BAYER (YIELD PREDICTION)

PURDUE UNIVERSITY | The Data Mine

TAs: Shivani Kogta, Suhaib Ahmed Khan | Students: Alizeh Naveed, Dhanush Biddala, Ethan Herrera, Hanwen Zheng, Marlon Valverde, Prateek Garg, Scott Mitchell, Shashank Mukkera, Xindi Wang

INTRODUCTION & GOAL

With a growing population, it is essential to improve crop yield to meet people's daily demands. In this project, we are attempting to predict maize yield using genotypic markers and phenotypic values by comparing each model using machine learning.

Goals:

1. identify genetic markers or environmental features that have a significant influence on the desire phenotypes using these models
2. Utilize the genetic marker information to adjust future generations to get a higher yield from their crops.

RESULTS

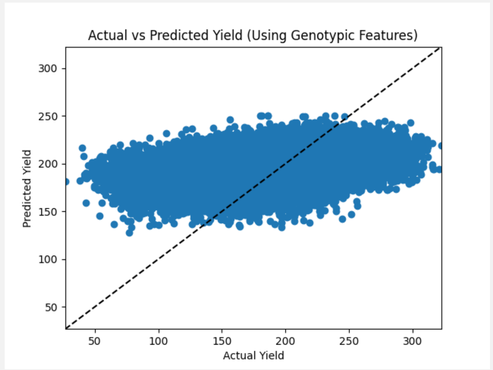

Figure 2: Scatter plot for the predicted yield (y-axis) versus the actual yield (x-axis) for the Lasso Regression Model using only Genetic Markers. (0.123 Adjusted R-squared)

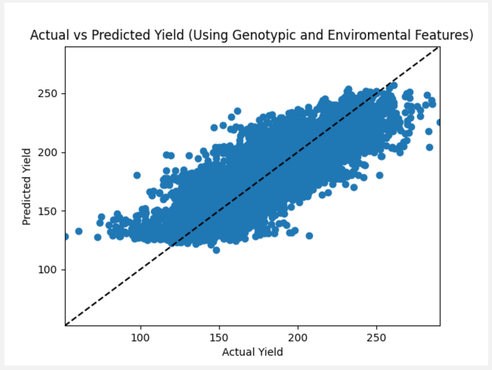

Figure 3: Scatter plot of the predicted yield (y-axis) versus the actual yield (x-axis) for the Lasso Regression Model using Genetic Markers and Enviromental Features. (0.569 Adjusted R-squared)

RESULTS

Model	Adjusted R-Squared	Mean-Squared Error	Mean-Absolute Error
Lasso	0.123	1182.976	26.699
Ridge	0.134	1067.929	25.351
Neural Net	0.141	1058.159	27.452
Elastic-Net	0.115	1082.862	25.527

Figure 5: Table showing the results of 4 regression models (Lasso. Ridge, Neural Net, and Elastic Net) alongside evaluation metrics

STEPS TO CREATE CENTRALIZED DATABASE

1	Divide the imputed genotypic markers by population
2	Join the imputed genotypic markers with the phenotypes
3	Implement the joined database into a new SQL database

Figure 1: Data preprocessing steps for increased security and a convenient source to draw out training and test sets.

BACKGROUND ON DATA & METHODOLOGY

Data Set

- Clusters comprised *male / female* heterotic groups
- Each population has many progenies with the following data:
  - Genetic markers imputed from genomic information of parents
  - Phenotypic data for each location
    - Measure of *Yield*, *Moisture*, *Plant Height* , etc.

Methodology

- Only 2% missing data after imputation
- Imputation of *"Null"* **genetic markers** values with *"0"*
- Filling **phenotypic** *"Null"* values with ***mean*** values
- Conversion of **CSV files** to **SQL database** *(Fig. 1)*
  - Data Security and Ease Of Access

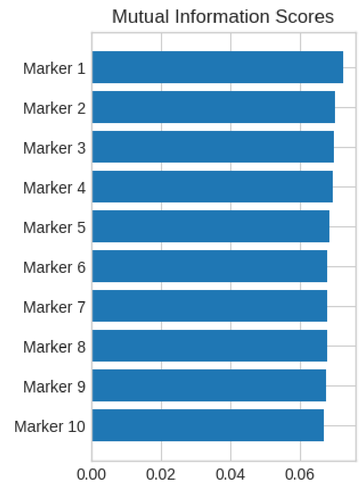

Figure 4: Mutual information scores show which genetic markers are most closely related to the yield.

FUTURE GOALS

- Make our models more accurate and applicable to larger databases.
- Explore Deep Neural Network.
- Add environmental data to the existing models to make an advanced model for more accurate prediction.
- Explore alternate feature selection.

REFERENCES

Jacobson, A., L. Lian, S. Zhong, and R. Bernardo. 2014. General combining ability model for genomewide selection in a biparental cross. Crop Sci. 54:895–905. doi:10.2135/crop- sci2013.11.0774

Acknowledgements: Sofia Brandariz Zerboni, Viveka Gorla, Brian Dilkes, Bayer and the Purdue Data Mine