Africa Soil Property Prediction – Kaggle Challenge

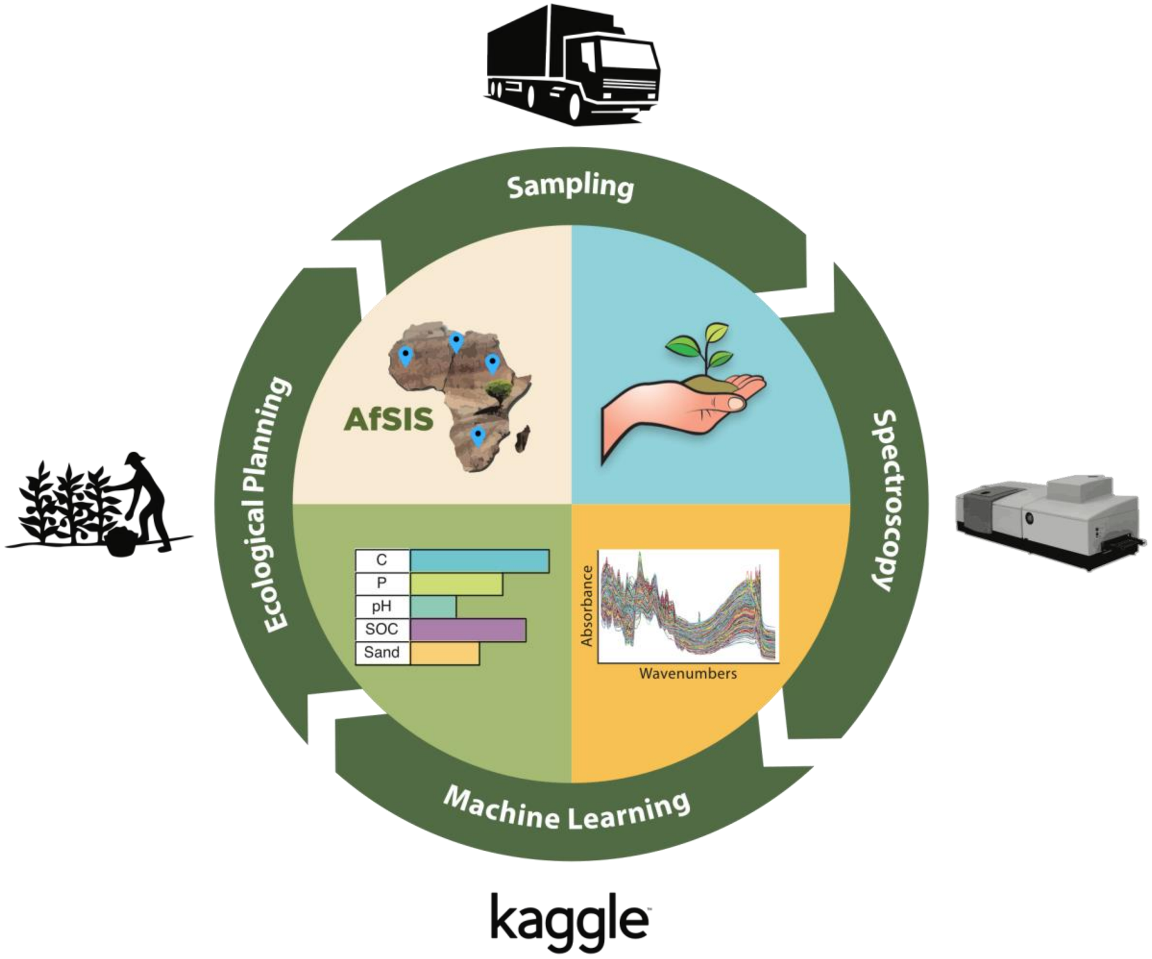
Graduate Project ECE/CS 498 DSG

Background

- Soil functional properties
  - Give information about primary productivity, nutrient and water retention, and resistance to soil erosion
  - Ecological planning
- Conventional reference tests to extract soil properties
  - Require much effort and time, slow, expensive, use chemicals
- Infrared Spectroscopy Approach
  - Have shown potential to provide a highly repeatable, rapid and low cost measurement of many soil functional properties
- Kaggle Challenge
  - Hosted a challenge to develop an accurate prediction model from spectroscopy data

Problem Statement

- 5 soil functional properties
  - Ca, P, pH, SOC, Sand
- Predict 5 target soil functional properties from infrared spectroscopy measurements

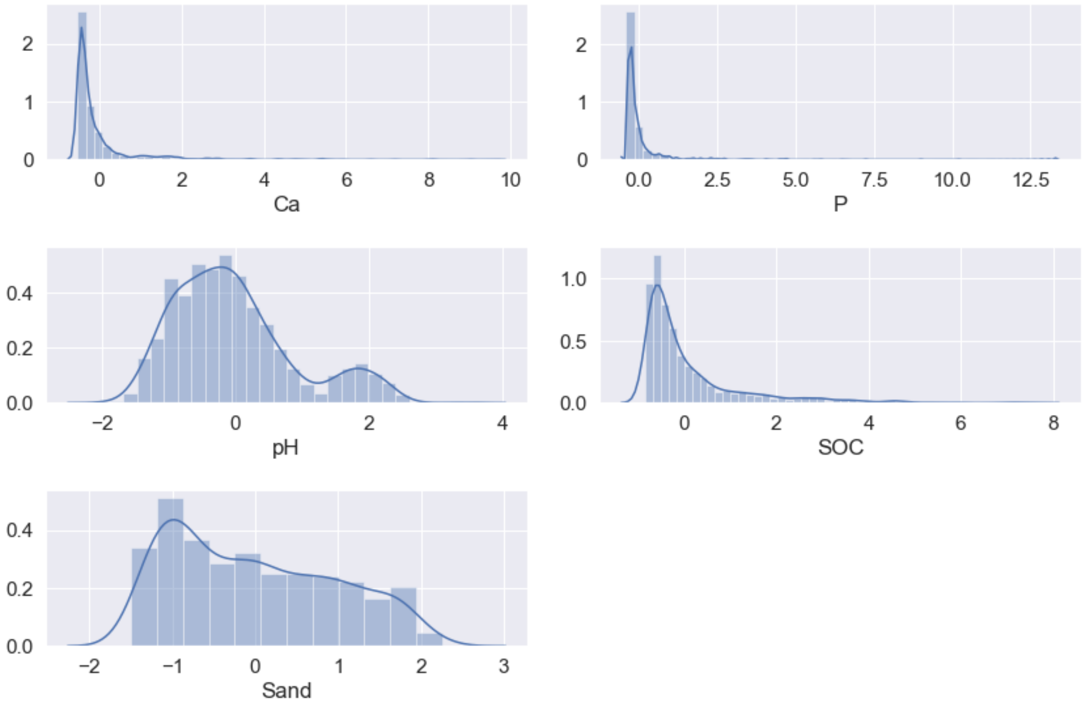


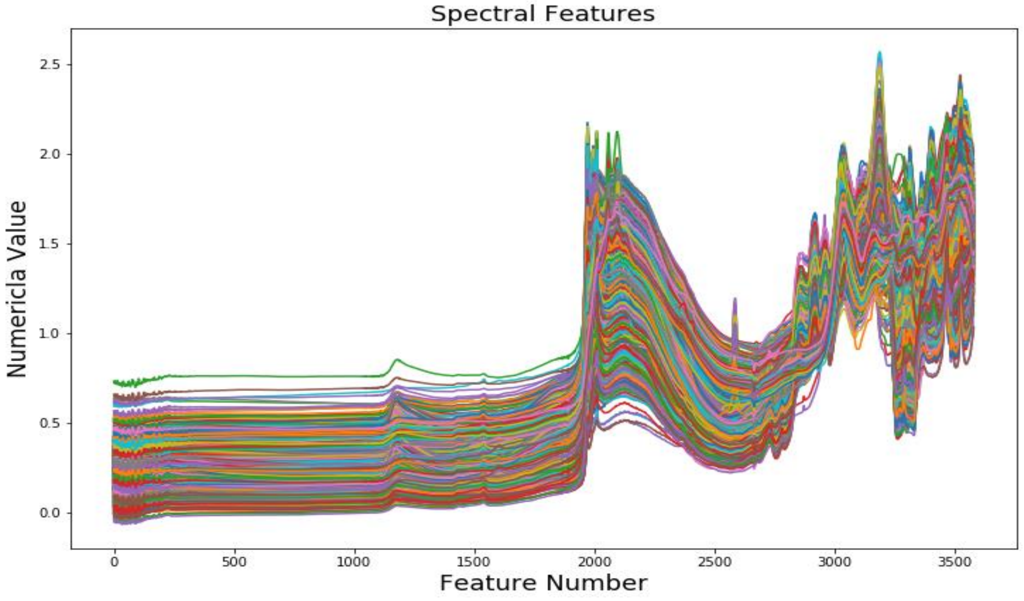
Data

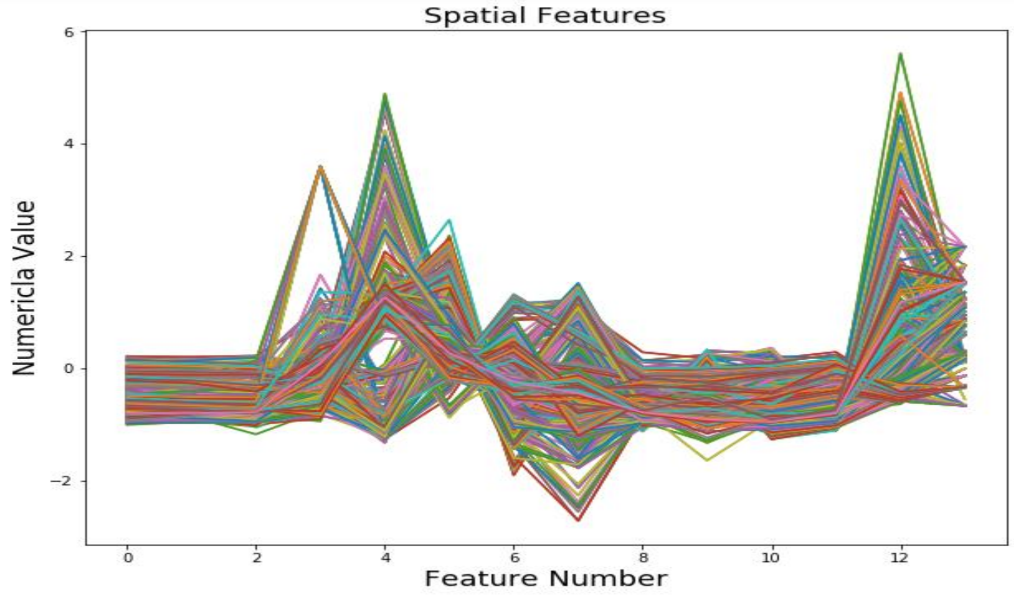
- Data Source: Kaggle competition
- Number of samples: 1157 (train data), 728 (test data)
- Number of features: 3594
  - 3578 – Spectral Features from spectroscopy – Numerical
  - 15 – Spatial Features from remote sensing data – Numerical
  - 1 – Depth of Soil – Categorical (topsoil & subsoil)
- Target Variables
  - SOC, pH, Ca, P, Sand are 5 target variables
  - Continuous/numerical in nature (Regression Problem)

Challenges

- High Dimensional Data
- Finding the relevant features
- Overfitting (Small number of samples)
- Low correlations with target variables



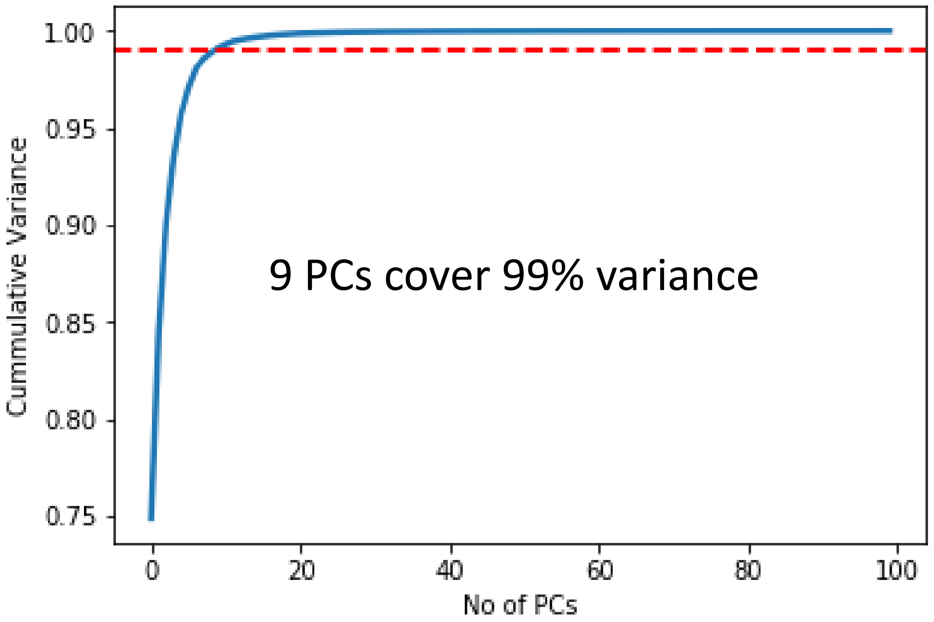


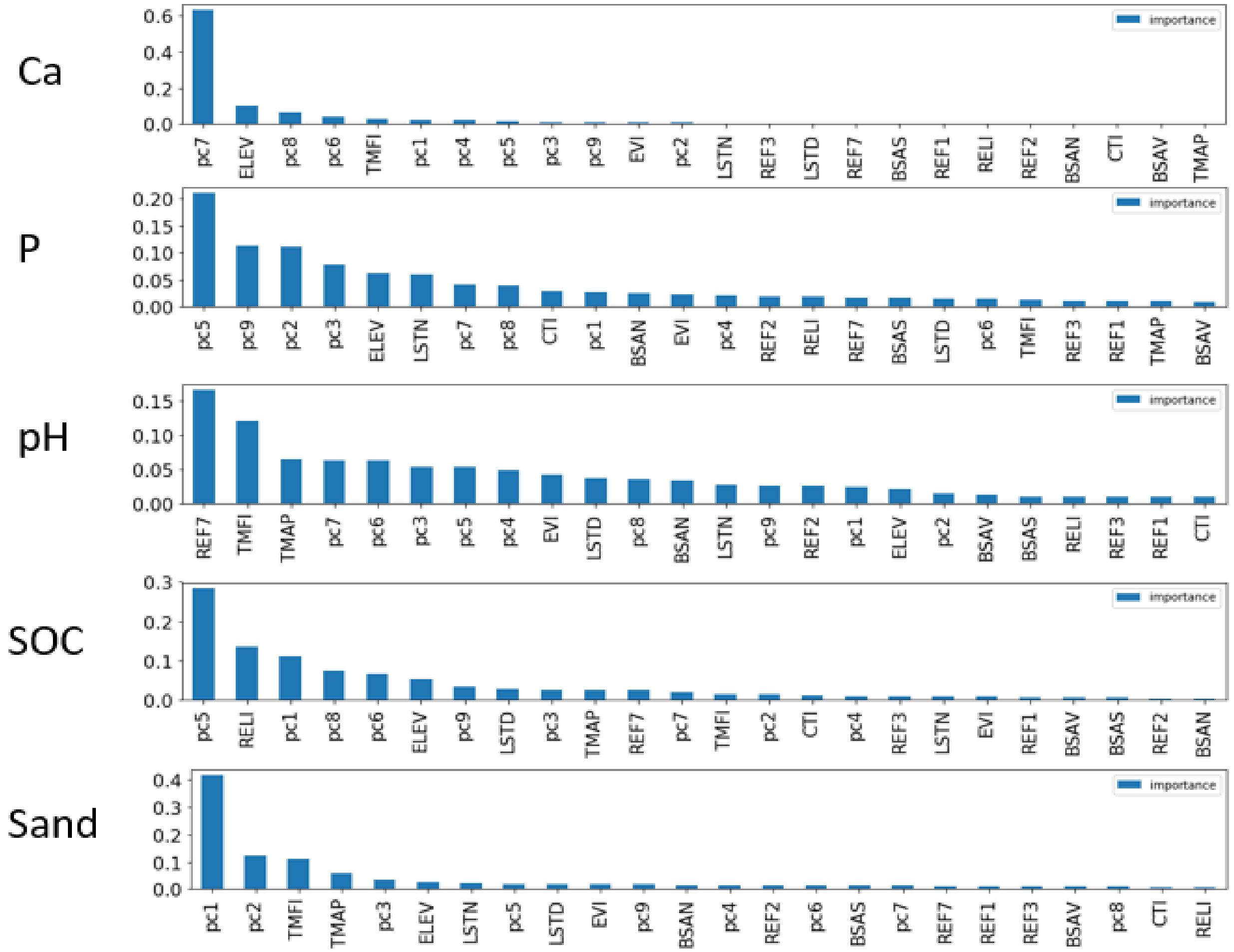


Solution approach

- Broad solution approach
  - Dimensionality Reduction
    - PCA
    - Feature Selection
  - Train Regression models
    - Linear, non-linear, ensemble
  - k-fold Cross validation
  - Model Selection
  - Hyperparameter tuning
  - Final Prediction
  - Comparison with existing solutions
  - Clustering

Solution approach

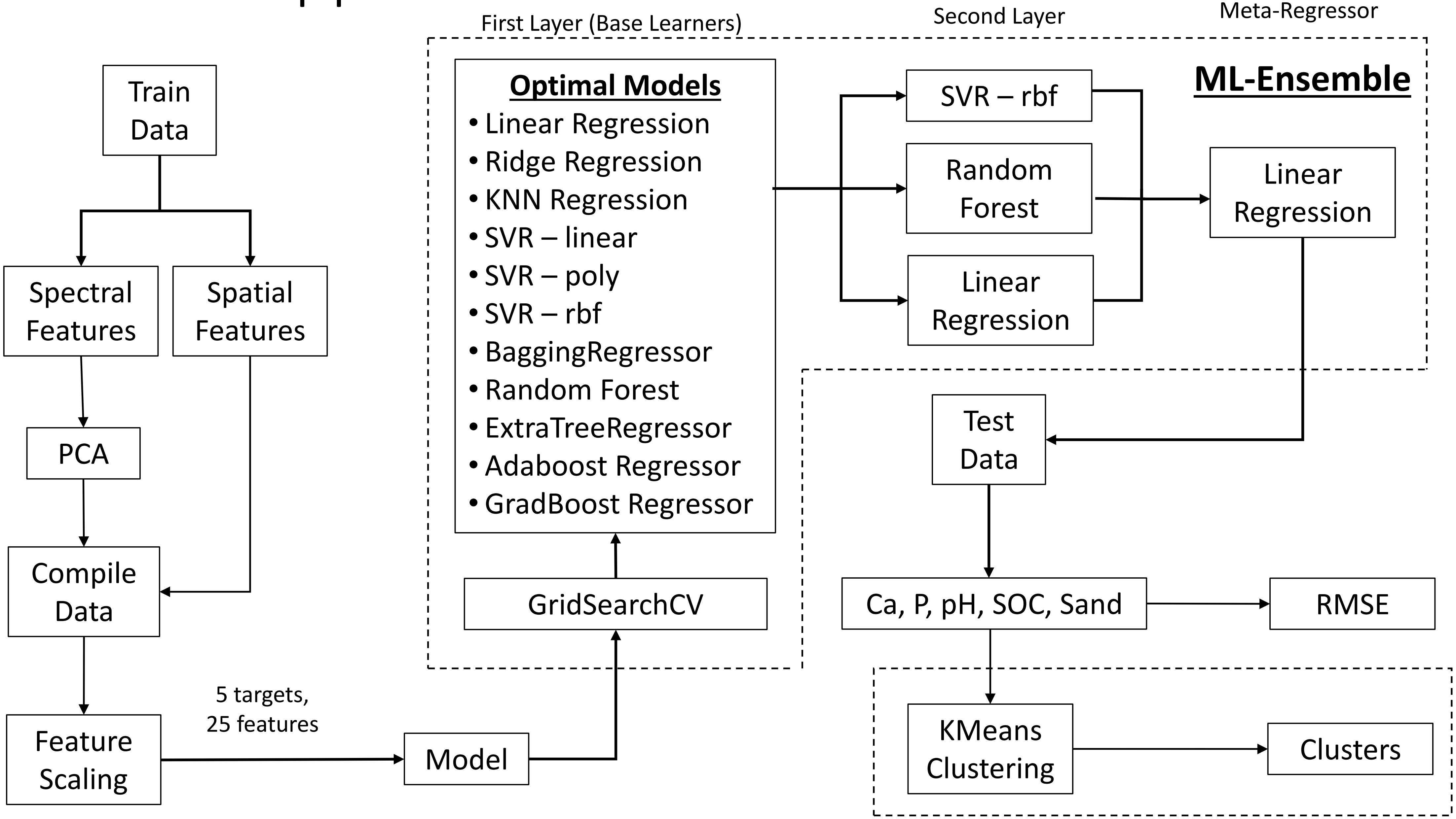




Relative Feature Importance from Random Forest Regressor

- Non-linear models performed better
- Feature Selection methods didn't work. Slightly improved results in case of linear models.
- PCA already reduced 3578 features to 9 PCs
- Decided to go with 25 features

Solution approach



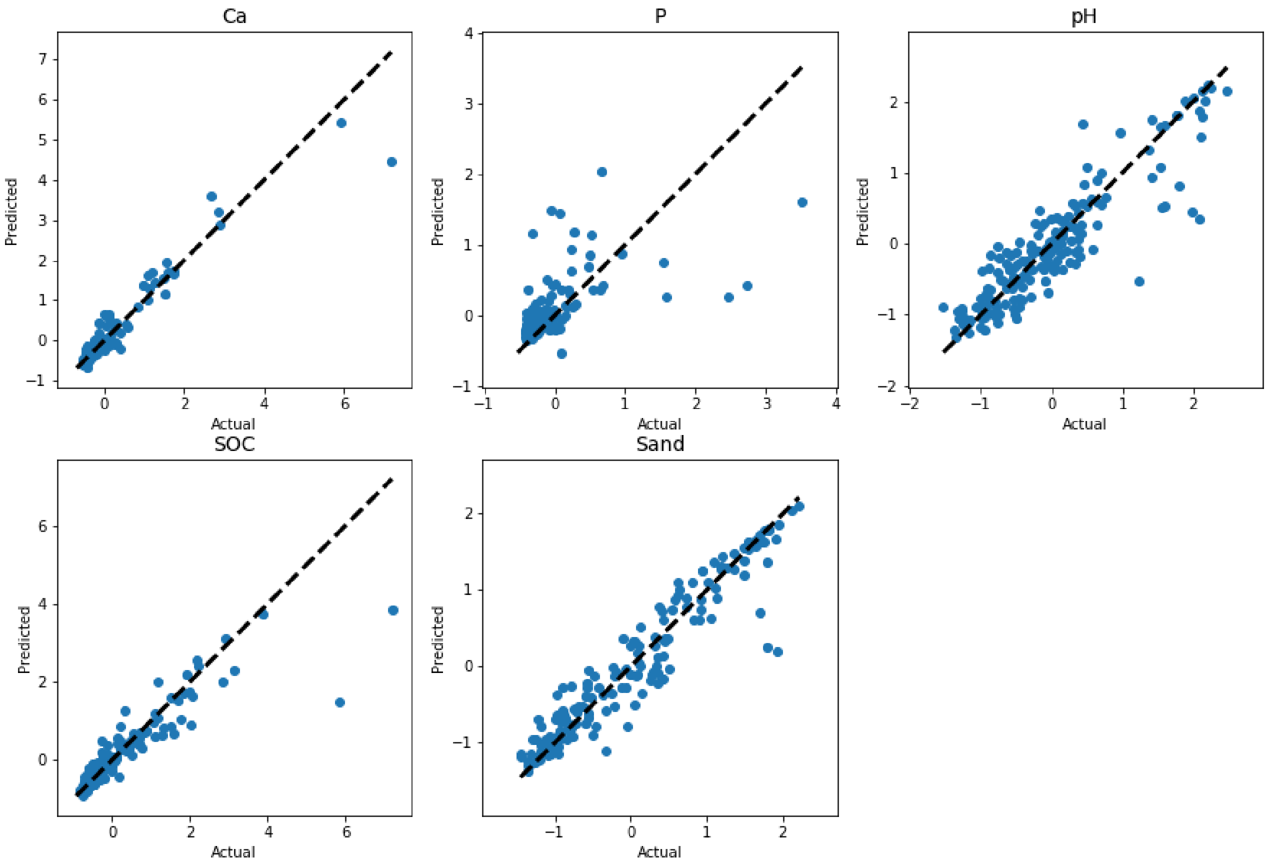
Baseline approach – existing methods

- Kaggle top solutions mentioned overfitting as a problem
- There is no signal model that provided best results.
- Some sort of ensemble was created from best performing models.
- Mean column wise RMSE was the scoring criteria

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

Regression Results

RMSE Scores	Ca	P	pH	SOC	Sand
LinearRegression	0.474129	0.509118	0.514583	0.626524	0.577974
Ridge	0.477349	0.509295	0.513086	0.627081	0.574445
KNN	0.416879	0.618948	0.524237	0.625840	0.386477
SVR_Linear	0.520894	0.490988	0.514074	0.662105	0.575716
SVR_Poly	0.344661	0.472887	0.456611	0.510234	0.426289
SVR_RBF	0.315481	0.496421	0.416152	0.432565	0.281697
BaggingRegressor	0.460271	0.360970	0.494940	0.604393	0.395441
RandomForestRegressor	0.306916	0.383201	0.454969	0.607998	0.356133
ExtraTreeRegressor	0.307119	0.387393	0.419065	0.522881	0.341521
AdaBoostRegressor	0.382999	0.395236	0.435237	0.492157	0.379222
GradientBoostingRegressor	0.385355	0.372973	0.451079	0.559635	0.342491
ML-Ensemble	0.285576	0.438138	0.398420	0.497032	0.313649



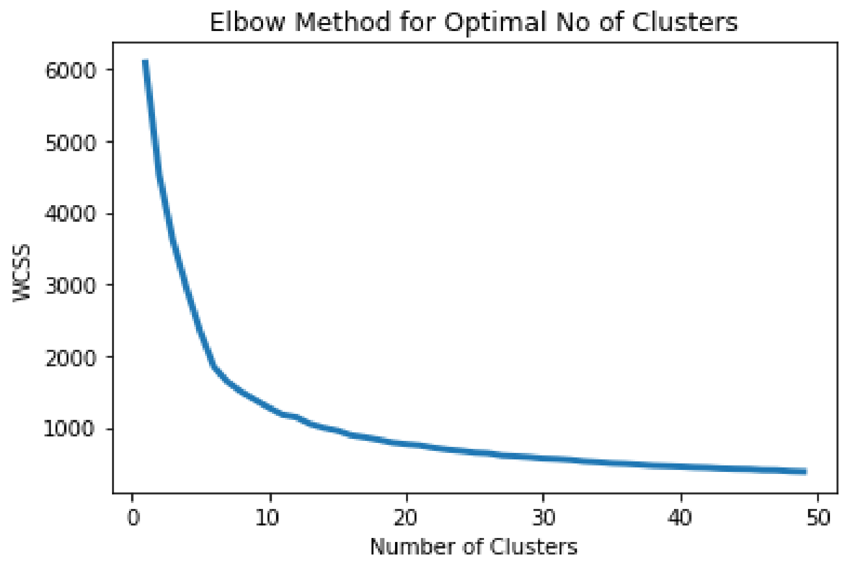
Kaggle Test Data:

Top MCRMSE Score = **0.46892**, Our MCRMSE Score = **0.5605**

Relationships among Target Variables

Clustering Results:

- There seemed to be 8 optimal clusters
- Grouped data by clusters and calculated means
- Gave interpretable Description for each cluster



Ca	P	pH	SOC	Sand	Cluster	Description
-0.377770	-0.219385	-0.449092	-0.388017	-0.074105	0	Low-Ca Low-P Neg-pH Low-SOC Mid-Sand
-0.040148	-0.124162	-0.574790	2.896514	-0.928389	1	Low-Ca Low-P Neg-pH High-SOC Low-Sand
0.747014	0.068563	1.537137	-0.304553	-0.389595	2	Mid-Ca Mid-P Pos-pH Low-SOC Mid-Sand
0.258263	3.141500	0.430292	0.317031	-0.174210	3	Mid-Ca Mid-P Pos-pH Mid-SOC Mid-Sand
4.859856	-0.190742	1.900154	2.483493	-0.967938	4	High-Ca Low-P Pos-pH High-SOC Low-Sand
-0.404145	-0.172640	-0.130470	-0.609670	1.282962	5	Low-Ca Low-P Neg-pH Low-SOC High-Sand
-0.137051	-0.209962	-0.449387	0.459577	-0.983348	6	Low-Ca Low-P Neg-pH Mid-SOC Low-Sand
1.040350	9.840077	0.638322	2.813268	-0.654799	7	Mid-Ca High-P Pos-pH High-SOC Low-Sand