7 On the Practical Relevance of Modern Machine Learning Algorithms for Credit Scoring Applications

Gero Szepannek
Stralsund University of Applied Sciences, Germany
gero.szepannek@fh-stralsund.de

Abstract

Although many new algorithms like e.g. support vector machines, boosting, random forests or neural networks have been proposed in the recent past logistic regression does still represent the gold standard in industrial praxis.

Benchmarking studies show the general superiority of flexible learning techniques that are able to detect complex structures. These studies typically restrict to the evaluation of one or several performance measures (like misclassification rate) and ignore further aspects of practical feasibility.

In this paper a critical investigation of pros and cons of modern machine learning techniques with respect to business requirements and their practical relevance is worked out. An exemplary case study based on credit scoring using random forests is executed.

Introduction

Although many new algorithms like e.g. support vector machines, boosting, random forests or neural networks have been proposed in the recent past there are several reasons why logistic regression does still represent the gold standard in industrial praxis:

1 Logistic regression is widely taught.

2 There are many software implementations of logistic regression available.

3 The resulting models are stable and no further parameter tuning is necessary/possible.

4 The results are easy to interpret.

While the first two reasons are historical and currently under change the third one refers to the necessity an additional parameter specification that is appropriate to the specific data situation. A wrong parameter specification will lead to suboptimal predictive power of the resulting model and can thus be considered as an operative risk from an economic point of view.

On the other hand, many benchmarking studies (cf. e.g. Szepannek et al., 2008, Szepannek et al., 2010, Lessmann et al., 2013) have shown some general superiority of modern flexible learning techniques in terms of quantitative performance measures, especially for complex data structures.

Figure 28 (left) shows a typical example in the credit scoring context: the relationship between age and default rate is nonlinear. A linear model like logistic regression will not be appropriate here. For

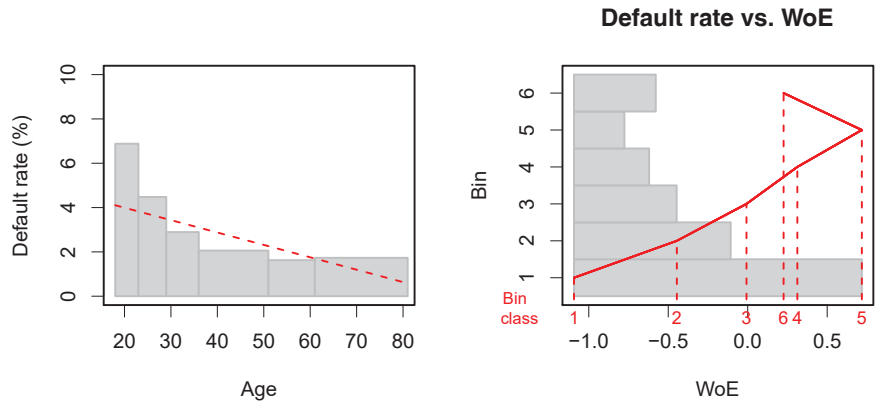

Figure 28: Example of a nonlinear relationship (left), binning and WoE transform (right).

this reason industrial praxis in credit scoring is a pre-binning of the data by the analyst. The binned variables are further used for modelling as dummy variables or transformed to **weights of evidence** (WoEs) where a new numeric variable

$$X_{WoE} = \log\left(\frac{f(x|Y=1)}{f(x|Y=0)}\right)$$

is created from the original variable X (cf. Fig. 28, right). A property of this transform is its (univariate) linearity in the logit of Y (cf. e.g. Szepannek, 2011). Such a pre-binning of the data not only allows to model nonlinear relationships using logistic regression and but also guarantees an implicit plausibility check of the data by the analyst which reduces operational risk.

But pre-binning still does not take into account nonlinear relationships between explanatory variables and moreover the manual nature of the process does still require some kind of additional automatized pre-processing if the number of variables is very large. In order to compare different methods academic benchmark studies typically set up automatized benchmark experiments and the comparison with an industrial use case is not the scope. As a consequence most benchmark studies will be biased towards algorithms that allow for a higher degree of automatization.

For this reason in this document logistic regression models both with and without pre-binning are constructed as a the baseline for comparisons. It has to be stated that manual pre-binning denotes some loss in scientific rigor which is accepted in order to increase practical relevance of the results.

An experiment is set up based on real world data from the credit scoring business context. Random forests (Breiman, 2001) are selected as they represent one of the most popular modern machine learning techniques. In contrast to typical benchmark studies not only the performance of an optimally tuned parameter set is analyzed but also the performance over the whole range of parameter combinations. As a consequence the study concentrates on random forests and no further classifiers are investigated.

In Subsec. 7.1 the parameters of random forests are discussed and in Subsec. 7.2 the experiment is presented. Several questions are analyzed to test the practical benefit of modern machine learning techniques:

- Potential benefit vs. the risk of a performance decrease
- Investigation of random forest parameters
- Cost benefit considerations: time to invest
- Evaluation of the improvements from a business perspective.
- The effect of the analyst
- Interpretation of the model

The results are given in Subsec. 7.3 and finally a summary is given in Subsec. 7.4.

7.1 Random Forests and Parameter Tuning

Random Forests (Breiman, 2001) are designed to avoid shortcomings w.r.t. bias and variance of either small or large single decision trees: a set of large trees (with small bias) is built on bootstrap samples. The variance of the final classifier is reduced by averaging the predictions of all trees. In addition, the single trees are further uncorrelated by allowing only for random subset of variables at each single split of each tree (cf. e.g. Segal, 2006). The **number of trees** as well as the **number of variables** that are considered for each split are two main parameters for random forest construction and optimized in most benchmark studies. A larger number of trees will generally increase computation time and improve performance but saturate. The number of randomly chosen variables should depend on the dimension of the data set: if it is too large the existence of dominant variables may lead to very similar trees and reduce the bootstrapping effect. On the other hand, selecting too few variables may be dangerous if the data consists of a large percentage of purely noisy variables w/o any information on the class.

As an extension to many studies the current experiment does not restrict on these two parameters but includes several additional parameters (for an overview see Boulesteix, 2012), namely the **minimum node size** and the **maximum number of terminal nodes**. Both parameters control the depth of single trees within the forest and should be chosen according to the philosophy of large unbiased trees.

In addition the bootstrap samples can be chosen either with or without **replacement**. A corresponding sample size w/o replacement is of $0.632\times$ the sample size. Strobl et al. (2007) describe the bias of variable importance for sampling with replacement. Finally, samples can be **balanced** w.r.t. the classes as tree splitting criteria typically are affected by the class proportions (cf. e.g. Bischl et al., 2016, Brown and Mues, 2012, Crone and Finlay, 2012). For this reason both balanced and unbalanced samples are investigated for random forest construction.

7.2 Case Study

An issue of credit scoring research is the general lack of real world data as they are confidential in general. For this study the freely available German Credit Data (Hoffmann, 1994) from the UCI

Additional debtors	Number of credits
Age	Other instalment plans
Amount	**People liable**
Credit history	**President residance**
Duration	Property
Duration employment	Purpose, product
Foreign worker	Savings
Housing, type of residence	Gender, family status
Instalment rate	Status checking account
Job	Telephone available

Table 7: Variables of the German credit data set.

Machine learning repository (Newman et al., 1998) is used. Based on one single data set the results should rather be considered as a case study without claim for generality. Bischl et al. (2016) try to overcome this issue by collecting a large number of data sets from other domains but transferability to the credit scoring context remains questionable.

The data consists of 1000 observations in two classes (default vs. non-default) which is quite few compared to real world applications. There are 20 explanatory variables. The prior probability of default is 0.3 which does not reflect typical unbalance of real world scoring data. In contrast to decision trees logistic regression is quite insensitive to class unbalance (Bischl et al., 2016, Brown and Mues, 2012, Crone and Finlay, 2012).

Table 7 summarizes the variables, **numeric** variables are in bold. Typical for business applications the variables are not all metric but most data comes along in categories further emphasizing the analyst's role in model building as the categories are w/o any order but have to be interpreted w.r.t. their meaning from a business point of view.

As a baseline logistic regression models are built with and without pre-binning of the variables. Binning is done based on manual plausibility checks of the splits generated by univariate decicsion trees with varying complexity parameters (Therneau and Atkinson, 1997). For the binned variables separate regression models are built either using dummies or WoEs.

The **Gini coefficient** $G = 2(AUC - 0.5)$ is used as a performance measure as it represents the most popular statistic to evaluate the performance of credit scoring systems in practise. Given the comparatively small sample size 10 fold cross validation is used for performance evaluation. Good practice would consist in an additional inner CV loop for parameter optimization of random forests (cf. e.g. Szepannek et al., 2010, Bischl et al., 2016). The focus of this study slightly differs from typical benchmark experiments as the entity of all models with different parameters is of interest rather than only the optimally parameterized one. For this reason no parameter optimization on an additional inner loop has been done here.

A total of 2304 random forests has been investigated based on an exhaustive parameter grid of the parameters given in table 8 (the default parameters are in bold, for the number of terminal nodes there is no restriction in the default parameterization of random forests, corresponding to a value of 500 in the setting).

In order to investigate its relevance for business a test on the achieved improvement is implemented according to Henking et al. (2006) using approximate normal distribution of the AUC with standard

Tuning Parameter	Values
Number of trees	$\{20, 50, 200, \mathbf{500}, 1000, 2000\}$
Number of variables | Split	$\{2, \mathbf{4}, 8, 16\}$
Min. node size	$\{\mathbf{1}, 5, 20, 50\}$
Max. number terminal nodes	$\{5, 10, 20, 50, 100, \mathbf{500}\}$
Sampling with replacement	$\{\mathbf{yes}, no\}$
Balanced class sizes	$\{yes, \mathbf{no}\}$

Table 8: Parameters for random forest optimization.

Pre-inning	Dummies	WoEs
Yes	57.13	57.80
No	59.06	**59.75**

Table 9: Results of logistic regression models.

error:

$$\hat{\sigma}_{AUC} = \sqrt{\frac{AUC(1-AUC) + (N_D - 1)(q_1 - AUC^2) + (N_{ND} - 1)(q_2 - AUC^2)}{N_D N_{ND}}}$$

and $q_1 = \frac{AUC}{2-AUC}$, $q_2 = \frac{2AUC^2}{1+AUC}$ as well as $N_{(N)D}$ the number of (non-)defaults in the sample.

7.3 Results

Logistic Regression: Table 9 shows the results of the logistic regression models. Both preliminary binning and WoE pre-transform of the binned data improve performance. Whereas improvement by binning might result from allowing to model nonlinearities the additional gain of using WoEs may be a consequence of the small data set as the use of dummy characteristics increases the degrees of freedom of the subsequent regression model.

Random forests and their parameterization: The best random forest model achieves a Gini coefficient of 61.53 which is no significant improvement (p value: 0.294). The optimal parameters are identical to the default ones except from the number of variables offered at each split being $2 < 4$ ($=$ floor of the $\sqrt{\cdot}$ of the number of variables in the data set). Figure 29 shows the frequency of model performance over all parameter combinations.

It can be seen that the relative improvement by using random forest models is quite small whereas the potential loss can be dramatic for a bad parameter set. Only 7.6% of the models improve logistic regression performance but 19.5% even led to significant performance decrease. This underlines the risk of blindly applying modern machine learning algorithms, instead a thorough understanding of the methodology is required. Interestingly, just using default parameterization already results in a (slight) performance increase (60.04, p value 0.465). In order to prevent from misinterpreting the results it has to be remarked that the frequencies in Figure 29 result from an experimental design and do not reflect a distribution as the figure might suggest. The results are further biased by the attempt to investigate a quite exhaustive parameter grid whereof some combinations turn out to be avoidable, leading to an analysis of the impact of the single parameters. Figure 30 shows the performance as a function of the parameters and levels in an OFAT sense: Mainly, the considerations of Subsec. 7.1 are confirmed as both small (flat) base single trees as well as a number of trees that is too small

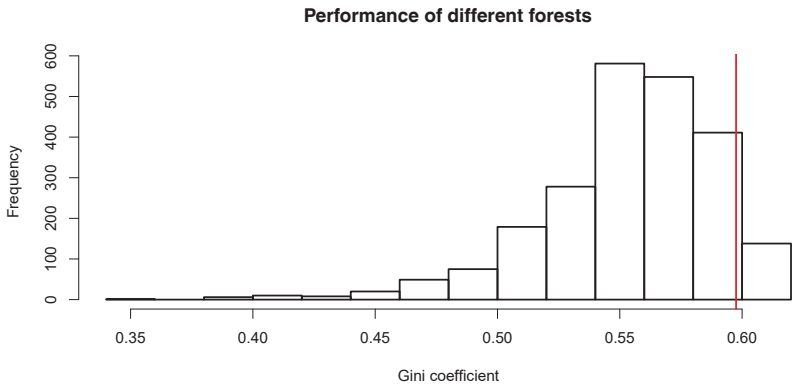

Figure 29: Results over all random forests (red line: logistic regression baseline).

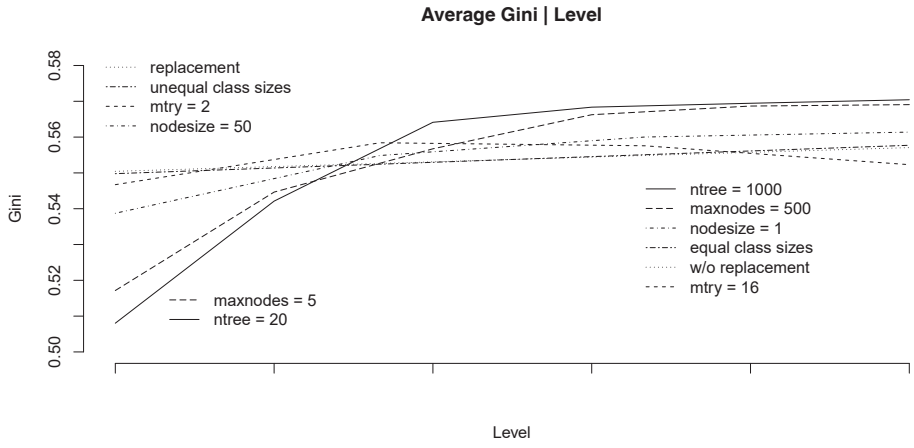

Figure 30: OFAT analysis of the performance by parameter levels.

result in strong performance decrease. Remarkably a number of four variables that is offered at each split on average gives the best results but the optimal model is obtained for two variables only. As a summary, some parameter values can be discarded for both theoretical considerations and empirical evidence.

In conclusion, the observed benefit of using random forests is relatively small and no significant improvement is obtained. But one may consider a business case of a population with 5% defaulters in total and a sum of 2bn. EUR of annual funding. In this case an improvement of only $+1\%$ rejected defaulters in will lead to a profit increase of 1mn EUR per year. The ROC curves for visual analytics as provided by many statistical software packages will not even allow to recognize any difference in the corresponding graphs.

Cost vs. benefit analysis: Different to academic research, time-to-market is typically an important business consideration which also holds for the credit scorecard modelling process. For this reason

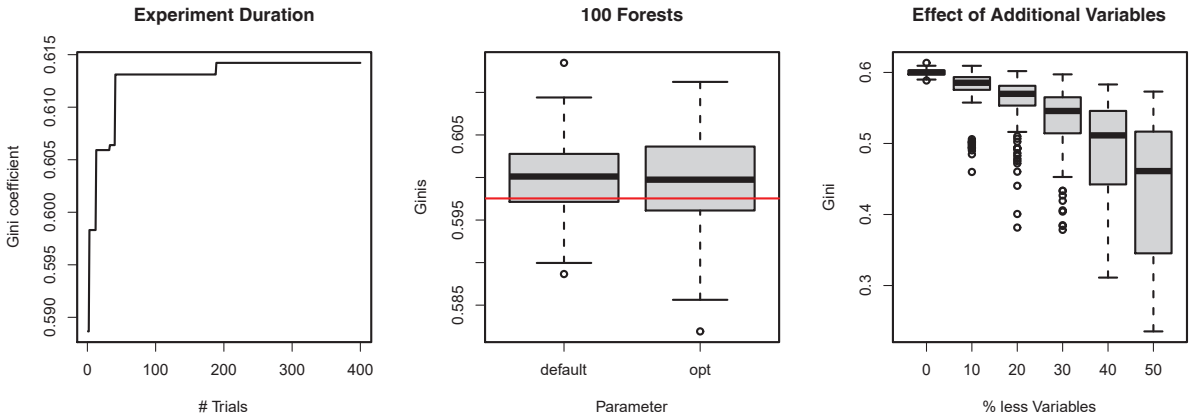

Figure 31: Performance vs. optimization time (left) and variance of forests (center) and performance decrease for lack of variables (right).

and based on the observed results from the last paragraph a parameter space (according to the previous grid but reduced by the identified implausible values for the tree and forest size) has been set up and a random search of additional 10000 forests in this parameter space has been set up in order to check the increase in performance with duration of the optimization process. (Please note that a naive random search can be speed up by intelligent search algorithms like genetic algorithms or iterated F-racing, cf. e.g. Bischl et al., 2016). Figure 31 (left) shows that the optimum is reached after only 400 iterations and not improved anymore which clearly indicates that the costs related to the additional parameter optimization process are comparatively low.

Astonishingly, it can be seen from the graph that the optimal performance from the first experiment (61.53) is not reached again within 10000 additional forests. This phenomenon is due to the ”randomness of random forests”: as the forests do depend on bootstrap samples two random forest models based on the same parameters will not be identical (see also Schäfer, 2006). Figure 31 (center) shows the performance variability of 100 random forests models for both the default parameters as well as the ”optimal” parameter set: the performance of the first experiment is not reached again, which again underlines the importance of the second test loop for parameter optimization in order to avoid overfitting (cf. Subsec. 7.2).

The higher variance of the best parameterization from experiment 1 can be explained by the additional randomness that results from selecting only two instead of four variables at each split which can increase or decrease performance depending on the selected variables. **A consequence, model selection should consist in taking into account both expectation and variance of the model performance estimation.** In order to win a competition like kaggle one may accept a higher variance to obtain increased upper performance quantiles whereas in a risk management context lower **quantiles of model performance estimation** will more relevant.

Model tuning vs. integration business knowledge: Improving models may not only be achieved by improving statistical modelling but also by identification of additional important explanatory variables. As the used data represents a real world example we can analyze what happened if some variables

were not available (to simulate the improvement that can obtained by new variables). Figure 31 (right) shows the loss in performance if a randomly selected percentage of variables were not available for model building. Please note the difference: as opposed to the random forest parameter the variables are not just removed for single splits within a tree but for the whole forest, here. The graph outlines the importance of identifying predictive characteristics: the benefit of an additional variable is much higher than the additional benefit obtained by model optimization. This step has to be considered as an important factor and emphasises the importance of a proper integration of business knowledge into the model building process.

Understanding the model: Finally, risk models not only have to be communicated to several directions (management, employees as well as its results towards the customers), there are also regulatory constraints often based on a natural mistrust towards black box algorithms. It should be remarked at this point that a **properly validated** and demonstrated superiority in model performance will lead to better estimates of risk which denotes one of the central targets of regulation.

Often the interpretability of regression coefficients (score points) are considered as advantageous to understand the key drivers of a model and to allow for its validation. The concept of variable importance (cf. e.g. Strobl et al., 2007) can be used to quantify the relevance of a variable within any classification model. Moreover the decrease in variance importance on out-of-time samples can be used for validation purposes and thus a decomposition of the black box.

7.4 Summary

The paper aims to bridge a gap between academic research concerning modern machine learning techniques and their business relevance for credit scoring applications. Random forests are investigated as one of today's most popular machine learning algorithms. The results are compared to logistic regression in a realistic setting. Several practical aspects are discussed like parameter tuning, business relevance of the improvements as well as cost vs. benefit aspects. In summary, the obtained benefits were comparatively small, but still of potentially large monetary impact. In addition the identification of new predictive characteristics has been demonstrated to be of great importance underlining the relevance of business knowledge integration in the modeling process.

As an important result, the simultaneous investigation of expectation and variance of classifier performance estimation has been worked out to be appropriate for model selection in a risk context potentially leading to a focus on quantiles.

Finally, variable importance has been presented as a tool to remove doubts concerning black box like algorithms w.r.t. regulatory constraints or for model validation purposes and improve the estimation of risks.

References

Bischl, B., Kuehn, T. and Szepannek, G.: On Class Imbalance Correction for Classification Algorithms in Credit Scoring. In Luebbecke, L., Koster, A., Letmathe, P., Madlenerm, R., Peis, B. and Walther, G. (eds.): *OR 2014, 37–43*, Springer, Heidelberg.

Boulesteix, A., Janitza, S., Kruppa, J. and König, I. (2012): Overview of Random Forest Methodology and Practical Guidance with Emphasis on Computational Biology and Bioinformatics. *Technical Report 129/2012*, Dept. Statistics, LMU Munich.

Breiman, L. (2001) Random Forests. *Machine Learning 45 (1)* 5–32.

Brown I., Mues C. (2012): An Experimental Comparison of Classification Algorithms for Imbalanced Credit Ccoring Data Sets, *Expert Systems with Applications 39 (3)*, 3446–3453.

Crone S., Finlay S. (2012): Instance Sampling in Credit Scoring: an empirical study of sample size and balancing, *International Journal of Forecasting 28 (1)*, 224–238.

Henking, A., Blum, C. and Fahrmeir, L. (2006): *Kreditrisikomessung. Statistische Grundlagen, Methoden und Modellierung*, Springer, Berlin.

Hoffmann, H. (1994): German Credit Data Set (Statlog) `http://archive.ics.uci.edu/ml/datasets/Statlog+\%28German+Credit+Data\%29`.

Lessmann S., Seow H., Baesens, B. and Thomas, L. (2013): Benchmarking State-of-the-art Classification Algorithms for Credit Scoring: A Ten-year Update. `http://www.business-school.ed.ac.uk/waf/crc\_archive/2013/42.pdf`.

Newman, D., Hettich, S., Blake, C. and Merz, C. (1998): UCI Repository of Machine Learning Database. `http://www.ics.uci.edu/~mlearn/MLRepository.html` University of California, Irvine, Dept. of Information and Computer Sciences.

Schäfer, M. (2006): Random Forests: A Case Study, Talk at 28. AG DANK, 27.10.2006, Dortmund.

Segal, M. (2004): Machine Learning Benchmarks and Random Forest Regression. *escholariship Univerity of California, Center for Bioinformatics and Molecular Biostatistics*. `http://escholarship.org/uc/item/35x3v9t4`.

Strobl, C., Boulesteix, A., Zeileis, A. and Hothorn, T. (2007): Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. *BMC Bioinformatics* 8–25.

Szepannek, G. (2011): Vortransformation in der Kreditantrags-Scoremodellierung. Talk at Data Mining Anwendertag, Heidelberg, `http://www.sas.com/reg/offer/de/datamining\_2011?page=download`.

Szepannek, G., Gruhne, M., Bischl, B., Krey, S., Harczos, T., Klefenz, F., Dittmar, C. and Weihs, C. (2010): Perceptually based Phoneme Recognition in Popular Music. In Locarek-Junge, H., Weihs,C. (eds.) *Classification as a Tool for Research*, Springer, Heidelberg, 751–758.

Szepannek, G., Schiffner, J., Wilson, J., Weihs, C. (2008): Local Modelling in Classification. In: Perner, P. (ed.): *Advances in Data Mining: Medical Applications, E-Commerce, Marketing, and Theoretical Aspects* Springer LNAI 5077, Berlin, 153–164.

Therneau, T., Atkinson, E. (1997): An Introduction to Recursive Partitioning using RPART Routines. *TR 61, Mayo Foundation*, `http://www.mayo.edu/hsr/techrpt/61.pdf`.