Expert Systems With Applications 176 (2021) 114835

Contents lists available at ScienceDirect

Expert Systems With Applications

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

A conservative approach for online credit scoring[☆]

Afshin Ashofteh [a,*], Jorge M. Bravo [b]

[a] NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa, Campus de Campolide, 1070-312 Lisboa, Portugal
[b] NOVA Information Management School (NOVA IMS) & Université Paris-Dauphine, Portugal

ARTICLE INFO

JEL Classifications:
C01
C41
C53
C58
E51
G21

Keywords:
Risk analysis
Online credit scoring
Big Data
Kruskal_Wallis statistic
Open banking
Machine learning

ABSTRACT

This research is aimed at the case of credit scoring in risk management and presents a novel machine learning method to be used for the default prediction of high-risk branches or customers. This study uses the Kruskal-Wallis non-parametric statistic to form a conservative credit-scoring model and to study the impact on modeling performance on the benefit of the credit provider. The findings show that the new credit scoring methodology represents a reasonable coefficient of determination and a very low false-negative rate. It is computationally less expensive with high accuracy with around 18% improvement in Recall/Sensitivity. Because of the recent perspective of continued credit/behavior scoring, our study suggests using this credit score for non-traditional data sources for online loan providers to allow them to study and reveal changes in client behavior over time and choose the reliable unbanked customers, based on their application data. This is the first study that develops an online non-parametric credit scoring system, which is able to reselect effective features automatically for continued credit evaluation and weigh them out by their level of contribution with a good diagnostic ability.

1. Introduction

Credit scoring involves the use of analytical methods to transform relevant data into numerical measures that inform and determine credit decisions. In recent years the use of credit scoring tools has expanded beyond their original purpose of assessing credit risk, such as establishing the initial and ongoing credit limits available to borrowers, assessing the risk-adjusted profitability of account relationships, and assisting in a range of loan servicing activities, including fraud detection, delinquency intervention, and loss mitigation (Thomas, 2000).

The critical role of the lending market in causing the latest global financial crisis has increased academic research, policy interest, and bank regulation in this area. The banking regulatory framework changes brought by the revised Basel Committee on Banking Supervision (BCBS) Accords (later adopted by national legislation in many countries and regions, for instance, the European Capital Requirement Directives and the US Regulatory Capital Rules) introduced stronger risk management requirements for banks, with capital requirements tightly coupled to estimated credit portfolio losses. The recently adopted IFRS9 and FASB's Current Expected Credit Loss (CECL) standards introduce revised expected credit loss or impairment calculation rules requiring financial institutions to calculate the expected loss for the banking book over the entire life of the exposures, conditional on macroeconomic factors, on a point-in-time basis, that is, recalibrating PDs where necessary to reflect the effects of the current economic conditions. Encouraged by regulators, banks devoted significant resources to develop an Internal Ratings Based approach (IRB) for the calculation of risk-weighted assets for credit risk to better support decisions when granting loans, to quantify expected credit losses, and to assign the mandatory economic capital (Chamboko & Bravo, 2020).[1]

To assess credit risk, in developed markets, lenders typically consider historical loan application and loan performance data collected regularly from a small number of sources on the basis of long-standing banking and credit relationships to develop credit-scoring models to evaluate the ability to repay, the willingness to repay, and identify fraud. The Edward Altman Z-score model for bankruptcy prediction and the FICO score for retail credit scoring are some of the oldest industry standards, which loan providers still use because of their high

☆ This article has been certified as Reproducible by the CodeOcean: https://codeocean.com/.

* Corresponding author.
E-mail addresses: aashofteh@novaims.unl.pt (A. Ashofteh), jbravo@novaims.unl.pt (J.M. Bravo).

[1] The recently approved BCBS (Basel IV) reforms of the standardized approach and of the CR-IRB approach for the calculation of risk-weighted assets for credit risk will limit the extent to which banks can reduce capital requirements through the use of internal models.

https://doi.org/10.1016/j.eswa.2021.114835
Received 1 July 2020; Received in revised form 1 March 2021; Accepted 1 March 2021
Available online 10 March 2021

interpretability (Baesens, Roesch, & Scheule, 2016).

These methods are less effective in emerging economies and among low-income unbanked segments of the population, who often do not have access to formal financing and/or do not earn regular labor income. To cope with these constraints and to improve credit risk assessment, banks and loan providers are increasingly using non-traditional data sets (e.g., mobile operators, utilities, retailers, and direct-sales companies data) to sophisticate their credit bureaus and credit rating services. This factor poses new challenges to credit scoring modelers since non-traditional data must typically be collected from different sources, and its volume is several times that of traditional sources. By pursuing this approach, lenders seek to have more accurate information and incentive to grow the credit market under a robust credit control framework. By increasing their use of these new data sources, they try to provide more lending to their public customers and get to analyze loan requests better, ultimately increasing the loan ratio and decreasing the decision time. People will then have more monthly disposable money for spending, which will contribute to the economy, but it can also create risks for financial institutions. Therefore, as shown in Fig. 1, non-traditional data sets provide the credit market a chance to manage different data sources to boost credit analysis outcomes and follow the stipulated recommendations of standards appropriately.

While this approach presents some opportunities, it also carries some challenges, which are represented in Fig. 1. First, most of the mobile phone features are redundant and do not contribute to representing credit risk. Second, banks and loan providers should follow the regulators, and the critical change factor in banking is regulation. It means that fluctuations in the economy and regulations could change the behavior of both banks and customers. Non-traditional data sets can reveal these fluctuations, but current methods are computationally expensive with a high false-negative rate. The third risk factor is the smart behavior of fraudsters, and it means conscious changes may exist in non-traditional data, which influence the contribution level of features in credit scoring. However, the current methods are not wise enough to renew credit scores over time. As a result, we need highly effective and computationally less expensive solutions to calculate an informative credit score for satisfying the accuracy expectation of financial institutions. Although there are a large number of techniques employed in the development of credit-scoring, empirical studies show that the false-negative rate obtained is still not good enough for non-traditional data sets.

In this paper, we introduce a novel time-dependent credit scoring method to identify good loans with a low false-negative rate. The method uses a two-step approach based on an initial Kruskal-Wallis non-parametric statistic analysis to form a computationally efficient credit-scoring model based on an artificial neural network, Logistic regression with Ridge penalty, Random Forests (RF), and Support Vector Machine (SVM) to learn the model and to assess model performance. The Kruskal-Wallis statistic is used for selecting the most prominent features for assessing loan default in credit risk management over time. This statistic is sensitive to events that are far from the credit scores of good

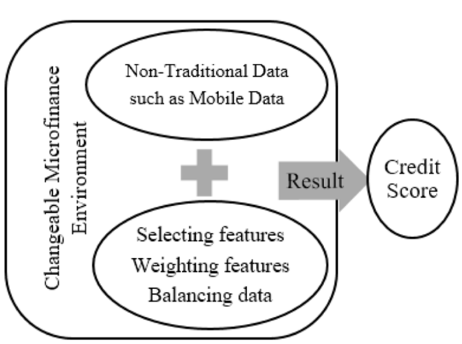


Fig. 1. Challenges in credit scoring.

clients, computationally less expensive and very simple to implement. Additionally, we introduce a credit scoring index that uses the Kruskal-Wallis statistic as a weight of feature to decrease the false-negative rate which is able to purify features and decrease the dimensions in real-time. This new credit scoring index may be particularly interesting for loan providers assessing loan applications from individuals without any credit history and based exclusively on non-traditional data analysis. Illustrative empirical results on the use of this novel time-dependent credit scoring method are provided considering a sample credit data-set. The empirical findings show that the new credit scoring methodology represents a reasonable coefficient of determination and a low false-negative rate. The accuracy is high and the model is computationally less expensive. This is the first study that develops a non-parametric credit scoring system that is able to reselect effective features for continued credit evaluation and weigh them out by their level of contribution with a good diagnostic ability.

The rest of the paper is structured as follows. In Section 2, we review credit-scoring models and new non-traditional data sources. Section 3 introduces the novel credit scoring method and discusses the calculation of the credit score. Then, using a credit risk data set, we compare the classification accuracy of credit scores with available features in Section 4. Furthermore, an artificial neural network model is dedicated to the new method to show the accuracy of predicting the probability of default. In Section 5, we discuss the main managerial and theoretical implications of this research. Finally, Section 6 contains some concluding remarks.

2. Literature review

2.1. Credit scoring models using traditional data sets

Traditional credit-scoring models applying single-period classification techniques (e.g., logit, probit) to classify credit customers into different risk groups and to estimate the probability of default are still the most popular data mining techniques used in the industry (Chamboko & Bravo, 2019a, 2019b). Altman (1968) pioneered this area by developing the Z-score discriminant analysis model based on five financial ratios to predict corporate bankruptcy. Since then, several techniques have been developed to help decision-makers and analysts in predicting financial failure by considering both traditional statistical methods and more sophisticated (e.g., advanced machine learning) modeling approaches and alternative sets of predictor features. Standard models using external ratings provided by external credit assessment institutions have also been successfully applied. The set of classification algorithms used in credit scoring includes individual classifiers and homogenous and heterogeneous ensembles.

Individual classifiers employing single statistical or operational research methods include linear and multiple discriminate analysis (DA), logistic regression (LR), probit analysis, linear and quadratic programming, and data envelopment analysis (see, e.g., Altman, Haldeman, & Narayanan, 1977; Zmijewski, 1984; Jones & Hensher, 2004; Premachandra, Bhabra, & Sueyoshi, 2009; Kwak, Shi, & Kou, 2012). Classifiers using machine learning methods such as neural network (NN), support vector machine (SVM), decision trees (DT), and genetic and evolutionary algorithms (GA) have also been investigated (see, e.g., Hand & Henley, 1997; Arminger, Enache, & Bonne, 1997; Baesens et al., 2003; Shin, Lee, & Kim, 2005; Lensberg, Eilifsen, & McKee, 2006; Erdogan, 2013; Kruppa, Schwarz, Arminger, & Ziegler, 2013; Acosta-González & Fernández-Rodríguez, 2014; Lessmann, Baesens, Seow, & Thomas, 2015; Butaru et al., 2016; Abellán & Castellano, 2017; Zhao et al., 2017). Homogenous ensembles typically employ one of the above individual classification methods with various samples or parameters to build base classifiers, which are subsequently combined using a majority voting rule or other frequentist or Bayesian integration methods (Feng, Xiao, Zhong, Qiu, & Dong, 2018). Recently, heterogeneous ensemble, which combines the prediction of base models created by alternative classification algorithms, often in a dynamic (adaptive or selective) way, has attracted much attention because of its superior predictive performance over homogeneous ensemble (see, e.g., Tsai, Hsu, & Yen, 2014; Xia, Liu, Da, & Xie, 2018). Some approaches model default as a dynamic (sequential) process (see, e.g., Volkov, Benoit, & Van den Poel, 2017).

Recent proposals in the field of credit scoring focuses on three dimensions: novel classification algorithms using dynamic ensembles, deep learning methods, dissimilarity space, associative memories, and probabilistic rough sets, novel performance measures, and the minimization of the decision-relevant costs, and statistical hypothesis tests. Xia et al. (2018) propose a novel heterogeneous ensemble credit model (named bstacking) that integrates the bagging algorithm with the stacking method. Feng et al. (2018) develop a new dynamic ensemble classification method for credit scoring based on soft probability in which classifiers are first selected based on their classification ability and the relative costs of Type I error and Type II error in the validation set and then combined to get an interval probability of default by using soft probability. Luo, Wu, and Wu (2017) investigate and compare the classification performance of deep belief networks (DBN) with Restricted Boltzmann Machines with that of popular credit scoring models such as LR, multi-layer perceptron, and SVM using credit default swaps data. Cleofas-Sánchez, García, Marqués, and Sánchez (2016) propose an alternative technique for financial distress prediction based on a specific type of neural network called hybrid associative classifier with translation (HACT). The HACT neural network is an associative memory that merges the encoding phase of the linear associator with the decoding phase of the Steinbuch's lern matrix to improve the performance of the classifier. Pławiak, Abdar, and Rajendra Acharya (2019) develop a new approach for credit scoring based on a deep genetic cascade ensemble of different SVM classifiers called Deep Genetic Cascade Ensembles of Classifiers (DGCEC) combining evolutionary computation, and ensemble and deep learning methods. García, Marqués, and Sánchez (2019a) address the problem of corporate bankruptcy prediction considering four linear classifiers (Fisher's linear discriminant, linear discriminant classifier, SVM, and LR) adopting a dissimilarity representation in which samples to be classified/predicted are derived from pairwise dissimilarities instead of being represented as usual by a set of features (explanatory variables), which defines a feature space. Maldonado, Peters, and Weber (2020) propose a methodology for credit scoring that minimizes the decision-relevant costs by classifying borrowers into three instead of two classes using the theory of three-way decisions with probabilistic rough sets. García, Fernández, Luengo, and Herrera (2010) perform an experimental analysis to compare scorecard performance.

Despite their popularity, credit scoring models can only provide an estimate of the lifetime probability of default for a loan but cannot identify the existence of cures and/or other competing transitions and their relationship to loan-level and macro covariates, and do not provide insight on the timing of default, the cure from the default, the time since default, and time to collateral repossession (Lessmann et al., 2015; Chamboko & Bravo, 2020). Survival models incorporating time-varying covariates such as macroeconomic conditions which affect performance on loan payment over time and the ability to forecast event occurrence (default, recovery, prepayment, foreclosure) in the next instant of time, given that the event has not occurred until that time, have proven to overperform traditional methods in empirical studies (see, e.g., Noh, Roh, & Han, 2005; Sarlija, Bensic, & Zekic-Susac, 2009; Tong, Mues, & Thomas, 2012; Bellotti & Crook, 2013; Castro, 2013; Chamboko & Bravo, 2016, 2019a, 2019b). A handful of studies have also used the same to model foreclosure on mortgages (Gerardi, Shapiro, & Willen, 2007) and cure from delinquency to current (Chamboko & Bravo, 2016, 2019a, 2019b; Ha, 2010; Ho Ha & Krishnan, 2012). The competing risks survival framework has also been used to model the competing risks of early payment and default on loan contracts (Deng, Quigley, & Order, 2000; Stepanova & Thomas, 2002).

2.2. Non-traditional data sets for credit scoring

Credit scoring models using non-traditional data sets are a cost-effective method of surveying personalities for risk management purposes of monetary institutions. It shifts credit scoring to high-tech to avoid the personal subjectivity of analysts or underwriters (Fensterstock, 2005). It also helps in increasing the speed and consistency of the application processes and allows financial firms to automate their processes (Rimmer, 2005).

Credit scoring and new technologies help loan providers shorten the processing time of loan applications and improve the allocation of resources (Jacobson & Roszbach, 2003). Additionally, it can aid insurance firms in making better predictions on claims and determining the interest rate which the firms should charge their consumers as well as the pricing of portfolios and products (Avery, Bostic, Calem, & Canner, 2000; Kellison & Wortham, 2003). Mobile phone data is a new Big Data source for smarter credit-scoring models, independent of the usual financial institutes databases. Mobile phones provide non-traditional data sources in the form of call detail records (CDR) and many other log files which can contribute to improving retail risk models not only for bank customers but also for the largely unbanked population who have no regular credit history.

Fig. 2 represents how non-traditional data is emerging in credit scoring by using parallel computing, distributed computing, and Big Data solutions. They represent how digitization in banking has gradually allowed financial institutions to use both Big Data and traditional credit records for managing risks. It shows how technology development is helping loan providers create value from several huge volumes of non-traditional data with increasing computation efficiency. Providing faster and consistent decisions for sub-prime customers with poor credit records, credit impairment, missing data in their credit histories, or difficulty in validating their income is another advantage of using non-traditional data in credit scoring modeling (Quittner, 2003). Despite its considerable benefits, lending to these groups is characterized as an inherent high risk due to a lack of collateral and information asymmetry. Some researchers use behavioral signatures in mobile phone data to predict default with an accuracy almost similar to that of credit-scoring methods that use financial history by Random Forest and Logistic regression (Bjorkegren & Grissen, 2018). Pedro, Proserpio, and Oliver (2015) developed MobiScore, a methodology that models the user's financial risk using data collected from mobile usage using gradient boosting, support vector machine, and linear regression models. Other studies have used boosted decision trees and Logistic regression to create a credit score for underbanked populations considering information about people's usage of various mobile apps to make conclusions about their mood and personality traits (Chittaranjan, Blom, & Gatica-Perez, 2013; Do & Gatica-Perez, 2010; Skyler, Eric, Isaac, & Felix, 2017; Verkasalo, López-Nicolás, Molina-Castillo, & Bouwman, 2010). As a result, mobile phone data and social network analytics were used in credit scoring applications showing that incorporating telco data has the potential to increase the value of credit scoring (Óskarsdóttir, Bravo, Sarraute, Vanthienen, & Baesens, 2019).

Current credit scoring methods are computationally expensive and face critical challenges such as drift and class imbalance, reject inference, outliers, data set shift, irrelevant features, and missing and noisy data. As we discussed in Fig. 1, the class imbalance problem and changes in the macro-finance environment and markets could potentially change the relationship between client characteristics and credit assessment results over time, causing concept drift in client credit assessments. Zhang and Liu (2019) proposed a novel multiple time scale ensemble classifier and a novel sample-based online class imbalance learning procedure to handle these two problems in the client credit assessment data streams. Because of the minority of delinquent customers, class distributions are highly imbalanced and represented skewed distributions. Although the topic of imbalanced classification has gathered the full attention of researchers during the last several years, such as the

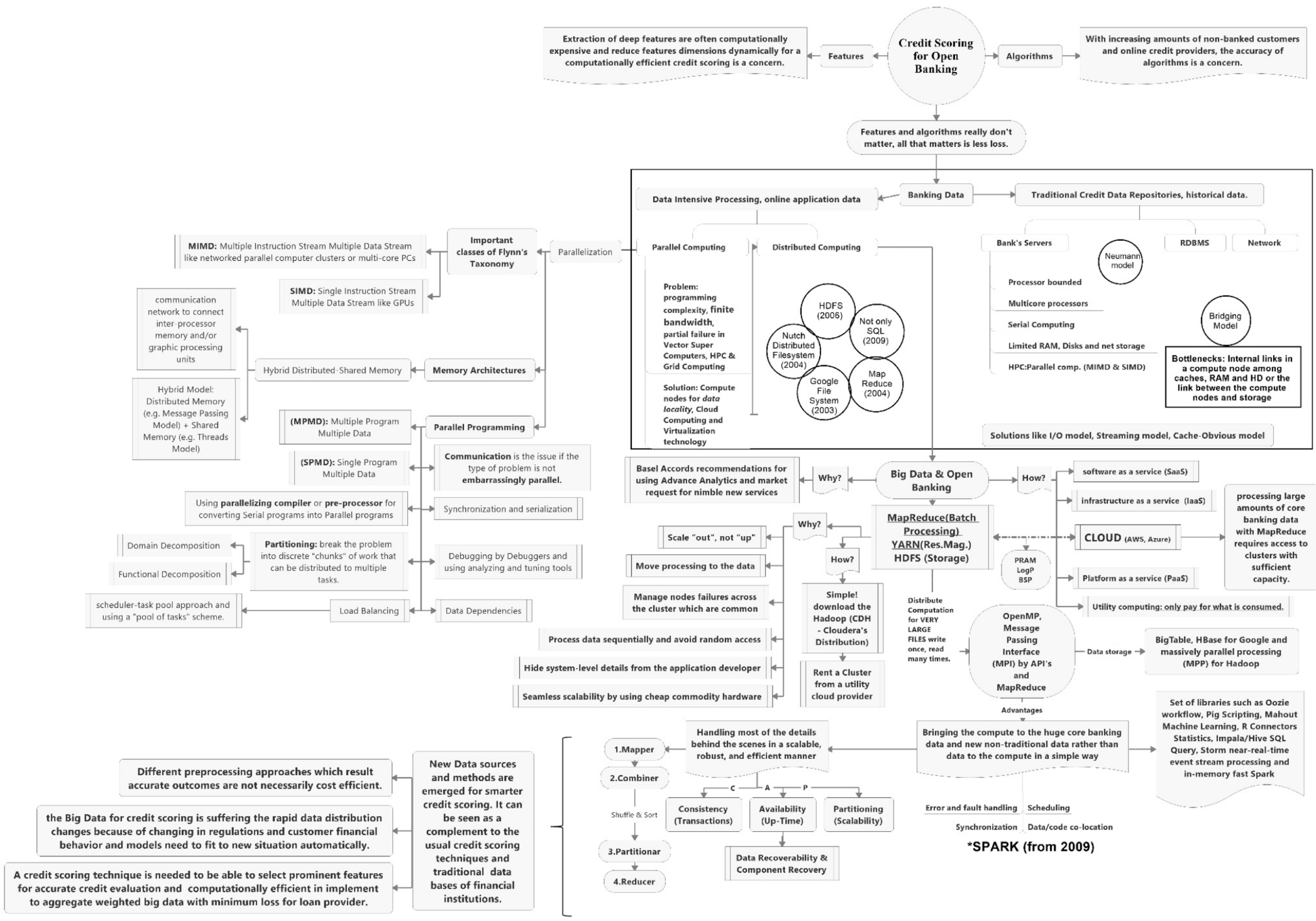


Fig. 2. Emergence of non-traditional data analysis in credit scoring.

cost-sensitive learning technique by Douzas and Bacao (2018), the emergence of mobile phone data brings new problems and challenges for the class imbalance problem in credit scoring, especially for unbanked individuals. The misclassification costs of false-negative cases are typically much higher than those associated with the non-default or non-bankrupt (negative) class. García, Marqués, Sánchez, and Ochoa-Domínguez (2019b) investigate the potential links between the performance of several classifier ensembles and the positive (default or bankrupt) sample types using 14 different real-life financial databases.

Credit scoring models are typically built on historical performances, which means that only accepted requests are used in estimating the probability of default, which may cause sample bias and reduce predictive performances. Recent proposals in the field of credit scoring adopt reject inference methods, i.e., use the information encompassed in the rejected samples in combination with accepted samples. Li, Tian, Li, Zhou, and Yang (2017) propose a new method in reject inference using the machine learning technique of Semi-supervised Support Vector Machines (SSVM) to classify the status of rejected borrowers and empirically investigate the performance of the new method. Xia (2019) propose a novel reject inference model (named OD-LightGBM) that combines a recent outlier detection algorithm (i.e., isolation forest) and state-of-the-art gradient boosting decision tree (GBDT) classifier (LightGBM). Nyitrai and Virág (2019) handle the problem of outliers in credit scoring and examine the impact of outliers winsorized at different levels. Gicić and Subasi (2019) address the problem of class imbalance in microcredit data and propose a new microcredit scoring model based on synthetic minority oversampling technique (SMOTE) for data pre-processing and ensemble classifiers. Óskarsdóttir et al. (2019) used the undersampling method to reduce class imbalance for training sets of data, which shows the intention to reduce the size of the majority class when applying these analytics techniques. Tian, Yong, and Luo (2018) propose a new method using the state-of-the-art kernel-free Fuzzy Quadratic Surface Support Vector Machine (FQSSVM) to infer the statuses of the rejected applicants and solve the outlier problem in credit assessment. Liu and Pan (2018) propose a new hybrid classifier based on fuzzy-rough instance selection to minimize the negative influence on the classification accuracy of using the wrong number of clusters or poor starting points of each cluster. García, Marqués, and Sánchez (2012) empirically investigate whether the application of filtering algorithms leads to an increase in the accuracy of instance-based classifiers in the context of credit risk assessment. The authors consider 20 different algorithms and 8 credit databases and conclude that that the filtered sets perform significantly better than the non-preprocessed training sets when using the nearest neighbor decision rule and that some techniques are most robust and accurate when confronted with noisy credit data. Recently, sound statistical and machine learning procedures that are computationally scalable to massive non-traditional datasets have been proposed (Jordan, 2013). Examples are subsampling-based approaches (Kleiner, Talwalkar, Sarkar, & Jordan, 2014; Kruppa et al., 2013; Liang, Cheng, Song, Park, & Yang, 2013; Ma, Mahoney, Yu, & Yu Ma, 2015; Maclaurin & Adams, 2015), divide and conquer approaches (Song & Liang, 2015), and online updating approaches (Schifano, Wu, Wang, Yan, & Chen, 2016).

3. Dynamic nonparametric credit score

The performance of a classification system can be improved by picking up the optimized features of mobile phones and decreasing the complexity of the model in the preprocessing stage. Many methods have been developed for choosing significant features with high information, such as the Kruskal-Wallis method (Saeys, Inza, & Larrañaga, 2007). The Kruskal-Wallis test as a nonparametric approach is useful to select informative features for loan default in credit risk management. Because it is sensitive to events that are far from the credit scores of good clients, we use the Kruskal-Wallis non-parametric statistic in our proposed method, which is computationally less expensive, and very simple to implement. Additionally, we introduce a new credit scoring approach that is able to purify features and decrease the dimensions in real-time. We will describe it in the following subsection 3.1. Next, we use the Kruskal-Wallis statistic measures as a weights of features to decrease the false-negative rate and improve the model's accuracy, which will be discussed in subsection 3.2, and finally in subsection 3.3, we combine these two steps to introduce a new credit scoring formula.

3.1. Kruskal-Wallis statistic for online features reduction

Let us define the null hypothesis that a feature does not contain discriminative information to detect default possibility in loan requests; otherwise, it is an informative feature and will be selected for contributing in the credit scoring. An assumption for this test is that the samples from the credit scores of good clients and credit scores of new clients are independent random samples from continuous distributions. In addition, we consider the time as an index of weights in our credit score method to ensure that the distributions of the training dataset of existing and new customers have the same shape at the time of analysis in this proposed online learning environment.

The computational procedure of the test can be considered as follows. Let $X_{ijt}X_{ijt}$ denote an observation of feature jj from the clientii at time tt. If we let N_t be the total number of credit scores, which is equal to the total number of customers at time tt, then by X_{ij}'sX_{ij}'s at time t, we will have a matrix $X^t{}_{N\times K}X^t{}_{N_t\times K_t}$ with N_t rows as the number of clients and K_t as the number of features at time tt. Loans in banking credit risk literature usually divide into three categories: "Good" for good loans, "Medium" for substandard loans/doubtful loans, and "Bad" for loss loans. The Kruskal-Wallis test is appropriate for these kinds of categorical variables with three or more groups. However, most of the available datasets for credit purposes merge the first and second groups as bad loans to create a label attribute with two values, for instance, zero for the good loans and one for the bad loans.

We denote by S the number of groups in the Kruskal Wallis statistic, n_{ijt} the number of observations in group i at time t, and vector $Y_{N\times 1}Y_{N_t\times 1}$ the label vector considering, for instance, three possible labels for each customer ("Good," "Substandard or Doubtful," and "Loss").

Computationally, if $R_{ij}R_{ijt}$ is the rank assigned to the $j^{th}j^{th}$ feature of $i^{th}i^{th}$ client at time t, then the Kruskal-Wallis statistic for $j^{th}j^{th}$ feature at time tt for N_t customers is

$$H_j^t = \frac{12}{N_t(N_t+1)}\left(\sum_{i=1}^{S}\frac{R_{ijt}^2}{n_{ijt}}\right) - 3(N_t+1)\ \ j = 1,2,\cdots,K_t \tag{1}$$

Then

$$H_j^t \xrightarrow{d} \chi_1^2 \text{ in distribution } H_j^t \xrightarrow{d} \chi_1^2 \text{ in distribution}$$

where $\chi_1^2\chi_1^2$ is the $\chi^2\ \chi^2$ distribution with one degree of freedom. The null hypothesis will be rejected if the computed value of $H_j^tH_j^t$ for each jj from 1 to K_t exceeds the value of chi-square for reselected confidence level and 1 degree of freedom. Simply, in the credit scoring scenario, we define $\gamma_j^t\gamma_j^t$ to find a balanced measure with less complexity to be followed with zero as a baseline and negative-positive values for making decisions about features. We make γ_j^t based on following $\tau_j^t\tau_j^t$.

Let

$$\tau_j^t = 1 - \frac{H_j^t}{H_j^t+\chi_1^2} = \frac{\chi_1^2}{H_j^t+\chi_1^2}, \tag{2}$$

$$0 \le \tau_j^t \le 1;\ \text{for all j}, j = 1,2,\cdots,k$$

Hence, we obtain that

$$-0.5 \le \tau_j^t - 0.5 \le 0.5$$

Then we define a nonparametric measure $\gamma\gamma$ for feature j at time t as $\gamma_j^t = \tau_j^t - 0.5\gamma_j^t = \tau_j^t - 0.5$, therefore

$$\gamma_j^t = \frac{\chi_1^2 - H_j^t}{2\times(\chi_1^2+H_j^t)} \tag{3}$$

and

$$-0.5 \le \gamma_j^t \le 0.5. \tag{4}$$

If $\gamma_j^t\gamma_j^t$ is positive, then the feature is not able to differentiate the classes, and if it is negative, then the feature could be used for the modeling phase to determine the creditworthiness of an applicant. It is clear that this change is only superficial to make it easier to understand, programing and represent it as a control chart in a dashboard. Now, we study the behavior of $\gamma_j^t\gamma_j^t$ by looking at its distribution function.

Let $\gamma_j^t\gamma_j^t$ be as defined in formula (3) and denote by $F_{\gamma_j^t}(y)F_{\gamma_j^t}(y)$ be the corresponding cumulative distribution function (cdf)

$$\begin{aligned} F_{\gamma_j^t}(y) &= P\left(\gamma_j^t \le y\right) = 1 - P\left(\frac{\chi_1^2 - H_j^t}{2\times\left(\chi_1^2+H_j^t\right)} > y\right) = 1 - P\Bigg(H_j^t \\ &< \frac{(1-2y)\chi_1^2}{(1+2y)}\Bigg) F_{\gamma_j^t}(y) = P\left(\gamma_j^t \le y\right) = 1 - P\left(\frac{\chi_1^2 - H_j^t}{2\times\left(\chi_1^2+H_j^t\right)} > y\right) \\ &= 1 - P\left(H_j^t < \frac{(1-2y)\chi_1^2}{(1+2y)}\right) \end{aligned} \tag{5}$$

or, equivalently,

$$F_{\gamma_j^t}(y) = 1 - F_H\left[\left(\frac{0.5-y}{0.5+y}\right)\times\chi_1^2\right] F_{\gamma_j^t}(y) = 1 - F_H\left[\left(\frac{0.5-y}{0.5+y}\right)\times\chi_1^2\right] \tag{6}$$

where F_HF_H is the cumulative distribution function of the Kruskal-Wallis statistic and $\left(\frac{0.5-y}{0.5+y}\right)\times\chi_1^2 > 0\left(\frac{0.5-y}{0.5+y}\right)\times\chi_1^2 > 0$ for any values of $y\in(-0.5, 0.5)y\in(-0.5, 0.5)$.

Furthermore, because $H\rightarrow x_1^2 H\rightarrow x_1^2$ in distribution, the density function of $\gamma_j^t\gamma_j^t$ will be

$$\begin{aligned} f_{\gamma_j^t}(y) &= \frac{dF_{\gamma_j^t}(y)}{dy} = -\frac{d\left(\left(\frac{0.5-y}{0.5+y}\right)\times\chi_1^2\right)}{dy}\times f_H\left(\left(\frac{0.5-y}{0.5+y}\right)\times\chi_1^2\right) f_{\gamma_j^t}(y) \\ &= \frac{dF_{\gamma_j^t}(y)}{dy} = -\frac{d\left(\left(\frac{0.5-y}{0.5+y}\right)\times\chi_1^2\right)}{dy}\times f_H\left(\left(\frac{0.5-y}{0.5+y}\right)\times\chi_1^2\right) \end{aligned} \tag{7}$$

or, equivalently,

$$f_{\gamma_j}(y) = \sqrt{\frac{\chi_1^2}{2\pi(0.25-y^2)(0.5+y)^2}}\times Exp\left\{-\left[\left(\frac{0.5-y}{1+2y}\right)\times\chi_1^2\right]\right\},\quad -0.5 < y < 0.5 \tag{8}$$

With 95% confidence level $\chi_{1,0.05}^2 = 3.841\chi_{1,0.05}^2 = 3.841$ and the $\gamma_j^t\gamma_j^t$ density function is

$$f_{\gamma_j^t}(y) = 0.7818 \times \sqrt{\frac{1}{(0.25 - y^2)(0.5 + y)^2}} \times Exp\left[\frac{3.841 \times (y - 0.5)}{1 + 2y}\right], \quad -0.5 < y < 0.5 \tag{9}$$

The density function of $\gamma_j^t\gamma_j^t$ is shown below (Fig. 3.).

From Fig. 3, it is clear that the majority of probability density based on the area under the density function curve for statistically significant features is between −0.3 and 0. We can expect that most of the magnificent fluctuations in the effectiveness of features on our credit score model will happen in the small probability area under the density function between −0.3 and 0. Therefore, if we consider only the sign of $\gamma_j^t\gamma_j^t$ in a flag attribute, instead of its value to select the informative features based on positive signs, we can manage memory better and eliminate unnecessary features in our computer program with only a Boolean type attribute, which only needs one byte of memory, the smallest unit addressable with the CPU in compare with, for instance, decimal which needs 12 bytes. It means less memory will be occupied and data transfer in the network will be optimized, especially for massive non-traditional datasets with numerous features.

3.2. Kruskal-Wallis statistic for empowering features

In this step, we use $\gamma_j^t\gamma_j^t$ in our credit score model to boost the effective features. It was shown that $\gamma_j^t\gamma_j^t$ signs can be a part of credit scoring as an indicator for feature selection. Now, we define a transformed $\gamma_j^t\gamma_j^t$ to be used as a transformation weight for the $k^{th}k^{th}$ feature to improve the performance of classification and credit scoring.

Let us define $w_k^t w_k^t$ as the transformation weight of feature kk at time tt:

$$w_j^t = \begin{cases} 2 \times \left|\gamma_j^t\right| & \text{for}\gamma_j^t < 0 \\ 0 & \text{for}\gamma_j^t > 0 \end{cases}, \quad j = 1, 2, \cdots, N \tag{10}$$

Then, $\varphi\varphi$ as the impact factor of feature kk at time tt will be

$$\varphi_j^t = \begin{cases} \dfrac{w_j^t}{\sum_{j=1}^{k} w_j^t} & \text{for } w_j^t > 0 \\ 0 & \text{for } w_j^t = 0 \end{cases}, \quad \sum_{j=1}^{N} \varphi_j^t = 1. \tag{11}$$

As we discussed, for a 95% confidence level, if $\left\{H_j^t|H_j^t > \chi^2_{1,0.05} = 3.84\right\}\left\{H_j^t|H_j^t > \chi^2_{1,0.05} = 3.84\right\}$ then we reject the null hypothesis. It means that the feature contains discriminative information to detect default possibility in loan requests. Furthermore, as shown in Fig. 3, important fluctuations in the effectiveness of features in credit scoring happens when $\gamma_j^t\gamma_j^t$ is between −0.3 and 0. Now, the equivalent values of $\gamma_j^t\gamma_j^t$ and $w_j^t w_j^t$ to values of $H_j^t H_j^t$ is shown below in Table 1.

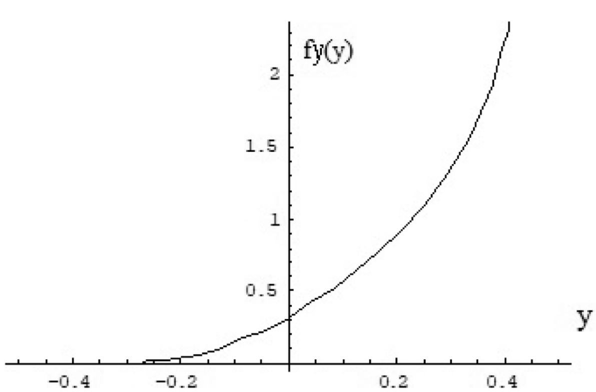


Fig. 3. The density plot of $\gamma_j^t\gamma_j^t$.

Table 1 illustrates that $w_j^t w_j^t$ and $\varphi_j^t\varphi_j^t$ react impressively to the changes of $\gamma_j^t\gamma_j^t$ in this interval. Therefore, $w_j^t w_j^t$ close to one shows the high capability of the feature to recognize the credit level of the customer. A trend towards zero means the probability of the feature's influence is decreasing. Monitoring the trend of $w_j^t w_j^t$ for different t could show the level of the model's reliability during the time. If it decreases to zero randomly and sequentially then it can be concluded that the reliability of the trained model based on those features is decreasing, and it is a warning sign to renew the model by using a new set of features. It helps to identify for how long our scoring model keeps up with the initial performance and to discover the right time of redoing the training step with a new set of features. It would also be useful to monitor $w_j^t w_j^t$ by considering zero for excluding the feature from the credit score model. In addition to the above mentioned advantages, we use $w_j^t w_j^t$ at time t as the power of attribute j to improve the accuracy of the model and to decrease the false-negative rate.

For this purpose, we consider all attributes for defaulted customers within the training dataset to the power of corresponding $w_j^t w_j^t$ in the training stage. The useful attributes with the higher $H_j^t H_j^t$ will have $w_j^t w_j^t$ closer to one. It means these optimal features will experience less change than features with a lower ability to determine the loan default. As a result, in the training stage, we can make distinguish between attributes of good and bad loans based on their power of contribution in the model. In fact, we are saving the high-performance attributes, and we use the less important attributes as labels for bad loans. By using the features with high $w_j^t w_j^t$ for making the model and the features with low $w_j^t w_j^t$ to differentiate between good and bad loans the model's performance is expected to improve.

3.3. Credit score formulation

Now, we introduce our new credit risk index (CRI) for attributes with interval or ratio scale of the i^{th} client at time tt by using the geometric mean as following:

$$CRI_{it} = \prod_{j=1}^{k} \left[\frac{(x_{ij}^t)^{w_j^t}}{(\overline{x}_j^t)} \times 100\right]^{\varphi_j^t}, \quad i = 1, 2, \cdots, N \tag{12}$$

where $\overline{x}_j^t\overline{x}_j^t$is the standard profile of attribute jj extracted from the data of good clients. If we investigate these credit score characteristics by using the most desirable axioms of the axiomatic approach to the index number theory, it satisfies most of them[2]. Thus, this credit index formula as a homogeneous symmetric average can be calculated as an accurate aggregate measure, and it is able to renew features dynamically and weighted out by $\varphi_j^t\varphi_j^t$ as their impact factors. The pseudo-code of the proposed methodology is listed in Table 2.

In the training stage, obviously, we will have a very large area under the AUC curve because the values of bad loans in the training set have experienced the shift based on $w_j^t w_j^t$. However, in the test stage, we are using an independent out-of-sample dataset, and we actually have no idea which loan is the default. As shown in Table 2, we first apply blindly the transformation for all attributes of all clients in the test dataset, including good and bad loans. This will shift the attributes of good loans

[2] Axiomatic approach to the index number theory such as the Positivity test, Continuity Test, Identity Test, Homogeneity Test for Period t, Homogeneity Test for Period zero, Commodity Reversal Test, Invariance to Changes in the Units of Measurement or the Commensurability Test, Time Reversal Test, Circularity or Transitivity Test, Mean Value Test, Monotonicity Test with Respect to Period t and Monotonicity Test with Respect to Period zero.

Table 1
$H_j^tH_j^t$ values and its equivalent in $\gamma_j^t\gamma_j^t$, $w_j^tw_j^t$ and $\varphi_j^t\varphi_j^t$.

H_j^t	0	1	2	3	3.5	4	5	10	20	40	100	400	700
γ_j^t	0.5	0.29	0.15	0.06	0.02	−0.01	−0.06	−0.22	−0.33	−0.41	−0.46	−0.49	−0.495
w_j^t	0	0	0	0	0	0.020	0.131	0.445	0.678	0.825	0.926	0.981	0.989
φ_j^t	0	0	0	0	0	0.004	0.026	0.089	0.136	0.165	0.185	0.196	0.198

Table 2
Pseudo Code of Proposed Methodology.

```
INPUT attributes and Status variable;
OUTPUT weighted attributes, CRI;
STATEXPLORE attributes;
CHANGE[outliers] = FALSE; {outliers are important in credit scoring.}
SET χ^2_{1,0.05} = 3.84χ^2_{1,0.05} = 3.84;
FOR each attribute DO
  The Kruskal-Wallis.test;
  IF (Scale[attribute] is not Nominal) THEN
    IF (KW.statistic > χ^2_{1,0.05}χ^2_{1,0.05}) THEN
      IF (data = train.set and Default = True) THEN
        SET Gama = (χ^2_{1,0.05}χ^2_{1,0.05} − KW.statistic)/(2*(χ^2_{1,0.05}χ^2_{1,0.05} + KW.
statistic))
        SET W = 2 * ABS(Gama);
        ET attribute.value = POWER [attribute.value, W]
        IF (MISSING(attribute) = TRUE) THEN attribute = attribute
        IF (MISSING(attribute) = TRUE) THEN attribute = AVERAGE(attribute)
        IF (attribute = 0) THEN attribute = AVERAGE(attribute)
      ELSEIF (data = test.set for ALL) THEN
        SET Gama = (χ^2_{1,0.05}χ^2_{1,0.05} − KW.statistic)/(2*(χ^2_{1,0.05}χ^2_{1,0.05} + KW.
statistic))
        SET W = 2*ABS(Gama);
        SET attribute.value = POWER [attribute.value, W]
    ELSE
    SET attribute = excluded; {equivalent to Set W = 0; Set Phi = 0;}
  Else
  SET attribute = unchanged; {equivalent to Set W = 1; Set Phi = 1}
END DO
# computing CRI
FOR each attribute DO
  IF (Default = False) THEN Mean_default_NO = AVERAGE(attribute)
  SET Phi = W/SUM(W's);
  SET attribute_CRI = POWER [(attribute.value.W/Mean_default_NO × 100),
Phi]
END DO
FOR each client DO
  CRI = Multiply(ALL attribute_CRI's)
END DO
```

to bad loans and potentially could decrease the true positive rate, but it will also decrease the false-negative rate dramatically. It could be interesting for loan providers, especially when they want to offer a loan to clients without any credit history and only based on Big Data analysis. In this case, it is beneficial if we can detect the separated sections of good and bad customers and struggle to detect good customers from the muddy intersection of good and bad loans in the dataset, where there is high similarity in attributes of different categories. Additionally, using the CRI as an aggregate of features with an interval/ratio scale will significantly decrease the required computation for modeling.

4. Experimental design

In this section we illustrate the use of our online credit scoring method, using two public loan datasets. One, "German Credit Data" which is a small dataset with 6377 observations and another "Lending club loan data" dataset with more than two million observations (2,260,668). It would be suitable to compare the performance of this credit scoring methodology in different situations.

Without loss of generality, we assume for simplicity that tt equals one and compute $\gamma_j^1\gamma_j^1$ and $w_j^1w_j^1$. In this paper, we use receiver operating curves (ROC) to show the statistical performance of the models. In the ROC chart, the horizontal axis represents the specificity, and the vertical axis shows the sensitivity. The greater the area between the curve and the baseline, the better the feature performance in default prediction. After investigating the characteristics of the new credit score model, we employed the area ratio of ROC curves to compare the classification accuracy and evaluate how well this credit scoring model performs. The data sets will randomly be divided into two groups, 65% for model training and the other 35% to apply different algorithms to the novel credit scoring methodology.

4.1. Small data set

4.1.1. Data description

We obtained the data from 'German Credit Data' and removed unnecessary features from it. We consider the following 13 explanatory variables. "V1: Seniority" for Job seniority (year), "V2: Home" for type of homeownership, "V3: Time" for time of requested loan, "V4: Age" for client age, "V5: Marital" for marital status, "V6: Records" for existence of records, "V7: Job" for type of job, "V8: Expenses" for amount of expenses, "V9: Income" for amount of income, "V10: Assets" for amount of assets, "V11: Debt" for amount of debt, "V12: Amount" for amount requested of loan and "V13: Price" for price of goods. Among the total 6377 observations, 2217 (34.8%) are loan requests with default payment according to the Basel accords definition, which have three or more late payments that imply default. The response variable is a binary variable named "Status," which represents loan default (No = 1, Yes = 2). The dataset is anonymized, and does not contain personal information. Descriptive statistics of attributes are presented in Appendix 1.

4.1.2. Results of small data-set

The results are organized in two parts starting with the Kruskal-Wallis statistics, $w_k^1w_k^1$ and $\varphi_i^1\varphi_i^1$ calculation to establish the artificial neural network model based on propose methodology. Subsequently, the results are detailed, first in terms of model comparison and then in terms of computation performance.

The $\varphi\varphi$ gives positive values to statistically significant features based on their detection power of default and gives others zero value, which means excluding those features from the model. The $\varphi_i^1\varphi_i^1$ based on $w_k^1w_k^1$ is shown below (Table 3). In this dataset, all features are statistically significant.

From Table 4, The Kruskal-Wallis test for credit risk index (CRI) resulted in a p-value of less than 0.0001, which means that CRI is significantly able to differentiate the categories of good and default loans.

First, we use Logistic regression, which is a simple classifier and performs very well for credit scoring as a benchmark. It can produce a probabilistic estimation of the binary response variable, and it is a prevalent method for credit scoring. Thus, model A1 in Table 5 shows the results of Logistic regression for the original variables. Similarly, the artificial neural network for the original variables is represented by A2 in Table 5. We use these two models to estimate the improvement of predictive accuracy and the performance of the classification of proposed models, shown in the same Table by A3 and C.

Therefore, model A3 is built with the weighted explanatory variables (see Section 4.1.1), as well as model A4, with a combination of CRI and nominal variables. We propose CRI in model A4 as a candidate for

Table 3
K-W, $w_k^1 w_k^1$ and.$\varphi_i^1 \varphi_i^1$.

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13
K-W	1361	1191	1795	1539	1103	131	1923	556	983	136	669	419	1175
w_k^1	0.994	1	0.995	0.995	1	1	1	0.986	0.992	0.945	0.988	0.981	0.993
φ_i^1	0.0772	0.0776	0.0773	0.0772	0.0776	0.0776	0.0776	0.0766	0.0770	0.0734	0.0767	0.0762	0.0771

Table 4
Credit score validation.

CREDIT RISK INDEX	N	MEAN	MEDIAN	MEAN RANK	K-W
GOOD LOANS	4160	1.64	1.28	3301.09	44.35
DEFAULT LOANS	2217	1.27	0.29	2978.67	P-VALUE = 0.0

representing variables with a ratio scale, that is, the reduction in the features dimension and efficient computation resource management. Models A3 and A4 show the main effects of each transformation on variables in accuracy and computation efficiency of the proposed credit scoring algorithms and models A1 and A2 combine unchanged variables to categorize loans into groups of default and non-default to show how much the new models are able to improve the accuracy and performance of classification.

As is common practice, in credit scoring, statistical model performance is measured by the area under the receiver operating curve (AUC), and it is represented in Tables 5 and 6.

In Table 6, the model A3 includes all variables belonging to bad loans to the power of $w_k^1 w_k^1$ except for the nominal features. We consider all nominal variables and CRI in model A4 as features of the model. As is evident in Table 6, model A4 did not provide more significant results than model A3 except for less computation. It could be important to deal with non-traditional datasets such as mobile data and Big Data sources of credit scoring.

In this credit scoring case, error rates are not the appropriate criteria to evaluate the performance of the credit score model because most clients are classified into creditable customers (93.6%). From Tables 9 and 10, it is clear that there is not a significant difference in the performance of the same models. However, by considering the area under the ROC curve as an essential factor of credit risk cost, models A1 and A2 perform the worst, of which models including the proposed methodology (models A3, A4) perform the best, not only in accuracy but also in computational efficiency. To offer a loan based on non-traditional data analysis, the benefit of correctly identifying a defaulter plays a prominent role, and it is interesting to see that having only CRI and the pre-calculation of weighted features on the bad loans section of the data-set allows discriminating potentially better clients.

Furthermore, by using the area ratio in the test data, the classification result shows almost the same performance of models A3 and A4; however model A4 yields a better performance in computation time that is a critical factor in the performance of parallel and distributed computing for non-traditional datasets.

4.2. Big Data set

4.2.1. Data description

Lending club loan data contains complete loan data for all loans through 2007–2018, each loan includes applicant information provided by the applicant as well as the current loan status (Current, Late, Fully Paid, etc.) and latest payment information. We found two versions of this dataset; one contains loans issued through the 2007–2015 and another version through 2012–2018. As a result, we combined these two datasets and removed the duplicates to obtain a complete dataset from 2007 to 2018 with maximum possible cases.

4.2.2. Aplication data

We consider the following application data as the following numeric attributes.

The "loan_amnt" for the listed amount of the loan applied for by the borrower with any possible reductions in the loan amount with the credit department by the time. The "emp_length" for employment length in years with possible values between 0 and 10, where 0 means less than one year and 10 means ten or more years. In the original dataset, the employment length is a combination of numbers, characters, plus signs that are converted to the numbers by the codes available in Appendix 2. The "annual_inc", for the self-reported annual income provided by the borrower during registration, "dti" as a ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income. The "delinq_2yrs" for the number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years, the "revol_util" for revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit, the "total_acc" for the total number of credit lines currently in the borrower's credit file, the "int_rate" for the interest rate on the loan, and finally, one categorical attribute "term_month", which stands for the number of payments on the loan with values in months, and can be either 36 or 60. We changed all the numeric attributes to float and years to double format and all rates in percentage in PySpark as in Appendix 2.

Table 5
Classification table and statistical model performance (AUC) for original variables.

ORIGINAL VARIABLES		PREDICTED NEGATIVE	PREDICTED POSITIVE	PERCENT CORRECT	AUC	TIME (MILLISECOND)
A1: LOGISTIC REGRESSION	**TRAINING NEGATIVE**	2757	161	94.5%	0.925	–
	POSITIVE	403	1120	73.5%		
	OVERALL PERCENT	71.1%	28.8%	87.3%		
	HOLDOUT NEGATIVE	1159	83	93.3%		
	POSITIVE	194	500	70.1%		
	OVERALL PERCENT	69.9%	30.1%	85.69%		
A2: NEURAL NETWORK	**TRAINING NEGATIVE**	2540	174	93.6%	0.926	840
	POSITIVE	401	938	70.1%		
	OVERALL PERCENT	72.6%	27.4%	85.8%		
	HOLDOUT NEGATIVE	1370	76	94.7%		
	POSITIVE	197	681	77.6%		
	OVERALL PERCENT	67.4%	32.6%	88.3%		

Table 6
Classification table and statistical model performance (AUC) for weighted variables and CRI.

ARTIFICIAL NEURAL NETWORK		PREDICTED NEGATIVE	PREDICTED POSITIVE	PERCENT CORRECT	AUC	TIME (MILLISECOND)
A3: WIEGHTED VARIABLES	**TRAINING NEGATIVE**	2715	0	100%	0.999	320
	POSITIVE	0	1339	100%		
	OVERALL PERCENT	67.0%	33.0%	100%		
	HOLDOUT NEGATIVE	514	932	35.5%		
	POSITIVE	2	876	99.8%		
	OVERALL PERCENT	22.2%	77.8%	59.8%		
A4: CRI AND CATEGORICAL VARIABLES	**TRAINING NEGATIVE**	2714	0	100%	0.999	80
	POSITIVE	0	1339	100%		
	OVERALL PERCENT	67.0%	33.0%	100%		
	HOLDOUT NEGATIVE	504	942	34.9%		
	POSITIVE	1	877	99.9%		
	OVERALL PERCENT	21.7%	78.3%	59.4%		

4.2.3. Behavioral data

The selected behavioral data are all categorical, and we convert them into dummy variables in PySpark to be able to contribute to the modeling stage. These attributes are “home_ownership” which is the homeownership status provided by the borrower during registration or obtained from the credit report with values: RENT, OWN, MORTGAGE, and OTHER. “Purpose” as a category provided by the borrower for the loan request, “addr_state” for the state provided by the borrower in the loan application, “verification_status” for indicates if income was verified by LC, not verified, or if the income source was verified. We mapped multiple levels if verification_status attribute into the one-factor level as is shown in appendix PySpark codes. Finally, “application_type” which indicates whether the loan is an individual application or a joint application with two co-borrowers.

The label variable is “default_loan” with TRUE value (code 1) for default loans with values of “Default”, “Charged Off”, “Late (31–120 days)”, “Late (16–30 days)”, and FALSE (code 0) for non-default loans for “Fully Paid” loans.

4.2.4. New measures

We created two new measures to be considered in credit scoring models. For length of credit in years, “credit_length_in_years” is computed by subtracting the issue year from the earliest year. The issue year is extracted from the issue date, and the earliest year is also substring of the date that the borrower’s earliest reported credit line was opened. Additionally, we want to know the fraction of the initial loan amount that has been reimbursed and to evaluate the loan provider profit and loss according to model results. Therefore, we created a new column named “remain” by subtracting “loan payments” from the “total loan amount”. This will represent the amount of money earned or lost per loan and the outstanding loan balance.

4.2.5. Model and train and test datasets

With the datasets featurized, credit scoring models are built using binary classifiers with a k-fold cross-validation on a dataset with 1,048,575 observations, which contains 217,930 default loans (True or 1), and 830,645 good loans (False or 0). The method of leave-one-out cross-validation is used to examine the between-sample variation of default prediction. The available data are divided into 10 disjoint subsets, and the models are trained on 9 of these subsets and the models selection criterion evaluated on the unused subset. This procedure is then repeated for all combinations of subsets by Python API of Apache Spark, as is presented in Appendix 2. Leave-one-out cross-validation helps the algorithm to use all data as both training and validation, and consider the mean of the model selection criterion computed over the unused subset in each fold for better accuracy estimation.

Furthermore, Logistic regression works well for many business applications, which often have a simple decision boundary. Moreover, because of its simplicity, it is less prone to overfitting than flexible methods such as decision trees. Further, as we will show, variables that contribute to overfitting might be eliminated using Lasso or Ridge regularisation, without compromising out-of-sample accuracy. In this case, the Ridge method presented better performance than Lasso, and the following Logistic regressions are based on the Ridge penalty with elastic net regularization zero and regparam 0.3 as the best hyperparameters. In addition to the Logistic regression classifier as an industry standard for building credit scoring models, other binary classifiers such as random forests and linear support vector machines are used for the empirical analysis. Although they are more complex and powerful than Logistic regression in an application, the interpretability of these models could not be guaranteed as well as Logistic regression outputs.

4.2.6. Results of big data-set

The results of the Kruskal-Wallis statistics, w_k^1 and φ_i^1 is presented in Table 7, based on the propose methodology and the CRI is calculated and checked for performance as is shown in Fig. 4. As we can see, all numerical attributes have the K-W statistic greater than $\chi^2_{1,0.05}$ and they are statistically significant. Therefore, we consider all of them for the next step, otherwise, the program specifies the power of zero for insignificant attributes, that is not the case here.

We start with the default flag rows of the training set and calculate the new numerical attribute k with powering to w_k^1. For non-default loans, we consider the same value in the original attribute for the new one and impute the missing values with the average. For computing the CRI, we have to impute the zero values in each numeric attribute with average to avoid a null result when multiplying by the CRI. By using the average, we nullify the effects of those specific zero values and extract the information of the other attributes in the benefit of CRI. Moreover, as we discussed in Section 3, we need to hold out an unseen portion of the dataset to apply blindly the transformation to all attributes of all clients, including good and bad loans, and use it as the main part of our proposed algorithm. The model will apply this transformation for every new customer after deploying. As we do not have the label for this test set, therefore we apply the mentioned transformations to all attributes and create the new features. The test dataset featurized obtains 223,722 observations.

Now, the new attributes in the train and test datasets are ready to compute the CRI based on Formula 12. We need to calculate the average of each attribute for the good loans in the training dataset and use φ_i^1 from the Table 7.

From Fig. 4, it is clear that CRI shows an appropriate performance to

Table 7
K-W, $w_k^1 w_k^1$ and.$\varphi_i^1 \varphi_i^1$.

	Term months	Loan amnt	Emp length	Annual inc	dti	Delinq 2yrs	Revol util	Total acc	Credit length in years	Int rate	Remain
K-W	39,784	4870	134	4718	13,055	342	2760	63	1685	69,573	449,682
w_k^1	1	0.998	0.944	0.998	0.999	0.978	0.997	0.887	0.995	1	1
φ_i^1	0.0926	0.0924	0.0874	0.0924	0.0925	0.0906	0.0923	0.0822	0.0922	0.0926	0.0926

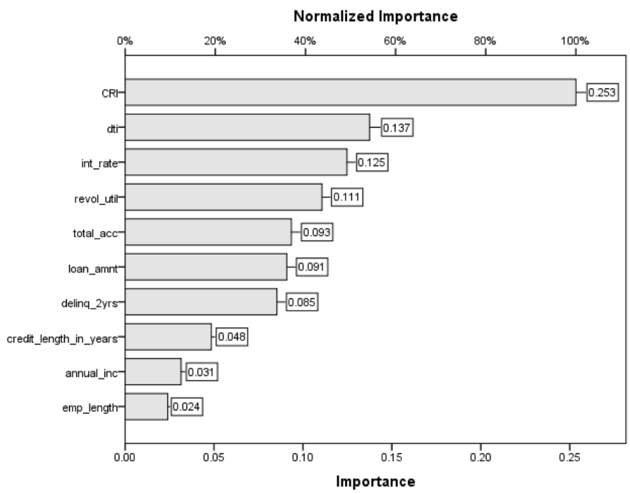


Fig. 4. Normalized importance of attributes in modeling stage.

differentiate the categories of good and bad loans and it could be considered as a candidate to contribute to the modeling phase for improving the accuracy. The resulted dataset consists of categorical and numerical attributes, for both original and transformed values, label variable for the train dataset, and an indicator variable to divide the dataset into test and train. As we discussed, the selected algorithms will be implemented using recursive partitioning with ten-fold cross-validation on the featurized training set to tune the models. The area under the ROC curve (AUC) and recall/sensitivity are computed to evaluate the model's performance in each scenario. The loss amount is also calculated in each scenario to evaluate the model's ability to reduce credit losses. Subsequently, the results of the proposed methodology are detailed, first for each algorithm in Tables 8–10, then a summary is represented in Table 11 for both statistical and financial performance.

We start with the Logistic regression with the Ridge penalty on the original dataset as a benchmark. The model B1 in Table 8, represents the results of Logistic regression for the original variables to investigate the possible improvements in the performance of the classification of proposed models, shown in the same Table by B2 and B3. The run time of all three models are very close and around 6.6 min by PySpark.

This dataset is representing a very high-risk scenario with a high false-negative rate. This situation is riskier in comparison with the case of the small dataset in the previous section. The amount of loss shows the benefit of correctly identifying a defaulter by the proposed algorithm, and the model with weighted attributes and CRI allows discriminating the bad loans and minimizing the loss. The Logistic regression model with new features could migrate the delinquent customers to the reject area and reduce the loan delinquency rate and subsequently the loss amount. As a trade-off between model sensitivity and specificity, AUC shows almost the same performance among the various scenarios. However, our new Logistic regression obtained a higher sensitivity (i.e. 0.92) in comparison with the normal model with a sensitivity of 0.73. Overall, the model B1 performs the worst, of which B2 and B3 including phi and CRI features perform best.

For the second algorithm, we use random forests classifier as an ensemble method. We consider 3, 5, and 10 decision trees to construct each forest and jointly decide upon the credit score. The model B4 in Table 9, represents the results of random forests classifier for the original variables to investigate the possible improvements in the performance of the classification of proposed models, shown in the same Table by B5 and B6. The run time of all three models are very close and around 30 min by PySpark.

Table 9 demonstrates that there is a significant difference in the performance of the three scenarios. The normal random forests model is not able to predict any unseen sample from samples it has seen during training and it was not able to model the distribution of loan situations, whereas the model B6, which is a random forests classifier with a combination of phi features and CRI, perform significantly better. This shows the usefulness of the new algorithm to improve the performance of the models. Moreover, it successfully maximize the profit and minimize the operational costs and risks as another benefit of this new approach by focusing on correctly identifying a defaulter.

The Linear support vector machine (SVM) algorithm is considered as another sophisticated supervised machine learning technique with expected higher accuracy than Logistic regression. As our data set is large, SVM has high training time (1.05 h) compare to other algorithms. The results are very similar to the Random forest, however, the amount of loss is dramatically higher than the two last techniques. Regardless of improvement in the performance of the model, with the help of our proposed algorithm, the amount of loss is not improved as much as other classifiers.

Finally, Table 11 represents the best model of each classifier, and again, the new Logistic regression B3 produces the best-performing model. It has also high interpretability and has been applying for credit scoring as an industry standard for many years. Model B3 has the highest profit, followed by models B6 and B8. It is interesting to see that having new features in all of these models produce decent profits, whereas the Normal models of the same algorithms do not, at least not when compared to the proposed algorithm.

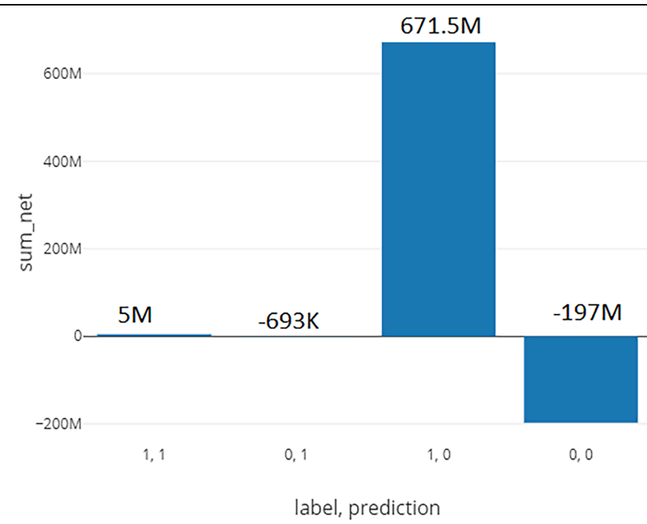
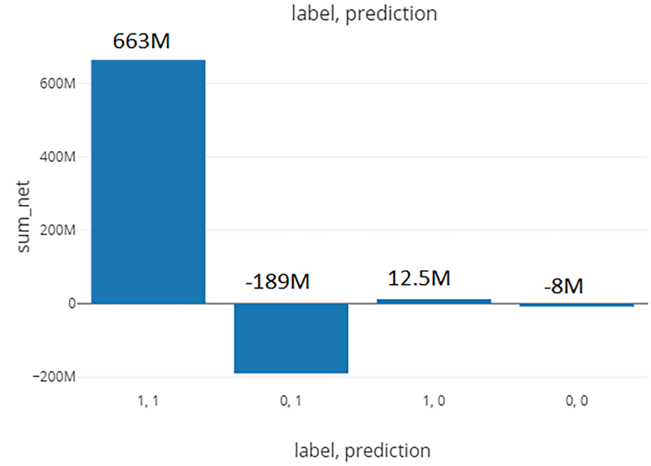
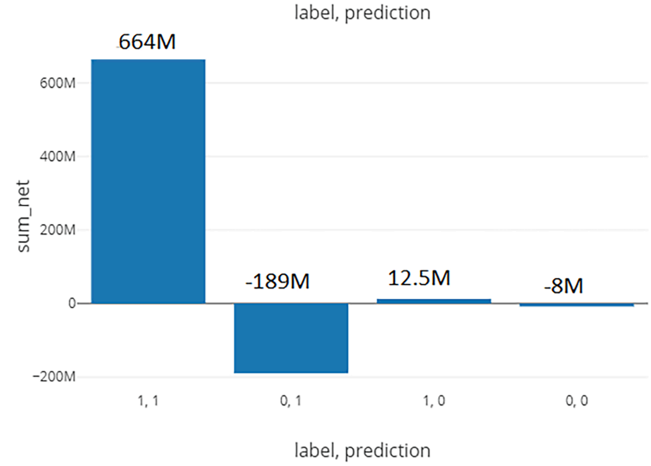
5. Impact of research

This section identifies various levels of impact based on the research findings.

5.1. Confidentiality and privacy

Transferring the sensitive data from data warehouses of financial institutions to different machines and nodes for parallel or distributed computation is always affected by privacy concerns. Financial service providers try to enhance trust in their systems as a fundamental policy of client rights and maintaining the confidentiality of personally identifiable information is crucial. Additionally, there are some standards and regulations, such as the General Data Protection Regulation (GDPR) in the European Union. The result of this study shows that an index of features can be calculated as an aggregation before data distributing for the mapping stage of the MapReduce algorithm. There is an ethical concern in data anonymization as well because of outliers, which mostly belong to well-known special customers. This indexing can guarantee the confidentiality of sensitive data to provide easier access to parallel computing tasks.

Table 8
Classification table for Logistic Regression with Ridge Penalty. In the table 0, 1 indicate non-default loan and default loan/payment arriars, respectively, and LOSS stands for the substraction of net remain of (0,0) from (1,0) combinations.

LOGISTIC REGRESSION WITH THE RIDGE PENALTY						
		PRED. NEGATIVE	PRED. POSITIVE	% CORRECT	METRICS	SUM NET (REMAIN)
B1: NORMAL MODEL	**NEG**	162,516	245	99.85%	AUC = 0.69	
	POS	60,649	312	0.51%	RECALL = 0.73LOSS = (474.5) M	
	OV. %	99.75%	0.25%	72.78%		
B2: PHI MODEL	**NEG**	12,296	150,465	7.55%	AUC = 0.67	
	POS	1124	59,837	98.16%	RECALL = 0.92LOSS = (4.5) M	
	OV. %	6.00%	94.00%	32.24%		
B3: PHI PLUS CRI MODEL	**NEG**	12,530	150,231	7.70%	AUC = 0.67	
	POS	1126	59,835	98.15%	RECALL = 0.92LOSS = (4.5) M	
	OV. %	6.10%	93.90%	32.35%		

5.2. Financial inclusion

Models based on non-traditional data sources such as mobile phone data in the form of call detail records or mobile phone log files, which are examples of Big Data sources, typically suffer from complexity and time-consuming sophisticated algorithms. Despite facilitating credit access to people without historical financial data, the models should be highly accurate to fulfill the expectations of loan providers. Using conservative models for these new sources of data or high-risk situations such as pandemics or economic crisis can help loan providers offer even small credits to underbanked populations, young people, patients, and immigrants, enhancing the assessment of whether the new clients are creditworthy.

Table 9
Classification table for Random Forest Classifier. In the table 0, 1 indicate non-default loan and default loan/payment arriars, respectively, and LOSS stands for the substraction of net remain of (0,0) from (1,0) combinations.

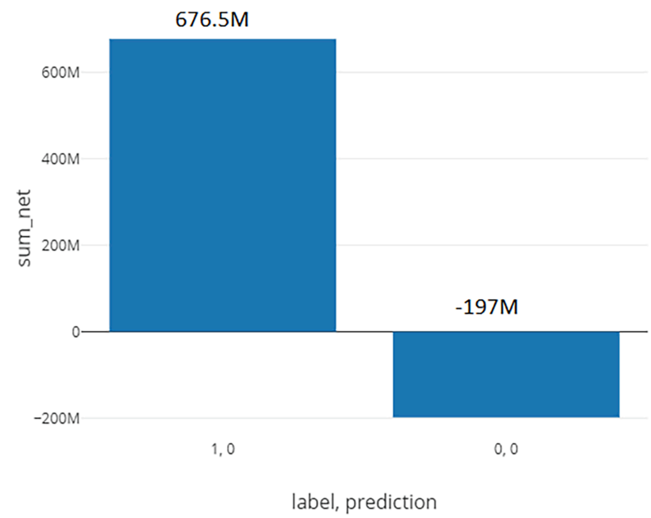
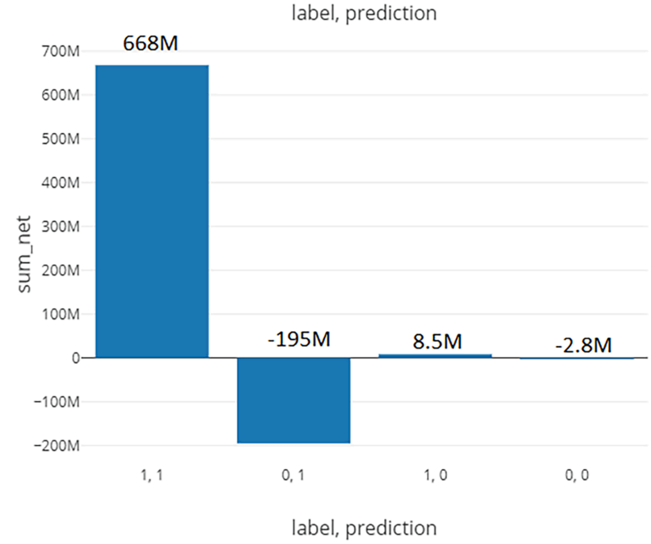
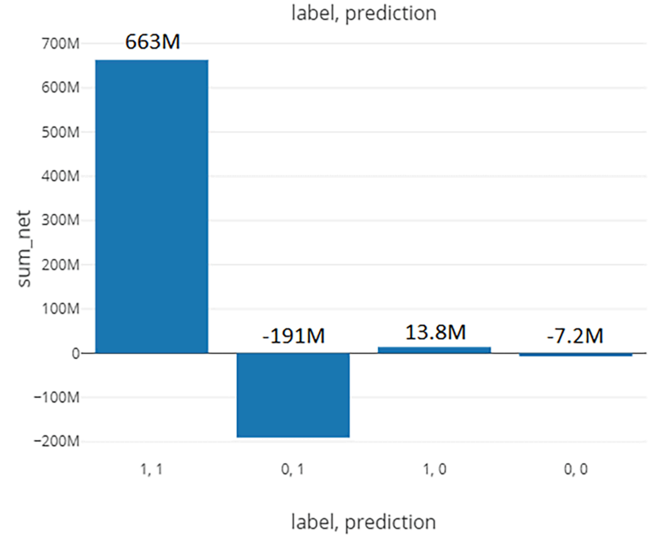
RANDOM FORESTS CLASSIFIER		PRED. NEGATIVE	PRED. POSITIVE	% CORRECT	METRICS	SUM NET (REMAIN)
B4: NORMAL MODEL	**NEG**	162,761	0	100%	AUC = 0.61 RECALL = 0.72LOSS = (479) M	
	POS	60,961	0	0%		
	OV. %	100%	0%	72.75%		
B5: PHI MODEL	**NEG**	1999	160,762	1.23%	AUC = 0.48 RECALL = 0.75LOSS = (5.7) M	
	POS	666	60,295	98.91%		
	OV. %	1.19%	98.81%	27.84%		
B6: PHI PLUS CRI MODEL	**NEG**	6841	155,920	4.20%	AUC = 0.62 RECALL = 0.87LOSS = (6.6) M	
	POS	1002	59,959	98.36%		
	OV. %	3.51%	96.49%	29.86%		

Table 10
Classification table for linear support vector machine In the table 0, 1 indicate non-default loan and default loan/payment arriars, respectively, and LOSS stands for the substraction of net remain of (0,0) from (1,0) combinations.

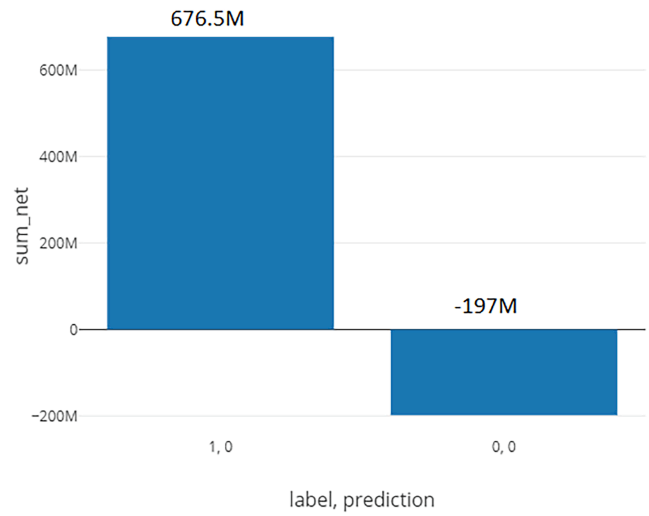
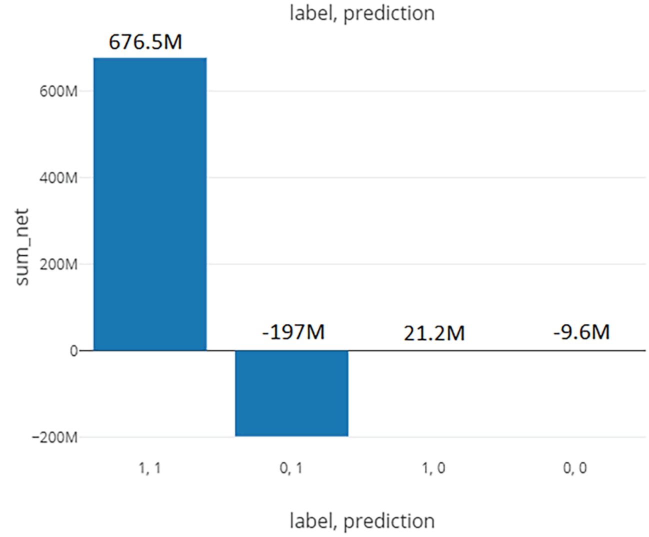
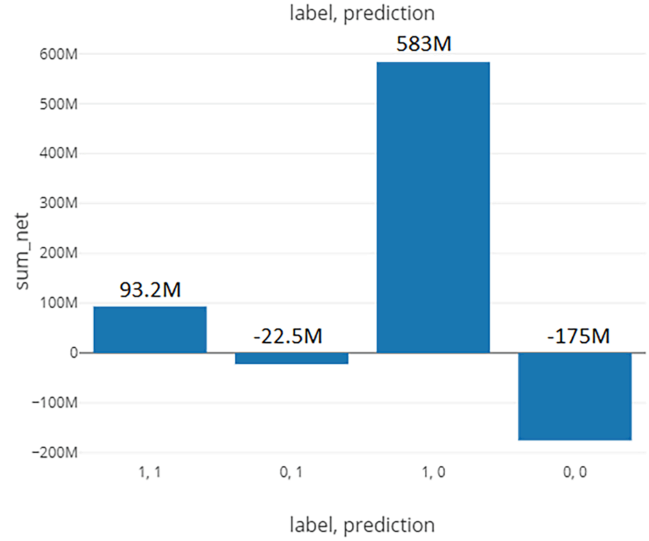
LINEAR SUPPORT VECTOR MACHINE						
		PRED. NEGATIVE	PRED. POSITIVE	% CORRECT	METRICS	SUM NET (REMAIN)
B7: NORMAL MODEL	**NEG**	162,761	0	100.00%	AUC = 0.50 RECALL = 0.72LOSS = (479.5) M	
	POS	60,961	0	0.00%		
	OV. %	100.00%	0.00%	72.75%		
B8: PHI MODEL	**NEG**	11	162,750	0.01%	AUC = 0.50 RECALL = 0.91LOSS = (11.6) M	
	POS	1	60,960	100.00%		
	OV. %	0.01%	99.99%	27.25%		
B9: PHI PLUS CRI MODEL	**NEG**	149,196	13,565	91.67%	AUC = 0.52 RECALL = 0.73LOSS = (408) M	
	POS	54,141	6820	11.19%		
	OV. %	90.89%	9.11%	69.74%		

Table 11
The best model of each classifier.

Classifier	Model ID	Feature group	AUC	Recall/ Sensitivity	Loss
Logistic regression	B3	PHI, CRI	0.67	0.92	4.5 M
Random forests	B6	PHI, CRI	0.62	0.87	6.6 M
Support vector machine	B8	PHI	0.50	0.91	11.6 M

5.3. Compliance risk impact

Committee on payment and settlement systems in key consideration 3-4-7 of principles for financial market infrastructures[3] and its explanatory notes declares that the financial systems should have clearly defined procedures for the management of credit and liquidity risks. It should specify the respective responsibilities of the system operator and the participants and provide appropriate incentives to manage and contain those risks. Credit-scoring models should also be considered in provisions and capital buffers calculation by financial institutions according to standards such as the Basel Accords and IFRS9. In this research, we illustrated how credit scoring has to be conservatively formulated to propagate in non-traditional datasets with the potential high-risk of the false-negative rate to detect default. This insight paves the way for loan providers to be able to use new sources of data more soundly and solidly and try to adapt to new emerged technologies without risk management concerns.

6. Conclusion

This study described a non-parametric statistics approach to assess credit candidate applicants' profiles and continued credit scoring based on non-traditional data. The approach uses a two-step approach based on an initial Kruskal-Wallis analysis and a neural network to learn the model. It introduced a novel credit scoring methodology that reselects significant highly informative features and weighted out by their level of contribution in predicting credit categories of loans to be used in modeling phase. This new credit scoring uses the Kruskal-Wallis non-parametric test, which enables it to be used for two or more categories. Therefore, categories could be "good loans" and "default loans" or even more than two categories such as "good," "doubtful," and "bad" loans as recommended by Basel Accords. The proposed credit risk index is computationally less expensive with reasonable accuracy in comparison with current computationally expensive hybrid algorithms in credit scoring or fixed-weight models in scorecards. The advantages of this approach could be summarized as following:

- Occupying less memory and transferring optimized data in the network by using only the sign of $\gamma_j^t \gamma_j^t$ as a flag attribute.
- Having a warning sign to renew the model by a new set of features based on the values and trend of $w_j^t w_j^t$.
- Improving the performance of the model and decreasing the false-negative rate.
- Using the CRI as a complementary feature with an interval/ratio scale.

In the classification accuracy, the results showed that this credit scoring method is more informative and conservative. It is able to predict default probability showing good performance with AUC = 0.99 for small dataset and unchanged AUC = 0.67 for big dataset with 18% improvement in Recall and Sensitivity. Thus, this credit index formula as a homogeneous symmetric average is an accurate aggregate measure, able to renew features dynamically and weighted out the attributes as their impact factors. It is suitable for traditional and non-traditional data sets such as regular loan data repositories or new mobile phone datasets, especially where selecting and extracting the information of features in one aggregated measure is needed for online credit scoring.

[3] An FMI should establish explicit rules and procedures that fully address any credit losses it may face as a result of any individual or combined default among its participants with respect to any of their obligations to the FMI (2012).

CRediT authorship contribution statement

Afshin Ashofteh: Conceptualization, Methodology, Software, Formal analysis, Writing - original draft, Writing - review & editing. **Jorge M. Bravo:** Conceptualization, Methodology, Resources, Validation, Writing - original draft, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.eswa.2021.114835.

References

Abellán, J., & Castellano, J. G. (2017). A comparative study on base classifiers in ensemble methods for credit scoring. *Expert Systems with Applications, 73*, 1–10. https://doi.org/10.1016/j.eswa.2016.12.020

Acosta-González, E., & Fernández-Rodríguez, F. (2014). Forecasting financial failure of firms via genetic algorithms. *Computational Economics, 43*(2), 133–157. https://doi.org/10.1007/s10614-013-9392-9

Altman, E. I. (1968). financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance, 23*(4), 589. https://doi.org/10.2307/2978933

Altman, E. I., Haldeman, R. G., & Narayanan, P. (1977). ZETATM analysis A new model to identify bankruptcy risk of corporations. *Journal of Banking and Finance, 1*(1), 29–54. https://doi.org/10.1016/0378-4266(77)90017-6

Arminger, G., Enache, D., & Bonne, T. (1997). Analyzing credit risk data: A comparison of logistic discrimination, classification tree analysis, and feedforward networks. *Computational Statistics, 12*(2), 293–310. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4801.

Avery, R. B., Bostic, R. W., Calem, P. S., & Canner, G. B. (2000). Credit scoring: Statistical issues and evidence from credit-bureau files. *Real Estate Economics, 28*(3), 523–547. https://doi.org/10.1111/1540-6229.00811

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society, 54*(6), 627–635. https://doi.org/10.1057/palgrave.jors.2601545

Baesens, Bart, Roesch, D., & Scheule, H. (2016). Credit Risk Analytics: Measurement Techniques, Applications, and Examples in SAS - Bart Baesens, Daniel Roesch, Harald Scheule - Google Books. Retrieved from <https://books.google.com/books?hl=en&lr=&id=ornsDAAAQBAJ&oi=fnd&pg=PR11&dq=Credit+risk+analytics:+Measurement+techniques,+applications,+and+examples+in+SAS&ots=NHbvLqOdKi&sig=RshAgweYNZ8JI0Monkk137s8Y60>.

Bellotti, T., & Crook, J. (2013). Forecasting and stress testing credit card default using dynamic models. *International Journal of Forecasting, 29*(4), 563–574. https://doi.org/10.1016/j.ijforecast.2013.04.003

Bjorkegren, D., & Grissen, D. (2018). Behavior Revealed in Mobile Phone Usage Predicts Loan Repayment. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2611775

Butaru, F., Chen, Q., Clark, B., Das, S., Lo, A. W., & Siddique, A. (2016). Risk and risk management in the credit card industry. *Journal of Banking and Finance, 72*, 218–239. https://doi.org/10.1016/j.jbankfin.2016.07.015

Castro, V. (2013). Macroeconomic determinants of the credit risk in the banking system: The case of the GIPSI. *Economic Modelling, 31*(1), 672–683. https://doi.org/10.1016/j.econmod.2013.01.027

Chamboko, R., & Bravo, J. M. (2016). On the modelling of prognosis from delinquency to normal performance on retail consumer loans. *Risk Management, 18*(4), 264–287. https://doi.org/10.1057/s41283-016-0006-4

Chamboko, R., & Bravo, J. M. (2019a). Frailty correlated default on retail consumer loans in Zimbabwe. *International Journal of Applied Decision Sciences, 12*(3), 257–270. https://doi.org/10.1504/IJADS.2019.100436

Chamboko, R., & Bravo, J. M. V. (2019b). Modelling and forecasting recurrent recovery events on consumer loans. *International Journal of Applied Decision Sciences, 12*(3), 271–287. https://doi.org/10.1504/IJADS.2019.100440

Chamboko, R., & Bravo, J. M. (2020). A multi-state approach to modelling intermediate events and multiple mortgage loan outcomes. *Risks, 8*(2), 64. https://doi.org/10.3390/risks8020064

Chittaranjan, G., Blom, J., & Gatica-Perez, D. (2013). Mining large-scale smartphone data for personality studies. In *Personal and Ubiquitous Computing* (pp. 433–450). Springer. https://doi.org/10.1007/s00779-011-0490-1.

Cleofas-Sánchez, L., García, V., Marqués, A. I., & Sánchez, J. S. (2016). Financial distress prediction using the hybrid associative memory with translation. *Applied Soft Computing Journal, 44*, 144–152. https://doi.org/10.1016/j.asoc.2016.04.005

Deng, Y., Quigley, J. M., & Order, R. (2000). Mortgage terminations, heterogeneity and the exercise of mortgage options. *Econometrica, 68*(2), 275–307. https://doi.org/10.1111/1468-0262.00110

Do, T.-M.-T., & Gatica-Perez, D. (2010). By their apps you shall understand them: mining large-scale patterns of mobile phone usage. *Telecommunications Policy, 24*(ii), 27: 1–27:10. https://doi.org/10.1145/1899475.1899502

Douzas, G., & Bacao, F. (2018). Effective data generation for imbalanced learning using Conditional Generative Adversarial Networks Improving Artificial Data Generation View project Factors impacting on mobile payment adoption on user perspective View project Georgios Douzas Effective d. *Expert Systems With Applications, 91*, 464–471. https://doi.org/10.1016/j.eswa.2017.09.030

Erdogan, B. E. (2013). Prediction of bankruptcy using support vector machines: An application to bank bankruptcy. *Journal of Statistical Computation and Simulation, 83* (8), 1543–1555. https://doi.org/10.1080/00949655.2012.666550

Feng, X., Xiao, Z., Zhong, B., Qiu, J., & Dong, Y. (2018). Dynamic ensemble classification for credit scoring using soft probability. *Applied Soft Computing Journal, 65*, 139–151. https://doi.org/10.1016/j.asoc.2018.01.021

Fensterstock, A. (2005). Credit scoring and the next step. *Business Credit, 107*, 46.

García, S., Fernández, A., Luengo, J., & Herrera, F. (2010). Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. *Information Sciences, 180*(10), 2044–2064. https://doi.org/10.1016/j.ins.2009.12.010

García, V., Marqués, A. I., & Sánchez, J. S. (2012). On the use of data filtering techniques for credit risk prediction with instance-based models. *Expert Systems with Applications, 39*(18), 13267–13276. https://doi.org/10.1016/j.eswa.2012.05.075

García, V., Marqués, A. I., & Sánchez, J. S. (2019a). Exploring the synergetic effects of sample types on the performance of ensembles for credit risk and corporate bankruptcy prediction. *Information Fusion, 47*, 88–101. https://doi.org/10.1016/j.inffus.2018.07.004

García, V., Marqués, A. I., Sánchez, J. S., & Ochoa-Domínguez, H. J. (2019b). Dissimilarity-based linear models for corporate Bankruptcy prediction. *Computational Economics, 53*(3), 1019–1031. https://doi.org/10.1007/s10614-017-9783-4

Gerardi, K., Shapiro, A., & Willen, P. (2007). Subprime outcomes: Risky mortgages, homeownership and foreclosure.

Gicić, A., & Subasi, A. (2019). Credit scoring for a microcredit data set using the synthetic minority oversampling technique and ensemble classifiers. *Expert Systems, 36*(2), e12363. https://doi.org/10.1111/exsy.v36.210.1111/exsy.12363

Ha, S. H. (2010). Behavioral assessment of recoverable credit of retailer's customers. *Information Sciences, 180*(19), 3703–3717. https://doi.org/10.1016/j.ins.2010.06.012

Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society. Series A: Statistics in Society, 160*(3), 523–541. https://doi.org/10.1111/j.1467-985X.1997.00078.x

Ho Ha, S., & Krishnan, R. (2012). Predicting repayment of the credit card debt. *Computers and Operations Research, 39*(4), 765–773. https://doi.org/10.1016/j.cor.2010.10.032

Jacobson, T., & Roszbach, K. (2003). Bank lending policy, credit scoring and value-at-risk. *Journal of Banking and Finance, 27*(4), 615–633. https://doi.org/10.1016/S0378-4266(01)00254-0

Jones, S., & Hensher, D. A. (2004). In *Predicting firm financial distress: A mixed logit model. Accounting Review*. American Accounting Association. https://doi.org/10.2308/accr.2004.79.4.1011.

Jordan, M. I. (2013). On statistics, computation and scalability. *Bernoulli, 19*(4), 1378–1390. https://doi.org/10.3150/12-BEJSP17

Kellison, B., & Wortham, G. (2003). Bureau of Business Research • McCombs School of Business • The University of Texas at Austin SPECIAL ISSUE. Retrieved from <https://repositories.lib.utexas.edu/handle/2152/15187>.

Kleiner, A., Talwalkar, A., Sarkar, P., & Jordan, M. I. (2014). A scalable bootstrap for massive data. *Journal of the Royal Statistical Society. Series B: Statistical Methodology, 76*(4), 795–816. https://doi.org/10.1111/rssb:12050

Kruppa, J., Schwarz, A., Arminger, G., & Ziegler, A. (2013). Consumer credit risk: Individual probability estimates using machine learning. *Expert Systems with Applications, 40*(13), 5125–5131. https://doi.org/10.1016/j.eswa.2013.03.019

Kwak, W., Shi, Y., & Kou, G. (2012). Bankruptcy prediction for Korean firms after the 1997 financial crisis: Using a multiple criteria linear programming data mining approach. *Review of Quantitative Finance and Accounting, 38*(4), 441–453. https://doi.org/10.1007/s11156-011-0238-z

Lensberg, T., Eilifsen, A., & McKee, T. E. (2006). Bankruptcy theory development and classification via genetic programming. *European Journal of Operational Research, 169*(2), 677–697. https://doi.org/10.1016/j.ejor.2004.06.013

Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research, 247*(1), 124–136. https://doi.org/10.1016/j.ejor.2015.05.030

Li, Z., Tian, Y., Li, K., Zhou, F., & Yang, W. (2017). Reject inference in credit scoring using Semi-supervised Support Vector Machines. *Expert Systems with Applications, 74*, 105–114. https://doi.org/10.1016/j.eswa.2017.01.011

Liang, F., Cheng, Y., Song, Q., Park, J., & Yang, P. (2013). A resampling-based stochastic approximation method for analysis of large geostatistical data. *Journal of the American Statistical Association, 108*(501), 325–339. https://doi.org/10.1080/01621459.2012.746061

Liu, Z. F., & Pan, S. (2018). Fuzzy-Rough Instance Selection Combined with Effective Classifiers in Credit Scoring. *Neural Processing Letters, 47*(1), 193–202. https://doi.org/10.1007/s11063-017-9641-3

Luo, C., Wu, D., & Wu, D. (2017). A deep learning approach for credit scoring using credit default swaps. *Engineering Applications of Artificial Intelligence, 65*, 465–470. https://doi.org/10.1016/j.engappai.2016.12.002

Ma, P., Mahoney, M. W., Yu, B., & Yu Ma, B. (2015). A Statistical Perspective on Algorithmic Leveraging. *Journal of Machine Learning Research, 16*. https://doi.org/10.5555/2789272.2831141

Maclaurin, D., & Adams, R. P. (2015). Firefly Monte Carlo: Exact MCMC with subsets of data. In *IJCAI International Joint Conference on Artificial Intelligence* (pp. 4289–4295).

Maldonado, S., Peters, G., & Weber, R. (2020). Credit scoring using three-way decisions with probabilistic rough sets. *Information Sciences, 507*, 700–714. https://doi.org/10.1016/j.ins.2018.08.001

Noh, H. J., Roh, T. H., & Han, I. (2005). Prognostic personal credit risk model considering censored information. *Expert Systems with Applications, 28*(4), 753–762. https://doi.org/10.1016/j.eswa.2004.12.032

Nyitrai, T., & Virág, M. (2019). The effects of handling outliers on the performance of bankruptcy prediction models. *Socio-Economic Planning Sciences, 67*, 34–42. https://doi.org/10.1016/j.seps.2018.08.004

Óskarsdóttir, M., Bravo, C., Sarraute, C., Vanthienen, J., & Baesens, B. (2019). The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics. *Applied Soft Computing Journal, 74*, 26–39. https://doi.org/10.1016/j.asoc.2018.10.004

Pedro, J. S., Proserpio, D., & Oliver, N. (2015). Mobiscore: Towards universal credit scoring from mobile phone data. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (pp. 195–207). Springer Verlag. https://doi.org/10.1007/978-3-319-20267-9_16.

Pławiak, P., Abdar, M., & Rajendra Acharya, U. (2019). Application of new deep genetic cascade ensemble of SVM classifiers to predict the Australian credit scoring. *Applied Soft Computing Journal, 84*, 105740. https://doi.org/10.1016/j.asoc.2019.105740

Premachandra, I. M., Bhabra, G. S., & Sueyoshi, T. (2009). DEA as a tool for bankruptcy assessment: A comparative study with logistic regression technique. *European Journal of Operational Research, 193*(2), 412–424. https://doi.org/10.1016/j.ejor.2007.11.036

Quittner, J. (2003). Credit cards: Sub-prime's tech dilemma: With delinquencies and charge-offs on the rise, the industry examines the role of automated decisioning. *Bank Technology, 16*(1), 19–23.

Rimmer, J. (2005). Contemporary Changes in Credit Scoring. Credit Control, 26(4), 56–60. Retrieved from <http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=17602788&site=ehost-live>.

Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. Bioinformatics. Retrieved from <https://academic.oup.com/bioinformatics/article-abstract/23/19/2507/185254>.

Sarlija, N., Bensic, M., & Zekic-Susac, M. (2009). Comparison procedure of predicting the time to default in behavioural scoring. *Expert Systems with Applications, 36*(5), 8778–8788. https://doi.org/10.1016/j.eswa.2008.11.042

Schifano, E. D., Wu, J., Wang, C., Yan, J., & Chen, M. H. (2016). Online Updating of Statistical Inference in the Big Data Setting. *Technometrics, 58*(3), 393–403. https://doi.org/10.1080/00401706.2016.1142900

Shin, K. S., Lee, T. S., & Kim, H. J. (2005). An application of support vector machines in bankruptcy prediction model. *Expert Systems with Applications, 28*(1), 127–135. https://doi.org/10.1016/j.eswa.2004.08.009

Skyler, S., Eric, M., Isaac, M., & Felix, K. (2017). Mobile phone-based Credit Scoring. NetMob2017.

Song, Q., & Liang, F. (2015). A split-and-merge Bayesian variable selection approach for ultrahigh dimensional regression. *Journal of the Royal Statistical Society. Series B: Statistical Methodology, 77*(5), 947–972. https://doi.org/10.1111/rssb.12095

Stepanova, M., & Thomas, L. (2002). Survival analysis methods for personal loan data. *Operations Research, 50*(2), 277–289. https://doi.org/10.1287/opre.50.2.277.426

Thomas, L. C. (2000). A survey of credit and behavioural scoring: Forecasting financial risk of lending to consumers. *International Journal of Forecasting, 16*(2), 149–172. https://doi.org/10.1016/S0169-2070(00)00034-0

Tian, Y., Yong, Z., & Luo, J. (2018). A new approach for reject inference in credit scoring using kernel-free fuzzy quadratic surface support vector machines. *Applied Soft Computing Journal, 73*, 96–105. https://doi.org/10.1016/j.asoc.2018.08.021

Tong, E. N. C., Mues, C., & Thomas, L. C. (2012). Mixture cure models in credit scoring: If and when borrowers default. *European Journal of Operational Research, 218*(1), 132–139. https://doi.org/10.1016/j.ejor.2011.10.007

Verkasalo, H., López-Nicolás, C., Molina-Castillo, F. J., & Bouwman, H. (2010). Analysis of users and non-users of smartphone applications. *Telematics and Informatics, 27*(3), 242–255. https://doi.org/10.1016/j.tele.2009.11.001

Volkov, A., Benoit, D. F., & Van den Poel, D. (2017). Incorporating sequential information in bankruptcy prediction with predictors based on Markov for discrimination. *Decision Support Systems, 98*, 59–68. https://doi.org/10.1016/j.dss.2017.04.008

Xia, Y. (2019). A novel reject inference model using outlier detection and gradient boosting technique in peer-to-peer lending. *IEEE Access, 7*, 92893–92907. https://doi.org/10.1109/Access.628763910.1109/ACCESS.2019.2927602

Xia, Y., Liu, C., Da, B., & Xie, F. (2018). A novel heterogeneous ensemble credit scoring model based on bstacking approach. *Expert Systems with Applications, 93*, 182–199. https://doi.org/10.1016/j.eswa.2017.10.022

Zhang, H., & Liu, Q. (2019). Online learning method for drift and imbalance problem in client credit assessment. *Symmetry, 11*(7), 890. https://doi.org/10.3390/sym11070890

Zhao, D., Huang, C., Wei, Y., Yu, F., Wang, M., & Chen, H. (2017). An effective computational model for bankruptcy prediction using kernel extreme learning machine approach. *Computational Economics, 49*(2), 325–341. https://doi.org/10.1007/s10614-016-9562-7

Zmijewski, M. E. (1984). Methodological issues related to the estimation of financial distress prediction models. *Journal of Accounting Research, 22*, 59. https://doi.org/10.2307/2490859