IJEW 国际工程期刊 International Journal of Engineering Works انجینئرنگ کاموں کا بین الاقوامی جرنل

International Journal of Engineering Works
ISSN-p: 2521-2419
ISSN-e: 2409-2770
Vol. 7, Issue 01, PP. 58-61, January 2020
https://www.ijew.io/
https://doi.org/10.34259/ijew.20.7015861

Indirect Gross Calorific Value Prediction using Random Forest

Waqas Ahmed[1], Khan Muhammad[2]
[1]Msc Mining Enginnering Student at UET Peshawar Pakistan
[2]Assistant Professor Department of Mining Engineering UET Peshawar
waqas.Ahmed@uetpeshawar.edu.pk[1], khan.m@uetpeshawar.edu.pk[2]
Received: 14 January, Revised: 21 January, Accepted: 25 January

***Abstract*—**During operation of coal-based power plants, frequent calorific Value measurement is necessary. Previously, Artificial Intelligence based models have been developed for instant calorific value calculation based on proximate analyses or ultimate analyses or combination of both. In this paper, random forest was used for comparison of all the three methods and computing relative analyses parameters importance. This study uses well known USGS coal qual dataset. In this work, 10-fold validation strategy and R-squared was used as validation strategy and performance metric respectively. Ultimate analyses (R-squared = 0.9984) performed slightly better than proximate analyses (R-squared = 0.9861) or combination of both (R-squared = 0.9982). Lastly, carbon was found to be the most important feature in all models.

***Keywords*—** Gross Calorific Value, Random Forest, Proximate Analyses, Ultimate Analyses, Neural Networks.

I. Introduction

Energy demand of world is increasing and at the present time, whole world energy demand is mostly met by fossil-based fuels [1], [2]. The demand for coal which is most common among fossil-based fuels due to its huge abundance and greater life cycle has increased [3], [4]. Coal which supplies about 40-45% of the planet's energy needs [4], [5] is expected to remain the dominant energy source for the near future due to its financial advantages [5], [6]. During operation of coal based power plants, frequent calorific Value (the heat capacity of a unit weight of coal after burning completely [7], [8]) measurement is necessary because it decides the quality of coal [7]. Usually bomb calorimeter test method is used to measure calorific value however, such method is expensive, destructive, and needs an advanced technician. For that purpose, various researchers have tried to predict calorific value indirectly from proximate analysis [2], [6], [9]–[11] or ultimate analysis [5] or combination of both [12]–[16]. However, task of comparing all the three methods, computing relevant importance of analyses parameters in each method was not carried out previously. Such work is important to get better insight into various indirect calorific value prediction methods. In this study, Random Forest (RF) is used for this task. This paper uses samples derived from coal database (coalqual version 3.0) which was derived from U.S. Geological Survey (USGS) of Energy Resource Program in such a way that validation rating of data was not "Incomplete" or "Suspect" for both proximate and ultimate analyses parameters. The data is available at ncrdspublic.er.usgs.gov/coalqual/. For comparison of all the three methods, validation method and performance metric used was popular "10-fold cross validation" [17] and "R-squared". Analysis parameters importance was computed based on their homogeneity values while splitting inside random forest.

In the next section a background of Random Forest is presented followed by Coalqual dataset description. Next, methodology is detailed. Lastly, results are discussed followed by conclusion section.

II. Random Forest

The random forest (developed by Breiman in 2001 [18]) is very popular machine learning algorithm. Machine learning is branch of Artificial Intelligence consisting of supervised learning, unsupervised learning, semi-supervised learning, and reinforced learning [19]–[21]. Task at hand is "regression" which comes under category of supervised learning. Regression is estimating/learning relationship between dependent variable (calorific value in this case) and independent variables (Analysis parameters in this case). For regression, various algorithms are available such as neural networks, support vector machines, nearest neighbours, decision trees, and random forest etc [19]–[21].

All regression algorithms try to find mapping between independent and dependent variables and the main task of machine learning engineer is to find best mapping/validation accuracy [19]–[21]. Best validation accuracy is achieved by tuning hyper-parameters of respective regression algorithm. Hyper-parameters are initial parameters which are set before training process of algorithm starts [19]–[21]. Out of all regression algorithms, Random Forest (RF) is categorized as ensembling method which combines decision trees and use their mean value as final predicted value [18]–[21]. In RF, multiple decision trees are fitted on different training datasets generated using bootstrapping (A resampling technique to generate multiple random realizations) of original dataset [18]–

[21]. To understand RF, explanation of decision trees is presented.

Decision tree is well known method in which upside-down tree is constructed by dividing training dataset into subsets using best independent variables sequentially one after the other [18]–[21]. Best independent variable is one which gives greater homogeneity and is calculated using various algorithms such as Iterative Dichotomiser 3 (ID3) or Classification and Regression Trees (CART). In ID3, best independent variable computation is done in three steps:

A. Standard deviation of output variable calculation for complete dataset

$$S = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$

Where x is dependent variable, $\bar{x}$ is mean of output variable, n is total number of samples, and S is Standard deviation of dependent variable for complete dataset.

B. Subsets from dataset are formed such that each subset contains single independent/input variable (analysis parameters) and dependent variable. For each subset, standard deviation is computed.

$$S(T, X) = \sum_{c \in X} P(c) S(c)$$

Where T is dependent variable, X is independent variable, c is different classes of that independent variable, P(c) is probability of class c in that independent variable, S(c) is standard deviation of dependent variable associated with class c of the independent variable.

C. For each subset, standard deviation reduction (SDR) value is computed by subtracting respective standard deviation from whole dataset standard deviation.

$$SDR(T, X) = S(T) - S(T, X)$$

Finally, best independent variable is the one whose subset gives maximum SDR value.

Best independent attribute computed for complete dataset is placed as root node followed by intermediate nodes which represent best independent attributes computed for subset of data created sequentially one after other using values of best independent attributes. Dividing data into subset of data continues till all data is processed. However, in such cases validation accuracies reduce. For that reason, hyper-parameters within decision trees such as Max_depth, Min_samples_split, Max_features, Min_samples_leaf are used to improve validation accuracy of decision tree. Main task of machine learning engineer is to find best values for these hyper parameters. Usually this task is done by trying out all combinations of ranges of hyper parameter values (Grid search).

As explained, mean of decision trees output gives RF prediction value. Depending on homogeneity of each independent variable, random forest also gives measure of variables importance.

III. Dataset Generation

Dataset used in this study was derived from coal database (coalqual version 3.0) which was derived from U.S. Geological Survey (USGS) of Energy Resource Program. Coalqual database contains proximate (Ash, Fixed Carbon, Volatile Matter, Moisture), and ultimate analyses (Hydrogen, Carbon, Nitrogen, Sulphur, and Oxygen) along with Gross Calorific Value results obtained from various ranks of coal including anthracite, bituminous, sub-bituminous, and lignite. All analyses carried out were in accordance to ASTM standards. This study includes samples which do not have "Incomplete" or "Suspect" validation rating for their respective analyses parameters. Summary statistics of data used is presented in Table I.

TABLE I. Summary Statistics of coalqual data

Category	**Variable Name**	**Summary Statistics**			
		mean	Std	min	max
Output	GCV (BTU)	11551	2314	3790	15193
Proximate Analysis	Moisture	8.233	10.05	0.4	52.5
	FC	47.93	11.31	4.1	87
	Ash	11.74	7.29	0.9	54.7
	VM	32.1	6.498	3	55.7
Ultimate Analysis	Hydrogen	4.331	0.799	0.21	9.12
	Carbon	65	12.5	22.5	88.2
	Nitrogen	1.282	0.343	0.2	5.6
	Oxygen	7.463	3.099	-0.9	18.96
	Sulphur	1.952	1.805	0.09	20.9

IV. Methodology

In this study, three different random forest models having 1) proximate analyses, 2) ultimate analyses, and 3) combination of both as input were developed in three steps. First, hyper-parameters were searched; second, models were developed and results were reported using 10-fold validation; third, feature importance was carried out for each model.

For each model, data was first split into 80:20 for training and testing/hyper-parameters tuning respectively. After splitting, hyper-parameter search was done in grid manner where each RF model was built for all combination of search ranges as shown in Table II. For each RF model, best hyper-parameter values were those which gave best accuracy on 20% test data.

TABLE II. HYPER-PARAMETERS RANGES TO BE SEARCHED

Hyper parameters	Range
Min_samples_leaf	(1 to 50)
Max_depth	(4 to 10) or Complete
Min_samples_split	(2 to 500)

Min_samples_leaf searched from 1 to 50 with increment of 3, Max_depth searched from 4 to 10 with increment of 2, Min_samples_split searched from 2 to 500 with increment of 10.

In second step, complete data was split into ten equal subsets where each model was trained on 9 parts and tested on remaining 10^{th} part. This process was repeated 10 times where each time 10^{th} part used for validation was different. In this way, all data was used for training as well as testing. Final value of performance metric was average of 10 values. Finally, feature importance of each input variable in all the three models was found using their respective homogeneity values.

V. RESULTS AND DISCUSSION

Main task in RF based models was finding optimum hyper-parameters values. It was found that increasing Max_depth value and decreasing Min_samples_leaf, Min_samples_split values were responsible for poor validation accuracy and vice versa. Optimum hyper-parameters values found for all models were same and is presented in Table III.

TABLE III. OPTIMUM HYPER-PARAMETERS FOUND

Hyper parameters	Range
Min_samples_leaf	10
Max_depth	Complete
Min_samples_split	22

RF model results using 10-fold validation are presented in Table IV and suggest that ultimate analyses performed slightly better than proximate analyses.

TABLE IV. 10-FOLD VALIDATION RESULTS

Model	R-squared
RF (Proximate)	0.9861
RF (Ultimate)	0.9984
RF (Proximate + Ultimate)	0.9982

As shown in Fig I, II, and III, carbon was most important feature for each model having ultimate or proximate or combination of both as inputs respectively.

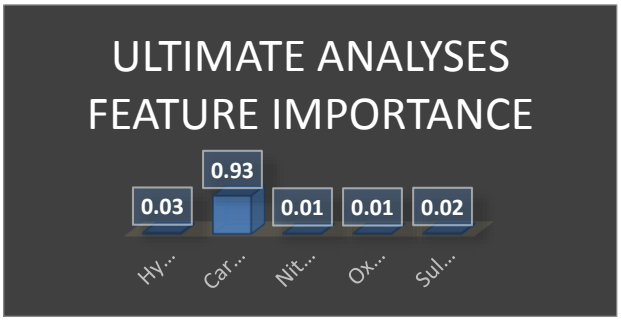

Figure I Ultimate Analyses Feature Importance

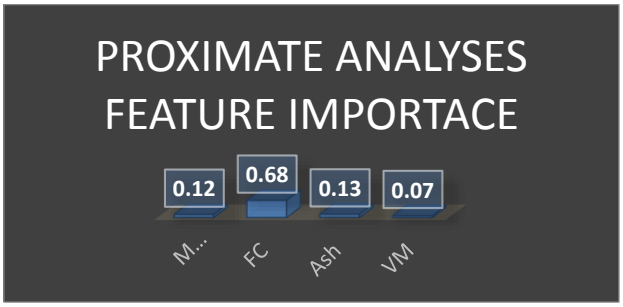

Figure II Proximate Analyses Feature Importance

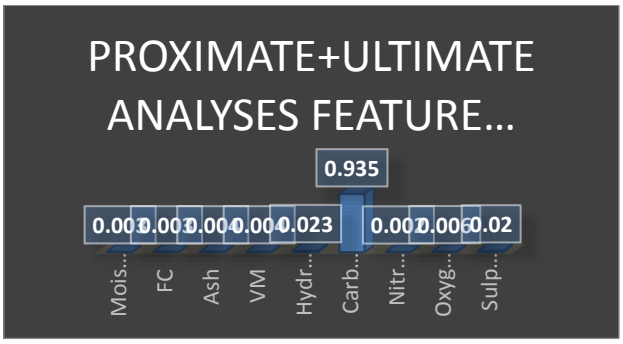

Figure III Proximate + Ultimate Analyses Feature Importance

CONCLUSION

This paper predicts calorific value using 1) proximate analyses parameters, 2) ultimate analyses parameters, and 3) proximate + ultimate analyses parameters. Results suggested that ultimate analyses performed slightly better than proximate analyses. Carbon was computed to be the most important feature in indirect determination of calorific value.

REFERENCES

[1] [M. Leahy, J. L. Barden, B. T. Murphy, N. Slater-thompson, and D. Peterson, "International energy outlook 2013," United States of America, 2013.

[2] A. V. Akkaya, "Proximate analysis based multiple regression models for higher heating value estimation of low rank coals," Fuel Process. Technol., vol. 90, no. 2, pp. 165–170, 2009.

[3] K. F. De Souza, C. H. Sampaio, and J. A. T. Kussler, "Washability curves for the lower coal seams in Candiota Mine - Brazil," Fuel Process. Technol., vol. 96, pp. 140–149, 2012.

[4] O. Sivrikaya, "Cleaning study of a low-rank lignite with DMS, Reichert spiral and flotation," Fuel, vol. 119, pp. 252–258, 2014.

[5] I. Yilmaz, N. Y. Erik, and O. Kaynar, "Different types of learning algorithms of artificial neural network (ANN) models for prediction of

gross calorific value (GCV) of coals," Sci. Res. Essays, vol. 5, no. 16, pp. 2242–2249, 2010.

[6] P. Tan, C. Zhang, J. Xia, Q. Y. Fang, and G. Chen, "Estimation of higher heating value of coal based on proximate analysis using support vector regression," Fuel Process. Technol., vol. 138, pp. 298–304, 2015.

[7] S. U. Patel et al., "Estimation of gross calorific value of coals using artificial neural networks," Fuel, vol. 86, no. 3, pp. 334–344, 2007.

[8] M. J. F. Llorente and J. E. C. García, "Suitability of thermo-chemical corrections for determining gross calorific value in biomass," Thermochim. Acta, vol. 468, no. 1–2, pp. 101–107, 2008.

[9] Q. Feng, J. Zhang, X. Zhang, and S. Wen, "Proximate analysis based prediction of gross calorific value of coals: A comparison of support vector machine, alternating conditional expectation and artificial neural network," Fuel Process. Technol., vol. 129, pp. 120–129, 2015.

[10] J. Akhtar, N. Sheikh, and S. Munir, "Linear regression-based correlations for estimation of high heating values of Pakistani lignite coals," Energy Sources, Part A Recover. Util. Environ. Eff., vol. 39, no. 10, pp. 1063–1070, 2017.

[11] M. Açıkkar and O. Sivrikaya, "Artificial neural networks for estimation of the gross calorific value of Turkish lignite coals," no. Imsec, pp. 1075–1079, 2018.

[12] S. Mesroghli, E. Jorjani, and S. Chehreh Chelgani, "Estimation of gross calorific value based on coal analysis using regression and artificial neural networks," Int. J. Coal Geol., vol. 79, no. 1–2, pp. 49–54, 2009.

[13] S. C. Chelgani, S. Mesroghli, and J. C. Hower, "Simultaneous prediction of coal rank parameters based on ultimate analysis using regression and artificial neural network," Int. J. Coal Geol., vol. 83, no. 1, pp. 31–34, 2010.

[14] N. Y. Erik and I. Yilmaz, "On the use of conventional and soft computing models for prediction of gross calorific value (GCV) of coal," Int. J. Coal Prep. Util., vol. 31, no. 1, pp. 32–59, 2011.

[15] S. S. Matin and S. C. Chelgani, "Estimation of coal gross calorific value based on various analyses by random forest method," Fuel, vol. 177, pp. 274–278, 2016.

[16] X. Wen, S. Jian, and J. Wang, "Prediction models of calorific value of coal based on wavelet neural networks," Fuel, vol. 199, pp. 512–522, 2017.

[17] S. Yadav and S. Shukla, "Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification," Proc. - 6th Int. Adv. Comput. Conf. IACC 2016, no. Cv, pp. 78–83, 2016.

[18] L. Breiman, "Random Forest Draft," pp. 1–33, 2001.

[19] S. J. Russell et al., Prentice Hall - Artificial Intelligence A Modern Approach. 2001.

[20] E. Sugawara and H. Nikaido, pattern recognition and machine learning, vol. 58, no. 12. 2014.

[21] A. Peña Yañez, "Elements of statistical learning," Rev. Esp. Enferm. Apar. Dig., vol. 26, no. 4, pp. 505–516, 1967.