Contents lists available at GrowingScience

International Journal of Data and Network Science

homepage: www.GrowingScience.com/ijds

Valuation and assessment of customers in banking industry using data mining techniques

Behrooz Asareh[a] and M.R. Ghaeli[a*]

[a]*Department of Commerce and Business Administration, New Westminster, BC, Canada*

C H R O N I C L E

Article history:
Received: September 17, 2018
Received in revised format: October 29, 2018
Accepted:December31, 2018
Available online:
December31, 2018

Keywords:
Data mining
Logistic regression
Bank customer

A B S T R A C T

One of the primary concerns in most financial institutions to have an appropriate method for ranking customers. Bank customers are the primary sources of creating income and the success of banking industry depends on how to select good customers for allocation of loans. This paper uses Decision Tree, K-nearest neighbor (KNN), Support Vector Machine (SVM), Naive Bayes, and Logistic Regression for data categorization to estimate credit ranking of bank customers in one of major banks in Middle East. The results indicate that Logistic Regression was considered as the best method for ranking customers with the precision of 76.17% while Decision Tree was considered as the weakest technique with the precision of 73.30%.

1. Introduction

Customers' assessment is essential in providing a clear picture of the status and the ability of the client to timely fulfil their obligations, preventing them from overuse of resources and falling into the financial crisis (Bosch & Steffen, 2011). Credit valuation is a process in which each borrower is assigned a quantity that represents an estimate of his/her future performance in repaying his/her loan (Durand, 1941). Validation models are mathematical models that, according to the characteristics of the borrower, compute a score that reflects the probability of defaulting the borrower and classify borrowers in certain categories of risk (Altman, 2000; Durand, 1941). What matters to banks is to assess the likelihood of non-repayment before the facility is granted and to choose a group to ensure that their commitment is guaranteed at due time (Fensterstock, 2003; Elmer & Borowski, 1988). The existence of high numbers of outstanding banking system demands, which today has become a concern for government officials in addition to the concern of bank managers, is largely influenced by the lack of credibility of the customers (Fensterstock, 2003). Another benefit of validation includes: increasing efficiency and speed, improving bank profitability and collecting statistical information, linking potential to mechanized scoring systems and providing similar information, eliminating most of the fraud and minimizing credit risk, allowing access to

* Corresponding author.
E-mail address: rghaeli@nyit.edu (M. R. Ghaeli)

doi: 10.5267/j.ijdns.2018.12.006

information on daily basis and access to information infrastructure and easy data management for industries related to collecting receivables and improving the ratio of non-settled debts (Thomas, 2000). Duff and Einig (2009) performed a survey on demand for corporate bond ratings provided by credit ratings agencies (CRAs) and the method on how issuers select CRAs using some interview. The study identified the principal source of demand for rating information could be considered to reduce agency conflicts between issuers and investors.

Today, in order to validate customers, systems such as credit scoring and credit rating have been developed. Credit scoring is a system by which banks and credit institutions, using the information of the present and past applicants, evaluate and reward customers with the possibility of not repaying the loan (Lee & Chen, 2005). In other words, credit scoring means the probability of default in the future. This method of credit rating is unbiased and based on quantitative statistics and information, while the old methods for assessing customers were mainly mentally based on the views of the authorities. Credit rating is, in fact, a way of identifying and agreeing to lend to low risk applicants and avoiding lending to high risk applicants through their rankings (Limsombunchai et al., 2005). Today, in the credit industry, neural networks have become one of the most accurate tools for credit analysis among other tools (Malhotra & Malhotra, 2002). Desai et al. (1996) examined the capabilities of neural networks and common statistical techniques such as linear regression analysis in constructing credit scoring models. The survey accomplished by West (2000) indicates that neural networks could improve the accuracy of scoring. They also maintained that linear regression analysis was an excellent alternative to neural networks while the decision tree and the nearest neighbor model and linear audit analysis did not produce promising and encouraging results (Louzada et al., 2012).

Credit scoring methods are performed quantitatively and qualitatively. In qualitative analysis, credit scoring is closely related to the ability of credit department officials. However, according to quantitative analysis, it is possible to determine the probability of non-return of the principal and the benefit of the facility through its distribution function (Ong et al., 2005). Most quantitative risk-taking patterns have similar semantic frameworks, but the differences that arise in implementing these models are due to the method of estimating the main parameters of the available information (Thomas, 2000).

Credit risk is considered as the main cause of bank failures. The proper functioning of credit risk management of banks and credit institutions will depend on identifying the inherent factors of risk in lending operations (Min & Lee, 2008). Banks, with the establishment of a proper credit risk management system, can take the necessary measures to eliminate or reduce credit risk. In this regard, banks, by classifying credit and not accepting loans and inappropriate credits, protect themselves from accepting additional risks. Without the proper credit risk management system, the effect of the loss of banking operations will be unpredictable. Therefore, a credible customer credit rating system can be used to identify, measure and manage credit risk in a convenient and efficient manner (Papaikonomou, 2010).

Today, experts acknowledge the development of global trade through the development of e-commerce, as well as the use of e-banking and the use of appropriate techniques, models and tools for the successful presence of financial institutions and banks in the field of competition and global trade. Banks in today's competitive business environment face a lot of problems with environmental and economic changes that have become more tangible with the development of technology and the bulking and complexity of activities, including the phenomenon risk. In fact, risks are inherent in banking and financial activities. Since it is theoretically possible to eliminate risk, it should be managed as the only possible solution. By studying in the field of risk management, we will encounter various techniques, methods and tools that will require them to identify the type of risk and appropriate tools for measuring and reducing risk.

Data mining has been widely used for risk management in the banking industry. Bank managers need to be aware that customers who are buying and selling are trustworthy and reliable, providing new customers with credit cards, extending existing customers' credit, and agreeing to give loans. If

information is not available about customers, they can make decisions by accepting some risks. Data mining can help reduce the risk of banks by determining which customers are willing to pay their debts. Validation is one of the tools for financial risk management. Validation can be very important for a lender when it comes to borrowing. The history of good and bad lenders can be used to provide good and bad loan applicants (Fisher, 1936; Frawley et al., 1992).

Data mining can also deduce the validity of borrowers' behavior on installment payments, mortgages, and credit card loans using properties such as credit history, usage period, and residence time. The rating helps the lender evaluate the customer and decides whether a customer is a good candidate for a loan and is risk averse. Customers who collaborate with the bank for some periods of time are in good shape, and those with high incomes are more likely to get a loan than those who are new customers and do not have a record in the bank or those who receive low income. Examining and measuring customer's credit in credit institutions today is one of the most important financial decisions. In the past, the decision to grant credit to natural or legal inistitutions seeking credit was often the responsibility of a qualified individual or a group of experts, and was carried out by the departments concerned with monetary and credit affairs. Since the judging methods are time consuming, costly and subjective, they do not have the scientific credibility and reliability. To this end, financial institutions should design credit-worthy credit rating systems based on models and models. Modern systems of measurement of customer credit are based on mechanized processes in which certain privileges are given to some of the important credit characteristics of customers (Horrigan, 1968; Koh et al., 2004).

2. The proposed method

The data are collected from 768 customers of one of banks located in province of Khuzestan, Iran during the year of 2017. This paper examines the results of data analysis and the data mining process is examined and finally the results of the data mining techniques are used to analyze the data. This paper uses Decision Tree, K-nearest neighbor (KNN), Support Vector Machine (SVM), Naive Bayes, and Logistic Regression for data categorization (Goukasian & Seaman, 2009). To analyze the characteristics of bank's customers, which is an influential factor in the validation of bank customers, the proposed study uses classification algorithms which describes knowledge and extracts rules in relation to a set of information. Since the algorithms presented in this study have been selected for classification, they should be evaluated by classification criteria. In order to evaluate the performance of the algorithms, the actual classification of customers is compared with the classification performed by the software and the ability of the algorithms in customer validation is examined. First, let's define some of the necessary factors as follows,

TN: represents the number of records whose real category is negative and the categorization algorithm also correctly categorizes them as negative.
TP: represents the number of records whose real category is positive and the categorization algorithm also recognizes them positively.
FP: represents the number of records whose real category is negative and the category classification algorithm recognizes positively.
FN: represents the number of records whose real category is positive and the algorithm has categorized their category negatively.

In the proposed solution, for the purpose of checking the accuracy and the most important criterion for determining the efficiency of an algorithm, the precision (accuracy) or categorization rate is used, which measures the accuracy of the entire category. In fact, this criterion is the most popular and general criterion for calculating the efficiency of categorization algorithms, which indicates that the classed category correctly categorizes several percent of the entire set of experimental records as follows,

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}}. \tag{1}$$

The Recall (x) criteria measures the accuracy of the category x with respect to all records. The Precision (x) criterion shows the accuracy of the x-category classification in terms of the case that the x tag is suggested for the record by the classifier.

Refresh: The number of true positive samples of the system on total number of positive samples.
Accuracy: The number of true positive samples of the system on total number of positive predicted by the system.

$$precision\ (Class = Yes) = \frac{TP}{TP + FP'}, \tag{2}$$

$$recall\ (Class = Yes) = \frac{TP}{TP + FN}. \tag{3}$$

In addition, we may use a hybrid of Eq. (2) and Eq. (3) as follows,

$$F_{measure} = \frac{2 \times precision \times recall}{precision + recall}. \tag{3}$$

To increase the validity of the model in this research, a validation model is used. In this way, the whole set of data is considered both for training and for testing. In this method, the initial data are randomly divided into k subsets of approximately equal size. In repeating i, the D_i part acts as a test set and the remaining parts are used for prediction. The proposed method of this paper uses some personal characteristics as well as their accounts' specifications of the participants such as Account Balance, No of Credits at this Bank, Credit Amount, Duration in month, Value Savings/Stocks, Length of current employment, Duration in Current address, Age and Creditability. Table 1 demonstrates the summary of the statistical data. The criteria presented in Table 2 are used to evaluate the performance of the classification system.

Table 1
The summary of the statistical observations

Row	Title	Description	Type
1	Account Balance	Status of existing checking account	Numeric
2	No of Credits at this Bank	Number of existing loans at the bank	Numeric
3	Credit Amount	The amount of loan given to customer	Numeric
4	Duration in month	The number of month for the payment of loan	Numeric
5	Value Savings/Stocks	Balance of customer's bank account	Numeric
6	Length of current employment	The length of customer's employment	Numeric
7	Duration in Current address	The number of months customer presently resides in the present address	Numeric
8	Age	Age in years	Numeric
9	Creditability	Classification variable	

Table 2
The criteria used for classification

Formula	Description
$TPR = \frac{TP}{TP + FN}$	The proportion of true positive (TP) cases that are properly categorized.
$TNR = \frac{TN}{TN + FP}$	The proportion of negative cases that are properly categorized.
$FPR = \frac{FP}{FP + TN}$	The proportion of negative cases classified as positive.
$FNR = \frac{FN}{FN + TP}$	The proportion of positive cases classified as negative.
$P = \frac{FP}{FP + TP}$	The ratio of falsified positive numbers to total positive results (both true positive and false positive)
$P = \frac{TN}{TN + FN}$	The ratio of true negative to total number of total truly negative and false negative
$AC = \frac{TP + TN}{TP + TN + FN + FP}$	The ratio of the correct results (both positive and negative) to the whole society
$AC = 2 \times \frac{P \times TPR}{P + TPR}$	$F\_measure$

3. The results

In this section, we present the results of the implementation of different methods for evaluating customer credit ranking.

3.1. Decision Tree

Decision tree use different techniques for customer ranking and Table 3 demonstrates them in details.

Table 3
The summary of the algorithm used for the proposed method

Classification algorithm	Accuracy	Test option
W-Random Forest	74.87%	10-Fold Cross Validation
Decision Tree	73.3%	10-Fold Cross Validation
J48	74.74%	10-Fold Cross Validation
W-J48 graft	74.73%	10-Fold Cross Validation
Random Forest	73.82%	10-Fold Cross Validation
W-REPTree	73.97%	10-Fold Cross Validation
Decision Stump	71.87%	10-Fold Cross Validation
W-LADTree	66.3%	10-Fold Cross Validation
2-BFTree	74.35%	10-Fold Cross Validation

Evaluation of models obtained from the modeling stage to validate bankers and the effectiveness of the results of the accuracy of the prediction of each group record from the dataset using two algorithms (J48-graft) and (decision tree) are shown next. Table 4 presents the results of the decision tree.

Table 4
The results of decision tree

	Truly bad	Truly good	Class precision
Predicted bad	170	107	61.37%
Predicted good	98	393	80.04%
Class recall	63.43%	78.60%	

Based on the information given in Table 4 we have

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \times 100 = 73.3\%$$

$$Precision\ (true\ Bad) = \frac{TP}{TP + FP} \times 100 = 61.37\%$$

$$Precision\ (true\ Good) = \frac{FP}{TP + FP} \times 100 = 80.04\%$$

$$Recall\ (true\ Bad) = \frac{TP}{TP + FN} \times 100 = 63.43\%$$

$$Recall\ (true\ Good) = \frac{FN}{TP + FN} \times 100 = 78.6\%$$

$$F\_measure = (2 \times Precision\ \times Recall\ )/(Precision\ + Recall\ )$$

$$F\_measure\ (true\ Bad\ ) = 40.97$$

$$F\_measure\ (true\ Good) = 85.33$$

As we can observe, in 393 observation the model has predicted good customers properly and in 98 cases it mistakenly anticipated, so the predicted success rate is 61.37%. It also accurately detects 170 cases of the bad customers, and in 107 cases it mistakenly predicted the bad customers, and the success rate in the prediction is 80.04%. The accuracy of the model is 73.3%, which is an acceptable value.

Moreover, the results of the implementation of the W-J48 graft is given in Table 5 as follows,

Table 5
The results of W-J48 graft

	Truly bad	Truly good	Class precision
Predicted bad	155	81	65.68%
Predicted good	113	419	78.76%
Class recall	57.84%	83.80%	

Based on the information of Table 5 we now have

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \times 100 = 74.74\%$$
$$Precision\ (true\ Bad) = \frac{TP}{TP + FP} \times 100 = 65.68\%$$
$$Precision\ (true\ Good) = \frac{FP}{TP + FP} \times 100 = 78.76\%$$
$$Recall\ (true\ Bad) = \frac{TP}{TP + FN} \times 100 = 57.84\%$$
$$Recall\ (true\ Good) = \frac{FN}{TP + FN} \times 100 = 83.8\%$$

As we can observe, the accuracy of the model is 74.74%.

3.2. KNN method

The k nearest neighboring (KNN) algorithm is a supervised education algorithm. In general, this algorithm is used in two ways: to estimate the density function of distribution of training data and to classify test data based on education patterns. The results of estimating the accuracy of the prediction of each group record from the dataset using the algorithm (KNN) are shown in Table 6.

Table 6
The results of the implementation of KNN

	Truly bad	Truly good	Class precision
Predicted bad	163	90	64.43%
Predicted good	105	410	79.61%
Class recall	60.82%	82.00%	

$$Accuracy = (TP + TN)/(TP + FN + FP + TN) \times 100 = 74.62$$
$$Precision\ (true\ Bad\ ) = \frac{TP}{TP + FP} \times 100 = 64.43\%$$
$$Precision\ (true\ Good) = \frac{FP}{TP + FP} \times 100 = 79.61\%$$
$$Recall\ (true\ Bad\ ) = \frac{TP}{TP + FN} \times 100 = 60.82\%$$
$$Recall\ (true\ Good) = \frac{FN}{TP + FN} \times 100 = 82\%$$

As we can see, the accuracy of the model is 74.62%.

3.3. SVM method

One of the main advantages of SVM method is that the solution is unique, and the other advantage of this method is that it does not depend on the number of educational samples and can work well with the number of features and the number of examples. Classification with this method is known as one of the most effective categorization methods compared with the other machine learning algorithms. As mentioned earlier, SVM was considered as one of the most successful classifiers (due to the simple idea, based on which it was expressed and its proper function) to address this issue for the final classification of the customers. The results obtained from the backup machine can be seen in Table 7.

Table 7
The results of the implementation of SVM

	Truly bad	Truly good	Class precision
Predicted bad	127	47	72.99%
Predicted good	141	453	76.26%
Class recall	47.39%	90.60%	

Thus, we reach an accuracy of the 75.52% for the implementation of SVM which is calculated as follows,

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \times 100 = 75.52\%$$

$$Precision\ (true\ Bad\ ) = \frac{TP}{TP + FP} \times 100 = 72.99\%$$

$$Precision\ (true\ Good) = \frac{FP}{TP + FP} \times 100 = 76.26\%$$

$$Recall\ (true\ Bad\ ) = \frac{TP}{TP + FN} \times 100 = 47.39\%$$

$$Recall\ (true\ Good) = \frac{FN}{TP + FN} \times 100 = 90.60\%$$

3.4. Naive Bayes

The Naive Bayes algorithm is a simple method for categorizing phenomena based on the probability of occurrence. Based on the inherent features of probability (especially the probability of subscription), the outcome of the initial training will provide good results. Despite the design issues and the assumptions that apply to this method, the method is appropriate for categorizing most issues in the real world. Table 8 shows the results of the implementation of Naive Bayes.

Table 8
The results of the implementation of Naive Bayes

	Truly bad	Truly good	Class precision
Predicted bad	159	79	66.81%
Predicted good	109	421	79.43%
Class recall	59.33%	84.20%	

In this method, we obtain an accuracy of the 75.52% for the implementation of Naive Bayes which is calculated as follows,

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \times 100 = 75.52\%$$

$$Precision\ (true\ Bad\ ) = \frac{TP}{TP + FP} \times 100 = 66.81\%$$

$$Precision\ (true\ Good) = \frac{FP}{TP + FP} \times 100 = 79.43\%$$

$$Recall\ (true\ Bad\ ) = \frac{TP}{TP + FN} \times 100 = 59.33\%$$

$$Recall\ (true\ Good) = \frac{FN}{TP + FN} \times 100 = 84.2\%$$

3.5. Logistics Regression

The Logistic Regression algorithm (Dong et al., 2010) is a regression statistical model for binary dependent variables such as death-life, healthy-sick, etc. This model can be considered as a generalized linear model which uses the logit function as a link function and the error follows a polynomial distribution. The use of this technique was mainly used at the beginning of the emergence of medical applications for the likelihood of occurrence of a disease. But today it is widely used in all fields of science. Table 9 demonstrates the results of our survey.

Table 9
The results of the implementation of Logistic Regression

	Truly bad	Truly good	Class precision
Predicted bad	149	64	69.95%
Predicted good	119	436	78.56%
Class recall	55.60%	87.20%	

In this method, we obtain an accuracy of the 76.17% for the implementation of Logistic Regression which is calculated as follows,

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \times 100 = 76.17\%$$

$$Precision\ (true\ Bad\ ) = \frac{TP}{TP + FP} \times 100 = 69.95\%$$

$$Precision\ (true\ Good) = \frac{FP}{TP + FP} \times 100 = 78.56\%$$

$$Recall\ (true\ Bad\ ) = \frac{TP}{TP + FN} \times 100 = 55.6\%$$

$$Recall\ (true\ Good) = \frac{FN}{TP + FN} \times 100 = 87.2\%$$

In summary, we can compare the performance of different methods in terms of Accuracy. Fig. 1 shows the results of the survey,

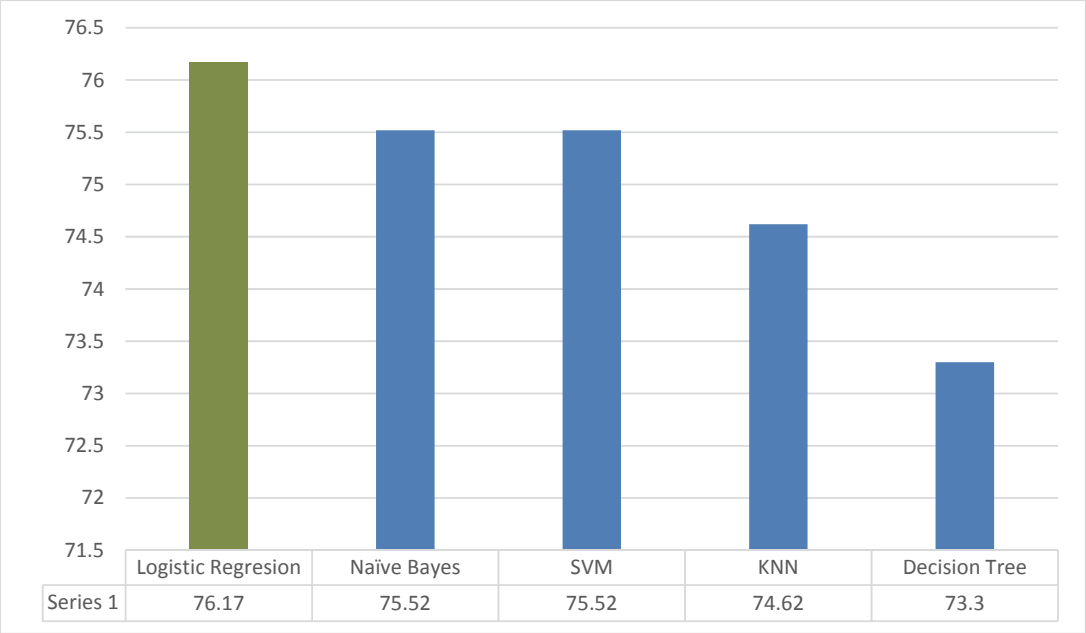

Fig. 1. The summary of the results of Accuracy for different method

4. Discussion and conclusion

Banks play a key role for the development of financial services and designing and implementing a credit rating system plays an important role for the development of financial services. Employing customer validation models can have an effective role in validating customers. Using data mining algorithms, a new method has been presented for customer diagnosis and validation. After interviewing experts in banking, we have identified various variables affecting customer validation and using the categorization method in data mining have been presented for customer ranking. This paper has used Decision Tree, K-nearest neighbor (KNN), Support Vector Machine (SVM), Naive Bayes, and Logistic Regression for data categorization to estimate credit ranking of bank customers in one of major banks in Middle East. The results indicate that Logistic Regression has been considered as the best method for ranking customers with the precision of 76.17% while Decision Tree was considered as the weakest technique with the precision of 73.30%.

References

Altman, E. I. (2000). Predicting financial distress of companies: revisiting the Z-score and ZETA models. *Stern School of Business, New York University*, 9-12.

Bosch, O., & Steffen, S. (2011). On syndicate composition, corporate structure and the certification effect of credit ratings. *Journal of Banking & Finance, 35*(2), 290-299.

Desai, V. S., Crook, J. N., & Overstreet Jr, G. A. (1996). A comparison of neural networks and linear scoring models in the credit union environment. *European Journal of Operational Research, 95*(1), 24-37.

Dong, G., Lai, K. K., & Yen, J. (2010). Credit scorecard based on logistic regression with random coefficients. *Procedia Computer Science, 1*(1), 2463-2468.

Duff, A., & Einig, S. (2009). Understanding credit ratings quality: Evidence from UK debt market participants. *The British Accounting Review, 41*(2), 107-119.

Durand, D. (1941). *Risk elements in consumer installment financing*: National Bureau of Economic Research, New York.

Elmer, P. J., & Borowski, D. M. (1988). An expert system and neural networks approach to financial analysis. *Financial Management, 12*, 66-76.

Fensterstock, A. (2003). Credit scoring basics. *BUSINESS CREDIT-NEW YORK THEN COLUMBIA MD-, 105*(3), 10-14.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics, 7*(2), 179-188.

Frawley, W. J., Piatetsky-Shapiro, G., & Matheus, C. J. (1992). Knowledge discovery in databases: An overview. *AI magazine, 13*(3), 57.

Goukasian, L., & Seaman, S. L. (2009). Comparison of classification models for predicting equipment lease and loan default. *The Journal of Equipment Lease Financing (Online), 27*(1), C1.

Horrigan, J. O. (1968). A short history of financial ratio analysis. *The Accounting Review, 43*(2), 284-294.

Koh, H. C., Tan ,W. C., & Peng, G. C. (2004). Credit scoring using data mining techniques. *Singapore Management Review, 26*(2), 25.

Lee, T.-S., & Chen, I.-F. (2005). A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. *Expert Systems with Applications, 28*(4), 743-752.

Limsombunchai, V., Gan, C., & Lee, M. (2005). An analysis of credit scoring for agricultural loans in Thailand.

Louzada, F., Ferreira-Silva, P. H., & Diniz, C. A. (2012). On the impact of disproportional samples in credit scoring models: An application to a Brazilian bank data. *Expert Systems with Applications, 39*(9), 8071-8078.

Malhotra, R., & Malhotra, D. K. (2002). Differentiating between good credits and bad credits using neuro-fuzzy systems. *European Journal of Operational Research, 136*(1), 190-211.

Min, J. H., & Lee, Y.-C. (2008). A practical approach to credit scoring. *Expert Systems with Applications, 35*(4), 1762-1770.

Ong, C.-S., Huang, J.-J., & Tzeng, G.-H. (2005). Building credit scoring models using genetic programming. *Expert Systems with Applications, 29*(1), 41-47.

Papaikonomou, V. L. (2010). Credit rating agencies and global financial crisis: Need for a paradigm shift in financial market regulation. *Studies in Economics and Finance, 27*(2), 161-174.

Thomas, L. C. (2000). A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International journal of forecasting, 16*(2), 149-172.

West, D. (2000). Neural network credit scoring models. *Computers & Operations Research, 27*(11-12), 1131-1152.