Optimizing Crop Portfolios with Machine Learning

My Role
Quantitative Researcher

Project Brief
I completed an independent capstone project where I leveraged different machine learning models to identify the optimal crop portfolio of soybeans. I analyzed over 34,000 rows of data to identify the top five highest-yielding crops under different weather, temperature and environmental conditions.

Problem Statement

How might we help farmers maximize their crop yields given limited farmland?

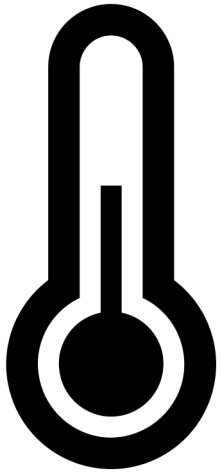
Understanding the Problem Space

Food supply optimization is essential due to growing global populations, rising temperatures, and diminishing farmland availability.

There is a global call to increase productivity and sustainable food production in the agriculture industry due to growing populations and alarming hunger rates around the globe.

As of 2021, the global population is quickly approaching **8 billion people**, and it is projected to increase to **9.7 billion people** in the next 20 years.[1]

The urgency of this problem is exacerbated by the alarming statistics surrounding global hunger, where close to **8.9%** of the global population are underfed.[2]



1. United Nations. (n.d.-a). *Global Issues: Population*. https://www.un.org/en/global- issues/population.

2. United Nations. (n.d.-b). *Sustainable Development Goals: Goal 2: Zero Hunger.* https://www.un.org/sustainabledevelopment/hunger/.

Project Goal

Leverage data analytics and machine learning tools to identify high-performing crop varieties and efficiently allocate farmland at our target farm while balancing risk.

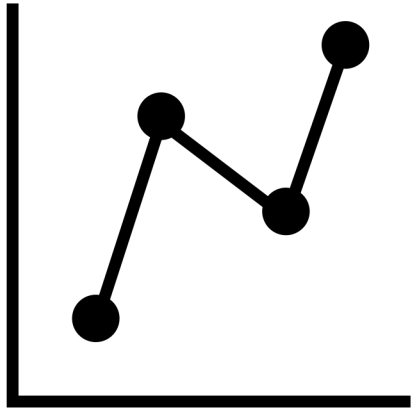
Literature Review

Growing climate change negatively impacts farming outcomes, necessitating an increase in strategic, data-driven interventions.

Without any remedial interventions, growing temperatures cause heat stress, shorten the growing season, and decrease total crop yields.[2]

Researchers at the University of Florence underscore the need to leverage analytics to react to climate change and plan against land shortage and other food security risks.

However, current crop simulation tools are insufficient as they often do not account for abiotic and biotic stressors — such as extreme climate events and pests.[3]



2. Gobin, A.. (2010). Modelling climate impacts on crop yields in Belgium. Climate Research, 44(1), 55-68.

3. Bindi, M., Palosuo, T., Trnka, M., & Semenov, M. (2015). Modelling climate change impacts for food security. Climate Research, 65, 3-5.

Literature Review

Increases in data collection combined with technological advances allow farms to adjust their strategies to account for climate variability.

Climate forecast data has become more widely available, and farms now have access to precision farming technologies that allow them to leverage field data, yield maps, and automated guidance systems.[4]

A study conducted Liu et al. in 2020 used a multivariate regression model to point out that while the sole impact of climate change negatively affects crop yields, data-driven crop management has a greater impact on wheat yield than climate change.[5]

4. Langemeier, M., & Shockley, J. (2019). Impact of Emerging Technology on Farm Management. Choices, 34(2), 1-6.

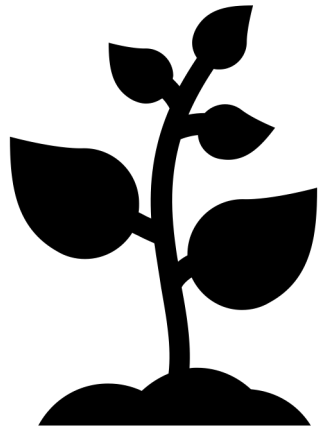
5. Liu Y., Zhang J., Ge Q. (2020). The optimization of wheat yield through adaptive crop management in a changing climate: evidence from China. Journal of the Science of Food and Agriculture. https://doi.org/10.1002/jsfa.10993.

Our Solution

Incorporate climate data in our machine learning model to predict optimal crop yields under different weather conditions.

By using readily available weather data that captures rising trends in global temperatures, we can now account for climate variability when creating our predictive model.

This would identify the highest performing crops that not only make the most use out of limited farmland, but also thrive under unstable weather conditions.



Descriptive Analytics

We should aim to produce 60 bushels per acre at our target farm.

The data set contained **118 unique locations** and **182 unique soybean varieties**.

Using R programming, I plotted the distribution of variety yields in a histogram, which reveals that the yields are normally distributed with a mean slightly **above 60 bushels per acre**.

Based on this, I concluded that 60 bushels per acre is an appropriate goal for the optimal portfolio at our target farm.

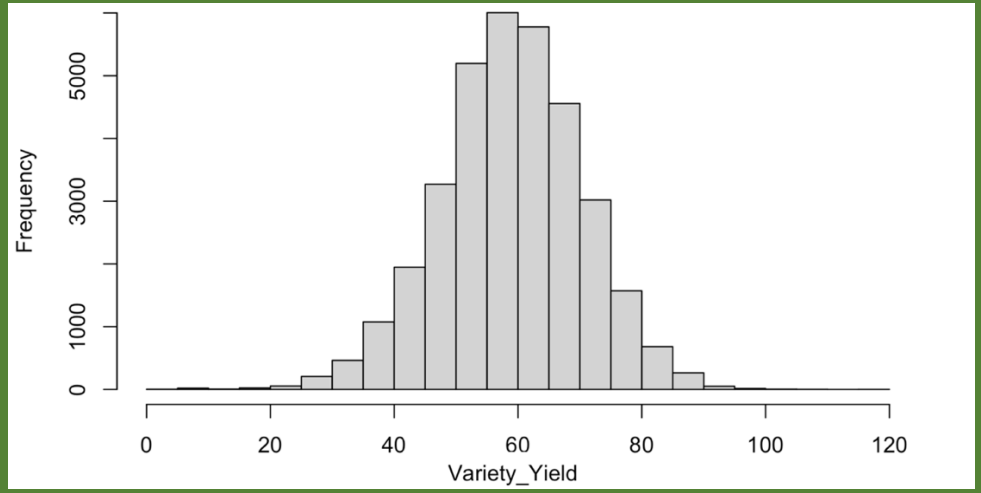


Figure 1: Frequency Distribution of all Soybean Yields in the Dataset

Descriptive Analytics

Knowing the precise location of our target farm within our dataset will help us make more accurate predictions.

Of the **34,213 farm observations** in the dataset, the majority were **concentrated in the Midwest**, and our test observation — identified by the green triangle — is located in the western part of the group.

Using K-Clusters, the observations were divided into **20 different groups based on their locations** — our test observations was categorized with **the fourth cluster**.

Grouping the data this way will help me account for weather variability at our test observation farm using data from nearby farms when I execute predictive analytics.

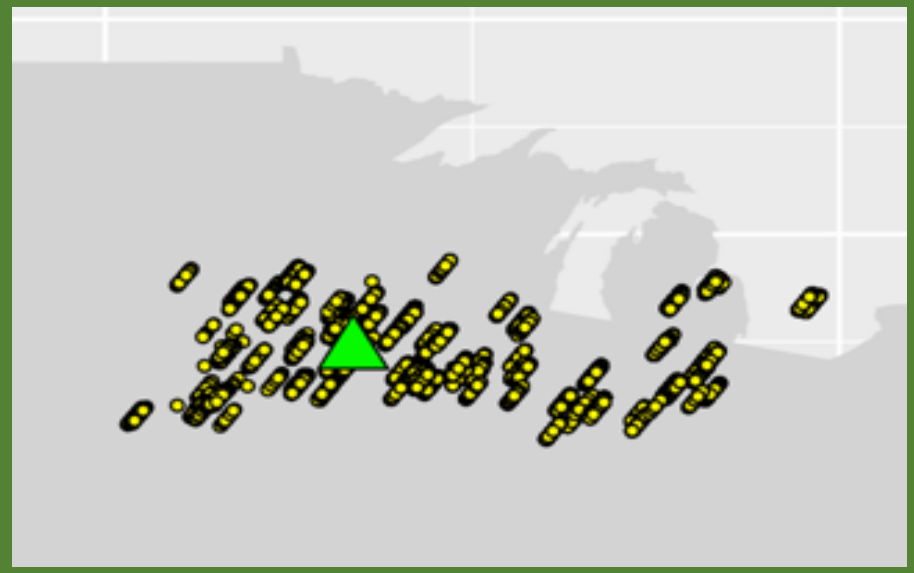


Figure 1: Farm Locations in the U.S.

Predictive Analytics

Of the 7 tested machine learning models, the Random Forest algorithm yielded the most accurate predictions.

Method	Regression Tree	Bagging	Random Forest	Boosted	Neural Network
RSME(bushels/acre)	9.6799	8.4966	8.1496	9.8637	58.4453

While I did run models using Linear Regression and LASSO, their Root Mean Square Error (RSME) values were so large that I did not consider them for my final model.

Predictive Analytics

Running the Random Forest Model on our target data set reveals the average yield of the different soybeans at our farm.

Within the test data set I isolated those observations in the 4th cluster that are geographically closest to the target farm. This new dataset had **2,329 observations**.

I then standardized the location and soil conditions to those of the target farm while keeping the weather conditions varied, which will allow the model to account for location-specific variations in weather when making predictions on soybean yields.

After running the model, I was able to predict the mean yield for each variety under varied weather conditions.

Prescriptive Analytics

Using a mean-risk ratio heuristic, I identified the top five soybeans that have the highest average yields and the lowest risk.

Variety	V32	V95	V181	V97	V48
Mean-Risk Ratio	328.38	245.02	229.56	224.20	216.60

While I want to select soybean varieties that produce a high yield, I also want to pick those that are low in risk. Picking those varieties with the largest yield/risk ratio which would ensure a higher yield and lower risk when growing those soybean varieties.

Prescriptive Analytics

With non-linear optimization, I found the optimal percentage of farmland to allocate to each of our five soybean varieties.

Variety	V32	V95	V181	V97	V48
Farmland Allocation	21%	25%	17%	22%	15%

Farmers are concerned about the impact weather volatility can have on their produce yields. Therefore, the objective of our optimization model was to minimize the risk associated with each of the selected soybean varieties, while adding a constraint that the expected yield must exceed 60 bushels/acre.

Using this soybean portfolio at the target farm would yield approximately 60.36 bushels per acre, which sufficiently meets the threshold we established at the start of this report.