WP 2017-15
October 2017

Working Paper

Charles H. Dyson School of Applied Economics and Management
Cornell University, Ithaca, New York 14853-7801 USA

Can Machine Learning Improve Prediction? An Application with Farm Survey Data

Jennifer Ifft

Coauthors are Ryan Kuhns (Farmer Mac) and Kevin Patrick (USDA ERS)

The views expressed herein are those of the authors and do not necessarily reflect the views of Farmer Mac or the US Department of Agriculture. Senior authorship is shared.

Can Machine Learning Improve Prediction? An Application with Farm Survey Data

October 6, 2017

Abstract

Businesses, researchers and policymakers in the agricultural and food sector regularly make use of large public, private and administrative datasets. These datasets are often used for prediction, including forecasting, public policy targeting, as well as management and financial research. Machine learning has the potential to substantially improve prediction with these datasets, but has thus far been underutilized by agricultural economists. In this study we demonstrate and evaluate several machine learning models for predicting demand for new credit with the 2014 Agricultural Resource Management Survey (ARMS). Many of the machine learning models used are shown to have much stronger predictive power than standard econometric approaches. We also use feature selection to show which variables are most useful for predicting demand for new credit. If correctly applied and interpreted, machine learning approaches can improve prediction with large datasets, with substantial benefits for research as well as public and private savings. However, careful implementation and evaluation is essential in realizing these benefits, as is the role of management researchers in interpreting results.

Keywords: Machine learning, prediction, Agricultural Resource Management Survey, farm debt, credit demand

1 Introduction

There is an ever-increasing number of large datasets to that can be used by businesses, government, and academic researchers to solve challenges facing agriculture, food and the environment. Sonka et al. (2014) emphasizes that both firms and and government will have to make fundamental management and organizational changes for the benefits of ‘Big Data’ to be fully realized for the ag sector. Researchers too will have to adapt to this new environment to take advantage of ongoing advances in data science and analytics. Management and warehousing of this data is an ongoing challenge (Woodard, 2016), and there serious issues with maintaining the the privacy and security of farm data (Sykuta et al., 2016). Along with these challenges, researchers must also adapt to methodological advances that have the potential to improve economics research. Many standard econometric models are not designed to take advantage of large datasets with detailed information for each observation, i.e. each farm, customer, plot of land.

Improved prediction has many practical business and policy uses as well as research applications. Businesses spend substantial resources predicting demand and targeting potential customers, and machine learning is widely used in private industry for prediction (Einav and Levin, 2014). ‘Big data’ firms are currently competing to solve agriculture and food sector problems through data analytics and new technology (Sparapani, 2017). There are also public uses for machine learning and prediction. For example, statistical agencies must collect data from farms, which have a wide range of response rates (Weber and Clay, 2013). Better targeting could save public resources while improving official statistics. Some research has also suggested that ‘big data’ can improve USDA forecasts (Tack et al., 2017).

In this study we consider the potential of machine learning to improve prediction of demand for new credit, using a dataset that is well-known to agricultural economists: the Agricultural Resource Management Survey (ARMS). ARMS data is used for a variety of official statistics, forecasting, and economic research, all of which could benefit from advances in machine learning. While there are limitations to use of machine learning for inference (Bellemare, 2013), there are many uses, including analysis of heterogeneous treatment effects.

While the focus of our analysis is demonstrating how machine learning can be used with a food and agriculture dataset to predict demand for new credit, future extensions could also take the machine learning methods demonstrated here for statistical inference applications, i.e. (Wager and Athey, 2017). Machine learning can be combined with ‘big data’ to improve agricultural forecasts or the targeting of government programs.

We provide a step-by-step guide on how setup a machine learning problem and how these methods can be evaluated and compared to standard econometric models. Throughout we address commonly-held concerns regarding machine learning, with a focus on overfitting of data, and explain how well-established methods can be used to mitigate these concerns. After comparing the benefits and drawbacks of machine learning methods relative to standard methods for our research question, we provide examples of using machine learning for prediction for private and public applications in food and agriculture.

1.1 Background

In this study we apply machine learning methods to ARMS in order to predict if a farm applied for new financing. The U.S. farm sector is entering its fourth year of declining income, and demand for credit faces upward pressure. As liquidity built up during high-income years is depleted, more farms may require additional lines of credit to cover operating expenses. Better predictions of farms desiring additional financing enables agricultural finance industry participants to better understand the characteristics of their potential customers and meet their needs. Additionally, the results inform the industry segments demanding greater financing and where credit constraints might occur.

There are some consistent descriptive findings in what U.S. farm and farm operator characteristics are related to debt use. Dairy and poultry operations have higher levels of credit use, while crop farms are less leveraged on average. Commercial farms and farms with younger primary operators also have higher levels of debt use (Ifft et al., 2014). Operator objectives may also drive demand for credit, for example operators may demand more credit because they want to increase the size of their operation or farm ‘full time’. For example, many U.S. dairy farms used credit to fund investments to expand capacity over recent decades (MacDonald et al., 2007). While it is well-established that farm financing requirements vary by production specialization as well as operator age, these general relationships may not be sufficient to accurately predict new credit demand.

Studies that use farm-level data to model credit demand are rare, with Katchova (2005) being the only published paper (to the best of our knowledge) to use U.S. farm-level data to model characteristics of farms that use credit. Katchova (2005) used 2001 ARMS data to explore determinants of various credit decisions, including use of credit and level of credit. This paper illustrates one approach to addressing truncation in modeling credit demand, by separately estimating the discrete decision to use any credit and the decision on amount of credit to use. Key factors found to influence the decision to use credit across farm types are gross farm income, risk management strategies, operator age, and risk aversion.

Prior to Katchova (2005), studies relied on bank data or farm data from outside the U.S. to estimate credit demand. More recently, Fecke et al. (2016) modeled individual loan amounts using data from a German bank and identified many factors that influence loan amount, including loan terms, value of farm production, and business expectations. They also note that sample selection bias is a common issue in the consumer credit choice literature as well as their study. Future research on the decision to apply for a loan is recommended. Using farm survey data from Ireland, Howley and Dillon (2012) found that in addition to the standard relationships between as farm size and operator age with debt levels, motivations, such as business or lifestyle-orientation for farming, also drive debt use.

The 2014 ARMS included research questions that asked respondents to indicate whether a respondent applied for new financing. The newly available data allows us to categorize whether farm operations applied for new financing, and determine if the demand for new credit can be predicted given other observable data about the operation. As a starting point we use a typical model[1] commonly used in econometric studies, logistic regression, to predict if each operation applied for financing. In order to demonstrate the potential benefits of machine learning methods for applied economics and management researchers, we explain the typical machine learning project process and terminology. We then employ an additional 10 machine learning algorithms to classify whether a farm operation responded to the 2014 ARMS survey indicating that they had applied for new financing. These are then compared to the 'literature driven logistic regression' and a dummy model that always predicts the outcome of the most prevalent class. An analysis of each of the ten classification methods used suggests machine learning can often lead to more accurate predictions and is a useful tool for econometric research in food and agriculture.

2 Data

The data used in this study comes from the 2014 Agricultural Resource Management Survey (ARMS). ARMS is an annual survey that is the USDA's primary source of information on U.S. farm businesses' financial performance and position, production practices, and resource use. The survey enables a broader understanding of the U.S. farm sector by including questions about the farm business along with questions on the demographics and economic well-being of the primary farm operator's household. The survey is constructed to be representative for the continental United States and to enable estimates at the state-level for the top agricultural States – typically the 15 states with highest levels of agricultural production. For 2014 the sample size was increased to allow state-level estimates for the top 25 states.

Beyond the typical questions asked in the ARMS survey, the USDA asks additional research questions that are included for just one year or are repeated sporadically. In 2014 the additional research questions focused on the debt portion of the farm's balance sheet, specifically in regards to applying for new loans or lines of credit. Section K of the 2014 ARMS survey included the following questions:

Question 7: Did you apply for any new loans or line of credit for agricultural purposes in 2014? (Yes/No)

Question 7a: Was a request for credit or loan application for agricultural purposes either turned down or were you not given as much credit as you applied for in 2014? (Yes/No)

Question 8: What was the MAIN reason you did not apply for any new loans or line of credit for agricultural purposes in 2014?

We focus our research on question 7 regarding whether the farm operator applied for any new loans or lines of credit for agricultural purposes in 2014. Of the 29,733 usable responses in the 2014 ARMS sample, all but 1,132 (3.8 percent) answered this question [2]. 32 percent (9,226 farm operators) answered affirmatively that they did apply for a new loan in 2014. Our variable excludes existing real estate and machinery (non-real estate) loans, as well as existing lines of operating credit that do not require reapplication. Banks may provide a line of operating credit that covers several years, typically secured by farm real estate. However generally operating loans are provided on a one-year basis and require annual reapplication.

There are differences in the characteristics of the farms and farm operators that applied for a new loan (which we will refer as credit applicants) and farm operators that did not apply for a new loan (which we will refer as non-applicants) in 2014. Similar to other research, these groups vary by demographic characteristics including age and sex, but have similar educational attainment. Farm characteristics including the commodity specialization, acres operated, and the farm's geographic location are also related to demand for credit, as well as financial characteristics of the farm business and the farm household. The number of surveyed farms in each category and the respective share of credit applicants are reported in table 1.

Perhaps the starkest contrast between credit applicants and non-applicants is by farm size, as defined by gross cash farm income (farm sales). More than half of credit applicants had sales greater than $350,000. Less than 20 percent of non-applicants reached that sales level. The difference between the two groups increases as the sales benchmark increases. 25 percent of credit applicants had more than $1,000,000 in sales compared to less than 8 percent for non-applicants.

3 The Prediction Pipeline

To illustrate how machine learning can be applied to common agricultural datasets and prediction problems, we follow a 'prediction pipeline' frequently used in the machine learning literature (Foster et al., 2016). We first define the question as a machine learning problem. Then, we explore and prepare the data for modeling. The next step is method selection, where models useful for answering the machine learning question are chosen. The final step is evaluating each model's performance. For each step in this prediction pipeline, we provide necessary context and language to compare to standard development of an econometric model.

There are many statistical software packages available to estimate machine learning models. We use the sci-kit learn Python package, which has functionality to carry-out a machine learning pipeline and includes many ML routines. R is another typical software used for econometric studies and it has several machine learning packages are available, including the Caret package. A variety of other statistical or programming software also offer the ability to implement multiple machine learning models.

In addition to choosing the models to use and evaluating their performance, many models require a user to choose several hyperparameters which govern how the model will fit the data. Throughout the model selection section we have highlighted important hyperparameters. Often referred to as tuning parameters, the number and purpose of these parameters varies by machine learning algorithm, but generally affects the degree to which a model under- or over- fits the data. For example, in Lasso or Ridge regression the strength of the regularization penalty must be determined. A weak regularization penalty results in a less sparse model, potentially resulting in overfitting. On the other hand, a penalty that is too strong could result in underfitting. Tree-based models including random forest, require the number of variables randomly selected to build each tree, the number of trees fit, and the maximum depth of each tree are commonly tuned. Allowing deeper trees that split on more variables could overfit the data, but splitting on too few variables could also underfit it. Because models that are under- or over- fit are unlikely to generalize to new data well, model hyperparameters can typically be tuned empirically by gauging the impact on out-of-sample predictive performance. In practice this can involve a grid-search over the relevant parameter space.

Many good resources, including the texts cited in this article, provide guidance on the exact tuning parameter choices required for different machine learning approaches. Ultimately, the choice of tuning parameter value can have a substantial impact on each model's performance and should be determined empirically from the data using cross-validation or another method for gauging the impact of tuning parameter choice on out-of-sample model performance. After determining the appropriate tuning parameters for each machine learning algorithm, each of the tuned models can then be compared via cross-validation to determine the model with the best predictive performance for the research question at hand.

3.1 Defining the Machine Learning Problem

When attempting to solve a prediction problem using machine learning techniques, it is essential to explicitly pose the question of interest as a machine learning problem. This is analogous to moving from a research question to an empirical strategy in standard econometric analysis. The type of problem will dictate the needed data and guide the model selection. Prediction of continuous variables like the amount of credit demanded use one set of machine learning tools, while categorical variables like whether or not a farmer applied for a new loan use another. [3]

Because our goal is to predict the farms that will apply for a new loan, we are interested in separating our data observations into one of two groups. In machine learning terminology, this is a binary classification problem. We are trying to classify farmers into one of two groups: new credit applicants or non-applicants. There are numerous machine learning methods that are well suited to tackle this problem and we outline several in the model selection section below. Defining the problem also enables use to gather and prepare the data needed as inputs for the machine learning methods.

3.2 Data Preparation

Having defined the question as a binary classification prediction problem with the goal of classifying farmers as either new credit applicants or non-applicants, we can create what is known as the label and features in machine learning literature. The 'label', or y-variable, is a binary variable that takes a value of 1 (or true) if the farm operation applied for a new loan in 2014 and (0 or false) otherwise. The explanatory x-variables are the 'features' that may help to predict the label. Unlike statistical inference where the estimated coefficients are important, accurate prediction is the goal. Hence many of the issues associated with explanatory variable selection for inference do not apply. Instead it is often preferable to include many more features and transformations of features including creating interactions or aggregations of variables. Having quality data is essential to the success of the machine learning models. We choose features that describe the primary farm operator, the farm, and the farm household. This includes variables related the age of the operator, the size and type of the farm, as well as how reliant they are on income from farming. We also exploit a question in the 2014 ARMS questionnaire on the primary farm operator's risk preferences. The full set of features (or variables) used is reported in table A1.

3.3 Model Selection

Having defined the research question as a machine learning problem and selected and transformed the data, we need to choose the machine learning methods that will perform the prediction. The type of problem dictates the appropriate methods for empirical testing. Our problem is a 'classification prediction problem', i.e., we want to classify a farm operator as either applied for a new or did not apply for a new loan. Therefore, we need to select machine learning methods that are capable of solving a classification problem. While there are many applicable supervised machine learning methods, we choose 10 common approaches not typically used in the standard econometric literature allowing the relative benefits of using machine learning methods for agricultural prediction problems to be illustrated. The machine learning models fall into six broad categories: generalized linear models, nearest neighbors, support vector machines, naive Bayes models, ensemble models, and a dummy classifier model. Each model is described below along with their potential strengths and weaknesses in solving the problem.

The most basic family of models is known as generalized linear models. These models will perform well if the target variable or ‘label’ can be approximated by a linear combination of the feature variables. Many commonly used models in the agricultural finance literature including ordinary least squares and logistic regression models fall into this category of models. From the large pool of generalized linear models, we choose four models. The first is a simple logistic regression (literature based logistic regression) without any added penalization for dimension reduction. Because we are using a relatively large feature set and interested in the performance of standard econometric techniques relative to machine learning approaches, we use the prevailing literature to guide variable selection for this model[4].

We then build on the standard logit model by encouraging sparsity via several regularization techniques. The goal when introducing sparsity is to determine if there any features that can be removed from the model, while still producing ’good’ forecasts. The use of regularization or penalization for model complexity has a long history and we apply two of the classic examples of regularization, Ridge and Lasso, to the logit model. Ridge penalizes model complexity in the form of a penalty function of the sum of squared betas. In practice, this means that a small move in the value of beta away from zero is not very ’costly’, but as beta increases the penalty becomes more binding. Therefore, Ridge tends to encourage many features with small betas rather than completely zeroing out a feature (Hastie et al., 2009). On the other hand, Lasso penalizes model complexity in the form of a penalty function of the sum of absolute values of betas. In practice it is more ’expensive’ to move the value of a beta away from zero, but once that move away from zero occurs, the additional ’cost’ of increasing beta is linear. Lasso is therefore more likely to zero out betas, effectively removing those features from the model (Hastie et al., 2009). Because both of these regularization models have penalties that involve sums of estimated coefficients, the coefficients scale and therefore the scale of the underlying feature need to be similar. A common approach for solving this issue, which we use, is to normalize all features to mean of zero and variance of 1.

Like many machine learning algorithms, the performance of Ridge and Lasso logistic regression is dependent on the chosen value for the each model's hyperparameter(s). For the aformentioned regularization approaches, the relative strength of either the Ridge or Lasso regularization penalty must be determined. A stronger penalty encourages greater sparsity, while a weaker penalty results in less regularization. In some cases a value of a model's hyperparameter(s) is set at a chosen level a priori. However, typically the value(s) are chosen empirically from the data in order to promote model performance.

The literature-based logistic regression uses a parsimonious set of variables based on standard econometric practice and the Ridge and Lasso models use a regularized subset of variables from the full set of features in A1. The remaining models use all features, as these machine learning models are designed to be able to take advantage of a larger set of information–features or variables–than in normal statistical models. This approach uses our dataset with each model using the standard methods or approach that each model was developed for. To enhance our comparison, we will also run the following models with the limited set of variables selected for the literature-based logistic regression.

The final linear model algorithm chosen is stochastic gradient descent (SGD). SGD is a relatively simple algorithm for optimizing generalized linear models, which tends to be computationally efficient(Bottou, 2010). We implement a logistic regression based version of the SGD algorithm. The SGD model uses an iterative optimization algorithm where the model starts with an initial set of parameters and iterative changes are made until the objective function is minimized. This process is repeated with the initial starting parameters shuffled. This model has been shown to perform well when given large data sets as it can be easily processed using parallel computation.

The k-nearest neighbor method (KNN) is another simple classification technique, which seeks to make predictions using information from other data points local to the observation being predicted (Hastie et al., 2009). Rather than fitting a model, this method requires two things: the number of nearest neighbors k to use and a distance function. The distance function is used to determine how similar/dissimilar the features of the data point to be predicted are relative to those of the other known observations. The k closest points are then used form a prediction. For classification problems, the prediction typically takes the form of the most commonly observed class outcome out of the k closest observations. The number of nearest neighbors, k, used is typically chosen to minimize prediction error. Because neighbors are judged on the basis of feature distance, the K-nearest neighbors approach is particularly sensitive to the scaling of features. If variables aren't moved to a common scale, differences between observations for features with broader scales will tend to dominate the distance calculations. As with the regularized logit models, we solve this scale issue by normalizing all features.

Support vector machines (SVM) represent a class of often used machine learning models, which can be linear or nonlinear. The SVM algorithm attempts to split the data into two classes by fitting an optimal hyperplane out of the possible lines (support vectors) that could be used to split the data. In cases where the classes are completely separable, the algorithm's goal is to maximize the margin between the decision boundary and support vectors. A larger margin means slight changes in the data are unlikely to result in incorrect classification. Therefore, maximizing the margin seeks to minimize potential prediction errors. In practice, the outcome classes are often not completely separable and the SVM algorithm must trade off a larger margin with the cost of misclassifying some existing observations as the margin increases. There are no set tolerances for erroneously classifying existing observations; rather the misclassification error tolerance is treated as a hyperparameter. Multiple strengths of the penalty associated with misclassifications can be tested in order to determine the penalty level that maximizes the model's accuracy (James et al., 2013). Kernel techniques allow the SVM algorithm to use a variety of feature basis expansions in order to incorporate nonlinearities into the decision boundary (Hastie et al., 2009). Given the algorithm's flexibility it is commonly used. However, a disadvantage of using support vector machines is that the scores are based on the margin or distance from the optimal hyperplane and are not probability estimates in the event predicted class probabilities are required. In addition to the misclassification tolerance, the type of kernel transformation used must also be determined. Although kernel techniques allow a variety of nonlinear SVM classifiers to be considered, we illustrate the modeling approach by requiring the fit hyperplane to be linear.

Slightly more complex than generalized linear models, k-nearest neighbors, or support vector machines, though still fairly simple, naive Bayes models are a family of supervised machine learning models that employ Bayes theorem of conditional updating to make predictions. In addition to using Bayesian updating, Naive Bayes models also make the the added 'naive' assumption that the features are independent (Kuhn and Johnson, 2013). From this family of models, we choose the Gaussian naive Bayes (GNB), which further assumes that the likelihood of the features follows the normal distribution (Kuhn and Johnson, 2013). For obvious reasons, this model will not perform well if the feature variables are not independent or if the likelihood function of the features is not normally distributed. One of the benefits of GNB models is that they tend to perform well even with relatively small amounts of training data. It is also computationally efficient which means it can be implemented quickly compared to more complex models.

Thus far we have described individual machine learning approaches. However, dating back to the seminal work of Bates and Granger (1969) a large body of research on forecast accuracy has found that combining forecasts from multiple models can result in more accurate forecasts. In this spirit, ensemble or weighted machine learning models combine forecasts from numerous base models. The goal of these models is to take advantage of the benefits of each base model while reducing the drawbacks from any single model. To illustrate common ensemble approaches, we compare four ensemble models that take different approaches to combining base models.

The first is bootstrap aggregation (bagging), which creates several data sets from the training data via resampling and fits the learning model to each one (Hastie et al., 2009). The model's prediction reflects either the mean or mode from the predictions made on each individual bootstrapped samples. By combining the information across the bootstrapped samples, the variance portion of prediction error can be reduced (Hastie et al., 2009). Since bagging is designed to improve model performance by reducing the variance component of prediction error it tends to result in greater performance improvements from models that have inherently greater variability, such as classification trees (Hastie et al., 2009). However, we use two slightly different approaches to combining tree-based models. To demonstrate the potential benefits of bagging, we use a bagged-KNN model. Along with the choice of k required by the underlying KNN model, bagged models also require the number of bagged datasets to be chosen.

Forests of randomized trees, often referred to as a random forest model (RF) take the concept of bagging a step further by randomizing the subset of features available to build each tree. Randomizing the features used in each tree results in less correlation between models on each resampled dataset, which should allow for further reduction in the variance component of prediction error relative to simple bagging (Hastie et al., 2009). For each tree, the data is split into two groups based on a particular feature that best splits the data between positive and negative cases. Each new subset of data is split again based on another feature that best splits the data. This is performed for each tree until additional splits are not found to improve the individual tree. The results are then averaged across all the trees. The ensemble model's final performance can be affected by the number of bagged datasets used. However, it tends to be most sensitive to the choice of the number of features randomly selected to use in the creation of each tree and this hyperparemeter is commonly tuned to improve predictive performance. By only considering a random subset of featuers, RF models have the potential to introduce bias, but they may result in a preferable model due to the reduced variation from averaging a diverse set of decision trees. Since each tree can be fit independently, this model is also easy to perform using parallel processing making it useful for extremely large data sets.

We also test a variant on RF models, known as extremely randomized trees (ERT). ERT models go a step further, randomizing both the subset of features and splitting thresholds

(Hastie et al., 2009). This method can increase the overall bias of the results, but tends to reduce the variance over the standard RF approach. Depending on exact the tradeoff between increased bias and error reduction, this can reduce prediction error. Again, the ability to fit each of the trees independently allows ERT models to take advantage of parallel processing.

The final ensemble model we consider is called boosting. Unlike bagged-KNN, RF and ERT type models where the base models are learned independently, boosting methods perform the base methods sequentially while seeking to minimize added bias at each step. This ensemble technique typically seeks to to combine numerous relatively weak prediction models to produce a model that has more predictive power than any of the individual models. We specifically choose gradient tree boosting (GTB), which iteratively adds decision trees in stages. After each stage, the observations predicted incorrectly are given greater weight and a new tree is fit; the weighted combination of the predictions is then combined to result in the final ensemble prediction(Hastie et al., 2009). In this way, the model seeks to gradually learn from the data to improve predictive accuracy. The learning rate, one of the model's primary hyperparameters, controls the weighting used to combine each stages predictions, while the number of boosting steps controls the number of iterative steps undertaken. Ultimately there is a tradeoff between the two parameters, smaller learning rate values result typically require greater number of boosting iterations (Hastie et al., 2009). GTB tends to perform well when the features are heterogeneous i.e., binary, categorical, and continuous feature variables. Although not a problem with our data, the iterative nature of this method means that parallel processing is difficult which means scaling this model up to accommodate large data sets is problematic.

The final method we report results for is a dummy classifier used as a baseline for all of the other models. For this method the prediction is the most frequent outcome from the training set. In our data, the majority of respondents did not apply for a loan. Therefore, this model would always a predict the farm operator did not apply for a new loan. This type of model is useful as a baseline line for which all other models are compared.

There are numerous other models that could have feasibly been considered, but the included models were chosen to demonstrate the breadth of machine learning techniques, differences in complexity and potential benefits and costs to using different approaches. After we've defined the problem, prepared the data, and selected competing models, we now can evaluate the performance of each model. The next section covers the metrics and procedure we follow for evaluation.

3.4 Evaluation

In most analysis using ARMS data the focus is on inference rather than prediction. Accordingly, the focus is on the economic interpretation and statistical significance of estimated regression coefficients. However, our emphasis is demonstrating the benefit of machine learning methods to successfully predict the farms that indicated they applied for new credit in the 2014 ARMS data. Therefore, we analyze the predictive accuracy of each method model considered.

Given that most econometric and machine learning methods minimize some measure of inaccuracy, evaluating predictive accuracy on the same data used to fit the model, called in-sample prediction, results in overly optimistic accuracy estimates. This is often referred to as over-fitting the model. Over-fit models tend to generalize poorly, resulting in poor predictive performance when applied to other data. Therefore, we follow standard practice in the forecasting and machine learning literature and base our analysis on out-of-sample rather than in-sample predictions. To accomplish this we split our original data into a 'training data' set used to fit the model and then apply the trained model to the 'test data' in order to evaluate its accuracy. While there are many methods of assigning observation to the test data, we elect to use repeated 'stratified k-fold cross-validation' because it allows us to evaluate the model using all observations. Repeating stratified k-fold cross-validation, allows us to better understand distribution of the resulting accuracy metrics.

In k-fold cross-validation, the data set is randomly assigned to k equally sized subsets called folds. Stratified k-fold cross-validation adds a constraint to the random fold assignment, requiring that the subsets preserve the proportion of observations observed in each class of the response variable in the full data set. The model is then fit *k* times. Each time *k*-1 of the folds are used to fit the model and the left out fold is used to evaluate model accuracy. K-fold cross-validation is typically preferred to simple out-of-sample testing where some percentage of the data is held out for predictive accuracy testing because with k-fold cross-validation each observation is used to evaluate the model's accuracy. Repeated stratified k-fold cross-validation repeats this process *r* times with the data randomly assigned to new k-fold subsets each time. This has the benefit of reducing the impact of the random fold assignments on the model's accuracy. Since repeated stratified k-fold cross-validation results in $r*k$ measures of model accuracy, it also allows us to gauge whether differences in model accuracy metrics are statistically significant.

Ultimately, the decision on the number of folds used to split the data and requires consideration between computational resources, as well as the bias/variance trade-off associated with estimated accuracy statistics. As *k* increases, the proportion of data used to fit each model increases, resulting in lower potential bias in estimated accuracy measures (James et al., 2013; Kuhn and Johnson, 2013). In the special case where *k* equals the number of observations, known as leave one out cross-validation (LOOCV), the difference between the size of the training data and original sample is small, resulting in little bias. However, the approach is very computationally intensive, particularly in larger data sets[5]. Additionally, the use of LOOCV tends to result in higher variance (James et al., 2013).

The error inherent in our estimates of a model's accuracy reflect both bias and variance. Therefore the choice of validation strategy should reflect the trade-off between bias and variance. Using a smaller value of *k* will result in additional bias but less variance. Depending on the exact trade-off this can potentially improve our estimates of a model's predictive performance. Although there are no set rules, 5- or 10- fold cross-validation has been shown to result in accuracy estimates with bias and variance that are not too high and are therefore commonly used (James et al., 2013; Kuhn and Johnson, 2013). We choose to use 10-fold cross-validation to evaluate each model in our analysis. There are also no set rules for choosing the number of repeats to perform. Increasing r, the number of times cross-validation is repeated, reduces the variance resulting from the random assignment to each of the k folds. However, this again increases the computational resources need to estimate the models. We choose to repeat the stratified 10-fold cross validation 10 times.

Although the terminology used can sometimes differ, the methods used to evaluate the predictive accuracy of machine learning models aligns with that used in the forecast evaluation literature. True positive (TP) cases are where the model correctly predicted a farm applied for an application and true negatives (TN) are where the model was able to correctly discern the farm operation did not apply for new credit. A model's overall accuracy, or percentage of correctly predicted outcomes, is calculated as the sum of these correctly predicted cases to the total number of observations.

$$Accuracy = \frac{True\ positive + True\ negative}{All\ predictions} \tag{1}$$

.

While accuracy is often used to provide a high-level overview of a model's predictive ability, it does not account for the relative frequency of the categorical outcomes or the ability to make correct predictions by chance. In the 2014 ARMS data, 32.3 percent of respondents indicated they had applied for a new loan or line of credit. As a result, a simple model assuming no farms applied for a new loan, would have an accuracy rate of 67.3 percent. It is clear that each model's accuracy needs to be viewed in comparison to some baseline. With this in mind, it can therefore be more informative to use a measure of accuracy that takes into account the expected accuracy given the prevalence of the event of interest in the confusion matrix (Kuhn and Johnson, 2013).

We use the 'kappa statistic', which takes into account the possibility of predicting the correct outcome by chance, as an alternative measure of overall accuracy (Cohen, 1960). To calculate kappa each model's observed accuracy (O) is scaled by the probability of predicting the outcome correctly by chance (E)[6]:

$$kappa = (\frac{O - E}{1 - E}). \tag{2}$$

The kappa statistic measures the agreement between the predicted and observed outcomes on a scale between -100 and 100 [7] [8]; however, in practice values range between 0, which signifies no predictive ability and 100, which indicates perfect agreement between the predictions and outcomes.

Because the goal is to predict farm operations that applied for credit, we also consider each model's ability to discern between applicants and non-applicants. A model's recall, also commonly referred to as sensitivity, is a measure of its ability to correctly predict the event of interest having occurred in the sample of observations where the event actually occurred. In the context of predicting credit applications, recall measures each model's ability to predict that a farm applied for a new loan among the 9,226 operations that were actually observed as having applied. It can be interpreted as the percent of farms that were correctly predicted would apply for a loan out of the total that actually applied. A recall value of 80 percent means the model was able to select 80 percent of the people that actually applied for a loan. Specificity is a related accuracy metric, which measures the ability to detect non-events in the observations that did not have the event of interest occur. Therefore, we use specificity as a gauge of the ability to correctly classify non-applicants as having not applied for new financing. As shown in equations 3 and 4, recall and specificity can be calculated from the confusion matrix as the number of true positives (negatives) relative to observed positives (negatives).

$$Recall(sensitivity) = \frac{True\ positives}{True\ positives + False\ negatives} \tag{3}$$

$$Specificity = \frac{True\ negatives}{True\ negatives + False\ positives} \tag{4}$$

While sensitivity and specificity are useful in assessing model accuracy, they are conditioned on the event of interest, in our case having applied for credit, having occurred or not occurred (Kuhn and Johnson, 2013). However, most often models are used to predict an event outcome without having prior knowledge of the event class the observation will actually end up in. Positive predictive value (PPV), also called precision, is a measure of the unconditional probability of the event occurring, while negative predictive value (NPV) is the unconditional probability of the event of interest having not occurred. In our case, precision measures the accuracy of the model when it has predicted a farm applied for a new loan. A precision value of 80 percent means that out of the farms the model predicted applied for a loan, it was correct 80 percent of the time. PPV and NPV are easily calculated directly from the confusion matrix as shown in the equations below.

$$Positive\ predictive\ value(precision) = \frac{True\ positives}{True\ positives + False\ positives} \tag{5}$$

$$Negative\ predictive\ value = \frac{True\ negatives}{True\ negatives + False\ negatives} \tag{6}$$

It is common to use the previously discussed metrics to determine whether one modeling approach results in more accurate predictions than another. Statistical tests can be used to gauge whether these differences are statistically significant. However, care must be used in choosing the appropriate test. When comparing accuracy metrics estimated via cross-validation, the results from each fold are unlikely to be independent. Therefore, statistical tests for dependent or paired sample must be used. T-tests for paired samples present one option; however, they assume the variables being tested are normally distributed. By definition, most of the metrics used to compare binary classification models are bounded between zero and one. Rather than make an assumption on their distribution, we choose to the Wilcoxon Signed Rank test, a nonparametric test for matched samples, to compare statistical differences in model accuracy.

After decision tree based models including random forest and gradient tree boosting are created, it is possible to construct the importance of each feature. Features importance is a relative score determined by the amount that splitting at that feature node in the decision tree improves some predictive metric. The metric we use is the average Gini impurity proposed in James et al. (2013). Since feature importance is relative, the values can be ranked and compared to determine the most important features. While feature importance rankings can provide information on the relative importance of each variable in making predictions, feature importance cannot be interpreted the same way as statistical significance in a typical regression model. In fact, a feature can have a low feature importance score not because it is a bad predictor, but because it is highly correlated to another feature and therefore does not add much to the prediction. Instead, the ranking of each feature indicates its relative importance for prediction for the particular model used. Hence highly ranked-features are important indicators of how each feature (independent or x-variable is the analog for standard econometric analysis) contributes to prediction, but a low rank doesn't necessarily mean that a particular variable is not a determinant of the label (dependent variable or y-variable in standard econometric analysis).

4 Results

The metrics described in the evaluation section are reported in tables 2 and 3. We use these metrics to evaluate the success of each model in predicting whether a farm operation applied for credit in 2014. Table 2 summarizes accuracy metrics results, by reporting the average metric by model for each of the metrics outlined in the evaluation section. Focusing first on each model's overall predictive ability, the accuracy statistics suggest that applying the more complex ensemble machine learning methods to ARMS data can improve the ability to predict credit demand compared to the baseline literature driven logistic regression, but not necessarily so. The results reported in table 3 are estimated using only the limited literature-based set of features for all models.

The baseline logistic regression model was able to correctly predict whether or not a farm operation applied for new credit 70 percent of the time. This outperformed the dummy model, which always predicted a farm did not apply for new credit, but the difference was not statistically significantly at any standard test level. Additionally, both forms of regularized logistic regression, LASSO and Ridge, had higher accuracy than the dummy classifier or literature-based model. Some but not all of the other machine learning models considered resulted in statistically significant improvements in prediction accuracy, which were often economically meaningful. In particular, the stochastic gradient descent (SDG), gradient tree boosting (GTB), bagged k-nearest neighbors (bagged-KNN), Linear Support Vector Machine, and K-nearest neighbor models were each able to correctly predict at least 73 percent of outcomes, representing a 3 to 5 percent improvement in predictive accuracy relative to the literature based logistic regression.

Comparing the model using the kappa statistic, which corrects for the likelihood of correct predictions due to random chance, results in nearly the same rank order of model accuracy. Interestingly, the gap in relative accuracy between logistic regression and the more accurate machine learning methods widens once the role of chance is taken into account. Because more than two-thirds of the ARMS sample reported not having applied for a new loan, there is intuitively a greater likelihood of having been correct by chance when predicting an observation was a non-applicant. Therefore, models predicting a greater number of non-applicants could appear more accurate. By analyzing the reported specificity, sensitivity and precision statistics, we can get a better sense of each model's ability to discern between applicants and non-applicants in more detail.

The specificity results suggest all models except the Gaussian Naive Bayes (GNB) were able to correctly predict a greater proportion of the actual non-applicants as compared to the literature driven logistic regression. The random forest (RF) and extremely randomized tree (EFT) models were particularly adept at correctly assigning actual non-applicants to the non-applicant group, correctly identifying more than 90 percent of all non-applicants. Comparing the models negative predictive value provides further insight into how often each model's prediction of a farm operation being a non-applicant was true.

Each of the linear regression models (LASSO, Ridge, SGD) had a negative predictive value of about 85 percent, suggesting about 15 percent of time an farm was predicted to be a non-applicant, the operation actually had applied for a loan. By comparing both specificity and negative predictive value, it is clear that most of the models able to hone in on non-applicants better than the literature-based logit. In contrast, the literature based logistic regression model had a somewhat larger gap between specificity and negative predictive value. This suggests the literature based approach was able to correctly identify non-applicants at the cost of being more likely to incorrectly predict that actual applicants did not apply for a loan.

While there are more observations in the underlying data where the respondents did not apply for a loan, being able to correctly identify applicants is likely of greater interest to industry participants and policy makers. For example, a government agency looking to target a loan program, policymaker looking to understand operations who could be credit constrained, or financial institution looking to lend to farmers would all be interested in correctly identifying those operations that applied for loans. Analyzing recall (sensitivity) and precision (positive predictive value) allow us to determine which models are best at classifying the outcome class of interest. In comparison to most of the models' relatively high specificity, the recall metric has lower values and a greater spread, ranging between 0.51 to 0.69 for all but the dummy model. This suggests the models had a more difficult time classifying the less common applicant observations. Again the literature based logistic regression model performance falls toward the bottom of the model pack, identifying just under 58 percent of applicants who applied for credit. Not only were many of the machine learning models statistically significantly more likely to identify actual applicants, but they were typically able to do so with statistically significantly greater precision.

The choice of model ultimately depends on the research or business/policy objective. All but the RF, ERT, and dummy models were able to identify applicants with higher recall and precision. In particular, the linear machine learning models performed relatively well in terms of both of these metrics. For the aforementioned government agency, policymaker or financial institution looking to understand applicants, these machine learning models could be used to gain economically meaningful improvements in targeting farms likely to apply for loans. In contrast, if the goal was to identify non-applicants to survey about the reasons they did not use debt, the RF, ERT models models may be preferable. However, they identifying actual non-applicants at a high rate, partially through an increased number of false negatives. If the user wanted to avoid contacting the additional farms falsely predicted to be a non-applicant potentially due to budget constrains, the GTB, bagged-KNN, KNN or linear models represent alternatives that also performed better than the literature model in terms of specificity, while also having higher negative predictive value than the RF and ERT models.

In table 3 we report the results of estimating all of our models with the set of featured used in the literature based model only. While the other models are designed to handle a large number features, or high-dimensional data more generally, this comparison allows us to see if ML algorithms would help with prediction even with smaller datasets by learning more from the data. Overall our results are consistent - most of our ML models are better at prediction than the literature based logistic regression. The LASSO and Ridge regressions do not provide much of an advantage in this approach, which is unsurprising given that with the same (relatively small) set of variables, the respective sparsity constraints become less important and the models would be expected to be very similar to a logistic regression. In contrast, the RF model performs statistically significantly better than the literature-based logistic regression in this scenario, despite failing to do so for most metrics when using the entire feature set. The difference likely reflects the interplay the RF algorithm's use of a random subset of features to build each tree and the data being used. In cases where relatively few of the available features are important for model performance, it is possible that many of the trees are built using random subsets excluding these variables. Using a smaller set of variables predetermined using past economic literature or alternative feature selection methods, may result in improved performance as more trees are built with informative predictors.

Tree based ensemble models including random forest (RF) and gradient tree boosting (GTB) were among the most successful models at minimizing the number of false positives and false negatives (precision and specificity). These models have the additional benefit of outputting the importance of each feature to the model results. We report the RF and GTB feature rankings in appendix tables A2 and A3, respectively. The most important features in predicting if a farmer applied for a new loan in 2014 primarily related to the size of the farm both in terms of gross cash income and in acres owned. Gross cash income was ranked as the most important feature for the RF model and the second most important feature for the GTB model. Differentiating the size of farms even further by determining if a farm had sales above or below $150,000 ranked highly. The level of debt, specifically prior long-term nonreal estate debt was the most important feature in the GBT model and the fourth for RF indicating that prior use of nonreal estate debt is a potentially useful feature to determine future interest in debt use.

Perhaps one of the most interesting findings is a farm operators reliance on off-farm income interacted with the county unemployment rate in the prior year was ranked as one of the top ten most important features for both models. This could indicate that household financial stress could push farm operators to apply for new credit or reflect the broader relationship between the farm economy and the rural non-farm economy. As discussed in the implementation section, a feature that has a low feature importance score does not necessary mean that that feature is not related to applying for a new loan–that feature may just be highly correlated with another feature and therefore does not add much to predictive power of the model. Still, some of the features that are typically included in a literature based model of credit use including location of the farm, demographics of the primary farm operator, and even commodity specialization of the farm, were not ranked as highly important features.

We are able to show that machine learning provides a set of potentially useful prediction tools when applied to a standard farm survey dataset, although the prediction outcomes varied based on the method. Regardless of the accuracy metric considered, most of the machine learning approaches were found to perform statistically significantly better than the comparison literature driven logistic regression model. This remains true when restricting the machine learning models to the same set of variables used to build the literature driven logistic regression. At the same time, a machine learning algorithm will not necessarily result in improved predictive performance. The RF model failed to outperform the literature driven model when using the entire feature set, but outperformed it when restricted to features deemed important by past economic research.

By utilizing machine learning, firms, policymakers and researchers may be able to improve the accuracy of their predictions. However, the results also emphasize the importance of not blindly applying a machine learning model, but considering the appropriate data, testing different models, choosing appropriate tuning parameters and understanding the benefits and drawbacks to each model in the context of the research objective. In the end, the best model depends on the prediction outcome that is most important to person or organization that will use the model. Someone that is interested in accurately targeting users, say to mail a letter and not have much waste, may prefer a model with high precision like the RF model. However, if someone is looking to reach as many potential new credit applicants, they may choose a model like GNB that had better recall. Most statistical packages for machine learning allow for use of and comparison between multiple models.

5 Conclusion

In this study we demonstrate how machine learning methods can be used to improve prediction with a large farm survey dataset. Many of the more complex machine learning methods used performed substantially better than our standard econometric model (logit) at predicting credit demand. This is even true for some models when a limited set of features or variables are used, which doesn't take advantage the ability of machine learning models to use much more information than standard econometric models. However, it is important to note that if the more complex methods are not feasible due to computing power, data availability or other issues, than the best approach may be using standard econometric modeling. Researchers will need to weigh their research question or business/policy objectives with available computing resources and data to determine the preferred methodology.

Through detailing how machine learning methods are typically implemented, we illustrate how machine learning can be used transparently and avoid over-fitting. Our approach can be used for applying machine learning methods to ARMS data as well as policy, food and business datasets. Machine learning models are already widely used in industry and banking, for example by credit card companies to evaluate applicants. However, researchers in food and agricultural economics are just beginning to use the more advance machine learning methods that are currently widely available.

There many ways which improved prediction can have private and public benefits. Many federal agencies undertake official forecasts, which may be improved by machine learning methods. When targeting is necessary, say for policies aimed towards specific groups, such as beginning farmers, or for survey completion, machine learning methods could be used with administrative data to lower costs. Many firms have large private databases and sales data, in addition to access to public datasets. More accurate targeting of potential customers and or prediction of demand could improve profitability of agribusinesses. Farm lenders could use our approach to better predict which farms are more likely face challenges with loan repayment.

In addition to improving prediction, other methodological issues surrounding research using ARMS and other food and agricultural datasets may benefit from machine learning methods. Machine learning provides a data-driven method to improve variable selection for statistical models, which is increasingly challenging when using large datasets for research. Another issue is that while many variables commonly used in research have imputation for missing responses (i.e. Morehart et al. (2014)). Machine learning may improve imputation, in cases when imputation is desirable. Given that these models can handle a larger number of variables, machine learning methods could also accommodate a less restrictive approach– explicitly including raw survey responses and variables indicating missing observations in statistical models.

As indicated by our findings, machine learning does not always improve prediction, let alone the challenges related to statistical inference. This is an important topic for future research. Machine learning alone cannot help researchers claim causal identification. However, at a minimum, machine learning can improve inference through facilitating use of large datasets and increasing options for robustness testing. Machine learning may also be useful in estimation of heterogeneous treatment effects, or understanding how research findings apply to different groups. Machine learning methods have some limitations, as with all methodologies. One serious concern is p-hacking or use of 'data mining' in economic research. Transparent use of these methods and testing of different approaches, such as in this paper, can mitigate these concerns. Many machine learning statistical packages allow for different models to be implemented.

In this study, we demonstrate how methodological advances–machine learning models–can be applied to improve prediction. In addition to improving forecasting capabilities, machine learning can provide a variety of improvements to current methodologies being used in applied research in the farm and food sector. While machine learning is not a panacea to the methodological challenges of prediction and statistical inference, there many benefits to its application to research using ARMS and other food and farm data sets. To be able to fully take advantage of new methodologies, researchers, firms and policymakers will need to change the way that data is collected and managed. This and many other concerns raised by Sonka et al. (2014) and others will need to be addressed before the promises of 'big data' can be fully realized for food and agriculture. Future studies should explore additional uses for machine learning as well as the broader challenges in managing large datasets.

References

Bates, J. and Granger, C. (1969). The combination of forecasts. *Operations Research*, 20(4).

Bellemare, M. (2013). Big dumb data? http://marcfbellemare.com/wordpress/8859.

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT sysmposium*. International Conference on Computational Statistics.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, XX(1).

Einav, L. and Levin, J. (2014). The data revolution and economic analysis. *Innovation Policy and the Economy*, 14(1):1–24.

Fecke, W., Fecke, W., Feil, J.-H., Feil, J.-H., Musshoff, O., and Musshoff, O. (2016). Determinants of loan demand in agriculture: empirical evidence from germany. *Agricultural Finance Review*, 76(4):462–476.

Foster, I., Ghani, R., Jarmin, R., Kreuter, F., and Lane, J. (2016). *Big Data and Social Science, A Practical Guide to Methods and Tools*. CRC Press, Boca Raton, Florida.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning Data Mining, Inference and Prediction*. Springer, New York, New York.

Howley, P. and Dillon, E. (2012). Modelling the effect of farming attitudes on farm credit use: a case study from ireland. *Agricultural Finance Review*, 72(3):456–470.

Ifft, J., Patrick, K., and Novini, A. (2014). Debt use by us farm businesses, 1992-2011. Technical report, United States Department of Agriculture, Economic Research Service.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer, New York, New York.

Katchova, A. L. (2005). Factors affecting farm credit use. *Agricultural Finance Review*, 65(2):17–29.

Kuhn, M. and Johnson, K. (2013). *Applied Predictive Modeling*. Springer, New York, New York.

MacDonald, J. M., O'Donoghue, E., McBride, W., Nehring, R. F., Sandretto, C. L., and Mosheim, R. (2007). Profits, costs, and the changing structure of dairy farming. Technical report, United States Department of Agriculture, Economic Research Service.

Morehart, M., Milkove, D., Xu, Y., et al. (2014). Multivariate farm debt imputation in the agricultural resource management survey (arms). In *2014 Annual Meeting, July 27-29, 2014, Minneapolis, Minnesota*. Agricultural and Applied Economics Association.

Sonka, S. et al. (2014). Big data and the ag sector: More than lots of numbers. *International Food and Agribusiness Management Review*, 17(1):1–20.

Sparapani, T. (2017). How big data and tech will improve agriculture, from farm to table. https://www.forbes.com/sites/timsparapani/2017/03/23/how-big-data-and-tech-will-improve-agriculture-from-farm-to-table/.

Sykuta, M. E. et al. (2016). Big data in agriculture: Property rights, privacy and competition in ag data services. *International Food and Agribusiness Management Review*, 19(A).

Tack, J., Coble, K. H., Johansson, R., Harri, A., and Barnett, B. (2017). The potential implications of'big ag data'for usda forecasts.

Wager, S. and Athey, S. (2017). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, (just-accepted).

Weber, J. G. and Clay, D. M. (2013). Who does not respond to the agricultural resource management survey and does it matter? *American journal of agricultural economics*, 95(3):755–771.

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.

6 Tables

Table 1: Summary Statistics

	Number	Share credit applicants*
Commodity Specialization		
Corn	2,802	49%
Soybean	2,295	43%
Wheat	725	39%
Cotton	308	53%
Specialty Crop	2,963	25%
Other Crop	6,993	31%
Cattle & Calve	8,598	25%
Dairy	1,700	50%
Hog	385	45%
Poultry & Egg	1,526	31%
Other Livestock	1,438	16%
Age		
< = 34	1,136	52%
35-44	2,569	47%
45-54	5,646	39%
55-64	11,115	33%
>= 65	9,267	21%
Acres Owned		
< 1%	2,537	46%
1% - 20%	2,632	54%
20% - 40%	2,731	51%
40% - 60%	2,751	46%
60% - 80%	2,551	40%
80% - 100%	2,110	39%
> 100%	14,421	17%
Education		
Less Than High School	1,747	30%
High School	11,410	31%
Some College	8,174	36%
College	8,402	31%
Sales		
Low-sales Small Farms	16,504	17%
Moderate-sales Small Farms	4,420	40%
Midsize Farms	4,923	52%
Smaller Million Dollar Farms	3,220	60%
Larger Million Dollar Farms	666	61%
Total	29,733	32%

Note: Survey weights are not applied

*Non-respondents excluded from calculation

Table 2: Results by Method: All features used

Method	**Accuracy**	**Kappa**	**Precision**	**Negative Predictive Value**	**Recall (Sensitivity)**	**Specificity**
Comparison Models						
Literature Based Logistic Regression	0.7004	0.1882	0.6528	0.8099	0.5795	0.8531
Linear Models						
LASSO Logistic Regression	0.7482***	0.3861***	0.7131***	0.8513***	0.6812***	0.8695***
Ridge Logistic Regression	0.7482***	0.3861***	0.7131***	0.8513***	0.6812***	0.8695***
Stochastic Gradient Descent	0.7296***	0.3862***	0.6954***	0.8549***	0.6951***	0.855***
Naïve Bayes models						
Gaussian naïve Bayes	0.6487	0.3293***	0.6761***	0.8536***	0.697***	0.841
Ensamble models						
Random forest	0.6916	0.07	0.7355***	0.8013	0.5266	0.9095***
Extremely randomized trees	0.684	0.0343	0.7178***	0.7957	0.5128	0.9035***
Gradient tree boosting	0.7398***	0.3362***	0.7096***	0.84***	0.6508***	0.8731***
Bagged K-nearest neighbor	0.7304***	0.306***	0.6965***	0.8336***	0.636***	0.868***
Other models						
Linear Support Vector Machine	0.7487***	0.384***	0.7145***	0.8508***	0.6792***	0.8707***
K-nearest neighbor	0.731***	0.3183***	0.694***	0.836***	0.6438***	0.8649***
Dummy model	0.6774	0	0	0.6774	0	1***

* Metric for model is statistically significantly better (higher) than the literature based logistic regression model at the =0.10 level.

** Metric for model is statistically significantly better (higher) than the literature based logistic regression model at the =0.05 level.

*** Metric for model is statistically significantly better (higher) than the literature based logistic regression model at the =0.01 level.

The best (highest) value of each metric is in bold.

Table 3: Results by Method: Selected features used

Method	Accuracy	Kappa	Precision	Negative Predictive Value	Recall (Sensitivity)	Specificity
Comparison Models						
Literature Based Logistic Regression	0.7004	0.1882	0.6528	0.8099	0.5795	0.8531
Linear Models						
LASSO Logistic Regression	0.7004	0.1881	0.6528	0.8099	0.5794	0.8532
Ridge Logistic Regression	0.7004	0.1882	0.6528	0.8099	0.5795	0.8532**
Stochastic Gradient Descent	0.6937	0.259***	0.6447	0.8234***	0.6231***	0.8363
Naïve Bayes models						
Gaussian naïve Bayes	0.6218	0.2053***	0.5996	0.8133**	0.6118***	0.8055
Ensamble models						
Random forest	0.7346***	0.3542***	0.6947***	0.8442***	0.6665***	0.8605***
Extremely randomized trees	0.6786	0.007	0.6852***	0.7849	0.5026	0.8742***
Gradient tree boosting	0.7056***	0.1853	0.6716***	0.8115**	0.5776	0.8655***
Bagged K-nearest neighbor	0.7056***	0.2003***	0.6651***	0.8129***	0.5844***	0.8598***
Other models						
Linear Support Vector Machine	0.7005	0.1786	0.6558***	0.8087	0.5746	0.8563***
K-nearest neighbor	0.7054***	0.2453***	0.655*	0.82***	0.6096***	0.8471
Dummy model	0.6774	0	0	0.6774	0	1***

* Metric for model is statistically significantly better (higher) than the literature based logistic regression model at the =0.10 level.

** Metric for model is statistically significantly better (higher) than the literature based logistic regression model at the =0.05 level.

*** Metric for model is statistically significantly better (higher) than the literature based logistic regression model at the =0.01 level.

The best (highest) value of each metric is in bold.

7 Appendix

Table A1: Features Used in Machine Learning Models

Feature	Feature description
GCFI*	Gross cash farm income
FARMHHI	Total farm household income
TOTOFI*	Off-farm household income
FamilyFarm*	Family farm (Yes/No)
AL*	Alabama farm (Yes/No)
AR*	Arkansas farm (Yes/No)
CA*	California farm (Yes/No)
FL*	Florida farm (Yes/No)
GA*	Georgia farm (Yes/No)
IL*	Illinois farm (Yes/No)
IN*	Indiana farm (Yes/No)
IA*	Iowa farm (Yes/No)
KS*	Kansas farm (Yes/No)
KY*	Kentucky farm (Yes/No)
MI*	Michigan farm (Yes/No)
MN*	Minnesota farm (Yes/No)
MS*	Missssissippi farm (Yes/No)
MO*	Missouri farm (Yes/No)
NE*	Nebraska farm (Yes/No)
NC*	North Carolina farm (Yes/No)
ND*	North Dakota farm (Yes/No)
OH*	Ohio farm (Yes/No)
OK*	Oklahoma farm (Yes/No)
PA*	Pennsylvania farm (Yes/No)
SD*	South Dakota farm (Yes/No)
TX*	Texas farm (Yes/No)
WA*	Washington farm (Yes/No)
WI*	Wisonsin farm (Yes/No)
Northeast*	Residual Northeast region farm (Yes/No)
South*	Residual South region farm (Yes/No)
West*	Residual West farm (Yes/No)
AcresOwnedPercent	Percent of operated acres that are owned
AcresCroplandPercent	Percent of operated acres that are cropland
AcresOwned	Total acres owned
CroplandAcres	Total cropland acres
AcresOwnedPercentLT01	Category 1 acres owned
AcresOwnedPercentBtw01_20	Category 2 acres owned
AcresOwnedPercentBtw20_40	Category 3 acres owned
AcresOwnedPercentBtw40_60	Category 4 acres owned
AcresOwnedPercentBtw60_80	Category 5 acres owned
AcresOwnedPercentBtw80_100	Category 6 acres owned
AcresOwnedPercentGT100	Category 7 acres owned

WheatFarm*	Wheat specialized farm
CornFarm*	Corn specialized farm
SoybeanFarm*	Soybean specialized farm
CottonFarm*	Cotton specialized farm
OtherCropFarm*	Other crop specialized farm
SpecialtyCropFarm*	Specialty crop specialized farm
CattleCalveFarm*	Cattle and calve specialized farm
HogFarm*	Hog specialized farm
PoultryEggFarm*	Poultry and egg specialized farm
DairyFarm*	Dairy specialized farm
OtherLivestockFarm*	Other livestock specialized farm
OperatorAge*	Primary operator's age
OperatorsAllYoung*	Are some operators <35 (Yes/No)
OperatorsSomeYoung*	Are all operators <35 (Yes/No)
LowSalesSmallFarm	Gross cash farm income <150,000
ModerateSalesSmallFarm	Gross cash farm income between 150,000 and 350,000
MidsizeFarm	Gross cash farm income between 350,000 and 1,000,000
SmallerMillionDollarFarm	Gross cash farm income between 1,000,000 and 5,000,000
LargerMillionDollarFarm	Gross cash farm income >=5,000,000
OperatorRetired	Primary operator retired? (Yes/No)
OperatorWorksOfffarm	Primary operator works off-farm? (Yes/No)
OperatorFemale	Primary operator female? (Yes/No)
OperatorEducSomeHS	Education category 1
OperatorEducHS	Education category 2
OperatorEducSomeCollege	Education category 3
OperatorEducCollege	Education category 4
AssetTotal	Total farm assets
AssetCurrent	Current farm assets
AssetNonCurrent	Noncurrent farm assets
AssetRealEstate	Real estate farm assets
RealDebt	Real estate debt
NonrealDebt	Nonreal estate debt
NonrealDebtShort	Short-term nonreal estate debt
NonrealDebtLong	Long-term nonreal estate debt
FCSloan	Has an FCS loan? (Yes/No)
FSAloan	Has an FSA loan? (Yes/No)
CommercialLoan	Has a commercial loan? (Yes/No)
LifeInsLoan	Has a life insurance loan? (Yes/No)
FarmerMacLoan	Has an Farm Mac loan? (Yes/No)
ImplementDealerLoan	Has an implement dealer loan? (Yes/No)
OtherLoan	Has an other loan? (Yes/No)
Metro2013	Farm in metro county 2013? (Yes/No)
UnemploymentRate2009	2009 county unemployment rate
UnemploymentRate2013	2013 county unemployment rate
UnemploymentRate2014	2014 county unemployment rate
UnemploymentRateChange13_14	2013 to 2014 county unemployment rate percent change

UnemploymentRateChange09_14	2009 to 2014 county unemployment rate percent change

*Used in literature-based models

Table A2: Feature Importance Scores for Random Forest

		Cumulative
	Feature	Feature
Feature	Importance	Importance
GCFI	7.11	7.11
AcresOwnedPercent	6.61	13.72
LowSalesSmallFarm	6.51	20.23
NonrealDebtLong	6.39	26.62
AcresOwnedPercentGT100	6.02	32.65
CroplandAcres	5.36	38.00
OFI_UnemploymentRate2014	4.73	42.73
OFI_UnemploymentRate2013	4.65	47.39
Farmbusiness	3.90	51.29
ImplementDealerLoan	3.51	54.80
OFI_UnemploymentRate2009	3.46	58.26
RealDebt	3.17	61.43
AssetTotal	2.67	64.11
CommercialLoan	2.58	66.69
FARMHHI	2.44	69.13
OperatorAge	2.40	71.54
AcresCroplandPercent	2.18	73.72
SmallerMillionDollarFarm	2.17	75.88
MidsizeFarm	2.16	78.04
HighRiskPref	1.65	79.69
OFI_Metro2013	1.59	81.28
FCSloan	1.53	82.81
OperatorWorksOfffarm	1.53	84.34
AverseRiskPref	1.27	85.61
AcresOwnedPercentBtw01_20	1.26	86.87
NonrealDebtShort	1.22	88.10
OperatorRetired	0.97	89.06
AcresOwnedPercentBtw20_40	0.85	89.91
DairyFarm	0.82	90.74
OperatorsSomeYoung	0.81	91.54
CattleCalveFarm	0.71	92.25
AcresOwned	0.61	92.86
LargerMillionDollarFarm	0.56	93.42
CornFarm	0.55	93.98
AcresOwnedPercentLT01	0.54	94.51
AcresOwnedPercentBtw40_60	0.50	95.01
TOTOFI	0.48	95.50
OperatorFemale	0.43	95.93
OperatorsAllYoung	0.33	96.26
OtherLoan	0.32	96.58
OtherLivestockFarm	0.29	96.87
UnemploymentRate2014	0.27	97.13

ModerateSalesSmallFarm	0.26	97.40
UnemploymentRate2013	0.24	97.63
FL	0.22	97.85
UnemploymentRate2009	0.21	98.06
Metro2013	0.20	98.26
FSAloan	0.18	98.44
SpecialtyCropFarm	0.18	98.61
MissingRiskPref	0.15	98.76
UnemploymentRateChange1314	0.13	98.90
IA	0.11	99.01
SoybeanFarm	0.10	99.11
UnemploymentRateChange0914	0.10	99.21
OperatorEducSomeCollege	0.10	99.30
AcresOwnedPercentBtw60_80	0.10	99.40
CA	0.06	99.46
OFI_UnemploymentRateChange1314	0.05	99.51
AcresOwnedPercentBtw80_100	0.04	99.55
OFI_UnemploymentRateChange0914	0.03	99.58
IN	0.03	99.62
AL	0.03	99.65
SD	0.03	99.68
CottonFarm	0.03	99.71
NE	0.03	99.74
TX	0.03	99.77
IL	0.03	99.80
GA	0.02	99.82
NeutralRiskPref	0.02	99.84
OperatorEducHS	0.02	99.85
MI	0.01	99.87
HogFarm	0.01	99.88
PoultryEggFarm	0.01	99.89
WA	0.01	99.90
WI	0.01	99.91
KS	0.01	99.93
West	0.01	99.94
OperatorEducCollege	0.01	99.95
FamilyFarm	0.01	99.96
WheatFarm	0.01	99.96
ND	0.01	99.97
FarmerMacLoan	0.01	99.97
OK	0.00	99.98
MS	0.00	99.98
MN	0.00	99.98
LifeInsLoan	0.00	99.99
Northeast	0.00	99.99
AR	0.00	99.99
NC	0.00	99.99

OperatorEducSomeHS	0.00	100.00
PA	0.00	100.00
OH	0.00	100.00
KY	0.00	100.00
South	0.00	100.00
MO	0.00	100.00
OtherCropFarm	0.00	100.00
LifeInsLoan	0.00	100.00
OtherCropFarm	0.00	100.00
South	0.00	100.00
OFI_UnemploymentRateChange09_14	0.00	100.00

Note: Feature importance scores are scaled by 100.

Table A3: Feature Importance Scores for Gradient Tree Boosting

Feature	Feature Importance	Cumulative Feature Importance
NonrealDebtLong	16.11	16.11
GCFI	13.10	29.21
AcresOwnedPercentGT100	12.72	41.93
RealDebt	8.60	50.53
AcresOwnedPercent	7.05	57.58
LowSalesSmallFarm	6.96	64.55
OFI_UnemploymentRate2013	5.37	69.92
AssetTotal	4.58	74.50
OperatorAge	4.05	78.55
SmallerMillionDollarFarm	3.83	82.37
NonrealDebtShort	3.71	86.08
FARMHHI	2.81	88.89
HighRiskPref	2.37	91.26
OFI_Metro2013	1.33	92.59
CommercialLoan	1.27	93.86
AcresOwnedPercentBtw01_20	1.12	94.98
CornFarm	0.91	95.90
CroplandAcres	0.59	96.49
OtherLoan	0.44	96.93
UnemploymentRate2009	0.39	97.32
OFI_UnemploymentRate2009	0.39	97.71
OFI_UnemploymentRate2014	0.34	98.05
AcresOwned	0.31	98.36
FCSloan	0.26	98.62
CA	0.26	98.88
FL	0.22	99.10
OperatorsSomeYoung	0.21	99.32
ModerateSalesSmallFarm	0.21	99.53
AcresCroplandPercent	0.20	99.73
UnemploymentRateChange0914	0.12	99.85
FSAloan	0.10	99.95
SoybeanFarm	0.04	99.99
UnemploymentRate2013	0.01	100.00
IA	0.00	100.00
TX	0.00	100.00
WA	0.00	100.00
WI	0.00	100.00
Northeast	0.00	100.00
South	0.00	100.00
PA	0.00	100.00
IN	0.00	100.00

West	0.00	100.00
IL	0.00	100.00
GA	0.00	100.00
SD	0.00	100.00
OK	0.00	100.00
KS	0.00	100.00
OH	0.00	100.00
TOTOFI	0.00	100.00
ND	0.00	100.00
FamilyFarm	0.00	100.00
AR	0.00	100.00
NC	0.00	100.00
AL	0.00	100.00
NE	0.00	100.00
MO	0.00	100.00
MS	0.00	100.00
MN	0.00	100.00
MI	0.00	100.00
KY	0.00	100.00
OFI_UnemploymentRateChange0914	0.00	100.00
AcresOwnedPercentLT01	0.00	100.00
ImplementDealerLoan	0.00	100.00
OperatorEducSomeHS	0.00	100.00
OperatorEducHS	0.00	100.00
OperatorEducSomeCollege	0.00	100.00
OperatorEducCollege	0.00	100.00
LifeInsLoan	0.00	100.00
FarmerMacLoan	0.00	100.00
Metro2013	0.00	100.00
OperatorWorksOfffarm	0.00	100.00
UnemploymentRate2014	0.00	100.00
UnemploymentRateChange1314	0.00	100.00
MissingRiskPref	0.00	100.00
NeutralRiskPref	0.00	100.00
AverseRiskPref	0.00	100.00
Farmbusiness	0.00	100.00
OperatorFemale	0.00	100.00
OperatorRetired	0.00	100.00
AcresOwnedPercentBtw20_40	0.00	100.00
OFI_UnemploymentRateChange1314	0.00	100.00
AcresOwnedPercentBtw40_60	0.00	100.00
AcresOwnedPercentBtw60_80	0.00	100.00
AcresOwnedPercentBtw80_100	0.00	100.00
WheatFarm	0.00	100.00
CottonFarm	0.00	100.00
OtherCropFarm	0.00	100.00
CattleCalveFarm	0.00	100.00

LargerMillionDollarFarm	0.00	100.00
HogFarm	0.00	100.00
PoultryEggFarm	0.00	100.00
DairyFarm	0.00	100.00
OtherLivestockFarm	0.00	100.00
OperatorsAllYoung	0.00	100.00
MidsizeFarm	0.00	100.00
SpecialtyCropFarm	0.00	100.00
LifeInsLoan	0.00	100.00
OtherCropFarm	0.00	100.00
South	0.00	100.00
OFI_UnemploymentRateChange09_14	0.00	100.00

Note: Feature importance scores are scaled by 100.

Notes

[1] Logistic regression and many other standard econometric models are technically simple machine learning models. We refer to logistic regression as 'literature-based logistic regression' to emphasize our comparison between existing or standard econometric methods and the advanced machine learning techniques we test.

[2] We also included people that reported a new loan in the debt table from 2014 as having applied for new agricultural financing, as high response rate was likely influenced by language in the survey that stated "response to this inquiry is required by law" may have influenced responses to debt-related questions

[3] In practice, many machine learning algorithms can be used for continuous or categorical variables. However, some algorithms are specific to dependent variable type.

[4] measures of income, location of farm (state/region), commodity specialization, farm operator demographics,as discussed in the Background section, as indicated in A1.

[5] For our data this would require each model to be estimated 28,601 times. Even if each iteration could be estimated in one minute, the 28,601 iterations for a given model would take nearly an entire day to run and it would take the better part of a a week to run all the models.

[6] The probability of predicting the outcome correctly by chance is calculated using the confusion matrix according to the formula $E = (\frac{TP+FP}{N})(\frac{TP+FN}{N}) + (\frac{FN+TN}{N})(\frac{FP+TN}{N})$.

[7] The kappa statistic is also often reported on a scale of -1 to 1, but we prefer to multiply by 100 so it is on the same scale as accuracy

[8] A negative value for kappa would indicate the model found a relationship between the input data and event outcome that predicted the opposite of what happens. In practice, machine learning techniques are designed find concordant relationships between input and output data so this is unlikely to occur.

OTHER A.E.M. WORKING PAPERS

WP No	Title	Fee (if applicable)	Author(s)
2017-14	Too Much to Eat It All: How Package Size Impacts Food Waste		Petit, O., Lunardo, R. and B. Rickard
2017-13	Why Secondary Towns Can Be Important For Poverty Reduction - A Migrant's Perspective		Ingalaere, B., Christiaensen, L., De Weerdt, J. and R. Kanbur
2017-12	Citizenship, Migration and Opportunity		Kanbur, R.
2017-11	The Digital Revolution and Targeting Public Expenditure for Poverty Reduction		Kanbur, R.
2017-10	What is the World Bank Good For? Global Public Goods and Global Institutions		Kanbur, R.
2017-09	Sub-Saharan Africa's Manufacturing Sector: Building Complexity		Bhorat, H., Kanbur, R., Rooney, C. and Steenkamp, F.
2017-08	Farmer Productivity By Age Over Eight U.S. Census Years		Tauer, L.W.
2017-07	Inequality Indices as Tests of Fairness		Kanbur, R. and Snell, A.
2017-06	The Great Chinese Inequality Turn Around		Kanbur, R., Wang, Y. and Zhang, X.
2017-05	A Supply Chain Impacts of Vegetable Demand Growth: The Case of Cabbage in the U.S.		Yeh, D., Nishi, I. and Gómez, M.
2017-04	A systems approach to carbon policy for fruit supply chains: Carbon-tax, innovation in storage technologies or land-sparing?		Alkhannan, F., Lee, J., Gómez, M. and Gao, H.
2017-03	An Evaluation of the Feedback Loops in the Poverty Focus of World Bank Operations		Fardoust, S., Kanbur, R., Luo, X., and Sundberg, M.
2017-02	Secondary Towns and Poverty Reduction: Refocusing the Urbanization Agenda		Christiaensen, L. and Kanbur, R.
2017-01	Structural Transformation and Income Distribution: Kuznets and Beyond		Kanbur, R.
2016-17	Multiple Certifications and Consumer Purchase Decisions: A Case Study of Willingness to Pay for Coffee in Germany		Basu, A., Grote, U., Hicks, R. and Stellmacher, T.
2016-16	Alternative Strategies to Manage Weather Risk in Perennial Fruit Crop Production		Ho, S., Ifft, J., Rickard, B. and Turvey, C.

Paper copies are being replaced by electronic Portable Document Files (PDFs). To request PDFs of AEM publications, write to (be sure to include your e-mail address): Publications, Department of Applied Economics and Management, Warren Hall, Cornell University, Ithaca, NY 14853-7801. If a fee is indicated, please include a check or money order made payable to Cornell University for the amount of your purchase. Visit our Web site (*http://dyson.cornell.edu/research/wp.php)* for a more complete list of recent bulletins.