



Machine learning-driven credit risk: a systemic review

Si Shi¹ · Rita Tse^{1,2} · Wuman Luo¹ · Stefano D'Addona³ · Giovanni Pau^{1,4} 

Received: 12 February 2020 / Accepted: 12 May 2020 / Published online: 11 July 2020
© The Author(s) 2020

Abstract

Credit risk assessment is at the core of modern economies. Traditionally, it is measured by statistical methods and manual auditing. Recent advances in financial artificial intelligence stemmed from a new wave of machine learning (ML)-driven credit risk models that gained tremendous attention from both industry and academia. In this paper, we systematically review a series of major research contributions (100 papers) over the past eight years using statistical, machine learning and deep learning techniques to address the problems of credit risk. Specifically, we propose a novel classification methodology for ML-driven credit risk algorithms and their performance ranking using public datasets. We further discuss the challenges including data imbalance, dataset inconsistency, model transparency, and inadequate utilization of deep learning models. The results of our review show that: 1) most deep learning models outperform classic machine learning and statistical algorithms in credit risk estimation, and 2) ensemble methods provide higher accuracy compared with single models. Finally, we present summary tables in terms of datasets and proposed models.

Keywords Credit risk · Machine learning · Deep learning · Statistical learning

1 Introduction

Machine learning advances heavily affected industry and academia in the past decades, ultimately transforming people's daily life. Artificial Intelligence (AI) has been applied to almost every human activity, including pattern recognition, image classification, business, agriculture, transportation, and finance. This paper focuses on machine learning applied to finance and credit risk estimation. Modern financial systems rely on credit and trust. Credit risk is a fundamental parameter that measures and predicts the default probabilities of a debtor. The correct estimation of credit risk is paramount for the entire system. Failing in the credit risk estimation can lead to systemic failures such as the sub-prime crisis of 2007–2009. Consequently, lenders devote large amounts of resources to predict the credit-worthiness of consumers and companies to develop appropriate lending strategies that minimize their risks. Historically, credit risk approaches use statistical methods such as Linear Discriminant Analysis [1] and Logistic Regression [2]. These methods, however, do not easily handle large datasets.

Advances in computing power and availability of large credit datasets paved the way to AI-Driven credit risk estimation algorithms such as traditional machine learning and deep learning. Conventional machine learning

✉ Giovanni Pau
giovanni.pau@unibo.it

Si Shi
si.shi@ipm.edu.mo

Rita Tse
ritatse@ipm.edu.mo

Wuman Luo
luowuman@ipm.edu.mo

Stefano D'Addona
daddona@uniroma1.it

¹ Faculty of Applied Sciences, Macao Polytechnic University, Macao SAR, 999078, China

² Engineering Research Centre of Applied Technology on Machine Translation and Artificial Intelligence of Ministry of Education, Macao Polytechnic University, Macao SAR, 999078, China

³ Department of Political Science, University of Roma Tre, 00146, Rome, Italy

⁴ Department of Computer Science and Engineering, University of California, Los Angeles, CA 90095, USA

⁵ UCLA Samueli Computer Science, University of California, Los Angeles, CA 90095, USA

techniques, e.g., k-Nearest Neighbor [11], Random Forest [12] and Support Vector Machines [13], are more effective and flexible than statistical methods. In particular, the vital branch of machine learning-deep learning techniques [14] applied to large credit risk data lake outperform their predecessors both in accuracy and efficiency.

This paper presents a systemic review of credit risk estimation algorithms. It analyzes both the major statistical approaches and AI-based techniques with a critical spirit. The aim is to provide a comprehensive overview of the current leading credit risk estimation technology, providing justification and connections between past and present works. This work proposes a novel taxonomy combining finance with machine learning techniques. In addition, this work ranks their performance in terms of accuracy and costs. This paper also discusses the challenges and possible solutions in terms of four aspects: data imbalance, dataset inconsistency, model transparency, and inadequate utilization of deep learning methods.

The remainder of the paper is organized as follows: the survey methodology will be discussed in Sect. 2. Section 3 introduces the principles of statistical learning, machine learning and deep learning. Section 4 analyzes credit risk-related applications in detail. In Sect. 5, presented algorithms are discussed and ranked by their performance against public datasets. Finally, results and current challenges are summarized in Sect. 6; while Sect. 7 concludes this work.

2 Survey methodology

2.1 Methodology

We applied PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Fig. 1) reviewing methodology in our paper. First, we adopted five searching platforms for our investigation: Google Scholar, ACM, IEEEExplore, Springerlink, and ScienceDirect. We used the keywords “machine learning” or “deep learning” combined with “credit risk” while searching. We got 1000 articles in total. Then, we applied a filtering algorithm considering the trade-off between publication year and citations to proceed. After removing 200 duplicate records, 100 ineligible records, and 50 incomplete articles, we obtained 650 screened records. Based on the relevance, we excluded 100 articles less related to the topic. After manually checking whether the paper has clear evaluation metrics, we further excluded another 50 papers. Finally, we kept 550 studies in terms of the relevancy to the research topic, precision of evaluation metrics, publication time, and number of citations as our source of reviewing.

Figure 1 depicts the PRISMA flow diagram

2.2 Inclusion and exclusion criteria

In this paper, we select three inclusion criteria: (1) the relevance of research topic, (2) the precision of evaluation metrics, (3) the publication year and citations. Moreover, the papers will be excluded if they are duplicated, incomplete, too early, low-related with the topic, having no clear metrics or comparatively low citations.

We show the whole workflow of the selection process in Fig. 1.

2.3 The datasets and approaches of the reviewed articles

The mainly used datasets by the papers under review are German and Australian public credit data from the UCI Machine Learning repository [15–16]. In addition, there exist some researches that discover and mine their own data. For example, Chee Kian Leong [17] uses data from a firm in [18]. Authors in [19–21] all employ their unique dataset. Those articles mainly emphasize the significance and the veracity of the original data.

We discuss the principles and application of the overall machine learning approaches. The traditional machine learning models for credit risk contain Support Vector Machines (SVMs) [22], k-Nearest Neighbor (k-NN) [23], Random Forests (RFs) [24], Decision Trees (DTs) [25–26], AdaBoost [27], Extreme Gradient Boost (XGBoost) [28], Stochastic Gradient Boosting (SGB) [29], Bagging [30], Extreme Learning Machine (ELM) [31] and GA (Genetic Algorithm) [32]. Neural network models generally belong to deep learning methods. Most of them include Convolutional Neural Networks (CNNs) [33], Deep Belief Neural Networks (DBNs) [34], Artificial Neural Networks (ANNs) [35], LSTM (Long Short-Term Memory) [36], Restricted Boltzmann Machines (RBMs) [37], Deep Multi-Layer Perceptron (DMLP) [38], and Recurrent Neural Networks (RNNs) [39].

Summary tables and bar charts regarding all the methods of the reviewed papers are provided.

2.4 Taxonomy

The taxonomy is shown as Fig. 2. We can divide it into two parts: the first is regarding computing technology and the second is credit risk application domain. The two parts are further categorized into subsections. These two parts are connected and fused with each other. All the right-side sub-domains include the left-side techniques, and all the techniques can be applied in the financial domains.