Using Machine Learning Algorithms to Identify Farms on the 2022 Census of Agriculture

Gavin Corral[1], Luca Sartore[2] ,Katherine Vande Pol[3], Denise Abreu[4], Linda J Young[5]

[1, 2, 4,5] National Agricultural Statistics Service, 1400 Independence Ave SW 20250 Washington D.C., USA
[2] National Institute of Statistical Sciences, 1750 K St NW 20006 Washington D.C., USA
[3]Smithfield Foods Inc., 4134 US-117 Rose Hill, North Carolina 28458, USA

Introduction

- NASS conducts numerous surveys and the Census of Agriculture every 5 years. The surveys and Census inform the agricultural reports and estimates released by USDA each year.
- NASS maintains a list frame, which is sourced from multiple entities and continuously updated.
- Records on the list frame are classified as active, inactive, and/or criteria.
- Screening through the National Agricultural Classification Survey (NACS) helps determine farm status for criteria records.
- **AS34s are criteria records that have not responded after numerous NACS attempts between Census cycles. Unlike other criteria records, AS34s typically have no data on our list frame.**
  - **Response rates near 0**
  - **Limited data makes modeling response propensity challenging**
- NASS utilizes machine learning techniques such as supervised and unsupervised learning for crop yield prediction, land use classification, and survey methodology optimization, to enhance data collection efficiency and accuracy.
- In this study, we consider several machine learning approaches for estimating record-level propensity to respond. We discuss the results, challenges, and implications for future research.

Methods

- **Data**: 74,040 AS34 records from the 2017 Census, divided into training (70%) and testing (30%) sets with equal farm and non-farm proportions.
- **Variables**: Demographics, geospatial, and record-specific information, excluding agricultural variables due to sparse reporting and lack of availability for 2017/2022 AS34s.
- **Machine Learning Models:** Random forest (RF), logistic regression (LR), neural network (NN), and support vector machine (SVM) models to predict AS34 farm status using R and Python
- **Model Development**: LR conducted in R using glm(), RF developed in R using randomForestSRC, NN and SVM trained in Python using scikit-learn and TensorFlow packages.
- **Model Evaluation**: Evaluated models using accuracy, sensitivity, specificity, precision, and area under the ROC curve (AU-ROC) on validation data, with sensitivity deemed the most important outcome.

Results

- **Model Performance:** The LR model showed all variables as significant with classifications based on probability cutoffs. NN prioritized variable importance by neuron weights, effectively highlighting key variables such as year and source of the record.
- **Best Model:** The NN model was identified as the best model due to its high sensitivity (80.5%) in detecting true farms, despite its lower specificity (45.3%). This was considered a suitable trade-off given the project goals.
- **Comparison with Other Models**: SVM showed the best outcomes for three metrics (accuracy, sensitivity, and specificity) among models. RF and LR provided comparable results, but the precision of the RF model was notably lower than that of the LR model.
- **Variable Importance:** The most influential variables identified were the year the record was added to the list frame, source of the record, the impact of the operation on commodity estimates, state, and sex of the primary producer.
- **Validation Outcomes**: SVM had the highest specificity, making it effective at minimizing false positives**.** In contrast, the NN, while less specific, was more adept at correctly identifying actual farm operations, aligning with the prioritized sensitivity metric.

Discussion

- **Challenges in Data Protection and Access**: The decentralized nature of the U.S. Federal Statistical System introduces complexities in data protection and access, as laws vary across agencies. This situation complicates the use of administrative and non-survey data for statistical purposes, impacting efforts to reduce respondent burden.
- **Improvement and Future Application of Models**: Advances in AI and data science offer opportunities to enhance model accuracy and generalize their application over time. Adjusting time-related variables and expanding data sources, including administrative and web-scraped data, could significantly improve model precision and utility.
- **Addressing Bias and Ensuring Representation**: It is vital to evaluate and adjust models to prevent bias, ensuring accurate representation of minority and female producers. Monitoring misclassification rates is essential for equitable data treatment and maintaining the integrity of statistical outputs.
- **Potential Broad Implications**: The developed models could influence strategies both before and after data collection, such as optimizing sample selection and reducing operational costs. These strategies demonstrate the potential for broader applications beyond the immediate census, enhancing overall data collection and analysis processes.

Model Evaluation

Metric	RF	SVM	NN	LR
Accuracy[1]	67.26%	72.90%	53.80%	66.40%
Sensitivity[2]	66.75%	13.80%	80.50%	67.30%
Specificity[3]	67.42%	96.40%	45.30%	66.20%
Precision[4]	27.05%	54.30%	31.70%	38.60%
AUC[5]	72.63%	72.50%	72.10%	71.80%

[1]Accuracy = number of correct predictions / total number of records.
[2]Sensitivity = number of farms correctly identified / number of true farms.
[3]Specificty = number of non-farms correctly identified / number of non-farms.
[4]Precision = number of farms correctly identified / number of records predicted as farms.
[5]Area under the Receiver Operating Curve = measure of the ability of the model to distinguish between farms and non-farms; range 0.5 to 1.0 (higher values indicate a better fit model).

Structure of the neural network used for identifying farms with active status

Layer type	Output neurons	Processed information
Input Layer	18	
Dropout at 25%	18	1.
Dense with linear activation	32	2.
Dense with rectified linear units	32	2.
Multiply	1088	(3., 4.)
Dropout at 50%	1088	5.
Dense with rectified linear units	4	6.
Dense rectified linear units	4	3.
Multiply (i.e., concatenate with interactions)	24	(7., 8.)
Dropout at 5%	24	9.
Dense with sigmoidal activation (in output)	1	10.