Yield Prediction Through Machine Learning

PURDUE UNIVERSITY The Data Mine

Ayden Bridges, Thirawat (Tam) Bureetes, Darren Iyer, Talia Jacobson, Shiva Konakanchi, Yingjun (Link) Lin, Ankush Maheshwari, Dingyan Shang, Karnika Soni, Nischay Uppal, Yiyao (Iris) Zhang

INTRODUCTION

There were major changes in the past that impacted corn yield, such as, double-cross hybrid in the late 30s, N-fertilizer in the mid-50s, single-cross hybrid in the 60s, genetically modified organism (GMO) in the mid-90s, and most recently, gene selection in 2010[1]. Modern gene-editing technologies like CRISPR-Cas9 open possibilities for researchers and breeders to select desirable traits for a higher yield. However, environmental factors influence crop yield and growth. These factors consist of temperature, precipitation, soil composition, and others. This project aims to utilize machine learning techniques to discover interactions between corn genetics and environmental conditions that impact yield.

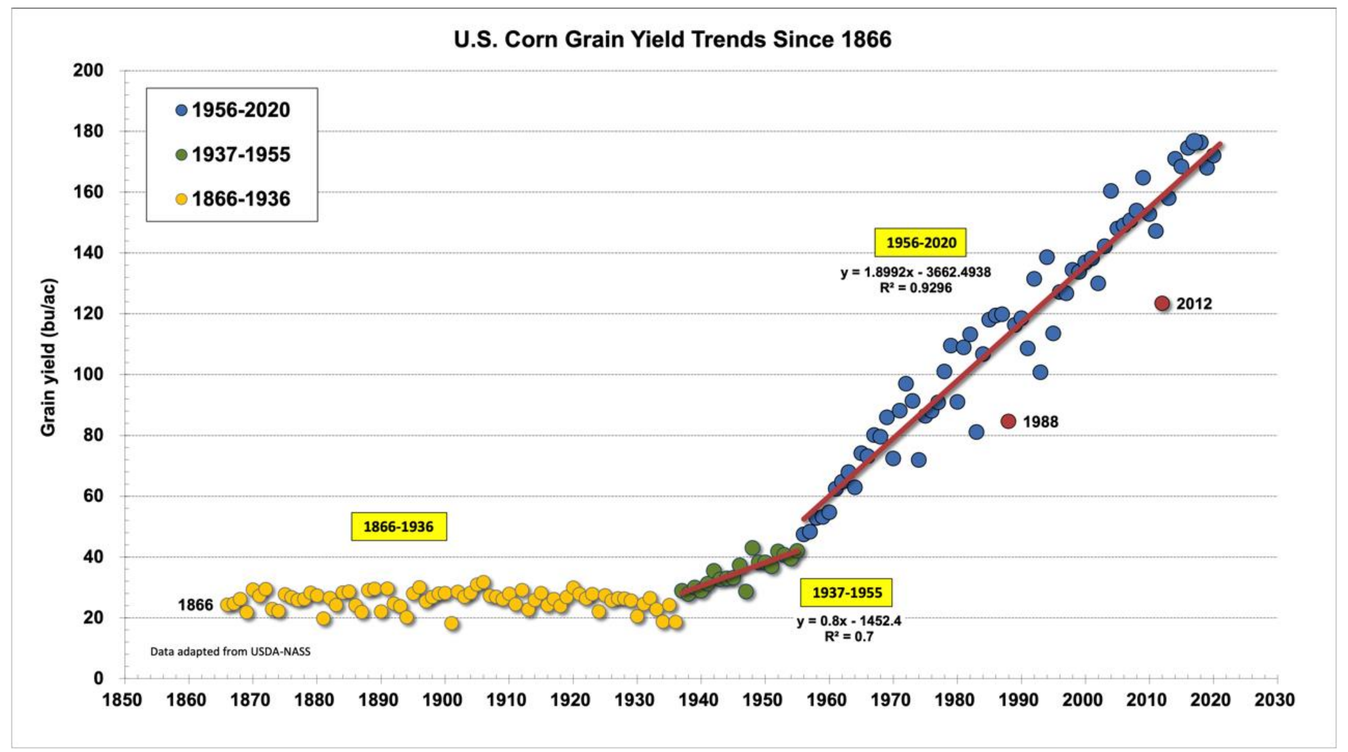

Figure 1 Historical Annual Corn Grain Yields in the U.S. since 1866. This data was derived from annual USFA-NASS Crop Production Reports.

RESULTS/CONCLUSIONS

Our team tested a variety of Machine Learning models. According to our results, it was found that the most accurate model was Lasso Regression. This model was able to predict yield based on genetic markers and environmental factors with an R-squared value of 0.256.

According to coefficients of the different models we tested, the environmental features that had the greatest impact on yield: Precipitation, Location, Temperature, Soil.

Model\Matrix	Dataset	Features	R-Square	MSE
LASSO Regression	Testing Set (unseen)	90 Geno, 84 Env	0.256	1,024
LASSO Regression #2	Testing set (Unseen)	90 geno, 48 Env	0.219	1,070
Elastic Net (Alpha = 0.5)	All 500 populations	2912 Geno, 11 Pheno	0.46	-
Deep Learning Model	All 500 populations	90 Geno, 84 Env	0.022	1,450
Linear Regression	One population	2912 Geno, 84 Env	-2.3e24	1.3e27

Figure 6: Comparison of prediction metrics from our models. R-Square and MSE refer to their respective values for Predicted Yield vs Actual Yield.

RESULTS/CONCLUSIONS

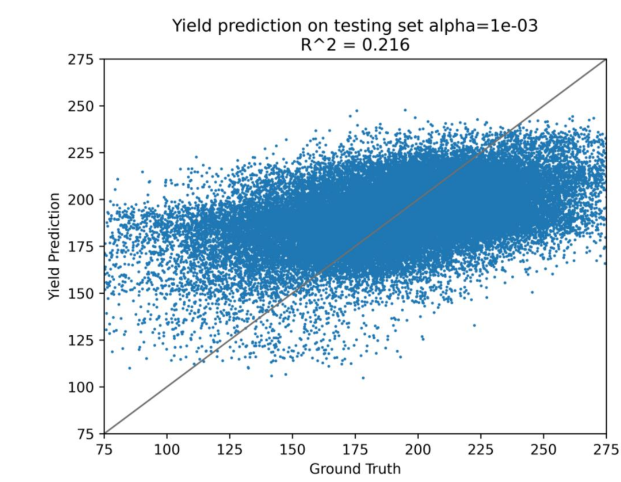

Figure 7: Yield prediction with Lasso model, trendline given for scale

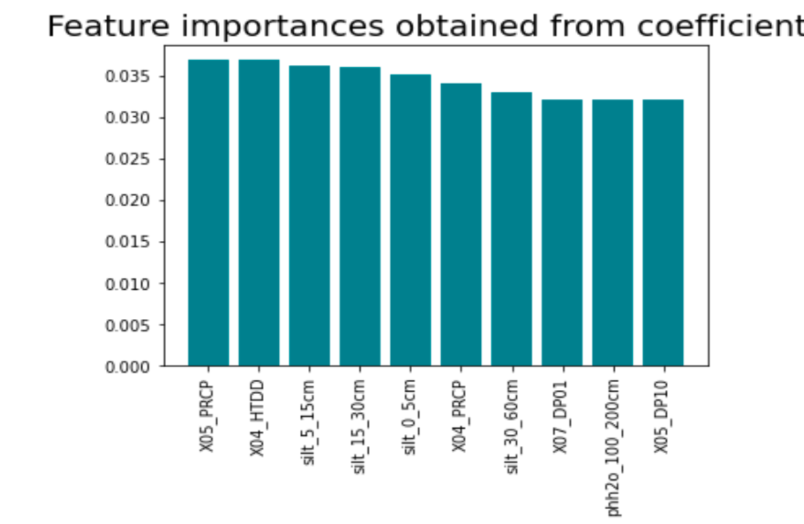

Figure 8: Histogram of the strength of each factor on the prediction

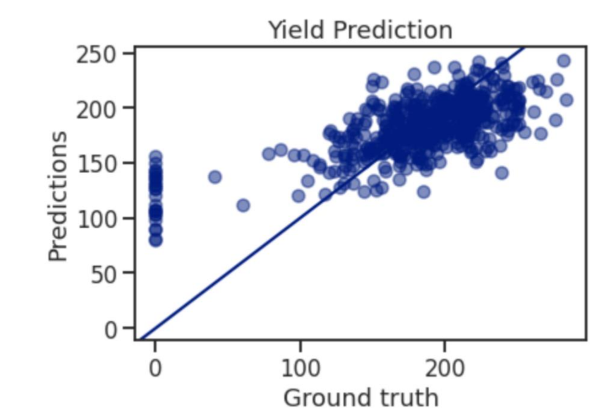

Figure 9: Yield prediction from our Elastic Net model

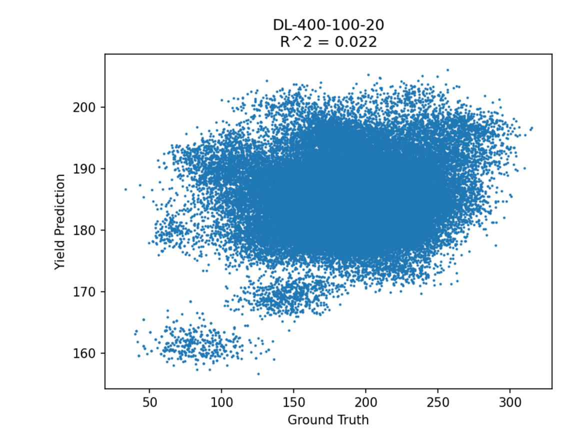

Figure 10: Yield prediction from our Deep Learning model

METHODOLOGY

DATASET

- Data was divided into two clusters with inbred lines bred as either male or female
- Collected from field trials across 18 states in the US between 2000 – 2008
- Phenotypic data collected includes yield, plant height, estimated relative maturity, etc.
- Genetic information includes genotyping of a list of important markers

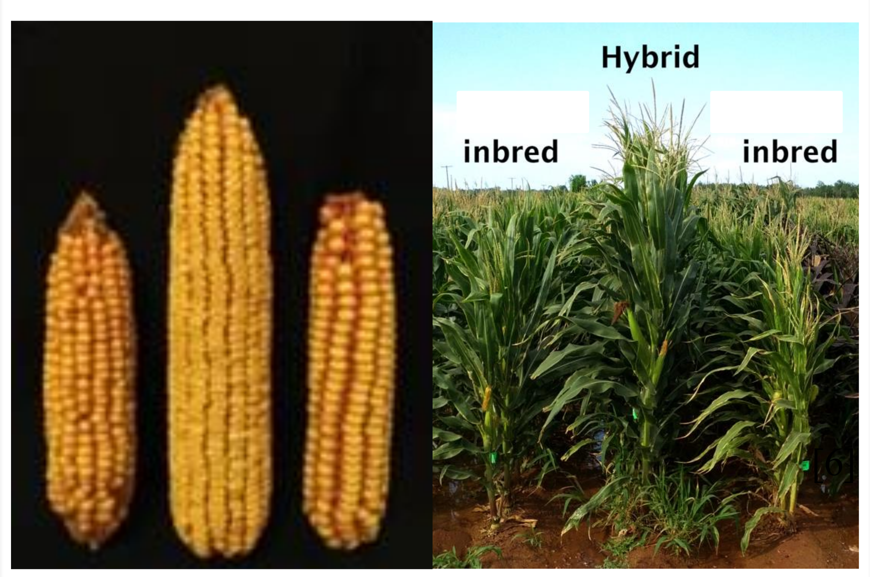

Figure 2 Inbred plants bred together to produce a hybrid plant with increased ear and plant height [3]

DATA PROCESSING

- Provided data was filtered into SQL databases for easier organization
- SQL data is difficult to modify and easy to add to using the established schema
- Having the data centrally located will make it easier for the next team as well

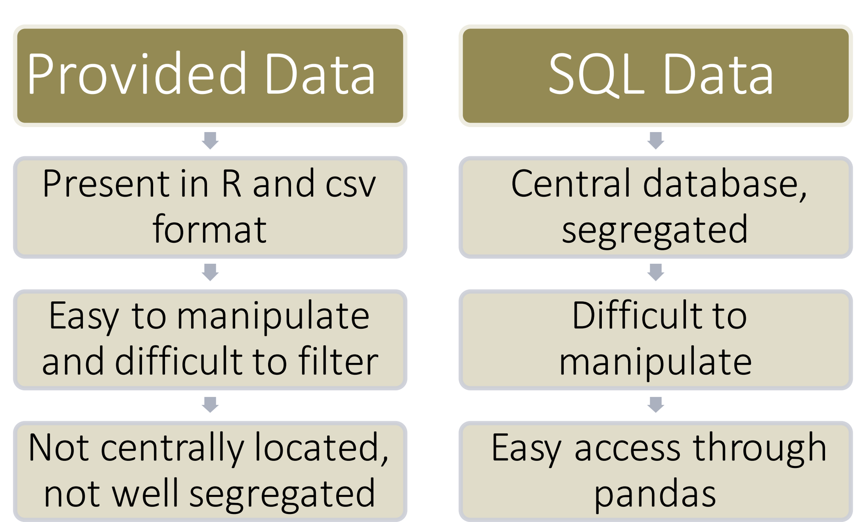

Figure 3 Data conversion chart from .csv to SQL

METHODOLOGY

PROCESS FLOW

- Genomic, weather, and soil data are utilized as features to predict yield
- Environmental data is imputed from weather stations close to crop locations through NOAA API[2]
- Missing genetic data is imputed through Beagle[4]
- Soil data is imputed from ISRIC SoilGrids API[5] and supplemented with KNN imputation

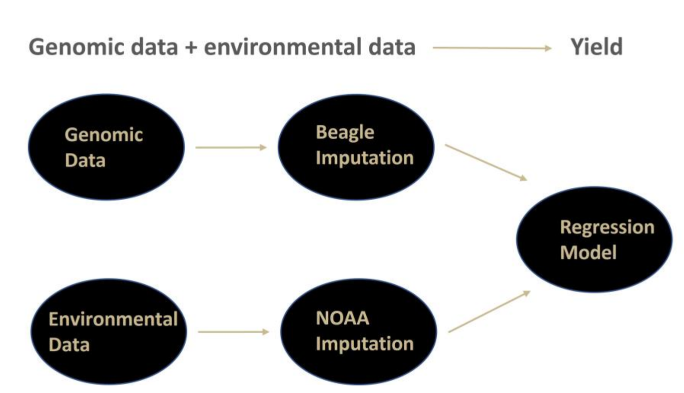

Figure 4 Imputation process for genomic, weather, and soil data

FEATURE AND MODEL SELECTION

- 90 genetic markers out of 2900 are selected based on a genomic data vs. yield study
- Environmental features are selected based on a feature multicollinearity study
- Based on literature review[7][8], LASSO regression model is selected as it might work well with data utilized in project

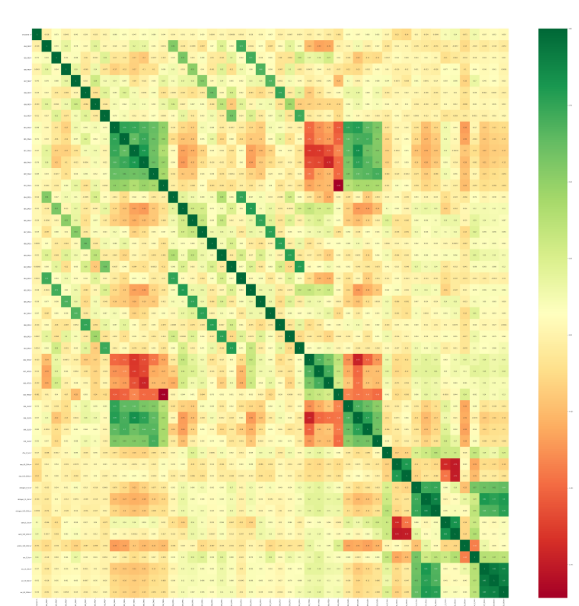

Figure 5 Correlation matrix for the features to eliminate multicollinearity issues in the model

FUTURE GOALS

- Try modeling larger and wider Deep Neural Network (DNN).
- Explore models, such as Stochastic Gradient Descent, along with custom ML models.
- Explore alternate feature selection.

ACKNOWLEDGEMENTS

- We would like to express special thanks to our mentors, Adam Scott, Brian Dilkes, and Sofia Brandariz, who guided us through the semesters and explained various concepts that helped us with our research and implementations

- We also would like to extend our gratitude to David Glass, Kevin Amstutz, and the rest of the Data Mine staff for providing us with technical assistance and helping us stay on track with our tasks

REFERENCES

[1] R.L. Nielsen, https://www.agry.purdue.edu/ext/corn/cornguy.html
[2] NOAA API, https://www.ncdc.noaa.gov/cdo-web/webservices/v2
[3] Texas A&M, Texas A&M releases new corn lines for use in commercial hybrids (tamu.edu)
[4] Beagle, https://faculty.washington.edu/browning/beagle/beagle.html
[5] SoilGrids API, https://faculty.washington.edu/browning/beagle/beagle.html
[6] DataMine Bayer Report, 2020-21
[7] Shahhosseini et al., https://doi.org/10.3389/fpls.2020.01120
[8] Klompenburg et al., https://doi.org/10.1016/j.compag.2020.105709