

Sorghum Composition Prediction

Ben Parr (bparr@cmu.edu)

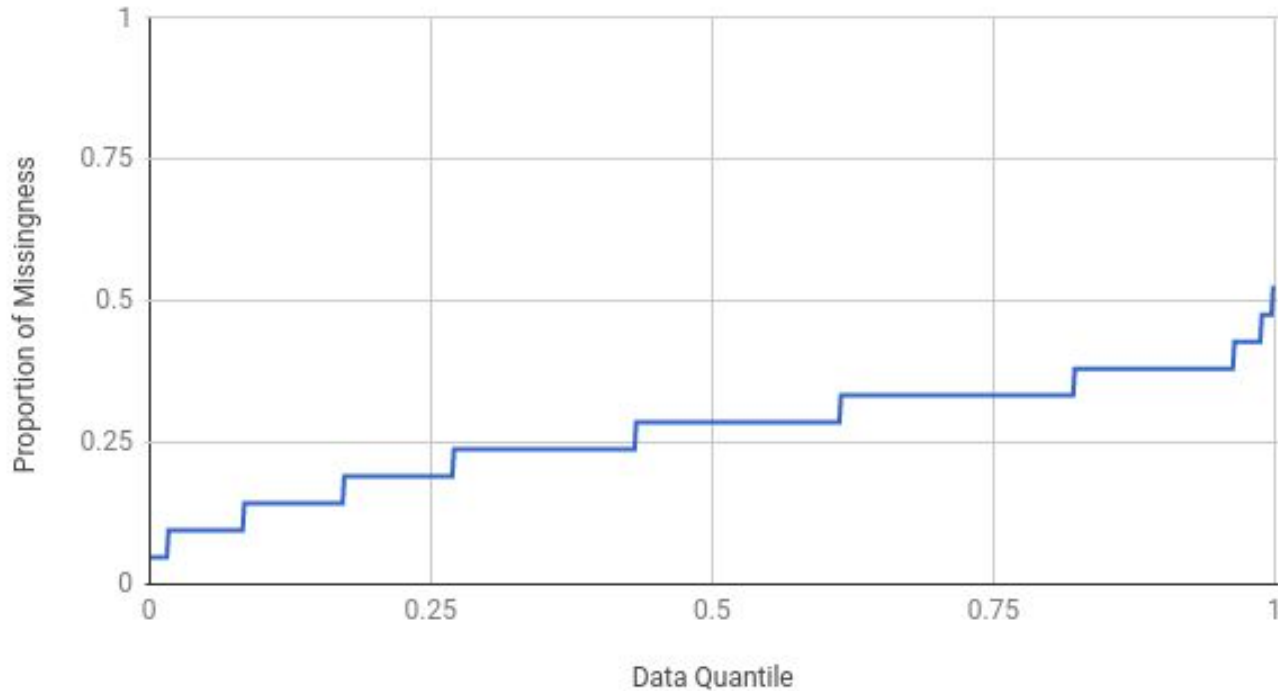
July 10, 2017

Methods

- Random Forests
- Inputs
 - Harvest features: Plant heights, wet weights and panicle lengths.
 - Robot features: Light interception, leaf area, vegetation index, etc.
 - Accession features: Sorghum type, photoperiod sensitive/insensitive, etc.
- Output: A single composition feature
 - A random forest was generated for each composition feature.

CDF of Missingness of the Inputs

Proportion of Missingness of Inputs vs. Data Quantile



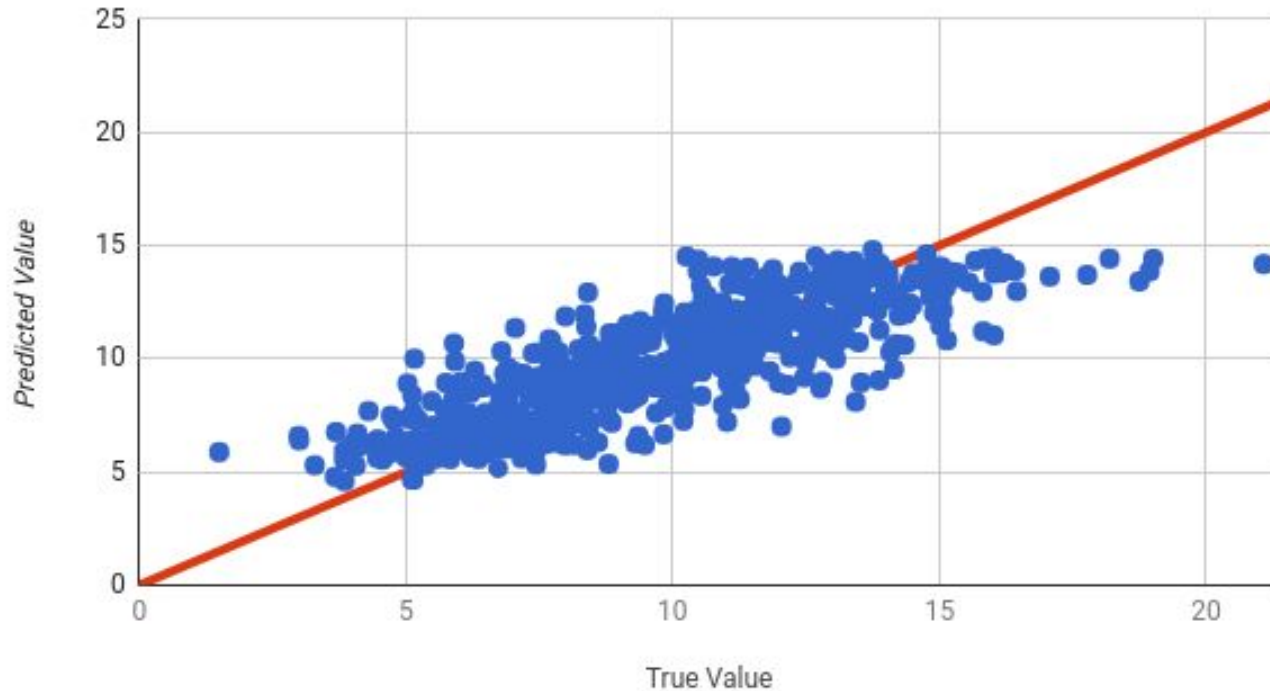
Random Forest R^2 Score from 10-fold CV Predictions

Feature	R^2 Score
Adj_CP	0.710
Crude_protein	0.665
WSC_Sugar	0.585
EE_Fat	0.575
NFC	0.563
SPCP	0.545
aNDFom	0.502
ADF	0.488
Cellulose	0.484
NEG_OARDC	0.482
NEM_OARDC	0.460
NEL3x_OARDC	0.438
Lignin	0.418
AD_ICP	0.400
TDN_OARDC	0.269
Hemicellulose	0.260
NEL3x_ADF	0.207
Starch	0.206
Dry_Matter	0.199
DCAD	0.091
Ash	0.053

Sorted by R^2 score.

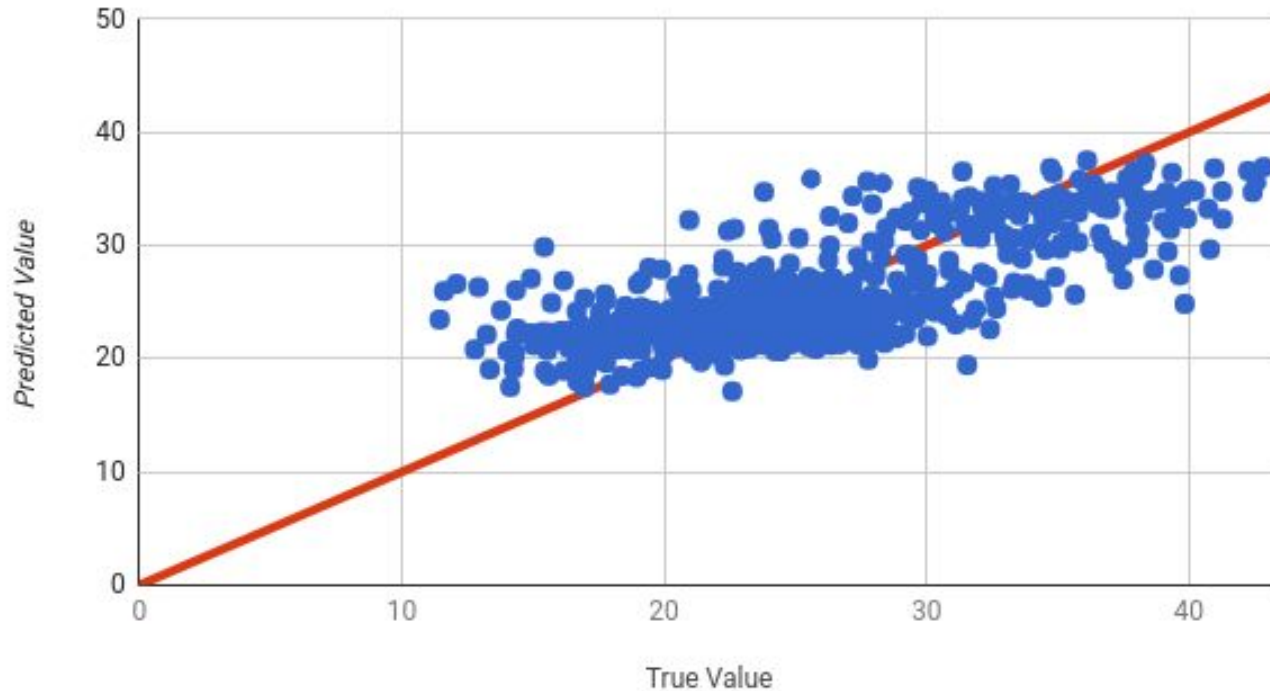
Predicting Adj_CP

Adj_CP: True vs. Predicted Value (r-squared of 0.710)



Predicting NFC Percentage

Cellulose: True vs. Predicted Value (r-squared of 0.563)



Discussion

- Unsatisfactory predictions for prioritized composition features
 - Cellulose, hemicellulose and lignin
- Will the aerial data improve prediction of the prioritized composition features?
 - For example, hyperspectral imaging gives view into chemical composition of plants.
- Dry Matter also performing poorly.
 - Is not correlated strongly with harvest height input, as originally expected.
 - Also, has a small range of values (89.26 - 93.98 kg).