Received: 13 December 2019 | Revised: 7 March 2020 | Accepted: 6 April 2020
DOI: 10.1002/spe.2842

SPECIAL ISSUE PAPER

WILEY

Comparative study of support vector machines and random forests machine learning algorithms on credit operation

Germanno Teles[1] | Joel J. P. C. Rodrigues[1,2,3] | Ricardo A. L. Rabêlo[2] | Sergei A. Kozlov[3]

[1]Instituto de Telecomunicações, Universidade da Beira Interior, Covilhã, Portugal
[2]Federal University of Piauí (UFPI), Teresina, Brazil
[3]ITMO University, St. Petersburg Russia

Correspondence
Joel J. P. C. Rodrigues, Federal University of Piauí, Campus Petrônio Portela, Av. Ministro Petrônio Portela, S/N - Bloco 8, Centro de Tecnologia 64049-550, Ininga, Teresina - PI, Brazil.
Email: joeljr@ieee.org

Funding information
Conselho Nacional de Desenvolvimento Científico e Tecnológico, Grant/Award Numbers: Grant No. 309335/2017-5, Process No. 200450/2015-8; Fundação para a Ciência e a Tecnologia, Grant/Award Number: Project UIDB/EEA/50008/2020; Government of Russian Federation, Grant/Award Number: Grant No. 08-08

Summary
Corporate insolvency has significant adverse effects on an economy. With the number of multinationals increasing rapidly, corporate bankruptcy can severely disrupt the global financial environment. However, multinationals do not fail instantaneously; objective strategies combined with a rigorous analysis of both qualitative and quantifiable data can go a long way in identifying an organization's financial risks. Recent advancements in information and communication technologies have made data collection and storage an easy task. The challenge becomes mining the appropriate data about a company's financial risks and implementing it in forecasting a company's insolvency probabilities. In recent years, machine learning has been incorporated into big data analytics owing to its massive success in learning complex models. Machine learning algorithms such as Support Vector Machines (SVM), Random Forests (RF), Artificial Neural Networks, Gaussian Processes, and Adaptive Learning have been used in the analysis of Big Data to predict the financial risks of companies. In this paper, credit scoring is explored with regards to data processed using the collateral as an independent variable. The obtained results indicate that RF algorithm is promising for use in credit risk management. This research shows the advantages of the RF approach over the SVM algorithm are its speed and operational simplicity, and SVM has the benefit of higher classification accuracy than RF. The paper compares the SVM and RF algorithms to forecast the recovered value in a credit task. The execution of the projected intelligent systems uses tests and algorithms for authentication of the projected model.

KEYWORDS
Big Data, credit operation, machine learning, random forests, support vector machines

1 | INTRODUCTION

In recent years, the use of Machine Learning algorithms has been integrated in most contemporary credit risk prediction methods. There are many Machine Learning algorithms that have been incorporated into credit risk prediction models including Artificial Neural Networks (ANN), Naive Bayes, Dimensionality Reduction Algorithms, gradient Boosting Algorithms, and Decision Trees.[1] Different credit risk assessment models have been developed for the various models.

The use of Machine Learning algorithms in credit risk prediction for financial institutions raises numerous ethical debates. In 2017, a public forum from National Commission on Informatics and Liberty was held with the key topic of discussion being how personal data is used as well as the problems associated with Big Data and specifically the General Data Protection Regulation directive.[2] People are definitely afraid of how personal data is used and the fact that with Machine Learning the decisional power on this data will be placed on algorithms only serves to heighten these fears.

Choosing the appropriate machine learning algorithm depends on a couple of factors such as the size of data, the quality required as well as the nature of the data being analyzed.[3] Several questions also have to be answered before the machine learning technique is selected, such as what is the desired result? What is the translation technique to be implemented on the mathematical algorithm to generate commands in the subject computer? or How long does the algorithm have to operate to acquire sufficient learning ?[4-6]

Deplying an effective credit system to predict the recovery value presents many challenges. Some of these challenges can be recurrent with specific solutions; however, the general principles of credit risk solutions still apply. Machine learning has been widely used in many of these solutions by programming computers to learn without any special instructions.

Support Vector Machines (SVM) is usually described to produce better outcomes than other classifiers.[7] The Random Forests (RF) algorithm has demonstrated to manipulate huge dimensional data properly and is relatively resistant to overfitting.[8]

This study then aims to investigate two machine learning algorithms, SVM and RF, on their applicability, efficiency, flexibility, and accuracy in credit risk assessment. In this comparative study, two machine learning techniques are modeled, that is, RF and SVM,[9] focus on credit risk scoring and the impacts of distinct machine learning models to identify defaults by lenders. The main contribution of this study describes the use of SVM algorithm and the RF algorithm in addressing issues of recovery value loan provision.[10] A sample dataset from a bank was implemented using the algorithms with the objective of learning to classify and comparing a credit operation (recovery value), with different kernels and kernel parameters. Results for RF and SVM will be analyzed and compared for varying sets of data. The results from the various kernels are tuned with appropriate parameter settings. Besides the technical questions regarding understanding machine learning algorithms, referrals to the various discussions associated with confidentiality arising from the use of personal data in these algorithms are also discussed. It has been observed that multiple approaches can be implemented to address the fundamental objective of identifying the choice of features/variables, the algorithm and the corresponding criteria. In this digital era of Big Data, transparency is of utmost importance.[11] It is necessary that terms used in this field are ethical, clear, transparent, and appropriately defined. It is necessary to implement appropriate strategies to train data based on deep learning,[12] and machine learning algorithms and their implementation must be monitored to ensure accuracy.

The paper goes further to evaluate the stability of the two models based on the variables chosen. The method used by banks in making loans decisions is unclear; however, implementation of classical linear models in banking systems is adequately documented. For this paper, the transparent elastic approach was used as a benchmark.

The remainder of the paper is organized as follows. Section 2 presents a theoretical background, introduces related works, and discusses the problem of credit risk assessment on classifying the credit-scoring approach with SVM and RF. Section 3 analyzes the results, hence demonstrating which classifier is superior in terms of accuracy and discusses the issues of SVM, and RF models and draws a comparison between those two models. Finally, the conclusion and future scope are provided in Section 4.

2 | BACKGROUND AND RELATED WORK

Machine Learning is a subset of Artificial Intelligence that focuses on the development of strategies, methods, and algorithms. Therefore, the development of algorithms that enable a computer system to learn from the provided data and execute tasks and activities of design with sampling data together with performing tests on the new data. The field of machine learning has some strong links to statistics in various ways. There are multiple approaches and methods created for machine learning tasks. Neural Network techniques are widely used but have various limitations with regards to generalization, developing prototypes that usually get over fit with data.[12] This can be attributed to optimization procedures implemented for specific statistical approaches and parameter selection to determine the most appropriate model possible. This problem of credit risk assessment can best be solved with machine learning algorithms,[13] as follows:

1. Supervised learning: This type is composed of a target output variable (dependent variable) which is to be forecast from a set of predictors (independent variable). From these sets of attributes, a function mapping inputs to target outputs is generated. The training procedure remains in progress until the method attains the target level of precision. Examples include: RF, regression, decision tree and k-nearest neighbors.[14]
2. Unsupervised learning for this algorithm, there is no target output variable to forecast. This algorithm is mainly used to cluster a certain population into various groups to enable the segmentation of customers in different groups for a certain intervention. Examples of unsupervised learning include: K-means and a priori algorithms.[15]
3. Reinforcement Learning for this algorithm, the computers are trained to make particular decisions. The computer is exposed to an environment in which it can consistently train itself by implementing trial and error methods. The environment also provides an addition to rewards, unique numerical values that the agent attempts to maximize over time. The machine in the end learns from previous experiences and attempts to obtain the most suitable knowledge to make appropriate decisions. The best example of a reinforcement learning algorithm is the Markov Decision Process.[16]

The concept of empirical data modeling is appropriate for numerous applications in the field of computer science.[17,18] Empirical data modeling involves an induction procedure to create a model of the scheme from which it can derive responses of the system which are to be tried or observed. The observed data is finite and is considered a sample. This sampling is not uniform, and because of the great dimensional nature of the data, the input will be sparsely distributed. The problem is therefore often misrepresented.

SVMs were initially introduced to machine learning algorithms by Boser, Guyon, and Vapnik in 1992.[19] SVMs have since become a popular algorithm, especially with regard to handwritten digit recognition. Currently, SVMs are a crucial component of all research related to Machine Learning and are now regarded as a primary example of kernel methods.[20]

SVMs are a family of machine learning algorithms that discriminatively classify variables that were initially defined by a different hyperplane. In simpler terms, an input of labeled training data, the algorithm produces an output that consists of an optimal hyperplane categorizing the data based on a set of objective conditions. On a two-dimensional plane, this hyperplane is a line dividing the plane into two classes. For example, given two characteristics of an individual such as height and hair length, the two characteristics (variables) would first be plotted in two n-dimensional spaces with each point having two coordinates (support vectors) (Figure 1).

SVMs have been widely used in various machine learning applications such as facial recognition, target recognition, object identification, speaker identification, and handwritten digit recognition.[5,21]

RF is a machine learning algorithm proposed by Breiman.[8] It involves the construction of a forecast ensemble consisting of a set of decision trees growing in arbitrarily identified subspaces of data. Breimans concepts

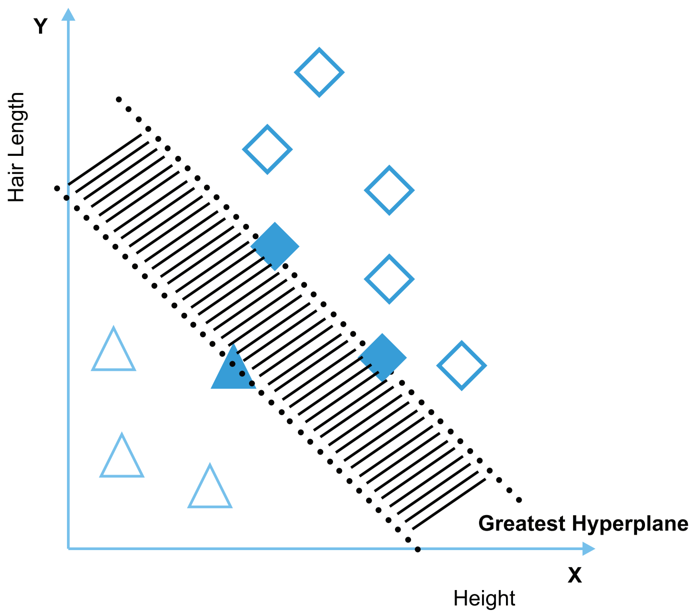

FIGURE 1 Illustration of support vector machine model

were significantly subject to the prior scholarly works of Amit and Geman (1997)[19] on geometrical feature selection, the random subspace method developed by Ho in 1998.[22] Dietterich's (2000) casual split selection approach was also assimilated in the development of the RF approach.[23] Various empirical studies have shown that RF are increasingly becoming a strong competitor to some state-of-the-art approaches such as boosting (developed by Freund and Shapire in 1996) and Support Vector Machines.[24] The popularity of RF can be attributed to the fact that besides being fast and easy to implement, they also produce precise predictions and they can handle numerous input variables without overfitting.[25] They are considered among the most accurate machine learning techniques in the market.

In Breimans random trees method, each tree forming the collection is developed by first randomly choosing a small set of input coordinates (features or variables) to split on at each node and then calculating the best split with regards to these variables in the trained set. The Classification and Regression Tree methodology is then applied to grow the tree to its extreme size without trimming. This scheme of randomizing subsets is combined with bagging to resample with replacement, the training set of data whenever a new distinct tree grows.[26] In simple terms, the algorithm creates multiple decision trees and combines them to obtain a more precise forecast. Although the working of this approach appears simple, there are many driving forces involved in the mechanism making it challenge to analyze. In fact, the mathematical properties of RF remain mostly unknown to date, and most theoretical studies have focused on remote parts or stylized forms of the procedure.[27] Nonetheless, the arithmetic mechanism of true RF is yet to be entirely implicit and is still being explored.

Throughout this study, RF is assumed to comprise of a training model $D_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ of independent and identically distributed (i.i.d.) $[0, 1]d \times R - valued$ random variables ($d \geq 2$) with a similar distribution to that of an independent simple pair (X, Y) satisfying $EY2 < \infty$. Space $[0, 1]^d$ has a standard Euclidean metric. For fixed $x \in [0, 1]d$, our objective is to approximate the regression function $r(x) = E[Y|X = x]$ by means of the data D_n. In other words, a regression function approximation rn is reliable if $E[r_n(X) - r(X)]^2 \Rightarrow 0$ as n $\Rightarrow \infty$.

2.1 | Study using the SVM

The objective of this approach is to implement the SVM learning technique to forecast the probability of loan defaults. The data needed training with each consisting of values for the set of input and output variables. The variables were chosen, and only those whose behavior was predictable were included. For the current comparative study, the data variables were considered the primary risk indicators and those borrowing were considered the subjects. The data collected for this study was from a bank, and it consisted solely of short-term loans since they make up the most significant share of loans. A dataset of 1890 credit files was sourced, and the subjects were classified as less risky or risky clients. The most constant variable is the probability of default together with a dummy variable, Y is equal to zero for a less risky client and one for risky clients. This means that $Y = 1$ when the repayment is delayed and $Y = 0$ when the payment is made in good time. SVM classifications were implemented to predict the potentials of class membership.

Before the SVM model was constructed, it was necessary to process the datasets in two standard deviation procedures. The data were split according to the average (mean) or assumed value, and all abnormal data was deleted. The final dataset consisted of 1890 samples with 1100 of them classified as good credits while 790 were classified as bad debts. For those in extremely bad credit positions, more than 3 years in default, they were classified as abnormal cases, and all their data was erased; this data represented only 50 samples. Since the quantity of the two sorts of samples is close, the SVM minimum requirements were met. The dataset is then divided into a training set and a test set. To show the learn and generalization capabilities of the SVM algorithm with regards to small samples, 30% (567 instances) of the sample data is chosen to build the SVM algorithm as a training sample. The remaining 70% (1323 instances) is randomly selected to test the generalization capability of the approach as a test dataset.

2.2 | Study using RF

The experiments for RF were conducted using Heuristic Lab and a modified RF Trees algorithm for classification. The key parameters to configure include: r, the ratio between 0 and 1; m, the number of variables; and nT, the number of trees. Appropriate selections of r and m impact the issue of noise tolerance in the training set meaning these parameters need careful adjustment. Empirical tests for this research showed that r, m, and nT are the most appropriate parameters. In order to tune RF parameters, r and m were selected arbitrarily, and nine runs were analyzed for the various trees. The trees ranged from 50 to 500 with an increment of 50 for each run. It was determined that changing the parameters has no significant influence on the test performance and thus two tests were chosen for in-depth analysis.

3 | RESULTS ANALYSIS AND DISCUSSION

From the above analysis, a sample set (x, y) was constructed where $x = 4$ and the dimension y acts as an attribution sample. For good credits, $y = 1$ and $y = -1$ for bad credits. If the inner kernel function selects the polynomial kernel function or any other function, SVM is able to acquire the results of estimated performance and distribution of support vectors. The inner product function used for SVM in this paper is given by Equation (1) .

$$K(x_1, x_i) = exp\left\{-\frac{|x - x_i|^2}{\sigma^2}\right\}. \tag{1}$$

Figure 2 shows the results of the categorization model on the subject data over the 10-fold cross-validation with $R = 0.3$ and $M = 0.5$ (for the first) $R = 0.66$ and $M = 0.3$ (for the second).

The modified RDF algorithm makes the study of impacts of developed trees and variables easier. The relationship between nT and the performance of the model is not clear, but it becomes better with 500 trees. Table 1 below illustrates the mean, SD, median, maximum and minimum of the various measures used in this experiment.

Both SVM and RF models produced high evaluation results in the receiver operating characteristic (ROC) curve. However, because the classifications were not balanced, this indicator could not be used independently. The high frequency of hits from the majority class creates a bias in the results. Therefore, an alternative indicator is required, the Percentage True Correctly Classified (PTCC), to identify the level of accuracy in the class of interest. A proper assessment of this aspect shows a significant variation between the two projected models, in the range of 60% to 80%, with the best result realized by the SVM algorithm (Table 2). Table 2 presents the classifier model (ful l training set), 992 out of 1322 occurrences are correctly classified by the algorithm with 75.03% of accuracy. The table also gives four different statistical error measurements that measure the degree of relationship among the predicted and actual.

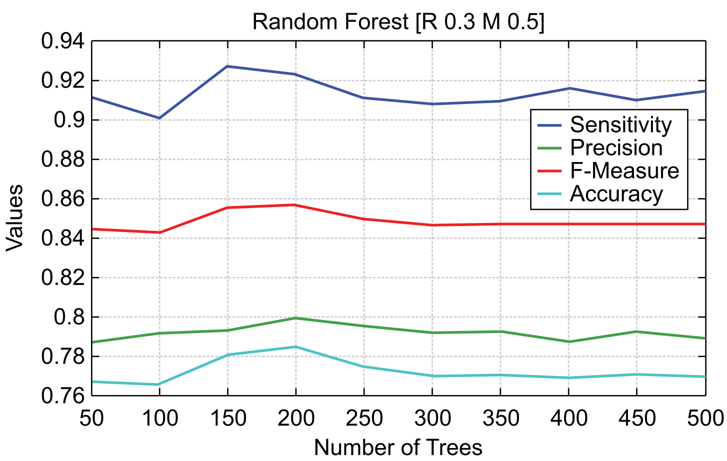

FIGURE 2 Results for the implantation of 10-fold cross validation for 10 runs ($r = 0.3$ and $M = 0.5$)

TABLE 1 Statistics from the test data after implementation of the 10-fold cross-validation for 10 runs ($r = 0.3, M = 0.5$)

	Mean	SD	Median	Maximum	Minimum
Sensitivity	0.916	0.007	0.914	0.931	0.901
Precision	0.80	0.004	0.79	0.8	0.788
f-Measure	0.851	0.004	0.852	0.86	0.844
Accuracy	0.781	0.006	0.773	0.741	0.766

TABLE 2 Stratified cross-validation for support vector machines

		Training set	Total
Percentage true		992	75.03%
Incorrectly classified incidences		330	24.97%
Mean absolute error	0.3672		
Root mean squared error	0.5844		
Relative absolute error	63.22%		
Root relative squared error	96.4556%		
Total number of incidences		1322	100%

TABLE 3 Quality control of support vector machine (SVM) and random forest (RF) machine learning algorithms when used with various dimensionality reduction techniques

		Accuracy	Precision	Sensitivity	F1 Score
PCA	SVM	98.34%	100%	91%	96%
	RF	98.2%	100%	91.2%	93%
LDA	SVM	97.8%	100%	94.6%	97%
	RF	98%	100%	94.6%	97%
ISOMAP	SVM	98%	100%	94.6%	96.85%
	RF	98%	100%	94.3%	97%
Kernel PCA	SVM	98%	100%	94.6%	97%
	RF	98%	100%	91.2%	94%

The SVM and RF classifiers were analyzed using the error confusion matrix approach, which is representative of the whole thematic classification. The error confusion matrix can be used to determine the overall accuracy as well as specific endmember accuracy.[28]

The results of the error confusion matrices show that by using the RF classifier, a few of the endmembers are misclassified. However, the SVM algorithm produces more accurate classification results compared with RF(Table 3). The classifier accuracy procedure requires that the confusion matrix be representative of the entire mapped data area.[29] The results for accurately classified, unclassified, and incorrectly classified data can be obtained from the error confusion matrices. The overall accuracy of each algorithm is obtained by dividing the total number of accurate classifications by the total number of variables in the error confusion matrix. An error confusion matrix with all non-major diagonals having values of zero means that the classifier is 100% accurate. SVM showed the greater accuracy in this test having 53 (1.66%) unclassified incidences with the RF classifier having 58 (1.8%). For each classifier to construct its classification model, the RF classifier on average took 2.7 seconds while SVM took 27 seconds.

These data was sourced from a lending bank institution. The dataset included 1890 instances that were classified into two categories: 1100 good credit and 790 "bad/ defaulted credit." The initial dataset consisted of 16 variables that were classified into 10 qualitative and 6 numerical as shown in Table 4. The dataset used for this study was however processed to convert the original into 16 numerical variables, with the number 16 being an output variable.

All the classifications were conducted using a 10-fold cross-validation approach using three different tools, Heuristic Lab, Weka, and Keel. Default configurations were selected to create and test the models for comparison purposes. The algorithm used with Weka was the SVM using Linear Kernel.

The problems of credit risks have been consistently addressed in conferences on artificial intelligence.[30] The primary concern has been that personal data is used and there is also a growing fear that an algorithm could replace human beings in decision making. These questions are genuine, and this paper emphasizes the appropriate algorithms with regards to decision-making in lending institutions. Algorithms can be implemented to simplify a process while at the same time increasing its fluidity as well as increasing speed.[11] Algorithms include a set of code modelled to achieve set objectives. For example, an algorithm designed to perform a recruitment process introduces several discriminatory conditions based on individual profiles. A similar approach is adopted by lending institutions when making lending decisions to banks.[31] It is therefore of the essence to understand the underlying challenges and find ways to manage the use of algorithms.

TABLE 4 Original variables

Nž	Variable	Type
1	Contract value	Qualitative
2	Balance value	Qualitative
3	Collateral value	Qualitative
4	Number of collaterals	Qualitative
5	Recovery value	Qualitative
6	Value transformation rate	Qualitative
7	Value transformation interest	Qualitative
8	Value rate overdue	Qualitative
9	Client size	Qualitative
10	Main value delay	Qualitative
11	Seniority level	Qualitative
12	Percent used	Numerical
13	Duration in months	Numerical
14	Duration in years	Numerical
15	Duration in days	Numerical
16	Delay in days	Numerical

This study's primary objective was to compare the SVM and RF machine learning techniques based on performance in lending institutions and more precisely in the determination of recovery value. The credit risk concept is characterized by a vagueness that complicates the formulation of a limited definition for its identifying risk factors, and suitable functional forms to approximate and forecast its value is no easy task.[27] In similar fashion, the range and complexity of the credit risk concept render traditional mathematical models obsolete learning techniques.[28] This study compared the performance two approaches, the SVM and the RF approach in addressing the problem of credit risk assessment. In the study, the variables were chosen with regards to the data collected from the banks' records. Despite the numerous capabilities of support vector machines and RF prediction algorithms, the concern of credit risk estimation has barely been tackled with regards to machine learning algorithms let alone a combination of them. The current study, therefore, fills existing gaps that permeate intelligent systems from extremely puzzling bank modelling issues. The primary focus was on the idea of insolvency as a characterization technique of the credit risk.

Internal factors have been consequently implemented to construct models that permit the forecasting of credit risks. This case study that incorporated support vector machines on bank data exhibited high levels of precision, effectiveness and swiftness of the data mining techniques when modelling recovery value as a credit risk. The implementation of support vector machine and RF approach made it possible to establish the riskiest factors and estimation of the subject risk through the training and learning processes, and the outcomes were also very consistent. Additionally, the binary results from this study highlighted the capability of the support vector machines, and RF approaches to authenticate the findings via a parallel and uninfluenced application on the set of data.

The SVM algorithm was determined to be the most appropriate solution for determining recovery value. The ROC curves for SVM were generated and compared as in Figure 3. Both SVM and RF exhibited good performance in ROC curves; however, SVM was marginally better.

4 | CONCLUSION AND FUTURE WORK

Managing credit risks is important for the success of lending institutions. It is therefore of the essence to develop an effective aid for credit decision-making processes. The results of this study indicate that the RF algorithm is promising for use in credit risk management research. The main advantage of the RF approach over the SVM algorithm is its simplicity of operation. The fact that RFs takes much less time to construct its model means that it is more desirable in computerized

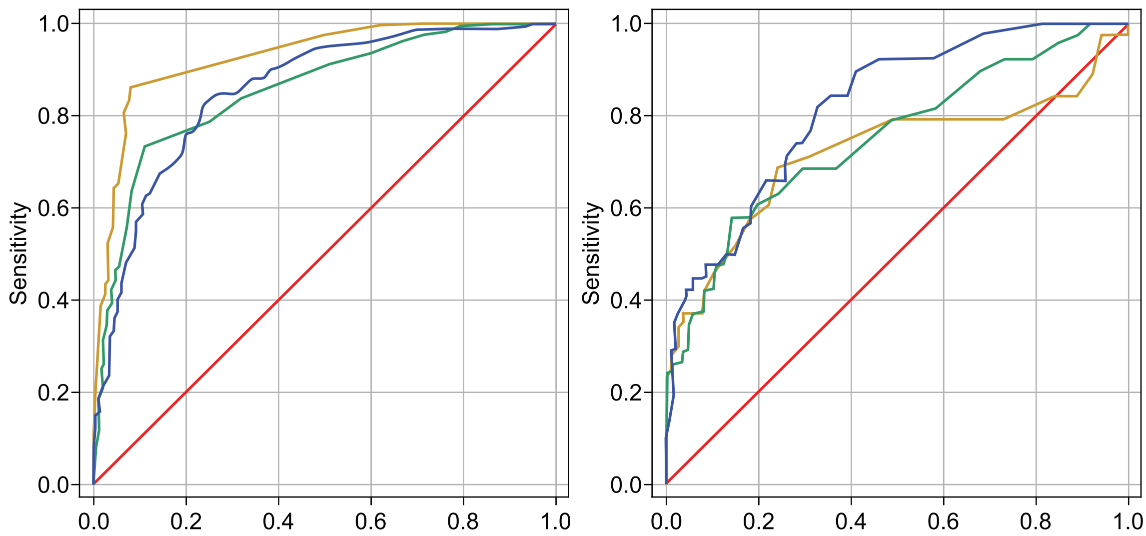

FIGURE 3 Receiver operating characteristic (ROC) curves for support vector machine (SVM) and random forests (RF). ROC curve for SVM (left) and ROC curve for RF (right)

environments. On the other hand, SVMs have the advantage of higher classification accuracy than RFs. A finding worth noting is the impact of injected randomness and the procedure of growing trees to produce optimal classification results. As has been experimentally proven, both algorithms are comparable, and each has a significant advantage over the other. It is, therefore, more advantageous to maximize on the advantages of each of these algorithms by using a hybrid model that will ensure the realization of highly accurate results that are also quite fast, rather than foregoing any of the advantages.

This being an initial study, it can further be enhanced to incorporate other predictive modelling approaches such as ANNs and Bayesian network model. Additionally, it is possible that the behaviors of loan default rates vary seasonally over a year due to circumstances that prevail over a year and the expenses tied to some specific dates throughout the year such as taxes and school fees, as well as seasonal variations in customer incomes. Future work will focus on developing more complex models accounting for seasonal variability.

More work can be done on how to grow the trees of RFs for optimal performance of decision trees, such as using multiple splitting ratios: 30/70, 40/60, 60/40, and 70/30. If the accuracies are fairly consistent then the algorithm is not sensitive to the number of samples. Hybrid models incorporating RF trees, SVMs, and other algorithms, also need to be thoroughly investigated and tested to develop more efficient systems. The future research opportunities that may be crucial for the prediction of credit risks may include the use of varying datasets, processing of the datasets to incorporate or remove various variables, research on the impact of each variable on the test performances, and modelling of varying problems.

ACKNOWLEDGEMENTS

This work was supported by FCT/MCTES through national funds and when applicable co-funded EU funds under the Project UIDB/EEA/50008/2020; by the Government of Russian Federation, Grant No. 08-08; by Brazilian National Council for Scientific and Technological Development (CNPq) via Grant No. 309335/2017-5; and by Ciência sem Fronteiras of CNPq, Brazil, through the Process No. 200450/2015-8.

ORCID

Joel J. P. C. Rodrigues https://orcid.org/0000-0001-8657-3800

REFERENCES

1. Lin WY, Hu YH, Tsai CF. Machine learning in financial crisis prediction: a survey. *IEEE Trans Syst Man Cybern Part C Appl Rev*. 2012;42(4):421-436. https://doi.org/10.1109/TSMCC.2011.2170420.
2. Stalla-Bourdillon S, Knight AE. Data Analytics and the GDPR: Friends or Foes? A Call for a Dynamic Approach to Data Protection Law. In: Leenes R, van Brakel R, Gutwirth S, de Hert P, eds. *Data Protection and Privacy: The Internet of Bodies*. Oxford: Hart Publishing; 2018;(March):1-22.

3. Wu X, Kumar V, Ross QJ, et al. Top 10 algorithms in data mining. *Knowl Inform Syst*. 2008;14(1):1-37. https://doi.org/10.1007/s10115-007-0114-2.
4. Crone SF. Finlay Steven. Instance sampling in credit scoring: an empirical study of sample size and balancing. *Int J Forecast*. 2012;28(1):224-238. https://doi.org/10.1016/J.IJFORECAST.2011.07.006.
5. Diederich J. *Rule Extraction from Support Vector Machines*. New York, NY: Springer; 2008.
6. Khondoker M, Dobson R, Skirrow C, Simmons A, Stahl D. A comparison of machine learning methods for classification using simulation with multiple real data examples from mental health studies. *Stat Methods Med Res*. 2016;25(5):1804-1823. https://doi.org/10.1177/0962280213502437.
7. Fassnacht FE, Latifi H, Stereńczak K, et al. Review of studies on tree species classification from remotely sensed data. *Remote Sens Env*. 2016;186:64-87. https://doi.org/10.1016/J.RSE.2016.08.013.
8. Breiman L. *Random Forests*. New York, NY: Kluwer Academic Publishers; 2001:5-32.
9. Altman EI. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *J Finance*. 1968;23(4):589-609. https://doi.org/10.2307/2329297.
10. Bekiroglu B, Takci H, Ekinci UC. Bank credit risk analysis with bayesian network decision. *IJAEST Int J Adv Eng Sci Technol*. 2011;9(2):273-279.
11. Antonakis AC, Sfakianakis ME. Assessing Naïve Bayes as a method for screening credit applicants. *J Appl Stat*. 2009;36(5):537-545. https://doi.org/10.1080/02664760802554263.
12. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recogn*. 1997;30(7):1145-1159. https://doi.org/10.1016/S0031-3203(96)00142-2.
13. Hachicha W, Ghorbel A. A survey of control-chart pattern-recognition literature (1991-2010) based on a new conceptual classification scheme. *Comput Ind Eng*. 2012;63(1):204-222. https://doi.org/10.1016/j.cie.2012.03.002.
14. Kabari LG, Nwachukwu EO. Credit risk evaluating system using decision tree – Neuro based model. *Int J Eng Res Tech*. 2013;2(6):2738-2745.
15. Pereira S, Portela F, Santos MF, Machado J, Abelha A. Predicting type of delivery by identification of obstetric risk factors through data mining. *Proc Comput Sci*. 2015;64:601-609. https://doi.org/10.1016/j.procs.2015.08.573.
16. Aven T. Risk assessment and risk management: review of recent advances on their foundation. *European J Operat Res*. 2016;253(1):1-13. https://doi.org/10.1016/j.ejor.2015.12.023.
17. Witzany J. *Credit Risk Management Pricing, Measurement, and Modeling*. New York, NY: Springer; 2017.
18. Fitch D. Structural equation modeling the use of a risk assessment instrument in child protective services. *Dec Supp Syst*. 2007;42(4):2137-2152. https://doi.org/10.1016/j.dss.2006.05.008.
19. Boser E, Vapnik N, Guyon IM. A training algorithm for optimal margin classifiers. Paper presented at: Proceedings of the 5th Annual Workshop on Computational Learning Theory; 1992:144-152; ACM.
20. Zhou H, Wang J, Wu J, Zhang L, Lei P, Chen X. Application of the hybrid SVM-KNN model for credit scoring. Paper presented at: Proceedings of the 2013 9th International Conference on Computational Intelligence and Security; 2013:174-177; IEEE.
21. Jiao L, Wang L, Gao X, Liu J, Wu F. Advances in natural computation. Paper presented at: Proceedings of the 2nd International Conference, ICNC 2006; September, 2006:24-28.
22. Davutyan N, Özar S. A credit scoring model for Turkey's micro & small enterprises (MSE's). Paper presented at: Proceedings of the 13th Annual ERF Conference; 2006:16-18; Kuwait.
23. Marqués AI, García V, Sánchez JS. Exploring the behaviour of base classifiers in credit scoring ensembles. *Expert Syst Appl*. 2012;39(11):10244-10250. https://doi.org/10.1016/j.eswa.2012.02.092.
24. Biau G. Analysis of a random forests model. *J Mach Learn Res*. 2012;13:1063-1095.
25. Sweeney D, AD, Williams T. *Statistics for Business and Economics*. 7th ed. London, UK: Thomson Learning EMEA; 2007.
26. Lessmann S, Baesens B, Seow HV, Thomas LC. Benchmarking state-of-the-art classification algorithms for credit scoring: an update of research. *European J Operat Res*. 2015;247(1):124-136. https://doi.org/10.1016/j.ejor.2015.05.030.
27. El-Shazly A. Financial distress and early warning signals: a non-parametric approach with application to Egypt. Paper presented at: Proceedings of the 9th Annual Conference of the Economic Research Forum, Sharjah, UAE; October, 2002:1-25; Emirates.
28. Provost Foster, Fawcett Tom. Analysis and visualization of classifier performance: comparison under imprecise class and cost distributions. Proceedings KDD-97; 1997:43-48.
29. Adams JB, Sabol DE, Kapos V, et al. Classification of multispectral images based on fractions of endmembers: application to land-cover change in the Brazilian Amazon. *Remote Sens Env*. 1995;52(2):137-154. https://doi.org/10.1016/0034-4257(94)00098-8.
30. Fawcett T. An introduction to ROC analysis. *Pattern Recog Lett*. 2006;27(8):861-874. https://doi.org/10.1016/j.patrec.2005.10.010.
31. Boardman, JW. Automating spectral unmixing of AVIRIS data using convex geometry concepts; 1993.

How to cite this article: Teles G, Rodrigues JJPC, Rabêlo RAL, Kozlov SA. Comparative study of support vector machines and random forests machine learning algorithms on credit operation. *Softw Pract Exper*. 2021;51:2492–2500. https://doi.org/10.1002/spe.2842