International Food and Agribusiness Management Review
Volume 21 Issue 8, 2018; DOI: 10.22434/IFAMR2017.0098

Received: 29 September 2017 / Accepted: 13 July 2018

Can machine learning improve prediction – an application with farm survey data

RESEARCH ARTICLE

Jennifer Ifft[ⓘa], Ryan Kuhns[b], and Kevin Patrick[c]

[a]*Assistant Professor, Charles H. Dyson School of Applied Economics and Management, Cornell University, 451B Warren Hall, Ithaca, NY 14853, USA*

[b]*Economist, Farmer Mac, 1999 K Street NW, 4th Floor, Washington, DC 20006, USA*

[c]*Lead Technologist/Data Scientist, Booz Allen, 901 15th St NW, Washington, DC 20005, USA*

Abstract

Businesses, researchers, and policymakers in the agricultural and food sector regularly make use of large public, private, and administrative datasets for prediction, including forecasting, public policy targeting, and management research. Machine learning has the potential to substantially improve prediction with these datasets. In this study we demonstrate and evaluate several machine learning models for predicting demand for new credit with the 2014 Agricultural Resource Management Survey. Many, but not all, of the machine learning models used are shown to have stronger predictive power than standard econometric approaches. We provide a cost based model evaluation approach for managers to analyze returns to machine learning methods relative to standard econometric approaches. While there are potentially significant returns to machine learning methods, research objectives and firm-level costs are important considerations that in some cases may favor standard econometric approaches.

Keywords: machine learning, prediction, agricultural resource management survey, farm debt, credit demand
JEL code: Q14, C53

ⓘCorresponding author: jifft@cornell.edu

1. Introduction

There is an ever-increasing number of large datasets that can be used by businesses, government, and academic researchers to solve challenges facing agriculture, food, and the environment. Sonka (2014) emphasizes that both firms 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. Management and warehousing of this data is an ongoing challenge (Woodard, 2016), and there are serious issues with maintaining the privacy and security of farm data (Sykuta, 2016). Along with these challenges, researchers must also adapt to methodological advances that have the potential to improve management 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, or plot of land.

In this study we consider the potential of machine learning to improve the prediction of demand for new credit, using a dataset that is well-known to U.S. agricultural economists: the Agricultural Resource Management Survey (ARMS). ARMS data is used for a variety of official statistics, forecasting, and economic research. Improved prediction has many practical business and policy uses, as well as research applications. Machine learning and prediction may be useful to policymakers. When targeting is necessary, say for policies aimed towards specific groups, such as beginning farmers, machine learning methods could be used with administrative data to lower costs. 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 the United States Department of Agriculture (USDA) forecasts (Tack *et al*., 2017).

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). Many firms have large private databases and sales data, in addition to access to public datasets. More accurate targeting of potential customers or prediction of demand could improve profitability of agribusinesses. There are many examples of machine learning being used by food and agricultural businesses. 'Big data' firms are currently competing to solve agriculture and food sector problems through data analytics and new technology (Sparapani, 2017). John Deere recently acquired a company that is developing machine learning technology to identify weeds to selectively spray herbicide on (Graham, 2017). Walmart has developed a product that uses machine learning to identify how soon produce will spoil, which is believed to have improved produce quality and already saved $86 million in food waste costs (Musani, 2018). Machine learning has also made demonstrated improvements to general business processes, including customer service and relationships, hiring processes, and supply chains (Wellers *et al*., 2017). Coble *et al*. (2018) provide many additional examples of 'Big Data' applications for food and agriculture.

There are multiple ways that machine learning is used in the management and marketing literature. George *et al*. (2014) discuss how 'Big Data' has the potential to "transform management theory and practice". Part of this process includes how 'Big Data' and techniques such as machine learning allow research to move beyond a sole focus on *P*-values and establishing causality, to develop a broad base of evidence on management issues from a diverse range of sources ("cosilience"). Cui *et al*. (2006) used machine learning – Bayesian Networks with Evolutionary Programming – to estimate response to direct marketing with a large direct marketing dataset. Dzyabura and Hauser (2011) use machine learning techniques to test for whether consumers use heuristic decision rules, which has useful implications for marketing and demand modeling. Bajari *et al*. (2015) used several machine learning techniques to estimate demand for salty snacks with grocery store scanner data and found that the techniques were better at prediction than standard econometric models.

In addition to improving prediction, other methodological issues surrounding research using food and agricultural datasets may benefit from machine learning methods. Future research could apply the machine learning methods demonstrated here to statistical inference and estimation of heterogeneous treatment effects, i.e. (Wager and Athey, in press). Prediction with machine learning can also generally improve a variety of

econometric models, for example through variable selection or dimension reduction (Mullainathan and Spiess, 2017).

Another issue is that 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 by explicitly including raw survey responses and variables indicating missing observations in statistical models.

We provide a step-by-step guide on how to setup a machine learning problem and how these methods can be evaluated and compared to prediction with standard econometric models. Throughout, we address commonly-held concerns regarding machine learning, with a focus on over-fitting 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 a cost-based evaluation approach that would allow managers to analyze machine learning relative to standard econometric methods, based on the costs associated with incorrect predictions faced by their firm. We conclude with a discussion of other important considerations for firms and management researchers considering use of machine learning methods.

2. Background

In this study we apply machine learning methods to ARMS 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. Farm lenders could also use our approach to better predict which farms are more likely face challenges with loan repayment.

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 and 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. One exception, 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 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 dataset 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. 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 nine additional 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' model.

3. Data

The data used in this study comes from the 2014 Agricultural Resource Management Survey. 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.

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 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%) answered this question.[2] 32% (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 to as credit applicants) and farm operators that did not apply for a new loan (which we will refer to 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

[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. The high response rate was likely influenced by language in the survey that stated "response to this inquiry is required by law", which may have influenced responses to debt-related questions.

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% of non-applicants reached that sales level. The difference between the two groups increases as the sales benchmark increases. 25% of credit applicants had more than $1,000,000 in sales compared to less than 8% for non-applicants. To some degree, this may reflect that ARMS over samples large farms, who are well-known to use more credit. We do not use survey weights in our analysis, as not all machine learning algorithms can accommodate survey weights and our research objective is not to estimate parameters that are representative of all U.S. farms.

Table 1. Summary statistics.[1]

		Number	Share credit applicants (%)[2]
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

[1] Survey weights are not applied.

[2] Non-respondents excluded from calculation.

4. Model selection and evaluation

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). 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: new credit applicants or non-applicants. 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 variables and many machine learning algorithms are explicitly designed to provide greater predictive accuracy in this context.

To evaluate the benefits of applying machine learning to predict if a farmer applied for a loan, we evaluate the predictive performance of several machine learning models relative to standard econometric techniques. Many machine learning models are designed to be able to take advantage of a larger set of features than in normal statistical models; therefore, any performance improvements could reflect the models underlying ability to take advantage of this extra information. Additionally, the models could show more predictive power than standard econometric approaches given the same information. To enhance our comparison, we also evaluate the predictive performance of each machine learning model using the same subset of literature-guided variables used in our standard econometric approach.

In addition to choosing the models 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. 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. We select each of the tuning parameters for the models using out-of-sample cross-validation to determine the model with the best predictive performance for the research question at hand.

4.1 Model selection

While there are many applicable supervised machine learning methods, we choose 9 common approaches not typically used in the standard econometric literature to explore the relative benefits of using machine learning for agricultural prediction problems. The chosen machine learning models fall into five broad categories: generalized linear models, Bayesian models, ensemble models, support vector machines, and nearest neighbor models. Selected models from each group are described below along with their potential strengths and weaknesses.

- *Generalized linear models*

The most basic family of models is known as generalized linear models. Many commonly used models in the agricultural economics literature, including ordinary least squares and logistic regression models, fall into this category of models. To provide a baseline we first estimate a simple logistic regression, predicting the likelihood a farmer applied for a loan using the prevailing literature to guide variable selection for this model.[3]

We then build on the standard logit model by using several regularization techniques designed to determine if there are 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

[3] Measures of income, location of farm (state/region), commodity specialization, farm operator demographics, as discussed in the Background section and indicated in Table S1.

examples, Ridge and Lasso, to the logit model. The Ridge complexity penalty is a function of the sum of squared coefficients. Therefore, Ridge tends to encourage many features with small coefficients rather than completely zeroing out a feature (Hastie *et al*., 2009). On the other hand, Lasso's penalty is a function of the sum of absolute value of the coefficients. Lasso is therefore more likely to zero out coefficients, effectively removing those features from the model (Hastie *et al*., 2009). Because both regularization models have penalties that involve sums of estimated coefficients, the coefficients scale and therefore the scale of the underlying features 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. The performance of Ridge and Lasso logistic regression is dependent on the relative strength of hyperparameter governing either the Ridge or Lasso regularization penalty. A stronger penalty encourages greater sparsity, while a weaker penalty results in less regularization.

■ *Bayesian models*

Bayes models are a family of supervised machine learning models that employ Bayes theorem of conditional updating to make predictions. We use the naïve Bayes models, which also makes the added 'naïve' assumption that the features are independent (Kuhn and Johnson, 2013). From this family of models, we choose the Gaussian naïve Bayes, 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. An advantage of Bayes models is that they tend can perform well even if there is a relatively small amount of training data and are computationally efficient.

■ *Ensemble models*

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. To illustrate common ensemble approaches, we compare four ensemble models that take different approaches to combining base models.

The first is bootstrap aggregation (bagging), where the base model is fit to multiple re-samples of the training data (Hastie *et al*., 2009). The mean or mode of the individual bootstrap samples are then used as the model's prediction. We implement a bagged-k-nearest neighbor method model, which requires the number of bagged models along with the choice of k nearest neighbors to be chosen.

Bagging can improve model performance by reducing the variance component of prediction error. Therefore, it can have an even greater benefit when applied to models that have inherently greater variability, such as classification trees (Hastie *et al*., 2009). We use two common variants on bagging to combine tree based models. Random forest models (RF) take the concept of bagging a step further by randomizing the subset of features used to build each tree. This results in less correlation between the models on each re-sampled dataset, which should allow for further reduction in the variance component of prediction error relative to simple bagging (Hastie *et al*., 2009). We also test a variant on RF models, known as extremely randomized trees, that randomize both the subset of features and splitting thresholds (Hastie *et al*., 2009).

Each of these tree-based ensemble models fits multiple classification trees and averages the results. For each tree, the data is split into two groups based on the 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. RF tends to be most sensitive to the number of randomly selected features used to create each tree and this hyperparameter is commonly tuned to improve predictive performance.

The final ensemble model we consider uses a technique called boosting. Unlike bagging where the base models are learned independently, boosting methods apply the base model sequentially while seeking to minimize added bias at each step. We specifically choose gradient tree boosting, which iteratively adds decision trees in stages. After each stage, a new tree is built for the residual error remaining after the previous stage; 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 controls the weighting used to combine each stage's predictions, while the number of boosting steps controls the number of iterative steps undertaken. Ultimately there is a trade-off between the two parameters, smaller learning rate values typically require greater number of boosting iterations (Hastie *et al*., 2009).

■ *Support vector machine models*

The Support Vector Machine (SVM) algorithm splits the data into two classes by fitting an optimal decision boundary. While kernel techniques allow the SVM algorithm to incorporate a variety of nonlinearities, we implement a linear version. 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. In practice, the outcome classes are often not completely separable and the SVM algorithm must trade off the benefits a larger margin with the cost of misclassifying some existing observations as the margin increases. In practice, the tolerance for classifying existing observations incorrectly is treated as a hyperparameter.

■ *Nearest neighbor models*

The k-nearest neighbor method makes predictions using information from other data points local to the observation being predicted (Hastie *et al*., 2009). This model uses a distance function to determine how similar/dissimilar the features of the data point to be predicted are relative to those of the other known observations. The most commonly observed class among the k closest points is then used to form a prediction. Because neighbors are judged on the basis of feature distance, the k-nearest neighbors approach is sensitive to the scaling of features. As with the regularized logit models, we normalize all features so variables with broader scales do not dominate the distance calculations. The number of nearest neighbors, k, used is typically chosen to minimize prediction error.

4.2 Model evaluation

In many economic analyses 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 considered.

■ *Model evaluation strategy*

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. These over-fit models tend to generalize poorly, resulting in poor predictive performance when applied to other data. Therefore, we 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 use repeated stratified k-fold cross-validation.

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. Repeated stratified k-fold cross-validation repeats this process *r* times with the data randomly assigned to new k-fold subsets each time, reducing the impact of the random fold assignments on model evaluation.

The decision on the number of folds requires consideration between computational resources, as well as the bias/variance trade-off associated with the 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). On the other hand, using a smaller value of *k* will result in additional bias but less variance. Depending on the exact trade-off this can potentially improve 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 predictions 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 and to repeat the cross-validation process 10 times. This has the benefit of reducing the variance resulting from the random assignment to each of the *k* folds, without adding too much computational burden.

Since the same repeated 10-fold cross-validation process is applied to each model, the resulting 100 measures of model performance can be compared to gauge whether differences across the estimated models are statistically different. When comparing accuracy metrics estimated via cross-validation, the results from each model for a given fold are calculated on the same holdout observations. Therefore, statistical tests for dependent or paired sample must be used. Rather than make an assumption on each performance metric's distribution, we choose to use the Wilcoxon Signed Rank test, a nonparametric test for matched samples, to compare statistical differences in model accuracy.

■ *Model performance metrics*

The terminology used can sometimes differ, but the methods used to evaluate the predictive accuracy of machine learning models aligns with the forecast evaluation literature. In our analysis, true positives are the correctly predicted cases where a farm applied for an application and true negatives are the correctly predicted cases where the farm operation did not apply for new credit. Alternatively, false positives occur when the model incorrectly predicts that a farm applied for a loan and false negatives occur when the model incorrectly predicts that a farm did not apply for a loan.

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

$$Accuracy = \frac{True\ positives + True\ negatives}{All\ predictions} \quad (1)$$

Accuracy provides a high-level view of a model's accuracy, but it weights the ability to identify applicants and non-applicants equally well. Because our analysis' 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, measures the 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 can be interpreted as the percent of farms that were correctly predicted as applying for a loan out of the total that actually applied.

$$Recall(sensitivity) = \frac{True\ positive}{True\ positives + False\ negatives} \quad (2)$$

While recall is useful in assessing model accuracy, it is conditioned on the event of interest, in our case having applied for credit, having occurred or not occurred (Kuhn and Johnson, 2013). Precision, is a measure

of the unconditional probability of the model's prediction being correct. In our case, precision measures the percentage of times the model predicted a farm applied for a loan and the farm actually was a credit applicant.

$$Precision\ (Positive\ predictive\ value) = \frac{True\ positives}{True\ positives + False\ positives} \tag{3}$$

In each of the outlined model performance metrics, incorrect predictions are given an equal weight. An alternative is to account for differences between the cost of false negative and false positive predictions, which can be unequal in practice. In the case of an agricultural finance institution targeting whether farmers will apply for a loan, the cost of a false positive would be the marketing and customer acquisition costs. Alternatively, the cost of a false negative is the income that is not earned from the never targeted farm operation. Depending on the size of these costs, an agribusiness manager choosing between models may prefer a model that tilts its incorrect predictions toward false negatives or false positives. In practice, an agribusiness manager choosing between models could apply knowledge of their operations actual costs associated with each of these items; however, we illustrate this approach using the cost adjustment term lambda to weight inaccurate predictions. If lambda is between 0 and 1, the cost of a false negative is lower than a false positive. On the other hand, lambda values above one signal that the cost of a false negative is larger than a false positive.

$$C = \lambda \times False\ negatives + False\ positives \tag{4}$$

■ *Feature importance*

The importance of each feature can be constructed for decision tree based models, including random forest and gradient tree boosting. Feature importance is a relative score determined by the amount that splitting at that feature node in the decision tree improves some predictive metric. We use the average Gini impurity metric proposed in James *et al*. (2013). Feature importance scores 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, it cannot be interpreted the same way as statistical significance in a typical regression model. Feature importance should not be interpreted in a similar manner to regression coefficients or marginal effects, as they only measure the importance to prediction and are not a precise measure of impact on probabilities or other economic interpretations. 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 contributes to prediction, but a low rank does not necessarily mean that a particular variable is not a determinant of the outcome.

5. Results

The metrics described in the evaluation section are reported for each model in Table 2. We use these metrics to evaluate the success of each model in predicting whether a farm operation applied for credit in 2014 and provide context on how an agricultural finance institution might determine whether or not to use a machine learning model instead of a standard econometric approach to target customers. Table 2 summarizes the model evaluation metrics by averaging across the repeated cross-validation output. In the first three columns, each model was limited to the same features used in the literature based logistic regression, providing insight into whether the machine learning algorithms can glean greater predictive ability from the same data. For columns 4-6, the results are estimated using the expanded set of features to quantify the benefit of each machine learning model's ability to handle a large number of features. Finally, Table 3 is used to illustrate how an agribusiness manager could use their knowledge of the differential cost of inaccurate predictions to decide which model their business should use.

Table 2. Results by method.[1]

Model	Literature selected features			All features		
	Accuracy	Recall	Precision	Accuracy	Recall	Precision
Comparison models						
Literature based log. regression	0.7004	0.2386	0.5881	0.7004	0.2386	0.5881
Linear models						
LASSO logistic regression	0.7004	0.2386	**0.6430**	0.7482***	0.4916***	0.6437***
Ridge logistic regression	0.7004	0.2386	0.6435	0.7481***	0.4919***	0.6435***
Naıve Bayes models						
Gaussian naıve Bayes	0.6218	**0.5836*****	0.4357	0.6487	**0.8332*****	0.4754
Ensamble models						
Random forest	0.7324***	0.4711***	0.6106***	**0.7594*****	0.5537***	**0.6491*****
Extremely randomized trees	0.6998	0.418***	0.5454	0.7468***	0.5270***	0.6282***
Gradient tree boosting	**0.7337*****	0.4835***	0.6101***	0.7585***	0.5553***	0.6465***
Bagged K-nearest neighbor	0.7056***	0.2427***	0.6103***	0.7304***	0.3701***	0.6426***
Other models						
Linear support vector machine	0.6138	0.3850***	0.4542	0.6708	0.4729***	0.5275
K-nearest neighbor	0.6702	0.4192***	0.4870	0.7043***	0.4219***	0.5548

[1] *, **, *** Metric for model is statistically significantly better (higher) than the literature based logistic regression model 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. Model results by relative costs of false negatives and false positives.[1,2]

Model	λ value								
	0.01	0.1	0.2	0.5	1	2	5	10	100
Comparison models									
Literature based log. regression	**159.9**	**223.3**	**293.8**	505.1	857.3	1,561.60	3,674.70	7,196.40	70,587.90
Linear models									
LASSO logistic regression	255.9	298.1	345	485.7***	720.3***	1,189.3***	2,596.5***	4,941.7***	47,156.2***
Ridge logistic regression	256.2	298.4	345.3	485.9***	720.3***	1,189.1***	2,595.5***	4,939.5***	47,131.5***
Naıve Bayes models									
Gaussian naıve Bayes	852.3	866.2	881.6	927.7	1,004.70	1,158.6***	**1,620.3*****	**2,389.9*****	**16,241.8*****
Ensemble models									
Random forest	280.5	317.6	358.7	482.3***	**688.1*****	**1,099.9*****	2,335.1***	4,393.9***	41,451.4***
Extremely randomized trees	292.3	331.5	375.2	506.1	724.3***	1,160.6***	2,469.7***	4,651.4***	43,922.9***
Gradient tree boosting	284.5	321.4	362.5	485.6***	690.7***	1,101.0***	2,331.9***	4,383.4***	41,310.4***
Bagged K-nearest neighbor	195.8	248.1	306.2	**480.6*****	771.1***	1,352.3***	3,095.7***	6,001.5***	58,305***
Other models									
Linear support vector machine	437.3	481.1	529.7	675.6	918.7	1,405.0***	2,863.8***	5,295.1***	49,058.5***
K-nearest neighbor	317.7	365.7	419	579	845.7***	1,379.0***	2,979.1***	5,645.9***	53,648.3***

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

[2] Reported values measure the cost of false positive and false negative predictions according to the formula C = λ × false negatives + false positives. λ less than 1 implies false negatives are less costly than false negatives, λ=1 implies equivalent cost of false positives and false negatives, λ greater than 1 implies false positives are less costly than false negatives.

Focusing first on comparing each model while restricting the features to the subset from the literature based logistic regression model (marked with an * in Supplementary Table S1), enables one to ascertain the differences in each models' predictive ability due to model performance alone. The literature based regression was able to correctly classify a farm as having applied for credit or not in just over 70% of cases. When restricted to using the same data, some but not all the machine learning models improved on this, but the improvements tended to be relatively small ranging from 0.7 to 4.8 percentage points. However, given the goal of identifying credit applicants, recall and precision provide a better gauge of model performance.

The literature based logistic regression model was only able to recall on average just under 24% of the farms that applied for a loan. All the machine learning models except for the penalized logistic regressions were able to statistically significantly outperform the literature based logistic regression model. The improvement in identifying credit applicants also tended to be somewhat larger, with the machine learning algorithms identifying between 2 to 145% more credit applicants.

However, the machine learning algorithms identified more farm applicants in part by being more likely to indicate a farm would apply for a loan when they actually did not. The literature based logistic regression model's prediction that a farm would apply for a loan had an average precision of 59%. Only three of the nine machine learning algorithms tested, including random forest, gradient tree boosting and bagged k-nearest neighbors, were able to correctly label farms as applicants with a statistically significant improvement in precision.

Because the machine learning algorithms can handle a greater number of variables, using additional features can often lead to further performance improvements. Our results show that each machine learning model achieved higher average accuracy, recall and precision scores when applied to the expanded set of features. Corresponding to the improved performance metrics, a greater number of the machine learning algorithms outperformed the literature based logistic regression once they were applied to additional features. All nine machine learning models outperformed the literature based logistic regression in terms of recall at the 1 percent significance level, while six of the nine algorithms had higher precision at the 1% significance level.

In addition to more machine learning models performing statistically significantly better than the literature based logistic regression, the degree of improvement is also larger once the expanded data set is used. Each machine learning model identified a greater portion of credit applicants and the average recall scores were 56 to 252% better than the literature based logistic regression baseline. Not only were the machine learning models able to better identify applicants, the precision of these predictions was 3.6 to 12.7% better.

These results indicate an agricultural finance institution looking to target potential customers could benefit from the use of machine learning. However, the company would still need to determine the machine learning approach they should use. When using the expanded feature set, the Gaussian naïve Bayes approach had the highest average recall score, identifying more than 8 out of every 10 applicants. But the Gaussian naïve Bayes approach also had a lower average precision score than the literature based logistic regression and several other machine learning models. This would make the Gaussian naïve Bayes model a potentially good choice if the focus was identifying potential applicants, even at the cost of potentially having a high number of false positives. On the other hand, if falsely identifying additional credit applicants created a large cost burden, the ensemble machine learning models or LASSO and RIDGE logistic regression models would be good candidates since they have the highest average precision scores.

Directly considering the costs associated with inaccurate predictions provides an alternative method of weighting the trade-off between predictive recall and precision. Although a business might use estimates of its actual costs associated with false positives and false negatives, we use nine values of lambda ranging from 0.01 to 100 to arrive at a measure of the cost of each model's inaccuracy. Even though the literature based logistic regression does not identify as many applicants, it also avoids false positives. Therefore, when the cost of false negative predictions is smallest relative to false positives (lambda values of 0.01, 0.1,

and 0.2), such as when the cost of customer acquisition is high relative to a customer's potential value, the literature based logistic regression model tends to perform well using a cost based approach. But as the cost of foregone customer value (false negatives) relative to customer acquisition (false positives) increases, the costs associated with using the literature based logistic regression rise more quickly than the machine learning approaches.

In addition to providing the potential to improve predictive performance, several of the tested machine learning models provide the ability to glean information on the features most useful for prediction (Supplementary Table S2 and S3). Perhaps one of the most interesting findings related to feature importance 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 25 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 indicate that the feature is not related to applying for a new loan. The feature may just be highly correlated with another feature and therefore not add much to predictive power of the model. Still, some of the features 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. Other important features, such as farm size (assets, gross cash farm income, acres), operator age, and off farm income are consistent with the agricultural finance literature.

6. Conclusions

We are able to show that machine learning provides a set of potentially useful prediction tools when applied to a standard farm survey data set, although the prediction outcomes varied based on the method. Even when restricted to using the same literature selected features, several machine learning approaches were found to perform significantly better than the comparison literature driven logistic regression model. This suggests that machine learning models can lead to better predictions, even when applied to the same data used in typical econometric studies. While additional variables can improve performance of standard econometric approaches, at some point there will be convergence or degrees of freedom issues. When allowed to use more of the expansive set of features available in the ARMS data, the machine learning algorithms performed even better, highlighting the usefulness of machine learning as data sets expand. Accordingly, machine learning represents a potentially useful tool for agribusiness managers and policymakers to consider for prediction problems.

At the same time, whether a machine learning algorithm will improve predictive performance depends on how model performance is being evaluated. We explain several common measures of predictive performance and illustrate how they could be used by agribusiness managers to judge model performance. In our analysis each of the machine learning models had higher recall when allowed to learn from the expanded set of features. But in several cases, this came at the expense of precision, or a higher incidence of false positives. In the end, the best model for a given problem depends on data available and the prediction outcome that is most important to person or organization that will use the model.

As we demonstrated using the cost based model evaluation approach, a business that is interested in accurately targeting users, may prefer a model with high precision if the cost of customer acquisition is high compared to the customer's potential value. In our analysis this would favor the use of the literature driven logistic regression model. However, if the cost of customer acquisition is low, say the cost of mailing a flyer, relative to a customer's potential value, a model with better recall, like the Gaussian naïve Bayes model in our results, would be more appropriate.

While our results highlight how an agricultural firm could apply machine learning to improve predictive performance, our analysis also emphasizes the importance of not blindly applying machine learning models.

It is difficult to rank individual machine learning models, as the appropriate model in practice depends on business objectives, organizational costs or other model specific characteristics. Further, the additional cost of implementing additional models is typically small, as most software packages include several. Additionally, Bajari *et al*. (2015) showed that model results could be combined for superior prediction. While this approach is beyond the scope of this paper, it is a commonly used machine learning approach. Each of the machine learning models compared in this analysis have different strengths and weaknesses, that need to be considered in respect to research objectives. As we explained in the model selection section they are also governed by hyperparameters, which must be tuned to optimize predictive performance. Researchers and modelers remain key "inputs" in the modeling process by 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.

There are additional considerations for whether machine learning methods more beneficial than standard econometric methods. For some data sets, machine learning methods may be feasible in cases where standard approaches are not – for example for very large data sets or when the number of relevant product characteristics is larger than the number of products being analyzed. This study demonstrates that machine learning methods are useful for data sets such as ARMS that are not considered 'Big Data'. In some cases, computing power, data availability, or research objectives may lead the researcher to use standard econometric modeling. Human resources is another consideration. Data science and analytics are increasingly being taught in both undergraduate and graduate programs, but the benefit for some firms may outweigh the training or hiring costs. Likewise, the cost of computing power has declined and there are a variety of free open source software packages like Python and R that can be used to implement machine learning models. Ultimately, researchers will need to weigh their research question or business/policy objectives with available computing resources and data to determine the preferred methodology.

Researchers in food and agricultural economics may be "uniquely positioned" to take advantage of 'Big Data' to address many problems facing the sector (Coble *et al*., 2018). 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 other policy, food and business datasets. As discussed in the introduction, machine learning models are already widely used in industry and banking, for example by credit card companies to evaluate applicants credit worthiness. As indicated by our findings, machine learning does not always improve prediction, let alone the challenges related to statistical inference. 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.

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 and in many cases standard approaches may be preferable, there are 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 (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.

Supplementary material

Supplementary material can be found online at https://doi.org/10.22434/IFAMR2017.0098.

Table S1. Features used in machine learning models.
Table S2. Top 25 feature importance scores, random forest.
Table S3. Top 25 feature importance scores, gradient tree boosting.

Disclaimer

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

References

Bajari, P., D. Nekipelov, S.P. Ryan and M. Yang. 2015. Machine learning methods for demand estimation. *American Economic Review* 105(5): 481-85.

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

Coble, K.H., A.K. Mishra, S. Ferrell and T. Griffin, T. 2018. Big data in agriculture: a challenge for the future. *Applied Economic Perspectives and Policy* 40(1): 79-96.

Cui, G., M.L. Wong and H.-K. Lui. 2006. Machine learning for direct marketing response models: Bayesian networks with evolutionary programming. *Management Science* 52(4): 597-612.

Dzyabura, D. and J. R. Hauser. 2011. Active machine learning for consideration heuristics. *Marketing Science* 30(5): 801-819.

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

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

Foster, I., R. Ghani, R. Jarmin, F. Kreuter and J. Lane. 2016. *Big data and social science, a practical guide to methods and tools*. CRC Press, Boca Raton, Florida, USA.

George, G., M.R. Haas and A. Pentland. 2014. Big data and management. *Academy of Management Journal* 57(2): 321-326.

Graham, K. 2017. John Deere advancing machine learning in agriculture sector. Available at: http://tinyurl.com/y7d9z5n8.

Hastie, T., R. Tibshirani and J. Friedman. 2009. *The elements of statistical learning data mining, inference and prediction*. Springer, New York, NY, USA.

Howley, P. and E. Dillon. 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., K. Patrick and A. Novini. 2014. Debt use by us farm businesses, 1992-2011. Technical report, United States Department of Agriculture, Economic Research Service, Washington, DC, USA.

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

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

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

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

Morehart, M., D. Milkove, Y. Xu. 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. Available at: http://tinyurl.com/yb6xlsf2.

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

Musani, P. 2018. Eden: the tech that's bringing fresher groceries to you. Available at: http://tinyurl.com/y9kxxn6p.

Sonka, S. 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. Available at: http://tinyurl.com/ycgb9j97.

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

Tack, J., K.H. Coble, R. Johansson, A. Harri and B. Barnett. 2018. The potential implications of 'big ag data' for USDA forecasts. Available at: http://tinyurl.com/yasdvthj.

Wager, S. and S. Athey. In press. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, DOI: https://doi.org/10.1080/01621459.2017.1319839.

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

Wellers, D., T. Elliott and M. Noga. 2017. 8 ways machine learning is improving companies' work processes. *Harvard Business Review.* Available at: http://tinyurl.com/y92n5hhm.

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.