Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research

Stefan Lessmann , Bart Baesens , Hsin-Vonn Seow , Lyn C. Thomas

PII: S0377-2217(15)00420-8
DOI: 10.1016/j.ejor.2015.05.030
Reference: EOR 12954

To appear in: *European Journal of Operational Research*

Received date: 23 December 2013
Revised date: 9 March 2015
Accepted date: 11 May 2015

Please cite this article as: Stefan Lessmann , Bart Baesens , Hsin-Vonn Seow , Lyn C. Thomas , Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research, *European Journal of Operational Research* (2015), doi: 10.1016/j.ejor.2015.05.030

Highlights

- Large-scale benchmark of 41 classifiers across 8 real-word credit scoring data sets.
- Introduction of ensemble selection routines to the credit scoring community.
- Analysis of 6 established and novel indicators to measure scorecard accuracy.
- Assessment of the financial impact of different scorecards.

Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research

Stefan Lessmann[a,*], Bart Baesens[bc] , Hsin-Vonn Seow[d] , Lyn C. Thomas[c]

[a] *School of Business and Economics, Humboldt-University of Berlin*

[b] *Department of Decision Sciences & Information Management, Catholic University of Leuven*

[c] *School of Management, University of Southampton, Highfield, Southampton, SO17 1BJ, United Kingdom*

[d] *Nottingham University Business School, University of Nottingham-Malaysia Campus*

Abstract

Many years have passed since Baesens et al. published their benchmarking study of classification algorithms in credit scoring [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.]. The interest in prediction methods for scorecard development is unbroken. However, there have been several advancements including novel learning methods, performance measures and techniques to reliably compare different classifiers, which the credit scoring literature does not reflect. To close these research gaps, we update the study of Baesens et al. and compare several novel classification algorithms to the state-of-the-art in credit scoring. In addition, we examine the extent to which the assessment of alternative scorecards differs across established and novel indicators of predictive accuracy. Finally, we explore whether more accurate classifiers are managerial meaningful. Our study provides valuable insight for professionals and academics in credit scoring. It helps practitioners to stay abreast of technical advancements in predictive modeling. From an academic point of view, the study provides an independent assessment of recent scoring methods and offers a new baseline to which future approaches can be compared.

Keywords: Data Mining, Credit Scoring, OR in banking, Forecasting benchmark

* Corresponding author: Tel.: +49.30.2093.5742, Fax: +49.30.2093.5741, E-Mail: stefan.lessmann@hu-berlin.de.

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

[b] Department of Decision Sciences & Information Management, Catholic University of Leuven, Naamsestraat 69, B-3000 Leuven, Belgium

[c] School of Management, University of Southampton, Highfield, Southampton, SO17 1BJ, United Kingdom

[d] Nottingham University Business School, University of Nottingham-Malaysia Campus, Jalan Broga, 43500 Semenyih, Selangor Darul Ehsan, Malaysia

1 Introduction

Credit scoring is concerned with developing empirical models to support decision making in the retail credit business (Crook, et al., 2007). This sector is of considerable economic importance. For example, the volume of consumer loans held by banks in the US was $1,132bn in 2013; compared to $1,541bn in the corporate business.[1] In the UK, loans and mortgages to individuals were even higher than corporate loans in 2012 (£11,676m c.f. £10,388m).[2] These figures indicate that financial institutions require formal tools to inform lending decisions.

A credit score is a model-based estimate of the probability that a borrower will show some undesirable behavior in the future. In application scoring, for example, lenders employ predictive models, called scorecards, to estimate how likely an applicant is to default. Such PD (probability of default) scorecards are routinely developed using classification algorithms (e.g., Hand & Henley, 1997). Many studies have examined the accuracy of alternative classifiers. One of the most comprehensive classifier comparisons to date is the benchmarking study of Baesens, et al. (2003).

Albeit much research, we argue that the credit scoring literature does not reflect several recent advancements in predictive learning. For example, the development of selective multiple classifier systems that pool different algorithms and optimize their weighting through heuristic search represents an important trend in machine learning (e.g., Partalas, et al., 2010). Yet, no attempt has been made to systematically examine the potential of such approach for credit scoring. More generally, recent advancements concern three dimensions: i) novel classification algorithms to *develop* scorecards (e.g., extreme learning machines, rotation forest, etc.), ii) novel performance measures to *assess* scorecards (e.g., the *H*-measure or the partial Gini coefficient), and iii) statistical hypothesis tests to *compare* scorecard performance (e.g., García, et al., 2010). An analysis of the PD modeling literature confirms that these developments have received little attention in credit scoring, and reveals further limitations of previous studies; namely i) using few and/or small data sets, ii) not comparing different state-of-the-art classifiers to each other, and iii) using only a small set of conceptually similar accuracy indicators. We elaborate on these issues in Section 2.

The above research gaps warrant an update of Baesens, et al. (2003). Therefore, the motivation of this paper is to provide a holistic view of the state-of-the-art in predictive

[1] Data from the Federal Reserve Board, H8, Assets and Liabilities of Commercial Banks in the United States (http://www.federalreserve.gov/releases/h8/current/).

[2] Data from ONS Online, SDQ7: Assets, Liabilities and Transactions in Finance Leasing, Factoring and Credit Granting: 1st quarter 2012 (http://www.ons.gov.uk).

modeling and how it can support decision making in the retail credit business. In pursuing this objective, we make the following contributions: First, we perform a large scale benchmark of 41 classification methods across eight credit scoring data sets. Several of the classifiers are new to the community and for the first time assessed in credit scoring. Second, using the principles of cost-sensitive learning, we shed light on the link between the (statistical) accuracy of scorecard predictions and the business value of a scorecard. This offers some guidance whether deploying advanced ï more accurate ï classification models is economically sensible. Third, we examine the correspondence between empirical results obtained using different accuracy indicators. In particular, we clarify the reliability of scorecard comparisons in the light of recently identified limitations of the area under a receiver operating characteristics curve (Hand, 2009; Hand & Anagnostopoulos, 2013). Finally, we illustrate the use of advanced nonparametric testing procedures to secure empirical findings and, thereby, offer guidance how to organize future classifier comparisons.

In the remainder of the paper we first review related work in Section 2. We then summarize the classifiers that we compare (Section 3) and describe our experimental design (Section 4). Next, we discuss empirical results (Section 5). Section 6 concludes the paper. The online appendix[3] provides a detailed description of the classification algorithms and additional results.

2 Literature review

Much literature explores the development, application, and evaluation of predictive decision support models in the credit industry (see, Crook, et al., 2007; Kumar & Ravi, 2007 for reviews). Such models estimate credit worthiness based on a set of explanatory variables. Corporate risk models employ data from balance sheets, financial ratios, or macro-economic indicators, whereas retail models use data from application forms, customer demographics, and transactional data from the customer history (e.g., Thomas, 2010). The differences between the types of variables suggest that specific modeling challenges arise in consumer as opposed to corporate credit scoring. Thus, many studies focus on either the corporate or the retail business. The latter is the focus of this paper.

A variety of prediction tasks arise in consumer credit risk modeling. The Basel II Capital Accord requires financial institutions to estimate, respectively, the probability of default (PD), the exposure at default (EAD), and the loss given default (LGD). EAD and LGD models have recently become a popular research topic (e.g., Calabrese, 2014; Yao, et al., 2015). PD models

[3] Available at: (URL will be inserted by Elsevier when available)

are especially well researched and continue to attract much interest. Topical research question include, for example, how to update PD scorecards in the face of new information (Hofer, 2015; Sohn & Ju, 2014). The prevailing methods to develop PD models are classification and survival analysis. The latter facilitates estimating not only whether but also when a customer defaults (e.g., Tong, et al., 2012). In addition, a special type of survival model called mixture cure model facilitates predicting multiple events of interest; for example default and early repayment (e.g., Dirick, et al., 2015; Liu, et al., 2015). Classification analysis, on the other hand, represents the classic approach and benefits from an unmatched variety of modeling methods.

We concentrate on PD modeling using classification analysis. Table 1 examines previous work in this field. To confirm the need for an update of Baesens, et al. (2003), we focus on empirical classifier evaluations published in 2003 or thereafter and analyze three characteristics of such studies: the type of credit scoring data, the employed classification algorithms, and the indicators used to assess these algorithms. With respect to classification algorithms, Table 1 clarifies the extent to which advanced classifiers have been considered in the literature. We pay special attention to ensemble classifiers, which Baesens, et al. (2003) do not cover.

TABLE 1: ANALYSIS OF CLASSIFIER COMPARISONS IN RETAIL CREDIT SCORING

Retail credit scoring study (in chronological order)	Data*			Classifiers**					Evaluation***			
	No. of data sets	Observations/variables per data set		No. of classifiers	ANN	SVM	ENS	S-ENS	TM	AUC	*H*	ST
(Baesens, et al., 2003)	8	4,875	21	17	X	X			X	X		P
(Malhotra & Malhotra, 2003)	1	1,078	6	2	X				X			P
(Atish & Jerrold, 2004)	2	610	16	5	X				X	X		P
(He, et al., 2004)	1	5,000	65	4	X				X			
(Lee & Chen, 2005)	1	510	18	5	X				X			
(Hand, et al., 2005)	1	1,000	20	4	X		X					
(Ong, et al., 2005)	2	845	17	6	X				X			
(West, et al., 2005)	2	845	19	4	X		X		X			P
(Y.-M. Huang, et al., 2006)	1	10,000	n.a.	10	X				X			
(Lee, et al., 2006)	1	8,000	9	5	X				X			
(S.-T. Li, et al., 2006)	1	600	17	2	X	X			X			P
(Xiao, et al., 2006)	3	972	17	13	X	X	X		X			P
(C.-L. Huang, et al., 2007)	2	845	19	4		X			X			F
(Yang, 2007)	2	16,817	85	3		X			X			
(H. Abdou, et al., 2008)	1	581	20	6	X				X			A
(Sinha & Zhao, 2008)	1	220	13	7	X	X			X	X		A

(C.-F. Tsai & Wu, 2008)	3	793	16	3	X		X		X			P
(Xu, et al., 2009)	1	690	15	4		X			X			
(Yu, et al., 2008)	1	653	13	7			X	X	X			
(H. A. Abdou, 2009)	1	1,262	25	3					X			
(Bellotti & Crook, 2009)	1	25,000	34	4		X				X		
(Chen, et al., 2009)	1	2,000	15	5		X			X			
(Nanni & Lumini, 2009)	3	793	16	16	X	X	X		X	X		
(Ġuġterġi¦, et al., 2009)	1	581	84	2	X				X			
(M.-C. Tsai, et al., 2009)	1	1,877	14	4	X				X			Q
(Yu, et al., 2009)	3	959	16	10	X	X	X		X	X		P
(J. Zhang, et al., 2009)	1	1,000	102	4					X			
(Hsieh & Hung, 2010)	1	1,000	20	4	X	X	X			X		
(Martens, et al., 2010)	1	1,000	20	4		X			X			
(Twala, 2010)	2	845	18	5			X		X			
(Yu, et al., 2010)	1	1,225	14	8	X	X	X		X			P
(D. Zhang, et al., 2010)	2	845	17	11	X	X	X		X			
(Zhou, et al., 2010)	2	1,113	17	25	X	X	X	X	X			
(J. Li, et al., 2011)	2	845	17	11		X			X			
(Finlay, 2011)	2	104,649	47	18	X		X		X			P
(Ping & Yongheng, 2011)	2	845	17	4	X	X			X			
(Wang, et al., 2011)	3	643	17	13	X	X	X		X			
(Yap, et al., 2011)	1	2,765	4	3					X			
(Yu, et al., 2011)	2	845	17	23	X	X			X			
(Akkoc, 2012)	1	2,000	11	4	X				X	X		
(Brown & Mues, 2012)	5	2,582	30	9	X	X	X			X		F/P
(Hens & Tiwari, 2012)	2	845	19	4		X			X			
(S. Li, et al., 2012)	2	672	15	5		X	X		X			
(Marqués, et al., 2012a)	4	836	20	35	X	X	X		X			F/P
(Marqués, et al., 2012b)	4	836	20	17	X	X	X		X	X		F/P
(Kruppa, et al., 2013)	1	65,524	17	5			X			X		
(Abellán & Mantas, 2014)	3	793	16	5	X		X			X		A
(C.-F. Tsai, 2014)	3	793	16	21	X		X		X			F/P
Mean / counts	1.9	6,167	24	7.8	30	24	18	3	40	10	0	17

* We report the mean of observations and independent variables for studies that employ multiple data sets. Eight studies mix retail and corporate credit data. Table 1 considers the retail data sets only.

** Abbreviations have the following meaning: ANN=Artificial neural network, SVM=Support vector machine, ENS=Ensemble classifier, S-ENS=Selective Ensemble (e.g., Partalas, et al., 2010).

*** Abbreviations have the following meaning: TM=Threshold metric (e.g., classification error, true positive rate, costs, etc.), AUC=Area under receiver operating characteristics curve, *H*=H-measure (Hand, 2009), ST=Statistical hypothesis testing. We use the following codes to report the type of statistical test used for classifier comparisons: P=Pairwise comparison (e.g., paired *t*-test), A=Analysis of variance, F=Friedman test, F/P=Friedman test together with post-hoc test (e.g., Demġar, 2006), Q=Pressôs Q statistic.

Five conclusions emerge from Table 1. First, it is common practice to use a small number of data sets (1.9 on average), many of which contain only few cases and/or independent variables. This appears inappropriate. Using multiple data sets (e.g., data from different

companies) facilitates examining the robustness of a scorecard toward environmental conditions. Also, real-world credit data sets are typically large and high-dimensional. The data used in classifier comparisons should be similar to ensure the external validity of empirical results (e.g., Finlay, 2011; Hand, 2006).

Second, the number of classifiers per study varies considerably. This can be explained with common research setups. Studies with fewer classifiers propose a novel algorithm and compare it to some reference methods (e.g., Abellán & Mantas, 2014; Akkoc, 2012; Yang, 2007). Studies with several classifiers often pair algorithms and ensemble strategies in a factorial design (e.g., Marqués, et al., 2012a; Nanni & Lumini, 2009; Wang, et al., 2011). Both setups have limitations. The latter focuses on preselected methods and omits a systematic comparison of several state-of-the-art classifiers. Studies that introduce novel classifiers may be over-optimistic because i) the developers of a new method are more adept with their approach than external users, and ii) the new method may have been tuned more intensively than reference methods (Hand, 2006; Thomas, 2010). Independent benchmarks complement the other setups in that they compare many classifiers without prior hypotheses which method excels.

Third, most studies rely on a single performance measure or measures of the same type. In general, performance measures split into three types. Those that assess the discriminatory ability of the scorecard (e.g., AUC); those that assess the accuracy of the scorecardôs probability predictions (e.g., Brier Score) and those that assess the correctness of the scorecardsô categorical predictions (e.g., classification error). Different types of indicators embody a different notion of classifier performance. Few studies mix evaluation measures from different categories. For example, none of the reviewed studies uses the Brier Score to assess the accuracy of probabilistic predictions. This misses an important aspect of scorecard performance because financial institutions require PD estimates that are not only accurate but also well calibrated. Furthermore, no previous study uses the *H*-measure, although it overcomes conceptual shortcomings of the AUC (Hand, 2009). It is thus beneficial to also consider the *H*-measure in classifier comparisons and, more generally, to assess scorecards with conceptually different performance measures.

Fourth, statistical hypothesis testing is often neglected or employed inappropriately. Common mistakes include using parametric tests (e.g., the *t*-test) or performing multiple comparisons without controlling the family-wise error level (shown by a óPô in the last column of Table 1). The approaches are inappropriate because the assumptions of parametric tests are violated in classifier comparisons (Demġar, 2006). Similarly, pairwise comparisons

without *p*-value adjustment increase the actual probability of Type-I errors beyond the stated level of | (e.g., García, et al., 2010).

Last, only two studies employ selective ensembles and they use rather simple approaches (Yu, et al., 2008; Zhou, et al., 2010). Selective ensembles are an active field of research and have shown promising results (e.g., Partalas, et al., 2010). The lack of a systematic evaluation of selective ensembles in credit scoring might thus be an important research gap.

From the literature review, we conclude that an update of Baesens, et al. (2003) is needed. This study overcomes several of the above issues through i) conducting a large-scale comparison of many established and novel classifiers including selective ensembles, ii) using multiple data sets of considerable size, iii) considering several conceptually different performance criteria, and iv) using suitable statistical testing procedures.

3 Classification algorithms for scorecard construction

We illustrate the development of a credit scorecard in the context of application scoring. Let ● سي ώ๐ℎώ.ℎ ϐ ℎώ๔ ᴺ ᴙ⍩ be an *m*-dimensional vector with application characteristics, and let ώ ᴺ ظطρ⍞طρ be a binary variable that distinguishes good ώ سي ظطρ and bad loans ώ سي ططρ . A scorecard estimates the (posterior) probability ή ططʂ●ₜ that a default event will be observed for loan *i*; where ή ططʂ● is a shorthand form of ή ώ سي ططρʂ● . To decide on an application, a credit analyst compares the estimated default probability to a threshold †; approving the loan if ή ططʂ● جى †, and rejecting it otherwise. The task to estimate ή ططʂ● belongs to the field of classification (e.g., Hand & Henley, 1997). A scorecard is a classification model that results from applying a classification algorithm to a data set Ὸ سي ώₜℎ●ₜ ₹₥๐ of past loans.

This study compares 41 different classification algorithms. Our selection draws inspiration from previous studies (e.g., Baesens, et al., 2003; Finlay, 2011; Verbeke, et al., 2012) and covers several different approaches (linear/nonlinear, parametric/non-parametric, etc.). The algorithms split into individual and ensemble classifiers. The eventual scorecard consists of a single classification model in the first group. Ensemble classifiers integrate the prediction of multiple models, called base models. We distinguish homogeneous ensembles, which create the base models using the same algorithm, and heterogeneous ensembles, which employ different algorithms. Figure 1 illustrates the modeling process using different classifiers.

decision costs. Let ő ﻃﻲﻎﻃ be the opportunity costs that result from denying credit to a good risk. Similarly, let ő ﻲﻎﻃ be the costs of granting credit to a bad risk (e.g., net present value of EAD*LGD ï interests paid prior to default). Then, we can calculate the error costs of a scorecard, C(*s*), as:

$$\text{ő í سي ő ﻃﻲﻎﻃ } z \,\&02\, \text{ط ő ﻲﻎﻃ } z \,\&.\,2 \qquad (2)$$

Given that a scorecard produces probability estimates ὴ ﻃﻎ● , FPR and FNR depend on the threshold †. Bayesian decision theory suggests that an optimal threshold depends on the prior probabilities of good and bad risks and their corresponding misclassification costs (e.g., Viaene & Dedene, 2004). To cover different scenarios, we consider 25 cost ratios in the interval ő ﻃﻲﻎﻃ ╢ő ﻲﻎﻃ سي ρ╢ς╫ϑ╫ρ╢υπ, always assuming that it is more costly to grant credit to a bad risk than rejecting a good application (e.g., Thomas, et al., 2002). Note that fixing ő ﻃﻲﻎﻃ at one does not constrain generality (e.g., Hernández-Orallo, et al., 2011). For each cost setting and credit scoring data set, we i) compute the misclassification costs of a scorecard from (2), ii) estimate expected error costs through averaging over data sets, and iii) normalize costs such that they represent percentage improvements compared to LR. Figure 2 depicts the corresponding results.

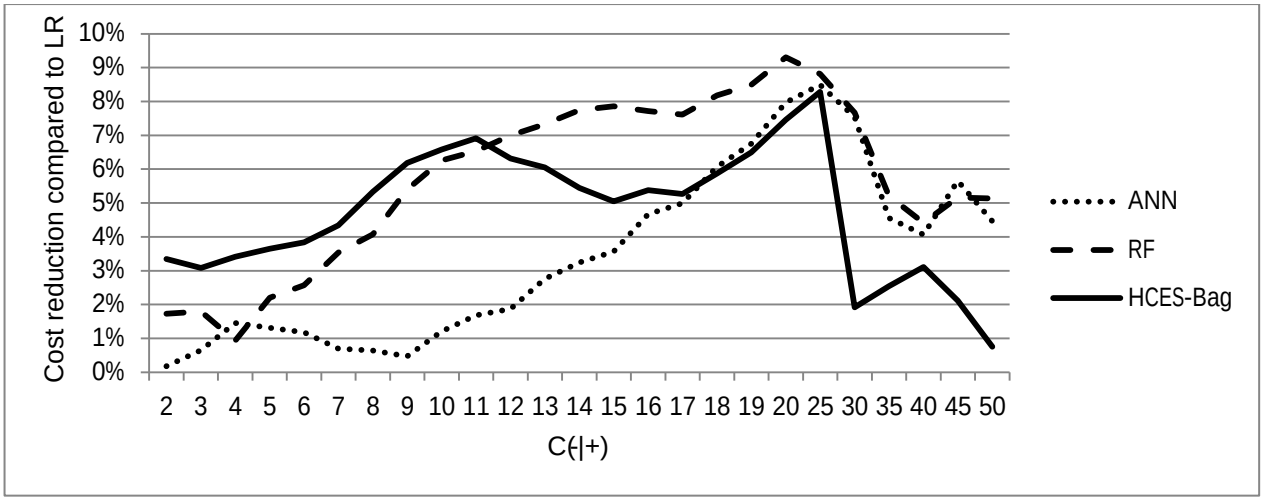

Figure 2: Expected percentage reduction in error costs compared to LR across different settings for ő ﻃﻲﻎﻃ assuming ő ﻃﻲﻎﻃ سي ρ and using a Bayes optimal threshold.

Figure 2 reveals that the considered classifiers can substantially reduce the error costs of a LR-based scorecard. For example, the average improvements (across all cost settings) of ANN, RF, and HCES-Bag over LR are, respectively, 3.4%, 5.7%, and 4.8%. Improvements of multiple percent are meaningful from a managerial point of view, especially when considering the large number of decisions that scorecards support in the financial industry. Another result is that the most accurate classifier, HCES-Bag loses its advantage when the cost of misclassifying bad credit risks increases. This shows that the link between (statistical)

accuracy and business value is far from perfect. The most accurate classifier does not necessarily give the most profitable scorecard.

RF and ANN achieve a larger cost reduction than HCES-Bag when misclassifying a bad risk is eleven and eighteen times more expensive than the opposite error, respectively. Using a Bayes-optimal threshold, higher costs of misclassifying a bad risk lower the threshold and thus the acceptance rate. Hence, incorrect rejections of actually good risks become the main determinant of the error costs of a scorecard. This suggests that the partial superiority of RF (and ANN) over HCES-Bag results from the latter producing too conservative predictions for clients with low credit risk. It could be interesting to examine whether this pattern persists if HCES-Bag were setup to minimize error costs directly (i.e., within ensemble selection). We leave this test to future research.

5.4 Correspondence of classifier performance across performance measures

Given that many previous studies have used a small number of accuracy indicators, it is interesting to examine the dependency of observed results on the chosen indicator. Moreover, such an analysis can add some empirical evidence to the recent debate whether and when the AUC is a suitable measure to compare different classifiers and retail scorecards in particular (e.g., Hand & Anagnostopoulos, 2013; Hernández-Orallo, et al., 2011).

Table 6 depicts the agreement of classifier rankings across accuracy indicators using *Kendall*ôs rank correlation coefficient. With respect to the AUC, we find that empirical results do not differ much between this measure and the *H*-measure (correlation: .93). Thus, if a credit analyst were to choose a scorecard among alternatives, the AUC and the *H*-measure would typically give similar recommendations. In fact, Table 6 supports generalizing this view even further. Pairwise correlations around .90 indicate high similarity between classifier ranks in terms of the KS and the PCC with those of the AUC and the *H*-measure. Despite substantial conceptual differences between these measures (e.g., local versus global assessment; see Section 4.3), they rank classifiers rather similarly. Therefore, it appears sufficient to use one of them in empirical classifier comparisons.

A different conclusion emerges for the BS and the PG. Using the same measurement approach as the AUC, the PG emphasizes the accuracy of a scorecard in the most important segment of the score distribution. Our results confirm that this captures a different aspect of performance. For example, the AUC is notably less correlated with the PG than with the *H*-measure. However, we observe the smallest correlation between the BS and the other measures. The BS is the only indicator that assesses the accuracy of probability estimates.

Table 6 reveals that this notion of performance contributes useful information to a classifier comparison over and above those captured in the AUC, PCC, *H*-measure, and KS.

Based on Table 6 we recommend that future studies use at least three performance measures: the AUC, the PG, and the BS, whereby one could replace the AUC with the *H*-measure. The PG and the BS offer an additional angle from which to examine predictive accuracy. Thus, they should routinely be part of scorecard comparisons.

TABLE 6: CORRELATION OF CLASSIFIER RANKINGS ACROSS PERFORMANCE MEASURES

	AUC	PCC	BS	H	PG	KS
AUC	1.00					
PCC	.88	1.00				
BS	.54	.54	1.00			
H	.93	.91	.56	1.00		
PG	.79	.72	.51	.76	1.00	
KS	.92	.89	.54	.91	.79	1.00

6 Conclusions

We set out to update Baesens, et al. (2003) and to explore the relative effectiveness of alternative classification algorithms in retail credit scoring. To that end, we compared 41 classifiers in terms of six performance measures across eight real-world credit scoring data sets. Our results suggest that several classifiers predict credit risk significantly more accurately than the industry standard LR. Especially heterogeneous ensembles classifiers perform well. We also provide some evidence that more accurate scorecards facilitate sizeable financial returns. Finally, we show that several common performance measures give similar signals as to which scorecard is most effective, and recommend the use of two rarely employed measures that contribute additional information.

Our study consolidates previous work in PD modeling and provides a holistic picture of the state-of-the-art in predictive modeling for retail scorecard development. This has implications for academia and industry. From an academic point of view, an important question is whether efforts into the development of novel scoring techniques are worthwhile. Our study provides some support but also raises concerns. We find some advanced methods to perform extremely well on our credit scoring data sets, but never observe the most recent classifiers to excel. ANNs perform better than ELMs, RF better than RotFor, and dynamic selective ensembles worse than almost all other classifiers. This may indicate that progress in the field has stalled (e.g., Hand, 2006), and that the focus of attention should move from PD

models to other modeling problems in the credit industry including data quality, scorecard recalibration, variable selection, and LGD/EAD modeling.

On the other hand, we do not expect the desire to develop better, more accurate scorecards to end any time soon. Likely, future papers will propose novel classifiers and the ñsearch for the silver bulletò (Thomas, 2010) will continue. An implication of our study is that such efforts must be accompanied by a rigorous assessment of the proposed method vis-à-vis challenging benchmarks. In particular, we recommend RF as benchmark against which to compare new classification algorithms. HCES-Bag might be even more difficult to outperform, but is not as easily available in standard software. Furthermore, we caution against the practice to compare a newly proposed classifier to LR (or some other individual classifier) only, which we still observe in the literature. LR is the industry standard and it is useful to examine how a new classifier compares to this approach. However, given the state-of-the-art, outperforming LR can no longer be accepted as a signal of methodological advancement.

An important question to be answered in future research is whether the characteristics of a classification algorithm and a data set facilitate appraising the classifierôs suitability for this data set a priori. We have identified classifiers that work well for PD modeling, but cannot *explain* their success. Nonetheless, our benchmark can be seen as a first step toward gaining explanatory insight in that it provides an empirical fundament for meta-analytic research. For example, gathering features of individual classifiers and characteristics of the credit scoring data sets, and using these as covariates in a regression framework to explain classifier performance (as dependent variable) could help to uncover the underlying drivers of classifier efficacy in credit scoring.

From a managerial perspective, it is important to reason whether the superior performance that we observe for some classifiers generalizes to real-world applications, and to what extent their adoption would increase returns. These questions are much debated in the literature (e.g., Finlay, 2011). From this study, we can add some points to the discussion.

First, we show that advancements in computer power, classifier learning, and statistical testing facilitate rigorous classifier comparisons. This does not guarantee external validity. Several concerns why laboratory experiments (as this one) may overestimate the advantage of advanced classifiers remain valid; and might be insurmountable (e.g., Hand, 2006). However, experimental designs with several cross-validation repetitions, different performance measures, and appropriate multiple-comparison procedures overcome some limitations of

previous studies and, thereby, provide stronger support that advanced classifiers have the potential to increase predictive accuracy not only in the laboratory but also in industry.

Second, our results facilitate some remarks related to the organizational acceptance of advanced classifiers. In particular, a lack of acceptance can result from concerns that much expertise is needed to handle such classifiers. Our results show that this is not the case. The accuracy differences that we observe result from a fully-automatic modeling approach. Consequently, certain advanced classifiers do not require human intervention to predict significantly more accurately than simpler alternatives. Furthermore, the current interest in Big Data indicates a shift toward a data-driven decision making paradigm among managers. This might further increase the acceptability of advanced scoring methods.

Finally, the business value of more accurate scorecard predictions is a crucial issue. Our preliminary simulation provides some evidence that the ñhigher (statistical) accuracy equals more profit equationò might hold. Furthermore, retail scorecards support a vast number of business decisions. Consider for example the credit card industry or scoring tasks in online settings. In such environments, one-time investments (e.g., for hardware, software, and user training) into a more elaborate scoring technique will pay-off in the long run when small but significant accuracy improvements are multiplied by hundreds of thousands of scorecard applications. The difficulties of introducing advanced scoring methods including ensemble models are more psychological than business related. Using a large number of models, a significant minority of which give contradictory answers, is counterintuitive to many business leaders. Such organizations will need to experiment fully before accepting a change from the historic industry standard procedures.

Regulatory frameworks and organizational acceptance constrain and sometimes prohibit the use of advanced scoring techniques today; at least for classic credit products. However, given the current interest in data-centric decision aids and the richness of online-mediated forms of credit granting, we foresee a bright future for advanced scoring methods in credit scoring.

References

Abdou, H., Pointon, J., & El-Masry, A. (2008). Neural nets versus conventional techniques in credit scoring in Egyptian banking. *Expert Systems with Applications, 35*, 1275-1292.

Abdou, H. A. (2009). Genetic programming for credit scoring: The case of Egyptian public sector banks. *Expert Systems with Applications, 36*, 11402-11417.

Abellán, J., & Mantas, C. J. (2014). Improving experimental studies about ensembles of classifiers for bankruptcy prediction and credit scoring. *Expert Systems with Applications, 41*, 3825-3830.

Akkoc, S. (2012). An empirical comparison of conventional techniques, neural networks and the three stage hybrid Adaptive Neuro Fuzzy Inference System (ANFIS) model for credit scoring analysis: The case of Turkish credit card data. *European Journal of Operational Research, 222*, 168-178.

Andreeva, G. (2006). European generic scoring models using survival analysis. *Journal of the Operational Research Society, 57*, 1180-1187.
Atish, P. S., & Jerrold, H. M. (2004). Evaluating and tuning predictive data mining models using receiver operating characteristic curves. *Journal of Management Information Systems, 21*, 249-280.
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*, 627-635.
Bellotti, T., & Crook, J. (2009). Support vector machines for credit scoring and discovery of significant features. *Expert Systems with Applications, 36*, 3302ï3308.
Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*, 123-140.
Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5-32.
Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications, 39*, 3446-3453.
Calabrese, R. (2014). Downturn loss given default: Mixture distribution estimation. *European Journal of Operational Research, 237*, 271-277.
Caruana, R., Munson, A., & Niculescu-Mizil, A. (2006). Getting the Most Out of Ensemble Selection. In *Proc. of the 6th Intern. Conf. on Data Mining* (pp. 828-833). Hong Kong, China: IEEE Computer Society.
Chen, W., Ma, C., & Ma, L. (2009). Mining the customer credit using hybrid support vector machine technique. *Expert Systems with Applications, 36*, 7611-7616.
Crone, S. F., Lessmann, S., & Stahlbock, R. (2006). The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing. *European Journal of Operational Research, 173*, 781-800.
Crook, J. N., Edelman, D. B., & Thomas, L. C. (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research, 183*, 1447-1465.
Demġar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research, 7*, 1-30.
Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning. *Neural Computation, 10*, 1895-1923.
Dirick, L., Claeskens, G., & Baesens, B. (2015). An Akaike information criterion for multiple event mixture cure models. *European Journal of Operational Research, 241*, 449-457.
Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*, 861-874.
Finlay, S. (2009). Credit scoring for profitability objectives. *European Journal of Operational Research, 202*, 528-537.
Finlay, S. (2011). Multiple classifier architectures and their application to credit risk assessment. *European Journal of Operational Research, 210*, 368-378.
Freund, Y., & Schapire, R. E. (1996). Experiments With a New Boosting Algorithm. In L. Saitta (Ed.), *Proc. of the 13th Intern. Conf. on Machine Learning* (pp. 148-156). Bari, Italy: Morgan Kaufmann.
Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis, 38*, 367-378.
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*, 2044-2064.
García, S., & Herrera, F. (2008). An extension on "Statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons. *Journal of Machine Learning Research, 9*, 2677-2694.
Gong, R., & Huang, S. H. (2012). A KolmogorovïSmirnov statistic based segmentation approach to learning from imbalanced datasets: With application in property refinance prediction. *Expert Systems with Applications, 39*, 6192-6200.
Guang-Bin, H., Lei, C., & Chee-Kheong, S. (2006). Universal approximation using incremental constructive feedforward networks with random hidden nodes. *IEEE Transactions on Neural Networks, 17*, 879-892.
Hall, M. A. (2000). Correlation-Based Feature Selection for Discrete and Numeric Class Machine Learning. In P. Langley (Ed.), *Proc. of the 17th Intern. Conf. on Machine Learning* (pp. 359-366). Stanford, CA, USA: Morgan Kaufmann
Hand, D. J. (2005). Good practice in retail credit scorecard assessment *Journal of the Operational Research Society, 56*, 1109-1117.
Hand, D. J. (2006). Classifier technology and the illusion of progress. *Statistical Science, 21*, 1-14.
Hand, D. J. (2009). Measuring classifier performance: A coherent alternative to the area under the ROC curve. *Machine Learning, 77*, 103-123.
Hand, D. J., & Anagnostopoulos, C. (2013). When is the area under the receiver operating characteristic curve an appropriate measure of classifier performance? *Pattern Recognition Letters, 34*, 492-495.
Hand, D. J., & Henley, W. E. (1997). Statistical classification models in consumer credit scoring: A review. *Journal of the Royal Statistical Society: Series A (General), 160*, 523-541.
Hand, D. J., Sohn, S. Y., & Kim, Y. (2005). Optimal bipartite scorecards. *Expert Systems with Applications, 29*, 684-690.

He, J., Shi, Y., & Xu, W. (2004). Classifications of Credit Cardholder Behavior by Using Multiple Criteria Non-linear Programming. In Y. Shi, W. Xu & Z. Chen (Eds.), *Data Mining and Knowledge Management, Chinese Academy of Sciences Symposium* (Vol. 3327, pp. 154-163). Beijing, China: Springer.

Hens, A. B., & Tiwari, M. K. (2012). Computational time reduction for credit scoring: An integrated approach based on support vector machine and stratified sampling method. *Expert Systems with Applications, 39*, 6774-6781.

Hernández-Orallo, J., Flach, P. A., & Ramirez, C. F. (2011). Brier Curves: A New Cost-Based Visualisation of Classifier Performance. In L. Getoor & T. Scheffer (Eds.), *Proc. of the 28th Intern. Conf. on Machine Learning* (pp. 585-592). Bellevue, WA, USA: Omnipress.

Hofer, V. (2015). Adapting a classification rule to local and global shift when only unlabelled data are available. *European Journal of Operational Research, 243*, 177-189.

Hsieh, N.-C., & Hung, L.-P. (2010). A data driven ensemble classifier for credit scoring analysis. *Expert Systems with Applications, 37*, 534-545.

Huang, C.-L., Chen, M.-C., & Wang, C.-J. (2007). Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications, 33*, 847-856.

Huang, Y.-M., Hunga, C.-M., & Jiau, H. C. (2006). Evaluation of neural networks and data mining methods on a credit assessment task for class imbalance problem. *Nonlinear Analysis: Real World Applications, 7*, 720-747.

Ko, A. H. R., Sabourin, R., & Britto, J. A. S. (2008). From dynamic classifier selection to dynamic ensemble selection. *Pattern Recognition, 41*, 1735-1748.

Kruppa, J., Schwarz, A., Arminger, G., & Ziegler, A. (2013). Consumer credit risk: Individual probability estimates using machine learning. *Expert Systems with Applications, 40*, 5125-5131.

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

Lee, T.-S., & Chen, I. F. (2005). A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. *Expert Systems with Applications, 28*, 743-752.

Lee, T.-S., Chiu, C.-C., Chou, Y.-C., & Lu, C.-J. (2006). Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. *Computational Statistics & Data Analysis, 50*, 1113-1130.

Li, J., Wei, L., Li, G., & Xu, W. (2011). An evolution strategy-based multiple kernels multi-criteria programming approach: The case of credit decision making. *Decision Support Systems, 51*, 292-298.

Li, S.-T., Shiue, W., & Huang, M.-H. (2006). The evaluation of consumer loans using support vector machines. *Expert Systems with Applications, 30*, 772-782.

Li, S., Tsang, I. W., & Chaudhari, N. S. (2012). Relevance vector machine based infinite decision agent ensemble learning for credit risk analysis. *Expert Systems with Applications, 39*, 4947-4953.

Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml/]. Irvine, CA: University of California, School of Information and Computer Science. Last accessed: 2015-02-16

Liu, F., Hua, Z., & Lim, A. (2015). Identifying future defaulters: A hierarchical Bayesian method. *European Journal of Operational Research, 241*, 202-211.

Malhotra, R., & Malhotra, D. K. (2003). Evaluating consumer loans using neural networks. *Omega, 31*, 83-96.

Marqués, A. I., García, V., & Sánchez, J. S. (2012a). Exploring the behaviour of base classifiers in credit scoring ensembles. *Expert Systems with Applications, 39*, 10244-10250.

Marqués, A. I., García, V., & Sánchez, J. S. (2012b). Two-level classifier ensembles for credit risk assessment. *Expert Systems with Applications, 39*, 10916-10922.

Martens, D., Van Gestel, T., De Backer, M., Haesen, R., Vanthienen, J., & Baesens, B. (2010). Credit rating prediction using Ant Colony Optimization. *Journal of the Operational Research Society, 61*, 561-573.

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

Ong, C.-S., Huanga, J.-J., & Tzeng, G.-H. (2005). Building credit scoring models using genetic programming *Expert Systems with Applications, 29*, 41-47.

Paleologo, G., Elisseeff, A., & Antonini, G. (2010). Subagging for credit scoring models. *European Journal of Operational Research, 201*, 490-499.

Partalas, I., Tsoumakas, G., & Vlahavas, I. (2009). Pruning an ensemble of classifiers via reinforcement learning. *Neurocomputing, 72*, 1900-1909.

Partalas, I., Tsoumakas, G., & Vlahavas, I. (2010). An ensemble uncertainty aware measure for directed hill climbing ensemble pruning. *Machine Learning, 81*, 257-282.

Ping, Y., & Yongheng, L. (2011). Neighborhood rough set and SVM based hybrid credit scoring classifier. *Expert Systems with Applications, 38*, 11300-11304.

Platt, J. C. (2000). Probabilities for Support Vector Machines. In A. Smola, P. Bartlett, B. Schölkopf & D. Schuurmans (Eds.), *Advances in Large Margin Classifiers* (pp. 61-74). Cambridge: MIT Press.

Pundir, S., & Seshadri, R. (2012). A novel concept of partial lorenz curve and partial gini index. *International Journal of Engineering ,Science and Innovative Technology, 1*, 296-301.

Rodriguez, 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*, 1619-1630.
Sinha, A. P., & Zhao, H. (2008). Incorporating domain knowledge into data mining classifiers: An application in indirect lending. *Decision Support Systems, 46*, 287-299.
So, M. M. C., & Thomas, L. C. (2011). Modelling the profitability of credit cards by Markov decision processes. *European Journal of Operational Research, 212*, 123-130.
Sohn, S. Y., & Ju, Y. H. (2014). Updating a credit-scoring model based on new attributes without realization of actual data. *European Journal of Operational Research, 234*, 119-126.
Ġuġterġiļ, M., Mramor, D., & Zupan, J. (2009). Consumer credit scoring models with limited data. *Expert Systems with Applications, 36*, 4736-4744.
Thomas, L. C. (2010). Consumer finance: Challenges for operational research. *Journal of the Operational Research Society, 61*, 41-52.
Thomas, L. C., Edelman, D. B., & Crook, J. N. (2002). *Credit Scoring and its Applications*. Philadelphia: Siam.
Tong, E. N. C., Mues, C., & Thomas, L. C. (2012). Mixture cure models in credit scoring: if and when borrowers default. *European Journal of Operational Research, 218*, 132-139.
Tsai, C.-F. (2014). Combining cluster analysis with classifier ensembles to predict financial distress. *Information Fusion, 16*, 46-58.
Tsai, C.-F., & Wu, J.-W. (2008). Using neural network ensembles for bankruptcy prediction and credit scoring. *Expert Systems with Applications, 34*, 2639-2649.
Tsai, M.-C., Lin, S.-P., Cheng, C.-C., & Lin, Y.-P. (2009). The consumer loan default predicting model - An application of DEA-DA and neural network. *Expert Systems with Applications, 36*, 11682-11690.
Twala, B. (2010). Multiple classifier application to credit risk assessment. *Expert Systems with Applications, 37*, 3326-3336.
Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research, 218*, 211-229.
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*, 505-513.
Viaene, S., & Dedene, G. (2004). Cost-sensitive learning and decision making revisited. *European Journal of Operational Research, 166*, 212-220.
Wang, G., Hao, J., Ma, J., & Jiang, H. (2011). A comparative assessment of ensemble learning for credit scoring. *Expert Systems with Applications, 38*, 223-230.
West, D., Dellana, S., & Qian, J. (2005). Neural network ensemble strategies for financial decision applications. *Computers & Operations Research, 32*, 2543-2559.
Woloszynski, T., & Kurzynski, M. (2011). A probabilistic model of classifier competence for dynamic ensemble selection. *Pattern Recognition, 44*, 2656-2668.
Xiao, W., Zhao, Q., & Fei, Q. (2006). A comparative study of data mining methods in consumer loans credit scoring management. *Journal of Systems Science and Systems Engineering, 15*, 419-435.
Xu, X., Zhou, C., & Wang, Z. (2009). Credit scoring algorithm based on link analysis ranking with support vector machine. *Expert Systems with Applications, 36*, 2625-2632.
Yang, Y. (2007). Adaptive credit scoring with kernel learning methods. *European Journal of Operational Research, 183*, 1521-1536.
Yao, X., Crook, J., & Andreeva, G. (2015). Support vector regression for loss given default modelling. *European Journal of Operational Research, 240*, 528-538.
Yap, B. W., Ong, S. H., & Husain, N. H. M. (2011). Using data mining to improve assessment of credit worthiness via credit scoring models. *Expert Systems with Applications, 38*, 13274-13283.
Yu, L., Wang, S., & Lai, K. K. (2008). Credit risk assessment with a multistage neural network ensemble learning approach. *Expert Systems with Applications, 34*, 1434-1444.
Yu, L., Wang, S., & Lai, K. K. (2009). An intelligent-agent-based fuzzy group decision making model for financial multicriteria decision support: The case of credit scoring. *European Journal of Operational Research, 195*, 942-959.
Yu, L., Yao, X., Wang, S., & Lai, K. K. (2011). Credit risk evaluation using a weighted least squares SVM classifier with design of experiment for parameter selection. *Expert Systems with Applications, 38*, 15392-15399.
Yu, L., Yue, W., Wang, S., & Lai, K. K. (2010). Support vector machine based multiagent ensemble learning for credit risk evaluation. *Expert Systems with Applications, 37*, 1351-1360.
Zhang, D., Zhou, X., Leung, S. C. H., & Zheng, J. (2010). Vertical bagging decision trees model for credit scoring. *Expert Systems with Applications, 37*, 7838-7843.
Zhang, J., Shi, Y., & Zhang, P. (2009). Several multi-criteria programming methods for classification. *Computers & Operations Research, 36*, 823-836.

Zhou, L., Lai, K. K., & Yu, L. (2010). Least Squares Support Vector Machines ensemble models for credit scoring. *Expert Systems with Applications, 37,* 127-133.

Benchmarking state-of-the-art classification algorithms for credit scoring: Classification model development and evaluation process