Combined Feature Selection and Rule Extraction for Credit Applicant Classification

Siham Akil[1(✉)], Sara Sekkate[2], and Abdellah Adib[1]

[1] Team: Data Science and Artificial Intelligence, Laboratory of Mathematics, Computer Science and Applications (LMCSA), Faculty of Sciences and Technologies, Mohammedia, Morocco
siham.ak2@gmail.com , abdellah.adib@fstm.ac.ma

[2] Higher National School of Arts and Crafts of Casablanca Hassan II University of Casablanca, Casablanca, Morocco

Abstract. A sensitive area such as credit risk assessment has always been a high priority and quite difficult for financial institutions to make financial decisions. In order to have a more relevant and accurate classification model, it is necessary for models to be interpreted and understood by financial professionals. The interaction and interpretation of the model becomes more limited as the number of features increases. In order to reduce the number of non significant features, feature selection must be implemented on the original data. The purpose of this study is not only to find a set of logical rules in an easily interpretable form for classification but also to select a set of informative features that can solve the classification problem. To this end, we present a combination of feature selection and rule extraction techniques. The experiment were carried out with the Australian Credit Dataset from the UCI Machine Learning Repository consisting of 13 attributes and a decision variable. The experimental results show that the used techniques are efficient in terms of interpretability, comprehensibility and accuracy.

Keywords: Credit scoring · Credit risk assessment · Rule extraction · Feature selection · Decision rules

1 Introduction

Almost all financial institutions – banks in particular – use credit scoring models to assess credit risk. These models employ a probabilistic measure of credit risk, often known as a risk rating or chance of default. The purpose of credit scoring is to ascertain how different borrower characteristics affect their likelihood of default, by using statistical methods that are calculated through software and

M. Ben Ahmed et al. (Eds.): SCA 2022, LNNS 629, pp. 97–104, 2023.
https://doi.org/10.1007/978-3-031-26852-6_9

historical data, typically the performance history of the borrowers or the history of the loans made. Additionally, they generate "scores," which are notes that quantify the default risk of present or prospective borrowers, allowing for the division of these distinct borrowers into risk classes: accepted or refused.

Support vector machines (SVM) [1], Random Forest (RF) [2], neural networks (NN) [3,4], and other machine learning techniques have been applied and have proven to be quite useful for credit risk assessment employing credit scoring models. However, some methodological approaches must be used in data processing to provide a performing model. As long as the number of features increases, the number of necessary calculations rises as well, residing that the interpretation and interactivity of the model is reduced. The solution is to implement two main techniques that solve these problems consisting of feature selection and rule extraction.

A combination of these two processes can be considered as a solution to these dimensionality and interpretability problems. Feature selection is the process of selecting the most consistent, non-redundant and narrowing down the set of features to those that are most relevant to the machine learning model [5]. It provides a simpler model, shorter training time, reduced variance thereby increasing accuracy and avoiding the curse of high dimensionality [6]. On the other hand, rule extraction is a technique that solves the black box problem of classification algorithms, their presumed complexity and the difficulty of fully understanding their underlying logic [7]. Hence, the need to present them in a more observable and understandable way, particularly in cases where the classification process is important, such as in sensitive areas like credit evaluation [8]. It consists in extracting simple, logical and understandable rules that can explain the behavior of machine learning models by revealing the internal knowledge of the trained models [9].

In this study, in order to ensure both a good interpretation and higher accuracy of credit scoring models and to improve the accuracy of the model, we focus on feature selection methods to retain only relevant features in credit evaluation and subsequently use rule extraction methods to extract logical rules and present them in a way that humans can easily interpret them. The rest of the paper is organized as follows: Sect. 2 summarizes the related work on credit scoring. The experimental setup and data description is given in Sect. 3. Results are reported and discussed in Sect. 4. Section 5 concludes the paper.

2 Related Works

Considerably, many studies have been conducted on feature selection for financial decision support, with the aim to reduce the original number of features and improve model performances [5]. Cheng-Lung Huang et al. [1] constructed three hybrid SVM-based approaches: SVM using grid search to optimize model parameters, SVM using grid search and F-score to select relevant features and SVM using Genetic Algorithm (GA) in order to both optimize and select relevant features. The authors in [2] used two Feature Selection (FS) techniques,

namely Random Forest (RF) and F-score, to select the most relevant features by combining them with the standard SVM, and subsequently compare them. Kozodoi, Nikita, et al. [10], designed a FS framework for credit scoring using GA NSGA-II which comprises three main steps: Fast non-dominated sorting, diversity preservation and population update. Yue Lie et al. [11] investigated existing GA approaches for FS, then they created and developed a method based on the usage of subpopulations with various types of data for fitness assessment. In order to deal with the problem of unbalanced data distribution, the authors of [12] aimed to evaluate the instances in terms of the entropy of their features. The authors in [13], examined multiple FS techniques such as Chi-Square, Information-Gain and Gain-Ratio and evaluated five Machine Learning (ML) classifiers such as Bayesian, Naïve Bayes, SVM, Decision Tree (C5.0) and RF with the aim to find the appropriate input predictors to build credit scoring models. For the credit scoring task, in conjunction with five ML classifiers, Akil et al. [5] have demonstrated the significance of filter and embedding feature selection strategies. These techniques were discovered to be faster, less to compute, and very efficient in producing an optimal subset that improved model performance. An effective credit scoring model requires a good trade-off between model accuracy and comprehensibility [14]. Indeed, classification algorithms are black boxes that do not offer good comprehensibility and interpretability [9]. In this context, various studies have been carried out on rule extraction for credit scoring problems [9,14]. In their study [15], Hruschka and Ebecken proposed a GA method for extracting rules from neural networks based on clustering the activation values of hidden units. To accurately categorize samples from a given data set, the authors first created a back-propagation network and trained it. This neural network model was then used to extract classification rules using the created Genetic Clustering Algorithm (CGA). Bazan et al. [16] aimed to compare the performance of a classifier based on computing all minimally consistent decision rules with a method based on two-layer learning. In the first learning layer, the collection of classifiers is derived from a portion of the original training dataset. Then, in the second layer, they guide the classifier using the model extracted from the created classifier based on the performance of the remaining training data. Hayashi et al. [17] considered a continuous Re-RX rule extraction method, in which an input unit was designed for each continuous attribute in the dataset and the discrete attributes were converted into a binary input string, with the aim of generating decision trees using J48graft that achieves both higher accuracy and good model interpretability. Sagi and Rokach [8] explored an approach where a decision forest is transformed into an interpretable decision tree. The methodology adopted consists of creating a set of rule conjunctions that represent the original decision forest; then they are hierarchically organized to build a new decision tree in order to find a compromise between accuracy and interpretability. Many authors have conducted extensive research to create robust credit scoring models using feature selection and other rule extraction techniques, but they have not yet been collectively studied in

conjunction. The above discussion shows that the search for an interpretable and robust credit scoring model remains a potentially important research area.

3 Background

3.1 Feature Selection

Feature selection is a crucial aspect in data preprocessing, useful on a variety of fronts such as curse dimensionality and reduction in training time. It aims at reducing the number of original features to an optimal subset, which can be used to provide equal or better results than the original set. For this study, we seek to explore the most relevant features to characterize the reliability of credit applicants. Below are the feature selection methods used in this work:

1- Analysis of variance: is a statistical test that measures the interdependence of two or more variables that differ significantly from each other, by calculating the F-test, which is the ratio of between-class variability to within-class variability that enables to determine the importance of a feature in the discriminant analysis. We compute the F-test value of a given feature as follows:

$$F = \frac{\sum_{j=1}^{J} N_j(\bar{x_J} - \bar{x})/(J-1)}{\delta^2} \tag{1}$$

where J is the number of classes, N_j is the number of instances in jth class, $\bar{x_J}$ is the mean of instances X in class j, $\bar{x_J}$ indicates the mean value for all instances, and δ^2 is the pool variance.

2- Kendall Rank Correlation : named after Maurice Kendall [18], who developed it in 1938, is a statistic measure that provides the ordinal connection between two quantities based on the τ coefficient calculated as follows [19]:

$$\tau = \frac{n_c - n_d}{\binom{n}{2}} \tag{2}$$

where n_c and n_d are the number of concordant and discordant pairs, respectively and $\binom{n}{k} = \frac{n(n-1)}{2}$ is the binomial coefficient for the number of items from n items. Kendall's τ coefficient of correlation varies between 1 and -1, it will take a value of 1 (high) when observations have an identical or similar rank, and a value of -1 (low) when the observations have a dissimilar rank between two features.

3- Random Forest: is a supervised ML algorithm, composed of many decision trees. In fact, each tree in the random forest can determine the importance of a feature based on its ability to increase node purity. The higher the increment in node purity, the higher the importance of the feature. This process is performed for each tree, averaged across all trees, and then normalized to 1. Therefore, the sum of the importance scores calculated by a random forest is equal to 1.

3- Gradient Boosting Decision Tree (GBDT): Feature selection using a gradient boosting decision tree falls into the category of embedded methods. After the construction of the boosted trees, an importance score is assigned to each feature. This score is calculated for each boosted decision tree based on the extent to which each feature separation point improves the performance measure, weighted by the number of observations for which the node is responsible [20]. The more an attribute is employed to form key decisions with decision trees, the upper is its relative importance.

3.2 Rule Extraction

1- Decision Tree: is a widely used data mining technique that is transparent and creates a set of production rules for decision making. It provides decision rules by extracting IF-THEN rules that contain potential information from the data. The IF-THEN rules make it easier to understand how the sample propagates through the tree during prediction. A rule is created for each path from the root to the leaf node to extract the rule from the decision tree. Each split criterion along the specified path is combined with the AND operator to form the antecedent (IF part) of the rule. The leaf node contains the class predictions that form the consequent part (THEN part) [21]. A decision tree can be transformed into an IF-THEN classification rule by tracing the path from the root node to each leaf node in the tree [22].

2- Random Forest: is a tree-based learning model composed of many decision trees [23]. When building individual trees, an attempt is made to use feature clustering and randomness to create an uncorrelated forest of trees where the predictions are more accurate than the predictions of the individual tree. Each node in each tree of a random forest can be transformed into an elementary rule. It leads to extract a large collection of rules from a set of trees that build and integrate multiple decision trees during the training phase, based on how often they appear [24]. Finally, the most common rules that represent robust and powerful data models are combined to form predictions.

4 Experimental Results and Discussion

In this work, Australian Credit Dataset, which is one of the most common datasets in the credit scoring field has been used to implement a combination of FS and rule extraction. The dataset has been obtained from the UCI Machine Learning Repository [25]. However, for confidentiality reasons, the names and values of all features are replaced with symbolic data. This includes 690 loan applicants with 14 features, including 383 creditworthy customers (Class 1) and 307 defaulting customers (Class 0). In this study, we randomly split this dataset into a 70% for training set and a 30% for testing set.

To evaluate our model, we use a very popular evaluation metric in ML to assess credit scoring models, that is, accuracy, which generally describes the performance of the model in all classes. Otherwise, the accuracy is the fraction of predictions that the model gets right.

Table 1. Reduced number of features.

Feature selection technique	Kendall Rank Correlation	ANOVA	RF	GBDT
Number of retained features	6	4	6	6

Table 1 shows the number of features held by each FS technique. Employing Kendall's rank correlation, ANOVA, RF, and GBDT, they have reduced the original number of the feature set by 7, 9, 7, and 7, respectively. Thus, only relevant features are selected and can be used for rule extraction.

From Table 2, we notice that the performance obtained using the feature selection techniques was almost similar and ranged from 77% to 88%. The best result was obtained using GBDT as a FS technique and Decision Tree (DT) as a classifier also devoted to rule extraction, as the decision tree algorithm produces accurate and interpretable models, it is also fast, both at the time of construction and application.

Table 2. Empirical results of our experiments.

Feature Selection technique	Rule Extraction method	Accuracy (%)	
		Train	Test
Kendall Rank Correlation	Decision Tree	85.3002	86.4734
	Random Forest	79.7101	77.2946
ANOVA	Decision Tree	86.9565	86.4734
	Random Forest	85.5072	86.9565
RF	Decision Tree	85.0931	86.4734
	Random Forest	85.7142	84.0579
GBDT	Decision Tree	86.3354	88.4057
	Random Forest	84.4720	85.9903

Table 3 shows that we obtained a total number of rules 7, 8, 7 and 8 respectively using DT and 24, 22, 23 and 24 using RF for Kendall Rank Correlation, ANOVA, RF and GBDT respectively. Subsequently, we eliminated all the insignificant rules that have an importance of 0, therefore we have an efficient, transparent and convincing system for financial professionals with fewer decision rules.

Interpretability of machine learning algorithms is necessary especially when the intended application involves important decisions. We argue that simplicity, stability and predictability are the minimum requirements for an interpretable model.

Table 3. Number of total and filtered rules.

Feature Selection technique	Rule Extraction method	Total number of rules	Number of filtered rules
Kendall Rank Correlation	DT	7	2
	RF	24	4
ANOVA	DT	8	3
	RF	22	6
RF	DT	7	2
	RF	23	6
GBDT	DT	8	2
	RF	24	4

5 Conclusion

In this paper, we examined four FS techniques in conjunction with two rule extraction methods for credit risk assessment using the Australian Credit dataset from the UCI ML repository. It was revealed that the combination of these techniques was a very important approach for credit scoring residing in a reduced number of features and fewer decision rules while maintaining a promising classification performance and provided a solution to the trade-off between model interpretability and accuracy. These results encourage us to pursue further research in this direction and in particular for the rule extraction techniques that deserve more attention.

Future studies could explore this approach further by using other FS and rule extraction techniques to improve the performance and comprehensibility of the models. We also aim to test our approach on other databases.

Acknowledgements. This work was supported by the Ministry of Higher Education, Scientific Research and Innovation, the Digital Development Agency (DDA) and the CNRST of Morocco (Alkhawarizmi/2020/01).

References

1. Huang, C.L., Chen, M.C., Wang, C.J.: Credit scoring with a data mining approach based on support vector machines. Expert Syst. Appl. **33**(4), 847–856 (2007)
2. Shi, J., Zhang, S.Y., Qiu, L.M.: Credit scoring by feature-weighted support vector machines. J. Zhejiang Univ. Sci. C **14**(3), 197–204 (2013)
3. Ha, V.S., Nguyen, H.N.: Credit scoring with a feature selection approach based deep learning. MATEC Web Conf. **54**, 1–5 (2016)
4. Munkhdalai, L., Namsrai, O.E., Ryu, K.H.: Credit scoring with deep learning. In: 4th International Conference on Information, System and Convergence Applications, pp.1–5 (2018)

5. Siham, A., Sara, S., Abdellah, A.: Feature selection based on machine learning for credit scoring : an evaluation of filter and embedded methods. In: 2021 International Conference on Innovations in Intelligent SysTems and Applications, INISTA 2021 - Proceedings (2021)
6. Rtayli, N., Enneya, N.: Selection features and support vector machine for credit card risk identification. Procedia Manufact. **46**, 941–948 (2020)
7. Friedman, J.H., Popescu, B.E.: Predictive learning via rule ensembles. Annal. Appl. Stat. **2**(3), 916–954 (2008)
8. Sagi, O., Rokach, L.: Explainable decision forest: transforming a decision forest into an interpretable tree. Inf. Fusion **61**, 124–138 (2020)
9. Hayashi, Y.: Application of a rule extraction algorithm family based on the Re-RX algorithm to financial credit risk assessment from a Pareto optimal perspective. Oper. Res. Perspectives **3**, 32–42 (2016)
10. Kozodoi, N., Lessmann, S., Papakonstantinou, K., Gatsoulis, Y., Baesens, B.: A multi-objective approach for profit-driven feature selection in credit scoring. Decis. Support Syst. **120**(March), 106–117 (2019)
11. Liu, Y., Ghandar, A., Theodoropoulos, G.: island model genetic algorithm for feature selection in non-traditional credit risk evaluation, pp. 2771–2778 (2019)
12. Carta, S., Ferreira, A., Recupero, D.R., Saia, M., Saia, R.: A combined entropy-based approach for a proactive credit scoring. Eng. Appl. Artif. Intell. **87**, 103292 (2020)
13. Trivedi, S.K.: A study on credit scoring modeling with different feature selection and machine learning approaches. Technol. Soc. **63**, 101413 (2020)
14. Mashayekhi, M., Gras, R.: Rule extraction from decision trees ensembles: new algorithms based on heuristic search and sparse group lasso methods. Int. J. Inf. Technol. Decision Making **16**(06), 1707–1727 (2017)
15. Hruschka, E.R., Ebecken, N.F.: Applying a clustering genetic algorithm for extracting rules from a supervised neural network. Proceed. Int. Joint Conf. Neural Netw. **3**(2), 407–412 (2000)
16. Bazan, J.G.: Classifiers Based on Two-Layered Learning. In: Tsumoto, S., Słowiński, R., Komorowski, J., Grzymała-Busse, J.W. (eds.) RSCTC 2004. LNCS (LNAI), vol. 3066, pp. 356–361. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-25929-9_42
17. Hayashi, Y., Oishi, T.: High accuracy-priority rule extraction for reconciling accuracy and interpretability in credit scoring. N. Gener. Comput. **36**(4), 393–418 (2018). https://doi.org/10.1007/s00354-018-0043-5
18. Kendall, M.G.: A new measure of rank correlation. Biometrika **30**(1/2), 81–93 (1938)
19. Kendall, M.G.: Rank correlation methods (1948)
20. Tian, Z., Xiao, J., Feng, H., Wei, Y.: Credit risk assessment based on gradient boosting decision tree. Procedia Comput. Sci. **174**, 150–160 (2020)
21. Griselda, L., Joaquín, A., et al.: Using decision trees to extract decision rules from police reports on road accidents. Procedia. Soc. Behav. Sci. **53**, 106–114 (2012)
22. Vasilev, N., Mincheva, Z., Nikolov, V.: Decision tree extraction using trained neural network. SMARTGREENS 2020 - Proceedings of the 9th International Conference on Smart Cities and Green ICT Systems, pp. 194–200 (2020)
23. Mashayekhi, M., Gras, R.: Rule extraction from random forest: the RF+HC methods. In: Barbosa, D., Milios, E. (eds.) CANADIAN AI 2015. LNCS (LNAI), vol. 9091, pp. 223–237. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-18356-5_20
24. Bénard, C.: SIRUS interpretable RF, vol. 130 (2021)
25. Dua, D., Graff, C.: UCI machine learning repository (2017)