Estimating DFT Calculated Voltage of Battery Materials : An Interpretable Machine Learning Approach

Rafat Ashraf Joy

Shahjalal University of Science and Technology,
Sylhet, Bangladesh
rafat.joy99@gmail.com

Abstract. This study deals with the.....

Keywords: Battery, Voltage, Machine learning

1 Introduction

The issue o

2 Methodology

2.1 Machine Learning Algorithms

2.2 Interpretable Machine Learning

Machine learning algorithms work through a series of complex math operations, and the complexities of why a particular model produces such outcomes are too complex for humans to comprehend. As a result, machine learning models are frequently referred to as "black boxes." Machine learning has recently been applied to high-stakes decision-making difficulties in domains such as financial risk management, healthcare and criminal justice. As a result, establishing trust and determining the sanity of the decision-making process that is taking place beneath the model is a critical question. Additionally, machine learning interpretability can lead to insights and knowledge discoveries that were previously unknown to humans.

Machine learning interpretability approaches are broadly classified into two types [2] : model-specific and model-agnostic. Model-specific interpretation methods are limited to a subset of model types. Interpreting the weights and bias of a linear regression, for example, is a model-specific interpretability method. Model agnostic methods, on the other hand, can be applied to any machine learning model, independent of its kind. In general, model agnostic approaches work by examining function input and output pairs.

3 Data

For our dataset, we have used the data adopted from [1]. The training data is originally extracted from Materials Project database. The material property features originally extracted from Materials Project database were: *space group, crystal lattice type, volume, capacity_grav, capacity_vol, energy_grav, energy_vol, max_frac, min_frac, min_instability, nsteps, numsites, average voltage, minimum voltage and maximum voltage.* Data cleaning and feature engineering were performed in the extracted data. The feature vectors were derived form elemental properties of each atomic constituents in a particular material and are adopted from [3]. The features used for training(X) are: *capacity_grav, capacity_vol, energy_vol, max_frac, min_frac, min_instability, nsteps, numsites* and the targets are: *average voltage, max voltage, min voltage* individually. Statistical normalization is used in order to normalize the data so that it is easier to train the machine learning models and they lie on a common scale. Finally, we have split the dataset into 80% training set and 20% test set.

4 Machine Learning Models

In this study, we have experimented with 7 machine learning algorithms. Among them, Gradient Boosted Trees and Random Forest Algorithm performed the best.

4.1 Random Forest

Random forest is a Supervised Learning approach for classification and regression problems that employs ensemble learning methods. It's a bagging method that was created to solve the problem of large variance in traditional decision trees. At training time, the Random Forest Regressor creates numerous decision trees with no correlation between them. A randomly split portion of the training data is used to generate each individual decision tree. The bagging approach of the random forest algorithm merges all discrete decision trees. Bagging is averaging the predictions of individual decision trees in a regression problem. As a result of tackling the overfitting problem, bagging produces more accurate predictions.

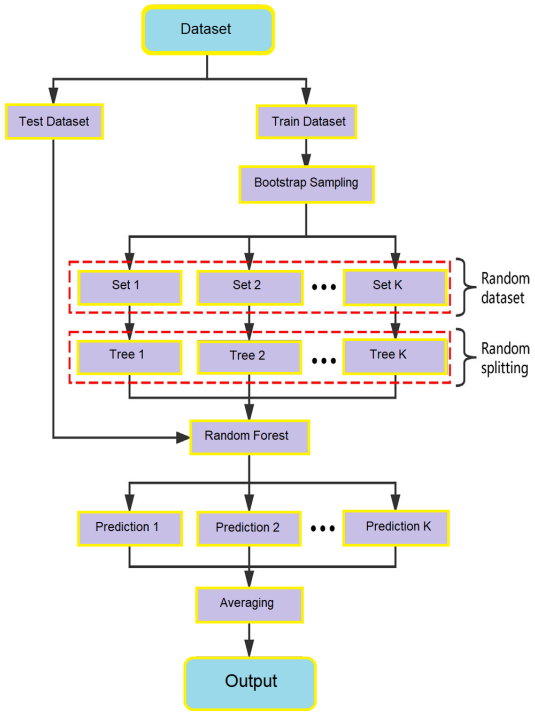

Fig. 1: Caption

4.2 XGBoost

5 Results

In this study, r squared value and RMSE were chosen as evaluation metrics.

The r squared value is a statistical indicator to measure the percentage of variance explained by a machine learning model. It is defined as:

$$r^2 = \frac{\sum_i (\widehat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} \tag{1}$$

where, y_i = the observed value, $\hat{y}_i$ = the predicted value and $\bar{y}$ = average output.

The root mean square error(RMSE) is the standard deviation of prediction errors. It is defined as:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \tag{2}$$

At first, seven machine learning models were trained on the training set. Among these algorithms, Xgboost and Random Forest performed the best. So, these two algorithms have been selected for further evaluation.

From our observation of SHAP feature importance, we eliminated the features(') that had very low impact on the model's decision making procedure. This elimination had improved the two model's performance metric, although by a narrow margin.

The following table shows the results of the Machine learning models on the test set which was seprated from train set before feeding the training data into machine learning algorithms.

Model Name	R2 score	RMSE Score
Xgboost	0.99	0.77
Random Forest		

Table 1: Caption

The following table shows the Hyper-parameters of the Xgboost model:

Hyperparameter	Value
min depth	7

Table 2: Caption

The following table shows the Hyper-parameters of the Random Forest model:

6 Interpretation

7 Conclusion and Future Work

References

1. Maphanga, R., Mokoena, T., Ratsoma, M.: Estimating dft calculated voltage using machine learning regression models. Materials Today: Proceedings **38** (2020). DOI 10.1016/j.matpr.2020.04.204
2. Molnar, C.: Interpretable Machine Learning (2019). https://christophm.github.io/interpretable-ml-book/
3. Ward, L., Agrawal, A., Choudhary, A., Wolverton, C.: A general-purpose machine learning framework for predicting properties of inorganic materials. npj Computational Materials **2** (2016). DOI 10.1038/npjcompumats.2016.28