# Gradient Boosting Model in Predicting Soybean Yield and Probabilistic Decision-making in Elite Selection

Hongyi Lin, Jiabei Yang, Yi Sun, Peizan Wang
6127 Pershing Ave, St. Louis, MO 63112, United States, hongyi.lin@wustl.edu

This paper applies gradient boosting regression to building yield prediction model for different soybean varieties with the training experiment data from previous years. The model can help forecast yields in relation to geographical information and variety features more accurately and also identify important drivers behind yield across seeds. The prediction accuracy of this model is validated by the training RMSLE of 0.15. To select true elites, this paper proposes a probabilistic approach and identified following 13 varieties as elites: V114545, V114556, V114564, V114565, V114585, V114589, V114649, V114655, V114685, V140364, V152300, V152312 and V152320 with a F-score of 0.57. We also propose a method to distinguish elite seed varieties from genetic information. Among over 2000 genetic marker locations, 91 SNPs are proved to be able to identify the true elites above, and the method could be generalized to provide a guidance for creating new varieties from parental varieties in the future.

*Key words:* gradient boosting regression, posterior probability, variational SNP identification

## 1. Introduction

Selecting the elite soybean varieties to commercialize can be challenging. In face of the unknown environmental variables ahead, growers need seed products that generate high yields stably. Our approach to the problem is to estimate a variety's probability of outperforming its corresponding benchmark across locations. In the meantime, we have grouped varieties into different RM bands so that varieties with longer RM would not dominate the prediction results. Using the probabilistic approach, we would be able to identify true elites that can withstand the test of different geographical conditions. In addition, type I error of each year is calculated based on the selection model and shed light on how to eliminate it.

Among the machine learning methods we used to build a reliable yield forecast model, the optimal one we found is Gradient Boosting Model (GBM). We incorporated all geographical information as external variables along with important internal feature like RM. Apart from that, we add more predictors - yield level cluster and variety matrix - to account for the variety characteristics that are not yet captured. Our model is quite robust with only 0.15 training RMSLE value as well a leading R2 value on the submission result dashboard. The predicted yield from the Gradient Boosting Model strongly supported our Elite variety selection. The result indicates that our Elite varieties not only possess relatively high yield but also make up a great portfolio in which the high yield and low variance are well-balanced.

With 13 varieties recognized as elites, we further explored the relationship between genetic pattern in SNPs and elite varieties. We proposed an approach to find SNPs that has greatest variation between elite and non-elite seed varieties. As a result, we identified 91 SNPs that has significant influence on seed's probability of being elite and could be jointly used to identify elites from SNP data of seed varieties with type I error preserved, along with the corresponding nucleotide combination at those SNPs.

## 2. Criteria used to select the seed varieties

An overview of the procedure we applied to select elite seed varieties is presented in Figure 1. In our final method of selecting elite seed varieties, we first build benchmarks at each tested location for a specific stage, and then select the seed varieties that are both being tested at most locations and outperform the most benchmarks at various tested locations to advance to the next stage. To eventually find elite seed varieties that will be successful after commercialization, we utilize the results of all the 3 stages of a certain class to find the seed varieties that have both high and stable yield across various tested locations and years.

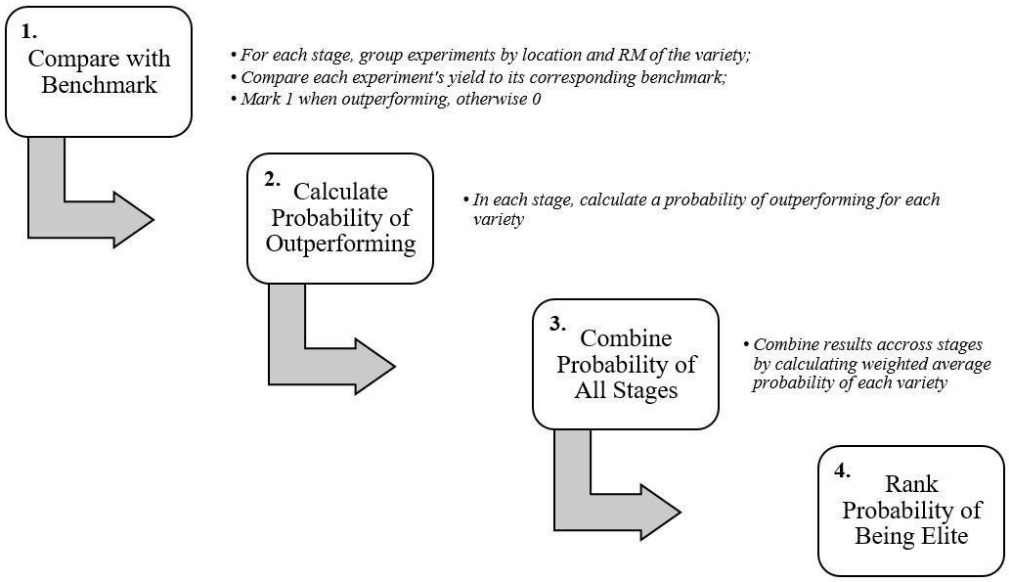


Figure 1: Overview of the process to select elite seed varieties

### 2.1 Building benchmarks for seed varieties at a specific location and stage

Provided the performance benchmarks (CHECK = True), corresponding RM and yield at various locations in different stages in both the original training dataset and the testing datasets, we want to use these information to build our benchmarks to be compared to. When we compare the yield of the seed varieties, we also want to control for the RM of the seed varieties, so that a high yield of a certain seed variety is not dominated by longer or shorter maturity period. Therefore, we build the benchmarks for tested varieties in following steps:

### 2.1.1 Group seed varieties by RM

Since we find that typically several performance benchmark varieties with different values of RM will be tested at a single location in a specific stage, at each location, we group the seed varieties that are grown there by their values of RM. The cutoff values are set to the RM of the performance benchmark varieties at that location; to have performance benchmark varieties in each group, we merge the two groups with smallest RM together, as illustrated by the gray arrow in Figure 2.

### 2.1.2 Build benchmarks by the average yield of performance benchmarks in each group

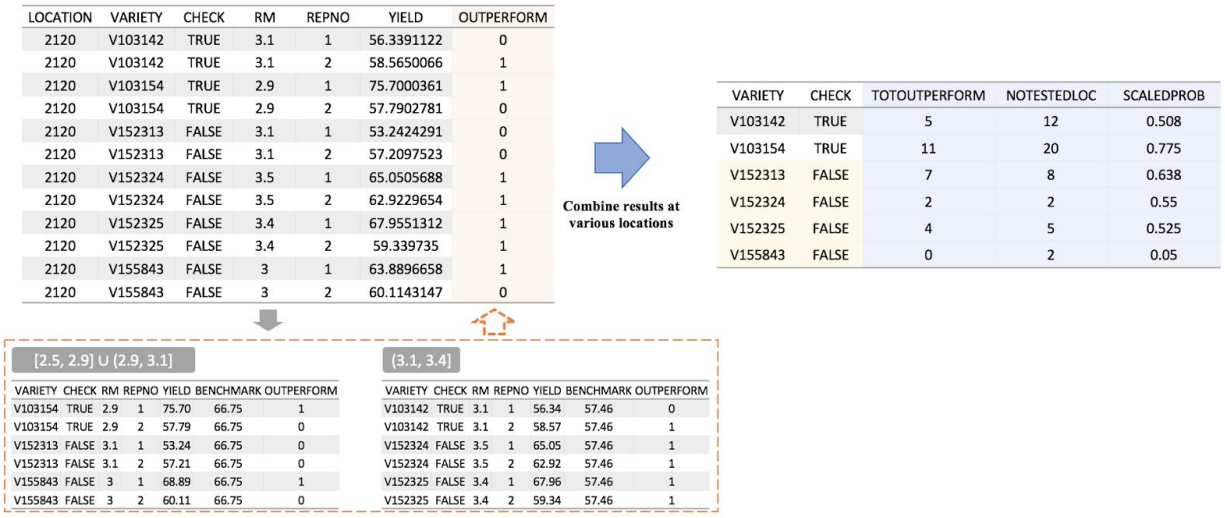


Figure 2: Procedure in detail to rank the seed varieties in each stage

In each group, we create a new variable, BENCHMARK, by calculating the average of the yield of performance benchmark seed varieties as indicated by the gray arrow in Figure 2. Sometimes there will be more than 1 performance benchmark variety in each group, which means these performance benchmark varieties take on the same value of RM, we assume that those varieties are grown in those locations for more stable and robust estimates for the benchmark yield, so we still take the average of the yield of all the performance benchmark varieties to build the benchmark for that group.

## 2.2 Criteria used to select outperforming seed varieties in each stage

We then take two steps to select outperforming seed varieties in each stage to advance to the next stage. Firstly, in each group of varieties with similar RM, we create a new indicator variable for every repeated

experiment, OUTPERFORM. OUTPERFORM will take on value 1 when the yield of a certain variety in that repeated experiment is greater than the corresponding benchmark, and will take on value 0 otherwise.

$$\texttt{OUTPERFORM} = \begin{cases} 1 & \texttt{YIELD} >= \texttt{BENCHMARK} \\ 0 & \texttt{YIELD} < \texttt{BENCHMARK} \end{cases}$$

We then link OUTPERFORM back to the original ungrouped data of a specific location as presented by the orange arrow in Figure 2.

Secondly, we merge the data across various locations to select the outperforming seed varieties in each stage as illustrated by the blue arrow in Figure 2. We argue that elite seed varieties should have the largest probability to outperform corresponding benchmarks at various locations. Therefore, for each variety, we count the number of times that the variety outperforms corresponding benchmarks, and store this information in a new variable, TOTOUTPERFORM; we also count the number of times it is tested across various locations, and store this information in a new variable NOTESTEDLOC. Note that a variety could be tested several times at the same location.

Using only the probability of outperforming corresponding benchmarks to decide whether or not a seed variety should advance to the next stage is not satisfactory. For instance, a variety that is tested 20 times at various locations and outperforms corresponding benchmarks 19 times will be worse than a variety that is tested only once across locations and outperform its benchmark for that tested time according to this criteria, since $\frac{19}{20} < \frac{1}{1}$. However, we may still think the former variety is better than the latter variety since the probability we get from 20 locations will be more robust than the probability we get from only 1 location. From this argument, we want to elevate the probabilities for those varieties that are tested at more locations relative to those are tested at only a few locations, so we create a new variable, SCALEDPROB:

$$\texttt{SCALEDPROB} = \frac{\frac{\texttt{TOTOUTPERFORM}}{\texttt{NOTESTEDLOC}} + \frac{\texttt{NOTESTEDLOC}}{max(\texttt{NOTESTEDLOC})}}{2}$$

We divide the sum of two probabilities by 2, so that the resulting quantity will still respect the range of 0 and 1 for a probability, and neither of the quantities in the numerator will play a larger role against the other. We could easily see that if some seed varieties have the same probabilities of outperforming corresponding benchmarks, the larger number of times a variety is tested at various locations, the higher chance this variety could be selected to advance to the next stage; or if some seed varieties were tested the same number of times across locations, the larger the probability a variety outperforms corresponding benchmarks, the higher chance this variety could be selected to advance to the next stage.

### 2.3 Selecting elite seed varieties for growers

We will describe two approaches to select elite seed varieties.

#### 2.3.1 Selecting elite seed varieties from stage to stage

To decide which elite seed varieties to advance to the next stage, we first discard the records of performance benchmark seed varieties in each stage since we are interested in non-benchmark seed varieties. We then sort all the varieties in a specific stage by SCALEDPROB, and depending on the sample sizes of different stages, we will recommend pick different proportions of top varieties to advance to the next stage. We recommend using top 30%, 60%, and 10% for stage I, II, and III respectively to advance to the next stage. The final selected elite seed varieties will be those advanced from Stage III. These proportions are manually picked now, but if we could know about the successful elites after commercialization, we could always optimize a better combination of proportions to be advanced to the next stage.

#### 2.3.2 Selecting elite seed varieties by overall performance along three stages

The approach above implicitly assumes that whether or not a variety should be selected to advance to the next stage only depends on the current stage, and the process simply ignores the performance of the varieties in previous stages. Here we propose that we could apply a weighted average of SCALEDPROB from the 3 stages, and pick the seed varieties with the largest weighted SCALEDPROB to be the final elites that we predict.

With the data from all the three stages, we first impute the SCALEDPROB for those varieties that failed to advance to stage II and stage III by 5 percentile of SCALEDPROB in the stage. We assume that previously eliminated seed varieties can at most perform better than 5% of the varieties that successfully advanced from the previous stage. For instance, if we want to impute the SCALEDPROB in stage II for the varieties that failed to advance from stage I, we will impute the value by 5 percentile of the computed SCALEDPROB for the varieties that successfully advanced to stage II. We hope to minimize the probability that we falsely eliminate a potential elite seed variety while trying not to introduce non-elite seed varieties to the final stage through this imputation.

After imputation, we calculate the weighted average of the SCALEDPROB in three stages. We then sort all the varieties by this weighted SCALEDPROB, pick the top seed varieties, and predict them to be the true elites. The weights for the 3 stages and the number of seed varieties to be selected as true elites could be optimized if we can know about the true elites in the previous years, the year of 2011 to 2013, and we can then apply these optimized quantities to class of 2014.

### 2.4 Results and discussion

We tried a wide range of possible values of the number of true elites to be selected and different combination of weights assigned to the 3 stages. The best result we have is when we pick the following 13 varieties to be true elites, V114545, V114556, V114564, V114565, V114585, V114589, V114649, V114655, V114685, V140364, V152300, V152312, V152320, and the weights assigned to the 3 stages being 0, 0, 100% respectively applied to the test stage data we are provided. The evaluation of our result is presented in Table 1.

| FMEASURE | ACCURACY | MATHEWSCC |
|---|---|---|
| 0.57 | 0.99 | 0.58 |

Table 1: Evaluation of true elites' selection

This result shows that the true elites to be selected will mostly or only depend on stage III data. This indicates the fact that our prediction of true elites will become more robust when the varieties are being tested at more locations in later stages.

To optimize the selection procedure, we will need the true elites in the previous years to train the parameters used in our process of defining elites, for example, the proportion of seed varieties to be selected to advance to the next stage in different stages, weights used in the last step, or the number of true elites to predict from our approach.

## 3. Estimates of Type I Errors

In this section, we are trying to estimate the type I error of previous years' data, which is, for each variety that is commercialized in a certain year, we want to see whether it is really successful (elite).

The following parts will show 1) how we regard a variety as a successful one, 2) the result of type I error estimation and 3) ways to reduce type I error.

### 3.1 Criteria of success

According to the original dataset, there is not a direct indication of which variety turns out to be a successful one, so we have to develop our own definition of 'success'. And since this definition will significantly affect the estimation of type I error, we want to make it reflect the actual result as good as possible.

In the previous section, we have discussed how to design a model which could predict whether a given variety in a certain year will be elite. And the test results are persuasive as well. Given that, we could easily figure out a way to describe whether a commercialized variety would be successful or not.

The way is to run the model on each year's commercialized varieties, namely the varieties that have passed stage three test in that year. And find out which varieties our model thinks would be successful in the following year. With that data, we could compare the fact that if a given variety is still successful after it has been selected to commercialization.

Since the model will give us a probability which indicates how likely a variety would become successful in the following years, we need to set a threshold to select out only a small portion of varieties and consider them as 'successful'. To do that, we first sorted the list of varieties and take the probability of the kth variety as the threshold, where k equals to the number of varieties that were selected to be commercialized in each year.

If all the commercialized varieties are considered successful, then the type I error rate will be 0 since they will be the first k items in the sorted list and are considered as successful ones.

### 3.2 Estimation Results

| Year | Count of Commercialized Varieties | Count of Successful Varieties | Type I Error |
|---|---|---|---|
| 2011 | 30 | 16 | 46.7% |
| 2012 | 31 | 28 | 9.67% |
| 2013 | 40 | 24 | 40.0% |
| 2014 | 32 | 27 | 15.6% |

Table 2: Type I Error of Each Year

The table shows that the estimation of Type I Error is not reliable since in some years the Type I Error is really low (only 9.67% in year 2012), however, in other years, almost half the the varieties are considered as Type I Error.

And what we can also tell is that there is no indication that the estimation of Type I Error is getting better, which means, if we keep selecting varieties in this way, no significant improvement could be achieved.

### 3.3 Recommendations

As we can see from the previous results, the current way to select successful varieties is not ideal in all cases. And based on our model, here are some suggestions to reduce type I error:

#### 3.3.1 Find a proper way to select benchmark varieties

During our exploration of the dataset, we find out that the benchmark varieties, who is checked true in the dataset, seem to have a big influence on deciding which one should be commercialized in the next year.

Although the way to select benchmark is not clear to us, we still came up with an approach made by us which is expected to have a better outcome:

The benchmark could be the variety whose performance is better than a majority (75%, 90%, etc) of others who have the same RM in the previous year's test at a given location.

### 3.3.2 Use genetic information

As we are trying to explain in the next section, there are certainly some patterns in the genetic information that could be a potential indicator of whether a variety would be successful or not. So using this information would definitely help to filter out varieties that don't have the potential of being a successful one, such like some varieties perform well in the previous year's test but lack most of the common patterns of other successful ones.

# 4. Methodology

## 4.1 Method to identify genetic patterns in elite seed varieties

Our method to identify genetic patterns in SNP of elite seed varieties is illustrated by the flowchart in Figure 3.

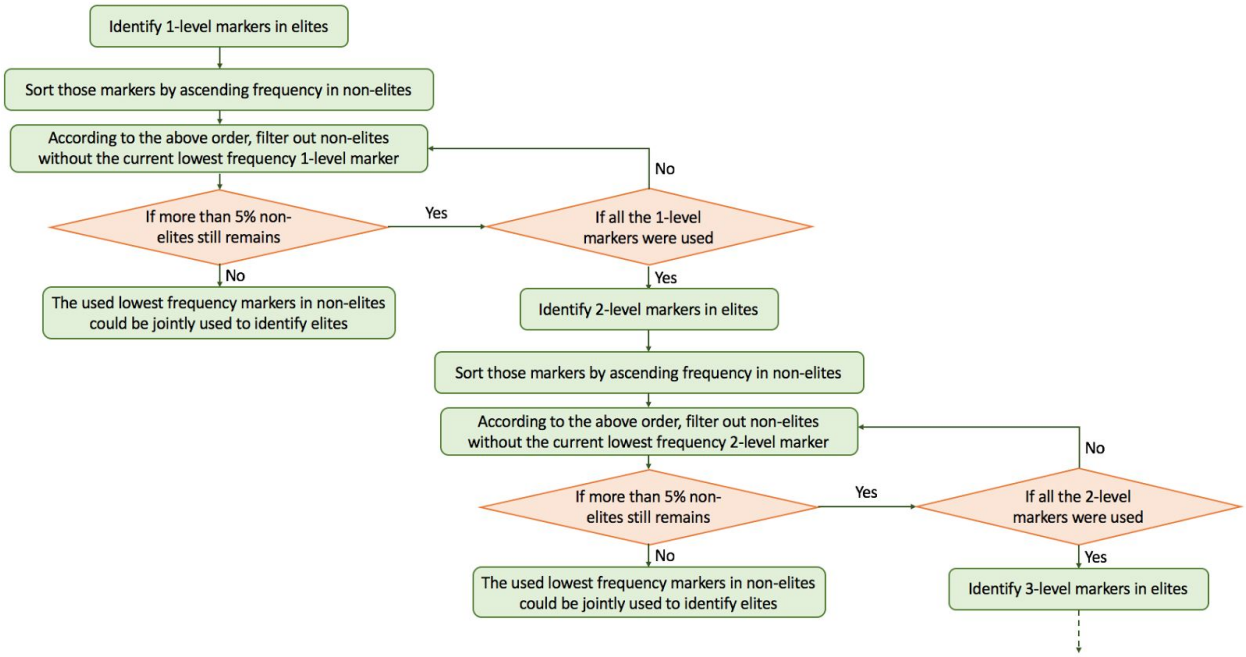


Figure 3: Procedure to identify genetic patterns in SNPs in elite seed varieties

Above all, we define the term "n-level marker" in a subset of varieties be the SNPs that can take on n possible nucleotide combinations in that subset. For example, 1-level marker means that this SNP doesn't have any variation and could only take on one possible nucleotide combination in that subset of varieties.

We divide our dataset of genetic information into two subsets of elites and non-elites. As presented in Figure 3, we first identify all the 1-level markers in elites.

Secondly, to distinguish elite and non-elite varieties, ideally, we would want non-elites could not have these 1-level markers identified in elites. However, this hardly happens so we hope to see low frequency of the nucleotide combination of the 1-level marker at the marker location in non-elites. Since the probability of the nucleotide combination of the 1-level markers in elites will be 1, we define the significant SNPs to be those with the lowest probabilities of the nucleotide combination of the 1-level markers identified in elites at those locations in non-elites. Therefore, we then sort all the 1-level markers identified in elites in ascending order of the frequency of the nucleotide combination at that marker location in non-elites.

Thirdly, we want to identify non-elites in the subset of non-elite data through finding those varieties that

don’t have the nucleotide combination of the 1-level marker. We start from the 1-level marker with lowest frequency in non-elites, filter out those varieties don’t take on the nucleotide combination of that marker and label them as non-elites. Loop through the 1-level markers until less than 5% of non-elite data could still not be identified as non-elites. “5%” is selected so that there will be less than probability of 0.05 that we label the variety as elite when it is actually non-elite, which is the type I error.

Fourthly, if after we loop through all the 1-level markers, there are still more than 5% non-elite data could not be identified as non-elite, we will start our analysis on 2-level markers. We can continue the process using similar reasoning, until we could identify 95% non-elites in the non-elite subset of data.

Finally, we can now identify elites by identifying those have all those markers that were looped through, by preserving the type I error that there will be 5% chance that we will label a variety as elite when it is actually non-elite.

### 4.2 Results and discussion

We apply this approach with the elites defined as the true elites we identified in Section 2, and present the result in Figure 4. In Figure 4, x-axis is all the SNPs in genetic information dataset, and every vertical line indicates a significant SNP, with the height of the line being the probability that a non-elite variety won’t have this SNP, so the higher the vertical line is, the more significant the corresponding SNP is.

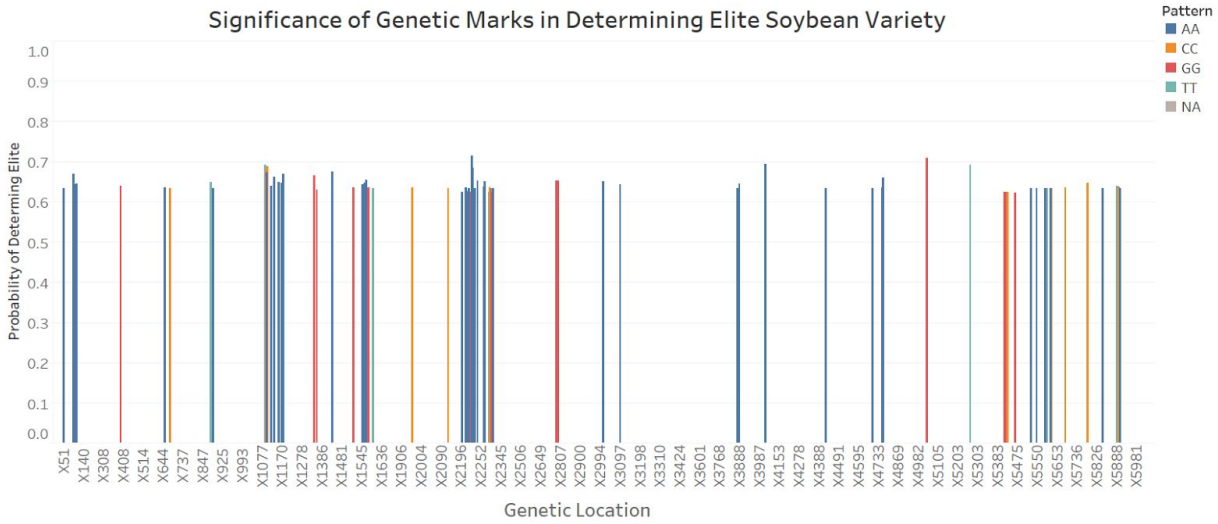


Figure 4: Significant SNPs to be jointly used to identify elites

From the above result, we find 91 significant SNPs and that 1) all the significant SNPs for identifying elite varieties are homozygous in this dataset and 2) the probabilities that a non-elite variety won't have these SNP are similar, providing a guidance in the creating new varieties from parental varieties in the future.

Our method could be easily generalized to larger genetic information dataset with more seed varieties or more elite seed varieties.

# 5. Quantitative results

We built a separate model to predict the soybean yield of 2015-2016 on the sample variety set which was given by the host in which contains the "Elite" varieties which we successfully selected above. The yield prediction model is Gradient Boosting Model(GBM) (implemented by the R-package: gbm). In addition, we use K-means, pre-processing and model parameter tuning to optimize the performance of the GBM. Eventually, the new model is trained with a combination of the original training dataset given by the organizer and all 3-stage datasets. And the result is predicted on the dataset which contains all of the sample varieties and the imputed 2015 geographic information. This new model gives us the other half of the answer to our overall objective which gives the predicted yield to the elites we found. Also, the relative high prediction yield of our selected elite supports our results of the elite selection process.

## 5.1 Yield prediction of elites

Based on the Gradient Boosting Model, we predicted the yield of all sample varieties. The summary of the prediction is presented in Figure 5.

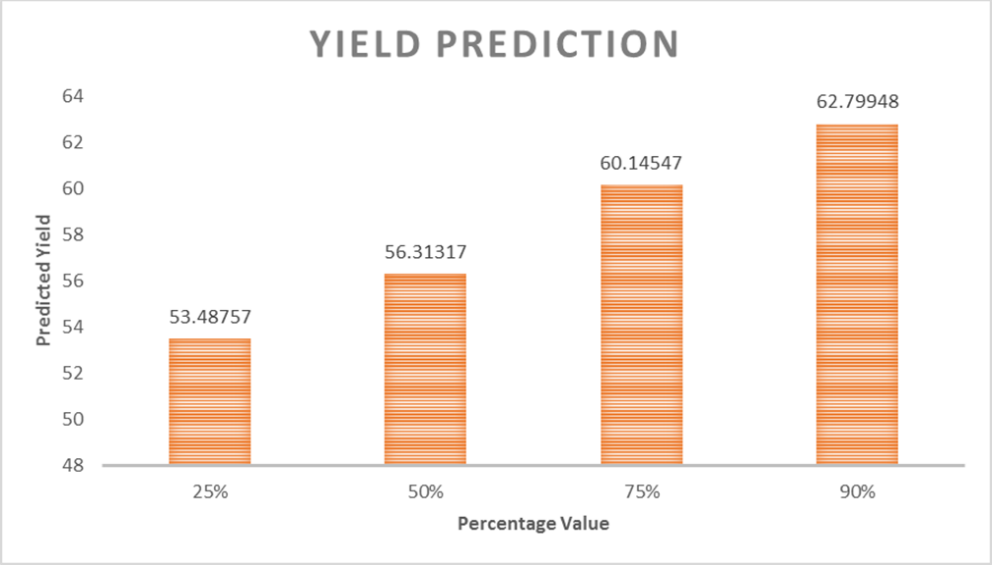


Figure 5: Percentage Value of Predicted Yield

The predicted yield of the elite varieties is presented in the table below. We find that among the 13 Elite, 11 of 13 has predicted yield higher than 75% of all sample varieties, and 8 of 13 has predicted yield higher than 90% of all sample varieties. Our yield prediction model supports our elite variety selection.

The predicted yield of the elite variety is presented in Table 2. We find that among the 13 Elite, 11 of 13 has predicted yield higher than 75% of all sample varieties, and 8 of 13 has predicted yield higher than 90% of all sample varieties. Our yield prediction model supports our elite variety selection.

| VARIETY_ID | PREDICTION | VARIETY_ID | PREDICTION |
|---|---|---|---|
| **V114564** | 66.66108053 | **V152300** | 63.61032317 |
| **V114565** | 66.47303753 | **V140364** | 61.43064506 |
| **V152320** | 64.79817763 | **V114585** | 61.42520173 |
| **V114589** | 64.71376262 | **V114556** | 60.50117771 |
| **V114655** | 64.39811759 | **V114545** | 59.96594522 |
| **V152312** | 64.18389591 | **V114649** | 58.01324667 |
| **V114685** | 63.71540086 | | |

Table 2: Predicted Yield of Elite Varieties

From Table 3 we could find that among this 8 most high-yield variety, the top 2 has largest predicted yield but with high variance. The following 6 has relatively high yield and low variance, compared with the average variance among sample variety which is 21.8.

| VARIETY_ID | PREDICTION | VARIANCE |
|---|---|---|
| **V114564** | 66.66108053 | 40.21177431 |
| **V114565** | 66.47303753 | 39.45898033 |
| **V152320** | 64.79817763 | 0.435812607 |
| **V114589** | 64.71376262 | 4.169078697 |
| **V114655** | 64.39811759 | 14.81584076 |
| **V152312** | 64.18389591 | 6.798070945 |
| **V114685** | 63.71540086 | 9.318599429 |
| **V152300** | 63.61032317 | 6.530267113 |

Table 3: Predicted Yield and Variance of Top 8 Elite Varieties

The good variety portfolio should be the balance of high yield – high variance variety and relatively high yield – low variance and the Elite variety we selected exactly form such a well-balanced portfolio.

### 5.2 Model optimization process

Our model optimization process is summarized in the following Figure 6.

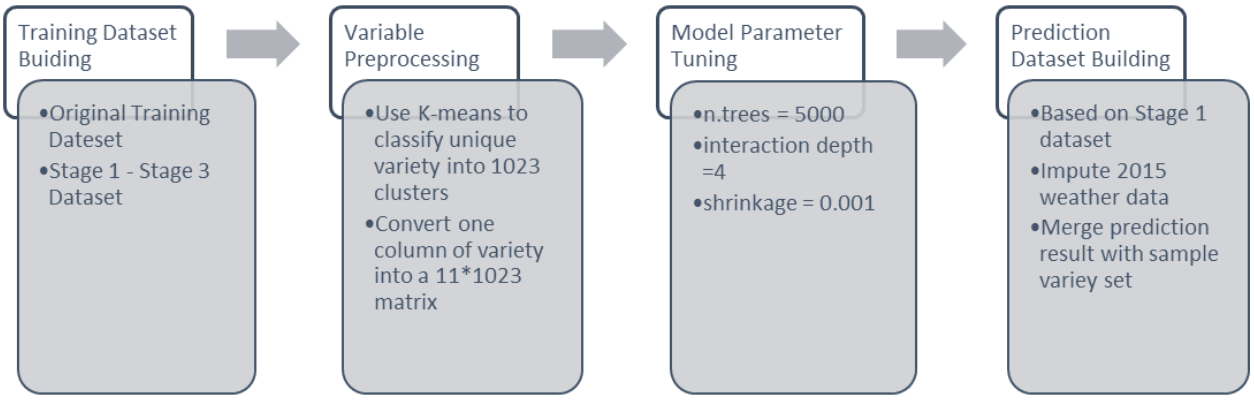


Figure 6: Model Optimization Process Summary

### 5.2.1 Data preparation

Based on the results of the constant experiment, we found that, to let the model learn as much as possible from the previous information, we should not only use the training dataset but also include the experiment information from Stage 1 to Stage 3. Also, due to the fact that GBM is not as influenced by the skewed scale of its variable as other models such as neural network or LASSO do , we feed the unscaled data into variables.

The yield forecast is built upon Stage 1 dataset which combined with imputed 2015 geographic information. Since we don't have the 2015 weather data, we take mean of the most recent 5-year geographic information as 2015's. And the reason we use Stage 1 dataset is that, first, our well-trained model should be able to output yield prediction on all varieties given in the sample variety set, and the Stage 1 dataset includes all of the sample variety. Second, the locations that covered in Stage 1 is more accurate. We originally joined the Stage 1 varieties with the 59 locations which appear in Stage 1 and predict on this joined dataset. However, after several attempts, we notice that the result predicts on 59 locations is never better than just simply predict on Stage 1 locations. We speculate that the actual location of 2015-2016 sowing is more similar with locations in Stage 1-3. As each variety of Stage 1 is planted at multiple locations, we took the median of the predicted yield on different locations as the final result to avoid the influence of extreme value.

### 5.2.2 Model selection and variable preprocessing

Besides the difference of geographic conditions, it is a legitimate inference that there must be some hidden factors within the difference of varieties which could influence the yield. Since we exclude genetic information from model building, we need to find another approach to representing the diversity among different variety at least family.

We first tried a series of models covered both parametric (Linear regression with/without polynomial variable, Lasso Regression) and nonparametric (Regression Tree, MARS, Neural Network) and we discovered that first, since one Family includes several Varieties that if we include VARIETY as a variable into our model, then all information in Family could also be covered. But on the contrary, if we just use FAMILY, then we would miss some information in Variety. Thus, we exclude Family in the further model building. Second, the Regression Tree Model includes the variety as a variable gave us relatively the best result. Thus, we reasonably deduce that the more sophisticated tree-based model would give us even better result and that lead us to use Gradient Boosting Model.

The boosting model provides us a lot of advantages. GBM can fit complex nonlinear relationships, and automatically handle interaction effects between predictors. Since the GBM is based on the hierarchical structure of tree model, it means that the response to one variable depends on values of variables on a higher place in the tree, so interactions between predictors are automatically modeled. After tuning the parameter, our GBM models contains 5000 trees, interaction depth equals to 4 and the learning rate is 0.001.

However, for GBM, the maximum level of a categorical variable is 1024, yet in our dataset, there are 11,141 unique varieties which require us to preprocess the variety to fit the constraint. We start with 2 approaches to deal with this problem.

#### 5.2.2.1 Use K-means to classify variety

In the training dataset, we use the median yield of one variety on different location as the yield of one

unique variety. Then we use K-means to classify varieties into several clusters by yield. Also, the prediction result getting better with the increase of cluster number. Thus, we classify all varieties into 1023 clusters and the "within-cluster-variation" is only 12.9. In the GBM, we would use the cluster number to represent multiple varieties.

#### 5.2.2.2 Convert variety into a 2-dimension matrix

We convert one column of 11,141 unique variety into a 2-dimension matrix and use a column number and a row number to represent a unique variety. We tried a number of column number and row number combinations, from 106*106 to 11*1023 and the result turned to be that with the matrix changing from a square to a narrow rectangle the predicted result generally getting better. Thus, we use 11*1023 matrix to represent all of the varieties.

Eventually, we combine this two approach and convert one column of Variety into 3 columns, one represents the cluster number, one represents the column number and one represents the row number of each variety.

## 5.3 Model outcome

### 5.3.1 The Relative Influence of variables

From the GBM, the relative influence of the top 10 variables is presented in the following table:

| Variable | Relative Influence |
|---|---|
| Cluster | 28.27 |
| SILT_TOP | 26.15 |
| Temperature | 6.73 |
| PH | 6.30 |
| PREC | 6.03 |
| Row Number | 4.58 |

| | |
|---|---|
| AREA | 4.29 |
| ORGANIC.MATTER | 4.05 |
| IRRIGATION | 3.97 |
| RAD | 3.26 |

Table 4: Relative Influence of Top 10 Variables from GBM

We notice that there are two variety-related variables in the top-10 list. Especially the cluster number of variety, it is the most influential variable. That finding is reasonable since we could expect the yield is mostly influenced by the characteristic of the variety itself, instead of the environment.

For the environment factors, we find out there are 3 soil-related variables: SILT, PH, and ORGANIC MATTER. Especially the SILT which indicate the percentages of medium size soil particles, it is the second most important variable. The Soil is such an important factor is understandable since the soil quality is the most directly influencing environmental factor for soybean. In the top five variables, the third is temperature and the fifth is related to precipitation. This result would suggest that the final yield of soybean should be a result of the synthetic action of soil, water, and temperature.

### 5.3.2 Model prediction accuracy measurement

One of our measurement of prediction accuracy is Root Mean Squared Logarithmic Error (RMSLE). The formula of RMSLE is:

$$\epsilon = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(p_i+1) - \log(a_i+1))^2}$$

n is the total number of observations in the dataset, pi is the prediction result, and ai is the actual response. This measurement would penalize underestimates more than overestimates. Since we don't want to underestimate yield and accidentally eliminate possible good variety, we choose to use RMSLE to measure prediction performance. Based on the GBM model, our training RMSLE is only 0.15 which

is a decent result.

The other one of our most important indicator of prediction accuracy is the R2 index on the Codalab. Although the result of GBM is below zero, yet the GBM provide us the highest score we could get, and that result is still at a leading position on the dashboard.

**5.4 Reflection and Discussion**

To predict the 2015-2016 yield of elite varieties and to justify our elite selection process, we build series of models - parametric and nonparametric methods - to predict yield. Eventually, the most reliable model turns out to be Gradient Boosting Model(GBM). After building complete training dataset and properly converting VARIETY variable, we reached our best result with predicting on the Stage 1 dataset combined with imputed 2015 weather information. Our model is quite robust with only 0.15 training RMSLE value. The predicted yield from the GBM supported our Elite variety selection. The result indicates that our Elite varieties not only possess relatively high yield but also make up a great portfolio in which the yield and variance of yield are well-balanced. Nonetheless, there are still several shortcomings.We infer that simply taking the mean of the most recent 5-year weather data as the future data might be arbitrary , and the locations of experiments for each variety is not precise enough. Besides, our procedure to convert Variety into the GBM usable form is not perfect which means we may still lose some variety information. The area that is yet to be optimized would be the starting point of the future study.

## 6. Team members


- Hongyi Lin, 6127 Pershing Ave, St. Louis, MO 63112 , hongyi.lin@wustl.edu
- Jiabei Yang, 1000 National Ave, Apt G52, San Bruno, CA 94066, bei9533@mail.harvard.edu
- Peizan Wang, 1000 National Ave, Apt G52, San Bruno, CA 94066, peizan.wang@sv.cmu.edu

• Yi Sun, 1031 Highlands Plaza Drive West Apt. 108, St. Louis, MO 63110, will.sun@wustl.edu

## 7. Supplementary materials

• github repo: https://github.com/danielwpz/ai-challenge-soybean