2010

Could decision trees help improve Farm Service Agency lending decisions?

Benjamin P. Foster
University of Louisville

Jozef Zurada
University of Louisville

Douglas K. Barney
Indiana University Southeast

Original Publication Information

Foster, Benjamin P., Jozef Zurada, and Douglas K. Barney. "Could Decision Trees Help Improve Farm Service Agency Lending Decisions?" 2010. *Journal of Management Information and Decision Sciences* (formerly *Academy of Information and Management Sciences Journal)* 13(1): 69-91.

COULD DECISION TREES HELP IMPROVE FARM SERVICE AGENCY LENDING DECISIONS?

Benjamin P. Foster, University of Louisville
Jozef Zurada, University of Louisville
Douglas K. Barney, Indiana University Southeast

ABSTRACT

This study examines whether a statistically derived decision tree could serve as a means to improve U.S.A. Farm Service Agency lending decisions. The study is a substantial extension and reanalysis of an earlier work by Barney, Graves and Johnson, (1999). Results indicate that a decision tree could be a valuable tool for Farm Service Agency employees in their lending decisions. The decision tree provides as good or better predictive accuracy than neural networks and logistic regression models at reasonable cutoff levels of Type II to Type I costs of lending. The decision tree also meets the transparency criteria for Farm Service Agency purposes by providing logical, understandable rules for lending decisions.

INTRODUCTION

The Farm Service Agency (FSA) directly loans, or guarantees loans to farmers totaling billions of dollars. The need for an understandable, accurate decision tool to assist FSA employees in their lending decisions is as great today as in the past. This article describes a substantial extension and reanalysis of an earlier work by Barney, Graves and Johnson (1999) examining Farmers Home Administration (FmHA) (predecessor of the FSA) lending decisions. Also, see Barney (1993) for a full description of the background and analysis. We do not recommend a loan classification system for immediate FSA use. Rather, we test whether a decision tree could potentially improve FSA lending practices, make lending decisions more transparent and be easily understood by applicants and the FSA staff. This study extends the earlier work by examining additional logistic regression models and neural networks and by investigating whether a decision tree could improve FSA lending decisions. This new analysis indicates that a decision tree could aid FSA employees in their lending decisions. The decision tree provides as good or better predictive accuracy than other methods, and provides logical, understandable rules for lending decisions.

Section 2 ties this study to the prior Barney, Graves and Johnson (1999) study, briefly reviews FSA lending, and summarizes relevant literature. Then, research methods are described in Section 3, followed by discussion of results in Section 4 and conclusions in Section 5.

LITERATURE REVIEW

Relationship of This Study to Barney, Graves and Johnson (1999)

The authors of the 1999 study used the newest methodology of that time (i.e. neural networks) to develop a model for FmHA use. This study investigates whether a better, possibly more accurate, fully transparent and interpretable methodology could now be applied by the FSA. This work extends the earlier work of Barney, Graves and Johnson (1999) by comparing a data mining technique, the decision tree, with the methodologies used in the 1999 study. Also, different logistic regression models and neural networks than those used in the 1999 study are developed.

Two factors are central to a technique's usefulness for the FSA: (1) ability to clearly and accurately categorize potential farm borrowers between those who will make scheduled debt payments and those who will not make timely debt payments, and (2) transparency and understandability to borrowers and FSA employees. The FSA is subject to the provisions of the Equal Credit Opportunity Act (1975) and therefore must be able to provide a clear explanation to borrower applicants when the FSA denies them a loan. The 1999 study found that the neural network produced predictive accuracy superior to criteria developed internally by the FmHA (FSA), criteria developed by Price Waterhouse, logistic regression and ordinary least-squares regression models. Even so, operation of the neural network model was not transparent to FSA employees and borrowers.

A neural network tends to work as a "black box" which would render lending decisions less subject to manipulation by loan applicants. However, that aspect of neural networks would make justifying a loan denial more difficult because FSA employees could not point to particular criteria as reasons for the denial. A decision tree may well serve as a lending decision tool as accurate as a neural network, but with the transparency of more traditional models and less subject to manipulation than the FSA model.

Also, the Barney, Graves and Johnson (1999) study concentrated entirely on two techniques: logistic regression and neural networks. In both methods they used all 14 input variables for building the models and testing their classification accuracy rates. The decision tree techniques and stepwise linear regression used in this study are classification and variable reduction techniques at the same time. Our best model, the chi-square decision tree, identified only four variables as relevant in predicting future loan payments, and pruned the remaining ten variables. Similarly, the stepwise linear regression method identified only three variables (out of 14) as significant. Because Barney, Graves and Johnson (1999) included all variables in his analyses, he developed a large neural network with a dozen neurons in the hidden layer. Such a large network can cause overtraining, i.e., memorizing the training patterns to produce almost perfect classification results on the training set, but less desirable performance on the test set. In this study, we used a small neural network with 2 neurons in the hidden layer to prevent overtraining.

Farm Service Agency Lending

What was once the Farmers Home Administration (FmHA) was merged into the Farm Service Agency (FSA), along with several other federal agencies, in 1995. While the name of the government entity changed, its function, at that time, remained basically unaltered (Farm Service Agency, 2006). Today, as in the early 1990s, the FSA is a lender of last resort for farmers. This means that the FSA will lend to individuals who are unable to obtain funding at reasonable terms from a commercial lender, (i.e. commercially risky borrowers).

Because the FSA is the "lender of last resort" it would expect higher default rates than commercial lenders. For example, the default rate was approximately 27.8% for loans from the early 1990s examined in this study. In contrast, general farm-level data from the Illinois Farm Business Farm Management Association from 1995 to 2002 contained a default rate of 0.567% (Katchova & Barry, 2005). Also, the Seventh Farm Credit District (Arkansas, Illinois, Indiana, Kentucky, Michigan, Minnesota, Missouri, North Dakota, Ohio, Tennessee, and Wisconsin) total loan accounting data base for 2001 contained a total default percentage of 1.83% (Featherstone, Roessler & Barry, 2006).

In comparison, according to Anne Steppe, a loan officer with the FSA, the FSA's direct loan default rate was 10.55% at the end of September 2008 and 16.1% at the end of April 2009 (per email communication on October 16, 2008 and phone conversation May 18, 2009). While this rate is certainly lower than the default rate in Barney's study, the rate is higher than that for other agricultural lenders, as would be expected from the lender of last resort. In addition, the FSA has experienced increased demand for its farm loans as a result of the 2008 lending/financial market crisis (per phone conversation with Tracy Jones, FSA Senior Loan Officer, Washington DC, May 13, 2009).

The FSA has two major farm borrowing plans. Originally, the FSA mission was to directly lend money to farm borrowers. More recently, the FSA has attempted to reduce its direct loan program and focus its activities more on guaranteeing loans made to farmers by commercial banks. Under the guaranteed loan program farmers start the loan process by requesting a loan from a commercial lender. If the commercial lender sees the loan as borderline, the lender then approaches the FSA about guaranteeing the loan. The FSA program will guarantee up to 95% of a farm loan.

The FSA has clearly moved away from making direct loans and emphasizes its guaranteed loan program. For example, at December 31, 1990 (shortly before data collection for the 1999 study) the FmHA held approximately $17 billion in direct loan debt, approximately 13% of all outstanding farm debt. At December 31, 2007, the FSA held approximately $5 billion in direct loan debt, approximately 2.3% of all outstanding farm debt. (Amounts calculated from information at http://www.ers.usda.gov/Data/FarmBalanceSheet/fbsdmu.htm.) Consequently, the relative overall importance of the FSA in direct agricultural lending has declined. However, the FSA continues to guarantee much outstanding farm debt.

> *From fiscal 2000 to 2004, 98,000 unique farmers and ranchers received 137,000 FSA direct and guaranteed loans totaling $16.3 billion. Direct programs accounted for only about one-fourth of all dollars obligated, but because of their lower average loan size accounted for half of all borrowers served. (Farm Service Agency, 2006, p. 25)*

The decision to guarantee a loan should require diligence by FSA employees similar to that expended in evaluating a direct loan. Thus, finding an adequate decision criteria/tool may be as important today as in the early 1990s.

Lending Criteria

Despite changes in the focus of FSA lending, discussed above, the process of direct lending at the FSA has undergone only minor changes since the original data was collected in the early 1990s. The FSA (FmHA) for decades used the same, primarily unaltered, form to collect farm financial data. This form, the FHP, provided some current balance sheet and projected income statement information. In 2005, new forms replaced the FHP nationwide. The FSA now uses the information on these two forms (FSA 2037 and FSA 2038) to develop the Farm Business Plan. The Farm Business Plan is very similar in content to the Farm and Home Plan, which it replaced. Both required considerable information about expected production operations (e.g. acres of corn, number of cows), revenues and expenses. To verify the reasonableness of the expense estimates on the Farm Business Plan, the FSA now also expects the borrower to provide up to five years of tax returns, if available. Lack of tax return data to support the expense estimates does not disqualify a borrower from receiving a loan and the tax returns are not otherwise used in the lending decision.

At the time of the Barney, Graves and Johnson (1999) study, the FmHA lending decision process was based on one number (a score for projected repayment ability) developed from actual and projected financial statements. Because projected repayment ability was based entirely on projected data, it was highly subject to manipulation. The FSA still uses only one number to make the loan decision, the Margin After Debt Service (MADS). This number is calculated in essentially the same manner as projected repayment ability. MADS is calculated by subtracting all projected operating and living expenses and next year's principal and interest payments from projected total farm income.

In the past, the FSA tried to change both the financial statements required of borrowers and the criteria used in the lending decision. In the late 1980s the FmHA attempted to switch to GAAP-based farm financial statements. Negative feedback from farmers (and from some FmHA employees) was so harsh that Congress passed a law forbidding the FmHA to use those statements further.

Also in the 1980s, the FmHA engaged Price Waterhouse to develop a lending model. After considerable time and expense, Price Waterhouse developed several credit screens, for different

types of loans. In addition, for several years the FmHA tested and used internally (not for making or denying loans, but solely for evaluation purposes) a four ratio evaluation model somewhat similar to the Price Waterhouse model. The FmHA never used the Price Waterhouse or internally developed models in its lending decisions.

Despite not adopting either the Price Waterhouse screening tool or its own internally generated model, the FSA evaluated these methods based on the FSA's two primary criteria: discriminatory power to separate borrowers who will repay FSA debt from those who will not, and transparency. Transparency, in essence, means that the decision criteria are understandable by both potential borrowers and the FSA local staff. Thus, the method used should provide clearly identified criteria for why a borrower received or was denied a loan.

Decision Trees as a Possible Improvement

Barney, Graves and Johnson (1999) examined the accuracy of different techniques/models at predicting whether farm borrowers would make farm loan payments as scheduled one year hence, based on data from the FHP and the past two years of repayment history. They found that a neural network could predict loan repayment (based on model accuracy measured in Type I, Type II, and total errors) better than the internally developed FmHA, Price Waterhouse, logistic regression, and ordinary-least-squares regression models.

Classification/predictive ability is an important criterion for any technique/model used. The previous discussion indicates that understandability of the loan decision process is also important to the FSA. Research with publicly traded companies has noted the same issue. Consequently, decision trees may be appealing because they produce easily interpretable results which could be understood by participants in the FSA lending process. For example, data mining literature specifically endorsed decision trees as an analytical method to generate easily understood and explained decisions in the form of if-then rules (Berry & Linoff, 1997; Kantardzic, 2003). Decision trees offer other advantages over alternative predictive methods, including that they do not require an excessive amount of computation, and unlike neural networks, easily identify the most important predictive variables (Berry & Linoff, 1997). If decision trees can be effective in predicting repayment or default on loans, they may be useful tools to help the FSA evaluate the ability of farmers to repay loans.

To attempt to find the best predictive techniques, prior research with public companies has compared several different methods, including decision trees. During the financial crisis of the late 1990s, critics of South Korean financial institutions' loan decisions believed that those decisions themselves determined whether a company survived or entered bankruptcy (Kyung, Chang & Lee, 1999). According to Kyung, Chang and Lee (1999), financial institutions' reliance on arbitrary judgment or a complicated statistical method would not satisfy business and political leaders who would prefer to hear well-defined, understandable decision rules for lending decisions. Consequently, they evaluated the predictive ability of a decision tree for data from corporations

listed on the Korea Stock Exchange. They concluded that the decision tree performed well, with substantially higher predictive accuracy rates than a multiple discriminant model under crisis conditions and slightly higher predictive accuracy under normal conditions.

Koh (2004) compared the ability of a logistic regression model, a neural network, and a decision tree to accurately classify 165 U.S. companies that became bankrupt from 1980 to 1987 and 165 matching U.S. companies. Similar to Kyung, Chang and Lee (1999), Koh (2004) observed better overall classification rates produced by the decision tree than the logistic regression model or neural network. Consequently, research in the corporate setting indicates that the decision tree technique may provide a viable alternative tool for loan screening by the FSA.

METHODS

Data Collection and Variables

The data set used in Barney (1993) and Barney, Graves and Johnson (1999) was collected from FSA employees (FmHA loan officers) randomly across the United States. Loan officers provided anonymous (borrower personal information was deleted) copies of FHPs. The data set and variables used in this study are the same as were used in the 1999 study. (See Barney, Graves & Johnson, 1999; Barney, 1993 for a more complete explanation of the variables and the data collection process used.)

The FHPs included financial operating results for 1990 and balance sheet balances at 1 January 1991. (Variables are defined in Table 1.) Whether the related borrowers made scheduled debt payments on 1 January 1992 was also noted by the loan officers. Lending officers reported a total of 261 observations. These observations were randomly divided into 196 training set observations and 65 test set observations. After eliminating 17 observations with incomplete data, the training set contained 184 observations (130, 70.7% repayments and 54, 29.3% defaults) and the test set contained 60 observations (46, 76.7% repayments and 14, 23.3% defaults).

Table 1[a] Prediction model variables[b]

Dependent Variable:	FmHA loan payment on 1 January, 1992 (PAY92) = 0 if missed, 1 if made
Independent Variables:	
Current Ratio (CR)	= 1991 Total current farm assets / 1991 Total current farm liabilities
Working Capital (WC)	= 1991 Total current farm assets - 1991 total current farm liabilities
Debt-to-Assets (DEBT/ASSETS)	= 1991 Total debts / 1991 Total assets

Table 1[a] Prediction model variables[b]	
Debt-to-Equity (DEBT/EQUITY)	= 1991 Total debts. / 1991 Total assets - 1991 Total debt + 400,000
Return on Farm Assets (RFA90)	= 1990 Total cash farm income from operations - operating expenses - family living expenses / 1990 Beginning total farm assets
Return on Equity (RRE90)	= 1990 Total cash farm income - operating expenses - interest expense - family living expenses / 1990 Total assets - 1990 Total debt + 400,000
Operating Proft Margin (OPM90)	= 1990 Total farm income - actual operating expenses - family living expenses / 1990 Total farm income
Projected Debt Repayment ratio (PDR91)	= Total debt and interest payments due on 1991 FHP / 1991 Projected total cash farm income + Non-farm income
Debt Repayment Ratio (DR90)	= Total debt and interest payments due on 1990 FHP / 1990 Total cash farm income + Non-farm income
Asset Turnover (AT90)	= 1990 Total cash farm income / 1990 Beginning total farm assets
Operating Expense (OE90)	= 1990 Total operating expenses[c] / 1990 Total farm income
Interest Expense (IE90)	= Total 1990 actual interest expense paid / Total 1990 farm income
Dummy Variable (REST90)	= 0 if restructured on 1 January, 1990; 1 otherwise
Dummy Variable (REST91)	= 0 if restructured on 1 January, 1991; 1 otherwise

a From Table 1 of (Barney, Graves, & Johnson, 1999)
b Unless stated otherwise, all ratios are calculated after restructuring and new loans.
c Unless stated otherwise, operating expenses do not include interest expense.

Analytical Methods

Logistic regression models, neural networks, and decision trees were used to analyze the data. A more detailed description of decision trees than the other techniques follows because use of the decision tree technique is the main extension provided by this study. Because many research studies involving use of categorical dependent variables have used logistic regression and neural networks, readers may see Press and Wilson (1978, Hosmer and Lemeshow (1989) for a complete description of logistic regression, and Hagan, Demuth and Beale (1996), Han and Kamber (2001), Giudici (2003), Kantardzic (2003) and SAS Enterprise Miner at http://www.sas.com) for a detailed and theoretical description of neural networks.

Logistic Regression

We will only briefly discuss logistic regression because many previous research studies with categorical dependent variables have used logistic regression. Logistic regression is included in

several statistical packages. We performed analysis using the Statistical Analysis System (SAS) which uses an iteratively reweighted least squares algorithm to compute maximum likelihood estimates of the regression parameters (SAS Institute, Inc. 1999). SAS uses the following model to classify farmers into the missed payment or made payment categories:

$$g(Y) = \ln [P(PAY92=0 \mid x) / P(PAY92 =1 \mid x)] = \beta_0 + \Sigma\beta_i x_i + \varepsilon \quad (1)$$

where: PAY92 = 0 if the farmer missed payment due January 1, 1992; and
1 if the farmer made payment due January 1, 1992.

The independent variables included in the analysis are denoted with the general expression, x.

Neural Networks

Popular data mining tools include neural networks. Neural networks have been used in a variety of business applications. Neural networks are simple computer programs that build mathematical models of the connections in the human brain by trial and error during data analysis. The computational property, the architecture of the network, and the learning property characterize neural network models (Hagan, Demuth & Beale, 1996).

The computational properties of a neural network are defined by the model of a neuron and weights connecting neurons. Typically, each neuron includes the summation node and the nonlinear activation function of the sigmoid $o = \frac{1}{1+\exp(-\lambda s)}$ form and/or hyperbolic tangent form $o = \frac{\exp(s)-\exp(-s)}{\exp(s)+\exp(-s)}$.

where s=**Wx** is the scalar output from a summation node;
l is the steepness of the activation function;
W is a weight matrix and **x** is an input vector.

In SAS Enterprise Miner, which was used in this simulation, the hyperbolic tangent and sigmoid are the default activation functions used in the hidden and output layers, respectively.

Neural networks are built from many neurons, organized in layers, because single neurons have limited capability. The typical neural network contains a hidden layer and an output layer. Using a numerical connection called a weight, each neuron in the hidden layer connects with every input and neuron in the output layer, if the neural network is fully connected. The strength of the connection and the relative importance of each input to the neuron are represented by the weights.

Because the network learns through repeated adjustment of the weights, they are crucial to neural networks' operation. Knowledge gained by the network during learning is encoded by the weights.

Neural networks come in several architectures. One of the most common architectures used in financial/accounting applications is the two-layer feed-forward network with error back-propagation. In such a network, signals propagate through the two layers from input to output.

Neural networks learn by experience from training patterns, typically in a supervised mode. A neural network is presented with many training patterns, one at a time. Each of the training patterns is marked by the class label of the dependent variable. After seeing enough of these patterns, the neural network builds the response model which reads in unclassified cases not seen during training, one at a time, and updates each with a predicted class.

Neural networks use a nonlinear activation function to model nonlinear behavior. Consequently, researchers often employ neural networks to solve sophisticated tasks and approximate functions in which relationships and interactions between variables are complex and nonlinear. One of the drawbacks of neural networks is the fact that the explicit mathematical equation estimated by the network to classify data is unknown; the neural network's knowledge is encoded in the numerical connections, called weights. Consequently, if/then rules that represent the relationships between inputs and outcomes cannot be easily constructed, making the produced results difficult to explain.

In our study we used a feed-forward network with back-propagation, default learning algorithm, and standard deviation normalization for input variables, all available in SAS Enterprise Miner. We tested several networks with different number of neurons in the hidden layer and one neuron in the output layer. The network with 2 neurons in the hidden layer apparently yielded the best classification results.

Decision Trees

Decision trees can also perform efficiently in classification tasks. Decision trees consist of flow-chart-like tree structures, where tests on the attributes are represented by nodes, conditions are represented by branches, and classes are reported in leaf nodes. Decision trees learn from input data in a supervised mode. For classification, the attribute values of an unknown sample are tested against the decision tree. The tree traces a path from a leaf node predicting a specific class back to the tree root for that sample.

Each unique path from the root to a leaf is represented by a rule. From the tree, if-then rules can easily be constructed to represent relationships between the dependent and independent variables. These rules can be very useful by providing insight into the model's operation and a compact explanation of the data. Reported at each node is the number of observations entering the node, the classification of the node, and the percent of cases correctly classified.

In decision trees, the type of splitting criteria available depends on the measurement level of the dependent variable. When the dependent variable is binary, the following three splitting

criteria are common: entropy reduction, Gini reduction and chi-squared test. One of the most common techniques for construction of entropy-based decision trees is the C4.5 algorithm which builds decision trees by a recursive, top-down, divide-and-conquer method (Quinlan 1993). The algorithm continually divides a data set into finer and finer clusters. The algorithm places the strongest predictive variable at the root of the tree.

The algorithm tries to produce pure clusters at the nodes by progressively reducing impurity in the original data set. Entropy (a concept borrowed from information theory) measures the impurity/information content in a cluster of data. The algorithm computes the gains in purity from all possible splits, and chooses a split that maximizes information gain. The process continues and the algorithm determines the least amount of splits to minimize the error rate on the training data set. Fewer splits, branches, and variables, produce a more understandable tree.

We now provide a brief introduction to the well-established concepts of entropy and information gain used to measure impurity. If a collection, *S*, contains positive (*yes*) and negative examples (*no*) of a target concept, the entropy of *S* in relation to that Boolean classification is:

$$Entropy(S) \equiv -p_{yes} \log_2 p_{yes} - p_{no} \log_2 p_{no} \tag{2}$$

In the equation, p_{yes} and p_{no} are the proportions of positive and negative examples in *S*, respectively.

The entropy of *S,* when the target attribute can take on *k* different values, is related to a *k*-wise classification defined as:

$$Entropy(S) \equiv \sum_{i=1}^{k} -p_i \log_2 p_i$$ in the entropy reduction method, and

$$Entropy(S) \equiv \left(1 - \sum_{i=1}^{k} (p_i)^2\right)$$ in the Gini reduction method.

Relative to a collection of examples *S*, *Gain(S, A),* the information gain of an attribute *A*, is defined as:

$$Gain(S, A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{S_v}{S} Entropy(S_v) \tag{3}$$

In the formula, *Values(A)* represents the set of all possible values for attribute *A*, while S_v represents the subset of *S* when attribute *A* has the value *v* (i.e., $S_v = \{s \in S | A(s) = v\}$

Chi-squared splitting criteria measure the reduction in variability of the target distribution in the branch (child) nodes. Specifically, the likelihood ratio Pearson chi-squared test statistic is a measure of association between the categories of the dependent variable and the branch nodes. This test statistic can be used to judge the worth of the split; it measures the difference between the observed cell counts and what would be expected if the branches and target classes were independent. We used a default significance level of 0.20 recommended by SAS for binary classification problems. (The 0.1 significance level produced exactly the same decision tree and the same classification rates for the training and test sets, whereas the 0.05 and 0.01 significance levels produced two simple trees with worse classification rates than the 0.2 significance level.

To summarize, logistic regression and neural networks embed their knowledge in their coefficients and weights, respectively, whereas knowledge in decision trees is represented in the form of linear and transparent rules. We discuss decision trees further in the following Results section. For a more thorough and comprehensive description of decision trees, see Giudici (2003; SAS Enterprise Miner at www.sas.com; Quinlan (1993); Dhar and Stein (1997; Kantardzic (2003).

RESULTS

Decision Tree

Because use of a decision tree is the focus of this study, we begin this section discussing results from the three decision tree methods. An advantage of using decision trees over neural networks is their ability to calculate the relative importance of input variables based on their predictive power and overall contribution to the classification tree (Breiman, Friedman, Olshen & Stone, 1984). The tree node incorporates the agreement between the surrogate split and the primary split in the calculation. The variable importance measure is scaled to be between 0 and 1 by dividing by the maximum importance. Thus, larger values indicate greater importance. Variables that do not appear in any primary or saved surrogate splits have importance equal to 0.

Table 2 presents the variables deemed important by the three decision tree methods. Panels A and B for the entropy reduction and Gini reduction methods, respectively, show that seven and ten variables, respectively, are important in those methods. The entropy reduction and Gini reduction methods consequently contain numerous splitting rules. In contrast, the results for the chi-squared test method, reported in Panel C, include only four important variables and relatively few splitting rules.

Table 2. Decision Tree - Relative Importance of Variables			
Panel A. Entropy reduction method			
Variable Name	Importance Value	Variable Role	Number of Splitting Rules Using the Variable
OE90	1.0	Input	4
REST90	0.798	Input	1
DEBT/ASSETS	0.62	Input	2
DEBT/EQUITY	0.62	Input	2
WORK_CAP	0.464	Input	1
AT90	0.458	Input	1
RFA90	0.349	Input	1
Remaining 7 variables	0.0	Rejected	0
Panel B. Gini reduction method			
Variable Name	Importance Value	Variable Role	Number of Splitting Rules Using the Variable
OE90	1.0	Input	2
REST90	0.906	Input	1
RRE	0.763	Input	2
DEBT/ASSETS	0.703	Input	2
RFA90	0.542	Input	1
IE90	0.528	Input	1
AT90	0.518	Input	1
DEBT/EQUITY	0.513	Input	1
DR90	0.458	Input	1
REST91	0.431	Input	1
Remaining 4 variables	0.0	Rejected	0
Panel C. Chi-square method			
Variable Name	Importance Value	Variable Role	Number of Splitting Rules Using the Variable
REST90	1.0	Input	1
OE90	0.951	Input	1
DEBT/ASSETS	0.528	Input	1
REST91	0.477	Input	1
Remaining 10 variables	0.0	Rejected	0

Table 2. Decision Tree - Relative Importance of Variables	
Dependent Variable: PAY92 = 0 if missed, 1 if made	
Independent Variables:	
REST90	= 0 if restructured on 1 January, 1990; 1 otherwise
REST91	= 0 if restructured on 1 January, 1991; 1 otherwise
DEBT/ASSETS	= 1991 Total debts/1991 Total assets
OE90	= 1990 Total operating expenses/1990 Total farm income
DEBT/EQUITY	= 1991 Total debts/(1991 Total assets - 1991 Total debt + 400,000)
WORK_CAP	= 1991 Total current farm assets - 1991 total current farm liabilities
AT90	= 1990 Total cash farm income/1990 Beginning total farm assets
IE90	= Total 1990 actual interest expense paid/Total 1990 farm income
RFA90	1990 Total cash farm income from operations - = operating expenses - family living expenses / 1990 Beginning total farm assets
RRE	1990 Total cash farm income from operations - = operating expenses - family living expenses / 1990 Total assets - 1990 Total debt + 400,000
DR90	= Total debt and interest payments due on 1990 FHP / 1990 Total cash farm income + Non-farm income

All three decision tree methods find that OE90, REST90, and DEBT/ASSETS are three of the four most powerful predictive variables. The methods disagree on what other variables are important. The chi-squared method found REST90 to contain the most predictive power. Thus, REST90 serves as the root of the chi-square tree. The relative importance of this variable is 1. Then OE90, DEBT/ASSETS, and REST91, in that order, were used in the tree. All the remaining ten variables have been pruned because their presence does not increase the overall classification accuracy of the tree.

All else equal, the simpler the decision tree and the fewer splitting rules, the better, particularly for FSA use. The chi-squared test method produced the simplest tree. However, predictive accuracy is an important criterion for potential users of decision trees. The decision trees developed on the training set were applied to the 60 test cases not included in the training set. Table 3 reports the predictive accuracy at different cutoff probabilities for these 60 observations overall, for the 14 defaulted loans, and the 46 paid loans.

Table 3. Classification Accuracy Rates for the Test Set by Different Decision Tree Methods: Counts and Percentages Classified Accurately for Different Cut-off Probabilities.

Cutoff probability [%]	DT Entropy reduction			DT Gini reduction			DT Chi square		
	O[1]	D[1]	P[1]	O	D	P	O	D	P
0	14	14	0	14	14	0	14	14	0
	23.3	100.0	0.0	23.3	100.0	0.0	23.3	100.0	0.0
10	29	14	15	34	9	25	14	14	0
	48.3	100.0	32.6	56.7	84.3	54.3	23.3	100.0	0.0
20	33	11	22	41	8	33	17	13	4
	55.0	78.6	47.8	68.3	57.1	55.0	28.3	92.9	8.7
30	34	10	24	42	8	34	48	8	40
	56.7	71.4	52.2	70.0	57.1	56.7	80.0	57.1	87.0
40	34	10	24	42	8	34	48	8	40
	56.7	71.4	52.2	70.0	57.1	56.7	80.0	57.1	87.0
50	48	7	41	44	7	37	50	7	43
	80.0	50.0	89.1	73.3	50.0	61.7	83.3	50.0	93.5
60	48	7	41	42	5	37	48	5	43
	80.0	50.0	89.1	70.0	35.7	61.7	80.0	35.7	93.5
70	46	5	41	42	5	37	48	5	43
	76.7	35.7	89.1	70.0	35.7	61.7	80.0	35.7	93.5
80	46	5	41	42	5	37	48	5	43
	76.7	35.7	89.1	70.0	35.7	61.7	80.0	35.7	93.5
90	47	5	42	46	5	41	48	5	43
	78.3	35.7	91.3	76.7	35.7	89.1	80.0	35.7	93.5

DT - Decision tree
Of a total of 60 cases divided into 14 defaulted loans and 46 paid loans, counts and percentages for: O – Overall, D – Defaulted, P – Paid

Overall, the chi-squared method classifies loans as accurately, or more accurately, than the other two decision tree methods at all reported cutoff levels above 20 percent. The Gini reduction method is more accurate at the 20 percent and 10 percent cutoff levels. A 50 percent cutoff implies that predicting a repayment is just as important as predicting a default; the cost associated with lending money to a farmer who does not repay (Type II error) is equal to the cost of not lending

money to a farmer who would repay the loan (Type I error). A 30 percent cutoff implies that a Type II error is more costly than a Type I error.

In a research note, Hsieh (1993) estimated that capital investors considered not correctly predicting an actual bankruptcy 3.242 times more costly than falsely predicting that a nonbankrupt firm would become bankrupt. She recommended using a cutoff percentage of .3085 for corporate bankruptcy predictions (Hsieh, 1993). While the loss function of equity investors is certainly different than that for FSA lending decisions, Hsieh's findings provide a reference to estimate appropriate cutoff percentages.

The FSA has a dual purpose in lending money: (1) providing support to farmers and (2) protecting taxpayer dollars through judicious lending decisions. Consequently, a 50 percent cutoff criterion may be appropriate. If the FSA mandate calls more heavily on protecting taxpayer funds, a lower cutoff percentage (perhaps 30 or 40 percent) would be more appropriate. Cutoff percentages above 50 percent would imply the unlikely assumption that denying loans to farmers who could repay their loans (Type I error) is more costly than lending money to farmers who do not repay (Type II error).

We discuss the results for the chi-squared test method in more detail because it produced the best overall classification results on the test set at cutoffs of 30% or greater with the least complex tree in terms of the number of leaves, the number of splits, and the depth of the tree. The tree is easy to understand because it uses only five rules and four variables to classify the data. The tree diagram with results for the training and test sets is shown in Figure 1.

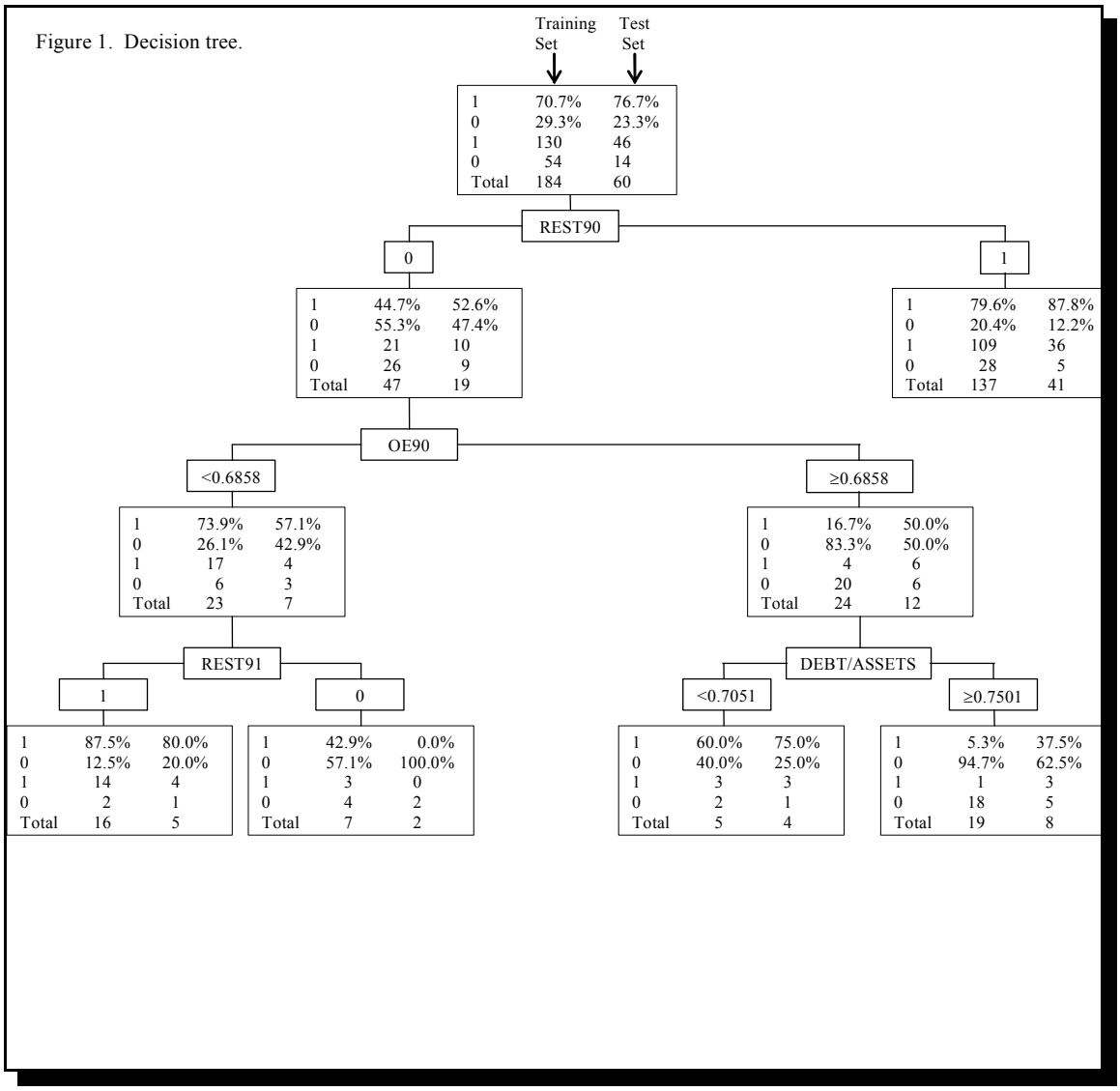
The decision branches and their split values in the tree make sense intuitively. The rules and classification rates produced by the tree for the training set follow. (N = number of cases entering the node). Remember that the dependent variable, payment of FmHA loan due on January 1, 1992, (PAY92) = 0 if missed, 1 if made, and REST90 and REST91 = 0 if farmer's FmHA debt was restructured and 1 if FmHA debt was not restructured in 1990 or 1991, respectively.

The tree generates five rules which use four variables only. As an example, the predicted values are calculated for a 50% cut-off. The tree first classifies loans to any farmers who did not restructure their farm debt on January 1, 1990 as expected to repay.

IF REST90 = 1
THEN Predicted value: 1

N	:	137 training cases	N	:	41 test cases
1	:	79.6% - 109 training cases	1	:	87.6% - 36 test cases
0	:	20.4% - 28 training cases	0	:	12.4% - 5 test cases

Figure 1. Decision tree.



As can be seen in Figure 1, farmers unable to make loan payments in 1990 (REST90 = 0) also faced difficulty paying off the loan due in 1992. (More than half of the farmers with REST90 = 0 in the training set defaulted on the 1992 payment.) For loans to these farmers, the tree examines their operating expense ratio first. If the operating expense ratio is less than 0.6858, whether farm debt was restructured in 1991 becomes the determining classification factor. Farmers who did not restructure debt in 1991 were predicted to repay in 1992 while farmers who restructured in 1991 were not expected to repay in 1992.

IF REST90 = 0 AND OE90 < 0.6858 AND REST91 = 1
THEN Predicted value: 1

N	:	16 training cases	N	:	5 test cases
1	:	87.5% - 14 training cases	1	:	80.0% - 4 test cases
0	:	12.5% - 2 training cases	0	:	20.0% - 1 test case

IF REST90 = 0 AND OE90 < 0.6858 AND REST91 = 0
THEN Predicted value: 0

N	:	7 training cases	N	:	2 test cases
1	:	42.9% - 3 training cases	1	:	0.0% - 0 test cases
0	:	57.1% - 4 training cases	0	:	100.0% - 2 test cases

If farmers restructured debt in 1990 (REST90 = 0) and exhibited operating expenses $\geq$ 0.6858 of farm income (OE90 $\geq$ 0.6858), the likelihood of not paying off the loan increases to about 83%. In this case, the debt to asset ratio becomes the determining classification factor. Such loans exhibiting DEBT/ASSETS < 0.7051 are predicted to make their 1992 debt repayment, while observations with DEBT/ASSETS $\geq$ 0.7051 are predicted to not repay their debt for 1992.

IF REST90 = 0 AND OE $\geq$ 0.6858 AND DEBT/ASSETS < 0.7051
THEN Predicted value: 1

N	:	5 training cases	N	:	4 test cases
1	:	60.0% - 3 training cases	1	:	75.0% - 3 test cases
0	:	40.0% - 2 training cases	0	:	25.0% - 1 test case

IF REST90 = 0 AND OE ³ 0.6858 AND DEBT/ASSETS ³ 0.7051
THEN Predicted value: 0

N	:	19 training cases	N	:	8 test cases
1	:	5.3% - 1 training case	1	:	37.5% - 3 test cases
0	:	94.7% - 18 training cases	0	:	62.5% - 5 test cases

Neural Network and Logistic Regression

To fully evaluate the predictive ability of the decision tree, the data was also analyzed to select a logistic regression model and neural network that produced the best predictive results. Unlike Barney, Graves and Johnson, (1999) who included all available variables in their logistic regression model, three variable selection methods available in SAS were used to find the best logistic regression model: forward, backward, and stepwise. In the forward selection method, the best one-variable model is first chosen. Then the method selects the best two-variable model among those that contain the first selected variable. The process continues until no additional variables have

a *p*-value less than the specified entry *p*-value known as a significance level. In the backward selection technique, the process begins with all variables included in a model. Variables are then removed from the model until only variables with a *p*-value less than a specified significance level remain.

The stepwise method is a modification of the forward selection method. The difference is that variables already selected for the model do not necessarily stay there. The stepwise process may remove any variable already in the model that is not associated with the dependent variable at the specified significance level. The process continues until none of the variables outside the model has a *p*-value less than the specified significance level and every variable in the model is significant at that level.

We analyzed the data using the three methods. The stepwise selection method, at a specified *p*-value of 0.05, identified a model including three significant variables, DEBT/ASSETS, REST90, and REST91, that produced the best overall classification results for the test set for any logistic regression model. Of several types of neural networks examined, the best classification results were produced by a two-layer, feed-forward network with back-propagation having two neurons in the hidden layer available in SAS Enterprise Miner. Table 4 presents output from the best logistic regression classification model and neural network selected.

Table 4. Logistic Regression Model Output

Panel A: Likelihood Ratio Test for Global Null Hypothesis: BETA=0

	Chi-Square	DF	Pr > ChiSq		
Likelihood Ratio	37.2599	3	<.0001		

Panel B: Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Wald Pr > ChiSq
Intercept	1	0.7643	0.7315	1.0916	0.2961
REST90	1	1.0551	0.3977	7.0366	0.0080
REST91	1	1.1170	0.4090	7.4596	0.0063
DEBT/ASSETS	1	-1.8589	0.6959	7.1354	0.0076

Dependent Variable:
PAY92 = 0 if missed, 1 if made
Independent Variables:
REST90 = 0 if restructured on 1 January, 1990; 1 otherwise
REST91 = 0 if restructured on 1 January, 1991; 1 otherwise
DEBT/ASSETS = 1991 Total debts/1991 Total assets

Evaluation/Comparison of Results

Of primary interest is the predictive ability of the analytical methods on the test set -- the 60 observations not included in the training set. The logistic regression model, neural network, and decision tree developed on the training set were applied to the 60 test cases. Table 5 reports the overall classification accuracy for at different cutoff percentages for the FmHA's internally developed criteria, and the criteria developed by Price Waterhouse, reported in the original studies by Barney (1993) and Barney, Graves and Johnson (1999). Table 5 also reports the classification accuracy rates for the overall test set, defaulted loans, and paid loans for the chi-squared test decision tree, neural network, and logistic regression model.

Table 5. Classification Accuracy Rates for the Test Set by Different Methods: Counts and Percentages Classified Accurately for Different Cut-off Probabilities

Cutoff probability [%]	FmHA[a b]	PW[a c]	LR[d]			NN[e]			DT[f]		
	O[g]	O	O	D[g]	P[g]	O	D	P	O	D	P
0	15	17	14	14	0	14	14	0	14	14	0
	25.0	28.3	23.3	100.0	0.0	23.3	100.0	0.0	23.3	100.0	0.0
10	16	25	15	14	1	22	12	10	14	14	0
	26.6	41.7	25.0	100.0	2.2	36.7	85.7	21.7	23.3	100.0	0.0
20	18	29	42	14	28	26	12	14	17	13	4
	30.0	48.3	70.0	100.0	60.9	43.3	85.7	30.4	28.3	92.9	8.7
30	20	30	46	12	34	46	5	41	48	8	40
	33.3	50.0	76.7	85.7	73.9	76.7	35.7	89.1	80.0	57.1	87.0
40	24	37	47	10	37	47	5	42	48	8	40
	40.0	61.7	78.3	71.4	80.4	78.3	35.7	91.3	80.0	57.1	87.0
50	28	39	49	9	40	47	5	42	50	7	43
	46.6	65.0	81.7	64.3	87.0	78.3	35.7	91.3	83.3	50.0	93.5
60	35	43	53	8	45	47	5	42	48	5	43
	58.3	71.7	88.3	57.1	97.8	78.3	35.7	91.3	80.0	35.7	93.5
70	43	43	51	5	46	47	5	42	48	5	43
	71.6	71.7	85.0	35.7	100.0	78.3	35.7	91.3	80.0	35.7	93.5
80	43	44	47	1	46	47	5	42	48	5	43
	71.6	73.3	78.3	7.1	100.0	78.3	35.7	91.3	80.0	35.7	93.5
90	45	44	46	0	46	46	4	42	48	5	43
	75.0	73.3	76.7	0.0	100.0	76.7	28.6	91.3	80.0	35.7	93.5

Table 5. Classification Accuracy Rates for the Test Set by Different Methods: Counts and Percentages Classified Accurately for Different Cut-off Probabilities

[a]Adapted from Table 17 in (Barney, Graves & Johnson, 1999)
[b]FmHA – Farmers Home Administration internally developed criteria in 1992
[c]PW – Price Waterhouse model developed for the FmHA
[d]LR - Logistic regression model
[e]NN - Neural network
[f]DT - Decision tree – Chi-square method
[g]Of a total of 60 cases divided int

The chi-squared test decision tree, neural network, and logistic regression model perform better (significantly) overall than the FmHA criteria at the 30 percent through 60 percent cutoffs. These methods are also significantly better than the Price Waterhouse selection criteria at the 30 percent and 40 percent cutoffs. The decision tree and neural network are significantly better at the 50% cut off. Table 6 presents the null hypothesis and the proportional z-statistics for comparisons of the overall accuracy rates of the techniques. The decision tree produces the highest overall classification accuracy rates for the 30, 40, and 50 percent cutoffs. However, the overall classification accuracy rates between the three analytical methods are not significantly different for cutoff percentages 30 percent and higher.

Table 6. Overall Classification Rate Comparisons Z-scores for the Test Set.

	Cutoff %									
Comparison:	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%
FmHA v. PW	-0.41	-1.74*	-2.05**	-1.86*	-2.38**	-2.03**	-1.54	-0.01	-0.21	0.21
FmHA v. DT	0.22	0.42	0.20	-5.16**	-4.47**	-4.21**	-2.57**	-1.07	-1.07	-0.66
FmHA v. LR	0.22	0.20	-4.38**	-4.78**	-4.27**	-4.01**	-3.71**	-1.78*	-0.85	-0.22
FmHA v. NN	0.22	-1.19	-1.51	-4.78**	-4.27**	-3.59**	-2.35**	-0.85	-0.85	-0.22
PW v. DT	0.63	2.15**	2.25**	-3.45**	-2.21**	-2.29**	-1.06	-1.06	-0.87	-0.87
PW v. LR	0.63	0.56	0.55	-3.04**	-1.98**	-1.62	-0.83	-0.83	-0.64	-0.43
PW v. NN	0.63	1.94**	-2.42**	-3.04**	-1.98**	-2.07**	-2.27**	1.77**	-0.64	-0.43
DT v. LR	0.00	-0.22	-4.57**	0.44	0.23	0.23	-1.24	-0.72	0.23	0.44
DT v. NN	0.00	-1.60	-1.71**	0.44	0.23	0.70	0.23	0.23	0.23	0.44
LR v. NN	0.00	-1.39	2.95**	0.00	0.00	0.47	1.47	0.95	0.00	0.00

Note: Z-score for null hypothesis that: (the proportion properly classified by the first method mentioned – the proportion properly classified by the second method mentioned) = 0.
* Significant at $p \leq 0.05$.
** Significant at $p \leq 0.01$

A weakness of the decision tree is that the technique predicts relatively poorly for very low probability cutoffs, those that consider the cost of a missed payment (Type II error) extremely high compared to the cost of not lending to a farmer who could repay the loan (Type I error). The Price Waterhouse model, logistic regression model, and neural network all performed significantly better than the decision tree at the 20 percent and/or 10 percent cutoff probability. However, given the mission of the FSA, a cutoff percentage lower than 30% would not likely be considered. Another weakness could be that the overall accuracy rates for the decision tree (80.0 and 83.3 percent at the 30 percent and 50 percent cut off probabilities, respectively), while relatively high compared to other methods, are not much higher than the 76.7 percent of loans in the test set that were repaid. A naïve, but unrealistic assumption that all loans will be repaid would produce a 76.7 percent overall classification accuracy rate. The decision tree achieves its accuracy rates while properly classifying 50.0 and 57.1 percent of loans that are not repaid at the 50 and 30 percent cut off probabilities, respectively.

CONCLUSION

The aim of this study is not to recommend a loan classification system for immediate FSA use. Rather, we build, test, and present a viable and transparent model, the decision tree, which could potentially improve FSA lending practices, making lending decisions more transparent and easily understood by applicants and the FSA staff. With loan default percentages varying over time, we discuss classification accuracy rates at several possible cut-offs. At the most likely relevant cut off percentages, a decision tree, neural network, or logistic regression model would significantly improve classification accuracy rates over the internally developed FmHA (FSA) criteria and perform better than the criteria developed by Price Waterhouse at much government expense. While the chi-squared test decision tree performs comparatively as well as the neural network and logistic regression model, its clarity when used in practice is a major advantage.

Once the decision tree determines the variables indicative of loan repayment or default and determines the appropriate cutoff point for those variables, the tree accounts for relevant possible combinations of those variables. In this manner, the decision tree accounts for all possible input observations and provides clear, understandable predictions (more so than other analytical methods). Then, the model or its user can determine into which group a loan application falls to predict repayment or default. FSA employees, farmers, and legislators could all understand the decision rules and evaluate the results of lending decisions based on those rules.

The decision tree technique should be considered in any revision of the FSA lending program because of its great potential to improve the FSA's lending practices and make them more transparent. Analysis with a more recent and larger data set would be an appropriate extension of this study as would performing more tests and implementing k-fold cross-validation to obtain more reliable and unbiased classification error estimates. The decision tree could be updated annually

based on actual repayment data from recent years. Assembling national data on repayment and default rates by farmers would be essential to improving and maintaining the system.

REFERENCES

Barney, D.K. (1993). *The farmers home administration and farm debt failure prediction*. Ph.D Dissertation, University of Mississippi.

Barney, D.K., O.F. Graves & J.D. Johnson (1999). The farmers home administration and farm debt failure prediction. *Journal of Accounting and Public Policy*. 18(2), 99-139.

Berry, M. & G. Linoff (1997). *Data mining techniques: For marketing, sales, and customer support*. New York: John Wiley and Sons.

Breiman, L., J.H. Friedman, R.A. Olshen & C.J. Stone (1984). *Classification and regression trees*. Chapman and Hall.

Dhar, V. & R. Stein (1997). *Seven methods for transforming corporate data into business intelligence*. Prentice Hall.

Farm Service Agency, from http://www.fsa.usda.gov/dafl/default.htm.

Farm Service Agency (2006). *Report to Congress: Evaluating the relative cost effectiveness of the Farm Service Agency's farm loan programs*. United States Department of Agriculture, from http://www.fsa.usda.gov/Internet/FSA_File/farm_loan_study_august_06.pdf.

Featherstone, A.M., L.M. Roessler & P.J. Barry (2006). Determining the probability of default and risk-rating class for loans in the seventh farm credit district portfolio. *Review of Agricultural Economics*, 28(1), 4-23.

Giudici, P. (2003). *Applied data mining: Statistical methods for business and industry*. Chichester, West Sussex, England: John Wiley & Sons.

Hagan, M.T., H.B. Demuth & M. Beale (1996). *Neural network design*. PWS Publishing Company.

Han, J. & M. Kamber (2001). *Data mining: Concepts and techniques*. San Francisco, CA: Morgan Kaufmann Publishers.

Hosmer, D.W. & S. Lemeshow (1989). *Applied logistic regression*. New York: Wiley.

Hsieh S-J. (1993). A note on the optimal cutoff point in bankruptcy prediction models. *Journal of Business Finance & Accounting*, 20(3), 457-464.

Kantardzic, M. (2003). *Data mining: Concepts, models, methods, and Algorithms*. IEEE Press/Wiley.

Katchova, A.L. & P.J. Barry (2005). Credit risk models and agricultural lending. A*merican Journal of Agricultural Economics*, 87, 195-206.

Koh, H.C. (2004). Going-concern prediction using data mining techniques. *Managerial Auditing Journal*, 19(3), 462-476.

Kyung, S., T.N. Chang & G. Lee (1999). Dynamics of modeling in data mining: Interpretive approach to bankruptcy prediction. *Journal of Management Information Systems,* 16(1), 63-85.

Press, S.J. & S. Wilson (1978). Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association*. December, 699-705.

Quinlan, J.R. (1993). *C4.5: Programs for machine learning*. San Mateo, California: Morgan Kaufman Publishers.

SAS Institute, Inc. (1999).