STATISTICAL AND DATA MINING METHODS IN CREDIT SCORING

Alireza Hooman¹, Mohana Omidi², G. Marthandan¹, Wan Fadzilah Wan Yusoff¹, Sasan Karamizadeh³

¹: Multimedia University, Malaysia
²: University of Social Welfare and Rehabilitation Science, Iran
³: Universiti Teknologi Malaysia

ABSTRACT

The growing interest in the credit industry resulted in credit scoring being developed as an essential component, especially in the credit department of banks dealing with huge sums of credit data. When a bank or a credit corporation is assessing a credit application request, they will have to decide whether to approve or deny the request. This necessitates the utilization of credit scoring. Although pioneers attempt to cover risks via interest rate, current investigation on financial conditions of different sections of society revealed that interest could not replace risk assessment; which means that credit risk needs its own specialized assessment. With the assistance of sorting methods, credit scoring simplifies the decision-making process. It is almost impossible to analyze this large amount of data in the context of manpower and economy, although the data mining technique helps alleviate this complexity. Nowadays, there are a lot of data mining methodologies being utilized to manage credit scoring. However, each method has its advantages and limitations, and there has not been a comprehensive approach in determining the most utilized data mining technique in the context of credit scoring. The major goal of this paper is to provide a complete literature survey on applied data mining methods, such as discriminant analysis, logistic regression, K-nearest neighbor, Bayesian classifier, decision tree, neural network, survival analysis, fuzzy rule-based system, support vector machine, and hybrid methods. These findings will assist researchers in realizing the most suitable approach in evaluating credit scores, pinpoint limitations, enhance them, and propose new approaches with improved capabilities. Finally, the limitations of the new approaches are discussed, and further suitable methods are recommended.

JEL Classifications: G240

Keywords: Credit Scoring, Data Mining, Feature selection, Classification

Corresponding Author’s Email Address: alireza_hooman@yahoo.com

INTRODUCTION

Competition in the credit industry is driven by the increasing demand for consumer credit. Subsequently, credit managers are required to design and develop mechanisms that will allow them to analyze credit data, which would help minimize error and processing times. Both the increasing quantity of non-performing loans and the financial crisis render credit scoring highly salient (Akkoc, 2012). Credit scoring is defined as a technique that
assist those granting credits in deciding whether and applicant should or should not be granted credit, based on characteristics such as income, marital status, and age (M. C. Chen & Huang, 2003). According to (Paleologo, Elisseeff, & Antonini, 2010), there are several types of scoring methods: applications, behavioral, collection scoring, and fraud detection. Ravi (2008) developed three primary applications for credit scoring, which are credit card applications mortgage and small business lending applications (Ravi, 2008).

Recently, credit risk evaluation has been conducted using several proposed quantitative methods. Data mining methods remain popular, due to it being able to determine practical knowledge from a given database and transforming it into usable and valuable information. In the event a company or an individual fail in meeting loan payments, it could prove to be very costly for both the stockholders and the banks involved. This is due to the fact that the banks will lose what they have already extended to the applicant, while stockholders stand to lose their equities. The problem has been always the issue of prediction prior to or after the fact of an event. As shown by experimental researches, certain classifications are capable of providing some hints. However, these could result in multiple values and conclusions, mostly from size, form, and informational organization. Therefore, selecting the appropriate strategy is more imperative. This research intends to survey the literature on data mining systems that were or are used to evaluate credit risk from the 2000 - 2015. This paper will increase the awareness of researcher’s vis-à-vis the current existing methods, its major goals, and respective limitations, which will lead to the proposal of methods that are more efficient.

**RESEARCH METHODOLOGY**

Journal databases were researched, namely Scopus, Science Direct, Springer, Emerald, J store, IEEE Explore, Academic Search Premier, and John Wiley. Other journals researched include European Journal of Operational Research, Expert Systems with Applications, Computer Engineering and Applications, Journal of Banking & Finance, Computers & Operations Research, Journal of Empirical Finance, Mathematics and Economics, Journal of Business Economics and Management, and international IEEE conference papers. Some textbooks on data mining and credit were used as well. The keywords include Credit Scoring, Credit Rating, Credit Risk, Data Mining, Neural Network, Classification, Bayesian Classifier, Decision Tree, Logistic Regression, Discriminant Analysis, K-Nearest Neighbor, Fuzzy Rule-Based System, Support Vector Machine, and Hybrid Model and Survival Analysis. The survey was conducted in two stages, where the first stage included the selection of journal papers associated with credit scoring, which was done by reviewing titles and abstracts. At this stage, about 200 journal papers were identified, containing descriptions related to credit scoring and data mining. This information was then extracted from journals and books for 2000 - 2015.

**LITERATURE SURVEY**

**Data Mining**

During periods of data blast, specific corporations will provide and gather enormous capacity of daily data. Determining valuable information from the databank and converting data into actionable outcomes is the most glaring difficulty encountered by corporations. Data mining is the progression of investigation and examination, by automatic or semi-automatic methods, of great amounts of records in order to determine significant configurations and guidelines having meanings (Kirkos, Spathis, & Manolopoulos, 2007). Data mining are
either supervised or unsupervised, depending on the information the analyst wishes to use, or is available on samples constituting the data matrix. Nowadays, data mining is an indispensable tool in a decision support system, and plays a key role in market segmentation, customer services, fraud detection, credit and behavior scoring, and benchmarking (Giudici, 2001; Thomas, 2000). Figure 1 shows the classification of data mining methods and the application of sub-categories of data mining in evaluating credit scoring based on literature.

Some researchers (Ince & Aktan, 2009) attempted to compare some of the data mining methods and answer questions such as:

1. Is there any difference of classification accuracy among the data mining techniques?
2. Could the estimated probability of default produced from data mining methods represent the real probability of default?

**FIGURE 1: MAJOR STATISTICAL AND DATA MINING METHODS USING IN CREDIT SCORING**

Discriminant Analysis

Fisher, in early 1936, studied the concept of discriminant analysis as a substitute to logistic regression, with the assumption that the explanatory variables track the multivariate normal distribution, and has a matrix with a common variance-covariance. This technique is utilized to classify the observations into two different classes (Fisher, 1936). The key disadvantage of this technique is the assumptions, which borders on the unrealistic. Nevertheless, this technique has very robust assumptions on the variables, so the forecast power is rather weak. Furthermore, this technique does not provide awareness vis-à-vis the relative acts of the variables. Since 1974, discriminant analysis has been utilized in credit scoring. Post-literature review, the arguments that are in favor
of linear discriminant analysis (LDA) are: (a) Most efficient technique for credit scoring purposes; (b) Easy to implement and to interpret; (c) Most efficient technique when apply to large sample. However, LDA is not without its shortcomings, and some are listed as: (a) Need for statistical assumptions; (b) Need for ordered categorical variables; (c) Outliers sensitivity (Doumpos & Zopounidis, 2014a; Ince & Aktan, 2009; Kambal, Osman, Taha, Mohammed, & Mohammed, 2013; H. Li, Sun, Li, & Yan, 2012).

Logistic Regression

Logistic regression is described as another type of regression that is linear. This model is able to forecast the distinct result of a group of variables that could be dichotomous, discrete, continuous, or a combination of either (Joanes, 1993). Overall, the dependent variable is normally dichotomous. The benefits of this method are that the logistic regression does not assume that the association between the dependent and the independent variables are linear; it does not need a set of variables that are normally distributed. Logistic regression has been extensively used in credit scoring applications (Bensic, Sarlija, & Zekic-Susac, 2005; Nguyen, 2015). Also, logistic regression is regarded to be the most accurate model between conventional models (Shi, Wang, Qi, & Cheng, 2015). This method does not assume that the link between the dependent and independent variables are linear, and it does not require variables to be distributed in a linear manner. The major benefit of this method is that it can generate a simple probabilistic formula for classification. The disadvantages are that LR is not able to deal with the problems of non-linear and cooperative effects of explanatory variables properly.

Probit Regression

Probit regression is the next conventional technique exploited in credit scoring. Tsaih et al (2004) used Probit regression to develop a credit scoring model. They suggested N-tier architecture, integrated with the Model View Controller, which would model scoring to obey fluctuations in businesses. One of the advantages of the design is that less time is needed for communicating changes to scoring models. Moreover, the models can be tweaked at the pleasure of the managers. Wallace (1978; 1981) applied regression and multivariate Probit models to predict bond ratings for 106 new general obligation, and revenue bond issues in the state of Florida. These are some of the advantages of Probit regression: (a) Less time consuming to implement; (b) Easy to alter the models at any time. (Abdou & Pointon, 2011; Giudici, 2005; Tsaih, Liu, Liu, & Lien, 2004)

K-nearest Neighbor

K-nearest neighbor is described as a classifier that is nonparametric, which is learned via similarity. For every new observation, this model examines the pattern distance for the K nearest neighbors that are nearest to new observations, based on the space between the explanatory variables (Karamizadeh, Abdullah, Manaf, Zamani, & Hooman, 2013). The major advantage of this approach is that it is not required to establish predictive model prior to classification, while the disadvantages are the fact that KNN does not produce a simple classification probability formula, and its predictive accuracy is highly affected by the measure of distance and the cardinality k of the neighborhood. k-NN is favorable, due to: (a) Enables the modelling of irregularities in the risk function over the feature space; (b) K-NN method is superior to other nonparametric techniques, such as kernel methods, especially in cases where the data are multidimensional; (c) It is fairly intuitive and easily understood by managers, who are responsible for its implementation. The weaknesses are: (a) KNN fails in generating a
simplified classification probability formula; (b) Its predicted accuracy is affected by the distance and the cardinality of k within the neighborhood (Brown & Mues, 2012; H. Chen & Chen, 2010; Lahsasna, Ainon, & Teh, 2010).

Multivariate Adaptive Regression Splines (MARS)

Multivariate adaptive regression splines (MARS) are nonlinear parametric regression that forms additive relationships or involve interactions with smaller number of variables. In the form of rough sets, neural networks and other recent techniques, multivariate adaptive splines, and the so-called MARS techniques have piqued the interest of researchers. Lee & Chen (2005) highlighted these particular advantages: (a) nonrequirement of pre-assumptions, so that it can model complex non-linear relations amongst variables without strong modelling assumptions; (b) it automatically chooses vital variables, hence, it captures the relative significance of independent variables for the dependent ones once many potential independent variables are taken into account; (c) there is no long training process, and (d) easy to interpret. (Chuang & Lin, 2009; Li-hua, Jia-shan, & Feng, 2006)

Decision Tree

A Decision tree is described as a tree-like decision graph and its possible outcomes. The highest node in this tree is the root node, where a decision is made (presumably). In every node inside, a test is carried out on the input or attribute variable(s). Every branch that tracks the node goes to the outcome of the test. (Baesens et al., 2003) Utilized decision trees to plan the credit scoring model. Many other scholars have utilized decision trees in credit scoring, and explained the process in detail. A review of the literature shows the arguments in favor of Recursive Partitioning to be: (a) logical relationship, (b) easy to interpret, (c) efficiency, (d) flexibility and (e) high dimensionality issue can be avoided. In a review of the advantages of this method, the following stood out: (a) global complex decision regions can be approximated by the union of simpler local decision regions at various levels of the tree; (b) a sample is tested against certain subsets of classes, eliminating unnecessary computations; and (c) flexibility of choosing different subsets of features at different internal nodes of the tree such that the feature subset chosen optimally discriminates among the classes in that node (Bastos, 2007; Paleologo et al., 2010).

Survival Analysis

The survival analysis is a newly developed model for credit scoring. The previous models were able to distinguish the good and bad borrowers at the point of application, but this model is able to input the profitability of a borrower based on the customers’ lifetime and profit scoring performance (Baesens, Van Gestel, Stepanova, Van den Poel, & Vanthienen, 2005). SA is able to forecast time up till the event’s occurrence, unlike others that predict the probability of the occurrence. (Stepanova & Thomas, 2002) proposed that the survival analysis techniques are capable of regressing time design to nonpayment and keep it informed of time movements. This technique is comprehensive, as it takes into account the overall attitude towards credit scoring, as pointed out by (Doumpos & Zopounidis, 2014b). There are also other studies that implemented survival analyses methods and techniques, which are lately used for credit scoring (Beran & Djâ“dja, 2007; Sohn & Shin, 2006).
Fuzzy Rule-Based System

Fuzzy rule-based system help creditors develop rules that result in accurate credit scores. Most other models for credit scoring tend to concentrate on predicting the score without any explanation on the obtained outcome. The advantage of this model is that the fuzzy rules are able to handle both qualitative and quantitative factors, resulting in a huge amount of input. The scoring results will not be influenced by minute errors in measurement; the major implementation part of fuzzy rule-based methods has always been control problems (Hoffmann, Baesens, Mues, Van Gestel, & Vanthienen, 2007). There are many studies done on the influence of rule weights in fuzzy rule-based classification methods for credit scoring (Akkoc, 2012; Capotorti & Barbanera, 2012; Sohn & Shin, 2006; Yu, Wang, & Lai, 2009).

Neural Network

Artificial neural networks (ANN) can be described as a non-linear statistical model built to replicate human brain functions. They are compelling tools for undetermined data relationship modeling. Since the 1990s, ANNs have been extensively utilized in monetary forecast researches. Most of these papers report that forecast correctness of ANNs is superior to conventional statistical methods. Despite the fact that ANN is capable of being utilized effectively in numerous fields, it is not without shortcomings. ANN necessitates a time-consuming training course in developing the optimal model (Karamizadeh, Abdullah, Zamani, & Kherikhah, 2015). ANN has also been critiqued for the shortage of theory. There is no chance to illuminate the outcome created by ANN by other means, as the model is viewed as being a black box. NN are favorable due to: (a) Memory, ability to generalize, robustness, absence of any explicit problem description; (b) Ability to handle large amount of data; (c) Need for less statistical assumptions; (d) Nonparametric and nonlinear method. Some of the drawbacks associated with NNs are: (a) Issue when apply to small samples; (b) May incorporate irrelevant attribute; (c) Long training time, selection time; (d) Over fitting when apply to large dataset; (e) Hard to interpret and issue of trial and error process in some studies.

Support Vector Machine

Support vector machine is described as a technique of classification that was first introduced by. The key benefits of this model are based on the nonparametric case; SVM needs no data structure assumptions, unlike continuity and normal distribution. SVM is a novel method in the ground of data mining, which can be used to solve machine learning via optimization approaches, and is also a machine-learning procedure that mainly focuses on statistical learning principle, as per. Standard support vector machines have been utilized by to investigate the many units of data of major U.S. credit card banks. Support vector machine is developed upon the statistical learning concept, constructed from VC element and operational risk minimization, which are mainly on the basis of restricted illustration data, and determine the equilibrium among complication of the concept and learning capability to gain the greatest generalization capability. The benefit of this technique is that in the nonparametric case, SVM needs no data structure assumptions, for example, normal distribution or
continuity (Karamizadeh, Abdullah, Halimi, Shayan, & Rajabi, 2014). The disadvantage of this technique is that, it is hard to interpret the features and standard formulations that do not have specification of business constraints. (Harris, 2015)

Hybrid Models

The hybrid models are described as credit scoring models that have been designed by combining two or more pre-existing models. The creditor can then profit from having two or more models, and minimize weaknesses by mixing them with other models; however, these methods are too complicated for devise and execution. Some efficient credit scoring prototypes of hybrid methods have also been planned recently (Harris, 2015; Karamizadeh, Abdullah, & Zamani, 2013; Oreski, 2014). Certain scholars proposed that more accurate explanations are possible via hybrid techniques (I.-F. Chen, 2013; Harikrishna & Farquad, 2012; Hsieh, 2005; F. C. Li, 2009; Min, Lee, & Han, 2006; Oreski, 2014; Oreski, Oreski, & Oreski, 2012; Ping & Yongheng, 2011; Wang & Ma, 2012). Table 2.1 illustrate the combination of methods used in hybrid credit scoring in past researches, and specially apply feature selection as a part of some of these methods.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Feature selection</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANN</td>
<td>MARS</td>
<td>(Lee &amp; Chen, 2005)</td>
</tr>
<tr>
<td></td>
<td>LR &amp; GA</td>
<td>(Lin, 2009)</td>
</tr>
<tr>
<td>ANN &amp; DT</td>
<td>PCA &amp; LR</td>
<td>(Šušteršič, Mramor, &amp; Zupan, 2009)</td>
</tr>
<tr>
<td></td>
<td>GA</td>
<td>(Oreski et al., 2012)</td>
</tr>
<tr>
<td>SVM</td>
<td>SVM</td>
<td>(Bellotti &amp; Crook, 2009)</td>
</tr>
<tr>
<td>SVM</td>
<td>GA</td>
<td>(D. Zhang, Hifi, Chen, &amp; Ye, 2008)</td>
</tr>
<tr>
<td>VBDTM</td>
<td>Rough Set</td>
<td>(Defu Zhang, Zhou, Leung, &amp; Zheng, 2010)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Feature selection</th>
<th>Model parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANN</td>
<td>LDA</td>
<td>(Lee, Chiu, Lu, &amp; Chen, 2002)</td>
</tr>
<tr>
<td>MARS</td>
<td>ANN</td>
<td>(Lee &amp; Chen, 2005)</td>
</tr>
<tr>
<td>ANN,</td>
<td>MARS</td>
<td>(Chuang &amp; Lin, 2009)</td>
</tr>
</tbody>
</table>
COMPARISON OF DATA MINING METHODS IN CREDIT SCORING

During this research, the author did not come across any proof that one method is undisputedly superior over another. Some comparisons are shown from multiple perspectives. Figure 2 show the comparison of data mining methods in the context of publication frequency.

FIGURE 2: COMPARISON OF DATA MINING METHODS BY PUBLICATION.

<table>
<thead>
<tr>
<th>Data Mining Methods</th>
<th>Advantage</th>
<th>Disadvantage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Efficiency / accuracy/ robustness</td>
<td>Long training</td>
</tr>
<tr>
<td>Discriminant Analysis, Logistic Regression, Probit Regression, Neural Network, K.N.N, SVM, Decision Tree</td>
<td>Easy to implement</td>
<td>Black box property</td>
</tr>
<tr>
<td>Discriminant Analysis, Logistic Regression, Probit Regression</td>
<td>Easy to interpret</td>
<td>Genetic Algorithm, Fuzzy based model, Survival Analysis, MARS</td>
</tr>
<tr>
<td>Discriminant Analysis, Logistic Regression, Probit Regression, K.N.N, Decision Tree, MARS</td>
<td>Easy to alter models at any time</td>
<td>Neural Network, SVM</td>
</tr>
</tbody>
</table>

CONCLUSION

Credit scoring has become an essential component, given the growing interest in the credit industry, especially in the credit departments of banks that have to deal with a huge sum of credit data. It is almost impossible to
analyze this amount of data in terms of manpower and economics levels. This study reviewed the journals and articles that had utilized data mining techniques in evaluating the challenges of credit risk. Ten data mining techniques that were most commonly utilized in the evaluation of credit risk were examined from 2000 – 2015. Recently, it has been found that the support vector machine is the most widely used system. In order to improve its performance, the features associated with its subset needs to be determined or designed, and as such, many hybrid SVM type models have been recommended. Additionally, the hybrid models have been used in the past decade utilized the benefits associated with both its base models. Most of the recommended models are only able to classify customers or borrowers into two categories of “good” or “bad.” From the perspective of risk management, the prediction of the default probability of the credit applicant will be better than using the binary classification. Therefore, the models that are capable of estimating the default probability and are easy to understand and interpret have been proposed.

REFERENCES


