FarmView: Regression Analysis of 2016 Sorghum Composition

Ben Parr
bparr@cmu.edu

Sorghum bicolor

- Bioenergy sorghum breeders expect to double yield in the next five years alone.
- Drought-tolerant and highly productive grass.
- Diverse gene pool containing over 40,000 genetic varieties.

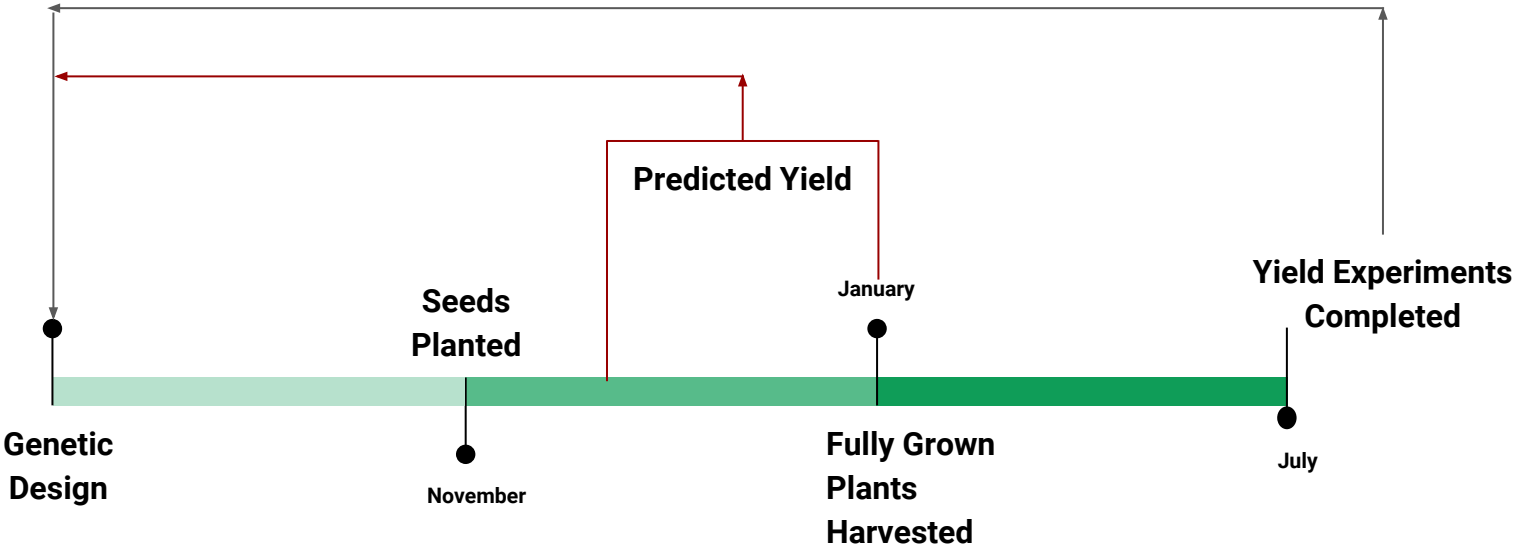
Ground Robot and Aerial Drone

- Ground robot traverses rows of sorghum.
- Aerial drone flies above field of sorghum.

Dataset

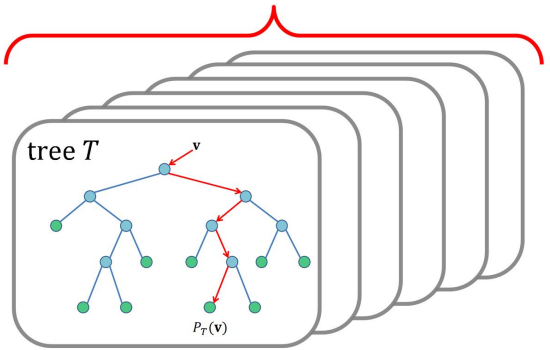
- Field cultivated in Pendleton, South Carolina in 2016.
- 698 subplots (samples).
- Inputs: Accession (e.g. country of origin), GPS field location, ground robot, aerial drone, harvest phenotypes (e.g. harvested plant weight).
- Outputs: Sorghum composition

Motivation


Predicted Yield
Seeds Planted
January
Yield Experiments Completed
Genetic Design
November
Fully Grown Plants Harvested
July

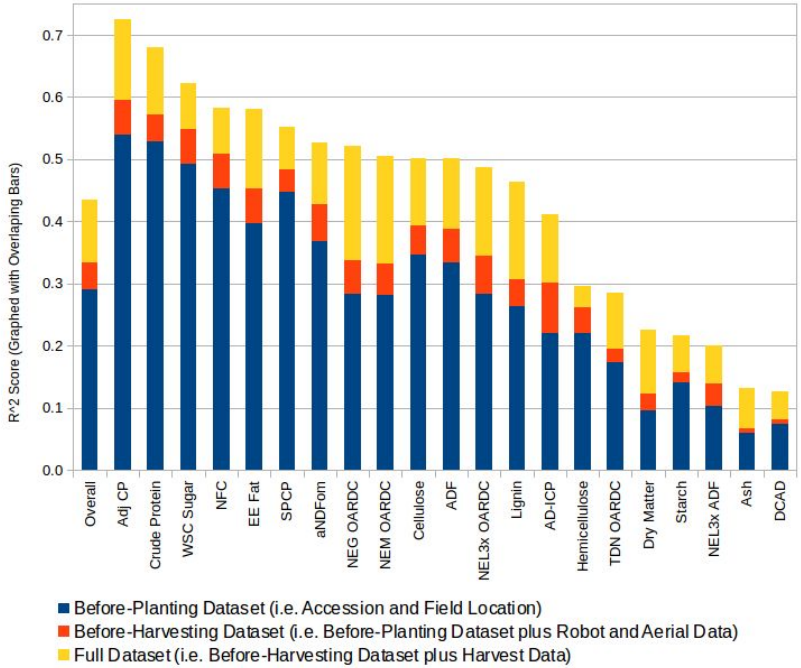
Best Regressor: Random Forest Regressor

- All regressors used 10-fold cross validation.
- Random forests are an ensemble method where the final prediction is the average of the outputs of its constructed regression decision trees.
- Random forests are fast and perform well in practice.



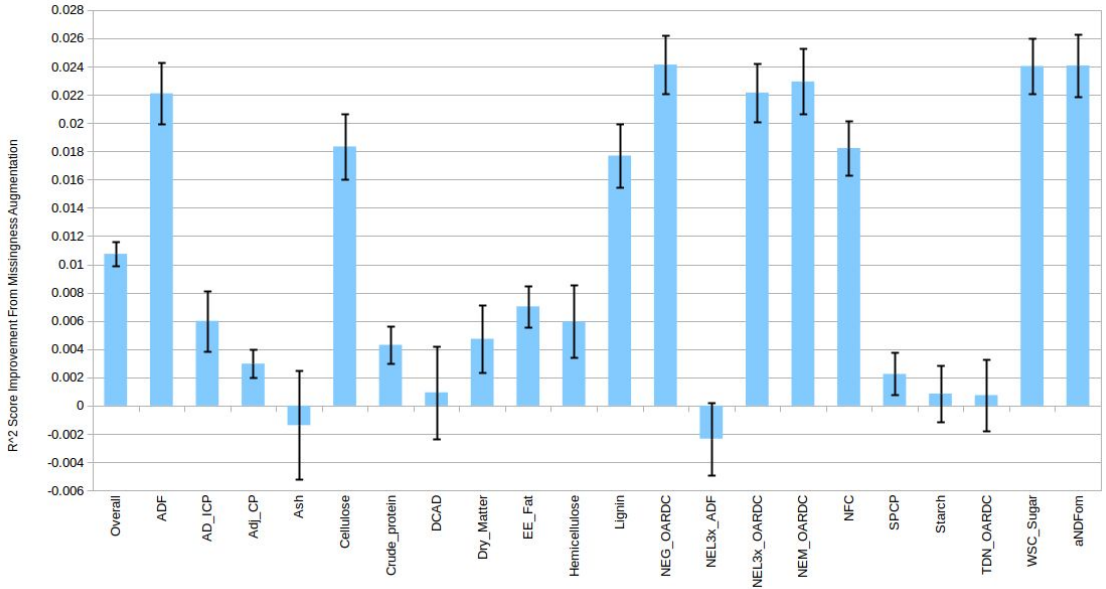
Overall and Individual Results

- Overall r^2 score of 0.436
  - Multi-dimensional r^2 score across all 21 output features.



Missingness Augmentation

- Over half of the samples are missing >24% of the input features.
- 2.5% increase in r^2 score by augmenting the training set with training samples that had selected values removed and replaced with missing values.



Acknowledgements

- Advisor: Dr. Artur Dubrawski.
- Saswati Ray and Simon Heath from AutonLab.
- U.S. Department of Energy: Breeding High Yielding Bioenergy Sorghum for the New Bioenergy Belt.