In-Season US Corn Acreage Forecasting Using Machine Learning

Walter Ac-Pangan, Nathan P. Hendricks
Department of Agricultural Economics, Kansas State University.

KANSAS STATE UNIVERSITY

Introduction

- The USDA is the primary source of public information in agriculture, and stakeholders (e.g., traders, farmers, and market analysts) adjust their expectations accordingly to USDA reports since it has been generally regarded as accurate and trustworthy.
- The Monthly World Agricultural Supply and Demand Estimates (WASDE) report is widely used by stakeholders. Most of these reports rely on statistical survey methods to gather data on production and usage, with information released on established dates throughout the year.
- USDA acreage estimation before harvest is mainly released in two reports. First is the "Prospective Plantings Report," which is released in late March. Second, is the "Acreage Report," which is published in late June.
- Studies have demonstrated that official forecasting released by government agencies has a significant impact on the markets and decisions on supply chains (Adjemian, 2012; Irwin et al., 1994; Isengildina et al., 2006; Isengildina-Massa et al., 2021).
- **Our study aims to develop and evaluate machine learning (ML) models to deliver accurate and timely updates of crop acreage forecasts in May using data on market and planting conditions.**
- Can publicly available data on planting conditions and markets up to May provide additional information to what was released in the Prospective Plantings Reports?

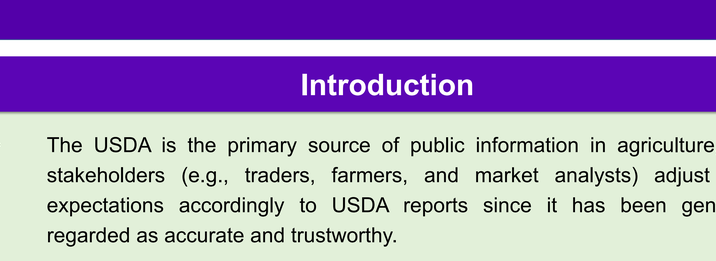

Figure 1: Crop Calendar and Report Released Timeline.

Data

- To estimate our model, we use panel data comprising 92 variables aggregated at the state level on an annual basis.
- All variables are grouped into broader categories: Acreage, Production, Yield, Fertilizer Prices, Futures Prices, Suitable Days for Field Work, Temperature, Precipitation, Soil Moisture Exposure, and Technological Innovation

Predictors:

- All the variables included aim to incorporate into the model factors the influence of crop planting decisions and enhance the model's ability to forecast acreage allocation.

Methods

- Models: OLS, Random Forest and Extreme Gradient Boosting

$$Change\_in\_Planting\_Intentions_{i,t} = \left(Weather_{(i,t)}, Prices_{(i,t)}, PriorProduction_{(i,t)}, Location_{(i,t)}, Trend_{(i,t)}\right)$$

$$Predicted\_Acres_{i,t} = PlantingIntentions_{i,t} - Change\_in\_Planting\_Intentions_{i,t}$$

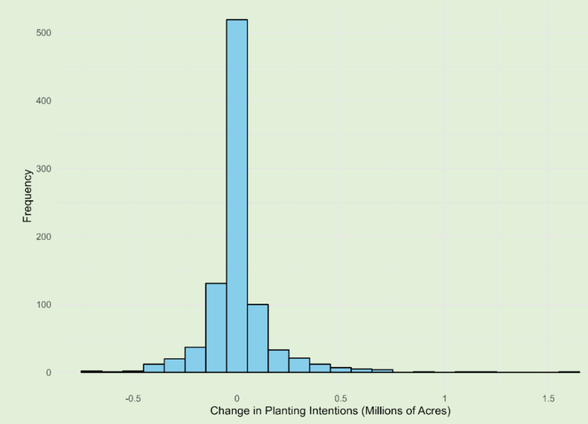

Figure 2: Change in Planting Intentions.

$$Acres_{i,t} = f\left(WeatherV_{(i,t)}, PricesV_{(i,t)}, PriorProductionV_{(i,t)}, LocationV_{(i,t)}, TrendV_{(i,t)}\right)$$

Out-of-Sample Validation Approach

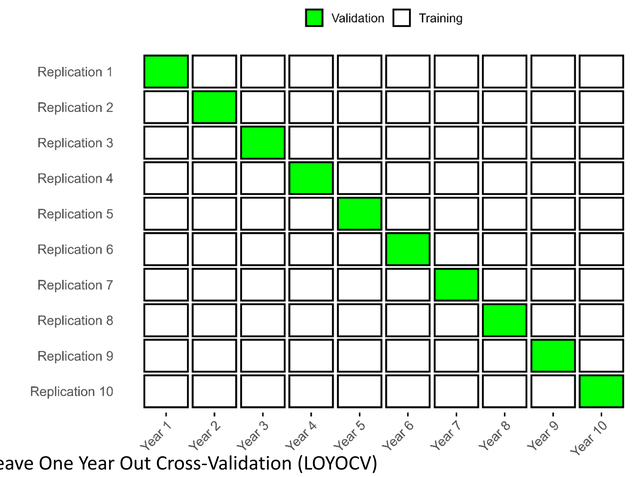

Figure 3. Leave One Year Out Cross-Validation (LOYOCV)

Results: Accuracy Measures

The most accurate Random Forest (RF) model achieves an average MAE of 33,440 acres across years and states when forecasting actual acreage planted. Compared to the USDA's June estimates, which are survey-based and have an MAE of 59,379 acres, the RF model reduces the forecasting error by 43%.

Table 2.1 Accuracy of Acreage Predictions when Forecasting Change in Planting Intentions

Rank	Model	RMSE	MAE	MAPE
1	RF	62,498	33,440	2.60
2	XGBoost	62,824	41,988	4.93
3	Survey June	163,310	59,379	4.32
4	Survey March	168,429	88,744	6.16
5	FEOLS	192,987	114,702	11.86

Variable Importance

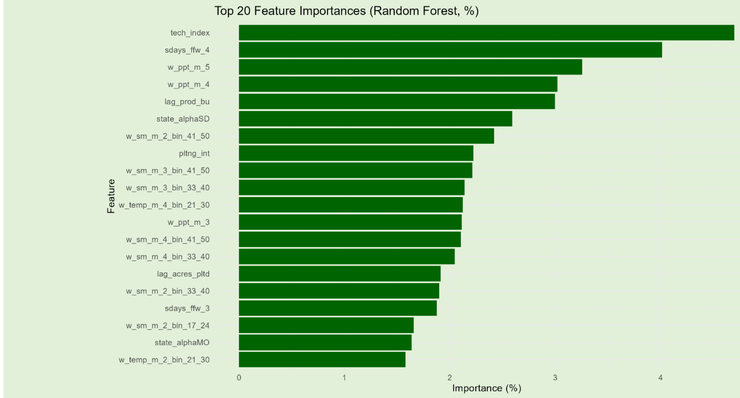

Figure 3. Variable Importance of Random Forest Model when forecasting Change in Planting Intentions

Conclusion

- The study provide 3 key findings

1. Using secondary data sources and incorporating the Planting Intentions Report information from USDA, we can generate acreage forecast that can outperform the accuracy level of the survey results released by USDA's Acreage Report, which is published on late June.

2. The Random Forest model shows that the most important variables for forecasting change in planting intentions are the Technology Index, Suitable Days for Fieldwork in April, Precipitation in May, and April.

3. Third, when assessing the value of the planting intentions survey, the results demonstrate the significant predictive value added by incorporating the USDA's Prospective Plantings Report into models that forecast acreage planted. Across models, the accuracy measures suggest that a model that does not incorporate the planting intentions information can decrease their accuracy from 30 to 48%.