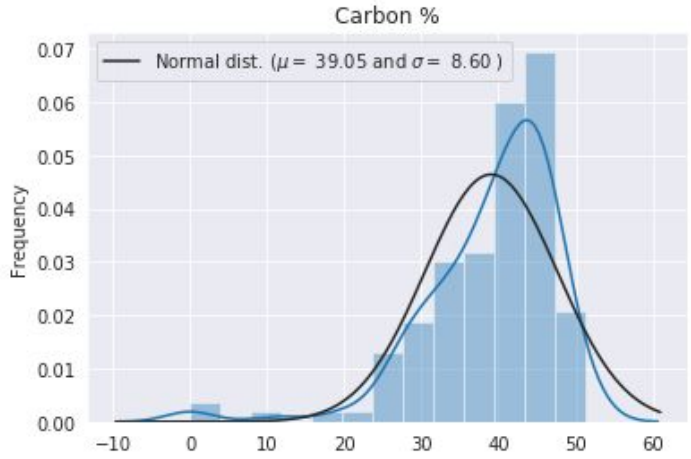
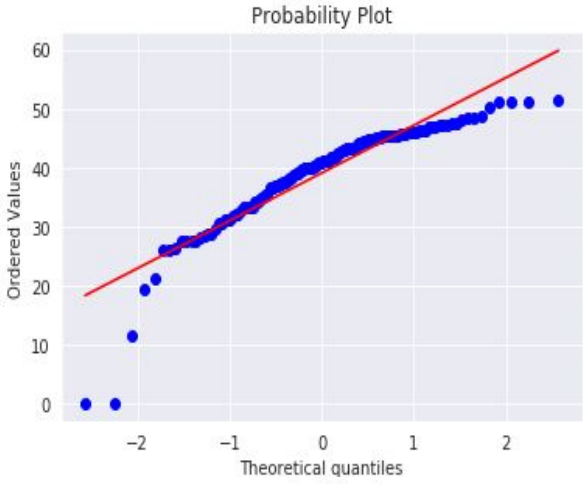
Prakriti Data Analytics Report

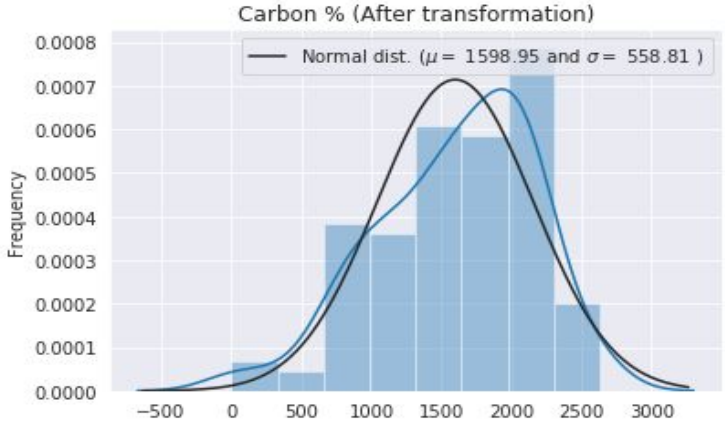
Team Name - KingSlayer
Members - Anik Dutta(E&ECE, IIT Kharagpur, Ph - 7365058272)
Samanway Sadhu(E&ECE, IIT Kharagpur, Ph - 9476306375)

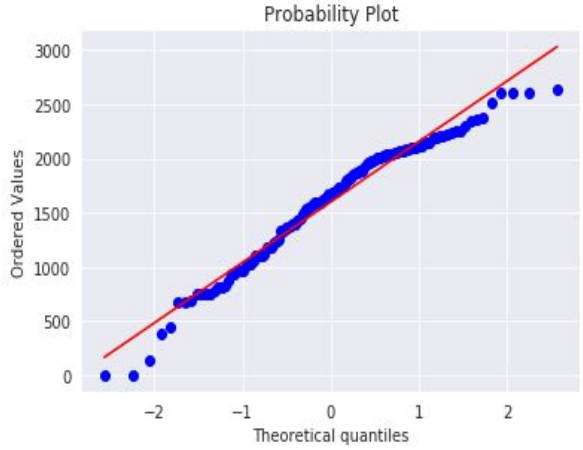
Pre-Processing

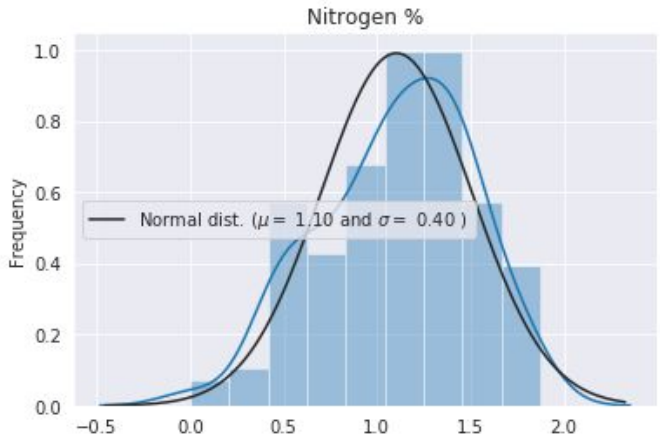
We first calculated the skewness and kurtosis of the target variables i.e. ‘TC (%)’ and ‘TN (%)’ and found they are both negatively(left) skewed with values -1.786 and -0.37 respectively. They had the kurtosis values 5.072 for Carbon and -0.41 for Nitrogen. We observed from the values and plotted graphs that that ‘TC (%)’ had a high skewness and kurtosis and needs to be pre-processed or else our model accuracy may be less. We used some of the common transformations used to process left-skewed data i.e. square, cube root, logarithmic and box-cox. We found that the square transformation gave the best results with new skewness and kurtosis values being -0.56 and -0.09 respectively. Since skewness and kurtosis of ‘TN (%)’ weren’t that high and the plots indicated that it was close to normal distribution, hence no transformation was used on them.

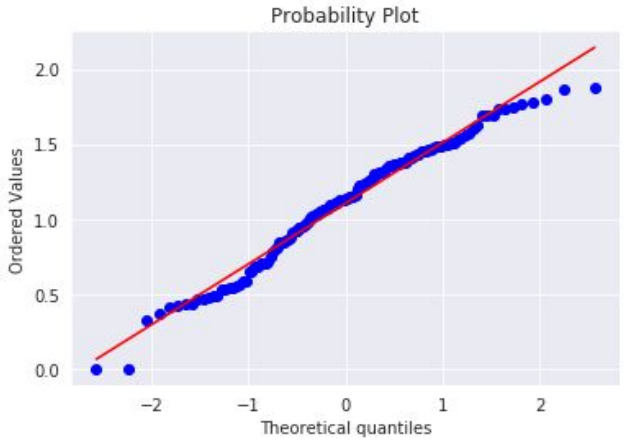












We then measured the skewness of the elemental data and found most of them had high skewness. We applied boxcox transformation on all the columns having absolute skewness value higher than 0.75 and removed their skewness.

	Skew
Si	2.778
Zn	2.144
Ti	1.693
Fe	1.326
Rb	1.186
Al	1.119
K	1.057
Ca	1.036
Mn	1.029
S	0.983
Sr	0.787

Before transformation

	Skew
Zn	1.512
K	0.283
Rb	0.232
Si	0.176
Ti	0.150
Al	0.136
Sr	0.095
S	0.038
Fe	0.010
Mn	-0.119
Ca	-0.223

After Transformation

We had 1 row of missing data where the target variable values were ‘n.d.’ so we simply dropped it.

Models and Experiments

We did 3 fold cross validation with the data. We experimented with the models Lasso Regression, Elastic Net Regression, Kernel Ridge Regression, Gradient Boosting Regressor, XGBoost and LightGBM. We also used a combination of the models like the average of models, stacked averaging regressor and voting regressor. Our initial model overfitted as we got a huge difference between the train RMSE and validation RMSE between the different folds of the data. We then started experimenting by changing the hyperparameters involved with each of the models. By hit and trial we arrived at a set of hyperparameters associated with each model which gave us almost equal train RMSE and validation RMSE values between the different folds of data indicating the models aren’t overfitting anymore. We found that in each of the cases of elemental data, spectroradiogram data and combination of both, we found the XGBoost model gave us the best results for the Carbon % in soil. We then repeated a similar experiment for the soil Nitrogen content and again arrived at the same conclusion that with XGBoost giving the best results.

Results

Carbon content of soil

- Elemental Data - XGBoost model
  3 fold validation scores -
  Train Score - [291.59363389 254.41492706 262.75286137]
  Validation Score - [374.83629686 433.39025346 518.78836308]
  RMSE on prediction - 4.810784582327536

- Spectroradiogram Data - XGBoost model
  3 fold validation scores -
  Train Score - [317.96870095 307.72115573 213.20143691]
  Validation Score - [276.49340911 317.82904647 573.96595678]
  RMSE on prediction - 5.275596277762709

- Combined Data - XGBoost model
  3 fold validation scores -

Train Score - [292.53954819 255.57826961 217.85789269]
Validation Score-[229.98331289 376.68307123 577.18478972]
RMSE on prediction - 4.959483783303534

Nitrogen Content of soil

- Elemental Data - XGBoost model
  3 fold validation scores
  Train Score - [0.30837253 0.31326145 0.26434348]
  Validation Score - [0.47693553 0.32947666 0.46893786]
  RMSE on prediction - 0.2927709176583464

- Spectroradiogram Data - XGBoost model
  3 fold cross validation scores
  Train RMSE - [0.31299339 0.38641805 0.30564135]
  Validation RMSE - [0.47964838 0.36446268 0.49786085]
  RMSE on prediction - 0.3359674924659507

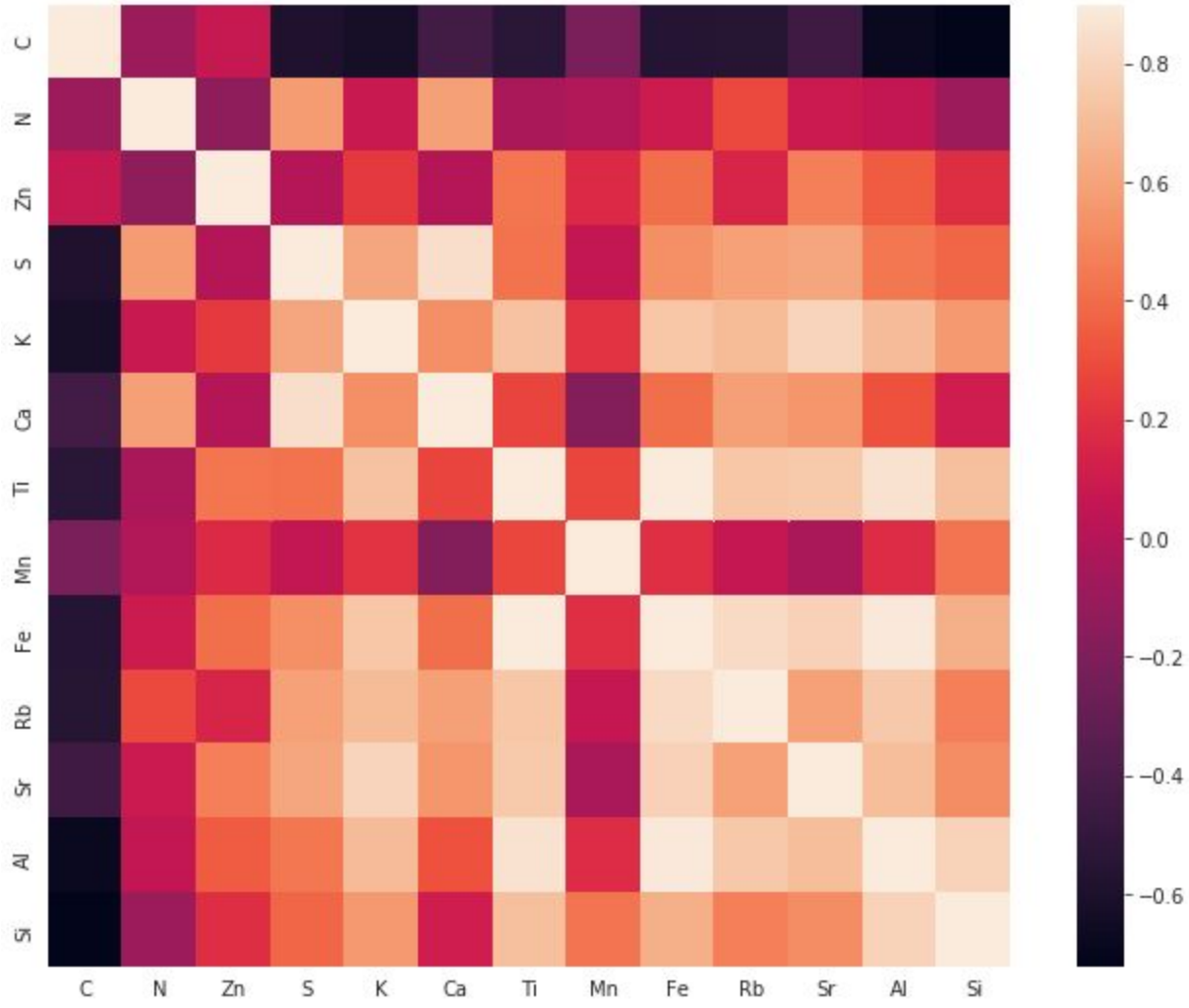
- Combined Data - XGBoost model
  3 fold validation scores
  Train Score - [0.3165421 0.3454295 0.28725418]
  Validation Score - [0.49860055 0.37992405 0.48902401]
  RMSE on prediction - 0.31652632121680657

Note:- For Carbon the Prediction RMSE values are much different from the 3 fold cross validation scores of train and validation set. This is because of the square transformation we did to remove skewness. The function for 3 fold cross validation returns the mean squared error on the transformed data. While during preding we take the RMSE with the actual data and predicted values. Since Nitrogen values weren't transformed we observe the RMSE and 3 fold cross validation scores quiet close to each other.

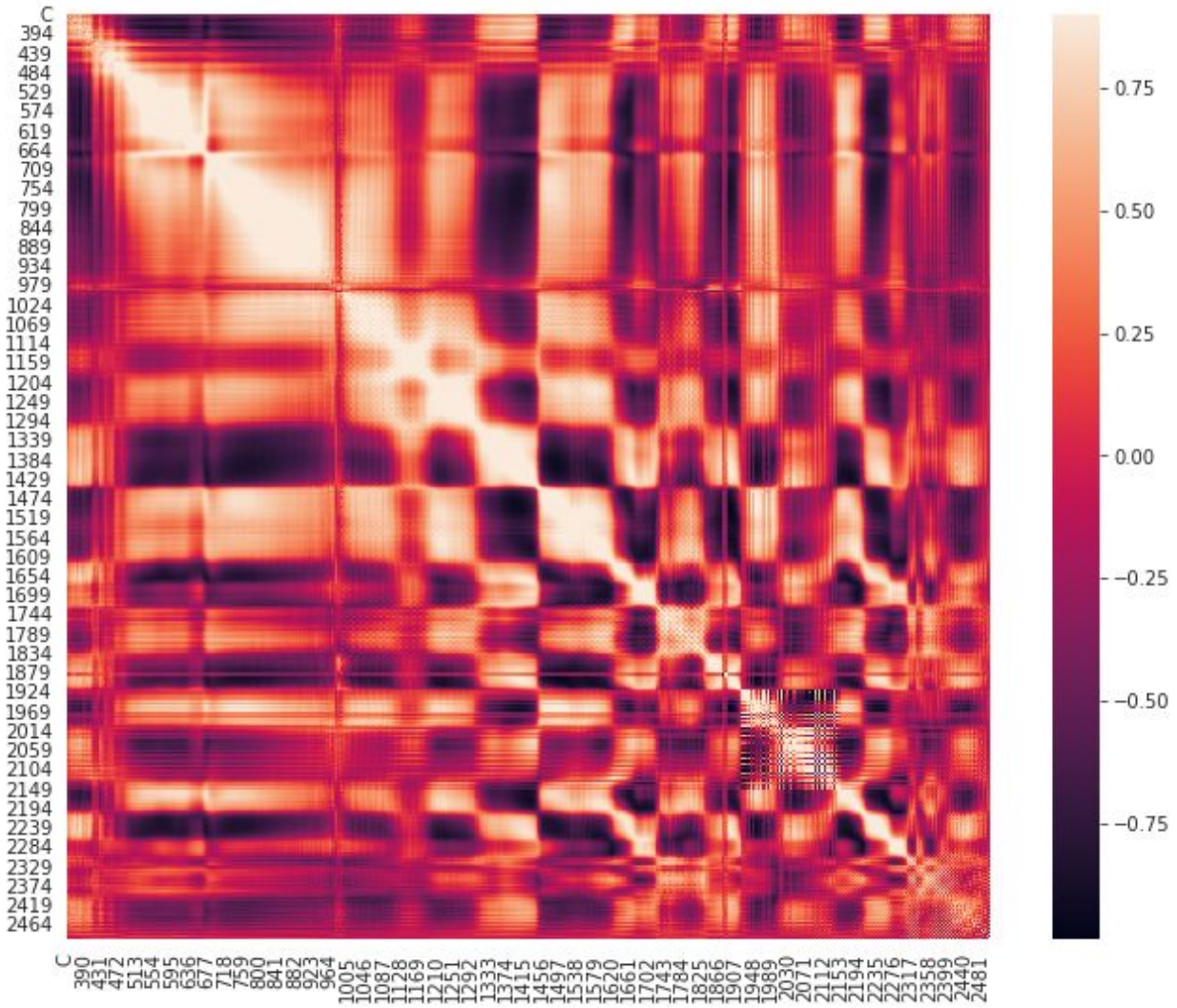
Feature-Target Correlation

Soil Carbon % has a negative correlation with most of the elements with their content in soil. Zn content in soil has a very insignificant effect on the Carbon content. It has a correlation 0.076. Si(-0.722), Al(-0.681), K(-0.622), S(-0.587), Fe(-0.562), Rb(-0.551) & Ti(-0.543) have high negative correlations with the Carbon content in soil. Sr(-0.450), Ca(-0.432) & Mn(-0.219) have a weak negative correlation with the Carbon content.

Soil Nitrogen % has insignificant correlation with most of the elemental data like, Fe(0.095), Sr(0.083), K(0.078), Al(0.057), Mn(-0.009), Ti(-0.036) & Si(-0.093). Ca(0.587) & S(0.571) have a strong positive correlation with Nitrogen content. While Rb(0.284) & Zn(-0.137) have weak correlation with it.



Correlation Matrix of Elemental data with target variables



Correlation matrix of Spectro-radiogram data with target variables.

As we can see from the diagram there is strong correlation among different wavelengths data which are close to each other. Hence all wavelengths close to each other have a similar correlation values with Carbon and Nitrogen content in soil. Wavelengths with 21xx have the highest positive correlation for Carbon and 22xx have the highest negative correlation.

Nitrogen doesn’t have a strong correlation with any wavelengths, 23xx wavelengths have a weak negative correlation while 16xx have a weak positive correlation.