

Africa Soil Property Prediction

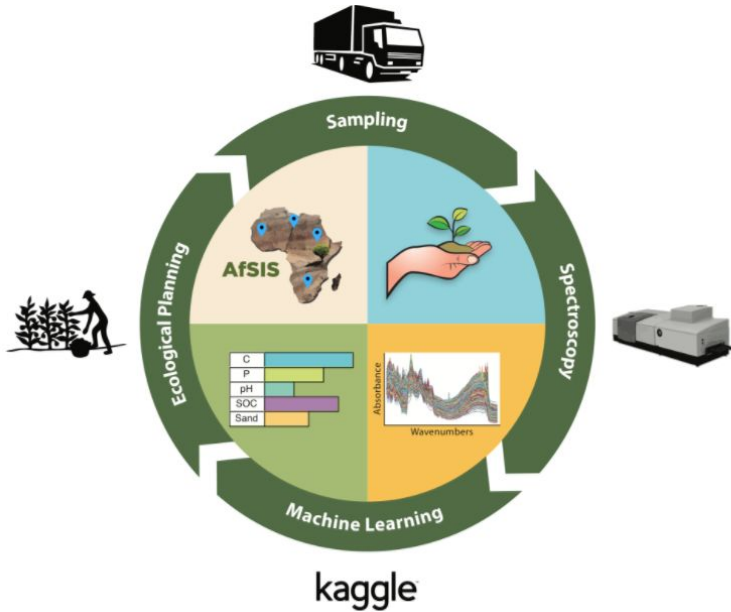
Jessica Sanders, Jennifer Casper, Carlos Castro

April 24th, 2017

W207 - Machine Learning



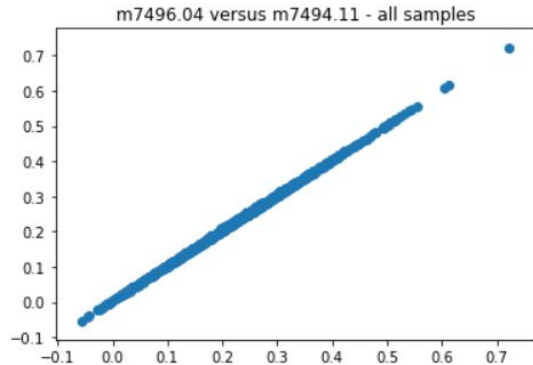
“Functional” properties can tell us how well soil will support an ecosystem



- Data: 1157 soil samples from 60 “Sentinel Landscapes” in Africa, each of which has:
 - **Mid-infrared light absorption measurements** for separate wavelengths.
 - Seventeen other predictor variables, such as soil depth, mean annual rainfall from sample area, etc.
- Challenge: Predict 5 target **soil functional properties**:
 - **SOC**: Soil organic carbon
 - **pH**: pH values
 - **Ca**: Calcium content
 - **P**: Phosphorus content
 - **Sand**: Sand content

The data presented with some challenges

Adjacent wavelength measurements (features) are **nearly co-linear** across the samples.



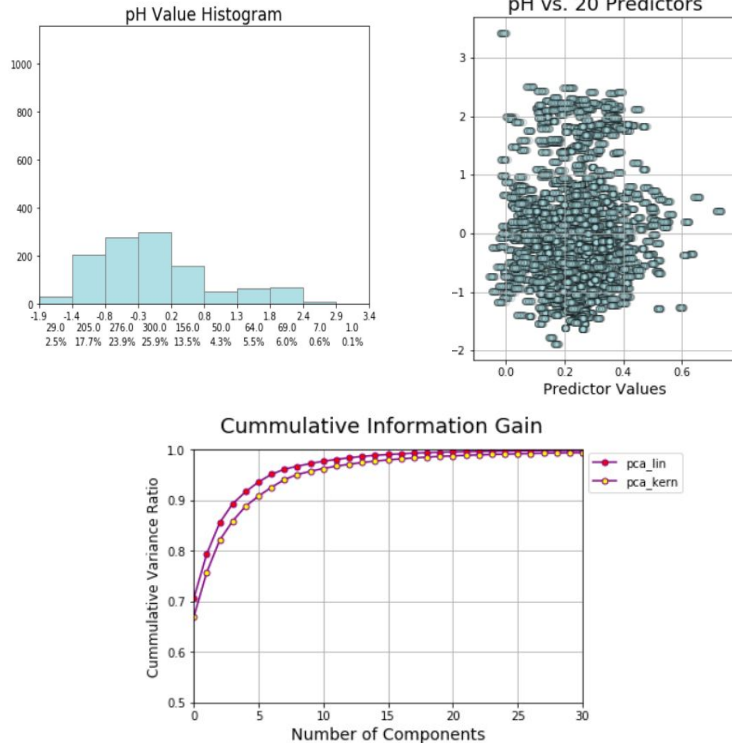
Training/test data was split along lines of Sentinel Landscapes, **and thus may be inherently different in some ways.**



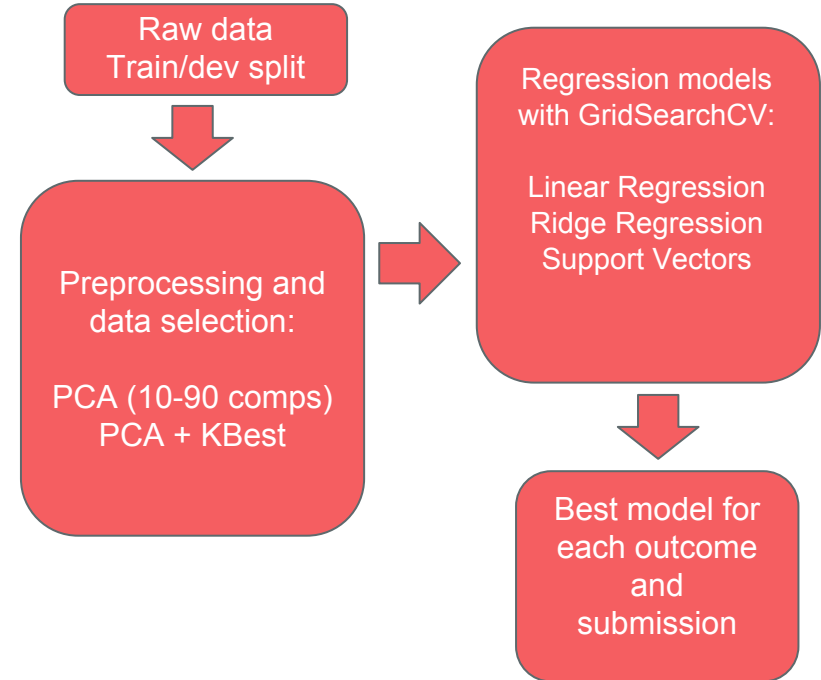
Number of features (3594) >> number of samples (1157)

Our process (Python/sklearn)

EDA and PCA experiments



Regression pipeline and model selection



Lessons Learned

- Tried many different models, including:
 - Linear Regression
 - Random Forests,
 - Neural Networks,
 - Support Vector Regression
 - K-Nearest Neighbors
 - ... and linear regression tended to win out.
- Found that the scoring was pretty sensitive to changes in the train/dev split
- Encoding the categorical depth predictor variable was not helpful
 - Phosphorus content MSE varied given depth
- Future options
 - Preprocessing alternatives
 - Split the data differently
 - Explore the other data collected - weather and spatial

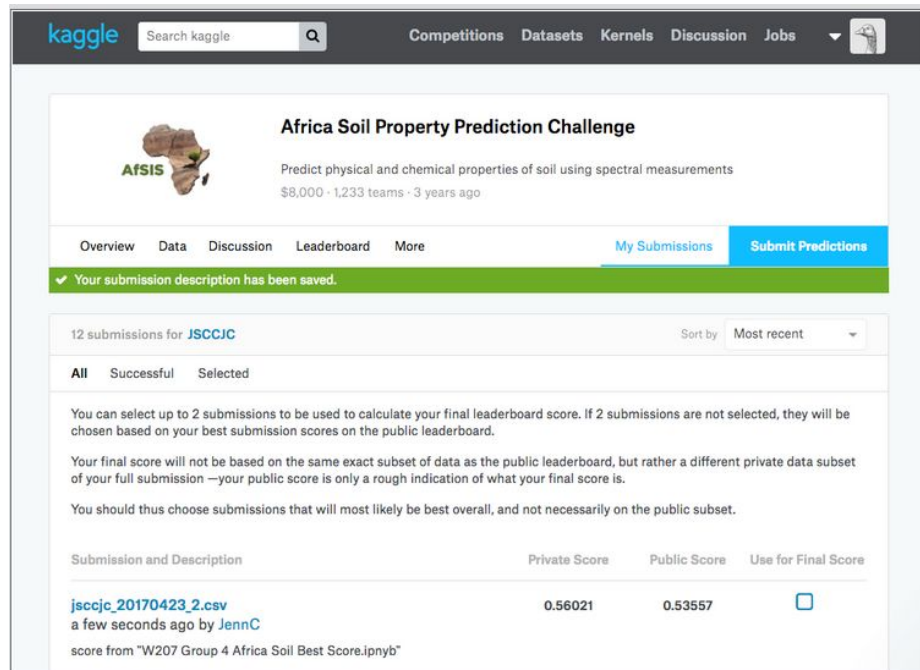
Results:

The scoring metric is the mean column root mean squared error:

$$\text{MCRMSE} = \frac{1}{5} \sum_{j=1}^5 \sqrt{\frac{1}{n} \sum_{i=1}^n (y_{ij} - \hat{y}_{ij})^2},$$

Our final score: 0.53557

The competition winner scored: 0.46892



The screenshot shows the Kaggle interface for the "Africa Soil Property Prediction Challenge". The page includes a navigation bar with links to Competitions, Datasets, Kernels, Discussion, and Jobs. The challenge title is "Africa Soil Property Prediction Challenge" with a sub-header "Predict physical and chemical properties of soil using spectral measurements". It mentions "\$8,000 · 1,233 teams · 3 years ago". The page has tabs for Overview, Data, Discussion, Leaderboard, and More. A green banner states "Your submission description has been saved." Below this, there's a section for "12 submissions for JSCCJC" with a "Sort by" dropdown set to "Most recent". A table lists submissions with columns: Submission and Description, Private Score, Public Score, and Use for Final Score. The submission "jscjc_20170423_2.csv" by "JennC" is highlighted, showing a Private Score of 0.56021 and a Public Score of 0.53557. A checkbox is present in the "Use for Final Score" column.

Submission and Description	Private Score	Public Score	Use for Final Score
jscjc_20170423_2.csv a few seconds ago by JennC score from "W207 Group 4 Africa Soil Best Score.ipnyb"	0.56021	0.53557	<input type="checkbox"/>