Soybean Variety Selection in Machine Learning Analysis

Agriculture is one of the most important areas or industries that are relevant to everyone in the world. The increasing population requires more and more food each year, and new techniques to study, analyze, optimize, and predict yield are common in agriculture. Moreover, seed variety selection is the key element of agriculture; it is one of the most prioritized farming sciences. With the help of statistical knowledge and methods, it is more convenient to make better seed selection decisions. In this report, I use the knowledge and technique in statistical data analysis and machine learning to analyze the data set of soybeans, which includes the data contained information of location, latitude, longitude, variety, yield, temperature, soil, weather, radiation, precipitation, and other variables. The data is from the real world, so the data cleaning and pre-processing procedure are necessary to eliminate missing or meaningless data, and choose the variables really matters. I examine the relationships among variables, make several plots, use a machine learning model to make predictions, and finally use Heuristics to select soybean varieties. This project has seven machine models: Linear Regression, LASSO, Regression Tree, Bagging, Random Forest, Boosted Trees, and Neural Network. I presented each model's output, discussed the results' meaning, and compared the models to find the optimal model to make a selection. In modeling and optimization, countless details can be improved or completed, and I tried several methods to promote the results. The Bagging trees model has the least mean squared error rate, so I used it to predict and select the seed varieties this way. Machine learning methods and portfolio analysis are the fundamental skills applicable to this project. It should be noted that the seed varieties selection project still has more space to develop and need more research.

Keywords: Soybean; Machine learning; Prediction; Optimization

1. Introduction

Farming is a fundamental area that feeds everyone in the world. In the modern world, it is crucial to select varieties to have a high yield and few risks, and we have advanced techniques to make the selection. This report uses the soybean data set to construct several machine learning models and then choose the promising varieties.

2. Literature Review

The studies on agriculture and farming using machine learning or other statistical technique are widely seen. Brown and Bergh (2020) reviewed the data-driven and decision-making methods used in agriculture and analyzed them. Specifically, there is also important research on soybean variety selection. Sanchez A, Frausto (2014) and Bustamante (2014) used machine learning methods in crop yield prediction, similar to this project's subject. Nalley (2009) and Barkley (2009) examined the portfolio analysis methods on vice variety to maximize profit, which is helpful in our third part of the project. Barkley and Peterson (2008) selected wheat variety using prediction technique and portfolio optimization, giving me some insights.

3. Methodology and Analysis

The data analysis is divided into three parts: Descriptive Analytics, Predictive Analytics, and Prescriptive Analytics.

3.1 Descriptive Analytics

```
> names(mydata)
 [1] "GrowingSeason"      "Location"           "Genetics"          "Experiment"         "Latitude"
 [6] "Longitude"          "Variety"            "Variety_Yield"     "Commercial_Yield"   "Yield_Difference"
[11] "Location_Yield"     "RelativeMaturity"   "Weather1"          "Weather2"           "Probability"
[16] "RelativeMaturity25" "Prob_IRR"           "Soil_Type"         "Temp_03"            "Temp_04"
[21] "Temp_05"            "Temp_06"            "Temp_07"           "Temp_08"            "Temp_09"
[26] "Median_Temp"        "Prec_03"            "Prec_04"           "Prec_05"            "Prec_06"
[31] "Prec_07"            "Prec_08"            "Prec_09"           "Median_Prec"        "Rad_03"
[36] "Rad_04"             "Rad_05"             "Rad_06"            "Rad_07"             "Rad_08"
[41] "Rad_09"             "Median_Rad"         "Density"           "Acres"              "PH1"
[46] "AWC1"               "Clay1"              "Silt1"             "Sand1"              "Sand2"
[51] "Silt2"              "Clay2"              "PH2"               "CEC"                "CE"
```

Fig 1: The Variables in the Data

Through Map, I plotted the latitudes and longitudes on a map to visualize each farm's location. According to Figure 1, we can see that most of the farms are located in the Midwest of the US, and our target farm has the largest distribution in the center of Iowa.

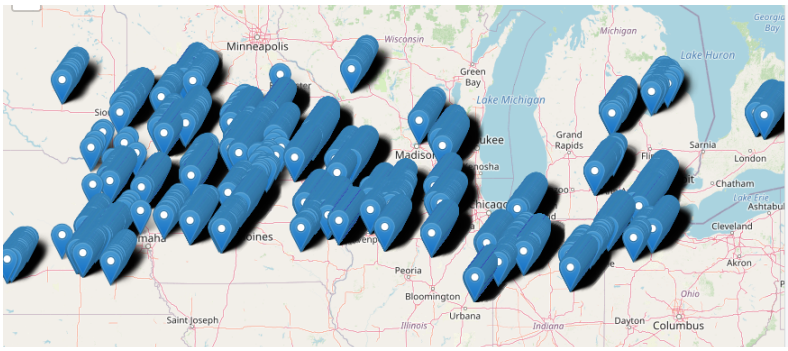

Fig 2: Plot of the Positions

Below I generated the frequency distribution for varieties; it can be seen that we don't have enough data for every variety to build a dedicated prediction. Fortunately, dozens of varieties have enough data to do research.

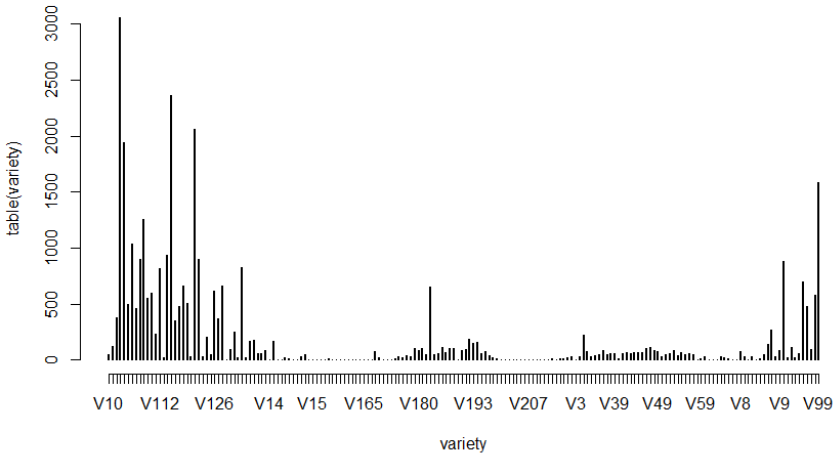

Fig 3: The Count of Varieties and Distribution Plot

Based on the conclusion above, it's not uncommon to guess the relationship between locations and varieties. Shown as the heatmap of variable frequency, it's clear that only some of the varieties are uniformly distributed in the locations. And some are grown relatively broadly in many locations, with the rest grown only in very few regions.

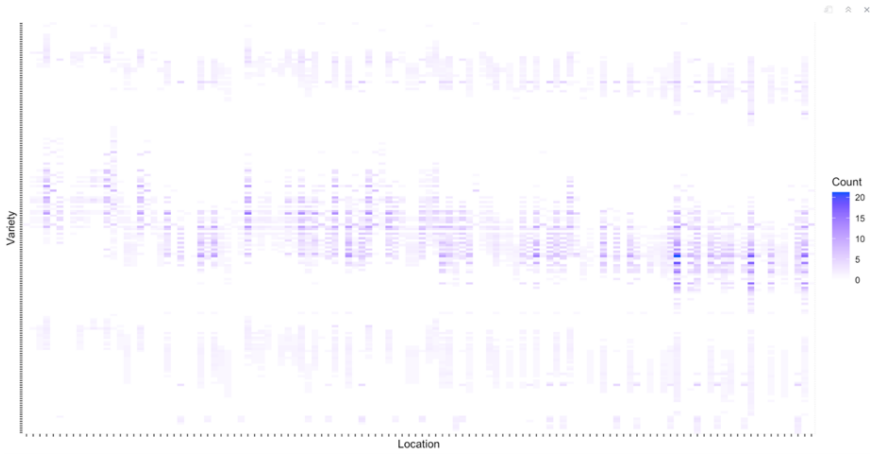

Fig 4: Heatmap of Variety in Different Regions

Then I looked for patterns in weather variables and explored relationships between locations and weather-related variables. The Weather1 variable is about Climate type based on temperature, precipitation, and solar radiation, and the weather2 variable is the season type. I found that the correlation between location and weather1 is 0.29, and the correlation between location and weather2 is 0.34. Below is the plot of the

distribution of the yield variables. We can know one realistic goal for the optimal portfolio at the target farm is to find portfolios that have yields higher than most varieties and have less risk.

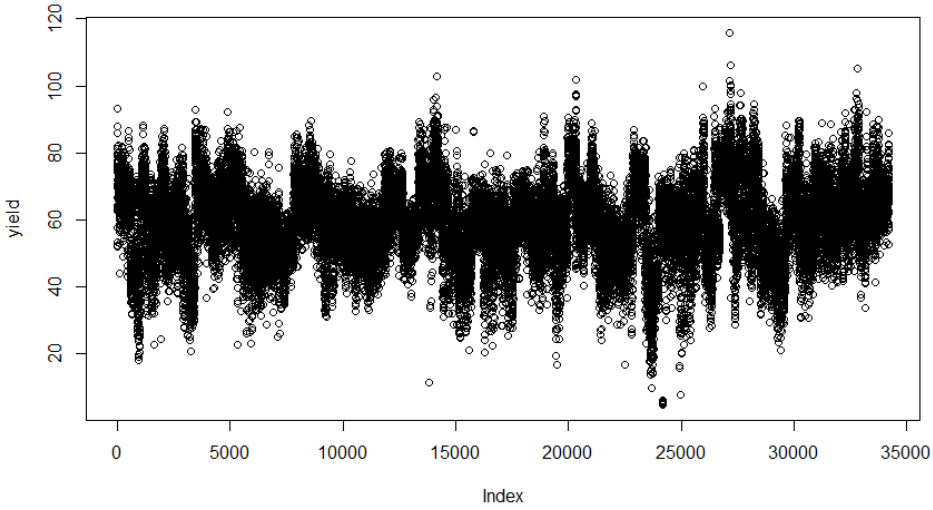

Fig 5: The Distribution Plot of the Yield Variable

3.2 Predictive Analytics

After cleaning and describing the data, I noticed that Variety_Yield is and Weather1 and Weather2 are categorized. I first divide the data into training and test data, then I used the following algorithms to analyze the factors affecting them and compared the error rates.

3.2.1 Linear Regression

The Linear regression model gives significant outcomes, which shows that the linear combinations of independent variables, including weather, soil, radiation, and others, and the model can explain the variation of response variable well. If the relation is linear, we can choose the variables with a significant coefficient to run another linear regression model. However, the linearity relation is not that common here.

```
Residual standard error: 8.509 on 1331 degrees of freedom
Multiple R-squared:  0.3799,   Adjusted R-squared:  0.3538
F-statistic: 14.56 on 56 and 1331 DF, p-value: < 2.2e-16
```

Fig 6: Output of Linear Regression

3.2.2 LASSO

Lasso is a more complex regression, and we should choose parameters that have the least error. But in this round, Lasso did not outperform simple regression

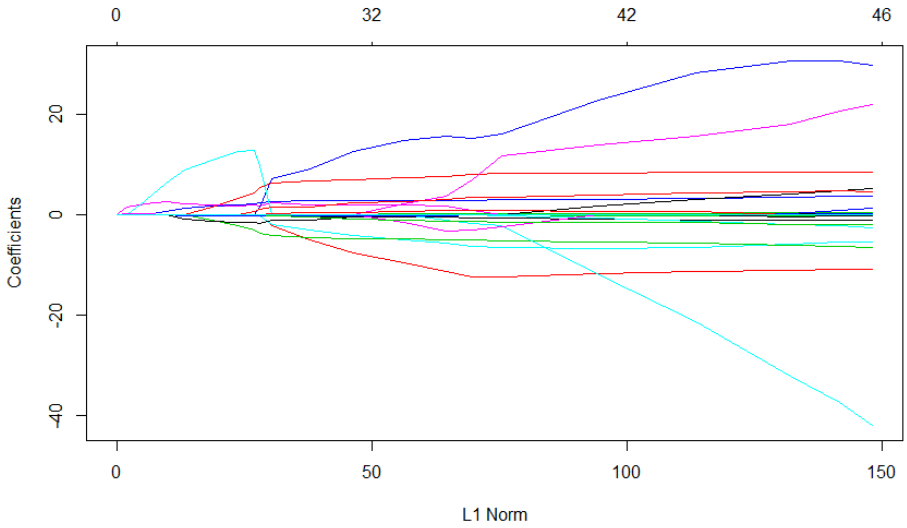

Fig 7: The Coefficients Plot of Lasso

3.2.3 Regression Tree & Bagging

The regression tree model gives branches based on the difference of predictors. We can see that precipitation, whether relative maturity, sand, and radiation are all significant predictors that significantly impact the model. In common words, the more sunshine and rainfall, the more plants grow. Bagging regression trees is also built to this data set, and we can be more certain of the results than the single tree below. The bagging method makes the model more robust.

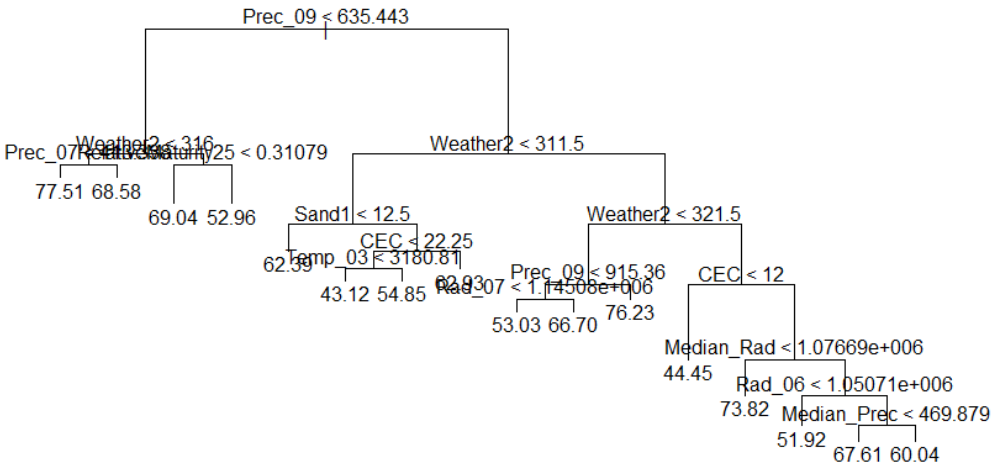

Fig 8: The Plot of Tree

3.2.4 Random Forest & Boosted Trees

According to Figure 9, random forest shows weather, precipitation, temperature, and radiation are the most important variables, similar to the simple tree model. I also applied the boosted trees technique in this project. Boosting can help improve tree models like bagging. The relative influence of the variables can be plotted from which we can see similar importance mentioned above.

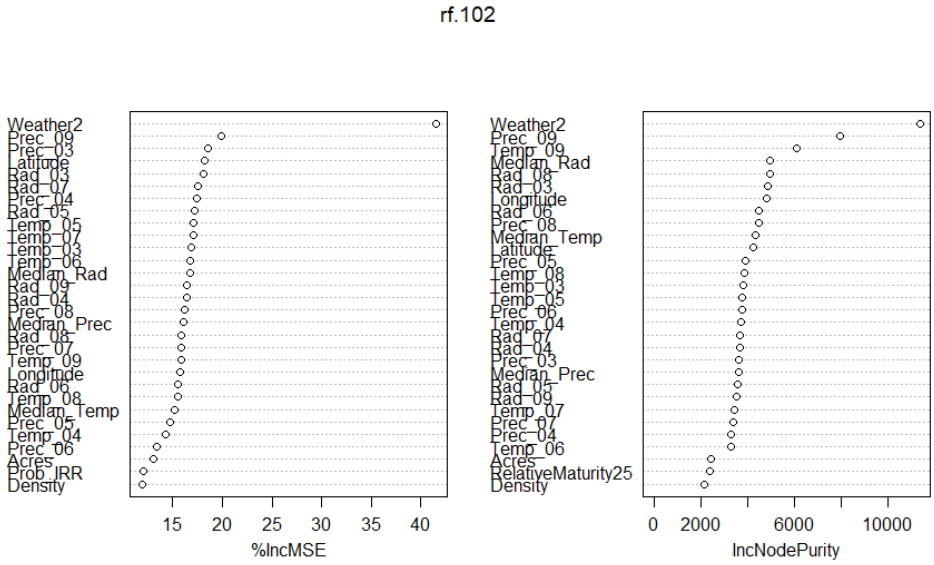

Fig 9: The Importance Plot of Random Forest

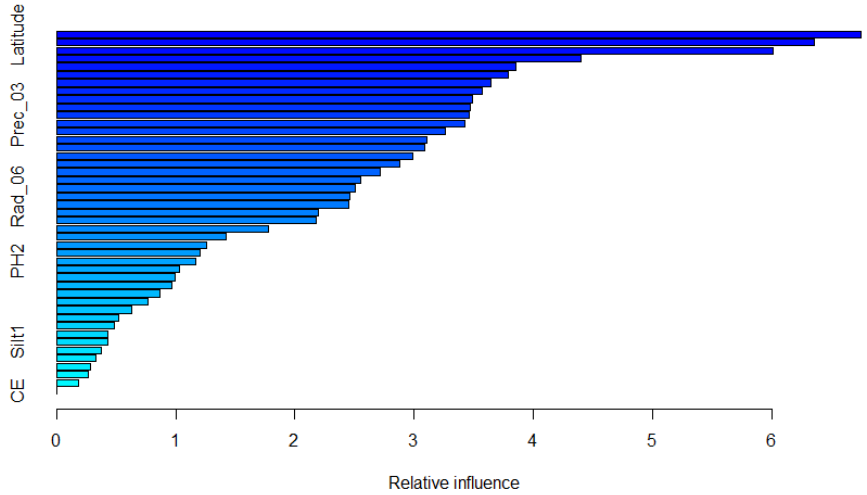

Fig 10: The Relative Influence Plot of Boosted Trees

3.2.5 Neural Network

The neural network can compute the relation between inputs and output. For this project, the variables have different coefficients on the impact to yield, and the model can be visually plotted.

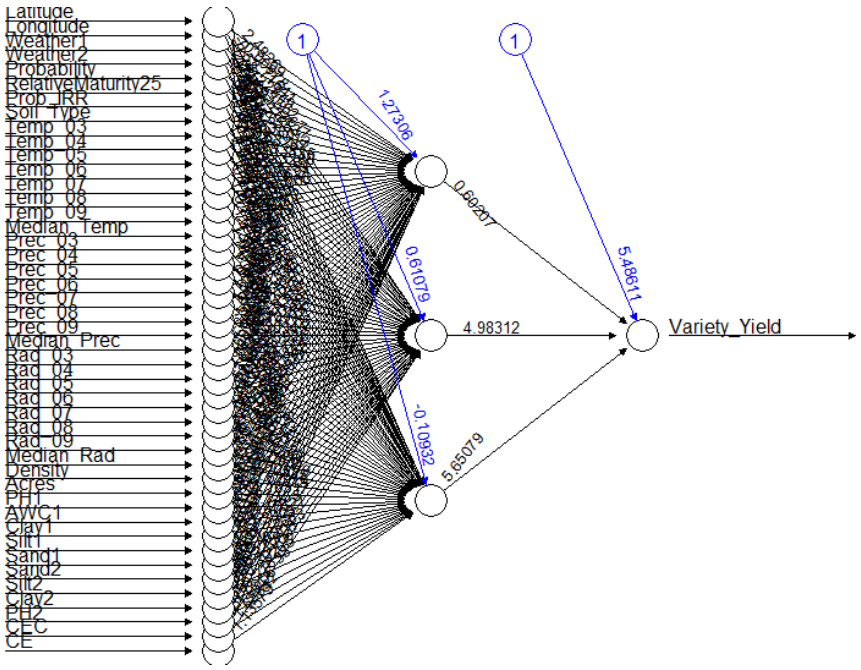

Fig 11: Neural Network Plot

3.2.6 Error Rate Comparation

As taught by the professor, machine learning models are evaluated by their prediction accuracy. At this time, MSE (Mean Square Error) was chosen as the measure to evaluate the models' accuracy. The smaller the MSE, the predicted value of the target is closer to the true value. Table 1 shows MSE values achieved by each of the models mentioned above, and we can see that the Bagging model has the least error rate, so I used it to do further work.

Table 1. MSE of Continuous Variable Models

Model	MSE
Linear Regression	78.177
LASSO	78.837
Regression Tree	63.360
Bagging	60.909
Random Forest	68.522

Boosted Trees	61.832
Neural Network	71.363

I repeated the steps above except changing some models because Weather1 and Weather2 are categorized. Compared with the error rate of other models in the test data, the Random Forest model is the most accurate one. Therefore, the Random Forest model was used to predict the Weather1 and Weather2 of the target farm, and the predictive value of them are 322, 211, respectively.

Table 2. Error Rate of Categorical Variable Models

Model	Error Rate of W1	Error Rate of W2
Logistic Regression	0.384	0.235
Classification Tree	0.348	0.277
Bagging	0.438	0.347
Random Forest	0.337	0.202
Boosted Trees	0.361	0.267
Neural Network	0.383	0.290
Support Vector Machine	0.378	0.210

3.3 Prescriptive Analytics

3.3.1 Naïve Heuristics

	Variety	mean
	<chr>	<dbl>
1	V82	78.5
2	V21	74.4
3	V19	73.0
4	V171	69.6
5	V22	69.5
6	V24	69.3
7	V41	66.1
8	V30	66.1
9	V38	65.7
10	V39	65.6

Fig 12: Predicted Yield under Naïve Heuristics

The Naïve Heuristics is used in the first step of prescriptive analytics. The top five

variety with the most average Variety Yield is V82, V21, V19, V171 and V22, so the Naïve Heuristics is to select the five variety and allocate 20 percent of the land for each variety. This method is simple and applicable, which invested only in highly productive varieties; thus, farmers can maximize their overall production.

2. Mean-Risk Heuristics

	Variety	mean	risk
	<chr>	<dbl>	<dbl>
1	V180	64.5	0.678
2	V180	64.5	0.678
3	V98	63.8	0.678
4	V98	63.8	0.678
5	V96	63.4	0.678
6	V99	61.7	0.678
7	V99	61.7	0.678
8	V95	61.5	0.678
9	V100	61.3	0.678
10	V111	61.3	0.678

... with 34,203 more r

Fig 13: Predicted Yield under Mean-Risk Heuristics

In this procedure, I took probability as the risk of the variety, which is the probability of growing soybeans in the site's nearby area. It is regarded as a risk because some high-productivity variety cannot be planted in a certain area, and we should consider this factor. The top five varieties with the max average Variety Yield and least risk are V180, V98, V96, V99 and V95, since the risk is the same and the mean yield differs slightly. Hence, the Mean-Risk Heuristics selects the five varieties and allocates 20 percent of the land for each variety.

4. Conclusion

In this project, I used the data in soybean farming to explore the factors that affect yield, which is important and extremely useful in agriculture and farming. The statistical methods can really help and improve the industry. The machine learning models for continuous variables show that the modeling for variety yield provides numerical relations with other predictors, and the Heuristics give recommended selection of varieties. The high-productivity varieties are selected to make the portfolio.

References

[1] Brown, D., Van den Bergh, I., de Bruin, S. et al. Data synthesis for crop variety evaluation. A review. Agron. Sustain. Dev. 40, 25 (2020).

[2] Godfray HCJ, Beddington JR, Crute IR, Haddad L, Lawrence D, Muir JF, et al. Food security: the challenge of feeding 9 billion people. Science. 2010;327(5967):812–818. pmid:20110467

[3] Ray DK, Ramankutty N, Mueller ND, West PC, Foley JA. Recent patterns of crop yield growth and stagnation. Nature communications. 2012; 3:1293. pmid:23250423

[4] Nalley LL, Barkley A, Watkins B, Hignight J. Enhancing farm profitability through portfolio analysis: the case of spatial rice variety selection. Journal of Agricultural and Applied Economics. 2009;41(03):641–652.

[5] Gonzalez-Sanchez A, Frausto-Solis J, Ojeda-Bustamante W. Predictive ability of machine learning methods for massive crop yield prediction. Spanish Journal of Agricultural Research. 2014;12(2):313–328.

[6] Barkley, A., Peterson. H. 2008. Wheat Variety Selection: An Application of Portfolio Theory to Improve Returns. Proceedings of the NCCC-134 Conference on Applied Commodity Price Analysis, Forecasting, and Market Risk Management. St. Louis, MO.