Bayer – Yield Prediction

Andrew Paul, Aryaman Sharma, Aryan Kaul, Jae Man Yim
Joon Hong Park, Lawrence Liu, Lucas Mori, Saksham Singh | TA: Prateek Garg

PURDUE UNIVERSITY The Data Mine

INTRODUCTION

By 2050, we're on track to have 10 billion people on our planet. That's a lot more food we'll need without much room to grow it since agriculture already uses half of our vegetated land and adds a quarter of the world's greenhouse gas emissions.

As such, it is instead more important to improve the crop yield to produce more food within the same, or less, area. This can be accomplished by identifying and growing the best possible seeds to optimize yield.

We are partnering with Bayer, a forefront innovator in the field of crop science, to investigate the application of machine learning in predicting maize yield. This research utilizes genotypic markers and environmental data, aiming to address the imminent food challenges that humanity is expected to face in the upcoming decades.

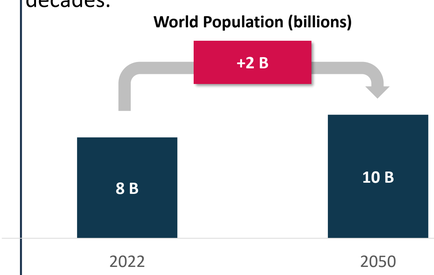

Figure 1 – World Population Projection - Source: "Creating a Sustainable Food Future" - World Resources Institute 2019

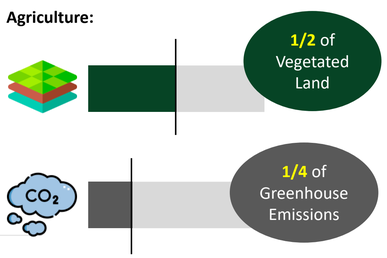

Figure 2 – Emissions/land landscape Source: "Creating a Sustainable Food Future" - World Resources Institute 2019

METHODOLOGY

Figure 3 – Main steps of our approach

Our Approach:

1. **Subset Dataset**: Worked with a smaller subset from our complete dataset by sampling 15 populations per year from cluster 1. Experimented with models including Neural Networks, Random Forest, XGBoost, LightGBM, and Linear Regression. Tree-based models outperformed others.
2. **Data Consolidation**: Merged data from both clusters and environmental data into one comprehensive dataframe.
3. **Handle Missing Values**: Filled missing genetic markers with zeros, considering but dismissing KNN imputation due to time constraints with our large dataset.
4. **Baseline Models**: Established performance baselines by training chosen models with default settings.
5. **Model Improvements**: Initial efforts to improve models through feature selection and hyperparameter tuning were not significantly successful.

About the data:

- **Genotypic Markers**: indicators that tell us if a plant has inherited certain traits from its parents. We mark them with -1, 0, or 1 to show different combinations of these traits.
- **Phenotypic Values**: Observable physical traits such as plant height, yield, and plant weight, among others.
- **Environmental Data**: Environmental conditions at specific locations and times, including year-specific metrics like precipitation, temperature, and soil characteristics.
- **Timeframe**: Data from 2000 to 2008.

Selected Final Models:

- **XGBoost**: An optimized distributed gradient boosting library designed for speed and performance, which uses decision trees as base learners.
- **LightGBM**: A fast, distributed, high-performance gradient boosting framework based on decision tree algorithms, used for ranking, classification, and other machine learning tasks.
- **Random Forest**: An ensemble learning method that builds multiple decision trees and merges their predictions, offering robustness and accuracy in various machine learning tasks.

Main Metric of evaluation:

- **R-squared Metric**: A statistical measure in regression analysis that represents the proportion of variance in the dependent variable that can be explained by the independent variable(s), used to assess the goodness of fit of a model.

RESULTS

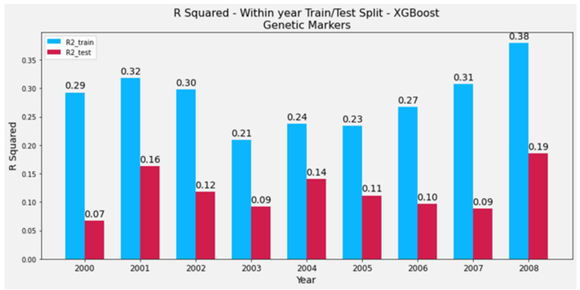

Figure 4
R squared results of a XGBoost model in a within each year train/test random split with only Genetic Markers as input to the model.

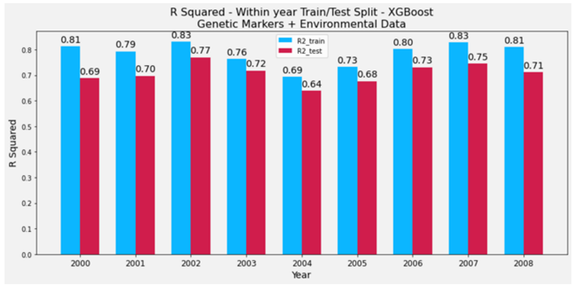

Figure 5
R squared results of a XGBoost model in a within each year train/test random split with Genotypic markers and Environmental data as input to the model.

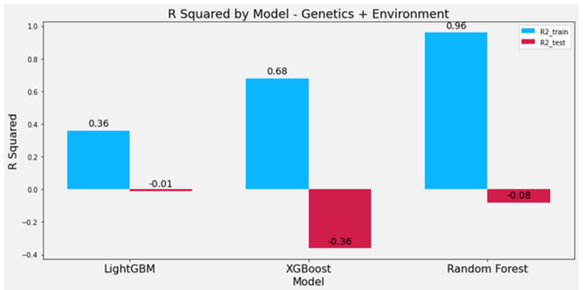

Figure 6
R squared results of the explored models in a 2000 – 2007 train / 2008 test setup with Genetic Markers and Environmental data as input.

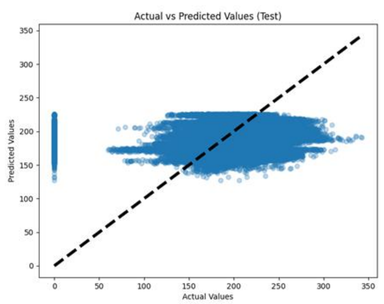

Figure 7
Actual vs. Predicted values of the LightGBM model – best performance on the test set in the 2000-2007 train / 2008 test setup with Genetic Markers and Environmental data as input.

CONCLUSIONS

- The annual random train/test split experiment revealed that the XGBoost model, a tree-based algorithm, accurately predicts corn yield by capturing genetic nuances, and incorporating environmental data further improves its predictive power (figure 4 and 5).
- In real-world scenarios requiring future predictions based on past data, including genotypic and environmental information, all the models we evaluated encountered challenges. Notably, LightGBM stood out by delivering the best performance on the test set (figure 6).
- Tree-based models, particularly Random Forest, XGBoost, and LightGBM, perform best for this task, considering both the full dataset and the 15 population/year subset.

FUTURE GOALS

- Utilize hyperparameter tuning and dimensionality reduction to deal with potential overfitting
- Explore other data imputation methods such as Beagle
- Explore statistical methods like LMM and RRBLUP as well as mixed models to improve prediction accuracy

REFERENCES

- https://www.uvm.edu/~statdhtx/StatPages/More_Stuff/RegToMean/RegToMean.html
- https://www.nature.com/scitable/knowledge/library/the-breeder-s-equation-24204828/
- https://passel2.unl.edu/view/lesson/c3ded390efbf/9

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

The Data Mine Corporate Partners Symposium 2024