Transparency and Predictive Power

Explaining Complex Classification Models

Gerhard Paass and Jörg Kindermann

GMD - German National Research Center for Information Technology
53754 Sankt Augustin, Germany, mailto: paass,kindermann@gmd.de

Abstract. Complex classification models like neural networks usually have lower errors than simple models. They often have very many interdependent parameters, whose effects no longer can be understood by the user. For many applications, especially in the financial industry, it is vital to understand the reasons why a classification model arrives at a specific decision. We propose to use the full model for the classification and explain its predictive distribution by an explanation model capturing its main functionality. For a real world credit scoring application we investigate a spectrum of explanation models of different type and complexity.

1 Introduction

During the last years the availability of business process data allowed the utilization of sophisticated prediction and classification models in enterprises. A prominent example is credit scoring, where a customer is classified as solvent or insolvent based on a vector $\mathbf{x}$ of features. The unknown model parameters are trained using data on credit histories of customers. There is a plethora of models starting from simple linear models to complex neural networks or support vector machines. Many experimental evaluations (e.g. [8]) showed that more complex procedures in general yield better results in terms of classification errors. These methods usually have many interdependent parameters whose meaning no longer can be understood by the user.

There is a trade-off between the classification reliability and transparency of approaches. It is, however, possible to use a complex *full model* optimized with respect to predictive power and one or more simplified *explanation models*, which approximate the predictions of the full model as well as possible. If the difference between both predictions is small we can explain the results of the full model by evaluating the simple structure of the explanation model.

In the literature the approximation of a neural network by a rule system was investigated [1]. Craven [6] proposed to use tree models for the same task. Becker et al. [2] explored the explanation properties of naive Bayesian classifiers while Madigan et al. [9] explained Bayes probabilistic networks. In 1997 we developed a new credit scoring procedure for the DSGV, a large group of German banks [8]. This paper describes the results of a follow-on project funded by DSGV where we analyzed which type of explanation model is best suited to explain the predictions of the full model.

2 Criteria for Selecting an Explanation Model

2.1 Notation

As *full model* we used a *Bayesian model* $p(y = 1|\mathbf{x}, \theta) = f(\mathbf{x}, \theta)$ for our credit scoring task. It predicts the *credit risk* for $\mathbf{x}$, i.e. the probability that a customer with feature vector $\mathbf{x}$ will not repay a credit ($y = 1$). Instead of estimating an '*optimal*' parameter θ we derived the *Bayesian posterior distribution* $p(\theta|\mathbf{y}, \mathbf{X}) = p(\mathbf{y}|\mathbf{X}, \theta)p(\theta) / \int p(\mathbf{y}|\mathbf{X}, \theta)p(\theta)d\theta$ of parameters given data $\mathbf{y}, \mathbf{X}$ and prior distribution $p(\theta)$. The posterior distribution describes the plausibility of different parameters. It is approximated by a sample $\{\theta_1, \dots, \theta_B\}$ of parameters, which is generated by Markov Chain Monte Carlo methods. The model is based on decision trees.

For a new input $\mathbf{x}_0$ we may calculate the risk values $p(y = 1|\mathbf{x}_0, \theta_b)$ with respect to all $\theta_b \in \{\theta_1, \dots, \theta_B\}$. These values approximate the *predictive distribution* describing the uncertainty of the credit risk. If we state the consequences of decisions in terms of a loss function, we should classify the case $\mathbf{x}_0$ such that the average expected loss is minimal. This implies that $\mathbf{x}_0$ is classified as insolvent if the mean risk is larger than some threshold π_{dec}

$$P(y = 1|\mathbf{x}_0) \approx \frac{1}{B} \sum_b p(y = 1|\mathbf{x}_0, \theta_b) > \pi_{dec} \quad (1)$$

A change in the loss function (e.g. interest rates) merely leads to a shift in π_{dec} .

In this framework an *explanation model* can be considered as a simplified model $p(y = 1|\mathbf{x}, \psi) = g(\mathbf{x}, \psi)$, which on the one hand should be easy to understand, and on the other hand should approximate the classifications of the full model as good as possible. To keep the situation simple only an optimal model $\hat{\psi}$ is derived, e.g. by maximizing the posterior density with respect to ψ .

2.2 Comprehensibility

The explanation of a classification model is based on the following factors:

- The agreement of the results with intuitive common *factual rationalizations*.
- Insight into the *internal "mechanics"* of the model.
- The *empirical validation* of results by simple statistics.

Undoubtedly an explanation by factual rationalizations is most convincing. The internal mechanics of a model is comprehensible only for models with simple structural features:

Analysis of Subpopulations. The data is partitioned into disjoint subsets according to the values of some features. Each subset is characterized by a simple "submodel" to enhance comprehension [7, p.185].

If-Then Rules. Rules are especially evident, if the conclusions correspond to the common lines of reasoning in the application.

Figure 1: Scorecard for six variables with up to three categories: low (L), medium (baseline) and high (H). The scores are the positive or negative differences to the baseline representing a_0 . The scores for a single case are indicated by lines.

Linear Relations. The behavior of linear models is simple to understand.

Independent Variables. Even nonlinear models are easier to understand, if each feature adds an independent term to the result.

Subsets of Variables. Comparison of the results for different subsets demonstrate the contradictory effects of the subsets for the specific case.

To compare the comprehensibility of explanation models, we mainly use the *number of parameters* of the explanation model. This criterion is increased or decreased according to the *transparency* as well as the *empirical validity*. The latter quantities are highly subjective and give only a rough indication of the relative performance of procedures.

2.3 Classification Similarity and Quality

It is the task of the explanation model to characterize the distribution of predicted risks $p(y = 1 | \mathbf{x}, \theta_b)$ of the full model. As the loss functions vary only to a small extent, the threshold π_{dec} is restricted to a small range. Only in this range the value of the decision risk $p(y = 1 | \mathbf{x}, \theta)$ affects the decision (c.f. (1)).

Outside₉₀: The main criterion for classification *dissimilarity* of the full model and the explanation model. the credit risk predicted by the explanation model is *not* between the 5%-quantile and the 95%-quantile of the predictive distribution (section 2.1) of the full model.

Err_{8.75}: As an additional criterion we assess the *classification error* of the explanation model itself. We determine the percentage of rejected ‘good’ loans when 8.75% of bad loans are accepted. The Err_{8.75}-value for the full model was 32.0.

3 Comparing Explanation Models for Credit Scoring

To improve the fit btw. explanation and full models we created *synthetic data* by calculating the credit risk for 11000 unclassified genuine input vectors. Together

Table 1. Results for Stepwise Linear Logistic Regression.

variable set	no. of significant . . .			classif. error	dissimilarity
	vars	interactions	param.	Err _{8.75} %	Outside ₉₀ %
$\mathcal{M}_6$	4		5	57.6	13.7
	3	6 of 15	10	56.6	12.9
$\mathcal{M}_{17}$	10		11	54.8	13.2
	4	11 of 136	16	51.8	13.1
$\mathcal{M}_{39}$	25		26	49.2	12.1
	9	16 of 741	26	50.8	12.5

with the 6000 original training instances this was used as training input for the explanation models. Each case was used twice for training with output 1 and 0 and sampling weights equal to the credit risk $p(y = 1|\mathbf{x}, \theta)$ and its complement.

We used three different sets of input features selected by domain experts: The set $\mathcal{M}_6$ of six most relevant variables, the set $\mathcal{M}_{17}$ containing additional medium important variables and a comprehensive set $\mathcal{M}_{39}$ [11, p. 135ff].

3.1 Linear Logistic Model

For a given feature vector $\mathbf{x} = (x_1, \dots, x_k)$ this procedure approximates the credit risk by $g(\mathbf{x}, \theta) = \exp(\eta)/(1 + \exp(\eta))$ with $\eta = \theta_0 + \sum_{i=1}^k \theta_i x_i$. This is a classical credit scoring model used in many analyses [3].

Transparency. If coefficient θ_i is positive, the risk grows as attribute x_i is increased. Psychological experiments show that linear relations are far simpler to comprehend than nonlinear relations [7, p.129]. However strongly correlated features and corresponding parameter values with high covariance reduce the interpretability of parameters.

Classification Similarity. We used stepwise regression to identify the most important parameters. Even for the synthetic training set only four of the six most important variables are significant (see table 1). In comparison with later results the dissimilarity and the classification error are rather high. In summary this indicates the presence of non-linearity in the full model. Therefore linear models are not useful for explaining credit risk for this application.

3.2 Scorecards

This approach assumes that each feature x_i has an independent additive effect. We divide the domain of each feature into m subsets, each of which gets a point *score* a_{ij} . The sum of independent scores for a specific case is the total score $y = a_0 + \sum_{i=1}^k \sum_{j=1}^{m-1} a_{ij} x_{ij}$ where $x_{ij} = 1$ in the j -th subset of x_i and 0 elsewhere. The a_{ij} can be estimated from training data using stepwise linear regression [12, p.220] to detect significant scores. Scorecards are widely used in the financial industry [13]. A scorecard for the set $\mathcal{M}_6$ is shown in figure 1.

Table 2. Results for Scorecards Estimated by Stepwise Regression.

variable set	no. of categ.	signif. param.	classif. Err _{8.75}	dissimilarity Outside ₉₀
$\mathcal{M}_6$	3	10	49.4	12.9
	5	20	47.5	12.4
$\mathcal{M}_{17}$	3	21	43.5	11.0
	5	40	41.9	10.2
$\mathcal{M}_{39}$	3	31	38.9	9.4
	5	53	37.9	8.2

Table 3. Results for Decision Trees for Different Sets of Features.

variable set	no. of rules	classif. Err _{8.75}	dissimilarity Outside ₉₀
$\mathcal{M}_6$	27	52.8	11.4
	146	45.5	10.4
$\mathcal{M}_{17}$	23	51.2	10.8
	116	40.3	8.8
$\mathcal{M}_{39}$	26	42.1	9.5
	115	37.6	7.2

Transparency. The transparency of decisions and the simplicity of the ‘mechanics’ of scorecards is very high. Psychological research shows that the use of scores corresponds to our intuitions [7, S.189]. In summary scorecards are good explanation models.

Classification Similarity. In spite of the simple model structure scorecards have a surprisingly good classification similarity and quality (table 2). Note that there is an additional drop if we either increase the number of features or the number of categories within features. Generalized versions of scorecards [4, S.249ff] did not improve the results.

3.3 Decision Trees

Decision trees are frequently used for credit scoring [10],[4]. In contrast to linear models and scorecards, decision trees can represent any functional relation arbitrary well. They have an inherent representation bias, as they are well-suited for step-like functions.

Transparency. Decision trees have a double advantage in terms of transparency. First the internal mechanics of the method is simple as the credit risk within subsets defined by a few splits is constant. Second the credit risk itself is determined as the empirical frequency of ‘bad’ cases in this subset, leading to a good empirical validation. The number of rules (nodes) in the tree can be controlled. Large trees no longer offer a good explanation, as the user can only cope with a limited number of rules. An additional problem arises from repeated splits with respect to a single feature, as this is more difficult to understand.

Classification Similarity. The classification dissimilarity in table 3 is roughly in the range of figures for scorecards. In summary the scorecards have slight advantages compared to decision trees.

3.4 Explanation Models in Leaves of Decision Trees

A natural extension is the combination of approaches. In this section we combine decision trees with the following Scorecards and One- and two-dimensional non-linear loess models as ‘leaf’-models. We first developed an integrated approach

Fig. 2. Classification dissimilarity of scorecards, decision trees and scorecards in the leaves of decision trees.

Fig. 3. Classification error of scorecards, decision trees and scorecards within decision trees.

which simultaneously determined the trees as well as the scorecards with the aim to reduce the overall error. It turned out that this algorithm didn't offer any advantages. Therefore we simply generated a tree of the desired size and subsequently estimated scorecards within the leaves using stepwise regression.

Transparency. A scorecard is easier to understand than a decision tree with the same number of parameters. As scorecard merely is the sum of independent feature scores. A decision tree usually develops highly interlocked subset definitions, whose multidimensional relation is difficult to comprehend. On the other hand small decision trees are easy to understand. The combination of small trees and scorecards also is relatively transparent.

Classification Similarity. In the figures 2 and 3 the classification errors and dissimilarities of scorecards, trees and scorecards within trees are compared in relation of the number of parameters. In terms of dissimilarity it turns out that for the variables in $\mathcal{M}_6$ trees work better than scorecards and scorecards in leaves. For the variables in $\mathcal{M}_{17}$ all tree approaches have similar performance up to about 40 parameters. Finally for the set $\mathcal{M}_{39}$ all three procedures are nearly identical. Taking into account that scorecards are somewhat easier to understand than pure trees this gives scorecards and scorecards within trees an advantage.

As a second model we used loess model within the leaves of trees. Loess is a nonparametric approximation approach which is generated by the superposition of many simple local models defined in a small subset [5, p.93ff]. For each leaf subset we determined loess models with respect to one or two variables. The

Fig. 4. Densities of predicted credit risk (left) and profit (right) for a case projected to the base population. Below are mean predictions of the subset-of variable models.

form of the curve gives a good idea of the resulting changes in credit risk if the feature is modified. However, in terms of classification error and similarity this approach performed worse than decision trees or scorecards.

3.5 Subset-of-Variables Models

The credit risk may be predicted by a model which includes only the variables of a specific group, e.g. profit. By comparing the prediction based on different variable groups we get a picture on the perhaps antagonistic effects of the groups. The model for this explanation may be a ‘black box’. Therefore we may use the best model at hand, in our case the Bayesian procedure.

Figure 4 shows in its upper part the density of the predicted credit risk as well as the profit of the full model for a specific case. The classification is determined by the mean value of the credit risk (vertical line), which is below the decision threshold (dotted line), indicating that the case is classified as ‘solvent’. Below are the mean predictions of four subset-of-variables models corresponding to the groups finance, capital, profit and liquidity. Finance as well as the liquidity indicate a higher credit risk, whereas the capital and profit suggest a lower risk. In that sense the height of the risk is ‘caused’ by finance and liquidity. The fact that the prediction is performed by an optimized model adapted to the training data gives a good empirical validation of the approach.

4 Summary

As a major result there was always some dissimilarity between the full Bayesian model and the simple explanation models for our data. Beside of this fact decision

trees, scorecards and scorecards within trees had roughly the same similarity to the full model.

Linear models have a transparency similar to decision trees, but only limited modelling capacities. If the number of parameters exceeds a certain level scorecards seem to be easier to understand than decision trees. They have, however, a lower empirical validity, as they require complex matrix operations while trees are based on counts in subsets. Scorecards within the leaves of a tree are easier to understand than trees.

Especially valuable are subset-of-variables models. Ignoring the internal mechanics they get explanatory power only from the variables used as inputs. A prerequisite is, however, that the user forms sensible groups of input variables.

The relative ranking of procedures may be different for a new domain. To arrive at good explanation models we should observe the following points:

- Consider the degree of nonlinearity, interaction and noise in the data.
- Evaluate comprehensibility and classification similarity of selected models.
- Ask potential users for an assessment and use good or optimal models.
- Develop an intuitive graphical representation and an interactive interface.

References

1. R. Andrews and J. Diederich, editors. *Rules and Networks*, Brisbane, 1996. Queensland University of Technology, Neurocomputing Research Centre.
2. B. Becker, R. Kohavi, and D. Sommerfield. Visualizing the simple bayesian classifier. In *KDD Workshop on Issues in the Integration of Data mining and Data Visualization*, 1997.
3. Steven V. Campbell. Workouts: Reorganize or liquidate? prediction model gives insights for small-business loans. *Commercial Lending Review*, 11:78–85, 1996.
4. J. M. Chambers and T. J. Hastie. *Statistical Models in S*, volume Monograph 35. Wadsworth & Brooks, Pacific Grove, California, 1992.
5. W. S. Cleveland. *Visualizing Data*. Hobart Press, Summit, New Jersey, 1993.
6. M. W. Craven. *Extracting Comprehensible Models From Trained Neural Networks*. PhD thesis, University of Wisconsin - Madison, 1996.
7. R. M. Hogarth. *Judgement and Choice*. Wiley, Chichester, 2nd edition, 1987.
8. J. Kindermann and G. Paass. Model switching for Bayesian classification trees with soft splits. In J. Zytkow and M. Quafafou, editors, *Principles of Data Mining and Knowledge Discovery*, pages 148–157. Springer, 1998.
9. D. Madigan, K. Mosurski, and R.G. Almond. Explanation in belief models. *Journal of Computational and Graphical Statistics*, 6:160–181, 1997.
10. E. Rosenberg and A. Gleit. Quantitative methods in credit management: A survey. *Operations Research*, 42:589–613, 1994.
11. C. Uthoff. *Erfolgsoptimale Kreditwürdigkeitsprüfung auf der Basis von Jahresabschlüssen und Wirtschaftsauskünften mit Künstlichen Neuronalen Netzen*. M & P Verlag für Wissenschaft und Forschung, Stuttgart, 1997.
12. W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S-Plus*. Springer Verlag, New York, 2nd edition, 1997.
13. Mark Zmiewski and Beverly Foster. Credit-scoring speeds small business loan processing. *Journal of Lending & Credit Risk Management*, 79(3):42–47, 1996.