The best of two worlds: Balancing model strength and comprehensibility in business failure prediction using spline-rule ensembles

Koen W. de Bock

The Best of Two Worlds: Balancing Model Strength and Comprehensibility in Business Failure Prediction Using Spline-Rule Ensembles

Koen W. De Bock

Audencia Business School

8 Route de la Jonelière, BP31222, 44312 Nantes, France

Corresponding author: Koen W. De Bock: kdebock@audencia.com, +33 (0)240373400

This article is published in **Expert Systems With Applications**.

Please cite as:

De Bock, K. W. (2017). The best of two worlds: Balancing model strength and comprehensibility in business failure prediction using spline-rule ensembles. *Expert Systems with Applications*, Vol. 90, 23-39.

Article doi: http://dx.doi.org/10.1016/j.eswa.2017.07.036

Abstract

Numerous organizations and companies rely upon business failure prediction to assess and minimize the risk of initiating business relationships with partners, clients, debtors or suppliers. Advances in research on business failure prediction have been largely dominated by algorithmic development and comparisons led by a focus on improvements in model accuracy. In this context, ensemble learning has recently emerged as a class of particularly well-performing methods, albeit often at the expense of increased model complexity. However, in practice, model choice is rarely based on predictive performance alone. Models should be comprehensible and justifiable to assess their compliance with common sense and business logic, and guarantee their acceptance throughout the organization. A promising ensemble classification algorithm that has been shown to reconcile performance and comprehensibility are rule ensembles. In this study, an extension entitled *spline-rule ensembles* is introduced and validated in the domain of business failure prediction. Spline-rule ensemble complement rules and linear terms found in conventional rule ensembles with smooth functions with the aim of better accommodating nonlinear simple effects of individual features on business failure. Experiments on a large selection of 21 datasets of European companies in various sectors and countries (i) demonstrate superior predictive performance of spline-rule ensembles over a set of well-established yet powerful benchmark methods, (ii) show the superiority of spline-rule ensembles over conventional rule ensembles and thus demonstrate the value of the incorporation of smoothing splines, (iii) investigate the impact of alternative term regularization procedures and (iv) illustrate the comprehensibility of the resulting models through a case study. In particular, the ability of the technique to reveal the extent and the way in which predictors impact business failure, and if and how variables interact, are exemplified.

Keywords

Bankruptcy prediction, business failure prediction, data mining, ensemble learning, model comprehensibility, penalized cubic regression splines, rule ensembles, spline-rule ensembles, risk management

1 Introduction

The global financial crisis of 2007-2008 and the subsequent economic recession initiated a wave of companies in financial distress. In the European Union alone, in 2009, over 178,000 companies became insolvent, an increase of 19% in comparison to 2008. While in the years following no significant change could be observed, figures for 2012 increased by another 9.1 percent in comparison to 2011 (Creditreform, 2014). Company insolvency or bankruptcy affects and thus represents a risk to all stakeholders involved, ranging from capital investors, creditors, suppliers, tax collection agencies, employees to customers.

In general, companies increasingly rely upon principles of enterprise risk management (ERM) to face and manage risks. ERM prescribes the development and execution of integrated strategies and processes to anticipate, face and overcome risks (Wu, Chen, & Olson, 2014; Wu & Olson, 2010; Wu, Olson, & Dolgui, 2015). ERM subdomains include investment risk evaluation (e.g. Wu, Zheng, & Olson, 2014), accounts receivable risk management (e.g. Baesens, et al., 2003; Lessmann, Baesens, Seow, & Thomas, 2015; Wu, Olson, & Luo, 2014) and vendor selection (e.g. Wu & Olson, 2010; Wu, Zhang, Wu, & Olson, 2010). Risk management often relies upon business intelligence nowadays, and more specifically, data mining (Wu, Chen, et al., 2014). In this context, this paper focuses on models for business failure prediction (BFP) that are widely used as early warning systems for financial distress or bankruptcy in partnering companies.

A BFP model generalizes the relation between business failure and a range of variables characterizing the company, its activities and performance in the past. Consider the following notation. T is a data set containing historical data, denoted the training data set, with information on n companies, described by a set of p predictive features x_1 to x_p and the binary outcome variable y that indicates whether a business failed (y=1) or survived (y=0). A BFP model is any function $F(\mathrm{x})$ that maps a given instance x to a conditional bankruptcy probability. Once estimated, the model allows the analyst to $score$, i.e., to produce estimations of future business failure for a new set of companies based upon their current profile and performance.

Since the 1960s, business failure prediction is an active domain of research. Over the years, different techniques for finding $F(\mathrm{x})$ have been introduced and compared. Globally, a distinction is often made between statistical and data mining techniques (Ravi Kumar & Ravi, 2007). The basis of the domain and the first category of techniques is formed by (Beaver, 1966) and (Altman, 1968). While both studies depend upon the usage of financial ratio's, the latter one was the first to deploy a multivariate statistical technique, linear discriminant analysis (LDA), to discriminate between failing and non-failing companies. Martin (1977) and Ohlson (1980) experimented with the usage of logistic regression for business failure prediction. Until today, both LDA and logistic regression remain popular candidate algorithms in industry to develop models for BFP and have served as benchmark algorithms in many comparative studies. Other statistical methods include probit regression (Grablowsky & Talley, 1981) and linear probability models (Meyer & Pifer, 1970).

By far, the majority of methodological contributions in the business failure prediction literature has focused upon methods originating from the data mining and machine learning literature. In this category one can cite artificial neural networks (Atiya, 2001; Pendharkar, 2005), decision trees (Frydman, Altman, & Kao, 1985), support vector machines (Li & Sun, 2011a), Bayesian networks (Sun & Shenoy, 2007), rough sets (McKee, 2003), k-nearest neighbors (Park & Han, 2002), association rules (Janssens, Wets, Brijs, & Vanhoof, 2005) and finally, ensemble learners (Li & Sun, 2011b). A comprehensive review of statistical and data mining techniques used for business failure prediction can be found in Ravi Kumar and Ravi (2007).

A special subcategory of data mining techniques which has received a growing amount of attention in BFP literature are ensemble learners. In recent years, the practice of combining predictions from single algorithms has become a popular topic in theoretical and applied research (Rodríguez, Kuncheva, & Alonso, 2006; van Wezel & Potharst, 2007; Xu, Krzyzak, & Suen, 1992). The predictions of ensemble learners are taken as combinations of the individual ensemble member predictions. The main factor defining the popularity of ensemble algorithms is their high level of predictive accuracy that has been observed within multiple comparative studies in various domains and applications (e.g. Bauer & Kohavi, 1999; Dietterich, 2000). An ensemble of individual prediction models is likely to generate better and more robust predictions than a single algorithm if accuracy and diversity are simultaneously present amongst the ensemble members. Several studies have demonstrated the strong performance of ensemble learners in the field of BFP (e.g. Verikas, Kalsyte, Bacauskiene, & Gelzinis, 2010).

Alternatively, techniques used for business failure prediction can be classified according to their ability to provide insight into the relationship between predictive features and business failure. It is often noted that in BFP literature predictive performance dominates as an evaluation criterion in benchmark studies, accuracy should not improve at the expense of model comprehensibility (Wu, 2010). As noted in Olson, Delen and Meng (2012), the *transparency* of data mining models for BFP is a highly desirable feature as (i) stakeholders share a need to understand the relative influence of financial and company-specific indicators on business failure, and (ii) increased comprehensibility makes models more *transportable;* i.e. easily applicable to new data sets or alternative business settings. While ensemble learners have received critical acclaim for their ability to generate accurate predictions, the practice of combining models introduces a level of complexity making such models difficult to understand. Similar to methods such as artificial neural networks, ensemble classifiers are sometimes criticized for their black box nature. Very few studies in BFP have evaluated ensemble methods in function of comprehensibility.

A promising technique designed to combine the merits of ensemble learners with a high degree of interpretability are *rule ensembles* (Friedman & Popescu, 2008). Similar to many ensemble learners, rule ensembles first generate a set of decision trees. However, in a subsequent phase, the technique decomposes trees into rules and only retains a compact set of rules derived from these trees through the application of regularized lasso regression. To account better for linear effects, the original features are also added as linear terms to the lasso regression. The simple structure of resulting models allows straightforward model interpretation, and the rule ensemble algorithm incorporates a number of additional instruments to gain insight into the model's functioning.

This study evaluates rule ensembles for business failure prediction and delivers a methodological contribution as a novel extension of the rule ensemble framework is proposed, entitled *spline-rule ensembles* (SRE). Spline-rule ensembles complement rules and linear terms by smooth terms (single-term penalized cubic regression splines) in order to better accommodate univariate, nonlinear relationships between the probability of bankruptcy, and individual explanatory variables.

The contributions of this paper are the following: (i) spline-rule ensembles are introduced to the field of BFP as a novel model category reconciling strong accuracy and advanced model interpretability, (ii) spline-rule ensembles are proposed as a natural extension of generic rule ensembles whereby smooth functions are added to rules and linear terms; (iii) experiments are conducted on a large set of 21 datasets containing information for European companies in various sectors to compare spline-rule ensemble to conventional rule ensembles and a set of benchmark algorithms in terms of several criteria of predictive performance, and (iv) through a case study, the comprehensibility of spline-rule ensembles is demonstrated.

The remainder of this article is structured as follows. In Section 2, an overview is given of related literature. In particular, the usage of ensemble learning in the domain of business failure prediction is addressed. Then, rule ensembles are explained in detail. In Section 3, spline-rule ensembles and their training process are introduced. Section 4 presents the experimental setup of this study whilst in Section 5, the results are described. This section first addresses the results of a benchmark study in terms of predictive performance (Section 5.1) and then presents the various deliverables of the rule ensemble technique that contribute to model insight (Section 5.2). A final Section concludes the study and addresses limitations to the study and directions for future research.

2 Related Literature

2.1 Ensemble learning for business failure prediction

Ensemble learners have become a popular algorithm choice in the field of business failure prediction over the past 10 years. The rationale of ensemble learning is straightforward: predictions of ensemble learners are taken as combinations of probability or class predictions delivered by multiple *ensemble members* or *base learners* (Kuncheva, 2004). An important factor explaining the popularity of ensemble algorithms at present is the strong predictive performance that is observed within multiple comparative studies (Sun, Li, Huang, & He, 2014). An ensemble of individual prediction models is likely to generate better and more robust predictions than a single algorithm when both accuracy and *diversity* are present amongst the ensemble members.

Applications of ensemble learning in BFP can be categorized according to whether the ensemble learner consists of ensemble members that belong to various algorithm classes, or whether it consists of multiple replications of a single algorithm. In the majority of ensemble learning applications in BFP literature, models originating from multiple algorithm classes are combined and the resulting ensemble learners are thus called *hybrid* ensembles. Early applications of hybrid ensembles in business failure prediction include an ensemble combining a multilayer perceptron, case-based reasoning and discriminant analyses through weighted averaging (Jo & Han, 1996) and the hybrid classifier proposed in Olmeda and Fernández (1997) consisting of a multilayer perceptron, linear discriminant analysis, logistic regression, MARS and a C4.5 decision tree. Other, more recent hybrid ensemble approaches can be found in Ravi, Kurniawan, Thai, and Kumar (2008) and Sun and Li (2008).

Other applications involve *homogeneous* ensemble classifiers, whereby a single base learner algorithm is chosen and replicated multiple times to constitute an ensemble. In this category, two classic approaches are bagging (Breiman, 1996) and boosting (Freund & Schapire, 1997). In the former, an ensemble is constructed by training ensemble members on bootstrap samples of the training data set. In the latter, in an iterative process, the algorithm is forced to attribute higher importance to observations that were misclassified during earlier rounds, either by reweighing or by resampling the training data set. Bagging and AdaBoost have been by far the most extensively researched homogeneous ensemble classifiers in the domain of BFP (e.g. Alfaro, García, Gámez, & Elizondo, 2008; Cortes, Martinez, & Rubio, 2007; Sun, Jia, & Li, 2011; West, Dellana, & Qian, 2005). More recently, the strong performance of bagging and AdaBoost was confirmed and found similar to random forests, another well-known homogenous ensemble learning algorithm (Barboza, Kimura, & Altman, 2017). Finally, Zięba, Tomczak, & Tomczak (2016) introduce a novel method to the domain entitled *extreme gradient boosting* and demonstrate its superiority over a large set of benchmark algorithms in the context of business failure prediction in Poland. For a comprehensive reviews on the usage of ensemble learning in the field of business failure prediction, see Verikas, et al. (2010) and Sun, et al. (2014).

2.2 Rule Ensembles

Rule ensembles (Friedman & Popescu, 2008) constitute a predictive method that combines principles of ensemble learning and semi- parametric regression. A rule ensemble derives simple rules from a training data set and then combines them linearly, as terms in an additive equation. The method belongs to the category of homogenous ensemble learners (Xie, Li, Ngai, & Ying, 2009).

Rule ensembles differ from other ensemble learning methods on three levels: (i) the members that constitute the ensemble, (ii) the combination rule used to combine individual predictions, and (iii) options for model interpretation. First, a rule ensemble is an ensemble of rules and linear terms. To come to this ensemble, the technique first generates a number of decision trees from the training data set T while subsequently, a library of a large number of *rules* is derived by, for every node within every tree (both interior and terminal), formulating the conditions that define the path down the tree to reach the node as a rule. Any rule r_j(x) takes the form

$$r_j(x) = \prod_{s_{jk} \neq s_k} I(x_k \in s_{jk}) \qquad (1)$$

i.e. a product of indicator functions whereby each indicator function $I(x_k \in s_{jk})$ represents a node condition and thus resolves to 1 if the value x_k falls within the interval or set s_{jk}. In other words, a rule resolves to 1 for companies for which all its conditions are true and to 0 otherwise. Note that the product function is limited to factors where the subset of values that define the condition is not equal to the entire value set S_k. Further, to enhance both accuracy and interpretability, the original variables $x_k;\ k = 1, \dots, p$ are added to the set of rules as additional candidate base learners for the final ensemble. As such, the model is better able to capture potential linear relationships between variables and the odds of business failure. Specifically, a new intermediate training data set *T'* is created by merging the rule set with the original training data set *T*. The final model is trained using *T'* and takes the form

$$F(x) = \hat{a}_0 + \sum_{j=1}^{q} \hat{a}_j r_j(x) + \sum_{k=1}^{p} \hat{b}_k l_k(x_k) \qquad (2)$$

with q the total number of rules generated and functions $l_k(x_k)$ denoting *winsorizations* of the original variables, i.e., truncations of the variable values below and above the βth and *(1-β)*th percentiles (indicated by δ_k^- and δ_k^+), respectively:

$$l_k(x_k) = min\ (\delta_k^+, max\ (\delta_k^-, x_k)). \qquad (3)$$

Second, rule ensembles differ in terms of the combination rule used to combine member predictions. Whilst many ensemble algorithms apply basic methods such as majority voting or averaging, rule ensembles train a regression model to this end. Specifically, to find coefficients $\hat{a}_j;\ j = 0, \dots, q$ and $\hat{b}_k;\ k = 1, \dots, p$ for this combination function, a linear regularized lasso-regression is applied. The advantages over using heuristic combination methods are twofold. First, the regularization enforces model *shrinkage*. Typically, many rule parameters will be set to 0, so that a large library of rules and linear terms is reduced to a smaller subset that is more easily interpretable. This makes the model more generalizable, leading to more accurate predictions and a better understanding of the data generation process. Second, selected terms (rules or linear terms) obtain a coefficient indicating whether it contributes positively or negatively to a prediction, and to which extent. The regularized lasso regression takes the form

$$\left(\{\hat{a}_j\}_0^q, \{\hat{b}_k\}_1^p\right) =$$

$$\underset{\{\hat{a}_j\}_0^q, \{\hat{b}_k\}_1^p}{arg\,min} \sum_{i=1}^n L\left(y_i, \hat{a}_0 + \sum_{j=1}^q \hat{a}_j r_j(x_i) + \sum_{k=1}^p \hat{b}_k l'_k(x_{ik})\right) + \lambda\left(\sum_{j=1}^q \left|\hat{a}_j\right| + \sum_{k=1}^p \left|\hat{b}_k\right|\right) \quad (4)$$

with λ denoting shrinkage parameter: larger values of λ will penalize the attribution of coefficient to less predictive rules or variables. As a result, many coefficients will be set to zero when the value of λ is increased.

Third, the method implements a number of instruments that make model interpretation straightforward. Apart from rule and linear terms and term coefficients, these indicators include variable importance measures, interaction strengths and partial dependence functions. These are discussed and illustrated in detail in Section 5.

Several advantages motivate the adoption of rule ensembles in business failure prediction. First, the technique has demonstrated highly competitive predictive performance (Friedman & Popescu, 2008). Second, unlike other ensemble methods, the resulting rule ensemble model is very easy to interpret. Third, data pre-processing such as feature selection can be omitted while model post-processing such as pruning is not necessary. Fourth, rules ensembles can easily handle high-dimensional data both in terms of number of observations and number of features.

3 Spline-Rule Ensembles

Rule ensembles are very flexible and due to their nature they automatically detect and accommodate two-way and higher-order interaction terms (through the inclusion and selection of rules) as well as linear terms (through the inclusion of winsorized linear terms). Non-linear correlations between individual variables and the outcome variable can also emerge in the model, but only in an indirect fashion, through an interplay of multiple rules.

Spline-rule ensembles are proposed in this study as a natural extension to conventional rule ensembles. Other than rules and linear terms, spline-rule ensembles introduce smooth functions of individual continuous variables as a third term category as a more direct strategy to accommodate non-linear effects in the model. In particular, penalized cubic regression splines (Wood, 2006) are chosen for smooth functions. Penalized cubic regression splines model the functional relation between the logit of the failure probability on a variable x by defining a set of u knots $\xi_1, \xi_2, \ldots, \xi_u$ on the range of the variable, and estimating a function that is built up of cubic polynomials between every pair of adjacent knots. Hence, $s(x)$ takes the form

$$s(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 h(x, \xi_1) + \cdots + \beta_{u+3} h(x, \xi_u)$$

$$with\ h(x, \xi) = \begin{cases} (x - \xi)^3 \text{ if } x > \ \xi \\ \quad 0 \text{ otherwise} \end{cases} \qquad (5)$$

A solution can be found by minimizing

$$\sum_{i=1}^{n} (y_i - \ s(x_i))^2 - \rho \int (s"(x))^2 dx \qquad (6)$$

where $s(x)$ is the cubic regression spline function and ρ is the smoothing parameter, a penalty term that is required to penalizes excessive curvature in the function. This study suggests an automated optimization of smoothness (i.e., parameter ρ) based upon the generalized cross-validation (GCV) criterion (Craven & Wahba, 1979; Wood, 2004).

Given the addition of smooth functions, the regularized regression in spline-rule ensembles takes the form

$$\left(\{\hat{a}_j\}_0^q, \{\hat{b}_k\}_1^p \{\hat{c}_g\}_1^u\right) =$$

$$\underset{\{\hat{a}_j\}_0^q, \{\hat{b}_k\}_1^p \{\hat{c}_k\}_1^u}{arg\,min} \sum_{i=1}^n L\left(y_i, \hat{a}_0 + \sum_{j=1}^q \hat{a}_j r_j(x_i) + \sum_{k=1}^p \hat{b}_k l'_k(x_{ik}) + \sum_{g=1}^u \hat{c}_g s_g(x_{ig})\right) + (1-\alpha)\left(\sum_{j=1}^q |\hat{a}_j| + \sum_{k=1}^p |\hat{b}_k| + \sum_{g=1}^u |\hat{c}_g|\right) + \alpha\left(\sum_{j=1}^q \hat{a}_j^2 + \sum_{k=1}^p \hat{b}_k^2 + \sum_{g=1}^u \hat{c}_g^2\right)$$

$$with \quad \alpha = \frac{\lambda_2}{\lambda_1+\lambda_2} \text{ and } 0 \leq \alpha \leq 1 \qquad (7)$$

Note that the regularized regression represented in equation (7) is the elastic net proposed by Zou and Hastie (2005) which can be seen as a generalization of ridge regression and lasso-regularized regression that combines the strengths and avoids weaknesses of both methods. The adoption of a more flexible form of regularization in spline-rule ensembles is inspired by the secondary objective of this study to experimentally compare alternative methods. Unlike ridge regression, lasso regression results in variable selection rather than mere parameter shrinkage, but has been found to underperform when multicollinearity occurs and is characterized by an undesirable degree of randomness when selecting one variable out of a group of correlating ones. Moreover, lasso regression does not perform well when there are more variables than observations which is a potential problem in rule ensembles since they prescribe term selection from a large library of rules and terms. Note that for $\alpha = 1$, equation 4 resolves to ridge regression and to lasso regression when $\alpha = 0$.

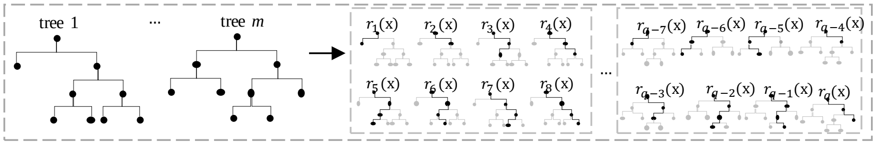
Figure 1 provides a schematic representation of the training process of spline-rule ensemble models.

Training process input: data set T with p predictor variables and outcome variable y for n observations

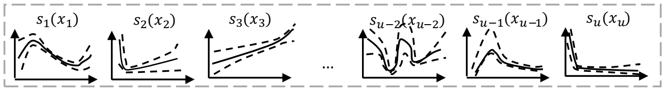
$$T = \begin{array}{c} \begin{array}{ccccc} x_1 & x_2 & & x_p & y \end{array} \\ \left[\begin{array}{ccccc} x_{11} & x_{12} & \cdots & x_{1p} & y_1 \\ x_{21} & x_{22} & \cdots & x_{2p} & y_2 \\ x_{31} & x_{32} & \cdots & x_{3p} & y_3 \\ \vdots & \vdots & \ddots & \cdots & y_n \\ x_{n1} & x_{n2} & \cdots & x_{np} & y_n \end{array}\right] \end{array}$$

Phase 1: prepare rules, splines and linear terms

Phase 1a: grow m trees and decompose into rules (i.e., rule membership indicators)



Phase 1b: estimate smooth functions (splines) for continuous features



Phase 1c: apply winsorization and normalization to variables to create linear terms

$$\mathrm{l'_k(x_k)} = 0.4\frac{\mathrm{l_k(x_k)}}{std(\,\mathrm{l_k(x_k)})} \text{ with } \mathrm{l_k(x_k)} = \min(\delta_k^+, \max(\delta_k^-, \mathrm{x_k}))\ \forall\, k \in \{1, \ldots, p\}$$

Phase 2: create term selection matrix T' by merging rule membership indicators $r_j(x)$ with splines $s_j(x_j)$ and original predictor variables x_k

$$T' = \begin{array}{c} \begin{array}{cccccccccccccc} r_1 & r_2 & & r_q & s_1 & s_2 & & s_u & l'_1 & l'_2 & & l'_p & y \end{array} \\ \left[\begin{array}{ccccccccccccc} r_{11} & r_{12} & \cdots & r_{1q} & s_{11} & s_{12} & \cdots & s_{1u} & x_{11} & x_{12} & \cdots & x_{1p} & y_1 \\ r_{21} & r_{22} & \cdots & r_{2q} & s_{21} & s_{22} & \cdots & s_{2u} & x_{21} & x_{22} & \cdots & x_{2p} & y_2 \\ r_{31} & r_{32} & \cdots & r_{3q} & s_{31} & s_{32} & \cdots & s_{3u} & x_{31} & x_{32} & \cdots & x_{3p} & y_3 \\ \vdots & \vdots & \ddots & \cdots & \vdots & \vdots & \ddots & \cdots & \vdots & \vdots & \ddots & \cdots & y_n \\ r_{n1} & r_{n2} & \cdots & r_{nq} & s_{n1} & s_{n2} & \cdots & s_{nu} & x_{n1} & x_{n2} & \cdots & x_{np} & y_n \end{array}\right] \end{array}$$

Phase 3: apply regularized regression to T'

$$\left(\{\hat{a}_j\}_1^q, \{\hat{b}_g\}_1^u \{\hat{c}_k\}_1^p\right) = \underset{\{\hat{a}_j\}_0^q, \{\hat{b}_g\}_1^u \{\hat{c}_k\}_1^p}{argmin} \sum_{i=1}^{n} \begin{array}{c} L\left(y_i, \hat{a}_0 + \sum_{j=1}^{q} \hat{a}_j r_j(\mathrm{x}_i) + \sum_{g=1}^{u} \hat{b}_g s_g(x_{ig}) + \sum_{k=1}^{p} \hat{c}_k l'_k(x_{ik})\right) \\ + \lambda\left(\sum_{j=1}^{q} |\hat{a}_j| + \sum_{g=1}^{u} |\hat{b}_g| + \sum_{k=1}^{p} |\hat{c}_k|\right) + (1-\lambda)\left(\sum_{j=1}^{q} \hat{a}_j^2 + \sum_{g=1}^{u} \hat{b}_g^2 + \sum_{k=1}^{p} \hat{c}_k^2\right) \end{array}$$

Training process output: rule ensemble model (with example)

$$\hat{F}(\mathrm{x}) = \hat{a}_0 + \sum_{j=1}^{q} \hat{a}_j r_j(\mathrm{x}) + \sum_{g=1}^{u} \hat{b}_g l'_g(x_g) + \sum_{k=1}^{p} \hat{c}_k l'_k(x_k)$$

E.g. $$\hat{F}(\mathrm{x}) = \hat{a}_0 + \hat{a}_2 r_2(\mathrm{x}) + \hat{a}_7 r_7(\mathrm{x}) + \hat{a}_{13} r_{13}(\mathrm{x}) c + \hat{b}_4 s_4(x_4) + \hat{c}_2 l'_2(x_2)$$

Figure 1: Schematic representation of spline-rule ensemble model training process.

As can be seen in Figure 1, the training process takes a data set as input and involves 3 subsequent phases: (i) training of classification trees, (ii) rule set derivation, (iii) evolved training data set creation and (iv) regularized regression. The output of the training process is a rule ensemble model of which the exact form depends upon model shrinkage, i.e. the terms selected through the regularized regression.

Figure 2 illustrates the training process of the spline-rule ensemble algorithm further by means of an example for a simplified data set containing 5 continuous variables (V1-V5). The figure shows how the term selection matrix is populated with rules, splines and linear terms. It also illustrates how rules are derived from decision trees. Linear terms are simple transformations of the input variables and are graphically represented by their class-conditional density curves. The final model has, through lasso regularization, out of a total of 26 (16 rules, 5 splines and 5 linear terms) terms, selected 5 (4 rules and one smooth term). Note that this example simultaneously illustrates the comprehensible nature of the model.

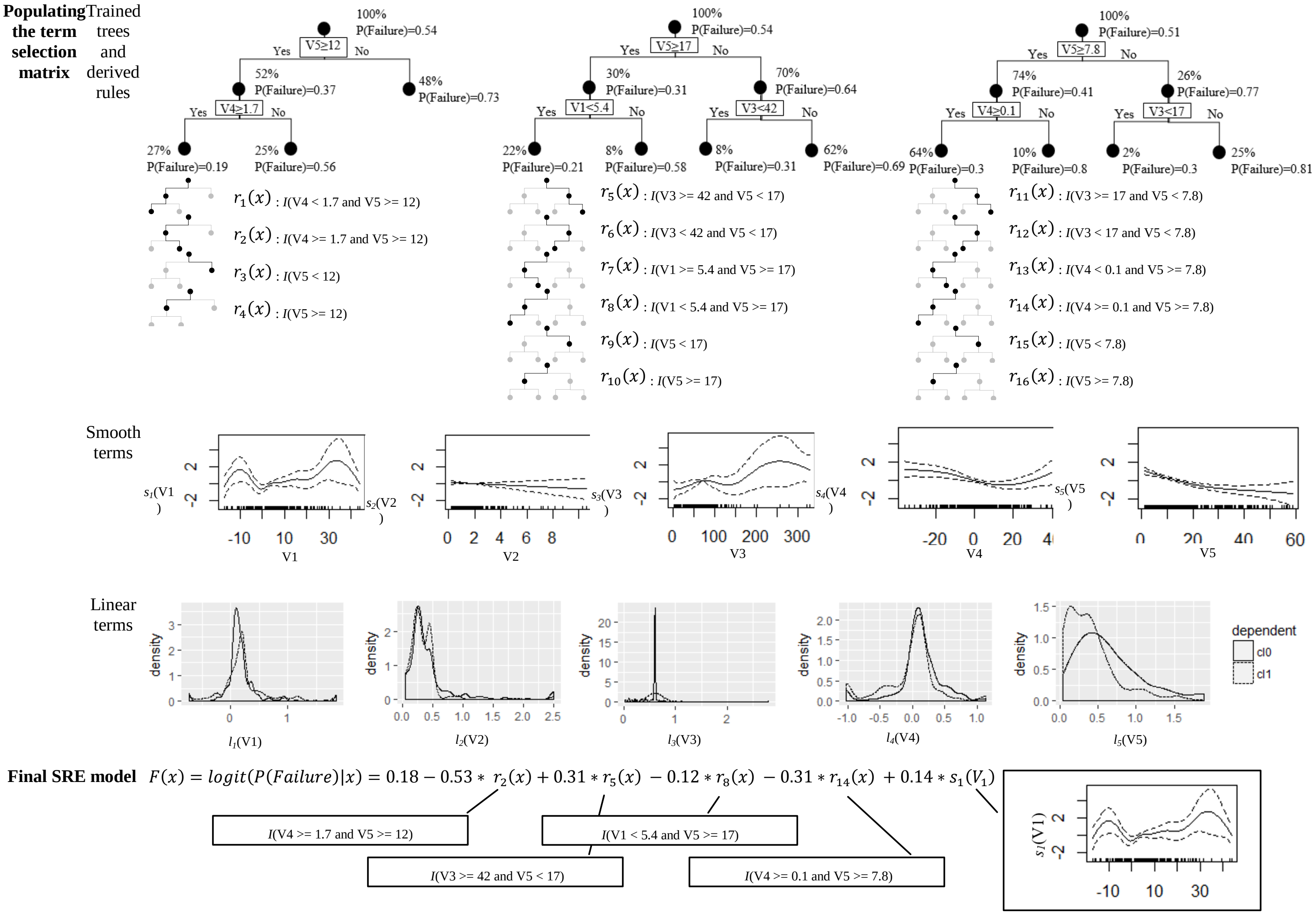


Figure 2: Illustration of the spline-rule ensemble training algorithm through application on a simplified data set.

4 Experimental Setup

4.1 Data Description

To assess and benchmark the accuracy and interpretability of spline-rule ensembles, experiments are set up using 21 datasets provided by two global data aggregators. These datasets contain information for a large selection of Belgian, French and Italian companies from various industries. Table 1 contains detailed information on the data sets considered in this study.

Data set	Country	SIC Bin	SIC Bin Definition	Nbr. of Companies	Nbr. of Features	Failure Rate
ds1	Belgium	15.000.000 <= SIC 8 < 18.000.000	Construction industries	9,976	108	4.54%
ds2	Belgium	20.000.000 <= SIC 8 < 40.000.000	Manufacturing	10,430	108	2.73%
ds3	Belgium	40.000.000 <= SIC 8 < 50.000.000	Transportation, communications and utilities	5,339	108	4.57%
ds4	Belgium	50.000.000 <= SIC 8 < 52.000.000	Wholesale trade	15,896	108	3.04%
ds5	Belgium	52.000.000 <= SIC 8 < 60.000.000	Retail trade	13,626	108	5.19%
ds6	Belgium	60.000.000 <= SIC 8 < 68.000.000	Finance, insurance and real estate	10,055	108	1.64%
ds7	Belgium	70.000.000 <= SIC 8 < 89.000.000	Service industries	20,364	108	2.73%
ds8	France	15.000.000 <= SIC 8 < 18.000.000	Construction industries	5,678	19	33.74%
ds9	France	20.000.000 <= SIC 8 < 40.000.000	Manufacturing	3,266	19	21.68%
ds10	France	40.000.000 <= SIC 8 < 50.000.000	Transportation, communications and utilities	1,787	19	16.96%
ds11	France	50.000.000 <= SIC 8 < 52.000.000	Wholesale trade	3,337	19	17.44%
ds12	France	52.000.000 <= SIC 8 < 60.000.000	Retail trade	6,450	19	23.55%
ds13	France	60.000.000 <= SIC 8 < 68.000.000	Finance, insurance and real estate	2,874	19	6.51%
ds14	France	70.000.000 <= SIC 8 < 89.000.000	Service industries	8,576	19	15.24%
ds15	Italy	15.000.000 <= SIC 8 < 18.000.000	Construction industries	3,801	19	14.29%
ds16	Italy	20.000.000 <= SIC 8 < 40.000.000	Manufacturing	5,093	19	12.84%
ds17	Italy	40.000.000 <= SIC 8 < 50.000.000	Transportation, communications and utilities	1,837	19	10.02%
ds18	Italy	50.000.000 <= SIC 8 < 52.000.000	Wholesale trade	3,671	19	12.45%
ds19	Italy	52.000.000 <= SIC 8 < 60.000.000	Retail trade	3,309	19	9.34%
ds20	Italy	60.000.000 <= SIC 8 < 68.000.000	Finance, insurance and real estate	3,732	19	4.02%
ds21	Italy	70.000.000 <= SIC 8 < 89.000.000	Service industries	6,579	19	5.46%

Table 1: Data set characteristics

Note that companies are assigned to industry categories based upon their 8-digit Standard Industry Code (SIC). As such, this study follows the recommendation made by several authors to train models for predicting business failure using single-industry samples (Brigham & Gapenski, 1994; Dimitras, Zanakis, & Zopounidis, 1996; McGurr & DeVaney, 1998). Numerous studies have focused on sector-specific BFP (e.g. Doumpos, Andriosopoulos, Galariotis, Makridou, & Zopounidis, 2017; Lanine & Vennet, 2006) whereas the inclusion of multiple data sets from several countries enhances the generalizability of results.

The data sets consist of companies with an obligation to publish consolidated annual accounts and contains information that describes their history. This information is used to model the dependent variable, a binary business failure indicator (1=business failure; 0= survival) that was measured over an observation time horizon of 12 months as shown in Figure 3. Note that the timeline deviates for the data sets describing Belgian companies versus the ones describing French and Italian companies.

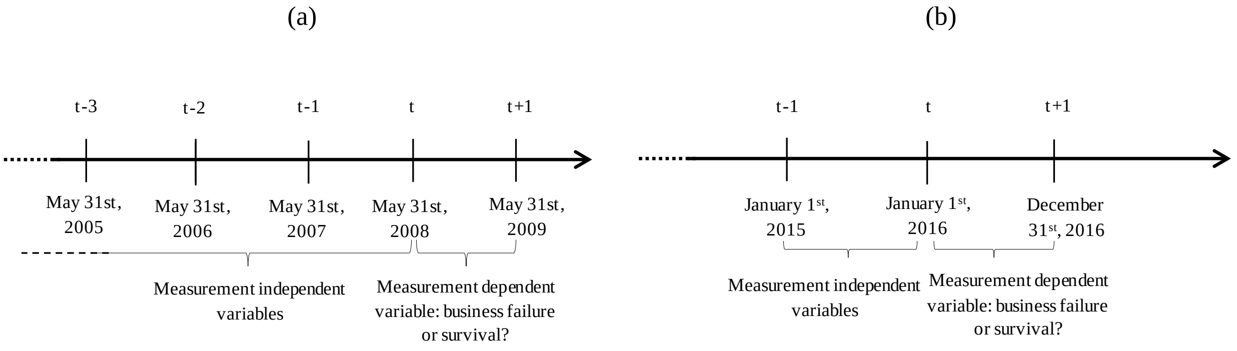

(a)
(b)
t-3
t-2
t-1
t
t+1
May 31st, 2005
May 31st, 2006
May 31st, 2007
May 31st, 2008
May 31st, 2009
Measurement independent variables
Measurement dependent variable: business failure or survival?
t-1
t
t+1
January 1st, 2015
January 1st, 2016
December 31st, 2016
Measurement independent variables
Measurement dependent variable: business failure or survival?

Figure 3: Data collection time lines. Figure (a) applies to data for Belgian companies (ds1 – ds7) while figure (b) applies to data for French and Italian companies (ds8-ds21)

To predict business failure, several independent variables were collected and created. For the Belgian data sets, these variables can be categorized into three categories: (i) financial ratios, (ii) payment promptness indicators and (iii) firmographics. For the French and Italian data sets, variables are limited to financial ratios. Financial ratios and variables related to cash flow have since long been the most important category of predictors used in business failure prediction (McGurr & DeVaney, 1998). The ratios considered in this study can be classified further into liquidity ratios (1a), long-term solvency ratios (1b), asset management ratios (1c) and profitability ratios (1d), analogous to (Ross, Westerfield, Jordan, & Roberts, 2002). For the Belgian datasets (ds1-ds7), two additional variable categories have been included: promptness of payment behavior, i.e. how well and timely a company pays its amounts due to the tax authority, social security authority and selected suppliers, and firmographics; a category that groups a number of features describing the company (e.g. company age, industry category, legal form and number of employees) and the company directors.

Tables 2 and 3 provide an overview of all features included in the data sets.

Variable category	Variable name	Description
1. Financial ratios		
1a. Liquidity ratios	*Cash ratio t-i*	Cash ratio: cash and cash equivalent assets / total liabilities, at time *t-i*
	Current ratio t-i	Current ratio: current assets / current liabilities, at time *t-i*
	NWC2TA ratio t-i	Net working capital to total assets ratio: (current assets - current liabilities) / total assets, at time *t-i*
	Quick ratio t-i	Quick ratio: (current assets - inventories) / current liabilities, at time *t-i*
1b. Long-term solvency ratios	*Debt ratio t-i*	Debt ratio: total liabilities / total assets, at time *t-i*
	Debt2worth ratio t-i	Debt to net worth ratio: total debt / (total assets - total liabilities), at time *t-i*
	Solvency ratio t-i	Solvency ratio: net profit after taxes / total liabilities, at time *t-i*
	Times interest earned ratio t-i	Times interest earned ratio: EBITDA / total financial charges, at time *t-i*
	Avg. collection period ratio t-i	Average collection period ratio: (average accounts receivable / sales revenue) * 365, at time *t-i*
1c. Turnover ratios	*Debtor turnover ratio t-i*	Debtor turnover ratio: net credit sales / average accounts receivable, at time *t-i*
	Fixed-asset turnover t-i	Fixed-asset turnover: sales / average net fixed assets, at time *t-i*
	Inventory turnover t-i	Inventory turnover: cost of goods sold / average inventory, at time *t-i*
	Asset turnover t-i	Asset turnover: net sales revenue / average total assets, at time *t-i*
1d. Profitability ratios	*Gross profit margin t-i*	Gross profit margin: profit before tax / revenue, at time *t-i*
	Profit margin t-i	Profit margin: profit after tax / revenue, at time *t-i*
	ROA t-i	Return on assets (ROA): net income before tax / total assets, at time *t-i*
	ROE t-i	Return on equity (ROE): net income after tax / equity, at time *t-i*
	ROI t-i	Return on investment (ROI): net income after interest and tax / total assets, at time *t-i*
2. Payment promptness indicators	*Social security dues t-i*	Amounts due to social security authority, at time *t-i*
	Tax dues t-i	Amounts due to tax authority, at time *t-i*
	Nbr. protested bills [t-j;t]	Number of protested bills in period *[t-j;t]*
	Nbr. summons [t-j;t]	Number of social security summons in period *[t-j;t]*
	Overdue balance [t-j;t]	Total current overdue balance in period *[t-j;t]*
	Pct. late payments [t-j;t]	Percentage reported transactions with late payment in period *[t-j;t]*
	Pct. late payments cat. k [t-j;t]	Percentage of reported transactions with late payment in payment delay category *k* in period *[t-j;t]*
3. Firmographics	*Avg. director age*	Average age of the directors and owners
	Domestic purchases only	Dummy indicator for exclusive domestic purchases
	Domestic sales only	Dummy indicator for exclusive domestic sales
	Move recency	Days since last change of business address
	Nbr. directors	Number of directors and/or owners
	Nbr. new directors	Number of directors appointed during last 12 months
	Nbr. resigned directors	Number of directors who resigned during last 12 months
	Nbr. directors with stock	Number of directors and/or owners holding stock
	Nbr. employees	Number of employees
	Nbr. directors (fail hist.)	Number of directors previously employed in a company that failed
	Nbr. directors (oob hist.)	Number of directors previously employed in a company that went out of business
	Years in business	Company age (total number of years of business activity)
	Legal form code	Legal form code

Table 2: Variable descriptions for datasets ds1 until ds7: Belgian companies. Year count indices i∈{0,1,2} and j∈{1,2} are used to indicate at which moment in time, or for which time interval, certain variables are calculated. Additionally, payment delay categories k;k∈{1,2,3,4,5,6} in the variable *Pct. late payments cat. k [t-j;t]* are coded as 1=up to 30 days ; 2=from 31 to 60 days ; 3=from 61 to 90 days; 4= from 91 to 120 days; 5= from 121 to 180 days and 6=more than 180 days.

Variable category	Variable name	Description
Financial ratios		
1a. Liquidity ratios	*Cash ratio*	Cash ratio: cash and cash equivalent assets / total liabilities
	Current ratio	Current ratio: current assets / current liabilities
	NWC2TA ratio	Net working capital to total assets ratio: (current assets - current liabilities) / total assets
	Quick ratio	Quick ratio: (current assets - inventories) / current liabilities
1b. Long-term solvency ratios	*Debt ratio*	Debt ratio: total liabilities / total assets
	Debt2worth ratio	Debt to net worth ratio: total debt / (total assets - total liabilities)
	Solvency ratio	Solvency ratio: net profit after taxes / total liabilities
	Times interest earned ratio	Times interest earned ratio: EBITDA / total financial charges
	Avg. collection period ratio	Average collection period ratio: (average accounts receivable / sales revenue) * 365
1c. Turnover ratios	*Debtor turnover ratio*	Debtor turnover ratio: net credit sales / average accounts receivable
	Fixed-asset turnover	Fixed-asset turnover: sales / average net fixed assets
	Inventory turnover	Inventory turnover: cost of goods sold / average inventory
	Asset turnover	Asset turnover: net sales revenue / average total assets
1d. Profitability ratios	*Gross profit margin*	Gross profit margin: profit before tax / revenue
	Profit margin	Profit margin: profit after tax / revenue
	ROA	Return on assets (ROA): net income before tax / total assets
	ROE	Return on equity (ROE): net income after tax / equity
	ROI	Return on investment (ROI): net income after interest and tax / total assets

Table 3: Variable descriptions for datasets ds8 to ds21: French and Italian companies.

A seen in Table 2, for the Belgian sets, most variables are available in a number of variations to take into account their evolution over time. In Table 2, time-varying variables are distinguished by a year count indicator which represents the year in which they are calculated, using the most recent information available at that time. Time index *t* denotes the end of the independent variable collection period, i.e. May 31st 2008. Consequently, for example, the variable *ROI t-1* provides the return on investment calculated using the most recent information available on May 31st 2007, i.e. using annual account information for the year 2006. A set of variables that belong to the payment promptness category are measured over time intervals, dating either one or two years back prior to time *t*. For example, the variable *Nbr. summons [t-2;t]* counts the number of social security summons during a two-year period until May 31st, 2008.

The data sets underwent a number of preprocessing steps. First, outlier detection and treatment was applied, consistent with previous literature (Bou-Hamad, Larocque, & Ben-Ameur, 2011; Chava & Jarrow, 2004). In particular, *winsorization* was applied: variables' values are truncated below the 2.5th and above the 97.5th percentile. Note that this winsorization is an implicit element in the spline-rule ensemble algorithm (see Section 3) and was separately applied to the training data for benchmark algorithms. Second, feature selection is another important data preprocessing step, and considered good practice in the domain of bankruptcy prediction (Abellán & Castellano, 2017; Tsai, 2009). Therefore, correlation-based feature selection (Hall, 2000) is applied, a basic filter feature selection approach that has seen prior applications in BFP literature (e.g. Tsai, 2009). As explained in the next Section, some methods that are known to be especially sensitive to the inclusion of uninformative or correlating features are implemented with an additional, wrapper-based feature selection; Finally, undersampling, a strategy known to reduce the negative impact that class imbalance exerts on many predictive methods (Weiss, 2004) and nowadays common practice in business failure prediction (Kotsiantis, Tzelepis, Koumanakos, & Tampakas, 2007) is applied.

4.2 Experimental Settings

Experimental results are all based on a ten-fold cross-validation that is repeated 10 times, in line with other studies on BFP. In ten-fold cross-validation, a data set is divided in 10 parts of equal size, while stratified random sampling is applied in order to maintain the original class distributions. Each data part serves as test set once, while the remaining data parts are stacked to form a training set. This results in 10 measurements of model performance. Note that undersampling of the training data sets is applied after the division of the data for the cross-validation, and that winsorization and feature selection are also repeated for each fold, whereby truncation percentiles and feature subsets are determined on the training dataset and applied on the corresponding test set.

To assess the predictive performance of rule ensembles, the method is compared to three groups of benchmark algorithms. A first set includes uncombined (i.e., non-ensemble) techniques with a proven track record in BFP and leading to highly interpretable models, i.e. logistic regression, linear discriminant analysis (LDA), quadratic discriminant analysis (QDA) and a C4.5 decision tree. A second category includes uncombined algorithms that have been popular choices in BFP literature but that lead to models that are difficult to interpret. These methods are neural networks (specifically, multi-layer perceptrons (MLP)), support vector machines (SVMs) and k-nearest neighbors (kNN). A third set includes the ensemble algorithms bagging, AdaBoost and random forests which have also been applied to BFP before (see Section 2) and are known to result in complex models with low comprehensibility. Algorithms are implemented in R (spline-rule ensemble, rule ensemble, random forest, bagging, AdaBoost), SAS (logistic regression, kNN, MLP, LDA and QDA), WEKA (C4.5) and Python/LIBSVM (SVM). Penalized cubic regression spline estimation depends on the *mgcv* R package (Wood, 2004) and regularized regression (ridge, lasso and elastic net) was implemented using the *glmnet* R package (Friedman, Hastie, & Tibshirani, 2010). Logistic regression, LDA and QDA are implemented with forward wrapper feature selection, while the C4.5 decision trees are pruned to reduce model overfitting. Given that these steps are commonly used in tandem with these algorithms, this makes for a fairer and more challenging comparison. SVM is implemented using a linear kernel function and its regularization parameter, the *soft margin constant*, is determined through grid search; MLP is implemented with one hidden layer and in kNN, *k* is set to 5. Note that for the latter two parameters, multiple values were tested and the ones leading to the best performance over all metrics were retained. All ensemble algorithms contain 100 members. For rule ensembles and spline-rule ensembles, this translates to initially training 100 trees from which rules are then derived and selected. An important additional parameter defining rule complexity in rule and spline-rule ensembles is average tree depth (the number of terminal nodes). This parameter was set to 9 in this study, allowing for the discovery of higher-order interactions. Penalized cubic regression spline estimation depends on specification of the knots for each variable (parameter *u*, and values $\xi_1, \xi_2, \dots, \xi_u$ as defined in Section 3). *u* is set to 10 and the knots' values are automatically determined in order to ensure equally-sized intervals. Finally, unless specified differently, spline-rule and rule ensembles are configured with regularized lasso regression.

It is worth noting that this study and its experimental setup overcomes many of the problems associated with benchmark studies in BFP (Balcaen & Ooghe, 2006). First, while undersampling is applied to the training data in order to increase models' performance, models are evaluated on representative, unbalanced test data sets. This contrasts with the sample biases introduced by the oversampling of failed companies and the arbitrary matched pair sampling of failing and non-failing companies. Second, predictive information is not limited to financial account information and financial ratios only. Third, variable selection is applied as a wrapper and only to those techniques sensitive to the inclusion of uninformative or correlating variables.

4.3 Evaluation Criteria

As business failure is modeled as a problem of binary classification, failing and non-failing companies can be classified correctly or incorrectly, which leads to a 2-dimensional confusion matrix as shown by Table 4.

	Predicted class	
Real class	*Business survival*	*Busines failure*
Business survival	*tn (true negative)*	*fp (false positive)*
Business failure	*fn (false negative)*	*tp (true positive)*

Table 4: Confusion matrix

Accuracy, or the percentage of correctly classified instances, is a conventional evaluation criterion in BFP studies (Balcaen & Ooghe, 2006). Using the notation from Table 3; it is calculated as $(tn + tp)/(tp + fp + tn + fn)$.

While accuracy is a straightforward and intuitive measure and the most widely used metric to evaluate the predictive performance of business failure prediction models, it has been criticized for a number of reasons. First, it is unreliable in a situation of class imbalance (Weiss, 2004). As an example, consider a naïve decision rule assigning the majority class to all test set instances, which would exhibit an inflated accuracy above 50% despite the absence of any discriminative power. Second, data resampling applied to reduce class imbalance also harms the reliability of accuracy estimations, albeit in a different way, as the model is forced to focus on failing companies at the expense of non-failing companies. On a balanced test sample, this would lead to an overly pessimistic accuracy level. Third, accuracy does not take into account predicted class membership probabilities but instead requires setting a cut-off to convert posterior probabilities or predicted scores to class. Accuracy can vary severely depending on the choice of this cut-off (Balcaen & Ooghe, 2006).

An alternative performance metric that circumvents these drawbacks is the *Area Under the Receiver Operating Characteristics curve* (AUC or AUROC). Several authors (e.g. Langley, 2000; Provost, Fawcett, & Kohavi, 2000) advocate AUC as an objective performance criterion, well-suited for the comparison of classifier performance. Unlike accuracy, it evaluates the ability of a classifier to distinguish between the two classes based on the predicted class membership probabilities, and is therefore suitable for imbalanced classification problems such as business failure prediction. While not as commonly used as accuracy in BFP studies, AUC starts to emerge as a viable alternative (e.g. Bou-Hamad, et al., 2011; Nanni & Lumini, 2009).

An expression for AUC can be derived from the confusion matrix. Using the definitions of the true positive rate; $tpr = \frac{tp}{tp+fn}$ and false positive rate; $fpr = \frac{fp}{fp+tn}$, and trivially parameterizing these expressions to acknowledge their dependence upon the choice of cutoff *t*, required to translate real-valued predictions to class predictions, the AUC can be expressed as

$$AUC = \int tpr(t)\big(-fpr'(t)\big)dt \qquad (8)$$

Finally, as observed by Balcaen and Ooghe (2006), misclassification costs associated with type I and type II errors are not equal in BFP. For example, for a financial institution, the inability of a model to timely predict the bankruptcy of a lending company could entail severe financial losses, while the cost associated with wrongfully flagging a company as potential risk would typically be limited (e.g. to the cost of an in-depth screening, or the loss of the contribution if the contract is cancelled). The evaluation and benchmarking of classifiers should take this into account. Similar to Chen and Ribeiro (2013), we therefore evaluate in terms of *expected misclassification cost* using multiple cost ratios. Expected misclassification cost (EMC) is given by the following formula:

$$EMC = \frac{C_{fn} * fn + C_{fp} * fp}{n} \qquad (9)$$

Where n is the total number of observations, C_{fn} denotes the cost associated with falsely predicting survival for a failing company, and C_{fp} the cost associated with falsely predicting business failure for a healthy company. The expected misclassification cost measure of a BFP model can be interpreted as the average cost that will be incurred from using the model to score one company. To facilitate the analyses, C_{fp} is assumed to be 1 and $C_{fn} > C_{fp}$. Three cost ratios ((C_{fn}/C_{fp}) are considered in this study: 2, 5 and 10.

Following Demšar (2006), in order to statistically compare algorithm performance over multiple data sets, Wilcoxon signed rank tests (Wilcoxon, 1945) are computed for comparisons involving two algorithms, while Friedman non-parametric ANOVA's (Friedman, 1937) are considered for comparisons involving more than two algorithms. Both tests are based upon the average rank of the performance measures of the algorithms considered, taken over all data sets. For the Friedman test, post-hoc tests are administered using the test statistic for comparing methods *i* and *j* are obtained as

$$z = \frac{(R_i - R_j)}{\sqrt{\frac{k(k+1)}{6N}}} \qquad (10)$$

where R_i is the average rank of method I, k is the number of algorithms and N the number of datasets. The probabilities associated with these statistics, obtained from the standard-normal distribution are compared to corrected values of α in order to account for family-wise error, introduced through making multiple algorithm comparisons. In this study, Hommel's procedure (Hommel, 1988) is used to this end.

5 Results

To assess the potential of rule-spline ensembles for business failure prediction, this Section focuses on model performance benchmarking and, subsequently, on detailing and illustrating the interpretability instruments inherent to rule ensembles through a case study on one chosen data set.

5.1 Predictive Performance

The assessment of the predictive performance of spline-rule ensembles for business failure prediction consists of three parts. First, as spline-rule ensembles are an extension of the rule ensemble framework as proposed by Friedman & Popescu (2008), the performance of both algorithms is compared. Second, the sensitivity of spline-rule ensembles to the choice of the regularization method used is investigated by comparing three alternative regularization methods: ridge regression, lasso regularized regression and elastic net regularization. Third, spline-rule ensembles are compared to established methods in BFP literature. Please note that this section reports aggregated performance indicators and results of statistical tests. The full cross-validated results (average cross-validation performance values and standard errors per data set, per algorithm) are available from the author upon request.

A first question involves whether the addition of smoothing spline terms in spline-rule ensembles enhances performance in business failure prediction over standard rule ensembles. Table 5 summarizes the statistical comparison of spline-rule ensembles versus rule ensembles. It reports the results of Wilcoxon signed rank tests for every metric considered, based upon cross-validated results in terms of accuracy (ACC), AUC and misclassification rate for cost ratios 2, 5 and 10 (denoted EMC_2, EMC_5 and EMC_{10}, respectively), over all data sets. The table also reports wins, losses and ties counts which summarizes pairwise comparisons, over the 21 datasets, of both algorithms in terms of their average cross-validated performance.

Metric	Evaluation method		*SRE vs. RE*
ACC	*Wilcoxon signed-ranks test*	*T-statistic*	103 (df =20)
		Significance	n.s. (p=0.332)
	Wins / losses / ties		13/8/0
AUC	*Wilcoxon signed-ranks test*	*T-statistic*	19 (df=20)
		Significance	*** (p=0.0004)
	Wins / losses / ties		18/3/0
EMC_2	*Wilcoxon signed-ranks test*	*T-statistic*	0 (df=20)
		Significance	*** (p<0.0001)
	Wins / losses / ties		21/0/0
EMC_5	*Wilcoxon signed-ranks test*	*T-statistic*	0 (df=20)
		Significance	*** (p<0.0001)
	Wins / losses / ties		21/0/0
EMC_{10}	*Wilcoxon signed-ranks test*	*T-statistic*	0 (df=20)
		Significance	*** (p<0.0001)
	Wins / losses / ties		21/0/0

Table 5: Statistical comparison of spline-rule ensembles (SRE) versus conventional rule ensembles (RE). ACC=accuracy, AUC=area under the ROC curve and EMC_θ = expected misclassification cost based upon cost ratio θ. n.s. = not significant; * indicates a significant difference at the 99% confidence level (α=0.01)**

These results clearly show the dominance of spline-rule ensembles over conventional rule ensembles for business failure prediction for all metrics considered, but accuracy. As AUC and expected misclassification cost metrics are considered more appropriate model evaluation metrics in the context of BFP where different types of errors are associated with different costs, it is concluded that extending the conventional rule ensemble framework to include smooth terms as a third category of candidate terms next to rules and winsorized linear terms.

A second experiment involves in investigation into the sensitivity of spline-rule ensembles to the adoption of alternative regularization schemes: ridge regression, lasso regression and regularized regression through the elastic net. Table 6 shows results of global Friedman tests per metric, as well as average algorithm ranks on which the Friedman test is based.

Metric	**Evaluation method**			***SRE (Lasso)***	***SRE (Elastic Net)***	***SRE (Ridge)***
ACC	*Friedman test*	*Chi-quare statistic*	*2.5714 (df=2)*			
		Significance	*n.s. (p=0.2765)*			
	Average ranks			2.1429	1.7143	2.1429
AUC	*Friedman test*	*Chi-quare statistic*	1.2381 (df=2)			
		Significance	n.s. (p=0.5385)			
	Average ranks			2.1905	1.8571	1.9524
EMC_2	*Friedman test*	*Chi-quare statistic*	0.09524 (df=2)			
		Significance	n.s. (p=0.9535)			
	Average ranks			1.9524	2	2.0476
EMC_5	*Friedman test*	*Chi-quare statistic*	0.09524 (df=2)			
		Significance	n.s. (p=0.9535)			
	Average ranks			2.0476	1.9524	2
EMC_{10}	*Friedman test*	*Chi-quaret statistic*	0.09524 (df=2)			
		Significance	n.s. (p=0.9535)			
	Average ranks			2	1.9524	2.0476

Table 6: Statistical comparison of alternative regularization schemes for spline-rule ensembles (SRE). ACC=accuracy, AUC=area under the ROC curve and EMC_θ = expected misclassification cost based upon cost ratio θ.

Variation comparisons demonstrate that no overall difference between regularization schemes could be identified in this study’s setting. This holds for all evaluation metrics considered. However, it is crucial to understand that in the case of ridge regression parameter shrinkage does not involve term selection, which substantially compromises model subsequent model interpretation. Moreover, as expected, regularization through the elastic net resulted in significantly larger models than lasso regression. An optimization exercise in function of both model performance and interpretability suggests a preference for lasso regularization.

A third comparison involves a benchmark of spline-rule ensembles to alternative, established methods in the domain of business failure prediction. Table 7 shows the results for a comparison to 5 established benchmark algorithms: multi-layer perceptron, support vector machines, logistic regression, linear – and quadratic discriminant analysis

Metric	Evaluation method			*SRE*	*Multi-Layer Perceptron*	*Support Vector Machines*	*Logistic Regression*	*Linear Discriminant Analysis*	*Quadratic Discriminant Analysis*
ACC	*Friedman test*	*Chi-square statistic*	16.075 (df=5)						
		Significance	*** (p=0.0066)						
	Average ranks			3.3333	4.6190	3.8571	3.3810	**2.3810**	3.4286
	Post-hoc sign. (SRE vs. benchmark)			-	**n.s. (adj. p=0.2723)**	**n.s. (adj. p=0.9343)**	**n.s. (adj. p=0.9343)**	**n.s. (adj. p=0.7326)**	**n.s. (adj. p=0.9343)**
	Wins / losses / ties (SRE vs. benchmark)			-	15/6/0	13/8/0	7/14/0	6/15/0	12/9/0
AUC	*Friedman test*	*Chi-square statistic*	65.109 (df=5)						
		Significance	*** (p<0.0001)						
	Average ranks			**1**	4	4.9524	2.9048	3.2871	4.8571
	Post-hoc sign. (SRE vs. benchmark)			-	***** (adj. p<0.0001)**	***** (adj. p<0.0001)**	***** (adj. p=0.0087)**	***** (adj. p=0.0009)**	***** (adj. p<0.0001)**
	Wins / losses / ties (SRE vs. benchmark)			-	21/0/0	21/0/0	21/0/0	21/0/0	21/0/0
EMC$_2$	*Friedman test*	*Chi-square statistic*	59.721 (df=5)						
		Significance	*** (p<0.0001)						
	Average ranks			**1.1905**	3.5238	4.9048	3.0476	3.2857	5.0476
	Post-hoc sign. (SRE vs. benchmark)			-	***** (adj. p=0.0007)**	***** (adj. p<0.0001)**	**** (adj. p=0.0104)**	***** (adj. p=0.0032)**	***** (adj. p<0.0001)**
	Wins / losses / ties (SRE vs. benchmark)			-	21/0/0	20/1/0	20/1/0	20/1/0	21/0/0
EMC$_5$	*Friedman test*	*Chi-square statistic*	62.061 (df=7)						
		Significance	*** (p<0.0001)						
	Average ranks			**1.0952**	3.5714	4.9523	3.1429	3.2381	5
	Post-hoc sign. (SRE vs. benchmark)			-	***** (adj. p=0.0002)**	***** (adj. p<0.0001)**	***** (adj. p=0.0043)**	***** (adj. p=0.0023)**	***** (adj. p<0.0001)**
	Wins / losses / ties (SRE vs. benchmark)			-	21/0/0	21/0/0	20/1/0	20/1/0	21/0/0
EMC$_{10}$	*Friedman test*	*Chi-square statistic*	63.122 (df=5)						
		Significance	*** (p=0.0000)						
	Average ranks			**1**	3.8095	5	3.0952	3.2857	4.8095
	Post-hoc sign. (SRE vs. benchmark)				***** (adj. p<0.0001)**	***** (adj. p<0.0001)**	***** (adj. p=0.0031)**	***** (adj. p=0.0009)**	***** (adj. p<0.0001)**
	Wins / losses / ties (SRE vs. benchmark)			-	19/2/0	21/0/0	21/0/0	21/0/0	21/0/0

Table 7: Predictive performance benchmarking: SRE versus non-ensemble (uncombined) classifiers. ACC=accuracy, AUC=area under the ROC curve and EMC$_\theta$ = expected misclassification cost based upon cost ratio *θ*. For each metric, the lowest (i.e., most favorable) average rank over all data sets is indicated in bold face type. n.s. = not significant; ** and * indicate significant differences at the 95% and 99% significance levels, respectively.**

A number of observations are made from Table 7. First, Friedman tests indicate that for all metrics, significant differences emerge between algorithms when average performance ranks are compared. Second, these results clearly demonstrate the dominance of spline-rule ensembles over uncombined algorithms. For all metrics except accuracy, spline-rule ensembles significantly outperform all benchmark algorithms. In terms of AUC and the three misclassification cost metrics, this is clearly shown by the post-hoc test results, average ranks, and further illustrated by the wins/losses/ties counts. In terms of accuracy, LDA demonstrates a lower (hence, more favorable) average rank than spline-rule ensembles, but the post-hoc tests indicate that the difference is not significant. The best performing benchmark algorithm in terms of average ranks for AUC and EMC is logistic regression. This is somewhat unexpected, as some authors have criticized logistic regression for its low predictive performance (Li & Sun, 2011b).

Metric	Evaluation method			*SRE*	*AdaBoost*	*Bagging*	*Random Forest*
ACC	*Friedman test*	*Chi-square statistic*	14.714 (df=3)				
		Significance	*** (p=0.0021)				
	Average ranks			**1.5714**	2.9048	2.8095	2.7143
	Post-hoc sign. (SRE vs. benchmark)			-	*** (adj. p=0.0008)	*** (adj. p=0.0094)	** (adj. p=0.0165)
	Wins / losses / ties (SRE vs. benchmark)			-	17/4/0	18/3/0	16/5/0
AUC	*Friedman test*	*Chi-square statistic*	57.057 (df=3)				
		Significance	*** (p<0.0000)				
	Average ranks			**1**	4	2.381	2.6191
	Post-hoc sign. (SRE vs. benchmark)			-	*** (adj. p<0.0000)	*** (adj. p=0.0011)	*** (adj. p=0.0002)
	Wins / losses / ties (SRE vs. benchmark)			-	21/0/0	21/0/0	21/0/0
EMC$_2$	*Friedman test*	*Chi-square statistic*	27 (df=3)				
		Significance	*** (p<0.0000)				
	Average ranks			**1.7143**	3.6667	2.52381	2.0952
	Post-hoc sign. (SRE vs. benchmark)			-	*** (adj. p<0.0000)	n.s. (adj. p=0.1295)	n.s. (p=0.3390)
	Wins / losses / ties (SRE vs. benchmark)			-	19/2/0	17/4/0	12/9/0
EMC$_5$	*Friedman test*	*Chi-square statistic*	44.371 (df=3)				
		Significance	*** (p<0.0000)				
	Average ranks			**1.1905**	3.8095	2.7143	2.2857
	Post-hoc sign. (SRE vs. benchmark)			-	*** (adj. p=0.0000)	*** (adj. p=0.005)	** (adj. p=0.0119)
	Wins / losses / ties (SRE vs. benchmark)			-	21/0/0	21/0/0	17/4/0
EMC$_{10}$	*Friedman test*	*Chi-square statistic*	47 (df=3)				
		Significance	*** (p<0.0000)				
	Average ranks			**1.0476**	3.7619	2.5238	2.6667
	Post-hoc sign. (SRE vs. benchmark)			-	*** (adj. p<0.0000)	*** (adj. p=0.0008)	*** (adj. p=0.0002)
	Wins / losses / ties (SRE vs. benchmark)			-	21/0/0	21/0/0	20/1/0

Table 8: Predictive performance benchmarking: SRE versus ensemble (uncombined) classifiers. ACC=accuracy, AUC=area under the ROC curve and EMC$_\theta$ = expected misclassification cost based upon cost ratio *θ*. For each metric, the lowest (i.e., most favorable) average rank over all data sets is indicated in bold face type. n.s. = not significant; ** and * indicate significant differences at the 95% and 99% significance levels, respectively.**

Finally, Table 8 reports results from a comparison of SRE to three ensemble algorithms: AdaBoost, Bagging, and Random Forests. The Friedman test results again reveal significant differences between algorithms for all evaluation criteria. Spline-rule ensembles are dominant across all metrics in terms of average ranks, in a comparison to AdaBoost, Bagging and Random Forest. Statistically, this superiority is confirmed by post-hoc tests, except in the case of misclassification cost with a cost ratio of 2 where the spline-rule ensembles do not significantly outperform Bagging and Random Forest.

In summary, these results demonstrate the highly competitive performance of spline-rule ensembles over conventional rule ensembles, and a large set of benchmark algorithms.

5.2 Model interpretation case study: business failure prediction in the Belgian services sector

In this Section, the comprehensibility of spline-rule ensembles is demonstrated by means of a case study focusing on business failure prediction in the service sector in Belgium. To this end, a spline rule ensemble is trained on the corresponding data set listed in Table 1 (*ds7*). This setting was chosen for two reasons: (i) the prevalence of the services sector in Belgium (accounting for about 77% of economic activity in 2016; (European Commission, 2016)), and (ii) the fact that this particular data set is the largest, in terms of number of companies and number of company characteristics.

The case study illustrates how the spline-rule ensemble model offers a high degree of comprehensibility through the model itself, and by demonstrating how additional insights can be delivered through calculation of variable importance scores and deriving partial dependence functions.

5.2.1 Model visualization and interpretation

Spline-rule ensembles, analogous to rule ensembles, derive their interpretability from three model characteristics: (i) the simplicity of their candidate member classifiers (i.e., rules, splines and linear terms), (ii) their simple linear combination and (iii) the shrinkage resulting from the selection procedure to which they are submitted. Consequently, the most obvious source of insights into the model's functioning is the model itself, i.e. the rules and terms that are selected by the regularized linear regression. Moreover, several measures reflect the relative influence and importance of the terms in the rule ensemble model. The first category are the *term coefficients*, i.e. the parameters $\hat{a}_j;\ j = 0, \dots, q$ and $\hat{b}_k;\ k = 1, \dots, p$ of the linear regularized regression that have received non-zero values. For a rule, a term coefficient reflects the relative influence a term has upon the logit transformation of the probability to fail if its conditions are met. Second, *rule support* refers to the percentage of training observations for which a rule holds, i.e., for which all rule conditions are true. Finally, for any rule j, the *rule importance* RI_j is then calculated as

$$RI_j = |\hat{\mathrm{a}}_j|.\sqrt{s_j(1 - s_j)} \qquad (11)$$

where s_j represents the rule support. For a linear predictor $\mathrm{x_k}$, a *linear term importance* $\mathrm{LI_k(x_k)}$ is obtained as

$$LI_k(x_k) = |\hat{b}_k|.std(l_k(x_k)) \qquad (12)$$

wherein $std(l_k(x_k))$ is the standard deviation of $l_k(x_k)$ and similarly, for cubic regression splines, a spline term importance $\mathrm{SI_k(x_k)}$ is calculated through

$$SI_k(x_k) = |\hat{c}_k|.std(s_k(x_k)) \qquad (13)$$

wherein $std(s_k(x_k))$ is the standard deviation of $s_k(x_k)$. Both importance measures are comparable, as they correspond to an absolute value of the coefficient of the respective standardized term (Friedman & Popescu, 2008).

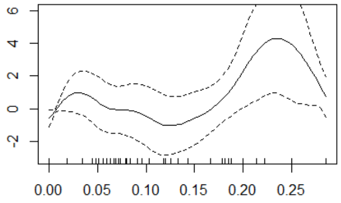
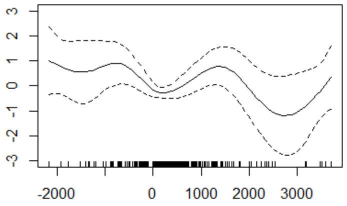
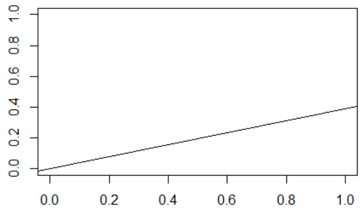
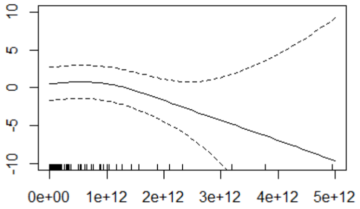
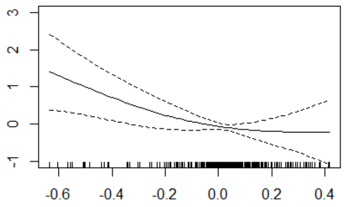
Term	Type	Term/spline visualization or rule conditions	Coefficient	Rule support	Term importance
1	Rule	*ROA t >= -24.67 %* *Nbr. summons [t-1;t] < 1*	-0.9532	0.804	100
2	Spline	*s(Pct. late payments cat. 3 [t-1;t])* 	0.3197	-	65.07
3	Rule	*Pct. late payments [t-2;t] < 65.83 %* *Nbr. summons [t-2;t] < 1*	-0.469	0.6349	59.66
4	Rule	*Years in business >= 1.7329* *Move recency < 699 days*	0.3992	0.3273	49.5
5	Rule	*Nbr. summons [t-2;t] < 1* *Move recency >= 699 days*	-0.3415	0.4928	45.12
6	Rule	*Move recency >= 487.5 days* *Nbr. summons [t-1;t] < 1*	-0.3089	0.5809	40.27
7	Spline	*s(Solvency ratio t)* 	0.3274	0.393	34.00
8	Linear term	*Pct. late payments [t-2;t]* 	0.3876	-	31.56
9	Spline	*s(Cash ratio t-1)* 		-	29.67
10	Spline	*s(ROI t-2)* 	0.2629	-	21.48
11	Rule	*Cash ratio t >= 55%* *Nbr. summons [t-1;t] < 1*	-0.1442	0.4766	19.03

Table 9: The spline-rule ensemble model: terms, term types, rule conditions and visualization of splines and linear terms; coefficients, rule support and term importance

Table 9 shows the selected terms in the current setting, their coefficients, rule supports and importance measures. Note that model terms have been sorted by their importance. From the table, the following observations emerge. First, the spline-rule ensemble showcases the capability of a spline-rule ensembles to account for different types of effects as the model contains 11 terms in total and combines rules (6), splines (4) and one linear term. Nonlinear effects have been identified for a variable on late payment behavior (term 2), solvency ratio (term 7), cash ratio (term 9) and return on investment (term 10), while a linear effect was found for another, more general variable on reported late payments (term 8). Second, all rules contain multiple conditions, indicating the possible presence of interaction effects. Interaction effects are analyzed in detail at a later stage. All rules, except one, decrease failure probability when fulfilled and include a condition on a payment promptness variable. Term 4 is an exception to both observations. Third, in general, closer inspection of splines, rule conditions and the linear term reveals that all effects are intuitive. In line with what can be reasonably expected, conditions based upon financial ratios specify left-discrete value intervals (hence, associate higher values with reduced failure risk) while conditions based upon creditworthiness variables most often specify right-discrete value intervals (i.e., associate timely payment behavior with lower risk).

5.2.2 Variable Importance Measures

Data sets in BFP typically consist of many predictors belonging to different variable categories. The relative importance of variables and variable categories is usually of great importance to financial analysts. Relative variable importances can be easily derived from the rule set by calculating v*ariable importance measures* $VI_k(x_k)$ which attribute higher value to variables appearing (i) more frequently and (ii) in more influential rules than others:

$$VI_k(x_k) = LI_k(x_k) + SI_k(x_k) + \sum_{\substack{j=1 \\ x_k \in r_j}}^{q} RI_j / p_j \qquad (14)$$

where the first part $LI_k(x_k)$ is the importance of the linear term devoted to variable x_k (receiving a value of zero if no such term was selected by the model) and the second term sums the rule importance measures of rules RI_j containing x_k and divides each by p_j, the number of variables featuring in the respective rule j. Similarly, the *variable category importance measure* for variable set x_c is defined as the sum of constituent variable importance measures:

$$CI_c(x_c) = \sum_{x_k \in x_c}\big(VI_k(x_k)\big). \qquad (15)$$

Rank	Variable	Variable category	Description	Variable importance
1	*Nbr. summons [t-1;t]*	Payment promptness	Number of social security summons in period [t-1;t]	100.00
2	*Move recency*	Firmographics	Days since last change of business address	91.12
3	*Nbr. summons [t-2;t]*	Payment promptness	Number of social security summons in period [t-2;t]	72.27
4	*ROA t*	Financial ratios	Return on assets (ROA): net income before tax / total assets	62.77
5	*Pct. late payments [t-2;t]*	Payment promptness	Percentage reported transactions with late payment in period [t-2;t]	37.45
6	*Years in business*	Firmographics	Company age (years)	31.07
7	*Cash ratio t*	Financial ratios	Cash ratio: cash and cash equivalent assets / total liabilities,	11.95
8	*Pct. late payments cat. 3 [t-1;t]*	Payment promptness	Percentage of reported transactions with late payment in payment delay category 3 in period *[t-1;t]*	0.31
9	*Solvency ratio t*	Financial ratios	Solvency ratio: (net profit after taxes) / total liabilities	0.16
10	*Cash ratio t-1*	Financial ratios	Cash ratio: cash and cash equivalent assets / total liabilities,	0.14
11	*ROI t-2*	Financial ratios	Return on investment (ROI): net income after interest and tax / total assets, at time *t-2*	0.10

Table 10: Variable importance measures.

Table 10 presents variable importance measures for all variables in the model. Measures have been rescaled so that the most important variable receives a score of 100. In total, 11 variables appear in the spline-rule ensemble model. Amongst selected variables, 5 financial ratios, 4 variables related to payment promptness and 2 firmographics emerge. The single most influential predictor is the number of social security summons counted over a period of one year. Further, the distribution of variable importance measures is highly skewed. 7 variables exhibit scores above 10 while the remaining 4 predictors demonstrate values below 0.5. Finally, payment promptness variables emerge as the most important variable category with a variable category importance measure of 210.03, followed by firmographics (122.19) and financial ratios (75.02).

5.2.3 Partial Dependence Plots

Partial dependence functions are a generic method for unveiling the nature of the dependence of a predictive model $F(\mathrm{x})$ on a selection of one or more predictor variables (Hastie, Tibshirani, & Friedman, 2001). For a subset of variables x_s they are estimated from the data as

$$\hat{F}_s(x_s) = \frac{1}{n}\sum_{i=1}^{n} F(x_s, x_{i\backslash s}) \qquad (16)$$

where *n* is the number of observations and $\mathrm{x}_{i\backslash s}$ represents the values of all variables *not* occurring in variable set x_s for observation *i*. Hence, partial dependence functions isolate the effect of the variables in subset x_s by taking into account an averaged effect of the other variables.

Figure 4 shows the partial dependence *plots* for the variables in the model, ordered by importance (see Table 10). Note that these plots are based upon partial dependence functions populating x_s with only one variable at a time. As such, they reveal the nature of the relationship between a single variable and the log odds of business failure.

Figure 4 demonstrates that the majority of partial dependence functions are non-linear and in most cases monotonically in- or decreasing. There are two exceptions to this latter observation: the partial dependence functions for the variables *Pct. late payments cat. 3 [t-1;t]* and *Solvency ratio t*. that demonstrate a more complex trend. In summary, the probability of failure increases in the presence of social security summons (*Nbr. Summons [t-1;t]* and *Nbr. Summons [t-2;t]*), with an increasing percentage of late payments (*Pct. late payments [t-2;t]*) and with company age (*Years in business*). On the other hand, the risk decreases as the number of days since the last change of address increases (*Move recency*), for larger values of the return on assets (*ROA t*), cash ratio (*Cash ratio t-1* and *Cash ratio t*) and return on investment (*ROI t-2*). The variable solvency ratio (*Solvency ratio t*) defines higher risk at the extreme ends of its distribution and boasts the only non-monotonic partial dependence function in the selection.

It is interesting to compare these partial dependence functions (and their graphical representations) with the model described earlier. The smooth functions can be easily recognized, and the partial dependence function for the variable *Pct. late payments [t-2;t]* is a composite of the linear term and the rule that feature the variable.

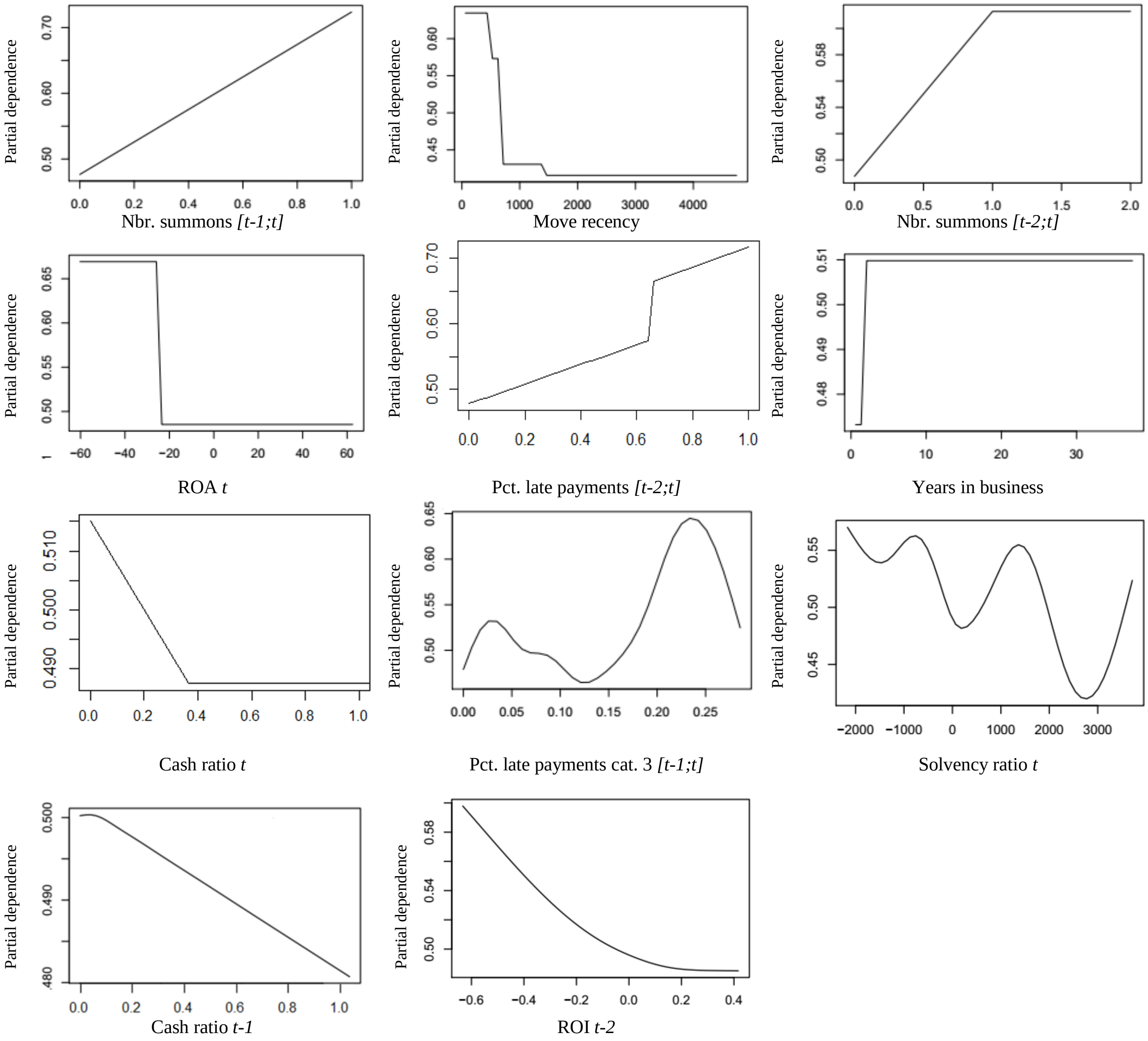


Figure 4: Partial dependence plots for selected variables

5.2.4 Variable Interactions

Partial dependence functions can be utilized further to analyze variable interactions. In particular, it can be insightful to identify variables that are involved in interactions, the specific variables they interact with as well as the degree, strength and functional form of these interaction effects. A measure for the strength of the interaction effect between variables x_j and x_k , H_{jk}^2, can be obtained by

$$H_{jk}^2 = \sum_{i=1}^{n} \left[\hat{F}_{jk}\left(x_{ij}, x_{ik}\right) - \hat{F}_j\left(x_{ij}\right) - \hat{F}_k(x_{ik})\right]^2 / \sum_{i=1}^{n} \hat{F}_{jk}^2(x_{ij}, x_{ik}). \quad (17)$$

Figure 5 visualizes interaction effects and interaction strength values for all interaction effects present in the model.

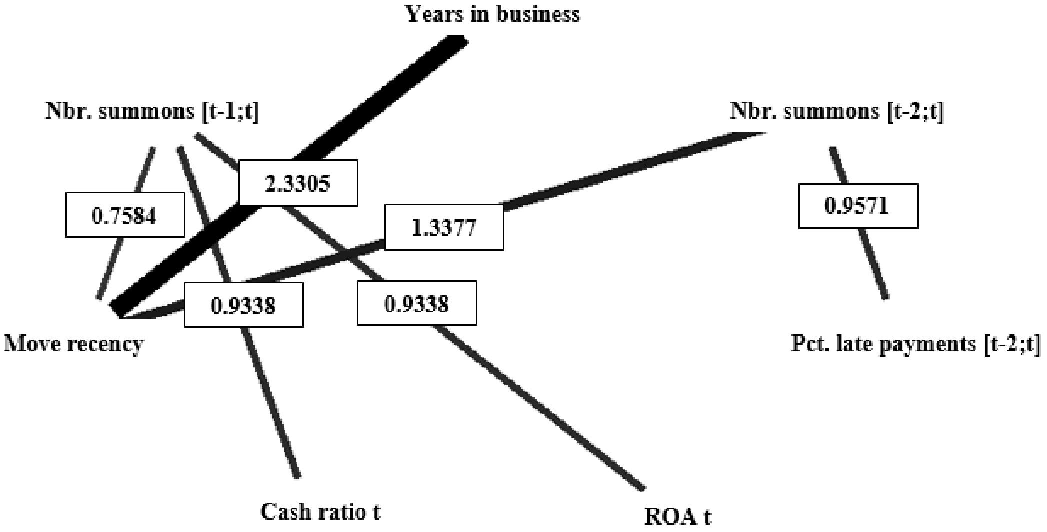


Figure 5: Visualization of interaction effects and interaction strengths

These results indicate that the variables *Nbr. Summons [t-1;t]* and *Move recency* are both involved in 3 interactions while *Nbr. Summons [t-2;t]* is involved in two interactions. Higher-order interactions can be identified using extensions of formula (11), but in the current setting, no such interaction effects were discovered.

A final step in the analysis of interaction effects involves the investigation of the nature of interaction effects. This is easily achieved by plotting partial dependence functions $\hat{F}_{jk}\left(x_{ij}, x_{ik}\right)$ for couples of variables that have been found to significantly interact.

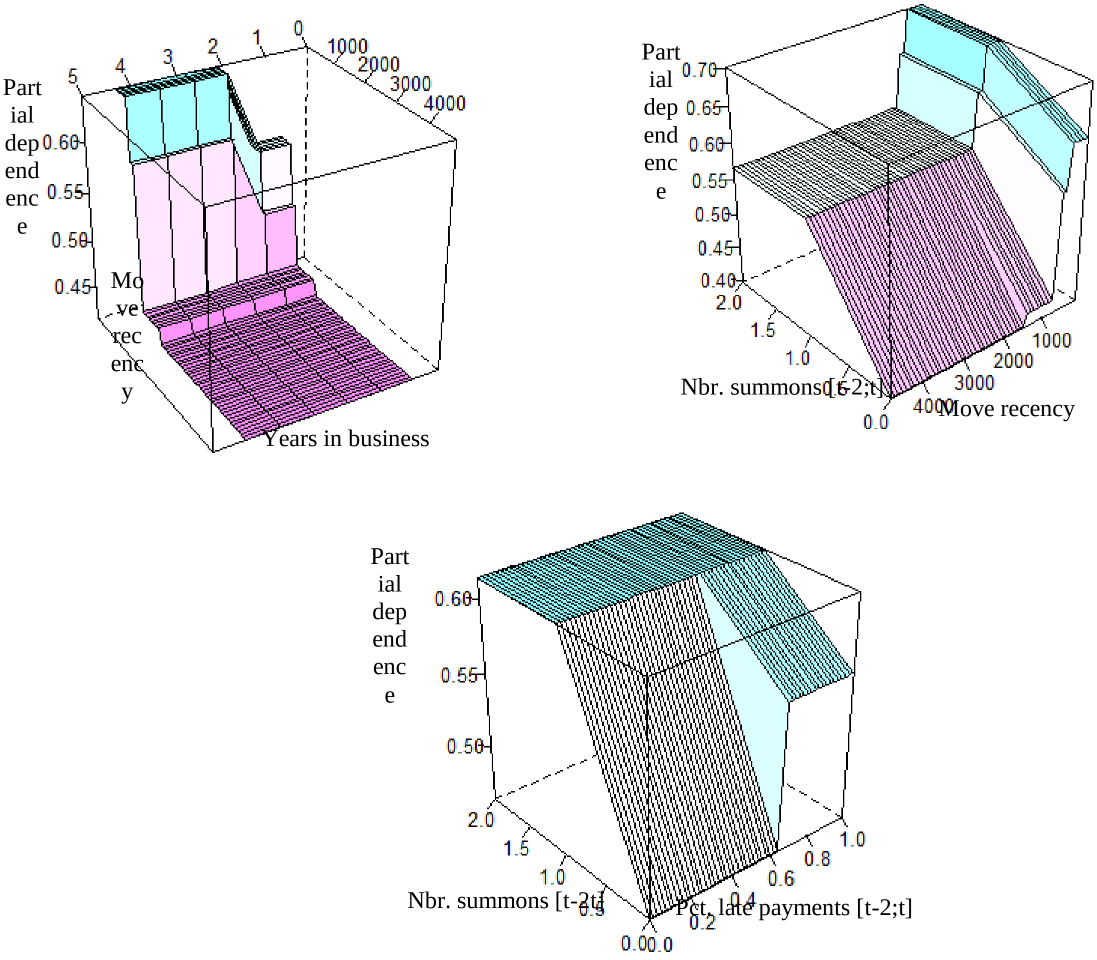


Figure 6: Graphical representation of partial dependence plots for 3 most important interaction effects

Figure 6 shows these partial dependence plots for the 3 most important interaction effects: *Move recency – Years in business*, *Nbr. Summons [t-2;t] – Move recency* and *Nbr. Summons [t-2;t]* and *Pct late payments [t-2,t]*. In the first, the interaction effect dictates that a recent move increases the failure probabity (i.e., the main effect of *Move recency*), but that this effect is substantially less pronounced when the company is young. This partial dependence plot also demonstrates that the variables *Years in business* has no impact when the company has not recently moved. The second interaction effect shows that the main effect of Move recency weakens for larger values of . *Nbr. Summons [t-2;t]*, the number of social security summons in the past two years. Lastly, the last interaction effect dictates that there is a positive relation between the percentage of late payments and company risk, but once at least one social security summon is observed, this effect is cancelled out.

6 Conclusions and Study Limitations

In the context of a growing interest in principles and practice of risk management, and specifically, enterprise risk management (ERM), companies turn to business intelligence and data mining tools to help them anticipate, face and overcome many types of risk. In this study, rule ensembles and a novel variation, spline-rule ensembles, are introduced and benchmarked in the domain of business failure prediction, a key tool for assessing and minimizing risks associated in business relations. Spline-rule ensembles extend the rule ensembles framework by introducing penalized cubic regression splines as a third term category in order to better accommodate simple, nonlinear effects. Due to the model's simplicity and regularization, spline-rule ensembles combine the strong performance of ensemble learning whilst offering a high degree of model interpretability and thus avoiding increased model complexity and more difficult model interpretation, a pitfall often associated with ensemble learning methods. Straightforward model interpretation is a quality typically more associated with uncombined (non-ensemble) methods. As such, spline-rule ensembles can be seen to offer the better of two worlds. To train a spline-rule ensemble model, in a first phase, a large set of rules are derived from decision trees and splines are trained, while in a second phase, ensemble selection is applied through regularized linear regression. As such, compact and insightful, yet powerful models are obtained. Both characteristics are investigated in the domain of business failure prediction. First, an experimental evaluation of the method demonstrates the superiority of spline-rule ensembles over conventional rule ensembles, and the method's ability to outperform several well-established, yet powerful methods in the field. Second, the method's integrated mechanisms to extract insights from the model are exemplified through a case study focusing on business failure prediction in the services sector in Belgium These include the rule model itself (rules, splines, linear terms and model coefficients), rule importances, variable importance measures, partial dependence functions and plots and interaction strengths.

Certain limitations can be identified for the current setup and the technique under consideration. First, while this study clearly demonstrates the versatility of spline-rule ensembles in terms of explaining underlying mechanisms and relationships within a business failure model, it does not link back these model insights to variable effects discovered in prior research. Future research should address a comparison between methods, and studies, in terms of the drivers of business failure and the nature of their influence. Second, the setup does not allow to assess the extent to which a spline-rule ensemble model is capable of unveiling the true data generation process and relations between variables, and how it compares to other techniques in this respect. To this end, experiments on simulated data would be more appropriate.

Acknowledgements

The author thanks B. Vindevogel and the anonymous data aggregator for providing part of the data used in this study and the editor and reviewers for their insightful comments.

References

Abellán, J., & Castellano, J. G. (2017). A comparative study on base classifiers in ensemble methods for credit scoring. Expert Systems with Applications, 73, 1-10.

Alfaro, E., García, N., Gámez, M., & Elizondo, D. (2008). Bankruptcy forecasting: An empirical comparison of AdaBoost and neural networks. Decision Support Systems, 45 (1), 110-122.

Altman, E. I. (1968). Financial Ratios, Discriminant Analysis and Prediction of Corporate Bankruptcy. Journal of Finance, 23 (4), 589-609.

Atiya, A. F. (2001). Bankruptcy prediction for credit risk using neural networks: A survey and new results. IEEE Transactions on Neural Networks, 12 (4), 929-935.

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. Journal of the Operational Research Society, 54 (6), 627-635.

Balcaen, S., & Ooghe, H. (2006). 35 years of studies on business failure: an overview of the classic statistical methodologies and their related problems. The British Accounting Review, 38 (1), 63-93.

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

Bauer, E., & Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. Machine Learning, 36 (1-2), 105-139.

Beaver, W. H. (1966). Financial Ratios as Predictors of Failure. Journal of Accounting Research, 4, 71-111.

Bou-Hamad, I., Larocque, D., & Ben-Ameur, H. (2011). Discrete-time survival trees and forests with time-varying covariates: application to bankruptcy data. Statistical Modelling, 11 (5), 429-446.

Breiman, L. (1996). Bagging predictors. Machine Learning, 24 (2), 123-140.

Brigham, E. F., & Gapenski, L. C. (1994). Financial Management: Theory and Practice, 7th. ed. Orlando, FL.: The Dryden Press.

Chava, S., & Jarrow, R. A. (2004). Bankruptcy Prediction with Industry Effects. Review of Finance, 8, 537-569.

Chen, N., & Ribeiro, B. (2013). A Consensus Approach for Combining Multiple Classifiers in Cost-Sensitivey Bankruptcy Prediction. In M. Tomassini (Ed.), ICANNGA 2013: Springer-Verlag.

Cortes, E. A., Martinez, M. G., & Rubio, N. G. (2007). A boosting approach for corporate failure prediction. Applied Intelligence, 27 (1), 29-37.

Craven, P., & Wahba, G. (1979). Smoothing noisy data with spline functions - estimating the correct degree of smoothing by the method of generalized cross-validation. Numerical Mathematics, 31, 377-403.

Creditreform Wirtschaftsforschung. (2014). Unternehmensinsolvenzen in Europa - Jahr 2013/14. Available online via http://www.creditreform.de/aktuelles/wirtschaftsforschung/insolvenzen-in-europa.html

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research, 7, 1-30.

Dietterich, T. G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. Machine Learning, 40 (2), 139-157.

Dimitras, A. I., Zanakis, S. H., & Zopounidis, C. (1996). A survey of business failures with an emphasis on prediction methods and industrial applications. European Journal of Operational Research, 90 (3), 487-513.

Doumpos, M., Andriosopoulos, K., Galariotis, E., Makridou, G., & Zopounidis, C. (2017). Corporate failure prediction in the European energy sector: A multicriteria approach and the effect of country characteristics. European Journal of Operational Research, 262 (1), 347-360.

European Commission. (2016). Country Report Belgium 2016. Available online via http://ec.europa.eu/europe2020/pdf/csr2016/cr2016_belgium_en.pdf

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 55 (1), 119-139.

Friedman, J. H., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. 2010, 33 (1), 22.

Friedman, J. H., & Popescu, B. E. (2008). Predictive learning via rule ensembles. Annals of Applied Statistics, 2 (3), 916-954.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of the American Statistical Association, 32 (200), 675-701.

Frydman, H., Altman, E. I., & Kao, D.-L. (1985). Introducing Recursive Partitioning for Financial Classification: The Case of Financial Distress. Journal of Finance, 40 (1), 269-291.

Grablowsky, B. J., & Talley, W. K. (1981). Probit and Discriminant Factors for Classifying Credit Applicants: A Comparison. Journal of Economics and Business, 33, 254-261.

Hall, M. A. (2000). Correlation-based feature selection for discrete and numeric class machine learning. In P. Langley (Ed.), Seventeenth International Conference on Machine Learning (ICML 2000): Morgan Kauffman.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). The Elements of Statistical Learning: Data Mining, Inference and Prediction. New York: Springer-Verlag.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. Biometrika, 75 (2), 383-386.

Janssens, D., Wets, G., Brijs, T., & Vanhoof, K. (2005). Adapting the CBA algorithm by means of intensity of implication. Information Sciences, 173 (4), 305-318.

Jo, H., & Han, I. (1996). Integration of case-based forecasting, neural network, and discriminant analysis for bankruptcy prediction. Expert Systems with Applications, 11 (4), 415-422.

Kotsiantis, S., Tzelepis, D., Koumanakos, E., & Tampakas, V. (2007). Selective costing voting for bankruptcy prediction. International Journal of Knowledge-based and Intelligent Engineering Systems, 11 (2), 115-127.

Kuncheva, L. I. (2004). Combining pattern classifiers: methods and algorithms. Hoboken, New Jersey: John Wiley & Sons.

Langley, P. (2000). Crafting papers on Machine Learning. In P. Langley (Ed.), 17th International Conference on Machine Learning (ICML 2000) (pp. 1207 - 1216). Stanford, CA.: Stanford University.

Lanine, G., & Vennet, R. V. (2006). Failure prediction in the Russian bank sector with logit and trait recognition models. Expert Systems with Applications, 30 (3), 463-478.

Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research, 247 (1), 124-136.

Li, H., & Sun, J. (2011a). On performance of case-based reasoning in Chinese business failure prediction from sensitivity, specificity, positive and negative values. Applied Soft Computing, 11 (1), 460-467.

Li, H., & Sun, J. (2011b). Principal component case-based reasoning ensemble for business failure prediction. Information & Management, 48 (6), 220-227.

Martin, D. (1977). Early warning of bank failure: A logit regression approach. Journal of Banking & Finance, 1 (3), 249-276.

McGurr, P. T., & DeVaney, S. A. (1998). Predicting Business Failure of Retail Firms: An Analysis Using Mixed Industry Models. Journal of Business Research, 43 (169-176).

McKee, T. E. (2003). Rough sets bankruptcy prediction models versus auditor signalling rates. Journal of Forecasting, 22 (8), 569-586.

Meyer, P. A., & Pifer, H. W. (1970). Prediction of Bank Failures. Journal of Finance, 25 (853-868).

Nanni, L., & Lumini, A. (2009). An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring. Expert Systems with Applications, 36 (2, Part 2), 3028-3033.

Ohlson, J. A. (1980). Financial Ratios and the Probabilistic Prediction of Bankruptcy. Journal of Accounting Research, 18 (1), 109-131.

Olmeda, I., & Fernández, E. (1997). Hybrid Classifiers for Financial Multicriteria Decision Making: The Case of Bankruptcy Prediction. Computational Economics, 10 (4), 317-335.

Olson, D. L., Delen, D., & Meng, Y. (2012). Comparative Analysis of Data Mining Methods for Bankruptcy Prediction. Decision Support Systems, 52, 464-473.

Park, C.-S., & Han, I. (2002). A case-based reasoning with the feature weights derived by analytic hierarchy process for bankruptcy prediction. Expert Systems with Applications, 23 (3), 255-264.

Pendharkar, P. C. (2005). A threshold-varying artificial neural network approach for classification and its application to bankruptcy prediction problem. Computers & Operations Research, 32 (10), 2561-2582.

Provost, F., Fawcett, T., & Kohavi, R. (2000). The Case against Accuracy Estimation for Comparing Induction Algorithms. In J. Shavlik (Ed.), 15th International Conference on Machine Learning (ICML 1998) (pp. 445-453). Madison, Wisconsin.: Morgan Kaufman.

Ravi Kumar, P., & Ravi, V. (2007). Bankruptcy prediction in banks and firms via statistical and intelligent techniques – A review. European Journal of Operational Research, 180 (1), 1-28.

Ravi, V., Kurniawan, H., Thai, P. N. K., & Kumar, P. R. (2008). Soft computing system for bank performance prediction. Applied Soft Computing, 8 (1), 305-315.

Rodríguez, J. J., Kuncheva, L. I., & Alonso, C. J. (2006). Rotation forest: A new classifier ensemble method. Ieee Transactions on Pattern Analysis and Machine Intelligence, 28 (10), 1619-1630.

Ross, S. A., Westerfield, R. W., Jordan, B. D., & Roberts, G. S. (2002). Fundamentals of Corporate Finance, Fourth Edition. Toronto, Canada: McGraw-Hill Ryerson.

Sun, J., Jia, M.-y., & Li, H. (2011). AdaBoost ensemble for financial distress prediction: An empirical comparison with data from Chinese listed companies. Expert Systems with Applications, 38 (8), 9305-9312.

Sun, J., & Li, H. (2008). Data mining method for listed companies' financial distress prediction. Knowledge-Based Systems, 21 (1), 1-5.

Sun, J., Li, H., Huang, Q.-H., & He, K.-Y. (2014). Predicting financial distress and corporate failure: A review from the state-of-the-art definitions, modeling, sampling, and featuring approaches. Knowledge-Based Systems, 57, 41-56.

Sun, L., & Shenoy, P. P. (2007). Using Bayesian networks for bankruptcy prediction: Some methodological issues. European Journal of Operational Research, 180 (2), 738-753.

Tsai, C.-F. (2009). Feature selection in bankruptcy prediction. Knowledge-Based Systems, 22 (2), 120-127.

van Wezel, M., & Potharst, R. (2007). Improved customer choice predictions using ensemble methods. European Journal of Operational Research, 181 (1), 436-452.

Verikas, A., Kalsyte, Z., Bacauskiene, M., & Gelzinis, A. (2010). Hybrid and Ensemble-Based Soft Computing Techniques in Bankruptcy prediction: A Survey. Soft Computing, 14, 995-1010.

Weiss, G. M. (2004). Mining with rarity: a unifying framework. SIGKDD Explorations, 6 (1), 315-354.

West, D., Dellana, S., & Qian, J. (2005). Neural network ensemble strategies for financial decision applications. Computers & Operations Research, 32 (10), 2543-2559.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. Biometrics bulletin, 1 (6), 80-83.

Wood, S. N. (2004). Stable and Efficient Multiple Smoothing Parameter Estimation for Generalized Additive Models. Journal of the American Statistical Association, 99 (467), 673-686.

Wood, S. N. (2006). Generalized Additive Models: An Introduction with R. Boca Raton, FL.: Chapman & Hall/CRC.

Wu, D. D., Chen, S.-H., & Olson, D. L. (2014). Business intelligence in risk management: Some recent progresses. Information Sciences, 256, 1-7.

Wu, D. D., & Olson, D. L. (2010). Enterprise risk management: a DEA VaR approach in vendor selection. International Journal of Production Research, 48 (16), 4919-4932.

Wu, D. D., Olson, D. L., & Dolgui, A. (2015). Decision making in enterprise risk management: A review and introduction to special issue. Omega, 57, Part A, 1-4.

Wu, D. D., Olson, D. L., & Luo, C. (2014). A Decision Support Approach for Accounts Receivable Risk Management. IEEE Transactions on Systems Man Cybernetics-Systems, 44 (12), 1624-1632.

Wu, D. D., Zhang, Y., Wu, D., & Olson, D. L. (2010). Fuzzy multi-objective programming for supplier selection and risk modeling: A possibility approach. European Journal of Operational Research, 200 (3), 774-787.

Wu, D. D., Zheng, L., & Olson, D. L. (2014). A Decision Support Approach for Online Stock Forum Sentiment Analysis. IEEE Transactions on Systems Man Cybernetics-Systems, 44 (8), 1077-1087.

Wu, W.-W. (2010). Beyond Business Failure Prediction. Expert Systems with Applications, 37, 2371-2376.

Xie, Y. Y., Li, X., Ngai, E. W. T., & Ying, W. Y. (2009). Customer churn prediction using improved balanced random forests. Expert Systems with Applications, 36 (3), 5445-5449.

Xu, L., Krzyzak, A., & Suen, C. Y. (1992). Methods of combining multiple classifiers and their applications to handwriting recognition. IEEE Transactions on Systems, Man and Cybernetics, 22 (3), 418-435.

Zięba, M., Tomczak, S. K., & Tomczak, J. M. (2016). Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. Expert Systems with Applications, 58, 93-101.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67 (2), 301-320.