A Multi-Objective Approach for Profit-Driven Feature Selection in Credit Scoring

Nikita Kozodoi[a,b], Stefan Lessmann[a], Konstantinos Papakonstantinou[b], Bart Baesens[c]

[a] *Humboldt University of Berlin, Berlin, Germany*
[b] *Kreditech Holding, Hamburg, Germany*
[c] *Catholic University of Leuven, Leuven, Belgium*

Abstract

In credit scoring, feature selection aims at removing irrelevant data to improve the performance of the scorecard and its interpretability. Standard techniques treat feature selection as a single-objective task and rely on statistical criteria such as correlation. Recent studies suggest that using profit-based indicators may improve the quality of scoring models for businesses. We extend the use of profit measures to feature selection and develop a multi-objective wrapper framework based on the NSGA-II genetic algorithm with two fitness functions: the Expected Maximum Profit (EMP) and the number of features. Experiments on multiple credit scoring data sets demonstrate that the proposed approach develops scorecards that can yield a higher expected profit using fewer features than conventional feature selection strategies.

Keywords: feature selection, multi-objective optimization, credit scoring, profit maximization, genetic algorithm

1. Introduction

Credit scoring refers to the use of statistical models that guide managerial decisions in the retail credit sector [11]. This sector has gained a considerable economic value: in 2017, consumer credit outstandings reached €1,195 billion in EU[1]. In the US, the total outstanding consumer credit amount exceeded $3,831 billion[2]. At the same time, the delinquency rate on consumer loans by commercial banks experienced a growth of more than 11% since 2015[3]. The rise of default rates emphasizes the importance of accurately deciding upon loan provisioning, which is a task of credit scoring. To distinguish defaulters and non-defaulters, financial institutions deploy binary scoring models (i.e., scorecards) that predict the probability of default (PD) – an applicant's willingness and ability to repay debt.

Data-driven models, which are used to score applicants, require financial institutions to face costs of gathering and storing large amounts of data on customer behavior. At the same time, companies are required to comply with regulations (i.e., the Basel Accords and IFRS 9) that enforce comprehensible scoring models. By removing irrelevant and redundant features, feature selection can reduce costs and improve the model performance and its comprehensibility (interpretability).

Feature selection can be considered as a multi-objective problem with conflicting goals. In credit scoring, these goals are: increasing the model profitability, reducing the data acquisition costs and improving the interpretability of the model. Yet, most existing approaches in machine learning literature treat feature selection as a single-objective task [5, 10, 39].

Standard feature selection techniques use statistical criteria to identify the optimal subset of features. Recent credit scoring literature criticized a widespread practice of using standard performance measures such as area under the receiver operating characteristic curve (AUC) for evaluating scoring models [17] and call for profit-based performance indicators [14, 36]. This finding stresses the importance of using value-oriented feature selection strategies that identify the optimal subset of features in a profit-maximizing manner.

The goal of this paper is to design a feature selection framework for credit scoring that overcomes some of the drawbacks of traditional feature selection techniques. The proposed method selects features in a profit-maximizing manner rather than relying on statistical measures and addresses both profitability and comprehensibility with multi-criteria optimization. We use the recently developed Expected Maximum Profit (EMP) measure to evaluate the model profitability [36]. We also consider the number of features as an indicator of model comprehensibility and data-related costs: minimizing the number of features reduces costs on data acquisition and storage and makes the model more comprehensible [28]. To simultaneously address both objectives, we employ a multi-objective feature selection framework based on the non-dominated sorting-based genetic algorithm (NSGA-II) [12] with two fitness functions: EMP and the number of features. The proposed method generates a frontier of non-dominated solutions, which represents a trade-off between two objectives and can, therefore, aid decision-makers in selecting a suitable solution. To validate the effectiveness of our approach, we conduct empirical experiments on ten real-world credit scoring data sets.

The contribution of this paper is three-fold. First, we intro-

[1] Source: https://www.ca-consumerfinance.com/en/Espaces/Press-corner/Panorama-du-credit-a-la-consommation-en-Europe/Overview-of-consumer-credit-in-Europe-in-2016-Strong-growth-in-the-European-consumer-credit-market

[2] Source: https://www.federalreserve.gov/releases/g19/current/

[3] Source: https://fred.stlouisfed.org/series/DRCLACBS

duce a profit maximization framework to the feature selection stage by using the EMP measure as a fitness function. Second, we employ a multi-objective feature selection approach based on the NSGA-II algorithm that was not previously considered in the credit scoring literature. Third, we provide empirical evidence that the proposed multi-objective feature selection method identifies feature subsets that deliver the same or higher expected profit using fewer features than conventional feature selection strategies. The results of our study imply that the standard practice of using single-objective feature selection methods misses promising solutions that can be identified by the suggested multi-criteria feature selection framework.

The remainder of this paper is organized as follows. Section 2 reviews related literature on feature selection methods and describes previous work on profit-driven credit scoring. In Section 3, we present and explain the proposed multi-objective feature selection framework. Section 4 describes our experimental setup and presents the empirical results. In Section 5, we discuss the main conclusions of our study.

2. Theoretical Background

2.1. Feature Selection

Feature selection is a dimensionality reduction technique that aims at selecting a subset of features from the input data by removing irrelevant, redundant or noisy features while maintaining the model performance [15]. Feature selection methods split into three groups: filters, wrappers and embedded methods [16].

Filters perform feature selection based on some general data characteristics before training the model. On the first stage, all features are ranked according to a certain criterion that describes the relevance of a particular feature. Popular measures include feature-to-target correlation [5], information gain [10], Fisher score [8] and others. In the second stage, a certain percentage of the top-ranked features is selected, whereas features with lower importance are dropped from the model. Compared to other feature selection strategies, filters are fast and efficient. However, they were shown to perform poorly in benchmark studies [16].

Wrappers are algorithms that iteratively process different feature subsets and select the optimal subset based on the model performance. Since evaluating all possible feature combinations is computationally expensive, research has suggested multiple heuristic search strategies. Popular approaches are sequential forward selection (SFS) and sequential backward selection (SBS) [16]. SFS starts with an empty model and iteratively adds features, selecting the one which brings the largest performance gain, whereas SBS starts with a full set of features and eliminates those contributing the least to the model performance. The search is continued until there is no further improvement. Another strategy relies on evolutionary algorithms such as genetic algorithms (GA), particle swarm optimization (PSO) and others [40]. GAs operate on a population of individuals, where each individual represents a model with binary genes indicating the inclusion of specific features. At each generation, a new population is created by selecting individuals according to their fitness (model performance), recombining them together and undergoing mutation. The model with the highest fitness is selected after running the algorithm for multiple generations.

Embedded methods conduct feature selection simultaneously with the model training. One of the popular approaches is L1-regularized regression that performs feature selection by assigning zero coefficients to irrelevant features in the process of the model development [37]. The main drawback of embedded methods is that they can only be applied within a specific model class.

Most existing feature selection techniques consider feature selection as a single-objective task. However, conflicting goals of feature selection (optimizing the model performance and minimizing the number of selected features) suggest that it can be treated as a multi-objective optimization problem. The literature on multi-objective feature selection is limited compared to the research on conventional single-objective techniques. Nevertheless, there exists a number of attempts to employ the multi-criteria optimization frameworks.

One of the approaches to perform multi-criteria feature selection is to convert a problem into a single-objective task by aggregating the weighted objectives into a single fitness function. For instance, Boln-Canedo and colleagues propose adding a new term to the evaluation function of well-known filter methods such as correlation-based feature selection, Minimal-Redundancy-Maximal-Relevance and RelieF [6, 7]. The new term represents a number of features or their cost, which ensures that two objectives are included in the fitness function. A major downside of this approach is the requirement to explicitly assign weights to objectives, which is a challenging task given uncertainty and different scales of the objectives.

Another approach to account for multiple objectives is to consider a single-objective optimization problem with a budget constraint. In some studies, researchers suggest minimizing the number of features given that a certain level of performance is achieved [3, 29], whereas others optimize predictive performance under the budget constraint for the cost of included features [26]. Both these directions require setting a specific threshold to introduce a budget constraint, either for the model performance or for the number of used features. Therefore, the application of this approach is problematic in cases with no hard budget constraints.

A more promising strategy is to consider objectives separately and look for a set of non-dominated solutions that are optimal in terms of multiple objectives instead of focusing on a single solution. The set of non-dominated points is also known as the Pareto efficient frontier and represents points, for which one can not improve on one objective without decreasing the other. Literature proposed multi-objective modifications of the well-known evolutionary algorithms such as GA and PSO that rely on multiple fitness functions to perform a search of the non-dominated solutions. Emmanouilidis et al. used a two-objective genetic algorithm to perform feature selection that minimizes the number of features and optimizes the error rate or RMSE for classification and regression on different data sets [13]. More recent studies use modified versions of multi-objective genetic

algorithms including the Strength Pareto Evolutionary Algorithm (SEPA-II) and the Non-Dominated Sorting Genetic Algorithm (NSGA-II) [32, 19] to perform feature selection with the same objectives. Research has also suggested using other evolutionary algorithms such as PSO [38] and Artificial Bee Colony (ABC) [18].

To the best of our knowledge, research on value-driven feature selection in credit scoring is currently limited to a single-objective embedded regularization framework for support vector machines (SVM) [23, 24]. Recent benchmarking studies in credit scoring have shown that SVM performs poorly in comparison with other classifiers [21]. Given these results, developing a profit-driven feature selection approach that is not limited to SVM and optimizes both profitability and model comprehensibility contributes to the literature.

2.2. Profit-Oriented Credit Scoring

The credit scoring task is commonly expressed as a classification problem, where a predictive model learns to differentiate between *bad* risks (defaulters) and *good* risks (repayers). Traditional machine learning algorithms are designed to optimize statistical measures such as mean squared error. In recent years, credit scoring literature proposed different strategies to introduce the profit maximization to the scorecard development. One approach is to modify the target variable to reflect profitability. For instance, Serrano-Cinca et al. suggest using the internal rate of return based on the loan interest [30]. Finlay proposes estimating a contribution of each applicant to the profit of the financial institution [14]. Both these measures imply replacing a binary default indicator by a continuous target variable and therefore transform a classification problem into a regression task.

Another approach toward profit scoring is based on using profit-related performance measures for model selection. Recently, Verbraken and colleagues suggested the Expected Maximum Profit (EMP) measure [36]. The calculation of EMP is based on costs and benefits that arise as a result of the actions the company undertakes. To illustrate the calculation process, we follow their notation and label defaulters as class 0 and non-defaulters as class 1. The scorecard assigns a score to each applicant that expresses the probability of default. Applicants are then considered as *bad* risks and rejected if the estimated credit score exceeds a cutoff value t. Table 1 provides a confusion matrix with the corresponding class probabilities, where π_i are prior probabilities of *good* and *bad* loans, and $F_i(t)$ are predicted cumulative density functions of the scores of class i.

The EMP measure assumes that in the basic scenario no scoring mechanism is implemented and therefore all loans are granted. Hence, if an applicant is predicted as a *good* risk, no additional costs or benefits are observed. In contrast, if an applicant is predicted to be a defaulter, the company faces cost C in case of an incorrect prediction and gets benefit B from an accurate prediction. The methodology to calculate parameters B and C was developed by [9].

Parameter B is the benefit from correctly identifying a *bad* risk. By not providing a loan to a defaulter, the company saves money that would be lost in case of issuing the loan. This amount is the expected loss in case of default:

$$B = \frac{\text{LGD} \cdot \text{EAD}}{A}, \tag{1}$$

where LGD refers to the loss given default, EAD is the exposure at default, and A is the principal of the loan [25]. Since recovery rates for defaulted loans vary heavily [31], B is considered as a random variable, which can take values between 0 and 1. The following probability distribution is assumed:

- $B = 0$ with probability p_0 (a customer repays the entire loan)
- $B = 1$ with probability p_1 (a customer defaults on the entire loan)
- B follows a uniform distribution in (0, 1) with $F(B) = 1 - p_0 - p_1$

Parameter C is the cost of the incorrect classification of *good* risks. By rejecting a *good* customer, the company looses money that could be earned as return on investment:

$$C = \text{ROI} = \frac{I}{A}, \tag{2}$$

where I is the total interest. Verbraken et al. [36] treat parameter C as constant and that we follow their approach in this paper. Given these parameters, the EMP measure can be computed as:

$$\text{EMP} = \int_0^1 \Big[B \cdot \pi_0 F_0(t) - C \cdot \pi_1 F_1(t)\Big] f(B) d(B) \tag{3}$$

EMP can be interpreted as the incremental profit from deciding on credit applications using a scorecard compared to a baseline scenario where credits are granted without screening. In this paper, we use the EMP criterion to measure the scorecard profitability and rely on it as one of the optimization objectives.

The literature on profit-oriented credit scoring focuses on model selection and parameter estimation but does not consider the feature selection stage. Current research on profit-driven feature selection in credit scoring is limited to embedded regularization framework for SVMs [23, 24]. This paper proposes a model-agnostic profit-driven feature selection approach that optimizes both profitability and model comprehensibility.

Table 1
Confusion Matrix with Costs

	Predicted Label	
Actual Label	Bad risk	Good risk
Bad risk	$\pi_0 F_0(t)$ benefit: B	$\pi_0(1 - F_0(t))$ cost: 0
Good risk	$\pi_1 F_1(t)$ cost: C	$\pi_1(1 - F_1(t))$ cost: 0

3. Proposed Profit-Driven Feature Selection Approach

We treat feature selection as a multi-objective problem with two goals: a) maximizing the performance of the scorecard; b) minimizing the number of used features used by the model. We propose a wrapper method based on the binary multi-objective nondominated sorting based genetic algorithm (NSGA-II) with two fitness functions: EMP and number of features. The suggested approach addresses two issues with traditional feature selection techniques in credit scoring: it relies on a profit-driven indicator rather than statistical performance measures and addresses both profitability and model comprehensibility by employing multi-objective optimization.

NSGA-II is a multi-objective evolutionary algorithm developed by [12] to address disadvantages of the previous version of NSGA [33]. NSGA-II is designed to solve multi-objective optimization problems by finding a set of non-dominated solutions which form the efficient Pareto frontier. Experiments on different test problems have shown that NSGA-II is able to maintain a better spread of solutions and convergence compared to some other multi-objective optimizers [12].

NSGA-II consists of three main stages: fast non-dominated sorting, diversity preservation and population update. First, the initial population of n individuals is generated with random gene values. In the case of feature selection, each individual represents a set of features included in the predictive model. We code a population of individuals with a set of binary genes with each gene representing the inclusion of a certain feature in the scorecard.

Second, we compute fitness values for the considered objective functions. For each individual in the current population, we construct a scoring model with a different set of features, which is defined by the gene values of these individuals. We evaluate the performance of the scorecard in terms of EMP and store EMP and the number of selected features as two fitness values.

On the next stage, the population goes through the usual genetic operators: selection, crossover and mutation. The selection is performed with a binary tournament method based on the crowded comparison operator. First, we sort the population by a non-domination rank – the number of individuals dominated by a given solution in terms of the considered objective functions. Next, individuals with the same non-domination ranks are sorted by their crowding distance – the average distance of two solutions on either side of this individual along each of the objectives. Next, one-point crossover is applied to the remaining population. Gene values of the child are computed as a weighted average of the gene values of the parents. In a binary NSGA-II, which is the focus of this paper, a one-point crossover operator simply copies parents' genes if they are the same and randomly chooses a binary value for the conflicting genes. Finally, each gene of the child is flipped with a mutation probability m. These operations are performed until the size of the offspring population reaches n.

After applying all genetic operations, both parents and children are merged into the new population of size $2n$ to ensure elitism. The population is again sorted according to the non-domination and crowding distance. After the sorting is complete, only the top n individuals are selected to proceed to the next stage. This approach helps the algorithm to construct a uniformly spread-out Pareto-optimal frontier by eliminating solutions that are either dominated or located in the crowded regions of the frontier.

Table 2
Credit Scoring Data Sets

Data Label	Sample Size	Num. Features*	Default Rate
australian	690	42	0.4449
german	1,000	61	0.3000
thomas	1,225	28	0.2637
bene1	3,123	83	0.3333
hmeq	5,960	20	0.1995
bene2	7,190	28	0.3000
uk	30,000	51	0.0400
lending club	43,344	206	0.1351
pakdd	50,000	373	0.2608
gmsc	150,000	68	0.0668

* *Number of features after data preprocessing (see Section 4.2)*

The NSGA-II algorithm was previously used for feature selection in fields not related to credit risk. The fitness functions considered in the literature are the number of features and statistical performance measures such as error rate or mean squared error [19, 27, 32]. In this paper, we rely on the NSGA-II algorithm to perform multi-objective feature selection for credit scoring. The central novelty of our framework is the use of a profit measure as one of the fitness functions as well as the area of application.

4. Experimental Results

4.1. Data Description

The empirical evaluations are based on ten retail credit scoring data sets coming from different sources. Data sets *australian* and *german* stem from the UCI Repository [22]. The data sets *pakdd*, *lending club* and *gmsc* were provided by different companies for the data mining competitions on PAKDD and Kaggle platforms. Data sets *bene1*, *bene2* and *uk* were collected from financial institutions in the Benelux and UK [1]. The *thomas* data set is provided by [34]. Finally, *hmeq* is a data set on home equity loans collected by [2].

Each of the data sets has a unique set of features describing the loan applicant (e.g., gender, income) and loan characteristics (e.g., amount, duration). Some data sets also include information on previous loans of the applicant. The target variable is a binary indicator of whether the customer has repaid the loan or not. Table 2 summarizes the main characteristics of the data sets.

As suggested by Table 2, most of the data sets are imbalanced: default rate fluctuates between 4% and 44%. The sample size and number of features varies significantly across the data sets, which suggests that we use a heterogeneous data library for further analysis.

Table 3
Parameter Grid

Method*	Parameter	Candidate values
LR	–	–
L1	cost	$2^{-10}, 2^{-9.5}, 2^{-9}, \ldots, 2^{10}$
XG	nrounds	10, 25, 50, 100, 250, 500, 1000, 2500
	eta	0.01, 0.03, 0.05
	max. depth	1, 3, 5

* *Abbreviations: LR = logistic regression, L1 = L1-regularized LR, XG = extreme gradient boosting*

4.2. Experimental Setup

Our modeling pipeline consists of several stages. First, each data set is pre-processed in the same way. We impute missing values with means for continuous features and with most frequent values for categorical features. Next, we encode all categorical features with $k-1$ dummies, where k is the number of unique categories.

After preprocessing, the data sets are randomly partitioned into two subsets: training sample (70% cases) and holdout sample (30%). On the training set, we use 4-fold cross-validation to perform feature selection. Next, we use the whole training set to train scorecards with the identified feature subsets and evaluate their performance on the holdout data.

Before performing feature selection, we use a subset of the training data to tune meta-parameters of the base classifiers. For each of the considered classification algorithms, we perform a learning curve analysis to select a suitable sample size by gradually increasing the percentage of the training sample until the model performance in terms of EMP stops improving. Next, we use the corresponding subset to perform parameter tuning using grid search [4]. The full parameter grid is presented in Table 3.

As base classifiers, we use three algorithms: logistic regression, L1-regularized logistic regression and extreme gradient boosting. This allows us to check the robustness of feature selection results across different predictive algorithms and see whether internal feature selection in models such as L1-regularized regression diminishes the value of the proposed wrapper approach.

After identifying suitable parameter values, we perform feature selection with the suggested multi-objective framework. The parameters of NSGA-II (number of generations and population size) were set to 200 based on experiments on subsets of training data. To evaluate the performance of the proposed algorithm, we compare it to three traditional feature selection strategies: SFS, SBS and single-objective GA. We also use a scorecard that relies on a full set of features as a benchmark. To ensure a fair comparison, we set the number of generations and number of individuals for the simple GA to the same values as for the NSGA-II. All three single-objective benchmarks use the EMP measure as a fitness function. We only consider wrapper methods as benchmarks because of their superior performance compared to other feature selection strategies [16].

4.3. Empirical Results

Figure 1 presents the graph matrix with the performance of the considered feature selection methods on all ten data sets. The Pareto frontier identified by the NSGA-II algorithm is depicted with red markers, whereas blue points represent the single-objective benchmarks. The black cross marks the baseline solution which is based on a full model without feature selection. In this section, we focus on the results of experiments where logistic regression is used as a base model for all methods. Logistic regression is still widely used in practice [20] despite that other algorithms have been shown to predict credit risks more accurately [21]. Results for other base classifiers are given in Figures A.1 – B.1 in the Appendix.

Results indicate that the size of the NSGA-II Pareto frontier varies across the data library from having just 2 solutions (*thomas* and *bene1*) to 20 feature subsets (*pakdd*). The small size of the Pareto frontier can be explained by two reasons: first, no candidate solutions with a larger number of features demonstrate better performance during cross-validation; second, some solutions become dominated when evaluating their quality on the holdout data and are therefore dropped from the frontier. Hence, NSGA-II frontiers are likely to contain fewer solutions on data sets with lower dimensionality and stronger differences in data distribution between the training and holdout samples.

Overall, the points on NSGA-II frontiers usually populate regions with a smaller number of features compared to benchmarks. Single-objective methods optimize predictive performance but do not account for the number of features. This does not motivate the algorithm to select smaller feature subsets. Nevertheless, sequential forward selection chooses fewer features compared to sequential backward elimination on all ten data sets.

To evaluate the quality of the frontiers and compare them with the single objective benchmarks, we look at the performance of the considered feature selection methods in Table 4. To facilitate comparison, on each of the Pareto frontiers we select one solution that achieves the best performance in terms of EMP (the upper-right point). Then, we compare this solution with single-objective benchmarks in terms of EMP and k (number of features).

As Table 4 suggests, the best-performing NSGA-II solution is based on fewer features compared to the single-objective solutions in 7 out of 10 cases and achieves a higher expected profit in half of the data sets. There is only one data set where one of the benchmarks identifies a solution which has both higher EMP and a lower complexity (*gmsc*). Tables with the performance of feature selection methods using other base classifiers produce similar results (see Appendix).

To further extend the comparison, consider the example Pareto frontier depicted in Figure 2. Here, the task is to minimize objective I while maximizing objective II. The frontier is represented by points A to E, whereas points F, G and H are external solutions. Point H is dominated by points A to D on the Pareto frontier because they perform better in two objectives. Points F and G demonstrate better performance in terms of objective II compared to the best solution from the Pareto frontier

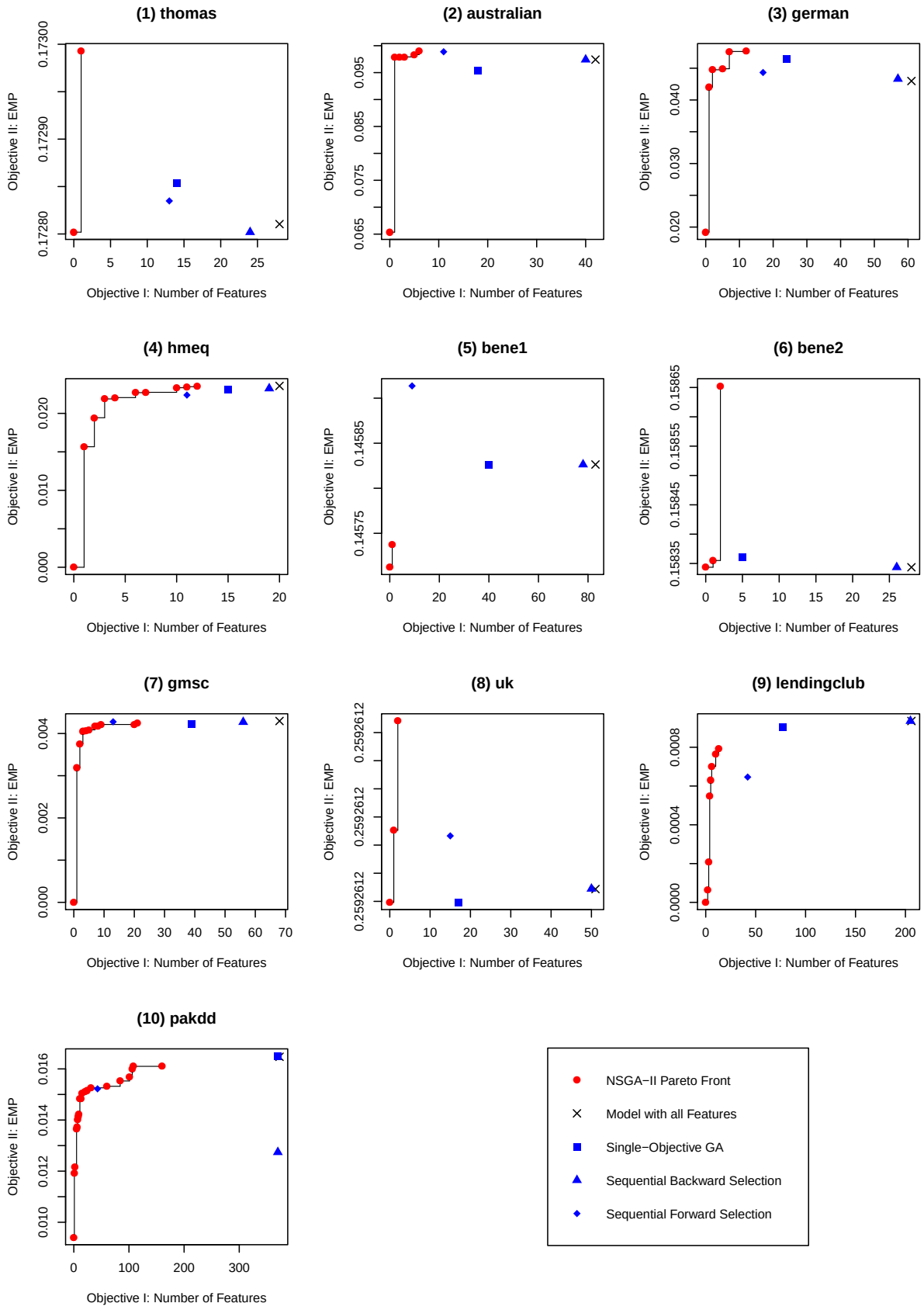

Fig. 1. *Performance of Feature Selection Methods.* Each diagram in the graph matrix depicts results on a single data set. The Pareto frontier outputted by the NSGA-II algorithm is depicted with red points, whereas the blue markers refer to benchmarks. LR is used as a base classifier.

Table 4
Performance of Feature Selection Methods*

Data	NSGA-II**		GA		SBS		SFS		Full Model	
	EMP	k***	EMP	k	EMP	k	EMP	k	EMP	k
australian	0.0990	6	0.0953	18	0.0974	40	0.0989	11	0.0974	42
german	0.0477	12	0.0465	24	0.0433	57	0.0443	17	0.0430	61
thomas	0.1730	1	0.1729	14	0.1728	24	0.1728	13	0.1728	28
bene1	0.1457	1	0.1458	40	0.1458	78	0.1458	9	0.1458	83
hmeq	0.0235	12	0.0231	15	0.0232	19	0.0224	11	0.0236	20
bene2	0.1587	2	0.1584	5	0.1583	26	0.1584	5	0.1583	28
uk	0.2593	2	0.2593	17	0.2593	50	0.2593	15	0.2593	51
lending club	0.0008	13	0.0009	77	0.0009	205	0.0006	42	0.0009	206
pakdd	0.0161	160	0.0165	370	0.0127	370	0.0152	43	0.0165	373
gmsc	0.0042	21	0.0042	39	0.0043	56	0.0043	13	0.0043	68

* *Results in this table use logistic regression as a base classifier. Results for other models are given in the Appendix. EMP is rounded to four digits after the decimal point.*

** *We consider a single solution on the NSGA-II frontier, which has the highest EMP and uses the maximal number of features*

*** *k refers to the number of selected features used to construct the model.*

(point E). However, there is a crucial difference between these points. Solution G does not dominate any points on the frontier – it achieves better performance in objective II only by deteriorating on objective I. At the same time, point F achieves better performance in both objectives compared to points D and E on the frontier. Therefore, F dominates these solutions.

Based on the example above, we define three metrics. Let S_1 be a share of data sets where all single-objective benchmarks are dominated by points on the Pareto frontier resulting from the NSGA-II algorithm (e.g., point H). If satisfied, this condition indicates a clear advantage of the multi-objective feature selection over the benchmarks, since they can not achieve better performance in any of the objectives. Next, let S_2 indicate a share of data sets with a weaker condition: none of the benchmarks dominates the solution on the Pareto front. Here, benchmarks may either be dominated by the solutions on the frontier (e.g., point H) or achieve better EMP than solutions on the frontier, but only if they use more features (e.g., point G). Finally, let S_3 be a share of data sets where one or more benchmarks dominates at least one solution on the frontier. This condition corresponds to point F from the aforementioned example and demonstrates an advantage of the single-objective benchmarks. We compute shares S_1, S_2 and S_3 separately for each base classifier. The results are given in Table 5.

As Table 5 suggests, all single-objective benchmarks are dominated by the best point on the NSGA-II frontier on 60% data sets for L1 and on 50% cases for LR and XG. In other words, NSGA-II identifies a feature subset that has a higher profitability and contains fewer features compared to the considered conventional single-objective strategies on at least half of the data sets.

In most of the remaining cases, single-objective benchmarks can outperform the best multi-objective solution in terms of EMP only if they use more features. This is observed for four remaining data sets when using LR or L1 and for five data sets for XG. In this case, solutions on the frontier identified by our method are still non-dominated by benchmarks and represent a trade-off between model comprehensibility and profitability in the regions where fewer features are used. Feature subsets selected by the single-objective benchmarks could serve as a possible extension of the frontier.

Table 5
Aggregated Results

Base Classifier	S_1	S_2	S_3
Logistic regression	50%	90%	10%
L1-regularized LR	60%	100%	0%
Gradient boosting	50%	100%	0%

From the business perspective, solutions on the NSGA-II frontier may be more attractive for companies even if the scorecards are characterized by a lower profitability but based on a significantly smaller amount of data. For instance, NSGA-II achieves EMP of 0.0161 on *pakdd* data using 160 features, whereas single-objective GA identifies a subset of 370 features that obtains EMP of 0.0165. Here, relying on a multi-objective algorithm results in a 2% drop in EMP but also eliminates 57% of features. It is then the task of a risk analyst to decide whether a drop in profitability would be compensated by reducing the costs of collecting and storing the data on customer behavior.

Taking both objectives into account, solutions lying on the NSGA-II frontier are not dominated by any of the benchmarks in 90% to 100% cases depending on the base model. There is a single case (*gmsc* data with LR), where one of the single-objective methods dominates some solutions on the frontier. This indicates a good performance of the proposed multi-objective feature selection algorithm.

5. Conclusion

This paper introduces a multi-objective profit-driven framework for feature selection in credit scoring. We use the recently developed EMP measure and the number of features as two fitness functions for the wrapper-based feature selection to address both profitability and comprehensibility. Multi-objective

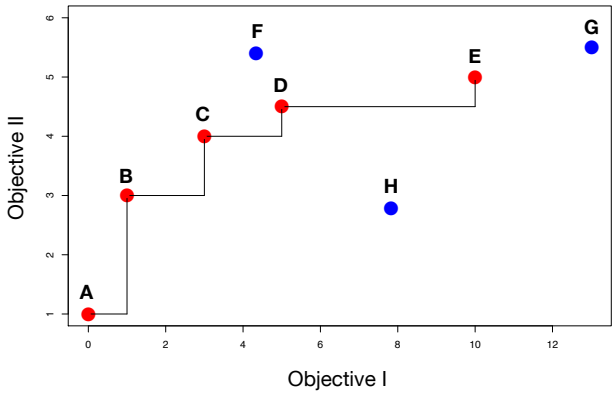

Fig. 2. *Example Multi-Objective Optimization.* The task is to minimize objective I while maximizing objective II. Points A – E represent solutions on the efficient frontier, points G, F and H are external solutions. Compared to the frontier, H is dominated by points A to D, G is a non-dominated point, and F dominates solutions D and E on the frontier.

optimization is performed with the genetic algorithm NSGA-II. We evaluate the effectiveness of our approach by running empirical experiments on ten real-world retail credit scoring data sets.

Empirical results indicate that the proposed multi-objective feature selection framework performs highly competitive compared to the conventional feature selection strategies. The developed approach identifies feature subsets that yield the same or higher expected profit using fewer features than single-objective benchmarks on at least half of the data sets. Depending on a base classifier, solutions selected by the NSGA-II are not dominated by any of the single-objective benchmarks in 90% to 100% of cases. The results imply that previous work in ignoring the two objectives of feature selection in credit scoring has missed promising solutions that can be identified using the suggested framework.

In addition to demonstrating a superior performance, the suggested multi-objective method serves as a tool to find a trade-off in two conflicting objectives: comprehensibility and profitability of the model. By comparing the non-dominated solutions on the efficient frontier, risk managers can select a suitable subset of features depending on their business context.

Future research could pursue several directions. Recent literature suggested new multi-criteria optimization methods that could replace the NSGA-II algorithm in the profit-driven feature selection framework. For instance, Hancer and colleagues suggest using multi-objective artificial bee colony optimization [18]; Zhang et al. apply multi-criteria particle swarm optimization to perform feature selection [41]. A benchmarking study with different evolutionary algorithms would shed more light on identifying a suitable optimizer for the profit-driven feature selection in credit scoring.

Another promising avenue would be to use the developed feature selection approach in other business applications. One of the possible domains is customer churn. Verbraken and colleagues developed a similar EMP measure for customer churn models [35], which could serve as one of the objectives for the feature selection algorithm.

References

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

[2] Baesens, B., Roesch, D., & Scheule, H. (2016). *Credit Risk Analytics: Measurement Techniques, Applications, and Examples in SAS*. John Wiley & Sons.

[3] Bentez-Pea, S., Blanquero, R., Carrizosa, E., & Ramrez-Cobo, P. (2018). Cost-sensitive Feature Selection for Support Vector Machines. *Computers & Operations Research*.

[4] Bergstra, J. S., Bardenet, R., Bengio, Y., & Kgl, B. (2011). Algorithms for hyper-parameter optimization. In *Advances in Neural Information Processing Systems* (pp. 2546-2554).

[5] Boln-Canedo, V., Snchez-Maroo, N., & Alonso-Betanzos, A. (2013). A review of feature selection methods on synthetic data. *Knowledge and Information Systems, 34*(3), 483-519.

[6] Boln-Canedo, V., Snchez-Maroo, N., Alonso-Betanzos, A., Bentez, J. M., & Herrera, F. (2014). A review of microarray datasets and applied feature selection methods. *Information Sciences, 282*, 111-135.

[7] Boln-Canedo, V., Snchez-Maroo, N., & Alonso-Betanzos, A. (2015). Recent advances and emerging challenges of feature selection in the context of big data. *Knowledge-Based Systems, 86*, 33-45.

[8] Bonev, B., Escolano, F., & Cazorla, M. (2008). Feature selection, mutual information, and the classification of high-dimensional patterns. *Pattern Analysis and Applications, 11*(3-4), 309-319.

[9] Bravo, C., Maldonado, S., & Weber, R. (2013). Granting and managing loans for micro-entrepreneurs: New developments and practical experiences. *European Journal of Operational Research, 227*(2), 358-366.

[10] Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering, 40*(1), 16-28.

[11] Crook, J. N., Edelman, D. B., & Thomas, L. C. (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research, 183*(3), 1447-1465.

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

[13] Emmanouilidis, C., Hunter, A., MacIntyre, J., Cox, C. (1999). Selecting features in neurofuzzy modelling by multiobjective genetic algorithms. In *Proceedings of the 9th International Conference on Artificial Neural Networks* (pp. 4387-4392). Washington, D.C.

[14] Finlay, S. (2010). Credit scoring for profitability objectives. *European Journal of Operational Research, 202*(2), 528-537. Chicago

[15] Guyon, I., & Elisseeff, A. (2003). An introduction to feature and feature selection. *Journal of Machine Learning Research, 3*, 1157-1182.

[16] Guyon, I., Gunn, S., Nikravesh, M., & Zadeh, L. A. (2006). *Feature Extraction: Foundations and Applications (Studies in Fuzziness and Soft Computing)*: Springer-Verlag.

[17] Hand, D. J. (2005). Good practice in retail credit scorecard assessment. *Journal of the Operational Research Society, 56*(9), 1109-1117.

[18] Hancer, E., Xue, B., Zhang, M., Karaboga, D., & Akay, B. (2018). Pareto front feature selection based on artificial bee colony optimization. *Information Sciences, 422*, 462-479.

[19] Hamdani, T. M., Won, J. M., Alimi, A. M., & Karray, F. (2007, April). Multi-objective feature selection with NSGA II. In *International Conference on Adaptive and Natural Computing Algorithms* (pp. 240-247). Springer, Berlin, Heidelberg.

[20] Jung, K. M., Thomas, L. C., & So, M. C. (2015). When to rebuild or when to adjust scorecards. *Journal of the Operational Research Society, 66*(10), 1656-1668.

[21] 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.

[22] Lichman, M. (2013). *UCI Machine Learning Repository <http://archive.ics.uci.edu/ml/>*. Irvine, CA: School of Information and Computer Science, University of California. Accessed 2018-09-01.

[23] Maldonado, S., Bravo, C., Lopez, J., & Perez, J. (2017). Integrated framework for profit-based feature selection and SVM classification in credit scoring. *Decision Support Systems, 104*, 113-121.

[24] Maldonado, S., Prez, J., & Bravo, C. (2017). Cost-based feature selection for support vector machines: An application in credit scoring. *European Journal of Operational Research, 261*(2), 656-665.

[25] Mays, E., & Lynas, N. (2004). *Credit scoring for risk managers: The handbook for lenders*. Ohio: Thomson/South-Western.

[26] Min, F., Hu, Q., & Zhu, W. (2014). Feature selection with test cost constraint. *International Journal of Approximate Reasoning, 55*(1), 167-179.

[27] Oliveira, L. S., Sabourin, R., Bortolozzi, F., & Suen, C. Y. (2002). Feature selection using multi-objective genetic algorithms for handwritten digit recognition. In *Proceedings of the 16th International Conference on Pattern Recognition* (pp. 240-247). IEEE.

[28] Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W., & Wallach, H. (2017). Manipulating and measuring model interpretability. In *NIPS 2017 Transparent and Interpretable Machine Learning in Safety Critical Environments Workshop.*.

[29] Saeedi, R., Schimert, B., & Ghasemzadeh, H. (2014). Cost-sensitive feature selection for on-body sensor localization. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (pp. 833-842). ACM.

[30] Serrano-Cinca, C., & Gutirrez-Nieto, B. (2016). The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (P2P) lending. *Decision Support Systems, 89*, 113-122.

[31] Somers, M., & Whittaker, J. (2007). Quantile regression for modelling distributions of profit and loss. *European Journal of Operational Research, 183*(3), 1477-1487.

[32] Soto, A. J., Cecchini, R. L., Vazquez, G. E., & Ponzoni, I. (2009). Multi-objective feature selection in QSAR using a machine learning approach. *QSAR & Combinatorial Science, 28*(1112), 1509-1523.

[33] Srinivas, N., & Deb, K. (1994). Muiltiobjective optimization using nondominated sorting in genetic algorithms. *Evolutionary Computation, 2*(3), 221-248.

[34] Thomas, L. C., Edelman, D. B., Crook, J. N. (2002) *Credit Scoring and its Applications*. Philadelphia: SIAM.

[35] Verbraken, T., Verbeke, W., & Baesens, B. (2013). A novel profit maximizing metric for measuring classification performance of customer churn prediction models. *IEEE Transactions on Knowledge and Data Engineering, 25*(5), 961-973.

[36] Verbraken, T., Bravo, C., Weber, R., & Baesens, B. (2014). Development and application of consumer credit scoring models using profit-based classification measures. *European Journal of Operational Research, 238*(2), 505-513.

[37] Vidaurre, D., Bielza, C., & Larraaga, P. (2013). A survey of L1 regression. *International Statistical Review, 81*(3), 361-387.

[38] Xue, B., Zhang, M., & Browne, W. N. (2013). Particle swarm optimization for feature selection in classification: A multi-objective approach. *IEEE Transactions on Cybernetics, 43*(6), 1656-1671.

[39] Xue, B., Zhang, M., Browne, W. N., & Yao, X. (2016). A survey on evolutionary computation approaches to feature selection. *IEEE Transactions on Evolutionary Computation, 20*(4), 606-626.

[40] Yang, J., & Honavar, V. (1998). Feature subset selection using a genetic algorithm. In *Feature Extraction, Construction and Selection* (pp. 117-136). Springer, Boston, MA.

[41] Zhang, Y., Gong, D. W., & Cheng, J. (2017). Multi-objective particle swarm optimization approach for cost-based feature selection in classification. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 14*(1), 64-75.

Appendix A. Empirical Results using L1 Model

Table A.1
Performance of Feature Selection Methods: L1

Data	NSGA-II**		GA		SBS		SFS		Full Model	
	EMP	k***	EMP	k	EMP	k	EMP	k	EMP	k
australian	0.1035	4	0.1029	22	0.1032	41	0.1029	10	0.1032	42
german	0.0369	26	0.0397	24	0.0366	60	0.0390	12	0.0358	61
thomas	0.1729	4	0.1728	16	0.1728	21	0.1728	5	0.1728	28
bene1	0.1460	2	0.1460	41	0.1457	82	0.1460	5	0.1457	83
hmeq	0.0206	13	0.0199	16	0.0205	18	0.0186	7	0.0198	20
bene2	0.1589	1	0.1584	12	0.1583	27	0.1591	2	0.1583	28
uk	0.2597	21	0.2593	24	0.2593	47	0.2593	1	0.2593	51
lending club	0.0010	42	0.0009	93	0.0009	205	0.0007	5	0.0010	206
pakdd	0.0160	214	0.0159	201	0.0158	371	0.0135	4	0.0156	373
gmsc	0.0045	17	0.0045	39	0.0045	65	0.0043	10	0.0044	68

* *Results in this table use L1 as a base classifier. EMP is rounded to four digits after the decimal point.*

** *We consider a single solution on the NSGA-II frontier, which has the highest EMP and uses the maximal number of features*

*** *k refers to the number of selected features used to construct the model.*

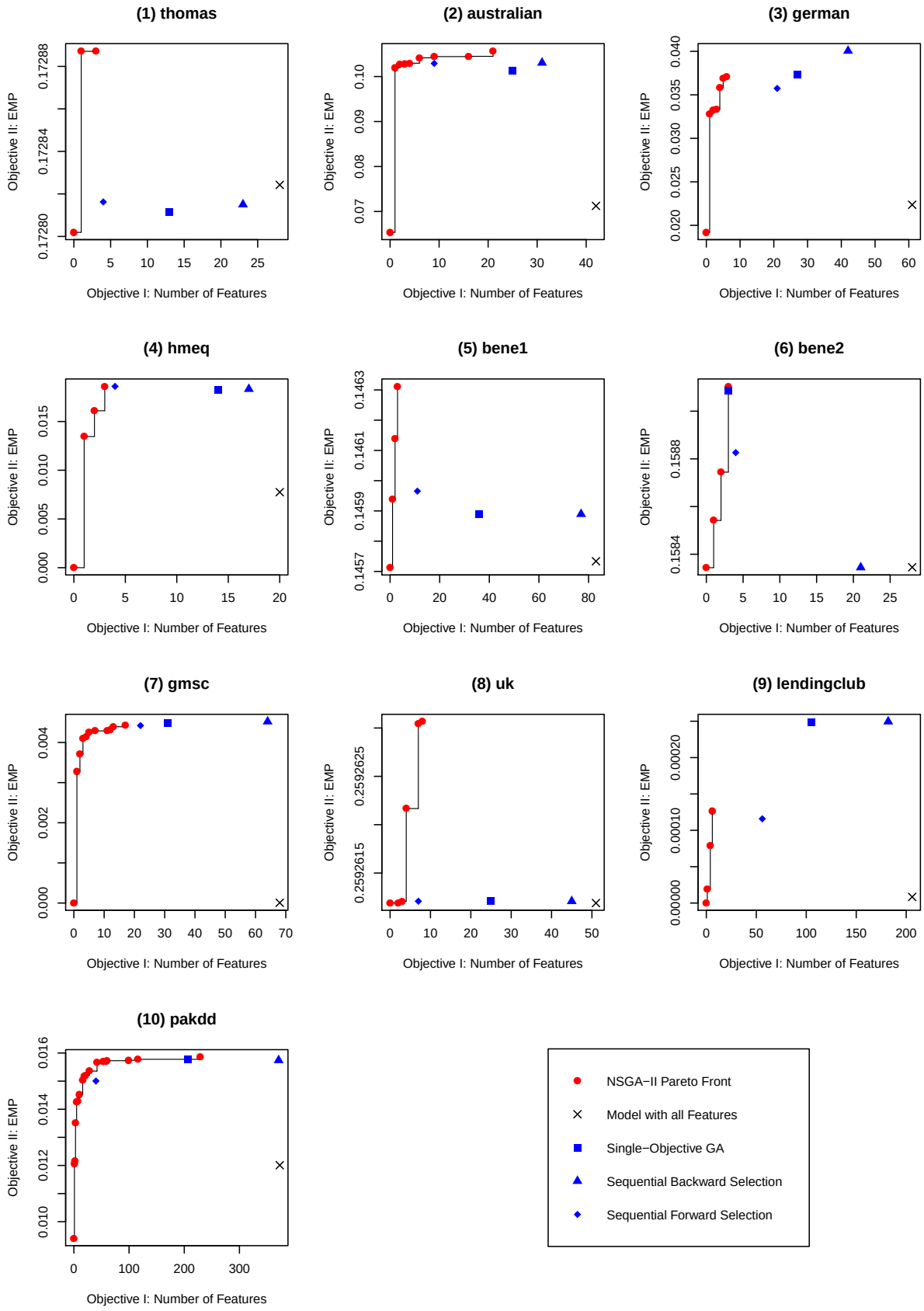

Fig. A.1. *Performance of Feature Selection Methods.* Each diagram in the graph matrix depicts results on a single data set. The Pareto frontier outputted by the NSGA-II algorithm is depicted with red points, whereas the blue markers refer to benchmarks. L1 is used as a base classifier.

Appendix B. Empirical Results using XG Model

Table B.1
Performance of Feature Selection Methods: XG

Data	NSGA-II**		GA		SBS		SFS		Full Model	
	EMP	k***	EMP	k	EMP	k	EMP	k	EMP	k
australian	0.1060	3	0.1065	17	0.1056	36	0.1060	11	0.1055	42
german	0.0393	2	0.0391	25	0.0411	51	0.0330	11	0.0392	61
thomas	0.1731	10	0.1728	12	0.1728	23	0.1728	3	0.1728	28
bene1	0.1457	1	0.1457	42	0.1457	80	0.1459	4	0.1457	83
hmeq	0.0422	19	0.0418	15	0.0415	19	0.0399	12	0.0418	20
bene2	0.1583	1	0.1583	10	0.1583	28	0.1583	3	0.1583	28
uk	0.2593	1	0.2593	18	0.2593	47	0.2593	5	0.2593	51
lending club	0.0008	12	0.0007	107	0.0007	197	0.0008	13	0.0009	206
pakdd	0.0168	203	0.0165	180	0.0164	366	0.0157	14	0.0166	373
gmsc	0.0046	24	0.0045	34	0.0046	66	0.0044	14	0.0045	68

* *Results in this table use XG as a base classifier. EMP is rounded to four digits after the decimal point.*

** *We consider a single solution on the NSGA-II frontier, which has the highest EMP and uses the maximal number of features*

*** *k refers to the number of selected features used to construct the model.*

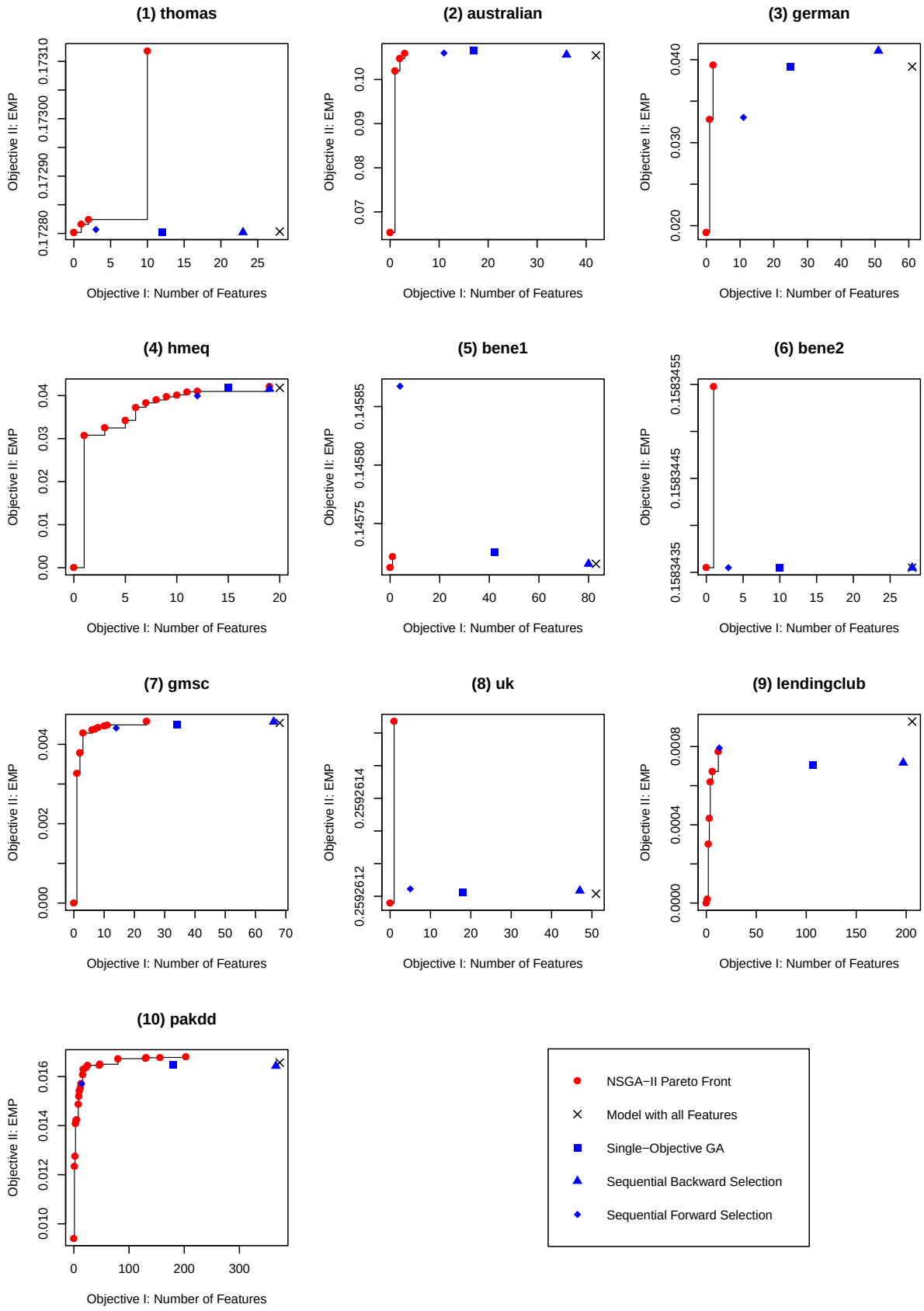

Fig. B.1. *Performance of Feature Selection Methods.* Each diagram in the graph matrix depicts results on a single data set. The Pareto frontier outputted by the NSGA-II algorithm is depicted with red points, whereas the blue markers refer to benchmarks. XG is used as a base classifier.