Biochar Properties Prediction Model Development Project

Julius Choi
April 17th, 2019

I. Definition

Project Overview

Biochar is a porous carbon-rich material produced from biomass-waste such as agricultural waste and food waste through a thermochemical process like gasification or pyrolysis[1]. The biochar has a potential application for waste-water treatment, catalysts, electrodes, carbon sequestration agent, soil amendment agents and energy fuel[2].

To design and produce the biochar for the specific application is highly dependent on the physicochemical properties of the biochar[1]. The process parameters (e.g., temperature, pressure, time) and intrinsic properties of biomass (e.g., element composition) affect the properties of the biochar[2]. Therefore, the accurate prediction of the properties of the biochar produced from the different process condition and biomass waste will be helpful for the determination of the potential application of biochar.

To date, several researches regarding the prediction of biochar properties were reported. For example, support vector machine was implemented to predict the biochar yield[3] and an artificial neural network to predict heating value[4] of biochar from the biomass waste properties. However, to the best our knowledge no one attempted to predict several properties together such as surface area and molar composition.

Therefore, in this project, we will try to develop the machine learning model to predict those properties: surface area, yield and molar composition of biochar which are essential properties for the development of wastewater treatment process.

Problem Statement

This project is focused on the development of a correlation for the prediction of biochar properties from a reaction condition and biomass waste properties. The successful model will remove the need for time-consuming and expensive process to test all mentioned properties. Given a dataset collected from literatures [2-7], I will develop the model to predict physicochemical properties of biochar derived from different reaction condition and different raw biomass waste: carbon, oxygen, nitrogen and hydrogen molar

composition, yield and surface area. An artificial neural network, multitarget regresser incorporated with support vector regressor and GradientBoostingRegressor will be used for this project. This model can be used to determine the potential application of products.

Metrics

A prediction model will be developed using several models including an artificial neural network (ANN), a decision tree and support vector machine. These models will be trained with training data (70% of the input data set). These models will be evaluated and optimized to minimize mean absolute error (MAE), mean square error (MSE) and Coefficient of determination (R^2). Based on these performance evaluations, the best model will be selected. Finally, predictions will be made on the test data set and will be evaluated.

As mentioned above, MAE, MSE and R^2 will be used for performance evaluation.

AAPRE measures the relative absolute deviation from the dataset

$$MAE = \frac{1}{N}\sum_{I=1}^{N}\left|Y_i^{pred} - Y_i^{exp}\right|$$

RMSE represents the data dispersion

$$MSE = \frac{\sum_{I=1}^{N}\left(Y_i^{pred} - Y_i^{exp}\right)^2}{N}$$

R2 measures the degree of fitting between the predicted value and experimental data. The closer value of R2 to 1, the better the model fits the experimental data.

$$R^2 = 1 - \frac{\sum_{I=1}^{N}\left(Y_i^{exp} - Y_i^{pred}\right)^2}{\sum_{I=1}^{N}\left(Y_i^{exp} - Y_{ave}^{exp}\right)^2}$$

$Y_i^{exp}, Y_i^{pred}, Y_{ave}^{exp}$ represent the data collected from literature, the predicted value and the average of the corresponding data value from literature, respectively.

II. Analysis

Data Exploration

The data is collected from several literatures in Google Scholar. Several biomass waste were chosen to provide the effect of different elemental composition on biochar properties. Also, several different reaction conditions for same biomass waste were

required to provide the effect of reaction condition on the final product properties. The goal of regression is to predict the properties and yield of biochar.

Input variables:

1. Temperature: number of reaction temperature for production of biochar

2. Heating rate: number of heating rate for production of biochar

3. Holding time: number of time to hold the biomass waste in the reactor.

4. C molar contents: continuous

5. H molar contents: continuous

6. N molar contents: continuous

7. O molar contents: continuous

Output variables:

1. Yield: continuous

2. Surface area: continuous

3. C: continuous

4. H: continuous

5. N: continuous

6. O: continuous

The dataset has 13 columns and 43 rows with seven features and six dependent variables. Raw data is available in Table 1.

Table 1 Sample data of reaction condition and properties of biomass waste and biochar

Reaction condition			Biomass waste properties				Biochar properties					
Temp (oC)	Heating rate (oC/min)	holding time (h)	C (%)	H (%)	N (%)	O (%)	Yield (%)	Surface Area (m2/g)	C (%)	H (%)	N (%)	O (%)
350	20	5	48.2	5.88	0.22	45.5	35	0.5	71.6	3.88	0.25	23.2
450	20	3	48.2	5.88	0.22	45.5	33	1.6	77.6	3.52	0.27	17.3
600	20	2	48.2	5.88	0.22	45.5	21	256	84.7	1.83	0.3	11.3
500	10	2	40.87	5.4	0.27	53.46	36	73	60	1.68	0.25	38.07
300	10	2	45.51	6.17	0.15	47.83	43.7	0	69.13	4.85	0.39	24.36
450	10	2	45.51	6.17	0.15	47.83	28.5	12.9	83.62	3.24	0.17	11.46

600	10	2	45.51	6.17	0.15	47.83	22.7	401	81.81	2.17	0.73	14.03
300	10	2	45.82	6.25	0.36	47.18	33.4	5.2	69.5	4.2	0.9	24.36
450	10	2	45.82	6.25	0.36	47.18	28	13.6	78.6	3.52	0.92	15.46
600	10	2	45.82	6.25	0.36	47.18	26.5	388.3	76.45	2.93	0.79	18.33
300	10	2	46.52	6.11	0.2	46.89	40.5	1.3	66.2	4.7	0.4	27.72
450	10	2	46.52	6.11	0.2	46.89	26.3	10.2	76.89	3.55	0.23	18.11
600	10	2	46.52	6.11	0.2	46.89	24	375.5	80.89	2.43	0.15	14.87
350	5	2	35	5.2	0.5	39.4	91	0.83	50.81	3.7	0.73	17.8
550	5	2	35	5.2	0.5	39.4	55	5.24	53.2	2.11	0.75	6.88
350	5	2	45.22	6.34	1.15	46.94	70	0.69	62.37	4.64	1.65	22.58
550	5	2	45.22	6.34	1.15	46.94	42	0.18	72.09	2.85	1.36	12.95
350	5	2	53.5	6.6	1.5	35.5	65	0.43	67.89	4.42	0.53	24.19
550	5	2	53.5	6.6	1.5	35.5	38	1.56	80.7	2.94	0.57	11.79
350	5	2	37.1	5.6	0.5	44.7	70	0.97	61.85	4.43	1.13	20.47
550	5	2	37.1	5.6	0.5	44.7	42.5	0.75	67.84	2.45	1.12	9.93
350	5	2	41	6.1	3.6	39.7	98	1.18	40.41	3.55	2.96	12.2
550	5	2	41	6.1	3.6	39.7	58	8.72	37.38	1.56	2.32	6.5
350	5	2	49.7	6.2	0.7	42.6	67	1.08	68.47	4.6	0.52	23.93
550	5	2	49.7	6.2	0.7	42.6	40	23.1	83.8	2.75	0.82	9.76
350	5	2	41.8	5.7	1.09	37.81	70	1.02	63.53	4.08	0.94	24.2
550	5	2	41.8	5.7	1.09	37.81	42	0.72	74.46	2.33	0.96	12.55
300	10	2	43.71	5.97	1.49	48.82	48	7.03	57.71	4.53	2.14	35.61
500	10	2	43.71	5.97	1.49	48.82	33	12.56	61.58	2.82	1.86	33.75
700	10	2	43.71	5.97	1.49	48.82	24	5.24	64.43	1.09	1.21	33.27
300	10	2	42.41	5.67	2.76	33.58	48	3.56	61.31	8.88	4.51	25.3
500	10	2	42.41	5.67	2.76	33.58	33	9.02	44.77	9.08	5.69	40.46
700	10	2	42.41	5.67	2.76	33.58	28	8.95	40.05	9.67	6.74	43.54
400	5	2	28.49	4.55	3.87	34.7	64.7	70.29	31.92	1.74	2.35	13.56
800	5	2	28.49	4.55	3.87	34.7	39.3	44.49	40.51	0.7	2	10.02
600	5	2	28.49	4.55	3.87	34.7	47.8	61.81	36.64	0.99	2.31	11.76
600	10	2	33.94	4.53	2.35	46.89	45.5	336	30.81	1.38	2.74	39.87
600	10	2	36.06	3.43	1.23	55.82	36.3	2.6	50.78	2.08	1.83	36.7
600	5	0	46.55	6.087	4.49	42.39	34.4	8.8	59.24	2.445	4.63	33.65
650	5	2	60.87	3.99	1.67	18.92	20	7	63.55	1.59	1.26	9.23
800	10	1	46.84	8.18	3.86	40.74	25	0.2	63.78	8.18	3.86	40.74
400	15	2	22.52	3.36	2.06	64.82	68	19.8	27.32	2.35	2.2	9.34
450	20	2	48.33	6.12	9.12	35.82	20	0	63.36	3.56	7.11	25.6
400	5	1	38.91	4.74	1.37	35.31	30	87	45.5	4.4	1.1	49
500	5	1	38.91	4.74	1.37	35.31	28	96	51.9	3.2	1.3	43.6
600	5	1	38.91	4.74	1.37	35.31	25	60	65.3	2.9	1.7	30.1
550	40	0	44	5.46	1.61	48.8	28	0	73.7	1.89	1.82	21.6

According to Table 2, Surface area has the largest standard deviation (107.6). This means that the precise prediction of surface area might be challenging.

Table 2 Statistics of data set

	Temp	Heating rate	Residence time	C	H	N	O	Yield	Surface Area	C	H	N	O
Avearge	493.6	9.3	1.9	42.7	5.7	1.6	42.6	42.0	51.6	61.6	3.5	1.7	22.6
Std	131.5	6.3	0.7	7.0	0.9	1.7	7.5	18.4	107.6	15.3	2.0	1.6	11.4
min	300.0	5.0	0.0	22.5	3.4	0.2	18.9	20.0	0.0	27.3	0.7	0.2	6.5
max	800.0	40.0	5.0	60.9	8.2	9.1	64.8	98.0	401.0	84.7	9.7	7.1	49.0

Exploratory Visualization

Using linear regression in statmodels, the six dependent variables were predicted from the reaction condition and biomass properties. The predicted dependent variables were plotted against each actual dependent variable, which is the important physichochemical properties of biochar to see the appropriate linear or non-linear relationships for the prediction of the properties of biochar. As shown in Figure 1, coefficient of determination (R2) is between 0.207 and 0.905. Accordingly, It can be inferred that there might be a loosely linear relationship between predicted values and actual values, but not satisfactory for any meaningful prediction using the experimental condition and data (reaction condition and biomass properties) meaning that the linear model cannot accurately predict biochar properties. Therefore, we need to develop a nonlinear empirical model using data-driven modelling technique to predict biochar properties based on the reaction condition and biomass waste properties. Furthermore, To predict six dependent variables together, the multi-target regressor was applied.

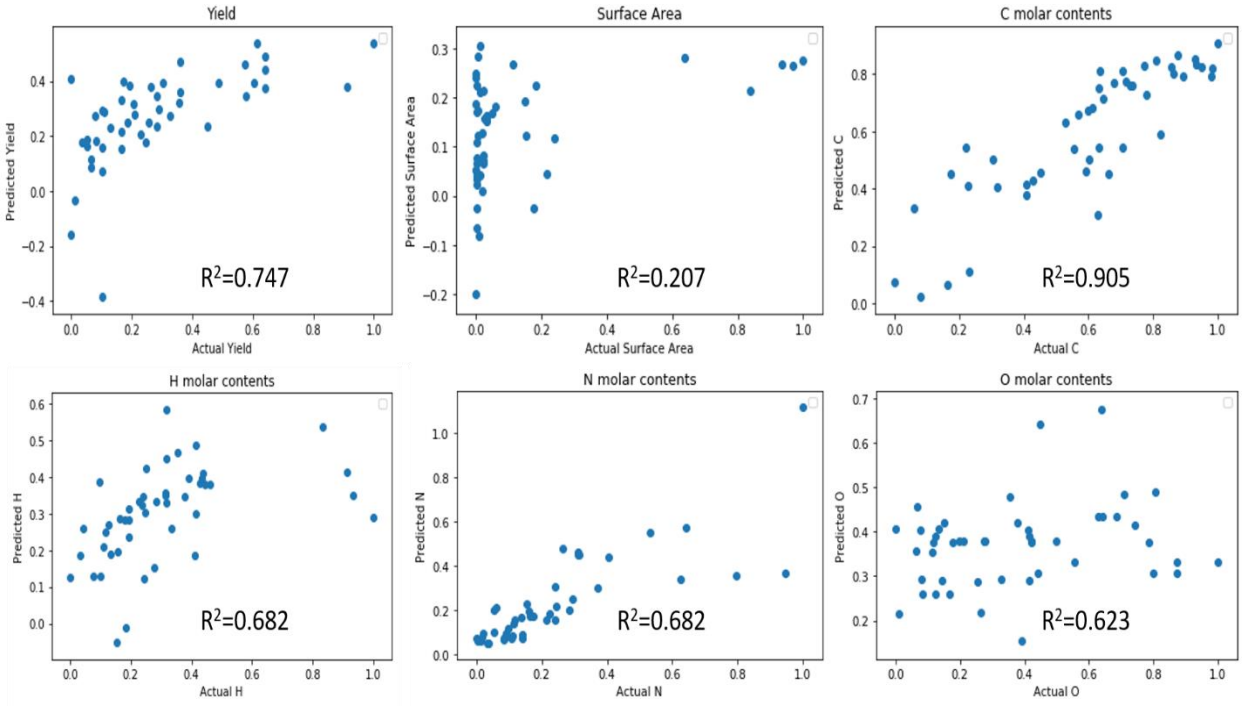

Figure 1 Data exploration by linear regression

Algorithms and Techniques

I intended to use artificial neural network (ANN), MultiOutputRegressors incorporated with support vector regressor (SVR), and GradientBoostingRegressor (GBR). ANN emulates the network on neuron of human brain. One advantage of ANN is the ability to learn mathematical model through training without the need to determine the mathematical relationship. However, it intend to converge into local minima. Whereas, SVR tends to find a global minima which make the advantage of SVT over ANN. GradientBoostingRegressor is one form of ensemble methods composed of weak prediction models. This regressor is used with a decision tree which is suitable for the prediction of biochar with specific properties. MultiOutputRegressor is for fitting one regressor for each target. This strategy make a parallelization of the regressor to fit multi-target. SVR As I mentioned before, our objective is to develop a model for the prediction of biochar properties such as surface area, elemental composition and yield using the data for reaction condition and the properties of raw materials. Therefore, we need to use those algorithms which is suitable for multivariative and multitarget problems. The following parameters can be tuned for optimization of model

- ANN: Number of layers, Layers types and activation function
- MultiOutputRegressor: For fitting one regressor for each target. Parallelization of the regressors to fit multi target.
- SVR: Random_state, learning_rate, kernel, parameter C
- GBR: random_state, learning_rate, min_samples_split, min_samples_leaf, max_depth, max_features

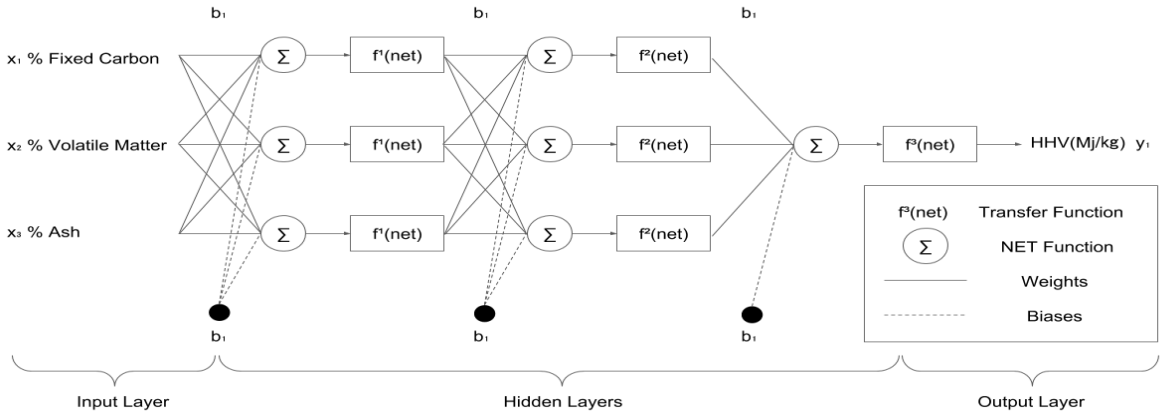

Figure 2 Schematic representation of ANN[1]

[1] Bioresource Technology 234 (2017) 122-130

Benchmark

Below figure shows performances of tried models in terms of MSE, MAE and R2_score. I adopted the ANN model, which is composed of one input layer, two hidden layer and one output layer from this Journal as one of benchmark model[2]. Furthermore, SVR and GBR model were applied to model the data. As shown in Figure 3, GBR model has the lowest MSE, MAE and the highest R2_score on training dataset and testing dataset. This result means that GBR model is the most appropriate model to predict results regarding the provided features. Therefore, GBR model was selected as the most promising benchmark model. However, overfitting of model to the training dataset resulted into high variance error. Accordingly, I tried to improve the model with hyperparameter tunining.

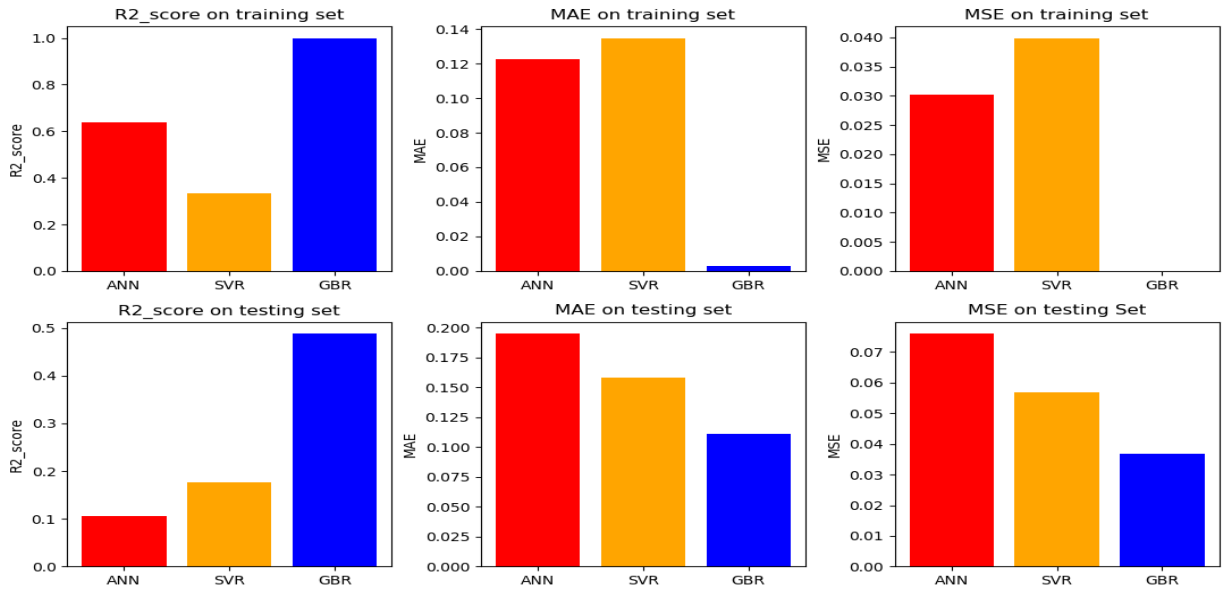

Figure 3 Metrics from benchmark models

III. Methodology

Data Preprocessing

The data was splited into features and target columns. Also, to normalize the variables, all the data was normalized using min-max normalization. The next step of data preprocessing was to split the data into training set and testing sets to check the feasibility of the model. In this project, 70% of the prepared data was randomly selected as the training set and 30 % was chosen as the testing set. All data preprocessing steps implemented for this project can be referenced for details in an attached python file.

Implementation

The implementation process is composed of six main stages:

1. Data preprocessing
   The data was explored checking number of features and dataset. Relationship between features and dependent variables was explored. The data was splitted into features and target column then normalized. Dataset was randomly divided into training set (70% of dataset) and validation set (30% of dataset).
2. Model evaluation
   Each algorithm is evaluated on training set and validation set using the default parameter. After reviewing metrics, one model, with the highest R2 score and lowest MSE and MAE, was chosen to further improvement.
3. Model training stage
   Using GridSearch, model was trained on training set to get improved parameter for model
4. Model validation stage
   The model trained from training stage was validated on the testing set
5. Model evaluation based on metrics
   The metrics on training and testing was review to see if the error is high bias error or high variance error.
6. Model tuning and step back to 2
   Based on the evaluation results, hyperparameter was changed or more dataset was collected to improve the model
7. Determine the best model
   The final model with the improved performance without any bias error and variance error but having the highest R2 score and lowest MAE and MSE was chosen. Predicted values were compared with the actual values.

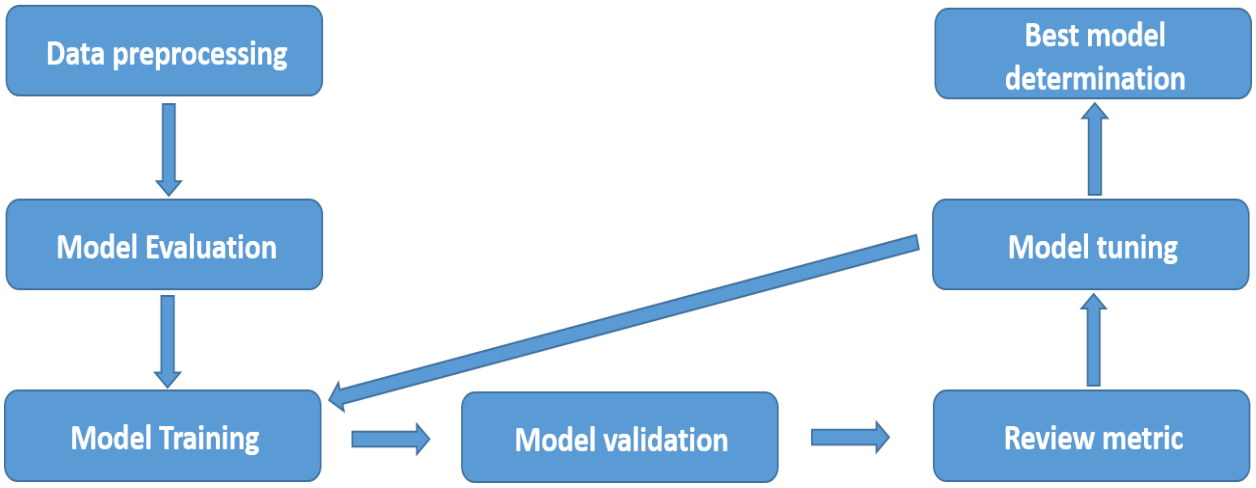

Figure 4 Process work flow

Refinement

As mentioned in Benchmark, the model had the high variance error due to the overfitting on training set. To overcome this issue and get higher R2_score, I tried to tune several parameters: learning_rate, min_samples_split, min_samples_leaf, max_depth, max_features, which normally lead to overfitting. Tested values are represented in Table 3. However, all parameter except for min_sample_leaf had no effect or a negative effect on R2 score. Therefore, the number of min_sample_leaf was tuned only using Gridsearch algorithm. Figure 5 shows that R2_score on testing test increased from 0.49 to 0.69, while, no big difference was observed on training set result. In the future, we need to collect more sample data to get more training data to overcome this high variance error.

Table 3 List of tested parameters

Parameter	Description	Values tested	Best Value
Learning_rate	Learning rate	4, 1, 0.1, 0.05, 0.02	Default
Min_samples_split	The minimum number of samples required to split an internal node	1,2,3,4,5	Default
Min_samples_leaf	The minimum number of samples required to be at a leaf node	1,2,3,4,5,6,7	4
Max_depth	Maximum depth of a tree	1,2,3,4,5	Default
Max_features	The number of features for the best split	1, 0.3, 0.1	Default

Figure 5 R2_score of optimized and unoptimized model

IV. Results

Model Evaluation and Validation

We applied ANN, SVR and GBR model to fit the data. GBR showed the best performance among other models so tried to tune the parameter to improve fitting. The metrics are calculated using sklearn-library which made the model performance calculation reliable. Finally, we got the best model which was improved by 40.8% in terms of R2_score compared with unturned benchmark model. As shown in Figure 6, this final model predicted each properties of biochar with R2 score between 0.58 and 0.91. Especially, this model showed the highest reliability for the prediction of hydrogen molar contents (R2_score=0.91). Therefore, we can say this model fulfilled the purpose of this project.

Furthermore, the final model was tested with different random state values to confirm the stability of the model. As shown in Table 4, R2_score is almost constant regardless of the random state values. So, we can conclude that the model is robust.

Table 4 R2_score with different random state

Random state	R2_score
3	0.6802
10	0.6806
50	0.6816
500	0.6808

Justification

Even though this model made a significant improvement over the benchmark model and can predict the properties of biochar with R2_score of 0.7, it must be improved more up to at least 0.9. It turned out that lack of training dataset lead to high variance error. However, it is still challenging to collect more data which meet our criteria. In the near future, I believe the more reliable dataset will be available, then, the model with high reliability with at least 0.9 R2_score will be developed.

V. Conclusion

Free-Form Visualization

The prediction values versus actual values from the model was visualized in Figure 6

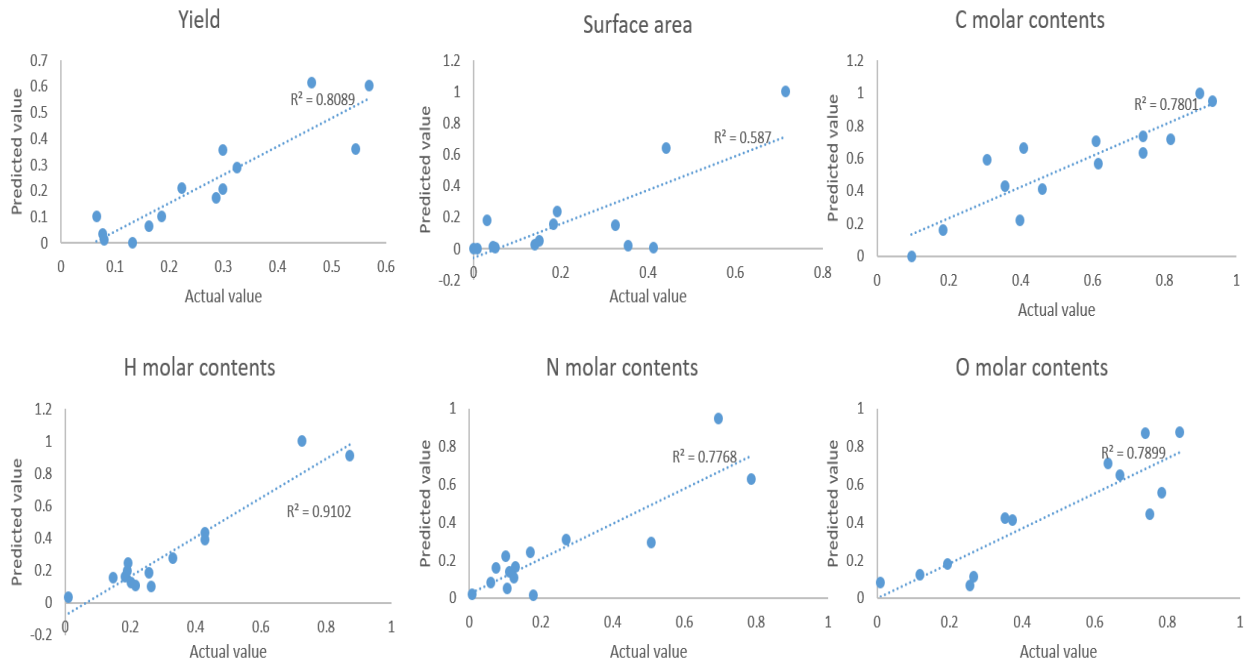

Figure 6 Plots of predicted values from test data set

Reflection

The process for this project can be summarized into

1. Problem identification and relevant dataset collection
2. Data preprocessing
3. Proper model selection and a benchmark creation for regresser
4. The regresser training and testing to develop the best model by tuning parameter

The most difficult part was step 1, step 3 and step 4. It was really challenging to find the literature having all information I looked for. Also, because the most journal was not open access made a limitation for access to the related information. Furthermore, I had to learn abouth the multi targer regressor and be familiarize myself with multi-target regressor that I had not used before this project. Also, turning parameter was quite challenging. I relied on my intuition and trial-error approach. This means that the process was the most time consuming.

The most interesting aspects of this project is that this framework can be used for identification of the relationship between the parameter and the prediction of properties of new biochar. Furthermore, I believe that this concept will be helpful for the development of tool for the investigation of the properties of materials besides biochar with available dataset.

Improvement

In this model, we can predict the value of dependent variables and identify the relationship between combined features and dependent variables. However, in reality, to improve the properties of biochar or materials experimentally, we also need to identify the effect of single feature or synergistic effect of two or three features on the predicted properties. So, I need to apply proper algorithms to figure out this relationship. Furthermore, I also need to develop the way to classify the potential application of biochar based on the anticipated biochar properties. The combination of the developed regresser for the prediction of properties and the classifier for application will lead to the development of the predictor of new product application from raw material. However, I still face the lack of the dataset. For example, although, we anticipate that there is the relationship between the elemental contents and the application of biochar (e.g., battery or adsorbent), the data is still insufficient to prove this concept. I believe this bottleneck will be overcame soon.

References

[1] C. Liu, H. Wang, A.M. Karim, J. Sun, Y. Wang, Catalytic fast pyrolysis of lignocellulosic biomass, Chemical Society Reviews, 43 (2014) 7594-7623.
[2] W.-J. Liu, H. Jiang, H.-Q. Yu, Development of Biochar-Based Functional Materials: Toward a Sustainable Platform Carbon Material, Chemical Reviews, 115 (2015) 12251-12285.
[3] H. Cao, Y. Xin, Q. Yuan, Prediction of biochar yield from cattle manure pyrolysis via least squares support vector machine intelligent approach, Bioresource Technology, 202 (2016) 158-164.
[4] H. Uzun, Z. Yıldız, J.L. Goldfarb, S. Ceylan, Improved prediction of higher heating value of biomass using an artificial neural network model based on proximate analysis, Bioresource Technology, 234 (2017) 122-130.
[5] Z. Ding, Y. Wan, X. Hu, S. Wang, A.R. Zimmerman, B. Gao, Sorption of lead and methylene blue onto hickory biochars from different pyrolysis temperatures: Importance of physicochemical properties, Journal of Industrial and Engineering Chemistry, 37 (2016) 261-267.
[6] S. Dawood, T.K. Sen, C. Phan, Adsorption removal of Methylene Blue (MB) dye from aqueous solution by bio-char prepared from Eucalyptus sheathiana bark: kinetic, equilibrium, mechanism, thermodynamic and process design, Desalination and Water Treatment, 57 (2016) 28964-28980.
[7] J. Fang, B. Gao, A. Mosa, L. Zhan, Chemical activation of hickory and peanut hull hydrochars for removal of lead and methylene blue from aqueous solutions, Chemical Speciation & Bioavailability, 29 (2017) 197-204.