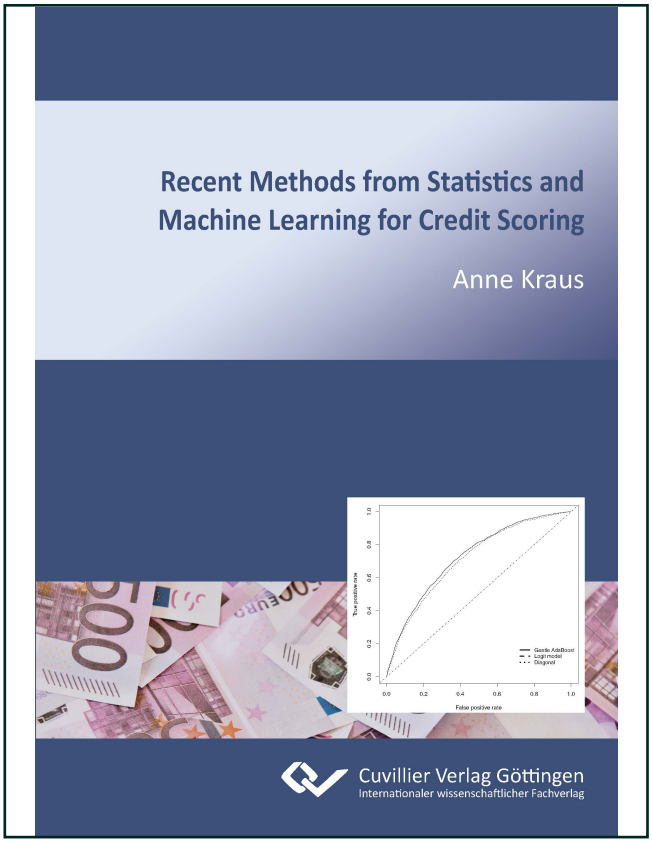
Anne Kraus (Autor)

Recent Methods from Statistics and Machine Learning for Credit Scoring



https://cuvillier.de/de/shop/publications/6703

Chapter 1

Introduction

1.1 Credit Scoring

Customers apply for consumer credit loans for many reasons, including flat-panel tvs, long-distance trips, luxury cars, relocation and family needs. The amount of consumer credits is lower than for real estate financing and involves several thousand euros, normally less than a hundred thousand. In this kind of credit business, personal information of the applicants are the main basis for the assessment of risk. This is in contrast to real estate financing, where higher credit amounts, lower interest rates and equity capital are the characteristics.

Since customer demand for personal loans has increased in the last decades, the consumer credit market evolved to become an important sector in the financial field and today represents a high-volume business. The UK and Germany are the countries with the largest total amount of consumer credits in Europe (Thomas, 2009). The German retail bank, from which the data is used in this thesis, receives about a million applications per year. These developments in the retail credit market requires automatic, fast and consistent decisions and processes to handle the huge amount of applications. The use of credit scoring models is now a key component in retail banking.

The development of so-called scorecards therefore represents the core competence of a retail bank's risk management when assessing the creditworthiness of an individual. The credit scorecards are embedded in the whole decision process, where other parts like budget account, the examination of the revenue and expenditure accounts and policy rules, are also relevant. These rules lead directly to the rejection of an applicant. Logistic regression is the main method applied in the banking sector to develop the scoring models. The performance of these models is essential, and improvements in the predictive accuracy can lead to significant future savings for the retail bank. The analysis of different models and algorithms to develop scorecards is therefore substantial for credit banks.

Since the market is changing rapidly, new statistical and mathematical methods are required for optimizing the scoring problem to decide on the question of whom to lend credit to. In recent years, many quantitative techniques have been used to examine predictive

power in credit scoring. Discriminant analysis, linear regression, logistic regression, neural networks, K-nearest neighbors, support vector machines and classification trees cover the range of different surveys (Thomas et al., 2005). An overview of publications is given in Thomas (2000) and Crook et al. (2007). For instance, Baesens et al. (2003) compare different classification techniques for credit scoring data where neural networks and least-squares support vector machines yield good results, but the classical logistic regression model still performs very well for credit scoring. In the meantime, new developments in statistics and machine learning raise the question of wether or not newer algorithms would perform better in credit scoring than the standard logistic regression model.

The area under the curve (AUC), based on the receiver operating characteristic (ROC) curve, is the most widely applied performance measure in practice for evaluating scoring models. However, classification algorithms are not necessarily optimal with respect to the AUC measure. This is the initial point for the idea to introduce the AUC as direct optimization criterion in this thesis. Since the AUC is a sum of step functions, the Nelder–Mead algorithm (Nelder and Mead, 1965) is used for the optimization, which represents a derivative-free and direct search method for unconstrained optimization of non-smooth functions. Moreover, the Wilcoxon statistic (Wilcoxon, 1945) is used for calculating the AUC measure. This novel approach is presented for the retail credit scoring case, and the properties of the algorithm are analyzed within a simulation study.

Recursive partitioning methods are very popular in many scientific fields, like bioinformatics or genetics, and represent a non-parametric approach for classification and regression problems. Since classification trees show rather poor results for prediction, random forests are especially prominent representatives of recursive partitioning algorithms leading to high predictive accuracy. Random forests belong to the ensemble methods, that overcome the instability of single classifiers by combining a whole set of single trees. Since the main task of this thesis is to improve the predictive accuracy of scoring models, standard recursive partitioning methods, and especially recent methodological improvements of the algorithms, are evaluated for the credit scoring problem. The new extensions overcome the problems of biased variable selection and overfitting in classification trees and random forests. In addition, an interesting approach is to combine classification trees with the classical logit model. This model-based recursive partitioning approach is also evaluated for the retail credit data with respect to the AUC performance.

The idea of boosting methods is to combine many weak classifiers in order to achieve a high classification performance with a strong classifier. Boosting algorithms are representatives of the machine learning area and offer a huge amount of different approaches. Apart from the classical AdaBoost with classification trees as so-called weak learners, it is possible to estimate a logit model by using linear base learners and the negative binomial log-likelihood loss within the component-wise gradient boosting. This boosting framework comes along with variable selection within the algorithm and more interpretability for the results. Since boosting methods are applied to different problems in other scientific fields, the aim here is to analyze these new developments in the retail credit sector for improving scoring models.

For the evaluation of recent methods from statistics and machine learning for the credit scoring case, the investigations in this thesis are based on data from a German bank. I am grateful to this bank for providing this data. Many surveys in this scientific field investigate well-known and often used public data sets for their empirical evaluation. A German credit data set with a 1,000 observations or a Japanese retail credit risk portfolio are prominent examples.[1]

As outlined above, the main focus of this thesis is the improvement of scoring models in the retail banking sector. It is a statistical point of view on credit scoring with concentration on recent methodological developments. The presented credit scoring data is evaluated with many different statistical methods rather than focusing on one specific model. The AUC is highlighted as performance measure and direct objective function. Moreover, different performance measures are presented for the analysis. In addition to the evaluation and the performance comparison of the presented algorithms, another aim is to stress the pros and cons of the proposed methods and to discuss them in the credit scoring context.

1.2 Scope of the Work

The scope of this work is to benchmark recent methods from statistics and machine learning for creating scoring models in order to maximize the AUC measure. To analyze the topic, this thesis is structured as follows:

Chapter 2 starts with the description of the AUC as a measure of performance in credit scoring, gives a short overview of further performance measures and continues with an overview of the credit scoring data used for the evaluation.

Chapter 3 presents a short introduction to the classical logit model, followed by a short summary of the classical scorecard development process in the retail banking practice.

Chapter 4 introduces a novel approach that uses the AUC as direct optimization criterion, and also includes some theoretical considerations and a simulation study for analyzing the properties of the proposed procedure.

Chapter 5 applies the generalized additive model for the credit scoring case as advancement of the classical logit model and representative of the well-known classical methods.

In Chapter 6 and 7, new methods from machine learning are investigated for the credit scoring case. In Chapter 6, classification trees, random forests and model-based recursive partitioning algorithms are presented and evaluated in the credit scoring context.

[1] Both available at http://archive.ics.uci.edu/ml/ of the University of California-Irvine (UCI) Repository.

In Chapter 7, the evaluation continues with various boosting methods where different loss functions and base learners are investigated. The AUC is especially used as a loss function within the boosting framework and analyzed for the credit scoring case.

Finally, the most important results are highlighted in the Chapter 8 summary. Concluding remarks and recommendations are given in the final chapter by additionally outlining issues for further research.

Computational aspects are given in the Appendix B where statistical software details are presented (B.1), and exemplary R-Codes are provided for the AUC approach and the simulation study in Chapter 4 (B.2). Additional graphics and result tables are included in the Appendix A.

Parts of Chapters 2 to 4 and 7 are published in an article in the Journal of Risk Model Validation (Kraus and Küchenhoff, 2014).

Chapter 2

Measures of Performance and Data Description

2.1 Receiver Operating Characteristic and Area under the Curve

A great variety of performance measures exists in credit scoring. Kullback divergence, Kolmogorov–Smirnov statistic (KS) and information value denote some of the measures used in this area (Anderson, 2007). The H measure (Hand, 2009) or the error rate are other performance measures used to evaluate scoring models. In this thesis, the focus is on the Receiver Operating Characteristic (ROC) and the related area under the curve (AUC) as the most widely applied performance measures in credit scoring practice. By far, these measures, including the Gini coefficient respectively, are the most important criteria in the retail banking sector. Further measures for discrimination, probability prediction, and categorical forecasts are also presented. The AUC is used as a measure of predictive accuracy for evaluating the different methods and algorithms in the credit scoring context. It is especially introduced as direct optimization criterion.

Primarily, ROC graphs have a long history of describing the tradeoff between hit rates and false alarm rates in signal detection theory (Swets, 1996; Swets et al., 2000). Thereafter, ROC graphs have been mainly used in medical decision making. Pepe (2003) considers, for example, the accuracy of a diagnostic test for the binary variable of the disease status (disease and non-disease). In recent years, the ROC analysis and the AUC have been increasingly used in the evaluation of machine learning algorithms (Bradley, 1997). According to Provost and Fawcett (1997), simple classification accuracy is often a poor metric for measuring performance so the ROC analysis gained more impact. For instance, one-dimensional summary measures, like the overall misclassification rate or odds ratio, can lead to misleading results and are rarely used in practice (Pepe et al., 2006). Since the consequences of false-negative and false-positive errors are hard to quantify, it is

common to draw both dimensions (tp rate and fp rate) into account (Pepe et al., 2006).

The survey of the credit scoring case denotes a binary problem regarding whether the customers default in a specific time period or the applicants pay regularly (cf. Section 2.2). The following two classes are examined:

- *default*, i.e., a customer fails to pay installments and gets the third past due notice during the period of 18 months after taking out the loan
- *non-default*, i.e., a customer pays regular installments during the period of 18 months after taking out the loan.

Due to this two class problem, two classes are used for the description of the ROC graph. For the multi-class ROC graphs, I reference the explanation of Hand and Till (2001). The following descriptions are based on Pepe (2003) and Fawcett (2006).
For classification models with a discrete outcome for prediction, different cases arise. If a default is correctly classified and predicted as a default, it is a *true positive*; while a non-default wrongly predicted as a default is counted as a *false positive*. Accordingly, the following parameters are computed:

$$tp\ rate = \frac{defaults\ correctly\ classified\ (tp)}{total\ defaults\ (p)} \tag{2.1}$$

$$fp\ rate = \frac{non\ defaults\ incorrectly\ classified\ (fp)}{total\ non\ defaults\ (n)} \tag{2.2}$$

Plotting the fraction of the correctly classified defaults (tp rate - TPR) versus the incorrectly classified non-defaults (fp rate - FPR) gives rise to the ROC graph. Figure 2.1 shows an example of ROC curves for three different default criteria. The diagonal describes a model with no predictive information, while the perfect model would imply a curve directly tending to the point $(0, 1)$. The perfect model would imply that there exists a score where all defaults have scores below this value, and the non-defaults above this value, respectively.

In the analysis, the main interest is on classifiers that do not produce a discrete good or bad decision, but a probability of default. Thresholds are used to create the ROC graph for these scoring classifiers. The threshold denotes a binary classifier for each special score, so for each threshold value one point in the ROC space can be evaluated. An important advantage of ROC curves is that they are insensitive to changes in the proportion of defaults to non-defaults (Fawcett, 2006).

Supplementary to the curves, numerical indices for the ROC graphs are often used. An important measure is the area under the ROC curve (AUC), which ranges between 0 and 1. As the name indicates this denotes the area under the curve presented in Figure 2.1. A perfect model has an AUC value of 1, while an uninformative scoring classifier has a value of 0.5, respectively. Normally scoring systems in practice have a value in-between.

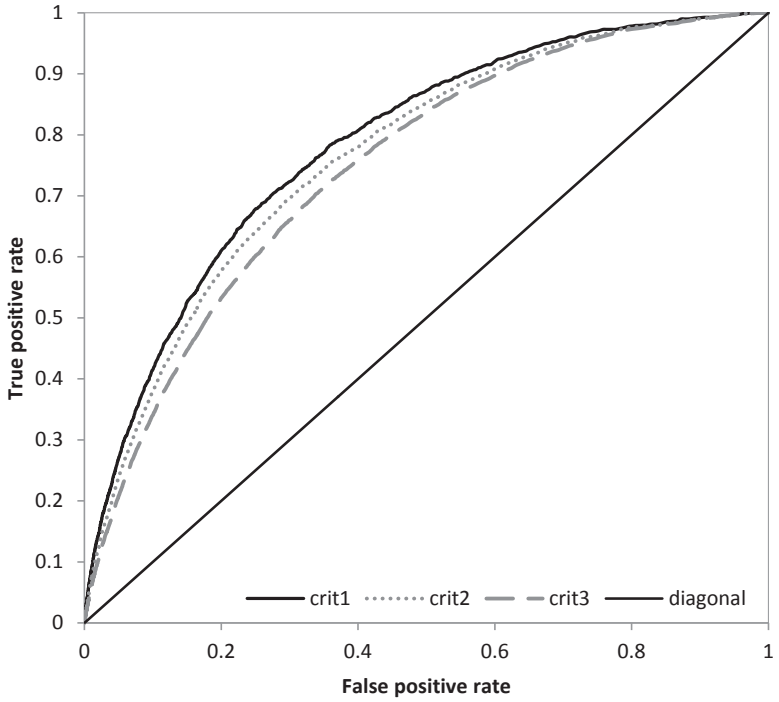


Figure 2.1: Example for ROC curves with three different criterions of default (crit1, crit2, crit3) by plotting the true positive rate versus the false positive rate.

An equivalent transformation to the AUC denotes the Gini coefficient with the definition $Gini = 2 \cdot AUC - 1$.

Since the AUC measure is used as well as direct optimization criterion, it is important to highlight the relationship to the well-studied Wilcoxon statistic (Hanley and McNeil, 1982). Assuming two samples of non-defaults n_{nd} and defaults n_d, all possible comparisons of the scores from each sample correspond to the following rule:

$$S(x_{nd}, x_d) = \begin{cases} 1 & \text{if } x_{nd} > x_d \\ 0.5 & \text{if } x_{nd} = x_d \\ 0 & \text{if } x_{nd} < x_d \end{cases} \tag{2.3}$$

where x_{nd} and x_d are the scores from the non-defaults and defaults, respectively. By averaging over the comparisons, the AUC can be written as follows:

$$AUC = \frac{1}{n_{nd} \cdot n_d} \sum_{1}^{n_{nd}} \sum_{1}^{n_d} S(x_{nd}, x_d) \tag{2.4}$$

This definition of the AUC value will be used for the AUC approach proposed in Chapter 4. Further explanations, especially concerning the confidence intervals, can be found in Hanley and McNeil (1982). Frunza (2013) covers the Gini coefficient in the context of credit risk model validation. For the analysis of areas under the ROC curves, DeLong et al.

(1988) propose a nonparametric approach by using the theory of generalized U-statistics to test whether the difference is significant.

The main objectives for a bank when lending consumer credits are to maximize the profit and reduce losses if a customer can not repay a loan. The cut-off defines the threshold denoting which customers are accepted and rejected, respectively. The ROC curve can also be used to display the influence of changes in the cut-off levels on various business measures of the loan portfolio like profit or loss (Thomas, 2009). But the business perspective goes beyond the scope of this thesis, since the focus lies on the statistical methods and the statistical model estimation with new algorithms. For that reason, I reference Thomas (2009) for further readings in the business aspects of the ROC curve.

In spite of its popularity, critical aspects remain for the AUC measure, such as the fact that it ignores the probability values and only considers the order of the scores (Ferri et al., 2005). Therefore, further performance measures are evaluated for the current credit scoring case.

Hand (2009) discusses some weaknesses of the AUC, such as the incoherency in terms of misclassification costs, and proposes the H measure as an alternative. He derives a linear relationship between AUC and expected minimum loss, where the expectation is taken over a distribution of the misclassification cost parameter that depends on the model under consideration (Flach et al., 2011). The H measure is derived by replacing this distribution with a Beta$(2, 2)$ distribution.

The KS curve charts the empirical cumulative distribution function percentages for defaults and non-defaults against the score. The KS statistic is defined by the maximum absolute difference between the two curves (Anderson, 2007). The above mentioned measures investigate the discrimination of the scoring models.

The minimum error rate (MER) is used to measure categorical forecasts. The performance metrics on comparing the classifier's predicted classes and their true labels are described previously in the context of the ROC-curve. A widely used summary of the so-called misclassification counts is the error rate (ER), which defines the total misclassification count (the sum of the false negatives FN and false positives FP) divided by the number of observations, i.e., $ER = \frac{FN+FP}{n}$. The error rate depends on a special threshold t to produce class labels since most of the classifiers produce a probability prediction. The minimum error rate (MER) corresponds to the value of t that achieves the minimum ER(t) over the test dataset (Anagnostopoulos et al., 2012).

To consider the probability prediction of a scorecard, the Brier score (Brier, 1950) is analyzed. The original definition of Brier (1950) is

$$BS = \frac{1}{n} \sum_{j=1}^{r} \sum_{i=1}^{n} (p_{ij} - Y_{ij})^2 \tag{2.5}$$

where p_{ij} denotes the forecast probabilities, Y_{ij} takes the value 0 or 1 according to whether the event occured in class j or not and r defines the possible classes ($r = 2$ for default and non-default). The Brier score defines, therefore, the expected squared difference of the

predicted probability and the response variable. The lower the Brier score of a model, the better is the predictive performance. However, the use of the Brier score is critical since it depends on direct probability predictions.

2.2 Data Description

The data used in this thesis was provided by a German retail bank that specializes in lending consumer credits. The application process of this bank is fully automated, thus offering a very large database with a wide range of information and a long data history. For the analysis, I concentrate on a part of the overall portfolio that guarantees representative results and also protects the bank's business and financial confidentiality.

The focus lies in the application scoring data because the decision of whether to grant credit to an applicant is the key issue in retail banking. If the promise of a loan is given and the money is paid out to a customer, the credit becomes part of the bank portfolio. As a result, the bank cannot cancel the decision. Behavioral scoring is important for forecasting the portfolio risk and is, therefore, relevant for the capital resources demanded by Basel II. Since the information is derived from a real application process, the data quality is very good. Because the customers are interested in taking out a loan, they are willing to divulge their personal information. The verification process inside the bank, where bank employees review the information, guarantees the accuracy of the data. Customers who provided wrong specifications do not receive credit unless the information is corrected. Missing data, therefore, poses a minor problem for the presented credit scoring case.

On the one hand, the most important variables are personal data from the customers, including income, marital status, age and debt position. Information from credit agencies completes the data set while applicants contractually permit the bank to obtain this data.

On the other hand, the payment history of consumers is essential for credit scoring, because this information enables the estimation of credit scoring models. Twenty six attributes are considered regarding process-related limitations for this specific bank. While variables like contract period or the amount of credit are good attributes to help predict a credit default, the decision of whether to grant credit has to be finalized before the customer decides about these variables. I note that these limitations present realistic empirical conditions for this bank. In other business models, for instance, loans are granted after the length of the loan is determined by the customer.

From the 26 attributes, four variables are categorical, while the other 22 covariates are of numerical order. For building credit scoring models, it is quite common to categorize the variables by building classes. Some explanations are given in Section 3.2 by describing the scorecard development process. If in the following variables are denoted as 'classified' or 'categorized', this always refers to the 22 numerical covariates. The remaining four variables are originally categorical and do not need further classification.

For model estimation, there are various options to define the outcome variable. A withdrawal of the credit can be denoted via a default as well as a past due notice. Moreover, it is essential to return to past events for observing the default. This time span can vary

considerably, for example, from 12 to 30 months.
The selected time span is always a tradeoff between actuality and the rate of defaults. For the analysis, the event of sending a third past due notice is assessed as the definition of default and a time span of 18 months is chosen (cf. Section 2.1).

Loans that are past due for more than 90 days can be classified as default as per the Basel II definition (Basel Committee on Banking Supervision, 2004). Since a time span of 12 months is often used for modeling predictive power, the choice of the time span depends on the portfolio structure. For differing customers who fail at the beginning and those failing after several months, survival analysis approaches are used for building scoring models. Banasik et al. (1999) and Stepanova and Thomas (2002) cover these topics in detail.

Due to confidentiality reasons, the descriptive analyses for the variables are not presented. For the evaluation, three different samples for training, test, and validation are used (shown in Table 2.1).

Data	All	Defaults	Defaults in %
trainORG	65818	1697	2.58
test	138189	3634	2.63
validation	72251	1878	2.60

Table 2.1: Data sets for training, test, and validation analyses, where the test and validation data represent out-of-time samples.

The default rates vary around 2.6%, a level that is characteristically low for the credit scoring market. As mentioned above, this quantity is an important reason for choosing a horizon of 18 months and the event of sending the third past due notice. As all three samples are drawn from different time intervals, the investigations on the test and validation samples are out of time.

For test purposes, additionally two samples are generated with randomly created default rates, as shown in Table 2.2. In practice, this procedure is common. The non-defaults are reduced to receive the specified default rate of 20 and 50%. These samples are generated from the original training sample (trainORG) of 65,818 applications.

Training data sets	All	Defaults in %
train020	8485	20
train050	3394	50

Table 2.2: Training data sets with simulated default rates of 20% and 50% drawn from the original training sample (trainORG) by randomly reducing the non-defaults.

The new methods of machine learning presented in this thesis are trained on a training sample, while the different tuning parameters are tuned according to the outcomes on the test sample. The validation sample is finally used for validating the results.