**3.2.1 Bandgap prediction utilizing supervised machine learning model**

In the process of most optimal Bandgap prediction, 36 data were collected and divided into two data sets based on band gap energy (eV) where Eg = 3.4 (eV) Each data set contains 18 data points, and each data point includes the value of Bandgap and the features of transmittance and resistivity. Linear Regression (LR) and Random forest (RF) models were constructed for a subset of the original dataset in order to find the optimal bandgap with transmittance as high as possible and resistivity as low as possible. Generally speaking, the linear regression model is effective for datasets of small volume. Furthermore, the results of linear regression have excellent interpretability and are able to give the understanding and interpretation of each variable based on the coefficients, which is beneficial for decision analysis. These properties make linear regression models occupy an irreplaceable position in physics, materials science, and finance. Similarly, the random forest model was chosen as the prediction model for this experiment because of its high accuracy and generality. In the meanwhile, the superiority of the random forest is reflected by its introduction of randomness, which is less prone to overfitting. The Radom forest model also has good resistance to noise and the ability to maintain considerable accuracy if a large fraction of the features missing.

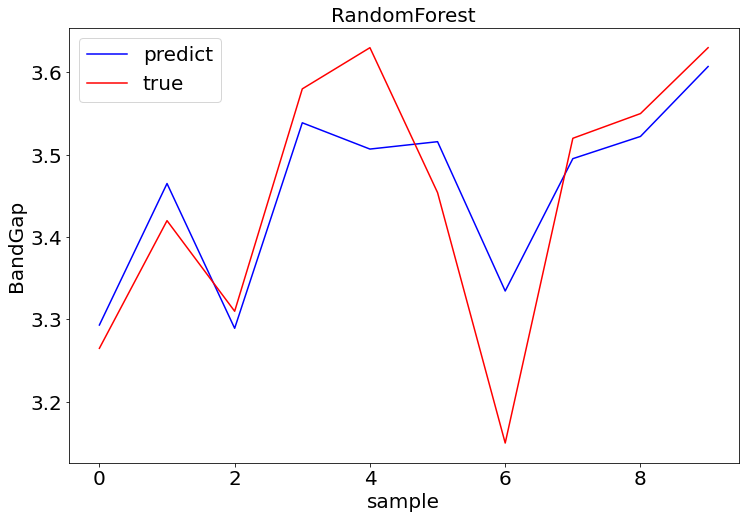
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  | | --- | --- | --- | --- | |  | Predicted Band gap | Transmittance% (input) | Resistivity(input) | | RF | 3.27117583 | 88.00 | 0.0068000000 | |  | 3.27421083 | 90.00 | 0.0000310000 | |  | 3.26466 | 93.50 | 0.0098000000 | |  | 3.26928 | 95.00 | 0.0009000000 | | LR | 3.28309243 | 88.00 | 0.0068000000 | |  | 3.29685819 | 90.00 | 0.0000310000 | |  | 3.305733 | 93.50 | 0.0098000000 | |  | 3.29083353 | 95.00 | 0.0009000000 | |

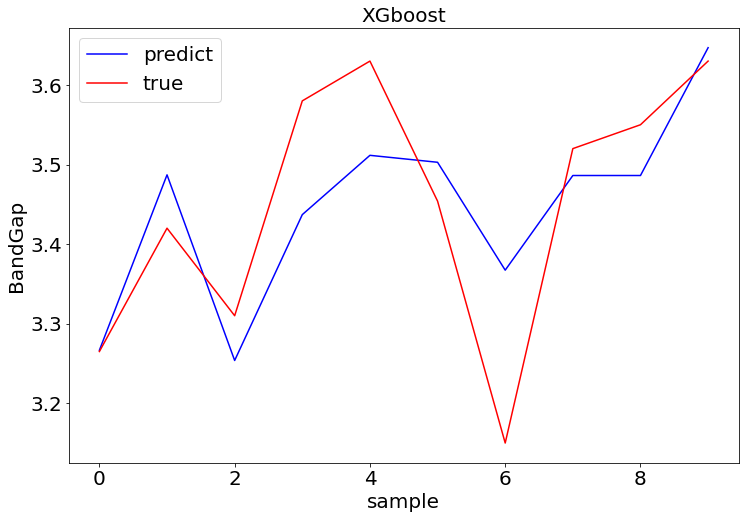
*Table 1, Eg < 3.4 eV*

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted Band gap | Transmittance% (input) | Resistivity(input) |
| RF | 3.464 | 88.00 | 0.0068000000 |
|  | 3.58749667 | 90.00 | 0.0000310000 |
|  | 3.5597 | 93.50 | 0.0098000000 |
|  | 3.6182 | 95.00 | 0.0009000000 |
| LR | 3.47126281 | 88.00 | 0.0068000000 |
|  | 3.59914814 | 90.00 | 0.0000310000 |
|  | 3.41657875 | 93.50 | 0.0098000000 |
|  | 3.58437188 | 95.00 | 0.0009000000 |

*Table 2, Eg > 3.4 eV*

After importing the two datasets into the linear regression (LR) and random forest (RF) regression models, the models were trained and the cross-validation method was adapted to validate the models, that is, the data were used repeatedly, and the obtained sample data were sliced and combined into different training and testing sets, and the training set was used to train the models, while the testing set was used to evaluate how well the models predicted. In this way, the predicted and actual bandgap comparison plots are plotted.





**3.2.2 Evaluating Various Supervised ML Models.**

In Section 3.2.1, two models, Linear Regression(LR) and Random forest(RF), were trained to predict the optimal Bandgap value. Building on previous experiments, we experimented with more supervised learning models, Xgboost and LightGBM. After importing two datasets to four different models and training and predicting them respectively, the model performance is evaluated by calculating the evaluation indicator of the training and test datasets, which are model accuracy R2 and root mean square (RMSE).

|  |  |  |
| --- | --- | --- |
| model | model accuracy (r2) | Root Mean Square Error (RMSE) |
| LR | 0.363 | 0. 123 |
| RF | 0.884 | 0.053 |
| XGBoost | 0.976 | 0.024 |
| Lightgbm |  | 0. 160 |

(next should be the analysis of the result, however, the r2 of Lightgbm is negative, meaning the data is not linearly dependent, probably due to the small volume of the dataset, so this section remains unresolved and needs to be improved)

**3.2.3 SHAP Analysis to Investigate Relative Impact of Features on Target**

Four kinds of classifier are used to predict the x value:

* Linear Regression
* Random Forest
* XG Boost
* LightGBM

This waterfall graph shows how the features contributing the output of the prediction of selected model. That is, the average model output based on each feature over the dataset we passed. The features are both in blue, which mean that, they give a lower expectation of the output of the model, and in red, which mean that, they give a higher expectation of the output of the model. Both features are in blue because our model is kind of overestimated.

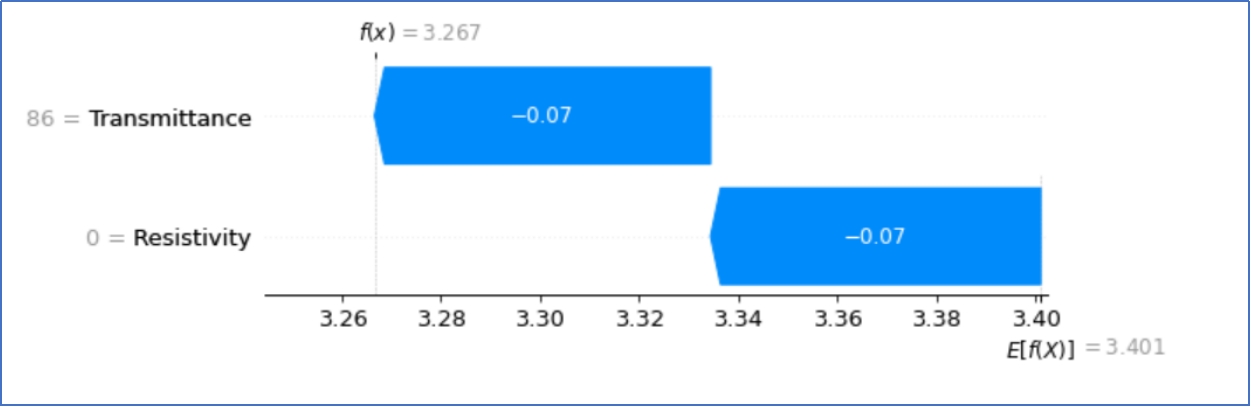


Fig 1 Waterfall graph – XGB Classifier

This is the plot show the base value of each feature and how it will react on the model.



Fig 2 Force Plot – XGB Classifier – Visualize the prediction’s explanation with a force plot

The force graph only shows part of the model. Here, by rotating 90 angle and stack them horizontally, we will have a more detailed sight for the entire dataset.

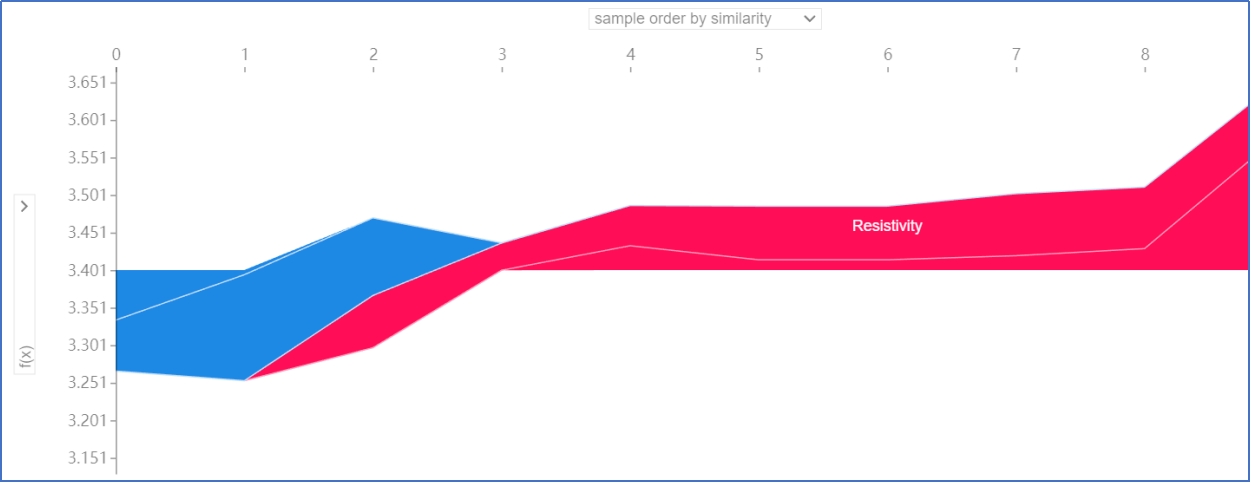


Fig 3 Force Plot – Entire Model

The bee swarm can give us an outlook of features involved in this model. It can give us an idea which feature has more influence on the model. The bee swarm plot sorts the features by their influence on the model, that is, the sum of SHAP value magnitudes of samples. We can see that the most important features are the ones with the highest SHAP values while the SHAP values show the distribution of the impacts of the features.

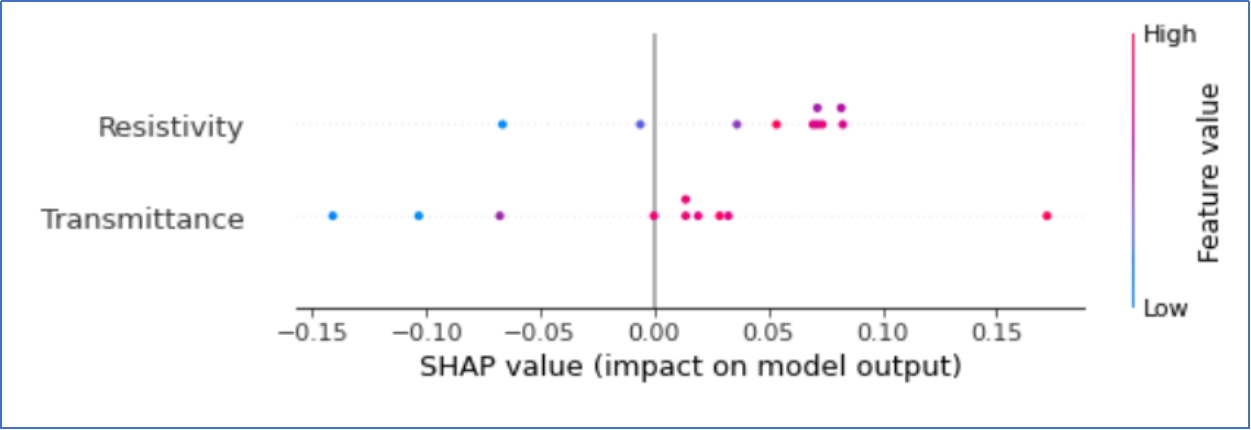


Fig 4 Bee Swarm – Outlook the features involved

This standard bar plot shows the mean absolute value of the SHAP value of both resistivity and transmittance.

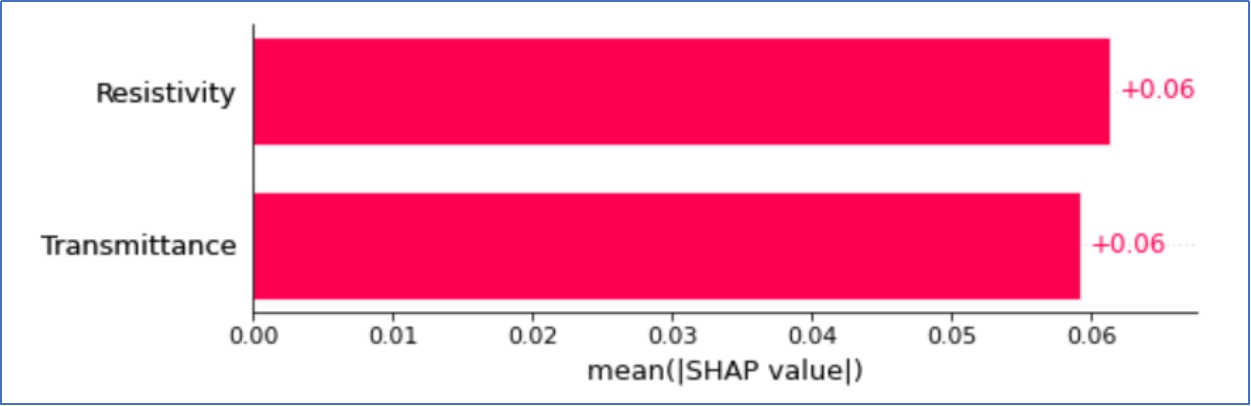
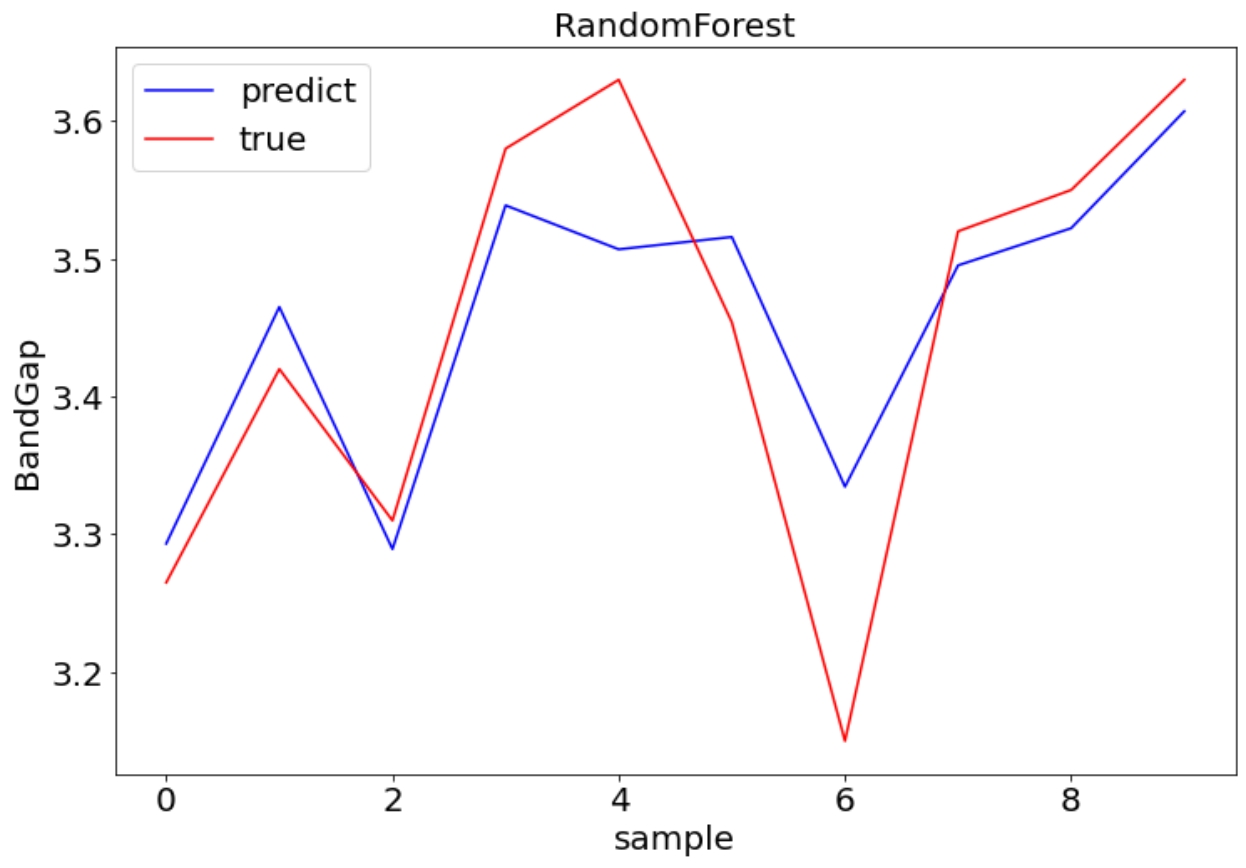
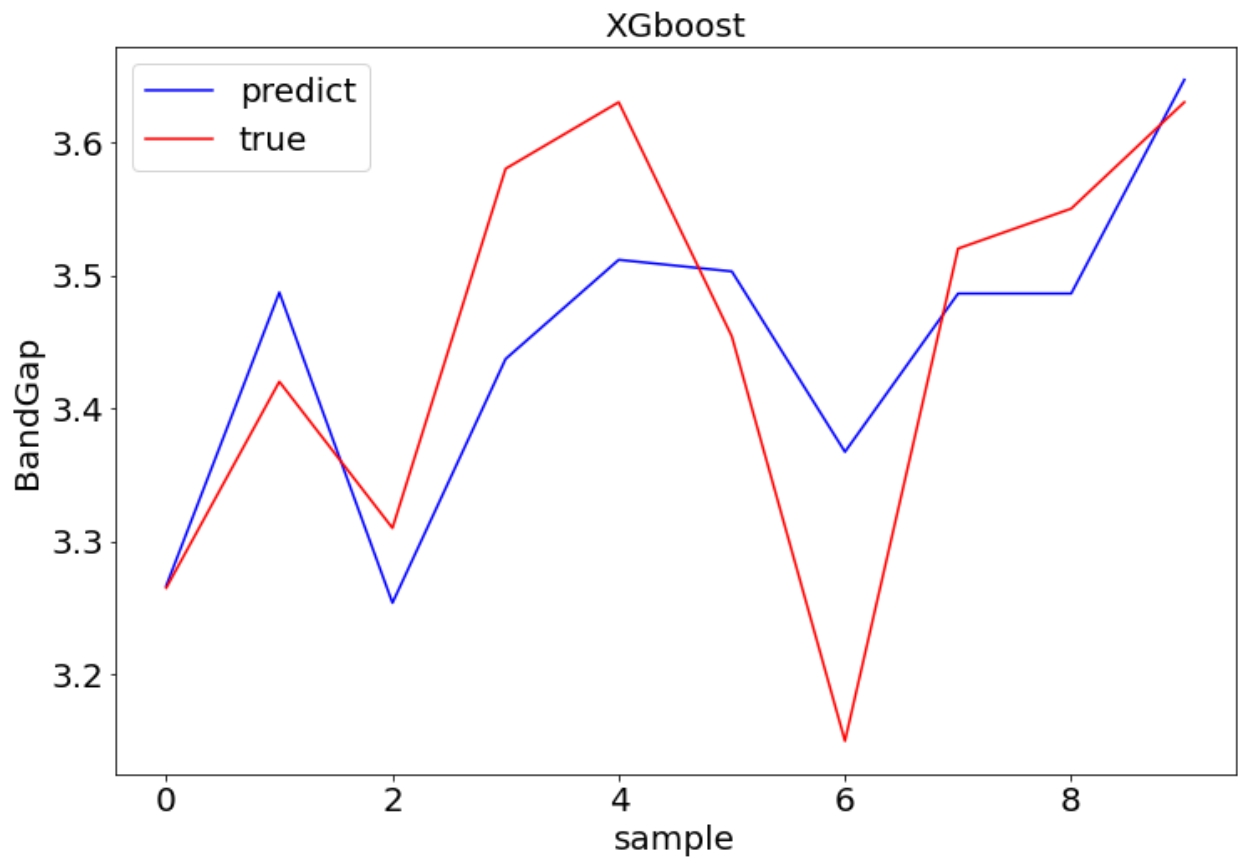


Fig 5 Bar Plot – Mean Absolute Value of the SHAP value

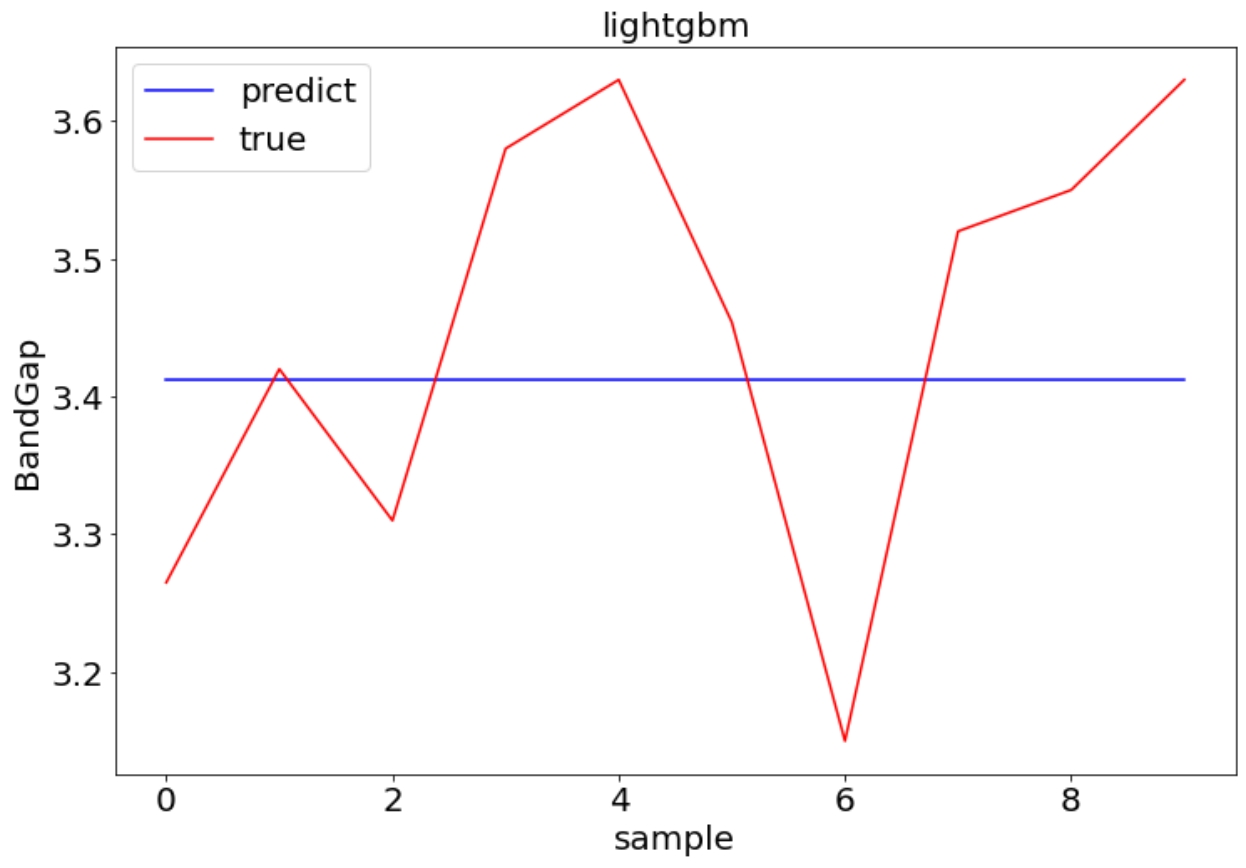
Possibly used Image:



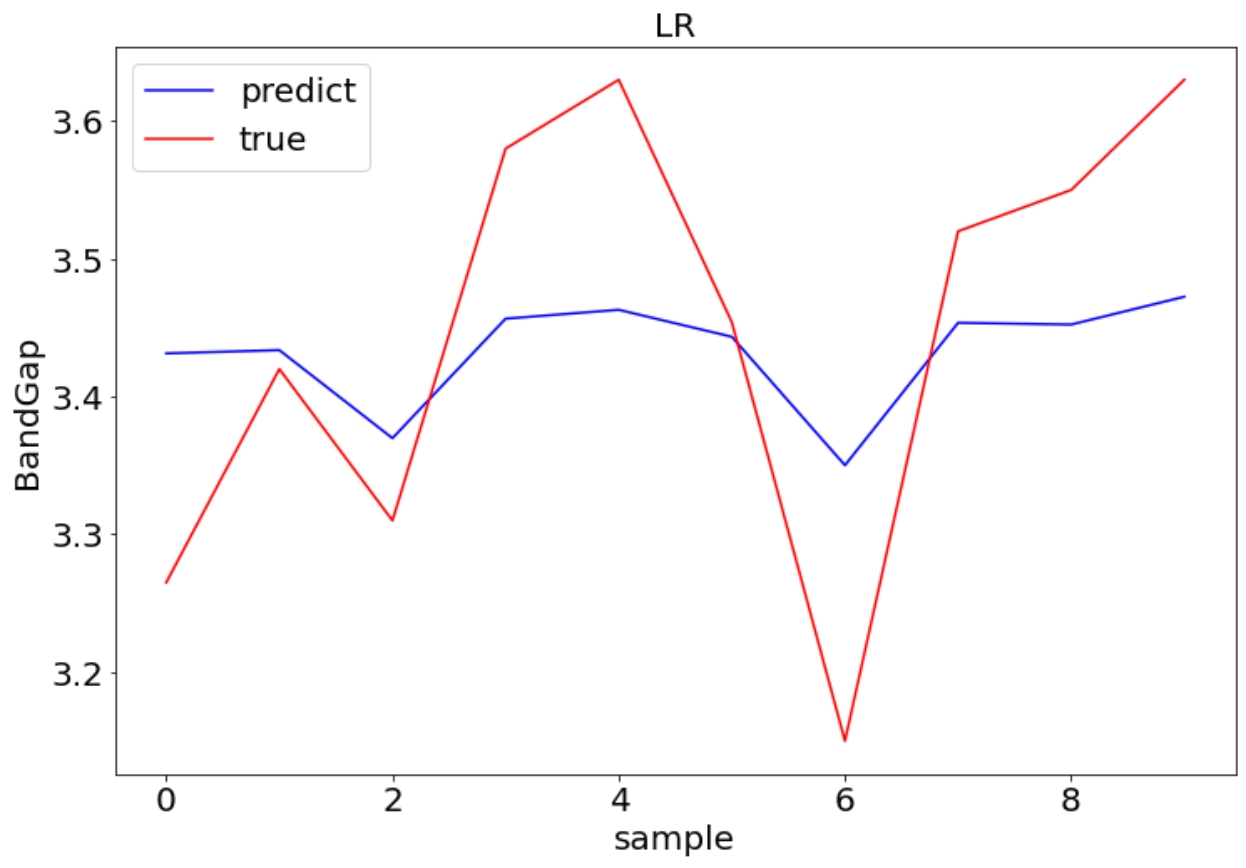
The difference between predicted and true value with random forest



The difference between predict and true value with XGboost



The difference between predict and true value with lightgbm



The difference between predict and true value with linear regression