Application of Machine Learning Approaches to Predict Membership in ARMS NOL Frame

Bayazid H. Sarkar, Peter Quan, Andrew Dau

Sampling and Frame Development Section
National Agricultural Statistics Service, USDA

FCSM Annual Conference, October 2024

Disclaimer

The findings and conclusions in this report are those of the authors and should not be construed to represent any official USDA or U.S. Government determination or policy.

Background

The Agriculture Resource Management Survey (ARMS) is administered annually by the United States Department of Agriculture's National Agricultural Statistics Services (NASS) to ascertain :

- U.S. farm and ranch production practices
- Resource use
- Economic information

Background Cont'd

The survey is administered in three phases:

- Phase 1 is a screener for Phases II and III
- Phase II collects production practices data
- Phase III collects farm, economic, and operator characteristics data

ARMS Phase III utilizes a dual frame design:

- List Frame
- Area Frame

Background Cont'd

For the Area Frame component, the June Area Survey (JAS) is used to identify operators that are not on the ARMS List Frame:

- These specific operations comprise the ARMS Area Frame Not on List (ARMS NOL) sampling frame.

Using the JAS to compile the ARMS NOL sampling frame is an important step to -

- Maintain sampling frame integrity
- Complement the list frame incompleteness
- Obtain accurate and complete farm population survey indicators

Five ML Methods

Machine Learning (ML) approaches will be used in this project to conduct screening level evaluations of the ARMS NOL sampling frame. We want to predict membership of a farm - whether it is ARMS NOL or not. Five Machine Learning supervised classification methods were explored for this project:

1. Random Forest
2. Gradient Boosting
3. Logistic Regression
4. Support Vector Machine
5. Neural Network

Dataset

1. Dataset for this initial phase of this project was obtained from 2023 Texas JAS segments.
2. Total record count is about 3,000.
3. Select administrative and geographical variables were removed.
4. Training and Testing data were split in approximately 60:40 ratio.
5. It is an imbalanced dataset in terms of response variable: (ARMS NOL vs. NOT ARMS NOL).
6. In the Training dataset positive ARMS NOL cases were oversampled to make it approximately balanced.

Random Forest

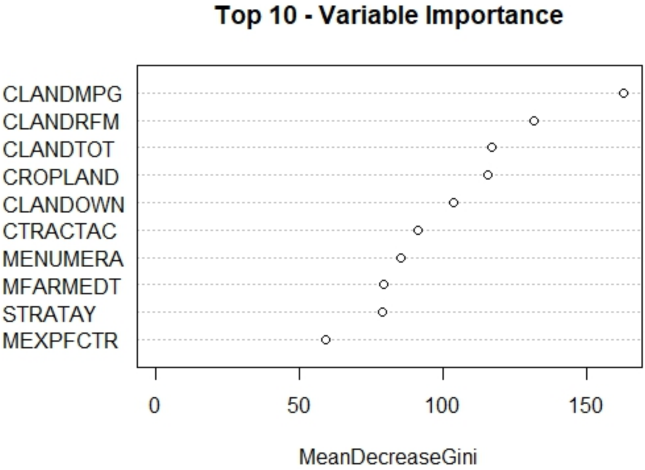
Random Forest (RF) is an ensemble learning technique. In a 2022 publication, Rezaei and Jabbari characterized-

1. Random forests as a group of decision trees working together on a specific prediction.
2. The outcome is determined by the predictions from the majority of these trees.

"caret" package in R was used to implement this method. Three repeated ten fold cross-validation were used.

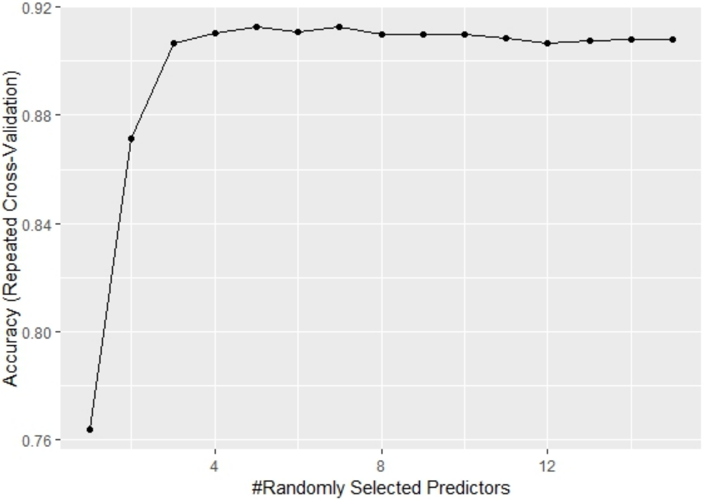
Random Forest

Important predictor variables



Random Forest

Repeated Cross Validation- tuning parameter is selected through Grid Search: metric accuracy



Random Forest

The Confusion Matrix and Statistics are based on Test Data:

		Actual	
		0	1
Predicted	0	647	106
	1	189	186

1. Accuracy = 0.7385
2. Kappa = 0.3761
3. Sensitivity = 0.6370
4. Specificity = 0.7739
5. Balanced Accuracy = 0.7055
6. Positive Class = 1

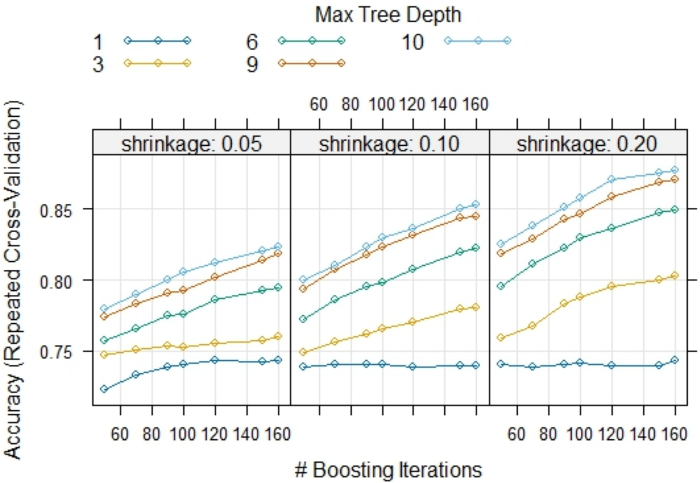
Gradient Boosting Method:GBM

In their 2021 paper, Belyadi and Haghighat explain that gradient boosting is an ensemble supervised machine learning technique that integrates several weak learners to form a final model-

1. This approach involves training the models sequentially, where more emphasis is placed on instances with incorrect predictions, which helps to progressively reduce the loss function.
2. The weak learners' predictions are evaluated against the actual outcomes, and the resulting difference indicates the model's error rate.
3. "caret" package in R was used to implement Stochastic Gradient Boosting method (GBM). Three repeated ten fold cross-validation were used.

Gradient Boosting Method

Repeated Cross Validation: metric accuracy



GBM

The Confusion Matrix and Statistics are based on Test Data

		Actual	
		0	1
Predicted	0	635	109
	1	201	183

1. Accuracy = 0.7252
2. Kappa = 0.3504
3. Sensitivity = 0.6267
4. Specificity = 0.7596
5. Balanced Accuracy = 0.6931
6. Positive Class = 1

Logistic Regression

The Logistic regression model is subset of a broad class of models known as generalized linear models (GLM).

1. Logistic regression models a relationship between predictor variables and a binary response variable.
2. Link function is Logit.
3. It is a supervised machine learning algorithm.

“caret” package in R was used to implement this method.

Ease of Interpretation

The model coeffcients of logistic regressions are easier to interpret using odds ratio.

1. The regression coefficient for land rented to others is -0.827 , we can say that an extra acre of land renting to others decreases the odds of being in ARMS NOL by a factor of 0.44.
2. The regression coefficient for owning land is 0.155, we can say that an extra acre of owning land increases the odds of being in ARMS NOL by a factor of 1.17.

Logistic Regression

The Confusion Matrix and Statistics are based on Test Data

		Actual	
		0	1
Predicted	0	384	39
	1	452	253

1. Accuracy = 0.5647
2. Kappa = 0.2231
3. Sensitivity = 0.8664
4. Specificity = 0.4593
5. Balanced Accuracy = 0.6629
6. Positive Class = 1

Support Vector Machine

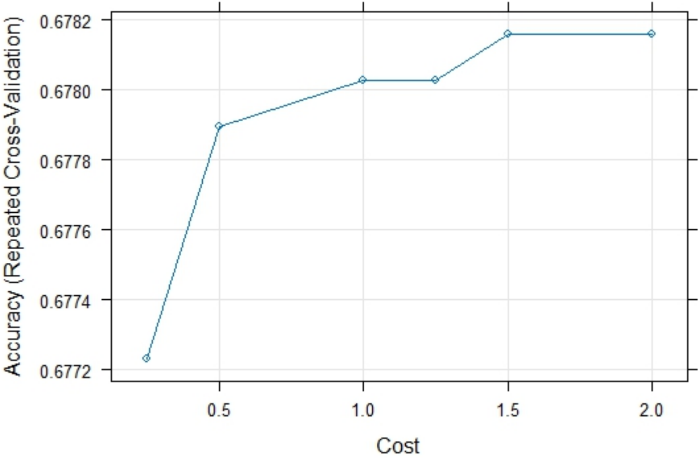
A Support Vector Machine (SVM) is a supervised learning algorithm employed for various classification and regression tasks, such as image processing, agriculture, text analytics, disease detection in medical applications, etc. According to MathWorks -

1. The goal of the SVM algorithm is to identify a hyperplane that most effectively separates data points of one class from those of another.
2. This"best" hyperplane is characterized by having the largest margin between the two classes.

"caret" package in R was used to implement this method. Three repeated ten fold cross-validation were used.

Support Vector Machine

Repeated Cross Validation: metric accuracy



SVM

The Confusion Matrix and Statistics are based on Test Data

		Actual	
		0	1
Predicted	0	429	57
	1	407	235

1. Accuracy = 0.5887
2. Kappa = 0.2287
3. Sensitivity = 0.8048
4. Specificity = 0.5132
5. Balanced Accuracy = 0.6590
6. Positive Class = 1

Neural Network-NNET

According to IBM - A Neural Network (NNET) is a machine learning program, or model, that makes decisions in a manner similar to the human brain.

Every neural network consists of

1. layers of nodes or artificial neurons—an input layer
2. one or more hidden layers
3. an output layer

“nnet” package in R was used to implement this method.

Neural Network-NNET

Architechture of NNET for ARMS NOL Data

1. 23 nodes in the input layer
2. 1 hidden layer and 10 nodes in the hidden layer
3. 1 node in the output layer
4. activation function is logistic

NNET

The Confusion Matrix and Statistics are based on Test Data

		Actual	
		0	1
Predicted	0	486	65
	1	350	227

1. Accuracy = 0.6321
2. Kappa = 0.2723
3. Sensitivity = 0.7774
4. Specificity = 0.5813
5. Balanced Accuracy = 0.6794
6. Positive Class = 1

Summary Metrics of the 5 ML Methods

ML Method	Sensitivity	Specificity	Accuracy
RF	0.6370	0.7739	0.7385
GBM	0.6267	0.7596	0.7252
Logistic	0.8664	0.4593	0.5647
SVM	0.8048	0.5132	0.5887
NNET	0.7774	0.5813	0.6321

The Performance of Five Methods

1. RF method achieved the highest accuracy and specificicity.
2. GBM method showed the second highest accuracy.
3. Logistic regression demonstrated the highest sensitivity; however, its specificity is the lowest.
4. SVM achieved the second highest sensitivity, but its specificity is the second lowest.
5. NNET showed moderate specificity and sensitivity.

Future Work

1. Expand the scope of this work to other states
2. Combine multiple years of data for states where NOL population count is small
3. Apply to resolve the cases when there is discrepency of name and address between multiple frames
4. Add additional JAS variables

References

1. Rezaei, Nima, and Parnian Jabbari. “Random forests in R” Immunoinformatics of Cancers: Practical Machine Learning Approaches Using R, Academic Press,2022, pp. 169-179.
2. Belyadi, Hoss, and Alireza Haghighat. “Supervised learning” Machine Learning Guide for Oil and Gas Using Python, Gulf Professional Publishing,2021, pp. 169-295.
3. Support Vector Machine(SVM), MathWorks, 9 Sep. 2024, https://www.mathworks.com/discovery/support-vector-machine.html.
4. What is a neural network, IBM, 9 Sep. 2024, https://www.ibm.com/topics/neural-networks.

THANK YOU!