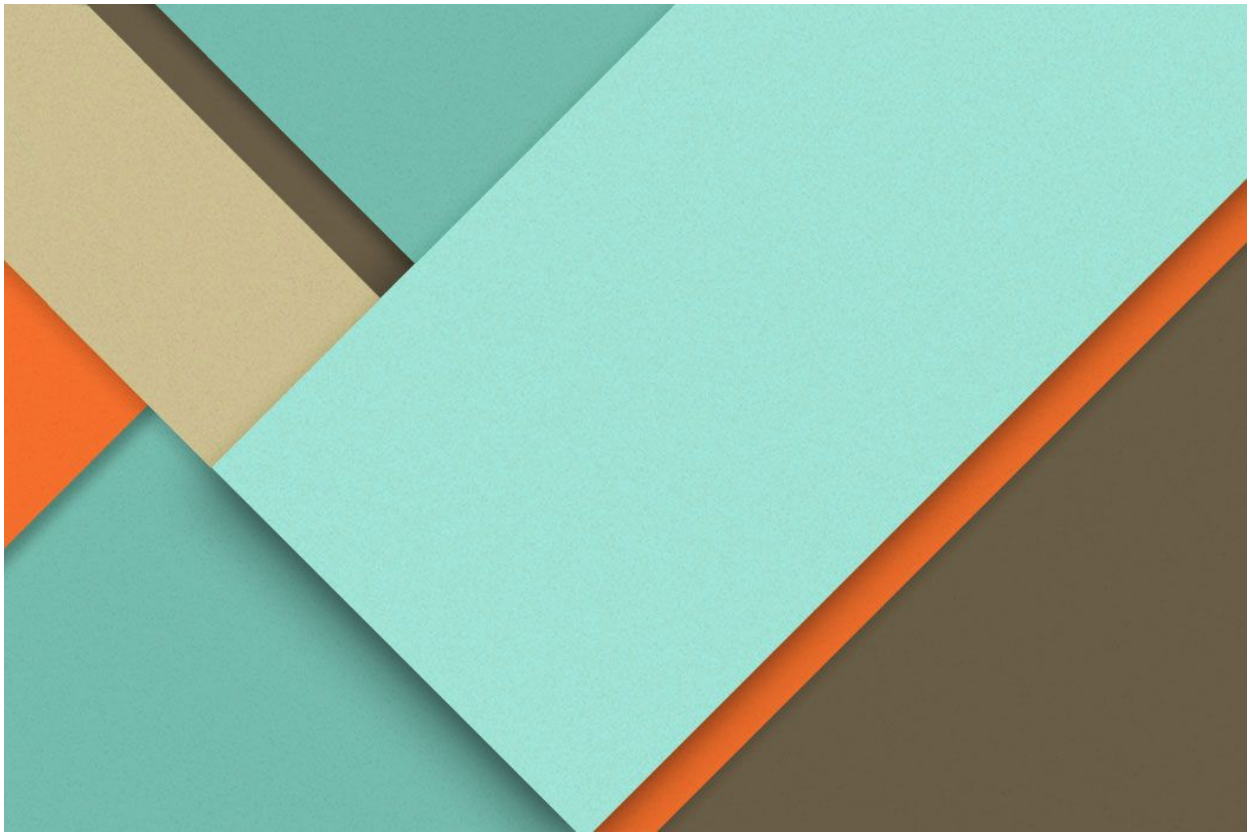

Africa Soil Property Prediction

Progress Report

Jessica Sanders

Carlos Castro

Jennifer Casper

W207

School of Information

UC Berkeley

Overview

This document provides a brief status update on the challenge of predicting 5 properties from African soil. The details of the challenge can be found here.

Status

These weeks, we've spent significant amount of time doing data exploration and experimenting with different models and transformations. It was key to gain insights on the shape of the data in order to understand potential approaches to solve the challenge.

Considering we used these weeks for planning and exploration, below we outline some of the next steps we've decided to execute before the project deadline.

Data

The dataset contains 1157 examples for train and dev data. Each example consists of 3593 numerical features, and 5 numeric predictors, each representing a different property of the soil.

Infrastructure

Given the low number of samples and high feature count in the data, before applying regression, we estimate that most of the work will be in pre-processing steps, such as transformations, feature selection, etc. This is an indication that we'll need to run multiple experiments with many pipeline combinations.

In order to experiment and make decisions effectively, we consider that one crucial step is to set up some basic infrastructure that allows us to simply specify different methods for each step of the pipeline and the infrastructure will take care of testing and comparing combinations.

We are currently laying the foundation of that infrastructure, which interestingly is simply a set of methods that provide regressors, feature selectors, etc, along with a method that orchestrates the training, testing and score reporting with the different combinations.

Methods

While we have only tried a handful of methods for the different work areas, we outline them here for reference

Standardization

Given that we're dealing with numeric features, we intend to standardize the data using a number of techniques:

- Z-score Standardization: subtracting the mean and dividing by the standard deviation
- 0-1 scaling
- Dividing by the range

Transformation

We intend to do more data analysis, and based on it experiment with transforms such as the following:

- Log-transform
- Exponentiation
- Wavelet transform
- Derivatives

Feature Selection

Given that in this dataset the number of features greatly exceeds the number of samples, feature selection will be key to the project. Some feature selection techniques we intend to apply and experiment with are:

- PCA
- K-Best
- TruncateSVR

Regression

Finally, once we've groomed the features and they are ready for regression, we'll experiment with different techniques such as:

- Support Vector Regression
- Knn regression
- Neural Nets
- Random Forests
- Lasso Regression

Other Challenges

We expect to encounter challenges with resources necessary to execute techniques, depending on the number of features chosen within the data. It will be a balancing act throughout experimentation with trade-off decisions documented.