Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 7, July 2021: 10977 - 10994

Research Article

# Process Innovation for Credit Scoring Using Machine-Learning Approach for Small Financial Institutions

Sakchai Suthipipat

**Abstract**

Lending is important activity for overall economy, in which it helps fund investment for entrepreneurs to produce goods and services and also helps speed up consumption for the economy. However, credit default, which is the credit to the borrowers who cannot pay back the loan, can create draw back to the economy and cause higher cost of borrowing to all borrowers as financial institutions will increase interest to cover loss from default customers. Then, managing credit default risk is the key success for financial institutions and credit scoring is one of the tools that financial institutions use to manage their credit default risk for consumer loans. Machine Learning with supervised learning technique has been used to develop credit scoring model to classify good customers from default customers for many years. However, due to its complexity and less friendly than other techniques i.e. statistic or judgement method, the use of machine learning to build credit scoring model is limited to only large financial institutions, especially in Thailand market. This study aims to focus on building credit scoring model using supervised learning for medium to small financial institutions in Thailand, in which there are more limitations than large financial institutions in terms of size and quality of credit dataset. This study also focus on imbalanced data problem between majority and minority class of the dataset, which normally number of good customers always dominates number of default customers.

***Keywords:*** *supervised learning technique for credit scoring, imbalanced data problem*

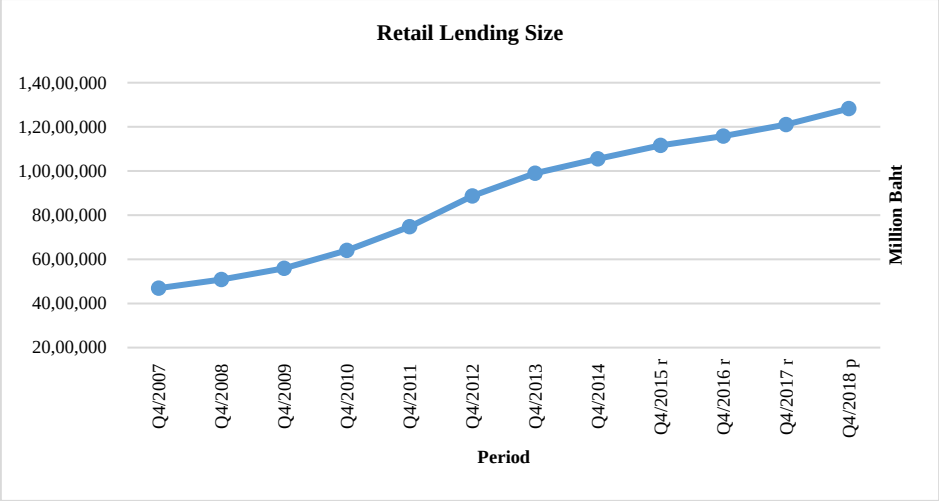
## Introduction

### Background and Rationale

Lending is essential to the economy of a country that helps individual or corporations to gain access to capital for their production of goods or services**.** It also help stimulate demand for consumption. Financial institutions like commercial banks are primary vehicles that perform lending activities. Non commercial banks, so called non banks, such as cooperatives, credit card companies, leasing or hire-purchase companies, and personal loan companies are also performing lending functions in the economy**.** They all are providing various type of loans including retail lending i.e. mortgage loans for buying houses, hire purchase loans for buying vehicles, credit card loans, personal loans. According to the Bank of Thailand, by the end of 2018, size of retail lending in Thailand was approximately 12**.**8 trillion baht, accounting for 78**%** of gross domestic product **(**gross domestic product or GDP**)**, comprising of loans by commercial banks of 11**.**35 trillion baht and more than 1**.**65 trillion baht by non banks.

**Fig**. **1**: **Size of Retail Lending in Thailand**

Technopreneurship and Innovation Management, Chulalongkorn University, Bangkok, Thailand




*Source: Bank of Thailand*

Counterparty credit default risk is one of the key problems in lending business. The risk is defined as the tendency that the borrower will be unable to meet their obligations once credit is granted. Thus adequate information about borrowers are needed in order to minimize counterparty credit risk from the business. However, as there is still information asymmetry, where two parties have unequal knowledge about each other's' information, especially for the lender side. This problem will lead to adverse selection or the problem of distinguishing between good customers from bad customers. This problem will lead to Lemon Market problem, where the lender will charge higher interest rates to cover the uncertainty about their counterparty's credit default risk. Thus economic inefficiency arises. According to the Bank of Thailand, by the end of 2018, non-performing loans or NPL in the commercial banking system stood at 443.387 million baht, or 3% of gross domestic product. NPLs for retail loans accounted for one fourth of total NPLs in the system. Financial institutions, thus, need to be considerate when granting credit to avoid losses from default customers.

Proper credit risk management will increase financial institutions' profit and helps sustain overall economy. There are 2 approaches to evaluate counterparty credit risk for financial institutions, namely judgmental approach and systematic approach. Judgmental analysis is a method of approving loan based on lender's judgment using their past experiences. Neither will this process employ any algorithms nor empirical process to determine credit. The 5 C's of credit which includes character, capacity, capital, collateral and conditions has been used as qualitative analysis to determine credit risk for judgmental approach. Credit scoring, a mathematically-based tools that ranks the borrower by the probability that default risk of may arise, is one of the tools used to determine credit on judgmental-based basis. The approach is suitable for large businesses and governments, than small corporations or individuals. Systematic approach is more suitable to evaluate credit risk for small businesses and individuals as there is a large number of customers. This method is less costly and the result is more consistent than judgmental approach. Only 1% enhancement on the accuracy of credit scoring system, would significantly improve the profits of financial institutions (Armaki et al. (2017).

Credit scoring has been introduced since 1950s using historical data to develop a statistic model (Thomas, Edelman, and Crook 2002). Linear discriminant and logistic regression were statistic techniques used to develop the credit scoring model. The statistic credit scoring techniques are often criticized due to their strong model assumptions that requires linear relationship between dependent and independent variables (Lee et al. 2006; Vojtek and Koeenda 2006). If the relationship between both variables are non-linear, the model accuracy will be significantly deteriorate (Lacher et al. 1995; Lee and Chen 2005).

Because of new computing technology has been continuously developed, machine learning credit scoring techniques are introduced to improve performance of traditional statistic credit scoring techniques **(**Lee, Chiu et al**.** 2006**).** Supervised learning technique is widely used for classification problem between good and bad loans . The most commonly-used algorithms for supervised learning are artificial neural networks **(**Lee et al**.** 2002**)**, genetic programming **(**Ong, Huang, and Tzeng 2005**)**, decision tree **(**Lee et al**.** 2006**)**, and *k*-nearest neighbor **(**Henley and Hand 1996**)**, etc**.** Machine learning has been proven to have better prediction performance than statistical techniques **(**Eddy and Bakar 2017; Fausett 1994; Crook, Edelman, and Thomas 2007; Huang et al**.** 2004; Ong, Huang, and Tzeng 2005**)** and it also support large-scale data calculations**.**

**Research Scope**

1**.** This research is a quantitative research applying credit dataset from the real world for credit scoring model development by using machine learning technology

2**.** The credit dataset was collected from a non-commercial bank, which provides hire-purchase loans to retail customers in Greater Bangkok and in other major metropolitan areas

**Research Objectives**

1**)** Study supervised learning techniques to develop credit scoring model and apply them to the credit scoring model development process that maximize the accuracy of the model

2**)** Develop credit scoring models that are suitable for the business environment

3**)** Test the accuracy of the credit scoring model

4**)** Study how to implement the credit scoring model in the business, study how to protect intellectual property, and evaluate the feasibility for commercializing the process

**Research methodologies**

This research has methods and procedures as follows**:**

Step **1:** To review the relevant literature

Step **2:** To provide credit dataset required to develop the model

Step **3:** To study the ML algorithm used in the model development process

Step **4:** To build up model

Step **5:** To test the performance of the model

Step **6:** To Implement the model in actual operation

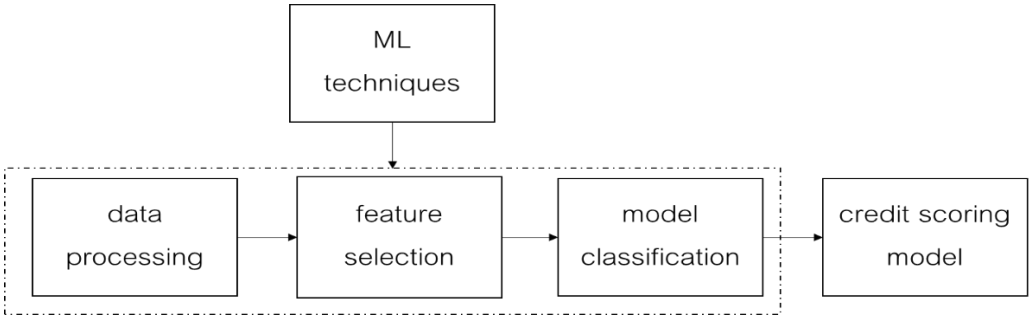
Step **7:** To study the acceptance of the innovation using Technology Acceptance model (TAM) and study the commercial feasibility and methods of protect intellectual property

**Academic and practical contributions**

1. Academic contribution
    - To create knowledge regarding the development of credit scoring models for non-commercial banks
    - To apply the process organizations by taking the model as a model to develop, and to build on the existing knowledge of innovation and development of the Credit Scoring model
    - To improve the predictive performance of credit scoring to help reduce credit risk
2. Practical contribution
    - User level: to develop a tool to assess customer credit risk that help enhance credit decision
    - Industrial level: for corporate executives, lending operator can use the model to enhance their capabilities in the competition in the lending business

**Conceptual framework**

**Fig**. **2**: **Conceptual Framework**




Source**:** Researcher

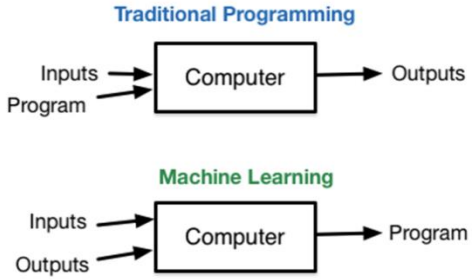
This considers the process of developing a credit scoring model, which has three key steps**:** 1**)** data processing, 2**)** feature selection, and 3**)** model classification.

## Literature Review

### Machine Learning

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to learn from data, identify patterns, make decision based on algorithms and improve from experience without being intervened by human or explicitly programmed. Machine learning focuses on the development of computer programs that can access and learn from dataset. Unlike traditional programming, which refers to any programs manually created that feeds input data and runs on a computer to produce the output, machine learning are the process that input and output (or labels) are used with algorithms on a computer to create the program. The idea of the difference between traditional programming and machine learning is shown as below figure.

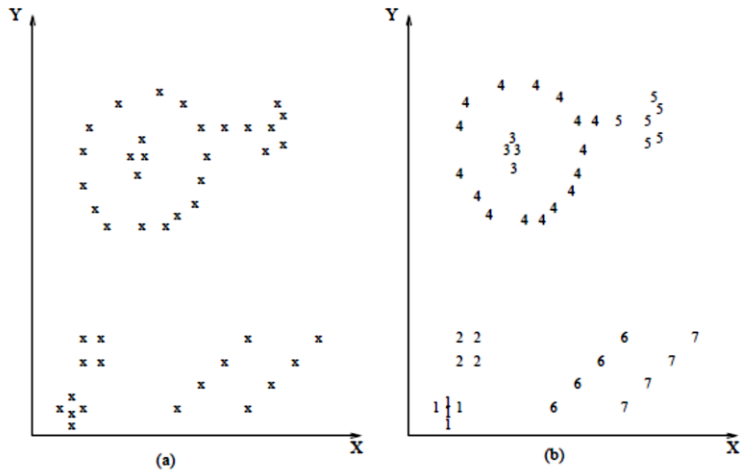
**Fig 3**: **Difference Between Traditional Programming and Machine Learning**




Machine leaning can be categorized into 2 approaches, supervised and unsurprised learning. Supervised learning is a machine learning approach that's defined by its use of output or labels. These labelled datasets are used to train algorithms to classify testing data or making predictive outcomes. Using labeled inputs and outputs, the model can measure its accuracy and learn over time. Supervised learning is used to solve classification problems.

Unsupervised learning uses machine learning algorithms to cluster unlabeled data sets. These algorithms discover hidden patterns in data without the need for labels, unlike supervised learning. Unsupervised learning can be used to solve clustering problems.

**Fig 4 Sample of Clustering**




**Credit scoring**

Credit scoring **(**Mester, 1997b**)** is a statistical method used to predict the probability that a loan applicant will default or become delinquent. It is already widely used for consumer lending and is becoming more commonly used in mortgage lending. To build a scoring model, or "scorecard," developers analyze historical data on the performance of previously made loans to determine which borrower characteristics are useful in predicting whether theloan performed well. A well-designed model should give a higher percentage of high scores to borrowers whose loans will perform well and a higher percentage of low scores to borrowers whose loans won't perform well. But no model is perfect, and some bad accounts will receive higher scores than some good accounts

Credit scoring system is a computerized process producing a score according to various relevant characteristics of the borrower, such as income, profession, age, wealth, previous loans, etc. The five C's of credit is widely used by lender to evaluate loan applicant's creditworthiness by considering a borrower's character, capacity to make payments, loan conditions, available capital and collateral.

Statistical credit scoring model was firstly used since the 1950'. to manage and diversify borrowers default risks **(**Thomas, Edelman, and Crook 2002**).** Popular techniques used to create credit scoring models are linear discriminant, analysis, and logistic regression, all of which have limitations from its assumption that require linear relationship between dependent and independent variables**.** Statistical models also have problems arising from independent variables having relationships with each other creating multicolinearity problems, in addition to the limitation from the size of historical data. These problems have impacted to prediction accuracy of statistical credit scoring model (Lacher et al**.** 1995; Lee and Chen 2005**)** Supervised learning is introduced to reduce the limitations of statistical models and have been used for improving performance from the original statistical model **(**Eddy and Bakar 2017; Fausett 1994; Crook, Edelman, and Thomas 2007; Huang et al**.** 2004; Ong, Huang, and Tzeng 2005**)**. Supervised learning approaches with artificial neural networks **(**Lee et al**.** 2002**)** genetic programming **(**Ong, Huang, and Tzeng 2005**)** decision tree **(**Lee et al**.** 2006**)** and k-nearest neighbor **(**Henley and Hand 1996**)** are among techniques that is sued to construct credit scoring model and they have proven with better predictive performance**.**

Based on a review of credit scoring literature using machine learning, the researcher searched the Web of Science database on the web, www.webofknowledge.com since March 2019, to analyze common interest concerning credit scoring problems, including to search for the research gap using the keyword "Credit Scoring" along with the associated keywords, divided into four categories of credit scoring problems, which are**:**

1**)** Common problems of machine learning such as data mining, machine learning, artificial intelligence and evolutionary computing**.**

2**)** Problems with model development such as classification, algorithms, classification techniques, supervised, unsupervised, clustering and predictive modeling etc**.**

3**)** Problems with feature selection such as feature selection, variable selection, parameter optimization, attribute selection

4**)** Data preparation problems or data processing such as samples selection, reject inference, imbalanced data, imbalanced problem, unbalanced data and transfer learning

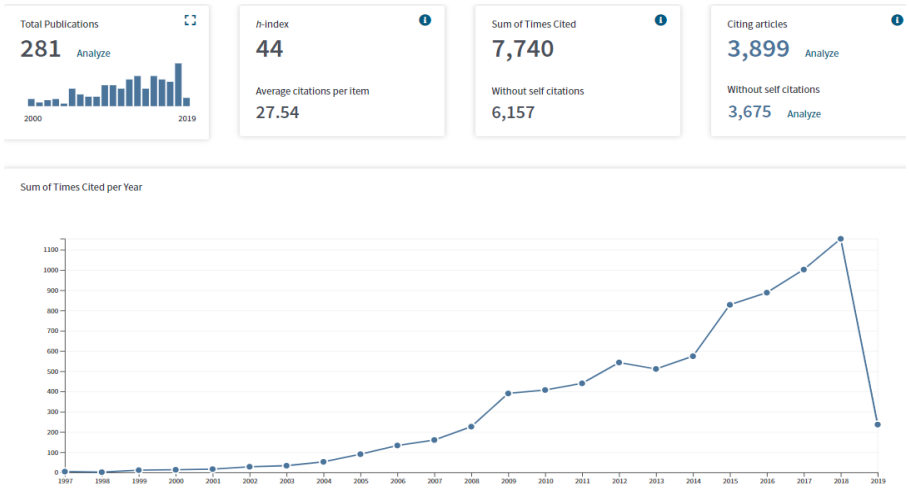
The result from the research has evidenced that most researches focus on the problems associated with the development of credit scoring model**.** The keywords "Classification" and "Algorithms" are mostly found from the search with 281 and 179 journals, respectively. Study with "Data Mining" keyword has been found in 71 publications. "Classification problem" are referred more than 7,740 times in journals concerning credit scoring, especially in 2018, with more than 1,000 times reference, mostly published on Computer Science Artificial Intelligence journal with more than 139 studies, followed by Operational Research Development Science journal, with the details shown in the figure below**:**

**Table 1: Search Result of Development of Credit Scoring Model from database of Web of Science**

| Keywords **(**search with **"**credit scoring**")** | No**.** of Search Result |
|---|---|
| **Generic keywords:** | |
| - data mining | 71 |
| - machine learning | 68 |
| - artificial intelligence | 25 |
| - evolutionary computing | 1 |
| **Model classification:** | |
| - classification | 281 |
| - algorithms | 179 |
| - clustering | 30 |
| - classification techniques | 17 |
| - supervised | 17 |
| - unsupervised | 4 |
| - predictive modeling | 2 |
| **Feature selection:** | |
| - feature selection | 47 |
| - variable selection | 17 |
| - parameter optimization | 3 |
| - attribute selection | 2 |
| **Data processing:** | |
| - sample selection | 22 |
| - reject inference | 19 |
| - imbalanced data | 10 |
| - imbalanced problem | 8 |
| - unbalanced data | 6 |
| - transfer learning | 1 |

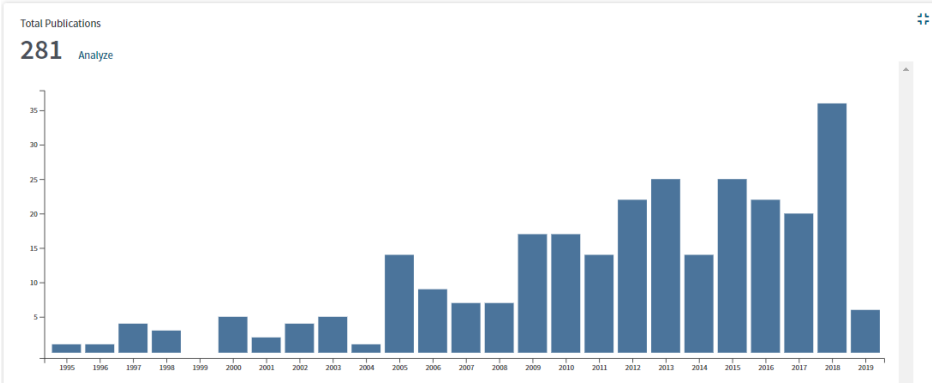
Source**:** web of science, collected on 10 March 2019

**Fig 5: Search Result of "Classification" with "Credit Scoring" by Time Cited per Year**



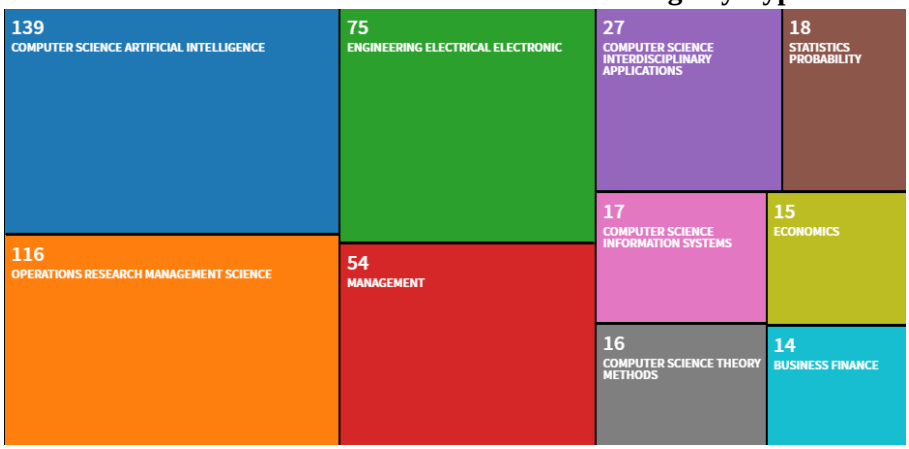

Source**:** web of science, collected on 10 March 2019

**Fig 6: Search Result of "Classification" with "Credit Scoring" by Year of Publication**




Source**:** web of science, collected on 10 March 2019

**Fig 7: Search Result of "Classification" with "Credit Scoring" by Type of Publication**




Source**:** web of science, collected on 10 March 2019

**Credit Scoring Model Development Process**

The development of credit scoring model **(**Siddiqi 2012**)**, consists of three main steps**:** 1**)** data preparation, 2**)** feature selection and 3**)** modeling and performance Measurement.

The first step is data processing using historical credit data **(**Chi and Hsu 2012**)**, which includes borrowers' attributes such as gender, age, marital status, occupation, together with other information such as past repayment history, sources of income for repayment, collateral and credit conditions, etc**.** To prepare as an independent variables, Eddy and Bakar **(**2017**)**, there should be information in four dimensions regarding 1**)** financial information 2**)** personal qualifications 3**)** occupation information 4**)**

borrowing behavior with financial institutions such as information from credit bureaus, which is in line with the 5Cs principles. It is noteworthy that there is no related research focusing on the importance of attributes (Abdou and Pointon 2011). This is somehow because the financial institutions do not want to disclose which variables are crucial in their credit decision from ethical perspective. The model performance depends on the large amount of data on both good debt and bad debt (Chi and Hsu 2012). The sample population should have at least 1,000 samples, of which 500 samples for good debt or bad debt each. The more credit dataset, the better the model performance (Mester 1997b). The historical dataset is divided into two groups which are training set and test set, respectively (Lee et al. 2006). This process will consume more than 80% of credit scoring model development (Piramuthu 1999b).

Since model development process requires large sample size of historical data, it is a barrier for new enterprise with limited past credit dataset to enter the credit business. The problem from imbalanced data also arise as it is common that the good debt population is significantly more than bad debt population. The credit dataset also have bias from gaining only borrowers whose loans already are approved by financial institutions and omitting the population that are rejected during the application process. This bias tends to favor the prediction of good debt more than bad debt, though bad debt prediction is the key issues in credit scoring model development. There are two approaches to address the issue from imbalanced data: 1) data level management by over-sampling and under-sampling to make both good and bad data balanced (Crane and Finlay 2012, Khemakherm, Ban Said, and Boujelpene 2018) and 2) algorithmic level by using learning algorithms to construct balanced dataset (Adam et al. 2010, Zeng and Sao 2009). The problem arising from not having populations whose rejected during credit decision can be addressed by reject inference, in which the dataset will include the applicant whose get denied to construct using "missing data mechanism" (Crook and Banasik 2004, Banasik, Crook, and Thomas 2003; Hand and Henley 1997).

Feature selection is the second process in credit scoring model development. This is the process of selecting the best subset variables from the whole set of variables (Dash and Liu 1997), eliminating the irrelevant variables. This process will help speed up the modeling process and reduces the costs incurred from model calculations (Edla et al 2018). Many techniques such as filter and wrapper (Somol et al. 2005) are used in this process (George 2000).

Model development by classification techniques is the last step for credit scoring model development. This process will divide dataset into two groups; training and testing set. The process will employ supervised learning technique using algorithms to learn the patterns from training dataset with specific labels, and construct model. After model is constructed, testing dataset is used to test the model performance. Algorithms widely used in supervised techniques are decision tree , artificial neural network or ANN (Desai, Crook, and Overstreet Jr 1996; Desai et al. 1997, Malhotra and Malhotra 2003, Jensen 1992, Piramuthu 1999a), genetic algorithms or GA (Marques, Garcia, and Sanchez 2013). The study of Wang et. Al (2001) suggests that there is no superior algorithms, while Yu, Wang and Lai (2008), Hung and Chen (2009) suggest that ensemble technique deliver better model performance. Bagging (Breiman 1996) and boosting (Freund and Schapire 1997) are the ensemble techniques that make random sampling and generate several training data sets, instead of only one training dataset, and use one classifier to construct the model with final decision made on average or voting of every models constructed from each subset of training dataset.

Hybridization technique or hybrid model, which is the mix between classification (supervised learning) and clustering (unsupervised learning) is used to enhance model performance (Armaki et al. 2017) using Australia and German credit datasets with 99.71% and 99.8% accuracy. The result is compared with other techniques are shown in table below.

**Table 2: Comparison of Accuracy by Credit Scoring Model Techniques from Australia Dataset**

| No. | Model | Accuracy | Year | Author(s) |
|---|---|---|---|---|
| 1 | (KNN-NN-SVMPSO)-(DL)-(DBSCAN) | 99.71 | 2017 | Armaki et al. (2017) |
| 2 | MC-LR (Intersection) | 99.11 | 2013 | Tsai and Hsu (2013) |
| 3 | Hybrid SOM-KM-NN | 97.98 | 2005 | Hsieh (2005) |
| 4 | ANN | 97.32 | 2008 | Tsai and Wu (2008) |
| 5 | AMMLP | 92.75 | 2011 | Marcano-Cedeno et al. (2011) |
| 6 | Gaussian classifier | 92.60 | 2005 | Somol et al. (2005) |
| 7 | VBDTM | 91.97 | 2010 | Zhang et al. (2010) |
| 8 | Hybrid NN | 91.61 | 2014 | Tsai and Hung (2014) |
| 9 | PSO-SVM | 91.03 | 2008 | Lin et al. (2008) |
| 10 | LS-SVM | 90.40 | 2003 | Baesens et al. (2003) |
| 11 | MLP | 90.20 | 2008 | Tsai (2008) |
| 12 | Parallel Random Forest | 89.40 | 2016 | Van Sang, Nam, and Nhan (2016) |
| 13 | 25GP | 89.17 | 2006 | Huang, Tzeng, and Ong (2006) |
| 14 | DeepSVM | 88.98 | 2016 | Qi et al. (2016) |
| 15 | Genetic Fuzzy classifier | 88.60 | 2010 | Lahsasna, Ainon, and Wah (2010) |
| 16 | Genetic programming | 88.27 | 2005 | Ong, Huang, and Tzeng (2005) |
| 17 | RS-Bagging DT | 88.17 | 2012 | Wang et al. (2012) |
| 18 | GNG+MARS | 88.10 | 2016 | Ala'raj and Abbod (2016) |
| 19 | RBF-SVM | 87.52 | 2011 | Ping and Yongheng (2011) |
| 20 | SVDD+FSVM | 87.25 | 2016 | Shi and Xu (2016) |
| 20 | Mixture-of-experts network | 87.25 | 2000 | West (2000) |
| 22 | RS-LMNC | 87.05 | 2009 | Nanni and Lumini (2009) |
| 23 | Adopted CBA | 86.96 | 2006 | Lan et al. (2006) |
| 24 | SVM+GA | 86.90 | 2007 | Huang, Chen, and Wang (2007) |
| 25 | ECSC | 86.86 | 2016 | Xiao, Xiao, and Wang (2016) |
| 26 | GR-GA-SVM | 86.84 | 2010 | Liu, Fu, and Lin (2010) |
| 27 | Bayes | 86.70 | 2007 | Hoffmann et al. (2007) |
| 28 | CLC | 86.52 | 2009 | Luo, Cheng, and Hsieh (2009) |
| 28 | LDA + SVM | 86.52 | 2010 | Chen and Li (2010) |
| 30 | LibSVM | 86.38 | 2008 | Peng et al. (2008) |
| 31 | FA-MLP | 86.08 | 2009 | Tsai (2009) |
| 32 | SVM | 85.70 | 2007 | Martens et al. (2007) |

Source: (Armaki et al. 2017)

**Table 3: Comparison of Accuracy by Credit Scoring Model Techniques from German Dataset**

| No. | Model | Accuracy | Year | Author(s) Year |
|---|---|---|---|---|
| 1 | (KNN-NN-SVMPSO)-(DL)-(DBSCAN) | 99.80 | 2017 | Armaki et al. (2017) |
| 2 | MC-LR (Intersection) | 99.18 | 2013 | Tsai and Hsu (2013) |
| 3 | Hybrid SOM-KM-NN | 98.46 | 2005 | Hsieh (2005) |
| 4 | MLP+FS | 97.20 | 2011 | Silva and Analide (2011) |
| 5 | LibSVM | 94.00 | 2008 | Peng et al. (2008) |
| 6 | Hybrid NN | 87.45 | 2014 | Tsai and Hung (2014) |
| 7 | CLC | 84.80 | 2009 | Luo, Cheng, and Hsieh (2009) |
| 8 | AMMLP | 84.67 | 2011 | Marcano-Cedeno et al. (2011) |

| | | | | |
|---|---|---|---|---|
| 9 | Gaussian classifier | 83.80 | 2005 | Somol et al. (2005) |
| 10 | DeepSVM | 83.70 | 2016 | Qi et al. (2016) |
| 11 | VBDTM | 81.64 | 2010 | Zhang et al. (2010) |
| 12 | PSO-SVM | 81.62 | 2008 | Lin et al. (2008) |
| 13 | 25GP | 79.49 | 2006 | Huang, Chen, and Wang (2007) |
| 14 | MLP | 79.11 | 2008 | Tsai (2008) |
| 15 | GNG+MARS | 79.00 | 2016 | Ala'raj and Abbod (2016) |
| 16 | ANN | 78.97 | 2008 | Tsai and Wu (2008) |
| 17 | HGA-NN | 78.90 | 2012 | Oreski, Oreski, and Oreski (2012) |
| 18 | FA-MLP | 78.76 | 2009 | Tsai (2009) |
| 19 | RS-Bagging DT | 78.36 | 2012 | Wang et al. (2012) |
| 20 | SVM+GA | 77.92 | 2007 | Huang, Chen, and Wang (2007) |
| 21 | Genetic programming | 77.34 | 2005 | Ong, Huang, and Tzeng (2005) |
| 22 | SVDD+FSVM | 77.30 | 2016 | Shi and Xu (2016) |
| 23 | LDA + SVM | 76.70 | 2010 | Chen and Li (2010) |
| 24 | RBF-SVM | 76.60 | 2011 | Ping and Yongheng (2011) |
| 25 | Mixture-of-experts network | 76.30 | 2000 | West (2000) |
| 26 | Parallel Random Forest | 76.20 | 2016 | Van Sang, Nam, and Nhan (2016) |
| 27 | Bayes | 76.00 | 2007 | Hoffmann et al. (2007) |
| 28 | GR-GA-SVM | 75.75 | 2010 | Liu, Fu, and Lin (2010) |
| 29 | Genetic Fuzzy classifier | 75.00 | 2010 | Lahsasna, Ainon, and Wah (2010) |
| 30 | RS-LMNC | 74.67 | 2009 | Nanni and Lumini (2009) |
| 31 | LS-SVM | 74.60 | 2003 | Baesens et al. (2003) |
| 32 | Adopted CBA | 74.40 | 2006 | Lan et al. (2006) |
| 33 | ECSC | 70.60 | 2016 | Xiao, Xiao, and Wang (2016) |

Source: Armaki et al. (2017)

**Model Performance Measurement (Model Accuracy)**

Once the model is constructed, the model performance is tested using data from testing set**.** The most popular method of performance testing is the confusion matrix and area under the curve, or AUC, is the pattern of the confusion matrix shown in Table 5**.**

**Table 4 Confusion Matrix**

| | | Prediction Value | |
|---|---|---|---|
| | | Ans Positive | Ans Negative |
| True Value | Ans Positive | true positive | false negative |
| | Ans Negative | false positive | true negative |

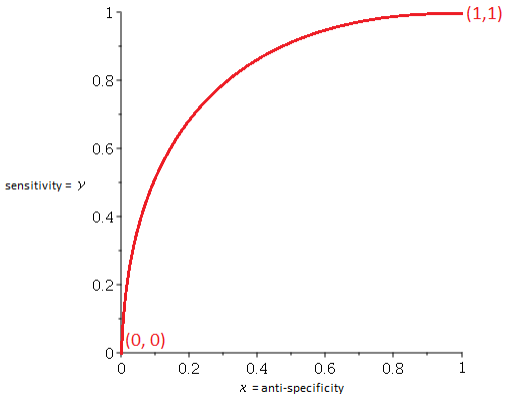
true positive **(**TP**)** is the number of positive data predicted as positive
true negative **(**TN**)** is the number of negative data predicted as negative
false positive **(**FP**)** is the number of negative data predicted as position
false negative **(**FN**)** is the number of positive data predicted as negative

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

The model performance can be measured using accuracy from the confusion matrix. The more the accuracy, the better the model performance.

ROC curve or receiver operating characteristic curve is another model performance measurement for credit scoring model development. The ROC curve is a graph showing the performance of a classification model at all classification thresholds. This curve are measured the relationship between two parameters: true positive rate (sensitivity) and false positive rate (specificity). The larger the area under the curve, the better the performance, since the performance of the model should have a high sensitivity and a high specificity**.** High specificity will result in low false positive rate**.** The ROC curve is shown in the below figure.

**Fig 8 Receiver Operating Characteristic (ROC) Curve**




This research uses confusion matrix to measure the model performance,**.**

**Research Methodology**

This research is a quantitative research**.** It is experimental research from secondary data to be tested using machine learning techniques to study and construct a development process of credit scoring model which is efficient and can be used in the real way**.**

**Table 5: Credit Scoring Sample Set to be used in this study**

| Sample set | Proportion of Good/Bad Debt | No**.** of Sample Set | No**.** of Ordinal | No**.** of Nominal | Total Attributes |
|---|---|---|---|---|---|

| | | | Attributes | Attributes | |
|---|---|---|---|---|---|
| Sample set from case study company | 8,652/2760 | 11,412 | 8 | 7 | 15 |

Source**:** Researcher

**Table 6: Details of Attributes**

| No**.** | Attribute Name | Attribute Type | Remarks |
|---|---|---|---|
| 1 | Gender | Text | Male/Female |
| 2 | Age | Number | |
| 3 | Marital status | Text | Single/Married/Divorce/Widow |
| 4 | Nationality | Text | |
| 5 | Zip code | Text | |
| 6 | Position | Text | |
| 7 | Occupation | Text | |
| 8 | Size of Monthly Income | Number | |
| 9 | Percentage of down payment | Number | |
| 10 | Type of collateral | Text | Motorcycle brand |
| 11 | Sub type of collateral | Text | Motorcycle model |
| 12 | Loan size | Number | |
| 13 | Contract period | Number | |
| 14 | Installment to revenue ratio | Number | |
| 15 | Interest rate | Number | |

**Table 7: Attributes Classified by 5Cs of Credit**

| Characters | Capacity | Capital | Collateral | Conditions |
|---|---|---|---|---|
| 1. Gender<br>2. Age<br>3. Marital Status<br>4. Nationality<br>5. Zip code<br>6. Occupation<br>7. Position | 8. Size of Monthly Income | 9. Percentage of Down Payment | 10. Motorcycle Brand<br>11. Motorcycle Model | 12. Loan Size<br>13. Loan Period<br>14. Installment to revenue ratio<br>15. Interest rate |

**Table 8: Detail of Labels**

| No**.** | Label Name | Attribute Type | Remarks |
|---|---|---|---|
| 1 | Default status | Text | 0 is Good Debt<br>1 is Bad Debt |

**The scope of the research and the sample**

This research will use data from credit from a credit company in Thailand, case study, which is credit data from 1 November 2017 to 31 December 2017. The identity of borrowers cannot be identified. The variables are in line with the principle of the 5 c of credit including information of demographic, income, capital, collateral and loan conditions. Supervised learning is used to classify good and bad debts.

**Sample credit dataset used in this study**

The dataset of 11,445 loan transactions is used in the study, of which 8,652 records are good debts and 2,750 records are bad debts.

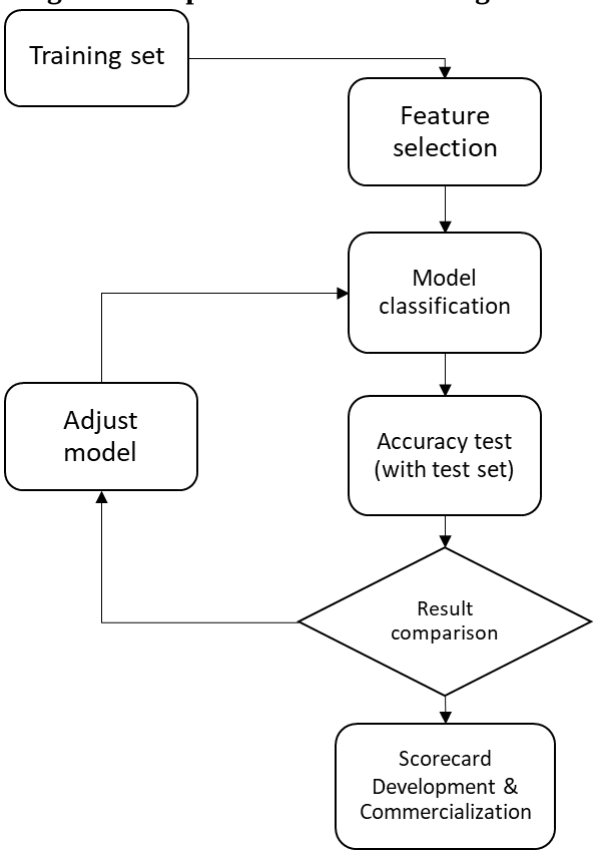
**Machine Learning Tool**

This study will use Rapidminer program to construct credit scoring model.

**The operational research process:**

The research will be carried out as follows**:**

**Fig 9: Development of Credit Scoring Model**




1**)** Study how to prepare the datasets used in the data processing procedure;

2**)** Study the factors of suitable sample size which will result in the best model performance;

3**)** Study factors for managing the imbalanced date problem of dataset between good debt and bad debt, which resulted in the best model performance;

4**)** Study techniques for feature selection

5**)** Study of the model development techniques using supervised learning from different types of classifier that resulted in best performance

6**)** Study factor techniques for model development by an ensemble method was used that resulted in superior performance over the single classifier model**.**

Each step is detailed in the following steps**:**

1**)** Clean up the data and properly structure the dataset to be used in the credit modeling;

2**)** Test the increase and decrease of the dataset size to affect the accuracy of the credit scoring model;

3**)** Test the process that handle the imbalanced data problem

4**)** Test the use of wrapper and filter techniques that resulted in the highest accuracy from the model;

5**)** Construct models from different classifiers

6**)** Develop the model by using an ensemble classifier with boosting and bagging techniques to compare the accuracy with the two using a single classifier**.**

## References

1. Abdou, Hussein A, and John Pointon. *2011.* 'Credit scoring, statistical techniques and evaluation criteria*:* a review of the literature', *Intelligent Systems in Accounting, Finance Management, 18: 59-88.*

2. Adam, Asrul, Ibrahim Shapiai, Zuwairie Ibrahim, Marzuki Khalid, Lim Chun Chew, Lee Wen Jau, and Junzo Watada. *2010.* "A Modified Artificial Neural Network Learning Algorithm for Imbalanced Data Set Problem." In *2010 2nd International Conference on Computational Intelligence, Communication Systems and Networks, 44-48.* IEEE.

3. Aggelidis, Vassilios P, and Prodromos D Chatzoglou. *2009.* 'Using a modified technology acceptance model in hospitals', *International journal of medical informatics, 78: 115-26.*

4. Ajzen, Icek, and Martin Fishbein. *1980.* 'Understanding attitudes and predicting social behaviour'.

5. Akhavein, Jalal, W Scott Frame, and Lawrence White. *2005.* 'The diffusion of financial innovations*:* An examination of the adoption of small business credit scoring by large banking organizations', *The Journal of Business, 78: 577-96.*

6. Ala'raj, Maher, and Maysam F *%*J Expert Systems with Applications Abbod. *2016.* 'A new hybrid ensemble credit scoring model based on classifiers consensus system approach', *64: 36-55.*

7. Armaki, Ali Ghasemy, Mir Feiz Fallah, Mahmoud Alborzi, and Amir Mohammadzadeh. *2017.* 'A Hybrid Meta-Learner Technique for Credit Scoring of Banks*'* Customers', *Engineering, Technology Applied Science Research, 7: 2073-82.*

8. Baesens, Bart, Tony Van Gestel, Stijn Viaene, Maria Stepanova, Johan Suykens, and Jan Vanthienen. *2003.* 'Benchmarking state-of-the-art classification algorithms for credit scoring', *Journal of the Operational Research Society, 54: 627-35.*

9. Banasik, John, Jonathan Crook, and Lyn Thomas. *2003.* 'Sample selection bias in credit scoring models', *Journal of the Operational Research Society, 54: 822-32.*

10. Bequé, Artem, and Stefan Lessmann. *2017.* 'Extreme learning machines for credit scoring*:* An empirical evaluation', *Expert Systems with Applications, 86: 42-53.*

11. Bofondi, Marcello, and Francesca *%*J Review of Industrial Organization Lotti. *2006.* 'Innovation in the retail banking industry*:* the diffusion of credit scoring', *Review of Industrial Organization, 28: 343-58.*

12. Breiman, Leo. *1996.* 'Bagging predictors', *Machine learning, 24: 123-40.*

13. Castillo, Flor, Kenric Marshall, James Green, and Arthur Kordon. *2003.* "A methodology for combining symbolic regression and design of experiments to improve empirical model building." In *Genetic and Evolutionary Computation Conference, 1975-85.* Springer.

14. Chen, Fei-Long, and Feng-Chia Li. *2010.* 'Combination of feature selection approaches with SVM in credit scoring', *Expert Systems with Applications, 37: 4902-09.*

15. Chi, Bo-Wen, and Chiun-Chieh Hsu. *2012.* 'A hybrid approach to integrate genetic algorithm into dual scoring model in enhancing the performance of credit scoring model', *Expert Systems with Applications, 39: 2650-61.*

16. Crone, Sven F, and Steven Finlay. *2012.* 'Instance sampling in credit scoring*:* An empirical study of sample size and balancing', *International Journal of Forecasting, 28: 224-38.*

17. Crook, Jonathan, and John Banasik. *2004.* 'Does reject inference really improve the performance of application scoring models?', *Journal of Banking Finance, 28: 857-74.*

18. Crook, Jonathan N, David B Edelman, and Lyn C Thomas. *2007.* 'Recent developments in consumer credit risk assessment', *European Journal of Operational Research, 183: 1447-65.*

19. Dash, Manoranjan, and Huan Liu. *1997.* 'Feature selection for classification', *Intelligent data analysis, 1: 131-56.*

20. Davis, Fred D, Richard P Bagozzi, and Paul R *%*J Management science Warshaw. *1989.* 'User acceptance of computer technology*:* a comparison of two theoretical models', *Management science,*

*35: 982-1003.*

21. Desai, Vijay S, Daniel G Conway, Jonathan N Crook, and George A Overstreet Jr. *1997.* 'Credit-scoring models in the credit-union environment using neural networks and genetic algorithms', *IMA Journal of Management Mathematics, 8: 323-46.*

22. Desai, Vijay S, Jonathan N Crook, and George A Overstreet Jr. *1996.* 'A comparison of neural networks and linear scoring models in the credit union environment', *European Journal of Operational Research, 95: 24-37.*

23. Eddy, Yosi Lizar, and Engku Muhammad Nazri Engku Abu Bakar. *2017.* 'Credit scoring models*:* Techniques and issues'.

24. Edla, D. R., D. Tripathi, R. Cheruku, and V. Kuppili. *2018.* 'An Efficient Multi-layer Ensemble Framework with BPSOGSA-Based Feature Selection for Credit Scoring Data Analysis', *Arabian Journal for Science and Engineering, 43: 6909-28.*

25. Fausett, Laurene V. *1994. Fundamentals of neural networks: architectures, algorithms, and applications (*prentice-Hall Englewood Cliffs*).*

26. Fensterstock, Albert. *2005.* 'Credit scoring and the next step', *Business credit, 107: 46-49.*

27. Féraud, Raphael, and Fabrice Clérot. *2002.* 'A methodology to explain neural network classification', *Neural Networks, 15: 237-46.*

28. Frame, W Scott, Larry D Wall, and Lawrence J White. *2018.* 'Technological Change and Financial Innovation in Banking*:* Some Implications for Fintech'.

29. Freund, Yoav, and Robert E Schapire. *1997.* 'A decision-theoretic generalization of on-line learning and an application to boosting', *Journal of computer system sciences, 55: 119-39.*

30. George, Edward I. *2000.* 'The variable selection problem', *Journal of the American Statistical Association, 95: 1304-08.*

31. Hand, David J, and William E Henley. *1997.* 'Statistical classification methods in consumer credit scoring*:* a review', *Journal of the Royal Statistical Society: Series A, 160: 523-41.*

32. Henley, WE, and David J Hand. *1996.* 'A k-nearest-neighbour classifier for assessing consumer credit risk', *The statistician: 77-95.*

33. Hens, Akhil Bandhu, and Manoj Kumar Tiwari. *2012.* 'Computational time reduction for credit scoring*:* An integrated approach based on support vector machine and stratified sampling method', *Expert Systems with Applications, 39: 6774-81.*

34. Hoffmann, Frank, Bart Baesens, Christophe Mues, Tony Van Gestel, and Jan Vanthienen. *2007.* 'Inferring descriptive and approximate fuzzy rules for credit scoring using evolutionary algorithms', *European Journal of Operational Research, 177: 540-55.*

35. Hsieh, Nan-Chen. *2005.* 'Hybrid mining approach in the design of credit scoring models', *Expert Systems with Applications, 28: 655-65.*

36. Huang, Cheng-Lung, Mu-Chen Chen, and Chieh-Jen *%*J Expert systems with applications Wang. *2007.* 'Credit scoring with a data mining approach based on support vector machines', *33: 847-56.*

37. Huang, Guang-Bin, Qin-Yu Zhu, and Chee-Kheong Siew. *2006.* 'Extreme learning machine*:* Theory and applications', *Neurocomputing, 70: 489-501.*

38. Huang, Jih-Jeng, Gwo-Hshiung Tzeng, and Chorng-Shyong Ong. *2006.* 'Two-stage genetic programming *(2*SGP*)* for the credit scoring model', *Applied Mathematics Computational Economics, 174: 1039-53.*

39. Huang, Zan, Hsinchun Chen, Chia-Jung Hsu, Wun-Hwa Chen, and Soushan Wu. *2004.* 'Credit rating analysis with support vector machines and neural networks*:* a market comparative study', *Decision support systems, 37: 543-58.*

40. Hung, Chihli, and Jing-Hong Chen. *2009.* 'A selective ensemble based on expected probabilities for bankruptcy prediction', *Expert Systems with Applications, 36: 5297-303.*

41. Jensen, Herbert L. *1992.* 'Using neural networks for credit scoring', *Managerial finance, 18: 15-26.*

42. Jung, Ki Mun, Lyn C Thomas, and Mee Chi So. *2015.* 'When to rebuild or when to adjust scorecards', *Journal of the Operational Research Society, 66: 1656-68.*

43. Khemakhem, S., F. Ben Said, and Y. Boujelbene. *2018.* 'Credit risk assessment for unbalanced datasets based on data mining, artificial neural network and support vector machines', *Journal of Modelling in Management, 13: 932-51.*

44. Lacher, R Christopher, Pamela K Coats, Shanker C Sharma, and L Franklin Fant. *1995.* 'A neural network for classifying the financial health of a firm', *European Journal of Operational Research, 85: 53-65.*

45. Lahsasna, Adel, Raja N Ainon, and Teh Y Wah. *2010.* 'Enhancement of transparency and accuracy of credit scoring models through genetic fuzzy classifier', *Maejo International Journal of Science Technology, 4: 136-58.*

46. Lan, Yu, Davy Janssens, Guoqing Chen, and Geert Wets. *2006.* 'Improving associative classification by incorporating novel interestingness measures', *Expert Systems with Applications, 31: 184-92.*

47. Lee, Tian-Shyug, and I-Fei Chen. *2005.* 'A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines', *Expert Systems with Applications, 28: 743-52.*

48. Lee, Tian-Shyug, Chih-Chou Chiu, Yu-Chao Chou, and Chi-Jie Lu. *2006.* 'Mining the customer credit using classification and regression tree and multivariate adaptive regression splines', *Computational Statistics Data Analysis, 50: 1113-30.*

49. Lee, Tian-Shyug, Chih-Chou Chiu, Chi-Jie Lu, and I-Fei Chen. *2002.* 'Credit scoring using the hybrid neural discriminant technique', *Expert Systems with Applications, 23: 245-54.*

50. Lewis, Edward M. *1992. An introduction to credit scoring (*Fair, Isaac and Company*).*

51. Li, Wei, Shuai Ding, Yi Chen, and Shanlin Yang. *2018. "*A Transfer Learning Approach for Credit Scoring*."* In *International Conference on Applications and Techniques in Cyber Security and Intelligence, 64-73.* Springer.

52. Lin, Shih-Wei, Kuo-Ching Ying, Shih-Chieh Chen, and Zne-Jung Lee. *2008.* 'Particle swarm optimization for parameter determination and feature selection of support vector machines', *Expert Systems with Applications, 35: 1817-24.*

53. Liu, Xiaoyong, Hui Fu, and Weiwei Lin. *2010.* 'A modified support vector machine model for credit scoring', *International Journal of Computational Intelligence Systems, 3: 797-804.*

54. Luo, Shu-Ting, Bor-Wen Cheng, and Chun-Hung Hsieh. *2009.* 'Prediction model building with clustering-launched classification and support vector machines in credit scoring', *Expert Systems with Applications, 36: 7562-66.*

55. Malhotra, Rashmi, and Davinder K Malhotra. *2003.* 'Evaluating consumer loans using neural networks', *Omega, 31: 83-96.*

56. Mapoka, Kenneth, Howard Masebu, and Tranos Zuva. *2013.* 'Mathematical models and algorithms challenges', *International Journal of Control Theory Computer Modelling, 3: 21-28.*

57. Marcano-Cedeno, Alexis, A Marin-De-La-Barcena, Juan Jiménez-Trillo, JA Pinuela, and Diego Andina. *2011.* 'Artificial metaplasticity neural network applied to credit scoring', *International journal of neural systems, 21: 311-17.*

58. Marqués, A. I., V. García, and J. S. Sánchez. *2012.* 'Exploring the behaviour of base classifiers in credit scoring ensembles', *Expert Systems with Applications, 39: 10244-50.*

59. Marques, AI, Vicente García, and José Salvador Sánchez. *2013.* 'A literature review on the application of evolutionary computing to credit scoring', *Journal of the Operational Research Society, 64: 1384-99.*

60. Martens, David, Bart Baesens, Tony Van Gestel, and Jan Vanthienen. *2007.* 'Comprehensible credit scoring models using rule extraction from support vector machines', *European Journal of Operational Research, 183: 1466-76.*

61. Mester, Loretta J. *1997*a. 'Measuring efficiency at US banks*:* Accounting for heterogeneity is important', *European Journal of Operational Research, 98: 230-42.*

62. ———. *1997*b. 'What*'*s the point of credit scoring?', *Business review, 3: 3-16.*

63. Nanni, Loris, and Alessandra Lumini. *2009.* 'An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring', *Expert Systems with Applications, 36: 3028-33.*

64. Nath, Ravinder, Balaji Rajagopalan, and Randy Ryker. *1997.* 'Determining the saliency of input variables in neural network classifiers', *Computers Operations Research, 24: 767-73.*

65. Ong, Chorng-Shyong, Jih-Jeng Huang, and Gwo-Hshiung Tzeng. *2005.* 'Building credit scoring models using genetic programming', *Expert Systems with Applications, 29: 41-47.*

66. Oreski, Stjepan, Dijana Oreski, and Goran Oreski. *2012.* 'Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment', *Expert Systems with Applications, 39: 12605-17.*

67. Pacelli, Vincenzo, and Michele Azzollini. *2011.* 'An artificial neural network approach for credit risk management', *Journal of Intelligent Learning Systems Applications, 3: 103.*

68. Papouskova, Monika, and Petr Hajek. *2019.* 'Two-stage consumer credit risk modelling using heterogeneous ensemble learning', *Decision support systems, 118: 33-45.*

69. Peng, Yi, Gang Kou, Yong Shi, and Zhengxin Chen. *2008.* 'A multi-criteria convex quadratic programming model for credit data analysis', *Decision support systems, 44: 1016-30.*

70. Ping, Yao, and Lu Yongheng. *2011.* 'Neighborhood rough set and SVM based hybrid credit scoring classifier', *Expert Systems with Applications, 38: 11300-04.*

71. Piramuthu, Selwyn. *1999*a. 'Financial credit-risk evaluation with neural and neurofuzzy systems', *European Journal of Operational Research, 112: 310-21.*

72. ———. *1999*b. *"*The Hausdorff distance measure for feature selection in learning applications*."* In *Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences. 1999. HICSS-32. Abstracts and CD-ROM of Full Papers, 6* pp*.:* IEEE.

73. Qi, Zhiquan, Bo Wang, Yingjie Tian, and Peng Zhang. *2016.* 'When ensemble learning meets deep learning*:* a new deep support vector machine for classification', *Knowledge-Based Systems, 107: 54-60.*

74. Schumpeter, Joseph A. *1934.* 'Change and the Entrepreneur', *Essays of JA Schumpeter.*

75. Shi, Jian, and Benlian Xu. *2016.* 'Credit scoring by fuzzy support vector machines with a novel membership function', *Journal of Risk Financial Management, 9: 13.*

76. Siddiqi, N. *2012. Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring (*Wiley*).*

77. Silva, Fábio, and Cesar Analide. *2011.* 'Information asset analysis*:* credit scoring and credit suggestion', *International Journal of Electronic Business, 9: 203-18.*

78. Somol, Petr, Bart Baesens, Pavel Pudil, and Jan Vanthienen. *2005.* 'Filter-versus wrapper-based feature selection for credit scoring', *International Journal of Intelligent Systems, 20: 985-99.*

79. Thomas, LC, DB Edelman, and JN Crook. *2002. "*Credit Scoring and Its Applications. Philadelphia*:* SIAM Monographs on Mathematical Modeling and Computation, *248* p*."* In*.:* ISBN *978-0-898714-83-8.*

80. Tsai, Chih-Fong. *2008.* 'Financial decision support using neural networks and support vector machines', *25: 380-93.*

81. ———. *2009.* 'Feature selection in bankruptcy prediction', *Knowledge-Based Systems, 22: 120-27.*

82. Tsai, Chih-Fong, and Chihli Hung. *2014.* 'Modeling credit scoring using neural network ensembles', *Kybernetes, 43: 1114-23.*

83. Tsai, Chih-Fong, and Jhen-Wei Wu. *2008.* 'Using neural network ensembles for bankruptcy prediction and credit scoring', *Expert Systems with Applications, 34: 2639-49.*

84. Tsai, Chih Fong, and Yu Feng Hsu. *2013.* 'A meta-learning framework for bankruptcy prediction', *Journal of Forecasting, 32: 167-79.*

85. Van Sang, Ha, Nguyen Ha Nam, and Nguyen Duc Nhan. *2016.* 'A novel credit scoring prediction model based on Feature Selection approach and parallel random forest', *Indian Journal of Science Technology, 9: 1-6.*

86. Varga, Gyorgy. *2009.* 'Investment decision in a new credit score system', *FCE Consulting: Brazil.*

87. Vojtek, Martin, and Evžen Koèenda. *2006.* 'Credit-scoring methods', *Czech Journal of Economics Finance, 56: 152-67.*

88. Wang, Gang, Jinxing Hao, Jian Ma, and Hongbing Jiang. *2011.* 'A comparative assessment of ensemble learning for credit scoring', *Expert Systems with Applications, 38: 223-30.*

89. Wang, Gang, Jian Ma, Lihua Huang, and Kaiquan Xu. *2012.* 'Two credit scoring models based on dual strategy ensemble trees', *Knowledge-Based Systems, 26: 61-68.*

90. West, David. *2000.* 'Neural network credit scoring models', *Computers & Operations Research, 27: 1131-52.*

91. Xiao, Guorong. *2011.* 'Data Processing Model of Bank Credit Evaluation System', *JSW, 6: 1241-47.*

92. Xiao, Hongshan, Zhi Xiao, and Yu Wang. *2016.* 'Ensemble classification based on supervised clustering for credit scoring', *Applied Soft Computing, 43: 73-86.*

93. Yang, Yingxu. *2007.* 'Adaptive credit scoring with kernel learning methods', *European Journal of Operational Research, 183: 1521-36.*

94. Yu, Lean, Shouyang Wang, and Kin Keung Lai. *2008.* 'Credit risk assessment with a multistage neural network ensemble learning approach', *Expert Systems with Applications, 34: 1434-44.*

95. Zeng, Zhi-Qiang, and Ji Gao. *2009. "*Improving SVM classification with imbalance data set.*"* In *International Conference on Neural Information Processing, 389-98.* Springer.

96. Zhang, Defu, Xiyue Zhou, Stephen CH Leung, and Jiemin Zheng. *2010.* 'Vertical bagging decision trees model for credit scoring', *Expert Systems with Applications, 37: 7838-43.*