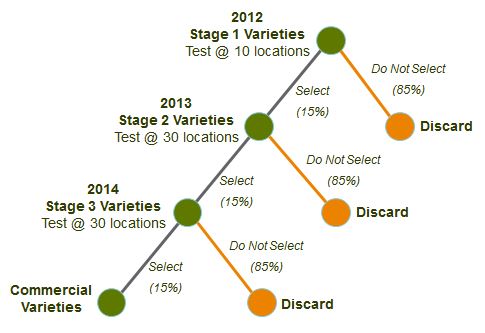
**FINAL PROBLEM STATEMENT**

Syngenta, one of the foremost seed biotech companies in the world, is Host Sponsor for the 2017 INFORMS OR & Analytics Student Team Competition. Syngenta executives seek analytics help in determining which soybean varieties to commercialize; these are critical decisions that help ensure Syngenta products perform to customers’ expectations. Below is the final problem statement.

**Syngenta: Better Decisions in Commercializing Seed Varieties**

*Have you ever made a purchase decision based on good data and sound research, only to find that the item you selected did not perform as expected? How could you have used the available information differently to make a better choice? Once you have that figured out, help us to use our data differently to make better decisions to ensure that our products perform to our customers’ expectations.*

**The Problem**



To become a commercial soybean variety, an experimental variety must pass through a series of “stage gates” (Figure 1).This is analogous to a student earning a passing grade in a course of study. Each year, after the data from the yield tests are analyzed, the decision is made to continue testing the variety (pass) or to discard it (fail). The final stage gate is the decision to commercialize the variety. The year that the decision to commercialize is made is analogous to its “graduation year”.

The example in Figure 1 shows the testing and selection scheme for the “graduating class” of 2014. In this example, several hundred experimental varieties were evaluated at up to 10 locations in 2012. After the experiments were harvested and the data were analyzed, approximately 15% of the varieties were selected to advance to the next year of testing, while the remaining 85% of the varieties were discarded.

In 2013, the selected varieties were evaluated at up to 30 locations with the top performing 15% of varieties being selected for the final year of evaluation.

Following testing in 2014, the top performing 15% of varieties graduated to commercial status and were sold the next year to growers as soybean varieties.

After commercialization, the success of a variety is defined by performance in growers’ fields, which is correlated to the number of bags of seed that are sold. Although we would like to think that all newly commercialized varieties are great, the reality is that we have some type 1 errors; soybean varieties that should not have been commercialized. Identifying these underperforming varieties prior to commercialization will help ensure that commercial varieties will perform as expected, as well as help us to use our resources more efficiently.

**The Challenge**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Class of 2011 | Class of 2012 | Class of 2013 | Class of 2014 |
| First Year Yield Test | 2009 | 2010 | 2011 | 2012 |
| Second Year Yield Test | 2010 | 2011 | 2012 | 2013 |
| Third Year Yield Test | 2011 | 2012 | 2013 | 2014 |
| Sales Volume of Graduates | 2012-2013 | 2013-2014 | 2014-2015 | Predict 2015-2016 |

Imagine that it is November of 2014 and that you are responsible for selecting varieties for commercial release. You have data from the current testing year (2014), as well as variety performance data from previous years. Given all of the available data, your challenge is to predict the potential sales volume of each variety tested in the final year of the class of 2014. In addition, you should identify the individuals from the 2014 class that should graduate to commercial sales. You may select no more than 15% of the varieties tested for commercialization. Keep in mind that the objective is to only select elite varieties for commercialization, so the number may be smaller than 15%.

**Training Data**

To develop a model to predict sales potential for the varieties within the class of 2014, you will be provided with data from the experiments associated with the classes of 2011, 2012, and 2013 as well as data from the experiments where experimental varieties from the same families were tested. In addition, participants will be provided with the sales volume for the graduates of each class by year.

**Data Description**

Each participant will receive two files. File and column descriptions are as follows:

* **EXPERIMENT DATA** – contains all of the data for the challenge.
  + YEAR – the year that the experiment was conducted.
  + EXPERIMENT – Experimental varieties of similar relative maturity are tested together in experiments. In the first year of yield testing, experiments often contain closely related experimental varieties, with the goal of selecting the best representatives of a family. In the second and third year of testing, varieties from different families are tested together to determine which varieties will be advanced to the next year or commercialized. In addition to the experimental varieties, designated “check” varieties are contained in the experiments for comparison.
  + LOCATION – Experiments are grown at many locations, depending on the stage of testing.  Individual varieties may respond differently to different sets of environmental conditions.  One of the reasons that varieties are tested over multiple years is to see how varieties will respond to a larger population of environments.  For the purpose of this challenge, we are assuming that the yield trial locations are representative of the market that the varieties will be sold in.  You may, however, find that some testing locations are more predictive than others as to the future performance of a variety.
  + VARIETY – the designation of the individual variety that is being evaluated in the experiment. From a botanical perspective, a variety is group of soybean plants that are genetically identical. They are selected for characteristics that are desirable to a grower (yield and agronomic traits). The seeds harvested from a soybean variety will be genetically identical from one generation to the next.
  + FAMILY – identifies the “breeding population” from which a variety was derived. Members of a breeding population are highly related to each other since they are derived from the same parents. Many representatives from a breeding population are typically tested together every year with the goal of selecting the best representative of the population.
  + CHECK – commercial soybean varieties that are used as performance benchmarks in yield trials. Check varieties are typically elite commercial varieties that are used as benchmarks to measure experimental variety performance. Since the check varieties are already being sold, an experimental variety needs to outperform the check varieties to be considered to move to the next stage of testing. After an experimental variety graduates to commercial, it may become a check in the following years.
  + RM – Soybean Relative Maturity – Soybean varieties are affected by day length throughout the growing season. Day length triggers soybean plants to produce seed during the summer and to mature in the fall. Soybean varieties are assigned a relative maturity number (e.g. 2.5) which reflects differences in amount of time it takes individual varieties to reach physiological maturity. For example, a 2.5 RM variety matures relatively later than a 2.1 RM variety. Historical data show late maturing varieties have greater yields than early maturing varieties, so it is important to account for this effect.
  + REPNO – replication number. Soybean yield experiments are typically replicated. Data from the individual replicates are included in this dataset.
  + YIELD – the amount of grain per unit of land that a soybean variety produces. Grain yield in soybeans in the United States is measured in bushels per acre (bu/a).
  + CLASS\_OF – the final year that a soybean variety is tested prior to commercialization.
  + GRAD – varieties that graduate to commercialization following their final year of experimental evaluation.
  + BAGSOLD – the number of bags of seed sold in the second year after commercialization. High relative sales volume in the second year of sales is associated with the superiority of a variety relative to other choices in the marketplace.
* **EVALUATION SET** – lists the experimental varieties in the class of 2014. These are the varieties to be evaluated in the challenge (see above). The columns are as follows (see above for descriptions):
  + CLASS\_OF
  + VARIETY
  + FAMILY
  + RM

**Evaluation Criteria**

The challenge is to identify patterns in the data that identify elite experimental varieties and expose the non-elite varieties prior to commercialization. Entries will be judged by the clarity of the solution, the technical strength of the methodology, the uniqueness of the approach, and the degree to which the data support your conclusions.

**Glossary of Terms**

**Advancement** – the decision to evaluate a variety to the next stage of testing or commercialization.

**Agronomic traits** – characteristics of the variety other than yield that are valued by soybean growers.  The ability of a plant to tolerate diseases, the ability to remain upright during periods of inclement weather and ease of harvest can all be considered agronomic traits.

**Commercial variety** – soybean varieties that are being sold in the marketplace.

**Commercialize** – the decision to begin selling a variety.

**Elite** – commercial varieties that have relatively large sales volume during their lifetime. For the purpose of the challenge, elite varieties account for a minimum of 1 million bags sold in the second year of sales.

**Experimental variety** – a non-commercial soybean variety that is being evaluated in yield trials.

**Grower** – farmers or farm managers that select, purchase, and plant commercial soybean varieties.

**Performance** – the amount of grain per unit of land that a soybean variety produces.  Grain yield in soybeans in the United States is measured in bushels per acre (bu/a).

**Selection** – the act of choosing a variety for advancement or commercialization.

**Stage** – the current testing level of a variety.  It is analogous to grade levels in school.

**Stage gate** – the requirements of a variety to qualify to a higher stage of testing or commercialization.

**Type 1 error** – a false positive. In this case, it refers to advancing or commercializing a variety that does not actually deserve to advance.

**Yield Test** – the experiment in which an experimental soybean variety is grown where grain production per variety is the primary characteristic of performance.

\_\_\_\_

If class is blank, grad will be blank. Bags sold = 0.

If grad = NO, then bags sold = 0.

Varieties without a class DO NOT appear in those with classes. However, families without a class may appear in those with classes.

If Class of 2011, there are only Check = TRUE and no Check = False for years 2012, 2013 and 2014 (years AFTER graduation). The only check = TRUE appears in the years after graduation. But for class of 2014, CHECK=TRUE for 2014’s experiments.

Experiments before the first yield test – Grad = NO.

A variety can be in the class of 2014 but not have any experiments in 2014 🡪 kicked out before that

534 experiments in total: 534 with check is true, 512 with check is false.

Panel data: Grad, family, RM, bags sold will be the same for each variety.

Varieties of checks vs non checks – mostly no overlap

Possible plots: grad, check