# Predicting Crop Quality: A Tour of Regression Models

Blake Bullwinkel

M.S. in Data Science

jbullwinkel@fas.harvard.edu

# The Goal

*"Given the provided data sets, derive a model that would predict the assessment scores [of crops] as accurately as possible using relevant features or predictor variables"*

# Approach

1. Collate the three Excel sheets into a single dataframe
2. Split the data into train (80%) and test (20%) sets
3. Build a series of regression models
4. Choose the model with the highest $R^2$ on the test set
5. Predict “Assessment Score” on the entire dataset

# Data Pre-Processing

- Append Sowing Date, Latitude, Elevation, Soil Parameter A, Soil Parameter B and Amount Fertilizer Applied Site Data to Crop Grain data by matching Site ID

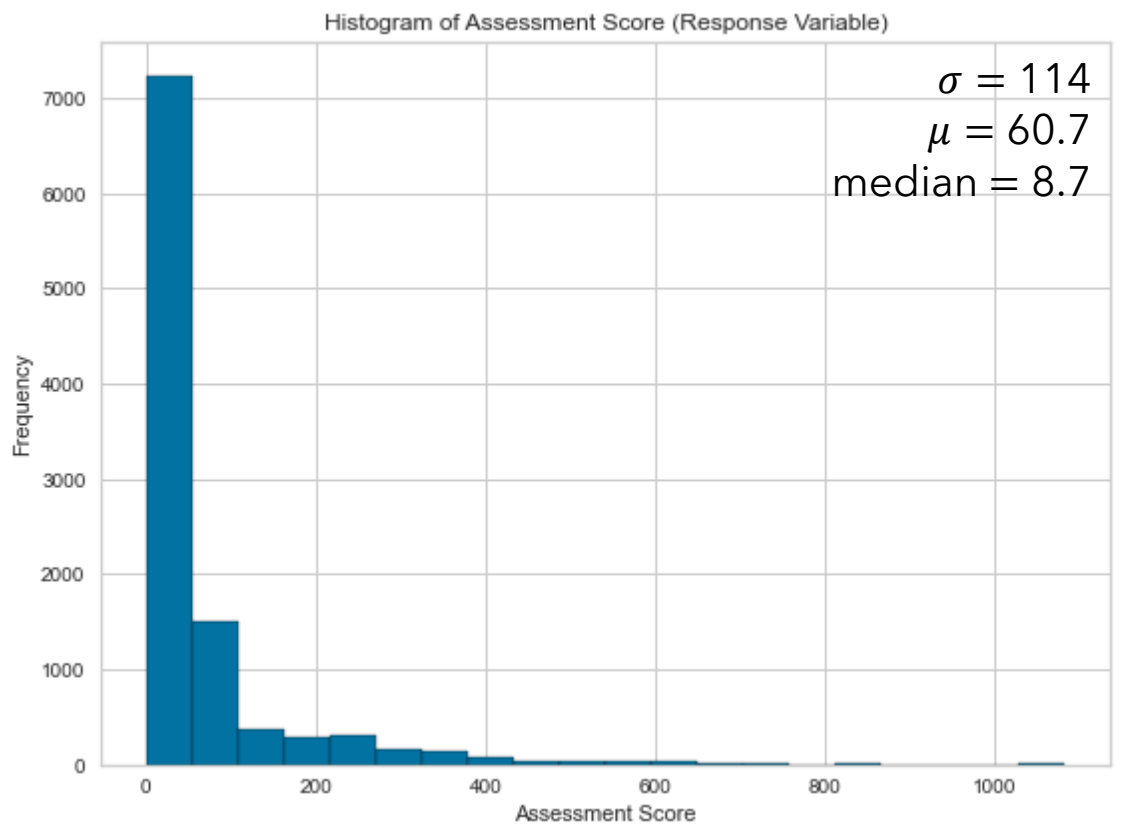
- Incorporate weather data by averaging each weather variable A-F from site Sowing Date to crop Assessment Date

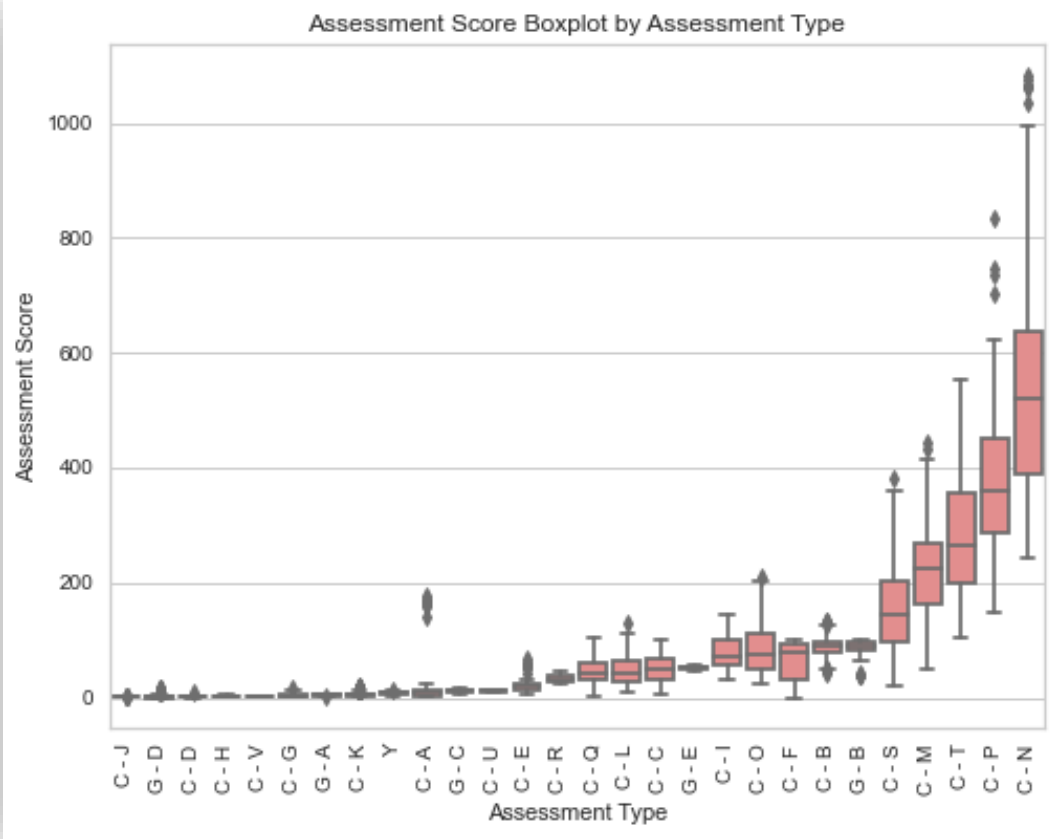
- Generate dummy variables for categorical predictors Variety and Assessment Type (one-hot encoding)

# 1. Baseline Model

Train $R^2$: 0.8521
Test $R^2$: 0.8340

- Multiple linear regression on all features

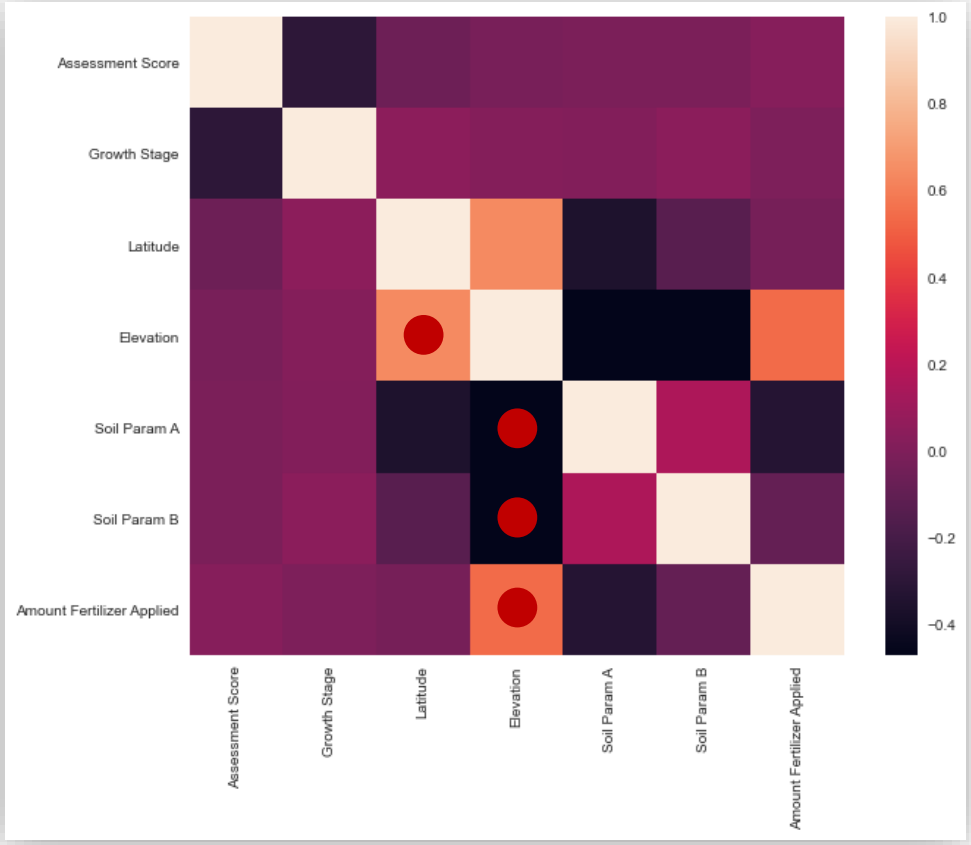


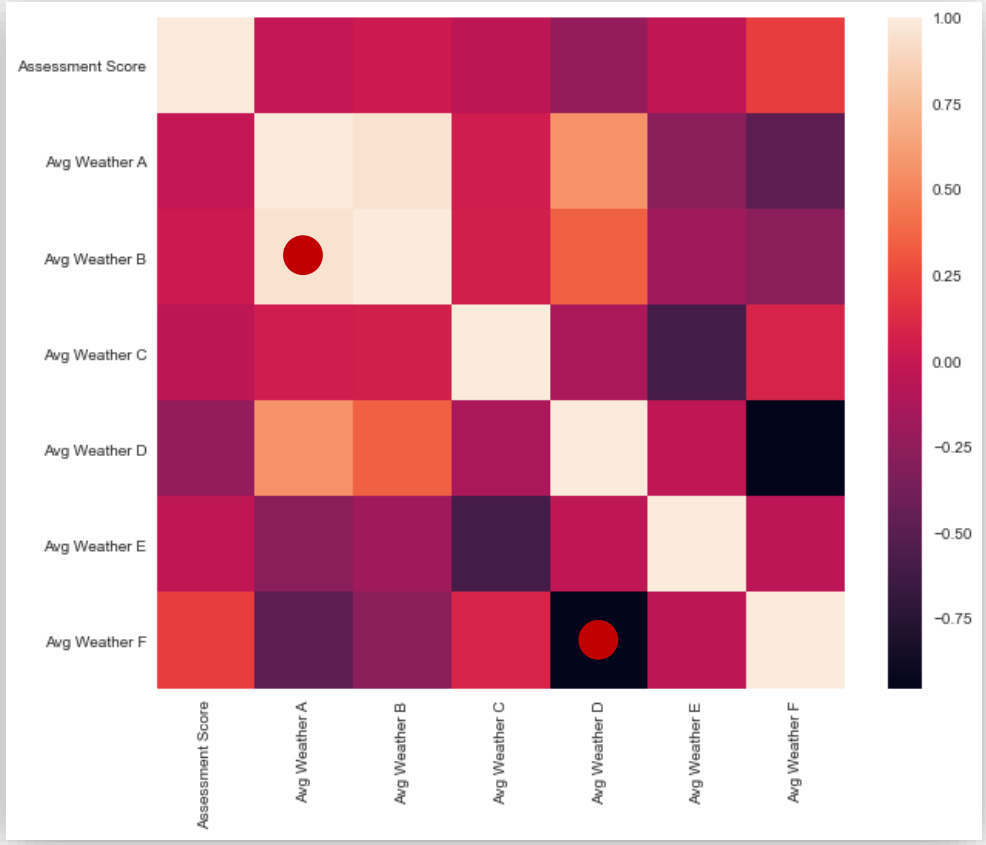



# 2. Incorporating Interactions

Train $R^2$: 0.8534
Test $R^2$: 0.8357

- Multiple linear regression on all features with interaction terms

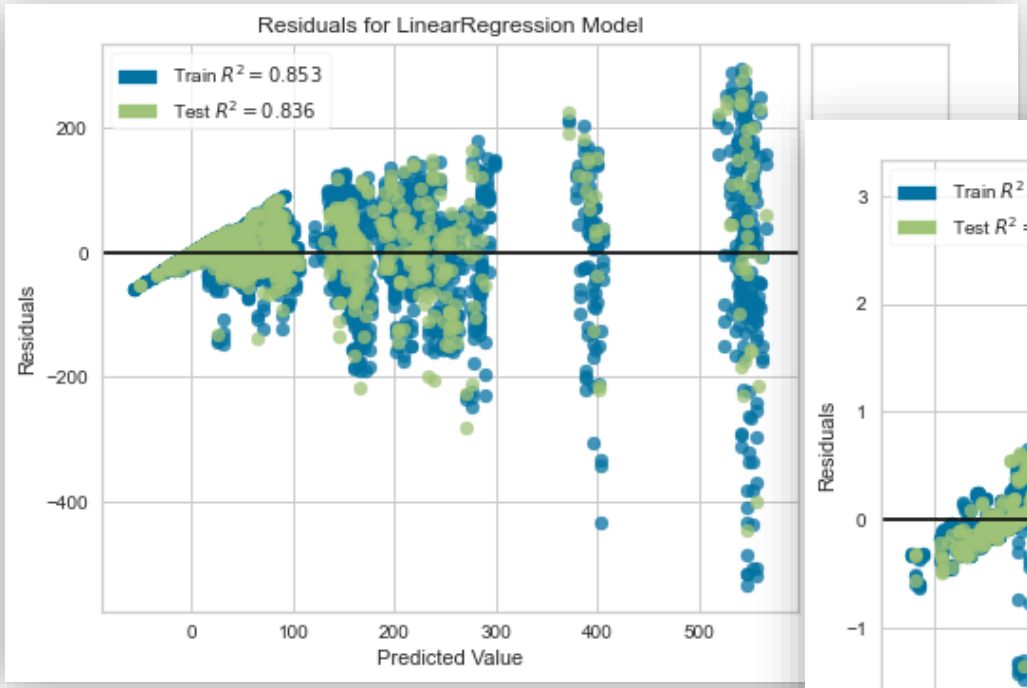


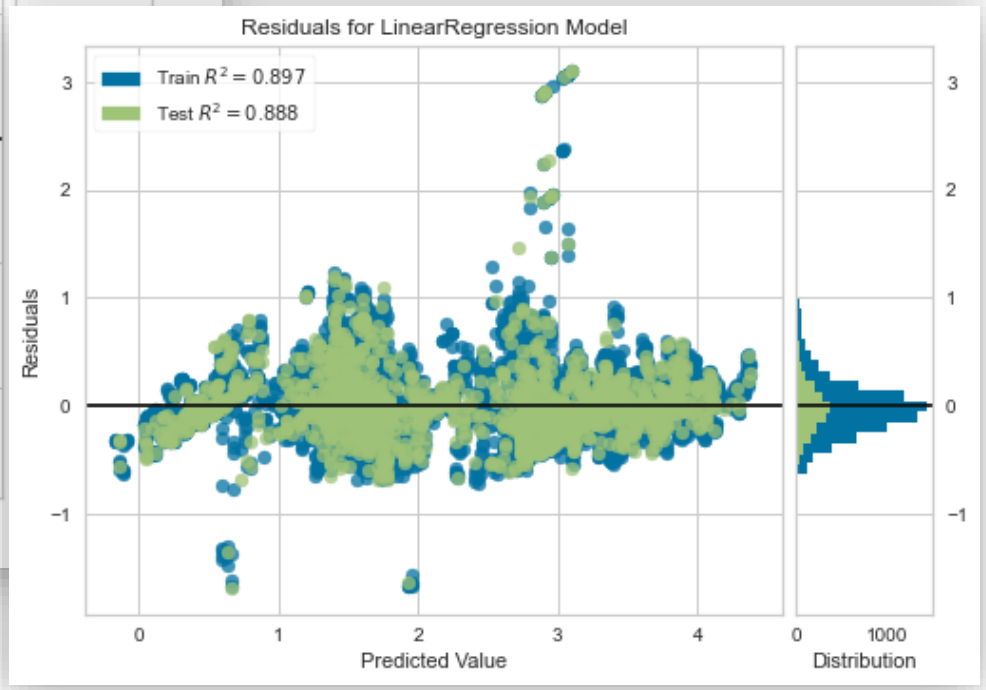



# 3. Transforming the Response

Train $R^2$: 0.8967
Test $R^2$: 0.8886

- Multiple linear regression on all features, interaction terms and Yeo-Johnson transformed response with $\lambda = -0.1467$.



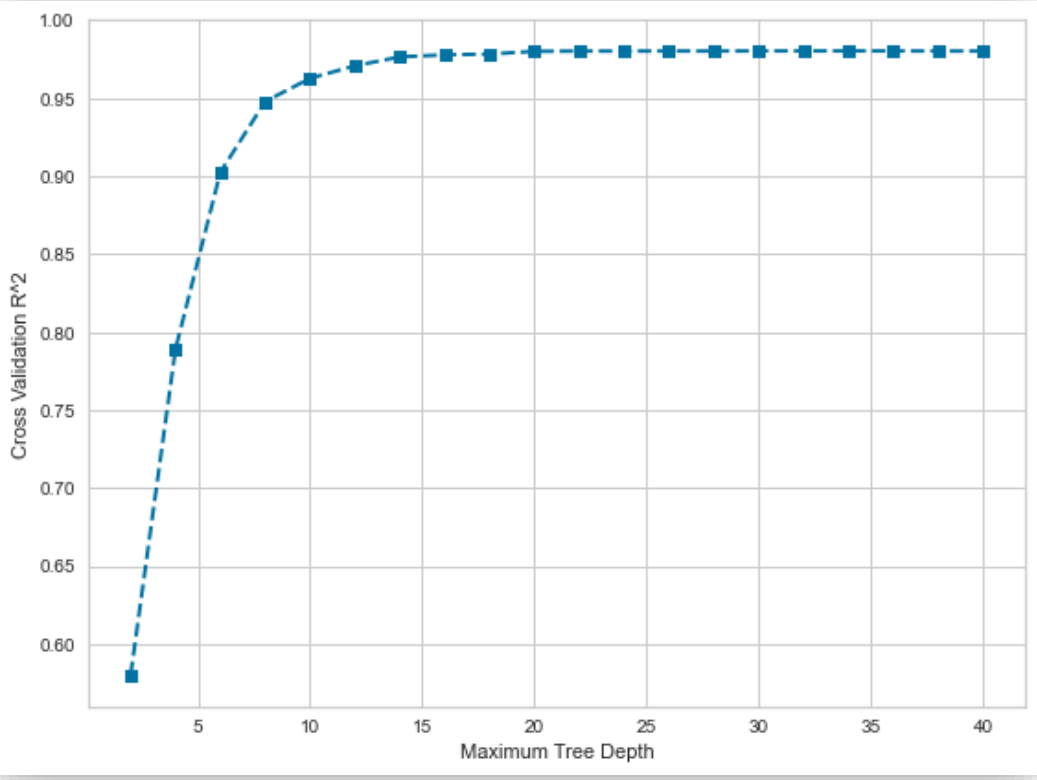

Model coefficients:

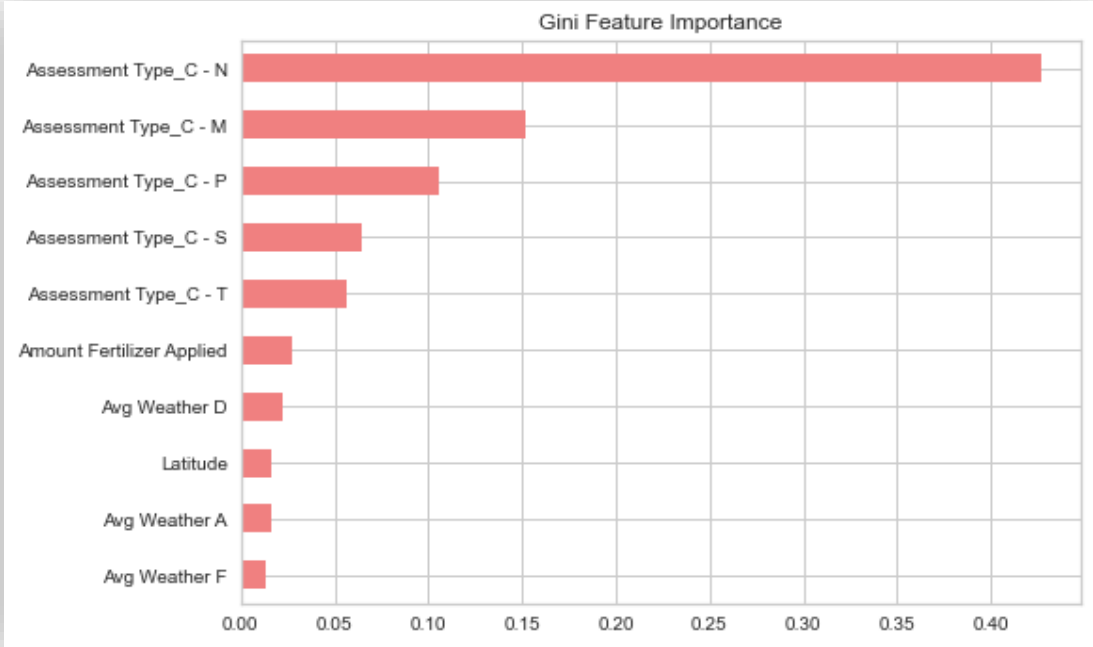
Assessment Type C-N = 1.517
Assessment Type C-M = 1.510
Assessment Type C-P = 1.308
Assessment Type C-T = 1.138
Assessment Type C-S = 0.890
…

# 4. Decision Tree

Train $R^2$: 0.9900
Test $R^2$: 0.9810

- Decision tree on all features and maximum tree depth hyperparameter tuned by 10-fold cross-validation

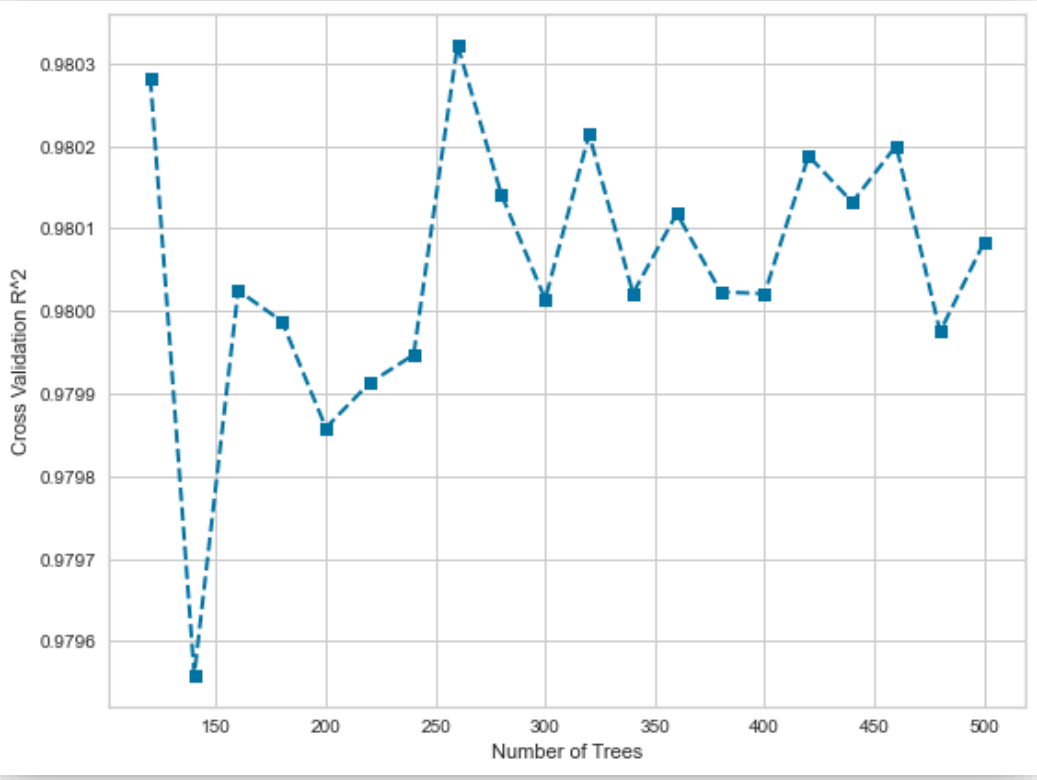


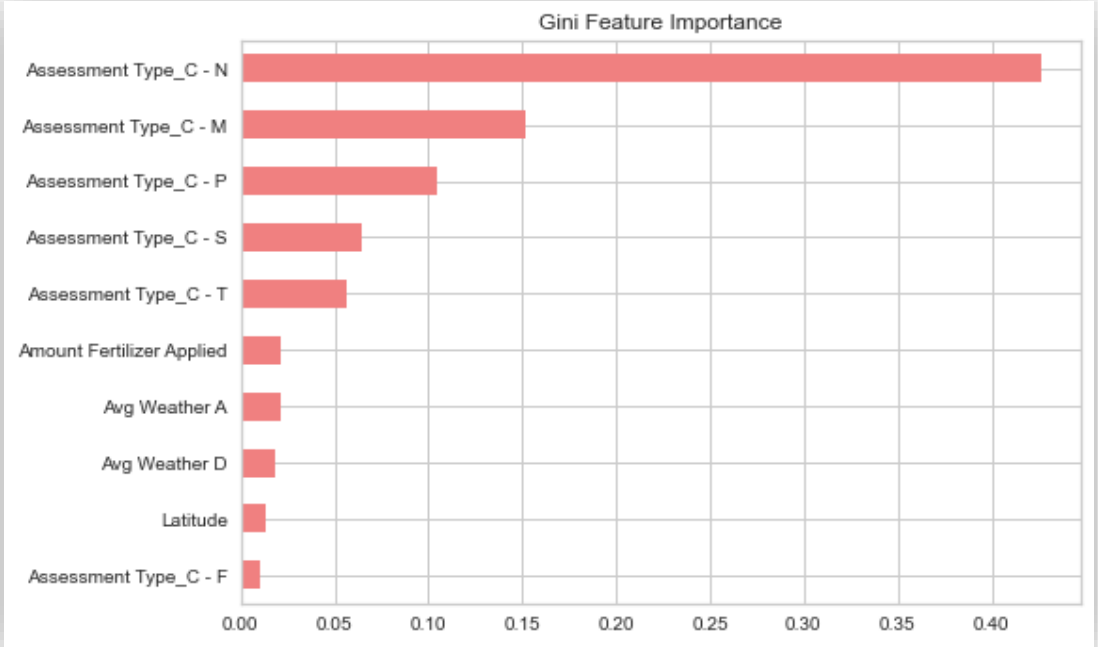



# 5. Random Forest

Train $R^2$: 0.9899
Test $R^2$: 0.9815

- Random forest on all features and number of trees hyperparameter tuned by 10-fold cross-validation



# Conclusions

- Most Assessment Scores are relatively low, while there is a small number of much higher values
- Assessment Type is the most useful feature for distinguishing between these low and high Assessment Scores
- Assessment Types C-N, C-M, C-P, C-S, and C-T are the best predictors; Amount Fertilizer Applied, Latitude, and weather are also useful
- Linear regression provides good predictions with precise interpretability; tree-based regression models achieve excellent predictions with relative interpretability