BECK'S HYBRIDS – PLOT LOSS TOOL

Julia Burroughs, Steven Doyle, Katie Hanagan ,Mark Jin, Nathan John, William Kao, Kindyl King, Ashwin Krishnaswamy,
Alexandra Loaiza, Nikhita Anantha Madhavan, Rohith Nadimpally, Ava Riaziat, Audrey Shirley, Michael Wu, Tianyi Zhang

PURDUE UNIVERSITY

INTRODUCTION

What is Beck's Hybrids?

Beck's Hybrids is the largest family-owned, retail seed company in the United States, serving farmers primarily in the Midwest. Beck's utilizes the best of genetics and trait technologies to provide customers with the best in seed quality and field performance.

Motivation

Beck's large-scale commercial corn breeding program is fueled by years of data from over 500,000 experimental plots per year. Every year breeders make decisions based on yield and phenotype of hybrid tests. It can be catastrophic to throw away a significant number of plots due to a variety of external factors, such as weather events, soil type, and slope within a field. Beck's provided us with a Prism dataset including whether plot was ultimately discarded in 2020. Our goal was to predict these discarded plots based on publicly-available Weather, Soils, and Elevation data. Understanding the effects of environmental variability, limiting field variability, and choosing promising plot locations is critical for Beck's to gain the greatest return on investment for their research efforts.

Objectives

- Aggregate publicly-available spatial data with Beck's historical research data
- Identify land features that are associated with a risk of plot loss
- Create a tool that identifies the best locations to plant research plots

Our approach involved two separate analyses. We assessed historical Weather data to understand conditions that led to plot loss, then analyzed Soils & Elevation data, creating a classification model that can predict if a location will be discarded.

WEATHER

Data Used:
- Visual Crossing Weather Data
- NOAA Weather Data

What We Looked For:
- Trends in weather data that correlate with plot loss
- Which plots are discarded in different locations
- How to incorporate public weather data with the Beck's data
- What weather attributes affect risk of plot loss

Zones:
- Since weather is relatively constant throughout a city, our analyses are between locations rather than individual plots.

Model:
- XGBoost (Extreme Gradient Boosted Decision Tree)

Results:

The box plots reveal how plot status changes with the following variables: maximum temperature, minimum temperature and precipitation. These variables were determined to be the most influential parameters according to a feature importance matrix

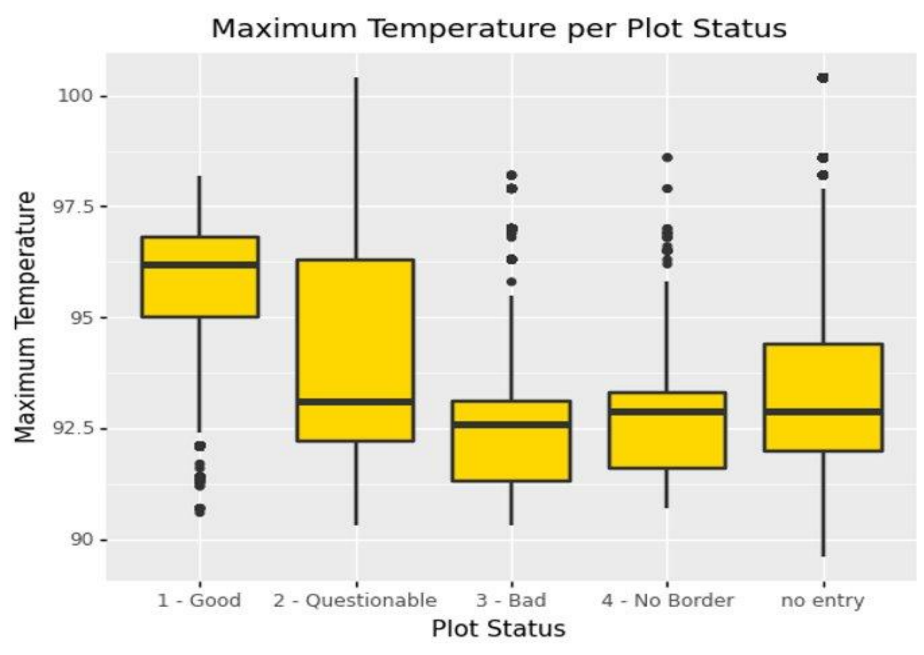

Figure 1a: Plots that experienced a higher maximum temperature are less likely to be discarded

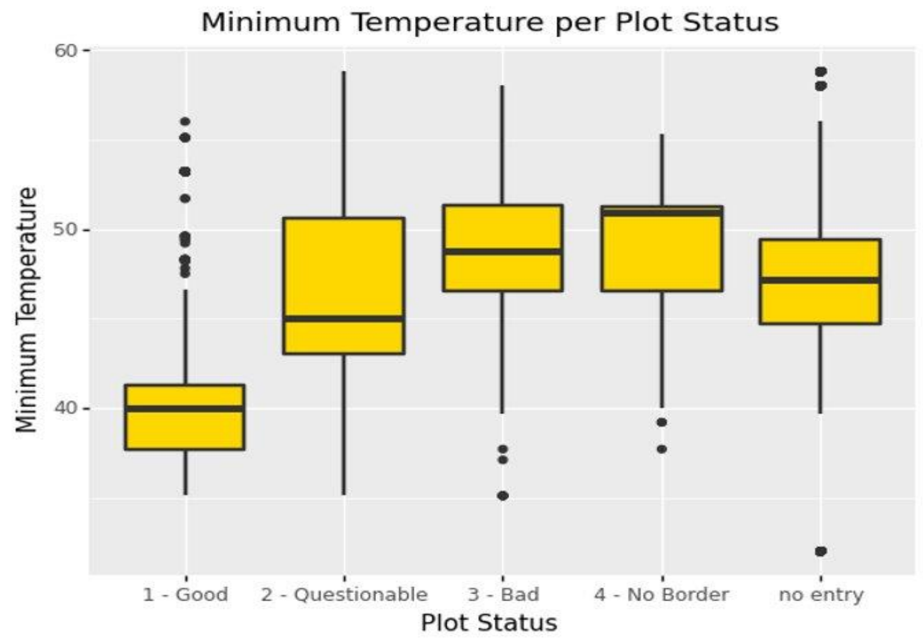

Figure 1b: Plots that experienced a lower minimum temperature are less likely to be discarded

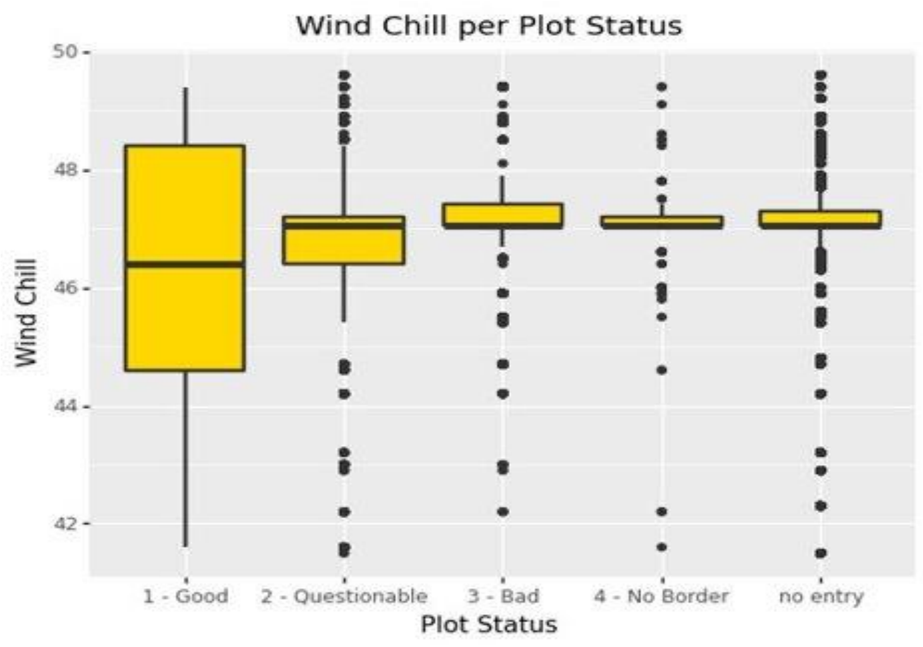

Figure 1c: Plots that experienced more variation in average wind chill are less likely to be discarded

SOILS & ELEVATION

Data:
- Soil Survey Geographic Database (SSURGO)
- Light Detection and Ranging (LiDAR)-based Digital Elevation Models

Preprocessing:
- Elevation data is available on a more granular level than each plot, so we aggregate the data to have one measurement per plot.
- Terrain attributes for each plot were normalized to the mean of a 1-kilometer radius around the field in order to consider the surrounding landscape.

Models:
- Multiple Factor Analysis (MFA)
- Random Forest

Results:

- 85% accuracy in classifying a single plot in Indiana as discarded or not discarded
- 89% classification accuracy for plots in IN, IA, MN, and IL using only soil data
- Most important features to the classification model:
  - Available Water Storage 0-150cm, 0-100cm, 0-50cm, and 0-25cm
  - Normalized field elevation to a 1km buffer

CONCLUSIONS

- Available water storage and normalized elevation encompassed most of the variability in a combined analysis of soil and elevation data
- The data is naturally very imbalanced because the data from most plots is kept, while less are discarded. Oversampling from the discarded plots in our training data set would address this problem and improve the accuracy of our model.
- Elevation features are most informative when looking at a larger surrounding landscape outside of the field boundary.
- Temperature and wind chill are the most important predictors on weather variables based on a XGBoost Decision Tree Model.

FUTURE GOALS

- Expand analyses beyond Indiana to all states where Beck's plants experimental plots
- Estimate plot boundary data to allow for flexibility in handling new locations
- Oversample discarded data to improve accuracy
- Holistically analyze insights from weather, soils, and elevation together

ACKNOWLEDGEMENTS

Thank you to our Beck's Hybrids Corporate Mentors: Steve Becker, Brad Fruth, Brooke Gajownik, Will Hirschfeld, Brandi Payne for guiding our project implementation with industry knowledge and personal wisdom.

Thank you to our Purdue Faculty Mentors: Dr. Jason Ackerson and Dr. Darrell Schulze for sharing your expertise in soil sciences and the computational methods for analyzing our data.

Thank you to the Purdue Data Mine staff, including Dr. Mark Daniel Ward and Maggie Betz, for facilitating the Corporate Partners program and investing in each student on the team.

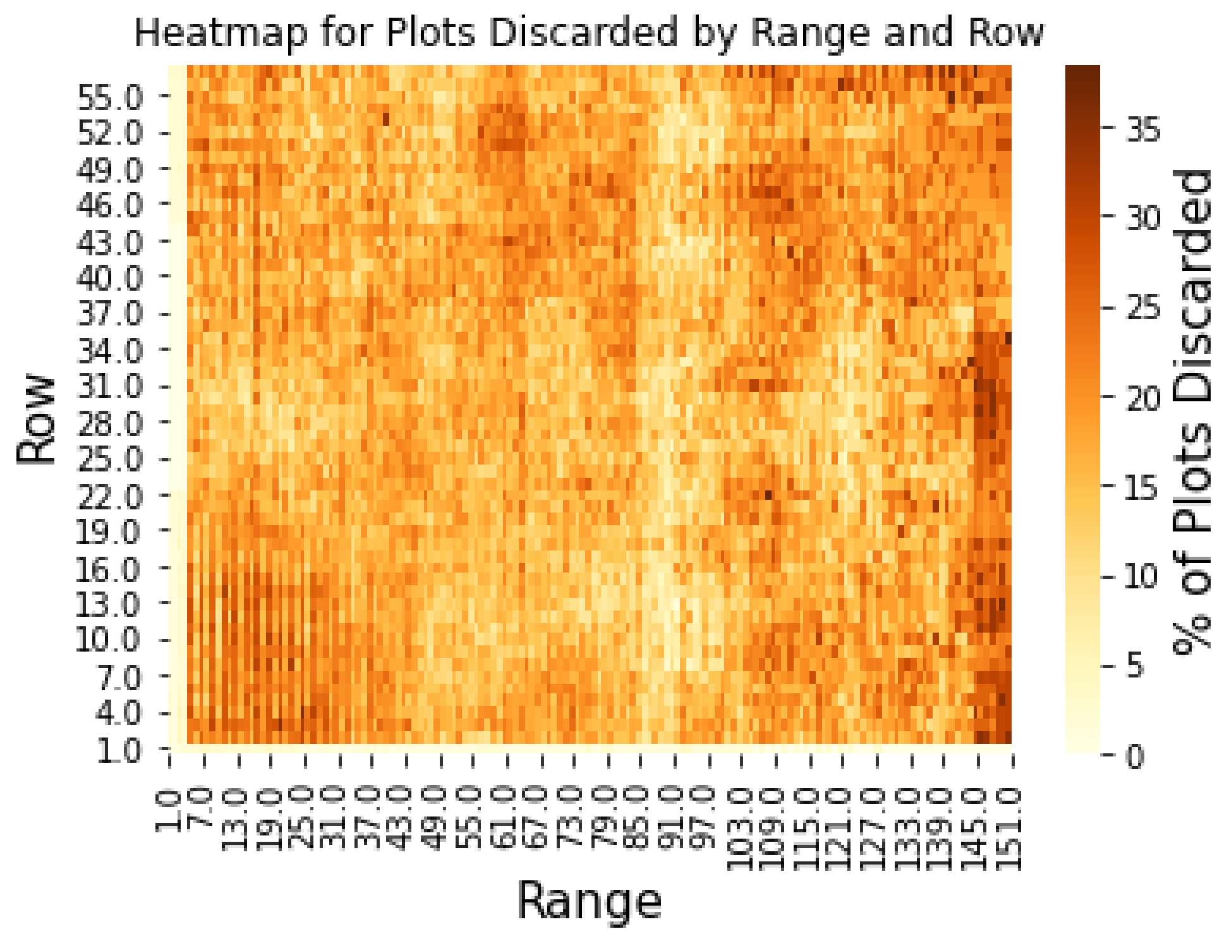

Figure 2: Heatmap

- Percentage of plots discarded varies based on row, range combinations in the field
- Using Prism dataset – all of Beck's 2020 plots
- Plots in the interior are discarded at a much lower rate than exterior plots

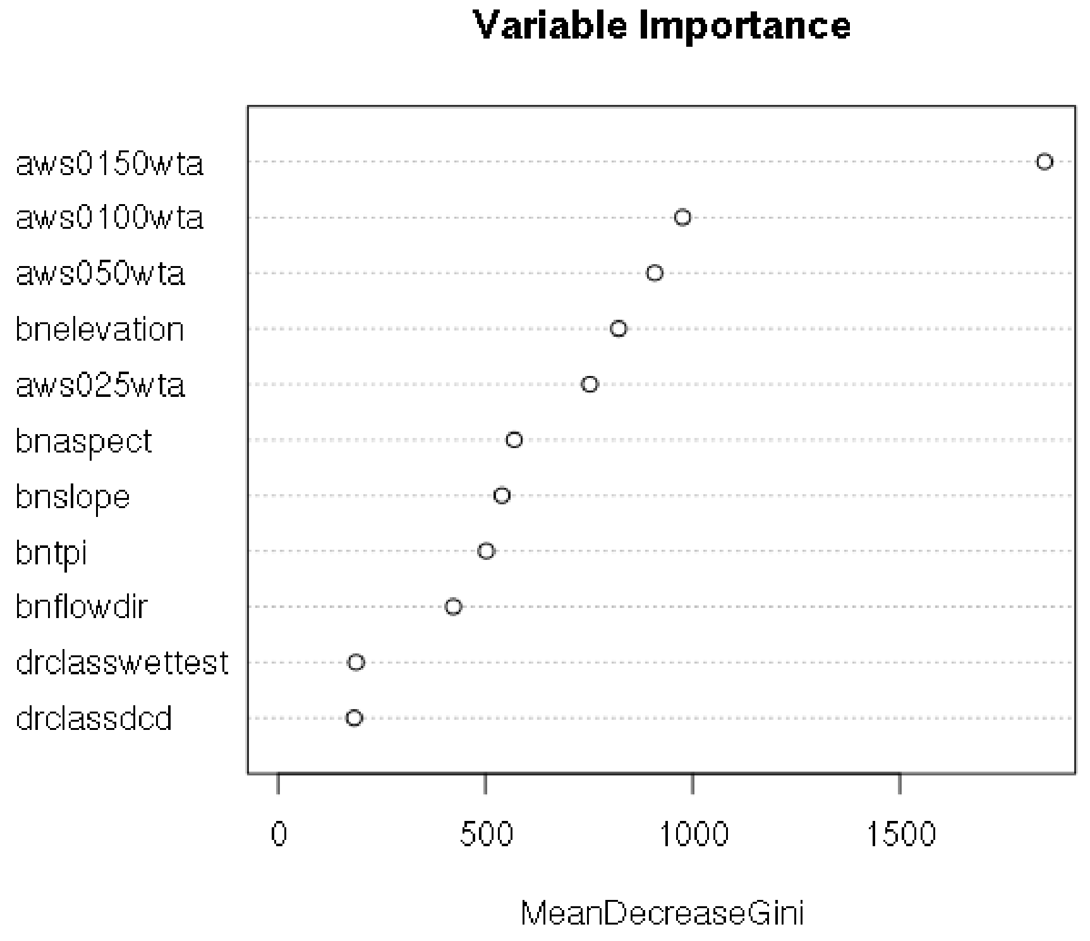

Figure 3: Mean decrease in node impurity. Larger values correspond to higher importance in the random forest.

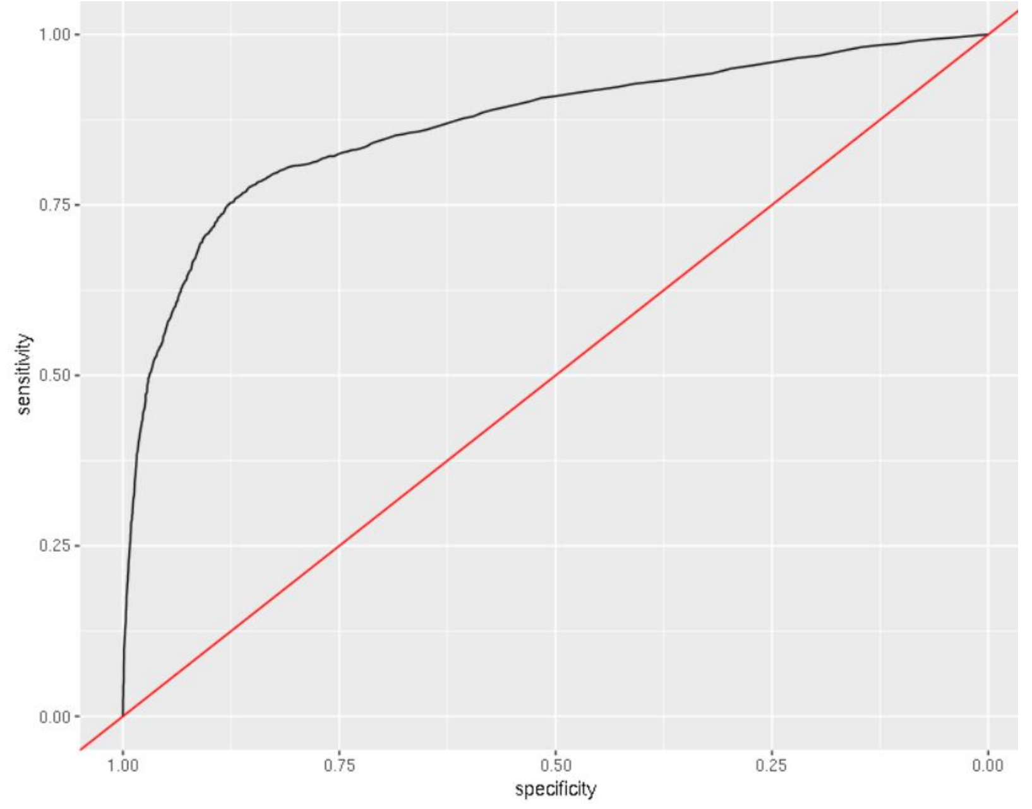

Figure 4: ROC curve. Sensitivity (True Positives) v Specificity (True Negatives). Best shape is toward top left.

The Data Mine Corporate Partners Symposium 2021