Modeling Crop Gene-Environment Interaction Under Climate Change

Arnav Sehgal, Aditya Bose, Seth DeWhitt, Soumil Verma, Kabeer Khubbar, Melanie Ruesch, Jack Hansen

PURDUE UNIVERSITY® | The Data Mine

INTRODUCTION

• Bayer provided data from past corn hybrid experiments in the Midwest

• We wanted to explore how different genetic traits interreacted with their environment

• Goal: Understand how different types of corn yield change as climate change affects agriculture

Feature Importances
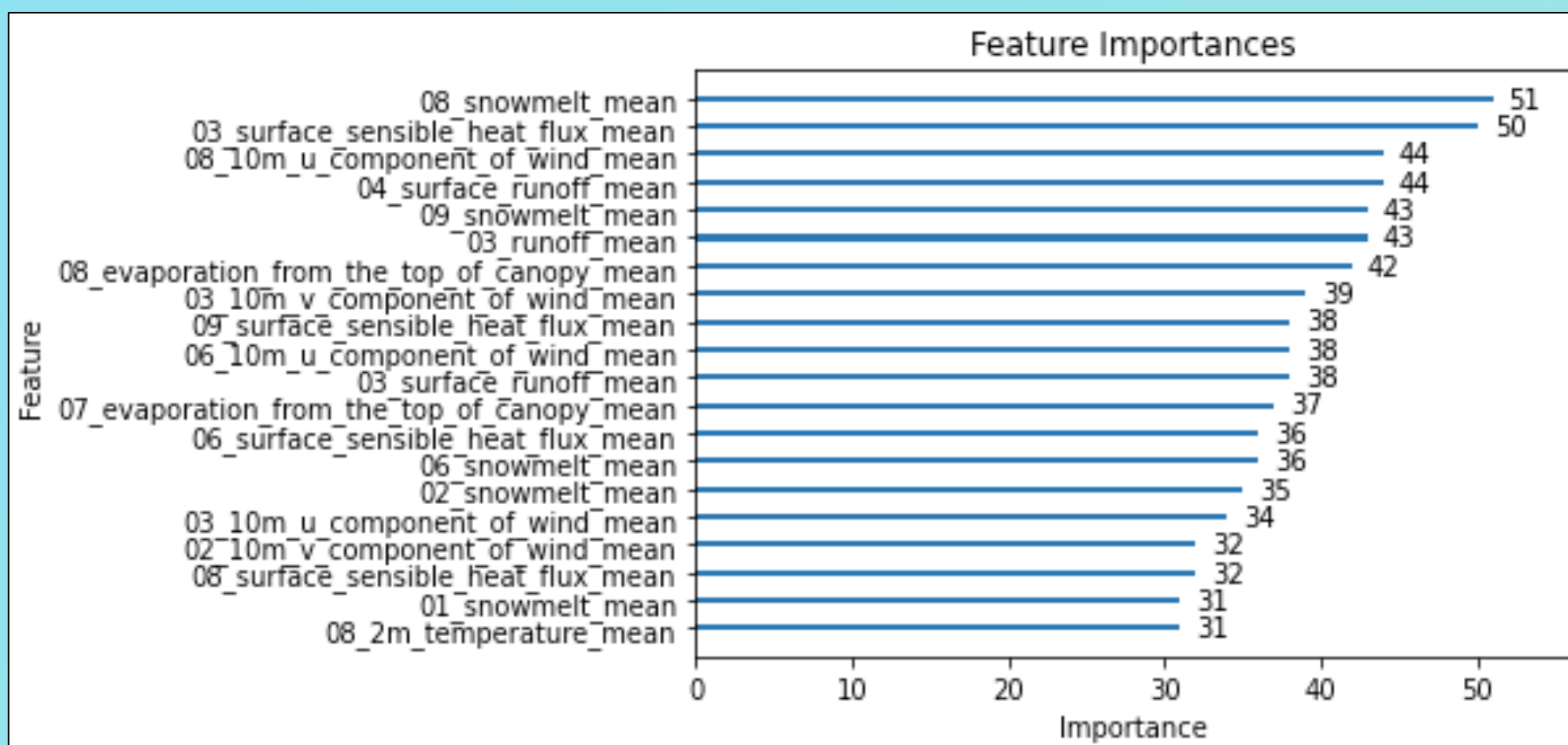
DATA OVERVIEW

Historical Dataset: [2000-2008]

• Started with different genetic variables and their yield
• Joined data with longitude and latitude with fields
  o Organized by year/weather keys
• Ended with dataset that includes the coordinates of fields with their yield, genetic variables, and weather data

Projection Dataset: [2040, 2050, 2070, 2080, 2090]

• Made by linking historical weather variables to specific genetic lines by coordinates.

MODEL TRAINING

• Tested multiple regression models: Lasso, Ridge, Random Forest, K-Nearest Neighbors, and LightGBM

• LightGBM is a machine learning model trained using the LightGBM framework
  ▪ It is known for its speed and accuracy with large datasets

• Cut the dataset down to just the most important variables using Principle Component Analysis

• Ran model with optimized hyperparameters

• Training set: subset of a random split selection of years in our dataset used to train the model
• Validation set: subset aiding in model selection and hyperparameter tuning
• Test set: evaluates its performance on the model with unseen data of future years

	Train	Validation	Test
MSE	17.585	17.832	41.766
Pearson R	0.881	0.875	0.087

BAYER

MAPS

Predicted Average Yield (bushels/acre) of All Corn Hybrids in SSP585 Weather Projections for Midwest Counties, 2040-2090

CONCLUSIONS

• Collaborating with farmers allows Bayer to gain insights into their needs, which can lead to the development of innovative solutions tailored to address agricultural issues.

• Between 2040 and 2090, agricultural developments such as genetic changes promote a significant increase in harvested amounts. Improved data analytics and sustainable methods help to increase production. By 2090, these trends predict a large increase in mean harvested amounts, Due to weather patterns, the Great Lakes region will possibly become more favorable on average compared to the other midwestern states.

FUTURE WORK

• Creating a website or user interface to display this data

• Finding stronger trends between feature importances and how different weather trends weigh on different states

• Improving model accuracy, correlation, other work on data science side

2040
2070
2090
Mean Bushels Per Acre