Predicting Credit Demand with ARMS: A Machine Learning Approach[*]

Jennifer Ifft[†], Ryan Kuhns[‡], Kevin Patrick[§]

May 19, 2017

Selected Paper prepared for presentation at the 2017 Agricultural & Applied Economics Association Annual Meeting, Chicago, Illinois, July 30-August 1

[*] 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.

[†] Charles H. Dyson School of Applied Economics and Management, Cornell University, jifft@cornell.edu

[‡] Farmer Mac, rkuhns@farmermac.com

[§] USDA Economic Research Service, Washington DC, ktpatrick@ers.usda.gov

1 Introduction

The U.S. farm sector is entering its fourth year of declining net farm 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. The objective of this study is to demonstrate the benefit of applying machine learning methods to data from the USDA Agricultural Resource Management Survey (ARMS) in order to predict whether or not a farm applied for new financing. This will allow us to better predict which farm types will demand more credit in coming years. Better predictions of farms desiring additional financing will allow agricultural finance industry participants to better understand the characteristics of their potential customers and meet their needs. Additionally, the results can inform where industry segments are demanding greater financing and where credit constraints might occur.

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 our focus on this study is demonstrating how machine learning can be used with ARMS data and predicting demand for credit, future extensions could also take advantage of new approaches for machine learning for statistical inference. Further, various sources of detailed survey and market data for the farm sector are available and currently being used in research that might benefit from advances in machine learning methods.

The 2014 ARMS included research questions that asked respondents to indicate whether a respondent applied for new financing in 2014 calendar year. The newly available data allows us to categorize whether or not farm operations applied for new financing, and determine whether or not the demand for new credit can be predicted given other observable data about the operation. As a starting point we use a simple machine learning model, the logit model of binary classification variables, to predict whether or not each operation applied for financing. In order to demonstrate the potential benefits of machine learning methods, we explain the typical machine learning project process and terminology. We then employ an additional five machine learning algorithms to classify whether a farm operation responded to the 2014 ARMS survey indicating that they had applied for new financing. An analysis of each of the six classification methods used suggests machine learning methods can lead to accurate predictions and is a useful tool to add to the toolkit of econometricians and users of econometrician's predictions.

2 Background

Farm debt financing requirements vary by production specialization as well as operator objectives and preferences. Some 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). There are some consistent 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). While these relationships are well-established, they do predict how credit demand will respond to changes in farm sector or macroeconomic conditions.

Structural changes in the farm sector give rise to some methodological issues in modeling credit demand. Harris et al. (2009b) note that the number of farms holding any debt dropped from 60 percent in 1986 to 31 percent in 2007. Harris et al. (2009a) emphasizes the importance of addressing truncation in research on farm loan use. Similar to trends in value of production, Ifft et al. (2014) notes that debt use has drastically declined for low-sales family farms and increased for very large family farms. Choice of functional form is a general issue in modeling credit decisions. Some households are observed with no use of credit: this may be due to credit constraints or while others may simply not demand credit.

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 ARMS 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 separating estimating of discrete decision to use any credit and the decision on amount of credit to use. The decision to use credit is modeled using a logit model, and then five other machine learning models that account for truncation are used to estimate the level of current debt holdings and number of loans. 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), most studies relied on bank data 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.

Some recent US-based papers have used aggregated data to estimate structural models of credit demand by U.S. farms. Bampasidou et al. (2017) use ARMS data to create a state-level panel, to estimate return to farm assets using farm sector characteristics and macroeconomic variables. Hubbs and Kuethe (2017) model farm sector credit demand at the national level from 1978 to 2014 and find frequently periods of excess demand and supply. While these papers elucidate general trends and characteristics of credit demand, analysis at this level cannot take advantage of the predictive potential of key farm-level attributes that drive credit demand.

In addition to improving prediction, other methodological issues surrounding research using ARMS data may benefit from machine learning methods. A major challenge for researchers is how to take advantage of increasingly rich data sets to predict credit demand and other farm-level decisions, within a transparent modeling framework. ARMS has many different measures of debt use as well as hundreds of variables on farm characteristics. The richness of ARMS data is still underutilized for methods of variable selection. Machine learning provides a transparent method to improve variable selection for prediction models. Another issue is that while many variables commonly used in research have imputation for missing responses (i.e. Morehart et al. (2014)), machine learning methods that can use raw survey responses while including missing observations may provide more insight into key drivers of credit demand as well as other decisions that can be estimated using ARMS data.

3 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 additional questions focused on the debt portion of the farm's balance sheet, specifically 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 [1]. 32 percent (9,226 farm operators) answered affirmatively that they did apply for a new loan in 2014.

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.

Credit applicants are younger than non-applicants, with an an average age of 55 and 19 percent below age 45. Non-applicants by comparison had an average age of 61 with only 10 percent below age 45. As would be expected for a younger set of farm operators, credit applicants own a smaller share of the land they operate (66 percent) on average compared to non-applicants (126 percent)[2]. While they own a smaller share of land, credit applicants operate farms that are about twice the size on average (1,428 acres operated vs. 719 acres operated) compared to non-applicants.

Non-applicants were more than twice as likely to be female compared to applicants. About 4 percent of those that applied for new credit were female, compared to 9 percent of those that eschewed new credit. The differences are less stark when comparing the educational attainment of the two groups. Credit applicants had similar levels of education than non-applicants.

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.

Farms specializing in dairy, corn, hogs, soybeans and wheat were the most likely to apply for new credit in 2014. The geographic differences largely track with the primary commodity specializations in the state. For example, farm operators in states where corn and soybeans are a significant portion of production, including many of those in the Midwest, are more likely to have applied for new credit in 2014. Around 40 percent of respondents from Iowa, Nebraska, Indiana, South Dakota, and Illinois applied for a new line of credit in 2014. In contrast, farm operators from states that typically specialize in specialty crops, including Florida and California, were much less likely to have applied for new credit in 2014.

The financial characteristics of credit applicants differed in terms of profitability, solvency, efficiency, and liquidity. Credit applicants had more than twice the net cash farm income ($168,800) compared to non-applicants ($76,600) on average. Half of all credit applicants had an operating profit margin ratio of 4.7 or greater. By contrast the median operating profit margin ratio for non-applicants was -7.4. Credit applicants had more farm assets ($3.6m vs. $2.0m) and more farm debt ($761k vs. $147k). As a group, credit applicants were more leveraged with a debt-to-asset ratio of 0.21, while non-applicants were less leveraged, with a debt-to-asset ratio of 0.07.

The average farm operator that did not apply for credit in 2014 had less current assets than those that applied for credit, but was much more liquid. The average non-applicant had $251,200 in current assets compared to $612,500 for credit applicants. Conversely, the average non-applicant had 64 dollars in current assets for every dollar in current debt, while average credit applicant had nearly 9 dollars in current assets for each dollar of current debt.

4 Empirical Strategy

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. The final step is evaluation. For each step in this prediction pipeline, we provide necessary context and language to compare to standard development of an econometric model.

4.1 Defining the Machine Learning Problem

When attempting to solve a prediction problem using machine learning techniques, it is essential to explicitly pose the question one is trying to solve as a machine learning problem. The type of problem will dictate the needed data and guide the model selection. We define our specific problem as trying to predict the farms that will apply for a new loan. 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. Specifically defining the problem also enables use to gather and prepare the data needed as inputs for the machine learning methods. Data preparation is discussed next.

4.2 Data Preparation

Having defined the problem as a binary classification prediction problem where we want to classify 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 in our case, 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 inference where the estimated coefficients are important, accurate prediction is the goal, hence many of the issues associated with explanatory variable selection for inference can be avoided. 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. The full set of features (or variables) used is reported in table 2.

4.3 Model Selection

Having defined the research question as a machine learning problem and transformed the data, we need to choose the machine learning methods that will actually 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. There are many applicable supervised machine learning methods out of which we choose six methods that are commonly used, each with their own well-established strengths and weaknesses. These six models fall into three broad categories: generalized linear models, naive Bayes models, and ensemble models. 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, ridge regression, and lasso models fall into this category of models. From the large pool of generalized linear models, we choose the logistic regression (logit) and stochastic gradient descent models. The L1 regularized logistic regression model we use fits a logistic curve to the data. L1 regularized logistic regression models have sparse solutions relative to other models. If the goal is to reduce over-fitting, logistic regression might be produce favorable results. The stochastic gradient descent model is another relatively simple generalized linear model that is computationally efficient. The stochastic gradient descent model is an iterative optimization model 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 tends to perform well when given large data sets.

Slightly more complex than generalized linear models, though still fairly simple, naive Bayes models are a family of supervised machine learning models that employ Bayes theorem of conditional updating with the added 'naive' assumption that the features are independent. From this family of models, we choose the Gaussian naive Bayes which makes the added assumption that the likelihood of the features is Gaussian. 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 Gaussian. This model tends to perform well even with relatively small amounts of training data. It is also computationally efficient which means its fast compared to more complex models discussed below.

Ensemble or weighted 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. We chose three ensemble models that take different approaches to combining base models. The first is forests of randomized trees, often referred to as a random forest model. For this model, decision trees are created each with a random subset of the available features. 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 and the results are averaged. This method has the potential to introduce bias based on the selection of features for each tree, but may result in a preferable model due to the reduced variation from averaging a diverse set of decision trees. This model is also easy to perform using parallel processing making it useful for extremely large data sets.

A variant on forests of randomized trees that we test is extremely randomized trees.

Extremely randomized trees goes an additional step further in randomizing the subset of features, by also randomizing the splitting thresholds. This method can increase the overall bias of the results, but tends to reduce the variance over the standard forests of randomized trees. Another type of ensemble model is called boosting, where rather than random forest type models where the base models are performed independently, boosting methods perform the base methods sequentially and try to minimize the added bias at each new model step. This ensemble method is often used 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, which iteratively adds decision trees in stages. Gradient tree boosting tends to perform well when the features are heterogeneous i.e., binary, categorical, and continuous feature variables. 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.

There are numerous other models that could have feasibly been considered, but the chosen set of models give breath of complexity and each has potential benefits over others. 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.

4.4 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 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 a 'stratified k-fold cross-validation' because it allows us to evaluate the model using all observations.

In the case of a stratified k-fold cross-validation, the data set is broken up into k equally sized subsets called folds. It is called stratified because in addition to the equal size requirement, the subsets preserve the proportion of observations observed in each class 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.

Ultimately, the decision on the number of folds used to split the data involves a trade-off between computational resources and the bias associated with estimated accuracy statistics. As k increases the proportion of data used to fit the model increases, resulting in lower potential bias in estimated accuracy measures (Kuhn and Johnson, 2013). In the special case where k equals the number of observations, known as leave one out cross-validation, the difference between the size of the training data and original sample is small, resulting in little bias. However, this approach is very computationally intensive, particularly in larger data sets[3]. Although there are no set rules, 5- or 10- fold cross-validation are commonly used (Kuhn and Johnson, 2013). We choose to use 5-fold cross-validation to evaluate each model in our analysis.

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. Contingency tables, often referred to as a confusion matrix, are commonly used to analyze the ability to predict categorical outcome variables (Kuhn and Johnson, 2013). Because we are interested in the accuracy of predictions of a binary variable our analysis of each models accuracy uses a 2x2 confusion matrix as outlined in table 3. The matrix's diagonal includes the true positive (TP) cases where the model correctly predicted a farm applied for an application and true negatives (TN), where the model was able to correctly discern the farm operation did not apply for new credit. A models 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 to some baseline comparisons. 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)[4]:

$$kappa = (\frac{O - E}{1 - E}) * 100. \tag{2}$$

The kappa statistic measures the agreement between the predicted and observed outcomes on a scale between -100 and 100 [5] [6]; 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. As shown in equation 3, recall can be calculated from the confusion matrix as the number of true positives relative to observed positives.

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

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.

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

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.

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

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

5 Results and Discussion

The metrics described in the previous section are reported in tables 4 and 5. We use these metrics to evaluate the success of each model in predicting whether a farm operation applied for credit in 2014. Table 5 summarizes the confusion matrix results into the accuracy 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 an individual model, but not necessarily so. The logistic regression model was able to correctly predict whether or not a farm operation applied for new credit 77 percent of the time. The stochastic gradient descent (SGD) and extremely randomized trees (ERT) models both were unable to correctly predict as many outcomes as logistic regression. On the other hand the Gaussian naive Bayes (GNB), random forest, and gradient tree boosting models were each able to correctly predict more than 80 percent of outcomes.

Comparing the model using the kappa statistic, which corrects for the likelihood of correct predictions due to random chance, results in the same rank order of model accuracy. The logistic regression model is again more accurate than the SGD and ERT. Likewise, the GNB, gradient tree boosting and random forest models continue to be most accurate. Interestingly, the gap in relative accuracy between logistic regression and the three 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 but the SGD model were able to correctly predict that most of the actual non-applicants did not applied for a loan. The gradient tree boosting machine learning algorithm was particularly adept at correctly assigning actual non-applicants to the non-applicant group, correctly identifying nearly 100 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. The logistic regression, SGD, ERT and gradient tree boosting models had a negative predictive value between 76 and 80 percent, suggesting just more than one out of every five times an operation was predicted to be a non-applicant it was actually an applicant. By comparing both specificity and negative predictive value, it is clear that the gradient tree boosting algorithm is trading the ability to almost perfectly assign non-applicants to the non-applicant group for false negatives, where applicants are incorrectly identified as non-applicants. In contrast the random forest algorithm misclassified more of the actual non-applicants, but also was less 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. 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, recall is between 40 to 50 percent in all but two cases. This suggests the models had a more difficult time classifying the less common applicant observations. Again the logistic regression model's performance falls in the middle of the model pack, identifying 49 percent of applicants as applying for credit. The random forest method performs best, correctly predicting the application of roughly two-thirds of actual applicants. The random forest model also performs relatively well compared to logistic regression at avoiding false positives where non-applicants are predicted to have applied for a loan. Although the GNB and gradient tree boosting models each are better at avoiding false positives, they do so with the trade-off of failing to identify as many actual applicants.

We are able to show that machine learning is a potentially useful prediction tool set when applied to the rich ARMS financial data, though the prediction outcomes varied based on the method. The results emphasize the importance of not blindly applying a machine learning model, but testing different models and understanding the benefits and drawbacks to each model in the context of the research objective. Random forests perform best over most of the prediction metrics and were particularly adept at identifying non-applicants well and applicants better than most other tested models. Random forests are widely used both for prediction and inference, and our findings support the usefulness of this approach. 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 gradient tree boosting. However, if someone is looking to reach as many potential new credit applicants, they may choose a model that had better recall, as was the case for the stochastic gradient descent model.

6 Conclusion

In this study we show how machine learning methods can be used with ARMS data to predict demand for new credit. Some of the more complex machine learning methods can perform better than a standard econometric model (logit) at predicting credit demand. We provide a useful framework for future application of machine learning methods to ARMS data by explaining how machine learning methods are implemented. We illustrate how machine learning can be used transparently and describe how the implementation approaches avoid issues such as over-fitting. 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.

References

Bampasidou, M., Mishra, A. K., and Moss, C. B. (2017). Modeling debt choice in agriculture: the effect of endogenous asset values. *Agricultural Finance Review*, 77(1).

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

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.

Harris, J., Johnson, J., Dillard, J., Williams, R., Dubman, R., et al. (2009a). The debt finance landscape for us farming and farm businesses. *Electronic Outlook Report from the Economic Research Service*, (AIS-87).

Harris, J. M., Dillard, J., Erickson, K., Hallahan, C., et al. (2009b). Changes in debt patterns and financial structure of farm businesses: A double hurdle approach. In *2009 Annual Meeting, July 26-28, 2009, Milwaukee, Wisconsin*, number 49402. Agricultural and Applied Economics Association.

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.

Hubbs, T. and Kuethe, T. (2017). A disequilibrium evaluation of public intervention in agricultural credit markets. *Agricultural Finance Review*, 77(1).

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.

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.

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.

7 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: 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	Misssissippi 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

Table 3: Confusion Matrix Legend

	Predicted	
	Applicants	Non-applicants
Actual applicants	True positives (TP)	False negatives (FN)
Actual non-applicants	False positives (FP)	True negatives (TN)

Table 4: Confusion matrices

Method	Confusion matrices	
Linear models		
Logistic regression	4,566	4,660
	2,002	17,373
Stochastic gradient descent	7,731	1,495
	8,045	11,330
Naïve Bayes models		
Gaussian naïve Bayes	4,205	5,021
	551	18,824
Ensemble models		
Random forest	6,200	3,026
	1,553	17,822
Extremely randomized trees	3,839	5,387
	2,542	16,833
Gradient tree boosting	4,439	4,787
	96	19,279

Table 5: Results by Method

			Negative			
			Predictive	Recall		
Method	Accuracy	Precision	Value	(sensitivity)	Specificity	Kappa
Linear models						
Logistic regression	77%	70%	79%	49%	90%	42%
Stochastic gradient descent	67%	49%	88%	84%	58%	36%
Naïve Bayes models						
Gaussian naïve Bayes	81%	88%	79%	46%	97%	49%
Ensemble models						
Random forest	84%	80%	85%	67%	92%	62%
Extremely randomized trees	72%	60%	76%	42%	87%	31%
Gradient tree boosting	83%	98%	80%	48%	100%	55%

Notes

[1] We also included people that reported a new loan in the debt table from 2014 as having applied for a line of credit for agricultural purposes, 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

[2] Acres owned percent can be above 100 percent because the farm owns more acres than it operates, i.e., the farm rents out land

[3] 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 six models.

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

[5] 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

[6] 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.