Bankruptcy Prediction by Generalized Additive Models

Daniel Berg†

University of Oslo & Norwegian Computing Center

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.

Keywords: Bankruptcy Prediction, Generalized Additive Models, Default Horizon, Performance Depreciation, Multi-year model

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 continuous development of new financial derivatives. For example, the pricing of credit derivatives relies on good estimates of counterparty risk. Two kinds of models are commonly addressed in the literature. First, there are accounting based models, for example discriminant analysis and logistic regression models. Second, there are market based models, for example Merton models (e.g. the Moody's KMV 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. Consequently, 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 any exchange. Hence, our focus is on accounting based models.

Linear discriminant analysis models have been widely used. The popular Z-Score (Altman, 1968) is for example based on linear discriminant analysis. Generalized linear models, or multiple logistic regression models are also 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 popular alternatives, with the ability to incorporate a very large number of features in an adaptive nonlinear model, see for example Wilson and Sharda (1994). See also Altman and Narayanan (1997) for a survey of business failure classification models.

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

Our main objective 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 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 estimation of GAM and neural networks is more computationally demanding than for linear models, but with the rapidly increasing power of computers we expect an increasing application of such models in practice.

We develop several models using the same explanatory variables. To compare the models we use the validation methodology that is referred to as "out-of-sample" and "out-of-time" validation in Sobehart et al. (2000). The data set used is an extensive collection of annual financial statements of Norwegian limited liability firms in the period $1996 - 2000$ as well as the year of bankruptcy for all firms that filed for bankruptcy in the years $1996 - 2001$.

In addition to model comparison we examine the sensitivity of the GAM model to default horizon, and we test the depreciation of the prediction models, examining how the prediction power of a model depreciates as time goes by. 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 on statements from three consecutive years, is compared with a one-year model, developed on statements from one year only.

Section 2 describes the models we will examine. Section 3 describes the data set and the explanatory variables, while Section 4 discusses model development and validation methodologies. Section 5 compares the prediction power of various models, out-of-sample and out-of-time. Section 6 shows the sensitivity of a GAM model to default horizon, while Section 7 shows the depreciation rate of a GAM model. Section 8 compares the performance of a multi-year model and a one-year model and finally, Section 9 presents a summary of our findings.

2. Prediction Models

When handling bankruptcy data it is natural to label one of the categories as success (healthy) and the other as failure (default) and to score these as 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 sensibly applied. There are several such transformations, or link functions. We will only consider the logit link, $\varepsilon = \ln(\frac{\pi}{1-\pi})$, often denoted by $\varepsilon = \text{logit}(\pi)$.

2.1. Linear Discriminant Analysis

Linear discriminant analysis (LDA) is a multivariate statistical technique that leads to the development of a linear 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 LDA in real-world situations, since they are difficult to meet (Doumpos and Zopounidis, 1999).

2.2. Generalized Linear Models

Generalized Linear Models (GLM) is a generalization of the multiple regression model

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon, \tag{1}$$

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

$$f(y;\theta,\phi) = \exp\{[y\theta - b(\theta)]/a(\phi) + c(y,\phi)\}, \tag{2}$$

for some specific functions $a(\cdot), b(\cdot)$ and $c(\cdot)$, and parameters θ and ϕ. The GLM has the following features:

(a) The Y_i's $(i = 1, \dots, n)$ are independent random variables sharing the same form of distribution from the exponential family.
(b) The explanatory variables provide a set of linear predictors $\eta = \sum_j \beta_j X_j$ for $j = 1, \dots, p$.
(c) The link between 1 and 2 is that $g(\mu) = \eta$, where μ is the mean of Y. $g(\cdot)$ is called the link function of the model.

Two extensions of the multiple regression model (1) that characterize the GLM, 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. This link is determined by the distribution of the random term ϵ in (1). If ϵ is logistically distributed, we use the logit link and the GLM is referred to as the logit model. We will consider the binomial distribution and the logit link function $g(\mu) = \ln(\frac{\mu}{n-\mu})$. Re-expressing this function in terms of π instead of μ, we obtain $g(\pi) = \ln(\frac{\pi}{1-\pi})$ or $\pi = \frac{\exp(\eta)}{1+\exp(\eta)}$ (Krzanowski, 1998).

2.3. Generalized Additive Models

Hastie and Tibshirani (1986) proposed Generalized Additive Models (GAM). These models assume that the mean of the dependent variable depends on an additive predictor through a nonlinear link function. GAM extends the GLM by replacing the linear form $\sum_j X_j\beta_j$ with the additive form $\sum_j f_j(X_j)$. The linear regression step in GLM is replaced by a non-parametric additive regression step, where the data is used to determine the appropriate smooth function f. This is done through iterative smoothing operations and allows for various non-linear effects of the explanatory variables.

The logistic additive model, when applied to binary response data, takes the form $\ln(\frac{\pi}{1-\pi}) = \sum_j f_j(X_j)$ or $\pi = \frac{\exp(\sum_j f_j(X_j))}{1+\exp(\sum_j f_j(X_j))}$.

One of the main reasons for using GAM is that they do not involve strong assumptions about the relationship between two or more variables that is implicit in standard parametric regression. Such assumptions may force the fitted relationship away from its natural path at critical points. Also, since each of the individual additive terms are estimated using univariate smoothers, GAM avoids the problem of rapidly increasing variance for increasing dimensionality. This problem is referred to as the "curse of dimensionality" and is present in many nonparametric methods.

2.4. Feed-forward Neural Networks

We consider the supervised class of Neural Networks (NNs) called Multi-Layer Perceptrons. One hidden layer is used and no skip-layer connections are allowed. The probability of belonging to class k is then computed by

$$f_k(\mathbf{x}) = f_o\left(\alpha_k + \sum_{j=1}^{M} v_{jk} f_h(\beta_j + \sum_{i=1}^{N} w_{ij} x_i)\right), \tag{3}$$

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. The activation function, $f_h(x)$, of the hidden layer is always taken to be the logistic function $f_h(x) = \frac{\exp(x)}{1+\exp(x)}$, while the output activation function, $f_o(x)$, may either be logistic or linear (Aas et al., 1999). We use the logistic output activation function only.

A NN with no hidden layers is identical to the GLM, while a NN 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).

3. Data

Our data sets are extensive collections of annual financial statements for all limited liability firms registered at the Norwegian register for business enterprises over the years $1996-2000$. The 5 data sets all include a company identification number, explanatory variables examined and the year of bankruptcy. For firms that had not failed at the time of bankruptcy data extraction (2001), the year of bankruptcy was set to 0. When referring to, for example, a model developed from 1996 data with a 2 year default horizon, we mean a model developed from the 1996 financial statements, where a response variable Y is set to 1 if the year of bankruptcy was 1997 or 1998 and 0 otherwise.

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 bankruptcy prediction. Bankruptcy is a rare and extreme event. However, since we have such a large data set, 1% of $100,000$ firms is still $1,000$ firms, we are able to develop and validate models in a proper manner.

3.1. Explanatory variables

The choice of, and investigation of explanatory variables is not one of the objectives in this paper. There are several studies of properties, relationships and empirical selection of

Table 1. Explanatory variables employed and their definition. For the variables marked with an asterisk 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 credit 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)

explanatory variables, see for example Beaver (1966). The appropriate variables to use will vary with region and industry.

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 of the ratios. Through these first differences (the change since 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. In a preliminary analysis we removed variables that were not significant in any model. The remaining 13 variables and 10 first differences, i.e. 23 variables in total, are summarized in Table 3.1. First differences are included for variables marked with an asterisk.

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

4. Methodology

4.1. Model Development Framework

When developing models we include all the explanatory variables summarized in Table 3.1. In practice a stepwise procedure should be applied to only include explanatory variables that add significant predictive power to the model. Since we develop and test so many models such a stepwise procedure is too time-consuming.

We do not exclude variables that are highly correlated. The inclusion of highly correlated explanatory variables may cause problems in practice, but only if interpretations of the individual effects of the explanatory variables are attempted. When including highly correlated variables such interpretations should be avoided, due to the phenomena multicollinearity. However, if a model is constructed solely for the purpose of prediction, then multicollinearity will not be of concern.

When developing models we generally use 60% of the data set, randomly selected from the full data set and referred to as the training set. The remaining 40% is used for validation and is referred to as the out-of-sample test set.

4.2. Validation Framework

The performance statistics of 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 as out-of-sample validation (Sobehart et al., 2000).

If we develop a model from 1996 financial statements, using a two year default horizon, we are predicting probabilities that firms will fail during $1997 - 1998$. That means we can't build this model until 1999, when the 1998 data is available. The model can then be applied to 1998 financial statements, predicting default probabilities for $1999 - 2000$. But how good will the model perform on these 1998 data? Validating the model on 1998 data 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 consider so-called power curves, visually indicating the predictive performance of the various models. 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).

While power curves is a convenient way of 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 the model under investigation 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 A that lies above the power curve of a random model (the $45°$ line) and below the power curve of the model under investigation. The larger the area below the curve and above the $45°$ line, the better the model is doing overall. The maximum area B that can be enclosed above the $45°$ line is achieved by the perfect curve. This maximum area is equal to 0.5. The ratio, A/B is referred to as the Accuracy Ratio (AR). It summarizes the predictive power over the entire range of possible risk values and is a fraction between 0 and 1.

To compare models we employ a resampling scheme where several subsets are resampled, randomly, from the full test set. For each of these subsets the AR is calculated and a t-test is performed to determine if a model performs significantly better than another, at a certain confidence level. When validating models, we sample 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.

5. Model Comparison

We now present the results from a comparison of two year default horizon models. Linear discriminant analysis (LDA), generalized linear models (GLM), generalized additive models (GAM) and single-hidden-layer neural networks (NN) are compared. For the NN models we use a weight decay of 0.01. We use an accuracy ratio maximizing function to determine the optimal network size. The network size corresponds to the number of nodes in the hidden layer, M in Equation (3). The output function $f_o(x)$ is chosen to be logistic.

Table 2. 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

Table 3. Significance indicators stating whether or not a model performs significantly better than the models above. The combination 'TF' indicates that a model does and does not perform significantly better than the uppermost model 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

5.1. Out-of-sample Validation

We first perform out-of-sample validation. We develop one model from the 1996 training set and test this model on the 1996 out-of-sample test set. We then develop one model from the 1997 training set and test this model on the 1997 out-of-sample test set, and correspondingly for the 1998 and 1999 data sets. The results for the 1996 models are displayed in the left graph of Figure 1, 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 statistics, displayed in Table 5.1. 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 5.1 shows whether or not a model performs significantly better than the models above it in the table, from left to right, the uppermost model to the model directly above. The table includes the results of the $1996, 1997, 1998$ and 1999 models. For the 1996 models 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 the 1997 models 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.

5.2. Out-of-time Validation

The results from the out-of-sample validation are interesting, but not exactly what we are interested in. We seek the performance on future data, hence we perform out-of-time validation on the 1998 data. The resulting power curves are displayed in the right graph of Figure 1, and the corresponding AR statistics are displayed in Table 5.2. The significance indicator tells us whether or not a model performs significantly better than the models above it in the table. We see that GAM still significantly outperforms all the other models. LDA, GLM and NN do not differ significantly. At high risk levels it seems like the NN model performs almost as good as the GAM, but as we move towards lower risk levels the

Table 4. Accuracy Ratio means and standard deviations for various models. The significance indicator states whether or not a model is significantly better than the ones above. 1996 data, two year default horizon, **out-of-time** validation on 1998 data, 99.5% confidence level.

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

GAM model outperforms all the other models, including the NN. This nicely demonstrates the importance of examining the model at the appropriate levels of risk. Note that while it varies which model performs second best, the GAM model seems to perform best at all levels of risk.

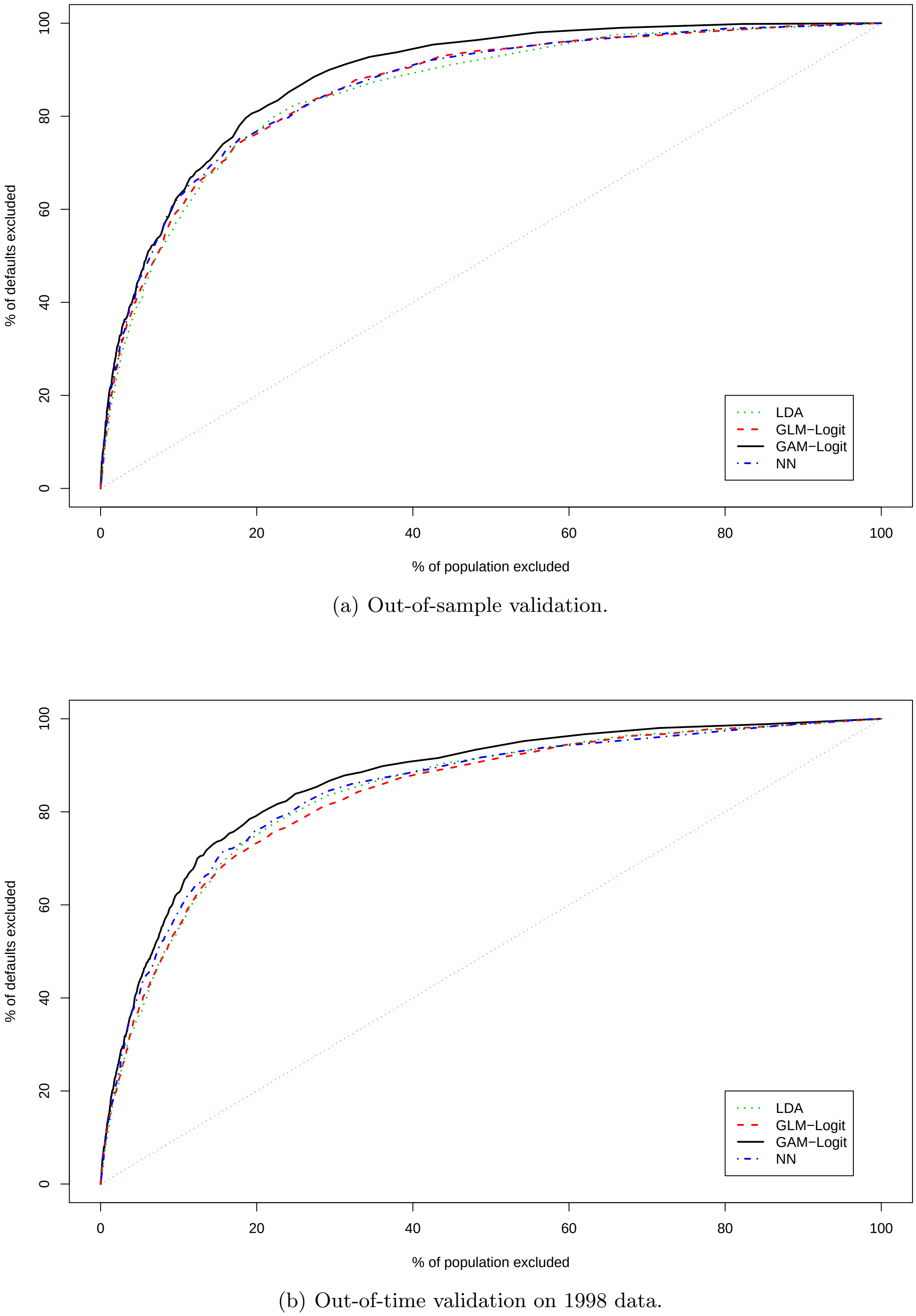
6. Default Horizon

The term default horizon refers to the time horizon for which the model tries to predict. A one year default horizon model will define firms that fail during the first year after model development as default, while a two year default horizon model will define firms that fail during the first two years as bankrupt.

There are several reasons why the default horizon is important to consider. When entering a credit derivative contract for example. The time period of the contract will determine the default horizon. Or say the intended 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 not as urgent and severe as those identified by the model. Alternatively, choosing a long default horizon will yield early warnings of distress, enabling preventive actions. Perhaps the best solution is to continuously use several models, each serving their own specific purpose.

We develop several GAM models on the same data set (1996), but with varying default horizons. In order to test a default horizon up to five years, we perform out-of-sample validation. Figure 2 shows the results and we clearly see that the performance is reduced as the default horizon is increased. 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 6 displays the results from the resampling procedure, and we see that the performance is reduced 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 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 good 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.



(a) Out-of-sample validation.

(b) Out-of-time validation on 1998 data.

Figure 1. Prediction power of LDA, GLM, NN and GAM models. 1996 data, two year default horizon, **out-of-sample** and **out-of-time** validation.

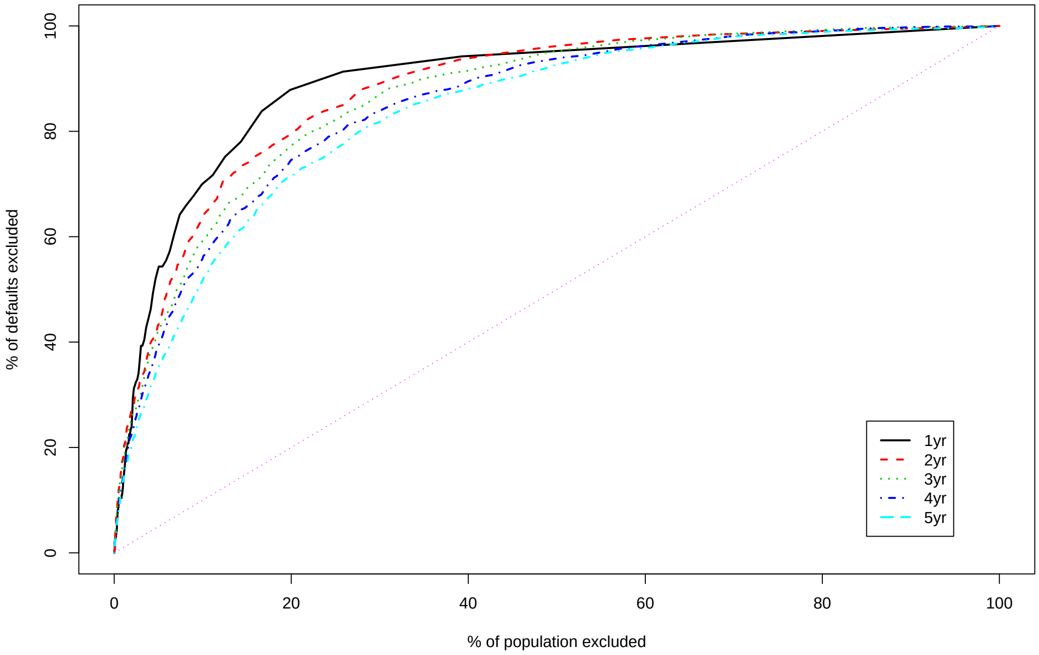


Figure 2. Power graph indicating discriminating power as default horizon varies. GAM models, 1996 data, **out-of-sample** validation.

Table 5. Accuracy Ratio means and standard deviations for GAM models as default horizon varies. 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.
5 years	0.672	0.02	-
4 years	0.701	0.03	T
3 years	0.732	0.02	TT
2 years	0.760	0.04	TTT
1 year	0.784	0.07	TTTT

Table 6. Accuracy Ratio means and standard deviations showing performance depreciation as time goes by. The significance indicator states whether or not a model performs significantly better than the models above. 1996 data, one year default horizon, 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

7. Performance Depreciation

When developing a model one might wish to keep this 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 we are to attempt to estimate the model risk. These are some reasons to examine the depreciation rate of bankruptcy prediction models as time goes by.

Let us look at how the performance of a one year default horizon model depreciates as time goes by. When performing out-of-time validation in Section 5.2 this was basically what we did. We built a two year default horizon GAM model on the 1996 data and tested its performance on 1998 data, that is 2 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. Table 7 shows us the AR statistics from the resampling procedure. 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 3 shows us the power curves for the models tested. This figure adds important information compared to the numbers in Table 7. An interesting property seen is that the performance stays quite good, even 4 years into the future for high risk levels. However, the figure shows that to maintain an exclusion of for example 80% of the defaults, the cut-off point will have to be drastically increased as time goes by. We also note that the greatest depreciation happens the first year. The model performs much better on out-of-sample 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 better than others at some risk levels but worse at other 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 when approximately $0-18\%$ of the population is excluded, while it performs better 2 years ahead if more than 18% of the population is excluded.

8. 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

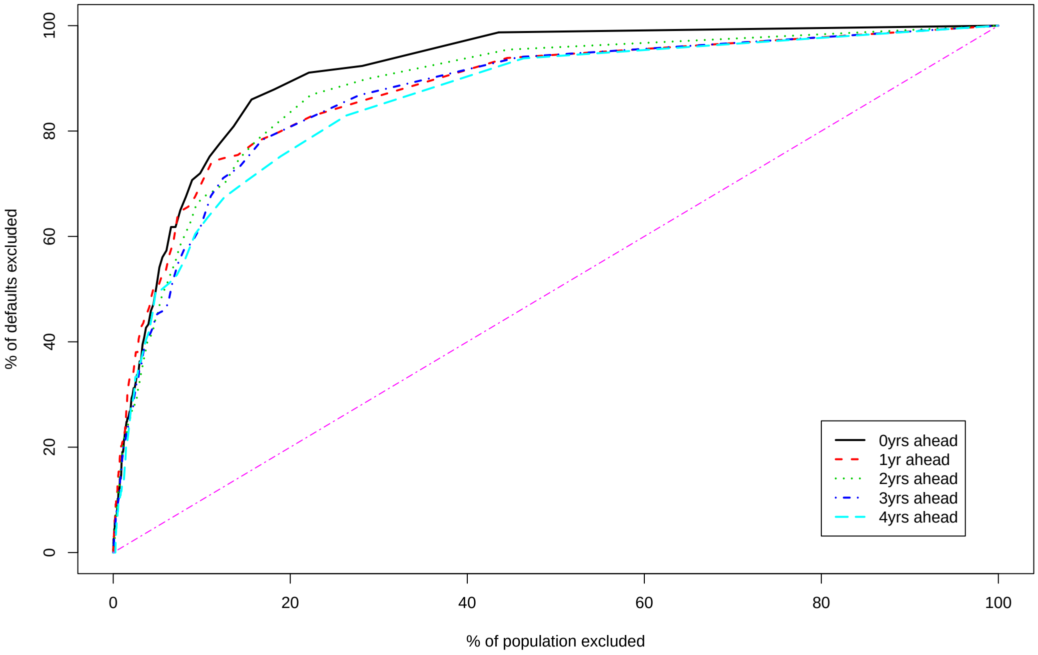


Figure 3. Predictive power depreciation $0 - 4$ years into the future for a GAM model. 1996 data, one year default horizon.

Table 7. Accuracy Ratio means and standard deviations for the M_{98} and M_{96-98} models. Standard deviations in parentheses. The significance indicator states whether or not the M_{96-98} model performs significantly better than the M_{98} model, with a confidence level of 99.5%. One year default horizon, **out-of-sample** and **out-of-time** validation.

Test Data	M_{96-98}	M_{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

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 as time goes by, we compare a one-year model with a multi-year model. The one-year model is built on 1998 data while the multi-year model is built on data from three consecutive years, $1996-1998$. Both models are one year default horizon GAM models. For the multi-year model we utilize much more data than for the one-year model. Henceforth we will refer to the multi-year model and the one-year model as M_{96-98} and M_{98} to ease notation. The subscript denotes the years of data used to develop the model.

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 is 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 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 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.

We developed several one year default horizon GAM models, one model for each year of data. Out-of-sample validation for each of these models shows that the out-of-sample performance 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.

We perform out-of-sample validation, testing the models on the 1998 data, and out-of-time validation on 1999 and 2000 data. Unfortunately we do not have data that enables us to test the performance of the multi-year model more than two years into the future. However, Table 8, still shows us interesting results. We see that M_{96-98} is more robust than M_{98}, as expected. The AR for M_{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_{96-98} performs significantly better than M_{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. The fact that the multi-year model outperforms the one-year model on the 1998 data is interesting. Apparently the $1996-1997$ data adds information about the 1998 out-of-sample test set, that the 1998 training set does not include.

9. Summary and discussion

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

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 can be used. One can define dummy variables, a number of variables each representing an interval of the values of the original variable, often referred to as binning. 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 a 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 or ridge regression on the dummies. The disadvantage is the process of defining the intervals for each dummy. This process can be subjective and cumbersome if not automated. 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. 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. Also, power curves adds valuable visual information about performance at different levels of risk. 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 how sensitive models are to the choice of default horizon. 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. This is very important to consider when deciding on desired levels of risk and cut-off points and also when estimating model risk. Figure 3 nicely demonstrates the need for cut-off adjustment as time goes by.

Finally we compared a one-year model, estimated from one year of data, with a multi-year model, estimated from three consecutive years of data. The multi-year model performed significantly better on out-of-sample test data and also on some out-of-time test data. The multi-year model seemed to be more robust, performing stable across the test data sets while the one-year model performed rather poor on one of the test data sets. The main reason for the multi-year model outperforming the one-year model is believed to be the size of the training data set. Unless there are good reasons to believe that the characteristics driving bankruptcies have changed, we argue that data from several years should be utilized.

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. We thank 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., R. Huseby, and M. Thune (1999, June). 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 P. Narayanan (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, Norges Bank.

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

Hastie, T. and R. Tibshirani (1986). Generalized additive models. *Statistical Science 1*, 297–318.

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.

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

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