Dept. of Math. University of Oslo
Statistical Research Report No. 1
ISSN 0806–3842 January 2005

Bankruptcy Prediction by Generalized Additive Models

Daniel Berg*†
Department of Mathematics,
University of Oslo, Norway.

Abstract

We compare several accounting based models for bankruptcy prediction. The models are developed and tested on large data sets containing annual financial statements for Norwegian limited liability firms. Out-of-sample and out-of-time validation shows that generalized additive models significantly outperform popular models like linear discriminant analysis, generalized linear models and neural networks at all levels of risk. Further, important issues like default horizon and performance depreciation are examined. We clearly see a performance depreciation as the default horizon is increased and as time goes by. Finally a multi-year model, developed on all available data from three consecutive years, is compared with a one-year model, developed on data from the most recent year only. The multi-year model exhibit a desirable robustness to yearly fluctuations that is not present in the one-year model.
J.E.L. Subject Classification: C13, C14, C44, C51, C52, G33.
Keywords: Bankruptcy Prediction, Generalized Additive Models, Default Horizon, Performance Depreciation, Multi-year model.

*Address for correspondence: Department of Mathematics, University of Oslo, P.O. Box 1053 Blindern, NO-0316 Oslo, Norway.
†E-mail: daniel@nr.no.

Contents

1 Introduction

Since the work of Beaver (1966) and Altman (1968), bankruptcy prediction have been studied actively by academics and practitioners. This field of risk management continues to be very active, much due to the continous development of new financial derivatives. For example, the pricing of credit derivatives rely on good estimates of counterparty risk. The literature on bankruptcy prediction is extensive. Many models have been proposed and tested empirically, often with contradictory conclusions.

There are two kinds of models that are commonly adressed in the literature. First, there are accounting based models, for example discriminant analysis and logistic regression models. Second, there is market based models, for example Merton or Black-Scholes Merton (BSM) models (e.g. the Moody's KMV public firm model). The market models are based on the value of a firm set by the market. Stock prices are commonly used as proxies for the value. Market based models require that firms are registered on a stock exchange and this is quite often not the case. In Norway, the majority of limited liability firms are not registered on the exchange. Hence, our focus is on accounting based models.

Linear discriminant analysis models have been widely used. Altmans popular Z-Score (Altman, 1968) is for example based on linear discriminant analysis. Generalized linear models, or multiple logistic regression models have also been popular. Ohlsons O-Score (Ohlson, 1980) is based on generalized linear models with the logit link function, also referred to as logit analysis. Neural network models are powerful and have become a popular alternative with the ability to incorporate a very large number of features in an adaptive nonlinear model (Kay and Titterington, 2000), see for example Wilson and Sharda (1994). See also Altman and Narayanan (1997) for a survey of business failure classification models.

The main objective of this report is to introduce generalized additive models, GAM, as a flexible non-parametric alternative for bankruptcy prediction and show that it performs significantly better than discriminant analysis, linear models and neural networks. GAM is a non-parametric generalization of the linear regression model. It replaces the usual linear function of a covariate with a sum of unspecified smooth functions, helping us discover potential non-linear shapes of covariate effects. The shape of the smooth function is determined by the data through iterative smoothing operations. The estimation of neural networks and generalized additive models is computationally more demanding than for linear models, but with the rapidly increasing power of computers we expect an increasing application of such models in practice.

All models are developed using the same explanatory variables, and we follow the validation methodology that is referred to as *out-of-sample* and *out-of-time* validation in Sobehart et al. (2000). We first perform a preliminary analysis, identifying which variables that are significant in the various models. Variables that prove to be insignificant in all models are excluded from further analysis. The remaining variables are included in all models. All financial ratios are defined as the deviance from their industry mean, and for the neural networks all variables are scaled to the range $[0, 1]$. The focus of this paper is not on explanatory variables, but rather on model performance, given a set of explanatory

variables. So the choice of explanatory variables and their characteristics is not discussed in detail.

The data set used is a collection of annual financial statements of Norwegian limited liability firms in the period $1995 - 2000$. Each year contains statements of approximately 100.000 firms. We also have access to a record of all firms that filed for bankruptcy in the years $1995 - 2001$.

In addition to the performance comparisons we examine the sensitivity of different models to *default horizon*. If a model is developed using financial statement data from e.g. 1996, a 1 year default horizon model would define a firm as default if it failed during 1997, while a 2 year default horizon model would define it as default if it defaulted during 1997 or 1998. Next, we test the *depreciation* of the prediction models, examining how the prediction power of a model depreciate $0 - 4$ years into the future. This is very important to consider when determining cut-off levels and also when considering model risk. Finally the performance of a *multi-year model*, developed from statements of all firms in the period $1996 - 1998$, is compared with a one-year model, developed from 1998 statements only. The multi-year model proves to be more robust than the one-year model.

The report is organized as follows. Section 2 describes the problem of bankruptcy prediction and the models we will examine. Section 3 summarizes the data set and the explanatory variables, while Section 4 discusses model development and validation methodologies. Section 5 performs a preliminary study of the performance for various choices *within* each model, that is we compare *linear* and *quadratic* discriminant analysis, then generalized linear models with *logit* link and *probit* link is compared, and finally generalized additive models with *logit* and *probit* link is compared. Section 6 compares the prediction power of the various models, out-of-sample and out-of-time. Section 7 presents results from a GAM model with varying default horizon. Section 8 shows how a GAM model depreciates after development. That is, how a model performs on data contemporary with that of development, how it performs on data one year after that of development and two, three and finally four years after. Section 9 compares the performance of a multi-year model and a one-year model. Finally, Section 10 presents a summary of our findings and suggestions for future work.

2 Prediction Models

When handling our bankruptcy data it is natural to label one of the categories as success (healthy) and the other as failure (default) and to assign these the values 0 and 1 respectively. A typical data set will have a series of ones and zeros as the response variable Y. Associated with each Y there will often be observations on a set of explanatory variables $X_1, X_2, \ldots, X_p$. A bank will typically have information on the earnings and debt of each customer.

Since Altman (1968) proposed to use Linear Discriminant Analysis (LDA) to predict bankruptcy, several contributions have been made to improve Altman's results, using different parametric, semiparametric and non-parametric models.

In contrast to normal-based regression models like the LDA, in which we wish to predict the value Y, given values for the explanatory variables, we will also be interested in predicting the probability π that $Y = 1$, given values for the explanatory variables (Krzanowski, 1998). Any probability is restricted to take values between 0 and 1, but a linear model can give rise to any value between $-\infty$ and ∞. It is thus necessary to transform π into a quantity that takes values in the interval $(-\infty, \infty)$ before a linear model can be applied. There are several such transformations, or link functions. We will examine the following two:

- The logit transformation: $\varepsilon = \ln(\frac{\pi}{1-\pi})$, often denoted by $\varepsilon = \text{logit}(\pi)$.
- The probit transformation: $\varepsilon = \Phi^{-1}(\pi)$, where $\Phi(\cdot)$ is the cumulative normal distribution function.

2.1 Discriminant Analysis

Discriminant analysis (DA) is often the first approach to consider when discriminating between different groups of objects (Doumpos and Zopounidis, 1999). DA is a multivariate statistical technique that leads to the development of a discriminant function maximizing the ratio of among-group to within-group variability, assuming that the variables follow a multivariate normal distribution and that the dispersion matrices of the groups are equal. Clearly, both of the assumptions pose a significant problem for the application of DA in real-world situations, since they are difficult to meet. The discriminant function can be linear or quadratic, and the corresponding models are referred to as Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA).

Some relevant details of LDA are as follows. We assume that the conditional density of the predictors in each class, denoted by $P(X|G)$, is multivariate Gaussian with each class having its own mean vector, but sharing a common covariance matrix. The density of class j is

$$\phi(X; \mu_j, \Sigma) = (2\pi)^{-p/2}|\Sigma|^{-1/2}\text{exp}(-\frac{1}{2}(X - \mu_j)^T\Sigma^{-1}(X - \mu_j)). \quad (1)$$

The class prior probabilities are $P(G = j) = \pi_j$. In this idealised setting, where everything is known, we can also obtain the ideal or Bayes optimal classifier. We will use Bayes' formula to flip the densities into class posterior probabilities $P(G|X)$. If the new observations to be classified arise from this same joint distribution, the rule

$$C(x) = j \text{ if } P(G = j|x) = \max_l P(G = l|x)$$

achieves the minimum misclassification rate. In this case we have

$$
\begin{aligned}
P(G=j|X=x) &= \frac{P(X=x|G=j)\cdot\pi_j}{P(X=x)} \\
&= \frac{\phi(x;\mu_j,\Sigma)\cdot\pi_j}{\sum_l \phi(x;\mu_l,\Sigma)\cdot\pi_l} \\
&= \frac{(2\pi)^{-p/2}|\Sigma|^{-1/2}\cdot\exp(-\frac{1}{2}x^T\Sigma^{-1}x)\cdot\exp(x^T\Sigma^{-1}\mu_j-\frac{1}{2}\mu_j^T\Sigma^{-1}\mu_j)\cdot\pi_j}{(2\pi)^{-p/2}|\Sigma|^{-1/2}\cdot\exp(-\frac{1}{2}x^T\Sigma^{-1}x)\cdot\sum_l\exp(x^T\Sigma^{-1}\mu_l-\frac{1}{2}\mu_l^T\Sigma^{-1}\mu_l)\cdot\pi_l} \\
&\propto \exp\left(x^T\Sigma^{-1}\mu_j-\frac{1}{2}\mu_j^T\Sigma^{-1}\mu_j+\log\pi_j\right) \\
&= \exp(x^T\beta_j+\epsilon_j) \\
&= \exp(-\delta_j), \qquad (2)
\end{aligned}
$$

where $\beta_j = \Sigma^{-1}\mu_j, \epsilon_j = \log\pi_j - \frac{1}{2}\mu_j^T\Sigma^{-1}\mu_j$ and δ_j denotes the discriminant function for class j. The equivalent rule is to classify to the class for which the $\delta_j(x)$ is smallest. The $\propto$ denotes proportionality. We are concerned only with the numerators since the denominators do not depend on the class label. Note also that the quadratic terms cancel. The decision boundary between class i and class j is defined as the set of points having equal posterior probability: $\{x \in R^p : P(G=i|x) = P(G=j|x)\}$. From (2) we see that in the case of LDA, this decision boundary is linear. In the case of QDA we do not assume that the classes share a common covariance matrix. Then, the quadratic terms will not cancel and the decision boundary will be quadratic.

The choice of DA as one of the models for comparison is based on its popularity among financial researchers in addressing financial classification problems such as bankruptcy prediction. This popularity of DA models is much due to the work of Altman (1968) and the relative ease with which these models can be implemented.

2.2 Generalized Linear Models

The class of Generalized Linear Models (GLM) was introduced as a generalization of the general linear model

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon, \qquad (3)$$

where ϵ has mean vector $\mathbf{0}$ and covariance matrix $\sigma^2\mathbf{I}$ (Krzanowski, 1998). The generalization makes use of the exponential family of distributions,

$$f(y;\theta,\phi) = \exp\{[y\theta - b(\theta)]/a(\phi) + c(y,\phi)\} \qquad (4)$$

for some specific functions $a(\cdot), b(\cdot)$ and $c(\cdot)$ and parameters θ and ϕ.

The GLM has the following features:

1. The $Y_i's (i = 1,\ldots,n)$ are independent random variables sharing the same form of distribution from the exponential family.

2. The explanatory variables provide a set of linear predictors $\eta_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}$ for $i = 1, \ldots, n$.

3. The link between 1 and 2 is that $g(\mu_i) = \eta_i$, where μ_i is the mean of Y_i for $i = 1, \ldots, n$. $g(\cdot)$ is called the link function of the model.

Two extensions of the general linear model (3) that characterize the generalized linear model, are its applicability to any member of the exponential family of distributions, and the presence of a link function when connecting the linear predictor η to the mean μ of Y. For a given problem the analyst may have to try more than one link function before deciding on the best model. However, a simplification is introduced if the chosen link function is the same as the function that defines the *canonical parameter* for the relevant distribution. Then the link function is called the canonical link (Krzanowski, 1998). Some standard distributions and their canonical links are given below.

- Binomial distribution: logit link $g(\mu) = \ln\left(\frac{\mu}{n-\mu}\right)$, probit link $g(\mu) = \Phi^{-1}\left(\frac{\mu}{n}\right)$
- Poisson distribution: log link $g(\mu) = \ln \mu$
- Normal distribution: identity link $g(\mu) = \mu$
- Gamma distribution: reciprocal link $g(\mu) = \frac{1}{\mu}$
- Inverse normal distribution: inverse square link $g(\mu) = \frac{1}{\mu^2}$

where Φ represents the cumulative normal distribution. We will consider the binomial distribution only. The two links are connected to the general linear model (3), where the distribution of the random term ϵ determines the link. If ϵ is normally distributed, we use the probit link and the general linear model is referred to as the probit model. If ϵ is logistically distributed, we use the logit link and the general linear model is referred to as the logit model.

First consider the probit model. The linear predictor is η_i and the model specified for the binomial parameter is $\pi_i = \Phi(\eta_i)$, which transforms the predictor from $(-\infty, \infty)$ to $(0, 1)$. The mean of the response variable, Y_i, from the binomial distribution is $\mu_i = n_i \pi_i$. Hence, the model can be re-expressed as $\mu_i = n_i \Phi(\eta_i)$ so that the linear predictor η_i is equal to $\Phi^{-1}(\mu_i/n_i)$. Thus, the link function is given by $g(\mu_i) = \Phi^{-1}(\mu_i/n_i)$, as stated above.

Choosing a logistic distribution, we would obtain the link function

$$g(\mu_i) = \ln\left(\frac{\mu_i}{n_i - \mu_i}\right).$$

Re-expressing this function in terms of π_i instead of μ_i, we obtain

$$g(\pi_i) = \ln\left(\frac{\pi_i}{1 - \pi_i}\right).$$

The choice of GLM with the logit link function is occasionally referred to as Logit Analysis (LA). We shall however refer to the models as GLM-Logit and GLM-Probit.

2.3 Generalized Additive Models

We will now discuss a further generalization of the linear regression model, Generalized Additive Models (GAM). This is an additive extension of the Generalized Linear Models. As mentioned in the introduction it is the aim of this paper to introduce this model as a superior alternative to the models commonly used today.

We briefly look again at the general linear model (3). This model makes a strong assumption about the dependence of $E(Y)$ on $X_1, \ldots, X_p$, namely that the dependence is linear in each of the predictors (Hastie and Tibshirani, 1991). There are many ways to generalize this linear regression model, one class of candidates is surface smoothers. They can be thought of as non-parametric proxies of the regression model and can be rather intuitively defined,

$$Y = f(X_1, \ldots, X_p) + \epsilon. \tag{5}$$

A problem that might arise with such models is that they may be difficult to interpret. How do we examine the effect of particular variables once we have fitted a complicated surface? For relatively low dimensional surfaces we can look at slices defined by conditioning on all but one of the variables, but this might be infeasible in higher dimensions (Hastie and Tibshirani, 1991). We will look at such slices when considering our model, as well as looking at how two and two variables relate to each other, holding all other variables constant.

2.3.1 Additive Models

The interpretation problem highlights an important feature of the linear model that has made it so popular for statistical inference: the linear model is additive in the predictor effects. Once we have fitted the linear model we can examine the predictor effects separately, in the absence of interactions. Additive models retain this important feature, they are additive in the predictor effects (Hastie and Tibshirani, 1991). An additive model is defined by

$$Y = \alpha + \sum_{j=1}^{p} f_j(X_j) + \epsilon \ , \tag{6}$$

where as before the errors ϵ are independent of the X_j's, $E(\epsilon) = 0$, $Var(\epsilon) = \sigma^2$ and $E(f_j(X_j)) = 0$. The f_js are arbitrary functions, one for each predictor.

The additive model has an a priori motivation as a data analytic tool. Since each variable is represented separately in (6), the model retains an important interpretive feature of the linear model: the variation of the fitted response surface holding all but one predictor fixed does not depend on the values of the other predictors. In practice this means that once the additive model is fitted to the data, we can plot the p coordinate functions separately

to examine the roles of the predictors in modeling the response. But this simplicity has a price. We must remember that an additive model almost always is an approximation to the true regression surface. But we hope that it is a useful one.

The estimated functions from an additive model are the analogies of the coefficients in linear regression. All the pitfalls encountered in interpreting linear regression models apply to additive models, and they can in many cases be expected to be more severe (Hastie and Tibshirani, 1991). It is important when using non-parametric additive models that we have much data, and this is indeed the case with our bankruptcy data.

2.3.2 Generalized Additive Models

The generalized additive models extend generalized linear models in the same way the additive model extends the linear regression model, that is, by replacing the linear form $\sum_j X_j\beta_j$ with the additive form $\sum_j f_j(X_j)$. The logistic additive model, when applied to binary response data, takes the form $\ln\left(\frac{\pi}{1-\pi}\right) = \sum_j f_j(X_j)$.

To compute the maximum likelihood estimates in a generalized linear model one would generally use some iterative-reweighted least-squares procedure. For the estimation of a generalized additive model, the linear regression step is replaced by a non-parametric additive regression step. The resulting algorithm is called local scoring, and is a minimizer of a penalized likelihood criterion (Hastie and Tibshirani, 1991).

Non-parametric models present, as we shall see, excellent results due to their robustness in detecting nonlinear relationships in the data. Conversely, they present a higher possibility of overfitting. Details of this and other properties of the generalized additive models can be found in Hastie and Tibshirani (1991).

2.4 Feed-forward Neural Networks

Neural networks (NNs) have received much attention in the field of bankruptcy prediction recent years. We will consider the class of NNs called 'feed-forward' NNs. These are sometimes also referred to as 'back-propagation NNs' or 'multi-layer perceptrons' (Ripley, 1996).

Feed-forward neural networks have units which have one-way connections to other units. The units can always be arranged in layers so that connections go from one layer to another. This is best seen graphically, see Figure 1. Each unit sums its inputs and adds a constant (the 'bias') to form a total input x_j and applies a function f_j to x_j to give output y_j. The links have weights w_{ij} which multiply the signals travelling along them by that factor. Thus a network such as Figure 1 represents the function

$$f_k(\mathbf{x}) = f_o\left(\alpha_k + \sum_{i=1}^{N} u_{ik}x_i + \sum_{j=1}^{M} v_{jk}f_h(\beta_j + \sum_{i=1}^{N} w_{ij}x_i)\right), \tag{7}$$

from inputs to outputs. Here N, M and K are the number of input nodes (i.e. the number of explanatory variables), the number of nodes in the hidden layer and the number of output nodes (i.e. the number of possible classes), respectively (Aas et al., 1999).

The general definition allows more than one hidden layer, and it also allows 'skip-layer' connections directly from input to output. It is also possible to avoid skip-layer connections in which case Equation (7) reduces to

$$f_k(\mathbf{x}) = f_o\left(\alpha_k + \sum_{j=1}^{M} v_{jk} f_h(\beta_j + \sum_{i=1}^{N} w_{ij} x_i)\right). \tag{8}$$

f_h and f_o are denoted activation functions. The function $f_h(x)$ of the hidden layer is always taken to be the logistic function (Aas et al., 1999):

$$f_h(x) = \frac{\exp(x)}{1+\exp(x)},$$

while the output function $f_o(x)$ may either be logistic or linear:

$$f_o(x) = \frac{\exp(x)}{1+\exp(x)} \text{ or } f_o(x) = x.$$

A neural network with no hidden layers is identical to the generalized linear model, while a neural network with one hidden layer, where the hidden layer uses nonlinear activation functions such as the logistic function is nonlinear in the parameters and corresponds to multivariate nonlinear logistic regression (Aas et al., 1999).

In practice the main issues are how the parameters, the weights, should be estimated, and how the architecture (the number of layers and the number of units in each, as well as which connections to include) is selected. The parameters may be estimated in at least three ways. Let $f_k(\mathbf{x}^p)$ and y_k^p be the value computed by the network and the true value (0 or 1) for the feature vector $\mathbf{x}^p$, respectively. If least squares fitting is used, we minimize

$$E = \sum_{p=1}^{P}\sum_{k=1}^{K}(y_k^p - f_k(\mathbf{x}^p))^2. \tag{9}$$

The second alternative is entropy (i.e. maximum conditional likelihood) fitting, where we minimize

$$E = \sum_{p=1}^{P}\sum_{K=1}^{K}\left[y_k^p \log\frac{y_k^p}{f_k(\mathbf{x}^p)} + (1-y_k^p)\log\frac{1-y_k^p}{1-f_k(\mathbf{x}^p)}\right]. \tag{10}$$

The last option is to use the softmax method. In this case the output function must be linear. The function to be minimized is

$$E = \sum_{p=1}^{P}\sum_{K=1}^{K} y_k^p \log\frac{y_k^p}{p_k^p}$$

where

$$p_k^p = \frac{\exp(f_k(\mathbf{x}^p))}{\sum_{j=1}^K \exp(f_j(\mathbf{x}^p))}.$$

For all three methods weight decay may be used. This means that instead of E we minimize

$$E + \lambda \left(\sum_{k=1}^{K} \alpha^2 + \sum_{k=1}^{K} \sum_{i=1}^{N} u_{ik}^2 + \sum_{k=1}^{K} \sum_{j=1}^{M} v_{jk}^2 + \sum_{j=1}^{M} \beta_j^2 + \sum_{i=1}^{N} \sum_{j=1}^{M} w_{ij}^2 \right). \tag{11}$$

The use of weight decay seems both to help the optimization process and to avoid overfitting. Suggestions have been made that $\lambda \in (0.01, 0.1)$ for the entropy fit (Aas et al., 1999).

A comprehensive discussion of neural networks can be found in Ripley (1996).

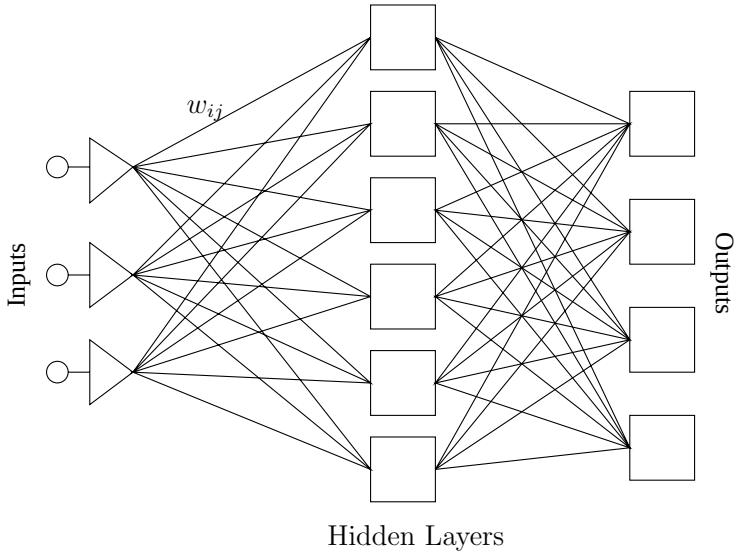


Figure 1: A generic feed-forward network with a single hidden layer.

3 Data

We have access to annual financial statements of all limited liability firms registered at the Norwegian register for business enterprises over the years $1995-2000$. We also have access to bankruptcy data from $1995-2001$ prepared by Dun & Bradstreet. These two data sets are merged and various variables are calculated. The resulting 5 data sets $D_{96}, D_{97}, D_{98}, D_{99}$ and D_{00} all include a company identification number, all explanatory variables examined and the year of bankruptcy. For firms that are not registered in the bankruptcy data set,

the year of bankruptcy is set to 'NA'. When referring to, for example, a model developed from 1996 data with a 2 year default horizon, we mean a model developed from the data set D_{96}, where a response variable is defined as 1 if the year of bankruptcy is 1997 or 1998 and 0 otherwise.

We remove the most extreme cases of the explanatory variables to prevent contaminated data from distorting the models. The appropriateness of this can be discussed, however our objective is to compare the predictive power of various models. If a model is to be used in practice by e.g. a bank, extreme cases should be incorporated in an appropriate manner. We also remove firms with total assets, revenue from operations, current liabilities, or book value of equity equal to zero, since such values will produce nulldivisions for some of the financial ratios considered. This is also prone to discussion. If a firm has, for example, current liabilities equal to zero then there is also no risk involved with this firm. Also, if a firm has total assets equal to zero, the company is bankrupt.

A particular feature of the data is the very small number of defaults. Of approximately 100.000 firms each year only about 1% defaulted the next year. This is representative of this kind of discrimination problem. Bankruptcy is a rare and extreme event. Suggestions have been made to ways of increasing the number of defaults when developing models, see for example Sobehart et al. (2000). Since we have such a large data set, 1% of 100.000 firms is still 1.000 firms, which is enough defaults to develop and validate models in an appropriate manner. When developing and testing models we split the data set used into two sets, a training set and a test set, containing 60% and 40% of the data, respectively, as discussed in Section 4.

3.1 Explanatory variables

The choice of, and investigation of explanatory variables is not the main objective of this paper. There are several studies of properties, relationships and empirical selection of explanatory variables, see for example Beaver (1966).

The explanatory variables considered here are found mainly in Bernhardsen (2001) and is a collection of financial ratios, an industry indicator, the number of auditor remarks and some first differences for the ratios. Through these first differences (the change from the previous year) we are able to utilize not only the most recent financial statement data of a firm, but also data from the previous year. The appropriate variables to use will depend on the data available, which will vary with region and industry. We remove variables that are not significant in any model and keep 13 variables and 10 first differences, i.e. 23 variables in total, summarized in Table 1.

All variables, except for the industry indicator, the number of auditor remarks and the first differences, are defined as their deviance from their industry mean. These industry means are trimmed for the most extreme 0.2% values. The variables will then reflect a firms risk compared to other firms within the same industry.

Table 1 summarizes variables that are included when developing models. For the variables marked with an asterix the first differences are also included.

Table 1: Explanatory variables employed and their definition. For the variables marked with an asterix the first differences are also investigated.

Variable	Definition
REVANM	No. of auditor remarks
AGE	Age of firm
EKA*	Equity share of total assets (solidity)
TKR*	Return on capital employed (profitability)
UBE*	Outstanding public dues to total assets
LEV*	Trade creditors to total assets
LIK*	Cash minus short term debt to revenue from operations (liquidity)
LDEB*	Consolidated long term liabilities to total assets
DIV*	Dividends to total assets
INDUSTRY	Which industry sector a firm belongs to
CurrentR*	Current assets to current liabilities (liquidity)
QuickR*	Current assets less inventory to current liabilities (liquidity)
RetAss*	Return on assets (profitability)

3.2 Industrial classification

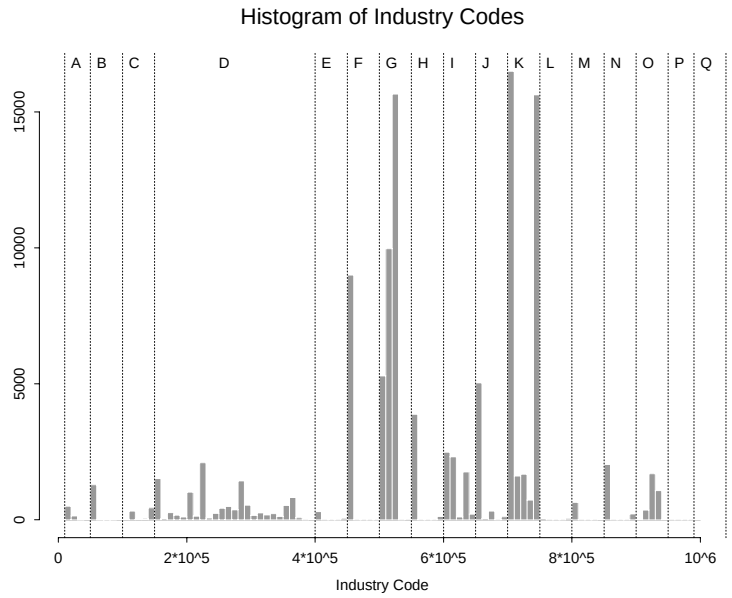
Benchmark values are not directly comparable over different industries (Bernhardsen, 2001). An attempt to overcome this problem is to include industry characteristics. We include an industry indicator.

Let us examine the composition of industries. We use Statistics Norway's standard for industrial classification, NACE, found at Statistics Norway's Internet pages, http://www.ssb.no. Table 2 summarizes the classification used, and Figure 2 displays a histogram of the industries represented in the financial statements of 1997. The letters in the figure refers to the classification in Table 2. We can see that our data set is dominated by firms from sector G and K. NACE divides each industry into several subindustries as well, which is why, in Figure 2, each sector include more than one bar.

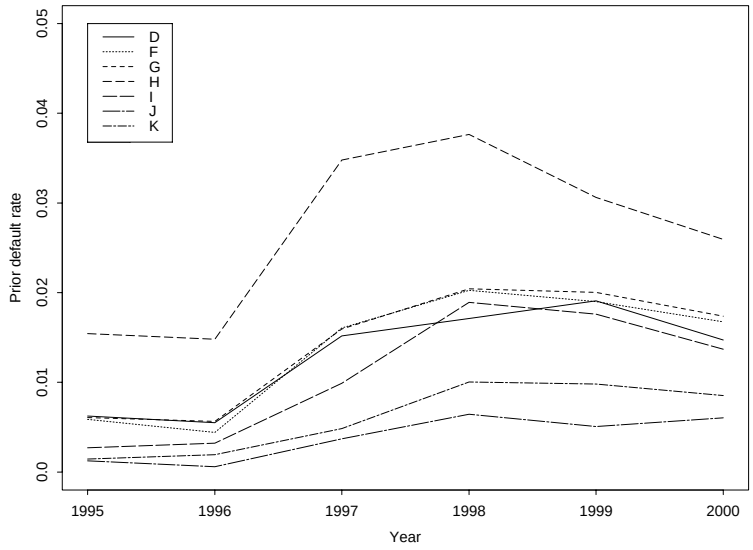
Figure 2 also display the default rates for the largest industries present in 1997, D, F, G, H, I, J and K in Table 2. We see that industry H, hotel and catering activity has a much higher default rate than other industries. We also note that industry J, financial services and insurance, has a very low default rate, as does industry K, property operations, commercial services and rental business. The industry indicator variable is thus expected to be very significant and indeed it proves to be.

4 Methodology

This section briefly describes the methodology for model development, model validation, the means for comparison between models, and the software we use for our analysis. The



(a) Histogram showing the composition of industries in 1997.



(b) Prior default rates for the largest industries.

Figure 2: Histogram of industry composition and prior default rates per industry.

Table 2: Industrial classification from Statistics Norway.

Sector	Description
A	Forestry and agriculture
B	Fishing
C	Mining and extraction
D	Industry
E	Water and power supply
F	Building and construction activity
G	Commodity trade, vehicle and domestic appliance repair
H	Hotel and catering activity
I	Transport and communication
J	Financial services and insurance
K	Property operations, commercial services and rental business
L	Public administration
M	Education
N	Health and social services
O	Other social and personal services
P	Salaried household work
Q	International organs and organizations

framework discussed here is applied when comparing the models in Section 6.

4.1 Model Development Framework

When developing models we include all the explanatory variables summarized in Table 1. We do not exclude variables that are not significant, or variables that are highly correlated with other variables. When developing a model for use in practice, a stepwise procedure should be applied where only explanatory variables that add significant predictive power is included in the model. Since we develop and test so many models such a stepwise procedure is too time-consuming.

The inclusion of highly correlated explanatory variables may cause problems in practice, but only if the individual effect of an explanatory variable is interpreted. When including highly correlated variables such interpretations should be avoided, due to the phenomena *multicollinearity*. By interpreting one explanatory variable's effect, separate from all other variables, there is a good chance of being mislead since the correlation structure is not easily interpreted. However, if a model is constructed solely for the purpose of prediction, not interpretation of each explanatory variable, then the multicollinearity will not be of concern.

When developing models we generally use 60% of the data set, randomly selected from the full data set. This *training* set is the same for all models. The remaining 40% will henceforth be referred to as the out-of-sample test set.

4.2 Validation Framework

The performance statistics of credit risk models can be highly sensitive to the data sample used for validation. To avoid embedding unwanted sample dependency, quantitative models should be validated on observations of firms that are not included in the sample used to build the model. This is referred to, by Sobehart et al. (2000), as *out-of-sample* validation.

Consider now a practical example. Suppose today is 1999. We are interested in building a 2 year default horizon model to use on the 1998 financial statements, predicting the probability that a firm will fail during $1999 - 2000$. We would then develop our model from 1996 data using a two year default horizon. But we would not be interested in how good this model performs on 1996 data, we would be interested in how good the model will perform on a forward going basis, that is how well will the model perform on the 1998 data, predicting default two years ahead from today. To compare models in this perspective, we test our models, developed from 1996 data, on 1998 data, to see which model best predicts bankruptcies in $1999 - 2000$. This is referred to as *out-of-time* validation and is the measure most interesting for practitioners. We investigate both out-of-sample and out-of-time validation.

To compare models we graphically look at so-called *power curves*, indicating the predictive performance of the various models. We also consider one of the metrics proposed in Sobehart et al. (2000), namely the *Accuracy Ratio (AR)*. We will explain both in this section. To be able to say whether or not a model performs significantly better than another,

we draw several small test sets randomly from the full test set. For each of the sampled test sets we calculate the accuracy ratio. We then perform a simple *t-test* to determine whether or not a model performs significantly better than another.

4.2.1 Power Curves

Power curves display the trade-off between Type I and Type II error for all possible values of the measure of interest. Type I and Type II errors are the errors of misclassifying a bankrupt firm as healthy and misclassifying a healthy firm as bankrupt, respectively. In statistical terms, power curves represent the cumulative probability distribution of default events for different default probabilities (Sobehart et al., 2000).

These and similar curves have many different names. Sobehart et al. (2000) call them Cumulative Accuracy Profiles (CAP) plots, while Hand and Henley (1997) call them Lorentz Diagrams. They are also referred to as ROC-curves, Sensitivity-Specificity Curves, Lift-Curves, Dubbed-Curves, Receiver-Operator Curves, etc.

Figure 3 displays a power curve. The solid line shows the performance of the model being evaluated. The way to interpret it is as follows. On the horizontal axis (% of population excluded) we have the probability of misclassifying a healthy firm as bankrupt (Type II error). If this probability is 1 it means we are classifying all firms as bankrupt and thus excluding 100% of the total population and our portfolio is empty. But at the same time we have also excluded 100% of the bankrupt firms, indicated on the vertical axis, and we find ourselves in the top right corner of Figure 3. On the vertical axis we have the probability of correctly classifying a bankrupt firm as bankrupt (1 - Type I error). If we, on the other hand, classify all firms as healthy, we will include 100% of the population, or exclude 0% of the population. At the same time, we are excluding 0% of the bankrupt firms and we now find ourselves in the bottom left corner of Figure 3. Other ways to interpret the curves is that they show us the proportions of good risks that are accepted (vertical axis) plotted against the proportion of bad risks that are accepted (horizontal axis). They also show the proportion of defaults identified/excluded (vertical axis) plotted against the cut-off point (horizontal axis).

A perfect model would follow the left vertical and top horizontal axes, accepting 100% of the good risks before accepting any of the bad risks, indicated by the dashed line in Figure 3. The naive case of classifying randomly at all thresholds would follow the diagonal line, indicated by the dotted line in Figure 3. A model that follows the bottom horizontal axis and the right vertical axis is only an inverted perfect model.

These curves allow us to assess a model at various potential cut-off points, showing important information of the performance at the desired level of risk and at which levels the model is more vulnerable. Some models may perform better than others in one interval but worse in another. A model may also perform overall worse than others but better at some specific cut-off level. If this is the risk level we desire, this will be the better model for us, even though there are models that perform better, over the entire range of risks.

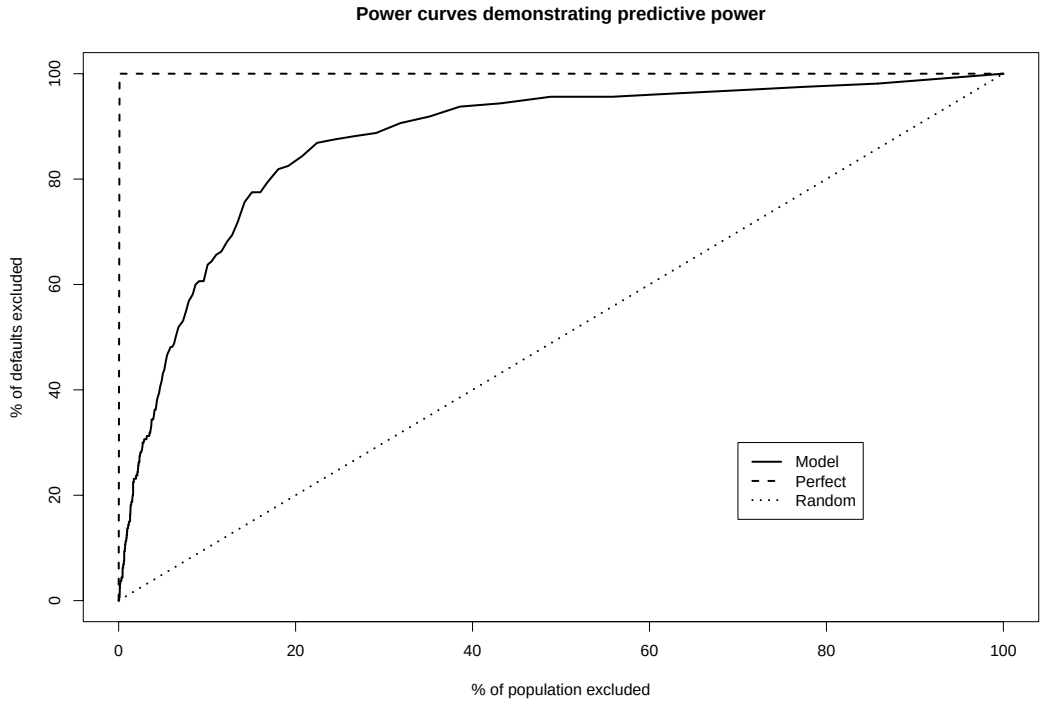


Figure 3: Power curves indicating discriminating power. The full line indicates the model under investigation, the dotted line the naive case of random classification and the dashed line a perfect model.

4.2.2 Accuracy Ratios

While power curves are convenient for visualizing model performance, it is often desirable to have a single measure that summarizes the predictive accuracy of each risk measure for both Type I and Type II errors into a single statistic. We employ one of the metrics proposed in Sobehart et al. (2000), namely the *Accuracy Ratio (AR)*. This metric is obtained by comparing the power curve of a model with that of the perfect model. The closer the power curve is to the perfect power curve, the better the model performs. To calculate the summary statistic we focus on the area that lies *above* the power curve of a random model (the 45° line) and *below* the power curve of the model under investigation, indicated by A in Figure 4. The more area below the curve and above the 45° line, the better the model is doing overall. The maximum area that can be enclosed above the 45° line is identified by the perfect curve, indicated by B in Figure 4. This maximum area is equal to 0.5.

Now, the ratio, A/B of the area between the models curve and the 45° line, A, to the area between the perfect curve and the 45° line, B, summarizes the predictive power over the entire range of possible risk values. This measure is referred to as the Accuracy Ratio (AR) (Sobehart et al., 2000), which is a fraction between 0 and 1. Models with ARs close to 0 display little advantage over a random model while those with ARs near 1 display

almost perfect predictive power. Figure 4 displays the accuracy ratio as the ratio of the shaded region in the graph on the left to the shaded region in the graph on the right, shown in the bottom graph.

In a loose sense, AR is similar to the commonly used *Kolmogorov-Smirnov (KS)* test designed to determine if a model is better than a random assignment of credit quality (Sobehart et al., 2000). However, AR is a global measure of the discrepancy between the power curves while the KS test focuses on the maximum discrepancy. Since the KS focuses only on a single maximum gap, it can be misleading in cases where two models behave quite differently for varying levels of risk.

Finally we mention that likelihood measures, for example the *Deviance*, will give us the same global measure of discrepancy as AR. We will however stick to ARs.

4.2.3 Resampling Scheme

When comparing models we employ a resampling scheme where several subsets are resampled, randomly, from the full validation set. For each of these subsets the AR is calculated and a simple t-test is performed to determine if a model performs significantly better than another, with a certain confidence level. We test the null hypothesis that the difference is not significant:

$$\begin{aligned} H_0 \quad &: \quad \overline{\mathrm{AR}_1} - \overline{\mathrm{AR}_2} = 0 \\ &\text{vs.} \\ H_1 \quad &: \quad \overline{\mathrm{AR}_1} - \overline{\mathrm{AR}_2} > 0, \end{aligned}$$

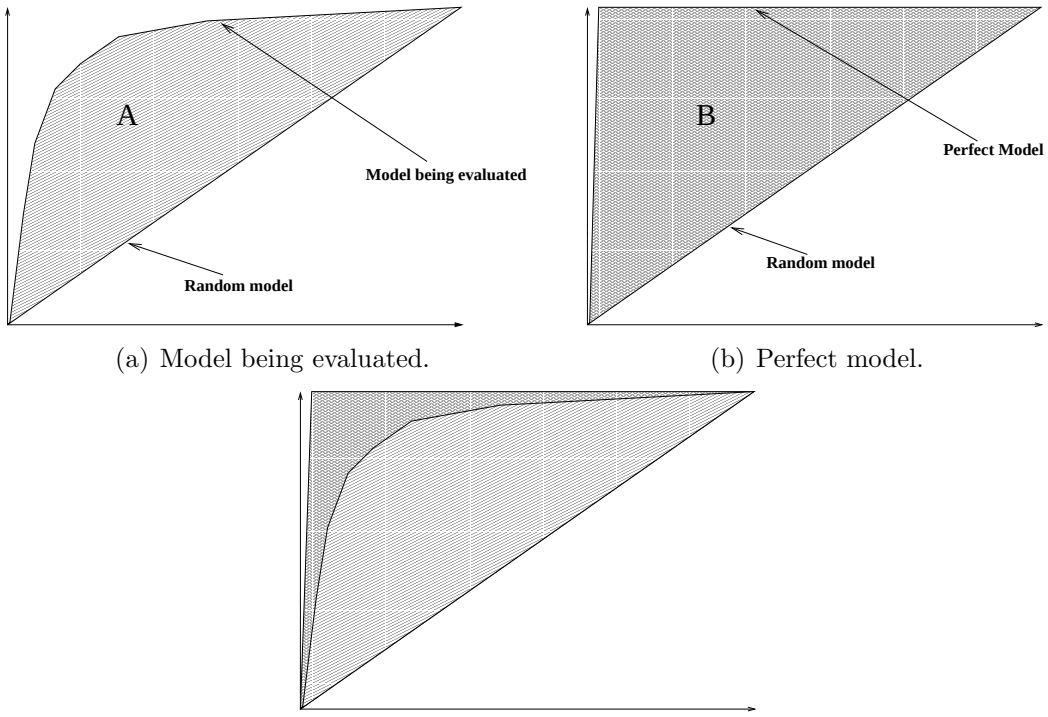
where $\overline{AR_1}$ and s_1^2 is the sample mean and variance of the ARs, respectively. To test this hypothesis we calculate the test statistic t^*:

$$t^* = \frac{\overline{\mathrm{AR}_1} - \overline{\mathrm{AR}_2}}{\sqrt{\frac{1}{n}(s_1^2 + s_2^2)}} \sim t_{df,\alpha} \ , \tag{12}$$

where *df* denotes the degrees of freedom for the Student-t distributed variable and α denotes the level of confidence. When validating models, we resample 100 subsets, each consisting of 5000 firms, hence we have 99 degrees of freedom for the Student-t distributed variable. We use a 99.5% confidence level. For a proper treatment of hypothesis testing and the Student-t distribution, see Walpole et al. (1998).

4.3 Software

R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment and can be considered as a different implementation of S. There are some important differences, but much code written for S runs unaltered under R.



(a) Model being evaluated.

(b) Perfect model.

(c) AR=A/B.

Figure 4: The accuracy ratio is the ratio of (A), the performance improvement over the naive random model of the model being evaluated to (B), the performance improvement over the naive random model of the perfect model. It is envisioned, in the bottom graph, as the ratio of the shaded region in the top left graph to the shaded region in the top right graph.

R provides a wide variety of statistical (linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, ...) and graphical techniques, and is highly extensible. The S language is often the vehicle of choice for research in statistical methodology, and R provides an Open Source route to participation in that activity.

One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed. Great care has been taken over the defaults for the minor design choices in graphics, but the user retains full control.

R is available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form. It compiles and runs out of the box on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux). It also compiles and runs on Windows 9x/ME/NT/2000/XP and MacOS.

When estimating and testing our models we use the R functions lda, qda, glm, gam and nnet for linear discriminant analysis, quadratic discriminant analysis, generalized linear models, generalized additive models and single-hidden-layer neural networks, respectively. The lda and qda functions are available in the MASS library, the glm and gam in the mgcv library and finally nnet is available in the nnet library.

5 Preliminary Study

This section presents some preliminary results from examining the various regression models separately. We compare the linear discriminant analysis (LDA) with the quadratic discriminant analysis (QDA). We then compare the generalized linear model with a logit link (GLM-Logit) with the generalized linear model with a probit link (GLM-Probit). Finally, we compare the generalized additive model with a logit link (GAM-Logit) with the generalized additive model with a probit link (GAM-Probit). The models are developed from 1996 data using a 2 year default horizon and tested on 1996 out-of-sample data. Table 3 shows the results. We see that the LDA significantly outperform the QDA. There is no significant difference in the performance of the Logit and Probit links neither for the GLM or GAM models. In the rest of this paper we will thus only consider the LDA, GLM-Logit and GAM-Logit models, along with single-hidden-layer neural networks.

6 Model Comparison

In this section we present the results from a comparison of two year default horizon models. Linear discriminant analysis (LDA), generalized linear models with a logit link function (GLM-Logit), generalized additive models with a logit link function (GAM-Logit) and single-hidden-layer neural networks (NN) are compared.

For the neural network models we use a single-hidden-layer network with a weight decay of 0.01, in line with suggestions in Aas et al. (1999). We use an accuracy ratio maximizing function to determine the optimal network size for each model. The network

Table 3: Accuracy Ratio means and standard deviations for various default prediction models. 1996 data, two year default horizon. Out-of-sample validation. The significance indicator states whether a model performs significantly better than its alternative (true or false). 99.5% confidence level.

Model	AR Mean	AR Std	Signif.
LDA	0.713	0.03	T
QDA	0.626	0.04	F
GLM-Logit	0.720	0.04	F
GLM-Probit	0.728	0.04	F
GAM-Logit	0.773	0.04	F
GAM-Probit	0.772	0.03	F

size corresponds to the number of nodes in the hidden layer, M in Equation (7). The output function $f_o(x)$ is chosen to be logistic and the parameters of the network are estimated using entropy fitting, explained in Section 2.4. We do not allow skip-layer connections.

For validation and comparison we follow the methodology outlined in Section 4.2. We assess power curves, obtained by testing the models on the full test data set, and accuracy ratios (ARs), obtained through resampling.

6.1 Out-of-sample Validation

First we perform out-of-sample validation. We develop one model from the 1996 training data set and test this model on the 1996 out-of-sample test set. We then develop one model from the 1997 training data set and test this model on the 1997 out-of-sample test set, and correspondingly for the 1998, 1999 and 2000 data sets.

The results for the 1996 model is displayed in Figure 5, showing the power curves of each model. The LDA, GLM and NN models seem to perform equally well, while the GAM model seems to outperform the others.

To confirm this visual impression we look at the sampled AR means, displayed in Table 4. We see that all models have approximately the same standard deviation and that the GAM model has a higher mean than the other models. Table 5 shows whether a model performs significantly better than than the other models above it in the table, going from the uppermost model in the table on the left to the model directly above on the right. The table includes results for the 1996, 1997, 1998 and the 1999 models. All tests of significance use a confidence level of 99.5%. For 1996 our visual impression from the power curves is confirmed. There is no significant difference between LDA, GLM and NN while GAM significantly outperforms the others. For 1997 the only difference from 1996 is that the GLM and NN models significantly outperform the LDA. For 1998 and 1999 the GLM does not perform significantly better than the LDA, but now the NN performs significantly better than the GLM. *For all years the GAM model, with a confidence level of* 99.5%,

performs significantly better than all other models tested.

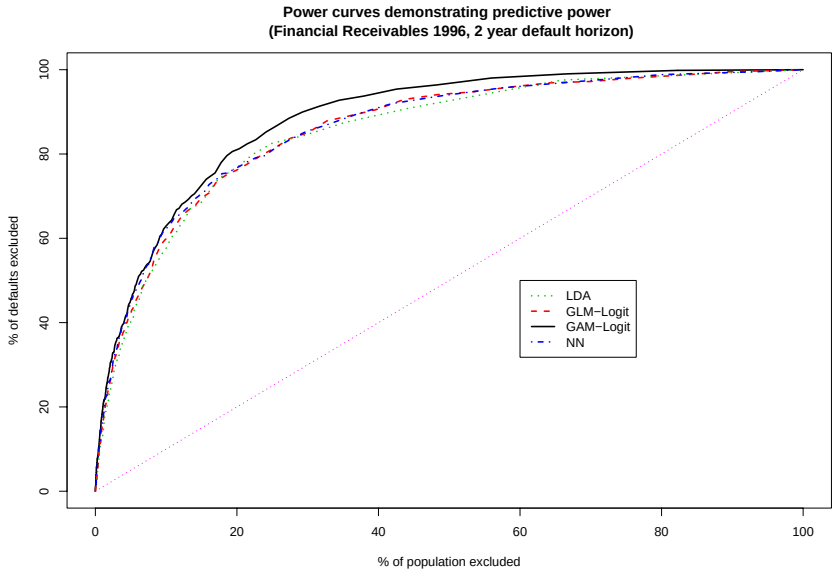


Figure 5: Prediction power of LDA, NN, GLM-Logit and GAM-Logit. 1996 data, two year default horizon, **out-of-sample** validation.

Table 4: Accuracy Ratio means and standard deviations for various default prediction models. 1996 data, two year default horizon, **out-of-sample** validation

Model	AR Mean	AR Std
LDA	0.713	0.03
GLM-Logit	0.720	0.04
NN	0.723	0.05
GAM-Logit	0.773	0.04

6.2 Out-of-time Validation

The results from the out-of-sample validation are interesting but not exactly what we are interested in. If we develop a model on 1996 data, with a default horizon of 2 years, we are interested in its performance on 1998 data. So we test the same 1996 model, but this time on 1998 data. Hence we are performing out-of-time validation on the 1998 data.

The resulting power curves are displayed in Figure 6, and the corresponding AR means, AR standard deviations and significance indicators are displayed in Table 6. The models in Table 6 are ordered by their AR means and the significance indicator tells us if a model

Table 5: Significance indicators stating whether or not a model for default prediction performed significantly better than the models above it in the table. The combination 'TF' indicates that a model did and did not perform significantly better than the uppermost model in the table and the model directly above it in the table, respectively. Two year default horizon, **out-of-sample** validation, 99.5% confidence level.

Model	1996	1997	1998	1999
LDA	-	-	-	-
GLM-Logit	F	T	F	F
NN	FF	TF	TT	TT
GAM-Logit	TTT	TTT	TTT	TTT

performs significantly better than the models above it in the table. *We see that the GAM model still significantly outperforms all the other models.* The LDA and GLM do not differ significantly, with a confidence level of 99.5%. At high risk levels we see that the NN model seems to perform as good as the GAM-Logit model, but as we move towards lower risk levels the GAM-Logit models outperforms the other models. This nicely demonstrates the importance of examining the models at the appropriate levels of risk, typically determined by the risk manager or the management group. Notice that the GAM model seems to perform best at all levels of risk.

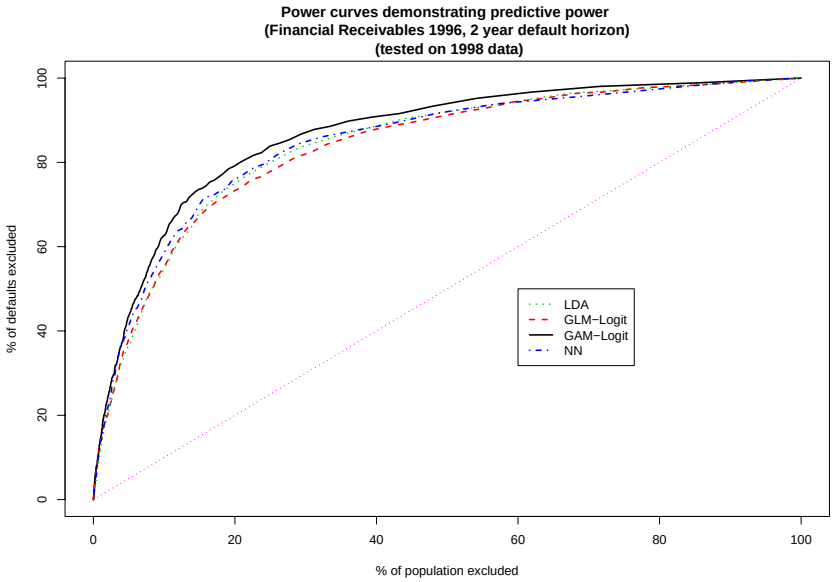


Figure 6: Prediction power of LDA, NN, GLM-Logit and GAM-Logit. 1996 data, two year default horizon, **out-of-time** validation on 1998 data.

Table 6: Accuracy Ratio means and standard deviations for various default prediction models. The significance indicator states whether or not a model is significantly better than the ones above it in the table. 1996 data, two-year default horizon, out-of-time validation on 1998 data, 99.5% confidence level.

Model	AR Mean	AR Std	Signif.
GLM-Logit	0.676	0.04	-
LDA	0.678	0.04	F
NN	0.695	0.04	TT
GAM-Logit	0.726	0.04	TTT

6.3 Non-linear effects

Let us examine the shape of some non-linear predictor effects estimated by the generalized additive model and compare them with the shapes estimated by the generalized linear model.

Figures 7 and 8 show estimated non-linear effects[1] of some explanatory variables for a 2 year default horizon model developed on 1996 data. We see the clear non-linear shapes. The figures indicate how a change in the value of a variable will affect the probability of default, the risk. For example, for the variable DIV, in the top left corner of Figure 7, we see a clear quadratic shape where, from the left towards the right, an increase in DIV will decrease the risk at first, but then a further increase will increase the risk again.

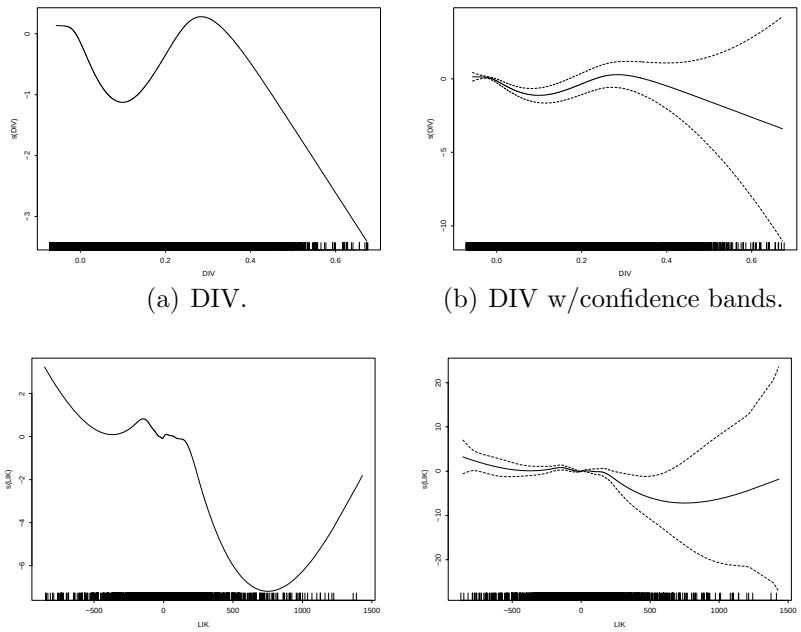
The vertical bars at the base of the plots indicate the intensity of observations in the various regions. In regions where the intensity of observations is high, we will see a high density of such bars. A high density of bars, or a high intensity of observations, means that our estimates have low uncertainties. A low density of bars means that there are few observations in this area and the uncertainty in the estimated shapes are thus very high. So if we look at DIV again we see that the region of the quadratic shape is estimated with low uncertainty since the density of bars in the bottom of the graph is very high while the decrease for large values of DIV is highly uncertain. This is also depicted in the graphs on the right of Figure 7, where confidence bands have been included. We see that the variance of the estimated shapes explode as the intensity of observations decreases. Note the scale on the vertical axis for the graphs on the right of Figure 7.

Despite the large uncertainty in some regions we see clear non-linearity in regions where we have low uncertainty. Some variables still seem to have nice, linear shapes in the regions of low uncertainty, for example UBE in Figure 8. For this variable we see the linear rise in risk as the value of UBE increases in the region of high observation intensity and low uncertainty. This result coincides with the estimate from the generalized linear model, depicted in the same figure. The quadratic shape in the right part of the plot is in a region of high uncertainty. Although the effects of each explanatory variable is not discussed

[1]These plots are obtained using the S-PLUS function plot.gam.

in this report we note that the risk increase as UBE increases makes sense since UBE is defined as outstanding public dues to total assets (see Table 1). For the variable RetAss however, the linear model is not able to detect the clear non-linear shape in a region of low uncertainty. This is a good example of why GAM will perform better than linear models.

Figure 9 visualizes the linear and non-linear effects on the risk of a firm, estimated by the GLM-Logit and GAM-Logit models, respectively. We see how the risk of a firm would change if two explanatory variables change simultaneously, holding all other variables constant. We have plotted the risk as a function of two of the most significant variables (DIV, RetAss) in the 1996, 2 year default horizon, model, and only for regions where the uncertainty is relatively low (compare with Figures 7 and 8).



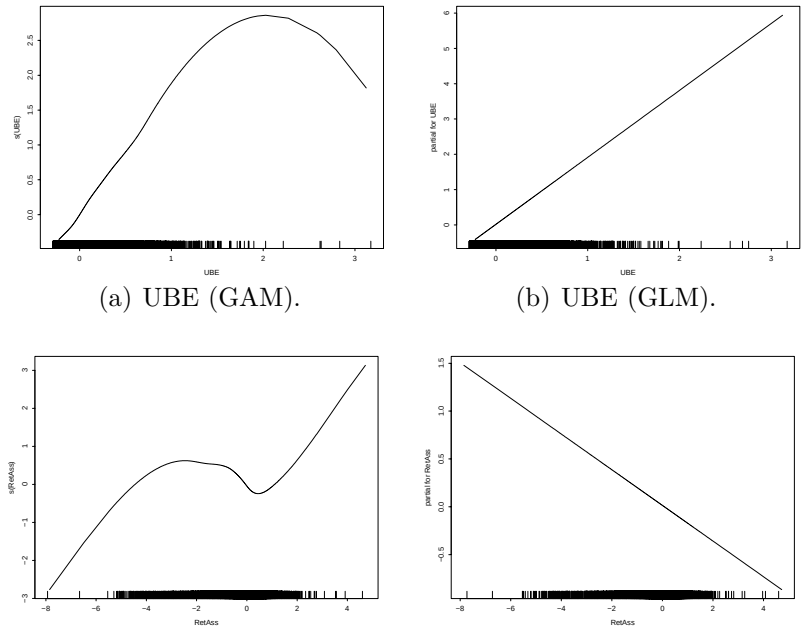
(a) DIV. (b) DIV w/confidence bands.

(c) LIK. (d) LIK w/confidence bands.

Figure 7: Estimated non-linear explanatory variable effects for the variables DIV and LIK. GAM model, 1996 data, two year default horizon.

7 Default Horizon

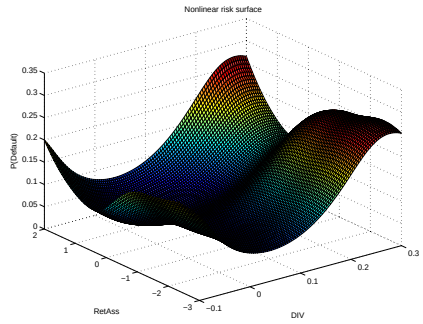
When entering credit derivative contracts the time period of the contract is specified. We then need models that are consistent with this time period. Say, for example, that this time period is specified to be two years. Then, for pricing purposes, we will be interested in the probability of a firm failing during the next two years. So we need a bankruptcy prediction model with a two year default horizon. Depending on the purpose of the model there are several reasons why the default horizon is important to consider. Say the intended



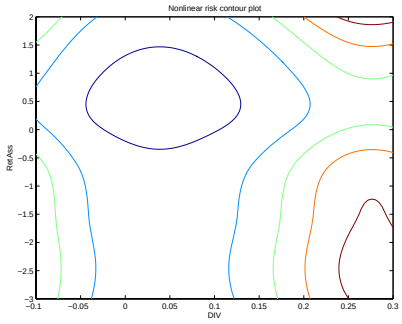
(a) UBE (GAM). (b) UBE (GLM).

(c) RetAss (GAM). (d) RetAss (GLM).

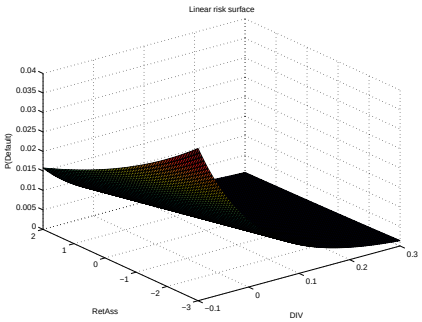
Figure 8: Estimated non-linear (GAM) and linear (GLM) shapes for the explanatory variables UBE and RetAss. Gam model, 1996 data, two year default horizon.



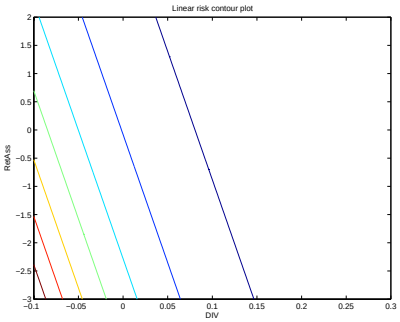
(a) Non-linear risk surface plot.



(b) Non-linear risk contour plot.



(c) Linear risk surface plot.



(d) Linear risk contour plot.

Figure 9: Linear and non-linear risk surface and contour plots. 1996 data, two year default horizon. The explanatory variables are DIV and RetAss. 1996 data, two year default horizon.

use is credit granting decision making. Then the highest prediction power is achieved by choosing a short default horizon, but this will 'hide' firms that are in distress but where this distress is not as urgent and severe as for those identified by the model. Alternatively, a long default horizon can be used to obtain early warnings of distress, enabling preventive actions. Perhaps the best solution is to continously use several models, each serving their own specific purpose.

We develop several GAM-Logit models on the same data set, but with varying default horizons. For the first model a firm is defined as default if it failed during the first year after model development, hence a 1 year default horizon. The second model defines a firm as default if it failed during the two first years, hence a 2 year default horizon, and so on. In order to test up to 5 year default horizon we perform out-of-sample validation. We expect different explanatory variables to prove significant for the various models since the variables that are strong in predicting short term distress in general will differ from the strongest variables for long term distress.

Figure 10 shows the predictive power of five GAM-Logit models, all developed on the 1996 data set and with $1-5$ year default horizon. We clearly see the performance depreciation as the default horizon increases. This is an expected, but nevertheless important result, and practitioners should keep this in mind when choosing a default horizon and assessing model risk. Table 7 displays the results from the resampling procedure, and we see that the performance depreciates significantly for each year added to the default horizon.

By looking at which explanatory variables prove most significant we find that the longer the default horizon the more variables proved significant. This is especially evident if we compare the 1 and 5 year default horizons. This indicates that signs of short term financial distress can be detected by looking at quite few variables. In general we can conclude that for longer default horizons the signs of distress are not so easily detected and much more complex interrelational structures are present. In such cases statistical models are crucial for detecting important information and insight regarding the riskiness of firms. We also note that for all models the strongest variables are the ones we expected would be dominating: number of accountant remarks, age, industry, outstanding public dues and trade credit.

8 Performance Depreciation

When developing a model for bankruptcy prediction the aim might for example be to keep this model and not develop a new model for some time. In this case it is very important to be aware of the depreciation rate of the model. If for example a bank wishes to exclude 80% of the defaults at all times the cut-off point needs to be adjusted as the model depreciates. This depreciation is also very important to consider if an estimate of model risk is to be attempted. These are some reasons to examine the depreciation of bankruptcy prediction models as time goes by.

Let us look at how the performance of a 1 year default horizon model depreciates as time goes by. When performing out-of-time validation in Section 6.2 this was basically

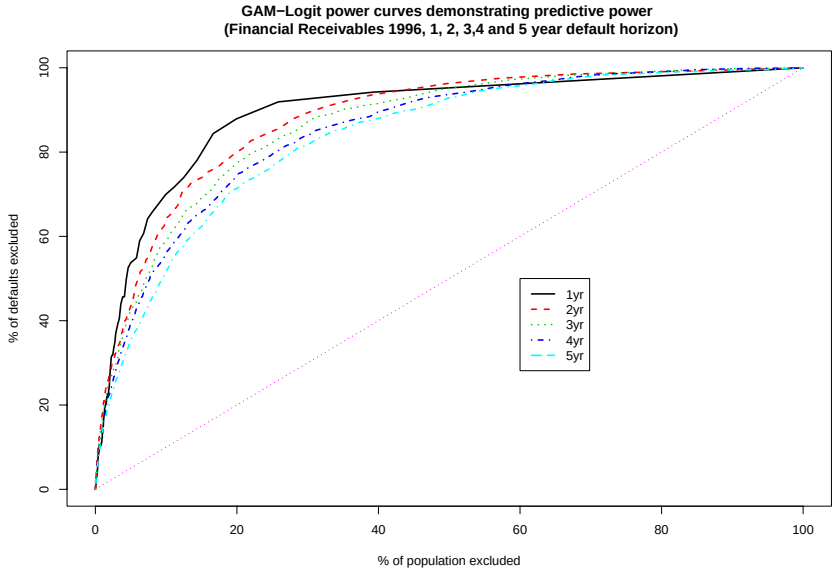


Figure 10: Power graph indicating discriminating power as default horizon varies. GAM-Logit models, 1996 data, **out-of-sample** validation.

Table 7: Accuracy Ratio means and standard deviations for GAM-Logit models. The significance indicator states whether or not a model performs significantly better than the models above it in the table. 1996 data, varying default horizons, **out-of-sample** validation, 99.5% confidence level.

Default Horizon	AR Mean	AR Std	Signif.
5yr	0.672	0.02	-
4yr	0.701	0.03	T
3yr	0.732	0.02	TT
2yr	0.760	0.04	TTT
1yr	0.784	0.07	TTTT

what we did. We built a 1 year default horizon GAM-Logit model on the 1996 training data and tested its performance on 1999 data, that is 3 years into the future. We now repeat this exercise and test the 1996 model on 1996, 1997, 1998, 1999 and 2000 data, that is $0-4$ years into the future. Figure 11 clearly shows the depreciation as time goes by. The predictive power 4 years into the future is much less than 0 and 1 years into the future.

Table 8 shows us the AR means and standard deviations of the models. We see that there is a big decrease in mean performance from 0 to 1 year ahead. We also see that there is no significant difference in performance on data 1 and 2 years ahead. From 2 to 3 and 4 years ahead we again see a significant decrease in performance.

Figure 11 shows us the power curves for the models tested. This figure adds important information compared to the numbers in Table 8. An interesting property seen is that the performance stays quite good, even 4 years into the future for low cut-off values, that is in the far left section of the power curves. In this section we are accepting a quite high level of risk. However, the figure shows that to maintain an exclusion of for example 80% of the defaults, the cut-off point will have to be increased drastically as time goes by. We also notice that the greatest deprecitation happens the first year. The model performs much better on contemporary data than on out-of-time data. This is a natural effect of overfitting. The depreciation from 3 to 4 years ahead is relatively small in comparison. The point discussed in Section 4.2, of models performing differently at different risk levels, is also nicely demonstrated. Consider the performance 1 and 2 years ahead, we see that the model seems to perform better 1 year ahead for high risk values, that is for approximately $0-18\%$ of the population excluded, while it performs better 2 years ahead for $> 18\%$ of the population excluded.

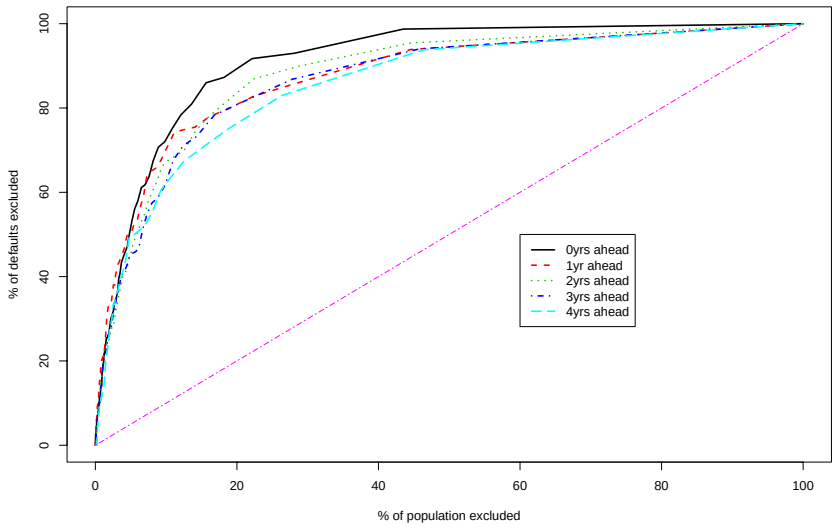


Figure 11: Predictive power depreciation $0-4$ years into the future, for a 1 year default horizon GAM-Logit model, developed from 1996 data.

Table 8: Accuracy Ratio means and standard deviations showing performance depreciation as time goes by. One year default horizon GAM-Logit model, developed from 1996 data. The significance indicator states if a model performs significantly better than the models above it in the table. 99.5% confidence level.

No. yrs into future	AR Mean	AR Std	Signif.
4	0.699	0.09	-
3	0.735	0.08	T
1	0.756	0.08	FT
2	0.770	0.07	FTT
0	0.824	0.06	TTTT

9 Multi-Year Model

We suspect that several actors in the market use only the most recent data when building bankruptcy prediction models. This is justified by the fact that the most recent data best reflect the characteristics of the data on which it will be used. But then the assumption is made that these characteristics change from year to year, and if this is true then the developed model will not be interesting anyway since it will only be applicable on contemporary data. So we must assume, unless we have good reason to believe otherwise, that the characteristics driving bankruptcy are constant. And if this is constant we should include as much data as possible when developing the model, since more data will give better estimates of default risk. Considering this and having seen the depreciation of models, we compare a one-year model built on 1998 data with a multi-year model built on data from three consecutive years, $1996 - 1998$. Both models are GAM-Logit with a one year default horizon. For the multi-year model we utilize the entire 1996 and 1997 data sets and the 1998 training set. This results in a training data set that is much larger than the data set used to develop the alternative one-year model (1998 training data only). Henceforth we will refer to the multi-year model and the one-year model as M^1_{96-98} and M^1_{98} to ease notation. The subscript denotes the years of data used to develop the model and the superscript denotes the default horizon.

There are several arguments to consider multi-year models, in addition to those already mentioned. We are able to utilize more data, giving our models a better basis for detecting signs of distress. The significance of variables in a multi-year model are less dependent on the macroeconomic conditions specific to one year. A model developed on one year of data only will build signs of distress, specific to that year only, into the model. A multi-year model on the other hand is expected to smooth out such year-specific effects. This way we would expect a multi-year model to be more robust and stable than a one-year model, making it interesting for practitioners, especially those who know there might be some years until a new model is developed.

Figure 12 shows power curves for 1 year default horizon models, developed from 1996, 1997, 1998, 1999 and 2000 training data and validated on 1996, 1997, 1998, 1999 and 2000 data, respectively. That is, we performed out-of-sample, contemporary validation each year. We see that the performance of such a standard model varies quite much from year to year. This justifies considering a multi-year model. We never know if next year will be a good or bad year for model development. By using several years of data we better guard ourselves against such yearly fluctuations.

Unfortunately we do not have a data set available that enables us to test the performance of this multi-year model more than 2 years into the future. However, Table 9, still shows us interesting results. We see that M^1_{96-98} is more robust than M^1_{98}, as expected. The AR for M^1_{98} falls quite low for the 1999 test data while the multi-year model performs well for all test sets. The resampling procedure shows that M^1_{96-98} performs significantly better than M^1_{98} on the 1998 and 1999 test data, with a confidence level of 99.5%. On the 2000 data there is no significant difference in performance.

Interesting future work will involve comparing the depreciation of such models further into the future and also with multi-year models spanning more than three years.

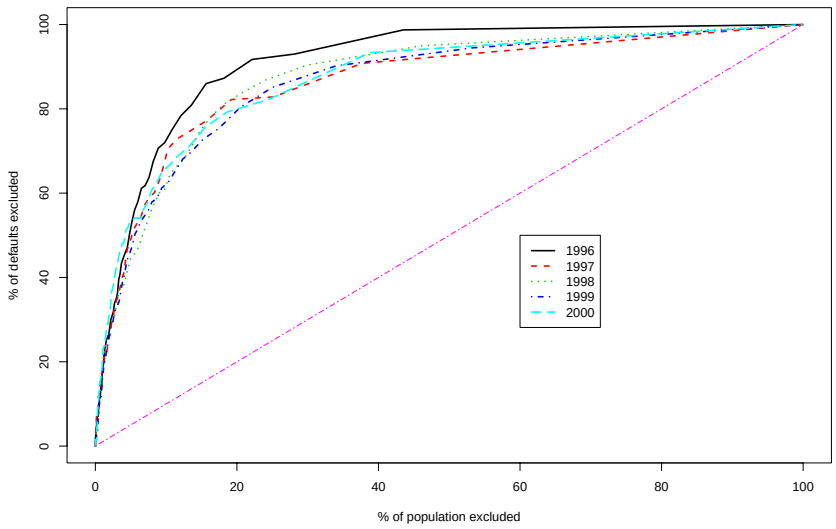


Figure 12: Power curves indicating predictive power of 1 year default horizon models developed on one year of data, out-of-sample validation.

10 Summary

We have shown, through out-of-sample and out-of-time validation, that generalized additive models significantly outperforms other methods like discriminant analysis, generalized linear models and neural networks.

Table 9: Accuracy Ratio means and standard deviations for a one-year model and a multi-year model, one year default horizon. Standard deviations in prentheses. The significance indicator states whether or not the M^1_{96-98} model performs significantly better than the M^1_{98} model, with a confidence level of 99.5%.

Test Data	M^1_{96-98}	M^1_{98}	Signif.
1998	0.780 (0.06)	0.752 (0.07)	T
1999	0.755 (0.07)	0.707 (0.09)	T
2000	0.759 (0.08)	0.747 (0.09)	F

If the IT system prevents the implementation of GAM models or the method is deemed non-intuitive and hard to justify to managers, an approximation called binning can be used. One can define dummy variables, a number of variables each representing an interval of the values of the original variable. For example $d_1 = \mathbf{1}_{\{DIV \leq 0\}}$, which means that d_1 will equal 1 if DIV is less than or equal to zero, and zero otherwise. Then simple linear- or ridge- regression is performed with all the dummies. This will be an approximation since it allows for non-linear effects. The advantage is that it is very easy to explain the effect and meaning of each variable and that once the dummies are defined all we need to do is apply simple linear regression on the dummies. The disadvantage is the process of manually defining the intervals for each dummy. This process is subjective and cumbersome. Also, the advantage of interpretation comes with a price, variables that are highly correlated must be excluded from the model to avoid multicollinearity problems.

We recommend further use of the out-of-time validation framework, employing resampling procedures, when presenting and comparing results from bankruptcy prediction models. The ability to say whether a model is significantly better than another, given a certain confidence level, is of uttermost importance and is best achieved by resampling. Power curves are visually highly informative and will be a valuable supplement to the accuracy ratios. In practice, one may only be interested in the performance for certain risk levels. In this case one can simply modify the AR-calculation to only consider the risk levels of interest.

Further we have shown, visually, the effect that the choice of default horizon has on models and their predictive power. This is important to consider when for example negotiating a credit derivative contract and for banks monitoring and actively managing their portfolios.

We also examined the depreciation rate of models, showing a large depreciation the first year that seemed to stabilize somewhat after the first year. The depreciation is very important to consider when deciding on desired level of risk and cut-off points, and also when estimating model risk.

Finally we compared a one-year model, estimated from 1998 data, with a multi-year model, estimated from three consecutive years of data, $1996 - 1998$. The multi-year model performed significantly better for 1998 out-of-sample validation and also for 1999 out-of-

time validation. For 2000 out-of-time validation, the two models did not differ significantly. The multi-year model seemed to be more robust, performing stably across the test data sets while the one-year model performed rather poor for 1999 out-of-time validation. The main reason for the multi-year model outperforming the one-year model is believed to be the size of the training data set, which is much larger for the multi-year model as it utilizes all data from 1996 and 1997, in addition to the 1998 training data. Unless there are good reasons to believe that the characteristics driving bankruptcies have changed, we argue that data from several years should be utilized.

Further work should involve further comparison of multi-year and one-year models. An analysis of the variation due to training and test set selection would also be interesting, for example through leave-one-out (or leave-several-out) cross validation. Also, other models could be implemented and tested on the same data sets, i.e. hazard models, the Z-Score and O-Score, time-series CUSUM models, etc. Finally, macroeconomic effects should be included in the multi-year models, preferably with several years of data.

Acknowledgements

This work is part of a Strategic Department Program called Statistical Analysis of Risk at the Department of Mathematics, University of Oslo and the Norwegian Computing Center. The author acknowledge the support and guidance of colleagues at the Norwegian Computing Center, in particular Assistant Research Director Kjersti Aas and Chief Research Scientist Xeni Kristine Dimakos. The author also thanks PhD student Sjur Westgaard at the Department of Industrial Economy and Technology Management, Norwegian University of Science and Technology, for providing the data set on which the entire work is based.

References

Aas, K., Huseby, R. B. and Thune, M. (1999), 'Data mining: A survey', Report, Norwegian Computing Centre. ISBN 82-539-0426-6.

Altman, E. I. (1968), 'Financial ratios, discriminant analysis and the prediction of corporate bankruptcy', *Journal of Finance* **23**, 589–609.

Altman, E. I. and Narayanan, P. (1997), 'An international survey of business failure classification models', *Financial Markets, Institutions and Instruments* **6**, 1–57.

Beaver, W. (1966), 'Financial ratios as predictors of failure. empirical research in accounting: Selected studies', *Journal of Accounting Research* **5**, 71–111.

Bernhardsen, E. (2001), A model of bankruptcy prediction, Technical report, Bank of Norway.

Doumpos, M. and Zopounidis, C. (1999), 'A multicriteria discrimination method for the prediction of financial distress: The case of Greece', *Multinational Finance Journal* **3**, 71–101.

Hand, D. and Henley, W. (1997), 'Statistical classification methods in consumer credit scoring: A review', *Journal of the Royal Statistical Society* **160**, 523–541.

Hastie, T. J. and Tibshirani, R. J. (1991), *Generalized Additive Models*, Chapman and Hall.

Kay, J. and Titterington, M., eds (2000), *Statistics and Neural Networks, Advances at the Interface*, Oxford University Press.

Krzanowski, W. J. (1998), *An Introduction to Statistical Modelling*, Arnold.

Ohlson, J. A. (1980), 'Financial ratios and the probabilistic prediction of bankruptcy', *Journal of Accounting Research* **18**, 109–131.

Ripley, B. D. (1996), *Pattern Recognition and Neural Networks*, Cambridge University Press, New York.

Sobehart, J. R., Keenan, S. and Stein, R. (2000), Rating methodology - benchmarking quantitative default risk models: A validation methodology, Technical report, Moody's Investors Service.

Walpole, R. E., Myers, R. H. and Myers, S. L. (1998), *Probability and Statistics*, Prentice Hall, Inc.

Wilson, R. and Sharda, R. (1994), 'Bankruptcy prediction using neural networks', *Decision Support Systems* **11**, 545–557.