NewNKSoy Final Report

Michael Ko, Patrick Nicholson, Jingzhuo Xu, Xueqi Zhao

Introduction

Our project is focused around the 2017 INFORMS O.R. & Analytics Student Team Competition. Commercial soybean selection, the process by which Syngenta chooses elite soybean varieties from a large set of experimental varieties and ultimately commercializes them, is a very thorough and drawn out process. A given soybean variety must make it through 3 stages of testing over 3 years in order to be commercialized. After the selection is complete, each variety's performance is evaluated based on the number of bags of seeds sold.

The purpose of this year's competition is to build a model and to predict the sales potential of different soybean varieties for seed biotech company Syngenta. More specifically, the model should be able to predict, among the "class of 2014" soybean varieties, which varieties to select for commercial release.

Description of the Data

We are given two datasets: one with the experiment data from previous years 2009-2014, and one with the experimental varieties in the class of 2014. We are analyzing the former dataset. It contains 258,253 observations with 12 columns for each experiment.

- YEAR – the year of the experiment (range 2009-2014)
- EXPERIMENT – experiment code; first digit denotes relative maturity of site (example 09YT000052; 512 unique values)
- LOCATION – the location of the experiment (example value 3210; 152 unique values)
- VARIETY – the soybean variety being tested (example V030090; 15,632 unique values)
- FAMILY – "breeding population" of variety (example FAM13986; 1,938 unique values)
- CHECK – denotes whether used as benchmark for other tested varieties (example TRUE)
- RM – relative maturity, 0.1 increase = 1 day of maturity (example 2.3; range 0-5.8)
- REPNO – indicates the replication number of the experiment (example 3; range 1-4)
- YIELD – production in bushels/acre (example 53.10325; range 0-124)
- CLASS_OF – final year tested before commercialized (example 2011; range 2011-2014)
- GRAD – indicates whether the variety graduated to commercialization (example YES)
- BAGSOLD– bags sold after commercialized (example 871,341; range 0-1,865,000)

Information about the year, variety, family, relative maturity (RM), and yield is provided consistently for the observations. In contrast, information about graduation (Class_Of, Grad, Bags Sold) is only given for less than 8.5% of the entire dataset, and it is "missing" for the rest of the observations since they are cut out before graduation. There are also some entries which do not make sense, such as a variety which graduated in 2013 but was experimented on in 2014; these corrupted entries will be removed from the models.

We notice that there is no location or experiment type for which all varieties have an experiment result from. For each location, there are on average only 753.8 unique varieties that have a test result from there, with the maximum being 11,410 (out of 15,632 total varieties). Similarly, for experiment types, there are on average 31.95 unique varieties that have run an experiment of

each type, with the maximum being 39. Figure 2 shows how yield varies between the location of the experiment. The x-axis is the location code so separate vertical lines represents for different locations.

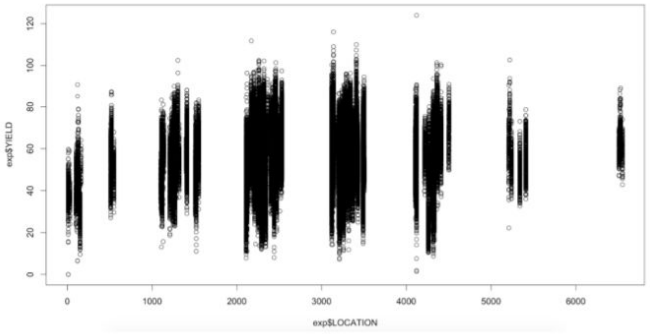


Figure 2: Plot of YIELD vs LOCATION, showing how different locations have different distributions of yields.

Descriptive Statistics and Preliminary Analysis

Our challenge is to predict the potential sales volume of each variety tested in the final year of the class of 2014. Given the experiment data, the output space could be BAGSOLD, or alternatively a Boolean variable indicating whether or not to commercialize the variety. Variables such as GRAD (whether or not the variety was commercialized) and YIELD may help us in making these predictions. After reading the data description thoroughly, we checked the correlation between the given factors, as shown in Figure 3. Thus, we tried several linear regression models on the data, analyzing and interpreting the coefficients and R-squared errors.

One of the main difficulties with this competition is that, out of all the data we are provided, only a very small subset relates to soybean varieties that were eventually commercialized and sold. Just 4,622 out of the 215,411 non-benchmark experiments relate to commercialized varieties. There are 42,842 “Check” experiments included, which are past experiments on elite varieties used as benchmarks. However, the data set does not include the number of bags sold for these benchmark varieties, making it more difficult to use them in a model. Ideally, there would be a small subset of the locations or experiments which prove "significant" in predicting BAGSOLD; however, we have yet to find this among our early linear models.

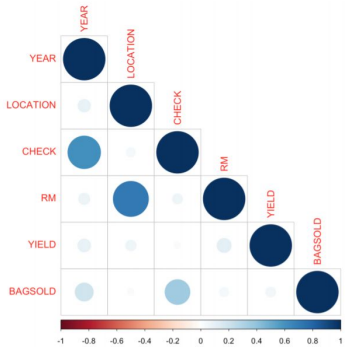


Figure 3: Correlation Plot

Initial Model Fitting and Testing

To avoid over-fitting, before building any models, we decided that for all models we would use k-fold validation with k=10. The training data set would be used to fit the model while the testing set would be used for validation. However, we did not limit ourselves to 10-fold validation. For example, another way to address over-fitting is to use regularization, which can penalize complex models and prevent over-fitting. Ultimately, the team used various validation methods in our models, although 10-fold cross validation was the most consistently used.

As a starting point for our model fitting, we fit a model on only the experiments on commercialized varieties. We did a linear regression looking at only those experiments with graduation and commercialization data completed, with BAGSOLD as the output space and YEAR, RM and YIELD as the predictors. We added another predictor LOC to represents the first digit of the LOCATION variable, which denotes the relative maturity of the experiment site. For example, LOC1 represents locations with LOCATION value beginning with "1." Using p-values to determine which predictors were significant in this model, it seemed that LOC1, LOC2, LOC3 and YEAR (p-value all close to 0) were significant, while RM and YIELD were not (p-values>0.5). The adjusted R-squared was very low at 0.09 because there were so few features to fully explain the model. And coupled with the very non-intuitive suggestion by the model that YIELD values do very little to predict BAGSOLD, it would seem that this model is incomplete.

A second preliminary approach was to treat this as a classification problem, and fit a model that would predict whether a given variety would be commercialized or not. Logistic regression immediately came to mind as a good method for dealing with classification. Fitting the model using the entire data set with YEAR, RM and YIELD as predictors and a Boolean GRAD as the output space, we found that YEAR and YIELD were the most significant. Using the same formula on a linear regression, we obtained the same results, suggesting that RM may truly be not related at all to whether a variety graduates or not. However, these models merely inform us a bit more on Syngenta's tendencies in picking graduates, and do not help with regards to predicting which varieties will sell the best. This is an important distinction; the goal here is to improve Syngenta's selection process, not to merely mirror it.

Tree-Based Methods and Bootstrapping

We decided that using decision trees would be an easily-interpretable model to continue with. Although trees do not have the same predictive accuracy as other techniques, using bootstrapping and random forests can significantly improve this performance. Random forests present an improvement over bootstrapping by decorrelating the trees. A number of decision trees will be created on bootstrapped training samples, but each time there is a split in the tree, only a small random subset of the predictors is considered. This prevents one strong predictor from dominating all of the bootstrapped trees, which would result in highly correlated predictions. We chose to use random forests because it is the same as bootstrapping, except it lowers variance by only selecting a subset of the features (James, 320).

We constructed a regression tree seen in Figure 4 with the subset of experiments with BAGSOLD, GRAD, and CLASS_OF information filled in. We use BAGSOLD as the output space and all relevant predictors in the feature space: YEAR, REPNO, YIELD, RM, LOC. We

notice that RM is clearly the most important variable here as it is primarily used for the branching. The max value for BAGSOLD is achieved through an RM < 2.05. RM values less than 3.45 are observed to obtain BAGSOLD values higher than the median value of 871,300. We use our tree on our test set and observe that the square root of the MSE is 149,351.4 which suggests our model's predictions are within 149,351.4 bags of the true median bags sold (James, 328).

Next we apply bootstrapping and random forests to the model, with m = 3 predictors that are considered for each split, and n = 500 trees. We view the importance of each variable in terms of node purity, which is a measure of the total decrease in purity (or decrease in Gini index) that results from splits over that variable, averaged over all trees. Thus RM is the most important, as seen in Figure 5, as accuracy of the model drops the most when RM is removed from the model.

The test set MSE for this random forests model turns out to be rather large: 16,714,678,749, with the square root value being 129,385.3 bags. Although this error value is not so bad, considering that the standard deviation of BAGSOLD among commercialized varieties is 427217.3 bags, we still should hope to do better, and thus we investigated some other models further.

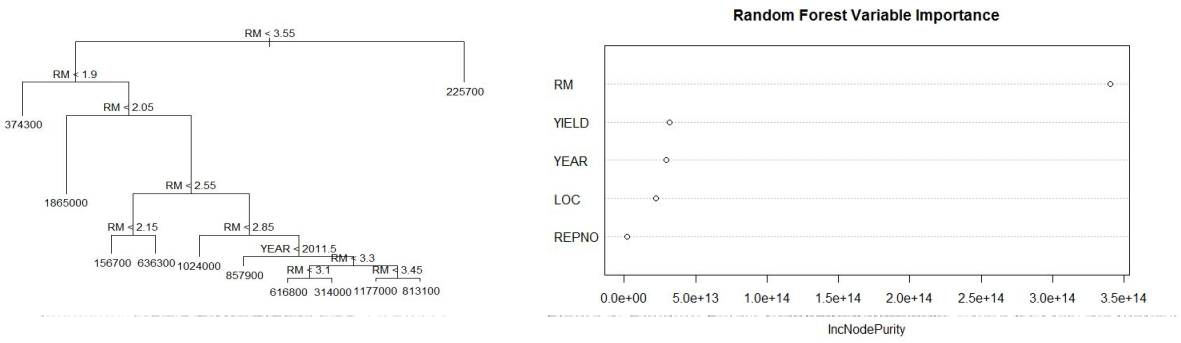


Figure 4 (left): Regression Tree with predictors: RM, YEAR, YIELD, LOC, REPNO & response: BAGSOLD
Figure 5 (right): Random Forest Importance measures, where RM has the highest variable importance

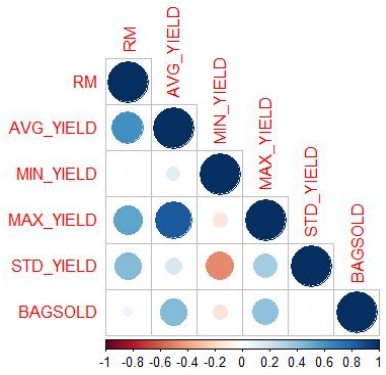
Feature Engineering Approaches

We figured some form of feature engineering or transformation would be necessary in order to somehow aggregate a single variety's performance over many experiments. The experiment data, in its initial form, had a row for each experiment result. Instead, we would prefer to have a row for each variety, as this would make it easier to fit models that would predict a single BAGSOLD value for each variety.

Ridge Regression Model

From our data analysis on the initial, unmodified data set, we noticed that two primary features which can significantly affect BAGSOLD are YIELD and RM. Because each variety has different values of YIELD for several experiments, the team decided to add AVG_YIELD, MIN_YIELD, MAX_YIELD and STD_YIELD as the rest of features in the model, aiming to characterize the YIELD distribution for each variety. These features represent the mean, minimum, maximum and standard deviation among the YIELD values for all experiments run on a variety. The transformed data set, with only one row per variety, is much smaller than the original; ultimately, we only used 33 rows, as there are only 33 varieties which were

commercialized (and thus have values for BAGSOLD). The correlation plot, as shown below, illustrated that AVG_YIELD and MAX_YIELD are positively correlated with BAGSOLD; however, MIN_YIELD has a negative relationship with BAGSOLD. While this result for MIN_YIELD may seem unintuitive, it could be that even elite varieties can have severely underperform on certain experiments.

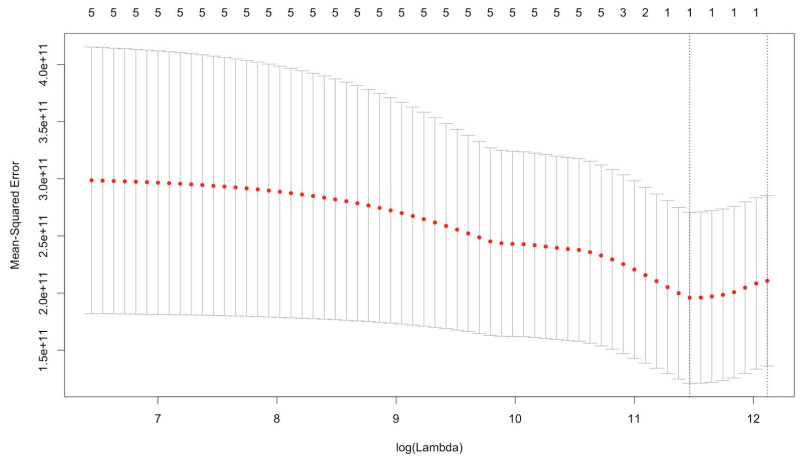


Before fitting the linear model with new input features, we normalized the dataset, transforming all features to the same scale. Then, we tried the simple linear model. From the summary of model fitting, we saw that none of the features are significant due to relatively large p-values. Furthermore, the adjusted R-squared has quite a small value with 0.3151 and rather large mean squared error with 1.049E11. We suspected the reason for such a bad result is that the input features suffered from multicollinearity. When multicollinearity occurs, least squares estimates are unbiased, but their variances are large so they may be far from the true value. In order to solve this problem, we performed ridge and lasso regression to add a degree of bias to the estimates, which can reduce the standard errors and give estimates that are more reliable.

We first did 10-fold cross validation to determine the optimal lambda, resulting in 10.5. Then we fit the ridge regression, summarizing the intercept and features' coefficients. The resulting model has heavy positive weight on AVG_YIELD (114662.77) compared to MAX_YIELD (86897.76), whereas RM(-45729.17), STD_YIELD (-71366.38), and MIN_YIELD(-39908.89) show negative effects on BAGSOLD. After predicting for the evaluation dataset, it has 0.98E11 mean squared error, which is smaller than that of simple linear regression model. Even though the result is fairly better, the ridge regression cannot zero out coefficients, which means we have to include all features in the model. It may be unreasonable if some features are irrelevant to the output space, leading us to fit a model using lasso regression.

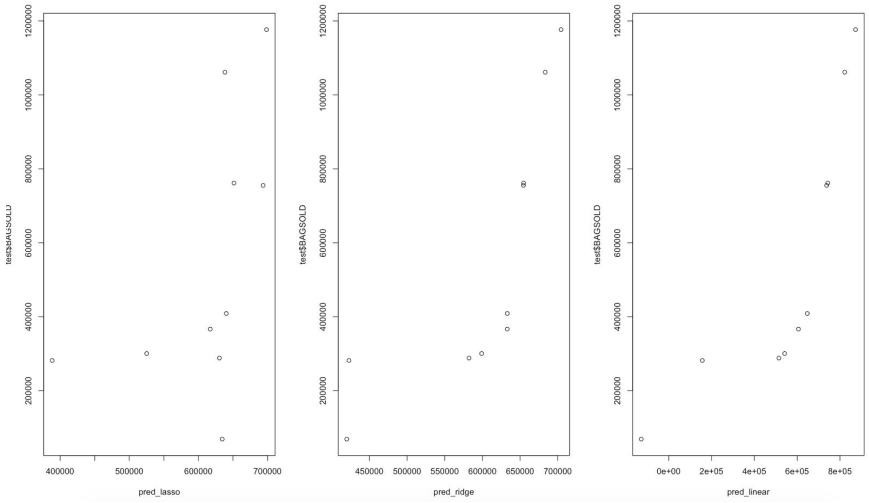
Lasso Regression Model

Unlike ridge regression, the lasso regression does both parameter shrinkage and variable selection automatically because it zeros out the coefficients of collinear variables. Again, we performed 10-fold cross validation for choosing the best lambda. The following figure compares the MSE versus different values of lambda and demonstrates the best lambda as 95416.72, which achieves lowest MSE for the model.



By fitting in the lasso regression model, we found that RM, MIN_YIELD, MAX_YIELD and STD_YIELD have been eliminated, and only AVG_YIELD (88766.42) has positive effect on BAGSOLD. The model has a 1.039E11 predicted mean squared error, which is slightly smaller than the results for linear regression but still higher than that for ridge regression.

The plots below compares the prediction results for the three models. The y-axis denotes the true BAGSOLD for varieties in the testset and the x-axis is their predicted BAGSOLD by the corresponding model. Since none of the models managed to achieve high accuracy with MSE that is low enough, we can examine the relative values instead of directly comparing the absolute values. For example, there are points in the lasso plot that have low true value but were predicted to have high sales (in the lower right corner), and they can be the type I error in practice. On the other hand, both ridge and linear assure that popular varieties are predicted with high sales so they will be selected for commercialization by these two models, and vice versa. Therefore, we would rather recommend using linear or ridge model with normalized predictors for commercialization decisions.



Utilizing Yield Means

One potential issue with engineering features such as AVG_YIELD and STD_YIELD was that we were not accounting for the different distributions of YIELD values for different experiment

types. Since all varieties were not tested using the same set of experiments, our lasso and ridge regression models could potentially be skewed towards varieties tested using experiments that tended to give higher yields, and perhaps were conducted in more favorable growing conditions.

Our first approach to handle this was to utilize varieties with CHECK = TRUE as benchmarks, and divide any result from a given experiment type by the mean of the CHECK variety results for that experiment. For example, if a variety yields 100 at an experiment where the average CHECK yield is 80, the yield value will be converted to 1.25, representing that the variety has a yield that was 125% of the CHECK mean. CHECK varieties are elite, already commercialized varieties that Syngenta includes in the experiments as a point of comparison. Thus, one would expect that a variety's performance relative to successful commercialized varieties should correlate with its own success once commercialized.

After applying this transformation, we noticed that the correlations between features were similar to those in the data set with AVG_YIELD; the maximum yield ratio ("yield ratio" is the term we use for percent yield relative to the experiment means) had the highest correlation with BAGSOLD at 0.3192, and the minimum yield ratio had the lowest at -0.4677. Once again, the simple linear model performed poorly; the best combination of features, as determined using the metric adjusted R-squared, still produced a test root mean squared error (RMSE) of 379060.6 bags after performing 10-fold cross validation. Ridge and lasso regression performed even worse, with test RMSE's of 416451.9 bags and 414715.8 bags, respectively.

Considering the possibility that the CHECK varieties might not be as great representatives of elite varieties as previously thought, we then tried to instead use the means of all experiments instead of the CHECK means. This would better make use of the entire data set, as it could show how the commercialized varieties compared in experiment performance to all tested varieties. Thus, even though we would only be fitting models on the commercialized varieties, we would not be completely discarding the information from non-commercialized varieties. For models fit on this data set, the best linear model had a test RMSE of 380918 bags, the ridge regression had a test RMSE of 408689.8 bags, and the lasso regression had a test RMSE of 414624.9 bags.

Time Series Data

Another approach was to create separate columns for results from each year of testing a variety. The intuition here was that in splitting yield information by year, we could try to recognize trends over time regarding the yields of these varieties, and highlight how a variety might perform differently on, for example, third-year experiments compared to first-year. Unfortunately, this approach merely created more missing data within the data set; among all varieties, 0.2% had no experiment results for their first year of testing, 96.6% had no second year results, and 99.1% had no third year results. Granted, this was mostly due to most varieties not graduating to the second or third year; however, even among commercialized varieties, 51.6% were missing results from at least one of their 3 years.

To deal with this missing data, we first tried fitting a generalized low rank model using the LowRankModels package in Julia. However, it seemed that despite trying several different loss functions and regularizers, our GLRM fit always resulted in a data set with unrealistic YIELD

values. Perhaps this was a case of converging to local minima, or simply having a data set with too many missing values for the GLRM to converge to a reasonable factorization.

We then tried directly predicting missing values using other features. Fitting linear models that predicted values for one year given values of another, we were able to approximate all missing entries within the data set. The best linear model on this subset, determined using best subset selection with adjusted R-squared, gave a test RMSE of 353353.8 bags, and included the maximum year 1 yield, minimum year 3 yield, and average yields from years 1 and 3. Fitting ridge and lasso did not yield better results, as both had test RMSE values over 400000.

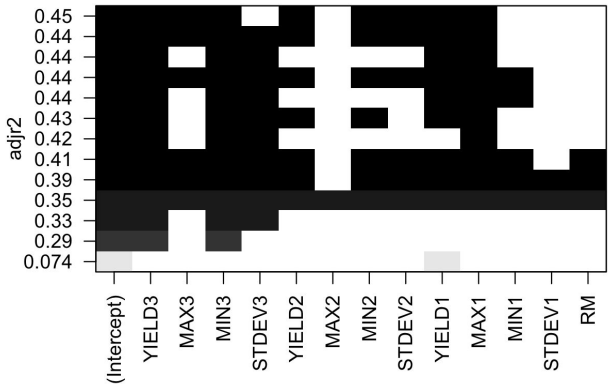


Figure9: Plot of best subset selection on data set with missing values filled using prediction. The Y axis is adjusted R-squared of each model, while the X axis is each feature (shaded if included in each model)

Conclusions

Surprisingly, the best performing model out of the ones we fit was one of our first, the random forest model. However, the conclusions drawn from this model largely conflict with what we discovered after applying feature engineering to the data set; namely, the random forest model considers RM to be the most important feature, while analysis of the data set after aggregating experiment results for each variety found almost no relationship between RM and BAGSOLD. Nonetheless, even the best of our models fit on the feature engineered data set, such as the ridge regression and the linear model on time series data, performed 2-3 times worse than the random forest model on test data sets.

As our results currently stand, we would likely not recommend Syngenta to utilize any of these models for decisions regarding Soybean commercialization. With slightly less than two months until the competition's January 30 deadline, we plan on focusing more on unsupervised learning approaches to the problem; although our early attempts at k-means and spectral clustering have not yielded great results so far, they seem more promising than our linear and tree-based models. Additionally, unsupervised methods allow us to fully utilize the entire data set, clustering varieties based upon YIELD, RM and LOCATION information while not requiring BAGSOLD information for every entry.

Works Cited

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: With Applications in R*. New York: Springer, 2013. Print.