space. If the set is linearly separable in this feature space, the classifier should be constructed as follows:

$$\omega^T \phi(x_i) + b \geq 1 \; if \; y_i = 1 \quad (2)$$
$$\omega^T \phi(x_i) + b \leq -1 \; if \; y_i = -1$$

Which is equivalent to

$$y_i(\omega^T \phi(x_i) + b) \geq -1 \; for \; i = 1, \ldots, N \quad (3)$$

In order to deal with data that are not linearly separable, the previous analysis can be generalized by introducing some nonnegative variables $\xi_i \geq 0$ such that (3) is modified to

$$y_i(\omega^T \phi(x_i) + b) \geq 1 - \xi_i \quad (4)$$

The nonzero ξ_i in (4) are those for which the data point x_i does not satisfy (3). Sum of ξ_i can be tretated as measurement of missclassification.

According to the structural risk minimization principle, the risk bound is minimized by formulating the following optimization problem:

$$Minimize \; \phi(\omega, b, \xi_i) = \frac{1}{2}\omega^T \omega + C \sum_{i=1}^{N} \xi_i \quad (5)$$

$$Subject \; to: y_i(\omega^T \phi(x_i) + b) \geq 1 - \xi_i \; for \; i = 1, .., N \; ; \; \xi \geq 0, for \; i = 1, \ldots, N$$

where C is a free regularization parameter controlling the trade-off between margin maximization and tolerable classification error. Searching the optimal hyper plane is a QP problem. By introducing a set of Lagrangian multipliers α_i and β_i for constraints, the primal problem becomes the task of finding the saddle point of the Lagrangian function,

$$L(\omega, b, \xi_i, \alpha_i, \beta_i) = \frac{1}{2}\omega^T \omega + C \sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} [\alpha_i \, y_i(\omega^T \phi(x_i) + b) - 1 - \xi_i] - \sum_{i=1}^{N} \beta_i \xi_i \quad (6)$$

Optimal solution for the weight vector is given by

$$\omega = \sum_{i=1}^{N_s} \alpha_i y_i \phi(x_i) \quad (7)$$

where *Ns* is the number of SVs.

Once the optimal pair (*w*, *b*) is determined, the decision function of SVM is obtained as

$$Z(x) = sign\left(\sum_{i=1}^{N_s} \alpha_i y_i K(x_i, x_j) + b\right) \quad (8)$$

where $K(x_i, x_j)$ is the kernel function in the input space that computes the inner product of two data points in the feature space. SVM is very popular classifier for credit scoring models and there is a lot of implementation using SVM or hybrid techniques with SVM included. Some examples are [38][39][40]. SVM is often used as classifier in hybrid models with PSO, GAs or Fuzzy.

5 Conclusion

Credit risk assessment is very important research field with wide application in the practice. Even if there is a hundreds of research, models and methods, it is still hard to say which model is the best or which classifier or which data mining technique is the best. Each model depends on particular data set or attributes set, so it is very important to develop flexible model which is adaptable to every dataset or attribute set. In order to have better accuracy of model every model should be tested by credit staff because their knowledge can help to improve our models.

ISBN: 978-960-474-317-9

References:

[1] Van Gestel, T. & Baesens,B. *Credit Risk Management:Basic Concepts: Financial Risk Components, Rating Analysis, Models, Economic and Regulatory Capital: Basic Concepts: Financial Risk Components, Rating Analysis, Models, Economic and Regulatory Capital*. Oxford University Press, 2008

[2] Thomas, L.C. & Edelman D.& Crook J. *Credit Scoring and its Applications* Society for industrial and Applied Mathematics, 2002

[3] Abrahams, C. & Zhang,M. *Credit risk assessment: the new lending system for borrowers, lenders, and investors.* New Jersey: John Wiley & Sons, Inc., 2009

[4] Mays, E. *Credit Risk Modeling: Design and Application.* Fitzroy Dearborn Publishers,. Chicago and London, 2002

[5] OeNB *Credit Approval Process and Credit Risk Management*, Oesterreichische Nationalbank (OeNB), 2004

[6] Fiedler E.(1971), *Measures of Credit Risk and Experience*,UMI, 1971

[7] Apostolik,R. & Donohue,C. & Went,P. (*Foundations of Banking Risk: An Overview of Banking, Banking Risks, and Risk-Based Banking Regulation.* New Jersey:John Wiley& Sons,Inc., 2009

[8] Yves Kodratoff, Ryszard S. Michalski, Machine Learning: An Artificial Intelligence Approach, Volume 3, 1990

[9] GUO Yingjian ,WU Chong, *Research on Credit Risk Assessment in Commercial Bank Based on Information Integration* Proceedings of 2009 International Conference on Management Science and Engineering, 2009

[10] D. Foust and A. Pressman (2008.), *"Credit Scores: Not-So-Magic Numbers,"* Business Week, February 7, 2008.

[11] Jing He & Yanchun Zhang & Yong Shi & Guangyan Huang, *Domain-Driven Classification Based on Multiple Criteria and Multiple Constraint-Level Programming for Intelligent Credit Scoring*. IEEE Transactions on Knowledge and Data Engineering, Vol. 22, No 6, 2010

[12] Fisher, R. A., The Use of Multiple Measurements in Taxonomic Problems, Annals of Eugenics 7 (2): 179–188., 1936

[13] J.R. Quilan, Induction of Decision Trees, Machine Learning, vol. 1, pp. 81-106, 1986.

[14] B.V. Dasarathy, *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*, ed. IEEE Computer Society Press, 1991.

[15] Marinakis,Y & Marinaki, M. &Doumpos,M. & Matsatsinis, N.& Zopounidis,C., *Optimization of nearest neighbor classifiers via metaheuristic algorithms for credit risk assessment*. Springer Science+Business Media, LLC, 2007

[16] Guo,H. & Gelfand,S.B. *"Classification Trees with Neural Network Feature Extraction,"* IEEE Trans. Neural Networks, vol. 3,pp. 923-933, 1992.

[17] Wang,Y. & Wang,S. & Lai,S.S., *A New Fuzzy* Support Vector Machine to Evaluate Credit Risk, IEEE Trans. Fuzzy Systems, vol. 13, no. 6, pp. 820-831, Dec. 2005.

[18] Laha,A., Building contextual classifiers by integrating fuzzy rule based classification technique and k-nn method for credit scoring" Advanced Engineering Informatics archive Volume 21 Issue 3, Pages 281-291, July, 2007

[19] Sun, J. and Li,H., Listed companies' financial distress prediction based on weighted majority voting combination of multiple classifiers, Expert Syst. Appl., vol. 35, pp. 818–827, 2008.

[20] Crook, J. N. (1996.) *Credit scoring: An* overview. Working paper series No. 96/13, British Association, Festival of Science. University of Birmingham, The University of Edinburgh., 1996

[21] Al Amari, A. The credit evaluation process and the role of credit scoring: A case study of Qatar. Ph.D. Thesis, University College Dublin., 2002

[22] I.-C. Yeh and C.-H. Lien, The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients, Expert Syst. Appl., vol. 36, pp. 2473–2480, 2009.

[23] Orgler, Y. E. Evaluation of Bank Consumer Loans with Credit Scoring Models. Journal of Bank Research 2 (1): 31-37., 1971

[24] T.-H. Lin, A cross model study of corporate financial distress prediction in Taiwan: Multiple discriminant analysis, logit, probit and neural networks models, Neurocomputing, vol. 72, pp. 3507–3516, 2009.

[25] M. Bensic, N. Sarlija, and M. Zekic-Susac, Modelling small-business credit scoring by using logistic regression, neural networks and decision trees, Intell. Syst. Account., Finance Manag., vol. 13, pp. 133–150, 2005.

[26] Hand, D. J., Henley, W. E., Statistical Classification Methods in Consumer Credit Scoring: A Review. Journal of the Royal Statistical Society: Series A (Statistics in Society) 160 (3): 523-541., 1997

[27] Abdou, H., Pointon, J., El Masry, A., Neural nets versus conventional techniques in credit scoring in Egyptian banking. Expert Systems with Applications 35 (3): 1275-1292., 2008

[28] Wei-Yang Lin, Ya-Han Hu, and Chih-Fong Tsai 2 *Machine Learning in Financial Crisis Prediction: A Survey* IEEE Transactions on systems, man, and Cybernetics VOL. 42, NO. 4, 421-436., JULY 2012

[29] S.S. Satchidananda and B.S. Jay. "Comparing decision trees with logistic regression for credit risk analysis," SUGI Asia conference, Mumbai, India, 2006

[30] Gately, E., Neural Networks for Financial Forecasting: Top Techniques for Designing and Applying the Latest Trading Systems. New York: John Wiley & Sons, Inc., 1996

[31] Cortes, C., & Vapnik, V., Support vector networks. Machine Learning, 20(3),273–297., 1995

[32] S. Oreski, D Oreski , G. Oreski. Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment, Expert Systems with Applications 39 12605–12617, 2012

[33] Khashman, A., Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes. Expert Systems with Applications, 37, 6233–6239., 2010

[34] Khashman, A. , Credit risk evaluation using neural networks: Emotional versus conventional models, Applied Soft Computing 11 (2011) 5477–5484, 2011

[35] L.Yu , S.Wang , K.K.Lai,, Credit risk assessment with a multistage neural network ensemble learning approach, Expert Systems with Applications 34 1434–1444, 2008

[36] C.-H. Wu, G.-H. Tzeng, Y.-J. Goo, and W.-C. Fang, A real-valued genetic algorithm to optimize the parameters of support vector machine for predicting bankruptcy,Expert Syst. Appl., vol. 32, no. 2, pp. 397– 408, 2007.

[37] E. Angelini, G. di Tollo, and A. Roli, A neural network approach for credit risk evaluation, Quart. Rev. Econ. Finance, vol. 48, pp. 733–755, 2008.

[38] T. Harris , Quantitative credit risk assessment using support vector machines: Broad versus Narrow default definitions, Expert Systems with Applications 40 4404–4413, 2013

[39] G. Wang, J.Mac, A hybrid ensemble approach for enterprise credit risk assessment based on Support Vector Machine, Expert Systems with Applications 39 (2012) 5325–5331, 2012

[40] C.-F. Tsai, Financial decision support using Neural Networks and support vector machines,"Expert Syst., vol. 25, no. 4, pp. 380–393, 2008.