Africa Soil Property Report – Code AI

Djordje Grozdic[1],Andjela Todorovic[2]
[1] Faculty of Electronic Engineering, University of Belgrade, Data Scientist at Grid Dynamics
[2] Faculty of Sciences and Mathematics, University of Nis, Software Engineer at Createsi GmbH

1 Project Goal

Low cost and rapid analysis of soil samples using infrared spectroscopy provide new opportunities for predicting soil functional properties. These soil functional properties are directly related to a soil's capacity to support essential ecosystem services such as primary productivity, nutrient and water retention, and resistance to soil erosion. If successful, it will provide digital mapping of soil properties in data sparse regions such as Africa and will improve mangament of agricultural and natural resources.

1.1. Problem Statment

To predict physical and chemical properties of soil using spectral measurements. Regression models for high dimensional data needs to be developed in order to predict five output variables namely Ca, P, pH, SOC and Sand content.

2 Data Description and Preprocessing

Training dataset contains 1,158 soil training samples and 789 testing samples. Each sample is represented by 3,564 features of mid-infrared absorbance measurements.

2.1. Evaluaion Metric

Each model is evaluated based on root mean squared error calcuated for each output variable. To calculate overall performance of a model, mean of RMSE errors of each output variables is taken. This is called as mean columnwise root mean squared error (MCRMSE)

2.2. Filtering the data

Both Savitzky-Golay and Convolution Box filtering has been applied to the data, and they have shown pretty similar results in signal filtering. (Figure 1)

2.3. Feature selection

To reduce the high dimension of data, feature selection was tested based on principal component analysis, standard deviation of original features and selection via random forest. None of the used techniques have shown expected performance, so due to the nature of the data we have decided only to perform cleaning and filtering.

3 Modeling Methods

As output variables are not correlated, independent models are created for each output separately. Different regression models have been discussed and compared, and gave the following results:

3.1. Gradient Boosting regression

MCRMSE: ([0.3848468514627918, 0.931737 8087122214, 0.48177209927763454, 0.45641 7834505394, 0.3880860199250725], 0.52857 21227766229)

3.2. Random Forrest regression

MCRMSE: ([0.4476500411683135, 0.964488 554989961, 0.49204065304193234, 0.521139 2510360906, 0.4459615792502725], 0.57425 6015897314)

3.3. Linear regression

MCRMSE: ([0.6502884935376078, 1.280896 7267503895, 0.540753018681597, 0.5692914 433449698, 0.608343151086622], 0.7299145 666802372)

3.4. Ridge regression

MCRMSE: ([0.5058262521723355, 0.896838 7427403517, 0.33159895094096886, 0.36549 56281119872, 0.3205795858422661], 0.4840 678319615819)

3.5. MLP regression

MCRMSE: ([0.5294926684394562, 1.043041 9832492213, 0.46695106406808223, 0.45088 061298507437, 0.4721632745121915], 0.592 5059206508051)

3.6. SVM regression

MCRMSE: ([0.5028311009894434, 1.038609 6955177184, 0.4518317914702387, 0.585172 6881454353, 0.4144762177200273], 0.59858 42987685727)

Observing the training error, it becomes clear that linear dimension reduction via PCA is not useful and suggests that data is more complicated in lower dimensions. Phosphorus training error for all models are higher than other variables indicating presence of noise in its measurement. So, methods robust against noise such as ensembling might improve the prediction for this output variable. Also, SVM and random forest performs equally well in training error. But SVM beats random forest in testing error proving its generalization capabilties in higher dimensions. (Figures 2 and 3)

4 Results and future notices

Results shows that SVM generalises well in higher dimensions. Also, there are some disparities between training error and testing error suggesting overfitting in modeling. Ensembling of models helped in reducing that disparity, but its useful only if computational cost is less. As for future work, new features will be considered based on the domain knowledge of spectral methods such as first derivative and second derivate of spectra.

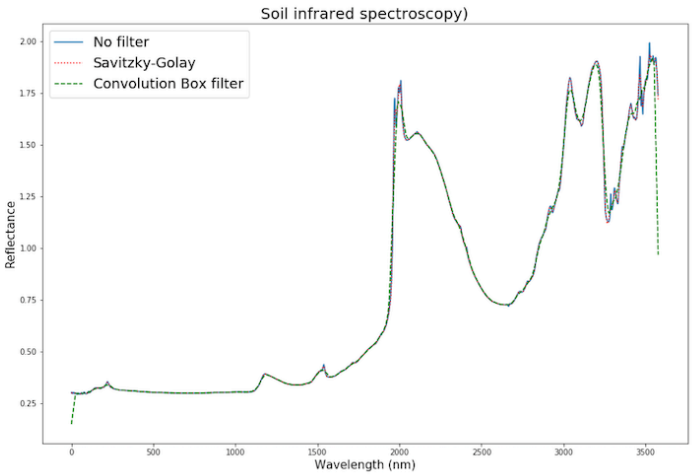

Figure 1. Filtering the data

	Ca	P	pH	SOC	Sand
Gradient Boosting Regression	0.384847	0.931738	0.481772	0.456418	0.388086
Random Forest Regression	0.447650	0.964489	0.492041	0.521139	0.445962
Linear Regression	0.650288	1.280897	0.540753	0.569291	0.608343
Ridge Regression	0.505826	0.896839	0.331599	0.365496	0.320580
MLP Regression	0.529493	1.043042	0.466951	0.450881	0.472163
SVM Regression	0.502831	1.038610	0.451832	0.585173	0.414476

Figures 2 and 3 (top and bottom).Comparing the models