The Data Mine Corporate Partners Symposium 2025

SYNGENTA: PREDICTIVE PRODUCT PLACEMENT

Team Members: Benny Zheng, Chatanya Sarin, Isabelle Lee, Mali Shaltouki Rizi, Rui Meng, Seth DeWhitt, Shlok Sheth, Amanda Huang, Yuehuan Fang

INTRO

Strategically positioning vegetable products in the most suitable markets is crucial for product success. The performance of products is influenced by a combination of genetics (G), environment (E), and management practices (M). This project aims to integrate machine learning models with GxExM to predict product placement.

BACKGROUND

- Sweet corn is a variety of maize grown for human consumption with a high sugar content, but population growth and climate change demand more stable, high-yielding varieties
- Genotype-by-environment (G×E) interaction affects performance under different conditions, making stability a desired trait in breeding
- Genomic prediction (GP) accelerates the process by predicting genetic potential, reducing breeding cycles, and increasing efficiency of selection. This makes possible the production of climate-tolerant, high-yielding sweet corn varieties

MATERIAL & METHODS

Data

1. **Phenotypic data:** Traits: field yield green weight for 1068 Hybrids, 41 Locations, 5 years
2. **Genotypic data:** Molecular markers for 18,000 of both male and female parents
3. **Environmental data:** Raw environmental data from all locations

Data Quality Control

1. **Phenotypic Data QC:** Remove entries with missing/zero values, exclude unreliable data, Utilize a linear mixed model to create a standardized residual graph to detect and remove outliers
2. **Genotypic Data QC:** Markers with >20% missing values, >2 alleles, >10% heterozygosity, or a minor allele frequency (MAF) <0.5% were excluded, Parental lines with >20% missing genotypic data were removed

MODEL DEVELOPMENT

Parametric Model

These models assume data follows a certain distribution with finite parameters.

Genomic Best Linear Unbiased Prediction (GBLUP)

A model that uses a genomic relationship matrix to predict breeding values.

Model: Y=u+GCA(female)+GCA(male)+e where GCA is General combining ability of the female parent (genomic relationship matrix)

Pearson Correlation: Correlation between the adjusted phenotypic value and predicted values

Method: Data split 80:20, Cross validation

Non-Parametric Model

These models make fewer assumptions about the data distribution and can adapt to complex patterns.

Light Gradient Boosting Machine (LightGBM)

A tree-based model that captures non-linear relationships and interactions between genomic predictors. It repeatedly evaluates the prediction error and builds more trees to improve its accuracy.

Method: Data split 80:20, Cross validation

Used parameters:

- Learning rate: 0.1
- Number of iterations:100

RESULTS

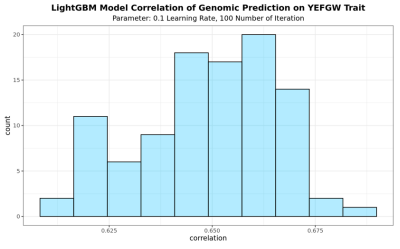

Figure 1: Distribution of Pearson correlation from running 100 trials of predictions on YEFGW (Field Yield Fresh Weight) using LightGBM.

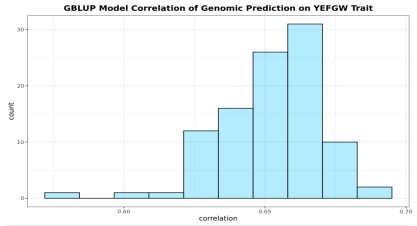

Figure 2: Distribution of Pearson correlation from running 100 trials of predictions on YEFGW (Field Yield Fresh Weight) using GBLUP.

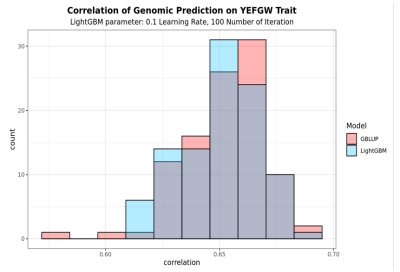

Figure 3: Comparison of Pearson correlation distribution from 100 trials of predictions on YEFGW (Field Yield Fresh Weight) between GBLUP and LightGBM.

Model	Pearson Correlation (r)	RMSE
LightGBM	0.649	1,361.45
GBLUP	0.651	1,783.82

Table 1: Comparison of Pearson correlation and RMSE between GBLUP and LightGBM.

ACKNOWLEDGMENTS

We would like to thank our TA Steven LaCroix, our mentor Dr. Jin Sherry Xiong, and the staff from The Data Mine & Syngenta for their support and guidance throughout this project!

CONCLUSIONS

- The average correlation of LightGBM and GBLUP is comparable
- LightGBM has a more stable correlation, as its correlation distribution has a smaller range, and it has a significantly less Root Mean Square error
- The small range in the correlation indicates that LightGBM has a higher precision and might be a better fit
- Increasing the learning rate of LightGBM to 0.05 or 0.01 and adding a higher number of iteration might improve the correlation significantly

FUTURE GOALS

If we were given more time, we would like to work on the following:

1. Adjust parameters in each model to optimize model performance
2. Test other models such as random forest and neural network and compare these model's performance with the ones with have
3. Evaluate other phenotypic traits, not just yield, with the existing models we have
4. Make inference about the missing phenotypic data by evaluating the other traits
5. Incorporate environmental factors as predictors along with genetic information in our prediction models
6. Develop a user interface to interact with our findings