JOURNAL OF APPLIED MATHEMATICS AND DECISION SCIENCES, 5(4), 215–234

Financial Applications of a Tabu Search Variable Selection Model

ZVI DREZNER zdrezner@fullerton.edu
College of Business and Economics, California State University, Fullerton, CA 92834.

GEORGE A. MARCOULIDES gmarcoulides@fullerton.edu
College of Business and Economics, California State University, Fullerton, CA 92834.

MARK HOVEN STOHS† mstohs@fullerton.edu
College of Business and Economics, California State University, Fullerton, CA 92834.

Abstract. We illustrate how a comparatively new technique, a Tabu search variable selection model [Drezner, Marcoulides and Salhi (1999)], can be applied efficiently within finance when the researcher must select a subset of variables from among the whole set of explanatory variables under consideration. Several types of problems in finance, including corporate and personal bankruptcy prediction, mortgage and credit scoring, and the selection of variables for the Arbitrage Pricing Model, require the researcher to select a subset of variables from a larger set. In order to demonstrate the usefulness of the Tabu search variable selection model, we: (1) illustrate its efficiency in comparison to the main alternative search procedures, such as stepwise regression and the Maximum R^2 procedure, and (2) show how a version of the Tabu search procedure may be implemented when attempting to predict corporate bankruptcy. We accomplish (2) by indicating that a Tabu Search procedure increases the predictability of corporate bankruptcy by up to 10 percentage points in comparison to Altman's (1968) Z-Score model.

Keywords: Statistical Forecasting, Variable Selection Techniques, Business Scoring Models, Finance, Bankruptcy/Financial Distress

1. Introduction

We argue that a comparatively new technique, the Tabu search variable selection model [Drezner, Marcoulides and Salhi (1999)], an extremely efficient means for selecting a subset of variables from among the whole set of explanatory variables under consideration, may be applied successfully to problems in finance. At least two classes of problems in finance require the selection of a subset of independent variables which yields the highest predictive power from among all possible subsets of the variables under consideration.

† Requests for reprints should be sent to Mark Stohs, College of Business and Economics, California State University, Fullerton, CA 92834.

One class includes specific problems in corporate, personal and real estate finance. For example, a widely used system in corporate finance is Altman's (1968) Z-score model for predicting corporate bankruptcy. Using discriminant analysis (DA), Altman selects a set of firm-specific variables intended to predict whether or not particular firms are likely to declare bankruptcy. Other problems in this class include the prediction of personal (consumer) bankruptcy, the prediction of bank loan defaults, commercial or residential mortgage scoring, and credit scoring. The practical scoring approach taken in these fields could readily be extended to the field of insurance.

The problem of selecting the appropriate factors for the Arbitrage Pricing Theory (APT) constitutes the second class by itself. As Roll (1988) notes, "the paucity of explanatory power represents a significant challenge to our science." He adds that *ex post* success for either the Capital Assets Pricing Model (CAPM) or the APT would suggest that the "R^2 should be close to 1.0." This is not to suggest that just any mix of factors for the APT, or any theory, satisfies the criteria of science. Instead, if additional factors provide more explanatory power, those factors are "acceptable" only if they are "indeed pervasive, non-diversifiable, and most important, are associated with additional risk premia" [Roll (1988)]. In short, factors (independent variables) ultimately selected for a theory must also be grounded in theory.

Debate over the best method for predicting the dependent variable centers around three issues: (1) which types of independent variables to use, (2) which statistical technique is most appropriate for the data at hand, and (3) which variable selection model is best for arriving at a finite set of independent variables from among a larger set of variables. As Altman (1968) notes with respect to predicting corporate bankruptcy, "the question becomes, which ratios are most important in detecting bankruptcy potential, what weights should be attached to those selected ratios, and how should the weights be objectively established?"

The first issue is best illustrated by the corporate bankruptcy models. Mossman Bell, Swartz, and Turtle (1998) consider the merits of models which rely on four different types of variables: financial statement ratios (as in Altman's Z-score), cash flows, stock returns and return standard deviations. The second issue focuses on whether regression analysis, DA, logit, probit, option-based models as in Shilton and Teall (1994) and based upon the original option-based models developed by Merton (1973) and Black and Scholes (1973), or some other statistical procedure provides the best predictive power.

The third issue considers which technique is most efficient for selecting among predictor variables. Using regression analysis alone, Drezner, Marcoulides and Salhi (1999) indicate seven techniques for selecting among

predictor variables: all possible regressions, forward selection, backward selection, stepwise selection, blockwise selection, maximum R^2 improvement and minimum R^2 improvement. Drezner et al. (1999) also demonstrate that the Tabu search procedure is more efficient than these seven (standard) regression techniques when selecting predictor variables.

In this paper we illustrate the efficiency of the Tabu search procedure as a method for selecting predictor variables for addressing problems within finance. The Tabu search model has the virtue of being independent of the first two issues. For instance, it can be readily used along with multiple regression, DA, logit, probit or simultaneous equations. The Tabu search model is also general because the distribution assumptions for the variables are minimal. Finally, any predictive system that is used in practice must rely on a finite subset of independent variables from among a larger set of variables. Since the Tabu selection process is most efficient at accomplishing this task, we recommend its use when constructing scoring systems in corporate, personal and real estate finance.

Whether implicit or explicit, the issue of selecting a subset of variables is important even in the corporate bankruptcy case. Mossman et al (1998) report that "little agreement exists regarding the best accounting ratios to determine likelihood of financial distress; [and that] more than 65 ratios [have been] used as predictors in previous literature." And Altman (1968) selects a final set of five variables which he argues do the "best overall job together in the prediction of corporate bankruptcy" from among an original list of 22 variables.

Section 2 reviews typical examples of predictive (scoring) models within finance in order to understand the appropriate use of the Tabu search selection procedure. Section 3 briefly reviews a specific version of the Tabu search model. This version is applied to some example data sets in Section 4 in order to illustrate its benefits in comparison to the traditional variable selection techniques. In Section 5, we apply a Tabu search procedure within the corporate bankruptcy setting. Section 6 concludes.

2. Bankruptcy Prediction and Mortgage Scoring Models

The literature about bankruptcy prediction and scoring models is extensive, but several key articles provide general overviews of the respective areas. Mossman et al (1998) examine a variety of corporate bankruptcy models. Domowitz and Sartain (1999) consider the consumer bankruptcy decision, though they do not select the variables which provide the highest predictive power. Asarnow and Edwards (1995) evaluate the success of predicting the expected loss on bank loans. Vandell (1993) reviews the mortgage default

research. And Mester (1997) surveys credit (card) scoring. In order to illustrate these systems in more detail, we comment briefly about these systems and then explain both Altman's Z-score and the Lehman Brothers default model as representative examples.

The information publicly available about prediction models varies dramatically. Altman's (1968) Z-score is published and used frequently in the corporate finance literature [see, for example, Mackie-Mason, 1990]. In contrast, Avery Bostic, Calem, and Canner (1996) report that "most credit history and application scoring systems are proprietary in nature, and the specific factors used and the risk weights assigned to these factors in establishing scores are not generally available to the public."

The number of factors considered and eventually used in a scoring system also varies. According to Mester (1997), some developers start with 50 or 60 variables in the development stage, but use only eight to 12 in the final scorecard. She adds that First Data Resources uses 48 factors in its final credit card scoring system.

Most of the business applications for these models are obvious. External stakeholders need to know the viability of corporations. Governmental regulations may be improved with more information about why individuals declare bankruptcy. Banks prosper when they are better able to predict which loans will default or become overdue. And mortgage and credit providers gain by distinguishing those borrowers who are likely to default on their loans or credit cards from those who are not. However, Avery et al (1996) explain five additional non-obvious uses, particularly for credit scoring, including: (1) monitoring the quality of portfolios, (2) evaluating the quality of mortgages for sale, (3) differentiating risk categories of loans for pricing decisions, (4) aiding the collection process, and (5) facilitating strategic planning decisions.

Altman's (1968) Z-score is one of the most well established models for predicting corporate bankruptcy. Even after 30 years it classifies the bankrupt from non-bankrupt firms during the year prior to bankruptcy better than the three leading competitors(Mossman, 1998). Altman considered a wide variety of financial ratios based upon firm-specific characteristics to arrive at the Z-score in equation (1):

$$\begin{aligned} Z - \text{Score} = \quad & 1.2 \times \frac{WC}{TA} + 1.4 \times \frac{RE}{TA} + \\ & 3.3 \times \frac{EBIT}{TA} + 0.6 \times \frac{MVE}{BVD} + 1.0 \times \frac{S}{TA} \end{aligned} \tag{1}$$

where WC is working capital (current assets minus current liabilities), RE is retained earnings, $EBIT$ is earnings before interest and taxes, MVE is

the market value of equity, S is sales, TA is total assets and BVD is the book value of debt. In this model, the *lower* the Z-score the higher the probability of bankruptcy. Firms with a Z-score lower than 1.81 have a high probability of failure within two years, while firms with Z-scores higher than 3.00 have a low failure rate.

Using the recent results of Mossman (1998), for instance, a sample of bankrupt firms from 1980 to 1991 has a mean of -0.028 for the working capital to total assets ratio, versus a mean of 0.314 for the non-bankrupt sample. The average Z-score for the bankrupt sample is 1.36, while the non-bankrupt mean is 4.12 . We adjust the Z-scores given in Mossman (1998) by 100 so that they are commensurate with our values as reported in Tables 3 and 4 below. The lower Z-score for the sample of bankrupt firms clearly indicates their higher probability of bankruptcy. The Z-score correctly classifies 84% of the firms in Mossman's sample.

As noted above, Altman (1968) initially compiles a list of 22 variables (financial ratios) to consider for inclusion in the final set of variables. He explicitly utilizes the following procedures to select his final list of five variables: "(1) observation of the statistical significance of various alternative functions including determination of the relative contributions of each independent variable; (2) evaluation of inter-correlations between the relevant variables; (3) observation of the predictive accuracy of the various profiles; and (4) judgment of the analyst." In the years following Altman's (1968) article, the selection techniques mentioned above became available, all of which attempt to reproduce procedures (1) - (3). It is likely, in other words, that a technique like the all-possible regressions selection technique would have produced an equation with even more predictive power than equation (1) above. And we argue that the Tabu search model adds at least incremental value to this process.

Pestre, Richardson and Webster (1992) explain the Lehman Brothers mortgage default model in detail. It is used to predict the probability of default of a pool of mortgages. They note that while mortgage delinquency ranges between 4.5% and 6% annually, loans entering default (foreclosure) have remained below 0.35% during the 1980s. A representative sample of about one million US home mortgages are used to construct the Lehman Brothers default model. The database includes fixed and dynamic characteristics, ranging from the details of each individual mortgage (e.g., zip code of residence, original loan-to-value ratio, purchase price, and date origination) to time series data (e.g., current loan balance, current delinquency status, the monthly unemployment series for each metropolitan statistical area (MSA), and the median house price by geographical area).

Table 1. Lehman Brothers Mortgage Default Model*

Risk Multipliers for Property and Loan Characteristics		
Variable	Subset of Variable	Risk Multiplier
Term	30-year	1.0
	15-year	0.8
Property Value	Average	1.0
	High; 3-5 times average	2.0
	Super High; 5+ times average	5.0
Occupancy	Owner	1.0
	Investor	1.7
Purpose	Purchase	1.0
	Refinance - no cash out	1.0
	Refinance - cash out	1.6
Property Type	Single family	1.0
	Condominium	1.0
Property Location	City Center	1.4
	Suburb	1.0

* Reproduced from Figure 15 of Pestre, Richardson and Webster (1992).

A logit technique is applied to the data to identify those characteristics which predict mortgage default. Pestre et al. (1992) note that the primary output of the model is a projected probability of default for each month a pool of mortgages is outstanding, with several equations roughly corresponding to Altman's Z-Score (equation (1) above).

The final determinants in the Lehman Brothers' model include four types of variables: (1) economic environment, (2) loan underwriting, (3) time and (4) property and loan characteristics, which include 12 variables altogether. While the number of variables initially considered for their model is not discussed, it is not difficult to imagine that up to 60 variables were candidates for inclusion in the final equation.

The model indicates that home price appreciation, an economic environment variable, is the most important in predicting mortgage foreclosure. For instance, Pestre et al (1992) state that "keeping all other loan characteristics constant, loans for which the underlying property value has fallen 20% since origination are projected to foreclose at 6.8 times the rate of loans for which property values are unchanged." In order to understand the procedure in more detail, consider Table 1, which presents the property and loan variables along with the corresponding risk multiplier (coefficient es-

timates in a regression analysis). The higher the risk multiplier, the more risky the mortgage loan. For example, properties with prices more than five times the median home price in the same MSA are five times more likely to foreclose than the average.

The Lehman Brothers model can readily be used by mortgage originating institutions both for monitoring their current portfolio of loans and for making the mortgage decision in the first place. This model is just one of several that currently exist, most of which are proprietary in nature. Naturally, institutions which rely upon these scoring systems continually update their data and search for new models which provide improved predictability. For each percentage of improvement in predictability, institutions may increase their profits. The practical use for improving these systems is apparent.

Researchers rarely explain their statistical procedures for narrowing down the number of variables. Altman's (1968) discussion of this procedure is an exception. Yet this process is crucial to the success of any given model. Improvements in the variable selection stage of the model-building process may increase the predictability of the model itself. We investigate how the Tabu search procedure allows for such improvement in the next section.

3. The Tabu Search Variable Selection Procedure

Econometric theory suggests that the variables in any prediction model (like regression modeling) be selected in accordance with the substantive theory under consideration. That is, researchers should set forth the regression model prior to data collection and statistical testing. To do otherwise, the argument goes, is to engage in "regression fishing." There is a long debate in econometrics over this issue, as discussed in Goldberger (1991) and more recently in *The Economist* (1998). However, testing various combinations of variables is appropriate and even necessary during the stage of theory building. Indeed, the point of constructing a theory is to explain regular patterns of behavior, and regression results are important for discerning regularities in behavior. This is the basic point of Arrow's (1951) comment that that "the choice between the alternative scientific tactics ... depends on the stage of formalization of the underlying theory. No dogmatism is possible..."

Two additional points reinforce the viability of searching for variables which provide strong predictive power. First, the practical use of such an approach is apparent in the uses considered herein, such as predicting bankruptcy and credit scoring. Second, while a researcher is extremely likely to obtain statistically significant results when running thousands of

regressions, additional statistical tests (e.g., out-of-sample testing) mitigate these problems. Simply, regression fishing is appropriate in some circumstances.

Of the seven general approaches mentioned by Drezner et al (1999) for selecting variables in a regression analysis, the all-possible regressions is clearly the most efficient (Berk, 1977). Its weakness, however, is that the number of all possible regressions with k independent variables equals 2^k. With just 40 variables under consideration, over one trillion different regressions must be run and the computational requirements are prohibitive. Even today's fast personal computers require a great deal of time to accomplish this task. For example, with an $n = 50$ and an initial set of variables of 26, the all-possible procedure requires over six hours versus only 4 seconds for the Tabu search procedure. For more details about the efficiency of the Tabu search procedure, see Drezner et al. (1999).

Commonly used search techniques (such as stepwise or max R^2 searches) examine the neighborhood of a set of variables and add or remove variables as long as the significance level of the neighboring set improves. The Tabu search procedure does not restrict the search to improving moves. The search may move to inferior solutions in the neighborhood of the current solution. As a result, a Tabu search allows the process to possibly exit local optimums when taking uphill moves. An addition of a certain variable may lead to a set with an inferior significance level, but additional changes of variables may lead the search to a set with a better significance level. To avoid cycling, Tabu search imposes a Tabu (prohibited) status to the parameters recently involved in the choice of the new solution.

The Tabu search procedure for multiple regression analysis requires the following seven definitions and parameters:

1. a criterion for the selection of independent variables,
2. a definition of the neighborhood,
3. a starting solution,
4. a definition of the Tabu list and its size,
5. a definition of an admissible subset,
6. search parameters, and
7. a stopping criterion.

Briefly consider each of these seven points. The *criterion* is the lowest significance level for the R^2 for the subset selected from among all possible

subsets. The *neighborhood* of a current solution is defined as all subsets of the current solution with one additional variable, one less variable, and those subsets for which one current variable is replaced by one not in the current solution set of variables. Given a total of k independent variables under consideration and $p < k$, a current (solution) subset of p independent variables has a neighborhood of size equal to $k - p$ possible additions, plus p possible removals, plus $p(k - p)$ possible replacements. The size of the neighborhood is thus $k + p(k - p)$ subsets. A *starting solution* is obtained by applying an algorithm which resembles the maximum R^2 improvement approach.

The *Tabu list* contains a list of variables which are not permitted to be used in a move. A move is adding, removing or replacing a variable. When a move is performed, the variable(s) in the move are added to the Tabu list. The *Tabu size* is set in advance by the researcher. For the experiments described below, the size is the larger of 10% of the neighborhood size and 10. When the length of the current Tabu list exceeds the predetermined size, the original member of the list is discarded in a FIFO (first in first out) manner. This means that we start with an empty Tabu list, and whenever a variable is involved in a move that is *not* better than the best known solution, it is put in the Tabu list and stays there until either a new better known solution is found (and the Tabu list emptied) or it becomes the "oldest" member in the list after the list reaches the pre-specified Tabu size. A move within the neighborhood is *admissible* if the variable(s) involved in the move are not on the Tabu list. The *stopping criterion* is that the search terminates when 30 consecutive Tabu search iterations do not produce a new best solution. The *search parameters* are the Tabu size and the number of consecutive iterations (set to 30 in our experiments) without an improvement which is used in the stopping criterion. Also, whenever a new best solution is found, the Tabu list is emptied. A complete description of the Tabu search procedure can be found in Drezner et al. (1999).

The flow of the Tabu search procedure we use is as follows:

1. An initial subset K of selected independent variables is generated.
2. The best current subset is $K_{\text{best}} = K$.
3. The iteration counter is set to $iter = 0$ (current iteration).
4. The neighborhood $N(K)$ of the subset K is created.
5. The significance levels $sig(K')\forall K' \in N(K)$ are evaluated.
6. If $sig(K') < sig(K_{\text{best}})$ for any $K' \in N(K)$, set $K_{\text{best}} = K'$. Go to step 8.

7. If for all $K' \in N(K)$: $sig(K') \geq sig(K_{\text{best}})$, choose the best admissible subset $K' \in N(K)$.
8. Set $K = K'$ and $iter = iter + 1$.
9. The Tabu list is updated. Go to Step 4 unless the stopping criterion is met.

4. Example Analyses

The Tabu search procedure was compared to two commonly-used regression selection procedures using simulated data: stepwise selection and the maximum R^2 improvement. The approach used in this paper to demonstrate the superiority of the Tabu search procedure is similar to that implemented by numerous other researchers: (1) utilize data for which there is a known correct model (i.e., an optimal set of predictor variables), and (2) determine whether the proposed procedure leads to the correct model (Coestner & Schoenberg, 1973; Herting & Costner, 1985). Particular attention was paid to ensure that the data sets used in the study were randomly generated with varying degrees of multicollinearity between the independent variables (because when there is no collinearity, selecting the best subset is a trivial matter). A total of ten different regression models were used to test the proposed procedure. The ten regression models consisted of models with the number of predictor variables ranging between 17 and 26. For all models examined, the sample size was set at 50 observations.

Table 2 compares the Tabu search results to the two commonly-used regression selection techniques. It is important to note that with the number of independent variables k ranging from 17 to 26, the problem have a maximum of over 67 million (226) possible solutions. As can be seen in Table 2, the Tabu search procedure provided the same set of optimal variables in all models examined. In contrast, the stepwise and maximum R^2 techniques selected the optimal solutions only 30% of the time. Just as importantly, these two common procedures selected different sets of variables as the solution set. For example, with k = 17, the optimal set of variables is (2, 6, 12, 17). The Tabu search procedure accurately selected this set. In contrast, the stepwise procedure selected the set (1, 4, 5, 7, 12, 13, 17) and the maximum R^2 selected the set (2, 5, 7, 12, 13, 17). It is important to note that variable #6 was not selected by the other procedures, even though it is in the optimal set, and that variables #5 and #13 were selected by the other procedures but are not in the optimal set.

Table 2. Comparison Between Tabu Search and Commonly-Used Variable Selection Procedures

Number[1] of Variables	Variables[2] in Optimal Solution	Tabu Procedure	Stepwise Procedure		Max R^2 Procedure	
			Include[3]	Exclude[4]	Include[3]	Exclude[4]
17	2, 6, 12, 17	identical[5]	1, 4, 5, 7, 13	2, 6	5, 7, 13	6
18	1, 6, 12, 13, 16, 17	identical	7	–	7	–
19	2, 6, 12, 13, 17	identical	1	–	8, 10, 15	–
20	1, 5, 7, 12, 13, 16, 17	identical	4	–	4, 8, 15	–
21	1, 6, 12, 13, 17, 18	identical	identical		identical	
22	1, 4, 6, 7, 9, 13, 17, 18, 22	identical	identical		identical	
23	1, 2, 3, 6, 12, 13, 17, 22	identical	4	–	4	–
24	1, 4, 5, 6, 9, 11, 12, 13, 16, 18, 22, 24	identical	identical		identical	
25	1, 3, 4, 6, 8, 10, 12, 13, 15, 17, 19, 22, 25	identical	5, 18, 23	3, 4, 8, 10, 15, 19	2, 7	1
26	1, 3, 6, 12, 13, 16, 17, 18, 20, 22, 24, 26	identical	5	–	8	–

[1]the number of variables each of the four selection procedures started with
[2]the set of variables in the optimal solution (the optimal set)
[3]the given procedure includes variables not in the optimal solution
[4]the given procedure excludes variables which are members of the optimal set
[5]'identical' indicates that the final set of variables is the same as the optimal set

5. Predicting Corporate Bankruptcy

In order to illustrate the potential benefits of using an efficient choice procedure, we compare Altman's (1968) Z-score method with the results of using a Tabu search procedure for selecting variables for predicting corporate bankruptcy. Note that we cannot replicate Altman's approach, for his original set of variables is unknown. Nonetheless, the comparison is appropriate, given the wide-spread use of the Z-Score model for predicting bankruptcy.

Sample firms are selected by identifying all bankrupt firms which meet our criteria along with a set of matching non-bankrupt firms, using the matching procedure detailed by Barber and Lyon (1997). Z-scores are calculated for each firm using equation (1) above. A Tabu search procedure is then used to select a subset of financial ratio variables which best predict bankruptcy from among a larger initial set of 20 variables, and use that subset of variables to calculate a new bankruptcy (Tabu prediction) score. Finally, we compare the success rate of the Tabu search results in predicting bankruptcy to the success rate from Altman's Z-score approach. The bankruptcy score based upon the Tabu search procedure predicts over 72% of the firms correctly, versus just under a 62% success rate for the Z-score.

Altman (1968) does not identify the original set of variables from which he selects his final five as represented in equation (1) above, but does indicate that he began with approximately 20 financial ratio variables. For our purposes, any well-accepted set of financial ratios may serve as the initial set of variables, from which the Tabu search procedure will select a subset as the best predictors of bankruptcy. The set of initial variables we use includes the original five in the Z-score, along with the remaining distinct (calculable) financial ratios listed in Table 3-2 of Brigham, Gapenski and Ehrhardt (1999). The definitions and descriptive statistics for our set of twenty initial variables are presented in Table 3 (the variable SIZE is used only for matching non-bankrupt to bankrupt firms).

5.1. Sample Selection

We select our sample firms using the following procedure. All relevant data is extracted for all bankrupt and liquidated firms from COMPUSTAT's most recent Research Annual file, which includes firms from 1978 to 1997. This time period is roughly twice that used by Mossman (1998), indicating that our data are not as homogeneous as his. Our time-period, along with other sample selection concerns, may account for the lower predictive power of the Z-Score and the Tabu Prediction Score as reported

Table 3. Variables Used in an Application of the Tabu Search Procedure to the Prediction of Corporate Bankruptcy

Panel A: Definitions of Initial Set of Twenty Variables and SIZE†

Variable	Definition
SIZE	Market value of equity = #Shares Outstanding x Price per Share
ALT1	Working Capital/Total Assets*
ALT2	Retained Earnings/Total Assets
ALT3	Basic Earning Power = EBIT/Total Assets
ALT4	Market Value of Equity/Book Value of Debt*
ALT5	Total Assets Turnover = Sales/Total Assets
CR	Current Ratio = Current Assets/Current Liabilities
QR	Quick Ratio = (Current Assets - Inventories)/Current Liabilities
INV X	Inventory Turnover = Sales/Inventories
DSO	Days Sales Outstanding = Receivables/(Annual Sales/360)
FAT	Fixed Assets Turnover = Sales/Net Fixed Assets
CAP REQ	Capital Requirement = Operating Capital/Sales
DEBT	Debt Ratio = Total Debt/Total Assets
TIE	Times-Interest-Earned = EBIT/Interest Charges
NOPAT	Net Operating Profit (Margin) After Taxes = EBIT(1-T)/Sales
PM	Profit Margin on Sales = EBIT/Sales
ROA	Return on Total Assets=Net Income Available to Shareholders/Total Assets*
ROE	Return on Equity = Net Income Available to Shareholders/Common Equity
PE	Price/Earnings Ratio = Price per Share/Earnings per Share*
CD OBL	Current Debt Obligation = Debt Due in One Year/SIZE
MB	Market-to-Book Ratio = Market Price per Share/Book Value per Share

†The variables ALT1-ALT5 denote the five Altman (1968) variables, while the remaining 15 variables are defined in Table 3-2 of Brigham, Gapenski and Ehrhardt (1999), with the exception of CD OBL, which is used as an alternative to TIE or the Debt Ratio. The SIZE variable is used only in matching non-bankrupt to bankrupt firms. Asterisks (*) indicate the variable was selected by the Tabu Search Procedure. The variable names are used in Panel B. The Z-Score is defined in Equation (1), while the Tabu Prediction Scores are defined in equations (2) and (3) in Section 5 above.

Panel B: Descriptive Statistics (see definitions in Panel A)

	Non-Bankrupt Firms					Bankrupt Firms				
Variable	Mean	Median	St Dev	MIN	MAX	Mean	Median	St Dev	MIN	MAX
Panel A: 46 Firms One Year Prior to Bankruptcy										
SIZE	44.56	12.75	80.96	0.89	433.43	24.23	7.82	54.59	0.19	348.06
ALT1	0.33	0.34	0.22	-0.30	0.89	0.16	0.26	0.37	-0.94	0.70
ALT2	0.05	0.22	0.57	-2.84	0.61	-0.12	0.16	0.78	-3.49	0.68
ALT3	0.07	0.09	0.17	-0.53	0.34	0.00	0.06	0.26	-0.99	0.35
ALT4	18.71	1.63	86.51	0.19	589.55	4.89	1.72	7.96	0.03	34.70
ALT5	1.68	1.49	1.30	0.48	8.67	1.93	1.77	1.49	0.00	8.33
CR	2.58	2.17	2.20	0.52	14.87	1.85	1.71	1.10	0.29	4.52
QR	1.46	1.07	2.06	0.39	14.30	1.05	0.88	0.83	0.09	4.04
INV X	11.15	5.53	19.91	1.46	91.91	10.84	5.29	18.78	0.03	118.15
DSO	60.23	55.18	35.00	10.22	191.66	153.49	49.00	682.94	4.75	4680.0
FAT	11.82	6.57	19.29	0.63	125.66	14.16	10.44	12.85	0.00	49.25
CAP REQ	0.51	0.44	0.31	0.06	1.78	4.78	0.36	26.99	0.00	183.37
DEBT	0.29	0.28	0.15	0.00	0.63	0.39	0.33	0.30	0.02	1.25
TIE	-76.77	2.52	579.14	-3916.3	137.00	5.03	1.80	16.88	-23.74	65.99
NOPAT	0.00	0.04	0.15	-0.80	0.18	-1.90	0.02	11.66	-78.89	0.48
PM	-0.01	0.03	0.15	-0.84	0.16	-2.20	0.01	12.90	-86.70	0.80
ROA	-0.01	0.04	0.15	-0.55	0.15	-0.06	0.02	0.29	-1.11	0.45
ROE	-0.15	0.07	0.99	-6.30	0.75	-0.61	0.05	2.29	-12.24	2.64
PE	76.84	8.32	394.14	-36.02	2680.2	5.84	3.78	29.19	-51.66	179.94
CD OBL	0.09	0.02	0.14	0.00	0.59	0.35	0.03	1.07	0.00	5.59
MB	2.80	1.24	4.08	0.32	20.74	1.47	1.11	6.81	-26.86	34.84
Z-Score	13.60	4.04	51.32	-0.26	355.63	4.89	3.86	5.65	-3.74	21.40
Tabu-Score	0.45	0.45	0.12	-0.08	0.68	0.55	0.50	0.17	0.30	1.07
Panel B: 139 Firms Two Years Prior to Bankruptcy										
SIZE	55.44	10.23	122.95	0.84	777.65	25.78	6.34	54.13	0.08	441.70
ALT1	0.23	0.26	0.29	-1.01	0.92	-0.06	0.08	0.58	-3.04	0.71
ALT2	-0.07	0.10	0.68	-4.45	0.73	-0.59	-0.15	1.82	-18.05	0.77
ALT3	0.05	0.07	0.18	-1.03	0.36	-0.16	-0.04	0.45	-2.41	0.31
ALT4	81.97	2.02	394.35	0.01	3838.62	4.76	1.16	13.02	0.00	136.72
ALT5	1.66	1.45	1.11	0.08	8.52	1.9	1.35	1.33	0.02	7.86
CR	2.59	1.78	5.11	0.08	56.73	1.55	1.15	1.81	0.05	18.22
QR	1.78	1.06	4.60	0.07	51.07	1.02	0.69	1.63	0.03	17.31
INV X	18.87	6.98	32.31	0.85	251.13	25.13	7.06	98.83	0.74	1141.50
DSO	55.22	52.30	38.26	3.06	271.74	6.49	50.14	64.68	0.00	415.08
FAT	10.52	5.55	20.64	0.30	211.36	9.54	5.83	12.72	0.06	95.38
CAP REQ	0.55	0.43	0.53	0.03	3.42	0.75	0.37	2.20	-1.35	23.65
DEBT	0.30	0.26	0.21	0.00	1.04	0.43	0.37	0.31	0.01	2.39
TIE	11.53	1.87	92.44	-645.00	726.75	-4.91	-0.70	28.85	-171.30	84.52
NOPAT	-0.05	0.03	0.77	-8.96	0.51	-0.20	-0.02	0.58	-4.01	1.41
PM	-0.07	0.02	0.77	-8.97	0.50	-0.27	-0.05	0.63	-4.37	1.22
ROA	-0.01	0.03	0.17	-1.03	0.27	-0.24	-0.10	0.45	-2.67	0.19
ROE	-0.15	0.06	0.80	-4.51	1.20	-2.59	-0.52	13.79	-161.50	0.38
PE	12.03	6.52	87.03	-538.67	681.25	-8.37	-0.33	204.81	-2365.3	356.57
CD OBL	0.37	0.03	2.96	0.00	34.95	1.54	0.05	10.92	0.00	126.76
MB	2.15	1.28	4.48	-.27	47.99	4.10	0.99	20.85	-34.44	229.91
Z-Score	51.18	3.85	237.49	-2.63	2,306.33	3.11	2.42	8.69	-32.02	81.85
Tabu-Score	0.41	0.40	0.14	-0.35	0.91	0.59	0.54	0.23	0.30	1.50

below. Relevant data includes all COMPUSTAT variables necessary to calculate the variables listed in Panel A of Table 3, along with SIC codes, dates and price per share and number of shares outstanding (COMPUSTAT variables #24 and # 25) which are used to calculate firm size for the purpose of selecting the matching non-bankrupt firms from the corresponding COMPUSTAT Industrial Annual File. The Research Annual file lists the date of bankruptcy or liquidation, hereafter referred to as the event year. Any firm which has complete data for the first year (event year -1), the second year (event year -2) or the first two years just prior to the event date remains in the sample. Naturally, with firms that eventually end up in bankruptcy, a large amount of data is missing from the years prior to bankruptcy. In addition, following Mossman (1998) we exclude financial firms.

From 1,002 firms listed as bankrupt or liquidated in the Research Annual file, 155 firms remain in the final sample, 13 firms with data for event year -1 only, 106 with data for event year -2 only, and 33 firms with data for both event years -1 and -2. If each firm-year is counted as an observation (or "separate" firm), the total number of observations for the bankrupt firms is 185, or 46 observations for event year -1 and 139 observations for event year -2 as indicated in the descriptive statistics in Table 3. The 155 bankrupt firms in the sample are then matched to non- bankrupt firms, yielding a total of 370 firms in the sample. Since our aim is to illustrate how the Tabu search procedure can be used to improve predictability, and not directly to suggest that our final set of bankruptcy prediction variables is a "perfect" set for predicting corporate bankruptcy, we do not focus on some of the data issues that would be relevant if we were to provide a full justification for this set of variables. Indeed, as discussed in Section II above, theoretical concerns are also relevant in the final selection of variables.

Variables are defined in Table 3, Panel A. Descriptive statistics for the SIZE variable, and the 20 initial variables used in the Tabu search procedure, and the resulting Z-Score and Tabu-Prediction Score are reported in Table 3, Panel B. We test for the impact of outliers on the outcome by using Cook's D, with the result that even though some outliers appear extreme, they have no statistically significant affect on the regression used to calculate the Tabu-generated bankruptcy scores. Cook's D is a useful overall measure of the impact of the i^{th} observation on all of the estimated regression coefficients, and is calculated by statistics packages such as SPSS. If a given observation has an abnormally large influence on a regression coefficient as measured by Cook's D, it is appropriate to omit that observation as an outlier [see Neter, Wasserman and Kutner (1989) for the appropriate calculation procedures]. Most importantly, some of the

Z- Scores are extremely high. It turns out that all Z-Scores greater than 100 correctly predict that the firms do not enter bankruptcy.

We know from Barber and Lyon (1997) that firm size as measured by the market value of the firm's equity biases cumulative abnormal returns used in event studies. This is why we match our firms by SIZE. Nonetheless, as clearly seen in Panel B of Table 3, non-bankrupt firms are significantly larger than the bankrupt firms to which they are matched. This occurs largely because we match firms as carefully as possible within a four-digit SIC code and the fact that bankrupt firms tend to be smaller (and younger) than the other firms in their industries. We use SIC codes for matching, rather than other variables like market-to-book, because of the impact of economy wide influences on whole industries. Future studies may benefit by examining the significance of firm size on predicting bankruptcy.

We use a Tabu search procedure to select the subset of variables from among the original 20 which best predict bankruptcy. This procedure is applied twice, once for the whole sample and once for the t-2 event year observations only. The second application is appropriate both because there are sufficient observations ($n = 278$) and because the uniformity of using data for one year only should, and does, increase predictability as a result of a higher degree of homogeneity among the sample. The dependent variable is a dummy variable, with 0 = a non-bankrupt firm and 1 = a bankrupt firm. Consider the two applications separately. For our first application of the Tabu search procedure as applied to the whole sample of 370 observations, the following two variables are selected as the best predictors:

1. Altman's 1^{st} variable, the ratio of working capital to total assets (ALT1), and

2. the firm's return on assets (ROA).

We run an OLS regression, using these two variables as the independent variables and the dummy (bankrupt/non-bankrupt) variable as the dependent variable, to generate a bankruptcy prediction score. The regression equation is the Tabu-generated prediction score:

$$\text{Tabu Prediction Score} = 0.500 - 0.217\mathbf{ALT1} - 0.278\mathbf{ROA} \qquad (2)$$

Using equation (2), we calculate a Tabu Prediction Score and the Z-Score according to equation (1) for each observation. To determine whether statistical differences exist between the bankrupt and non-bankrupt matched firms, we report the results of a paired two-sample t-test of the difference in means for both the Z-Score and the Tabu Prediction Score in Table

Table 4. Comparison of Means of Bankrupt versus Non-Bankrupt Tabu-Generated Scores and Z-Scores†

Panel A: Complete Sample of 185 Non-Bankrupt and 185 Bankrupt Firms Paired by SIC and SIZE

	Tabu-Prediction Score		Z-Score	
	Non-Bankrupt Firms	Bankrupt Firms	Non-Bankrupt Firms	Bankrupt Firms
Mean	0.438	0.562	41.835	3.553
Variance	0.015	0.039	43,225.283	65.260
N	185	185	185	185
Pearson Correlation	0.215		0.082	
t-statistic	-8.010		2.511	
P($T \leq t$) two-tail	0.000		0.013	

Panel B: t-2 Sample of 139 Non-Bankrupt and 139 Bankrupt Firms Paired by SIC and SIZE

	Tabu-Prediction Score		Z-Score	
	Non-Bankrupt Firms	Bankrupt Firms	Non-Bankrupt Firms	Bankrupt Firms
Mean	0.412	0.588	51.178	3.110
Variance	0.020	0.054	56,402.160	75.581
N	139	139	139	139
Pearson Correlation	0.205		0.104	
t-statistic	-8.506		2.394	
P($T \leq t$) two-tail	0.000		0.018	

†Paired Two Sample t-Test, with hypothesized mean difference = 0

4, Panel A. Both the Z-Scores and the Tabu Prediction Scores are significantly different across the bankrupt and non-bankrupt samples, though the difference in means is significant beyond the 0.01 level for the Tabu Prediction Score and only at the 0.05 level for the Z-Score.

The next step is to calculate the predictive success of both approaches. We conduct sensitivity analyses of the Z-Score by using values between 1.81 and 3 for the Z-Score, as indicated in Section 2 above, and find that 1.81 provides the optimal predictive power. This value of 1.81 is thus the cutoff point for the Z-Score and the optimal value for the Tabu Prediction Score. The success rates in predicting bankruptcy/non-bankruptcy are reported in Table 5, Panel A. For the whole sample, the Tabu Prediction Score correctly predicts 66.5% of the cases, while the Z-Score predicts only 60.5% of the cases correctly.

Table 5. Predictive Power of the a Tabu-Generated Prediction Score in Comparison to the Z-Score

Panel A: Complete Sample of 185 Bankrupt and 185 Non-Bankrupt Firms

	Z-Score		Tabu Prediction Score	
	# of Firms	Success Rate	# of Firms	Success Rate
Total	370		370	
Non-Bankrupt Firms				
Predicted Correctly	154	83.2%	167	90.3%
Bankrupt Firms				
Predicted Correctly	70	37.8%	79	42.7%
Total Correct Predictions	224	60.5%	246	66.5%

Panel B: Sample of 139 Bankrupt and 139 Non-Bankrupt Firms for the t-2 Event Year

	Z-Score		Tabu Prediction Score	
	# of Firms	Success Rate	# of Firms	Success Rate
Total	278		278	
Non-Bankrupt Firms				
Predicted Correctly	115	82.7%	128	92.1%
Bankrupt Firms				
Predicted Correctly	57	41.0%	73	52.5%
Total Correct Predictions	172	61.9%	201	72.3%

To determine whether the Tabu Selection procedure can outperform the Z-Score model even more, consider a more homogeneous sample of the t-2 observations only, the second application of the Tabu search procedure. For this second procedure, the Tabu model selects the following variables:

1. Altman's 1^{st} variable, the ratio of working capital to total assets (ALT1),
2. Altman's 2^{nd} variable, retained earnings scaled by total assets (ALT2),
3. Altman's 4^{th} variable, the market value of equity scaled by the book value of debt (ALT4),
4. days sales outstanding for inventory, DSO, and
5. the firm's return on assets, ROA.

The resulting regression equation is:

$$\begin{aligned} \text{Tabu Prediction Score} &= 0.401 - 0.251\mathbf{ALT1} + 0.055\mathbf{ALT2} \\ &- 0.0002\mathbf{ALT4} + 0.0016\mathbf{DSO} - 0.446\mathbf{ROA} \end{aligned} \tag{3}$$

The t-tests are reported in Table 4, Panel B. Again, the differences in means across the bankrupt and non-bankrupt samples for the Z-Score and the Tabu Prediction Score are statistically significant, with the Tabu Prediction Score significant at a higher level.

Finally, the success rates in predicting bankruptcy/non-bankruptcy are reported in Table 5, Panel B. For the whole sample, the Tabu Prediction Score correctly predicts 72.3% of the cases, while the Z-Score predicts only 61.9% of the cases correctly. This is a 10.4 percentage point difference between the two approaches, a difference well worth considering whether attempting to predict corporate or personal bankruptcy, mortgage or credit scores, or other related financial attributes.

6. Conclusion

This paper illustrates how the Tabu search procedure can be applied efficiently to problems in finance. The Tabu search was compared to two commonly-used regression selection procedures. The results indicate the superiority of the Tabu procedure over commonly used selection procedures. It is superior in two pertinent respects, because it selects the optimal set of explanatory variables: (1) more frequently than the other procedures, and (2) more efficiently. A Tabu search procedure can be applied readily to problems in finance, including both the selection of variables for the APT and the selection of variables for scoring and bankruptcy models in corporate, personal and real estate finance. We apply the Tabu procedure to the standard corporate bankruptcy prediction problem, and compare the success in predicting bankruptcy/non-bankruptcy to the predictive success when using the Z-Score. Results indicate that a prediction score, based upon the output variables of the Tabu search procedure, yields a 72.3% prediction success rate in comparison to only a 61.9% success rate for the Z-score.

References

1. Altman, E. (1968) "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy," *Journal of Finance*, Vol. 23, 589-609.
2. Arrow, K. J. (1951) "Mathematical Models in the Social Sciences," reprinted in Brodbeck, May, *Readings in the Philosophy of Social Sciences*, Macmillan Publishing Co., New York, 1968, 635-667.
3. Asarnow, E. and D. Edwards (1995) "Measuring Loss on Defaulted Bank Loans: 24-Year Study," *The Journal of Commercial Lending*, March, 11-23.

4. Avery, R. B., R. W. Bostic, P. S. Calem and G. B. Canner (1996) "Credit Risk, Credit Scoring, and the Performance of Home Mortgages," *Federal Reserve Bulletin*, July, 621-648.
5. Barber, B. M. and J. D. Lyon (1997) "Detecting Long-Run Abnormal Stock Returns: The Empirical Power and Specification of Test Statistics," *Journal of Financial Economics*, Vol. 43, 341-372.
6. Black, F. and M. Scholes (1973) "The Pricing of Options and Corporate Liabilities," *Journal of Political Economy*, Vol. 81, 637-59.
7. Brigham, E. F., L. C. Gapenski and M. C. Ehrhardt (1999) *Financial Management - Theory and Practice*, Ninth Edition, The Dryden Press, New York.
8. Costner, H. L. and R. Schoenberg (1973) "Diagnostic Indicator Ills in Multiple Indicator Models," in A. S. Goldberger & O.D. Duncan (Eds.), *Structural Equation Models in the Social Sciences*, New York, Seminar Press, 167-199.
9. Domowitz, I. and R. L. Sartain (1999) "Determinants of the Consumer Bankruptcy Decision," *Journal of Finance*, Vol. 54, 403-420.
10. Drezner, Z., G. A. Marcoulides and S. Salhi (1999) "Tabu Search Model Selection in Multiple Regression Analysis," *Communications of Statistics*, Vol. 28, 2, 349-367.
11. *The Economist* (1998) "The Perils of Prediction," August 1, 61-62.
12. Goldberger, A. S. (1991) *A Course in Econometrics*, Harvard University Press.
13. Herting, J. R. & H. L. Costner (1985) "Respecification in Multiple Indicator Models," in H.M. Blalock, Jr. (Ed.), *Causal Models in Social Sciences* (2nd ed.), Hawthorne, NY, Aldine, 321-393.
14. MacKie-Mason, J. K. "Do Taxes Affect Corporate Financing Decisions?" *Journal of Finance*, 45 (1990), 1471-1493.
15. Mester, L. J. (1997) "What's the Point of Credit Scoring?" *Business Review* (Federal Reserve Bank of Philadelphia), September/October, 3-16.
16. Merton, R. C. (1973) "Theory of Rational Option Pricing," *Bell Journal of Economics and Management Science*, Vol. 4, Spring, 141-83.
17. Mossman, C. E., G. G. Bell, L. M. Swartz and H. Turtle (1998) "An Empirical Comparison of Bankruptcy Models," *The Financial Review*, Vol. 33, 35-54.
18. Neter, J., William Wasserman and Michael H. Kutner (1989) *Applied Linear Regression Models*, Second Edition, Irwin.
19. Pestre, C., P. Richardson and C. Webster (1992) "The Lehman Brothers Mortgage Default Model and Credit-Adjusted Spread Framework," *Fixed Income Research*, The Lehman Brothers.
20. Roll, R. (1988) "R^2," *Journal of Finance*, Vol. 43, July, 541-566.
21. Shilton, L. and J. Teall (1994) "Option-Based Prediction of Commercial Mortgage Defaults," *The Journal of Real Estate Research*, Vol. 9, 219-236.
22. Vandell, K. D. (1993) "Handing Over the Keys: A Perspective on Mortgage Default Research," *Journal of the American Real Estate and Urban Economics Association*, Vol. 21, 211-246.