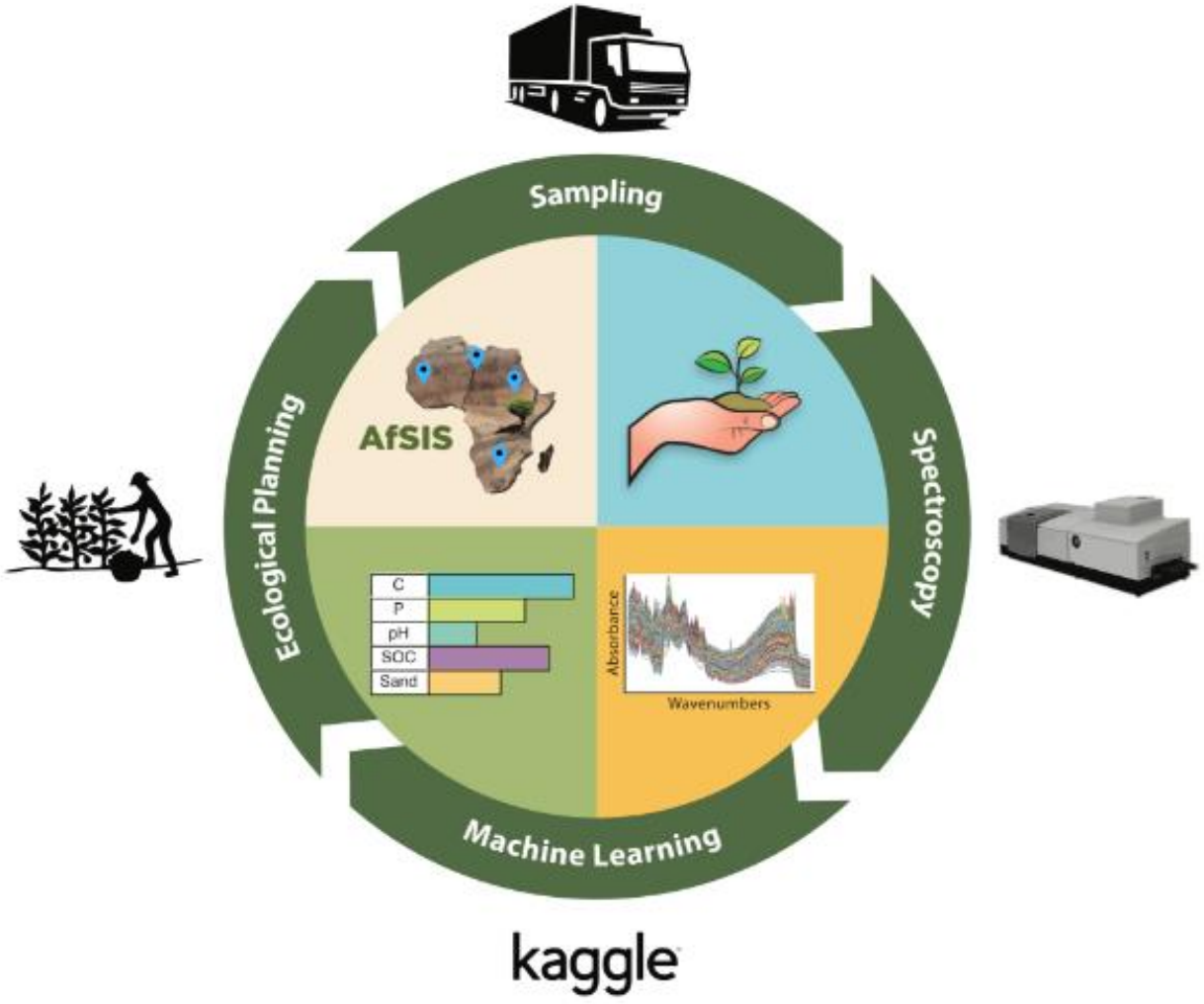
Africa Soil Property Prediction Challenge

(AFSIS)

A CASE STUDY REPORT



Mayank Aggarwal

Mayank953@gmail.com| GITHUB

ABOUT AFSIS

1. Problem Statement

This competition asks you to predict 5 target soil functional properties from diffuse reflectance infrared spectroscopy measurements.

2. INTRODUCTION TO PROJECT

A) CASE STUDY EXPLANATION

Advances in rapid, low cost analysis of soil samples using infrared spectroscopy, georeferencing of soil samples, and greater availability of earth remote sensing data provide new opportunities for predicting soil functional properties at unsampled locations. Soil functional properties are those properties related to a soil's capacity to support essential ecosystem services such as primary productivity, nutrient and water retention, and resistance to soil erosion.

Digital mapping of soil functional properties, especially in data sparse regions such as Africa, is important for planning sustainable agricultural intensification and natural resources management.

B) DOMAIN KNOWLEDGE

Diffuse reflectance infrared spectroscopy has shown potential in numerous studies to provide a highly repeatable, rapid and low cost measurement of many soil functional properties. The amount of light absorbed by a soil sample is measured, with minimal sample preparation, at hundreds of specific wavebands across a range of wavelengths to provide an infrared spectrum (Fig. 1). The measurement can be typically performed in about 30 seconds, in contrast to conventional reference tests, which are slow and expensive and use chemicals.

Conventional reference soil tests are calibrated to the infrared spectra on a subset of samples selected to span the diversity in soils in a given target geographical area. The calibration models are then used to predict the soil test values for the whole sample set. The predicted soil test values from georeferenced soil samples can in turn be calibrated to remote sensing covariates, which are recorded for every pixel at a fixed spatial resolution in an area, and the calibration model is then used to predict the soil test values for each pixel. The result is a digital map of the soil properties.

C) EVALUATION

Submissions are scored on MCRMSE (mean column wise root mean squared error):

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

where y and $\hat{y}$ are the actual and predicted values, respectively.

For each row in the dataset, the submission file should contain an identifier column (PIDN) and 5 prediction columns: Ca, P, pH, SOC, and Sand. PIDN, the sample identifier, should be copied from the first column of test data file. Ca, P, pH, SOC, and Sand are soil properties whose values you must predict.

D) DATASET PROVIDED

We were provided with 3 DataSets so as to begin our Analysis with. The 3 Data Sets Are as given below:-

- **train.csv** - the training set has 1158 rows.
- **test.csv** - the test set has 728 rows.
- **sample_submission.csv** - all zeros prediction, serving as a sample submission file in the correct format.

3. EXPLORATORY DATA ANALYSIS & DATA EXPLORATION

SOC, pH, Ca, P, Sand are the five target variables for predictions. The data have been monotonously transformed from the original measurements and thus include negative values.

Data Fields

- **PIDN**: unique soil sample identifier
- **SOC**: Soil organic carbon
- **pH**: pH values
- **Ca**: Mehlich-3 extractable Calcium
- **P**: Mehlich-3 extractable Phosphorus
- **Sand**: Sand content
- **m7497.96 - m599.76**: There are 3,578 mid-infrared absorbance measurements. For example, the "m7497.96" column is the absorbance at wavenumber 7497.96 cm-1. We suggest you to remove spectra CO2 bands which are in the region m2379.76 to m2352.76, but you do not have to.

- **Depth**: Depth of the soil sample (2 categories: "Topsoil", "Subsoil")

We have also included some potential spatial predictors from remote sensing data sources:-

1. **BSA:** average long-term Black Sky Albedo measurements from MODIS satellite images (BSAN = near-infrared, BSAS = shortwave, BSAV = visible)
2. **CTI:** compound topographic index calculated from Shuttle Radar Topography Mission elevation data
3. **ELEV:** Shuttle Radar Topography Mission elevation data
4. **EVI:** average long-term Enhanced Vegetation Index from MODIS satellite images.
5. **LST:** average long-term Land Surface Temperatures from MODIS satellite images (LSTD = day time temperature, LSTN = night time temperature)
6. **Ref:** average long-term Reflectance measurements from MODIS satellite images (Ref1 = blue, Ref2 = red, Ref3 = near-infrared, Ref7 = mid-infrared)
7. **Reli:** topographic Relief calculated from Shuttle Radar Topography mission elevation data
8. **TMAP & TMFI:** average long-term Tropical Rainfall Monitoring Mission data (TMAP = mean annual precipitation, TMFI = modified Fournier index)

INITIAL INSIGHTS-

One point of interest here is that it is suggested to remove the Spectra CO2 band given in the range.
We will try to obtain result by removing & keeping them to see whether it is advantageous or Disadvantageous. Further Spectral bands property include negative values as they have been transformed monotonously.
The provided Spacial predictors can either make the model more robust or add noise to it as the main property of the soil is predicted based on its Spectral Nature.

BART Example

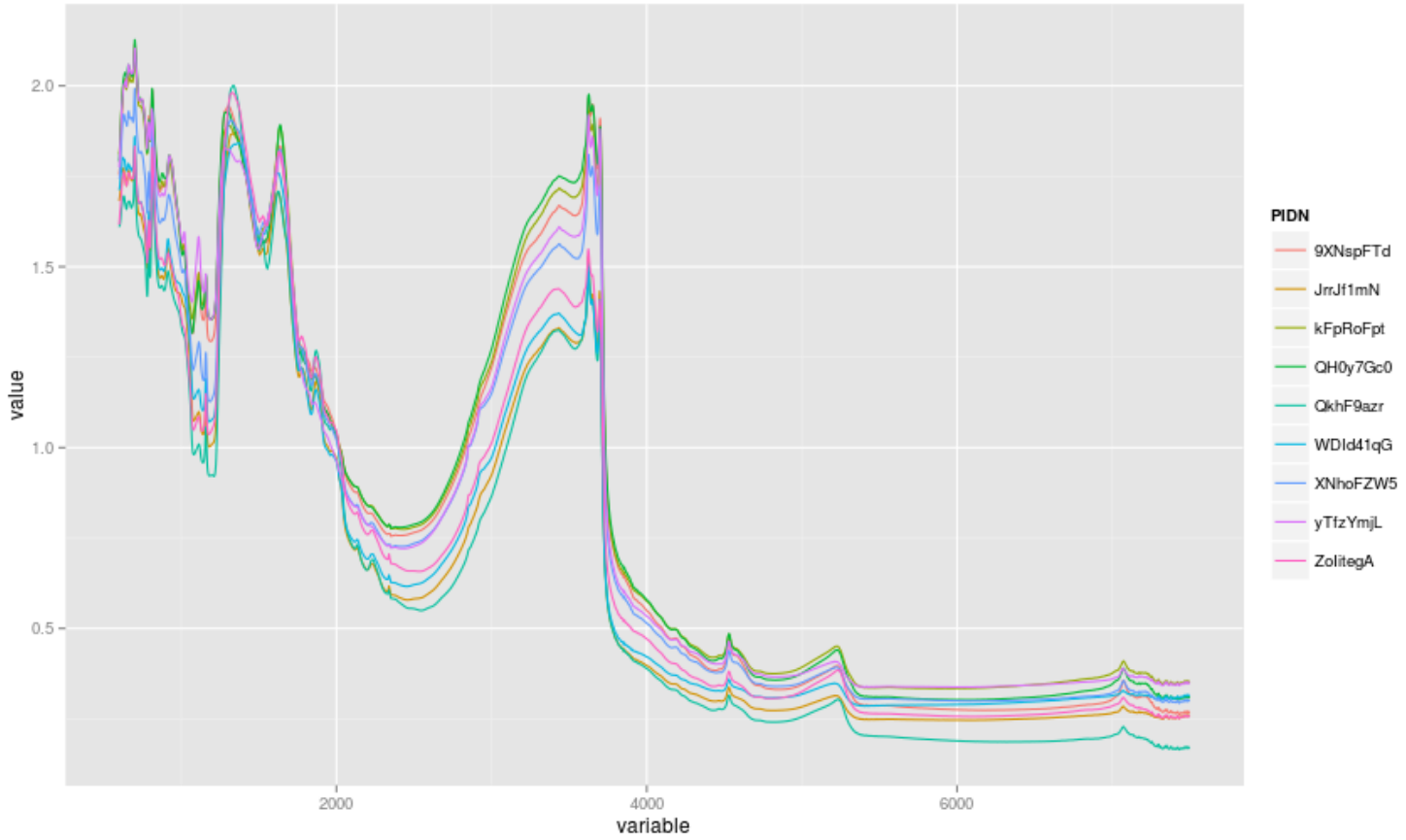
We have been provided with the Bayesian Additive Regression Trees solution.
The major thing to note here is that the first Derivative of Bands values is Taken so as to reduce the noise. Alternatively, many other filters based on the domain knowledge can be applied on the data to either decrease or increase the noise present.

Derivative is a high pass filter. It will accentuate most noise, if information content is mostly lower frequency.

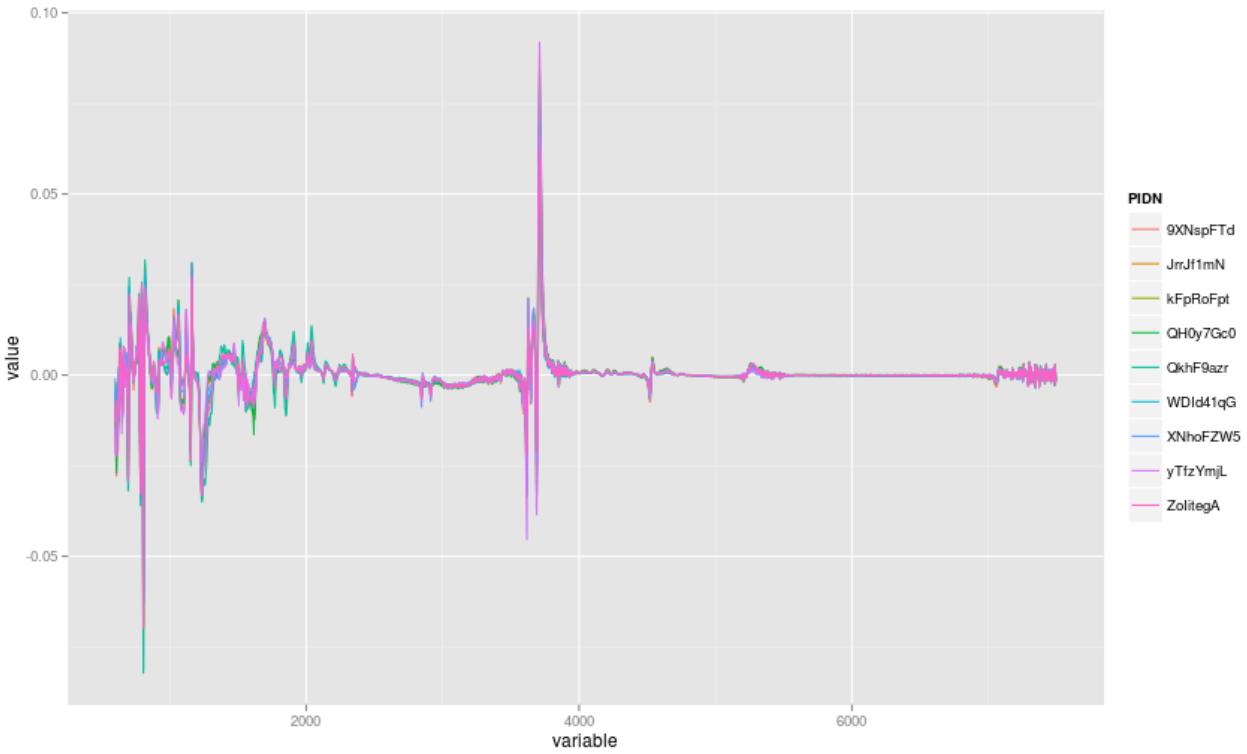
The derivative gets rid of smoothly varying components in the spectra & It will remove the envelope, leaving only tiny peak-like details.

A comparison is shown below on the first derivative of the MIR measurements & taking normal / Raw measurements.

RAW :-



FIRST DERIVATIVE



MISSING VALUE TREATEMENT

```
In [62]: train.columns[train.isnull().any()]
Out[62]: Index([], dtype='object')

In [63]: test.columns[train.isnull().any()]
Out[63]: Index([], dtype='object')
```

As it can be seen that there are no Missing values in the Dataset so no Treatment for that is required.

```
In [73]: train= pd.read_csv("training.csv")

In [74]: train.groupby('Depth').size()
Out[74]:
Depth
Subsoil    576
Topsoil    581
dtype: int64

In [75]: test.groupby('Depth').size()
Out[75]:
Depth
Subsoil    359
Topsoil    368
dtype: int64
```

Further the **Depth** variable present in the Data is Categorical & we will change it to float by using 1 & 0 to represent the category. New Variable is Depth1.

This will further help us in Apllying **PCA** as PCA requires all variable to be of Numeric Nature & not Categorical.

We have also used Derivative Method for Denoising of the MRI measurement as suggested in The BART algorithm.

We have also chosen to not select the variables other than MRI measurement as they were the main focus of study & soil property depend mainly on the reading obtained for different wavelength.

Further, a improved Domain Knowledge will help us provide better understanding on choice of features & the Denoising Filter etc.

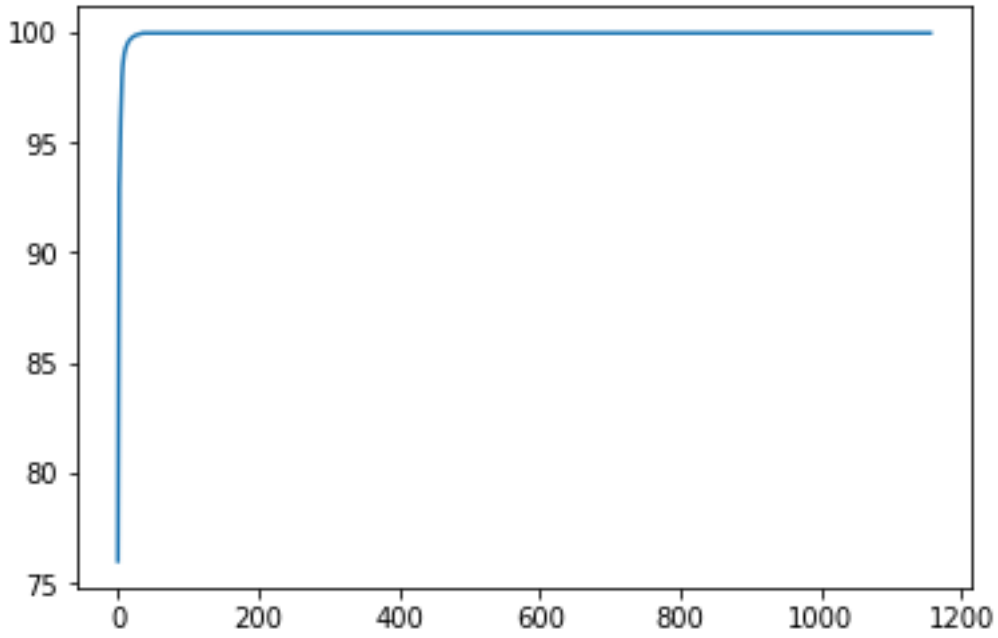
SUMMARY OF TARGET VARIABLE-

	TMFI	Ca	P	pH	SOC	\
count	1157.000000	1157.000000	1157.000000	1157.000000	1157.000000	
mean	0.746303	0.006442	-0.014524	-0.028543	0.080414	
std	0.825242	1.070541	0.995469	0.920224	1.141989	
min	-0.862741	-0.535828	-0.418309	-1.886946	-0.857863	
25%	0.056843	-0.451077	-0.345681	-0.717841	-0.615639	
50%	0.729111	-0.348682	-0.269595	-0.175376	-0.349974	
75%	1.414215	-0.042654	-0.089755	0.376442	0.275121	
max	2.976315	9.645815	13.266841	3.416117	7.619989	

	Sand
count	1157.000000
mean	-0.012646
std	0.988520
min	-1.493378
25%	-0.899649
50%	-0.134651
75%	0.786391
max	2.251685

[8 rows x 3598 columns]

Further Pre-processing has been done in the .py which is added in comments & is self-Explanatory.



MODEL BUILING

We have a normal test & train Data wherein We will be carrying out the model Training & testing. As we have kaggle Solution Checker available so we will train our Model on our whole Train Data & Cross Verify the result on the basis of Values provided by Kaggle on our Test Data.

There are various Pre-processing which can be done to our data so as to obtain different Result after which we may select the one with the best results.

In the Train Data, The result will depend on :-

1. Whether CO2 band which were suggested to removed are taken or not.
2. Whether Derivative is taken so as to denoise our MIR measurement or any other Filter is taken.
3. Whether or not locational Variable are taken/considered during Training.
4. Whether the training is done on Model after Carrying out the PCA as there are a large number of Variables.

After Running the model the best result were obtained by considering the below 2 models.

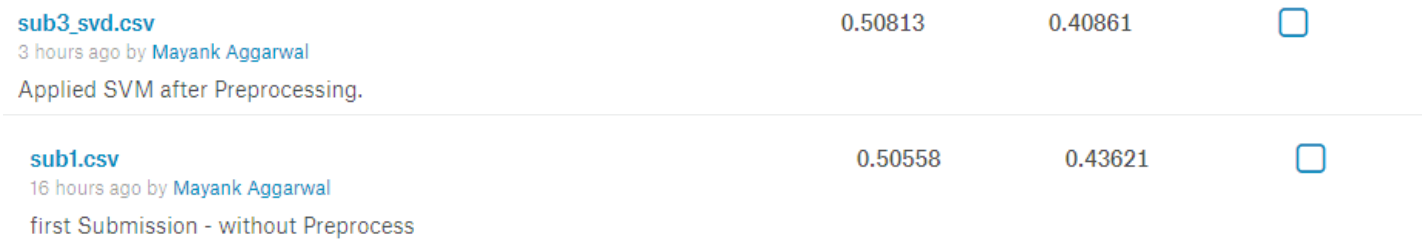
1. Random Forest
2. Svm

1. Random Forest :-

Random forest was initially applied onto the Train. It trained quite fast & the result obtained was.

2.SVM

Next , the SVM model was applied & the result obtained was as show.



The first score shown is the Evaluation metric on private leaderboard While the other one is the score on public Leaderboard.

Conclusion

By simple analysis & EDA, we were able to obtain a good Evaluation metric i.e. in Top 150 for both the scores .Further steps to apply more complex Neural Network is being Carried out so as to obtain a better accuracy.Also Domain Knowledge will be applied more extensively so as to extract more features which can get us better results.