# Cost-sensitive business failure prediction when misclassification costs are uncertain: A heterogeneous ensemble selection approach

Koen W. de Bock, Kristof Coussement, Stefan Lessmann

# Cost-Sensitive Business Failure Prediction When Misclassification Costs Are Uncertain: a Heterogeneous Ensemble Selection Approach

Koen W. De Bock[1], Kristof Coussement[2,3] and Stefan Lessmann[4]

[1] Audencia Business School, 8 Route de la Jonelière, F-44312, Nantes, France

[2] IESEG School of Management, 3Rue de la Digue, F-59000, Lille, France

[3] LEM-CNRS 9221, 3 Rue de la Digue, Lille, France

[4] School of Business and Economics, Humboldt-University of Berlin, Unter den Linden 6, D-10099 Berlin, Germany

E-mail addresses: kdebock@audencia.com (Koen W. De Bock), k.coussement@ieseg.fr (Kristof Coussement), stefan.lessmann@hu-berlin.de (Stefan Lessmann)

Corresponding author: Koen W. De Bock, Audencia Business School, 8 Route de la Jonelière, F-44312, Nantes, France, Tel.: +33 2 40 37 34 34

This article is published in the ***European Journal of Operational Research.***

Please cite as:

**De Bock, K. W., Coussement, K., & Lessmann, S. (2020). Cost-sensitive business failure prediction when misclassification costs are uncertain: A heterogeneous ensemble selection approach. European Journal of Operational Research, 285(2), 612–630.**

Article doi: https://doi.org/10.1016/j.ejor.2020.01.052

# Cost-Sensitive Business Failure Prediction When Misclassification Costs Are Uncertain: a Heterogeneous Ensemble Selection Approach


## Abstract

In order to assess risks associated with establishing relationships with corporate partners such as clients, suppliers, debtors or contractors, decision makers often turn to business failure prediction models. While a large body of literature has focused on optimizing and evaluating novel methods in terms of classification accuracy, recent research has acknowledged the existence of asymmetric misclassification costs associated with prediction errors and thus, advocates the usage of alternative evaluation metrics. However, these papers often assume a misclassification cost matrix to be known and fixed for both the training and the evaluation of models, whereas in reality these costs are often uncertain. This paper presents a methodological framework based upon heterogeneous ensemble selection and multi-objective optimization for cost-sensitive business failure prediction that accommodates uncertainty at the level of misclassification costs. The framework assumes unknown costs during model training and accommodates varying degrees of uncertainty during model deployment. Specifically, NSGA-II is deployed to optimize cost space resulting in a set of pareto-optimal ensemble classifiers where every learner minimizes expected misclassification cost for a specific range of cost ratios. An extensive set of experiments evaluates the method on multiple data sets and for different scenarios that reflect the extent to which cost ratios are known during model deployment. Results clearly demonstrate the ability of our method to minimize cost under the absence of exact knowledge of misclassification costs.



**Keywords:** Business failure prediction, cost-sensitive learning, ensemble selection, Brier curves, cost curves, cost uncertainty, multicriteria optimization, genetic algorithms, NSGA-II

# 1 Introduction

In the wake of the financial crisis of 2008 and the subsequent economic downturn, numerous companies experienced financial distress. Figures of insolvent companies rose up to 178,000 in the European Union alone. After a number of stable years, 2012 saw another year-on-year rise of 9.1 percent in bankruptcies (Creditreform, 2014). In Western Europe, insolvencies are expected to rise 3% in 2019 due to slowing economic growth, slowing world trade and unstable trade regulation (Bodnar, 2019). In this light, business failure prediction (BFP) will continue to play a significant role as an instrument for assessing the risk of corporate failure of collaborating companies. BFP models predict business failure or financial distress based upon all that is known about a company at a given moment in time. Such models first generalize the link between business failure and a range of variables characterizing the company, its activities and performance based upon historical data. Then, in a second stage, the model allows the risk analyst to produce estimations of future business failure for a new set of companies based upon their current profile and performance.

Numerous algorithms have been deployed for BFP models. Early approaches of Altman (1968), Martin (1977) and Ohlson and James (1980) predicted business failure using multivariate statistical methods on a variety of financial ratio's. More recent studies have focused on data mining methods. Examples are artificial neural networks (Pendharkar, 2005), support vector machines (Li & Sun, 2011a), Bayesian networks (Sun & Shenoy, 2007), decision trees (Frydman, Altman, & Kao, 1985) and ensemble classifiers (Li & Sun, 2011b). A comprehensive review of statistical and data mining techniques used for BFP can be found in (Ravi Kumar & Ravi, 2007).

Ensemble classifiers have evangelized the practice of combining predictions from individual models in BFP (Verikas et al., 2010). Ensemble classifiers use combinations of predictions generated by constituent models called ensemble *members* (Kuncheva & Rodriguez, 2007). The main factor defining the popularity of ensemble algorithms in the field of BFP is the strong prediction performance (Sun et al., 2014; Verikas et al., 2010). An ensemble of member classifiers is likely to generate better and more robust predictions than a single algorithm when accuracy and diversity are present amongst

the ensemble members. This paper focuses on BFP through *heterogeneous ensemble selection*, a subclass of ensemble classifier algorithms that seeks diversity through an interplay of member model variation and selective member fusion. When diversity is sought by combining ensemble members originating from different algorithms, *heterogeneous* ensemble classifiers are created. Many studies have demonstrated the added value of creating heterogeneous ensemble classifiers in BFP studies (Doumpos & Zopounidis, 2007; Ravi et al., 2008). In *ensemble selection*, a selective fusion rule excludes certain member models from the final ensemble classifier. The promise is that an elitist selection of a notably competent subcommittee of models could improve performance. In BFP, such an improvement of classification performance was demonstrated by Chen & Ribeiro (2013) who applied ensemble selection based on individual member performance and pairwise diversity.

BFP models can be evaluated through different performance metrics. Classification metrics such as accuracy and ranking metrics such as AUC are commonly reported in BFP studies. Despite their ease of interpretation, such metrics fail to recognize that the costs associated with the two types of errors (identifying a healthy company as a failing one, and vice versa) are rarely equal (Balcaen & Ooghe, 2006; Bauer & Agarwal, 2014). 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 a 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 consider the consequences of errors. An approach that is more in line with real-life usage of BFP models is offered through evaluation in terms of misclassification cost metrics, which accommodate unequal misclassification cost for different types of errors. While still less common nowadays, gradually more papers on BFP report expected misclassification cost (Bauer & Agarwal, 2014; Chen & Ribeiro, 2013; Kirkos, 2012; Pendharkar, 2008). When cost information is not only involved for model evaluation but also incorporated during model training and models are inherently designed to minimize misclassification cost, one enters the realm of *cost-sensitive learning* (Viaene & Dedene, 2005).

Cost-sensitive learning has been applied in BFP (e.g. Chen, Chen, & Ribeiro, 2015; Kirkos, 2012; Pendharkar, 2005). Corresponding studies accommodate asymmetric misclassification costs by involving them for model evaluation or by incorporating them in the model training phase. However, prior work in BFP assumes error costs to be known. This is an unrealistic assumption in many domains (Zadrozny & Elkan, 2001). Several reasons could introduce uncertainty on misclassification cost values or ratios, both at the moment when models are trained, and at the moment when they are deployed for scoring. First, while misclassification cost asymmetry in BFP is generally acknowledged, it is extremely challenging to estimate the exact values of misclassification costs since it is difficult to estimate the cost of a partnering company's failure (Kirkos, 2012). It is almost guaranteed that a bankruptcy will incur losses to associated companies, but the search for compensation is time- and cost-intensive, and its outcome highly uncertain (Kolay, Lemmon, & Tashjian, 2016). Second, the cost that the bankruptcy of a partnering company incurs is highly dependent on variables that could evolve over time: contract value, trade and contract terms, switching costs and external variables such as legal counsel costs, exchange and interest rates. When cost information is not known at the model training phase, the usage of traditional cost-sensitive algorithms is guaranteed to lead to a suboptimal solution.

Certain approaches such as RiskBoost (Johnson, Raeder, & Chawla, 2015) and cost-interval-sensitive support vector machines (CISVM; Liu & Zhou, 2010) have been specifically designed for scenarios of cost uncertainty. Such methods replace conventional classifiers with entirely new, purpose-built algorithms and could thus not be deployed to extend or leverage existing models, such as (heterogeneous) ensemble classifiers. Other approaches such as score calibration (Zadrozny & Elkan, 2002) or threshold varying (Hernández-Orallo, Flach, & Ferri, 2012) allow converting existing classification models to cost-sensitive learners when cost information is revealed *after* model training by transforming the model's predictions. Such approaches have two notable disadvantages: (i) they require cost information to be fully known when the model is deployed, i.e., at the scoring phase, and moreover (ii) they require additional training of a meta-model at the scoring phase that needs to be repeated whenever operating conditions (such as costs or cost ratios) evolve or vary among data

segments (Liu & Zhou, 2010). Finally, recent approaches in classifier optimization and selection (Cheng et al., 2019) have successfully deployed multicriteria optimization to directly or indirectly deal with uncertain operating conditions by optimizing multiple performance metrics simultaneously. Specifically, to build multi-purpose classifiers that perform well under various operating conditions, previous attempts have focused on optimizing classifier performance in ROC (Receiver Operating Characteristics Curve) space (e.g. Chatelain et al., 2010; Cheng et al., 2019; Zhao et al., 2018). While such approaches could be suitable for accommodating cost uncertainty, their potential for this task has not been formally evaluated and their potential for building heterogeneous ensemble classifiers been not been investigated.

To address these shortcomings, the overarching objective of this study is to raise the efficiency of BFP. We develop a novel modeling framework based on two design goals that overcome the shortcomings of existing approaches. First, we intend to build cost-sensitive ensemble models for BFP whilst accommodating cost uncertainty both at the training and at the scoring phase. We opt for heterogeneous ensemble classifiers and build upon their strong performance in the domain in the past. A second design objective is to develop a methodology that builds upon common classifier methods and requires no additional, computationally intensive analyses at the scoring phase.

To this end, we present and empirically validate a new methodological framework for building heterogeneous ensemble classifiers through ensemble selection as an approach for building cost-sensitive BFP models that accommodates unknown or uncertain misclassification costs. Specifically, during its training phase, the presented methodology involves the training of a heterogeneous library of models, followed by a cost-sensitive ensemble selection. Analogous to recent approaches in classifier optimization and selection in ROC space our ensemble selection phase implies a multicriteria optimization of *cost space* (Drummond & Holte, 2006), a classifier evaluation framework used to map a model's expected misclassification cost over a range of possible operating conditions, such as cost ratios. The result of the ensemble selection is a set of pareto-optimal ensemble classifiers obtained through multicriteria optimization, in cost space, and an *ensemble nomination curve* that maps

competence regions of these ensemble classifiers. This ensemble nomination curve allows, at the model scoring phase, to nominate one particular ensemble classifier that will deliver predictions. Depending on the remaining level of cost uncertainty at the model scoring phase, our framework prescribes alternative usages of the ensemble nomination curve. To validate the framework, extensive experiments are conducted on a large number of data sets collected for predicting business failure in various countries and sectors.

This study contributes to literature in several ways. Conceptually, our study is the first to acknowledge and address the common problem of cost uncertainty during both model training and deployment in any business analytics or risk analysis related predictive scoring application in general, and the BFP literature specifically, by means of an integrated modeling framework. Second, methodologically, our study introduces the practice of multicriteria optimization of cost space to the problem of heterogeneous ensemble selection. Additionally, it is the first to translate the result of this multicriteria optimization to an ensemble model selection framework in function of two alternative classifier performance measurement frameworks for cost-sensitive learning in cost space, i.e. cost curves (Drummond & Holte, 2006) and Brier curves (Hernández-Orallo, Flach, & Ramirez, 2011) which can help analysts select the best ensemble in function of cost uncertainty and available cost information. Finally, this study sets a benchmark for evaluating cost-sensitive classifiers under cost sensitivity by distinguishing between three levels of cost uncertainty.

## 2 Related literature

This subsection discusses related literature in three domains related to the approach presented in this study: ensemble learning in BFP, ensemble selection and finally, cost-sensitive learning under cost uncertainty.

### *2.1 Ensemble Learning for Business Failure Prediction*

The paper contributes to the literature of heterogeneous ensemble classifiers applied to BFP. Table 1 presents an overview of applications of heterogeneous ensemble classifier in BFP and is an extension to a literature overview by Verikas et al. (2010). It also includes selected applications of heterogeneous

ensemble learning in the domain of credit scoring. Since this domain is closely related to business failure prediction, both conceptually and methodologically, we believe it is important to extend our literature overview to credit scoring, especially since this domain has seen successful applications of ensemble selection recently.

A number of conclusions emerge from Table 1. First, several well-established algorithms in the BFP domain emerge as popular base learners for hybrid ensembles; most notably multi-layer perceptrons (MLP), support vector machines (SVM), linear discriminant analysis (LDA), logistic regression (LP) and decision trees (CART, C4.5). In recent credit scoring applications, a broader selection of base learners was incorporated. Second, with the exception of Lessmann et al. (2015) previous studies have not varied model parameters and the number of models considered in past approaches in BFP is limited. This study considers a substantially larger number of models, both through incorporating more ensemble member algorithms, and through varying model parameters. Third, the study by Chen & Ribeiro (2013) is the only one proposing a cost-sensitive method based on ensemble selection and is in that sense more closely related to the method presented here. However, their method does not accommodate cost uncertainty during model training, nor does it include a model evaluation under this realistic assumption. Finally, while experimental validations in previous studies only considered one data set, this study empirically compares models on a solid basis of 21 data sets, covering various industries and countries.

### *2.2 Ensemble Selection*

The practice of nominating the members of an ensemble model out of a larger pool or library of models is denoted *ensemble pruning* (Zhou, 2012) and in the context of heterogeneous ensembles, the term *ensemble selection* (ES) is often used (Caruana et al., 2004). Common motivations for ensemble selection include increased efficiency, since less storage space and computational resources are required for storing and operationalizing ensemble learners; comprehensibility since smaller ensembles could lead to less complex, and therefore more interpretable models, and most commonly, improved model performance (Zhou, 2012).

Prior ensemble selection algorithms differ in terms of (i) the selection approach used, (ii) whether ES is static or dynamic and (iii) the focal metrics during this process. First, many selection methods have been investigated. These include ordered aggregation (Martinez-Munoz, Hernandez-Lobato, & Suarez, 2009), clustering (Bakker & Heskes, 2003), probabilistic models (Woloszynski & Kurzynski, 2011; Woloszynski et al., 2012), and various optimization methods such as greedy forward selection (Caruana et al., 2004), nonlinear mathematical programming (Özöğür-Akyüz, Windeatt, & Smith, 2015) and evolutionary algorithms such as genetic algorithms. Second, a distinction is made between *static* and *dynamic* ensemble selection (Britto Jr, Sabourin, & Oliveira, 2014): In the former, selection occurs on a global level as part of the ensemble training, while in the latter, selection is dynamically applied on an instance-level during model scoring (dos Santos, Sabourin, & Maupin, 2008; Ko, Sabourin, & Britto, 2008). While dynamic ES was shown to increase performance in certain applications, a notable disadvantage is decreased efficiency at the scoring phase, since (i) a secondary part of the model training occurs when predictions are required and (ii) the full model pool should be stored. Finally, ensemble pruning approaches differ in terms of the metrics they optimize.

| Study | Application | Ensemble member algorithms | Parameter variation | # models | Ensemble selection | Cost sensitive model training | Cost-sensitive model evaluation | Cost uncertainty | # datasets and size |
|---|---|---|---|---|---|---|---|---|---|
| (Jo, Han, & Lee, 1997) | BFP | MLP, LDA, CBR | No | 3 | No | No | No | No | 1 (*n*=544) |
| (Olmeda & Fernández, 1997) | BFP | MLP, LDA, LR, MARS, C4.5 | No | 5 | No | No | No | No | 1 (*n*=66) |
| (Lin & McClean, 2001) | BFP | MLP, LR, LDA, C5.0 | No | 5 | No | No | No | No | 1 (*n*=1133) |
| (Kim & Yoo, 2006) | BFP | MLP, LR | No | 2 | No | No | No | No | 1 (*n*=4231) |
| (Hua et al., 2007) | BFP | SVM, LR | No | 2 | No | No | No | No | 1 (*n*=120) |
| (Ravi et al., 2008) | BFP | MLP, RBF, PNN, SVM, CART, FRB, PCA+MLP, PCA+RBF, PCA+PNN | No | 9 | No | No | No | No | 1 (*n*=1000) |
| (Sun & Li, 2008) | BFP | MLP, SVM, LDA, LR, CBR | No | 5 | No | No | No | No | 1 (*n*=270) |
| (Chen & Ribeiro, 2013) | BFP | kNN, MLP, SVM, NB, BLR, C4.5, ADT, RBF, LR, DT | No | 10 | Yes | Yes | Yes | No | 1 (*n*=37) |
| (Davalos et al., 2014) | BFP | kNN, MLP, C4.5, LDA, SVM | No | 5 | No | No | No | No | 1 (*n*=153) |
| (Lessmann et al., 2015) | Credit scoring | BN, CART, ELM, kNN, C4.5, LDA, SVM, LR, RBF, MLP, NB, VP, QDA, BAG, ADA, LMT, RF, RTF, SGB,ADT | Yes | 1141 | Yes | No | Yes | No | 8 (avg. *n*=30403) |
| (Ekinci & Erdal, 2017) | BFP | C4.5, BAG, VP, MB, RSM | No | 5 | No | No | No | No | 1 (*n*=1200) |
| (Xia et al., 2018) | Credit scoring | SVM, RF, XGB and GPC | No | n.a. | Yes | No | No | No | 4 (avg. *n* = 1438) |
| (Li et al., 2018) | Credit scoring | XGB,LR,DNN | No | 3 | No | No | No | No | 1 (*n* =80000) |
| (Papouskova & Hajek, 2019) | Credit risk modelling | FPA,CDT,HDT,C4.5,RET, AMT,M5P,RAT,LR,BN,SVM,NN,BAG, RTF,MB,ADA,LB,DR,RSM,SVR | No | 21 | Yes | No | Yes | No | 2 (avg. *n*=193785) |
| **This study** | **BFP** | **BAG, TBAG, SGB, RTF, RSM, RF, CART, C4.5, C4.4, LR, LDA, QDA, MLP, SVM, kNN, ADAC, C4.5+MC, C-RF, C-CART** | **Yes** | **200** | **Yes** | **Yes** | **Yes** | **Yes** | **21 (avg. *n*=6937)** |

**Table 1: Literature overview: applications of heterogeneous ensemble classifiers for business failure prediction. MLP=multilayer perceptron, LDA=linear discriminant analysis, QDA=quadratic discriminant analysis, RBF=radial basis function network CBR=case-based reasoning, LR=logistic regression, MARS=multivariate adaptive regression splines, ADT=alternating decision tree, DT=decision table, PNN= probabilistic neural network, FRB=fuzzy rule-based classifier, MB=MultiBoost, RSM=random subspace method, BAG=bagging, RF=random forests, TBAG=trimmed bagging, ADAC=AdaCost, kNN=k-nearest neighbors, MC=MetaCost, C-RF=cost-sensitive random forest, C-CART=cost-sensitive CART, VP=voted perceptron. PCA=principal component analysis, BLR=Bayesian logistic regression, NB=naïve Bayes, SVM=support vector machines, RTF=rotation forest, GPC=Gaussian process classifier, BN= Bayesian network, ELM=extreme learning machine, LMT=logistic model tree, XGB=XGBoost, ADT=alternating decision tree, FPA=forest penalizing attributes, CDT=credal decision tree, HDT=Hoeffding decision tree, RET=REPTree, M5P=M5P Tree, AMT=alternating model tree, RAT=random tree, M5P=M5 model tree, LOR=Broyden-Fletcher-Goldfarb-Shanno learning algorithm, LB=LogitBoost, DEC=decorate, SVR=support vector regression, DNN=deep neural network, DR=decorate**

| Study | Model type (EP/ES/CS/CO) | Optimization method(s) | Single or multicriteria optimization (SO/MO) | Optimization criteria | Integration of cost space / cost curve | Integration of cost space/ brier curve | Cost uncertainty addressed | Dataset domains | # datasets and size |
|---|---|---|---|---|---|---|---|---|---|
| (Provost & Fawcett, 2001) | CS | ROC convex hull | MO | FPR;TPR | No | No | Yes | Mixed (UCI) | 10 (avg. *n*=n.a.) |
| (Caruana, Munson, & Niculescu-Mizil, 2006) | ES | Greedy hillclimb search | SO | ACC,FSC,LFT,AUC,APR, BEP,RMS,MXE | No | No | No | Mixed | 11 (avg. *n*=22317) |
| (dos Santos, Sabourin, & Maupin, 2008) | EP | GA, NSGA-II | SO,MO | ACC, DIV | No | No | No | Mixed | 7 (avg. *n*=32492) |
| (Partalas, Tsoumakas, & Vlahavas, 2009) | ES | Reinforcement learning | MO | ACC | No | No | No | Mixed | 20 (avg. *n*=656) |
| (Chatelain et al., 2010) | CS,CO | NSGA-II | MO | FPR,TPR | No | No | Yes | Mixed (UCI); handwritten digit recognition | 7 (avg. *n*=558) |
| (dos Santos, 2012) | EP | GA, PSO, NSGA, NSGA-II, controlled elitist NSGA | SO,MO | ACC, DIV, DIM | No | No | No | Handwritten digit recognition | 2 (avg. *n*=99418) |
| (Levesque et al., 2012) | CO,EP | NSGA-II | MO | FPR,TPR | No | No | No | Mixed (UCI) | 6 (avg. *n*=563) |
| (Zhao et al., 2016) | CO | 3DCH-EMOA, NSGA-II, GDE3, SMS-EMOA, SPEA2, MOEA/D | MO | FPR,FNR,CCR | No | No | No | Mixed (UCI); spam classification | 20 (avg. *n*=802) |
| (Zhao et al., 2018) | CO | 3DCH-EMOA,3DFCH-EMOA,Two-Arch2, NSGA-III, MOEA/DD, RVEA, AR-MOEA, MPSO/D | MO | FPR,FNR,CCR | No | No | No | Mixed (UCI) | 14 (avg. *n*=792) |
| (Cheng et al., 2019) | CO | MOPA | MO | TPR,K-FPR | No | No | No | Mixed (UCI, libsvm) | 10 (avg. *n*=10674) |
| **This study** | **ES** | **NSGA-II** | **MO** | **FPR,FNR** | **Yes** | **Yes** | **Yes** | **BFP** | **21 (avg. *n*=6937)** |

**Table 2: Literature overview: optimization-based ensemble selection and classifier selection methods. EP=homogeneous ensemble pruning, ES=heterogeneous ensemble selection, CS=classifier selection, CO=classifier optimization. FPR=false positive rate, FNR=false negative rate, TPR=true positive rate, ACC=accuracy, FSC=F-score, LFT=lift, AUC=area under the ROC curve, APR=average precision, BEP= precision-recall break-even point, RMS=squared error, MXE=cross-entropy; DIV=ensemble diversity, CCR=classifier complexity ratio, DIM=ensemble size, K-FPR=partial range false positive rate. GA=genetic algorithm, 3DCH-EMOA=3D convex-hull-based evolutionary multiobjective algorithm; MOPA=multiobjective evolutionary algorithm for optimizing partial AUC, PSO=particle swarm optimization, NSGA=non-dominated sorting genetic algorithm,**

These metrics include measures of classification accuracy derived from the confusion matrix (Caruana et al., 2004; Sylvester & Chawla, 2006), ROC space (Levesque et al., 2012), or statistical measures (Caruana et al., 2004).

Other ensemble pruning approaches deploy multicriteria optimization in order to optimize several metrics simultaneously, such as measures of accuracy and ensemble diversity (dos Santos, Sabourin, & Maupin, 2008; Margineantu & Dietterich, 1997). While genetic algorithms (GA), and more specifically, multi-objective genetic algorithms (MOGA) have been used before in the setting of homogeneous ensemble pruning, they have not been deployed for heterogeneous ensemble selection. Moreover, most prior approaches focused on a selection in terms of accuracy, diversity, or both while to the best of our knowledge, no cost-sensitive applications exist in literature. This study contributes to the literature on GA-based ES through optimizing the entire cost space using a multicriteria approach with the purpose of creating cost-sensitive, heterogeneous ensemble classifiers that accommodate cost uncertainty.

### *2.3 Cost-Sensitive Learning for Uncertain Misclassification Costs*

Finally, this study contributes to literature on methodologies to tackle cost-sensitive learning when misclassification costs are not or not fully known during model training and/or scoring. A relatively limited number of methods belonging to different algorithmic paradigms has been proposed to deal with the scenario of cost uncertainty.

In Zadrozny & Elkan (2001), a method coined *cost-sensitive decision-making* is introduced for situations where costs are assumed instance-specific and known during model training, but unknown during model scoring. Their approach involves two components: the estimation of calibrated posterior probabilities and estimating instance value while applying a procedure for sample selection bias. Experiments in a setting of charitable donations showed improved performance over MetaCost (Domingos, 1999). Liu and Zhou (2010) adapt SVMs for scenarios in which cost information is provided in the form of an interval at training time. In an experimental validation, uniform probability distribution function for the cost intervals and the suggested *CISVM* algorithm is shown to outperform standard SVM and cost-sensitive SVM. A third approach by Wang and Tang (2012) assumes that exact cost information is missing, but that multiple cost matrices are given and involves the estimation of a

MiniMax classifier. The MiniMax classifier aims to minimize the maximum total cost over a set of equally likely cost matrices. An algorithm is presented that simplifies the estimation of the model to a number of standard cost-sensitive problems and sub-problems that only involve two cost matrices at a time. Both CISVM and Wang & Tang's (2012) MiniMax classifier assume some cost information to be known at training time. A fourth approach is *RiskBoost* (Johnson, Raeder, & Chawla, 2015), a variant of AdaBoost which iteratively assigns higher weights to instances that are misclassified by the member classifier with the highest risk, where risk denotes the expected cost of that classifier given a likelihood distribution over a range of cost ratios. Experiments demonstrated improved AUC performance over a set of UCI datasets.

Finally, our framework is related to a stream of approaches that pursue classifier or ensemble selection through evaluating and optimizing model performance in ROC space by means of multicriteria optimization (e.g. Cheng et al., 2019; Zhao et al., 2018). Classifier selection is different from ES in that it involves the selection of one single model out of a set of models at the scoring phase. Two approaches for classifier selection address cost uncertainty explicitly. First, Provost and Fawcett (2001) propose a method that suggests classifier selection through consulting the ROC convex hull (ROCCH) formed by a set of pre-trained models. Second, Chatelain et al. (2010) propose an evolutive model selection framework for SVM classifiers where hyperparameters are evolved using a multi-objective genetic algorithm in order to optimize classifiers in ROC space. Similar to Provost and Fawcett's ROCCH approach (Provost & Fawcett, 2001), the authors propose the concept of a ROC front: the set of Pareto-optimal SVM classifiers from which an optimal classifier is chosen during runtime. The authors provide suggestions on how this selection can be achieved based on whether cost information is available or not, but the method's experimental validation is limited to a comparison in terms of AUC. The method described by Chatelain et al. (2010) is similar to ours since it prescribes a multicriteria optimization of ROC space. They introduce the concept of a ROC front that allows the analyst to choose a model in function of a desired tradeoff of false and true positive rates. However, there are several fundamental differences. Our approach is a method for selecting heterogeneous ensemble classifiers from a pre-trained library of models while Chatelain et al. (2010) focus on SVM parameter optimization and model

selection. Second, instead of focusing on ROC space, our method optimizes cost space which allows for a more intuitive linkage of model performance to operating conditions. Third, instead of the ROC front, our approach depends on the determination of an ensemble nomination curve which is based on the concept of cost space. This allows for a more intuitive model selection based on an operating condition. Fourth, our method explicitly prescribes alternative usages depending on the degree of cost uncertainty during the model's scoring phase. Finally, our experiments are more elaborate since they evaluate the models over a larger pool of datasets and in terms of multiple performance criteria and cost uncertainty scenarios.

Table 2 provides an overview of related optimization-based ensemble selection, classifier selection and classifier optimization approaches.

## 3 Methodology

### *3.1 Cost and Brier Curves*

A metric used previously for evaluating BFP models in a cost-sensitive manner is *expected misclassification cost* (EMC) (Chen & Ribeiro, 2013). EMC involves an estimation of the average cost of using the model to classify one randomly chosen instance and can be written as:

$$EMC = p(-) * p(+|-) + p(+) * p(-|+) * \alpha \qquad (1)$$

In which $p(-)$ and $p(+)$ are the business survival and failure rates, respectively, $p(+|-)$ is the false positive rate, and $p(+|-)$ is the false negative rate, while α is the cost ratio, i.e. the ratio of the cost associated with a false negative error to the cost of a false positive error.

This study assumes uncertainty with respect to misclassification costs. Hence, in order to create cost-sensitive models under this condition, a more flexible framework is required. To this end, our method relies upon the notion of *cost space* in which *cost curves* (Drummond & Holte, 2006) and *Brier curves* (Hernández-Orallo, Flach, & Ramirez, 2011) can visualize a classifier's cost-sensitive performance over a range of operating conditions. From Equation (1), it is clear that EMC depends on factors related to the *scoring context*, i.e. the failure rate and the cost ratio on the one hand, and on the classification performance of the model, i.e. the false negative and false positive rates, on the other. Cost curves

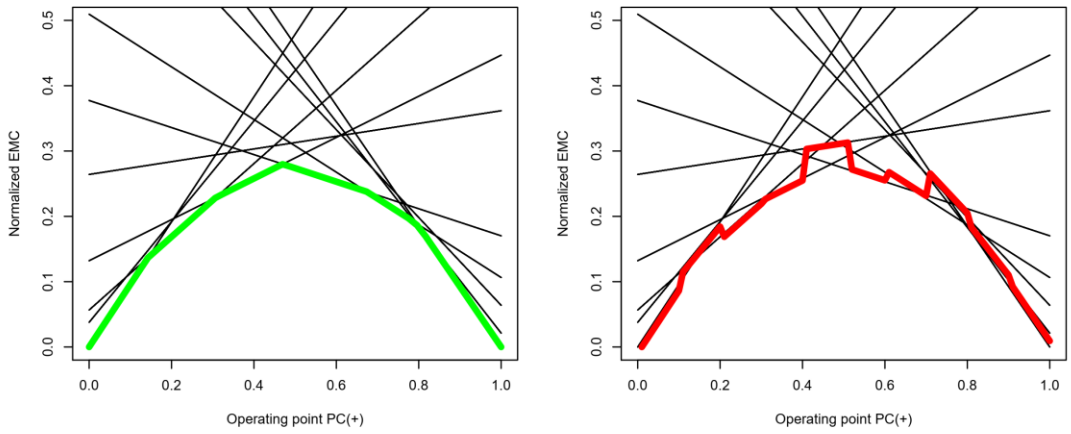
(Drummond & Holte, 2006) measure and visualize classifier cost performance over the full range of operating conditions, determined by misclassification cost ratios and class distributions. Hence, they accommodate uncertainty with respect to the scoring context, i.e. $\alpha$ and $p(-)$ in Equation (1). Specifically, the operating condition is coined 'probability costs' $(PC(+))$ and is quantified as a normalization of failure rate times the cost ratio:

$$PC(+) = \frac{\mathrm{p}(+)*\alpha}{\mathrm{p}(-)+\mathrm{p}(+)*\alpha} \qquad (2)$$

Expected misclassification costs (EMC), when normalized by dividing by the highest possible EMC and rewritten as a function of $PC(+)$, can thus be expressed as

$$EMC_{Norm}(PC(+)) = \big(p(-|+) - p(+|-)\big) * PC(+) + p(+|-) \qquad (3)$$

The cost curve is the lower envelope of all cost lines obtained for every possible threshold value used to convert numerical predictions into class predictions. This is shown in panel (a) of Figure 1.




**(a)** **(b)**

**Figure 1: Example of cost lines, a cost curve (a), and corresponding Brier curve (b) of a binary classifier**

One disadvantage of cost curves, when used to assess performance of a classifier that outputs continuous predictions that reflect prediction confidence, is that they assume optimal threshold choice to transform continuous scores into class predictions, which can prove difficult in reality. Therefore, Brier curves were proposed by Hernández-Orallo, Flach, and Ramirez (2011) as an alternative where probabilistic loss is calculated, i.e. EMC obtained through converting predicted posterior probabilities using the operating condition $PC(+)$ as a threshold. Figure 1(b) shows an example of the Brier curve.

Given the widespread adoption and simplicity of cost curve, as well as the more realistic performance measurement of the Brier curve, the CSMES framework, which we introduce below, adopts both approaches and leaves the choice to the analyst.

When no precise information is known on a classifier's operating condition (misclassification cost ratio and/or failure rate), it is insightful to consider a model's full performance profile in cost space. This is achieved by calculating the area under the cost curve (AUCC; Adams & Hand, 1999; Drummond & Holte, 2006), or equivalently, the area under the Brier curve (AUBC; Hernández-Orallo, Flach, & Ramirez, 2011), depending on the cost space framework that is considered. Both frameworks express global cost-sensitive performance without assuming a single specific cost ratio and are therefore relevant measures for evaluating model performance under high cost uncertainty at scoring time. When exact cost ratios remain unknown, but cost intervals or cost probability distributions are available at scoring time, AUCC and AUBC can be adapted to measure cost space performance for a restricted range of operating conditions, or for variable probabilities over this range. We denote these partial measures $pAUCC$ and $pAUBC$. The AUCC and pAUCC are given by the following equations:

$$AUCC = \int_0^1 EMC_{Norm}(x)dx \qquad (4)$$

$$pAUCC = \int_0^1 EMC_{Norm}(x) * Prob(PC(+) = x)dx \qquad (5)$$

Where $Prob(PC(+) = x)$ is a probability distribution function defined over the range of operating conditions $PC(+)$. The exact choice of this distribution depends on context. The area under the brier curve (AUBC) and partial area under the brier curve (pAUBC) can be calculated analogously.

### *3.2 Non-Dominated Sorting Genetic Algorithm (NSGA-II)*

The simultaneous minimization of the false positive rate and false negative rate to obtain a set of models that are suitable for deployment under varying operating conditions is tackled using a multi-objective GA. Purpose-built multicriteria optimization algorithms aim to identify the Pareto front, a set of solutions that are each optimal in their tradeoff between multiple objectives. Pareto-optimal solutions are solutions for which no objective function can be improved further without degrading performance on at least one other objective function. This study adopts a popular, widely used Pareto-based

evolutionary algorithm: NSGA-II or the fast elitist non-dominated sorting genetic algorithm (Deb et al., 2002). NSGA-II is recognized as a highly efficient algorithm for multicriteria optimization. First, it adopts *elitism*, meaning that over subsequent generations, the fittest solutions can be preserved. Second, it enforces diversity in terms of objective functions (and thus, dispersion over the Pareto front range) using the concept of *crowding distance* and incorporating this distance measure into the assessment of solution fitness. We kindly refer the reader to Appendix A and to Deb et al. (2002) for a detailed explanation of the NSGA-II algorithm.

### *3.3 Cost-Sensitive Multicriteria Ensemble Selection (CSMES)*

The method presented in this study for tackling cost-sensitive classification under cost uncertainty in a PFB context is denoted *Cost-Sensitive Multicriteria Ensemble Selection* (CSMES). The algorithm's training and scoring phases are visualized in Figure 2 and Figure 3 respectively. These procedures are explained in detail in the following subsections.

#### *3.3.1 CSMES Training Phase*

The CSMES model training phase involves three steps: (i) the creation of a library of models, (ii) the optimization of cost space and (iii) the derivation of an ensemble nomination curve. Note that misclassifications costs, or their ratio, are assumed unknown during the training phase.

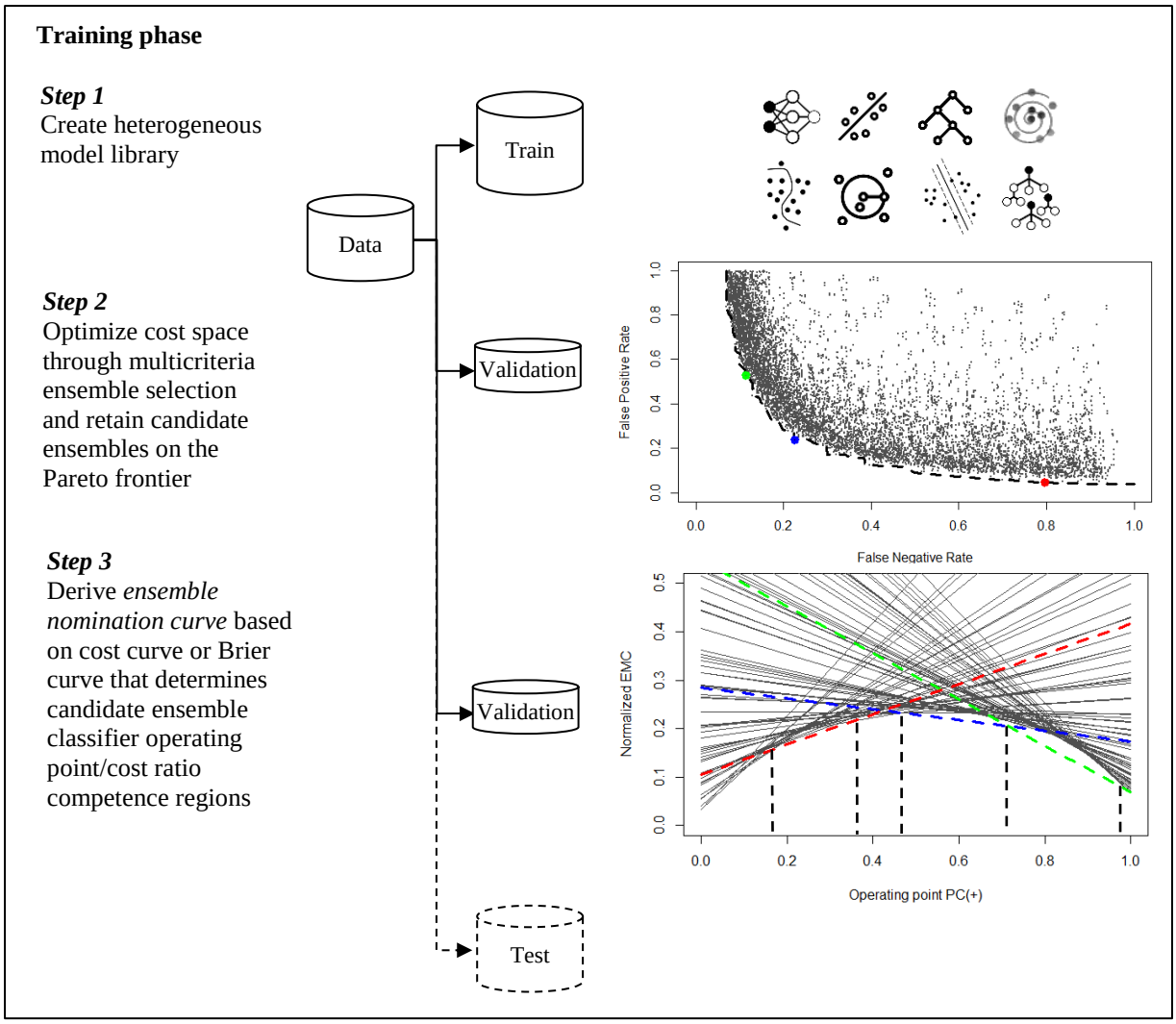



**Figure 2: Graphical representation of training phase of CSMES**

### 3.3.1.1 Model Library Creation

Analogous to other ensemble selection approaches (Caruana et al., 2004; Zhou, 2012), the first step of the algorithm involves the creation of a heterogeneous model library, where algorithms and their hyperparameters are varied to estimate multiple models using a training data set. The approach adopted in this study involves the inclusion of several well-known algorithms commonly available in analytical software environments. The exact selection adopted in the empirical validation of this study is revealed in Section 4.4.

### 3.3.1.2 Cost Space Optimization

The second step in the training phase of CSMES is heterogeneous ensemble selection. To this end, we adopt an approach similar to methods for classifier optimization and classifier selection (Chatelain

et al., 2010; Cheng et al., 2019; Levesque et al., 2012; Zhao et al., 2016). These studies suggest multicriteria optimization in order to evolve and select classifiers by optimizing ROC space, i.e. by simultaneously maximizing true positive rate and minimizing false positive rate. As such, this strategy can be easily applied to our objective, which is to select multiple ensemble classifiers that are optimal in cost space.

As discussed in Section 3.1; in cost space, a classifier's performance (EMC) depends on $p(+|-)$; the false positive rate, and $p(+|-)$; false negative rate and the operating condition. Through multicriteria optimization, and specifically NSGA-II, both the dimensions $p(+|-)$; and $p(+|-)$ are minimized simultaneously. Instead of obtaining a single optimal ensemble, a Pareto-optimal set of ensemble classifiers is obtained that each represent an optimal tradeoff between both metrics. In cost space, each Pareto-optimal ensemble is represented through a cost line. These candidate ensembles will thus each be optimal for a certain subrange of operating points (PC(+)). Note that in order to reduce the risk of overfitting, a validation data sample should be foreseen for this step. The Pareto-frontier obtained through optimizing both $p(+|-)$; and $p(+|-)$ corresponds to the lower envelope of cost curves.

Figure 2 illustrates how three Pareto-optimal classifiers (the colored dots in the upper plot and similarly colored dashed lines in the lower plot) are optimal in cost space: they minimize EMC for different operating point ranges (shown in the lower plot).

Brier curves relax the somewhat unrealistic assumption of optimal threshold choice through a simple choice rule for classification thresholds. Consequently, the calculation of EMC values that constitute the Brier curve framework differs slightly from those used in cost curves. In the case of the Brier curves, a cutoff is chosen equal to the operating condition, whilst in the case of the cost curve, the cutoff that minimizes EMC is assumed. However, in our approach, the optimization of cost space is identical for both frameworks. Since an arbitrary threshold of 0.5 is used to convert a candidate ensemble's predictions into class predictions which are then used to calculate $p(+|-)$; and $p(+|-)$, the influence of the threshold choice is cancelled in this stage, and a single optimization is needed to optimize both cost space frameworks simultaneously.

#### 3.3.1.3 Ensemble Nomination Curve Derivation

The third step in the CSMES training phase is the derivation of an *ensemble nomination curve*. This curve is the lower envelope of cost or Brier curves of all pareto-optimal ensemble classifiers obtained in the previous step. These cost or Brier curves are calculated for the validation sample. The ensemble nomination curve determines candidate ensemble classifier operating point/cost ratio competence regions, i.e., which candidate ensemble classifiers are optimal for which operating condition ranges. Figure 2 shows an example of a cost curve-based ensemble nomination curve and illustrates how different candidate ensembles are optimal for different operating conditions.

### *3.3.2 CSMES Scoring Phase*

The scoring phase of CSMES involves three steps: (i) determination of the degree of cost uncertainty, (ii) ensemble classifier nomination and (iii) model scoring. These are visualized in Figure 3.

#### 3.3.2.1 Degree of Cost Uncertainty Determination

This study aims at the conception of a cost-sensitive method for BFP to be deployed in situations of cost uncertainty. As detailed above, CSMES assumes cost uncertainty during model training (including the ensemble selection process), while during model deployment, no assumptions are made in terms of cost uncertainty. The first step of the scoring phase involves the determination of the degree of cost uncertainty remaining at the scoring phase. Three cost uncertainty scenarios are possible: either (i) *high cost uncertainty*: the uncertainty about the cost ratio remains when model predictions are due; (ii) *partial cost uncertainty*: there is still uncertainty, but a probability distribution is known over the range of cost ratios, and (iii) *no cost uncertainty*: the exact cost ratio is known.

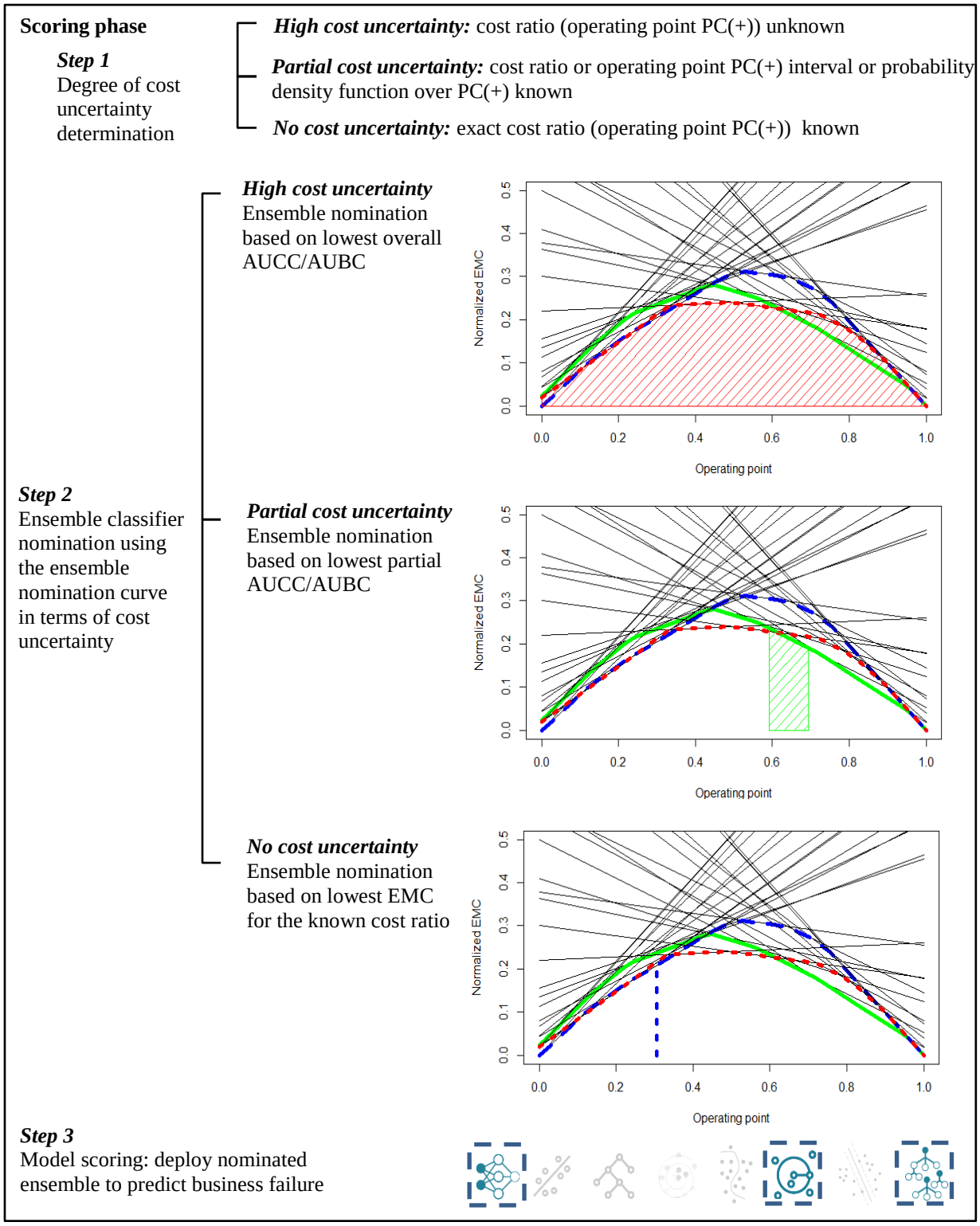



**Figure 3: Graphical representation of scoring phase of CSMES**

### 3.3.2.2 Ensemble Classifier Nomination

Depending on the degree of cost uncertainty, three alternative strategies for choosing the ensemble that will deliver predictions are proposed. To avoid any confusion with the term ensemble selection,

this step is henceforward denoted *ensemble classifier nomination*. First, under the assumption of absence of cost uncertainty, a cost ratio is known, thus $PC(+)$ (Equation 2) can be determined and the ensemble classifier nomination curve can be directly used to nominate to the optimal ensemble that minimizes EMC. Second, under the assumption of full or partial cost uncertainty a different strategy for ensemble nomination is required. For these settings, it is possible to evaluate a classifier's theoretical performance in cost space, i.e. over the range of operating conditions that are likely to occur at the scoring phase. Specifically, in the scenario of full cost uncertainty we rely upon the AUCC and AUBC measure that express a classifier's performance throughout cost space. The ensemble classifier with the best (smallest) overall AUCC or AUBC is selected. Analogously, in the scenario of partial cost uncertainty, pAUCC and pAUBC are relied upon to reveal which ensemble classifier performs best in a part of cost space, i.e. over a range of operating conditions. The three ensemble nomination strategies are visualized in Figure 3.

The subsequent and final step is trivial: the models constituting the nominated ensemble deliver individual predictions which are aggregated through averaging.

## 4 Empirical validation

### *4.1 Data*

To validate the CSMES framework, a benchmarking experiment with several datasets provided by two global data aggregators is conducted. These datasets contain information about a selection of French, Italian and Belgian companies that publish consolidated annual accounts, originating from various industries. Ample research has addressed BFP at a sector level (e.g. Doumpos et al., 2017; Lanine & Vennet, 2006) while other authors (Brigham & Gapenski, 1994; Dimitras, Zanakis, & Zopounidis, 1996; McGurr & DeVaney, 1998) have suggested to develop models for BFP using homogeneous samples in terms of sector. We believe that the inclusion of multiple data sets from several countries enhances the generalizability of the reported results. Table 3 contains detailed information on the 21 data sets considered in this study. Note that companies are classified into industry categories based upon their 8-digit Standard Industry Code (SIC).

| Dataset | Country | Industry | # features | # companies | Failure rate |
|---|---|---|---|---|---|
| *1* | France | Construction industries (15.000.000 <= SIC 8 < 18.000.000) | 19 | 5 678 | 33.74% |
| *2* | France | Manufacturing (20.000.000 <= SIC 8 < 40.000.000) | 19 | 3 266 | 21.68% |
| *3* | France | Transportation, communications and utilities (40.000.000 <= SIC 8 < 50.000.000) | 19 | 1 787 | 16.96% |
| *4* | France | Wholsale trade (50.000.000 <= SIC 8 < 52.000.000) | 19 | 3 337 | 17.44% |
| *5* | France | Retail trade (52.000.000 <= SIC 8 < 60.000.000) | 19 | 6 450 | 23.55% |
| *6* | France | Finance, insurance and real estate (60.000.000 <= SIC 8 < 68.000.000) | 19 | 2 874 | 6.51% |
| *7* | France | Service industries (70.000.000 <= SIC 8 < 89.000.000) | 19 | 8 576 | 15.24% |
| *8* | Italy | Construction industries (15.000.000 <= SIC 8 < 18.000.000) | 19 | 3 801 | 14.29% |
| *9* | Italy | Manufacturing (20.000.000 <= SIC 8 < 40.000.000) | 19 | 5 093 | 12.84% |
| *10* | Italy | Transportation, communications and utilities (40.000.000 <= SIC 8 < 50.000.000) | 19 | 1 837 | 10.02% |
| *11* | Italy | Wholsale trade (50.000.000 <= SIC 8 < 52.000.000) | 19 | 3 671 | 12.45% |
| *12* | Italy | Retail trade (52.000.000 <= SIC 8 < 60.000.000) | 19 | 3 309 | 9.34% |
| *13* | Italy | Finance, insurance and real estate (60.000.000 <= SIC 8 < 68.000.000) | 19 | 3 732 | 4.02% |
| *14* | Italy | Service industries (70.000.000 <= SIC 8 < 89.000.000) | 19 | 6 579 | 5.46% |
| *15* | Belgium | Construction industries (15.000.000 <= SIC 8 < 18.000.000) | 108 | 9 976 | 4.54% |
| *16* | Belgium | Manufacturing (20.000.000 <= SIC 8 < 40.000.000) | 108 | 10 430 | 2.73% |
| *17* | Belgium | Transportation, communications and utilities (40.000.000 <= SIC 8 < 50.000.000) | 108 | 5 339 | 4.57% |
| *18* | Belgium | Wholsale trade (50.000.000 <= SIC 8 < 52.000.000) | 108 | 15 896 | 3.04% |
| *19* | Belgium | Retail trade (52.000.000 <= SIC 8 < 60.000.000) | 108 | 13 626 | 5.19% |
| *20* | Belgium | Finance, insurance and real estate (60.000.000 <= SIC 8 < 68.000.000) | 108 | 10 055 | 1.64% |
| *21* | Belgium | Service industries (70.000.000 <= SIC 8 < 89.000.000) | 108 | 20 364 | 2.73% |

**Table 3: Dataset characteristics**

The outcome variable, a binary business failure indicator (1=business failure; 0= survival) indicates the event of bankruptcy over a time horizon of 12 months. The predictors common to all datasets are financial ratios and variables related to cash flow (McGurr & DeVaney, 1998). Analogous to (Ross et al., 2002), the ratios considered in this study can be classified into liquidity ratios, long-term solvency ratios, asset management ratios and profitability ratios. Belgian datasets include additional firmographics and variables related to payment timeliness. Appendix B provides a detailed overview of all predictors included in the data sets.

Three common preprocessing steps were applied to all data sets. The first step involves the detection and treatment of outlier values (Bou-Hamad, Larocque, & Ben-Ameur, 2011; Chava & Jarrow, 2004). To this end *winsorization* is applied: variables' ranges are reduced by truncating their values below the $2.5^{th}$ and above the $97.5^{th}$ percentiles. Second, feature selection, commonly considered good practice in the domain of bankruptcy prediction (Abellán & Castellano, 2017; Tsai, 2009) was applied. Specifically, t-test-based feature selection, a filter-based feature selection approach that compares group means and has seen prior applications in BFP literature (e.g. Tsai, 2009) was chosen. Features for which failing and healthy companies are significantly different ($\alpha$=0.05) are retained. Finally, as Table 3 shows, failure rates in the data sets range from 1.64 to 33.74 percent. To counter the potential negative

impact that class imbalance exerts on many predictive methods (Weiss, 2004), *undersampling*, a common practice in BFP (Kotsiantis et al., 2007) was applied to the training data sets through random removal of majority-class instances (healthy businesses) until classes are evenly distributed.

### *4.2 Evaluation Framework and Metrics*

As detailed in Section 3, this study assumes cost uncertainty during the model training phase, while in the scoring phase CSMES supports three scenarios of full, partial, and no cost uncertainty. Hence, CSMES and benchmark methods are evaluated under each assumption of cost uncertainty, and in a cost-sensitive manner. To this end, two categories of performance measures are considered: (i) misclassification cost (EMC), and (ii) aggregated cost space-measures (AUCC, AUBC, pAUCC and pAUBC). Figure 4 visualizes the evaluation framework, indicating the three evaluation scenarios with respect to cost uncertainty, as well as the performance metrics considered in each of these scenarios.

First, in each of the three cost uncertainty scenarios, a comparison is made in terms of EMC. The objective of this comparison is to assess methods in terms of the metric we are ultimately hoping to minimize, even when the exact costs associated with misclassifications are not known. An estimate of misclassification cost can be calculated for any model by simply simulating an evaluation condition, i.e. randomly drawing a specific cost ratio and calculating EMC as defined in Equation (3) in terms of this cost ratio. By repeating this process for different cost ratios and aggregating results, one obtains an estimate of a classifier's true cost-sensitive performance over a range of operating conditions. Hence, EMC is reported for all three scenarios of cost uncertainty. Experimental results are reported over a 10-fold cross-validation.

To cover a broad range of operating conditions, 10 cost ratios are randomly drawn from a range of 1 to 20 for each dataset and cross-validation fold**.** This cost ratio range was suggested previously by Chen & Ribeiro (2013). EMC is calculated as a function of the simulated cost ratios, which results in 100 (10 cross-validation folds times 10 cost ratios) EMC values per algorithm and dataset. The detailed procedure used to simulate cost ratios is provided in Appendix D. The aggregation of these results and a statistical comparison allows for a cost-sensitive evaluation of CSMES and the benchmark algorithms in cost space.

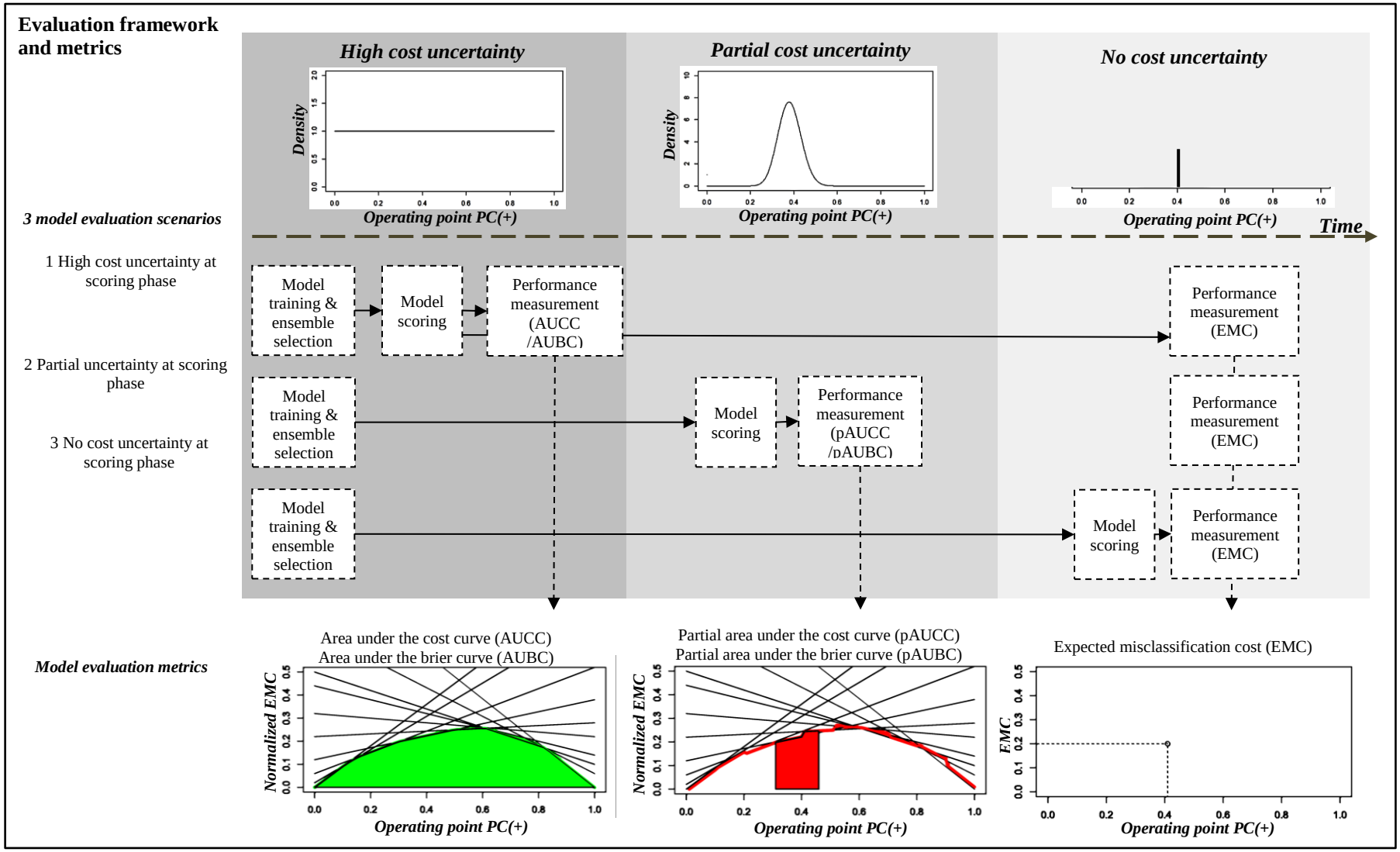



**Figure 4: Evaluation framework showing three model evaluation scenarios as a function of scoring phase cost uncertainty, and corresponding performance metrics**

Second, aggregated cost space measures are reported. These are more theoretical in nature since they do not assume a specific cost ratio, but instead summarize cost space performance by calculating the area (or partial area) under the cost curve or Brier curve. On the one hand, AUCC and AUBC (see Section 3.3.2.2) measure a classifier's performance throughout cost space. They are relevant evaluation measures when no cost information is assumed available at the scoring phase and are therefore reported for this scenario only. On the other hand, pAUCC and pAUBC measure a classifier's performance for a part of the cost space, i.e. for an interval of operating conditions and are therefore reported for the scenario of partial cost uncertainty. The underlying distribution over operating conditions (i.e., $Prob(PC(+) = x)$ in the equations of pAUCC and pAUBC; see Equations 9 and 10) should be chosen in terms of the particular nature of cost uncertainty that exists. In our experiments, the beta-distribution is chosen since it is well-suited to express varying degrees of certainty around expected operating point values that lie within the interval [0,1] (Johnson, Raeder, & Chawla, 2015). Similar to the procedure for randomly simulating hypothetical cost ratios, 10 beta-distributions for operating conditions are randomly generated per fold and dataset. The beta-distribution is characterized by two parameters, α and β that determine its shape. Instead of randomizing α and β directly, they are determined as a function of a desired (mean) operating condition, and a randomly generated desired standard deviation for the operating condition. The desired average operating conditions are derived from the simulated cost ratios, as described above, through application of Equation (3). The standard deviations are randomly chosen between 0.02 and 0.25 to cover various degrees of cost uncertainty. The result is a set of probability density functions that represent various degrees of cost ratio uncertainty around operating conditions that cover the entire range of $PC(+)$.

In the scenario in which there is no cost uncertainty at the scoring phase, the cost ratio used for ensemble classifier nomination in CSMES is consistently the same as the one used to calculate EMC values for model comparisons. When partial cost uncertainty is assumed for model scoring, cost ratios to calculate EMC values are derived from operating conditions (PC(+)) drawn from the probability density function of the simulated beta distribution that is assumed for the calculation of pAUCC and pAUBC values used for ensemble classifier nomination.

### *4.3 Benchmark algorithms*

To validate CSMES as a robust, cost-sensitive method in the presence of various degrees of cost uncertainty at the model deployment stage, it is compared to three sets of benchmark algorithms. First, a comparison is made to three algorithms that have been specifically designed as cost-sensitive classifiers when cost information is unknown during the model training stage: *CISVM* (Liu & Zhou, 2010), *RiskBoost* (Johnson, Raeder, & Chawla, 2015) and the minimax classifier by (Wang & Tang, 2012) (henceforward labeled *MiniMax*). While RiskBoost assumes full cost uncertainty, CISVM and MiniMax require some cost information to be provided. However, in this study they will be evaluated in the assumption of full cost uncertainty during model training, through providing a wide cost interval to CISVM, and a wide set of cost ratios to MiniMax.

The second set of benchmark algorithms are alternative heterogeneous ensemble classifier and ensemble selection approaches that are less complex in nature than CSMES, but built using the same model library. The purpose of this second comparison is to verify whether the performance of CISVM can be matched or surpassed by simpler strategies that either avoid optimization of cost space and an ensemble nomination step at the scoring phase, or that avoid ensemble selection altogether. One crucial benchmark (*Full*) produces predictions through simply averaging outputs of all models in the library. Other benchmarks in this category also take advance of cost-space-wide performance, but in a different way than CSMES. A variation on the full library ensemble is a weighted variant that uses AUCC or AUBC performance measures of individual models (*Weighted*). Three additional benchmarks heuristically select the single best model (*Best*), the top ten (*Top10*) and top twenty-five *(Top25*) of best performing models, respectively. Note that these ensemble selection strategies select models using area under the cost or Brier curve performance on a validation sample.

A third set of competitive benchmarks consists of ensemble or classifier selection approaches based on evolutionary algorithms The first benchmark (*GHS-ES*) is an adaptation of Caruana, Munson, and Niculescu-Mizil (2006)'s greedy hillclimb search approach, which was introduced as a versatile approach for heterogeneous ensemble selection where the analyst is interested in optimizing arbitrary performance metrics. We also, based on work and findings by dos Santos (2012) and dos Santos,

Sabourin, and Maupin (2008), include three ensemble selection strategies based on single-criterion and multi-criteria evolutionary optimization. In (dos Santos, 2012), it was found that for single-objective ensemble selection, genetic algorithms and particle swarm optimization outperformed other methods to maximize accuracy, and that for multi-criteria ensemble selection focusing on maximizing accuracy and diversity, NSGA-2 outperformed NSGA and controlled elitist NSGA. Several measures for ensemble diversity were compared, and ambiguity was found to provide better results. Based on these findings, we implement three benchmarks: (i) ensemble selection based on genetic algorithms, optimizing cost space through a minimization of AUCC or AUBC (*ES-GA*); (ii) a similar approach but based on optimization through particle swarm optimization (*ES-PSO*) and (iii) multicriteria ensemble selection using NSGA-II that optimizes cost space (AUCC/AUBC) and ensemble diversity (ambiguity) simultaneously (*MGA-ES*). Finally, we adopt the SVM parameter optimization and model selection approach by Chatelain et al. (2010) (*MGA-SVM-CS*).

### *4.4 Experimental Settings*

All experimental results are reported over a 10-fold cross-validation. In each fold, the 9 data parts not used for testing are further split evenly into a training sample, and a validation sample. In ensemble selection algorithms, it is common practice to select models on a data sample that was not involved in the training of the models in the model library (Caruana et al., 2004). Moreover, the availability of a validation sample allows for an optimization of model in terms of an arbitrarily chosen performance metric. Hence, all ensemble selection algorithms involved in the empirical benchmarking deploy the validation sample for model selection purposes, while for other algorithms, the validation sample is used for an exhaustive search for the best hyperparameter configuration.

Ensemble selection algorithms require the creation of a model library. For the empirical validation of the framework, both algorithms and their hyperparameters are varied to estimate multiple models that are commonly available in data analytics software environments and have been used in BFP before. Appendix C shows the algorithms that are included, the varied hyperparameters, and the value ranges over which hyperparameters are varied. Note that five model categories are included: homogeneous

ensemble learners, decision trees, data mining algorithms, statistical methods, and finally a number of conventional cost-sensitive algorithms.

CSMES optimizes cost space through NSGA-II configured for real-coded chromosomes, population sizes of 100 individuals and termination after 100 generations. This configuration also applies to the benchmarks based on evolutionary algorithms (PSO-ES, GA-ES, MGA-ES and MGA-SVM-CS). As no cost information is assumed known in this paper, CISVM focuses on a cost ratio range of 1 to 20, consistent with the range from which cost ratios are sampled for EMC calculations. Analogously, MiniMax is configured to focus on the same range and optimizes for 5 cost ratios: 1, 5, 10, 15 and 20. CISVM is based on a radial basis function kernel and its two parameters gamma and the regularization parameter *C*, are optimized through grid search. Depending on the cost space framework, the best model is selected in terms of AUCC or AUBC performance on the validation sample. Their values ranges are chosen identical to the ones for the SVM models in the model library (see Appendix C). The number of iterations in MiniMax and RiskBoost is set to 100.

Statistical comparisons of CSMES and benchmark algorithms are accomplished by the Friedman non-parametric anova (Friedman, 1937). This approach was recommended by Demšar (2006), and has subsequently been adopted in several studies that compare multiple classifiers across multiple data sets (e.g. Lessmann et al., 2015). The test ranks methods for every dataset using a metric of choice and uses the average ranks to determine whether they differ significantly. Pairwise post-hoc tests can be administered using the following test statistic for comparing algorithms *i* and *j*

$$Z = \frac{(R_i - R_j)}{\sqrt{\frac{k(k+1)}{6d}}} \qquad (6)$$

Where $k$ is the number of methods, $d$ the number of datasets and $R_j$ denotes the average rank of algorithm *j*. The probabilities associated with these statistics need to be corrected for family-wise error introduced by making multiple algorithm comparisons. In this study, Li's procedure (Li, 2008) is used to this end, as recommended by García et al. (2010).

CSMES, as well as all benchmark algorithms presented in Section 4.3 are implemented in R (R Core Team, 2019). Functions to implement CSMES are made publicly available in the new CSMES R package (De Bock, Coussement, & Lessmann, 2020) accessible via http://cran.r-project.org.

# 5 Results

This section first presents the results of an experimental comparison of CSMES to other algorithms in three scenarios: high, partial, and no cost uncertainty in BFP in terms of classification performance. Top-level Friedman test results are included in Appendix E. All global tests indicate the presence of significantly different performance levels between the compared algorithms and therefore, post-hoc comparisons that are discussed in detail in subsections 5.1, 5.2 and 5.3 are justified. A fourth subsection (Section 5.4) discusses a comparison of CSMES to other algorithms in terms of computational costs.

## *5.1 High Cost Uncertainty During Scoring Phase*

A first set of comparisons is made in the scenario of high cost uncertainty. This scenario implies that no cost ratio information is known during the prediction or soring phase. As outlined before, the ensemble nomination in CSMES is here based on the lowest area under the cost or Brier curve, depending on the cost space framework adopted. The first comparison is made between CSMES and alternative algorithms designed for cost-sensitive learning under cost uncertainty: CISVM, RiskBoost and MiniMax.

Table 4 presents average ranks and adjusted p-values of post-hoc pairwise comparison test results based on Li's procedure for the comparison of CSMES to CISVM, RiskBoost and MiniMax. A first comparison involves the generalized performance in cost space. For the cost curve-based model selection and evaluation, this is measured as the area under the cost curve (AUCC), while the area under the Brier curve (AUBC) is the equivalent for a Brier curve-based evaluation. CSMES outperforms all three benchmark methods in terms of AUCC. In terms of AUBC, CSMES outperforms MiniMax, but is, despite a lower average rank, not found to significantly outperform CISVM or RiskBoost.

| **Cost space paradigm** | **Evaluation metric** | | ***Algorithm*** | | | |
|---|---|---|---|---|---|---|
| | | | ***CISVM*** | ***RiskBoost*** | ***MiniMax*** | ***CSMES*** |
| *Cost curve* | AUCC | Avg. rank | 2.3801 | 2.7143 | 3.9048 | 1 |
| | | Adj. p-value | 0.0005*** | 0.000*** | 0.0000*** | |

| | | | | | | |
|---|---|---|---|---|---|---|
| | EMC | Avg. rank | 2.0476 | 2.9214 | 3.3333 | 1.6976 |
| | | Adj. p-value | 0.0000*** | 0.0000*** | 0.0000*** | |
| *Brier curve* | AUBC | Avg. rank | 2.2857 | 2.1905 | 3.8571 | 1.6667 |
| | | Adj. p-value | 0.1886 | 0.1202 | 0.0000*** | |
| | EMC | Avg. rank | 2.1857 | 2.9024 | 3.2881 | 1.6238 |
| | | Adj. p-value | 0.0000*** | 0.0000*** | 0.0000*** | |

**Table 4: Cost-uncertainty accommodating cost-sensitive benchmark results when cost ratios are unknown at scoring time (high cost uncertainty). '***' indicates a significant result at α=0.001.**

A second comparison considers simulated cost ratios and compares models based on the expected misclassification cost for these cost ratios. Here, it is clear that CSMES outperforms all three benchmarks, regardless of the cost space paradigm chosen. This latter result indicates that CSMES, where an optimal ensemble is nominated by evaluating overall cost space performance, minimizes misclassification costs effectively, even when there is no information about the misclassification cost ratio that applies. Based on average ranks, CISVM is the closest competitor.

Next, CSMES is compared to alternative ensemble and ensemble selection strategies that depend on the same heterogeneous model library. Given the increased complexity of CSMES in comparison to other ensemble selection strategies (most notably, the ensemble nomination step), it is important to verify whether alternative, simpler strategies for ensemble selection in cost space could match or even surpass the performance of CSMES. Table 5 presents the post-hoc results of this comparison.

| **Cost space paradigm** | **Evaluation metric** | | **Algorithm** | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | ***Full*** | ***Weighted*** | ***Best*** | ***Top10*** | ***Top25*** | ***CSMES*** |
| *Cost curve* | AUCC | Avg. rank | 2.2836 | 3.1905 | 4.0476 | 4.7142 | 5.0952 | 1.6667 |
| | | Adj. p-value | 0.2836 | 0.0103** | 0.0000*** | 0.0000*** | 0.0000*** | |
| | EMC | Avg. rank | 3.5119 | 3.9157 | 3.3024 | 3.6386 | 3.6886 | 2.9429 |
| | | Adj. p-value | 0.0000*** | 0.0000*** | 0.0000*** | 0.0000*** | 0.0000*** | |
| *Brier curve* | AUBC | Avg. rank | 3.4762 | 3.6190 | 3.6667 | 3.9528 | 4.8571 | 1.4285 |
| | | Adj. p-value | 0.0004*** | 0.0002*** | 0.0002*** | 0.0000*** | 0.0000*** | |
| | EMC | Avg. rank | 3.4438 | 3.9038 | 3.3171 | 3.5547 | 3.6505 | 3.13 |
| | | Adj. p-value | 0.0000*** | 0.0000*** | 0.0005*** | 0.0000*** | 0.0000*** | |

**Table 5: Ensemble selection benchmark results when cost ratios are unknown at scoring phase (high cost uncertainty). '**' indicates a significant result at α=0.05, while '***' indicates a significant result at α=0.001.**

The following conclusions emerge. First, in a cost-curve based evaluation, CSMES significantly outperforms basic ensemble and classifier selection strategies. The only exception is the full library ensemble which does not perform significantly worse in comparison to CSMES. Second, in a Brier curve-based evaluation, overall, CSMES dominates all other approaches in terms of global performance (AUBC). Third, when evaluating expected misclassification costs for specific cost ratios, the dominance

of CSMES becomes more pronounced, since it significantly outperforms all other approaches, both for the cost curve and the Brier curve.

| Cost space paradigm | Evaluation metric | | Algorithm | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | ***GA-ES*** | ***PSO-ES*** | ***GHS-ES*** | ***MGA-ES*** | ***MGA-SVM-CS*** | ***CSMES*** |
| *Cost curve* | AUCC | Avg. rank | 2.9524 | 2.9524 | 3.7619 | 3.4762 | 5.4762 | 2.3810 |
| | | Adj. p-value | 0.3223 | 0.3223 | 0.2414 | 0.7862 | 0.0000*** | |
| | EMC | Avg. rank | 3.0571 | 3.3752 | 3.2324 | 3.1776 | 4.6152 | 3.0571 |
| | | Adj. p-value | 0.0000*** | 0.0000*** | 0.0024** | 0.0369** | 0.0000*** | |
| *Brier curve* | AUBC | Avg. rank | 2.5714 | 3.2381 | 3.8571 | 1.7143 | 6 | 3.6190 |
| | | Adj. p-value | 0.1133 | 0.5894 | 0.6801 | 0.0024** | 0.0002*** | |
| | EMC | Avg. rank | 3.3067 | 3.4710 | 3.8981 | 3.1814 | 4.3276 | 2.8152 |
| | | Adj. p-value | 0.0000*** | 0.0000*** | 0.0000*** | 0.0000*** | 0.0000*** | |

**Table 6: Optimization-based ensemble and classifier selection benchmark results when cost ratios are unknown at scoring phase (high cost uncertainty). '**' indicates a significant result at α=0.05, while '***' indicates a significant result at α=0.001.**

Finally, Table 6 summarizes post-hoc comparison test results for CSMES and alternative optimization-based ensemble and classifier selection approaches. In terms of AUCC and AUBC, CSMES consistently outperforms *MGA-SVM-CS*, a benchmark explicitly designed for accommodating cost uncertainty. Multi-criteria ensemble selection simultaneously minimizing AUCC/AUBC and maximizing diversity (MGA-ES) is outperformed, while single-criterion optimization-based ensemble selection based on genetic algorithms, particle swarm optimization and greedy hillclimb search (GA-ES, PSO-ES and GHS-ES) are not outperformed. Similar results are found for brier curve-based evaluations. In terms of expected misclassification costs, the dominance of CSMES is evident: all benchmark algorithms are significantly outperformed; both for cost curve and brier-curve model evaluations.

In summary, since we believe a cost-ratio based evaluation corresponds closer to reality than theoretical measures such as AUCC and AUBC, it is fair to conclude that in a scenario of high cost uncertainty, CSMES proves its value and demonstrates that a multi-criterion optimization of cost space and an ensemble nomination based on generalized cost space performance is a solid strategy to tackle cost uncertainty when a model for business failure prediction is consulted to generate predictions.

### *5.2 Partial Cost Uncertainty During Scoring Phase*

The second scenario assumes a partial resolving of cost uncertainty at the stage where predictions are required. As explained earlier, this translates to the knowledge of a probability distribution over the

range of possible operating conditions. In CSMES, ensemble nomination is then based upon the performance of the candidate ensembles in a part of cost space, i.e. based on pAUCC or pAUBC. CSMES is again evaluated through a comparison to other cost uncertainty-accommodating methods (Table 7), alternative heuristic ensemble selection approaches (Table 8) and optimization-based ensemble and classifier selection approaches (Table 9). In the comparison to CISVM, RiskBoost and MiniMax, CSMES now consistently outperforms all benchmarks in a highly significant manner. In first instance, this applies to the theoretical measures of partial AUCC (pAUCC) and partial AUBC (pAUBC) that indicate how well methods perform given the distribution over cost space for which they also have been trained. Moreover, an evaluation for specific simulated cost ratios that have been randomly drawn from the probability distribution over operating conditions demonstrates that CSMES extends this superiority to a more specific and realistic, cost-ratio based comparison. A comparison to alternative ensemble and ensemble selection approaches leads to identical findings as in the scenario of full cost uncertainty. In terms of pAUCC, the full library ensemble, and the single-criterion ensemble selection based on genetic algorithms constitute strong competitors and do not perform significantly worse in comparison to CSMES. In a Brier curve-based evaluation, however, all alternative ensemble approaches result in worse performance. The comparison with respect to specific misclassification cost ratios is clearly in favor of CSMES. None of the alternative approaches matches the performance level of CSMES in terms of expected misclassification cost.

| **Cost space paradigm** | **Evaluation metric** | | ***Algorithm*** | | | |
|---|---|---|---|---|---|---|
| | | | ***CISVM*** | ***RiskBoost*** | ***MiniMax*** | ***CSMES*** |
| *Cost curve* | pAUCC | Avg. rank | 2.5833 | 2.4762 | 3.8786 | 1.0620 |
| | | Adj. p-value | 0.0000*** | 0.0000*** | 0.0000*** | |
| | EMC | Avg. rank | 2.0405 | 2.9143 | 3.3333 | 1.7119 |
| | | Adj. p-value | 0.0002*** | 0.0000*** | 0.0000*** | |
| *Brier curve* | pAUBC | Avg. rank | 2.5083 | 2.825 | 3.3512 | 1.3155 |
| | | Adj. p-value | 0.0000*** | 0.0000*** | 0.0000*** | |
| | EMC | Avg. rank | 2.1905 | 2.9 | 3.2857 | 1.6238 |
| | | Adj. p-value | 0.0000*** | 0.0000*** | 0.0000*** | |

**Table 7: Cost-uncertainty accommodating cost-sensitive benchmark results when cost ratios are partially known at scoring time (partial cost uncertainty). '***' indicates a significant result at $\alpha$=0.001.**

| Cost space paradigm | Evaluation metric | | Algorithm | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | ***Full*** | ***Weighted*** | ***Best*** | ***Top10*** | ***Top25*** | ***CSMES*** |
| *Cost curve* | pAUCC | Avg. rank | 3.4962 | 3.9107 | 3.2814 | 3.6164 | 3.6774 | 3.0179 |
| | | Adj. p-value | 0.0000*** | 0.0000*** | 0.0000*** | 0.0000*** | 0.0000*** | |
| | EMC | Avg. rank | 3.9107 | 3.4962 | 3.2814 | 3.6164 | 3.6774 | 3.0179 |
| | | Adj. p-value | 0.0000*** | 0.0000*** | 0.0000*** | 0.0000*** | 0.0000*** | |
| *Brier curve* | pAUBC | Avg. rank | 3.4669 | 3.9240 | 3.3086 | 3.5419 | 3.6505 | 3.1081 |
| | | Adj. p-value | 0.0000*** | 0.0000*** | 0.0005*** | 0.0000*** | 0.0000*** | |
| | EMC | Avg. rank | 3.4669 | 3.9240 | 3.3086 | 3.5419 | 3.6505 | 3.1081 |
| | | Adj. p-value | 0.0000*** | 0.0000*** | 0.0005*** | 0.0000*** | 0.0000*** | |

**Table 8: Ensemble selection benchmark results when cost ratios are partially known at scoring phase (partial cost uncertainty). '***' indicates a significant result at α=0.001.**

| Cost space paradigm | Evaluation metric | | Algorithm | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | ***GA-ES*** | ***PSO-ES*** | ***GHS-ES*** | ***MGA-ES*** | ***MGA-SVM-CS*** | ***CSMES*** |
| *Cost curve* | pAUCC | Avg. rank | 3.1176 | 3.1302 | 3.3905 | 3.2974 | 4.9143 | 3.15 |
| | | Adj. p-value | 0.6821 | 0.7321 | 0.0001*** | 0.0383** | 0.0000*** | |
| | EMC | Avg. rank | 3.5 | 3.3593 | 3.2098 | 3.1745 | 4.6114 | 3.145 |
| | | Adj. p-value | 0.0000*** | 0.0005*** | 0.4013 | 0.6091 | 0.0000*** | |
| *Brier curve* | pAUBC | Avg. rank | 2.415 | 2.9969 | 4.7031 | 3.0769 | 4.8645 | 2.9436 |
| | | Adj. p-value | 0.0000*** | 0.3556 | 0.0000*** | 0.0314** | 0.0000*** | |
| | EMC | Avg. rank | 3.2733 | 3 .4726 | 3.8848 | 3.1493 | 4.3386 | 2.8814 |
| | | Adj. p-value | 0.0000*** | 0.0000*** | 0.0000*** | 0.0000*** | 0.0000*** | |

**Table 9: Optimization-based ensemble and classifier selection benchmark results when cost ratios are partially known at scoring phase (partial cost uncertainty). '**' indicates a significant result at α=0.05, while '***' indicates a significant result at α=0.001.**

Finally, the comparison of CSMES to the more competitive optimization-based ensemble selection and classifier selection benchmarks (Table 9) shows a more nuanced result. On the one hand, in terms of partial AUCC and partial AUBC, CSMES outperforms both benchmarks based on multicriteria optimization (MGA-ES and MGA-SVM-CS) as well as ensemble selection based on greedy hillclimb search (GHS-ES). However, no clear-cut advantage over GA-based and PSO-based ensemble selection is found. On the other hand, in terms of expected misclassification costs, CSMES is superior over all benchmarks, both for cost-curve-based and brier-curve-based evaluation.

### *5.3 No Cost Uncertainty: Known Cost Ratios During Scoring Phase*

The final analysis involves the scenario of no cost uncertainty, which corresponds to settings where the cost ratio is known when business failure predictions are required. Under this assumption, only one evaluation metric is reported which is expected misclassification cost. CSMES significantly outperforms CICSVM, RiskBoost and MiniMax. (see Table 10), and all alternative heuristic ensemble and ensemble selection strategies, on the other (see Table 11). ). This observation holds for both the cost curve and the Brier curve model evaluation paradigm. Compared to optimization-based ensemble

selection and classifier selection approaches, CSMES consistently performs at least as well as the benchmark algorithms, and ,with the exception of ensemble selection based on greedy hillclimb searching minimizing AUCC, it outperforms all benchmark algorithms (Table 12). Hence, when misclassification cost ratios are not known at the time of model training, but become known at the time of model scoring, the multicriteria optimization of cost space, in tandem with ensemble nomination in terms of the applicable operating condition puts CSMES at a significant advantage as a method for performing heterogeneous ensemble selection.

| **Cost space paradigm** | **Evaluation metric** | | ***Algorithm*** | | | |
|---|---|---|---|---|---|---|
| | | | ***CISVM*** | ***RiskBoost*** | ***MiniMax*** | ***CSMES*** |
| *Cost curve* | EMC | Avg. rank | 2.0666 | 2.9071 | 3.3310 | 1.6952 |
| | | Adj. p-value | 0.0000*** | 0.0000*** | 0.0000*** | |
| *Brier curve* | EMC | Avg. rank | 2.1881 | 2.9 | 3.2857 | 1.6262 |
| | | Adj. p-value | 0.0000*** | 0.0000*** | 0.0000*** | |

**Table 10: Cost-uncertainty accommodating cost-sensitive benchmark results when cost ratios are known at scoring time (no cost uncertainty). '***' indicates a significant result at α=0.001.**

| **Cost space paradigm** | **Evaluation metric** | | **Algorithm** | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | ***Full*** | ***Weighted*** | ***Best*** | ***Top10*** | ***Top25*** | ***CSMES*** |
| *Cost curve* | EMC | Avg. rank | 3.4855 | 3.915 | 3.2881 | 3.6267 | 3.6826 | 3.0021 |
| | | Adj. p-value | 0.0000*** | 0.0000*** | 0.0005*** | 0.0000*** | 0.0000*** | |
| *Brier curve* | EMC | Avg. rank | 3.4660 | 3.9181 | 3.3105 | 3.5417 | 3.6481 | 3.1157 |
| | | Adj. p-value | 0.0000*** | 0.0000*** | 0.0007*** | 0.0000*** | 0.0000*** | |

**Table 11: Ensemble selection benchmark results when cost ratios are known at scoring phase (no cost uncertainty). '***' indicates a significant result at α=0.001.**

| **Cost space paradigm** | **Evaluation metric** | | **Algorithm** | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | ***GA-ES*** | ***PSO-ES*** | ***GHS-ES*** | ***MGA-ES*** | ***MGA-SVM-CS*** | ***CSMES*** |
| *Cost curve* | EMC | Avg. rank | 3.5052 | 3.3529 | 3.2140 | 3.1755 | 4.6145 | 3.1378 |
| | | Adj. p-value | 0.0000*** | 0.0004*** | 0.2781 | 0.5147 | 0.0000*** | |
| *Brier curve* | EMC | Avg. rank | 3.2504 | 3.4695 | 3.8814 | 3.1445 | 4.3407 | 2.9133 |
| | | Adj. p-value | 0.0000*** | 0.0000*** | 0.0000*** | 0.0000*** | 0.0000*** | |

**Table 12: Optimization-based ensemble and classifier selection benchmark results when cost ratios are known at scoring phase (no cost uncertainty). '**' indicates a significant result at α=0.05, while '***' indicates a significant result at α=0.001.**

### *5.4 Computational cost analysis*

The results discussed above demonstrate that CSMES consistently outperforms the non-ensemble methods considered in this study (i.e., CISVM and MiniMax) in terms of cost-sensitive performance metrics. However, ensemble selection depends on the training of a sizeable heterogeneous library of models and CSMES adds the component of multicriteria optimization which could result in an increased computational cost. To compare CSMES to the other methods in function of computational effort, an

analysis was conducted in terms of training phase runtimes. The setup of this comparison, as well as the full results are presented in Appendix F.

In summary, these results show that CSMES performs comparably to MGA-SVM-CS but is clearly outperformed by standalone algorithms designed for cost uncertainty: CISVM, MiniMax and RiskBoost. However, CSMES' dependence on multicriteria optimization and the derivation of an ensemble nomination curve does not result in a disadvantage in terms of training phase runtimes in comparison to alternative ensemble and classifier selection approaches. All ensemble selection benchmark algorithms that depend on optimization (GA-ES, PSO-ES, GHS-ES and MGA-ES) are significantly outperformed by CSMES. An explanation of this result is found in the nature of the metrics that are optimized by CSMES. False negative and false positive rates are less computationally expensive in comparison to AUCC and AUBS that are minimized in these four benchmarks.

## 6 Conclusion, Limitations and Directions for Future Research

There is growing agreement that the evaluation of BFP models should accommodate asymmetric misclassification costs. This has inspired researchers to adopt alternative evaluation metrics and indicates the relevance of cost-sensitive learning algorithms, designed to involve cost information during the model training stage. Unfortunately, the assumption that costs or cost ratios are known, or can be reliably estimated prior to model estimation, is often not realistic. This leads to a non-trivial challenge: a need for models that can be trained when cost information is not available, yet are more cost-conscious than existing algorithms in an attempt to reduce misclassification costs when these models are actually deployed.

This study proposes a novel method for heterogeneous ensemble selection that assume an absence of cost information during model training, and various degrees of remaining cost uncertainty during model scoring. Ensemble selection prescribes an informed selection of members from a library of models that originate from various algorithm classes, and usually involves an optimization towards one or more performance criteria. The approach presented in this study deploys multicriteria optimization through NSGA-II to optimize cost space and obtain a set of Pareto-equivalent ensemble candidates that

each represent a different trade-off of error types. The concept of an ensemble nomination front is introduced, which maps competence regions of candidate ensembles in cost space. This nomination front is an instrument that allows an analyst, when predictions are due, to select an appropriate ensemble model with respect to the cost ratio information that is known at that time, or any cost uncertainty that remains. Three ensemble nomination strategies are suggested for three scenarios of cost uncertainty at the time of model scoring.

An extensive experimental benchmark validates the presented framework on 21 datasets representing companies in various sectors and countries, and results are analyzed for three scenarios of cost uncertainty at the models' scoring phase. A comparison is made to three sets of benchmark methods. A first set involves alternative methods that have been introduces in literature for dealing with cost uncertainty. A second set of benchmarks is formed by alternative ensemble and ensemble selection approaches that are based on the same model library, and represent less complex strategies to optimize cost space. A third benchmark selection consists of ensemble selection and classifier selection approaches that, like CSMES, depend on optimization algorithms. The results demonstrate that our method outperforms all benchmark algorithms that have been explicitly proposed in literature to deal with cost uncertainty. Furthermore, CSMES performs competitively on metrics that reflect generalized performance in cost space overall. The results also show that CSMES outperforms all alternative ensemble and classifier selection approaches in terms of misclassification cost, based on specific misclassification cost ratios. Finally, while CSMES is outperformed by benchmarks that do not depend on the training of a heterogeneous model library in terms of computational cost, it does demonstrate shorter training runtimes in comparison to alternative approaches for ensemble selection based on optimization. In the light of these results we believe the presented method is a valuable contribution to the BFP domain since it is the first to address cost uncertainty through an integrated ensemble selection framework.

Several limitations can be identified relating to the presented approach and its empirical validation. First, the study does not provide recommendations on the choice between an ensemble nomination based on the cost curve versus the Brier curve. Since both frameworks have advantages and

disadvantages, we opted to accommodate both in our method to allow analysts to make a choice in terms of preferences and project requirements. Second, the presented method and its validation rely on the assumption that models and ensembles are fully trained at the training stage, that is, they are not fine-tuned or selected at the stage of model scoring. While this would substantially increase the computational cost, in some situations it could be feasible to postpone model training, ensemble selection, or a part thereof, to the moment when predictions are due. In that case, model training and ensemble selection could benefit from updated cost ratio information. Finally, the method has been validated for BFP, which is one application in business analytics where cost-sensitive model evaluation and cost uncertainty are relevant. Future research should validate the method in other domains.

## 7 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.

Adams, N. M., & Hand, D. J. (1999). Comparing classifiers when the misallocation costs are uncertain. *Pattern Recognition, 32*(7), 1139-1147.

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

Bakker, B., & Heskes, T. (2003). Clustering ensembles of neural network models. *Neural networks, 16*(2), 261-269.

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.

Bauer, J., & Agarwal, V. (2014). Are hazard models superior to traditional bankruptcy prediction approaches? A comprehensive test. *Journal of Banking & Finance, 40*, 432-442.

Bodnar, D. (2019, 17 October 2019). Insolvencies Are on the Rise in Western Europe, *CFO.com*.

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.

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

Breiman, L., Friedman, J. H., Olsen, R. A., & Stone, C. J. (1984). *Classification and regression trees*: Chapman & Hall / CRC.

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

Britto Jr, A. S., Sabourin, R., & Oliveira, L. E. (2014). Dynamic selection of classifiers—a comprehensive review. *Pattern Recognition, 47*(11), 3665-3680.

Caruana, R., Munson, A., & Niculescu-Mizil, A. (2006). Getting the Most Out of Ensemble Selection. In *Proceedings of the 6th International Conference on Data Mining (ICDM '06)*, (pp. 828-833) IEEE Computer Society.

Caruana, R., Niculescu-Mizil, A., Crew, G., & Ksikes, A. (2004). Ensemble selection from libraries of models. In *Proceedings of the 21st International Conference on Machine Learning*, (pp. 18-27) ACM.

Chatelain, C., Adam, S., Lecourtier, Y., Heutte, L., & Paquet, T. (2010). A multi-model selection framework for unknown and/or evolutive misclassification cost problems. *Pattern Recognition, 43*(3), 815-823.

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

Chen, A., Chen, N., & Ribeiro, B. (2015). Comparative study of classifier ensembles for cost-sensitive credit risk assessment. *Intelligent Data Analysis, 19*(1), 127-144.

Chen, C., Liaw, A., & Breiman, L. (2004). Using random forests to learn imbalanced data *Technical Report 666.*: Statistics Department of University of California, Berkeley.

Chen, N., & Ribeiro, B. (2013). A Consensus Approach for Combining Multiple Classifiers in Cost-Sensitive Bankruptcy Prediction. In *Proceedings of the International Conference on Adaptive and Natural Computing Algorithms (ICANNGA 2013)*, (pp. Springer-Verlag.

Cheng, F., Fu, G., Zhang, X., & Qiu, J. (2019). Multi-objective evolutionary algorithm for optimizing the partial area under the ROC curve. *Knowledge-Based Systems, 170*, 61-69.

Creditreform. (2014). Unternehmensinsolvenzen in Europa - Jahr 2013/14: Creditreform Wirtschaftsforschung.

Croux, C., Joossens, K., & Lemmens, A. (2007). Trimmed bagging. *Computational Statistics & Data Analysis, 52*(1), 362-368.

Davalos, S., Leng, F., Feroz, E. H., & Cao, Z. (2014). Designing an if-then rules-based ensemble of heterogeneous bankruptcy classifiers: a genetic algorithm approach. *Intelligent Systems in Accounting, Finance and Management, 21*(3), 129-153.

De Bock, K. W., Coussement, K., & Lessmann, S. (2020). CSMES: Cost-Sensitive Multi-Criteria Ensemble Selection and Other Classifiers for Cost-Sensitive Learning under Unknown Cost Conditions (R package version 1.0). Retrieved from https://CRAN.R-project.org/package=CSMES

Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation, 6*(2), 182-197.

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

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.

Domingos, P. (1999). Metacost: A general method for making classifiers cost-sensitive. In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 155-164) ACM.

dos Santos, E. M. (2012). Evolutionary algorithms applied to classifier ensemble selection. In *Proceedings of the 44th Brazilian Operations Research Symposium/16th Latin Ibero American Conference on Operations Research*, (pp.

dos Santos, E. M., Sabourin, R., & Maupin, P. (2008). A dynamic overproduce-and-choose strategy for the selection of classifier ensembles. *Pattern Recognition, 41*(10), 2993-3009.

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.

Doumpos, M., & Zopounidis, C. (2007). Model combination for credit risk assessment: A stacked generalization approach. *Annals of Operations Research, 151*(1), 289-306.

Drummond, C., & Holte, R. (2006). Cost curves: An improved method for visualizing classifier performance. *Machine Learning, 65*(1), 95-130.

Ekinci, A., & Erdal, H. İ. (2017). Forecasting Bank Failure: Base Learners, Ensembles and Hybrid Ensembles. *Computational Economics, 49*(4), 677-686.

Fan, W., Stolfo, S. J., Zhang, J., & Chan, P. K. (1999). AdaCost: Misclassification cost-sensitive boosting. In *Proceedings of the Sixteenth International Conference on Machine Learning (ICML '99)*, (pp. 97-105) Morgan Kauffman.

Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. In *Proceedings of the Thirteenth International Conference on Machine Learning (ICML 1996)*, (pp. 148–156). Bari, Italy: Morgan Kauffman.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis, 38*(4), 367-378.

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.

García, S., Fernández, A., Luengo, J., & Herrera, F. (2010). Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. *Information Sciences, 180*(10), 2044-2064.

Hernández-Orallo, J., Flach, P., & Ferri, C. (2012). A unified view of performance metrics: translating threshold choice into expected classification loss. *Journal of Machine Learning Research, 13*(Oct), 2813-2869.

Hernández-Orallo, J., Flach, P. A., & Ramirez, C. F. (2011). Brier Curves: a New Cost-Based Visualisation of Classifier Performance. In *Proceedings of the 28th International Conference on International Conference on Machine Learning (ICML'11)*, (pp. 585-592).

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(8), 832-844.

Hua, Z., Wang, Y., Xu, X., Zhang, B., & Liang, L. (2007). Predicting corporate financial distress based on integration of support vector machine and logistic regression. *Expert Systems with Applications, 33*(2), 434-440.

Jo, H., Han, I., & Lee, H. (1997). Bankruptcy prediction using case-based reasoning, neural networks, and discriminant analysis. *Expert Systems with Applications, 13*(2), 97-108.

Johnson, R. A., Raeder, T., & Chawla, N. V. (2015). Optimizing Classifiers for Hypothetical Scenarios. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2015)*, (pp. 264-276) Springer International Publishing.

Kim, M. H., & Yoo, P. D. (2006). A Semiparametric Model Approach to Financial Bankruptcy Prediction. In *Proceedings of the 2006 IEEE International Conference on Engineering of Intelligent Systems*, (pp. 1-6).

Kirkos, E. (2012). Assessing methodologies for intelligent bankruptcy prediction. *Artificial Intelligence Review*, 1-41.

Ko, A. H. R., Sabourin, R., & Britto, J. A. S. (2008). From dynamic classifier selection to dynamic ensemble selection. *Pattern Recognition, 41*(5), 1718-1731.

Kolay, M., Lemmon, M., & Tashjian, E. (2016). Spreading the Misery? Sources of Bankruptcy Spillover in the Supply Chain. *Journal of Financial and Quantitative Analysis, 51*(6), 1955-1990.

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., & Rodriguez, J. J. (2007). An experimental study on rotation forest ensembles. In *Proceedings of the International Workshop on Multiple Classifier Systems (MCS 2007)*, (pp. 459-468). Prague, Czech Republic: Springer-Verlag Berlin.

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.

Levesque, J.-C., Durand, A., Gagne, C., & Sabourin, R. (2012). Multi-objective evolutionary optimization for generating ensembles of classifiers in the ROC space. In *Proceedings of the 14th annual conference on Genetic and evolutionary computation*, (pp. 879-886). Philadelphia, Pennsylvania, USA: ACM.

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.

Li, J. (2008). A two-step rejection procedure for testing multiple hypotheses. *Journal of Statistical Planning and Inference, 138*(6), 1521-1527.

Li, W., Ding, S., Chen, Y., & Yang, S. (2018). Heterogeneous ensemble for default prediction of peer-to-peer lending in China. *IEEE Access, 6*, 54396-54406.

Lin, F. Y., & McClean, S. (2001). A data mining approach to the prediction of corporate failure. *Knowledge-Based Systems, 14*(3), 189-195.

Liu, X.-Y., & Zhou, Z.-H. (2010). Learning with cost intervals. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 403-412). Washington, DC, USA: ACM.

Margineantu, D., & Dietterich, T. (1997). Pruning Adaptive Boosting. In *Proceedings of the Fourteenth International Conference on Machine Learning (ICML '97)*, (pp. 211-218) Morgan Kaufmann Publishers Inc.

Martin, D. (1977). Early Warning of Bank Failure: A Logit Regression Approach. *Journal of Banking and Finance, 1*(3), 249-276.

Martinez-Munoz, G., Hernandez-Lobato, D., & Suarez, A. (2009). An Analysis of Ensemble Pruning Techniques Based on Ordered Aggregation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31*(2), 245-259.

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).

Ohlson, & James, A. (1980). Financial Ratios and the Probabilistic Prediction of Bankruptcy. *Journal of Accounting Research, 18*(1), 109.

Olmeda, I., & Fernández, E. (1997). Hybrid Classifiers for Financial Multicriteria Decision Making: The Case of Bankruptcy Prediction. *Computational Economics, 10*(4), 317-335.
Özöğür-Akyüz, S., Windeatt, T., & Smith, R. (2015). Pruning of Error Correcting Output Codes by optimization of accuracy–diversity trade off. *Machine Learning, 101*(1), 253-269.
Papouskova, M., & Hajek, P. (2019). Two-stage consumer credit risk modelling using heterogeneous ensemble learning. *Decision Support Systems, 118*, 33-45.
Partalas, I., Tsoumakas, G., & Vlahavas, I. (2009). Pruning an ensemble of classifiers via reinforcement learning. *Neurocomputing, 72*(7), 1900-1909.
Pendharkar, P. (2008). Misclassification cost minimizing fitness functions for genetic algorithm-based artificial neural network classifiers. *Journal of the Operational Research Society, 60*(8), 1123-1134.
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., & Domingos, P. (2003). Tree induction for probability-based ranking. *Machine Learning, 52*(3), 199-215.
Provost, F., & Fawcett, T. (2001). Robust Classification for Imprecise Environments. *Machine Learning, 42*(3), 203-231.
Quinlan, R. (1993). *C4.5: Programs for Machine Learning*. San Mateo, CA.: Morgan Kauffman Publishers.
R Core Team. (2019). R: A Language and Environment for Statistical Computing. Vienna, Austria. Retrieved from https://www.R-project.org
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., & Li, H. (2008). Listed companies' financial distress prediction based on weighted majority voting combination of multiple classifiers. *Expert Systems with Applications, 35*(3), 818-827.
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.
Sylvester, J., & Chawla, N. V. (2006). Evolutionary ensemble creation and thinning. In *Proceedings of the International Joint Conference on Neural Networks, 2006. IJCNN'06*, (pp. 5148-5155) IEEE.
Tsai, C.-F. (2009). Feature selection in bankruptcy prediction. *Knowledge-Based Systems, 22*(2), 120-127.
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.
Viaene, S., & Dedene, G. (2005). Cost-sensitive learning and decision making revisited. *European Journal of Operational Research, 166*(1), 212-220.
Wang, R., & Tang, K. (2012). Minimax Classifier for Uncertain Costs. *arXiv, 1205.0406*.
Weiss, G. M. (2004). Mining with rarity: a unifying framework. *SIGKDD Explorations, 6*(1), 315-354.
Woloszynski, T., & Kurzynski, M. (2011). A probabilistic model of classifier competence for dynamic ensemble selection. *Pattern Recognition, 44*(10-11), 2656-2668.
Woloszynski, T., Kurzynski, M., Podsiadlo, P., & Stachowiak, G. W. (2012). A measure of competence based on random classification for dynamic ensemble selection. *Information Fusion, 13*(3), 207-213.
Xia, Y., Liu, C., Da, B., & Xie, F. (2018). A novel heterogeneous ensemble credit scoring model based on bstacking approach. *Expert Systems with Applications, 93*, 182-199.
Zadrozny, B., & Elkan, C. (2001). Learning and making decisions when costs and probabilities are both unknown. In *Proceedings of the Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 204-213). San Francisco, California: ACM.
Zadrozny, B., & Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the 8th ACM SIGKDD international conference on Knowledge discovery and data mining*, (pp. 694-699). Edmonton, Alberta, Canada: ACM.
Zhao, J., Basto Fernandes, V., Jiao, L., Yevseyeva, I., Maulana, A., Li, R., Bäck, T., Tang, K., & T.M. Emmerich, M. (2016). Multiobjective optimization of classifiers by means of 3D convex-hull-based evolutionary algorithms. *Information Sciences, 367-368*, 80-104.

Zhao, J., Jiao, L., Liu, F., Basto Fernandes, V., Yevseyeva, I., Xia, S., & T.M. Emmerich, M. (2018). 3D fast convex-hull-based evolutionary multiobjective optimization algorithm. *Applied Soft Computing*, *67*, 322-336.
Zhou, Z.-H. (2012). *Ensemble Methods: Foundations and Algorithms*: Chapman & Hall/CRC.

## Appendix A: NSGA-II algorithm

The pseudocode of the NSGA-II algorithm (adapted from (Deb et al., 2002)) is given by:

**Parameters:** population size $N$; number of generations $M$; $p$ fitness functions $F_i$; $i=1,...,p$; mutation probability $p_m$ and crossover probability $p_c$

Initialize population: randomly generate $N$ real-valued chromosomes into $P_0$
$t=0$
**While** ($t < M$ and stopping criteria not met) **do**
    Evaluate fitness of each population chromosome for each fitness functions $F_i$; $i=1,...,p$
    Apply *non-dominated sorting* to $P_t$ to determine *fitness rank* and front membership of chromosomes
    Calculate *crowding distance* for each member chromosome of $P_t$
    Create offspring population $Q_t$ of size N from $P_t$ by applying the following operators:
        - *Parent selection* through *binary tournament selection* based on fitness rank and crowding distance and the *crowding selection operator*
        - *Crossover* using crossover probability $p_c$
        - *Mutation* using mutation probability $p_m$
    Apply *non-dominated sorting* to $P_t \cup \boldsymbol{O}_t$ to determine front membership and assign non-domination level (rank) to each chromosome
    Calculate *crowding distance* for each member chromosome of $P_t \cup \boldsymbol{O}_t$
    Create population of generation by selecting N best chromosomes from $P_t \cup \boldsymbol{O}_t$ using and the crowding selection operator
    t=t+1
**End While**

The NSGA-II algorithm initializes by randomly generating a starting population $P_0$ containing $N$ chromosomes. In the iterative process that follows, offspring populations are derived from parent populations whereby selection favours parents that are fitter in terms of the $p$ objective functions and more diverse in comparison to others. Specifically, fitness is assessed through the process of *non-dominated sorting,* which assigns instances to a hierarchy of fronts of equally fit individuals and domination ranks are awarded accordingly as fitness ranks. Then, with the aim of enforcing spread, individuals on each front are evaluated in terms of *crowding distance* that quantifies dispersion in terms of how they score on the $p$ objective functions. Both fitness rank and crowding distance influence parent selection, which takes the form of binary tournament selection. Comparison between individuals involves use of a crowding selection operator which favours individuals with lower (i.e., better) fitness rank or in case of a fitness draw, individuals with higher crowding distance. Offspring

and parent populations are combined, non-dominated sorting is applied again and finally, the new generation is formed by selection the best *N* chromosomes.

## Appendix B: Overview and description of dataset variables

This appendix provides an overview of the variables included in the datasets used in the experimental validation in this study.

| Variable class | Variable label | Variable description | Availability | |
|---|---|---|---|---|
| | | | Datasets 1-14 (Fra,Ita) | Datasets 15-21 (Bel) |
| **1. Financial ratios** | | | | |
| Liquidity ratios | *Cash ratio t-i* | Cash ratio: cash and cash equivalent assets / total liabilities, at time *t-i* | ✓ [1] | ✓ [2] |
| | *Current ratio t-i* | Current ratio: current assets / current liabilities, at time *t-i* | ✓ [1] | ✓ [2] |
| | *NWC2TA ratio t-i* | Net working capital to total assets ratio: (current assets - current liabilities) / total assets, at time *t-i* | ✓ [1] | ✓ [2] |
| | *Quick ratio t-i* | Quick ratio: (current assets - inventories) / current liabilities, at time *t-i* | ✓ [1] | ✓ [2] |
| Long-term solvency ratios | *Debt ratio t-i* | Debt ratio: total liabilities / total assets, at time *t-i* | ✓ [1] | ✓ [2] |
| | *Debt2worth ratio t-i* | Debt to net worth ratio: total debt / (total assets - total liabilities), at time *t-i* | ✓ [1] | ✓ [2] |
| | *Solvency ratio t-i* | Solvency ratio: net profit after taxes / total liabilities, at time *t-i* | ✓ [1] | ✓ [2] |
| | *Times interest earned ratio t-i* | Times interest earned ratio: EBITDA / total financial charges, at time *t-i* | ✓ [1] | ✓ [2] |
| | *Avg. collection period ratio t-i* | Average collection period ratio: (average accounts receivable / sales revenue ) * 365, at time *t-i* | ✓ [1] | ✓ [2] |
| Turnover ratios | *Debtor turnover ratio t-i* | Debtor turnover ratio: net credit sales / average accounts receivable, at time *t-i* | ✓ [1] | ✓ [2] |
| | *Fixed-asset turnover t-i* | Fixed-asset turnover: sales / average net fixed assets, at time *t-i* | ✓ [1] | ✓ [2] |
| | *Inventory turnover t-i* | Inventory turnover: cost of goods sold / average inventory, at time *t-i* | ✓ [1] | ✓ [2] |
| | *Asset turnover t-i* | Asset turnover: net sales revenue / average total assets, at time *t-i* | ✓ [1] | ✓ [2] |
| | *Profit margin t-i* | Profit margin: profit after tax / revenue, at time *t-i* | ✓ [1] | ✓ [2] |
| | *ROA t-i* | Return on assets (ROA): net income before tax / total assets, at time *t-i* | ✓ [1] | ✓ [2] |
| | *ROE t-i* | Return on equity (ROE): net income after tax / equity, at time *t-i* | ✓ [1] | ✓ [2] |
| | *ROI t-i* | Return on investment (ROI): net income after interest and tax / total assets, at time *t-i* | ✓ [1] | ✓ [2] |
| **2. Payment timeliness indicators** | *Social security dues t-i* | Amounts due to social security authority, at time *t-i* | | ✓ [2] |
| | *Tax dues t-i* | Amounts due to tax authority, at time *t-i* | | ✓ [2] |
| | *Nbr. protested bills [t-j;t]* | Number of protested bills in period *[t-j;t]* | | ✓ [2] |
| | *Nbr. summons [t-j;t]* | Number of social security summons in period *[t-j;t]* | | ✓ [2] |
| | *Overdue balance [t-j;t]* | Total current overdue balance in period *[t-j;t]* | | ✓ [2] |
| | *Pct. late payments [t-j;t]* | Percentage reported transactions with late payment in period *[t-j;t]* | | ✓ [2] |
| | *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]* | | ✓ [2] |
| **3. Firmo-graphics** | *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) | ☐ | ✓ |
| | *SIC bin* | Binned standard industry code (SIC 8) | ☐ | ✓ |
| | *Legal form code* | Legal form code | ☐ | ✓ |

**[1] $i$=0 for French and Italian companies (datasets 1 to 14); [2] $i \in \{0,1,2\}$ and $j \in \{1,2\}$ for Belgian companies (datasets 15 to 21).**

As can be seen in the table, in terms of variables the Belgian datasets (15-21) deviate from the Italian and French datasets (1-14) in three ways. First, they include two additional sets of variables:

variables that measure payment timeliness and firmographic variables. The former variables reflect how well and timely a company pays its amounts due to the tax authority, social security authority and selected suppliers while firmographics provide meta-information about the company (e.g. company age, industry category, legal form and number of employees) and its management. Second, most ratios and payment timeliness variables are calculated at different time points. Year count indices *i* and *j* are used to indicate at which moment in time, or for which time interval relative to time *t* 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. For example for Belgian dataset, time point *t* denotes the end of the independent variable collection period, i.e. May 31st 2008. As such, the variable *Current ratio t-1* provides current ratio calculated using the most recent information available on May 31st 2007, i.e. using annual account information for the year 2006. Similarly, a set of payment timeliness variables are measured over time intervals, dating either one or two years back prior to time point *t*. For example, the variable *Pct. late payments [t-2;t]* is the percentage of registered transactions for which payment was late, measure over a two-year period until May 31st, 2008. Finally, a different timeline was respected for the measurement of predictor and outcome variables, as shown in the following figure.

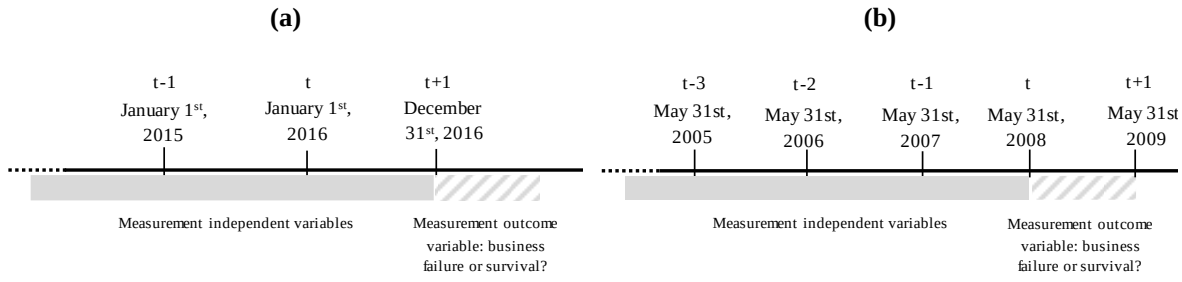



**Figure B.1: Data collection time lines**

Note that Figure (a) applies to data French and Italian companies (datasets 1-14) while Figure (b) applies to data for Belgian companies (datasets 15-21).

## Appendix C: Overview of model library algorithms, varied hyperparameters and values

| Method category | Algorithm | Varied hyperparameters and values |
|---|---|---|
| ***Homogeneous ensembles*** | Bagging (Breiman, 1996)<br>Trimmed bagging (Croux, Joossens, & Lemmens, 2007)<br>Stochastic gradient boosting (Friedman, 2002)<br>Rotation forest (Rodríguez, Kuncheva, & Alonso, 2006)<br>AdaBoost (Freund & Schapire, 1996)<br>Random subspace method (Ho, 1998)<br>Random forest (Breiman, 2001) | Ensemble size: 10,50,100<br>Random feature subset size: (5%,10%,25%,50%,75%)*#features, SQRT(#features) |
| ***Decision trees*** | CART (Breiman et al., 1984)<br>C4.5 (Quinlan, 1993)<br>C4.4 (Provost & Domingos, 2003) | Pruning = TRUE, FALSE<br>Minbucketsize = AUTO 2 4 6 8 10 20 40 60 80 100 250 500 750 1000 |
| ***Statistical models*** | Logistic regression<br>Linear discriminant analysis<br>Quadratic discriminant analysis | Variable selection: none, forward, backward, stepwise<br>variable selection entry & stay probabilities = 0.01, 0.05,0.1,0.15,...,0.95 |
| ***Data mining algorithms*** | Multi-layer perceptron | Number of hidden layers = 1,2,3,4,5,6,7,8,9,10 |
| | Support vector machines | Linear kernel: regularization parameter (soft margin constant C): 2**(-12,-6,0,6,12)<br><br>Radiant basis function kernel: regularization parameter (soft margin constant C): 2**(-12,-6,0,6,12) X gamma=2**(-13,-9,-6,-1) |
| | K-Nearest neighbours | Number of nearest neighbors :<br>1,5,10,50,100,150,200,300,400,500,600,700,800,900,1000,1500,2000,2500,3000,3500,4000 |
| ***Cost-sensitive classifiers*** | AdaCost (Fan et al., 1999)<br>Metacost - C4.5 (Domingos, 1999) | Ensemble size: 10,50,100<br>Cost ratio: 2,5,10 |
| | Cost-sensitive (weighted) random forest (Chen, Liaw, & Breiman, 2004) | Ensemble size: 10,50,100<br>Cost ratio: 2,5,10<br>Random feature subset size: (5%,10%,25%,50%,75%)*#features, SQRT(#features) |
| | Cost-senstive CART (Breiman et al., 1984) | Pruning = TRUE, FALSE<br>Minbucketsize = AUTO 2 4 6 8 10 20 40 60 80 100 250 500 750 1000<br>Cost ratio: 2,5,10 |

## Supplementary Materials

## Appendix D: Procedure for generation of operating points and operating point probability distributions

The procedure for simulation operating points (required for measuring expected misclassification costs) and probability density functions over operating points (required for measuring partial AUCC and partial AUBC performance estimations) is given by the following pseudocode:

Parameters: **number of datasets *d*, number of cross validations *ncv*, number of folds per cross-validation *nf*, number of evaluations per fold *ne*, $[\alpha_{min}, \alpha_{max}]$ is the desired range of cost ratios used for simulating operating points and $[sd_{min}, sd_{max}]$ is the desired range of standard deviations for probability distributions over operating points**

For **(*i* in 1 to *d*)** do
  For **(*j* in 1 to *ncv*)** do
    For **(*k* in 1 to *nf*)** do
      - **Determine failure and survival rates for the validation sample of dataset *i*, cross validation iteration *j* and fold *l*: $p_{i,,j,k,l}(-)$ and $p_{i,,j,k,l}(+)$**
      - **Determine $[PC_{i,j,k,l,min}(+), PC_{i,j,k,l,max}]$, the operating point range corresponding to $[\alpha_{min}, \alpha_{max}]$ using equation (2), $p_{i,,j,k,l}(-)$ and $p_{i,,j,k,l}(+)$.**
        - Illustrated through an example in Figure D.1 (a).

      For **(*l* in 1 to *ne*)** do
        - **Pick a random operating condition $PC_{i,j,k,l}(+)$ from interval $[PC_{i,j,k,l,min}(+), PC_{i,j,k,l,\max}(+)]$.**
          - Used for comparing algorithms in terms of expected misclassification cost (EMC) in evaluation scenarios 1 and 3
          - Illustrated in Figure D.1 (b).

        - **Simulate a Beta distribution** Beta$(a, b)$ as probability density function **for $PC(+)$.**
          - **Simulate a standard deviation value for $PC_{i,,j,k,l}(+)$ by generating a random number in the interval $[sd_{min}, sd_{max}]$ and denote it as $sd_{i,j,k,l}$**
          - **Determine $a_{i,j,k,l}$ and $b_{i,j,k,l}$, shape parameters of the beta distribution so that its mean is $PC_{i,j,k,l}(+)$ and its standard deviation is $sd_{i,j,k,l}$**
          - **Beta**$(a_{i,j,k,l}, b_{i,j,k,l})$ is used as probability density function over $PC(+)$ for comparing algorithms in terms of partial AUCC and partial AUBC in evaluation scenario 2.
          - Illustrated in Figure D.1 (c).

        - **Randomly draw an operating condition value $PC'_{i,j,k,l}(+)$ from** Beta$(a_{i,j,k,l}, b_{i,j,k,l})$.
          - Used for comparing algorithms in terms of expected misclassification cost (EMC) in evaluation scenario 2. Illustrated in Figure D.1 (d).

      Endfor
    Endfor
  Endfor
Endfor

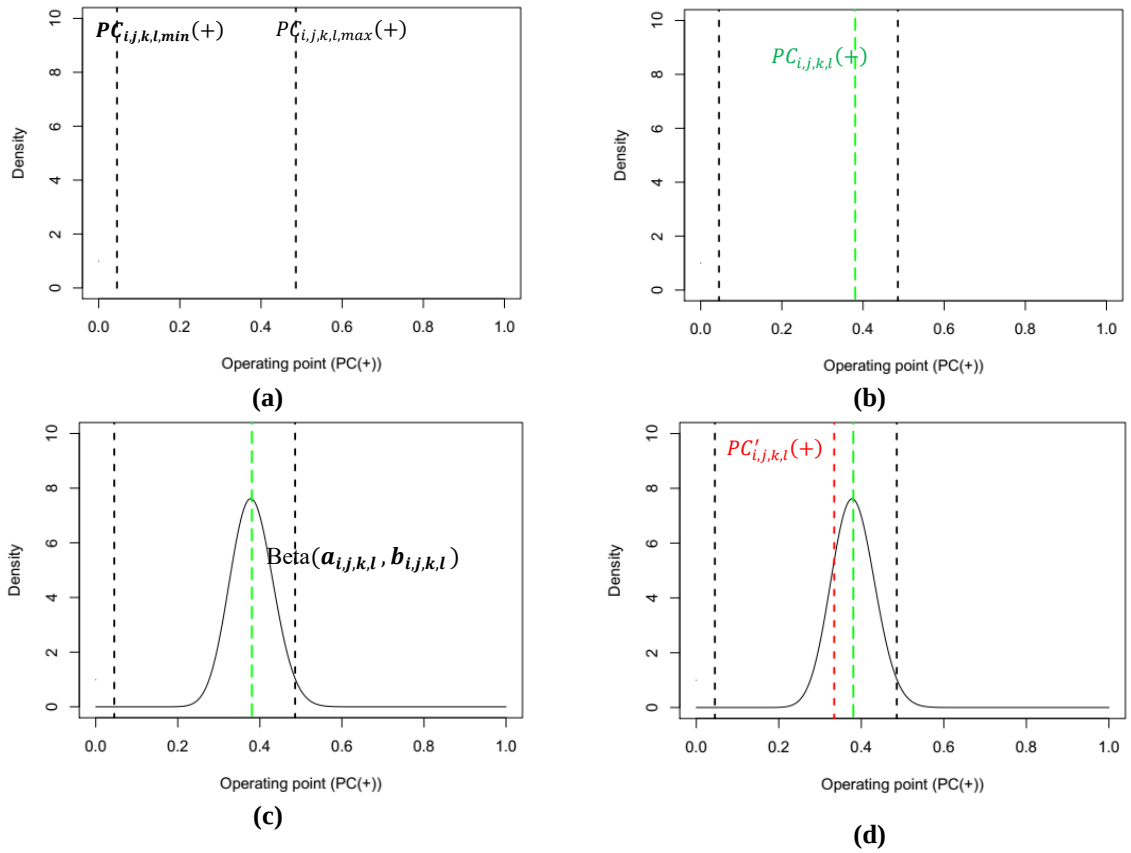



**Figure D.1: Examples of operating point range boundaries (a), target operating point (b), operating point probability density function (Beta distribution) (c) and operating point drawn from ), operating point probability density function (d)**

## Appendix E: Friedman non-parametric Anova test results

| Cost uncertainty scenario | Benchmarks | Algorithms | Cost space paradigm for ensemble selection and model evaluation | Evaluation metric | Friedman Chi-squared statistic | P-value | Significance |
|---|---|---|---|---|---|---|---|
| ***High*** | Cost-sensitive classifiers for cost uncertainty | CISVM, RiskBoost, MiniMax, CSMES | *Cost curve* | AUCC | 53.971 (df=3) | 1.138 e-11 | *** |
| | | | | EMC | 427.98 (df=3) | < 2.2 e-16 | *** |
| | | | *Brier curve* | AUBC | 39.914 (df=3) | 1.111 e-08 | *** |
| | | | | EMC | 412.82 (df=3) | < 2.2 e-16 | *** |
| | Alternative ensemble/ensemble selection strategies | Full, Weighted, Best, Top10, Top25, CSMES | *Cost curve* | AUCC | 56.048 (df=5) | 7.945 e-11 | *** |
| | | | | EMC | 346.31 (df=5) | <2.2 e-16 | *** |
| | | | *Brier curve* | AUBC | 20.456 (df=5) | 0.00103 | ** |
| | | | | EMC | 217.32 (df=5) | <2.2 e-16 | *** |
| | Optimization-based ensemble and classifier selection benchmarks | GA-ES, PSO-ES, GHS-ES, MGA-ES, MGA-SVM-CS, CSMES | *Cost curve* | AUCC | 34.959 (df=5) | 1.533 e-06 | *** |
| | | | | EMC | 979.67 (df=5) | <2.2 e-16 | *** |
| | | | *Brier curve* | AUBC | 63.068 (df=5) | 2.818 e-12 | *** |
| | | | | EMC | 871.22 (df=5) | <2.2 e-16 | *** |
| ***Partial*** | Cost-sensitive classifiers for cost uncertainty | CISVM, RiskBoost, MiniMax, CSMES | *Cost curve* | pAUCC | 1002 (df=3) | <2.2 e-16 | |
| | | | | EMC | 427.98 (df=3) | < 2.2 e-16 | *** |
| | | | *Brier curve* | pAUBC | 562.8 | < 2.2 e-16 | *** |
| | | | | EMC | 413.2 (df=3) | < 2.2 e-16 | *** |
| | Alternative ensemble/ensemble selection strategies | Full, Weighted, Best, Top10, Top25, CSMES | *Cost curve* | pAUCC | 1496.8 (df=5) | <2.2 e-16 | *** |
| | | | | EMC | 296.37 (df=5) | <2.2 e-16 | *** |
| | | | *Brier curve* | pAUBC | 255.67 (df=5) | <2.2 e-16 | *** |
| | | | | EMC | 237.33 (df=5) | <2.2 e-16 | *** |
| | Optimization-based ensemble and classifier selection benchmarks | GA-ES, PSO-ES, GHS-ES, MGA-ES, MGA-SVM-CS, CSMES | *Cost curve* | pAUCC | 1475.2 (df=5) | <2.2 e-16 | *** |
| | | | | EMC | 942.76 (df=5) | <2.2 e-16 | *** |
| | | | *Brier curve* | pAUBC | 3197.4 (df=5) | <2.2 e-16 | |
| | | | | EMC | 845.4 (df=5) | <2.2 e-16 | *** |
| ***None*** | Cost-sensitive classifiers for cost uncertainty | CISVM, RiskBoost, MiniMax, CSMES | *Cost curve* | EMC | 426.3 (df=3) | < 2.2 e-16 | *** |
| | | | *Brier curve* | EMC | 413.2 (df=3) | < 2.2 e-16 | *** |
| | Alternative ensemble/ensemble selection strategies | Full, Weighted, Best, Top10, Top25, CSMES | *Cost curve* | EMC | 308.76 (df=5) | <2.2 e-16 | *** |
| | | | *Brier curve* | EMC | 229.94 (df=5) | <2.2 e-16 | *** |
| | Alternative Optimization-based ensemble and classifier selection benchmarks | GA-ES, PSO-ES, GHS-ES, MGA-ES, MGA-SVM-CS, CSMES | *Cost curve* | EMC | 1480.8 (df=5) | <2.2 e-16 | *** |
| | | | *Brier curve* | EMC | 3172.7 (df=5) | <2.2 e-16 | *** |

**'**' indicates a significant result at α=0.05; '***' indicates a significant result at α=0.001.**

## Appendix F: Computational cost analysis

In this appendix, CSMES is compared to benchmark algorithms in terms of computational cost required to train models. Two comparisons are made in terms of model training runtimes. First, CSMES

is compared to all standalone algorithms designed for cost-sensitive learning under the assumption of cost uncertainty. This benchmark set involves CISVM, RiskBoost, MiniMax and MGA-SVM-CS. The second comparison considers CSMES and alternative ensemble and classifier selection algorithms that depend on the training of a (heterogenous) model library. Since all these selection strategies are applied to the same model libraries in this study, this comparison only involves runtimes of the model selection stages. For GA-ES, PSO-ES and GHS-ES this involves runtimes of single-criterion optimization schemes. For CSMES and MGA-ES this involves multi-criteria optimization (through NSGA-II), followed by the derivation of an ensemble nomination curve. Finally, for all ranking-based classifier and ensemble selection strategies (Full, Weighted, Best, Top10, Top25) this merely involves an assessment of all constituent classifiers in terms of AUCC or AUBS, depending on the cost space paradigm chosen for the evaluation. Since these ranking-based ensemble and classifier selection strategies all depend upon the same performance assessment and subsequent ranking of model library members they are reported as a single entity.

Experimental configurations are identical to the settings deployed for comparing model classification performance. Model training phase runtimes are recorded in seconds and statistically compared over 21 datasets and a 10-fold cross-validation. Moreover, these analyses make abstraction of the cost space paradigm deployed (cost space or Brier space) by averaging runtimes over both settings. All models are trained on a desktop computer equipped with an Intel Core I7-6700k processor with 8 threads clocked at 4.0 Ghz and 32 GB RAM. All classifiers involved in the comparisons are implemented in R (R Core Team, 2019). The model libraries upon which ensemble and classifier selection algorithms depend are built in R and SAS using a variety of packages and procedures. None of the classifiers and optimization algorithms deploy parallel computing.

The following table (Table F.13) present the results of Friedman tests (Friedman, 1937) for both comparisons depicted above.

| Benchmarks | Algorithms | Friedman Chi-squared statistic | P-value | Significance |
|---|---|---|---|---|
| Standalone cost-sensitive classifiers for cost uncertainty | CISVM, RiskBoost, MiniMax, MGA-SVM-CS, CSMES | 78.324 (df=4) | 4.441 e-16 | *** |
| Alternative ensemble/ensemble selection strategies based on heterogeneous model library | GA-ES, PSO-ES, GHS-ES, MGA-ES, Ranking-based benchmarks (Full, Weighted, Best, Top10, Top25), CSMES | 75.422 (df=5) | 7.55 e-15 | *** |

**Table F.13: Model train time duration comparison test results. '**' indicates a significant result at α=0.05; '***' indicates a significant result at α=0.001.**

These results indicate the existence of significant differences in training phase runtimes between the algorithms in the comparison. The following tables (Table F.14 and Table F.15) present average ranks (where lower ranks indicate shorter average runtimes) and post-hoc test results based on Li's procedure (Li, 2008) that provide more detailed insights in how CSMES compares to the benchmark algorithms.

| Evaluation metric | | Algorithm | | | | |
|---|---|---|---|---|---|---|
| | | ***CISVM*** | ***RiskBoost*** | ***MiniMax*** | ***MGA-SVM-CS*** | ***CSMES*** |
| Model training time (sec.) | Avg. rank | 1 | 2.1904 | 2.8571 | 4.8571 | 4.0952 |
| | Adj. p-value | 0.0000*** | 0.0001*** | 0.0125** | 0.1184 | |

**Table F.14: Model train time duration post-hoc tests: CSMES versus cost-uncertainty accommodating cost-sensitive benchmark algorithms. '**' indicates a significant result at α=0.05, while '***' indicates a significant result at α=0.001.**

| Evaluation metric | | Algorithm | | | | | |
|---|---|---|---|---|---|---|---|
| | | ***GA-ES*** | ***PSO-ES*** | ***GHS-ES*** | ***MGA-ES*** | ***Ranking-based benchmarks (Full, Weighted, Best, Top10, Top25)*** | ***CSMES*** |
| Model training time (sec.) | Avg. rank | 3.5714 | 4.5238 | 5.3810 | 4.3333 | 1.4762 | 1.7143 |
| | Adj. p-value | 0.0040** | 0.0000*** | 0.0000*** | 0.0383** | 0.6801 | |

**Table F.15: Model train time duration post-hoc tests: CSMES versus alternative ensemble/ensemble selection strategies based on heterogeneous model libraries. '**' indicates a significant result at α=0.05, while '***' indicates a significant result at α=0.001.**

The results in Table F.14 demonstrate that CSMES is significantly outperformed by CISVM, RiskBoost and MiniMax classifiers in terms of model training runtimes. Since CSMES depends on the training of a sizeable heterogeneous library of models and multicriteria optimization, it is unsurprising that non-ensemble classifiers have an advantage over CSMES in terms of model training time. This

observation does not hold for MGA-SVM-CS, a more complex benchmark explicitly designed for accommodating cost uncertainty that is characterized by multicriteria hyperparameter optimization and classifier selection.

Table F.15 compares CSMES to alternative ensemble and classifier strategies that, like CSMES, share dependence on the creation of a heterogeneous model library. These results demonstrate the competitive nature of CSMES in comparison to these benchmarks in terms of computational requirements. First, in comparison to single-criterion ensemble selection, CSMES is characterized by training phase runtimes that are significantly shorter. A likely explanation is that GA-ES, PSO-ES and GHS-ES deploy AUCC and AUBS as optimization criteria, which are far more complex and thus, computationally demanding, than false negative rate and false positive rates minimized by CSMES. A similar reasoning explains the advantage that CSMES holds over MGA-ES which adds the criterion of model ambiguity to AUCC and AUBS. Finally, in comparison to heuristic ensemble selection approaches that depend on a simple ordering of model library classifiers along AUCC or AUBS but involve no optimization, no significant difference could be established.

In summary, these results show that the added complexity of CSMES over standalone algorithms designed for cost uncertainty comes at a cost of increasing runtimes. However, in comparison to alternative ensemble or classifier selection approaches, CSMES' dependence on multicriteria optimization and the derivation of an ensemble nomination curve does not result in a disadvantage in terms of training runtimes.