International Food and Agribusiness Management Review
Volume 24, Issue 5, 2021; DOI: 10.22434/IFAMR2020.0019

Received: 10 February 2020 / Accepted: 5 April 2021

Agricultural loan delinquency prediction using machine learning methods

RESEARCH ARTICLE

Jian Chen[a], Ani L. Katchova[ⓘb] and Chenxi Zhou[c]

[a]Former PhD student, [b]Professor and Farm Income Enhancement Chair, Department of Agricultural, Environmental, and Development Economics, The Ohio State University, 2120 Fyffe Road, 250 Ag Admin Building, Columbus, OH 43210, USA

[c]PhD student, Department of Statistics, The Ohio State University, Cockins Hall, Room 305E, 1958 Neil Ave, Columbus, OH 43210, USA

Abstract

The recent economic downturn in the agricultural sector that started in 2013 has caused some concerns for farmers' repayment capacity, which raises the need for precise prediction of financial stress in the agricultural sector. Machine learning has been shown to improve predictions with large financial data, however, its application remains limited in the agricultural sector. In this study, we approximate financial stress by agricultural loan delinquency, and predict it by employing a logistic regression and several machine learning methods. The main datasets include the Call Reports and Summary of Deposits from the Federal Deposit Insurance Corporation (FDIC). Our results show that ensemble learning methods have the best performance in prediction accuracy, with improvement of 26 percentage points at most and that the Naïve Bayes classifier is the best method to maintain the lowest cost from false predictions when the failure of identifying potentially high-risk loans is very costly. From the perspective of banks, while there are important benefits to using machine learning, the bank-level costs are also important considerations that may lead to different choices of machine learning methods.

Keywords: agricultural credit, forecasting, machine learning, logistic regression
JEL code: C63, E37, G21, Q14

ⓘCorresponding author: katchova.1@osu.edu

1. Introduction

Financial stress has been of growing importance to the agricultural sector. The farm economy has been experiencing six years of lower farm incomes, stagnant land values and increased bankruptcy and delinquency rates (Katchova and Dinterman, 2018). In this environment, it has become increasingly important to understand the causes of agricultural financial stress and to predict their occurrence. Obtaining more accurate predictions of loan delinquencies would provide lenders with better guidance to adopt appropriate financial management strategies, such as decisions of whether to make extend more or less credit to farmers and how much to hold in reserves to prepare for a potential downturn. These prediction results can also help farmers to make more informed financial decisions on whether and how much credit to take to ensure the long-term survival of their farm businesses.

As a recent development in computer science and statistical techniques, prediction could be improved by using large datasets that provide detailed information for each observation. For example, Tack *et al.* (2017) suggest that the United States Department of Agriculture (USDA) forecasts can be improved with big data. However, there are some challenges in managing and using large public, private, and administrative datasets in the agricultural sector, such as the management and warehousing of the data, maintaining their privacy and security, and most importantly, the need to adapt new methodological advances (Woodard, 2016). Many standard econometric methods cannot take advantage of large datasets with comprehensive information: they impose strong assumptions on the data generating process; and they are unable to deal with non-linear relationships and high multi-collinearity among independent variables, potentially leading to large prediction errors.

Advanced empirical techniques, like machine learning algorithms, have the potential to substantially increase the research value of large datasets. In financial studies, researchers have been describing, studying, and predicting financial crises and mortgage defaults. Machine learning methods have been applied to predict financial crises and defaults of financial institutions using large financial datasets and these methods have been found to have superior performance for the purpose of prediction and corresponding policy making (Demyanyk and Hasan, 2010; Ekinci and Erdal, 2017; Iturriaga and Sanz, 2015; Ozturk *et al.*, 2016, Zhao *et al.*, 2017). However, studies in the agricultural sector that use improved empirical techniques have remained limited, especially from the perspective of banks.

In this study, we address the research questions of whether machine learning techniques provide more accurate and timely predictions of agricultural loan delinquency than a logistic regression and determine the factors that a bank should take into account when making choices of different prediction methods. Specifically, we approximate financial stress by agricultural loan delinquencies, and follow Ifft *et al.* (2018)'s framework to evaluate different machine learning methods and compare the machine learning results to prediction using a standard econometric model. We also analyze the benefits and costs of different machine learning methods, focusing on the costs of incorrect predictions by banks or regulatory agencies.

The main dataset used in the study is the Consolidated Reports of Condition and Income (i.e. Call Reports) published by the Federal Deposit Insurance Corporation (FDIC, Washington, DC, USA). Call Reports data serve a regulatory and public policy purpose, and contain plenty of financial and operational information from every national bank, state member bank, insured state nonmember bank and savings association. This dataset has been commonly used by the public, bank rating agencies, and researchers for analyses, regulatory purposes, and policy recommendations.

Our study makes several contributions to the existing literature by applying a relatively new technique of machine learning to agricultural loan delinquency prediction. First, unlike other studies on agricultural financial stress prediction, this study is from the perspective of banks and regulatory agencies and provides a more accurate technique framework to identify the risk of agricultural loans applying existing machine learning methods. Second, we compare different machine learning algorithms with a standard econometric model

by using various indicators to evaluate model performance in the context of agricultural loan delinquency prediction. As a relatively new technique to be implemented, we provide a discussion of the machine learning methodology in the area of agricultural loan prediction, as well as other important considerations for lenders and researchers to take into account when using these advanced techniques. While such methods and evaluation techniques have been used in other contexts, we apply them using the FDIC bank dataset to provide results and implications for banks or financial institutions. In terms of the implementation of machine learning in practice, our study provides a conservative perspective: not all advanced models show significant improvement and the cost of advanced methods could be high. Since improvement in prediction depends very much on the specific dataset, our results using aggregate-level data may not be applied directly to customer- or portfolio-level data, which will need further exploration.

The remaining paper is organized as follows: Section 2 presents an overview of modeling choices of financial crisis prediction, especially in the agricultural sector. In Section 3, several datasets on bank financial situations and agricultural features are presented. Sections 4, 5, and 6 describe the methodology to deal with the data imbalance issue, to implement predictions, and to evaluate model performance. In Section 7, results from different methods are presented and compared. In the last section, work carried out in the study is concluded.

2. Relevant research

A number of studies have researched the impacts of various factors on financial stress indicators like delinquency rate and bankruptcy, and most have employed logit or probit techniques to build their models. The adoption of a binary model design divides banks into two classes – failure and non-failure, and estimate a bank's probability of falling into one class or the other, given bank-specific characteristics (Cole and Gunther, 1998; Pantalone and Platt, 1987; Thomson, 1991). Different from a logit or probit model, the duration model has the capability to generate estimates of the probable time to failure, in addition to estimates of the probability of financial risk. Li *et al.* (2014) propose a Cox proportional hazard model to investigate the relationship between survival time and bank-level determinants of failure among agricultural and commercial banks during the latest Great recession, and they find that non-performing loans and interest rate risk significantly impaired banks' survival and financial health.

However, these traditional classification and regression methods such as logit or probit models, ordinary least squares regression, and generalized linear models often impose strong assumptions on the data-generating process and linear relationships between the dependent and independent variables. In addition, they have limitations when dealing with outliers, non-linear relationships, and variable selection.[1] Particularly, these traditional methods hardly address situations in which independent variables are highly collinear, leading to: (1) unreliable statistical inference given that the coefficients may have larger variance and may lose their significance; and (2) potential overfitting problem in out-of-sample prediction due to large variance and therefore large prediction errors for the model (Friedman *et al.*, 2001).[2]

Studies have shown that these methodological issues can benefit from machine learning methods. Mullainathan and Spiess (2017) argue that machine learning can generally improve the predictive performance of a variety of econometric methods through variable selection among complex functional structure that was not specified in advance, especially for datasets with a large number of variables. Furthermore, for raw survey responses and variables including missing observations, machine learning may improve imputation and accommodate a less restrictive approach for existing statistical methods (Morehart *et al.*, 2014). Machine learning can also

[1] One way to deal with a non-linear relationship is to add polynomial terms or interaction terms based on theoretical reasons or one-way plots between dependent and independent variables, but this would impose assumptions on the functional form and the error term. Another way is to apply non-parametric methods such as nearest neighborhood or regression trees. However, these non-parametric methods belong to the more general machine learning methods where parameters are to be optimized to reduce the overfitting problem given a dataset.

[2] In statistics, overfitting refers to the analysis that corresponds closely or exactly to a particular data set and may fail to fit additional data or predict future observations. An overfitted model is a statistical model that contains more parameters or independent variables than can be justified by the data. Overfitted models are often free of bias in the parameter estimators but have large sampling variances that lead to poor precision of the estimators in out-of-sample prediction. A best approximating model is achieved by properly balancing bias and variance (Burnham and Anderson, 2002).

lead to inferences from its complex algorithm, such as the lasso-based average treatment effect estimation and the random forest-based heterogeneous treatment effect estimation (Bloniarz *et al.*, 2016; Wager and Athey, 2017).

Most importantly, by being flexible with such common data issues, machine learning methods have been shown to improve prediction compared to traditional methods in the context of financial crises prediction. Demyanyk and Hasan (2010) review the literature on predicting the subprime mortgage crisis in the U.S., and suggest that the statistical methods need to be improved to better predict and analyze defaults and crises by including machine learning techniques. Iturriaga and Sanz (2015) employ neural networks to detect commercial bank failures in the U.S., and demonstrate that their model outperforms the traditional models of bankruptcy prediction with a 96% accuracy rate. Ekinci and Erdal (2017) demonstrate that hybrid ensemble machine learning models outperform conventional logistic and ensemble models when predicting bank failures in the Turkish banking system. Zhao *et al.* (2017) propose a kernel extreme learning machine, a two-step grid search strategy, to predict bankruptcy and obtain superior performance in classification accuracy.

In the agricultural sector, there are studies on financial stress from the perspective of farms (Dinterman *et al.*, 2018; Quaye *et al.*, 2017), but a limited number of studies have explored the potential of machine learning techniques. By using detailed farm-level survey data (from the Agricultural Resource Management Survey (ARMS)), Quaye *et al.* (2017) examine the factors and behaviors that affect Southeast US farmers' ability to meet their loan repayment obligations within the stipulated loan term. They use a probit method to study how borrower-specific, loan-specific, lender-specific, macroeconomic and climate characteristics affect delinquency. They find that farmers with larger farms, who have more insurance, higher net income, lower debt-to-asset ratio, single loans and those who take the majority of their loans from sources other than commercial banks are less likely to default on their loans. Temperature and precipitation are also shown to affect outcomes but with small magnitudes.

By using the ARMS dataset, Ifft *et al.* (2018) employ machine learning methods to predict new credit demand and demonstrate that advanced models have advantages and are more flexible than the logistic regression in prediction, especially for large datasets. While Ifft *et al.* (2018)'s study is from the perspective of borrowers using detailed farm-level data, there are no studies, to our knowledge, that investigate the risk from the perspective of banks using aggregate-level data. Thus, our paper fills this gap in the literature.

3. Data

Datasets in this article come from different sources. Information on agricultural loan delinquency rates were obtained through the FDIC's Consolidated Reports of Condition and Income (e.g. Call Reports). Delinquency volume is determined by summing the total value of loans whose repayment were more than 90 days late and in nonaccrual status, adding the value that has been charged off, and subtracting the value that has been recovered. Delinquency rate is the ratio of delinquent volume on agricultural loans to the total value of agricultural loans. Agricultural loans generally have different risks than other types of loans, such as real estate and consumer loans, which leads to lower delinquency rates according to historical data from Federal Reserve Banks. In financial crisis prediction literature, institutional crisis or bankruptcy is usually associated with a very high loan delinquency rate within the institution (Paulson *et al.*, 2013). Thus, the delinquency rate is an important indicator to determine the quality of the loan portfolio. A lower delinquency rate is always desirable, as it indicates that there are fewer loans in the lender's loan portfolio that have late outstanding debt. However, the absolute value of the delinquency rate is not necessarily of interest, but the delinquency rate should be compared to an industry average or among the loan portfolio of competitors to determine whether the loan portfolio of a bank has an 'acceptable' rate. We use a default rate of 3% as the cut-off value for high risk associated with agricultural loans because it is one standard deviation (0.24%) above the mean delinquency rate (0.06%). In addition, 3% delinquency rate is considered high risk for agricultural loans. The Federal Reserve Bank considers 3% as medium-high rate for commercial loans. We therefore transform the delinquency rate into a classification problem in the prediction framework.

Information used for predicting financial stress from the Call Reports include bank-level financial and operational variables.[3] Traditional proxies for the CAMELS[4], which refer to the five components of the regulatory rating system for a bank – capital adequacy, asset quality, management, earnings, and liquidity – have been demonstrated as important determinants of bank financial stress (Cole and White, 2012). Following Cole and White (2012) and Iturriaga and Sanz (2015), we select variables included in the Call Reports to calculate these five components and use them as the independent variables in the model (see Table 1 for the variables and their definitions). Other variables include bank size and market share[5]; the latter indicates the size of a bank in relation to the market and its competitors.

Agricultural characteristics, such as farm net income and interest expense, come from the Farm Income and Wealth Statistics (Dinterman *et al.*, 2018). Macroeconomic indicators, such as unemployment rate and interest rate from the Federal Reserve Economic Data and the Bureau of Labor Statistics, are hypothesized to play an important role when predicting financial stress (Li *et al.*, 2014). In order to link the bank-level information of agricultural loan delinquency and other operational variables with state-level agricultural characteristics and macroeconomic conditions, we use the branch-level geographical information[6] from the Summary of Deposits published by the FDIC to link them together. That is, we calculate the average state-level agricultural characteristics based on the specific locations of bank branches using as weights the amount of deposits owned by each branch. We use the following assumptions: (1) the financial situation of agricultural loans for a bank is affected by agricultural conditions in multiple states depending on where its branches are located; (2) a bank that has more deposits in one state would be more likely to be affected by agricultural conditions in this state, therefore we give this state a larger weight in weighted average calculation.[7]

Among the general types of risk in agriculture, we consider production risk and market risk by controlling for agricultural characteristics such as farm income and production expenses. We also consider institutional risk by controlling for bank-level financial indicators except for borrower level information (Komarek *et al.*, 2020).

The sample data we use consist of bank data, agricultural characteristics, and macroeconomic conditions one year prior to measuring the agricultural delinquency risk. One-year-ahead models have been frequently used (e.g. Barr and Siems, 1994; Li *et al.*, 2014; Tam and Kiang, 1992, Zhao *et al.* 2009). While some studies have estimated models with longer prediction periods, say 5 years, to capture cumulative financial stress, such models are often not sufficiently accurate. Unobserved factors are more likely to change over a longer time period, thus models with shorter prediction have more certainty (Zhao *et al.* 2009). Since we also would like to include as many years of data as possible in our panel, we will not examine models with longer prediction periods in our study.

The final sample covers the period from 1994 to 2017 since the Summary of Deposits started data collection in 1994. About 36.6% of the banks in the Call Reports do not have agricultural loans, and thus are not included in this study. In order to link the final sample with yearly observations in other datasets, we use observations of the fourth quarter in each year from the Call Reports to calculate delinquency rates and financial predictors. After excluding observations with missing key variables, there are 131,431 bank-year observations in the final

[3] Here the financial and operational variables are aggregated information at the bank level, i.e. the entire bank that usually has multiple branches in different locations.

[4] The Uniform Financial Rating System, informally known as the CAMEL ratings system, is used by regulators during on-site examinations to determine the bank financial conditions. It was introduced by U.S. regulators in November 1979; in 1996, CAMEL evolved into CAMELS, with the addition of a sixth component which summarizes the sensitivity to market risk.

[5] Bank size refers to the sum of deposits from all of its branches. The market share for each branch is calculated based on its proportion of deposits among all deposits (including those from different banks) in a county to indicate a branch's market share in a county's financial market considering competition from other banks. To link with bank-level data (our main dataset), we further calculate the average of branch-level market share weighted by branch deposit for a bank to indicate the entire bank's market share across U.S.

[6] A bank has multiple branches located in different counties or different states. The Summary of Deposits contains a bank's information of all its branches, i.e. locations and deposit sizes. This information can be used to aggregate and serve as weights when we link bank-level data with state-level information, because a bank usually has multiple branches located across different states and a simple average of its data across states will miss the fact that a bank may have more services in one state than the other.

[7] This is a strong assumption since we do not have information on agricultural loans or accounts in a branch, thus we use the entire deposits of a branch as an approximation.

sample. Table 1 provides definitions of selected variables for agricultural loan delinquency prediction and a summary of the available data at the bank level. The majority of banks in our sample have a low delinquency rate, and only 6.4% of the banks have high risk associated with agricultural loans. As we will discuss later, the disproportion of two categories in the dependent variable would be problematic if using the standard method for prediction. Supplementary Table S1 shows the correlation matrix of these independent variables.

4. The SMOTE technique

A binary dependent variable is called 'imbalanced' if the majority of the observations are labeled as one category (or the majority class) while only a few observations are labeled as the other category (or the minority class), usually the more important class or the class of most interest. The imbalance problem has been recognized in many fields (Guo *et al.*, 2008; Zhu *et al.*, 2019). Since traditional algorithms seek an accurate performance over the full range of observations, they tend to be overwhelmed by the majority class and ignore the minority class in prediction.[8] Therefore, machine learning techniques do not work well when applied to an imbalanced class dataset.

[8] One example is that if a dataset has 99% of the observations categorized as 1 and 1% categorized as 0, then a model simply predicting every observation as '1' regardless of any independent variables will have a predictive accuracy of 99%. However, such model would not provide any useful information.

Table 1. Definition and statistics of selected variables.

Dimension	Definition	Mean	Standard deviation
Agricultural loan delinquency	binary, critically high (>3%) or acceptably low (≤3%)	0.064	0.244
Capital adequacy	total equity / total asset	0.105	0.033
Asset quality	gross loans / total asset	0.608	0.150
	agricultural loans / total loans	0.117	0.151
	loan unearned income / total equity	0.010	0.045
Management	net income / net operating income	1.021	2.120
Earnings ability	net income / average total assets	0.010	0.009
	interest income / gross loans	0.103	0.058
	net operating income / total assets	0.010	0.009
	efficiency ratio	0.665	0.256
	return on assets	0.010	0.009
	return on equity	0.101	0.542
	total interest income / average earnings assets	0.066	0.018
	total interest expenses / average earning assets	0.024	0.013
Liquidity position	cash / total assets	0.063	0.059
	(cash + securities) / total assets	0.320	0.149
	(cash + FedFunds) / total assets	0.111	0.083
	gross loans / deposits	0.732	0.722
Other financial	weighted market share of all branches (deposit at the county level)	0.142	0.152
Agricultural characteristics (weighted by branch deposit at the state level, $ billions)	net farm income	2.201	1.900
	net cash income	2.717	2.222
	net value-added	3.777	2.962
	interest expenses	0.527	0.318
	production expenses	8.521	6.051
	cash receipts	9.026	6.753
Macro indicators	unemployment rate	0.057	0.015
	interest rate	0.027	0.022

Financial crisis is a rare event in real life and in statistical prediction (Demyanyk and Hasan, 2010; Dinterman *et al.*, 2018). In our study, only 6.4% of the banks are classified as high risk associated with agricultural loans (with a corresponding default rate on agricultural loans of above 3%, one standard deviation above the mean value). However, these banks are the class of most interest since it is the banks with high risk that are of most concern to be able to identify correctly ahead of time, like the financial crisis and bankruptcies in 2008. Misclassifying them as banks with low risk would potentially cause losses to both banks and their borrowers when financial stress occurs. In order to remedy this class imbalance problem in our dataset, we have chosen to change the class distribution by using over-sampling techniques. We use the Synthetic Minority Over-Sampling Technique (SMOTE), which generates synthetic minority observations to over-sample the minority class (Chawla *et al.*, 2002; Zhu *et al.*, 2019). Specifically, it forms new minority class observations by interpolating between several existing minority class observations using the *k*-nearest neighbor algorithm. Experiments show that SMOTE is superior to other over-sampling techniques in terms of computational efficiency and prediction accuracy of minority class (Chawla *et al.*, 2002). In the following section, we apply the SMOTE technique in a logistic regression and all machine learning methods to correct the imbalance issue.

5. Models

Our goal is to predict which banks will have high risk in terms of default rates associated with agricultural loans. We separate our data into two groups: banks with high risk and those with low risk. To assess the benefits of using machine learning methods for prediction, we follow Ifft *et al.* (2018) and compare the predictive performance of several machine learning methods to that of the standard econometric techniques. We also use the same subset of literature-guided variables for all methods.

Many machine learning algorithms require tuning the parameters that determine how the model fits the data. The choice of these parameters affects the extent to which the model will under-fit or over-fit the dataset. Since underfitting or overfitting models are less likely to be generalized well on new data, researchers typically tune model parameters by measuring the out-of-sample predictive performance. Typically, a grid-search approach is employed using the relevant parameter space (Ifft *et al.*, 2018). Throughout the model selection section, we highlight several important parameters in each model. We then select the value of each parameter using cross-validation to determine the most appropriate value that leads to the best out-of-sample predictive performance.

We have chosen seven common approaches among many applicable supervised machine learning methods to explore the relative benefits of advanced techniques for agricultural loan delinquency predictions. Specifically, we predict whether a bank will have a high delinquency rate on agricultural loans based on available bank and agricultural characteristics. The chosen algorithms fall into four categories: methods based on logistic regression, Naïve Bayes classifier, ensemble methods, and support vector machine methods. We describe these selected methods below together with their strengths and weaknesses.

The first method we use is the logistic regression. The logistic regression assumes that the logarithm of the conditional likelihood ratio is a linear combination of the independent variables, which is usually fit by maximizing the likelihood function. In this study, we use a logistic regression[9] as the baseline model and compare it with the other methods. However, with a large number of independent variables that may be collinear with each other, the logistic regression would not be able to drop redundant variables or select the most predictive information, and this may lead to large sampling variance and poor out-of-sample predictive performance, i.e. an overfitting of the model. To enhance the prediction accuracy and the interpretability of the model, we add a penalty term to the likelihood function. Adding different penalty terms leads to different methods and the magnitude of the penalty is determined by the complexity parameter. With a larger complexity parameter, there will be more shrinkage in the coefficients.

[9] The logistic regression here refers to a logistic regression where we minimize a given log loss using a maximum likelihood function algorithm. The logistic regression belongs to a more general machine learning algorithm where we can change parameters to adjust the loss function. The logistic regression is a special case of the machine learning method.

The Ridge penalty is the sum of the squared coefficients, which shrinks all coefficients by the same non-zero factor (Hoerl and Kennard, 1970). While the Ridge penalty reduces the standard error of coefficients and leads to better prediction, it does not remove variables that poorly predict the outcome. On the other hand, the Lasso penalty sums the absolute value for the coefficients, which tends to result in coefficients that are zero and to remove these irrelevant variables from the model (Tibshirani, 1996). However, if a group of highly correlated variables is present, the Lasso penalty tends to only select one of them and ignore the others, and, consequently, this could result in a higher prediction error. Zou and Hastie (2005) combined the Lasso and the Ridge penalties and proposed the Elastic Net to maintain the merits of both penalties and alleviate their shortcomings. Thus, the Elastic Net method can keep all variables that belong to the same group with a more holistic view of the information associated with the separate variables.

The second method we use are the Naïve Bayes classifiers which make predictions on the classification problem based on the Bayes' theorem and assume the independence of variables (Maron, 1961). Particularly, we have chosen the Gaussian Naïve Bayes, which also uses the assumption that the likelihood of the independent variables is normally distributed (Kuhn and Johnson, 2013). It simplifies the estimation dramatically by separately estimating the individual class, i.e. conditional marginal densities, using univariate Gaussians to represent these marginals. Theoretically, the method will perform poorly if the assumptions are not fulfilled, i.e. the likelihood function of the independent variables is not normal or the variables are not independent. However, the Naïve Bayes classifiers often outperform far more sophisticated alternatives in practice. Although the individual class density estimates may be biased, the bias might not hurt the posterior probabilities much (Friedman *et al.*, 2001).

The third method we use are the ensemble models. The idea behind ensemble learning is to build a prediction model with better performance by combining the strengths of a collection of simpler base models (Opitz and Maclin, 1999). Ensemble learning first develops a population of base learners from the training data and it then combines them to form the composite model (Friedman *et al.*, 2001). We use two common ensemble approaches to compare their performance of prediction: random forest and adaptive boosting.

Bagging is an ensemble algorithm designed to improve the stability and accuracy of machine learning algorithms by lowering the variance component of prediction error. When applied to decision tree methods, the model fits multiple re-samples of the training data. The model prediction uses the mode or the mean coming from the bootstrapped samples (Breiman, 1996). However, the problem is that if single trees perform poorly on predictions, bagging will not make the combination better. The random forest method improves the bagging by randomizing the subset of variables used to build each tree, resulting in less correlation among the models for each re-sample of the dataset. When the variance component is reduced for the prediction error, the random forest method is able to improve prediction accuracy relative to the simple bagging method (Barandiaran, 1998). The random forest method depends on the number of variables which are randomly selected and used in every tree and this parameter is set to improve the model prediction.

The other ensemble technique we use is boosting. In bagging, the base models are found independently, whereas boosting applies these base models in a sequence while aiming to minimize the additional bias in every step. Boosting algorithms iteratively learn weak classifiers by adding them to a final strong classifier. When base models or weak learners are added, they are typically weighted based on their accuracy. One boosting method we include is adaptive boosting, which up-weights observations that were misclassified before and down-weights observations that were classified correctly. In this way, it leads to strong learners with weighted addition of the weak learners (Freund *et al.*, 1999).

The fourth method we use is the support vector machine which divides the dataset into two classes while it fits an optimal decision boundary that maximizes the margin that is between the support vectors and the decision boundary. In linear classification, the support vector machine method separates two classes through a linear hyperplane based on all available characteristics. In addition to performing linear classification, support vector machines can also construct non-linear classifiers using kernel techniques by mapping the

dataset into a higher-dimensional characteristics' space where a hyperplane can be found to separate the two classes (Boser *et al.*, 1992). In reality, the classes are typically not perfectly separable and the SVM method needs to balance the advantage coming from a larger margin against the costs from misclassification of some observations when the margin grows. Therefore, the tolerance of an incorrect classification of some observations is estimated as a parameter. However, SVM may have the advantage of reacting favorably to sampling and thus delivering better results compared to the conventional logistic regression in high event regimes (Zhang and Trubey, 2019).

6. Model evaluation

The focus of this study is to evaluate the performance of machine learning methods in predicting banks that have high agricultural loan risk. Therefore, we turn to model predictive performance and report the importance of different predictors contributing to the prediction.

In-sample prediction, where predictive accuracy is evaluated using the same information to fit the model, tends to result in overly optimistic accuracy estimates. When a model captures too much noisy information, also called overfitting, it will generate poorly predictive performance on different data. Therefore, our analysis focuses on out-of-sample predictions instead of in-sample predictions. Specifically, we divide the dataset into training data that we use to fit the model, and after that we use the trained model with a test dataset to assess its out-of-sample predictive accuracy.

We randomly assign observations to the training (80%) and test data (20% of the observations). Next, we repeat this process 100 times to reduce the influence of the random assignments on evaluating the model, without adding too much computational burden. These 100 measures of model performance can then be tested for significant differences across models. To make sure the comparison across the methods is calculated on the same assigned observations, statistical tests must be used on paired samples. Following Ifft *et al.* (2018), we use the Wilcoxon Signed Rank test, which is a nonparametric test applied for the matched samples, to test for differences across methods in terms of accuracy.

In alignment with forecast evaluation literature, we use several performance indicators to evaluate prediction performance across several machine learning methods. In our context, true positives are those correctly predicted observations in which a bank has an agricultural delinquency rate that is higher than 3% and true negatives are the correctly predicted observations in which a bank has a delinquency rate lower than or equal to 3%. On the other hand, false positives happen when the model has an incorrect prediction that a bank has high risk when it has low risk and false negatives happen when the model has an incorrect prediction that a bank has low risk when in fact it has high risk.

An accuracy for a model shows the percentage of correct predictions as a proportion of all predictions. Accuracy is defined as the sum of true positives and true negatives divided by all predictions. One problem with accuracy is that it weights the model's ability to identify banks with high and low financial risk equally well. Since our focus is to identify banks with high risk associated with agricultural loans ahead of time, we consider the models' ability to discern between the two types of banks. Therefore, we have recall, which is also called sensitivity, which measures the model's ability to correctly predict the actual outcome of interest (high default rate) for the observations that this outcome occurred. In our analysis, recall captures the percentage of banks that are correctly predicted as having a high delinquency rate out of all banks that have a high delinquency rate.

Recall is useful in assessing the prediction accuracy of a model, but it is conditional on the outcome of interest. Recall is defined as true positives divided by the sum of true positives and false negatives. Precision is a measure of the unconditional probability that the model prediction is correct. In the context of delinquency, precision measures the proportion of times when the model predicts a bank has a high delinquency rate and the bank does. Precision is defined as true positives divided by the sum of true positives and false positives.

An alternative way is to consider the costs of false positive and also false negative predictions. In our analysis of a bank or a regulatory agency identifying banks' delinquency performance, the cost associated with a false positive is the cost from a false warning and potentially a corresponding adjustment. On the other hand, the cost associated with a false negative is the loss due to the failure to identify high risk and not take actions ahead of time. Depending on the size of these costs, a bank or a regulatory agency choosing across these models may select a model that favors incorrect prediction either toward false positives or false negatives. Following Ifft *et al.* (2018), we use the cost adjustment term λ as a weight for inaccurate predictions. If λ is from 0 to 1, the weight of a false negative is less than that of a false positive. Alternatively, λ values above 1 indicate the weight of a false negative is more than that of a false positive. Here, C is defined as:

$$C = \lambda \times \textit{false negatives} + \textit{false positives}$$

The literature shows that simple predictive accuracy is inappropriate when the dataset is imbalanced or the costs of different errors vary significantly. Instead, indicators such as the Receiver Operating Characteristic (ROC) curve, Precision Recall (PR) curve, and corresponding Areas Under the Curve (AUC), are more accurate ways to measure a model's performance (Chawla *et al.*, 2002; Guo *et al.*, 2008). In all previous indicators, the default tradeoff is 0.5, meaning that an estimated probability between 0 and 0.5 is categorized as a negative outcome (0) and an estimated probability that is higher than 0.5 is a positive outcome (1). Instead, the model can be more flexible when categorizing the probabilities for each class.

Using the ROC curve provides flexibility to choose and calibrate the threshold for how to interpret the estimated probabilities. The ROC curve is a standard technique for summarizing model performance over a range of tradeoffs between true positive and false positive error rates (Swets, 1988). The true positive rate describes how good the model is at predicting a bank with high default risk when it actually has high risk associated with agricultural loans. The false positive rate measures how often a bank with high delinquency risk is predicted when it actually has low risk. Thus, the ROC curve shows the tradeoff between power and type I error rate of a model. The area under the curve (AUC) is an accepted traditional performance metric calculated from a ROC curve, referring to the area under the ROC curve (Bradley, 1997; Duda *et al.*, 2001). AUC of different models can be compared directly: the higher the AUC is, the better the model is.

Similar to the ROC curve, the PR curve is a plot of the precision and the recall for different thresholds. Since the calculations of precision and recall do not use true negatives, the PR curve is only concerned with the correct prediction of banks that have high delinquency risk. Saito and Rehmsmeier (2015) show that the PR curve can be more informative than the ROC curve when evaluating models on imbalanced datasets: when a dataset is imbalanced, the ROC curve could be overly optimistic with respect to conclusions about the reliability of classification performance, while the PR curve provides an accurate prediction of future classification performance. Correspondingly, the AUC of the PR curve summarizes the integral of the area under the PR curve.

The importance of each variable can be reported from tree-based methods. Feature importance is represented by a relative score based on the extent that dividing at that feature node in the decision tree increases the prediction criterion, such as the average Gini impurity metric (James *et al.*, 2013). The scores based on feature importance can be ordered using rank and examined to find the most important variables in terms of the relative importance in making predictions. Ifft *et al.* (2018) show that feature importance rankings reported by machine learning methods cannot be interpreted similarly to the statistical significance reported in a regression model, since only reflect the importance of the prediction but are not exact measures of the impact on outcomes.

7. Results

Table 2 shows the performance metrics for the methods presented in the previous section. We employ these indicators to evaluate whether each method is successful in predicting whether a bank has high agricultural risk one year later and to provide context on how a bank may determine whether to use a machine learning

method to manage associated risk ahead of time. Table 2 presents the model evaluation indicators as an average across repeated outcomes. Each method is limited to the same independent variables used in the literature, which are articulated in Table 1. Table 3 shows how a bank could use the different costs of inaccurate predictions to decide which method to use.

Based on Table 2, the logistic regression is able to correctly classify a bank as whether or not it has high agricultural loan risk in only 61% of the cases. Several machine learning techniques improve their performance compared to the base method; the improvements tend to be in a wide range, from 1.7 to 25.9 percentage points. However, recall and precision provide a better measure of a model's predictive performance to identify high agricultural risks in banks.

The logistic regression on average recalls 68% of the banks that will have high delinquency risk. Penalization techniques, including Lasso, Ridge, Elastic Net, and Gaussian Naïve Bayes, among all machine learning methods, perform better than the logistic regression. The penalization techniques in identifying high risk tend to result in small improvements, i.e. less than 1 percentage point; while the Naïve Bayes classifier can identify 12.8 percentage points more high-risk banks than the baseline model.

Due to the imbalance class issue our data exhibits, the logistic regression's prediction that a bank would have high agricultural loan risk has an average precision of only 10%. That is, among our prediction of the class of interest, only 10% are correctly predicted. Three machine learning methods, which include random forest, adaptive boosting, and support vector machine, are able to correctly predict banks as having high risk with a statistically higher precision. The random forest shows the highest relative improvement (79.8%) compared to the logistic regression.

Given the imbalance of the data issue, areas under the ROC curve and the PR curve are more accurate indicators for model comparisons, because they allow flexible tradeoff values to be considered. When considering a full range of possible tradeoff values, the logistic regression has an area under the ROC curve of 0.695, which indicates better performance than a randomly predicting model whose area is 0.5. Among all machine learning techniques, support vector machine and ensemble learning can improve significantly their performance compared to the baseline model; ensemble learning shows better performance than the support vector machine method, with relative improvement of 6 to 10%. In terms of the area under the PR

Table 2. Performance metrics by method.[1]

	Accuracy (%)	Recall (%)	Precision (%)	ROC AUC[2]	PR AUC[2]
Baseline method					
Logistic regression	61.3	67.8	10.4	0.695	0.894
Logistic regression based methods					
LASSO log. regression	60.9	68.2***	10.3	0.694	0.894
Ridge log. regression	60.5	68.6***	10.3	0.694	0.894***
Elastic net log. regression	61.2	68.0***	10.4	0.695	0.894
Naïve Bayes classifier					
Gaussian Naïve Bayes	34.1	**80.6*****	7.3	0.622	**0.909*****
Ensemble methods					
Random forest	**87.2*****	31.2	**18.7*****	**0.763*****	0.884
Adaptive boosting	84.1***	36.4	16.1***	0.734***	0.889
Other methods					
Support vector machine	63.0***	66.2	14.1***	0.714***	0.891

[1] *** metric for this method is statistically significantly better (higher) than the logistic regression at the α=0.01 level. The best (highest) value of each metric is in bold.

[2] AUC = areas under the curve; PR = precision recall curve; ROC = receiver operating characteristic curve.

curve, results show a different picture: logistic regression with Ridge penalization and Gaussian Naïve Bayes exhibit significant improvement, while ensemble learning is not significantly better than the baseline model.

The above results indicate that a bank or regulatory agency seeking to identify high risk prior to potential financial stress could benefit if using machine learning methods. They still must decide which machine learning method to utilize, especially with data that shows imbalanced characteristics and when using different indicators or different types of errors. The Gaussian Naïve Bayes algorithm has the largest average recall score and can identify around 80% of banks with potential financial stress, but it also has a lower average precision score than the ensemble learning algorithm. Therefore, the Naïve Bayes would be the preferred model if the goal is to identify potential banks with high risk, but it comes with the cost of having more false positives. However, if sending wrong signals to banks with actual low risk creates a large cost burden, the random forest would be a better candidate since it has the highest average precision score. Compared with other recent studies, our findings indicate different choices based on a bank's decision on the indicators or purpose. Zhu *et al.* (2019) build a loan default prediction model with user loan data and find that random forest algorithm outperforms logistic regression, decision tree and the SVM. This is similar to our findings on certain indicators, but our results consider more machine learning algorithms as well as more thorough model performance indicators and thus lead to a different conclusion.[10]

Another perspective of deciding between predictive recall and precision is to directly consider the costs associated with inaccurate predictions (Ifft *et al.*, 2018), as shown in Table 3. In the real world, a bank or a regulatory agency can use their actual costs for false positives and false negatives, whereas we use different λs to measure the cost associated with model inaccuracy. Given a scenario where the model incorrectly predicts that a bank would not have a high agricultural delinquency rate (a false negative) is costlier than the scenario where the model incorrectly predicts that a bank will have a high agricultural delinquency rate (a false positive), λ value that is larger than 1 could be more realistic in our context. Therefore, we use seven λ values ranging from 1 to 1000. Our results in Table 3 show that the relative advantage of different machine learning methods depends on the specific λ values. When λ is less than or equal to 10, indicating that a false negative is no more than 10 times costlier of a false positive, ensemble learning and support vector machine lead to a lower cost than the logistic regression. When the λ value is equal to or larger than 50, indicating that a false negative is considered costlier than a false positive, Gaussian Naïve Bayes and logistic regression with penalization can significantly reduce the cost compared to the baseline model.

Beyond providing a potential to improve predictive performance, some machine learning techniques gather data on the variables most helpful for prediction, as shown in Table 4, which is from one realized model of the random forest method. Mean decrease in Gini is the average of a variable's total decrease in node impurity, weighted by the proportion of samples reaching that node in each individual decision tree in the random forest. A higher mean decrease in Gini indicates higher variable importance.

Our results show that the three most important variables are the proportion of agricultural loans to total loans, the ratio of net income to net operating income, and the interest rate in the prior year. An important finding for the feature importance is that none of the agricultural characteristics at the state level is among the top 10 variables most important for the predictions. This indicates that bank-level financial and operational conditions and macroeconomic indicators, such as interest rate, are more important than agricultural characteristics in terms of predicting the delinquency risk of a bank. A low feature importance score for a variable does not mean that the variable is not associated with high delinquency risk. The variable may be highly correlated with another independent variable and thus not add much to the predictive power.

[10] This study also uses the SMOTE technique to deal with the imbalance data issue. However, when comparing model performance, it does not consider the PR indicator, which could be more informative than the ROC curve when evaluating models on imbalanced datasets.

Table 3. Relative costs of false negatives and false positives.[1]

	λ value[2]						
	1	**5**	**10**	**50**	**100**	**500**	**1000**
Baseline method							
Logistic regression	9,498	11,478	13,953	33,753	58,503	256,503	504,003
Logistic regression based methods							
LASSO log. regression	9,598	11,555	14,001	33,566***	58,022***	253,674***	498,239***
Ridge log. regression	9,694	11,629	14,048	33,396***	57,581***	251,065***	492,920***
Elastic net log. regression	9,527	11,496	13,956	33,641***	58,247***	255,095***	501,155***
Naïve Bayes classifier							
Gaussian Naïve Bayes	16,173	17,366	18,857	**30,785*****	**45,695*****	**164,975*****	**314,075*****
Ensemble methods							
Random forest	**3,149*****	**7,383*****	**12,675*****	55,011	107,932	531,296	1,060,501
Adaptive boosting	3,909***	7,823***	12,715***	51,851	100,772	492,136	981,341
Other methods							
Support vector machine	5,908***	8,903***	13,245***	46,206	94,207	410,490	837,017

[1] *** metric for this method is statistically significantly better (lower) than the logistic regression at the α =0.01 level. The best (smallest) value of each metric is in bold.

[2] λ value is equal to or larger than 1.

Table 4. Top 10 feature importance, random forest.

Dimension	**Variables**	**Mean decrease Gini**
Asset quality	agricultural loans / total loans	11,914.813
Management	net income / net operating income	6,147.437
Macro indicators	interest rate	4,797.108
Earning ability	total interest income / average earnings assets	4,780.094
Other financial	market share	4,516.130
Earning ability	total interest expenses / average earning assets	4,385.970
Earning ability	interest income / gross loans	3,800.585
Asset quality	loan unearned income / total equity	3,673.076
Liquidity position	(cash + FedFunds) / total assets	3,472.261
Earning ability	efficiency ratio	3,274.823

8. Conclusions

We consider seven common machine learning approaches and compare their prediction performances with that of a logistic regression. Our results show that among all machine learning methods, random forest has the highest predictive accuracy, precision, and area under the ROC curve, while the Naïve Bayes classifier has the highest recall and area under the PR curve among all methods employed. Some machine learning methods, such as Lasso, Ridge, and Elastic net logistics, do not show significant improvement in most performance metrics compared to the logistic model, except in the performance of the recall indicator.

In general, a bank or regulatory agency looking to identify high risk prior to potential financial stress may benefit by using machine learning techniques, compared to the use of traditional econometric models. However, the bank or regulatory agency still has to decide which machine learning method to use, especially when the information shows imbalanced characteristics and when they focus on different indicators. For example, the Naïve Bayes classifier would be a preferred model if the goal is to identify potential banks with high risk,

even though it comes with the cost of having a relatively many false positives; ensemble method would be a better candidate if sending wrong signals to banks with actual low risk creates a large cost burden. Open-source statistical packages such as R and Python can implement these machine learning methods. If a bank analyst can implement these machine learning methods, she/he can also replicate the approach from this study, applying their own cost information as appropriate, with very little additional cost.

In our context, a scenario where the model incorrectly predicts that a bank would not have a high agricultural delinquency rate (a false negative) is costlier than a scenario where the model incorrectly predicts that a bank will have a high agricultural delinquency rate (a false positive). After we take the costs of false negatives and false positives into consideration and give two kinds of incorrect predictions different weights, we find that how we weigh them matters: ensemble learning is the best when the cost of a false negative is not more than 10 times of that of a false positive; while Gaussian Naïve Bayes is the best when the cost of a false negative is much larger. However, a bank may have its own cost information and using these techniques can get different results. Additionally, our results also imply that bank-level financial and operational conditions and macroeconomic indicators, such as interest rate, are more important than the publicly available aggregate agricultural characteristics when predicting the delinquency risk of a bank.

From the perspective of the bank, the machine learning methods applied in this study provide a more accurate technique framework to identify the risk of agricultural loans, so that the bank can take actions ahead of time, such as holding reserves for a potential crisis, or adjusting the proportion of agricultural loans versus other types of loans to reduce total risk. In terms of implementation of machine learning in the context of big data analysis, this study recommends a conservative perspective: not all advanced techniques show significant improvement, and the cost of advanced methods could be high. Since the improvement of prediction depends very much on the dataset, our results using aggregate-level data may not be able to be applied directly with customer- or portfolio-level data, which need further exploration for policy implications.

There are some limitations in this study. First, we could not find borrower or farm level data that can be merged with our bank level data since customer information is usually confidential. Therefore, debt-to-asset ratios, insurance coverage, and other financial risk indicators, which can typically be used to assess the borrowers' ability to meet their loan repayment obligations, could not be included in our models. Second, other information like crop prices and yields may be good predictors in our models. However, we do not have access to such data in this study. We believe that if such data were available for bank customers and borrowers and for the local agricultural economy, banks would be able to make more accurate predictions using our approach. Lastly, when transforming the delinquency rate into a classification problem using 3% as a threshold, we lose information on the banks' delinquency rate which creates the imbalance problem that we address through SMOTE. An alternative way is to use continuous models to predict the delinquency rate in order to assess financial stress. Future research would need to address this issue and to compare the advantages and disadvantages of different modeling choices.

Supplementary material

Supplementary material can be found online at https://doi.org/10.22434/IFAMR2020.0019

Table S1. Correlation matrix of selected variables.

References

Barandiaran, I. 1998. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20(8): 832-844.

Barr, R.S. and T.F. Siems. 1994. *Predicting bank failure using DEA to quantify management quality.* Financial Industry Studies Working Paper 94-1, Federal Reserve Bank of Dallas, Dallas, TX, USA.

Bloniarz, A., H. Liu, C.H. Zhang, J.S. Sekhon and B. Yu. 2016. Lasso adjustments of treatment effect estimates in randomized experiments. *Proceedings of the National Academy of Sciences* 113(27): 7383-7390.

Boser, B.E., I.M. Guyon and V.N. Vapnik. 1992. *A training algorithm for optimal margin classifiers*. In: Proceedings of the 5th Annual Workshop on Computational Learning Theory. Pittsburgh, PA, USA, pp. 144-152.

Bradley, A.P. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition* 30(7): 1145-1159.

Breiman, L. 1996. Bagging predictors. *Machine learning* 24(2): 123-140.

Burnham, K.P. and D.R. Anderson. 2002. *A practical information-theoretic approach*. *Model selection and multimodel inference*, 2nd edition. Springer, New York, NY, USA.

Chawla, N.V., K.W. Bowyer, L.O. Hall and W.P. Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* 16: 321-357.

Cole, R.A. and J.W. Gunther. 1998. Predicting bank failures: a comparison of on-and off-site monitoring systems. *Journal of Financial Services Research* 13(2): 103-117.

Cole, R.A. and L.J. White. 2012. Déjà vu all over again: the causes of US commercial bank failures this time around. *Journal of Financial Services Research* 42(1-2): 5-29.

Demyanyk, Y. and I. Hasan. 2010. Financial crises and bank failures: a review of prediction methods. *Omega* 38(5): 315-324.

Dinterman, R., A.L. Katchova and J.M. Harris. 2018. Financial stress and farm bankruptcies in US agriculture. *Agricultural Finance Review* 78(4): 441-456.

Duda, R.O., P.E. Hart and D.G. Stork. 2001. Pattern classification. *International Journal of Computational Intelligence and Applications* 1: 335-339.

Ekinci, A. and H.İ. Erdal. 2017. Forecasting bank failure: base learners, ensembles and hybrid ensembles. *Computational Economics* 49(4): 677-686.

Freund, Y., R. Schapire and N. Abe. 1999. A short introduction to boosting. *Journal of the Japanese Society for Artificial Intelligence* 14(771-780): 1612.

Friedman, J., T. Hastie and R. Tibshirani. 2001. *The elements of statistical learning.* Vol. 1. Springer, New York, NY, USA.

Guo, X., Y. Yin, C. Dong, G. Yang and G. Zhou. 2008. *On the class imbalance problem*. In: ICNC 2008: 4th International Conference on Natural Computation. 25-27 August 2007. Jinan, China, pp. 192-201.

Hoerl, A.E. and R.W. Kennard. 1970. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 12(1): 55-67.

Ifft, J.E., R. Kuhns and K.T. Patrick. 2018. Can machine learning improve prediction an application with farm survey data. *International Food and Agribusiness Management Review* 7: 1-16.

Iturriaga, F.J.L. and I.P. Sanz. 2015. Bankruptcy visualization and prediction using neural networks: a study of US commercial banks. *Expert Systems with Applications* 42(6): 2857-2869.

James, G., D. Witten, T. Hastie and R. Tibshirani. 2013. *An introduction to statistical learning.* Vol. 112. Springer, New York, NY, USA.

Katchova, A.L. and R. Dinterman. 2018. Evaluating financial stress and performance of beginning farmers during the agricultural downturn. *Agricultural Finance Review* 78(4): 457-469.

Komarek, A.M., A. De Pinto and V.H. Smith. 2020. A review of types of risks in agriculture: what we know and what we need to know. *Agricultural Systems* 178: 102738.

Kuhn, M. and K. Johnson. 2013. *Applied predictive modelling.* Vol. 26. Springer, New York, NY, USA.

Li, X., C.L. Escalante and J.E. Epperson. 2014. *Agricultural banking and bank failures of the late 2000s financial crisis: a survival analysis using Cox Proportional Hazard Model*. Selected paper prepared for presentation at the Southern Agricultural Economics Association (SAEA) Annual Meeting. 1-4 February 2014. Dallas, TX, USA.

Maron, M.E. 1961. Automatic indexing: an experimental inquiry. *Journal of the ACM* 8(3): 404-417.

Morehart, M., D. Milkove and Y. Xu. 2014. *Multivariate farm debt imputation in the agricultural resource management survey (ARMS).* Selected poster prepared for presentation at the Agricultural & Applied Economics Association's 2014 AAEA Annual Meeting. July 27-29, 2014. Minneapolis, MN, USA.

Mullainathan, S. and J. Spiess. 2017. Machine learning: an applied econometric approach. *Journal of Economic Perspectives* 31(2): 87-106.

Opitz, D. and R. Maclin. 1999. Popular ensemble methods: an empirical study. *Journal of Artificial Intelligence Research* 11: 169-198.

Ozturk, H., E. Namli and H.I. Erdal. 2016. Reducing overreliance on sovereign credit ratings: which model serves better? *Computational Economics* 48(1): 59-81.

Pantalone, C.C. and M.B. Platt. 1987. Predicting commercial bank failure since deregulation. *New England Economic Review* 7: 37-47.

Paulson, N., X. Li, C.L. Escalante, J.E. Epperson and L.F. Gunter. 2013. Agricultural lending and early warning models of bank failures for the late 2000s Great Recession. *Agricultural Finance Review* 73(1): 119-135.

Quaye, F., D.A. Nadolnyak and V. Hartarska. 2017. Factors affecting farm loan delinquency in the Southeastern USA. *Research in Applied Economics* 9(4): 75-92.

Saito, T. and M. Rehmsmeier. 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE* 10(3): e0118432.

Swets, J.A. 1988. Measuring the accuracy of diagnostic systems. *Science* 240(4857): 1285-1293.

Tack, J., K.H. Coble, R. Johansson, A. Harri and B. Barnett. 2017. The potential implications of 'Big Ag Data' for USDA forecasts. SSRN 2909215. https://dx.doi.org/10.2139/ssrn.2909215

Tam, K.Y. and M.Y. Kiang. 1992. Managerial applications of neural networks: the case of bank failure predictions. *Management Science* 38(7): 926-947.

Thomson, J.B. 1991. Predicting bank failures in the 1980s. *Federal Reserve Bank of Cleveland Economic Review* 27(1): 9-20.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B* 58(1): 267-288.

Wager, S. and S. Athey. 2017. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association* 113(523): 1228-1242.

Woodard, J.D. 2016. Data science and management for large scale empirical applications in agricultural and applied economics research. *Applied Economic Perspectives and Policy* 38(3): 373-388.

Zhang, Y. and P. Trubey. 2019. Machine learning and sampling scheme: an empirical study of money laundering detection. *Computational Economics* 54(3): 1043-1063.

Zhao, D., C. Huang, Y. Wei, F. Yu, M. Wang and H. Chen. 2017. An effective computational model for bankruptcy prediction using kernel extreme learning machine approach. *Computational Economics* 49(2): 325-341.

Zhao, H., A.P. Sinha and W. Ge. 2009. Effects of feature construction on classification performance: an empirical study in bank failure prediction. *Expert Systems with Applications* 36(2): 2633-2644.

Zhu, L., D. Qiu, D. Ergu, C. Ying and K. Liu. 2019. A study on predicting loan default based on the random forest algorithm. *Procedia Computer Science* 162: 503-513.

Zou, H. and T. Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B* 67(2): 301-320.