Machine Learning Approaches to soybean Yield Estimation and Seed Selection Using Historical Climate Data: A Case of Iowa State

Stephen Li; 489522

Abstract

Crop yield has always been the focal topic of researchers. Accurately predict the crop yield is curial to a table crop yield. This paper uses 4 machine learning tools, Regression Tree (RT), Random Forest (RF), Boosted Tree (BT), and Neural Network (NN), and a classical statistical approach, Multiple Linear Regression, to predict the soybean yield at the target farm under different weather situations. Compared 5 models, Random Forest (RF) is finally selected to make predictions because it returns a low error rate and requires less computational resources. This paper also uses an optimization tool to select proper varieties to minimize the risk, like portfolio analysis measured by the standard deviation of portfolio, while controlling the yield. The result is a weighted combination of 5 varieties. This paper demonstrates an applicable way of selecting seeds based on historical data by using machine learning methods.

Keywords: Machine Learning, Crop Yield, Optimization

1. Introduction & Literature Review

Agricultural issues have always been a focal issue of global concern because this is the most basic issue of people's livelihood. The United States is a large agricultural country, but it faces many problems and challenges. Among them, the issue of uncertain crop yields is very serious because after the industrialization, crops are grown at more concentrated places than in hunting and gathering society. Therefore, it is crucial to lower down the risk or variation of crops' yield reducing the effect of natural uncertainty. For example: On Aug. 17, 2020, the Iowa Soybean Association reported that the total area affected by the Derecho storm in Iowa Included 5.64 million acres of soybeans and 8.18 million acres of corn.

Some crop production statistics in many countries are calculated by ground-based observations and production reports. These approaches are costly, time-consuming, and inefficient. In addition, these data are collected too late for decision-making. (Reynolds et al., 2000) Therefore, continuous monitoring of crop conditions and early yield estimation and prediction is very important.

Remote sensing (RS) has been extensively adopted to predict/estimate crop yield at regional scales as well as individual fields. For that, the result of different planting times by peasants, the ripening and harvesting dates in different fields are not the same. These crops are characterized by different NDVI temporal profiles. Moreover, the time for each soybean to reach maximum yield and NDVI peak is not similar. (Bolton & Friedl, 2013) These conditions suggest that the second type of crop prediction methods for classic crops have some difficulties.

To solve the above limitations of the RS, simple regression model was built. Furthermore, there are many examples of different components of the CI algorithm in the agricultural RS literature, such as artificial neural networks, fuzzy-set theory, and genetic algorithms to estimate the yield of various crops. There are also some relatively new machine learning techniques, namely Gaussian Process Regression Support Vector Regression, Enhanced Regression Tree, and Random Forest Regression. Next, let's take a closer look at machine learning technology.

Machine learning is an effective empirical method of classification and prediction. Saha et al. (2021) show the promise for machine learning to significantly improve field-level flux predictions. Results show that regression-based ML models such as Random Forest can substantially improve temporal N2O flux predictions. Kuwata and Ishizaki (2015) used the deep learning method to estimate the corn yield in Illinois and obtained a high accuracy. Similarly, Barkley et al. (2014) used a regression model to predict wheat yield and epidemic diseases and weather as predictive factors.

Aghighi et al. (2018) introduced advanced ML techniques and compared their performance with some proposed conventional regression methods. That paper demonstrated that some advanced ML approaches can predict the silage maize yield and they are less sensitive to the inconsistency of NDVI time series

In this paper, I use historical data of the mid-west US to help the farmer to select seeds from various varieties at a target farm located in the state of Iowa by machine learning approaches. My goal is to select varieties with good yield and low risk (standard deviation) improving certainty.

2. Methodology and Analysis

2.1. Descriptive analysis

2.1.1. Data Description

Training data contains Three parts, attributes of experimental, attributes of varieties, and attributes of the sites. Attributes of experimental consist of experimental information, such as growing year, breeding group, and coordinates. Data were released by Syngenta R&D, which also provides variety data. Attributes of varieties including yield. Attributes of the sites consist of climate information, including weather and soil property. Data was collected or recorded by the European Centre of Medium-Range Weather Forecasts, CONUS, and ISRIC.

2.1.2. Preprocessing

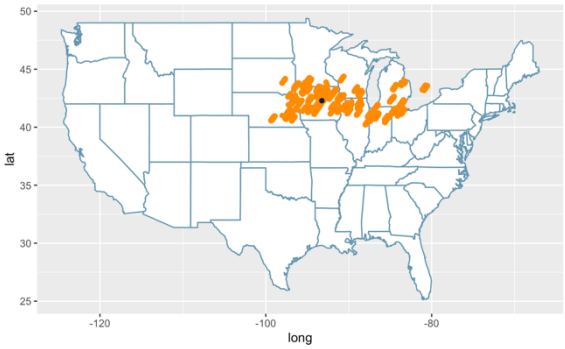
Among 58 columns in the original dataset, only 26 are kept predicting the yield. Because not all the experiments contain weather data of the previous year, we kept only the weather data of the current year when the experiment was done. Therefore,

we preprocessed to keep the current year's temperature, radiation, and precipitation. Also median of temperature, radiation, and precipitation are taking into prediction.

There are 182 different seed varieties in total. Taking 100 rows as the boundary, varieties containing less than 100 rows are considered insufficient variety, and more than 100 rows are considered sufficient variety. In this dataset, 132 varieties are insufficient, and 50 varieties are sufficient. In the predictive analysis, each variety with sufficient data will be put into the model separately, while insufficient data will be put into model together and treat variety as a category variable to reduce the bias and improve the accuracy. Therefore, a total of 51 models will be constructed.

2.1.3. Distribution of Experimental Locations

I plot the longitude and latitude data into maps to see where our experiments were conducted and where is our target farm located. The dark orange in figure 1 represents the experimental (predictive data) locations, while the dark point represents the target farm. Experimental farms spread out across 8 different states but all in the mid-west of the USA and the target farm is in the state of Iowa.



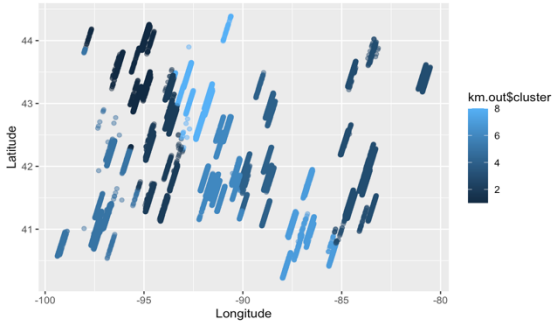


Figure1: Experimental locations distribution; Figure 2: Cluster result based on longitude and latitude data

Because It is important to make a prediction of yield with season weather data, while it is not feasible to accurately knowing the coming weather. I make a prediction on possible scenarios of weather on target farms and do optimization for seed selection. To do so, I do a clustering to cluster target location into groups and the historical weather scenarios of the cluster that the target farm belongs to will be considered as the possible weather conditions. K-means method is used to do

clustering. Figure 2 shows the result of clustering. A total of 8 clusters are decided because experimental locations are scattered at 8 different states. In figure 2, blue from light to dark represents cluster group from 1 to 8 respective.

The target farm is assigned to cluster 2 in which 4593 different weather scenarios (temperature, precipitation, and radiation) happened. Random 1000 historical weather conditions are chosen to be possible scenarios of the target farm. In this way, we narrow down the possible weather conditions from 34023 to 1000.

2.1.4. Correlation coefficients of the variables

Figure 3 shows the correlation coefficients of variables in dataset. CE and CEC are highly correlated. Density and Acres are highly correlated. Therefore, I will keep only one of two in each correlation while performing linear regression model.

2.2. Predictive analysis

I chose 5 prevailing predictive tools to predict the yield—Multiple Linear Regression, Regression Trees, Random Forest, Boosted Trees, and Neural Network.

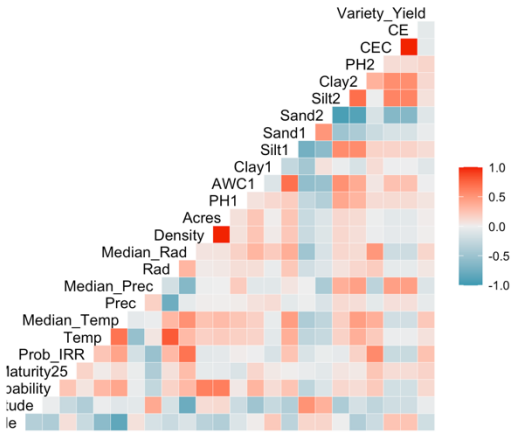


Figure 3: Correlation coefficients of variables in dataset

Among them, 50 varieties containing sufficient data will train each machine learning tool to see which performs best and apply it to make final optimization.

2.2.1. Validation

Validation set will be used to evaluate the accuracy of models. 80 percent of the data will be split out to train model and the rest 20 percent will be used to evaluate. In some cases, if there are parameters that should be decided while training, cross-validation method will be used. With regarding the measurement of accuracy or

mistake rate. I choose 4 measurements to have a comparison from multiple dimensions. I calculate Mean Absolute Error (MAE), Mean Square Error (MSE), Normal Mean Square Error (NMSE), and Root Mean Square Error (RMSE) to evaluate the accuracy of models.

2.2.2. Linear Regression

Linear Regression is a widely used traditional statistical method of prediction. Hight interpretability is the key advantage of the Linear Regression model. In this case, different from the validation set method I mentioned before, I put all data into training set, and take random 20 percent of data into the validation set. This is done because if I take 80 percent of data as the train set, there will be new categories appear on factor variable (Soil type) and it is not able to predict. Results see Table 1.

2.2.3. Regression Tree

Tree model is a commonly used machine learning model in real word and known by its good interpretability. Trees are more practical for users, and easier to build. Although It is not my priority to interpret the model, to reduce the effect of overfitting I test xerror corresponding to different complexity parameters and select the cp value with smallest xerror. Result see table 1.

2.2.4. Random Forest

Random Forest is another tree-structured machine learning method and can be considered as an improved version of the decision tree. It uses bootstrapping to sample the training data with replacement and make trees. Then make prediction by averaging the result or voting for classification problem. RF is known for it can handle huge input variables and missing values. In this case, although these advantages are not helping RF success in Soybean dataset because dataset is clear, and all the variables are helpful variables, it still outperforms the regression tree model. Two parameters should be decided in RF—number of trees and number of variables for splitting. I set the number of trees at default, 500, and the number of variables for splitting at one-third of predictive variables, 8. Result are showed in table 1.

2.2.5. Boosted Tree

Boosted Tree is another Tree-structured approach. BT was created to reduce the possible overfitting problem of trees. While Random Forest uses bagging to combine several trees, Boosted Tree uses boosting method to combine trees. BT build many trees and each of them focuses on the training error of previous trees to improve the predictability. Its fitting process is slower than RF, but it has high accuracy. In pruning the model, I use 10-fold cross validation to test the parameters combination. Due to the limitation of computational power and different variety has similar data structure, I randomly choose one variety to do cross-validation. The result is that 50 trees, 3 interaction depth, and 0.2 shrinkage returns the best accuracy. This combination is used to fit all the varieties.

2.2.6. Neural Network

Neural Network simulate the neural activities of the human brain. Neural Network has two parameters to decide, number of layers and number of nodes in each layer. Therefore, Neural Network contains infinite structure, and it is hard to find the best combination. It is also a problem that Neural Network overfits the data. In this case, after some trials, I choose to use a sample structure of 3 layers and 8 nodes in each layer, which performs the best result.

2.2.7. Result

After calculating the average MSE, MAE, NMSE, and RMSE of all varieties for each model, Multiple Linear Regression performs best among them. Table 1 shows the performance of each model. However, considering all dataset is put into training, the real performance of linear regression must be worse than what we get now.

Table 1: performance of 5 machine learning models

	MLR	RT	RF	BT	NN
MAE	5.78	6.48	6.01	5.82	6.27
MSE	58.23	73.88	63.2	58.20	81.54
NMSE	0.51	0.68	0.59	0.56	0.78
RMSE	7.53	8.46	7.88	7.55	8.75

Apart from MLR, the tree-structured model slightly outperforms the Neural network. Among three tree model, boosted tree return the lowest error rate and RF predicts slightly better than RT. Because boosted tree requires more computational power than random forest, while its improvement is limited, I choose random forest and boosted tree as the final models that will be used to predict yield in each scenario. Two models are selected because random forest can only input category variables with less than 53 factors.

Table2:example of weather-yield matrix

	V121	V127	V117	…	V133
1	54.36344	48.35987	56.72775	…	49.49500
4	55.81224	48.83761	57.54701	…	49.46519
…	…	…	…	…	…
1000	59.50376	50.48967	59.58408	…	50.79849

1000 weather scenarios are taken into prediction with each variety to have a yield matrix with variety type as x-axis and weather scenarios as y-axis. I perform one random forest model for 50 varieties with sufficient data, and one boosted tree for the rest varieties. In total, 51 models are built. A boosted tree is used for insufficient varieties because it contains 132 different varieties exceeding the limit of 53.

A sample of the weather-yield matrix is showed at table 2. Predicted possible maximum soybean yield is 68 and the possible minimum yield is around variety 154. With predicted soybean yield under different weather conditions, I will choose 5 soybean varieties with highest estimated yield to make optimization for farmers.

2.3. Prescriptive analysis

2.3.1. Method

I choose 5 varieties with the highest mean yield to construct a nonlinear optimization problem. Set each variety a weight, and the sum of them should be 1. The optimization function should minimize the risk while keeping a minimal yield of the mean of 5 varieties average yield. Like investment portfolio analysis, risk is measured

as the standard deviation of the portfolio. This is done in Python, using a package called scipy.

$$Portfolio\ Risk = \sum_{i=1:5, j=1:5} w_i * w_j * cor_{ij}$$

$$expected\ yield = \sum_{i=1:5} w_i * E_i$$

2.3.2. Result

5 varieties with the highest mean predicted yield under 1000 different weather conditions are V31, V181, V82, V98, V90. They have mean yield of 64, 62, 62, 61, 61 respectively. Thus, set the minimal boundary of yield as the mean of them, which is 62.3. The result is showed in table 3. 20% of the area should grow V31, 34% of the area should grow V181, and 46% of the area should grow V98, While V82 and V90 should be given up for this situation.

Table 3: optimal combination of 5 varieties

	V31	**V181**	**V82**	**V98**	**V90**
weight	0.200419	0.340299	0.0	0.459282	**0.0**

3. Conclusion

This paper uses historical soybean experimental data to predict the yield at the target farm which is located in Iowa with various seeds by 4 Machine learning models, RF, RT, NN, BT, and a classical statistic model, MLR. Comparing the results of each model, BT win out among them, with the lowest error measurement of MSE, MAE, NMSE, RMSE. Because while BT has a slightly lower error rate than RF, while it requires way more computational resources than RF, I choose RF as the predictive model. The RMSE of RF model is around 7.88. I predict each varieties' yield under 1000 different weather scenarios, V31, V181, V82, V98, V90 return the highest average yield. I Constructed a nonlinear optimization model and put the predicted data of 5 varieties to make a combination with lowest standard deviation. The result shows 20% of the area should grow V31, 34% of the area should grow V181, and 46% of the area should grow V98, While V82 and V90 should be given up for this situation.

Reference:

Aghighi, H., Azadbakht, M., Ashourloo, D., Shahrabi, H. S., & Radiom, S(2018). Machine Learning Regression Techniques for the Silage Maize Yield Prediction Using Time-Series Images of Landsat 8 OLI. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *11*(12), 4563–4577. https://doi.org/10.1109/jstars.2018.2823361

Bolton, D. K., & Friedl, M. A. (2013). Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics. *Agricultural and Forest Meteorology*, *173*, 74–84. https://doi.org/10.1016/j.agrformet.2013.01.007

Barkley, A., Tack, J., Nalley, L. L., Bergtold, J., Bowden, R., & Fritz, A. (2014). Weather, Disease, and Wheat Breeding Effects on Kansas Wheat Varietal Yields, 1985 to 2011. *Agronomy Journal*, *106*(1), 227–235. https://doi.org/10.2134/agronj2013.0388

Kuwata, K., & Shibasaki, R. (2015). Estimating crop yields with deep learning and remotely sensed data. *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. https://doi.org/10.1109/igarss.2015.7325900

Reynolds, C. A., Yitayew, M., Slack, D. C., Hutchinson, C. F., Huete, A., & Petersen, M. S. (2000). Estimating crop yields and production by integrating the FAO Crop Specific Water Balance model with real-time satellite data and ground-based ancillary data. *International Journal of Remote Sensing*, *21*(18), 3487–3508. https://doi.org/10.1080/014311600750037516

Saha, D., Basso, B., & Robertson, G. P. (2021). Machine learning improves predictions of agricultural nitrous oxide (N2O) emissions from intensively managed cropping systems. *Environmental Research Letters*, *16*(2), 024004. https://doi.org/10.1088/1748-9326/abd2f3