NewNKSoy Project Proposal

Patrick Nicholson (pkn7) and Michael Ko (mk872)

September 23, 2016

Commercial soybean selection, the process by which Syngenta chooses elite soybean varieties from a large set of experimental varieties and ultimately commercialized, 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. Despite such a thorough selection process, there have been soybean varieties which were commercialized in the past that have performed poorly. There is no good reason for this to be happening; non-elite varieties should not be taking the spot of well-performing ones. We believe that, using data analysis on some large data sets related to the soybean selection process and past soybean variety commercial performance, we can help minimize these selection mistakes. The overall objective is to create a model that will help us select the soybean varieties that will sell the most bags of seed once commercialized.

It is fairly clear that this project deals with an issue of utmost importance to Syngenta. Seed biotech is Syngenta's main industry, after all, and any seed biotech company relies on selling seed in order to be profitable. Thus, this project could help improve Syngenta's profitability in two ways; first, by lowering production costs through helping prevent non-elite soybean varieties from being commercialized, and second, by increasing revenue through helping the company better recognize and focus on truly elite soybean varieties. Potential side effects of this project, besides improving our ability to predict seed sales themselves, could be additional insights into Syngenta's testing practices. For example, in fitting this model, we may find that certain tests are more indicative of soybean success than others; in cases where we find little to no correlation between a certain test or experiment and soybean sales, Syngenta could simply do away with that test, thus reducing research and development costs. Regardless of the specific results, there are many ways in which this project will benefit Syngenta, both in a purely profits-focused sense and also in terms of providing new insights to Syngenta's soybean selection process.

While we cannot promise success with absolute certainty for this project, there are many aspects of the data available to us which suggest success is certainly attainable. The dataset we are working with contains experimental data from the soybean varieties that were tested for commercialization in the years 2011, 2012, and 2013, along with the sales volume for those that were commercialized. The data contains many features which lend themselves well towards classification, such as soybean family, relative maturity and variety, and experiment year, location and type. This provides us opportunities to try both models focusing on classification of the soybean varieties and those focusing on analysis of time series data. Having examples of past varieties that were both successful and unsuccessful over a three year period should provide a sufficiently large training set for fitting models on different samples, cross-validation, and ultimately predicting accurately on our test data of 2014 varieties. Whether variety success is tied mostly to the variety's characteristics, its performance on certain experiments, a combination of the two, or something slightly different, this data set should provide us the opportunity to use the tools necessary to succeed in this project.