Variable Selection in Data Mining: Building a Predictive Model for Bankruptcy

Author(s): Dean P. Foster and Robert A. Stine

Source: *Journal of the American Statistical Association*, Apr., 2004, Vol. 99, No. 466 (Apr., 2004), pp. 303-313

Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association

Stable URL: https://www.jstor.org/stable/27590387

Variable Selection in Data Mining: Building a Predictive Model for Bankruptcy

Dean P. FOSTER and Robert A. STINE

We predict the onset of personal bankruptcy using least squares regression. Although well publicized, only 2,244 bankruptcies occur in our dataset of 2.9 million months of credit-card activity. We use stepwise selection to find predictors of these from a mix of payment history, debt load, demographics, and their interactions. This combination of rare responses and over 67,000 possible predictors leads to a challenging modeling question: How does one separate coincidental from useful predictors? We show that three modifications turn stepwise regression into an effective methodology for predicting bankruptcy. Our version of stepwise regression (1) organizes calculations to accommodate interactions, (2) exploits modern decision theoretic criteria to choose predictors, and (3) conservatively estimates p-values to handle sparse data and a binary response. Omitting any one of these leads to poor performance. A final step in our procedure calibrates regression predictions. With these modifications, stepwise regression predicts bankruptcy as well as, if not better than, recently developed data-mining tools. When sorted, the largest 14,000 resulting predictions hold 1,000 of the 1,800 bankruptcies hidden in a validation sample of 2.3 million observations. If the cost of missing a bankruptcy is 200 times that of a false positive, our predictions incur less than 2/3 of the costs of classification errors produced by the tree-based classifier C4.5.

KEY WORDS: AIC; Bonferroni; C_p; Calibration; Hard thresholding; Risk inflation criterion (RIC); Step-down testing; Stepwise regression.

1. INTRODUCTION

Bankruptcy increases the cost of lending. Defaults take millions of dollars from the profits of creditors, who in turn pass costs back to consumers. With $500 billion in outstanding debt in the United States, credit-card portfolios have considerable exposure to default and lenders have cause for concern. During the economic expansion of the 1990s, default rates increased. Bankruptcy filings numbered 200,000 in 1978, but have climbed to an annual rate of about 1.5 million. Over 1% of households in the United States filed for bankruptcy in 1997, up nearly 75% from just a few years earlier (Gross and Souleles 2002).

Diverse explanations for increases in bankruptcy claims abound. Lenders have saturated the United States with applications for credit cards: About 75% of households in the United States now have at least one credit card. The expansive availability of easy access to credit has opened the market to more diverse, and perhaps more risky, borrowers. Another argument suggests societal attitudes toward bankruptcy have changed, with less stigma and legal penalties. Some states allow the borrower to declare bankruptcy while retaining substantial assets.

Rather than try to distinguish among such conjectures, our goal here is to predict the onset of personal bankruptcy. This problem is particularly challenging in many ways, and we devote much of our analysis to issues arising from trying to predict a rare event. Though increasingly common and expensive to creditors, personal bankruptcy remains rare, particularly when the timing of the event is considered. Our dataset holds longitudinal records for 244,000 active credit-card accounts. For the 12 months considered in this analysis (a subset of the time period described in Gross and Souleles 2002), about 1% of these accounts (2,244 accounts) defaulted. That amounts to 2,244 events in almost 3,000,000 months of activity.

Statistical models are not new to this area. To quantify risk, the credit industry makes extensive use of statistical modeling (Hand and Henley 1997; Hand, Blunt, Kelly, and Adams 2000; Thomas, Edelman, and Crook 2002). Creditors can turn to specialized companies such as Fair–Isaac for proprietary models that assess the riskiness of loan portfolios, and automated decision-support systems that incorporate statistical models are credited with saving millions of dollars (Curnow, Kochman, Meester, Sarkar, and Wilton 1997). These models are closely guarded corporate secrets.

In place of proprietary models, we use a fully automated stepwise regression. Our use of stepwise regression is expansive: We pick predictors from a pool of 67,160 variables. Although our model would surely benefit from more substantive expertise, our approach is fully automatic. Put bluntly, we build a stepwise regression from a large set of predictors expanded to include *all* pairwise interactions. These interactions are essential: A model with only two or three interactions predicts better out-of-sample than a model with the best 100 linear predictors. To select from so many predictors, we take great care to avoid selection bias—choosing predictors that look good in-sample but predict new data poorly.

It is well known that selection bias can overwhelm stepwise regression, leading to claims of a better fit than actually obtained (e.g., Rencher and Pun 1980; Freedman 1983). For example, in an orthogonal regression, criteria like the Akaike information criterion (AIC) or C_p choose any predictors that have an absolute t statistic that exceeds $\sqrt{2}$. If we imagine a "null model" in which no variable is useful, then the AIC would choose about 16% of them even though none is predictive (e.g., see Mallows 1973). Although overfitting is benign when choosing from 10 predictors, it becomes serious when choosing from 67,160.

To avoid overfitting, we combine a selection criterion developed in decision theory with a conservative estimate of significance. These results are easy to implement. They also run quickly because they avoid the need for computationally expensive methods like cross-validation to select predictors. We

Dean P. Foster is Associate Professor and Robert A. Stine is Professor (E-mail: *stine@wharton.upenn.edu*), Department of Statistics, The Wharton School of the University of Pennsylvania, Philadelphia, PA 19104-6340. The authors thank Nick Souleles of the Wharton Finance Department and The Wharton Financial Institutions Center for providing the raw data used in this analysis. The Financial Institutions Center also funded the authors during the initial development of this methodology. We also thank Mike Steele for introducing us to Bennett's inequality.

Journal of the American Statistical Association
June 2004, Vol. 99, No. 466, Applications and Case Studies
DOI 10.1198/016214504000000287

need only to change the method for computing standard errors and the "p-to-enter" parameter of stepwise regression. We follow these steps with a calibration that, in effect, estimates a link function. We consider two forms of calibration: one that imposes the logistic link and another that estimates the link nonparametrically.

To show that our methodology works, we use fivefold cross-validation. In each fold, we fit models to estimation samples with 600,000 cases and use these models to predict 2.3 million cases in the validation sample. In practice, we would model the full dataset, but to show that our approach works, we reserve a large validation sample. Ideally, we would compare our model to those used in the credit industry, but the combination of proprietary algorithms and issues of confidentiality make this impossible. So, we instead compare our results to those obtained by C4.5 and C5.0, which are classifiers developed in machine learning (see Quinlan 1993 and the associated commercial web page at *www.rulequest.com/see5-unix.html*). Classifier C5.0 is often called C4.5 plus boosting. Classifiers C4.5 and C5.0 build trees designed to minimize the costs of classification errors. We find that regression minimizes costs as well as, and generally better than, these classifiers, even though the classifiers take costs into account when building predictions (whereas regression does not). For cost ratios that imply that missing a bankruptcy is much more expensive than annoying a customer, stepwise regression produces a much smaller loss.

The lift chart in Figure 1 shows the ability of stepwise regression to predict bankruptcy. To motivate this figure, consider the challenge faced by a creditor who wants to target customers at risk of bankruptcy. The creditor knows the bankruptcy status of 600,000 accounts and wants to predict which of the 2.3 million accounts in another sample will default. The creditor plans to contact those most at risk of bankruptcy in hopes of changing their anticipated future behavior. Because of budget constraints, the creditor can afford to call, say, only 1,000 of the 2.3 million customers. If the creditor places calls at random, 1,000 calls on average contact fewer than one impending bankruptcy. Suppose instead of placing calls at random, the creditor sorts the accounts by the predictions from regression. Returning to Figure 1, the horizontal axis shows the proportion of the validation sample called and the vertical axis shows the proportion of bankruptcies found. The diagonal line in the plot shows the expected performance of random selection; the concave curve shows the lift provided by the calibrated regression model. The initial steep rise indicates its success in isolating the accounts with high chance of bankruptcy. If the creditor calls the 1,000 accounts with largest predictions, these calls reach 383 of the 1,786 bankruptcies scattered among the 2.3 million accounts—more than 20% of the accounts that go bankrupt. Contacting the top-scored 14,000 customers would identify more than 1,000 of the bankrupt accounts.

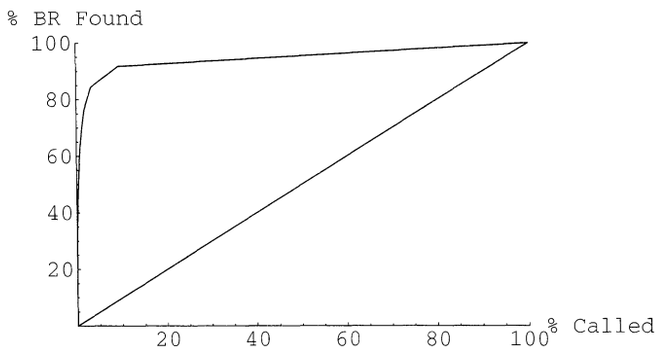

Figure 1. Lift Chart for the Regression Model That Uses 39 Predictors to Predict the Onset of Personal Bankruptcy. The chart shows the percentage of bankrupt customers in the validation data found when the validation observations are sorted by predicted scores. For example, the largest 1% of the predictions holds 60% of the bankruptcies. The diagonal line is the expected performance under a random sorting.

We have organized the remainder of this article as follows. Section 2 gives more details of our data. Section 3 explains our choice of least squares loss, and Section 4 elaborates on variable selection. Section 5 provides details needed to find estimates, standard errors, and p-values when selecting from so many possible predictors. Section 6 summarizes the performance of these models, including a comparison with C4.5 and C5.0. Section 7 offers a brief discussion of the role of automatic methodologies such as ours. Readers with an interest in trying these methods can find an implementation in SAS Enterprise Miner, a commercial collection of data-mining routines (Bastos, Wolfinger, Duling, and Auslender 2003).

2. DATA PREPARATION

The raw data from which we began present a longitudinal view of 280,000 credit-card accounts. These data blend accounts from several lenders, and care has been taken to merge common attributes (see Gross and Souleles 2002 for more details of the data). We obtained the data as part of a research project studying bankruptcy at the Wharton Financial Institutions Center. The time frequency of the data vary; some measurements are monthly (such as spending and payment history), whereas others are quarterly or annual. Other factors are fixed demographic characteristics, such as place of residence. Because of variations in the scope of data collection, we focus on a 12-month period during 1996–1997 that permits us to select predictors from a consistent set of variables. We number these months $t = 1, \ldots, 12$. We removed 35,906 accounts that appeared to be inactive or closed and continued with the remaining 244,094 accounts.

We next aligned the data as though collected at a single point in time. This alignment allows us to fit a common model with all 2,244 bankruptcies rather than different models in each month. Let Y_{it} denote the response for the ith account in month t and let $\mathbf{X}_{it}$ denote the corresponding predictors. We model the data as though the conditional mean of the response has finite memory

$$E(Y_{it}|\mathbf{X}_{i,t-1}, \mathbf{X}_{i,t-2}, \ldots) = E(Y_{it}|\mathbf{X}_{i,t-1}, \mathbf{X}_{i,t-2}, \ldots, \mathbf{X}_{i,t-\ell}) = \mu_{it}(\mathbf{X}_{i,t-1}, \mathbf{X}_{i,t-2}, \ldots, \mathbf{X}_{i,t-\ell}).$$

We set $\ell = 4$ because of a belief that bankruptcy events are sudden rather than long term; the data include four lags of all monthly predictors. (Other time horizons produced similar results.) We further assume stationarity and model the conditional mean as invariant over accounts and months during the studied year:

$$\mu_{it}(\mathbf{X}) = \mu(\mathbf{X}) \qquad \forall i, 1 \le t \le 12. \tag{1}$$

Monthly dummy variables allow the model to capture seasonal effects.

The assumption of stationarity allows us to pool the bankruptcy events to estimate the mean function. We treat the sequence of data for one account $Y_{i1}, \ldots, Y_{i12}$ as 12 uncorrelated responses. Similarly, we align all bankruptcies to the same point in time. Thus each nonbankrupt account contributes 12 observations to the final dataset, and each bankrupt account contributes at most 12. After alignment, the dataset of 244,094 accounts expands to 2,917,288 monthly responses Y_{it} and the accompanying lagged predictors. The analysis does not "know" that observations originate from one longitudinal sequence. Although this arrangement of the data simplifies the presentation of our methodology and allows us to pool all of the bankruptcy events, it does make it harder to predict bankruptcy. For example, consider an account that declares bankruptcy in month 7. It is an error to predict bankruptcy for any of the previous six observations even though these may presage the event.

Each observation at this stage consists of the 0/1 bankruptcy indicator and 255 predictors. Of these predictors, 72 originate as three lags of 24 other predictors. To capture trends or seasonal variation, we retain the month index t as a possible predictor. We treat the time trend as both a continuous predictor and as a collection of 12 seasonal dummy variables. We converted other categorical variables into dummy variables and we merged some of the indicators to reduce the number of predictors. For example, we heuristically grouped states into 10 geographic bins. Missing data in categorical variables simply define another category. We handle missing data in continuous predictors in a different manner as described next.

We treat any continuous variable with missing data as the interaction of an unobserved, complete variable with a "missingness" indicator. Jones (1996) offered a thorough analysis of this approach. For our data, 110 of the 255 time-aligned predictors have missing data. For each of these, we filled its missing values with the mean $\overline{X}_j$ of the observed cases and added an indicator, say B_j, to the dataset. This augmentation adds 110 more variables, giving a total of $110 + 255 = 365$ possible predictors.

The penultimate stage of preparations adds *all* second-order interactions to the set of possible predictors. These include interactions with missing data indicators and quadratics. The addition of interactions allows regression to incorporate second-order nonlinearity and subset differences, at a cost of a dramatic expansion in the number of candidate predictors. The addition of interactions expands the dataset by 365 squares and $\binom{365}{2} = 66{,}430$ interactions, for a total of 67,610 possible predictors. (Some of these predictors are redundant, such as the squares of binary indicators. Redundant terms are isolated and skipped in the calculations.) Because of the number of variables, this step is implicit; our code computes interactions as needed.

We treat interactions like any other predictor, violating the so-called principle of marginality (see, e.g., McCullagh and Nelder 1989 for a discussion of the use of interactions with and without lower order terms). The principle of marginality requires, for example, that a model that contains the interaction $X_j * X_k$ must also include both X_j and X_k. Our reasoning is simple: If the model benefits from having the base linear terms, then the selection procedure should find them. We also allow overlapping interactions of the form $X_j * X_{k_1}$ and $X_j * X_{k_2}$, unlike Gustafson (2000). In fact, our search for predictors of bankruptcy discovers many overlapping interactions.

Table 1. Number of Bankruptcies and Accounts in the Five Subsets That Define the Estimation and Validation Datasets Used in the Fivefold Reversed Cross-Validation

Subset	Bankrupt	Total accounts
$\mathcal{E}_1$	458	583,117
$\mathcal{E}_2$	442	584,028
$\mathcal{E}_3$	467	583,545
$\mathcal{E}_4$	431	583,262
$\mathcal{E}_5$	446	583,336
Total	2,244	2,917,288

NOTE: Each subset forms an estimation sample with the other four used for validation.

Finally, we randomly divided the time-aligned dataset into five subsets described in Table 1 for cross-validation. Each observation was assigned a random uniform value $U_i \overset{\text{iid}}{\sim} \text{U}[0, 1]$. Those observations for which $0 \le U_i \le .2$ went into the first subset, those with $.2 < U_i \le .4$ went into the second, and so forth. (As a result, the five splits do not have the same number of bankruptcies.) The results reported in the Introduction use the first of these five subsets for estimation (583,117 account months with 458 bankruptcies) and treat the other four as a validation sample (2,334,171 account months with 1,786 bankruptcies).

Our choice to reserve 80% for validation reflects our desire for an accurate assessment of the predictive ability of a model. Increasing the size of the estimation sample might lead to a better model, whereas decreasing the size of the validation sample makes it hard to recognize the benefit. Heuristically, our split implies that standard errors of the validation sums of squares are half those computed from the estimation samples. We did not shrink the estimation sample further because preliminary calculations indicated that the model fitting would deteriorate with smaller estimation samples; the number of bankruptcies becomes too small relative to the number of potential predictors. Subsequently, we denote the indices of observations in the jth estimation sample by $\mathcal{E}_j$ and those in the corresponding validation dataset by $\mathcal{V}_j$.

3. LOSS FUNCTIONS

Our goal for variable selection is to find the model with minimal Brier score. The Brier score of a model is the squared error of its predictions, summed over the validation data. Let $\widehat{Y}_{i,j}^{M(k)}$ denote the prediction of Y_i in $\mathcal{E}_j$ ($j \neq k$) given by fitting model M to estimation sample $\mathcal{E}_k$. The Brier score for M is

$$B(M) = \sum_{k=1}^{5} B(M, k),$$

$$B(M, k) = \sum_{j \neq k} \sum_{i=1}^{|\mathcal{E}_j|} \left(Y_{i,j} - \widehat{Y}_{i,j}^{M(k)}\right)^2. \tag{2}$$

The Brier score is one of many possible metrics for assessing a predictor; Hand (1997) discussed it and several others. Because we lack a known model for the probability of bankruptcy in

this application, the Brier score is a compromise that weights the contributions of all observations to the total error equally.

This equal weighting is an important consideration in modeling bankruptcy and differs from common practice. Typical concerns for statistical efficiency lead in a different direction, namely to the binomial loss associated with a logistic regression. Implemented as a weighted least squares, the fit of a logistic regression minimizes

$$W(\widehat{Y}) = \sum_i \frac{(Y_i - \widehat{Y}_i)^2}{\widehat{Y}_i(1-\widehat{Y}_i)}.$$

Because binomial variances are small near 0 and 1, this criterion places the most weight on the observations with estimated probabilities near these extremes. For bankruptcy, $\widehat{Y}_i \approx 0$ for essentially all of these highly weighted observations. Binomial weights put more emphasis on the vast majority of accounts that are unlikely to end in bankruptcy. Although binomial weights lead to efficient estimates in a model of *known* form, these weights need not be optimal for finding the right model. Hand and Vinciotti (2003) gave an illustration. Although we have no rigorous argument, common sense suggests placing more, not less, weight on the few bankruptcies when trying to find a model to explain these rare occurrences.

Economic considerations also motivate more balanced weights. As discussed by Hand and Vinciotti (2002, 2003), cases with extremely low or high credit risk stand out: The challenge for statistical model is to separate those that lie near the boundary for acceptance. In our context, accounts for which $\widehat{Y}_i \approx 0$ or 1 are generally easily recognized; the important cases to classify are those with nontrivial probabilities (say, $\widehat{Y}_i \geq .05$). Ideally, one would like to place the most weight on those observations near the decision boundary. Of course, however, $\widehat{Y}_i$ until we have chosen a model. As a compromise, least squares evenly weights all of the observations rather than emphasizing the extremes.

4. VARIABLE SELECTION

Our model for bankruptcy includes a link function. We assume that a monotone transformation of the conditional mean is linear in the predictors:

$$g(\mu_i) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{i,p}, \qquad \mu_i = E[Y_i|X_i].$$

Most, if not all, of the β_j are likely to be 0. Let $\gamma = \{j_1, \ldots, j_q\}$ denote the indices of a subset of $q = |\gamma|$ predictors, and write $\mathbf{X}_\gamma = [1, \mathbf{X}_{\gamma_i}, \ldots, \mathbf{X}_{\gamma_q}]$ for the $n \times (q+1)$ matrix of predictors identified by γ and the intercept. (We include an intercept in all models.) When selecting predictors, we initialize $g(\cdot)$ to the identity, $g(x) = x$. Stoker (1986) discussed the effects of a missing or misspecified link function on the estimated coefficients. Once we have identified the predictors, we estimate the inverse link $h(\cdot) = g^{-1}(\cdot)$.

Our method for variable selection mixes step-down thresholding with a robust p-value. With the identity link, the fit of the model with predictors γ is

$$\widehat{\mathbf{Y}}(\gamma) = \hat{\boldsymbol{\beta}}(\gamma)_0 + \sum_{j\in\gamma} \hat{\beta}_j(\gamma)\mathbf{X}_j,$$

with $\hat{\boldsymbol{\beta}}(\gamma) = (\mathbf{X}_\gamma \mathbf{X}_\gamma)^{-1}\mathbf{X}'_\gamma \mathbf{Y}$. We seek the subset γ that minimizes the mean squared error $\mathrm{MSE}(\gamma) = E\|\boldsymbol{\mu} - \widehat{\mathbf{Y}}(\gamma)\|^2$, where for vectors $\mathbf{x}$, $\|\mathbf{x}\|^2 = \sum x_i^2$. Notice that minimizing $\mathrm{MSE}(\gamma)$ is equivalent to minimizing the expected squared error when predicting an independent copy of the response, $\mathbf{Y}^* = \boldsymbol{\mu} + \boldsymbol{\epsilon}^*$ for $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)'$. Models that estimate $\boldsymbol{\mu}$ accurately also predict well out-of-sample.

4.1 Background

A common starting point for variable selection is to find an unbiased estimator of $\mathrm{MSE}(\gamma)$ and then choose the subset that minimizes this estimator. This is the path taken by the AIC (Akaike 1973) and C_p (Mallows 1973). The act of choosing the model with the smallest estimated mean squared error, however, leads to selection bias. The minimum of a collection of unbiased estimates is not unbiased. This effect is small when the AIC is used, for example, to select the order of a nested sequence of autoregressions or polynomials. The problem becomes magnified, however, when one compares many models of equal dimension (as noted in Mallows 1973). The resulting selection bias produces a criterion that chooses too many variables when none is in fact useful.

The literature contains a variety of alternatives to the AIC (as reviewed, e.g., in McQuarrie and Tsai 1998; Miller 2002). At the extreme, the Bonferroni criterion selects only those predictors that have a two-sided p-value smaller than α/p, where p is the number of predictors under consideration and α is the Type I error rate, typically 5%. In contrast to the AIC, the Bonferroni criterion selects on average only a fraction of one predictor under the null model. Because the p-values implied by the Bonferroni criterion can be so small (e.g., $.025/67{,}000 \approx .00000037$), many view this method as hopelessly conservative. The associated threshold, however, is not as large as might be expected (for $p = 67{,}000$ and $\alpha = .05$, the threshold for the z score is 4.95). Important predictors—those with z scores larger than 10, say—remain easy to detect.

The parsimonious models identified by the Bonferroni criterion have theoretical advantages. Results in statistical decision theory show that predictions from these models are optimal in a certain minimax sense. These optimality properties are typically associated with hard thresholding (Donoho and Johnstone 1994) or the risk inflation criterion (RIC; Foster and George 1994). We refer to both of these methodologies as the RIC. The RIC is usually described in the context of a Gaussian orthogonal regression with known error variance (think wavelets). In this idealized setting, the RIC selects X_j as a predictor if its squared z score $z_j^2 > 2\log p$. The resulting predictions obtain a type of minimax optimality. If q nonzero β_j are scattered among the p elements of $\boldsymbol{\beta}$, Foster and George (1994) showed that

$$\min_{\hat{\boldsymbol{\beta}}} \max_{\boldsymbol{\beta}} \frac{E\,\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2}{q\sigma^2} = 2\log p - o_p(\log p). \tag{3}$$

The minimum is taken over *all* estimators of $\boldsymbol{\beta}$ and is asymptotic in the number of predictors p, holding q fixed. The MSE of the model identified by the RIC is within a factor of $2\log p$ of that obtained by estimating the true model, and this is the best possible asymptotic performance. Miller (2002) discussed the RIC further.

What is less well known is that models chosen by the RIC are essentially the same as those identified by the Bonferroni

criterion. Let $\overline{\Phi}(x) = \int_x^\infty e^{-t^2/2}/\sqrt{2\pi}\, dt$ denote the upper tail of the standard normal distribution. The Bonferroni threshold $\overline{\Phi}^{-1}(\alpha/p)$ is asymptotically sandwiched for large p between $\sqrt{2\log p - \log\log p}$ and $\sqrt{2\log p}$. As a result, Bonferroni selection is asymptotically (in p) equivalent to the RIC, with the accuracy of the theorems in Foster and George (1994) or Donoho and Johnstone (1994). In fact, the RIC roughly corresponds to a Bonferroni threshold with $\alpha \approx .2$ for $100 \le p \le 100{,}000$ (Foster and Stine 2002).

4.2 Step-Down Testing

Although the RIC and Bonferroni criterion share the optimality of (3), the size of the threshold makes it unlikely that either can identify more subtle effects. The analysis igiven by Foster and George (1994) suggests that one can do better, at least in models in which more predictors affect the response. When the number of nonzero coefficients q is near zero, the minimax result (3) implies one can do little better than Bonferroni. With more nonzero coefficients, however, we can obtain some improvement. In particular, a modification of the proof of (3) yields the revised claim (Wang 2002)

$$\min_{\hat{\beta}} \max_{\beta} \frac{E\,\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2}{q\sigma^2} = 2\log\frac{p}{q} - o_p(\log p)$$

as long as the proportion of nonzero coefficients diminishes asymptotically, $q/p \to 0$ as $p \to \infty$ (see also Abramovich, Benjamini, Donoho, and Johnstone 2000; Johnstone and Silverman 2002).

This observation motivates a step-down rule for variable selection in the spirit of multiple testing as in Benjamini and Hochberg (1995) and Simes (1986). Rather than compare every p-value to α/p, a step-down rule gradually increases the threshold as significant features are discovered. Suppose that we have ordered the two-sided p-values obtained in an orthogonal regression with known error variance as $\hat{p}_{(1)} \le \hat{p}_{(2)} \le \cdots \le \hat{p}_{(p)}$. (It is unfortunate that p has become the common symbol for the number of available predictors; we hope that our use of $\hat{p}$ for p-values will avoid confusion.) We first compare the p-value of the most significant predictor to the Bonferroni threshold, α/p. If $\hat{p}_{(1)} < \alpha/p$, we add the associated predictor $X_{(1)}$ to the model. Otherwise we retain the null model. Assuming we added $X_{(1)}$, we then consider the second most significant predictor. In place of the Bonferroni bound, we compare $\hat{p}_{(2)}$ to the larger threshold $2\alpha/p$. The selection process stops if $p_{(2)} > 2\alpha/p$; otherwise we add $X_{(2)}$ to the model and continue to the third predictor. In general, the step-down rule implies that we

$$\text{add the } q\text{th most significant factor} \quad \Longleftrightarrow \quad \hat{p}_{(q)} \le \frac{q\alpha}{p}.$$

The resulting procedure can also be motivated by ideas in empirical Bayes estimation (George and Foster 2000) and information theory (Foster and Stine 1996). As summarized in Table 2 (which is discussed further in Sec. 6), the step-down procedure improves the predictive accuracy, albeit at the cost of a larger model. Although the gains are modest, the additional predictors found by the step-down rule indeed improve the predictions.

Remark. The theory underlying thresholding presumes orthogonal predictors, as in wavelet regression. The predictors available to model bankruptcy are, of course, collinear. Note, however, that an orthogonalization procedure, such as Gram–Schmidt, converts any set of linearly independent vectors into a sequence of orthogonal subspaces. That is the approach that we take. Given an ordering of the predictors, the stepwise algorithm removes each from the remaining covariates.

4.3 Bennett's Inequality

Step-down selection needs p-values. These are easy to obtain if $\hat{\beta}_j$ is normal; even for a binary response, we would expect the central limit theorem (CLT) to produce normal estimates with 600,000 observations. Because of our sparse data, however, high leverage points affect $\hat{\beta}_j$. Many of the interactions in our application are products of indicators, such as those for missing data, and are sparse. The CLT is of little help, and we have to turn to alternatives.

The stylized example in Figure 2 shows this problem in more detail. These simulated data illustrate the impact of combining a sparse binary response with sparse predictors. We simulated the predictor as $X_i \overset{\text{iid}}{\sim} \mathrm{N}(0, .025^2)$ for $i = 2, \ldots, 10{,}000$ with $X_1 = 1$. We independently simulated the response, with $P(Y_i = 1) = 1/1{,}000 = 1 - P(Y_i = 0)$. (The values plotted in Figure 2 are dithered vertically with normal noise.) Under this setup, the estimated slope $\hat{\beta}_1$ is about $Y_1 - .001$. Thus, $\hat{\beta}_1 \approx 1$ whenever $Y_1 = 1$, and a commonsense p-value for $\hat{\beta}_1$ for the data in Figure 2 is $1/1{,}000$. If we fit a least squares regression,

Table 2. Validation Sums of Squares (VSS) for the Fivefold Reversed Cross-Validation of Stepwise Regression Using the AIC, the Bonferroni Criterion, and Step-Down Testing

		AIC									
		Linear		Quadratic		Bonferroni		Step-down VSS*			
Data	No. of bankruptcies	q	VSS	q	VSS	q	VSS	q	Raw	LoR	PAV
$\mathcal{V}_1$	1,786	102	1,730	350	1,680	15	1,651	33	1,630	1,605	1,573
$\mathcal{V}_2$	1,802	99	1,747	350	1,648	15	1,664	34	1,647	1,555	1,573
$\mathcal{V}_3$	1,777	104	1,722	350	1,633	13	1,639	26	1,625	1,618	1,575
$\mathcal{V}_4$	1,813	95	1,756	350	1,674	23	1,673	38	1,653	1,584	1,609
$\mathcal{V}_5$	1,798	99	1,742	350	1,670	20	1,643	28	1,640	1,594	1,597
Avg	1,795		1,739		1,661		1,654		1,639	1,591	1,585
Improvement	0		56		134		141		156	204	210

NOTE: The columns headed by q indicate the number of predictors in the models identified by the AIC (over linear and quadratic predictors), the Bonferroni criterion, and step-down testing. The AIC search over interactions was halted at 350 predictors. (Table 3 summarizes key predictors.)

*The step-down testing is without adjustment (Raw), calibrated using logistic regression (LoR), and calibrated with the pool adjacent violators (PAV) algorithm.

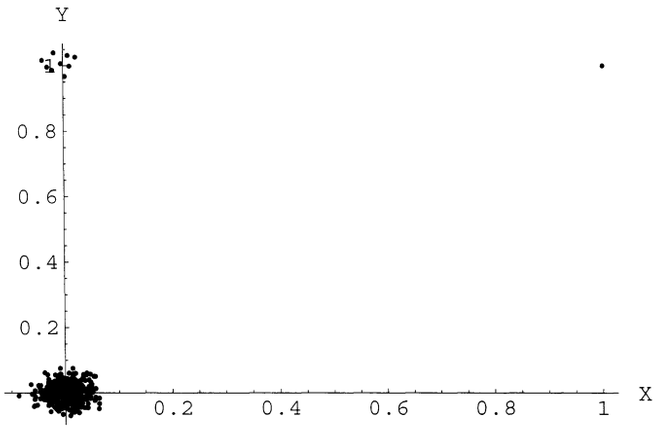

Figure 2. Simulated Data That Illustrates How a Single Highly Leveraged Observation Can Inflate the Significance of a Predictor in Sparse Data. The normal approximation suggests a p-value near $\overline{\Phi}^{-1}(14) \approx 0$, whereas 1/1,000 is appropriate. Bennett's inequality (4) conservatively bounds the p-value at 1/100.

the t ratio for $\hat{\beta}_1$ is about 14, grossly overstating the significance if we compute its p-value from the t distribution. The combination of a leveraged observation with the rare event $Y_1 = 1$ inflates the significance. The problem is not with the t ratio per se, but with our use of the t distribution to determine its significance.

Bennett's inequality avoids this inflated significance. For a bounded collection of n independent random variables $U_1, \ldots, U_n$ with $\sup |U_i| < M$, $E\,U_i = 0$, and $\sum_i E\,U_i^2 = 1$, Bennett (1962) showed that for $\tau > 0$,

$$P\left(\sum_i U_i \geq \tau\right) \leq \exp\left(\frac{\tau}{M} - \left(\frac{\tau}{M} + \frac{1}{M^2}\right)\log(1 + M\tau)\right) = \overline{G}(\tau, M). \tag{4}$$

If $M\tau$ is small, the upper bound is approximately $\exp(-\tau^2/2)$, close to the usual normal probability. When the dispersion is concentrated in just a few of the U_i, however, Bennett's bound accommodates the resulting Poisson-like variation. For the data in Figure 2, the use of Bennett's inequality as shown in Section 5.2 conservatively bounds the significance of the slope at about 1/100. Though obviously conservative, this bound is about as good as possible. The leverage of (X_1, Y_1) is .14, so only about 14% of the variation is nonnormal.

Thus, we select variables using a step-down rule applied to p-values determined by $\overline{G}$ rather than $\overline{\Phi}$. Section 5 supplies the important details of the calculations. In particular, we do not simply pick the variable with the smallest p-value.

4.4 Calibration

The last step in our modeling calibrates the predictions from stepwise regression. For our data, in-sample plots of Y_i on $\widehat{Y}_i$ reveal a lack of calibration. Cook and Olive (2001) used similar plots for visualizing a Box–Cox transformation. For our purposes, these plots show that $E(Y_i|\widehat{Y}_i) \neq \widehat{Y}_i$. As a remedy, we consider two easily computed estimates of the inverse link $h(\cdot)$. A logistic regression using the predictors identified by stepwise selection fixes the link and reweights the predictors. Alternatively, the pooled adjacent violators algorithm provides a nonparametric estimate of $h(\cdot)$ that retains $\widehat{Y}$ from the regression. Section 5 gives the details of the calculations.

5. IMPLEMENTATION

5.1 Modifying the Sweep Operator

Our implementation of stepwise regression modifies the standard sweep operator (Goodnight 1979; Thisted 1988). Selection from 67,160 variables with 600,000 observations breaks most implementations. For this section, assume that each variable has been centered and let $\mathbf{X}$ denote the $n \times q$ matrix of predictors in a regression model for $\mathbf{Y}$. The least squares estimator is $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ with estimated covariance matrix $\text{var}(\hat{\boldsymbol{\beta}}) = s^2(\mathbf{X}'\mathbf{X})^{-1}$ with $s^2 = \mathbf{e}'\mathbf{e}/(n-q)$ and residual vector $\mathbf{e} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$. The value of the sweep operator in stepwise regression becomes apparent when one seeks to add predictors to a regression. Suppose that the $p - q$ predictors that are not in the model comprise the array $\mathbf{Z}$ and denote the projection matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Sweeping $\mathbf{X}$ from the expanded cross-product matrix yields the transformation

$$\left[\begin{array}{c|c|c} \mathbf{Y}'\mathbf{Y} & \mathbf{Y}'\mathbf{X} & \mathbf{Y}'\mathbf{Z} \\ \hline \mathbf{X}'\mathbf{Y} & \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \hline \mathbf{Z}'\mathbf{Y} & \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} \end{array}\right] \Longrightarrow \left[\begin{array}{c|c|c} \mathbf{e}'\mathbf{e} & \hat{\boldsymbol{\beta}}' & \mathbf{e}'(\mathbf{I}-\mathbf{H})\mathbf{Z} \\ \hline \hat{\boldsymbol{\beta}} & (\mathbf{X}'\mathbf{X})^{-1} & \mathbf{X}'(\mathbf{I}-\mathbf{H})\mathbf{Z} \\ \hline \mathbf{Z}'(\mathbf{I}-\mathbf{H})\mathbf{e} & \mathbf{Z}'(\mathbf{I}-\mathbf{H})\mathbf{X} & \mathbf{Z}'(\mathbf{I}-\mathbf{H})\mathbf{Z} \end{array}\right].$$

The first row of the resulting array begins with the residual sum of squares and $\hat{\boldsymbol{\beta}}'$ followed by $\mathbf{e}'(\mathbf{I}-\mathbf{H})\mathbf{Z}$. This vector is $n - q$ times the estimated partial covariance between $\mathbf{Y}$ and $\mathbf{Z}_j$ given $\mathbf{X}$. The diagonal of $\mathbf{Z}'(\mathbf{I}-\mathbf{H})\mathbf{Z}$ determines the partial variances of $\mathbf{Z}$. The ratio

$$\frac{\mathbf{e}'(\mathbf{I}-\mathbf{H})\mathbf{Z}_j}{\sqrt{(\mathbf{e}'\mathbf{e})(\mathbf{Z}'(\mathbf{I}-\mathbf{H})\mathbf{Z})_{jj}}}$$

gives the estimated partial correlation. The predictor $\mathbf{Z}_j$ with the largest partial correlation offers the greatest reduction in the residual sum of squares and is the choice of the standard forward stepwise regression.

To handle so many predictors, we avoid computing the entire cross-product matrix; most of the elements are never used. Rather, we defer calculations and compute only those elements that are needed for identifying the next predictor. When considering the omitted predictors $\mathbf{Z}$, our implementation computes the full sweep for the relatively small augmented matrix $[\mathbf{Y}|\mathbf{X}]$, providing standard errors, slopes, and residuals $\mathbf{e}$. For evaluating the remaining predictors, the algorithm computes $\mathbf{e}'(\mathbf{I}-\mathbf{H})\mathbf{Z}_j$ and the diagonal of $\mathbf{Z}'(\mathbf{I}-\mathbf{H})\mathbf{Z}$. Combined with the sweep of $\mathbf{X}$, we have all of the information needed to find the next predictor. Although they offer the most in-sample improvement, predictors with large partial correlation may not, in fact, predict well out-of-sample. For modeling bankruptcy, partial correlations tend to find leverage points like that in Figure 2.

5.2 Finding p-Values

In place of partial correlations, our second modification of the usual forward search uses p-values determined by Bennett's inequality. Let $\mathbf{z}$ denote the vector defined by a predictor not in the model. We first sweep the current predictors from $\mathbf{z}$, forming $\tilde{\mathbf{z}} = (\mathbf{I} - \mathbf{H})\mathbf{z}$. The sweep operator includes coefficients so that this operation is done as a weighted sum of $q + 1$ vectors. The estimated slope if $\tilde{\mathbf{z}}$ is added to the model is

$$\hat{\boldsymbol{\beta}}_{\mathbf{z}} = \frac{\tilde{\mathbf{z}}'\mathbf{y}}{\tilde{\mathbf{z}}'\tilde{\mathbf{z}}} \qquad \text{with} \qquad \text{var}(\hat{\boldsymbol{\beta}}_{\mathbf{z}}) = \frac{\tilde{\mathbf{z}}'\mathbf{D}\tilde{\mathbf{z}}}{(\tilde{\mathbf{z}}'\tilde{\mathbf{z}})^2}, \tag{5}$$

where $\mathbf{D}$ is the $n \times n$ diagonal matrix with elements $D_{ii} = E\,Y_i(1 - E\,Y_i) = \text{var}(Y_i)$. We estimate $\text{var}(\hat{\boldsymbol{\beta}}_{\mathbf{z}})$ by substituting $\widehat{\mathbf{D}} = \text{diag}(\widehat{Y}_i(1 - \widehat{Y}_i))$ for $\mathbf{D}$ in (5), where $\widehat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$ is the current fit *without* $\tilde{\mathbf{z}}$, constrained so that $0 < \widehat{Y}_i < 1$.

To estimate the significance of $\tilde{\mathbf{z}}$, we bound its p-value using Bennett's inequality. Write the observed t ratio as

$$t_{\mathbf{z}} = \frac{\hat{\beta}_{\mathbf{z}}}{\text{SE}(\hat{\beta}_{\mathbf{z}})} = \sum_i \frac{\tilde{z}_i(Y_i - \widehat{Y}_i)}{(\tilde{\mathbf{z}}'\widehat{\mathbf{D}}\tilde{\mathbf{z}})^{1/2}}, \tag{6}$$

where SE denotes standard error. Conditional on the current fit and the null hypothesis that $\beta_{\mathbf{z}} = 0$, the summands in (6) meet the conditions for the U_i in (4). The bounding M from (4) is

$$M_{\mathbf{z}} = \frac{\max_i(|\tilde{z}_i|, \max(\widehat{Y}_i, 1 - \widehat{Y}_i))}{(\tilde{\mathbf{z}}'\widehat{\mathbf{D}}\tilde{\mathbf{z}})^{1/2}}$$

and Bennett's bound for the two-sided p-value is $\hat{p}_{\mathbf{z}} = 2\,\overline{G}(|t_{\mathbf{z}}|, M_{\mathbf{z}})$.

We now have all the pieces of our selection procedure. Assume that the model currently has q predictors and that they have been swept from the excluded $p - q$ predictors. For each excluded predictor, we bound its two-sided p-value with $\hat{p}_{\mathbf{z}} = 2\overline{G}(|\hat{\beta}_{\mathbf{z}}|/\text{SE}(\hat{\beta}_{\mathbf{z}}), M_{\mathbf{z}})$. If $\hat{p}_{\mathbf{z}} > \alpha q/p$ for every unused predictor, the step-down search halts. If several p-values are significant, we choose the predictor that promises the largest improvement in the fit as estimated by

$$\text{SS}(\mathbf{z}) = (\tilde{\mathbf{z}}'\tilde{\mathbf{z}})\big(|\hat{\beta}_{\mathbf{z}}| - \tau_{\mathbf{z}}\text{SE}(\hat{\beta}_{\mathbf{z}})\big)^2,$$

where $\tau_{\mathbf{z}}$ satisfies $2\overline{G}(\tau_{\mathbf{z}}, M_{\mathbf{z}}) = \alpha q/p$ and SS denotes sum of squares. The threshold $\tau_{\mathbf{z}}$ identifies the smallest value for $t_{\mathbf{z}}$ that would be judged significant for these values of α, q, p, and $M_{\mathbf{z}}$, analogous to the, say, 5% point in a t distribution. The $\text{SS}(\mathbf{z})$ is positive for predictors identified as significant by the Bennett p-value. One can think of $\text{SS}(\mathbf{z})$ as a "guaranteed" reduction in residual sum of squares in the sense of a confidence interval; it shrinks the effect attributed to a variable to the smallest value consistent with its estimated effect. This bound resembles soft thresholding of estimates with different variances (Donoho and Johnstone 1994), although we shrink the estimates only to order them, not once they have been added to the fit.

5.3 Calibration

The last step in our modeling calibrates the predictions. A predictor $\widehat{Y}_i$ is calibrated if $\widehat{Y}_i = E(Y_i|\widehat{Y}_i)$. Figure 3 plots the average probability of bankruptcy on the average predicted values in $\mathcal{E}_1$. To obtain this plot, we extracted the 3,382 observations in $\mathcal{E}_1$ for which $\widehat{Y}_i > .075$. As a simple smoothing procedure, we partitioned these cases into disjoint bins of 200 adjacent observations, identified by sorting the predicted values. Figure 3 plots the proportion of bankruptcies in these subsets versus the average of the predicted values. Vertical bars denote ± 2 standard errors, and the diagonal is the reference line $x = y$. If the predictions were calibrated, then most intervals would include the diagonal. They do not; the actual probability of bankruptcy trends away from the diagonal. For example, the average prediction in one bin is .20, but 39% of these 200 are bankruptcies.

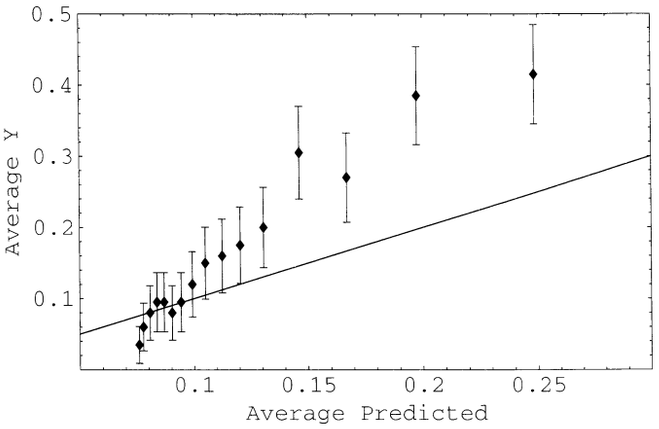

Figure 3. For the Estimation Sample $\mathcal{E}_1$, the Proportions of Bankrupt Accounts in Nonoverlapping Subsets Do Not Track the Corresponding Average Prediction. Error bars (± 2 SE) show that the deviations from the diagonal reference line indicate a systematic lack of calibration.

To calibrate these predictions, we consider two approaches. By assumption, $\mu_i = h(\boldsymbol{\beta}'\mathbf{X}_i)$. Our stepwise search has found a good set of predictors, but constrains $h(x) = x$. It is computationally intractable to remove this constraint fully, but two compromises work well in our data. For the first, we fix $h(x) = 1/(1 + e^{-x})$ and fit a logistic regression that reweights the predictors identified by the stepwise search. This approach constrains the link, but allows a reweighting of the predictors suited to the logistic link. For the second, we retain $\widehat{Y}$ but unconstrain the link, estimating $h(\cdot)$ by using the pool adjacent violators algorithm (PAV; Barlow, Bartholomew, Bremmer, and Brunk 1972). This algorithm fits a monotone function of Y on $\widehat{Y}$, estimating $h(\cdot)$. Both procedures represent one step of an iterative scheme. One could revise the logistic fit by altering the choice of predictors, perhaps dropping some or even adding others. Similarly, the PAV calibration could benefit from reweighting the predictors in $\widehat{Y}$ given an estimate of $h(\cdot)$, alternating the estimation of $h(\cdot)$ with the weighting of the predictors.

Our implementation of PAV is straightforward. We begin by sorting $\widehat{Y}$, merging ties into a weighted observation. Starting with the smallest, we group the sorted data into bins to obtain a monotone predictor. Suppose that we currently have a bin composed of observations that have predicted values that lie in the half-open interval $b_j < \widehat{Y}_i \le b_{j+1}$. Let $\hat{\pi}_j$ denote the proportion of bankruptcies in this bin and assume that the probabilities for the current fit are monotone increasing, $\hat{\pi}_1 \le \hat{\pi}_2 \le \cdots \le \hat{\pi}_j$. If the next observation, say Y_i, is a bankruptcy, close this bin and start another. If the next observation is not a bankruptcy, revise the range of the current bin to $(b_j, \widehat{Y}_i]$ and update $\hat{\pi}_j$. If $\hat{\pi}_j \ge \hat{\pi}_{j-1}$, we move on to Y_{i+1}. If instead $\hat{\pi}_j < \hat{\pi}_{j-1}$, we merge (recursively, as needed) the jth bin with its predecessors

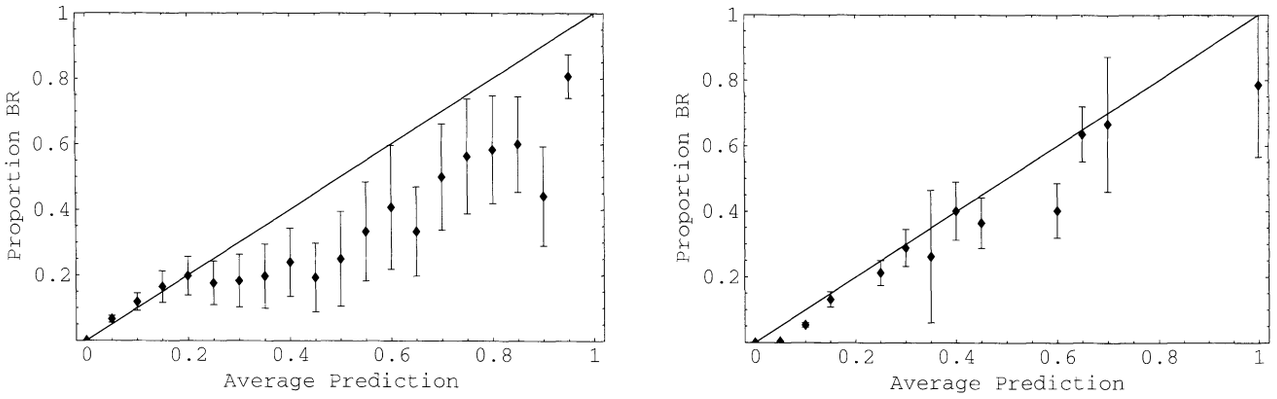

Figure 4. Calibration Improves the Predictions in $\mathcal{V}_1$ From Stepwise Regression. These plots show the average out-of-sample proportion of bankruptcy for given forecasted probabilities. Logistic regression was used to calibrate the predictions on the left and the pool adjacent violators algorithm was used to correct those on the right. Both out-of-sample predictions are closer to the diagonal reference line and more calibrated than the uncalibrated, in-sample predictions in Figure 3.

until the fit is monotone. This procedure yields a sequence of, say, K intervals with monotone estimates of the probability of bankruptcy,

$$\{[b_1, b_2], \hat{\pi}_1\}, \{(b_2, b_3], \hat{\pi}_2\}, \ldots, \{(b_K, b_{K+1}], \hat{\pi}_K\},$$
$$\hat{\pi}_1 \le \hat{\pi}_2 \le \cdots \le \hat{\pi}_K.$$

The sequence $\hat{\pi}_1, \ldots, \hat{\pi}_K$ amounts to a monotone, piecewise constant estimate of $h(\cdot)$,

$$\hat{h}(y) = \hat{\pi}_k \qquad \text{for } b_k < y \le b_{k+1}.$$

Figure 4 shows the effects of these calibrations when predicting the held-back validation sample $\mathcal{V}_1$. Logistic regression is calibrated at smaller probabilities, whereas PAV is more consistently calibrated over the full range. Compared to the uncalibrated predictions shown in Figure 3, both sets of calibrated predictions fall closer to the diagonal; the probability of bankruptcy is closer to the average prediction. Notice that Figure 4 shows out-of-sample data; the analog of Figure 3 for PAV would be uninteresting because the calibrated averages would fall, by construction, along a 45 line.

6. RESULTS

We used a "reversed" fivefold cross validation to evaluate our methodology and compare it to the machine-learning classifiers. As described in Section 2, we randomly partitioned the complete dataset of 2.9 million months of credit activity into five disjoint subsets. We then fit models to the estimation samples and obtained their out-of-sample predictions.

To set a baseline, we ran standard forward stepwise regression using the AIC to select from the 365 linear predictors. The results for these models confirm the need to use interactions in regression models. Table 2 shows the Brier scores, or validation sums of squares (VSS), when each regression predicts the corresponding validation sample. The AIC selects models with about 100 predictors, but its models find little structure and have much higher Brier scores than models that use interactions. To put these scores into perspective, observe that the Brier score obtained by predicting no bankruptcies (i.e., set $\widehat{Y}_i \equiv 0$) is the number of bankruptcies in each validation sample. Each reduction of the Brier score by 1, in a loose sense, represents finding one more bankruptcy. For example, the first validation sample $\mathcal{V}_1$ has 1,786 bankruptcies. The Brier score of the model selected by the AIC from linear predictors is 1,730, so this model "finds" 56 bankruptcies.

We next allowed forward stepwise regression to select from the expanded set of variables, including interactions. In this setting, the AIC keeps selecting more and more predictors. The AIC adds predictors to the model so long as the incremental squared t-ratio exceeds 2, corresponding to a p-value near .16. With 67,000 to choose from, we were unable to run the AIC to completion. Table 2 shows the VSS of models that use the first 350 predictors chosen by the AIC. At this point, each added predictor increases the resulting Brier score even though the AIC itself is falling. If run further, the VSS for the AIC would have been larger.

The Bonferroni approach (or hard thresholding) stops the variable selection much sooner, evidently leaving out some predictive variables but obtaining a much better Brier score. The average Brier score of Bonferroni selection is 1,654, whereas the Brier score of the larger models produced by step-down testing is 1,639. On average, step-down testing offers a modest 10% improvement over hard thresholding, at the expense of a larger model. Thus interactions generate more predictive fits that can be obtained by searching the original variables, as long as one stops the selection process before overfitting.

The results in Table 2 confirm the benefits of calibration. Calibration offers more than three times the improvement obtained by using step-down testing in place of Bonferroni. Whereas step-down testing reduces the average validation sum of squares by 15, calibration further reduces these errors by 48 when using logistic regression or 54 when using the PAV.

All of the predictors in the stepwise regressions found by our methodology are interactions (including six predictors that are squares). Table 3 identifies the most common of these; the interactions in this table appear in three or more of the five regressions in the cross-validation. Combined, these interactions comprise 61 of the 159 interactions in these models. The interactions are highly overlapping. Of these 61 interactions, two variables appear in 50: the number of credit cards held by the account holder and the number of times that this person let an

Table 3. Interactions That Appear in Three or More of the Five Stepwise Regression Models Obtained in the Fivefold Cross-Validation Analysis

Common interactions		Prevalence		Appears in
X_1	X_2	X_1	X_2	k models
Number of credit cards	Prior cards past due 60 days	35	36	5
Number of credit cards	Number of credit cards	35	35	5
Number of credit cards	Prior cards closed	35	20	4
Number of credit cards	Late charge in prior month	35	31	3
Number of credit cards	External flag unavailable	35	20	3
Number of credit cards	External credit flag 2	35	16	3
Number of credit cards	External credit flag 1	35	9	3
Prior cards past 60 days	Late charge in prior month	36	31	5
Prior cards past 60 days	Prior cards closed	36	20	5
Prior cards past 60 days	External flag unavailable	36	20	5
Prior cards past 60 days	Internal bank status code 2	36	8	3
Prior cards past 60 days	External credit flag 2	36	16	3
Prior cards past 60 days	External flag 1, prior quarter	36	5	3
Late charge prior month	Prior cards closed	31	20	5
Late charge prior month	Missing FICO score	31	7	3
External flag unavailable	External credit flag 3	20	4	3

NOTE: The shown prevalence indicates the number of interactions with that predictor among the 159 interactions in the five regression models. (Table 2 summarizes the fits of these models.)

account slip for 2 months. Predictors labeled "Internal" or "External" are proprietary creditor scores for the accounts. Table 3 also lists the prevalence of these predictors in interactions. For example, the number of credit cards held by an individual appears in 35 interactions (5 of which are squares); the number of times that the individual has delayed paying for 2 months appears in 36 interactions.

To see how well these predictions stack up to those of C4.5 and C5.0 (Quinlan 1993), we repeated the cross-validation with these classifiers. Unlike regression, C4.5 and C5.0 require as input the ratio ρ of the cost of missing a bankruptcy to the cost of annoying a good customer. Given ρ, C4.5 and C5.0 search for a tree that minimizes observed costs. Costs experienced in the industry are confidential, so we chose several illustrative values for ρ. We ran C5.0 with the default parameters, including 10 steps of boosting. The sizes of the resulting trees vary widely, growing with ρ. For $\rho = 2$, C4.5 produced trees with an average of 94 terminal nodes. For $\rho = 49$ and $\rho = 199$, the trees grow to 712 and 960 terminal nodes, respectively. Predictors in Table 3 appear frequently in the trees.

Instead of requiring a new model for each ρ, one regression handles any ρ. If calibrated, one obtains minimal costs by classifying as bankrupt any observation for which $\widehat{Y}_i > 1/(1+\rho)$. For example, suppose that the average cost of calling a good customer (in lost business, perhaps) is \$50, whereas each missed bankruptcy costs, on average, \$4,950. The cost ratio $\rho = 4{,}950/50 = 99$. How should we classify an observation for which $P(Y_i = 1) = 1/100 = 1/(1+\rho)$? If we classify this account as bankrupt, then our expected cost is $\$50(99/100) = \$49.50$. On the other hand, labeling this a good account produces the same expected cost, $\$4{,}950(1/100) = \$49.50$. Hence, to minimize expected costs, we classify observations as bankrupt once the predicted probability exceeds $1/(1+\rho)$.

Table 4 and Figure 5 compare the costs of the null predictor that classifies all accounts "good" to C4.5, C5.0, and regression, both with and without the PAV calibration. Larger cost ratios seem more natural in the context of modeling bankruptcy. The logit horizontal scale in Figure 5 associates areas in the figure with entropic losses rather than quadratic losses summarized in Table 2. Table 4 shows the costs generated by these predictors in each validation sample for two rather different cost ratios, $\rho = 199$ and $\rho = 4$. The table is thus useful for appreciating the variability of the costs; the costs generated by the trees are more variable across the samples.

All of the models beat the null classifier, albeit not by much when ρ approaches 1. Regression dominates C4.5 and C5.0 for larger cost ratios. The costs generated by the calibrated regression predictor, for example, are about half those of C5.0 for $\rho = 199$. Boosting was not helpful for high costs; C4.5 obtains lower costs in the two high-cost situations. Once ρ drops to

Table 4. Fivefold Cross-Validation of the Costs Obtained by a Null Model (predicting no bankruptcy), Tree Classifiers C4.5 and C5.0, and Stepwise Regression (Reg), and Calibrated Stepwise Regression (PAV)

	$\rho = 4$					$\rho = 199$				
Subset	Null	C4.5	C5.0	Reg	PAV	Null	C4.5	C5.0	Reg	PAV
$\mathcal{V}_1$	1,429	1,311	1,207	1,300	1,244	1,777	975	1,400	842	618
$\mathcal{V}_2$	1,441	1,341	1,223	1,306	1,232	1,792	1,095	1,182	897	650
$\mathcal{V}_3$	1,421	1,409	1,237	1,292	1,257	1,768	745	1,217	848	644
$\mathcal{V}_4$	1,450	1,344	1,218	1,305	1,289	1,803	1,140	1,213	849	624
$\mathcal{V}_5$	1,438	1,401	1,201	1,311	1,273	1,789	1,119	1,038	846	661
Avg	1,436	1,361	1,217	1,303	1,259	1,786	1,015	1,210	856	639

NOTE: The cost of missing a bankruptcy is assumed to be ρ times the cost of incorrectly classifying a good customer.

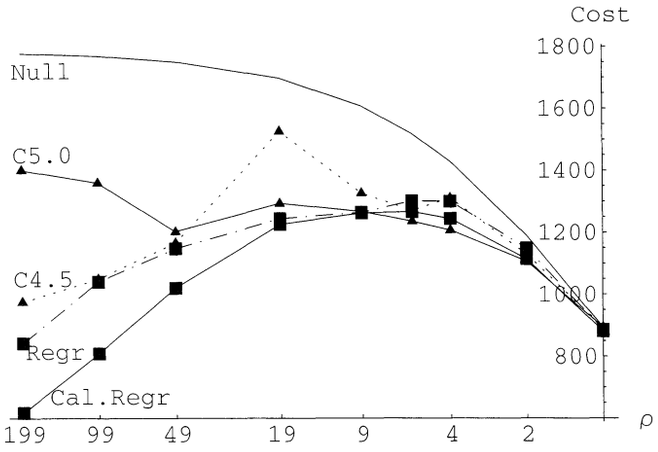

Figure 5. Average Costs Accumulated Over Each Fifth of the Data for the Null Model (solid line), Classification Trees (C4.5 and C5.0, triangles), and Regression (with and without calibration, boxes).

about 20, however, C5.0 performs about as well as regression and bests C4.5. With $\rho = 4$, C5.0 has the lowest costs.

7. DISCUSSION

Many important applications, such as diagnosing a disease, identifying profitable customers in a mailing list, or predicting bankruptcy, have become the province of computer science rather than statistics. The data-mining community has enthusiastically embraced neural nets and trees (e.g., see Fayyad and Uthurusamy 1996; Vol. 42, November 1999 issue of *Communications of the Association for Computing Machinery* offers several articles on knowledge discovery and machine learning). Our results demonstrate that, given some tuning, standard statistical methods are competitive. Like Leo Breiman, we urge statisticians to contribute to this larger modeling community (see Breiman's 2001 discussion of Cheng and Titterington 1994, and the work of Hastie, Tibshirani, and Friedman 2001).

Our results suggest two areas hold promise for obtaining better models of wide datasets with many predictors. Our step-down selection process does not overfit; the added variables reduce the out-of-sample error. Other variables not included may nonetheless be predictive; it would be useful to find methods that can identify more predictive features without overfitting. Also, the clear benefits of one step of calibration suggest that augmenting the selection process with an iterative calibration may produce larger gains. It appears that one could obtain further gains by alternating the estimation of the link with the choice of weights for combining the predictors.

On the substantive side, we hardly expect creditors to drop their familiar, hand-tooled models. That said, automated models can reveal whether data contain predictors missed by manual search. Although we have automated this choice from a host of predictors, such automation cannot find certain predictors, such as, say, an exponential smooth of some aspect of past payment history. It remains for well-informed judgment to add such features to the domain of predictors. An analyst often has a keen sense of which transformations and combinations of predictors will be effective, and these should be added to the search space. Our on-going research seeks ways to add this knowledge to the modeling process.

[Received March 2001. Revised January 2004.]

REFERENCES

Abramovich, F., Benjamini, Y., Donoho, D., and Johnstone, I. (2000), "Adapting to Unknown Sparsity by Controlling the False Discovery Rate," Technical Report 2000-19, Stanford University, Dept. of Statistics.

Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle," in *Second International Symposium on Information Theory*, eds. B. N. Petrov and F. Csàki, Budapest: Akademia Kiàdo, pp. 261–281.

Barlow, R. E., Bartholomew, D. J., Bremmer, J. M., and Brunk, H. D. (1972), *Statistical Inference Under Order Restrictions*, New York: Wiley.

Bastos, E., Wolfinger, R., Duling, D., and Auslender, L. (2003), "Data Mining Breast Cancer Clinical and Expression Data," submitted for publication.

Benjamini, Y., and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society*, Ser. B, 57, 289–300.

Bennett, G. (1962), "Probability Inequalities for the Sum of Independent Random Variables," *Journal of the American Statistical Association*, 57, 33–45.

Breiman, L. (2001), "Statistical Modeling: The Two Cultures," *Statistical Science*, 16, 199–215.

Cheng, B., and Titterington, D. M. (1994), "Neural Networks: A Review From a Statistical Perspective" (with discussion), *Statistical Science*, 9, 2–54.

Cook, R. D., and Olive, D. J. (2001), "A Note on Visualizing Response Transformations in Regression," *Technometrics*, 43, 443–449.

Curnow, G., Kochman, G., Meester, S., Sarkar, D., and Wilton, K. (1997), "Automating Credit and Collections Decisions at AT&T Capital Corporation," *Interfaces*, 27, 29–52.

Donoho, D. L., and Johnstone, I. M. (1994), "Ideal Spatial Adaptation by Wavelet Shrinkage," *Biometrika*, 81, 425–455.

Fayyad, U., and Uthurusamy, R. (1996), "Data Mining and Knowledge Discovery in Databases," *Communications of the Association for Computing Machinery*, 39, 24–34.

Foster, D. P., and George, E. I. (1994), "The Risk Inflation Criterion for Multiple Regression," *The Annals of Statistics*, 22, 1947–1975.

Foster, D. P., and Stine, R. A. (1996), "Variable Selection via Information Theory," Technical Report Discussion Paper 1180, Northwestern University, Center for Mathematical Studies in Economics and Management Science.

——— (2002), "Hard Thresholding, RIC, and Bonferroni Methods Are Equivalent," unpublished manuscript.

Freedman, D. A. (1983), "A Note on Screening Regression Equations," *The American Statistician*, 37, 152–155.

George, E. I., and Foster, D. P. (2000), "Calibration and Empirical Bayes Variable Selection," *Biometrika*, 87, 731–747.

Goodnight, J. H. (1979), "A Tutorial on the SWEEP Operator," *The American Statistician*, 33, 149–158.

Gross, D. B., and Souleles, N. S. (2002), "An Empirical Analysis of Personal Bankruptcy and Delinquency," *The Review of Financial Studies*, 15, 319–347.

Gustafson, P. (2000), "Bayesian Regression Modeling With Interactions and Smooth Effects," *Journal of the American Statistical Association*, 95, 795–806.

Hand, D. J. (1997), *Construction and Assessment of Classification Rules*, New York: Wiley.

Hand, D. J., Blunt, G., Kelly, M. G., and Adams, N. M. (2000), "Data Mining for Fun and Profit," *Statistical Science*, 15, 111–131.

Hand, D. J., and Henley, W. E. (1997), "Statistical Classification Methods in Consumer Credit Scoring: A Review," *Journal of the Royal Statistical Society*, Ser. A, 160, 523–541.

Hand, D. J., and Vinciotti, V. (2002), "Scorecard Construction With Unbalanced Class Sizes," technical report, Imperial College, London, Dept. of Mathematics.

——— (2003), "Local Versus Global Models for Classification Problems," *The American Statistician*, 57, 124–131.

Hastie, T., Tibshirani, R., and Friedman, J. (2001), *The Elements of Statistical Learning*, New York: Springer-Verlag.

Johnstone, I. M., and Silverman, B. W. (2002), "Empirical Bayes Selection of Wavelet Thresholds," technical report, Stanford University, Dept. of Statistics.

Jones, M. P. (1996), "Indicator and Stratification Methods for Missing Explanatory Variables in Multiple Linear Regression," *Journal of the American Statistical Association*, 91, 222–230.

Mallows, C. L. (1973), "Some Comments on C_p," *Technometrics*, 15, 661–675.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models* (2nd ed.), London: Chapman & Hall.

McQuarrie, A. D., and Tsai, C.-L. (1998), *Regression and Time Series Model Selection*, Singapore: World Scientific.

Miller, A. J. (2002), *Subset Selection in Regression* (2nd ed.), London: Chapman & Hall.

Quinlan, J. R. (1993), *C4.5: Programs for Machine Learning*, San Mateo, CA: Morgan Kaufmann.

Rencher, A. C., and Pun, F. C. (1980), "Inflation of R^2 in Best Subset Regression," *Technometrics*, 22, 49–53.

Simes, R. J. (1986), "An Improved Bonferroni Procedure for Multiple Tests of Significance," *Biometrika*, 73, 751–754.

Stoker, T. M. (1986), "Consistent Estimation of Scaled Coefficients," *Econometrica*, 54, 1461–1481.

Thisted, R. A. (1988), *Elements of Statistical Computing: Numerical Computation*, New York: Chapman & Hall.

Thomas, L. C., Edelman, D. B., and Crook, J. N. (2002), *Credit Scoring and Its Applications*, Philadelphia, PA: Society for Industrial and Applied Mathematics.

Wang, L. (2002), "Problems in Adaptive Variable Selection," unpublished Ph.D. thesis, University of Pennsylvania, Dept. of Statistics.