Commercialization of Soybean Seed Varieties

Cody Stancil (CAS9XX) • Gabriel Rushin (GHR3YS) • Nick Kim (NSK9RK)
UVA Data Science Institute | INFORMS OR & Analytics Student Team Competition

Data Mining Final Project Fall 2016

Abstract

The goal of this analysis was to help a biotech company called Syngenta select the soybean varieties that will be most productive upon commercialization. Syngenta products soybeans and commercializes them each year. Each soybean variety that makes it to the market has gone through a very selective process and passed a series of tests. Soybeans go through classes much like students in a school. Each year they are tested and only some will pass; the rest will be removed from consideration. After several years of testing, the soybean varieties reach a graduation year and a select percentage will be selected to finally be commercialized. Syngenta aims to commercialize only the one that will produce the best results for consumers. Therefore the main aim to this study is to use analytics to help Syngenta choose the best varieties and minimize false positives (seeds that test well and perform poorly).

We chose to build a deep learning model that would predict whether a variety should pass or not. However, we do not know which seeds passed in 2014 or which seeds *should have* passed. So we first built a model based on the data from 2009-2013 since we know how these varieties performed and if they were selected. We evaluated this model and got a sense of how the model performed when we used it to make our 2014 predictions. However, since we do not know which selection were the best ones in 2014 we do not know for sure how accurate our model was in predicting the best varieties. Lastly to make our sales volume predictions for the varieties we used past sales volume data and extrapolated the growth rates forward for each variety.

The conclusions reached from this conducted analysis are that (1) the deep learning approach to modeling the classification part of the problem achieved an AUC of 0.7392 and which almost beat the baseline standard set at 75% from previous research and (2) predictions for the numbers of bags sold with the selected varieties were done using a mixed effects model to account for the multi-level problem and random effects from location and family.

Syngenta's Problem

Descriptive and Normative Analysis

Syngenta is a biotech company that seeks to use analytics to commercialize only the best soybean varieties. Currently, soybeans move through a graduation process in order to be selected for commercialization. Each year, soybeans are tested and evaluated; some pass through to the next stage of testing, many fail and are discarded.

This rigorous and lengthy process is to ensure that the best performing and most profitable varieties will be chosen to meet customers' expectations.

For this study, Syngenta has provided us with a data set that has data on several classes of seeds across many years (2009-2014). For seeds from 2011-2013 we know which varieties graduated and how they performed in the market (judged by how many bags were sold). One goal of the study is to make a prediction for the class of 2014 of how many bags of seed will be sold for the varieties we choose to advance to be commercialized. This means another goal of the study is to select the best varieties of seeds.

Even though this is a careful selection process, not all of the best varieties are chosen; there are many selections that have tested well but do not perform well when they are commercialized and grown. These are type 1 errors or false positives and therefore should not have been commercialized. There are several factors that may be associated with selecting poorly performing seed varieties. First, varieties are tested in different locations at different stages of testing. The first stage may be grown in 10 locations and the second and third stage previously has been up to 30 locations. It is possible that those locations are not representative of the markets where the soybeans are being sold. Even if those markets are representative, the weather and testing conditions vary from year to year and therefore they may not be indicative of the year when the seeds are sold and grown. There may be many testing factors that are affecting the current situation: the nutrient levels of the tested fields compared to the fields of the actual farmers, the level of care in growing, the above mentioned factors, and pure randomness (our type 1 errors).

Using the available data, the ideal alternative situation is to more accurately predict the sales volume of the commercialized seeds. Additionally and most importantly, the forthcoming analysis will create a better seed selection process. Therefore, the seeds that are selected to graduate and become commercialized would at the end actually be the best varieties of all possibilities.

Stakeholders and Impact

There are two main stakeholders in this problem. First is Syngenta. Syngenta aims to remain as profitable as possible and provide the highest quality product to consumers. This will help retain and attract new consumers to their product. Therefore, if we can minimize the amount of false positive varieties that slip through and become commercialized, consumers will likely continue to purchase the seeds produced by Syngenta.

The second stakeholder the consumer. Many of the consumers rely on the product as a stream of income. If they buy a product that does not perform well in the fields, then they will potentially have wasted a lot of time and money. Syngenta understands this and wants to eliminate this harm and risk to the consumers.

Objectives, Data, and Metrics

Objectives

The analysis has two primary objectives. The scenario will be as though it is in the final stage of testing and select the seed varieties that should be released for commercialization. The primary step is to determine and classify which varieties in the final year of testing should end up ‘graduating’ to commercial production utilizing all of the variables at hand. The second step is to take those selected varieties from the first model and then make a prediction of the potential sales volume for each. At the end of the analyses, using the models and conclusions from them, a selection of the best set of seeds to be commercialized will be made containing no more than 15% of the provided number of options.

Data Description

The data provided for this analysis is provided by Syngenta and includes information on the past four classes of seeds. A training set is given that consists of the experimental data from 2009 - 2013 that includes the location of testing, the variety and breeding family of the seed, experiment number, replication number, among many other descriptive variables.

Additionally, the check variety is the elite commercial variety that is used as a benchmark that measures experimental variety performance. The last predictor is relative maturity which reflects differences in amount of time it takes individual varieties to reach physiological maturity. According to Syngenta, late maturing varieties have greater yields than early maturing varieties which makes it an important effect in the analyses. The actual sales volume data for each of these years is also provided.

Secondly, an evaluation data set that includes information about the varieties that were tested in 2014 is provided. This set details the varieties from which to make the selections from. The limit of 15% results in selecting no more than 5 of the 38 varieties listed in this set.

Metrics

The metric that will be used for the classification part of the problem in identifying the varieties that should graduate on to commercial production is ROC and AUC. This is a very simplistic way to visually understand and recognize the overall performance of the model that is built in a binary classification problem such as the one in this problem.

The second part of the problem involving a prediction of the number of bags sold for the varieties that are selected will involve a model building process with fixed and random effects that will compare models using the likelihood ratio test to ultimately find a final model.

Literature and Previous Work

While the applications of machine learning algorithms in the agricultural field exist, the purpose of this analysis is very specific - mainly, to increase the accuracy at which the superior varieties of soybean are selected correctly. Looking at literature to see what previous work has already been done in this realm of research turned up many helpful approaches.

Heslot et. al. in 2011 in their publication entitled “Genomic Selection in Plant Breeding: A Comparison of Models”, outlined a helpful approach. In their comparison of several machine learning methods, they found genomic selection could be based on a reduced set of models such as weighted Bayesian shrinkage regression, and random forest to capture non-additive effects. They also found linear combinations of different models as well as bagging and boosting methods not very helpful in improving accuracy. Their main conclusion was also a reiteration of one of the biggest problems in this analysis which is that there are large differences in accuracy between subpopulations within a dataset that cannot be explained purely by differences in phenotypic variance and size [1]. This is analogous to the soybean problem outlined by Syngenta which provided the motivation to try a mixed effects model involving both fixed and random effects.

An actual application of the mixed modeling approach was done by Useche, Barham and Foltz in their model of recent genetically modified corn adoption data using a mixed-multinomial logit model to estimate the effects of traits and farm characteristics on adoption outcomes to eventually show what is important in preferences for seed traits [5].

Another approach was taken by Zhang et. al. in 2009 to solve a very similar problem to Syngenta’s soybean selection process. In their prediction of soybean growth and development using artificial neural network (ANN) and statistical models, they modeled almost an analogous situation in which four years of field data and a fifth validation year were taken from experiments with a vegetative growth state and reproductive growth stages (including maturity, a feature in the dataset relevant as it was provided for this particular situation by Syngenta as well). Their conclusion was that the ANN method was the one that provided the greatest accuracy in predicting phenological events indicating that it can be applied in crop modeling [2]. This conclusion from their study provided motivation for a deep learning or artificial neural network approach in modeling the Syngenta problem.

In addition to Zhang et. al.'s results, a similar study was done that was looked into to see if similar results were achieved. It was done by Kaul, Hill and Walthall and was an identification of whether or not ANN models proved to be a superior methodology for accurately predicting corn and soybean yields. This was confirmed in their conclusions and was another study confirming the usefulness of neural networks and deep learning in modeling this problem [3].

A third study done by David Johnson was an assessment of pre- and within-season remotely sensed variables for forecasting corn and soybean yields in the United States. In this approach, Johnson attempted to account for the variables such as climate and seasonality in his approach to forecast yield. While Johnson's study analyzed data at a level much more granular than the one provided by Syngenta, his approach showed that regression tree-models for within-sample coefficients of determination (R^2) were 0.93 for both crops. In prediction for the test year of 2012, his derived models compared reasonably well against the outlined baseline official statistics with (R^2) 0.77 for corn and 0.71 for soybeans [4].

Hypotheses and Approach

Hypothesis

The hypothesis is that deep learning classification methods will be able to correctly classify the graduation class on the test set provided with an AUC of > 75%. Although Syngenta's current false positive rate is unknown, 75% AUC seems to be a roughly objective and conservative estimate for a well-performing classifier in this instance given the minimal data provided as well as the many underlying variables that cannot be controlled such as climate and location variables among other things. Provided in Johnson's study, the baseline official statistics were around this number as well, giving a general idea of what to aim for in terms of an objective measure of success in this crop forecasting problem [3].

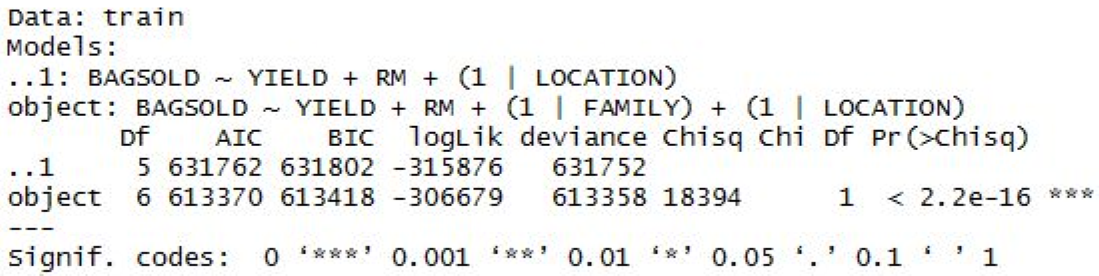
Methods

Keeping in mind the successful results from the literature review, the chosen methodologies were utilizing a deep learning approaching in the classification part of the hypothesis and then applying the selected varieties to a mixed effects model approach to account for the random effect that may be associated at different levels, namely, the ones involving the location that the crop was grown on and the family that the soybean variety comes from.

A deep learning modeling structure was chosen to classify the varieties by which one provided the best yields. The training set was the data from the class of 2013 up to the final testing year of 2013 and the test set was the final year of 2014 which was used to validate the model that was built. This modeling structure examined relationships between our variables and assign weights to those relationships. The variables that were explored were all of the ones available in the dataset. Those weights will be reassigned until an optimized loss function is found. The loss function chosen was cross entropy because this it penalizes false positives; this fits well with the business objective which is to allow less varieties through that test well but perform poorly.

The loss function to use was an available tuning option but also the choice of many other parameters to choose from such as the number of hidden layers and how many nodes are in those layers. There are too many possibilities to do an exhaustive search to find the globally optimal solution. Iterating through, tuning the models, and doing a grid search for the best model in the selection were hypothesized to achieve the best local results. The resulting selected model is the one that has the best and most consistent training and testing AUC and MSE values. The hidden layer parameters that created the best model were with two hidden layers between 5 and 10 nodes.

In forecasting a prediction for the number of bags sold for the selected varieties from the output of the deep learning model, a mixed effects model approach was used. A mixed model refers to the use of both fixed and random effects in the same analysis; random effects have levels that are not of primary interest, but are representative of a random selection from a much larger set of levels. The levels identified were at the individual variety level, the family level, and by location level. Starting with the initial fixed effects yield and relative maturity, the likelihood ratio test was done in determining whether or not the added random effect in question was contributing to the model.



```
Data: train
Models:
..1: BAGSOLD ~ YIELD + RM + (1 | LOCATION)
object: BAGSOLD ~ YIELD + RM + (1 | FAMILY) + (1 | LOCATION)
       Df    AIC    BIC  logLik deviance Chisq Chi Df Pr(>Chisq)
..1     5 631762 631802 -315876   631752
object  6 613370 613418 -306679   613358 18394      1  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 1. Exemplary output of one of the likelihood ratio tests done to consider which elements of the model were statistically significant

Evaluation Setup

The evaluation of the hypothesis was done in a multi-faceted manner. The first metric was the AUC for the deep learning model. Secondly taking a look at the TPR and FPR identified how well the predictions were done. This was done because the classes were imbalanced so taking a look at raw accuracy could have been misleading. Moreover a side objective was to determine how many false positives were identified by the model.

The prediction for the number of bags sold for these selected varieties was difficult to find a metric for but a baseline check was done against other similar varieties in the same family by location to check for any obvious outliers or in understanding anomalies in the predictions.

Results

The deep learning model classifies both “Yes” and “No” for any given variety since varieties are tested many times. So the deep learning model will not directly predict which varieties should graduate. To determine which seed varieties should graduate from the class of 2013 the sum of the “Yes” predictions were calculated after being aggregated by variety of the deep learning model. The top five varieties ranked by the “Yes” predictions were then selected as the target graduating group. The top 5 varieties represent 15% of the varieties of the class of 2013. We then compared those varieties to the 5 varieties that should have graduated. Our model selected these 5 seed varieties to graduate from the class of 2013:

V103142, V103150, V155843, V155842, V103163

The following seed varieties should have graduated based on the number of bags sold:

V139548, V156553,V152053, V156806, V152079

After we modeled on 2013 we made our predictions for 2014:

V114655, V114553, V114556, V114649, V152306

After the deep learning model was built and validated, it produced an AUC value of 73.92%.

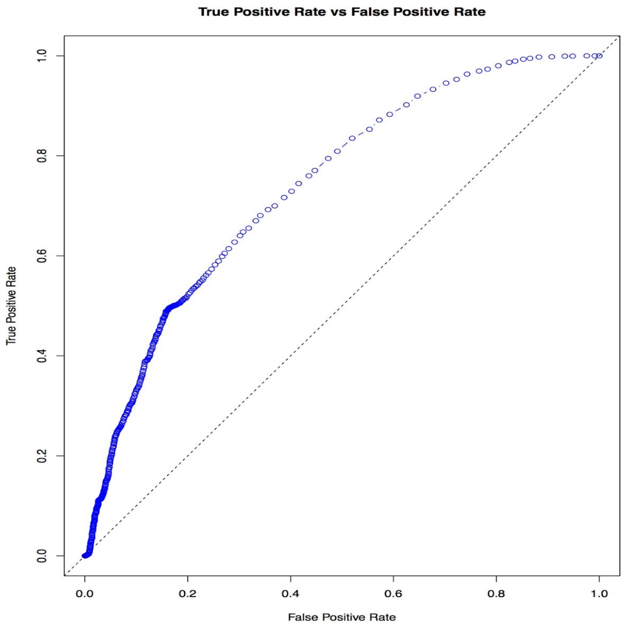
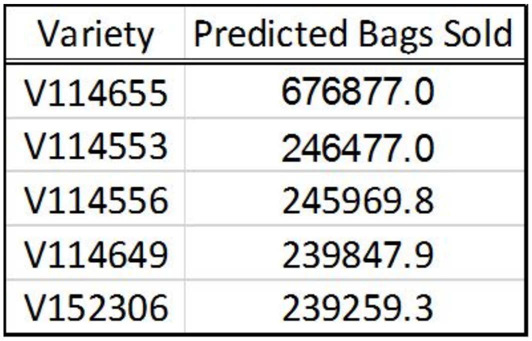


Figure 2. The receiver operating curve of the deep learning classification model.

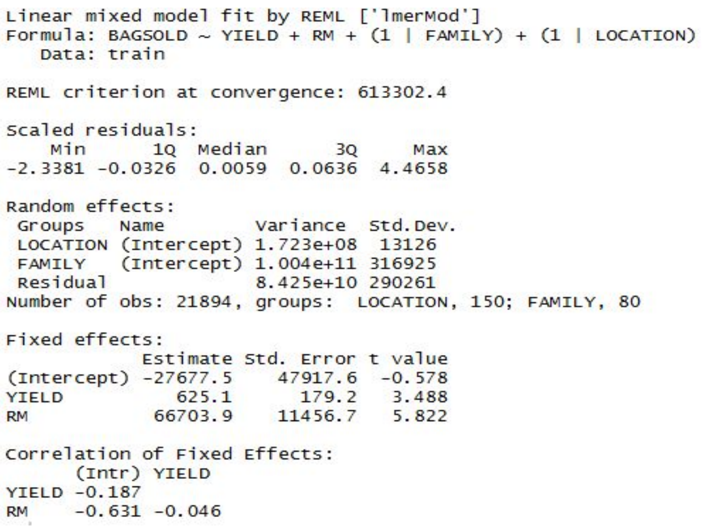
Using the selected varieties, the predicted number of bags sold in the following year 2015-2016 for these were 676877.0, 246477.0, 245969.8, 239847.9, and 239259.3.



Variety	Predicted Bags Sold
V114655	676877.0
V114553	246477.0
V114556	245969.8
V114649	239847.9
V152306	239259.3

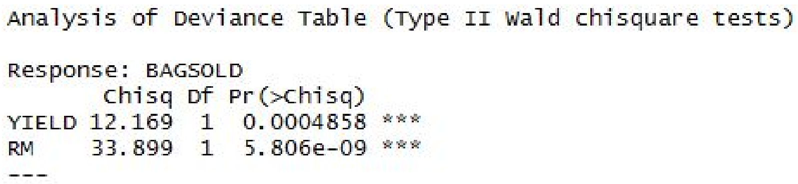
Figure 3. The final varieties that were selected and their corresponding predicted number of bags sold in the following year 2015-2016.

The final mixed effects model had the fixed effect terms Yield and Relative Maturity with the random effects Location and Family. With the best model in terms of prediction selected by comparing all the options for fixed and random effects utilizing the likelihood ratio test and thinking about the blocking effects, it was used to predict the number of bags sold for the varieties selected from the deep learning model.



```
Linear mixed model fit by REML ['lmerMod']
Formula: BAGSOLD ~ YIELD + RM + (1 | FAMILY) + (1 | LOCATION)
   Data: train

REML criterion at convergence: 613302.4

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.3381 -0.0326  0.0059  0.0636  4.4658

Random effects:
 Groups   Name        Variance  Std.Dev.
 LOCATION (Intercept) 1.723e+08  13126
 FAMILY   (Intercept) 1.004e+11 316925
 Residual             8.425e+10 290261
Number of obs: 21894, groups:  LOCATION, 150; FAMILY, 80

Fixed effects:
            Estimate Std. Error t value
(Intercept) -27677.5    47917.6  -0.578
YIELD          625.1      179.2   3.488
RM           66703.9    11456.7   5.822

Correlation of Fixed Effects:
      (Intr) YIELD
YIELD -0.187
RM    -0.631 -0.046
```

Figure 4. Summary output of the final resulting mixed model generated in R to predict the bags sold.



```
Analysis of Deviance Table (Type II Wald chisquare tests)

Response: BAGSOLD
       Chisq Df Pr(>Chisq)
YIELD 12.169  1  0.0004858 ***
RM    33.899  1  5.806e-09 ***
---
```

Figure 5. Output of the Wald Test for the final model

After the final model was selected, a Wald Test was conducted which allows one to see how confident one can be of the estimate of the effect of yield and relative maturity on bags sold, and the p-value shows that they are very good in doing so.

Conclusions and Further Research

While the hypothesis' objective was not met in terms of the AUC, it was very close to the baseline set at 75% AUC. In context, the AUC was a relatively successful measure for this deep learning model considering that climate effects and other standardization efforts were not made in the dataset provided as opposed to other studies such as by Johnson in 2014. Additionally, the predictions for the number of bags sold was done utilizing the model that was identified to be the best in terms of being able to model the levels of the problem including at the individual variety, family, and location levels. The predictions didn't seem to deviate very far from the other varieties in the same family which provided a somewhat safe or conservative prediction in that regard. This approach could be improved upon by using unsupervised clustering approaches to identify trends in the data or provide insights into underlying patterns between varieties or families that might have gone undetected in this research.

References

[1] N. Heslot, H.-P. Yang, M. Sorrells, and J.-L. Jannink, "Genomic Selection in Plant Breeding: A Comparison of Models," Alliance of Crop, Soil, and Environmental Science Societies Crop Science, Jun. 2011.

[2] J.-Q. Zhang, L.-X. Zhang, M.-H. Zhang, and C. Watson, "Prediction of Soybean Growth and Development Using Artificial Neural Network and Statistical Models," Acta Agronomica Sinica, vol. 35, no. 2, pp. 341–347, Feb. 2009.

[3] M. Kaul, R. L. Hill, and C. Walthall, "Artificial neural networks for corn and soybean yield prediction," *Agricultural Systems*, vol. 85, no. 1, pp. 1–18, Jul. 2005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0308521X04001398. Accessed: Dec. 15, 2016.

[4] D. M. Johnson, "An assessment of pre- and within-season remotely sensed variables for forecasting corn and soybean yields in the United States," Remote Sensing of Environment, vol. 141, pp. 116–128, Feb. 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0034425713003957. Accessed: Dec. 15, 2016.

[5] P. Useche, B. L. Barham, and J. D. Foltz, "Integrating technology traits and producer heterogeneity: A Mixed-Multinomial model of genetically modified corn adoption," *American Journal of Agricultural Economics*, vol. 91, no. 2, pp. 444–461, Jan. 2009. [Online]. Available: http://ajae.oxfordjournals.org/content/91/2/444.short.