ARTICLE IN PRESS

Journal of Business Research xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Journal of Business Research

journal homepage: www.elsevier.com/locate/jbusres

Big Data techniques to measure credit banking risk in home equity loans☆

A. Pérez-Martín*, A. Pérez-Torregrosa, M. Vaca

Department of Economic and Financial Studies, Miguel Hernández University of Elche, Spain

ARTICLE INFO

Keywords:
Credit scoring
Big Data
Monte Carlo
Data mining

ABSTRACT

Nowadays, the volume of databases that financial companies manage is so great that it has become necessary to address this problem, and the solution to this can be found in Big Data techniques applied to massive financial datasets for segmenting risk groups. In this paper, the presence of large datasets is approached through the development of some Monte Carlo experiments using known techniques and algorithms. In addition, a linear mixed model (LMM) has been implemented as a new incremental contribution to calculate the credit risk of financial companies. These computational experiments are developed with several combinations of dataset sizes and forms to cover a wide variety of cases. Results reveal that large datasets need Big Data techniques and algorithms that yield faster and unbiased estimators. Big Data can help to extract the value of data and thus better decisions can be made without the runtime component. Through these techniques, there would be less risk for financial companies when predicting which clients will be successful in their payments. Consequently, more people could have access to credit loans.

1. Introduction

Any credit rating system that enables the automatic assessment of the risk associated to a banking operation is called credit scoring. This risk may depend on several customer and credit characteristics, such as solvency, type of credit, maturity, loan amount, and other features inherent in financial operations. It is an objective system for approving credit that does not depend on the analyst's discretion.

In the 1960s, coinciding with the massive demand for credit cards, financial companies began applying credit scoring techniques as a means of assessing their exposure to risk insolvency (Altman, 1998). At the same time, the United States also began to develop and apply credit scoring techniques to assess credit risk assessment and to estimate the probability of default (Escalona Cortés, 2011).

Since 1970, credit scoring models had been based on statistical techniques, and particularly on discriminant analysis, which was generalized in 1990 (Gutierrez, 2007). However, with the development of better statistical resources and new advances in technology, it became necessary for financial institutions to carry out their risk assessments more effectively and efficiently.

Since the 1980s, due to the increase in credit demand and computational progress, credit scoring techniques have been extended to loans. In this context, some financial companies started to use different statistical techniques to optimize the differentiation between good and bad loans (Durand, 1941; Reichert, Cho, & Wagner, 1983).

In 2004, the recommendations of the Basel Committee (known as Basel II) on banking supervision appeared. Since then, the use of advanced methods of credit scoring have become a regulatory requirement for banks and financial institutions to improve the efficiency of capital allocation. In response to the global financial crisis, a new document (Basel III) appeared. This document introduced more changes and demands for financial companies regarding the control of borrowed capital and the ensuing increase in reserves based on their risk. As a result, improving the accuracy of credit risk evaluations has become a potential benefit to financial institutions.

In recent decades, although several different investigations have compared different methods for measuring risk, scientific literature has not solved the problem efficiently. As well as this, there has been a rise in financial operations, which has led to an increase in the volume of databases. All the methods analyzed in the scientific literature are suitable for the classification of good or bad credit, but all of them have advantages and disadvantages.

Nowadays, the volume of databases that financial companies manage is so great that it has become necessary to address this problem, and the solution to this can be found in Big Data techniques applied to massive financial datasets for segmenting risk groups. The presence of large datasets is approached through the development of some Monte Carlo experiments using known techniques and algorithms. These computational experiments reveal that large datasets need Big Data techniques and algorithms that yield faster and unbiased estimators.

☆ The authors acknowledge financial support provided by Conselleria de Educación, Generalitat Valenciana under grant GVA/2016/053.
* Corresponding author.
E-mail addresses: agustin.perez@umh.es (A. Pérez-Martín), agustin.perez01@goumh.umh.es (A. Pérez-Torregrosa), m.vaca@umh.es (M. Vaca).

https://doi.org/10.1016/j.jbusres.2018.02.008
Received 20 June 2017; Received in revised form 1 February 2018; Accepted 5 February 2018

They are developed with several combinations of dataset sizes and forms to cover a wide variety of cases.

Big Data can help to extract the value of data and thus better decisions can be made. Then, the high costs of the runtime, which would make the problem intractable, can be avoided. Through these techniques, financial companies would have less risk when predicting which clients will be successful in their payments. Consequently, more people could have access to credit loans.

This research area is very important for financial institutions because credit risk is 60% of a company's total risk. With the introduction of the 9th International Financial Reporting Standards (IFRS 9) in January 2018, financial companies will have to calculate expected losses from defaults over the 12 months following these financial instruments. In this case, a good method for estimating risk would mean lower expected losses and a higher profit for the company and maybe more credit loans.

There are various methods available for assessing credit risk. These range from a personalized study by an expert in risk analysis to different statistical and econometric methods of credit scoring. Nowadays, in a first step, it is not feasible to apply specific analyses to the study of home equity loans. Credit scoring methods are more efficient, objective, and consistent in their predictions, since they can be used to analyze and make quick and inexpensive decisions about many credit applications.

In line with some authors, credit scoring can be considered as a way to identify different groups within a population. One of the first proposals to solve this problem was introduced in statistics by Fisher (1936) using discriminant analysis and a multivariate statistical technique. Durand (1941) was the first to recognize that the same statistical techniques can be used to optimize the differentiation between good and bad loans.

The use of credit scoring models is not only the result of the generalization of credit, but also the result of the banking regulations and supervision introduced in the past three decades. Financial and credit institutions are subject to what is known as "prudential policy", which means the amount of equity must be maintained to ensure a smooth operation and to cover several risks which may arise, including credit risk (Trias, Carrascosa, Fernández, París, & Nebot, 2005).

Between the late 20th century and early 21st century, due to economic growth, consumer credit increased spectacularly. The need for financial institutions to increase their market share has become a reality today; the larger the volume of credit granted by a company, the greater its potential profits. However, this should be linked to an increase in quality, because otherwise the end result would be a significant deterioration in the income statement. Consequently, statistical methods for assessing credit risk have become increasingly important (Hand & Henley, 1997).

Since Basel II, the use of advanced methods of credit scoring has become a regulatory requirement for banks and financial institutions in order to improve the efficiency of capital allocation. Nevertheless, Basel III introduced stricter changes for controlling borrowed capital, where an increase in reserves occurs in financial institutions based on their risk. Improving the accuracy of credit risk evaluation is a potential benefit to financial institutions, even if it is only slight. Over the past decades, there have been different methods for measuring risk.

Nowadays, credit scoring models are based on mathematics, econometric techniques, and artificial intelligence (Ochoa, Galeano, & Agudelo, 2010; Canton, Rubio, & Blasco, 2010). Empirical studies by various authors present alternative approaches and compare different techniques and algorithms with the problem that they present different conclusions. These approaches include the following: decision trees used by Srinivasan and Kim (1987), Hand and Henley (1997), Galindo and Tamayo (2000), Huang, Hung, and Jiau (2006) or Lee, Chiu, Chou, and Lu (2006); the logistic regression technique described by Thomas (2000), Boj, Claramunt, Esteve, and Fortiana (2009b), and Alaraj and Abbod (2016); discriminant analysis used by Altman (1998), Yobas, Crook, and Ross (2000), and Boj, Claramunt, Esteve, and Fortiana (2009a); or the use of support vector machines by authors like Van Gestel, Baesens, Garcia, and Van Dijcke (2003), Liu, Frazier, and Kumar (2007) or Yu (2014). In our research, we try to address the problem by using several datasets with some Monte Carlo experiments.

All the methods analyzed in the scientific literature are suitable for the classification of good or bad credit, but all of them have advantages and disadvantages. The method or algorithm used depends on the structure of the data, the features used, the possibility of separating the classes by using these features, and the purpose of the classification of the data structure (Morales, Pérez-Martín, & Vaca, 2013, Baesens et al., 2003). Baesens et al. (2003) compare sixteen methods for credit risk evaluation based on eight datasets of different sizes and origin. They concluded that the experiments also indicated that many classification techniques yield performances which are quite competitive with each other. Only a few classification techniques were clearly inferior to the others, but they did not mention their computational efficiency. Yu, Yao, Wang, and Lai (2011) compare different methods for credit risk evaluation with two datasets (German and Australian UCI credit datasets). They concluded that weighted least squares support vector machine (LSSVM) classifier is the best for credit risk evaluation. They also indicated that the credit industry requires quick decisions, and in this sense, there should be a trade-off between computational performance and computational efficiency. In our paper, we carried out a computational measure as an incremental contribution.

Pérez-Martín and Vaca (2017) analyze quadratic discriminant analysis (QDA) and support vector machine with linear kernel (LSVM). With respect to effectiveness, LSVM is found to be the best method for estimating credit risk, but in terms of computational efficiency, LSVM takes longer than QDA to solve the same problem. For large datasets (5000 records) and a large number of explanatory variables, LSVM has better success rates. When the number of explanatory variables are equal or less than 10, differences are unnoticeable.

Scientific literature therefore has not solved the problem efficiently. As well as this, the increase in financial operations has led to an increase in the volume of databases. As a result, financial companies are faced with the problem of managing a huge volume of databases and the need to address this situation. Big Data techniques applied to massive financial datasets for segmenting risk groups is the solution. Big Data can help to extract the value of data and thus better decisions can be made. Then, the high costs of the runtime, which would make the problem intractable, can be avoided. Therefore, an automatic evaluation is necessary, using fast and adaptive techniques like machine learning, where the probability of default can be calculated with historical massive datasets in a reasonable period of time.

In this paper, eight methods for solving the problem of credit scoring in home equity loans are proposed. Firstly, measures are made of how a loan can be classified and how cost influences execution time. To evaluate this, different Monte Carlo simulation experiments are performed. Several (72×10^4) random datasets with different sizes are generated so that the result of a method does not depend on the data, and a linear mixed model (LMM) method is proposed as a new and better contribution.

The execution time component may be important when deciding whether to apply one method or another due to the massive volume of data. A computationally efficient method can be much more competitive, since it provides advantages in terms of time expected in resolving requests. The main goal in this study is to present and compare credit scoring methods that are effective and efficient.

In Section 2, the methods used in this research are outlined. In Section 3, simulation experiments are developed and several efficient measures are presented. Finally, in Sections 4 and 5, results, conclusions, and recommendations are given.