Expert Systems With Applications 200 (2022) 116889

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Assessing credit risk of commercial customers using hybrid machine learning algorithms

Marcos Roberto Machado, Salma Karray [*,1]

Ontario Tech University, 2000 Simcoe St N., L1G 0C5, Oshawa, ON, Canada

ARTICLE INFO

Keywords:
Analytics
Hybrid algorithm
Credit scoring
Risk assessment
Machine learning

ABSTRACT

Given the large amount of customer data available to financial companies, the use of traditional statistical approaches (e.g., regressions) to predict customers' credit scores may not provide the best predictive performance. Machine learning (ML) algorithms have been explored in the credit scoring literature to increase predictive power. In this paper, we predict commercial customers' credit scores using hybrid ML algorithms that combine unsupervised and supervised ML methods. We implement different approaches and compare the performance of the hybrid models to that of individual supervised ML models. We find that hybrid models outperform their individual counterparts in predicting commercial customers' credit scores. Further, while the existing literature ignores past credit scores, we find that the hybrid models' predictive performance is higher when these features are included.

1. Introduction

Credit risk is an important issue for commercial banks and financial institutions. Such risk exists because borrowers may fail to make the mandatory payments on their loans, causing large losses to the lenders. In 2019, the outstanding credit for business in Canada, the USA and the UK, were, respectively $2,262 billion, $15,243 billion and 18,582 million pounds (Bank of Canada, 2020; Bank of England, 2020; USA Federal Reserve, 2020). Therefore, financial institutions would highly benefit from models that can accurately predict credit risk.

Modeling customers' credit risk involves assessing their probability of default, which is the likelihood of customers' defection for a specific period of time. Such probability is commonly estimated using credit risk models based on the regulatory-evaluation approach presented in the Bank for International Settlements (2006) in order to identify "good" (low risk) and "bad" (high risk) customers (Thomas, Oliver, & Hand, 2005).

Credit-scoring models are used to estimate customers' worthiness to receive credit. The higher the credit score, the lower the risk and the better the terms of the loan that will be offered to the customer (Crook, Edelman, & Thomas, 2007). Lenders rely on credit scores to manage the loan from the early stages of pre-screening applications and setting the terms of the loan and the interest rate, to the later stages of managing and possibly terminating the account.

According to Altman and Saunders (1997), credit scoring models can be univariate or multivariate-based systems. In univariate systems, the decision-maker compares various key accounting ratios of potential borrowers with industry norms. In multivariate systems, the key accounting variables are combined and weighted to produce a credit risk model. In both systems, the credit score is estimated using information about the customers' payment history, types of credit taken, current debt and length of credit, as well as demographic and other behavioral information (Thomas et al., 2005). After the customers' credit risk scores are estimated, they are compared to preset values and customers are assigned to groups of different risk levels. Based on this information, customers who are allocated to the high-risk group could be denied loans or subjected to increased scrutiny.

Many statistical methods for developing credit-scoring systems have been used in the literature such as linear probability (Crook et al., 2007), logistic regression (Wiginton, 1980) and discriminant analysis (Altman, 1968). These methods are common because they are easy to implement and allow for inference. Despite these advantages, they do not perform well in terms of prediction power when dealing with large customer data sets. This is an important shortcoming given that a small improvement in prediction power can lead to substantial increase in profitability (Abellán & Castellano, 2017). Recently, improved computing power has presented new opportunities to assess customers' credit scores using machine learning (ML) algorithms which prediction power

* Corresponding author.
E-mail addresses: marcos.machado@ontariotechu.net (M.R. Machado), salma.karray@ontariotechu.ca (S. Karray).

[1] The auhor acknowledges the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) [funding reference number RGPIN-2020-05156 and 535756–18].

https://doi.org/10.1016/j.eswa.2022.116889
Received 15 May 2020; Received in revised form 5 February 2021; Accepted 11 March 2022
Available online 26 March 2022

can outperform classical statistical approaches (Bequé & Lessmann, 2017; Dahiya, Handa, & Singh, 2017; Xia, Liu, Da, & Xie, 2018).

Many applications of unsupervised and supervised learning algorithms have been proposed to predict credit risk. The commonly used frameworks to assess customers credit risk when a target is given (supervised learning) are support vector machines (SVM) (Sun, Li, Chang, & Huang, 2015; Xu, Zhou, & Wang, 2009), decision trees (DT) (Zhou, Si, & Fujita, 2017), random forest (RF) (Malekipirbazari & Aksakalli, 2015), and artificial neural network (ANN) (West, 2000). As for problems that do not have a given label, clustering algorithms are explored. The most used methods in the financial industry are the k-Means (Bao, Lianju, & Yue, 2019; Lim & Sohn, 2007; Luo, Cheng, & Hsieh, 2009) and DBSCAN models (Lisuwan, Boonserm, & Sinapiromsaran, 2017; Schubert, Sander, Ester, Kriegel, & Xu, 2017). However, to the best of our knowledge, DBSCAN does not have any application in the credit risk setting.

Lately, ensemble models that integrate different ML algorithms have been explored in the context of credit scoring. One of the common strategies has been to perform customer credit scoring prediction based on outcomes of other ML algorithms. A growing literature that uses an ensemble integrated strategy has shown increased prediction power in credit-scoring models (Ala'raj & Abbod, 2015; Ala'raj & Abbod, 2016; Cleofas-Sánchez, García, Marqués, & Sánchez, 2016; Lessmann, Baesens, Seow, & Thomas, 2015; Wang, Ma, Huang, & Xu, 2012; Xia et al., 2018). Specifically, Bao et al. (2019) and Lim and Sohn (2007) used hybrid models where a supervised algorithm (k-Means) is implemented to cluster retail customers before applying different unsupervised ML algorithms to predict whether customers are "good" or "bad" payers.

In this paper, we use a similar approach (unsupervised learning for clustering + ML algorithms). However, our work differs from these papers in many ways. In particular, we use data about commercial customers while they focused on retail customers. Also, we apply different clustering approaches (k-Means and DBSCAN), while Bao et al. (2019) and Lim and Sohn (2007) implemented k-Means alone. Finally, both of these studies explore a classification problem given their categorical target variable (default/non-default). In our data set, we have access to the numeric value of customers credit score, which allows us to predict credit risk using regression algorithms. From a managerial perspective, our approach can be more accurate since we are not restricted to tagging customers into pre-specified classes. Instead, we predict a continuous credit score value that can later be used to group customers into different classes (or not) depending on managers' needs.

The aim of this study is to develop hybrid credit risk models for commercial customers using a private data set from a large North American bank, which offers different products and services and operates globally. In a first stage, we apply different unsupervised learning algorithms (k-Means and DBSCAN) to cluster customers based on a set of features. Then, we implement different supervised learning algorithms commonly used in the literature to perform the final credit scoring prediction (Adaboost, gradient boosting (GB), support vector machine (SVM), DT, RF and ANN). As illustrated in Fig. 1, different combinations of supervised and unsupervised methods are tested in order to verify the validity of our hybrid approach. The predictive performance of these algorithms is compared along different metrics such as mean absolute error (MAE), explained variance (EV) and mean squared error (MSE). Also, past credit scores are included as additional features in the data set, which can impact the prediction accuracy of all individual and hybrid models.

By applying unsupervised algorithms, our approach can help reduce overfitting of predictive outputs, a problem generally associated with implementation of all supervised learning algorithms. Also, this paper presents a new application of ensemble models (supervised and unsupervised algorithms) for credit risk prediction to the context of North American commercial customers. Another contribution of this work consists of testing the importance of past score features in the modeling process through the inclusion of observed past credit scores information in the main data set.

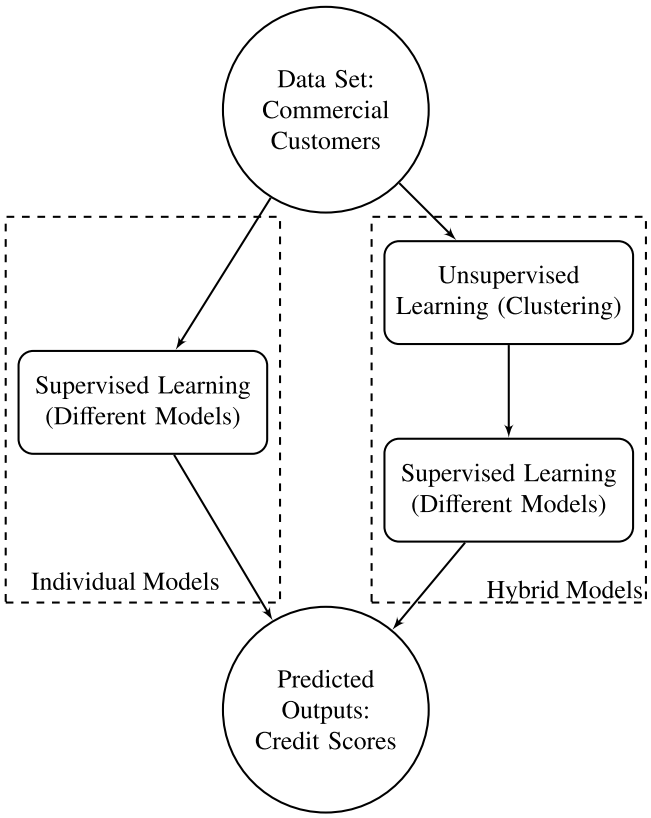

Fig. 1. Schematic framework for this study.

The paper is divided into six sections. Section 2 presents a literature review of credit risk models. Section 3 includes an overview of the main algorithms used in this study. Section 4 describes the experimental set up and data pre-processing treatments. Section 5 presents the results of implementing our proposed framework. Finally, Section 6 concludes and discusses areas for future research.

2. Literature review: Predictive modeling of credit risk

In this section, we discuss the literature that studied credit risk prediction. First, we introduce credit scoring models. Then, we discuss the literature that used a hybrid approach for credit risk prediction similar to the one we use in this paper. Next, since our study is focused on commercial customers, we discuss data challenges and differences between retail and commercial data sets. Finally, we showcase our contributions.

2.1. Credit risk assessment and prediction

Credit risk is the risk that a borrower defaults and does not honor their obligations to service debt (Van Gestel & Baesens, 2008). It is represented by the probability that a default event occurs (probability of default (PD)). According to Bank for International Settlements (2006), a default event on a debt obligation occurs if: 1. it is unlikely that the obligor will be able to repay its debt to the bank without giving up any pledged collateral, and 2. the obligor is more than 90 days past due on a material credit obligation. To measure credit risk, it is common to use the score function which can be a probability of a discrete outcome (default/non-default) or a continuous variable (Van Gestel & Baesens, 2008). Most scores are default risk scores related to a delinquency, default or bankruptcy probability. According to Siddiqi (2005), credit

scores for individuals or customer groups can be predicted using a statistical model (scorecard) that takes into account customer historical data and PD information. The predictive power of these models can be improved through the periodic monitoring of customer behavior.

The development and implementation of a credit scorecard model requires the adoption of an internal rating system within the organization and the application of credit risk models to estimate the customers' PD. This process involves different stages from gathering and preparing the relevant data to deploying and monitoring the final scorecard (Bequé & Lessmann, 2017). The literature provides many valuable resources for selecting explanatory features to be included in the model (Falangis & Glen, 2010), working with missing values (Florez-Lopez, 2010) or handling unbalanced data distribution (Paleologo, Elisseeff, & Antonini, 2010).

In the financial industry, different applications have been designed for the development of a PD/credit scorecard model and for the prediction and classification of credit scores. Lending technologies are defined by Moro and Fink (2013) as the approaches used by banks and financial institutions to differentiate between "bad" and "good" customers. They are categorized by two types: transaction lending that is based primarily on "hard" quantitative data, and relationship lending, which is based significantly on "soft" qualitative information. Under this categorization, transaction lending is generally focused on informationally transparent borrowers, while relationship lending is focused on opaque borrowers (Berger & Udell, 2006). For example, when assessing a new customer's credit risk, the bank or financial institution does not have any prior information about the customer. Therefore, the attribution of the labeled variable (probability of default/customer credit risk) has to be performed for the first time. In this case, relationship lending would have a higher weight in the final decision of whether or not to issue a loan to the customer. However, transaction lending can be more important than relationship lending when developing products and services or marketing campaigns that are tailored to the needs of customers, and when predicting and classifying customers' credit scores. In fact, recent research has shown that improvements in prediction power of credit score models can significantly reduce losses for lenders (Bao et al., 2019; Barboza, Kimura, & Altman, 2017; Bequé & Lessmann, 2017).

The literature focused on credit risk prediction has applied different methods. Notably, some studies used traditional statistical models, others have implemented machine learning algorithms, while a few have adopted a hybrid approach where data clustering is performed before prediction.

Traditional methods for predicting credit scores through the application of linear discriminate analysis and linear or logistic regression were presented by Altman (1968), Crook et al. (2007), Fisher (1936), Gurney (1997), Gurný and Gurny (2013), Li and Zhong (2012) and Wiginton (1980). In these studies, demographic variables such as age, education level, address and marital status for retail customers were considered, as well as behavioral features such as the number of active loans, the highest loan issued and the number of payment defaults.

Models that use machine and deep learning to predict credit scores in the retail setting were presented by Kozodoi, Lessmann, Papakonstantinou, Gatsoulis, and Baesens (2019), Liu, Xie, Zhao, Xie, and Liu (2019), Soui, Gasmi, Smiti, and Ghédira (2019) and Zhang, He, and Zhang (2019). Similar models were also implemented in the commercial setting (Barboza et al., 2017; Bequé & Lessmann, 2017; Kvamme, Sellereite, Aas, & Sjursen, 2018; Mai, Tian, Lee, & Ma, 2019; Pérez Martín, Pérez-Torregrosa, & Lamata, 2018). These models include transactional data such as the number of accounts and customer credit balance, in addition to variables describing the business features such as the number of employees and the geographic location of the company's offices (Barboza et al., 2017; Liang, Lu, Tsai, & Shih, 2016; Mai et al., 2019). Notably, Liu et al. (2019) and Mai et al. (2019) used and compared the efficiency of different classification algorithms to predict default within companies and Bequé and Lessmann (2017) tested extreme learning machines (ELM) to predict customers' credit scores.

When traditional statistical models are used for credit risk prediction (e.g., Bayesian statistics, logistic regressions and data mining), results are evaluated based on type I and II errors and hypotheses testing (e.g., Bravo, Maldonado, & Weber, 2013; Hand & Adams, 2014). However, machine and deep learning applications use a more diverse set of evaluation metrics such as accuracy, root mean squared error (RMSE) and area under the curve (AUC; Mai et al., 2019; Pérez Martín et al., 2018).

Next, we discuss studies that predicted credit risk using a hybrid approach where data clustering is performed before prediction.

2.2. Hybrid approaches for credit risk prediction

A few studies have used a hybrid method for credit risk prediction. This approach consists in clustering customers before predicting their credit category. In particular, Chandler and Ewert (1976) used a hybrid strategy by clustering data before performing predictions. In their study, different credit scoring systems were proposed for a sample of 2000 credit card and account bank customers. They segmented customers based on their gender before predicting credit categories using logistic regressions. Their results showed that this method improves the model's accuracy.

Similar to Chandler and Ewert (1976), other studies clustered retail customers based on pre-determined demographic variables before predicting customers' credit categories. For example, Banasik, Crook, and Thomas (1996) used demographic variables (e.g., marital and retirement status), then implemented classifiers (Logistic Regression (LR) and Linear Discriminant Analysis (LDA)) to predict whether customers are "good" or "bad" payers. Bijak and Thomas (2012) clustered customers based on their status, then applied different logistic regression tree classifiers to also predict "good" or "bad" payers. Finlay (2011) adopted a different approach by clustering customers using supervised learning techniques (e.g., K-Nearest Neighbor and Artificial Neural Network) before applying different classification methods using machine learning algorithms to perform prediction (e.g., LR, LDA and Adaboost). With the exception of Chandler and Ewert (1976), all of these studies showed that clustering prior to predicting whether customers are "good" or "bad" payers does not always improve prediction power (Banasik et al., 1996; Bijak & Thomas, 2012; Finlay, 2011).

Bao et al. (2019) and Lim and Sohn (2007) improved on these works by applying unsupervised learning techniques for clustering followed by classification ML algorithms for prediction of retail customers' credit categories. In particular, Lim and Sohn (2007) implemented the k-Means algorithm to define three different clusters, then applied an ANN classifier for prediction. Their results showed that the usage of hybrid models improves accuracy. However, they cautioned that the small size of the data set they used (about 1,000 cases) may have influenced their results. Similarly, Bao et al. (2019) implemented the k-Means algorithm to define clusters for retail customers based on missing (incomplete) data using a private Chinese data set. Then, they applied different classifiers to predict whether customers are "good" or "bad" loan applicants. They found that the combination of unsupervised (k-Means) and supervised machine learning algorithms improves the prediction performance of their models.

Table 1 presents a summary of this literature. It shows that our research differs from these studies as it is the only work to date that combines the following features: 1. it uses a data set focused on commercial (not retail) customers, 2. it applies different unsupervised learning algorithms for clustering (k-Means and DBSCAN), 3. it implements ML algorithms for prediction after data clustering, and 4. it explores a regression (not a classification) problem since our target is a continuous variable (credit score value).

2.3. Commercial data use and availability in credit risk prediction

The banking industry is formed by various banks and financial institutions that provide services to a diverse set of customers (retail

Table 1
Comparison of our paper to other studies that applied hybrid models for credit risk assessment.

	Chandler and Ewert (1976)	Banasik et al. (1996)	Lim and Sohn (2007)	Finlay (2011)	Bijak and Thomas (2012)	Bao et al. (2019)	Our study
Uses commercial customers' data set.							X
Implements ML algorithms.			X	X	X	X	X
Uses unsupervised learning (e.g., k-Means) for clustering.			X			X	X
Explores a regression problem.							X

and commercial customers). Retail customers are individual consumers while commercial customers comprise organizations of different sizes. In the context of credit score prediction, it is important to distinguish between retail and commercial customers. In fact, these groups are identified with different sets of variables, and the methods used to predict their credit risk differ as well (Thomas, 2010; Van Gestel & Baesens, 2008). For example, retail credit risk models incorporate demographic and behavioral features, whereas commercial credit risk models consider financial data (e.g., from balance sheets), classification data (company information, location, etc.) and financial ratios (Thomas, 2010).

The literature about credit risk assessment, specifically credit scoring, points out the challenge of finding real-world data, given the concern for keeping customers' credit information confidential (Bao et al., 2019; Barboza et al., 2017; Bequé & Lessmann, 2017; Kozodoi et al., 2019; Papouskova & Hajek, 2019). Access to such data is even more limited where commercial customers are concerned, as publicly available credit data sets and data science competition platforms do not make such data available (Dua & Graff, 2019; Kaggle, 2019). This research overcomes this challenge by assessing credit risk for commercial customers using a private data set provided by a North American commercial bank. The data set used in this study contains real-world financial, classification and transactional data as well as labeled information (i.e., current and past credit scores) of commercial customers over a period of three years.

A review of the literature that focuses on the prediction and classification of credit scores shows that many studies investigated retail customers' credit (Banasik et al., 1996; Bao et al., 2019; Bijak & Thomas, 2012; Chandler & Ewert, 1976; Finlay, 2011; Kozodoi et al., 2019; Kvamme et al., 2018; Lim & Sohn, 2007; Liu et al., 2019; Soui et al., 2019; Zhang et al., 2019), while others looked at commercial customers' credit (Barboza et al., 2017; Ben-David & Frank, 2009; Bequé & Lessmann, 2017; Liang et al., 2016; Mai et al., 2019; Vanneschi, Horn, Castelli, & Popovic, 2018). Research that explored credit scoring for retail customers usually relied on data sets from the UCI Machine Learning Repository (Dua & Graff, 2019), even when the purpose was only for validating research results (Bao et al., 2019; Bequé & Lessmann, 2017; Soui et al., 2019; Zhang et al., 2019), with few studies having used private data sets from specific markets (Bao et al., 2019; Kvamme et al., 2018; Liu et al., 2019).

As shown in the previous section, the literature that applied hybrid methods for credit risk prediction has solely focused on retail customers. The commercial setting has been explored using either traditional models or non-hybrid machine learning approaches. In particular, Bequé and Lessmann (2017) explored the applications of the extreme learning machines method and its performance compared with different ML algorithms in predicting customers' credit risk. They used both retail and commercial data. Their data set included a binary variable to identify commercial customers that presented default behavior in the past. However, they did not use labeled information nor a hybrid model that applies different ML algorithms to clustered data. Further, Ben-David and Frank (2009) compared different ML models to assess commercial customers' credit risk using data from an Israeli institution. Barboza et al. (2017) modeled North American companies' default using data spanning a 30-year period by implementing different ML algorithms. West (2000) applied different ANN models to test credit scoring prediction accuracy using real-world data from the German (retail) and Australian (commercial) markets. He found that both mixture-of-experts and radial basis function ANN models should be considered for credit-scoring predictions. Finally, Mai et al. (2019) proposed a model to assess credit risk adding text analysis through deep learning models and Liang et al. (2016) explored feature selection in credit-scoring model development using a 10-year data set containing information more than 500 Taiwanese companies. To our knowledge, our study is the first to use data about North American commercial customers with labeled information while applying a hybrid ML model.

2.4. Research contributions

Our review shows that most of the literature to date has been focused on credit risk prediction and classification for retail customers. Also, most studies in the literature explored applications of traditional and individual ML/DL models (e.g., ensemble and integration of ML/DL models or individual ML/DL models). However, the use of hybrid ML models has been very limited (see Table 1).

In particular, only Bao et al. (2019) and Lim and Sohn (2007) used hybrid methods for credit risk prediction that implement an unsupervised learning algorithm for clustering (k-Means) prior to applying machine learning algorithms for predicting credit risk. In this paper, we use a similar approach (unsupervised learning for clustering + ML algorithms). However, our work differs from these papers in many ways.

In fact, Lim and Sohn (2007)'s study relied on a small size data set, only used retail customer data, implemented k-Means for clustering and applied only an ANN classifier to predict retail customers' credit class ("good" vs. "bad" payers). In our paper, we use a large private data set, only commercial customers are considered, two different unsupervised methods for clustering are implemented (k-Means and DBSCAN) and multiple regressor algorithms are applied to predict customers' credit score (Adaboost, GB, DT, RF, SVM and ANN).

Similar to Bao et al. (2019), Lim and Sohn (2007) focused on retail customers and only used k-Means for clustering. However, they applied multiple ML algorithms for prediction (Logistic Regression, Decision Tree, Random Forest, Neural Networks, Gradient Goosting, Support Vector Machines, k-Nearest Neighbors) and considered a large data set. Our study differs from Bao et al. (2019) mainly in four different ways. First, we use data from a North American bank while they used data from the Chinese market. Further, we explore a commercial customer data set while they investigated retail customers. Second, we implement clustering in our hybrid models using k-Means and DBSCAN algorithms, while Bao et al. (2019) implemented k-Means alone. Third, concerning the clustering strategy, Bao et al. (2019) segmented their customers based on missing information in their data sets, while we use our target variable and other features in our data set. Fourth and finally, Bao et al. (2019) explored a classification problem since, in their data set, the target is a categorical variable (default/non-default). In our research, our target is a continuous variable (credit score numeric value). Therefore, we use ML regressor algorithms for prediction.

This study predicts credit scores of commercial customers at a North American commercial bank using a hybrid model, which combines unsupervised and supervised learning methods. To the best of our knowledge, this is the first work to date that combines the following features: 1. it uses a data set focused on commercial (not retail) customers, 2. it applies different unsupervised learning algorithms for clustering (k-Means and DBSCAN), 3. it implements ML algorithms for prediction after data clustering, and 4. it explores a regression (not a classification) problem since our target is a continuous variable (credit score value).

3. ML methods for classification and prediction

In this study, unsupervised learning is used through the implementation of k-Means and DBSCAN, and the individual supervised learning algorithms are tested through Adaboost, GB, DT, RF, SVM and ANN models. k-Means and DBSCAN were selected because of their efficiency and applicability (Dhanachandra, Manglem, & Chanu, 2015; Schubert et al., 2017). In the next section, we present a brief review of how these different algorithms (supervised and unsupervised learning) work, and provide their pseudocodes. We also briefly describe hybrid and ensemble methods.

3.1. k-Means

k-Means is a simple and efficient unsupervised learning algorithm used to cluster customers in k pre-defined clusters (AghaeiRad, Chen, & Ribeiro, 2017). The implementation of this clustering method follows the steps below (Bao et al., 2019):

1. Randomly select k cluster centers;
2. Assign each data point to its nearest cluster center;
3. Replace the original center with the position center in each cluster;
4. Relocate each data point to a new cluster which it is nearest to; and
5. Repeat steps 3 and 4 until no data point changes position or some convergence criterion is met.

There is only one main parameter in the k-Means model, that is, the number of clusters (k). This input can be chosen by intuition or using a cluster-predict methodology. For example, the Elbow method can help determine a good choice for k, based on the sum of squared distance between data points and their assigned clusters' centroids.

In this work, k-Means is used to segment customers into a number of clusters (defined with the Elbow method) using the target variable and the higher credit obtained by customers as features to split the data set into several subsets, based on which supervised machine learning models were built.

3.2. DBSCAN

The Density-Based Spatial Clustering of Applications with Noise (DBSCAN), first proposed by Ester, Kriegel, Sander, and Xu (1996) is the clustering algorithm most often used in the scientific literature. It is a non-parametric algorithm: given a set of points in some space, it groups points that are closely packed together (points with many nearby neighbors), marking the points that lie alone in low-density regions (whose nearest neighbors are too far away) as outliers.

The implementation of DBSCAN follows these steps (Schubert et al., 2017):

1. Define the points in the neighborhood (ϵ) of every point, and find the main points with more than minimum neighbors;
2. Identify the connected components, the set of main points on the neighbor graph, as the non-main points are ignored; and
3. Assign each non-main point to a nearby cluster if the cluster is a neighbor (ϵ), otherwise assign it to noise.

When implementing DBSCAN, a minimum number of samples per cluster must be defined, as well as the length of connections (and how this distance would be measured—for instance using the Euclidean distance). One of the advantages of this algorithm is that it is not necessary to identify the number of clusters in the modeling process. On the other hand, one of its disadvantages is that, depending on the data set and problem, some points cannot be reached, which makes the algorithm not entirely deterministic (Schubert et al., 2017).

3.3. AdaBoost regressor algorithm

Ada Boosting or Adaptive Boosting was the first practical boosting algorithm, initially proposed by Freund and Schapire (1996). This technique focuses on classification or regression problems and aims to convert a set of weak predictors into strong ones. The final prediction can be defined as:

$$F(x) = sign\left(\sum_{m=1}^{M} \theta_m f_m(x)\right), \quad (1)$$

where f_m are the weak classifiers and θ_m are their corresponding weights ($m = 1..M$). Therefore, the final prediction is given by the weighted combination of M weak classifiers.

Algorithm 1 presents the pseudocode on how the AdaBoosting algorithm works. It shows that, given a set of n points, where $x_i \in R^d$ and $y_i \in -1, 1$ (-1 represents the opposite class to 1), the weights are initialized for each point in step 1. Then, for all iterations, it is necessary to fit weak classifiers to the data set and select the one classifier with the lowest weighted classification error. For the lowest weighted classification error, it is necessary to calculate the weight for the m_{th} weak classifier. Then, the weight for each data point is updated. Finally, after M iterations, we can get the final prediction by summing up the weighted prediction of each classifier.

Algorithm 1: AdaBoosting: Pseudocode (Scikit Learn (2019), adapted by the authors).

Input: Data $\{(x_i, y_i)\}_{i=1}^{n}$

Step 1: Initialize the weight for each observation point: $w(x_i, y_i) = \frac{1}{n}$ for $i = 1, 2, ..., n$.

Step 2: for $m = 1$ to M:

- Fit weak classifiers to the data and select the one with the lowest weighted classification error: $\epsilon_m = E_{w_m}[1_{y \neq f(x)}]$;
- Calculate the weight for the m_{th} weak classifier: $\theta_m = \frac{1}{2} ln\left(\frac{1-\epsilon_m}{\epsilon_m}\right)$; and
- Update the weight for each data point as: $w_{m+1}(x_i, y_i) = \frac{w_m(x_i, y_i) exp[-\theta_m y_i f_m(x_i)]}{Z_m}$ where Z_m is a normalization factor to ensure the sum of all instance weights equals 1.

Step 3: Extract the final prediction by summing the weighted prediction of each classifier ($m = 1$ to M).

3.4. Gradient boosting (GB) algorithm

Gradient boosting models are used either for classification or regression. Breiman (1997) first presented the idea that boosting can be interpreted as an optimization algorithm for a given cost function. Later on, regression gradient boosting algorithms were implemented by different researchers (Elith, Leathwick, & Hastie, 2008; Friedman, 2000). Gradient boosting uses a strong learner – a classifier arbitrarily well correlated with the true classification – which is built from the combination of different weak learners, that is classifiers that are only slightly correlated with the true classification. This approach is applied in an iterative method (Li, 2017). By combining weak and

strong learners, the gradient boosting model's output is a prediction (classification or regression).

Algorithm 2 presents the pseudocode on how the gradient boosting algorithm works. As an input, it is necessary to have a given data set, $\{(x_i, y_i)\}_{i=1}^{n}$, and a loss function, $L(y_i, F(x_i))$. The algorithm is initiated by assuming that the average score is the predictor that minimizes the loss function. Next, the residuals (r_{im}) are measured for each instance m, which provides the gradient boosting calculation. Then, a regression is fitted over the residuals creating terminal regions (R_{jm}). After that, the minimum value of each defined terminal (γ_{jm}) is obtained by calculating the average value in each defined region. Finally, the constant (average) with which the iteration process started will be optimized in each integration time step to accurately obtain the final prediction ($F_m(x)$).

Algorithm 2: Gradient Boosting: Pseudocode (Scikit Learn (2019), adapted by the authors).

Input: Data $\{(x_i, y_i)\}_{i=1}^{n}$, and a differentiable loss function $L(y_i, F(x_i))$.

Step 1: Initialize model with a constant: $F_0(x) = \min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma)$.

Step 2: for $m = 1$ to M:

- Compute $r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}$ for $i = 1, 2, ..., n$;
- Fit a regression tree to the r_{im} values and create terminal regions R_{jm} for $j = 1, 2..., J_m$;
- For $j = 1, 2, ..., J_m$ compute $\gamma_{jm} = \min_{\gamma} \sum_{x_i \in R_{ij}} L(y_i, F_{m-1}(x_i) + \gamma)$; and
- Update $F_m(x) = F_{m-1}(x) + \nu \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$.

Step 3: Output $F_m(x)$.

3.5. Decision tree (DT) models

Decision tree models do not require powerful computations, are easy to understand and explain, handle missing information well, and work with numerical as well as nominal variables. However, they tend to suffer from overfitting problems (Bishop, 2006; Russell & Norvig, 2009). For these models, it is important to highlight the split concept, that is, when growing a tree, it is necessary to choose which feature will be selected to split the data in order to guarantee the best results possible. The method chosen to split the data should be applied in all processes until all data has been used. Decision tree models have been explored in the literature over the last few decades by applying various split methods (Dietterich, 2000; Fayyad & Irani, 1992; Kohavi & Quinlan, 2002; Quinlan, 1986). The most common metrics used to split data are gini impurity, information gain (entropy) and variance reduction. In this study, the algorithms were implemented using scikit-learn libraries which is based in the variance reduction of the gain of variance.

3.6. Random forest (RF) models

Random forest (RF) models are a category of additive models that make predictions by merging DT model outputs, which come from a sequence of base models. Formally, this class of models is defined in Eq. (2).

$$H(x) = \sum_{i=0}^{n} F_i(x), \tag{2}$$

where the final prediction model, $H(x)$, is described as the sum of the simple base models $F_i(x), i = 1..n$. It is important to mention that, in this case, each base classifier or regressor is a decision tree. RF models are built using trees that are formed by employing different subsamples of the data. A RF is an ensemble model because it uses multiple decision trees and a technique called bootstrap aggregation or bagging. This technique involves training each decision tree with the usage of different subsamples of the data where sampling is done with replacement (Negahban, 2019).

3.7. Support vector machines (SVM) models

The support vector machines (SVM) model was first proposed by Cleofas-Sánchez et al. (2016). It is another supervised learning algorithm that can be used to perform binary classification and regression predictions. This model has been extensively applied in the field of credit scoring, given its powerful predictive capabilities. In comparison with other similar models, this algorithm presents a superior solution for addressing the problem of sparsity in the data sets (Mai et al., 2019). The main idea is to project the data into a high-dimensional feature space and then find a hyperplane upheld by the support vectors to separate the two classes with a maximal margin. Based on the features of the support vectors, the label of the new input sample can be predicted (Bao et al., 2019). Different functions (kernels) can be used in SVM to map the input data into the high-dimensional feature space of higher orders (Zhou, Lai, & Yu, 2010).

3.8. Artificial neural network (ANN) models

Artificial neural network (ANN) models usually use a non-parametric approach that was inspired by the design of a neuron and mimics the human brain functions in terms of capturing complex relationships among inputs and outputs (Bishop, 1995). The most commonly-used type of ANN is the multi-layer perceptron (MLP), which consists of one input layer, one or more hidden layers and one output layer. ANN has been used for credit risk prediction (regressions) and classification problems (Bao et al., 2019; Mai et al., 2019; West, 2000). The ANN model begins passing the features of each customer to the input layer. These features are processed by the hidden layers, and then the output layer is reached, presenting the final prediction which is based on the weights. The latter is defined for each feature based on its relative importance. Finally, an activation function, such as sigmoid, collects all the weighted features to produce outputs. This process is repeated in several loops to reduce the error between the predicted and true class (Bao et al., 2019; Malhotra & Malhotra, 2003).

3.9. Hybrid and ensemble models

Both ensemble and hybrid ML models use an information fusion approach, albeit differently. On one hand, ensemble models present multiple but homogeneous combined individual models (Kajdanowicz, Kazienko, & Kraszewski, 2010). Usually, within the individual models, many merging algorithms are used to build a group of regressors or classifiers (Kuncheva, 2004). On the other hand, hybrid models combine completely different individual and heterogeneous ML techniques (Castillo, Melin, & Pedrycz, 2010; Corchado, Abraham, & de Carvalho, 2010).

Ensemble and hybrid ML models have become more popular in the last few years with applications spanning different fields such as medical diagnosis, bioinformatics, recommender systems, text/music classification, and relevant to this work, financial forecasting (Bao et al., 2019; Bergstra, Casagrande, Erhan, Eck, & Kégl, 2006; Castillo et al., 2010; De Caigny, Coussement, & De Bock, 2018; Kempa, Lasota, Telec, & Trawiński, 2011; Lim & Sohn, 2007; Okun, 2011). These types of ML models can handle high dimensional data sets and have demonstrated better performance than individual models (Brazdil, Giraud-Carrier, Soares, & Vilalta, 2009). Ensemble and hybrid ML models are also popular because, by combining individual models and extracting the best method, they mitigate possible weaknesses of individual models (Cios & Kurgan, 2002) and Wermter and Sun (2014).

Recently, hybrid ML models that combine diverse approaches by first clustering a pattern, then performing predictions have been proposed in many areas. For example, Luo (2020) implemented k-Means algorithms to cluster different weather profiles, then applied ANN to predict the day-ahead cooling demand. In their hybrid approach, Corizzo, Pio, Ceci, and Malerba (2019) proposed a new clustering

Table 2
Data set size and description of DPT (data pre-processing treatment).

Source of data	Data set size	Data Description
S1	314 × 702,818	Classification, financial and transaction data from 3 years of quarterly observations.
S2	67 × 243,251	Classification and financial data from 3 years of quarterly observations.
S3	54 × 702,816	Classification and financial data from 3 years of quarterly observations.

density-based algorithm (DENCAST). While similar to DBSCAN, it can provide better results when implemented in large-scale and high dimensional data in different settings (e.g., 20 million cases and 221 features). Bandara, Bergmeir, and Smyl (2020) studied hybrid models for time series problems. In their experiments, algorithms such as PAM and SNOB[2] have been tested in addition to k-Means and DBSCAN and applied in combination with predictive frameworks such as the ARIMA and ETS. Further, Lin, Yen, and Wang (2020) proposed a series of auto-encoders to extract customers' features and used k-Means to cluster customer behavior in order to predict upcoming trend videos in a time series model. Finally, Ceci, Corizzo, Japkowicz, Mignone, and Pio (2020) proposed ECHAD (Embedding-based CHAnge Detection) to cluster information in different time series for the smart grid industry. This approach leverages embedding techniques, one-class learning, and a dynamic detection approach that incrementally updates the learned model to reflect the new data distribution. These new hybrid methods are useful to analyze time series data.

In our paper, the data set we use contains cross-sectional data. As such, we will implement similar hybrid techniques to those used in the credit risk prediction literature (Bao et al., 2019; Bijak & Thomas, 2012; Finlay, 2011; Lim & Sohn, 2007). In particular, we will apply different unsupervised clustering algorithms (k-Means and DBSCAN). Since our data set can be clustered using DBSCAN, a variation of this algorithm for a large scale and high dimensional data set does not have to be tested (Corizzo et al., 2019). Then, we will implement different regressor models on the obtained cluster to predict credit scores.

4. Experimental set-up

4.1. Data and experimental design

This study explored the financial data set provided by a North American commercial bank, which originated from different credit bureaus. Table 2 describes the data used in this modeling process.

The experimental design of this research was selected to test and compare the credit risk prediction performance of individual models (supervised learning algorithms) and of hybrid models (unsupervised and supervised learning algorithms) for commercial customers using the data set provided by the financial institution. The same scenario was tested again including customers' past scores information. Cross-validation (5 × 2) was implemented in all supervised learning cases, splitting the data in training, testing and validation sets (Bao et al., 2019; Burez & den Poel, 2009; Dietterich, 1998). Finally, metrics of validation were extracted to compare the different models.

4.2. Data pre-processing

At the data pre-processing stage, the following tasks are performed:

1. In the labeled data set explored in this study, the target variable was provided in different ranges by different sources of data. The scores were then normalized in our data set following Equation (3) to obtain the credit scores values for each customer ranging between 0 and 1;

$$I_0 = \frac{I_0 - I_{min}}{I_{max} - I_{min}} \tag{3}$$

where I_0 is the instance of the score to be normalized, I_{min} is the minimum score observed in the distribution and I_{max} is the maximum score observed at that point in time;
2. The duplicated features, specifically the variables with the same meaning but different descriptions, were removed;
3. The categorical features were handled using one-hot encoder process and applying the usage of dummies/numerical characters in the final data frames;
4. The missing values were evaluated. When missing data represented more than 5% of cases, it was replaced by the mean, median or mode of the variable, and deleted from the final data frame otherwise;
5. The outliers were analyzed and dummy/flag variables were created to verify their behavior and to determine how important they were for the model;
6. Multicollinearity was tested using the measurement of the variance inflation factor (VIF). We found that no feature in the main data set was highly correlated, therefore no variable was deleted; and
7. Finally, correlation and scaling analyses were performed. In cases where a high correlation was found between features (correlation higher than 90%), they were dropped from the final data set.

4.3. Parameter settings

Different individual algorithms have been used in previous studies on credit risk prediction and classification (Barboza et al., 2017; Burez & den Poel, 2009; Kozodoi et al., 2019; Mai et al., 2019). A few studies also used ensemble and hybrid models (Bao et al., 2019; Dietterich, 2000). In this research, parameter settings were defined based on what these previous studies recommended. For instance, when required data was trained with learning rates 0.01, 0.025 and 0.05, the number of iterations tested were 500, 1000, 2500 and 5000, respectively. In all cases, the maximum number of features used was the maximum presented in the data set after the data pre-processing treatment. Specifically, for k-Means and DBSCAN, the number of clusters created (3) was defined based on the Elbow method. Parameters for each type of individual model were set in all implemented cases, with or without past scores information.

4.4. Evaluation criteria

In regressor models, different metrics for validation can be extracted and measured such as mean squared error (MSE), explained variance (EV), R^2 score, mean absolute error (MAE), mean squared log error (MSLE) and median absolute error (MeAE). These metrics have been used in the literature when supervised learning (regressor) models were implemented (Cornejo-Bueno et al., 2017; Martínez-Martínez et al., 2011; Siddiqi, 2005; Torre, Marelli, Embrechts, & Sudret, 2019). All these metrics are measuring the distance between the predicted credit scores and the observed ones, and can be used to compare and discuss the obtained results.

We also compare the performance of our models (regressors) using the 5 × 2 cross-validation paired t-test (Dietterich, 1998). This test is

[2] PAM is the most common k-medoid algorithm and is a partitioning method similar to k-Means. SNOB clusters information through classification and is based on an expectation–maximization algorithm and Bayesian principles.

chosen because it addresses the issues of other statistical tests, such as the resampled paired t-test and the k-fold cross-validated paired t-test, which violate the assumptions of the Student's t-test.[3] This test is performed as follows: 1. the data set is split into training and test samples (50% each) five times, 2. in each fold, a classifier is fit to the training set, and its performance is evaluated in the test set, 3. the training and test sets are rotated, and performances are computed, 4. the performance difference measures are calculated, 5. the mean and variance of these differences are computed, and 6. the variance is used to compute the t-statistic that can then be compared with a p-value. The t-statistic approximately follows a T distribution with 5 degrees of freedom under the null hypothesis that the two models evaluated have equal performance (Dietterich, 1998).

5. Results

The results are presented in two parts. First, we describe the predictive performance of each individual and hybrid model along the evaluation metrics considered in this study. Then, we provide a comparison and discussion of these results. Note that, for the hybrid models, implementations with k-Means and DBSCAN were performed.

The results were very similar across methods with k-Means presenting a slightly better performance in all hybrid models. For instance, the best predictions for the hybrid ML models using k-Means in the unsupervised part of the framework are given by the combination of k-Means + (DT/RF), with an EV of 99% and MSE of 0.00001. When using DBSCAN, the best metrics of validation are presented by the combination DBSCAN + (DT/RF), which has an EV of 98% and MSE of 0.00003. Similarly, the worst hybrid models in both frameworks are the ones that consider ANN in the supervised learning part of the method. Thus, we only include here results obtained with k-Means, while all results considering DBSCAN are presented in the supplementary materials (Tables S.1. and S.2.).

5.1. Predictive performance

All findings are presented in Tables 3 and 4. The results obtained for the models that exclude past credit scores are shown in Table 3, while those that include past credit scores are presented in Table 4. Each of these Tables contains the metrics of validation for each model category (individual vs. hybrid) to facilitate comparison among the different approaches. Comparisons across tables allows us to analyze the effect of including past score features on the results. Looking at the evaluation metrics for MSE, MAE, MSLE and MeAE, lower values indicate better model performance, while for EV and R^2, larger absolute values are preferred.

To analyze these results, we first start by looking at the individual models without the past score features. Table 3 shows that DT and RF models present the smallest MSE as well as the largest EV and R^2, followed by GB which has a low MSE and large EV and R2. Therefore, when past score features are omitted, DR, RF and GB provide the best performance, predicting almost 100% of the information. This result holds even when using cross-validation to ensure that no overfitting is occurring. The next-best performing method is SVM which shows an MSE of 0.00724 and 84% of the explained variance (EV). The worst prediction power is obtained by ANN, with the highest MSE (0.01710) and the smallest R^2 (0.01). This is in line with the literature that also found low prediction power of credit scoring using ANN (Bao et al., 2019; Bequé & Lessmann, 2017).

Looking at the hybrid models' performance when past scores are excluded, Table 3 shows that the combinations of k-Means with RF presents the best prediction power with the smallest MSE (0.00005) and the largest EV (0.99744). Comparing the performance of the different hybrid models, we find that the top three frameworks (k-Means + (DT/RF/Gradient Boosting)) have an average EV of approximately 99%. In contrast, the EV's average for the three least accurate models (k-Means + (Adaboost/SVM/ANN)) is 81%.

To compare the performance of individual and hybrid models, we contrast the metrics of validation for the individual models against the average indicators for the hybrid frameworks, which is calculated over the metrics for each cluster. As we can see in Table 3, the error for all hybrid methods is smaller than in the individual models. For instance, the combination k-Means + Gradient Boosting's error (MSE) is five times smaller than the Gradient Boosting's error. Further, even considering a relatively higher standard deviation for the given mean, the hybrid models' performance is more accurate than their individual versions. For example, Adaboost has an MSE of 0.00725, while in the hybrid approach (k-Means + Adaboost), the average MSE is 0.00186 with a standard deviation of 0.00111. Therefore, even considering a stress scenario for the variation, the average performance still presents better performance in the hybrid approach. However, some exceptions are noted for SVM and ANN where a few metrics of validation for clusters (and average) present better results in the individual models. For instance, the MSE for ANN is 0.01710, while it is four times higher for k-Means + ANN framework (average MSE across clusters = 0.04911). This poor performance can be attributed to the data set size and to SVM and ANN poor predictive performance in regression problems (Utkin, 2019; West, 2000).

Further, the same set of experiments were implemented after including past credit score features to the data set. In particular, we applied the same individual and hybrid models after adding two explanatory variables which consisted in the historic credit scores from a one-year and two-year period. Correlation and multicollinearity tests with these variables showed no need for feature removal.

Table 4 presents the metrics of validation for the individual models tested with both the one- and two-year past score features added to the models. The results indicate that the best prediction power is obtained with the DT, RF and GB models as indicated by an MSE of 0.00008, 0.00004 and 0.00036 and an EV of 0.99972, 0.99952 and 0.97908, respectively. On the other hand, the model with the worst prediction accuracy was ANN, again with MSE of 0.01719 and EV of 0.01. These findings are similar to the individual models applied to the data set excluding past score information (Table 3), albeit a slight improvement.

Table 4 also presents the prediction power of the hybrid models tested using the past credit score features. These results show that, considering all different clusters, RF models present the best prediction power with the smallest error as well as the largest EV and R^2. Similar to the results obtained without past credit score features (Table 3), Table 4 shows that the hybrid models provide lower errors than their individual counterparts. For instance, RF has an MSE of 0.00004, while k-Means + RF average MSE is four times lower (0.00001). Also, the best three hybrid models (k-Means + (RF/DT/Gradient Boosting)) present an average EV of approximately 99%. In contrast, the individual versions of these frameworks (RF/DT/Gradient Boosting) show an average EV of 98%.

Finally, comparing results obtained from hybrid models in Tables 3 and 4 shows better performance of these models when past score features are included in the analysis than when they are excluded. In fact, the average error (MSE) for all proposed models excluding past credit scores is about 0.00494 (Tables 3), while the MSE is 0.00491 for the models including past score features (Tables 4). Similar gains can be observed for all metrics extracted. For example, the average EV of each case's top three models (k-Means + (DT/RF/Gradient Boosting)) is approximately 98.58% when past credit scores are not considered and 98.62% when these features are included in the models. Furthermore, comparing the EV of the best predictive models (k-Means + RF) in Tables 3 and 4 shows that including past scores leads to an improvement

[3] The differences of the model performances are not normally distributed because the accuracies are not independent.

Table 3
Metrics for validation — Individual and Hybrid models excluding past score features.

Framework	Cluster	MSE	EV	MAE	MSLE	MeAE	R^2
Individual Models							
AdaBoost		0.00725	0.27943	0.23631	0.07418	0.00395	0.09192
Gradient Boosting		0.00031	0.96793	0.01288	0.00010	0.00790	0.97791
Decision Tree		0.00004	0.99632	0.00100	0.00002	0.00001	0.99730
Random Forest		0.00004	0.99649	0.00100	0.00002	0.00001	0.99749
SVM		0.00724	0.84431	0.08212	0.00333	0.08678	0.57921
ANN		0.01710	0.00100	0.10344	0.00719	0.07099	0.01000
Hybrid Models							
k-Means + Adaboost	1	0.00256	0.86877	0.04624	0.00134	0.04602	0.85506
	2	0.00242	0.73307	0.04476	0.00121	0.04410	0.72527
	3	0.00058	0.94436	0.02095	0.00027	0.02247	0.93995
	Mean	0.00186	0.84874	0.03732	0.00094	0.03753	0.84009
	SD	0.00111	0.10706	0.01420	0.00058	0.01308	0.10812
k-Means + Gradient Boosting	1	0.00015	0.99248	0.00631	0.00784	0.00375	0.99248
	2	0.00003	0.99780	0.00385	0.00001	0.00316	0.99780
	3	0.00001	0.99897	0.00027	0.00000	0.00012	0.99897
	Mean	0.00006	0.99642	0.00348	2.62163	0.00234	0.99642
	SD	0.00007	0.00346	0.00304	4.54077	0.00195	0.00346
k-Means + Decision Tree	1	0.00012	0.99435	0.00285	0.00006	0.00001	0.99435
	2	0.00002	0.99886	0.00011	0.00001	0.00000	0.99886
	3	0.00003	0.99799	0.00017	0.00001	0.00000	0.99799
	Mean	0.00005	0.99707	0.00104	0.00003	0.00000	0.99706
	SD	0.00005	0.00239	0.00156	0.00003	0.00000	0.00239
k-Means + Random Forest	1	0.00011	0.99438	0.00284	0.00006	0.00000	0.99438
	2	0.00001	0.99927	0.00009	0.00000	0.00000	0.99927
	3	0.00002	0.99867	0.00016	0.00001	0.00000	0.99866
	Mean	0.00005	0.99744	0.00103	0.00002	0.00000	0.99744
	SD	0.00006	0.00267	0.00157	0.00003	0.00000	0.00267
k-Means + SVM	1	0.00680	0.80023	0.07756	0.00317	0.09236	0.66611
	2	0.00797	0.87365	0.08698	0.00372	0.09873	0.47619
	3	0.00615	0.69063	0.07106	0.00273	0.09309	0.51910
	Mean	0.00697	0.78817	0.07853	0.00320	0.09502	0.55380
	SD	0.00092	0.09210	0.00801	0.00050	0.00321	0.09960
k-Means + ANN	1	0.02038	0.00000	0.10729	0.00923	0.07543	0.00000
	2	0.11417	0.00000	0.10336	0.00892	0.08625	0.00000
	3	0.01280	0.00000	0.09037	0.09037	0.00531	0.00000
	Mean	0.04911	0.00000	0.10034	0.03618	0.05566	0.00000
	SD	0.05646	0.00000	0.00885	0.04694	0.04394	0.00000

of about 0.2%. Therefore, these results indicate that the usage of past credit scores improves the accuracy of these models.

Besides extracting metrics of validation, a 5 × 2 cross-validation paired t-test was implemented for individual and hybrid ML frameworks. Sixty-six different pairs were formed for each case considered in our study (with and without past credit scores) for individual and hybrid ML models. The test's null hypothesis is that models compared in each pair have equal performance, and the chosen significance level is 5% (Dietterich, 1998; Witten, Frank, & Hall, 2011). The results show that among all 132 pairs tested, we could not reject the null hypothesis in only 4 cases (less than 0.5%). These pairs correspond to DT comparisons with RF in individual and hybrid approaches, meaning that these models' performance was not significantly different (e.g., p-value = 0.415 and t-Statistic of −3.899 for comparison of k-Means + DT with k-Means + RF). Details of all results for this test are included in the supplementary materials (Tables S.3. and S.4.). These findings show that the different modeling strategies considered in our paper can lead to significantly different results; hence the importance of implementing different modeling approaches when predicting credit risk.

5.2. Discussion

A series of observations and comparisons can be made based on the results presented in the previous section. First, individual models such as AdaBoost and ANN yield the worst prediction power mainly because of the size of the data set. According to Ala'raj and Abbod (2015), Desai, Crook, and Overstreet (1996), Scikit Learn (2019) and West (2000), these models perform better when dealing with a very large amount of information. A similar observation can be made for the hybrid model consisting of k-Means combined with ANN /Adaboost which gave poor performance on all evaluation metrics. This result supports previous findings in the literature dealing with these methods in credit score prediction (Bao et al., 2019; Bequé & Lessmann, 2017; Kvamme et al., 2018; Mai et al., 2019).

Second, comparing the prediction power of all models, we can see that RF, DT and GB were the best predictors. Even after submitting these models to a cross-validation system to avoid overfitting problems, we still find that almost 100% of the variance can be explained by these models.

Third, for all models explored (i.e., individual and hybrid models, with and without past score features), SVM presented good prediction power with EV or R^2 around 80% and MSE of 0.0003. This indicates that SVM is also a good tool to predict credit scores for commercial customers.

Fourth, comparisons of results obtained for individual and hybrid models show that each hybrid ML model outperformed its individual version. For instance, Adaboost, which shows an EV of about 30% when implemented alone, presents approximately 90% of explained variance when combined with k-Means.

Fifth, when hybrid models are implemented, comparison of the number of clusters created through both DBSCAN and k-Means based on the implementation of the elbow method shows the same results (three clusters). Each of the three clusters presented different evaluation metrics' values with some models performing better than others on some clusters. For example, in the case of k-Means + SVM including past score features (Table 4), the EV over the three clusters

Table 4
Metrics for validation — Individual and Hybrid models including past score features.

Framework	Cluster	MSE	EV	MAE	MSLE	MeAE	R^2
Individual Models							
AdaBoost		0.00705	0.30954	0.25917	0.07352	0.00351	0.09211
Gradient Boosting		0.00036	0.97908	0.01293	0.00018	0.00891	0.97608
Decision Tree		0.00008	0.99952	0.00005	0.00003	0.00001	0.99952
Random Forest		0.00004	0.99972	0.00006	0.00001	0.00009	0.99972
SVM		0.00722	0.84471	0.08210	0.00338	0.08682	0.57982
ANN		0.01719	0.01000	0.10347	0.00792	0.07097	0.00010
Hybrid Models							
k-Means + Adaboost	1	0.00199	0.90221	0.04133	0.00101	0.04323	0.88775
	2	0.00250	0.70514	0.04486	0.00129	0.04450	0.69192
	3	0.00024	0.97842	0.01284	0.00011	0.00894	0.97814
	Mean	0.00158	0.86192	0.03301	0.00080	0.03222	0.85260
	SD	0.00119	0.14103	0.01755	0.00061	0.02017	0.14631
k-Means + Gradient Boosting	1	0.00011	0.99469	0.00560	0.00005	0.00336	0.99469
	2	0.00003	0.99787	0.00348	0.00001	0.00286	0.99787
	3	0.00001	0.99883	0.00028	0.00000	0.00012	0.99882
	Mean	0.00005	0.99713	0.00312	0.00002	0.00211	0.99713
	SD	0.00005	0.00216	0.00268	0.00003	0.00174	0.00216
k-Means + Decision Tree	1	0.00000	0.99978	0.00007	0.00000	0.00000	0.99978
	2	0.00001	0.99905	0.00010	0.00001	0.00000	0.99905
	3	0.00002	0.99827	0.00013	0.00001	0.00000	0.99827
	Mean	0.00001	0.99903	0.00010	0.00000	0.00000	0.99903
	SD	0.00001	0.00075	0.00003	0.00000	0.00000	0.00075
k-Means + Random Forest	1	0.00000	0.99976	0.00007	0.00000	0.00000	0.99976
	2	0.00001	0.99939	0.00008	0.00000	0.00000	0.99939
	3	0.00001	0.99892	0.00013	0.00000	0.00000	0.99892
	Mean	0.00001	0.99936	0.00010	0.00000	0.00000	0.99936
	SD	0.00000	0.00042	0.00003	0.00000	0.00000	0.00042
k-Means + SVM	1	0.00679	0.80067	0.07750	0.00316	0.09332	0.66677
	2	0.00796	0.87385	0.08695	0.00371	0.09878	0.47665
	3	0.00615	0.69088	0.07108	0.00273	0.09325	0.51887
	Mean	0.00697	0.78846	0.07851	0.00320	0.09512	0.55410
	SD	0.00092	0.09209	0.00798	0.00049	0.00318	0.09984
k-Means + ANN	1	0.02038	0.00000	0.10729	0.00923	0.07543	0.00005
	2	0.41151	0.00000	0.13004	0.00009	0.08625	0.00000
	3	0.01280	0.00000	0.09037	0.00531	0.10325	0.00070
	Mean	0.14823	0.00000	0.10923	0.00488	0.08831	0.00025
	SD	0.22804	0.00000	0.01991	0.00459	0.01402	0.00039

was 0.80067, 0.87385 and 0.69088, respectively. These results may be improved for the cluster presenting the lowest explained variance. This can be accomplished, for example, by balancing data for the target, implementing a new clustering process, or by dimension-reduction and/or feature engineering.

Sixth and finally, comparing the results from all individual and hybrid models considering the effects of past scores, we find that adding these features improved the prediction power of all models. For example, the k-Means + RF hybrid model, which presents the best validation metrics in all experiments, shows better accuracy when past credit scores are used. This result challenges the literature about credit scoring prediction which has ignored historic credit score data. It indicates that such features should be considered in order to improve the accuracy of commercial customers' credit scoring prediction.

6. Conclusions

Financial institutions rely on credit risk assessment to process and manage loans. Customers' credit scores can provide valuable information to lenders about their customers' financial viability. This research implements individual and hybrid supervised machine learning (ML) models to predict commercial customers' credit scores. Hybrid ML models operate in two steps; first the data is clustered using a classifier method (k-Means and DBSCAN), then different ML models are applied to each of these clusters to obtain predictions. Six individual ML models are implemented in this study (Adaboost, GB, DT, RF, SVM, and ANN) and their prediction performances are compared with hybrid models. The unique data set used for this study is provided by a North American commercial bank and contains financial information about their commercial customers. While most literature to date utilizes data sets about individual (retail) bank customers, to our knowledge, this is the first study that focuses on credit score prediction for commercial customers using hybrid ML models.

Results show that the hybrid models (k-Means + (DT/RF)) outperform individual models in predicting commercial customers' credit scores. Since outliers and missing values are handled in this study, this finding can be explained by the fact that clustering allows for dimensionality reduction in the data which allows for improvement in prediction performance (Bao et al., 2019; Bequé & Lessmann, 2017). Results also indicate that considering past credit score features improves the prediction power of all models.

Based on the metrics of validations used in this work, lenders can implement different combinations of unsupervised and supervised learning algorithms to predict commercial customers' credit scores. Our results show that, in our setting, k-Means provides better results than DBSCAN when implemented in a hybrid model. Further, we find that a combination of k-Means with (DT/RF) provides the best prediction output for our data set. k-Means and SVM also present good prediction power with EV or R^2 around 80% and a very low MSE. Finally, Adaboost, which does not perform well alone with an EV of 0.279, presents a higher average EV over the obtained clusters of 0.84874 when integrated with k-Means.

These findings have interesting implications for banks and financial institutions in predicting commercial customers' credit scores. To achieve a higher prediction power when assessing credit risk, hybrid and/or ensemble models should be tested and implemented. Lenders

should also consider including past credit scores, as this information can improve the prediction power of their models. In particular, based on our results, developing and implementing a hybrid model formed by a k-Means and DT or RF, and considering past credit score features provides better prediction accuracy.

Finally, future work should focus on different parameter settings that can be used and tested, such as the number of leaves and trees (in DT and RF models) and the number of layers in the ANN models. Another extension of this study will be to establish classes of scores and implement classification models (instead of regressor algorithms) that can be compared with most benchmark studies in the literature (Bao et al., 2019; Barboza et al., 2017; Mai et al., 2019). Finally, future research can implement hybrid models using other clustering algorithms.

CRediT authorship contribution statement

Marcos Roberto Machado: Conceptualization, Methodology, Software, Data curation, Visualization, Investigation, Formal analysis, Writing – original draft. **Salma Karray:** Supervision, Funding acquisition, Resources, Conceptualization, Methodology, 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 material related to this article can be found online at https://doi.org/10.1016/j.eswa.2022.116889.

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.

AghaeiRad, A., Chen, N., & Ribeiro, B. (2017). Improve credit scoring using transfer of learned knowledge from self-organizing map. *Neural Computing and Applications*, *28*(6), 1329–1342.

Ala'raj, M., & Abbod, M. (2015). A systematic credit scoring model based on heterogeneous classifier ensembles. In *2015 International symposium on innovations in intelligent systems and applications* (pp. 1–7). IEEE.

Ala'raj, M., & Abbod, M. F. (2016). A new hybrid ensemble credit scoring model based on classifiers consensus system approach. *Expert Systems with Applications*, *64*, 36–55.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, *23*(4), 589–609.

Altman, E. I., & Saunders, A. (1997). Credit risk measurement: Developments over the last 20 years. *Journal of Banking & Finance*, *21*(11–12), 1721–1742.

Banasik, J., Crook, J., & Thomas, L. (1996). Does scoring a subpopulation make a difference. *International Review of Retail, Distribution and Consumer Research*, *6*(2), 180–195.

Bandara, K., Bergmeir, C., & Smyl, S. (2020). Forecasting across time series databases using recurrent neural networks on groups of similar series: A clustering approach. *Expert Systems with Applications*, *140*, 112896.

Bank for International Settlements (2006). *International convergence of capital measurement and capital standards a revised framework*. Basel, Switzerland: Bank for International Settlement, Press & Communications.

Bank of Canada (2020). Business credit: Descriptive statistics.

Bank of England (2020). Money and outstanding credit for business: Statistics.

Bao, W., Lianju, N., & Yue, K. (2019). Integration of unsupervised and supervised machine learning algorithms for credit risk assessment. *Expert Systems with Applications*, *128*, 301–315.

Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications*, *83*, 405–417.

Ben-David, A., & Frank, E. (2009). Accuracy of machine learning models versus "hand crafted" expert systems—A credit scoring case study. *Expert Systems with Applications*, *36*(3), 5264–5271.

Bequé, A., & Lessmann, S. (2017). Extreme learning machines for credit scoring: An empirical evaluation. *Expert Systems with Applications*, *86*, 42–53.

Berger, A. N., & Udell, G. F. (2006). A more complete conceptual framework for SME finance. *Journal of Banking & Finance*, *30*(11), 2945–2966.

Bergstra, J., Casagrande, N., Erhan, D., Eck, D., & Kégl, B. (2006). Aggregate features and ADABOOST for music classification. *Machine Learning*, *65*(2–3), 473–484.

Bijak, K., & Thomas, L. C. (2012). Does segmentation always improve model performance in credit scoring? *Expert Systems with Applications*, *39*(3), 2433–2442.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. New York, NY: Oxford University Press Inc.

Bishop, C. M. (2006). *Pattern recognition and machine learning. Information science and statistics*. Berlin, Germany: Springer-Verlag.

Bravo, C., Maldonado, S., & Weber, R. (2013). Granting and managing loans for micro-entrepreneurs: New developments and practical experiences. *European Journal of Operational Research*, *227*(2), 358–366.

Brazdil, P., Giraud-Carrier, C., Soares, C., & Vilalta, R. (2009). *Metalearning: applications to data mining* (1st ed.). Berlin, Germany: Springer.

Breiman, L. (1997). *Arcing the edge*. Berkeley, CA: University of California.

Burez, J., & den Poel, D. V. (2009). Handling class imbalance in customer churn prediction. *Expert Systems with Applications*, *36*(3, Part 1), 4626–4636.

Castillo, O., Melin, P., & Pedrycz, W. (2010). *Studies in fuzziness and soft computing, Hybrid intelligent systems: analysis and design*. Berlin, Germany: Springer.

Ceci, M., Corizzo, R., Japkowicz, N., Mignone, P., & Pio, G. (2020). ECHAD: Embedding-based change detection from multivariate time series in smart grids. *IEEE Access*, *8*, 156053–156066.

Chandler, G. G., & Ewert, D. C. (1976). *Discrimination on the basis of sex under the equal credit opportunity act*. (8), Citeseer.

Cios, K. J., & Kurgan, Ł. A. (2002). Hybrid inductive machine learning: An overview of CLIP algorithms. In L. C. Jain, & J. Kacprzyk (Eds.), *New learning paradigms in soft computing* (pp. 276–322). Heidelberg, Germany: Physica-Verlag.

Cleofas-Sánchez, L., García, V., Marqués, A., & Sánchez, J. (2016). Financial distress prediction using the hybrid associative memory with translation. *Applied Soft Computing*, *44*, 144–152.

Corchado, E., Abraham, A., & de Carvalho, A. (2010). Hybrid intelligent algorithms and applications. *Information Sciences*, *180*(14), 2633–2634, Including Special Section on Hybrid Intelligent Algorithms and Applications.

Corizzo, R., Pio, G., Ceci, M., & Malerba, D. (2019). DENCAST: Distributed density-based clustering for multi-target regression. *Journal of Big Data*, *6*(1), 1–27.

Cornejo-Bueno, L., Cuadra, L., Jiménez-Fernández, S., Acevedo-Rodríguez, J., Prieto, L., & Salcedo-Sanz, S. (2017). Wind power ramp events prediction with hybrid machine learning regression techniques and reanalysis data. *Energies*, *10*(11), 1–27.

Crook, J., Edelman, D., & Thomas, L. (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research*, *183*, 1447–1465.

Dahiya, S., Handa, S., & Singh, N. (2017). A feature selection enabled hybrid-bagging algorithm for credit risk evaluation. *Expert Systems*, *34*(6), e12217.

De Caigny, A., Coussement, K., & De Bock, K. W. (2018). A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. *European Journal of Operational Research*, *269*(2), 760–772.

Desai, V. S., Crook, J. N., & Overstreet, G. A. (1996). A comparison of neural networks and linear scoring models in the credit union environment. *European Journal of Operational Research*, *95*(1), 24–37.

Dhanachandra, N., Manglem, K., & Chanu, Y. J. (2015). Image segmentation using k-means clustering algorithm and subtractive clustering algorithm. *Procedia Computer Science*, *54*, 764–771.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, *10*(7), 1895–1923.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Multiple classifier systems* (pp. 1–15). Berlin, Germany: Springer.

Dua, D., & Graff, C. (2019). UCI Machine learning repository. Irvine, CA: University of california, school of information and computer science.

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, *77*(4), 802–813.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD'96, Proceedings of the second international conference on knowledge discovery and data mining* (pp. 226–231). Portland, Oregon: AAAI Press.

Falangis, K., & Glen, J. J. (2010). Heuristics for feature selection in mathematical programming discriminant analysis models. *Journal of the Operational Research Society*, *61*(5), 804–812.

Fayyad, U. M., & Irani, K. B. (1992). On the handling of continuous-valued attributes in decision tree generation. *Machine Learning*, *8*(1), 87–102.

Finlay, S. (2011). Multiple classifier architectures and their application to credit risk assessment. *European Journal of Operational Research*, *210*(2), 368–378.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7*, 179–188.

Florez-Lopez, R. (2010). Effects of missing data in credit risk scoring: A comparative analysis of methods to achieve robustness in the absence of sufficient data. *Journal of the Operational Research Society*, *61*(3), 486–501.

Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. In *ICML'96, Proceedings of the thirteenth international conference on international conference on machine learning* (pp. 148–156). San Francisco, CA: Morgan Kaufmann Publishers.

Friedman, J. H. (2000). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*, 1189–1232.
Gurney, K. (1997). *An introduction to neural networks*. Bristol, PA: Taylor & Francis.
Gurný, P., & Gurny, M. (2013). Comparison of credit scoring models on probability of default estimation for US banks. *Prague Economic Papers*, *22*, 163–181.
Hand, D., & Adams, N. (2014). Selection bias in credit scorecard evaluation. *Journal of the Operational Research Society*, *65*, 408–415.
Kaggle (2019). Kaggle.
Kajdanowicz, T., Kazienko, P., & Kraszewski, J. (2010). Boosting algorithm with sequence-loss cost function for structured prediction. In M. Graña Romay, E. Corchado, & M. T. Garcia Sebastian (Eds.), *Hybrid artificial intelligence systems* (pp. 573–580). Berlin, Germany: Springer.
Kempa, O., Lasota, T., Telec, Z., & Trawiński, B. (2011). Investigation of bagging ensembles of genetic neural networks and fuzzy systems for real estate appraisal. In N. T. Nguyen, C.-G. Kim, & A. Janiak (Eds.), *Intelligent information and database systems* (pp. 323–332). Berlin, Germany: Springer.
Kohavi, R., & Quinlan, J. R. (2002). Data mining tasks and methods: Classification–decision tree discovery. In W. Klosgen, & J. M. Zytkow (Eds.), *Handbook of data mining and knowledge discovery* (pp. 267–276). New York, NY: Oxford University Press.
Kozodoi, N., Lessmann, S., Papakonstantinou, K., Gatsoulis, Y., & Baesens, B. (2019). A multi-objective approach for profit-driven feature selection in credit scoring. *Decision Support Systems*, *120*, 106–117.
Kuncheva, L. I. (2004). *Combining pattern classifiers: methods and algorithms*. Hoboken, NY: Wiley-Interscience.
Kvamme, H., Sellereite, N., Aas, K., & Sjursen, S. (2018). Predicting mortgage default using convolutional neural networks. *Expert Systems with Applications*, *102*, 207–217.
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.
Li, C. (2017). *A gentle introduction to gradient boosting*. College of Computer and Information Science Northeastern University.
Li, X.-L., & Zhong, Y. (2012). An overview of personal credit scoring: Techniques and future work. *International Journal of Intelligence Science*, *2*(4), 181–189.
Liang, D., Lu, C.-C., Tsai, C.-F., & Shih, G.-A. (2016). Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study. *European Journal of Operational Research*, *252*(2), 561–572.
Lim, M. K., & Sohn, S. Y. (2007). Cluster-based dynamic scoring model. *Expert Systems with Applications*, *32*(2), 427–431.
Lin, Y., Yen, C., & Wang, J. (2020). Video popularity prediction: An autoencoder approach with clustering. *IEEE Access*, *8*, 129285–129299.
Lisuwan, P., Boonserm, P., & Sinapiromsaran, K. (2017). Extreme anomalous score clustering algorithm. In *ICIT 2017, Proceedings of the 2017 international conference on information technology* (pp. 66–70). New York, NY: Association for Computing Machinery.
Liu, C., Xie, J., Zhao, Q., Xie, Q., & Liu, C. (2019). Novel evolutionary multi-objective soft subspace clustering algorithm for credit risk assessment. *Expert Systems with Applications*, *138*, 112827.
Luo, X. (2020). A novel clustering-enhanced adaptive artificial neural network model for predicting day-ahead building cooling demand. *Journal of Building Engineering*, *32*, Article 101504.
Luo, S.-T., Cheng, B.-W., & Hsieh, C.-H. (2009). Prediction model building with clustering-launched classification and support vector machines in credit scoring. *Expert Systems with Applications*, *36*(4), 7562–7566.
Mai, F., Tian, S., Lee, C., & Ma, L. (2019). Deep learning models for bankruptcy prediction using textual disclosures. *European Journal of Operational Research*, *274*(2), 743–758.
Malekipirbazari, M., & Aksakalli, V. (2015). Risk assessment in social lending via random forests. *Expert Systems with Applications*, *42*(10), 4621–4631.
Malhotra, R., & Malhotra, D. (2003). Evaluating consumer loans using neural networks. *Omega*, *31*(2), 83–96.
Martínez-Martínez, J. M., Escandell-Montero, P., Soria-Olivas, E., Martín-Guerrero, J. D., Magdalena-Benedito, R., & Gómez-Sanchis, J. (2011). Regularized extreme learning machine for regression problems. *Neurocomputing*, *74*(17), 3716–3721.
Moro, A., & Fink, M. (2013). Loan managers' trust and credit access for SMEs. *Journal of Banking & Finance*, *37*(3), 927–936.
Negahban, A. (2019). Simulation-based estimation of the real demand in bike-sharing systems in the presence of censoring. *European Journal of Operational Research*, *277*(1), 317–332.
Okun, O. (2011). *Feature selection and ensemble methods for bioinformatics: algorithmic classification and implementations*. Hershey, PA: IGI Publishing, Information Science Reference.
Paleologo, G., Elisseeff, A., & Antonini, G. (2010). Subagging for credit scoring models. *European Journal of Operational Research*, *201*(2), 490–499.
Papouskova, M., & Hajek, P. (2019). Two-stage consumer credit risk modelling using heterogeneous ensemble learning. *Decision Support Systems*, *118*, 33–45.
Pérez Martín, A., Pérez-Torregrosa, A., & Lamata, M. (2018). Big data techniques to measure credit banking risk in home equity loans. *Journal of Business Research*, *89*, 448–454.
Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81–106.
Russell, S., & Norvig, P. (2009). *Artificial intelligence: a modern approach* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.
Schubert, E., Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (2017). DBSCAN Revisited, revisited: Why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems*, *42*(3).
Scikit Learn (2019). *Ensemble methods: gradient boosting*. Scikit Learn.
Siddiqi, N. (2005). *Credit risk scorecards: developing and implementing intelligent credit scoring*. North Carolina, USA: John Wiley & Sons.
Soui, M., Gasmi, I., Smiti, S., & Ghédira, K. (2019). Rule-based credit risk assessment model using multi-objective evolutionary algorithms. *Expert Systems with Applications*, *126*, 144–157.
Sun, J., Li, H., Chang, P.-C., & Huang, Q.-H. (2015). Dynamic credit scoring using B & B with incremental-SVM-ensemble. *Kybernetes*, *44*(4), 518–535.
Thomas, L. (2010). Consumer finance: Challenges for operational research. *Journal of the Operational Research Society*, *61*(1), 41–52.
Thomas, L. C., Oliver, R., & Hand, D. (2005). A survey of the issues in consumer credit modelling research. *Journal of the Operational Research Society*, *56*, 1006–1015.
Torre, E., Marelli, S., Embrechts, P., & Sudret, B. (2019). Data-driven polynomial chaos expansion for machine learning regression. *Journal of Computational Physics*, *388*, 601–623.
USA Federal Reserve (2020). Borrowing by businesses and households: Statistics.
Utkin, L. V. (2019). An imprecise extension of SVM-based machine learning models. *Neurocomputing*, *331*, 18–32.
Van Gestel, T., & Baesens, B. (2008). *Credit risk management–basic concepts: financial risk components, rating analysis, models, economic and regulatory capital*. New York, NY: Oxford University Press.
Vanneschi, L., Horn, D. M., Castelli, M., & Popovic, A. (2018). An artificial intelligence system for predicting customer default in e-commerce. *Expert Systems with Applications*, *104*, 1–21.
Wang, G., Ma, J., Huang, L., & Xu, K. (2012). Two credit scoring models based on dual strategy ensemble trees. *Knowledge-Based Systems*, *26*, 61–68.
Wermter, S., & Sun, R. (2014). *Hybrid neural systems*. Berlin, Germany: Springer.
West, D. (2000). Neural network credit scoring models. *Computers & Operations Research*, *27*(11), 1131–1152.
Wiginton, J. C. (1980). A note on the comparison of logit and discriminant models of consumer credit behavior. *Journal of Financial and Quantitative Analysis*, *15*(03), 757–770.
Witten, I. H., Frank, E., & Hall, M. A. (2011). *Morgan kaufmann series in data management systems, Data mining: practical machine learning tools and techniques* (3rd ed.). Amsterdam: Morgan Kaufmann.
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.
Xu, X., Zhou, C., & Wang, Z. (2009). Credit scoring algorithm based on link analysis ranking with support vector machine. *Expert Systems with Applications*, *36*(2, Part 2), 2625–2632.
Zhang, W., He, H., & Zhang, S. (2019). A novel multi-stage hybrid model with enhanced multi-population niche genetic algorithm: An application in credit scoring. *Expert Systems with Applications*, *121*, 221–232.
Zhou, L., Lai, K. K., & Yu, L. (2010). Least squares support vector machines ensemble models for credit scoring. *Expert Systems with Applications*, *37*(1), 127–133.
Zhou, L., Si, Y.-W., & Fujita, H. (2017). Predicting the listing statuses of Chinese-listed companies using decision trees combined with an improved filter feature selection method. *Knowledge-Based Systems*, *128*, 93–101.