Credit scoring: Discussion of methods and a case study

Dorte Kronborg and Tue Tjur
Department of Management Science and Statistics
Copenhagen Business School
Denmark

Bo Vincents
Department of Operations Management
Copenhagen Business School
Denmark

Estimated number of words: 11.000

Abstract

The scenario considered is that of a credit association, a bank or another financial institution which, on the basis of information about a new potential customer and historical data on many other customers, has to decide whether or not to give that customer a certain loan. We discuss three popular techniques: logistic regression, discriminant analysis and neural networks. We shall argue strongly in favour of the logistic regression. Discriminant analysis can be used, and for reasons that can be explained mathematically it will often result in approximately the same conclusions as a logistic regression. But the statistical assumptions are not appropriate in most cases, and the results given are not as directly interpretable as those of logistic regression. Neural network techniques, in their simplest form, suffer from the lack of statistical standard methods for verification of the model and tests for removal of covariates. This problem disappears to some extend when the neural networks are reformulated as proper statistical models, based on the type of functions that are considered in neural networks. But this results in a somewhat specialized class of non–linear regression models, which may be useful in situations where local peculiarities of the response function are in focus, but certainly not when the overall — usually monotone — effect of many more or less confounded covariates is the issue. We discuss, within the logistic regression framework, the handling of phenomena such as time trends and corruption of the historical data due to shifts of policy, censoring and/or interventions in highrisk customers' economy. Finally, we illustrate and support the theoretical considerations by a case study concerning mortgage loans in a Danish credit association.

Keywords: Credit scoring, discriminant analysis, logistic regression, neural network, event history analysis.

Address: Tue Tjur, MES, Copenhagen Business School, Solbjerg Plads 3 , DK-2000 Frederiksberg C, Denmark. Email: tuetjur@cbs.dk.

1 The problem

Suppose we have historical data on n customers, in the form of

covariates x_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, k$, and

responses y_i, $i = 1, \ldots, n$.

The responses are assumed to be binary, with the event "bankruptcy" coded as 1, "not bankruptcy" as 0. Bankruptcy in this context means the event that the customer, willingly or unwillingly, fails fully to repay the loan; thus, $y_i = 0$ means that customer i fullfils his contract.

The covariates are assumed to represent the information available to the financial institution about the customers. If, for a moment, we define the customers as persons (in many applications we would also have institutions, firms, married couples, etc.), the covariates could include informations like age, sex, marital status, income, housing expenses, certain household expenses, information about other loans and payment behaviour during the period of the loan, perhaps even payment behaviour in earlier periods with other loans. To this comes, in the case of a credit association, a lot of information about the value of the property, other mortgages, etc. When evaluating a new applicant we must make our decision based on a description af the customer in terms of covariate values $x_1, \ldots, x_k$ — often with a lot of "missing values", for instance information which is not meaningful for a new customer, or information that just happens not to be available right here and now.

The problem could be one of two, one, the decision of whether to accept or reject an application of a new loan, two, the decision of which action to take when repayments of an existing loan are defaulted. The decision of whether to give a new loan could be seen as equivalent to answering the question: "if we give this loan, what is the probability that this customer will go bankrupt?" Similarly, when repayments af an existing loan are defaulted an action could be based on the answer to the question " if we continue the loan, what is the probability that this customer will go bankrupt?"

The action to be taken in case of defaulted repayments could be:

1. Terminate the loan (with potential losses) via legal rules.
2. Replace the loan by a new loan with different (easier) payback conditions.

This is similar to the situation of determining new loan applications with termination being equivalent to rejection and replacement being equivalent to acceptance. Consequently, the problem of which action to take when repayments of an existing loan are defaulted, can be seen as a special case of treatment of new loan applications. Therefore, we have used the senario of new loan applications in our description of the theory. However, the case illustrating our findings is of the type "defaulted loans".

But let us take a critical look at all the over–simplifications we have already made. First of all, the event "bankruptcy" (and thereby the probability of this event) is not quite well–defined (see below), and even if it makes some sense it is only a small part of what we are interested in. What we really would like to know is something like the joint probability distribution of two variables related to the future behaviour of the customer, namely

— the amount of money we are going to lose if and when this customer is unable to pay back (set to zero if this never happens), and

— the time when this event takes place (set to anything, most naturally $+\infty$, if this never happens).

Here we have even made an additional simplification, because the deviation from a regular payment flow may very well become more complicated than accounted for by a single event, with several delayed or reduced rates. But in the exclusive situations where it makes sense, this bivariate distribution is the least we can do with if we want to make exact insurance mathematics type calculations. Together with information about (or qualified predictions of) the rate of interest, administration expenses, etc., a reliable estimate of this bivariate distribution would provide us with everything that is needed for a full analysis of the decision problem in terms of a comparison of the discounted expected loss with the discounted expected gain.

The presence of the variable "time of bankruptcy" raises another problem, which also is related to the historical data. What do we mean by "the customer going bankrupt"? The immediate interpretation is that "the customer goes bankrupt sooner or later", but this event is related to the future not only for our new customer, but also for a large portion of the customers in the historical data set. In fact, the response y_i is only observed for those (hopefully few) of them that already went bankrupt ($y_i = 1$), and for those that are no longer customers after full repayment ($y_i = 0$). In statistical terms, the problem is that the variable "time of bankruptcy" is *censored* at

the endpoint of the historical study. Methods for the handling of censored data are extensively studied and developed in a biostatistical context, and we shall return to this point in section 5. But in the present section, and in the discussion of the three main methods, we make the following — very restrictive — assumptions.

In order to make the responses y_i fully observable, we pretend that we are only interested in bankruptcy *within a certain period,* say the first year. Accordingly, we exclude from our historical data set all customers that have been customers for less than a year, and redefine the responses such that $y_i = 1$ means "customer i went bankrupt during the first year of his loan". Later bankrupts are ignored. Similarly, we rephrase our problem concerning the new potential customer to "what is the probability that this customer goes bankrupt within a year?".

This is not exactly the question we posed from the beginning. Nevertheless, it should be realized that we are much better off with an estimate of this probability than without any quantitative considerations at all. For short loans a time horizon of one year may be all that is needed. For longer loans a reasonable assumption may very well be that the expected loss associated with a bankruptcy during the first year is proportional to the total expected loss, or at least that there is a monotone relation between these two quantities. More generally, we can say that the difficult part of the problem is to combine the many covariate values to a single measure of credit worthiness. A rescaling of that measure or a monotone transformation of it does not matter, since we are basically only interested in the decision rules associated with that measure. By this we mean the decision rules of the type "reject loan if measure exceeds threshold value, accept otherwise". Once the general measure of credit worthiness is given, we can always decide which threshold value to operate with, simply by selecting a value that would have given acceptable decisions if the corresponding decision rule had been applied to all customers in the historical data set. We may (and probably should, in most cases) even adjust this method in order that also the proportion between the expected loss in case of bankruptcy and the expected gain in case of full repayment is taken into account. But this aspect, which has more to do with economics than with statistics, will be ignored in the following.

In addition, we make the standard assumption that all covariate values are observed for all customers in the historical data set as well as for the new potential customer. With reference to the total data set, this is usually

unrealistic. What we mean by this is, of course, that whenever a piece of information is missing, we must remove either the corresponding covariate or the corresponding customer from the present analysis. An immediate consequence is (since the new potential customer cannot be removed) that all covariates that are not observed for the new customer must be removed. In practice, this means that it may be necessary to use different historical data sets for different new customers. For the historical data it is natural to start by removing the covariates which (by common sense or by some statistical test) are irrelevant for the prediction of the probability of bankruptcy, in particular those with many missing values. We may also be forced to remove covariates that, although they seem to contribute significantly, are too sparsely observed. However, this problem is often encountered in the analysis of data of some complexity. We shall assume in the following three sections that all these tedious compromises have been made in advance, in order that we may focus our attention on a fully observed rectangular data set.

2 Logistic regression – the forwards method

The problem is to estimate the probability of bankruptcy, say

$$P(y = 1) = p(x_1, \ldots, x_k),$$

where $x_1, \ldots, x_k$ are the covariate values for the new customer. Since we want a method that can handle any covariate pattern, we can also say that we want to estimate the function p, which to an arbitrary covariate pattern assigns the probability that a customer with this pattern goes bankrupt within the first year.

If it was not for the fact that the responses are binary rather than numeric, this looks very much like a standard multiple regression problem. If, for a moment, we imagine that the y's were some "degrees of credit worthiness" on a continuous scale that could be observed after a year, a standard solution to this problem would be to consider the hierarchy of regression models of the form

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i,$$

describing the outcomes y_i as sums of a linear combination of the covariates and normal independent errors ϵ_i with mean zero and common variance σ^2.

However, the fact that the responses are binary does not prevent this. The standard statistical analogue to regression models, when the responses are binary, is *logistic regression* or *logit–linear modelling*. The only difference between ordinary linear regression for normal variables and logistic or logit–linear regression for binary variables is that the expression of the *expected response* as a linear combination of covariates is replaced with an expression of the logit–*transformed probability of positive response* as a linear combination of covariates; in our case,

$$\text{logit } (p(x_1, \ldots, x_k)) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k,$$

where the function

$$\text{logit } (p) = \log\left(\frac{p}{1-p}\right),$$

is the simplest choice of a function that "stretches" the probability interval]0,1[to the whole real axis. Other choices are possible (for example the inverse to the c.d.f. of the normal distribution, often called the probit–transform), but the logit function turns out to have some desirable algebraic properties in this context, among which we would like to emphasize two properties related to the interpretation of the model and its maximum likelihood estimates:

(1) As recently pointed out by Alan Lucas (2001), the maximum likelihood estimates of the individual bankrupt probabilities (the fitted values) have the following property. If the model includes a factor F, then for any level f of that factor the average of the estimated bankrupt probabilities over the corresponding set of customers equals the relative frequency of bankrupts in that subset. For example, if there are different types of loans involved and the model takes this into account, then for each type the average of the fitted bankrupt probabilities will equal the actual relative frequency of bankrupts for that type. For a quantitative covariate, say the customers age, we have the similar property that if the covariate is included in the model as a simple linear term β*age, then the average age of bankrupters equals the weighted average of ages over the whole population, when the estimated bankkrupt probabilities are taken as weights. These exclusive properties of the logistic regression model follow from its interpretation as an exponential family. Actually, the likelihood equations *are* essentially all equations of the types mentioned above, equating sufficient statistics with their expectations under the model.

(2) In many applications bankruptcy is a rare event. For a data set consisting of, say, a million customers with only 1000 bankrupts, it is tempting to reduce the data size by construction of an artificial data set, consisting of all the bankrupters and a small randomly drawn portion (say 1%, i.e. around 10000) of the non–bankrupters. This will hardly affect the accuracy of the conclusions, because the shortage of bankrupters will be the dominating error source in all matters regarding the difference between bankrupters and non–bankrupters. Another exclusive property of the logit model is that it is essentially unaffected by such a reduction of the data set; namely in the very precise sense that if a logit model holds for the original data set, then the reduced data set can be described by the same model with the same parameters, except that the constant term β_0 should (of course) be corrected to account for the artificially increased probability of bankrupt. This means that inference based on such a reduced data set can easily be translated to valid inference about the original data.

Just as in ordinary multiple regression the technique is to identify and estimate a model which is as simple as possible, but still exhaustive enough to explain the significant relations between covariates and responses in the historical data set. In this process, we can — with few and mainly technical modifications — draw on the whole classical machinery of ordinary multiple regression, including tests for the removal of terms from the model (which are likelihood ratio tests based on the χ^2–approximation, not F–tests), the introduction of interactions, product or polynomial terms as required, the grouping or transformation of covariates, etc. When a satisfactory model is found, the prediction of a new customer's response is just a matter of inserting his covariates in the formula for the probability of bankruptcy.

3 Discriminant analysis – a backwards method

Consider, in the historical data set, the two k–dimensional populations of covariate values constituted by the customers that went bankrupt and those that did not. In the standard expositions of discriminant analysis, these two populations are assumed to be multivariate normal with the same covariance matrix Σ and different mean vectors μ_1 (for those that went bankrupt) and μ_0 (for those that did not).

In the present context, this is a "backwards" model, in the sense that the

responses y_i are regarded as fixed and the covariates as random. But it has one advantage over the logistic regression model, which probably explains why the method has become (though not why it still is) so very popular in credit scoring, namely that it is computationally more simple. Whereas the estimation in a logistic regression model requires numerical maximization of the likelihood function, the estimates in a discriminant analysis model can be computed explicitly. The maximum likelihood estimates of the two mean vectors are simple co–ordinatwise averages of the covariate values of the respective populations, and the maximum likelihood estimate of the covariance matrix (with standard correction for bias) is a weighted average of the empirical covariance matrices of the two k–dimensional samples.

A relevant criticism of this model is that the assumptions are very restrictive and hardly ever satisfied in practice. In the applications that we have in mind, many of the covariates are binary (e.g. gender), and many others will have to be derived from classifications in three or more unordered categories, which means that they must enter the linear expressions as "dummies" (indicators for group membership), which are again binary. This makes normality quite unrealistic.

Even in a situation where all covariates are proper quantitative measures, the natural normality assumption would usually be normality of the *whole* population, not of the two subpopulations defined by the response. It appears rather naive to assume that "bankrupters" and "non–bankrupters" are so fundamentally different species that one could — in principle and if one had data enough — identify the two normal components in the marginal distribution of the covariates, without observing the responses at all. It seems much more realistic to assume that bankruptcy is a random event, influenced by the covariates, but certainly not with such a "backwards" impact on the distribution of the covariates.

Another drawback of the discriminant analysis model is that it does not give an immediate answer to the original question concerning the probability of bankruptcy for a new potential customer. It does, however, give a function that can be used to discriminate between the two populations, namely the proportion between the two (estimated) normal densities, or its logarithm.

Formally, the discriminant analysis model reduces the original problem to the following, provided that the parameter estimates are accurate enough to take the role of true parameter values. The set $(x_1, \ldots, x_k)$ of covariates for our new customer is known to be a random vector from a normal distribution

with covariance matrix Σ and with a mean that is either μ_1 or μ_0. Which of them is it? Or, more precisely, what is the probability that it is the one with mean μ_1?

An assignment of a probability to this event requires a Bayesian formulation, which will be explained below. But the problem of deciding (whatever that means) which of the two populations the covariate set comes from is a more fundamental statistical problem, which — as agreed on by all schools of statistics, including the Bayesian — should be solved by consideration of the *likelihood ratio*, the proportion between the two densities at the point $\mathbf{x} = (x_1, \ldots, x_k)$. Intuitively, the idea is that if the two densities happen to be approximately equal in the observed point $\mathbf{x}$, then we cannot not say more about the new customer than we could before $\mathbf{x}$ was observed; on the other hand, if the density with a mean of μ_1 is, say, more than 20 times greater than the density with a mean of μ_0, then we have a strong indication of a bankruptcy; and vice versa.

Let φ_1 and φ_0 denote the densities for the normal distributions of the covariates for bankruptcies and non–bankruptcies, respectively. Then, by a straightforward calculation, the logarithm to the proportion between the two densities at $\mathbf{x} = (x_1, \ldots, x_k)$ is

$$\log \frac{\varphi_1(\mathbf{x})}{\varphi_0(\mathbf{x})} = \left(\mathbf{x} - \frac{\mu_1 + \mu_0}{2} \right)' \Sigma^{-1}(\mu_1 - \mu_0),$$

(where, by convention, $\mathbf{x}$, μ_1 and μ_0 are regarded as $n \times 1$ columns when matrices are multiplied). Consequently, our decision of whether to accept or reject the new customer should be based on this quantity or, equivalently, on the linear function $\mathbf{x}'\Sigma^{-1}(\mu_1 - \mu_0)$ of the covariates. If this linear combination exceeds some threshold value we should reject the loan application, otherwise we should accept. The decision as to which threshold value to apply does not follow from this, but as we have noticed before, this problem can be solved through an examination of the hypothetical historical consequences of decision rules based on different threshold values.

One way of assigning a concrete probability to the event that the new customer goes bankrupt is by assigning a Bayesian prior probability p_{tot} to this event. This probability should be interpreted as a "total" or *unconditional* probability, as opposed to those discussed earlier which are conditional on the new customer's covariates. An estimate of this probability p_{tot} is difficult

to obtain, because it should in principle be estimated as a (historical) frequency of bankruptcies in the population of loan applicants, including those that were rejected and for whom it does not even make sense to consider the event "bankruptcy". Anyway, if we can find a reasonable value for p_{tot}, we have in principle specified the total joint distribution of covariates and response for the new potential customer. The distribution of the response is given by p_{tot} , and the conditional distribution of the covariates, given the response, is multivariate normal, as specified by the model and the parameters Σ, μ_1 and μ_0. A straightforward computation (an application of Bayes' formula) gives the following relation between the conditional probability $p = P(y = 1 \mid x_1, \ldots, x_k)$ of bankruptcy, given the covariates, and the unconditional probability p_{tot} of this event,

$$\frac{p}{1-p} = \frac{p_{\text{tot}}}{1-p_{\text{tot}}} \times \frac{\varphi_1(\mathbf{x})}{\varphi_0(\mathbf{x})},$$

or

$$\frac{p}{1-p} = \exp\left(\log\frac{p_{\text{tot}}}{1-p_{\text{tot}}} + \left(\mathbf{x} - \frac{\mu_1+\mu_0}{2}\right)' \Sigma^{-1}(\mu_1-\mu_0)\right),$$

or

$$\text{logit}(p) = \text{logit}(p_{\text{tot}}) + \left(\mathbf{x} - \frac{\mu_1+\mu_0}{2}\right)' \Sigma^{-1}\left(\mu_1-\mu_0\right).$$

This formula can be used for the computation of p when p_{tot} is known, and is useful if we want to assign a probability to the event that the new customer goes bankrupt.

A second — and perhaps more interesting — consequence of this formula is that the logistic regression model has an interpretation as a conditional model in a "super model" based on the discriminant analysis model. By the super model is meant the model considered above, when p_{tot} is given the role of an unknown parameter rather than a subjective prior probability. Hence, the model states that any customer goes bankrupt with probability p_{tot}, and the conditional distribution of the covariates, given this event, is multivariate normal as specified by the discriminant analysis model. This is the description of the super model from the point of view of a stepwise observation scheme where the responses y_i are observed first in their marginal distribution, and then the covariates are observed in their conditional distribution, given the responses. The above computations illustrate how it appears from the point of view of a "forwards" observation scheme, where the covariates

are observed first in their marginal distribution (which is a mixture of two normal distributions), and thereafter the responses are observed in their conditional distribution, given the covariates. The expression for logit(p) above shows that the conditional model in the last step coincides with the logistic regression model considered in section 2, with parameters

$$\beta_0 = \text{logit}(p_{\text{tot}}) - \left(\frac{\mu_1 + \mu_0}{2}\right)' \Sigma^{-1} (\mu_1 - \mu_0),$$

and

$$(\beta_1, \ldots, \beta_k) = \Sigma^{-1} (\mu_1 - \mu_0).$$

This explains why the conclusions resulting from the logistic regression very often can be reproduced with high accuracy by the corresponding discriminant analysis. Indeed, if the marginal distribution of the covariates contains very little information about the parameters of interest (which essentially means that the two normal components cannot be identified), almost all the information lies in the second step, which is the standard logistic regression model. The interpretation of the linear expression

$$\beta_1 x_1 + \cdots + \beta_k x_k = \mathbf{x}' \Sigma^{-1} (\mu_1 - \mu_0),$$

as the discriminating function on which the decisions should be based holds for both models. Even the expression for the probability of bankruptcy, given the covariates, is the same for the two models, only with a slightly different interpretation of p_{tot}.

However, this is only an excuse for the discriminant analysis model, not a recommendation of it. If normality holds, the discriminant analysis model should be preferred (cf. Efron 1975). But normality is absurd in most of the situations we have in mind, in particular normality of the two subpopulations defined by the response, and it is easy to construct examples where things go entirely wrong because the normality assumption does not hold. If one simulates, for the case $n = 1000$ and $k = 1$, the x–values as independent normal with mean 0 and a standard deviation of 10, and thereafter generates the responses y_i according to a logistic regression model with $\beta_0 = 0$ and $\beta_1 = 1$, the fit of a logistic regression will give an estimate of β_1 around 1 ± 0.2, whereas the discriminant analysis model will give an estimate close to 0.4. Other convincing arguments in this direction can be found in Press and Wilson (1978). For these reasons, we cannot recommend discriminant analysis of credit scoring data in general.

A final remark (in favour or against, the choice is yours) about discriminant analysis is the following. As noticed by Fisher (1938, see also Anderson 1984), the vector $(\hat{\beta}_1, \ldots, \hat{\beta}_k)$ of estimated coefficients for the discriminating linear function is proportional to the vector of estimated coefficients in an ordinary least squares multiple regression of the binary response y on the covariates $(x_1, \ldots, x_k)$. This peculiar result implies that the set of decision rules resulting from the discriminant analysis coincides with the set of decision rules of this "dummy" multiple regression. This may be of particular interest to SAS users, because PROC GLM — as opposed to PROC DISCRIM — can generate dummies from factors or products of factors ("CLASS terms") in a model formula. But as a statistical method, this regression appears rather naive, and it is easy to imagine situations where this model will result in conclusions that are more complicated than necessary. Think of a case where a single factor on two levels is very dominating, in the sense that all bankruptcies are on level 1 of that factor. Almost inevitably, this will result in significant interaction of that factor with all other factors of interest in the least squares analysis. Whereas a logistic regression model can easily incorporate such a "dominating" factor, acting additively with all other effects.

4 Neural networks – the black box method

The term "neural networks" covers a very large class of models and algorithms. For more general expositions of neural network modelling from a statistics point of view, we refer to Ripley (1994) and Sarle (1994). We shall consider only a few of the simplest neural network models and discuss their relation to the logistic regression model.

4.1 The neural network without hidden layers

To start with a triviality, a neural network with "no hidden layers" can be described by the approximate functional relationship

$$y_i \approx F\left(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}\right),$$

where the function F (the "activation function") is usually (and will in the following be) taken as the inverse logit function $F(x) = \frac{e^x}{1+e^x}$. In its most primitive form, this is nothing but an approximate description of the way the

responses y_i depend on the covariates $x_{i1}, \ldots, x_{ik}$. The relationship is usually estimated ("trained") by ordinary least squares or by minimization of some other simple measure of distance between the "targets" y_i and the "outputs" $F(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})$. The models ability to predict is then tested by cross validation, i.e. by comparison of predicted and observed values in a new data set, the "test set". Since no other methods for model verification are available, cross validation plays a much more important role in neural networks than in ordinary statistical practice. Quite often, it is necessary to split the data set randomly in two.

From a statistical point of view, it is natural to interpret the functional relationship as a statistical model, for example — in the context of the present paper — to think of the expression on the right hand side as the *probability* of getting the value $y_i = 1$ for a customer with covariates $x_{i1}, \ldots, x_{ik}$, and, accordingly, use maximum likelihood rather than least squares. With this modification we have, with all reservations concerning other interpretations that are beyond our level of perception:

The neural network with no hidden layers coincides with the logistic regression model.

4.2 The neural network with one hidden layer consisting of a single neuron

This is the model specified by

$$y_i \approx F\left(\alpha_0 + \alpha_1 F\left(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}\right)\right).$$

Again, the function F (which occurs twice in quite different contexts) is assumed to be the inverse of the logit function. Other functions could be used here, but since a standard assumption is that these functions are increasing and bounded, it makes very little difference in the following discussion. And again, we prefer to think of the right hand side as the probability of the response 1.

The main difference between this model and the logistic regression model is that the range of the right hand side is a proper subinterval of the unit interval. For $\alpha_1 > 0$, the right hand side is an increasing function of the linear combination $\beta_0 + \beta_1 x_{i1} + \ldots \beta_k x_{ik}$. For large positive values of this linear combination the value of the right hand side comes close to $F(\alpha_0 + \alpha_1) < 1$, for

large negative values it comes close to $F(\alpha_0) > 0$. This behaviour represents a well known modification of the logistic regression model to situations where some "background—randomness" implies that the response is not asymptotically deterministic for large absolute values of the linear combination of covariates. In the context of credit scoring it would mean that even the most well behaved customer has probability at least p_1 of going bankrupt, and even the most unreliable customer has probability at least p_0 of not going bankrupt, where p_1 and p_0 are (small but) positive. Another simple modification of the logistic regression model that takes this into account can be constructed as follows. Imagine that "preliminary" responses y_i^* are generated by a standard logit–linear model. But the final responses are generated from the preliminary responses by a mechanism that changes a 0 to a 1 with probability p_1, and changes a 1 to a 0 with probability p_0. The expression for the probability of bankruptcy in this model is easily seen to be

$$P(y_i = 1 \mid x_{i1}, \ldots, x_{ik}) = (1 - p_0)\, F(\ldots) + p_1\, (1 - F(\ldots))\,,$$

where ... stands for the usual linear combination of the covariates. As functions of this linear combination, these functions are similar in shape to the functions that can occur on the right hand side of the neural network model. But the neural network model in this case is certainly an alternative which is worth considering.

However, it is not quite fair to use the term neural network for the model with a single neuron in the hidden layer, because the whole idea of neural networks is to build the model by recursive use of two operations, the formation of *linear combinations* of "inputs" from the previous layer, and the *transformation* by the "activation function" to produce the "output", serving as "input" to the next layer. In this respect, the model with a single neuron in the layer before the last one is a degenerate model.

Thus, the smallest nontrivial model which has all the characteristic features of a neural network is

4.3 The neural network with one hidden layer consisting of two neurons

This model can be written

$$y_i \approx F(\alpha_0 + \alpha_1 F(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}) \\ + \alpha_2 F(\gamma_0 + \gamma_1 x_{i1} + \cdots + \gamma_k x_{ik})),$$

where, again, we assume that F is the inverse logit function and the right hand side is interpreted as the probability of the event $y_i = 1$.

The most remarkable feature of this model is, perhaps, its complexity. Even in the case of a single covariate x, the 7–parameter family of functions of the form

$$f(x) = F\left(\alpha_0 + \alpha_1 F\left(\beta_0 + \beta_1 x\right) + \alpha_2 F\left(\gamma_0 + \gamma_1 x\right)\right),$$

includes — just as an example — functions that increase from the asymptotic minimum at $-\infty$ to a global maximum, then decrease to a local minimum, and finally increase to the asymptotic value at $+\infty$. In the case with two covariates things become even more complicated. The typical function of the form

$$f(x_1,x_2) = \\ F\left(\alpha_0 + \alpha_1 F\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2\right) + \alpha_2 F\left(\gamma_0 + \gamma_1 x_1 + \gamma_2 x_2\right)\right),$$

has four different asymptotic values for (x_1, x_2) escaping towards the horizon in different directions. It is not a triviality to discuss how the shape of this function depends on its nine parameters, and it is almost impossible to imagine what happens when two or more layers with several neurons are allowed.

In conventional statistics, there is a very hesitative attitude to the use of models as complicated as this. The reason for this is not so much the computational difficulties — they can be overcome — but the fact that the whole purpose of a statistical analysis is to explain the data as the result of a process involving two components, the *systematic variation*, represented by the statistical model and its parameters, and the *random variation*, represented by the actual outcome of the random model (in its most concrete form, the error terms in a regression model). The impossibility of the drawing of a sharp borderline between randomness and complexity — as most recently emphasized by chaos theory, and more implicitly contained in the traditional

concept of "overfit" — makes it not only desirable, but necessary, for the statistician to avoid complexity as far as possible.

A simple illustrative example from more traditional statistics comes from polynomial regression. If an ordinary, linear regression model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ fails to give a satisfactory description of data, it may be because there is some curvature that can not be catched by a linear function. A simple way to include this is by quadratic regression, $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$. If this is not good enough, a third degree term can be added, and so on. For every term of higher degree we add, the fit becomes better in the sense that the square sum of the differences between observed and fitted values becomes smaller. But uncritical continuation of this process will obviously result (most extremely when degree $n-1$ is reached) in a more and more perfect fit of a function which becomes more and more useless for extrapolation. If one fits an unnecessarily complicated model to a "training set", its ability to predict correctly in a "test set" will usually be poor. This is the kind of problems we would expect to run into with the neural network models, in particular those with several layers consisting of many neurons.

Of course, this is not a principal criticism of the neural network models, nor of polynomial regression models of degree 10. What we are saying is just that there are so many other more simple modifications that one can make of the basic logistic regression model, that it is difficult to imagine situations where we would end up with something as complicated as this. But the possibility exists, of course, and in section 6 we shall try to fit a neural network model with one hidden layer of two neurons, to see what comes out of it.

5 Time trends, censoring and interventions

Until now we have concentrated on predicting whether or not a customer goes bankrupt within a certain time period. Historical data do not only provide information on whether the customer payed the debt or not, but will often include more detailed information about the repayment. Generally, registrations of the date of repayment, the age of the loan, customers changing financial ability, e.g. getting married or divorced or — in order to reduce a potential loss — interventions, such as reduced rates for customers unable to pay the full rate, are possible registrations.

The statistical method for dealing with this sort of more detailed informa-

tion is known as event history analysis or single/multiple spell analysis, well known in e.g. applied labour economics (Heckman and Singer (1985) and Lancaster (1990)), and synonymous with survival analysis, extensively used in biostatistics (Andersen et al. (1993)). In the present context the idea is to describe the waiting time, T, from duedate until payment. This can be infinite and the customer is bankrupted, but as discussed in Section 1, it is convenient to use the term bankruptcy if the customer fails to pay within a time period of a given length, say T_0. The hazard, $\lambda(t)$, for T plays a central role in event history analysis. $\lambda(t)$ is defined as the conditional density for T at t, given $T > t$, — that is, $\lambda(t)dt$ is the probability of payment within an interval of length dt immediately after t given that the customer has not payed at time t. The cumulative distribution function, F, for the waiting time may be uniquely characterized by the hazard function

$$1 - F(t) = P(T > t) = \exp(-\int_0^t \lambda(u)du).$$

Notice that the term hazard is awkwardly used here for the "risk" of payment. The influence of the covariates is frequently modelled as

$$\lambda(t, x_1, \ldots, x_k) = \lambda_0(t) \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k),$$

known as Cox's proportional hazards model (Cox (1972)). $\lambda_0(t)$ is the common underlying hazard function for all customers. On a logarithmic scale the hazard is modelled as a linear combination of the covariates. Assuming that the hazard function is constant over time, say $\lambda_0(t) \equiv \lambda$, the model reduces to the exponential regression model for waiting times.

The real advantage of using event history methods in the analysis of historical credit data is that it is possible to include not only the information on when (if ever within the time limit) the customer payed the loan but also time–varying covariates and censoring of observations can be modelled. Right censored data appear frequently. Waiting times for customers not paying before T_0 are censored observations, but also e.g. interventions from the financial institution such as reducing the size of payment can be dealt with as censored observations. Important time–varying covariates may be marital and/or job status.

In Section 2 we discussed the logistic regression model for the probability of going bankrupt, $\mathrm{P}(T > T_0)$. Using the log(–log) transformation instead of the logit transformation for the probability of bankruptcy modelled as

above the influence of the covariates is seen to be linear. The constant term in the log(–log) regression then reflects the integrated underlying common hazard for bankruptcy; λT_0, in the model with constant hazards. However, customers' payment behaviour can hardly be described by a constant hazards model. Constant hazard may be appropriate over small time intervals, but the hazard for payment may be expected to decrease with time. Piecewise constant hazards can be handled by dividing the time period into disjoint intervals defined by $0 < t_1 < \cdots < t_m < T_0$. Then the probability of bankruptcy can be written as a product of conditional probabilities

$$P(T > T_0) = P(T > t_1)P(T > t_2 \mid T > t_1) \ldots P(T > T_0 \mid T > t_m),$$

where each factor on the right hand side is of the same form as $P(T > T_0)$ above. Moreover, the influence of the covariates on the overall probability of bankruptcy is additive on the log(–log) scale, and the constant term is a weighted sum of common underlying hazards. In this way event history models and models with binary outcome are closely related as long as the covariates are well defined from the start of the period of observation and the influence of the covariates is assumed not to change with time.

A further advantage is that it is possible to introduce piecewise constant coefficients, β's. For instance, some customers could be in arrears with the payment simply because of human mistakes or absentmindedness and will pay immediately after becoming aware of the mistake. These people definitely have charateristica different from those being in real financial trouble, and still not having payed after, say, three months. Analysis of the conditional probabilities above, either by a logistic, a log(–log) or a Cox's regression model makes it possible successively to point out customers repaying in the near future. With this kind of information, the financial institution will be able to determine which of the customers it will be worth offering special attention.

6 A case study

6.1 The data

In 1795 a fire broke out in Copenhagen and a quarter of the city burnt down. The fire was followed by an acute need for credit to rebuild the lost

homes, and this need occasioned in 1797 the establishment of the first Danish mortgage credit institution, later to become a part of BRF-Kredit, the third largest mortgage credit institution in Denmark.

For the past two centuries the Danish mortgage credit market has been controlled and regulated through strict acts with the three latest large revisions in 1970, 1980 and 1989. In 1970 the standard mortgage system was introduced allowing the borrower to raise loans from a single institution based on first and second mortgage lending. In 1980 a reform established the current principle of financing up to 80% of a property's value through first mortgage credit from one of the, at that point in time, four authorized first mortgage credit institutions (FMCI). The remaining 20% must be financed through other sources (e.g. banks, private resources, etc.). In 1989 authorization of FMCIs was liberalized allowing establishment of new FMCIs including foreign investors.

The main principle of first mortgage credit to private housing is as follows: Up to 80% of the cash selling value of the property can be financed through the FMCIs. The credit is typically established through 30 years fixed interest rate annuity loans with quarterly settling periods. The capital for the loan is raised by the credit institute through the issuing and selling of bonds with a fixed integer interest rate around the actual market interest rate, and the price of the bonds determine the effective interest rate of the annuity loan. If the client wants to pay back the loan before it is end of the instalment, this is done by buying the bonds at the size of the de facto price at the payback time.

Two factors influence the size of the first mortgage relative to the sales value of the property:

1. The bonds are negotiable on the open stock market and the price will of course alternate with the market interest rate (e.g. the bond prices increase when market interest decreases and vice versa). This means that if the client wants to pay back his loan before the end of the instalment (typically when selling the property) he may face increased debt due to a decreased market interest rate (one can at a premium rate take precaution against this situation when establishing the loan, but as that will cost a premium rate, far from all make that guard).

2. As the FCMI-loans have a very long settling period, many houseowners use the opportunity given by increasing houseprices to raise capital for

consumer expenditures through constantly having FCMI loans on 80% of the actual property sales value and still have some of the remaining 20% financed from banks or similar sources.

Thus if houseprices and/or market interest drops significantly the quoted value of the FCMI loan may become higher than the sales value of the property!

The international economic boom in the beginning of the 1980s resulted in a strong growth in house prices in Denmark. From 1982 to 1986 the average house prices increased by 83%. The boom combined with the above-mentioned tradition of financing consumer expenditures through additional FCMI loans resulted in an overheated economy. In June 1986 the conservative-liberal government therefore introduced two law complexes to reduce the speed of the economy. One complex consisted of increased taxes, rates and dues and a compulsory saving, the other reduced the tax reduction of private interest expenditures (like mortgage interest) from typically 70% to 48%. Because of these initiatives house prices dropped within the next 18 months with approximately 30%. In addition the internationally influenced decreasing market interest was intensified resulting in a decrease of market interest from 19% in January 1983 to 10% in June 1986.

As a result a large portion of the Danish houses was mortgaged much higher than the sales value of the property. This created two problems:

1. Quite a few house owners (especially new owners) could not afford their mortgage expenses due to the effect on their private economy of the two law complexes.

2. A large amount of house owners trying to sell their house experienced that the value of the house was far below the quoted value of the total mortgage credit.

Consequently, the housing market faced a boom in numbers of sales by order of the court with high losses for the FMCIs as a result.

In 1990 BRF-Kredit introduced a program to reduce their loss on compulsory sales. The clients (customers) of the institute could be divided into the following 4 groups:

Group 1: Customers repaying the loan prompt on schedule.
Group 2: Customers repaying the loan, but sometimes late of schedule.
Group 3: Customers not able to repay the loan unless they got some respite and use of this respite to restructure their economy.
Group 4: Customers unable to repay the loan.

The potential losses on the mortgage loans did of course occur in group 3 and 4 and studies showed that the earlier these customers were identified and contacted to either establish respite (group 3) or compulsory sale (group 4) the smaller (if any) the losses would be.

Part of the restructuring program consisted therefore in a study to build a prediction model for immediate identification of the members of group 3 and 4 among those not paying a given settling period due. The task consisted of the following: "The first day after the quarter due date to identify group 3 and 4 members among those having not paid the instalment".

Due to the Danish rules of electronic registration of people and the rules of FMCIs the loans were solely registrated with objective information like: name, age and sex of the customer, size, age, interest, settling period and past payment history of the loan, size, location, type and age of the property.

Part of this information form the base of the case study of this paper. We have selected data from two settling periods to illustrate our conclusions. The first data set consists of those loans not paid due in the 4'th settling period 1989 (due day 31'st of December). The second consists accordingly of the past due of the 3'rd settling period 1990 (due day 30'st of September). The 1989 data are used to build the prediction models and the 1990 data are used to test the prediction models. All data are made anonymous and only part of the available information is brought into use. Loans with a past due payment later than 6 months are considered as bankrupted.

In December 1989, 7941 private house loans were not paid due. Hereof 776 (9.7%) failed to pay within 6 months. The equivalent September 1990 data consisted of 7699 not paid due out of which 1114 (14.5%) was not paid after 6 months. For the purpose of this paper we have selected the following information about the loans:

Social, a factor on four levels defining whether the loan is guaranteed by a single individual or a couple, two persons — males and/or females.

Singles are grouped according to gender and couples are grouped according to whether they do have same address or not.

Age, the age of the primary mortage holder. The age is registered in years.

County, defining in which county the property is situated. The factor has 16 levels.

Property, defining whether the property is a house or an appartment.

Mode, an indicator for whether the mode of payment has been changed from automatic to manuel or not.

Debt, the total outstanding debt in BRF-Kredit.

Instalment, the size of the total instalment in BRF-Kredit the given settling period.

Percent, defined as the percentage of the property's value financed through FMCI's.

Respite, defining whether or not the customer did ask for respite before the duedate.

Trade conditions, a time factor describing the market trend constructed according to when the loan is raised: Loans raised before 1986, loans raised from 1986 until end 1988 and loans raised after 1988.

Term before, an indicator showing whether the instalment immediately before the present was payed duetime.

6.2 A logistic regression model

For the use in the logistic regression analysis it turned out to be convenient to categorize the age of the mortage holder into four groups namely below 30 years, 30–40 years, 40–50 years and more than 50 years of age. Preliminary marginal analyses showed that all covariates, except for the total instalments (χ^2=0.84, df=1, p=36.8%) and whether or not the customer did ask for respite (χ^2=2.31, df=1, p=12.8%), had significant effect on whether the customer went bankrupt or not. While the degree of the influence of the trade factor (χ^2=8.49, df=2, p=1.4%) and the total outstanding debt (χ^2=4.59, df=1, p=3.2%) were modest, the influence of the remaining factors were highly significant. The above numbers in brackets are the calculated –2log likelihood ratio test statistic, the degrees of freedom in the corresponding χ^2 –distribution and the resulting test probability for the hypothesis of no influence.

The multiple logistic regression analysis was performed by a forward/backward procedure. For a start, two factor interactions were successively added to the model with all factors included as main effects. Only the interactions between age and the social factor respectively age and the type of property were significant. After adding these two interactions to the model, succesive elimination of covariates revealed that trade conditions not affects the probability of bankruptcy (χ^2=0.013, df=1, p=90.8%). Furthermore, Percent (χ^2=0.822, df=1, p=36.5%), Debt (χ^2=0.018, df=1, p=89.3%) and Instalment (χ^2=2.10, df=1, p=14.8%) were non–significant and removed from the model. The remaining factors were all highly significant (p<0.0005) and the final model became

$$(1) \qquad \text{logit}\,(p) = \beta_0 + \beta_{age} + \beta_{social} + \beta_{property} + \beta_{mode} + \beta_{county} + \beta_{respite} + \beta_{term\,before} + \beta_{age*property} + \beta_{age*social}.$$

The adequacy of the above model was supported by the Hosmer and Lemeshow goodness–of–fit statistic. The test statistic was 12.8, which evaluated in a $\chi^2(8)$ –distribution corresponds to p=0.12. Some estimated parameters from the above model are seen in table 1. Due to overparametrization the parameters are relative to a reference category. The parametrizations are indicated in the subscripts to the $\hat{\beta}$'s , e.g. $\hat{\beta}_{mode=changed}$ is the estimated effect on the logit transformed probability of bankruptcy of *changing* the mode of payment from automatic to manuel compared to *not changing* the mode of payment. The interpretation is that changing the mode of payment increased the probability of bankruptcy. Similary, those having asked for respite had a higher probability of bankruptcy, while having payed late the instalment before decreased (!) the probability of getting bankrupted.

The estimates of the factor county and the interaction terms are crowded out. The estimates of the county parameters did reflect that the county ef-

TABLE 1
Selected estimates, standard errors (SE), odds ratios and corresponding 95% confidence limits.

	Estimate	SE	Odds Ratio	95% Confidence limits
$\hat{\beta}_{mode=changed}$	0.553	0.085	$\hat{OR}_{changed/not\,changed} = 1.74$	(1.46 , 2.05)
$\hat{\beta}_{respite=no}$	-0.204	0.093	$\hat{OR}_{not\,asked/asked} = 0.816$	(0.680 , 0.978)
$\hat{\beta}_{termbefore=late\,payment}$	-0.746	0.091	$\hat{OR}_{late/duetime} = 0.474$	(0.397 , 0.567)

TABLE 2
Classification results using the logistic regression model (1)

Historical data							
	Correctly predicted as:		Incorrectly predicted as:		Percentages		
Probability level	Bankruptcy	Non-bankruptcy	Bankruptcy	Non-bankruptcy	Correct	False positive	False negative
0.10	534	3181	2447	206	58.3	82.1	6.1
0.20	198	5022	606	542	82.0	75.4	9.7
0.30	80	5503	125	660	87.7	61.0	10.7
Test data							
	Correctly predicted as:		Incorrectly predicted as:		Percentages		
Probability level	Bankruptcy	Non-bankruptcy	Bankruptcy	Non-bankruptcy	Correct	False positive	False negative
0.10	510	2986	2117	448	57.7	80.6	13.1
0.20	192	4602	501	766	79.1	72.3	14.3
0.30	57	4999	104	901	83.4	64.6	15.3

fect can be due to geographical economic variations, probably variations in the unemployment rate. Further, some main tendencies were that couples with different addresses had an increased probability of going bankrupt for all age groups except for the oldest customers. Younger customers in appartments, between 30 and 40 years of age, had higher risk of going bankrupt than older customers.

For classification and validation purposes the predicted probabilities of bankruptcy were calculated, both for customers in the historical data set and for customers in the test data set. Based on the predicted probabilities, customers may be classified as bankrupts or non-bankrupts, according to whether the predicted probability exceeds a given level or not. In table 2 the classification results are given for each of the probability levels 0.10, 0.20 and 0.30. The results obtained from the test data set are almost similar to the results for the historical data set, the percentage of correct classifications being only a few percent lower than for the historical data set.

6.3 Comparison of methods

Suppose we have a method for identification of bankrupters which, on the basis of experience obtained from the training set, suggests some way of scoring the costumers' risks of going bankrupt as a function of their covariates.

Figure 1a

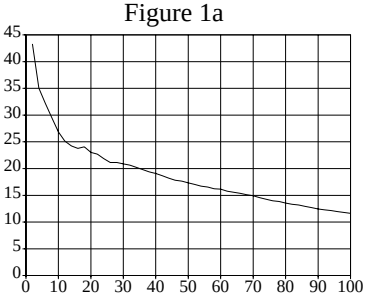

Figure 1b

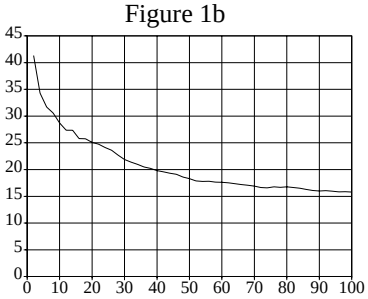

Figure 2a

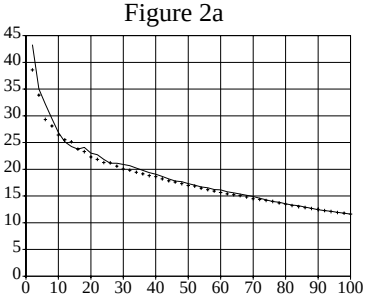

Figure 2b

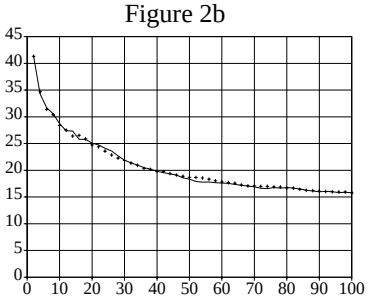

Figure 3a

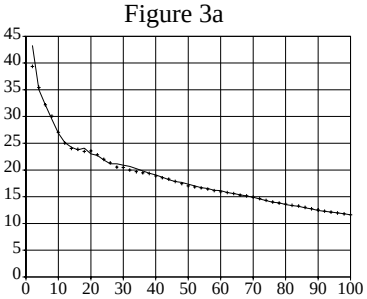

Figure 3b

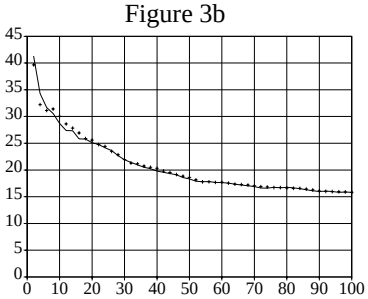

Figure 4a

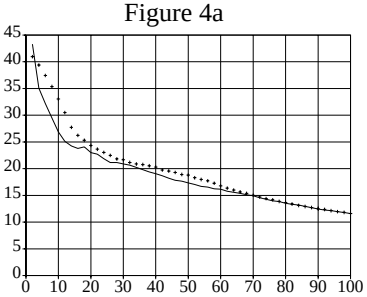

Figure 4b

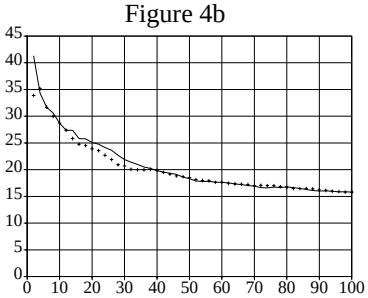

Figure 1-4. Illustrations to section 6.3.

This may be in terms of an estimated probability of bankruptcy, or the logit of this. A simple way of illustrating the method's ability to pick out the bankrupters before the non–bankrupters is by a "performance plot" which, for any percentage P between 0 and 100 shows how many of the highest scored P % that actually went bankrupt. This is just a way of displaying the whole continuum of tables of the form given in table 2 for the selected score threshold values 0.10, 0.20 and 0.30. However, for cosmetic reason we use only the values P =2%, 4% , ..., 98%, 100%. Notice that this way of evaluating a method depends only on the ordering of the customers by scores. For example, the performance plot for a discriminant analysis model is independent of the prior probability occurring in the formula for the probability of bankruptcy.

Figure 1a and 1b show the performance plots for the training set and the test set, respectively, for the logistic regression model (1) described in section 6.2. Just to make sure that the definition is clear, the fact that the point (35,20) is (almost) on the broken line means that among the 35% highest scored customers, 20% actually went bankrupt.

What is really interesting here is, of course, the plot for the test set (figure 1b), since a high predictive power in the training set may be due to overfit. As in section 6.2, the scores for the test set are computed with parameters estimated from the training set.

We have drawn the plot for this model as a broken line. For the other models considered, only the breakpoints will be marked, with the broken line for the present model overlayed as a sort of common baseline.

Three other models were considered.

(2) A logistic regression model with the same factors as in (1), but with omission of the two interaction terms. These interactions were strongly, but not astronomically significant (p=0.004 for age∗social, p=0.0003 for age∗property), and it is of some interest to see how sensitive the performance is to this kind of "oversimplification". For the training set (figure 2a), the simplified model seems to perform worse than the model with interactions, but for the test set (figure 2b) the difference is less pronounced. The reason may be that the estimation error for the many interaction parameters — most of which are probably insignificant, in some sense — makes the procedure based on the model without interactions more stable. This supports the idea that overfit is a serious error source in this kind of problems. Though

the conclusion is not so clear in this case, the general recommendation should certainly be to make the model as simple as possible, by removal of all terms that are not unambiguously significant.

(3) An ordinary discriminant analysis of the factor dummies occuring in model (1) gave the results shown by figure 3a and 3b. For the training set, this shows almost the same performance as the corresponding logistic regression. For the test set there are some small differences, but they do not point uniquely in any direction. Hence, in spite of the model assumptions (which are really absurd here), the discriminant analysis performs well in this case. However, since there is no obvious way of anticipating this phenomenon other than performing the logistic regression, we cannot recommend the method in general (cf. section 3).

(4) A neural network model with one hidden layer consisting of two neurons was fitted, with both neurons having the same linear structure as the logistic regression model (1). The fit was difficult to obtain, because the log–likelihood is not too well behaved. With Fisher's scoring method, the parameters had a tendency to dissappear towards infinity, probably because there are many "boundary" models that fit well. By suitable adjustment of the rate of divergence, different kinds of models (including logistic regression models, and models describing the probability by step functions of linear combinations of the covariates) can come up as as limiting models. For this reason, we worked with a weakly penalized likelihood. In Bayesian terms, we took a rather flat normal distribution as the prior and computed the maximum of the posterior density instead of the maximum for the likelihood. Moreover, it was necessary to start the algorithm with initial values obtained by a more primitive search by a controlled random walk. This was tryed several times, and the result presented here is the one that gave the highest (unpenalized) likelihood. The model has 38+38+3=79 parameters, and the gain in likelihood over the logistic regression model (1) with 38 parameters was 136.5, which (on 41 degrees of freedom) corresponds to a formal p–value of the order of magnitude 10^{-12}. Hence, there is indication of a strongly significant improvement over the logistic model, and this is indeed supported by figure 4a. But figure 4b strongly suggests that this is an artefact due to overfit. For the test set, the neural network model performs worse than any other model we have considered.

6.4 Event history analysis

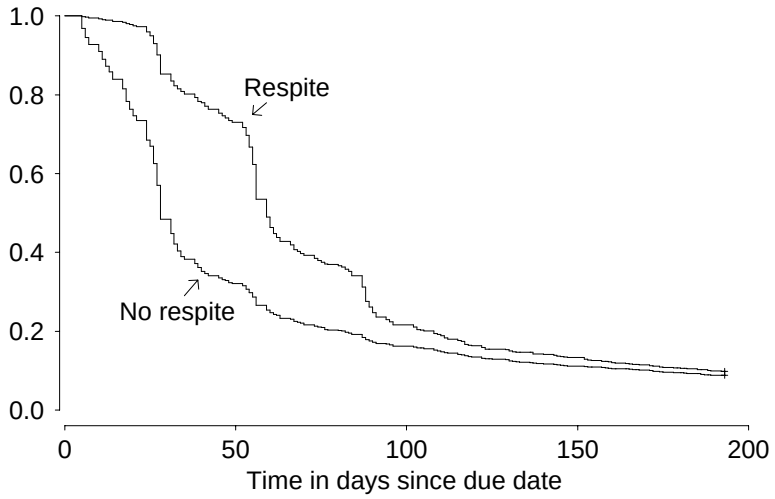

Figure 5. Empirical waiting time probability for customers having asked for respite and customers not having asked for respite.

As described in Section 5 time trends may be studied by analysis of data sets originating from partitioning the time axis in smaller intervals and successively conditioning with the event that the given customer has still not payed the instalment. The non-parametric maximum likelihood estimator for the probability that the waiting time from duedate until payment exceeds t, $T > t$, known as the Kaplan–Meier estimator, is seen in figure 5. Here the Kaplan–Meier estimator is calculated after stratification with respect to whether or not the customer did ask for respite before duedate.

From figure 5 it is seen that only a few of those having asked for respite pay in the weeks just after duedate whereas several of those not having asked for respite pay within a few weeks. Contrary, those having asked for respite pay later. If we only consider the part of the customers who still have not

payed after 30 days from duedate, the probability of bankruptcy still depends significantly on whether there has been asked for respite or not. However, when modelling the above conditional probability of bankruptcy by the logistic regression model (1) the estimated effect becomes larger and the sign changes. The estimated parameter is $\hat{\beta}_{respite=no} = 0.406$ with corresponding $\hat{OR}_{not\,asked/asked} = 1.50$ (1.24, 1.81). So, here we have an example of a time dependent effect of a covariate. The implication for the financial institution is that customers not having asked for respite and not having payed after 30 days from duedate are worth paying special attention, even though the overall effect was an increased risk of bankruptcy for those having asked for respite.

7 Conclusion

This paper has examined three of the most popular methods for credit scoring: logistic regression, discriminant analysis and neural networks. The paper discusses those applications where the given credit evaluation can be viewed as a sample from a population, of which the observations have characteristics known to the credit association through historical data around credit behaviour. The credit scoring can be either for the evaluation of new applications or the evaluation of existing loans.

Though many papers (e.g. Rosenberg and Gleit (1994), Hand and Henley (1997)) have presented surveys of methods for automatic credit scoring, we have missed a systematic evaluation of which methods that from a strict mathematical view lead to the best basis for decision. In the present paper we have examined the principles and assumptions behind the three methods and their mathematical and statistical implications regarding simplicity, interpretation of explaining functions and stability of prediction. The result is as expected (see also Wiginton (1980), Press and Wilson (1978), Hand and Henley (1997), Lucas (2001)) that logistic regression demonstrates best performance in all three areas.

The paper additionally discusses the opportunities in censoring and interventions and the use of time–varying coefficients for successive evaluation of loans (or customers) in arrears with payment. The benefit of using time–varying coefficients can improve the decision base significantly as exemplified in the illustrating case from BRF-Kredit where the coefficient of the covariate

respite changes sign when time passes 30 days.

In summary we conclude that regarding credit scoring logistic regression is superior to discriminant analysis and neural networks though the latter two methods in given applications may show performance matching logistic regression.

Regarding neural networks the analysis shows that the models, even in simple versions, contain very complex prediction functions. As a consequence of this these models have a build–in tendency to overfit, which makes it necessary, in practice, to split the data set in a training set and a test set.

8 References

Andersen, P.K., Borgan, Ø., Gill, R.D. and Keiding, N. (1993). *Statistical models based on counting processes.* Springer, New York.

Anderson, T.W. (1984). *An introduction to multivariate statistical analysis.* Second edition, Wiley, New York.

Cox, D.R. (1972). Regression models and life-tables (with discussion). *J. R. Statist. Soc. B*, **34**, 187-220.

Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *J. Amer. Statist. Assoc.* **70**, 892-898.

Fisher, R.A. (1938). The statistical utilization of multiple measurements. *Annals of Eugenics.* Vol. VIII, Pt.IV, 376-386.

Hand, D.J. and Henley, W.E. (1997). Statistical classification methods in cosumer credit scoring: A review. *J. R. Statist. Soc. A*, **160**, 523-541.

Heckman, J.J. and Singer, B. (1985). *Longitudinal analysis of labour market data.* Cambridge University Press, Cambridge.

Lancaster, T. (1990). *The econometric analysis of transition data.* Cambridge University Press, Cambridge.

Lucas, A. (2001). Statistical challenges in credit card issuing. *Applied Stochastic Models in Business and Industry.* **17**, 83-92.

Press and Wilson (1978). Choosing between logistic regression and discriminant analysis. *J. Amer. Statist. Assoc.* **73**, 699-705.

Ripley, B.D. (1994). Neural networks and related methods for classification (with discussion). *J. Roy. Statist. Soc. B*, **56**, 409-456.

Rosenberg, E. and Gleit, A. (1994). Quantitative methods in credit management: A survey. *Operations Research*, **42**, p589-613.

Sarle, W.S. (1994). Neural networks and statistical models. *Proceedings of the Nineteenth Annual SAS Users Group International Conference*, SAS Institute Inc., Cary, NC.

Wiginton, J.C. (1980). A note on the comparison of the logit and discriminant models of consumer credit behaviour. *J. Finan. Quant. Anal.*, **15**, p757-770.