A comparative study of forecasting corporate credit ratings using neural networks, support vector machines, and decision trees

Parisa Golbayani[a], Ionuţ Florescu[b], Rupak Chatterjee[c]

[a] *Financial Engineering, School of Business, Stevens Institute of Technology, United States*
[b] *Hanlon Financial Systems Center, Financial Engineering, School of Business, Stevens Institute of Technology, United States*
[c] *Department of Physics, Center for Quantum Science and Engineering, Schaefer School of Engineering and Science, Stevens Institute of Technology, United States*

ARTICLE INFO

ABSTRACT

Credit ratings are one of the primary keys that reflect the level of riskiness and reliability of corporations to meet their financial obligations. Rating agencies tend to take extended periods of time to provide new ratings and update older ones. Therefore, credit scoring assessments using artificial intelligence has gained a lot of interest in recent years. Successful machine learning methods can provide rapid analysis of credit scores while updating older ones on a daily time scale. Related studies have shown that neural networks and support vector machines outperform other techniques by providing better prediction accuracy. The purpose of this paper is two fold. First, we provide a survey and a comparative analysis of results from literature applying machine learning techniques to predict credit rating. Second, we apply ourselves four machine learning techniques deemed useful from previous studies (Bagged Decision Trees, Random Forest, support vector machine and Multilayer Perceptron) to the same datasets. We evaluate the results using a 10-fold cross validation technique. The results of the experiment for the datasets chosen show superior performance for decision tree based models. In addition to the conventional accuracy measure of classifiers, we introduce a measure of accuracy based on notches called "Notch Distance" to analyze the performance of the above classifiers in the specific context of credit rating. This measure tells us how far the predictions are from the true ratings. We further compare the performance of three major rating agencies, Standard & Poors, Moody's and Fitch where we show that the difference in their ratings is comparable with the decision tree prediction versus the actual rating on the test dataset.

1. Introduction

Machine learning models have been widely applied on a range of applications in financial fields due to their capability of detecting embedded trends in financial data. For example, asset prices follow a random, non linear dynamic pattern due to many factors such as political events, economic conditions and traders behavior (Huang, Nakamori, & Wang, 2005). Despite this non linear nature of price movements, machine learning techniques provide an accurate estimation of asset prices (Kim, 2003; Huang et al., 2005; Kanchymalay, Salim, Sukprasert, Krishnan, & Hashim, 2017; Ganguli & Dunnmon, 2017; Culkin & Das, 2017; Feng, He, & Polson, 2018; Gu, Kelly, & Xiu, 2019; Sambasivan & Das, 2017). In addition to asset pricing, machine learning techniques gained a lot of interest in credit rating prediction problem. Credit ratings have been used as an informative indicators of the level of riskiness of companies and bonds (Huang, Chen, Hsu, Chen, & Wu, 2004; Lee, 2007). They reflect the likelihood that a counterparty will default

on his/her financial obligation (Martens et al., 2010). Rating agencies such as Standard and Poor's and Moodys analyze various aspects of companies financial status to come up with these credit ratings (Huang et al., 2004). Such an assessment is an expensive complicated process often taking months since it involves many experts to analyze different variables that reflect the underlying corporations' reliability (Hajek & Michalak, 2013). One solution to reduce this financial and time cost is to come up with a predictive quantitative model based on the historical financial information of a company.

In this study we provide a comparative analysis of the literature of corporate credit rating prediction problem. We focus on modern statistical and Artificial Intelligence techniques. We review and summarize the performance of techniques that were demonstrated to be the most accurate in the literature. We implement these most accurate techniques and compare their performance when analyzing three different sectors of the US economy. To our knowledge, this is one of the first papers that implements corporate rating models on the same data and thus compares algorithms' performance on an equal footing.

Corporate credit rating is a different problem than the typical image classification techniques. Thus, we introduce a new evaluation measure: the *Notch Distance* in Section 4.2 on page 22. This measure allows us to better quantify and distinguish between performance of various machine learning methods.

This paper is organized as follows. In Section 2, we provide a review of prior literature on the credit rating problem. Section 3 contains an explanation of each machine learning method studied in order to predict corporate credit ratings. Section 4 explains the nature of the input and output data sets in this study as well as the evaluation methodology used. Experimental results and analysis are provided in Section 5. Final conclusions are drawn in Section 6, together with future applications of the methodology.

2. Prior methods on the credit rating problem using machine learning techniques

Bond Rating. The majority of literature on the credit rating problem is focused on predicting bond's probability of default and thus bond rating (Saha & Waheed, 2017). Garavaglia (Garavaglia, 1991) simulated S&P corporate bond rating using the unidirectional version of the counter-propagation neural network. This neural network model introduced by Robert Hetch-Nielsen. Chaveesuk et al. (Chaveesuk, Srivaree-Ratana, & Smith, 1999) studied three supervised neural networks concepts, backpropagation, radial basis function and learning vector quantization. However, they concluded, neither a neural network nor a regression model performs well on bond rating assignments. Shin and Han (Shin & Han, 2001) combined inductive learning approach based on decision trees with case indexing process in order to build a knowledge-based system that can classify Korean corporate bond ratings. Kim & Han (Kim & Han, 2001) clustered financial data among bond rating groups. They used competitive artificial neural network to generate the centroid values of the clusters. They perform a case-indexing study based on these cluster information and combine it with the classification technique in order to present a hybrid model for forecasting corporate bond ratings. They apply their method on 167 financial variables including five categorical variable and 162 financial ratios. After performing some statistical tests such as ANOVA, factor analysis and step-wise method, they select 13 financial ratios as the final input data. Abdou (Abdou, 2009) studied feed forward neural network to evaluate consumer loans within the Egyptian private banks. Lessman et al. Lessmann, Baesens, Seow, and Thomas, 2015 provided a deep analysis of machine learning techniques used in forecasting retail credit scoring. They studied 41 classification algorithms used to predict probability of default for retail loans. The authors used multiple datasets and criteria for evaluation and concluded that their hybrid method is superior to other methods. Gangolf et al. Gangolf, Dochow, Schmidt, and Tamisier, 2016 provided a comparison of statistical and artificial intelligence techniques used for forecasting probability of default for bonds. In case of predicting probability of default, Angelini et al. (Angelini, di Tollo, & Roli, 2008) studied two different neural networks to evaluate credit risk of Italian small businesses. They showed that neural network architectures can be useful for estimating the probability of default of borrowers. Tsai et al. (Tsai & Wu, 2008) studied the application of multilayer perceptron network on bankruptcy prediction and credit risk evaluation. Their analysis proved that single neural network provides accurate results compared to multi-layers classifiers. Addo et al. (Addo, Guegan, & Hassani, 2018) studied seven binary classifiers (elastic net, random forest, gradient boosting model and four different neural network architectures) in predicting loan default probability. They implemented these models on 117019 data points in the year 2016 and 2017 provided from European banks. They used 60% of data as training set, 20% as validation and 20% as the test set. They compared the performance of these models on a selected subset of important features based on variable importance of each model. They concluded three important points. First, that decision tree based models outperform multilayer artificial neural networks. Second, that standard performance criteria such as AIC or R2 are not sufficient enough to compare different models. Third, that the selected subset of important features based on variable importance of models, does not necessarily provide stable results.

Corporate credit rating. In this work we are concerned with *corporate credit rating*. Although, the two problems are related, the bond ratings are not necessarily the same as the issuing company's corporate rating due to the priority of claim of various bonds. There are many more bonds than there are corporations and thus the data availability is much more limited. When rating publicly traded corporations the input is based generally on the 10-Q and 10-K statements. However, bonds are traded and thus bond values can be marked to market far more frequently than quarterly. Furthermore, while the basic bond structure is similar for all issuing companies regardless of sector, the quarterly statement issued by a bank is fundamentally different from that issued by a company from say the Consumer Staples sector. Thus papers dedicated to corporate credit rating are generally constructing models for companies from a specific sector. The lack of large data sets is thus typical of this domain and this makes the credit rating problem challenging.

Methodologically, statistical models and artificial intelligence approaches are used in the field of credit rating prediction. In this section, we provide a review of prior and relevant literature of models used. We start from standard models and later present more recent and complex models. We summarize these studies in Tables 1–3. We are basing our subsequent analysis on the accuracy numbers in these tables to determine which models will be implemented on the financial statements data.

Table 1

Comparison results for ANN approach.

Reference	Year	Num. of Input Var.	Input Data	Credit rating Provider	Accuracy
(Garavaglia, 1991)	1991	87	797 Companies	S&P	85%
(Moody & Utans, 1994)	1994	10	196 Indudtrial Firms	S&P	36%
(Huang et al., 2004)	2004	5	36 Commercial Banks	S&P	80.00%
(Huang et al., 2004)	2004	6 & 16	Taiwan Dataset	Taiwan Ratings Corporation	75.68%
(Kim, 2005)	2005		1080 US companies	S&P	84%
(Brabazon & O'Neill, 2006)	2006	8	791 Non-financial US Firms	S&P	85%
(Cao et al., 2006)	2006	17	US Manufactoring Companies	S&P	80.28%
(Lee, 2007)	2007	10	3017 Korean Companies	Korea Information Service, Inc	59.93%
(Yu et al., 2008)	2008	13	Jap. Consumer Credit Card	Application Approvals	88.08%
(Yu et al., 2008)	2008	12	60 UK Corporations	UK Ratings	85.87%
(Khashman, 2010)	2010	24	German Dataset	German Ratings	85.9%
(Kim & Sohn, 2010)	2010	10	Korean SMEs	Korean Default Rates	64.23%
(Hájek, 2012)	2012	52	852 US Mining Companies	S&P	80%
(Hájek, 2012)	2012	14	154 US Municipalities	Moody's	71.62%
(West, 2000)	2000	24	German and Australian Data	German and Australian Rating	error:0.2243
(Khemakhem & Boujelbene, 2015)	2015	15	86 Tunisian companies	Tunisian credit rating	82.55%
(Zhao et al., 2015)	2015	20	German Credit Dataset	German Credit Ratings	87%
(Hájek & Olej, 2014)	2014	20	Investment and non-investment grade	S&P	88.44%
(Addo et al., 2018)	2018	181 and 10	117019 companies asking for a loan	European Bank	RMSE: 0.044
(Daniel et al., 2019)	2019	17	Random US companies	S&P and Moody's	83%
(Wallis et al., 2019)	2019	27	308 companies from S&P 500	Moody's	63.6%

Table 2

Comparison results for SVM approach.

Reference	Year	Num. of Input Var.	Input Data	Credit rating Provider	Accuracy
(Huang et al., 2004)	2004	14	36 Commercial Banks	S&P	80.00%
(Huang et al., 2004)	2004	6	Taiwan Dataset	Taiwan Ratings Corporation	79.73%
(Cao et al., 2006)	2006	17	US Manufactoring Companies	S&P	84.61%
(Lee, 2007)	2007	10	3017 Korean Companies	Korea Information Service, Inc	67.22%
(Ye et al., 2008)	2008	33	Four US industries	Moody's	64%
(Yu et al., 2008)	2008	13	Jap. Consumer Credit Card	Application Approvals	79.91%
(Yu et al., 2008)	2008	12	60 UK Corporations	UK Ratings	77.84%
(Kim & Sohn, 2010)	2010	10	Korean SMEs	Korean Default Rates	66.16%
(Hájek & Olej, 2011)	2011	5	852 US firms	S&P	85.96%
(Hájek & Olej, 2011)	2011	12	452 Czech municipals	local experts	89.76%
(Kim & Ahn, 2012)	2012	14	Korea Dataset	Korea Ratings Corporation	67.98%
(Yeh et al., 2012)	2012	18	2470 Taiwanese companies	Taiwanese Rating	74.4%
(Hájek & Olej, 2014)	2014	20	Investment and non-investment grade	S&P	87.28%
(Pai et al., 2015)	2015	18	Taiwan Dataset	Taiwan Ratings Corporation	86%
(Wallis et al., 2019)	2019	27	308 companies from S&P 500	Moody's	60.1%

Table 3

Comparison results for Tree-based approach.

Reference	Year	Num. of Input Var.	Input Data	Credit rating Provider	Accuracy
(Shin & Han, 2001)	2001	12	Korean Companies	Korean Ratings	69.3%
(Koh et al., 2004)	2004				74.2%
(Ye et al., 2008)	2008	33	Manufacturing/Wholesale/Retail/Services	Moody's	47.1%
(Hájek & Olej, 2014)	2014	20	Investment and non-investment firms	S&P	86.13%
(Veronezi, 2016)	2016	33	Healthcare, Financial, Technology Sectors in USA	S&P and Moody's	>70%
(Addo et al., 2018)	2018	181 and 10	117019 companies asking for a loan from a bank	European Bank	RMSE: 0.044
(Wallis et al., 2019)	2019	27	308 companies from S&P 500	Moody's	64.6%
(Moscatelli et al., 2019)	2019	26	Italian non-financial firms	Italian Credit Register	>84%

Statistical Methods. Standard statistical techniques such as logistic regression analysis (LRA) and discriminant analysis (DA) have been used to construct credit scoring models (Horrigan, 1966; West, 1970; Kaplan & Urwitz, 1979; Kim, Weistroffer, & Redmond, 1993; Kamstra, Kennedy, & Suan, 2001; Oelerich & Poddig, 2006; Abdou, 2009; Akkoç, 2012; Abdou, Tsafack, Ntim, & Baker, 2016). Modified version of these techniques have been developed to improve credit scoring models. Hwang, Cheng, and Lee (2009) and Hwang, Chung, and Chu (2010) modify the usual ordered probit model by replacing the linear regression function with a semiparametric function to increase the choices of regression function. They apply this prediction method on four market driven variables, nineteen accounting variables and industry indicator variables. Their conclusion shows that an ordered semiparametric probit model

(OSPM) provides a better prediction than ordered linear probit model (OLPM). However, different assumptions behind most of these statistical approach, such as linearity, normality and distribution-dependency limit their predictive capability comparing to Artificial Intelligence techniques (Huang et al., 2004; Pai, Tan, & Hsu, 2015).

Nonetheless, recent papers are still using traditional statistics methods and new developments. Gogas, Papadimitriou, and Agrapetidou (2014) used ordered probit regression model to forecast long term bank credit rating. The benchmark was Fitch credit rating in 2012. Irmatova (2016) introduced a new rating model based on principal component analysis and k-means clustering algorithm to predict 30 country's rating. Karminsky and Khromova (2016) implemented an ordered probit regression model to predict bank ratings using the Bankscope database. Hajek, Olej, and Prochazka (2016) developed a methodology based on latent semantic analysis to extract topical content from company-related documents. The authors use naive bayesian network to forecast corporate credit ratings. Recently, (Petropoulos, Chatzis, & Xanthopoulos, 2016) developed a methodology based on Student's-t Hidden Markov Models (SHMMs) to investigate the temporal dimension and heavy tailed distribution of firm-related data.

We next present three prevalent AI techniques used for corporate credit rating: Artificial Neural Networks (ANN), Support Vector Machines (SVM) and Classification and Regression Trees (CART).

Artificial Neural Networks (ANN). Artificial Neural Networks draws attention from many researchers with its robustness, flexibility and higher accuracy (Dutta & Shekhar, 1988; Surkan & Singleton, 1990; Kim et al., 1993; Maher & Sen, 1997; Kwon, Han, & Lee, 1997; Brennan & Brabazon, 2004; Abdou, 2009; Abdou et al., 2016). West (West, 2000) implemented five neural network architectures: multi-layer perceptron, mixture-of-experts, radial basis function, learning vector quantization and fuzzy adaptive resonance. He compared the results of these neural networks with linear discriminant analysis, logistic regression, k nearest neighbor, kernel density estimation, and decision trees. They studied German and Australian credit scoring data and conclude that multi-layer perceptron is not the most accurate neural network architecture and logistic regression outperforms all neural network architectures. Kim (Kim, 2005) applied adaptive learning networks (ALN) which is a non-parametric model, on both financial and non financial variables to predict S&P credit ratings. Yu et al. (Yu, Wang, & Lai, 2008) proposed a six stage neural network ensemble learning model to assess a credit risk measurement on Japanese consumer credit card applications approval and UK corporations. Kashman (Khashman, 2010) investigated three neural networks based on back propagation learning algorithm on German Credit Approval dataset. The architecture of these neural networks are different according to various parameters used in the model such as hidden units, learning coefficients, momentum rate and random initial weight range. Pacelli and Azzollini (Pacelli & Azzollini, 2011) provided an overview of different types of neural networks used in credit rating literature.

Support Vector Machines. Among all artificial intelligence techniques, Support Vector Machine (SVM) has indicated a powerful classification ability (Cortes & Vapnik, 1995; Kim & Sohn, 2010; Vapnik, 2013; Xiao, Xiao, & Wang, 2016). However the standard SVM, addresses the two-class classification problems. For multi-class classification problems, SVM can be represented as several two class tasks such as one-against-one (OAO-SVM) (Schölkopf, Burges, & Smola, 1999), one-against-all (OAA-SVM) (Bottou et al., 1994) and direct acyclic graph (DAG-SVM) (Platt, Cristianini, & Shawe-Taylor, 2000; Wang, Yuan, Liu, Yu, & Li, 2009). Cao et al. (Cao, Guan, & Jingqing, 2006) studied all these three approaches on US companies from the manufacturing sector. Hajek & Olej (Hájek & Olej, 2011) studied Support Vector Machines with supervised learning as well as kernel-based approach with semi-supervised learning in order to predict corporate and municipal credit ratings. Kim & Ahn (Kim & Ahn, 2012) studied a new type classifier based on Support Vector Machine, called (OMSVM) which applies ordinal partwise partitioning (OPP) to extend the binary SVM.

Decision Trees. According to the literature, some researchers focus on the higher level of interpretability and feature extraction flexibility of decision tree-based approach in multi-class classification problems (Lee, Chiu, Chou, & Lu, 2006; Kao, Chiu, & Chiu, 2012; Abdou et al., 2016). The fundamental of these techniques are introduced first by (Breiman, Friedman, Olshen, & Stone, 1984). Paleologo et al. (Paleologo, Elisseeff, & Antonini, 2010) proposed a missing data imputation method and ensemble classification technique, subagging, particularly for unbalanced credit rating data sets. The proposed subagging technique is based on different models such as decision trees and support vector machines. Khandani et al. (Khandani, Kim, & Lo, 2010) studied general classification and regression tree technique (CART) on a combination of traditional credit factors and consumer banking transactions to predict consumer credit risk. Pedro Veronezi (Veronezi, 2016) applied Random Forest(RF) and Multi Layer Perceptron (MLP) to predict corporate credit ratings using their financial data. First, he used RF to perform feature selection and then the result was fed to two different RF models. One model aims to predict the changes and another model predicts the direction. The result of these two models have been added to the input dataset. After creating a new dataset, two models were trained: a MLP model and a RF model. He concluded that RF outperforms MLP by providing a more accurate and stable results in a shorter period of time.

Hybrid Models. In addition to all these frequently used techniques, some researchers study other approaches to provide a credit scoring model. Peng et al. (Peng, Wang, Kou, & Shi, 2011) introduced three multiple criteria decision making (MCDM) methods to evaluate classification algorithms for financial risk prediction. Chen (Chen, 2012) investigated rough set theory (RST) approach to classify Asian banks' ratings. Some researchers take one step further and integrate multiple techniques to achieve a higher accuracy. Yeh et al. (Yeh, Lin, & Hsu, 2012) combined random forest feature selection with different approach such as rough set theory (RST) and SVM. They tested these techniques on 17 financial variables in addition to the market based information which has been obtained by Moody's KMV tool for each corporation. Wu & Hsu (Wu & Hsu, 2012) introduced an enhanced decision support model which is a combination of relevance vector machine and decision tree. Hajek (Hájek, 2012) used fuzzy rule-based system adapted by a feed-forward neural network to classify credit ratings of US municipalities and companies. Pai et al. (Pai et al., 2015) used Decision Tree Support Vector Machine (DTSVM) integrated to multiple feature selection strategies and rough set theory (RST) to predict credit ratings. Description for other hybrid techniques are available in (Chandra, Ravi, & Bose, 2009; Akkoç, 2012; Chen & Cheng, 2013).

Evaluation and Comparison. The Comparison of the discussed techniques with different parameters and architectures on several datasets, have been widely studied in credit scoring field. Huang et al. (Huang et al., 2004) studied support vector machines (SVM)

and back-propagation neural network (BNN) to obtain more accurate credit rating predictions for Taiwan financial institutes and United States commercial banks. Lee (Lee, 2007) used 5-fold cross-validation to find the optimal parameters of SVM, multiple discriminant analysis (MDA), case-based reasoning (CBR), and three-layer fully connected back-propagation neural networks (BPNs) on Korean Datasets. Ye et al. (Ye, Liu, & Li, 2008) compared the performance of four different approaches: Linear Regression, Bagged decision tree with Laplace smoothing, SVM and Proximal SVM (PSVM) over four industries manufacturing, wholesale, retail and services. Ryser & Denzler (Ryser & Denzler, 2009) used k-fold cross validation procedure to compare the performance of different credit rating models. They concluded that although non-parametric approaches such as random forest, neural network and generalized additive models still outperform the parametric models, they may overfit during the training process. Hajek & Michalak (Hajek & Michalak, 2013) studied the performance of supervised machine learning techniques such as Multilayer Perceptron (MLP), Support Vector Machines (SVM), radial Basis Function Neural Network (RBF), Naive Bayes (NB), Random Forest (RF), Linear Discriminant Classifier (LDC) and Nearest Mean Classifier (NMC) on both US and European non-financial companies. They focused on 81 financial and non-financial variables of 852 US companies and 43 financial and non-financial variables of 244 European companies. They reduced the dimension of data by applying feature selection techniques such as wrapper algorithms. They concluded that US credit ratings are more affected by the size of companies and market value ratios while the European ratings depend more on profitability and leverage ratios. Zhong et al. (Zhong, Miao, Shen, & Feng, 2014) provided a comparison study over reliability, overfitness and error distribution of different learning algorithms such as SVM, Back propagation neural network, Extreme Learning Machine (ELM) and Incremental Extreme Learning (IELM). They considered 6 financial variables on both S&P and Moody's rating categories to classify the US corporate credit ratings.

Bahrammirzaee (Bahrammirzaee, 2010) provided a comparative analysis of the application of three categories of techniques: artificial neural networks, expert systems and hybrid intelligent systems on credit evaluation. Their analysis categorized articles where artificial neural network outperform other methods as well as articles where neural network performs as same as or even worse than other traditional methods such as decision trees and linear regression. Dima et al. (Dima & Vasilache, 2016) compared the performance of neural networks with Bayesian approach on a large sample of companies (3000 companies) applying for credit to the same bank in Romania. Khemkhem et al. (Khemakhem & Boujelbene, 2015) implemented neural networks and linear discriminant analysis to forecast Tunisian corporate credit ratings. They concluded neural networks outperform discriminant analysis. Hajek et al. (Hájek & Olej, 2014) compared the performance of artificial immune classification algorithms (AIRS1, AIRS2, AIRS2-p and CSCA) and several well-known machine learning techniques (MLP, Decision tree, SVM and RBF) on investment and non-investment grade firms. Both groups of input variables were collected for 520 U.S. companies in the year 2010, 195 classified as investment grade and 325 as non-investment grade. They used accuracy and misclassification cost as performance measures of the studied classifiers. They also provided confusion matrix for each classifier (TP, TN, FP, FN). First, they conclude that AIRS2 outperforms AIRS1, AIRS2-p and CSCA on the test set based on the studied measures. They also show that MLP and Decision tree are the best models by providing the highest accuracy and the lowest misclassification cost, respectively.

Wallis et al. (Wallis, Kumar, & Gepp, 2019) provided a comparative analysis of several most popular machine learning techniques (multinomial logistic regression, linear discriminant analysis, regularized discriminant analysis, artificial neural network, support vector machine, gaussian process classifier, random forest, gradient boosting machine) to predict Moody's lone term credit rating. They studied these models on 308 of the S&P 500 companies from January 2016 to November 2017. They concluded the top three performing models were random forest, artificial neural network and support vector machines where random forest outperforms the others with 64.6% accuracy.

Moscatelli et al. (Moscatelli, Narizzano, Parlapiano, & Viggiano, 2019) applied random forest, gradient boosted trees, logistic regression, penalized logistic regression and linear discriminant analysis to predict default probability of Italian non-financial firms from 2011–2017. They studied economic and financial ratios as potential indicators of corporate defaults. They also added a set of credit behavioral indicators on the firm-bank relationship to the studied financial ratios. They concluded random forest outperform all other models in predicting default/non-default credit rating.

A summary of discussed approach including number of financial variables, input dataset, output data set provider and the corresponding accuracy is provided in tables below. Table 1 refers to ANN-based approach. Table 2 addresses SVM-based models and Table 3 represents Decision tree-based methods.

3. Analytical methods

Supervised learning methods are based on learning from observations. Let $X \subset R^n$ and Y be the input and output space respectively. The output space identifies the classification type. When $Y = \{-1, +1\}$, the problem is a binary classification. If $Y = \{1, 2, ..., m\}$, the problem is a multiple class classification. $Y \subset R^n$ leads to a regression type problems. During the learning process, the classifier assumes there exists a hidden function $Y = f(x)$ in a training set $L = \{(x_i, y_i)\}_{i=1}^{N}$. The classifier attempts to find a prediction for this function such that the error is minimized.

There are multiple ways to determine the exact nature of function f. In computer science and statistics these are generically named machine learning algorithms. We are focusing an a few specific machine learning algorithms described in this section. The previous section and the summary Tables 1–3 detail these methods as the most popular in credit rating problem. This is the reason we chose them for implementation.

3.1. Bagged Decision Trees (BDT)

Let L be the learning/training set consists of N instances. Bagging or bootstrap aggregation algorithm creates B bootstrap samples where each sample consists of $n < N$ randomly chosen instances but with replacement from L. Each of these B bootstrap samples train the same classifier which in this case is a decision tree. In order to predict an outcome for each instance of the testing set, the new case must be run down each of the B decision trees. The prediction is obtained by a maximum number of votes among the B classifiers.

This ensemble structure combines predictions of multiple classifiers to provide a better performance than unique learner. The combination of learners in bagging structure reduces the risk of classification error (variance) for those unstable high variance classifiers such as decision trees.

In Bagged Decision Trees (BDT), the possibility of over-fitting of individual tree is less concerned. For this reason, the individual decision trees are grown deep and the trees are not pruned. Bagged Decision tree has only two hyper-parameters, the number of trees and the number of samples. However, they may have many similarities which leads to a high correlation in predictions. The structure of Random Forest will address this problem. A more detailed description of these methods can be found in (Breiman et al., 1984).

3.2. Random Forest (RF)

Random Forest (RF) is another algorithm based on ensemble of classification trees which is developed by Breiman (Breiman et al., 1984; Breiman, 2001). The only difference between RF and BDT is that RF takes one extra step. In addition to taking the random subset of data, it also chooses randomly a subset of X at each node and calculate the best split at that node only within the given subset of X. This structure provides uncorrelated or weakly correlated predictions. Also, there is no pruning step, which means all the trees of the forest are grown deep. RF has only two hyper-parameters, the number of variables in the random subset at each node and the number of trees in the forest. Moreover, RF ranks variables by the importance of a variable based on the classification accuracy, while considering the interaction between variables.

3.3. ANN: Multi Layer Perceptron (MLP)

A perceptron is a linear classifier that separates two classes with a straight line. Let X be the input vector and y be a perceptron that produces a single output:

$$y = \Phi\left(\sum_{i=1}^{n} w_i x_i + b\right) = \Phi(w^T x + b)$$

where w is a vector of input weights, b is a bias and Φ denotes a non-linear activation function.

A multilayer perceptron is a deep, artificial neural network composed of more than one perceptron. In this NN, the input layer receives the input vector, the output layer makes a decision and provide a prediction, and in between these two, there exists a number of hidden layers. The input signal moves forward from the input layer to the output layer, through the hidden layers. Let l represents the number of input units, m represents the number of hidden units and n represents the number of output units. Each hidden layer unit is defined as:

$$h_j = f\left(\sum_{i \to j} w_{ij}^{(1)} x_i + w_{0j}^{(1)}\right) = f(w^{(1)} x)$$

where f is an activation function on hidden layers, $i = \{1, 2, ..., l\}$ and $j = \{1, 2, ..., m\}$. The common choice for f is a logistic function $f(z) = \frac{e^z}{1 + e^z}$. this hoice is important and depends on the nature of the data.

Similarly, the output units are denoted as:

$$y_k = g\left(\sum_{j \to k} w_{jk}^{(2)} h_j + w_{0k}^{(2)}\right)$$

where g is another activation function and $k = \{1, 2, ..., n\}$. Again, the common choices are logistic functions.

Considering the layers:

$$y_k(x, (w^{(1)}, w^{(2)})) = g\left(w_{0k}^{(2)} + \sum_{j=1}^{m} w_{jk}^{(2)} f\left(\sum_{i=1}^{l} w_{ij}^{(1)} x_i + w_{0j}^{(1)}\right)\right)$$

The number of parameters required to build this neural network is:

$$l.\, m + m + m.\, n + n = (l + 1)m + (m + 1)n$$

So depending on the weights and the choice of activation functions, different values of outputs are obtainable.

When the decision of the output layer is computed against the real outputs, the error will be calculated. The parameters of the model, the weights and the bias terms, will be adjusted during the training in order to minimize the error.

$$\Delta w = \eta^* e_k * x_{jk}$$

where e_k represents the error on the output unit k and x_{jk} is the input that caused the error and η is the learning rate which defines how much to change the weight to correct the error.

Partial derivative of the error w.r.t. the weights and biases are back propagated through the MLP. Therefore the change of weights in the hidden layer follows the previous formula except the error term which will be computed as:

$$e_j = w_j * e_k * d(y_k)$$

where d is the transfer derivative. The network keeps doing these forward and backward passes until the error can go no lower. This state is called convergence.

3.4. Support Vector Machines (SVM)

SVM is an algorithm that implements non linear boundaries between classes by transforming the input data into a high dimensional space. This mapping into a new space is a task of kernel functions which make the input data set linearly separable. In the new space, SVM constructs a maximal margin hyperplane which provides a maximum separation between output classes. The training observations that are closest to the maximal margin hyperplane are called support vectors. It introduced first by Vapnik in a form of quadratic optimization problem with bound constrains and one linear equality constraints. The standard SVM classifier is only applicable on binary type classification problems.

$$\begin{cases} w^T \Phi(x_i) + b \geqslant +1 & \text{if } y_i = +1 \\ w^T \Phi(x_i) + b \leqslant -1 & \text{if } y_i = -1 \end{cases}$$

such that w and b represent the weight vector and bias respectively. $\Phi(): R^n \rightarrow R^{n_k}$ represents a mapping function and $w^T \Phi(x_i) + b = 0$ the optimal separating hyperplane.

The formula above is equivalent to

$$y_i[w^T \Phi(x_i) + b] \geqslant +1$$

for $\forall\, i = \{1, 2, ..., N\}$

This formula constructs two hyperplanes on the opposite sides of the optimal hyperplane with the total margin size of $\frac{2}{\|w\|^2}$.

The classifier takes a decision then based on a decision function

$$sgn(w^T \Phi(x_i) + b)$$

In order to solve the linearly non-separable problems, it is general to use a slack variable ξ which allows a miss-classification error. The optimization problem containing the weighted miss-classification error will be formed as:

$$\min_{w,b,\xi} \quad \frac{1}{2} w w^T + C \sum_{i=1}^{N} \xi_i$$

$$\text{subject to} \quad \begin{cases} y_i[w^T \Phi(x_i) + b] \geqslant 1 - \xi_i & i = 1, 2, ..., N \\ \xi_i \geqslant 0 & i = 1, 2, ..., N \end{cases}$$

where $C \in R^+$ is a tuning hyperparameter that weights the importance of miss-classification errors. This optimization problem is solvable using Lagrangian where the Lagrangian multiplier α exists for each training observation. The non-zero α's represent the support vectors in the training set.

Therefore, the formula above leads to a dual problem:

$$\max_{\alpha} \quad \frac{1}{2} \alpha^T Q \alpha - e^T \alpha$$

$$\text{subject to} \quad \begin{cases} 0 \leqslant \alpha_i \leqslant C & i = 1, 2, ..., N \\ y^T \alpha = 0 \end{cases}$$

where e is the vector of ones, $Q_{ij} = y_i y_j K(x_i, x_j)$ is a $N*N$ positive semi-definite matrix and $K(x_i, x_j) = \Phi(x_i)^T \Phi(x_j)$ is a kernel function. The choice of kernel functions depends on the type of the problem in hand. The linear kernel function is $K(x_i, x_j) = x_i^T x_j$, the radial basis kernel function is $K(x_i, x_j) = exp\{-\gamma \|x_i - x_j\|^2\}$, the polynomial kernel function with degree d is $K(x_i, x_j) = (\gamma x_i^T x_j + r)^d$ and the sigmoid kernel is $K(x_i, x_j) = tanh\{\gamma x_i^T x_j + r\}$

The final SVM classifier will be in a form of:

$$sgn\left(\sum_{i}^{N} \alpha_i y_i K(x, x_i) + b \right)$$

3.5. Multi Class Support Vector Machines

The conventional SVM explained above is a binary classifier. For multi-class classification problems, we need to modify its structure. There are different approaches that extends binary SVM structure to solve the multi-class classification problems, two of which have been explained and used in this study.

3.5.1. One versus All approach

Consider an M-class classification problem. The problem can be decomposed into M binary subproblems. One-against-all approach constructs M binary SVM classifiers, each of which separates one class from the other M-1 classes. Let N be the number of training samples $(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)$. The i-th SVM classifier is trained with all the training examples of the i-th class with positive labels, and all the others with negative labels (Liu & Zheng, 2005; Kim & Ahn, 2012).

3.5.2. One versus One approach

This approach constructs binary SVM classifiers for all pairs of classes; in total, there are $\frac{M(M-1)}{2}$ pairs. For each given pair, binary SVM classifier follows the optimization problem above to maximize the margin between classes (Kim & Ahn, 2012).

4. Research methodology

In order to have a fair assessment of the methodologies developed to predict credit rating, we need to have an equal evaluation footing for the methods. To do so we need to use the same dataset with the same structure of training, validation and testing. Second we need to have a fair measure for performance evaluation. We introduce a so called Notch measure in Section 4.2 that we believe it is appropriate for assessing the performance of credit rating algorithms. In our applications we focus on methods described in Section 3. These are the methods used by the majority of credit rating literature (see Tables 1–3).

4.1. Data sets

Previous Studies show that machine learning techniques are able to predict corporate credit ratings using their historical financial variables (Huang et al., 2004; Hájek, 2012; Pai et al., 2015). In this study, we consider 52 stocks of financial sector, 28 stocks of energy sector and 44 stocks of healthcare sector. The input data set covers these corporate historical financial variables from 1990–2018 for financial sector and from 2009 to 2018 for energy and healthcare sectors. These variables are taken from both Bloomberg and Compustat and have been merged together to reduce the number of missing values. Among all available variables, we selected 16 financial variables for financial sector and 20 variables for both energy and healthcare sectors in order to predict the credit ratings. The primary factors in the selection of these variables are the availability of data and the influence of a variable on the credit rating based on previous studies (Huang et al., 2004). The output data which are the corresponded credit ratings are taken from S&P and it contains 19 rating classes starting from AAA to CC. These variables are listed in Tables 4 and 5. Table 6 and Fig. 1 present the distribution of each sector with respect to rating classes.

4.2. Evaluation methodology

The typical way to evaluate a machine learning model is the standard accuracy as the proportion of observations correctly classified in the test data. For example, the column "Accuracy" in Tables 1–3 present a summary of these percentages in prior work. A

Table 4
Financial variables used in financial sector.

	Variables
V_1	Asset/Equity
V_2	Total Common Equity
V_3	Total Asset
V_4	Total Invested Capital
V_5	Total Debt/Total Equity
V_6	Total Debt/Total Asset
V_7	Total Liabilities
V_8	Sgort and Long term Debt
V_9	Long Term Borrow
V_{10}	Return on Asset
V_{11}	Debt/Market Cap
V_{12}	Operating Margin
V_{13}	IS-OPER-INC
V_{14}	Net Income
V_{15}	Profit Margin
V_{16}	EPS-FOR-RATIOS

Table 5
Financial ratios used in energy and healthcare sectors.

	Ratio
R_1	Debt/EBITDA
R_2	FFO/Total Debt
R_3	EBITDA/Interest
R_4	FFO/Interest
R_5	CFO/Debt
R_6	FFO/Net Profit
R_7	NWC/Revenue
R_8	Current Asset/Current Liabilities
R_9	(FFO + Cash)/Current Liabilities
R_{10}	EBITDA/Revenues
R_{11}	Cash/Total Debt
R_{12}	Total Debt/Tangible Net worth
R_{13}	Total Debt/Revenue
R_{14}	Debt/Capital
R_{15}	Cash/Asset
R_{16}	Total Fixed Capital/Total Fixed Assets
R_{17}	Equity/Asset
R_{18}	NWC/Total Assets
R_{19}	Retained Earnings/Total Assets
R_{20}	EBITDA/Total Assets

Table 6
Number of companies in each class of rating on December 2018.

Rating classes	Financial sector	Energy sector	Healthcare sector
AAA	1	1	6
AA+	3	1	1
AA	12	1	5
AA-	24	4	5
A+	30	4	10
A	34	7	16
A-	36	11	18
BBB+	33	15	19
BBB	18	18	20
BBB-	12	16	17
BB+	6	10	14
BB	4	9	13
BB-	2	5	6
B+	1	3	3
B	1	1	2
B-	1		
CCC+	1		
CCC	1		
CCC-			
CC	1		
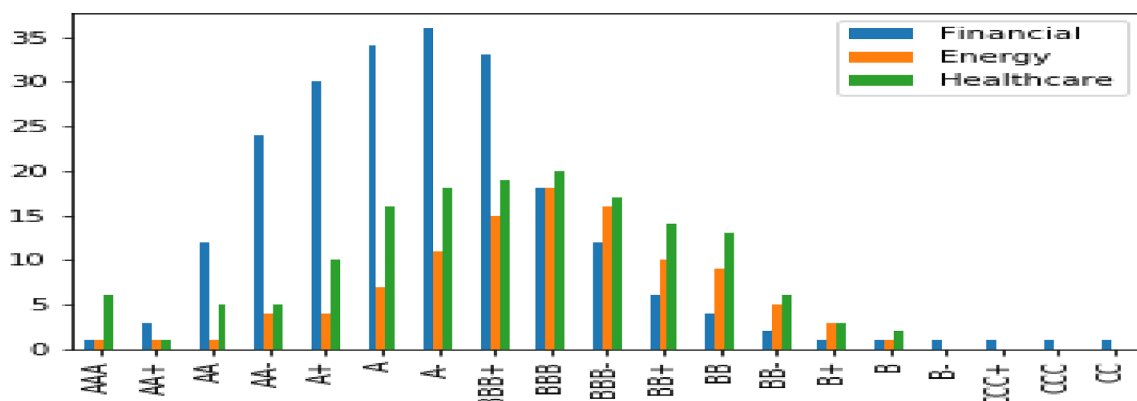
Fig. 1. Distribution of credit ratings for each sector.

quick inspection of these numbers appear to say that accurate predictions based on publicly available financial variables using machine learning techniques is achievable. However, predicting credit rating is a problem that goes beyond most other classification problems.

First, the credit rating data is temporal. That is, a prediction has to be done at a point in time and data from future quarters may not be used in prediction. Most techniques divide data into 80% training and 20% test data sampled at random which means invariably that future observations will be used in the model development. We will investigate the models performance when the test data is made of future observations.

Second issue is that corporate credit ratings do not change very often. Due to this, in about 90% of observations the credit rating does not change with respect to the previous quarter and in only 10% of observations the credit rating changes. Therefore, creating a simple model that just predicts the same credit rating as the previous quarter will have a 90% accuracy! This beats the majority of the results in Tables 1–3. Thus, we believe it is crucial to investigate the performance of the classifier on those observations where the rating is changing from the previous quarter.

Third issue is that the proportion of observations correctly classified tells us nothing about how far from the true rating were the inaccurate predictions. We quantify the distance between predictions and real ratings using what is known in literature as *Notches* distance. To create a numerical value of the distance we sort the ratings from best (AAA) to worst (CC). We give a numerical value to each rating (e.g., AAA is 1).

Definition 1 (*Notch Distance*). To formulate this distance, we let y denote the true rating of an observation. Let $\hat{y}$ denote the prediction given by a particular model. Then define a random variable: Notch Distance as $Y = \hat{y} - y$. To asses the distribution of this variable for a particular model that makes predictions $\hat{y}_1, \ldots, \hat{y}_N$, we calculate F the *frequency* of the notch i value:

$$F(i) = \sum_{k \in N} I_{(\hat{y}_k - y_k = i)}/N$$

where N is the total number of observations in the test set, $i \in \{\ldots, -1, 0, 1, \ldots\}$, and I is the indicator function taking value 1 every time the condition is satisfied and 0 otherwise.

In the definition above $F(0)$ is the proportion of observations predicted correctly. To create a measure of accuracy we can use the random variable created. Specifically, we can calculate the expected value:

$$DC = \mathbf{E}[Y] = \sum_i i * F(i)$$

and we may regard it as a *Dissimilarity Coefficient*. The numerical value is the expected notch distance corresponding to a particular method. In a similar way we may define an *Absolute Dissimilarity Coefficient*

$$ADC = \mathbf{E}[|Y|] = \sum_i |i| * F(i)$$

We can of course calculate the standard deviation of the variable Y and that would be a measure of how far the typical prediction is from the actual rating.

Theorem 1 (*Jensen's Inequality*). *Let X be a random variable with finite expectation, and let g: R → R be a convex function whose domain includes the range of X. Then,*

$$E[g(X)] \geqslant g(E[X])$$

If the function g is strictly convex, then the inequality holds with equality if and only if X is degenerate. As a consequence of Jensen's inequality,

$$E[|Zt|]^2 \leqslant E[Z^2] = \sigma^2.$$

Both these measures will be influenced by the percentage of the correct predictions (value 0). To have a more accurate measure in terms of average notch distance when the prediction is wrong we can calculate a conditional expectation:

$$\mathbf{E}[Y|Y \neq 0] = \sum_{i \neq 0} i * \frac{F(i)}{\sum_{j \neq 0} F(j)}$$

This conditional expectation eliminates the correct prediction and its expectation should be a better measure of how many notches we expect the algorithm to be off when the prediction fails.

5. Experiment results and analysis

In this section we study the performance of four Machine learning techniques: Bagged Decision Tree (BDT), Random Forest (RF), Multiple Layer Perceptron (MLP) and Support Vector Machine (SVM) techniques to predict corporate credit ratings. Figs. 2–4 illustrate some examples of the type of results obtained. Figures depict the training set (blue circle), test set (red circle) and predictions (green x) for various stocks for three methods: Bagged Decision Tree, Random Forest, and MLP.
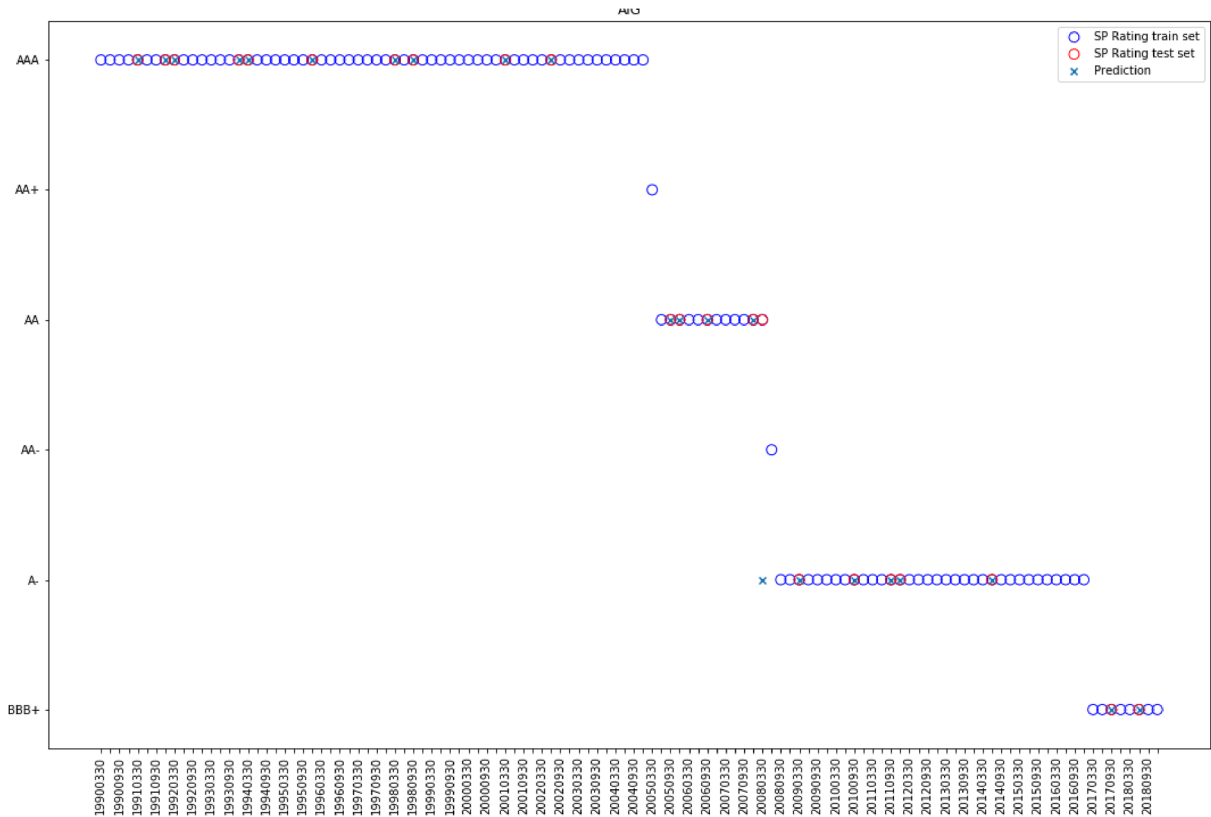
Fig. 2. Train set, Test set and Predictions of AIG using Bagged Decision Tree.

5.1. Analysis of prediction accuracy for different machine learning models

To address overfitting and obtain a more accurate results, we use a, so called, 10-fold cross validation procedure. Specifically, each sector data is split into three subsets, a training set, a validation set and a test set. For example, for financial sector dataset, a training set of 70%, which is 4220 data points, a validation set of 10% which is 603 data points and a test set of 20% which is 1206 data points of input data have been considered. Table 7 presents prediction accuracy of each of these techniques based on 10-fold cross validation procedure. The SVM technique is a binary type classification technique. However, in our case we have several categories that need to be classified not just two. Thus we use two different SVM-based methods to classify the rating categories (Kim & Ahn, 2012).

Table 7 contains the type of results that are reported by the majority of articles applying Machine Learning algorithms to the credit rating problem. According to the results we obtain it is clear that Bagged Decision Tree and Random Forest outperformed MLP and SVM-based approaches. This type of performance was observed for all datasets and periods considered. Based on the Tables 1 and 2, most of the related studies on S&P credit scores, prove that neural network and support vector machine obtain more than 80% prediction accuracy on US firms. Only one study obtained 36% accuracy of neural network on industrial firms Moody and Utans, 1994.

For a financial analysis it is important to have a measure of how far the prediction is when the algorithm fails. To this end recall we introduced the Notch Distance in Definition 1 on page 22.

A summary of the distribution of the notch distance Y is presented in Table 8. As expected values for $F(0)$ are exactly the prediction accuracy percentages in Table 7.

Table 9 provides expected value and standard deviation of the notch distance for every model. The columns represent the three sectors we are analyzing: financial sector, energy sector and healthcare sector. First we look at the expected value as an indication of symmetry of prediction. The standard deviation and expected absolute value are measure of how far the prediction is from the actual rating.

The results in the table generally indicate that Bagged decision tree and Random Forest methods have a better performance compared to the ANN and SVM. The largest proportion of predictions is accurate thus the notches distance is zero. We next calculate the conditional expectation of the notches measurement value when we are wrong. Table 10 shows this conditional expected values and the conditional standard deviation.

Looking at the standard deviation and expected absolute for the unconditional Notches distance in Table 9 it would appear that the performance for all sectors is roughly similar. However, looking at the conditional numbers in Table 10 we see that healthcare
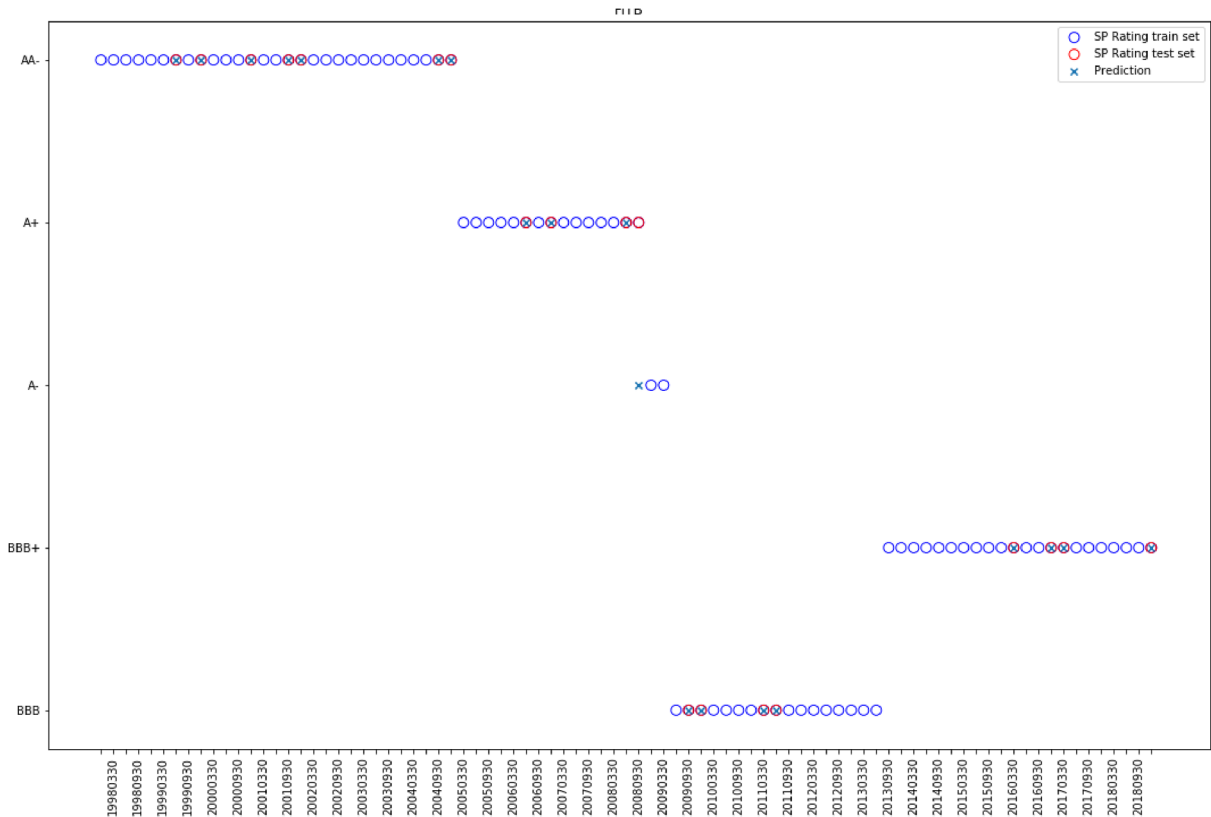
Fig. 3. Train set, Test set and Predictions of FITB using Random Forest.

and perhaps energy sectors are predicted better than the financial sector. The numerical values in Table 10 also provide information about how far the prediction is when it is wrong. For example the Bagged decision tree is on the average 0.81 notches away from the true value.

5.2. Comparing the performance of the various Machine Learning Credit Ratings with the differences between different credit rating agencies

In the previous section we compared the accuracy of the prediction of Machine learning models with the corporate ratings provided by Standard and Poor's, Moody's and Fitch. These three credit agencies do not always agree on the ratings of companies. Therefore, the variation of their ratings as measured by Notch distance is useful in order to assess a measure of an acceptable standard deviation of a specific credit rating.

In Table 11 we present the same statistics as in the previous table using quarterly ratings from each of the rating agencies. We can see that the notch distance between the rating agencies shows less agreement between themselves than most of the Machine learning methods considered in this paper. This should give some confidence in the machine learning methods in this paper.

5.3. Capturing rating changes

Generally, credit ratings do not change very much from quarter to quarter. Thus for example if a dataset contains 90% of quarters in which ratings did not change from the previous quarter, a vary naive model that would predict the same rating as the one from the previous quarter would have 90% accuracy! Thus, the most important issue when predicting credit rating based on information available is about predicting a rating change when the actual rating will change. Table 12 shows the percentage of credit rating changes that have been captured by each model. These numbers are actually much lower than the precision of the models. They are larger than we expected, however, the performance of the Bagged decision tree and random forest is not that far from the MLP and SVM in capturing rating changes. These ideas will be investigated in future work as predicting changes is the most important part of the credit analysis.

6. Conclusion and future directions

In this work we first provide a survey of current literature using machine learning techniques to predict corporate credit ratings. The survey pointed toward support vector machine (SVM) and artificial neural network (ANN) as methods that produce the most

Fig. 4. Train set, Test set and Predictions of BK using MLP.

Table 7

Accuracy of different techniques.

	Financial sector	Energy sector	Healthcare sector
Bagged Decision Tree	84.21%	82.11%	83.90%
Random Forest	82.83%	84.45%	82.97%
Multi Layer Perceptron	73.95%	78.19%	76.63%
One vs. One-SVM	42.12%	75.31%	71.29%
One vs. All-SVM	40.14%	59.17%	61.89%

Table 8

Computed notches for all stocks in three sectors.

Sector	Methods	Zero	One Absolute	Greater than One Absolute
Financial	Bagged Decision Tree	84.21%	12.01%	3.78%
	Random Forest	82.83%	12.08%	5.09%
	Multi Layer Perceptron	73.95%	15.92%	10.13%
	One vs. One-SVM	42.12%	32.13%	25.75%
	One vs. All-SVM	40.14%	32.28%	27.58%
Energy	Bagged Decision Tree	82.11%	13.98%	3.91%
	Random Forest	84.45%	11.85%	3.70%
	Multi Layer Perceptron	78.19%	15.48%	6.33%
	One vs. One-SVM	75.31%	16.98%	7.71%
	One vs. All-SVM	59.17%	20.64%	20.19%
Healthcare	Bagged Decision Tree	83.90%	13.51%	2.59%
	Random Forest	82.97%	11.74%	5.29%
	Multi Layer Perceptron	76.63%	16.08%	7.29%
	One vs. One-SVM	71.29%	16.03%	12.68%
	One vs. All-SVM	61.89%	24.50%	13.61%

Table 9
Expected values and standard deviation of all notches in three sectors.

Sector	Methods	$E[X]$	$\sigma(X)$	$E[|X|]$
Financial	Bagged Decision Tree	0.0501	0.7614	0.2719
	Random Forest	0.0068	0.7028	0.2731
	Multi Layer Perceptron	0.1381	1.0152	0.5114
	One vs. One-SVM	0.2661	1.6917	1.4584
	One vs. All-SVM	0.6263	1.4562	1.4699
Energy	Bagged Decision Tree	0.0980	0.5817	0.2851
	Random Forest	−0.0594	0.5892	0.3212
	Multi Layer Perceptron	−0.0800	0.8421	0.3858
	One vs. One-SVM	−0.0685	0.7224	0.3481
	One vs. All-SVM	−0.0066	0.9117	0.6491
Healthcare	Bagged Decision Tree	0.1975	0.3512	0.2959
	Random Forest	−0.2154	0.6317	0.3223
	Multi Layer Perceptron	−0.0531	0.6854	0.4486
	One vs. One-SVM	0.1191	0.7515	0.6531
	One vs. All-SVM	0.2461	0.9683	0.8372

Table 10
Conditional expected values and standard deviation of all notches in three sectors.

Sector	Methods	$E[X|X \neq 0]$	$\sigma(X|X \neq 0)$	$E[|X||X \neq 0]$
Financial	Bagged Decision Tree	-0.0018	0.8619	1.2795
	Random Forest	-0.0489	0.9037	1.1451
	Multi Layer Perceptron	0.2391	1.3294	1.3912
	One vs. One-SVM	0.0911	1.8011	2.0594
	One vs. All-SVM	0.7216	1.4991	1.9171
Energy	Bagged Decision Tree	0.0770	0.3735	0.8264
	Random Forest	−0.0885	0.3010	0.7587
	Multi Layer Perceptron	−0.1317	0.5615	0.9991
	One vs. One-SVM	−0.2415	0.3512	0.9387
	One vs. All-SVM	-0.5131	0.4411	1.2853
Healthcare	Bagged Decision Tree	0.1973	0.2218	0.5313
	Random Forest	−0.4516	0.3054	0.6214
	Multi Layer Perceptron	−0.2712	0.3614	0.7778
	One vs. One-SVM	−0.0710	0.5784	1.0496
	One vs. All-SVM	0.4052	0.6413	1.2384

Table 11
Notch distances for the three rating agencies S&P, Moody's, and Fitch quarter.

	$E[X]$	$\sigma(X)$	$E[|X|]$	$E[X|X \neq 0]$	$\sigma(X|X \neq 0)$	$E[|X||X \neq 0]$
S&P and Moody's	0.33	1.23	0.76	0.54	1.53	1.43
S&P and Fitch	−0.05	0.95	0.64	−0.11	1.32	1.23
Moody's and Fitch	−0.30	1.02	0.75	−0.51	1.28	1.27

Table 12
Captured changes.

	Financial sector	Energy sector	Healthcare sector
Bagged Decision Tree	67.64%	88.88%	77.27%
Random Forest	71.56%	77.77%	83.33%
Multi Layer Perceptron	53.92%	77.77%	66.66%
One vs. One-SVM	29.41%	77.77%	58.33%
One vs. All-SVM	25.49%	33.33%	41.66%

successful results. We implemented these methods and two other methods based on decision trees which in addition have the ability to select best features for the classification problems. We applied all methods for the same datasets spanning 2009 to 2018 for companies from two different sectors: energy, and healthcare and another dataset spanning 1990 to 2018 for companies from

financial sector. All our results indicate that decision tree methods are in fact outperforming SVM and MLP. To check the results we introduced a Notches Distance concept. This measure helps us quantify how far the predictions are from the true values when the algorithm fails to produce the correct prediction. Using this new measure the results were unchanged. Since this is a new measure we need to have a benchmark for the results we obtained. To this end we calculated the Notch Distance using ratings provided by S&P, Moody's and Fitch on the same companies. We found that the best algorithms produce a Notch Distance from the true ratings which is comparable with the distance between ratings produced by different rating agencies on the same company.

CRediT authorship contribution statement

Parisa Golbayani: Methodology, Software, Investigation, Writing - original draft, Formal analysis. **Ionuț Florescu:** Conceptualization, Investigation, Validation, Writing - review & editing, Supervision, Funding acquisition, Formal analysis. **Rupak Chatterjee:** Conceptualization, Validation, Writing - review & editing, Formal analysis.

References

Abdou, H. A. (2009). An evaluation of alternative scoring models in private banking. *The Journal of Risk Finance, 10*(1), 38–53.

Abdou, H. A., Tsafack, M. D. D., Ntim, C. G., & Baker, R. D. (2016). Predicting creditworthiness in retail banking with limited scoring data. *Knowledge-Based Systems, 103*, 89–103.

Addo, P. M., Guegan, D., & Hassani, B. (2018). Credit risk analysis using machine and deep learning models. *Risks, 6*(2), 38.

Akkoç, S. (2012). An empirical comparison of conventional techniques, neural networks and the three stage hybrid adaptive neuro fuzzy inference system (anfis) model for credit scoring analysis: The case of turkish credit card data. *European Journal of Operational Research, 222*(1), 168–178.

Angelini, E., di Tollo, G., & Roli, A. (2008). A neural network approach for credit risk evaluation. *The Quarterly Review of Economics and Finance, 48*(4), 733–755.

Bahrammirzaee, A. (2010). A comparative survey of artificial intelligence applications in finance: Artificial neural networks, expert system and hybrid intelligent systems. *Neural Computing and Applications, 19*(8), 1165–1195.

Bottou, L., Cortes, C., Denker, J. S., Drucker, H., Guyon, I., Jackel, L. D., et al. (1994). Comparison of classifier methods: A case study in handwritten digit recognition. In *Pattern recognition, 1994. Vol. 2-Conference B: Computer vision & image processing, Proceedings of the 12th IAPR International. Conference on, volume 2* (pp. 77–82). IEEE.

Brabazon, A., & O'Neill, M. (2006). Credit classification using grammatical evolution. *Informatica, 30*(3).

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees.* Monterey, CA: Wadsworth.

Brennan, D., & Brabazon, A. (2004). Corporate bond rating using neural networks. In *IC-AI*.

Cao, L., Guan, L. K., & Jingqing, Z. (2006). Bond rating using support vector machine. *Intelligent Data Analysis, 10*(3), 285–296.

Chandra, D. K., Ravi, V., & Bose, I. (2009). Failure prediction of dotcom companies using hybrid intelligent techniques. *Expert Systems with Applications, 36*(3), 4830–4837.

Chaveesuk, R., Srivaree-Ratana, C., & Smith, A. E. (1999). Alternative neural network approaches to corporate bond rating. *Journal of Engineering Valuation and Cost Analysis, 2*(2), 117–131.

Chen, Y.-S. (2012). Classifying credit ratings for asian banks using integrating feature selection and the cpda-based rough sets approach. *Knowledge-Based Systems, 26*, 259–270.

Chen, Y.-S., & Cheng, C.-H. (2013). Hybrid models based on rough set classifiers for setting credit rating decision rules in the global banking industry. *Knowledge-Based Systems, 39*, 224–239.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297.

Culkin, R., & Das, S. R. (2017). Machine learning in finance: The case of deep learning for option pricing. *Journal of Investment Management, 15*(4).

Daniel, R., Hančlová, J., el Woujoud Bousselmi, H., et al. (2019). Corporate rating forecasting using artificial intelligence statistical techniques. *Investment Management & Financial Innovations, 16*(2), 295.

Dima, A. M., & Vasilache, S. (2016). Companies default prediction using neural networks. *Romanian Journal of Economic Forecasting, 19*(3), 127.

Dutta, S., & Shekhar, S. (1988). Bond rating: A non-conservative application of neural networks. *Proceedings of the IEEE international conference on neural networks: Vol. 2,* (pp. 443–450). Los Alamitos: IEEE Press.

Feng, G., He, J., & Polson, N. G. (2018). *Deep learning for predicting asset returns*. arXiv preprint arXiv:1804.09314.

Gangolf, C., Dochow, R., Schmidt, G., & Tamisier, T. (2016). Automated credit rating prediction in a competitive framework. *RAIRO-Operations Research, 50*(4–5), 749–765.

Ganguli, S., & Dunnmon, J. (2017). *Machine learning for better models for predicting bond prices*. arXiv preprint arXiv:1705.01142.

Garavaglia, S. (1991). An application of a counter-propagation neural network: Simulating the standard and poor's corporate bond rating system. In *First international conference on artificial intelligence applications on Wall Street, 1991. Proceedings* (pp. 278–287). IEEE.

Gogas, P., Papadimitriou, T., & Agrapetidou, A. (2014). Forecasting bank credit ratings. *The Journal of Risk Finance, 15*(2), 195–209.

Gu, S., Kelly, B. T., & Xiu, D. (2019). Empirical asset pricing via machine learning. *31st Australasian finance and banking conference 2018*.

Hájek, P. (2012). Credit rating analysis using adaptive fuzzy rule-based systems: An industry-specific approach. *Central European Journal of Operations Research, 20*(3), 421–434.

Hajek, P., Olej, V., & Prochazka, O. (2016). Predicting corporate credit ratings using content analysis of annual reports–a naïve bayesian network approach. In *FinanceCom 2016* (pp. 47–61). Springer.

Hajek, P., & Michalak, K. (2013). Feature selection in corporate credit rating prediction. *Knowledge-Based Systems, 51*, 72–84.

Hájek, P., & Olej, V. (2011). Credit rating modelling by kernel-based approaches with supervised and semi-supervised learning. *Neural Computing and Applications, 20*(6), 761–773.

Hájek, P., & Olej, V. (2014). Predicting firms' credit ratings using ensembles of artificial immune systems and machine learning–an over-sampling approach. *IFIP international conference on artificial intelligence applications and innovations* (pp. 29–38). Springer.

Horrigan, J. O. (1966). The determination of long-term credit standing with financial ratios. *Journal of Accounting Research,* 44–62.

Huang, Z., Chen, H., Hsu, C.-J., Chen, W.-H., & Wu, S. (2004). Credit rating analysis with support vector machines and neural networks: A market comparative study. *Decision Support Systems, 37*(4), 543–558.

Huang, W., Nakamori, Y., & Wang, S.-Y. (2005). Forecasting stock market movement direction with support vector machine. *Computers & Operations Research, 32*(10), 2513–2522.

Hwang, R.-C., Cheng, K., & Lee, C.-F. (2009). On multiple-class prediction of issuer credit ratings. *Applied Stochastic Models in Business and Industry, 25*(5), 535–550.

Hwang, R.-C., Chung, H., & Chu, C. (2010). Predicting issuer credit ratings using a semiparametric method. *Journal of Empirical Finance, 17*(1), 120–137.

Irmatova, E. (2016). *Relarm: A rating model based on relative pca attributes and k-means clustering.* arXiv preprint arXiv:1608.06416.

Kamstra, M., Kennedy, P., & Suan, T.-K. (2001). Combining bond rating forecasts using logit. *Financial Review, 36*(2), 75–96.

Kanchymalay, K., Salim, N., Sukprasert, A., Krishnan, R., & Hashim, U. R. (2017). Multivariate time series forecasting of crude palm oil price using machine learning techniques. *IOP conference series: Materials science and engineering: Vol. 226,* (pp. 012117–). IOP Publishing.

Kao, L.-J., Chiu, C.-C., & Chiu, F.-Y. (2012). A bayesian latent variable model with classification and regression tree approach for behavior and credit scoring. *Knowledge-Based Systems, 36*, 245–252.

Kaplan, R. S., & Urwitz, G. (1979). Statistical models of bond ratings: A methodological inquiry. *Journal of Business,* 231–261.

Karminsky, A. M., & Khromova, E. (2016). Extended modeling of banks' credit ratings. *Procedia Computer Science, 91*, 201–210.

Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance, 34*(11), 2767–2787.

Khashman, A. (2010). Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes. *Expert Systems with Applications, 37*(9), 6233–6239.

Khemakhem, S., & Boujelbene, Y. (2015). Credit risk prediction: A comparative study between discriminant analysis and the neural network approach. *Accounting and Management Information Systems, 14*(1), 60.

Kim, K.-J. (2003). Financial time series forecasting using support vector machines. *Neurocomputing, 55*(1–2), 307–319.

Kim, K. S. (2005). Predicting bond ratings using publicly available information. *Expert Systems with Applications, 29*(1), 75–81.

Kim, K.-J., & Ahn, H. (2012). A corporate credit rating model using multi-class support vector machines with an ordinal pairwise partitioning approach. *Computers & Operations Research, 39*(8), 1800–1811.

Kim, K.-S., & Han, I. (2001). The cluster-indexing method for case-based reasoning using self-organizing maps and learning vector quantization for bond rating cases. *Expert Systems with Applications, 21*(3), 147–156.

Kim, H. S., & Sohn, S. Y. (2010). Support vector machines for default prediction of smes based on technology credit. *European Journal of Operational Research, 201*(3), 838–846.

Kim, J. W., Weistroffer, H. R., & Redmond, R. T. (1993). Expert systems for bond rating: A comparative analysis of statistical, rule-based and neural network systems. *Expert Systems, 10*(3), 167–172.

Koh, H. C., Tan, W. C., & Peng, G. C. (2004). Credit scoring using data mining techniques. *Singapore Management Review, 26*(2), 25.

Kwon, Y. S., Han, I., & Lee, K. C. (1997). Ordinal pairwise partitioning (opp) approach to neural networks training in bond rating. *International Journal of Intelligent Systems in Accounting Finance and Management, 6*, 23–40.

Lee, Y.-C. (2007). Application of support vector machines to corporate credit rating prediction. *Expert Systems with Applications, 33*(1), 67–74.

Lee, T.-S., Chiu, C.-C., Chou, Y.-C., & Lu, C.-J. (2006). Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. *Computational Statistics & Data Analysis, 50*(4), 1113–1130.

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.

Liu, Y., & Zheng, Y. F. (2005). One-against-all multi-class svm classification using reliability measures. *Neural Networks, 2*, 849–854.

Maher, J. J., & Sen, T. K. (1997). Predicting bond ratings using neural networks: A comparison with logistic regression. *Intelligent Systems in Accounting, Finance & Management, 6*(1), 59–72.

Martens, D., Van Gestel, T., De Backer, M., Haesen, R., Vanthienen, J., & Baesens, B. (2010). Credit rating prediction using ant colony optimization. *Journal of the Operational Research Society, 61*(4), 561–573.

Moody, J. & Utans, J. (1994). Architecture selection strategies for neural networks: Application to corporate bond rating prediction. In *Neural networks in the capital markets* (pp. 277–300). Citeseer.

Moscatelli, M., Narizzano, S., Parlapiano, F., Viggiano, G., et al. (2019). *Corporate default forecasting with machine learning*. Technical report, Bank of Italy, Economic Research and International Relations Area.

Oelerich, A., & Poddig, T. (2006). Evaluation of rating systems. *Expert Systems with Applications, 30*(3), 437–447.

Pacelli, V., & Azzollini, M. (2011). An artificial neural network approach for credit risk management. *Journal of Intelligent Learning Systems and Applications, 3*(02), 103.

Pai, P.-F., Tan, Y.-S., & Hsu, M.-F. (2015). Credit rating analysis by the decision-tree support vector machine with ensemble strategies. *International Journal of Fuzzy Systems, 17*(4), 521–530.

Paleologo, G., Elisseeff, A., & Antonini, G. (2010). Subagging for credit scoring models. *European Journal of Operational Research, 201*(2), 490–499.

Peng, Y., Wang, G., Kou, G., & Shi, Y. (2011). An empirical study of classification algorithm evaluation for financial risk prediction. *Applied Soft Computing, 11*(2), 2906–2915.

Petropoulos, A., Chatzis, S. P., & Xanthopoulos, S. (2016). A novel corporate credit rating system based on student'st hidden markov models. *Expert Systems with Applications, 53*, 87–105.

Platt, J. C., Cristianini, N., & Shawe-Taylor, J. (2000). Large margin dags for multiclass classification. *Advances in Neural Information Processing Systems,* 547–553.

Ryser, M., & Denzler, S. (2009). Selecting credit rating models: A cross-validation-based comparison of discriminatory power. *Financial Markets and Portfolio Management, 23*(2), 187–203.

Saha, S., & Waheed, S. (2017). Credit risk of bank customers can be predicted from customer's attribute using neural network. *International Journal of Computer Applications, 975*, 8887.

Sambasivan, R., & Das, S. (2017). A statistical machine learning approach to yield curve forecasting. In *2017 International conference on computational intelligence in data science (ICCIDS)* (pp. 1–6). IEEE.

Schölkopf, B., Burges, C. J., & Smola, A. J. (1999). *Advances in kernel methods: Support vector learning.* MIT Press.

Shin, K.-S., & Han, I. (2001). A case-based approach using inductive indexing for corporate bond rating. *Decision Support Systems, 32*(1), 41–52.

Surkan, A. J., & Singleton, J. C. (1990). Neural networks for bond rating improved by multiple hidden layers. In *1990 IJCNN international joint conference on neural networks, 1990* (pp. 157–162). IEEE.

Tsai, C.-F., & Wu, J.-W. (2008). Using neural network ensembles for bankruptcy prediction and credit scoring. *Expert Systems with Applications, 34*(4), 2639–2649.

Vapnik, V. (2013). *The nature of statistical learning theory.* Springer Science & Business Media.

Veronezi, P. H. (2016). *Corporate credit rating prediction using machine learning techniques.* Master's thesis, Stevens Institute of Technology.

Wallis, M., Kumar, K., & Gepp, A. (2019). Credit rating forecasting using machine learning techniques. *Managerial perspectives on intelligent big data analytics* (pp. 180–198). IGI Global.

Wang, A., Yuan, W., Liu, J., Yu, Z., & Li, H. (2009). A novel pattern recognition algorithm: Combining art network with svm to reconstruct a multi-class classifier. *Computers & Mathematics with Applications, 57*(11–12), 1908–1914.

West, R. R. (1970). An alternative approach to predicting corporate bond ratings. *Journal of Accounting Research,* 118–125.

West, D. (2000). Neural network credit scoring models. *Computers & Operations Research, 27*(11–12), 1131–1152.

Wu, T.-C., & Hsu, M.-F. (2012). Credit risk assessment and decision making by a fusion approach. *Knowledge-Based Systems, 35*, 102–110.

Xiao, H., Xiao, Z., & Wang, Y. (2016). Ensemble classification based on supervised clustering for credit scoring. *Applied Soft Computing, 43*, 73–86.

Ye, Y., Liu, S., & Li, J. (2008). A multiclass machine learning approach to credit rating prediction. In *2008 International symposiums on information processing (ISIP)* (pp. 57–61). IEEE.

Yeh, C.-C., Lin, F., & Hsu, C.-Y. (2012). A hybrid kmv model, random forests and rough set theory approach for credit rating. *Knowledge-Based Systems, 33*, 166–172.

Yu, L., Wang, S., & Lai, K. K. (2008). Credit risk assessment with a multistage neural network ensemble learning approach. *Expert Systems with Applications, 34*(2), 1434–1444.

Zhao, Z., Xu, S., Kang, B. H., Kabir, M. M. J., Liu, Y., & Wasinger, R. (2015). Investigation and improvement of multi-layer perceptron neural networks for credit scoring. *Expert Systems with Applications, 42*(7), 3508–3516.

Zhong, H., Miao, C., Shen, Z., & Feng, Y. (2014). Comparing the learning effectiveness of bp, elm, i-elm, and svm for corporate credit ratings. *Neurocomputing, 128*, 285–295.