Machine Learning Algorithms Based Credit Risk Assessment Models

Dr. Olcay Erdoğan
Independent Researcher
Email: olcayerdogan960@gmail.com
DOI: 10.23918/ICABEP2019p40

Abstract

This paper describes an approach for a credit risk evaluation based on machine learning classifiers. The constructed credit rating models were on sample data that consists of financial ratios from 356 enterprises that are listed on the Istanbul Stock Exchange. The applied methods are k- nearest neighbor, support vector machines and decision trees. This research develops models to evaluate the credit risk of the companies obtained from the financial statement of enterprises. The study supports building a balanced financial environment by reducing the cost of bankruptcy and help to determine the firms which are appropriate for the credit loan.

Keywords: CRA, credit risk, machine learning.

1 INTRODUCTION

One of the important research topics in finance is credit risk evaluation and bankruptcy prediction, models. Minimization of credit risk by effective credit risk evaluation tools such as credit rısk assessment models are necessary for every financial institution.

CRA play an important role in the analysis of the financial situation of companies. It is also helpful in providing useful input for hedging practices (Gestel and Baesens, 2009). Credit rating activities form reliable and stable financial markets within the economy. They also facilitate the outsourcing of the economy and the merge of domestic markets with international markets. Rating activities restrict the general risk level of the economy, increase the efficiency of financial transactions, and provide more efficient finance for growth. Through credit rating, it is possible for domestic enterprises to protect their credibility. Primarily, rating activities promote the strength of financial structures and the restriction of risks. They also provide savings in the cost of deposits, diversification in credit interest rates according to risks, and an increase in reliance on the financial system without the government guarantee. Besides these, rating activities improve relations with international finance environments and reduce the cost of outsourcing (Babuşçu and Hazar, 2007).

The paper is organized as follows. In Section 2 the credit risk assessment process is presented. Section 3 briefly describes the machine learning techniques. Section 4 gives the analysis results Section 5 concludes with the contributions of the study to CRA and directions for future research.

2 CREDIT RISK ASSESSMENT PROCESS

The process involves a decision either to extend credit risk or to refuse credit. The situation is shown as a decision problem in Figure 2.1. The requirement is to consider the benefit of taking the credit risk by extending credit against the potential loss (Brown & Moles, 2014).

FIGURE 2. 1 Decision-making Process. Reprinted from "*Credit Risk Management*" by Brown, K., & Moles, P., 2014.

The stages of CRA start with the financial assessment. The first step presents the analysis associated with a credit analyst who identifies the financial health of an institution. The analysis of the competitive position and operating environment of a firm helps to determine the risk level. Also, management and other qualitative factors are taken into consideration to determine the risk level of the company (Cabbar, 2006). The credit analyst would work on the financial reports to determine if the earnings and cash flows are sufficient to cover the debt. The analyst also

examines the firm's leverage, access to the capital markets and whether it has the flexibility to borrow money (Crouhy et al., 2001). The financial ratios give information about the profitability and interest coverage of the issuer, asset protection and cash flow adequacy

3 METHODOLOGY

The models were implemented employing WEKA machine learning framework to obtain the required algorithms. The dataset used in the models are involving 9 financial ratios used is derived from balance and income statement of enterprises functioning in Istanbul Stock Exchange.

Metrics, used as an accuracy identifier in machine learning algorithms such as TP (True Positive) and MCC rates, were selected to evaluate classification performance.

i i i i TruePositives TPR TruePositives FalseNegatives □ □

Here FN is the number of ''positive'' cases that are incorrectly classified as ''negative''. TP and TN, naturally, represent the numbers of correctly identified ''positive'' and ''negative'' cases.

The MCC is a correlation coefficient between the observed and predicted binary classifications; it returns a value between −1 and +1. A coefficient of +1 represents a perfect prediction and −1 indicates total disagreement between prediction and observation (Powers, 2003). The Matthews correlation coefficient (MCC) is used in machine learning as a measure of classifications. For investment, speculative and below investment groups the value is close to +1 which indicates the perfect prediction.

3.1 Machine Learning Techniques

The machine learning studies on constructing computer programs that can be improved with experience. In recent years, many successful machine learning applications have been developed and employed for risk assessment models. We've performed the analysis by SVM, k-NN and DT methods which are mostly employed methods for CRA.

3.1.1 Support Vector Machine (SVM)

Support Vector Machine is one of the machine learning techniques that has the capability of classifying big data. An optimal separating hyperplane that separates the data with the maximal margin is constructed by solving an optimization problem. The solution to the problem gives out a subset of training patterns that lie closest to the boundary. Classification of the decision surface of a SVM is given in general by

SVMxsignwxb

where is a mapping in some feature space *F*. The parameters are such that they minimize an upper bound on the expected risk (Chow and Cho, 2007). Figure 3.1 shows a hyper plane that separates two classes with the maximal margin.: n F □ □ □wF and bR □□

The support vector machine tries to find the optimal separating hyperplane between the groups by maximizing the margin between them. Points lying on the boundaries are referred to as support vectors, while the middle of the margin is called the optimal separating hyperplane (Yu, 2008). SVMs classify by forming an *N*-dimensional hyper plane that can separate the data into two categories.

Studies point out that SVM is an efficient classification method and it is superior to most of the other methods in many experiments such as text categorization and face or fingerprint identification (Yu, 2008). Wang and Lai (2005) propose a fuzzy support vector machine to

discriminate good and bad customers and found out that new fuzzy support vector machine has more classification ability

FIGURE 3.1 Linear Separable Support Vector Machine. Reprinted from"A hybrid ensemble approach for enterprise credit risk assessment based on Support Vector Machine" by Wang, G. & Ma, J., 2012. *Expert Systems with Applications*, *39*(5), 5325-5331

3.1.2 K-Nearest Neighbor

The *k*-nearest-neighbor (KNN) method is a nonparametric approach, which classifies a data instance by considering only the *k*-most similar data instances in the training set. The algorithm accepts that the n-dimensional space involves all instances. Figure 3.2 shows the shape of this decision surface induced by 1-nearest neighbor the entire instance space.

FIGURE 3.2 Decision Surface Induced by 1-Nearest Neighbor. Reprinted from "*Machine Learning"* by Mitchell, T. M., 1997. Copyright by McGraw-Hill

The k-nearest neighbor algorithm used to calculate the mean value of the k nearest training examples rather than calculate their most common value (Mitchell, 1997).

3.1.3 Decision Trees (DT)

Decision trees are a non-parametric data mining technique where the root node contains all training observations and the training data are recursively partitioned by values of the input variables until reaching the leaf nodes where the classification decision is made for all

observations contained (Zhang and Hardle, 2010). An instance is classified by starting at the root node of the tree, testing the attribute specified by this node, then moving down the tree branch corresponding to the attribute in the given example. This process is repeated for the subtree rooted new node (Mitchell, 1997). Figure 3.3 illustrates the decision tree structure.

FIGURE 3.3 A Decision Tree Structure

4 EXPERIMENTAL RESULTS

We implemented Decision Tree algorithms (REP Tree, Random Tree and Random Forest), Support Vector Machines and k- Nearest Neighbor Algorithms on the obtained dataset of enterprises' financial information.

The results represent the ability of REP Tree, Random Tree and Random Forest models to classify the data. The Random Forest model had the highest accuracy of classification with 99.44% of the instances formed of the companies of ISE (Istanbul Stock Exchange). In addition, REP tree model was comparable to Random Forest with misclassification of 1.40% is shown by table 4.1.

TABLE 4.1 Comparisons of the Employed Models. REP Tree	Random Tree		Random Forest	SVM	k-NN
Correctly Classified Instances	353	345	354	300	273

Incorrectly Classified Instances	3	11	2	56	83
Kappa statistic	0.9859	0.9486	0.9906	0.717	0.6177
Mean absolute error	0.0111	0.0206	0.0324	0.2865	0.1581
Root mean squared error	0.0747	0.1435	0.0778	0.372	0.3924
Relative absolute error	0.0277	0.0515	0.0809	0.7156	0.3949
Root relative squared error	0.1669	0.3209	0.1739	0.8318	0.8775
Total Number of Instances	356	356	356	356	356

Model 1: The MCC is a measure between the observed and predicted binary classifications; it returns a value between −1 and +1. A coefficient of +1 represents a perfect prediction and −1 indicates total disagreement between prediction and observation (Powers, 2007). The Matthews correlation coefficient (MCC) is used in machine learning as a measure of the quality of classifications. For investment, speculative and below investment groups the value is close to +1 which indicates the perfect prediction.

Random forests system can be explained in two steps: First, T subsets are generated by the selection of data from the original sample. Secondly, a tree is built from each subset using random subspace feature selection to generate partitions which reduce correlation between trees in the forest (Breiman, 2001). In this study, the Random Forest model had the highest accuracy of classification with 99.44%. The following tables Table 4.2, Table 4.3 and Table 4.4 represent the DT results.

TABLE 4.2 REP Tree Classification Results

TP Rate	FP Rate	Precision	Recall	MCC	Class
1	0.016	0.983	1	0.969	INVESTMENT
1	0	1	1	1	SPECULATIVE
0.978	0	1	0.978	0.988	BELOW

TABLE 4.3 Random Tree

TP Rate	FP Rate	Precision	Recall	MCC	Class
0.971	0.022	0.977	0.971	0.949	INVESTMENT
0.936	0.013	0.917	0.936	0.915	SPECULATIVE
0.978	0.014	0.978	0.978	0.964	BELOW

TABLE 4.4 Random Forest

TP Rate	FP Rate	Precision	Recall	MCC	Class
1	0.011	0.989	1	0.989	INVESTMENT
1	0	1	1	1	SPECULATIVE
0.985	0	1	0.985	0.988	BELOW

The obtained TP Rate values can be considered as a good result (all were over 0.9). This indicates that instances were classified successfully.

Model 2: In the SVM based model, for investment and below investment groups the value is close to +1 which indicates a good prediction. But for speculative level companies, MCC gives no better than random prediction. Table 4.5 shows the MCC values which are close to +1 which means an accurate prediction investment and below investment levels.

TABLE 4.5 SVM Results

TP Rate	FP Rate	Precision	Recall	MCC	Class
0.994	0.209	0.82	0.994	0.799	INVESTMENT
0	0	0	0	0	SPECULATIVE
0.941	0.081	0.876	0.941	0.849	BELOW

Model 3: For the results of the k-NN algorithm, MCC value shows that for investment and below investment groups the value is close to +1 which indicates a good prediction but less accurate than other models. But for speculative level companies, the prediction is not accurate. Table 4.6 gives an accurate result in below investment level.

TABLE 4.6 kNN Results

TP Rate	FP Rate	Precision	Recall	MCC	Class
0.822	0.154	0.836	0.822	0.668	INVESTMENT
0.426	0.129	0.333	0.426	0.268	SPECULATIVE
0.815	0.068	0.88	0.815	0.759	BELOW

We compare the results by error rates and check the accuracy of each model. Dataset has 356 samples with 10 features classified in investment, speculative and non-investment risk levels. Table 4.7 shows the distribution of levels as 174 companies in investment, 47 are speculative and 135 of them are in the non-investment level.

TABLE 4.7 Comparison of the Employed Methods

MODEL	SVM	kNN	Random Forest	Random tree	REP Tree
Accuracy	84.27	76.69	99.44	96.91	99.16

Finally, we monitored the credit risk level of healthy firms and also defaulted companies. Based on the results discussed it is possible to conclude that the decision trees are suitable tools for the prediction of future company credit rates.

Empirical results revealed that the Random Forest model had the highest accuracy of classification with 99.44% of the instances. The model we proposed has a satisfying performance with 90.46 %.

5 CONCLUSIONS

Through the accurate assessment of credit risk, it is possible for domestic enterprises to protect their credibility. CRA is providing a framework for the risk management strategy to future investors. In this study, we explore an approach for machine learning driven credit risk evaluation using five distinct methods and found out that the Random Forest model had the highest accuracy of classification. Our study has to be supported by the larger datasets including more enterprises in order to have a higher efficiency of the prediction systems.

REFERENCES

Brown, K., & Moles, P. (2014). Credit Risk Management. Edinburg: Edinburgh Business School.

Danenas, P., & Garsva, G. (2015). Selection of Support Vector Machines based classifiers for credit risk domain. Expert Systems With Applications, 42(6), 3194-3204. doi: 10.1016/j.eswa.2014.12.001

Babuşçu, Ş & Hazar, A., 2007, SPK: Kredi Derecelendirme Uzmanlığı Sınavlarına Hazırlık, Akademi Consulting and Training, 41, 301.

Breiman, L. (2001). Random forests. Machine Learning, 45, 5–32.

Brown, K., & Moles, P. (2014). Credit Risk Management. Edinburg: Edinburgh Business School.

Cabbar, H. (2006). Management Of Credit Risk In The Firms (M.Sc.). Marmara University.

Chow, T. W. S., & Cho, S.-Y. (2007). Neural networks and computing: learning algorithms and applications. London: Hackensack, NJ: Imperial College Press; Distributed by World Scientific.

Crouhy, M., Galai, D., & Mark, R. (2001). Prototype risk rating system. Journal of banking & finance, 25(1), 47-95..

Gestel, T. van, & Baesens, B. (2009). Credit risk management: basic concepts: financial risk components, rating analysis, models, economic and regulatory capital. Oxford ; New York: Oxford University Press.