*Submitted to Syngenta Crop Challenge in Analytics Award*

FOR INTERNAL USE ONLY (leave blank)

Submission Number:

Team Number:

Boosted Yield Prediction

Edima Udo

76 Westgate Blvd, Toronto, Ontario, M3H 1P2, Canada, edimaudo@gmail.com

You

1. Introduction

Commercial corn is processed into multiple food and industrial products. It is widely known as one of the world’s most important crops. Each year, plant breeders create new corn products, known as experimental hybrids, by crossing two “parents” together. The parents are known as inbreds and the development of the inbreds takes up the bulk of a corn breeding program. Most of that effort is spent evaluating the inbreds by crossing to another inbred, called a “tester.”

It is a plant breeder’s job to identify the best parent combinations by creating experimental hybrids and assessing the hybrids’ performance by “testing” it in multiple environments to identify the hybrids that perform best. Historically, identifying the best hybrids has been by trial and error, with breeders testing their experimental hybrids in a diverse set of locations and measuring their performance, then selecting the highest yielding hybrids. The process of selecting the correct parent combinations and testing the experimental hybrids can take many years and is inefficient, simply due to the number of potential parent combinations to create and test.

The objective of the challenge of the is to estimate yield performance of the cross between inbred and tester combinations in a given holdout set. Specifically, we are asking for the mean yield performance of each inbred by tester combination in the holdout set. The measure of success is accuracy of the predicted values in the test set based on root mean squared error.

To achieve this objective multiple machine learning models model were evaluated to see which one would produce the most accurate model based on the data provided.

1. Methodology and Theory

The approach to solving the problem involved:

- Reviewing the data dictionary provided

- Performing exploratory data analysis

- Model development

- Model Selection

**Dictionary Review**

The dataset contains the observed yield (consistently scaled to an internal benchmark) for a large set of corn hybrids tested across multiple environments between 2016 and 2018. These hybrids are created through the crossing of 593 unique inbreds and 496 unique testers. Creating a two-way table of means with inbreds as rows and testers as columns results in a data table with approximately 96% missing values. Each row contains the year and location ID of the observation. Additionally, each row includes a cluster value for each inbred and tester. This represents the genetic grouping of the inbreds and testers and has been determined using internal methods. Inbreds and testers are not treated any differently when clustering, so a shared cluster value indicates genetic similarity regardless of whether a parent is defined as an inbred or a tester.

The dataset has seven columns which include:

YEAR: Year grown

LOCATION ID: for each location

INBRED ID: for Inbred

INBRED\_CLUSTER: Cluster association for each inbred which denotes genetic grouping

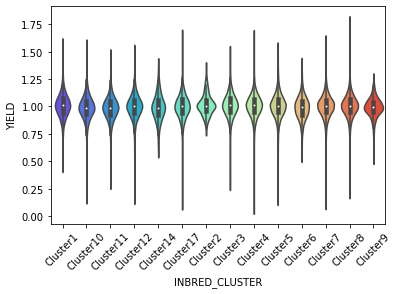
TESTER ID: for Tester TESTER\_CLUSTER Cluster association for each tester which denotes genetic grouping

YIELD: The performance of the Line and Tester combination.

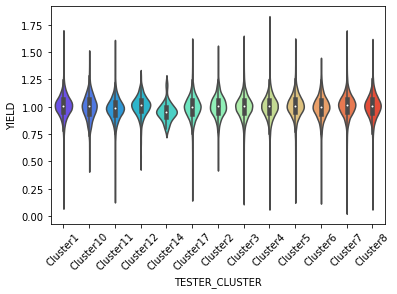
The test dataset does not have the LOCATION ID and YIELD columns.

**Exploratory Analysis**

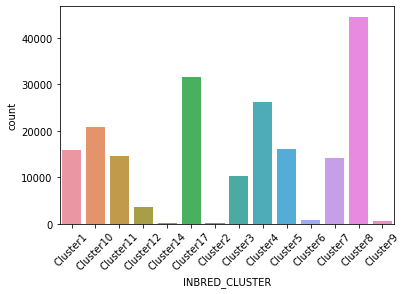
This involved taking a deep dive into the data and looking for patterns and insights. The training data contains 199476 rows. The test data has 1200 rows. The inbred data has 14 clusters while the Tester data has 13 clusters. The yield variance with regard to the inbred cluster is shown in figure 1. There is a large concentration around 1. Cluster 4 has the largest variance. In figure 2 it shows the yield variance with regards to the test cluster. Cluster 4 has the largest variance. Figures 3 and 4 show that Cluster 17 is the most predominant cluster in both the inbred and test clusters.



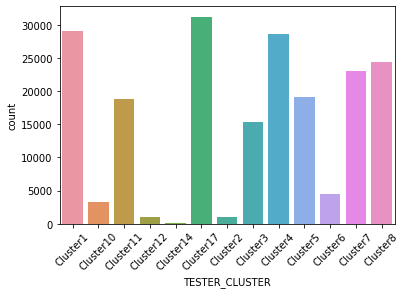
**Figure 1: Inbred cluster and Yield**



**Figure 2: Tester cluster and Yield**

****

**Figure 3: Inbred cluster Count**

****

**Figure 4: Tester cluster Count**

**Model Development**

1. Quantitative Results

In accordance.

1. Conclusion

Provide a summary of key findings.

1. Team Members

Edima Udo, 76 Westgate Blvd, Toronto, M3H 1P2, edimaudo@gmail.com

1. Supplementary Materials (Optional)

Supplementary

1. References

Syngenta. 2019. Submission template example. *Journal of Syngenta Research*. 1(1)