Agricultural Loan Delinquency Prediction Using Machine Learning Methods

Jian Chen[1], Ani Katchova[2], Chenxi Zhou[3]

Selected Paper prepared for presentation at the 2019 Agricultural & Applied Economics Association Annual Meeting, Atlanta, GA, July 21-23

[1] Ph.D. researcher, Department of Agricultural, Environmental, and Development Economics, The Ohio State University. chen.4797@buckeyemail.osu.edu.

[2] Associate Professor and Farm Income Enhancement Chair, Department of Agricultural, Environmental, and Development Economics, The Ohio State University. katchova.1@osu.edu.

[3] Ph.D. researcher, Department of Statistics, The Ohio State University.

Agricultural Loan Delinquency Prediction Using Machine Learning Methods

Abstract

The recent downturn in the agricultural sector since 2013, measured by declines in net farm income, land values, and prices of key agricultural products, has caused some concerns for farmers' repayment capacity. This, in turn, raises the need for precise prediction of financial stress in the agricultural sector. Machine learning has the potential to improve the predictions with large financial datasets. Yet, the application of machine learning remains limited in the agricultural sector. In this article, we approximate financial stress by agricultural loan delinquency, and predict it by employing both a standard logistic regression and several machine learning methods. The main datasets include the Call Reports and Summary of Deposits from Federal Deposit Insurance Corporation (FDIC). Our results show that ensemble learning methods can improve the prediction accuracy by 26 percentage points at most. Our results also show 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 bank is very costly. From the perspective of banks, while there are potentially significant benefits to machine learning in terms of changes in practice beforehand, research objectives and bank-level costs are important considerations that in some cases may favor different choices of machine learning methods.

Keywords: agricultural loan delinquency, prediction, machine learning, logistic regression

Recent trends of declining farm incomes and commodity prices have brought concerns that the agricultural downturn may significantly affect farmers' repayment capacity, therefore posing concerns for the agricultural sector. Prices for key agricultural products have fallen since 2013, while input costs have not fallen by much. Consequently, farmers have had lower revenues and thinner profit margins. Even though the debt-to-asset ratio at the farm level remains low—between 11 and 15 percent from 2000 until 2015—compared with over 20 percent during the 1980s farm crisis, cash flow is of particular importance and points to potential issues with liquidity and repayment capacity (Katchova and Dinterman, 2018). In this environment, it has become increasingly important to understand the causes of financial stress in the agricultural sector and to predict their occurrence. Obtaining precise predictions of loan delinquencies would avoid unnecessary concerns and allow both lenders and farmers to adopt appropriate financial management strategies to ensure the long-term survival of farm businesses.

In this article, we approximate financial stress by agricultural loan delinquencies and consider the potential of machine learning to improve the prediction of delinquency among agricultural loans, mainly using the Consolidated Reports of Condition and Income (i.e. Call Reports) published by the Federal Deposit Insurance Corporation (FDIC). 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. The dataset has been commonly used by the public, bank rating agencies and researchers for analyses, regulatory purposes and policy recommendations.

Large datasets, like FDIC Call Reports, provide detailed information for each observation and would benefit the analysis only if used appropriately. Tack et al. (2017) suggest that "Big Data" can improve the United States Department of Agriculture (USDA) forecasts. Yet, as Ifft et al.

(2018) point out, large public, private, and administrative datasets in the agricultural sector, face some challenges. These challenges include management and warehousing of the data (Woodard, 2016), maintaining its privacy and security (Sykuta, 2016), and most importantly, the need to adapt to methodological advances. Many standard econometric methods are not designed to take advantage of large datasets with detailed information (Ifft et al., 2018). They impose strong assumptions on the data generating process. They are also 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 methods, have the potential to substantially improve the research value of large datasets. Most financial datasets with detailed information typically exhibit non-linear relationships and high correlation among hundreds of variables. Given these features in a financial context, machine learning methods have been applied to crisis prediction and have turned out to be superior to the traditional methods for the purpose of prediction and corresponding policy making (Demyanyk and Hasan, 2010; Du Jardin, 2010; Iturriaga et al., 2015). However, studies in the agricultural sector using improved empirical techniques remain limited, especially from the perspective of banks.

In order to fill this gap, we follow Ifft et al. (2018)'s framework to set up a machine learning problem of delinquency on agricultural loans, evaluate different machine learning methods, and compare the machine learning results to prediction using the standard econometric models. We also address common concerns regarding machine learning, such as over-fitting of the data, and explain how well-established methods can be used to mitigate these concerns. Finally, by employing the cost-based evaluation approach established by Ifft et al. (2018), we analyze the benefits and drawbacks of different machine learning methods, based on the costs associated with incorrect predictions faced by banks or regulatory agencies. We conclude with a discussion of insights for machine learning methodology and big data analysis in the area of agricultural prediction and other important considerations for lenders and researchers considering the use of these advanced techniques.

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 logit or probit model, the duration model has the capability to generate estimates of the probable time of 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 methods often impose strong assumptions on the data-generating process and linear relationships between dependent and independent variables. Also, they have limitations when dealing with outliers, non-linear relationships and variable selection, etc. Particularly, they hardly address situations in which independent variables are highly correlated, leading to large variance and thus large prediction errors for the model (Friedman et al., 2001).

Studies have shown that these methodological issues may benefit from machine learning methods. Mullainathan and Spiess (2017) argue that machine learning can generally improve the performance of a variety of econometric methods through variable selection, 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 (Ifft et al., 2018). In order to establish the causal effect, Wager and Athey (2017) develop a causal forest for estimating heterogeneous treatment effects based on the random forest algorithm, a machine learning technique.

Most importantly, by being flexible with such common data issues, machine learning methods have been demonstrated 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 et al. (2015) employ a specific machine learning technique, the neural network method, to detect commercial bank failure in the U.S., and demonstrate that their model outperforms the traditional models of bankruptcy prediction with a 96% accuracy rate.

In the agricultural sector, there are studies on financial stress from the perspective of farms (Dinterman et al., 2018; Quaye et al., 2017), but limited number of studies have delved into 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. And they find that farmers with larger farms, who have more insurance, higher net income, lower debt-to-asset ratio, single loans and those that 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 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 standard logit model 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, investigating the risk from the perspective of banks using the aggregate-level data; thus, the purpose of our paper is to fill this gap.

Data Description

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 involve 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. Therefore, we use the default rate of 3% as the cut-off value for high risk associated with agricultural loans and transform the delinquency rate into a classification problem in the prediction framework.

Other information used for predicting financial stress from the Call Reports include bank-level financial and operational variables. For example, traditional proxies for the CAMELS, 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). Other variables include bank size and market share.

Agricultural characteristics, such as farm net income and interest expense, come from Farm Income and Wealth Statistics (Dinterman et al., 2018). Macroeconomic indicators, such as unemployment rate and interest rate from Federal Reserve Economic Data and Bureau of Labor Statistics, are also thought 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 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 a bank's branches using as weights the amount of deposits owned by each branch.

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 & Siems, 1994; Li et al., 2014; Tam & Kiang, 1992, Zhao et al. 2009). While some studies have built models with longer prediction periods, such models are often not accurate enough to be practically useful and will not be examined 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 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 sample. Table 1 provides the definition 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, these characteristics of imbalance would be problematic if using the standard method for prediction.

[Table 1 here]

Imbalance Dataset and SMOTE Technique

The imbalance problem has been recognized in many fields (Guo et al., 2008). A dataset is imbalanced if a majority of the observations are labeled as one class while only a few observations are labeled as the other class, usually the more important class or the class of most interest. It is also the case that the cost of misclassifying an interesting (minority) observation as a normal (majority) observation is often much higher than the cost of the reverse error. Since traditional algorithms seek an accurate performance over a full range of observations, they tend to be overwhelmed by the majority class and ignore the minority class. Therefore, standard machine learning techniques do not work well when applied to an imbalanced class dataset.

In our case, while 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%). However, these banks is the class of our most interest since it is important to identify them and take measures ahead of time. And 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 such class imbalance problem in our dataset, we have chosen to change class distribution by using over-sampling techniques. One of the most common over-sampling techniques is Synthetic Minority Over-Sampling Technique (SMOTE), which generates synthetic minority observations to over-sample the minority class (Chawla et al., 2002). Specifically, it forms new minority class observations by interpolating between several existing minority class observations using the *k*-nearest neighbors 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).

Method Selection and Evaluation

Our goal is to predict the banks that 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 evaluate the potential benefits of applying machine learning techniques to prediction, we compare the predictive performance of several machine learning methods to that of the standard econometric technique. 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 affect the degree to which a model under- or over-fits the data. Since under- or over-fitting models are less likely to be generalized well on new data, researchers typically tune model parameters by measuring the out-of-sample predictive performance. In practice, a grid-search approach is employed over the relevant parameter space. 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.

A List of Methods

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. We describe these selected methods below together with their potential strengths and weaknesses.

Logistic Regression Based Methods. Standard logistic regression is widely used to model a binary dependent variable and assumes 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 article, we use basic logistic regression as the baseline model and compare with the other methods. However, one drawback of the basic logistic regression is that it suffers from the overfitting issue. In order 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 strength of the penalization is dictated by the complexity parameter. The larger the complexity parameter is, the more shrinkage of the coefficients there is.

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 is the sum of the absolute value of the coefficients, using which is more likely to zero out coefficients and effectively removes those irrelevant variables from the model (Tibshirani, R., 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, as a consequence, could result in a higher prediction error. Zou and Hastie (2005) combined the Lasso and the Ridge penalties and proposed the Elastic Net, in order to maintain the merits of both penalties and alleviate their shortcomings.

Naïve Bayes Classifier. Naïve Bayes classifiers make predictions on classification problem based on the Bayes' theorem and assume the independence of variables (Maron, 1961). Particularly, we have chosen the Gaussian Naïve Bayes, which further assumes that the likelihood of independent variables follows the normal distribution (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 not perform well if assumptions do not meet, i.e. variables are not independent or the likelihood function of variables is not normally distributed. However, 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 as much (Friedman et al., 2001).

Ensemble Methods. 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; 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 reducing the variance component of prediction error. When applied to decision tree methods, the model fits multiple re-samples of the training data. The mean or mode of the individual bootstrap samples are then used as the model’s prediction (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 takes the concept of bagging a step further by randomizing the subset of variables used to build each tree, resulting in less correlation between the models on each re-sampled dataset. By reducing the variance component of prediction error, the random forest method is able to improve prediction accuracy relative to the simple bagging method (Barandiaran, I., 1998). The random forest method tends to be sensitive to the number of randomly selected variables used to create each tree and this parameter is commonly tuned to improve its predictive performance.

The other ensemble technique we use is boosting. Unlike bagging where the base models are learned independently, boosting applies the base models sequentially while seeking to minimize added bias at each 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).

Support Vector Machine. The support vector machine algorithm splits the data into two classes by fitting an optimal decision boundary that maximizes the margin between the decision boundary and support vectors. 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 practice, the outcome classes are often not completely separable and the SVM algorithm must trade off the benefits of a larger margin with the cost of misclassifying some existing observations as the margin increases. Therefore, the tolerance for classifying existing observations incorrectly is treated as a parameter.

Model Evaluation

In many econometric analyses, the goal is statistical inference rather than prediction. Thus the focus is economic interpretation along with statistical significance of estimated regression coefficients. Yet, the focus of this study is to evaluate the performance of machine learning methods in predicting banks that have high agricultural loan risk.

Model evaluation strategy. 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 other data. So we based our analysis on out-of-sample instead of in-sample predictions. Specifically, we split the original dataset into a "training data" set used to fit the model, and later apply the trained model to the "test data" in order to evaluate its out-of-sample predictive accuracy.

We randomly assign observations to the training (80 percent) and test data (20 percent of the observations). Next, we repeat this process 100 times, in order to reduce the impact of the random assignments on model evaluation, without adding too much computational burden. The resulting 100 measures of model performance can be compared to gauge whether differences across the estimated models are statistically significant. In order to make sure the comparison across each method is calculated on the same assigned observations, statistical tests must be used on paired samples. We use the Wilcoxon Signed Rank test, a nonparametric test for matched samples, to compare statistical differences across methods in terms of accuracy, without making assumptions about each performance metric's distribution.

Model Performance Metrics. In alignment with forecast evaluation literature, we use several performance indicators to evaluate prediction performance across machine learning methods. In our context, true positives are the correctly predicted cases where a bank has an agricultural delinquency rate that is higher than 3% and true negatives are the correctly predicted cases where the bank has a delinquency rate lower than or equal to 3%. Alternatively, false positives occur when the model incorrectly predicts a bank with high risk when it has low risk and false negatives occur when the model incorrectly predicts that a bank has low risk when in fact it has high risk.

A model's accuracy shows the percent of correctly predicted outcomes out of all the predictions.

$$Accuracy = \frac{True\ positives + True\ negatives}{All\ predictions}$$

One problem with accuracy is that it weights a 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, also referred to as sensitivity, which measures a model's ability to correctly predict the event of interest (high default rate) having occurred in the sample of observations where the event actually occurs. In our analysis, recall captures the percent of banks that are correctly predicted as having a high delinquency rate out of all banks that have a high delinquency rate.

$$Recall\ (sensitivity) = \frac{True\ positives}{True\ positives + False\ negatives}$$

While recall is useful in assessing a model's prediction accuracy, it is conditioned on the event of interest. Precision is a measure of the unconditional probability of a model's prediction being correct. In the context of delinquency, precision measures the percent of times the model predicts a bank has a high delinquency rate and the bank actually does.

$$Precision\ (Positive\ predictive\ value) = \frac{True\ positives}{True\ positives + False\ positives}$$

An alternative way is to consider the costs of false negative and false positive predictions, which can be different in practice. In our analysis of a bank or a regulatory agency identifying banks' delinquency performance, the cost of a false positive would be the cost from a false warning and potentially a corresponding adjustment. Alternatively, the cost of a false negative is the loss due to 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 between models may prefer a model that tilts its incorrect predictions toward false negatives or false positives. Following Ifft et al. (2018), we use the cost adjustment term λ to weight inaccurate predictions. If λ is between 0 and 1, the weight of a false negative is lower than that of a false positive. On the other hand, λ values above 1 signal that the weight of a false negative is larger than that of a false positive.

$$\mathrm{C} = \lambda * False\ negatives + False\ positives$$

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 the Area 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 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 the 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 involves high risk associated with agricultural loans. The false positive rate measures how often a bank with high delinquency risk is predicted when it actually involves 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 (Duda, Hart, & Stork, 2001; Bradley, 1997; Lee, 2000). 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.

Feature Importance

The importance of each variable can be reported from tree-based methods. Feature importance is a relative score determined by the amount that splitting at that feature node in the decision tree improves some predictive metric, such as the average Gini impurity metric (James et al., 2013). Feature importance scores can be ranked and compared to determine 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 in the same way as statistical significance reported in a standard regression model, since they only measure the importance to prediction and are not a precise measure of impact on outcomes. For example, a variable with low feature importance score is a bad predictor only because it is highly correlated with another variable and thus does not add much to the prediction. However, this does not mean that the variable is not a determinant of the outcome.

Results

Table 2 shows the performance metrics described in the evaluation section for each method. We employ these aforementioned indicators to evaluate the success of each method in predicting whether a bank involves high agricultural risk 1 year later and to provide context on how a bank may determine whether to use a machine learning method instead of a standard econometric approach to manage associated risk ahead of time. Table 2 summarizes the model evaluation indicators by averaging across the repeated outputs. Each method is limited to the same independent variables used in the literature, which are articulated in Table 1. Table 3 illustrates how a bank could use its knowledge of the differential costs of inaccurate predictions to decide which method to use.

[Table 2 here]

Based on Table 2, the standard logistic regression is able to correctly classify a bank as involving high agricultural loan risk or not in just 61% of cases. When restricted to using the same data, some but not all the machine learning techniques improve compared to the base method; the improvements tend to be in a variety of ranges, from 1.7 to 25.9 percentage points. However, given the goal of identifying high agricultural risk in banks, recall and precision provide a better vision of a model's predictive performance.

The standard logistic regression is able to recall on average 68% of the banks that will involve high delinquency risk. Penalization techniques, including Lasso, Ridge, and Elastic Net, and Gaussian Naïve Bayes, among all the other machine learning methods, are able to significantly outperform the standard logistic regression. The improvement of penalization techniques in identifying high risk tend to be small, i.e. less than 1 percentage point; while 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 standard logistic regression's prediction that a bank would involve high agricultural loan risk has an average precision of only 10%. That is, among our prediction of the class of interest, just 10% are correctly labeled. Three of the seven machine learning algorithms tested, including random forest, adaptive boosting, and support vector machine, are able to correctly label banks as involving high risk with a statistically significant improvement in precision. The random forest shows the highest relative improvement (79.8%) compared to standard logistic regression.

Given the imbalance data issue, areas under the ROC curve and the PR curve are more accurate indicators for model comparison, because they allow flexible tradeoff values to be considered. When considering a full range of possible tradeoff values, the standard logistic regression has area under the ROC curve as 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 compared to the baseline model; ensemble learning shows better performance than support vector machine, with relative improvement 6% - 10%. In terms of the area under the PR curve, results show a different picture: logistic regression with Ridge penalization and Gaussian Naïve Bayes exhibit significant improvement, while ensemble learning is not better than the baseline model.

The above results indicate a bank or regulatory agency looking to identify high risk prior to potential financial stress could benefit from the use of machine learning methods. However, they still have to determine which machine learning approach to use, especially when the data shows imbalanced characteristics and when we focus on different indicators or different types of errors. Gaussian Naïve Bayes algorithm has the highest average recall score, identifying around 8 out of every 10 banks with potential financial stress, but it also has a lower average precision score than ensemble learning algorithm. This would make the Naïve Bayes a potentially good choice if the focus is identifying potential banks with high risk, even at the cost of having a relatively high number of false positives. On the other hand, 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.

An alternative perspective of weighting the trade-off 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 might use estimates of its actual costs associated with false positives and false negatives, whereas we employ different values of λ to arrive at a measure of the cost of each model's inaccuracy. Given that the 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 larger than 1 could be more realistic in our context. Therefore, we use seven λ values ranging from 1 to 1,000. Our results in Table 3 show that the relative advantage of different machine learning methods depends on specific λ values. When λ is smaller 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 lower cost than the logistic regression; when λ value is equal to or larger than 50, indicating that a false negative is weigh costlier than a false positive, Gaussian Naïve Bayes and logistic regression with penalization can significantly reduce cost than the baseline model.

[Table 3 here]

Beyond providing the potential to improve predictive performance, some machine learning techniques have the ability to gather information on the variables most useful 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.

[Table 4 here]

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. Perhaps one of the most interesting findings related to feature importance is that none of the agricultural characteristics at the state level is showing 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 status on the farmers' side in terms of predicting the delinquency risk of a bank. As discussed in the model evaluation section, a variable that has a low feature importance score does not necessarily indicate that the variable is not related to high delinquency risk. The variable may just be highly correlated with another variable and therefore not add much information to predictive power.

Conclusion

Recent downturns in the agricultural sector since 2013 have caused some concerns for farmers' repayment capacity and thus raise the need for precise prediction of financial stress in the agricultural sector. Machine learning has the potential to improve the predictions with large financial datasets. Yet, application of machine learning remains limited in the agricultural sector. In this article, we fill this gap by exploring the potential of machine learning to improve the prediction of delinquency rate among agricultural loans. We use data from FDIC Call Reports, Summary of Deposits, Farm Income and Wealth Statistics, Federal Reserve Economic Data, and Bureau of Labor Statistics, and choose 7 common machine learning approaches and compare their prediction performances to the standard logistic regression. Focusing on out-of-sample prediction, we evaluate the models' capacity in predicting the likelihood of a bank having a high delinquency rate one year later. Aligning with the forecast evaluation literature, we use performance metrics such as prediction accuracy, recall, precision, areas under the ROC and PR curves to evaluate the prediction performance across models.

Our results show that among all machine learning methods, random forest has the highest predictive accuracy, precision, and area under the ROC curve, and 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 of the performance metrics compared to the logistic model, except in the performance of recall indicator.

In general, a bank or regulatory agency looking to identify high risk prior to potential financial stress could benefit from the use of machine learning techniques, compared to the use of traditional econometric models. However, the bank or regulatory agency still has to determine which machine learning approach to use, especially when the data shows imbalanced characteristics and when they focus on different indicators. For example, Naïve Bayes classifier is a potentially good choice if the focus is identifying potential banks with high risk, even at the cost of having a relatively high number of false positives; ensemble method would be a better candidate if sending wrong signals to banks with actual low risk creates a large cost burden.

In our context, the 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). 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. Additionally, our results also imply that bank-level financial and operational conditions and macroeconomic indicators, such as interest rate, could be more important than agricultural status on the farmers' side when predicting the delinquency risk of a bank.

From the perspective of the bank, this article provides a more accurate technique framework to identify the risk for agricultural loans, so that the bank can take actions ahead of time, such as holding reserves for potential crisis, or adjusting the proportion of agricultural loans versus other types of loans to reduce total risk. In terms of the implementation of machine learning in the context of big data analysis, the article 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 on the aggregate-level data may not be able to be applied directly for customer- or portfolio- level data, which needs further exploration for policy implications.

References

Barandiaran, I., 1998. The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence* 20(8).

Boser, B.E., Guyon, I.M. and Vapnik, V.N., 1992. A training algorithm for optimal margin classifiers. *In Proceedings of the fifth annual workshop on Computational learning theory*:144-152, ACM.

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

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

Cole, R.A. and Gunther, J.W., 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 White, L.J., 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 Hasan, I., 2010. Financial crises and bank failures: A review of prediction methods. *Omega* 38(5): 315-324.

Du Jardin, P., 2010. Predicting bankruptcy using neural networks and other classification methods: The influence of variable selection techniques on model accuracy. *Neurocomputing* 73(10-12): 2047-2060.

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

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*: 1189-1232.

Friedman, J., Hastie, T. and Tibshirani, R., 2001. *The elements of statistical learning*, vol. 1. New York: Springer series in statistics.

Guo, X., Yin, Y., Dong, C., Yang, G. and Zhou, G., 2008. On the class imbalance problem. *In 2008 Fourth international conference on natural computation*, vol. 4: 192-201, IEEE.

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

Ifft, J.E., Kuhns, R. and Patrick, K.T., 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 Sanz, I.P., 2015. Bankruptcy visualization and prediction using neural networks: A study of US commercial banks. *Expert Systems with applications* 42(6): 2857-2869.

James, G., Witten, D., Hastie, T. and Tibshirani, R., 2013. *An introduction to statistical learning*, vol. 112. New York: springer.

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

Leo, B., Friedman, J.H., Olshen, R.A. and Stone, C.J., 1984. Classification and regression trees. *Wadsworth International Group* 37(15): 237-251.

Li, X., Escalante, C.L. and Epperson, J.E., 2014. Agricultural banking and bank failures of the Late 2000s financial crisis: A Survival analysis using Cox proportional Hazard Model. No. 1374-2016-109396: 3.

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

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

Opitz, D. and Maclin, R., 1999. Popular ensemble methods: An empirical study. *Journal of artificial intelligence research* 11: 169-198.

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

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

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

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

Zou, H. and Hastie, T., 2005. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)* 67(2): 301-320.

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
Agriculture 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

Table 2: Performance Metrics by Method

	Accuracy(%)	**Recall(%)**	**Precision (%)**	**ROC AUC**	**PR AUC**
Baseline Method					
Standard log. 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

*, **, *** Metric for method is statistically significantly better (higher) than the standard logistic regression at the α=0.10, α=0.05, and α=0.01 levels, respectively; the best (highest) value of each metric is in bold.

Table 3: Relative Costs of False Negatives and False Positives

	λ value						
	1	**5**	**10**	**50**	**100**	**500**	**1,000**
Baseline Method							
Standard log. 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

*, **, *** metric for method is statistically significantly better (lower) than the literature based logistic regression at the α=0.10, α =0.05, and α =0.01 levels, respectively; the best (smallest) value of each metric is in bold.

Table 4: Top 10 Feature Importance, random forest

Dimension	**Variables**	**Mean Decrease Gini**
Asset Quality	Agricultural Loans/Total Loans	11914.813
Management	Net Income/Net Operating Income	6147.437
Macro Indicators	Interest Rate	4797.108
Earning Ability	Total Interest Income/Average Earnings Assets	4780.094
Other Financial	Market share	4516.130
Earning Ability	Total Interest Expenses/Average Earning Assets	4385.970
Earning Ability	Interest Income/Gross Loans	3800.585
Asset Quality	Loan_Unearned Income/Total Equity	3673.076
Liquidity Position	(Cash + FedFunds)/Total Assets	3472.261
Earning Ability	Efficiency Ratio	3274.823