Soybean Midterm Project Report

Patrick Nicholson, Michael Ko, Xueqi Zhao, Angela Xu

October 29, 2016

1 Introduction

Our project is focused around the 2017 INFORMS O.R. & Analytics Student Team Competition. The competition this year is focused on creating a model 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.

2 Description of the Data

We are given two datasets: one with the experiment data from previous years from Year 2009 to Year 2014, and one with the experimental varieties in the class of 2014. We are working with the former dataset. It contains 258,253 observations with 12 features for each experiment (including replications of the experiment; some experiments were done up to four times). Information about the year of the experiment, the variety and family being tested, soybean relative maturity (abbreviated "RM"), and yield of grain produced are provided consistently for the observations. There are 15,632 soybean varieties and 1938 families, or populations from which the varieties originate. Relative maturity is a number indicating the amount of time it takes to reach physiological maturity relative to other varieties, and has historically been shown to be related to yield. Additionally, there is a column labeled "CHECK" indicating whether the variety is an elite variety that has already been commercialized and therefore not being tested. These elite soybean varieties are used to measure the experimental variety performance of the other varieties in the dataset. Furthermore, information is given about whether each variety graduates to commercialization, the final year it is tested before commercialized, and the number of bags of seed sold in the second year after commercialization.

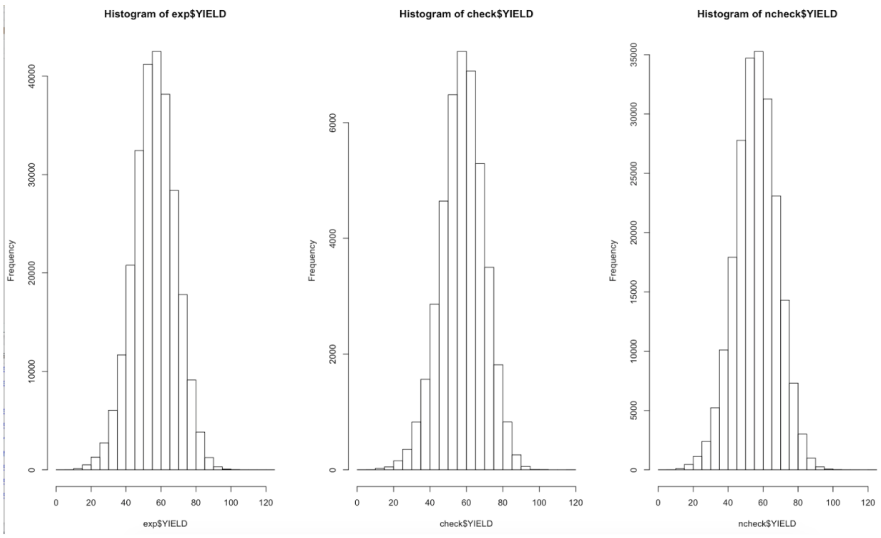

Figure 1: Histograms for YIELD in (1) the whole data set, (2) only rows where CHECK is True, and (3) only rows where CHECK is false.

This graduation and commercialization information is only given for less than 8.5% of the entire dataset, and it is "missing" for the rest of the observations. By inspection we also see that there are a small number of observations that do not make sense in the data: for example, the variety labeled V139548 failed to graduate in 2013 and had 0 bags sold, and yet was experimented

on in 2014. Additionally, there are rows for experiments done in 2014 that have "TRUE" for the "CHECK" column, and are also identified as "CLASS OF 2014". From the problem statement, "TRUE" for "CHECK" should only happen after the graduation year. This corrupted data will be removed from the models.

For each experiment, information on location and experiment type are provided. Across the entire data set, there are a total of 512 unique experiment types, and 152 unique locations. However, 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. 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.

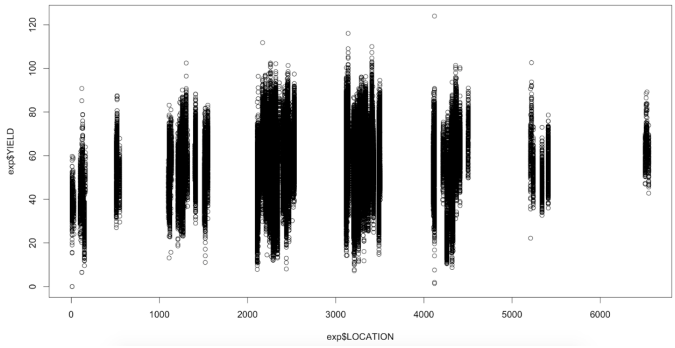

Figure 2: Plot of YIELD vs LOCATION, showing how different locations have different distributions of yields.

3 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 could be 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. By reading the data description thoroughly, we found that factors LOCATION, RM, and FAMILY may have correlation with the predictor YIELD. 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 4622 out of the 215411 non-benchmark experiments relate to commercialized varieties. There are 42842 "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.

4 Model Fitting and Testing

For the purpose of avoiding overfitting of the data, before fitting any models, the team decided to first split the data into training set and testing set, and applied k-fold validation with k=10. The training data set was used to estimate the response while the testing set was used to validate the training process on the training set. During the cross validation process, different sub-data sets will be used as test data, resulting in a more conservative estimate of generalization and ensuring a higher accuracy of the prediction model. However, we will not be limiting ourselves to 10-fold validation. For example, another way to address over-fitting is to use regularization. Regularization controls the penalty for complexity and forces the magnitudes of the parameters to be smaller (shrinking the hypothesis space), which will prevent over-fitting. Ultimately, the team will be using

variety of validation methods in the next step.

As mentioned before, out of the two provided data sets, we will use the experiment data to build our model, and then evaluate the prediction accuracy and test effectiveness based on the given evaluation data set. Bias and variance are two sources of errors that can take place as we build the predictive model. Achieving a balance between bias and variance - by reducing the variance and tolerating some bias - can lead to a better and more effective predictive model.

As a starting point for our model fitting, we simply fit a model on only the experiments on commercialized varieties. We did a linear regression, with BAGSOLD as the output space and YEAR, RM and YIELD as the predictors. Using p-values to determine which predictors were significant in this model, it seemed that RM (p-value of 0.00206) and YEAR (p-value of 0.01085) were significant, while YIELD (with a p-value over 0.1) was not. Out of trying all other combinations of these three predictors in simple linear regressions, the model using all three had the highest adjusted R-squared, at 0.004193. This is still an extremely low value, though, and coupled with the very non-intuitive suggestion by the model that YIELD values do very little to predict BAGSOLD, it would seem that these models are non useful.

A second 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 this time. 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.

5 Future Development

Now that we have analyzed the data fairly thoroughly, our next steps largely revolve around finding the ideal method for dealing with this atypical data set. A particular issue involves finding a way to compare experiment results across varieties, since there is not a lot of overlap from variety to variety in terms of which experiments were performed. For example, out of all experiment locations, only one had more than half the varieties take part in experiments there. Thus, we are going to somehow have to find a way to incorporate smaller, location-specific experiment trends into one larger model. One way could be to try to identify locations that have a large covariance, and consider experiment results from co-varying locations as being the same feature. This could decrease the overall number of different kinds of experiments significantly, and allow us to then analyze which types of experiments are most significant by fitting models.

Another idea for dealing with the many different experiment types/locations would be to form some sort of adjacency matrix. For example, if variety X and variety Y took part in 3 of the same experiments, and their yields from those experiments are 98% similar, then this adjacency matrix would have a link or edge between X and Y with value 0.98. We could then use this matrix as features, or use some sort of unsupervised learning method like spectral clustering to classify elite and nonelite varieties using this matrix. The downside to this approach would be the interpretability of the model, and how to deal with overfitting and finding appropriate cutoffs within such a 'black box'-like model.

Even if we do not end up using an adjacency matrix, we will certainly have to perform some sort of feature engineering to create a more useful feature space. Potential ideas include grouping experiment yields by year and creating time series data, combining correlated locations and/or experiments together (as discussed earlier), or converting yield values to percentages of the benchmark value for each experiment.