Submitted to Syngenta Crop Challenge 2018

FOR INTERNAL USE ONLY (leave blank) Submission Number:

Team Number:

INSTRUCTIONS TO AUTHORS

• Use this template in either the .doc or .tex with style file format.

• Do not remove the INTERNAL USE ONLY section

• Please do not change the font size, margins, or headers

• Unless otherwise requested, do not include personal names or information

• Failing to do any of the above may result in a delay in evaluating your submission

Performance Predictions of Hybrid Corn Varieties in Untested Locations

Arindam Bhattacharya, Madhuri Gupta,

Jeff Parker

2311 N Campus Drive, Suite 1400, Evanston, IL 60208, USA

arindambhattacharya2017@u.northwestern.edu; madhurigupta2017@u.northwestern.edu;

jeffreyparker2017@u.northwestern.edu

You should replace this paragraph with your abstract. As we know, the challenge is to develop a quantitative framework for predicting hybrid performance in new, untested locations. You will first need to build and validate predictive models using observations from previous years. You will then use this model to predict the performance of a set of hybrids tested in 2017, without precisely knowing the environmental conditions. Following the standards for an academic publication, the entry should also provide the predictions of performance for 2017 test hybrids at given locations in the test dataset, a clear description of the methodology and theory used, and appropriate references. In the abstract, please briefly summarize your prediction result and key points.

*Key words*: Key words summarizing the main techniques used in your approach, for example: regression,

supervised learning, neural networks

1. Introduction

As corn is one of the most important crops in the United States, it is critical for the agriculture industry to innovate on new crop varieties. New varieties of corn should enhance the desirably qualities: durability in droughts, water consumption, required nutrients from the soil, resistance to pests, taste and of course yield produced. New varieties are developed by seed researchers by cross-breeding two existing varieties to create a hybrid. Hybrid varieties share genetic makeup in the same way a child shares the genetic makeup of his parents. However, cross-breeding is very costly to test in terms of both time and expense. Test hybrid varieties must be planted in selected test fields, then mature over the full plant life cycle – usually about X months. Then the yield must be measured and compared to the opportunity cost of growing an empirically successful variety. The experiment set-up takes additional care and oversight on the part of the farmer. Additionally, yield performance is highly variable and dependent in part on the weather, season and soil composition. A potentially high-yielding variety might never be identified if it is planted in a field with a particular soil composition or if the weather conditions not ideal. Finally, much like brothers and sisters are all different, hybrid varieties will have different genetic makeup from the parents and thus be slightly different with every breeding. Despite the randomness in the gene selection, cross-breeding has proved to be more digestible to the consumer than genetic modification.

With all the costs and variability in cross-breeding experimentation, forward thinking organizations are looking to data-driven methods to predict outcomes of experimental varieties under the highly variable conditions. To this end, the Syngenta Data Challenge seeks to create a data framework to predict hybrid yields performance over a competitive benchmark. To make these predictions, a framework can be broken into three sub-categories: 1) Performance, 2) Weather and 3) Genetic Makeup. Using time series machine learning techniques, weather variables were predicted in a roll-forward manner. The hybrid genetics variables were reduced into clusters. Soil and weather variables were also reduced into few variables using unsupervised learning techniques. Using location, the reduced soil variables (soil remains constant over time) and the reduced weather predicted variables models were developed to predict the yield and benchmark yield. Thus the yield difference could be calculated. Instead of predicting the yield difference as a continuous variable, predictions were made on a binary variable indicating whether or not the hybrid outperformed the benchmark. Then an additional model was trained to measure the magnitude of the performance for only those varieties that were predicted to outperform the benchmark.

As corn is one of the mostIn the introduction, you should include an overview of the approach. Throughout the document, appropriate citations, such this imaginary website of the American Mathematics Institute (2005) and figures, such as Figure 1, should be used. For pagination reasons, figures and tables may appear on different pages as does Figure 1. Remember, the entries will be evaluated mainly by how well predicted performance aligns with observed hybrid performance for 2017.

Bhattacharya, Gupta, Parker: Hybrid Corn Variety Performance Predictions

2 Article submitted to Syngenta Crop Challenge in Analytics Award

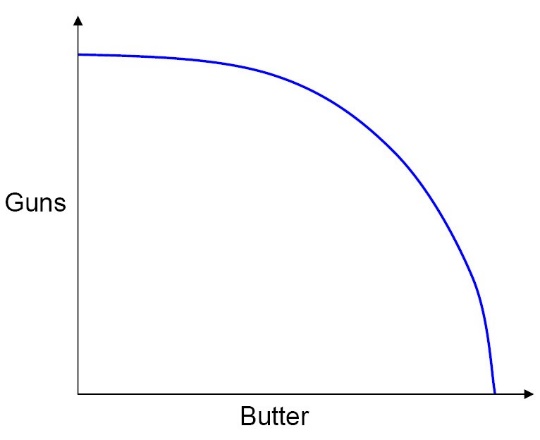


Figure 1. Production Possibilities Frontier.

In addition, criteria like novel ideas used to create predictive model, simplicity of the solution, evaluation of factors included in the decision process, clarity in the explanation and the quality and clarity of the finalist’s presentation at the 2018 INFORMS Conference on Business Analytics and Operations Research will be considered, as well. It is vital that you document your methodology in sufficient detail for evaluation.

2. Analysis of data

In this section, you should state any analysis you conduct before predicting the result. The analysis could include but not limit to the relation between features, the distribution of the features, the distribution of the target and any other necessary preliminary analysis. Figures, tables may be useful to communicate your thinking.

3. Methodology and theory

Present the methodology and theory of your approach. Charts, diagrams, flowcharts or other visualizations may be useful to communicate your thinking. Including discussion and considerations, such as the assumptions that were made, feature selection on weather, soil and genetic information, the scope, and early considerations, may provide a useful framing of your solution. Given the complexity and practical aspects of the problem, background information may be useful to include or reference. It is vital that you document your methodology in sufficient detail and clarity that it can be understood and evaluated.

Bhattacharya, Gupta, Parker: Hybrid Corn Variety Performance Predictions

Article submitted to Syngenta Crop Challenge in Analytics Award 3

4. Quantitative results

In accordance with your methodology, present quantitative results that justify your prediction result. You can use the validation dataset for 2016.

5. Team members

List all the team members associated with the submission, including the Corresponding Author. This section is required if there is more than one person on a team. A team’s solution should be submitted once (as opposed to each member of the team submitting the same solution individually).

• Team Member One (aka Corresponding Author), Address of Corresponding Author, Corre[spondingAuthor@gmail.com](mailto:spondingAuthor@gmail.com)

• Team Member Two, Address of Team Member Two, [TeamMemerTwo@company.com](mailto:TeamMemerTwo@company.com)

• Team Member Three, Address of Team Member Three, [TeamMemb](mailto:TeamMemberThree@nonprofit.org)[erThree@nonprofit.org](mailto:erThree@nonprofit.org)

6. Exogenous data sets (optional)

Since the data sets provided in the Challenge have geographic coordinates, researchers have the option to use additional geo-referenced data sources (i.e. ISRIC, VegScape, and Drought Monitor among others). Any additional datasets used must be available for public use and properly cited. In this section, please list any exogenous data used, along with a reference.

• Exogenous Data Set One: reference or publically available website to access the data set

• Exogenous Data Set Two: reference or publically available website to access the data set

• Exogenous Data Set Three: reference or publically available website to access the data set

7. Supplementary materials (optional)

The description of the entry in the submission template should be self-explanatory. If you upload supplementary files along with your submission, please provide a description of the files here.

• Supplementary File One Name: file one description

Corresponding Author: Your Short Title

4 Article submitted to Syngenta Crop Challenge in Analytics Award

• Supplementary File Two Name: file two description

Acknowledgments

Necessary acknowledgements to funding agencies or others should be made in this section. This section is optional.

References

American Mathematical Institute (2005) Better predictors of geospatial variability. Retrieved June 14, 2005, [www.mathematicsinstitute.org.](http://www.mathematicsinstitute.org.)