PREDICTING MATERIAL PROPERTIES WITH MACHINE LEARNING

COLBY WIRTH, LUCAS MATHESON, BENJAMIN FRANKLIN

Abstract. Launched in 2013 as part of the Materials Genome Initiative, the Materials Project is an open-source collaboration in which computer and materials scientists have aimed efforts at making advances in materials science. Traditional methods like Density Functional Theory (DFT) can be computationally intensive. Thus, the Materials Project seeks to expedite research by using machine learning (ML), artificial intelligence (AI) and other predictive measures to accelerate the development of new materials [1]. This project focuses on the prediction of the properties of inorganic compounds with ML. We pursued two primary objectives: (1) training models on experimental data sets containing non-theoretical, lab-measured data—to predict specific features of elements, and (2) evaluating model performance against theoretical data sets derived from computational estimations (e.g., DFT or ML-generated data). This project demonstrates the capabilities of utilizing regression models to predict a collection of molecules' properties.

Keywords: Machine Learning, Artificial Intelligence, Inorganic Chemistry, Materials Project

1. Introduction

Since the 1970s, Density Functional Theory has been a widely used technique for predicting the properties of chemical compounds. This method is based on the Universal Functional, a computational approach used to simulate a molecule's quantum mechanical properties. However, the Universal Functional is limited by its computational complexity, as calculating exact results for complex systems can be computationally intensive [6]. In order for DFT to produce exact or highly accurate results, it is believed that the computational complexity of such calculations would require quantum computation. These problems are called the Quantum Merlin Arthur (QMA) class of problems. These QMA problems are analogous to the NP class of problems in classical computing, which are defined to involve non-deterministic computation for verifying solutions [6].

Machine Learning (ML) offers an alternative approach to DFT. ML models provide predictions in relatively low-intensity computation. In this project, our goal was to predict various molecular properties with low degrees of variance and high degrees of accuracy through the implementation various ML models with Python Sci-Kit Learn libraries. In the following sections, we outline our pipeline to include data pre-processing, feature selection, model selection, and an analysis of the models' performances.

2. Literature Review

New Machine Learning Algorithm: Random Forest

Authored by Yanli Liu, Yourong Wang, and Jian Zhang, and published Springer Nature Journal in 2012, this research paper discusses the underlying mechanisms and mathematics of Random Forests in categorization and regression problems.

The most critical information that pertains to this current project lies with their analysis of the performances of Random Forrest Regression (RFR), Support Vector Regression (SVR) and Linear Regression Models. Their data showed for one regression task, that the RFR model outperformed the SVR and Linear Regression models in all metrics but Root Mean Squared Error (RMSE). In this case, the Linear Regressor was marginally more performative; yet all other metrics favored the RFR model [3]. These results would guide us when assessing the validity of our results which each experiment.

3. Data Set Overview and Preprocessing Pipeline

The data was retrieved directly from the Materials Project official website [1]. At the time of the writing of this paper, the data set contains 155,361 compounds tabulated with 70 features. From the 70 features, 12 were either duplicated data from another column, or data related to the meta-data of the database, thus they were removed from the set.

Of the remaining data, there were categorical data that could easily be converted to numerical data with one-hot encoding functions. Some models, like the RFR model for example, can handle categorical data; however, we established a uniform data set for all models with each target feature. Thus, all categorical data was converted to numeric. The remaining unprocessed data included nested dictionaries containing intricate data regarding the lattice structure of the molecules. This lattice structure data could be encapsulated with a single feature that generally represented the structure of the molecule while abstracting from the intricacies of this complicated data. This step was inherently lossy, however it drastically simplified the data set and maintained the goal of only using numeric features.

After converting the categorical data to numeric, there were fifteen attributes that contained null data. Any feature with more than 50% null values were dropped altogether. The remaining null values were imputed if and only if the models used to predict the values could not handle null values. Only features that would be used as predictors or target values would have their null values mutated. For each attribute, the distribution was analyzed. None of the features were normally distributed, rather they were skewed right or skewed left - consequently the mean values were skewed. Therefore, the median values served as a better representation of the data sets, and all null values were filled with *median* values.

To finalize the data pre-processing, the data was split into two subsets: theoretical and non-theoretical data. According to the Materials Project documentation, the non-theoretical molecules have been measured in a lab environment, where the theoretical molecules have not [5]. These serve as two different benchmarks to

measure our modes against: the non-theoretical molecules contain measured data, compared to the theoretical molecules that are composed of data from DFT other predictive models.

4. Methodology

Target Features:

Four target features were selected for model training. Predictor features were included only if they had a null-value rate below 5%, ensuring minimal modifications to the dataset.

1. Band-Gap: the property associated with the amount of energy required to move an electron from an atom's valence band to its conductive band. It is measured in electron vols (eV) [2].

2. Density: Density is typically calculated with the equation: $D = \frac{M}{V}$. However if the compound cannot be synthesized, then a mass and volume cannot be measured. Therefore, predictive methods must be employed.

3. Energy Above Hull: The energy above hull is a measure of a material's thermodynamic stability, calculated from a convex hull of formation energies and compositions. Materials on the convex hull are stable, while those above it are metastable or unstable, with larger values indicating potential instability and difficulty in synthesis [4].

4. Total Magnetization: The Total Magnetization property is the vector sum of the magnetic moments of all the individual particles or atoms within a system [7]. It is traditionally calculated with DFT.

Model Selection and Hyperparameter Tuning

For all target features, it was presumed that their relationships to the predictor features are non-linear, given the complex nature of the data. Across all target values a RFR model and a selection of one or more of the following regressor models were trained: Decision Tree Regressor (DTR), Support Vector Regressor (SVR), and KNN Regressor. The data from the DTR, KNN and SVR models can be found in the Appendix, as this paper focuses on the best performing models, the RFR models.

The hyperparameters for each model were tuned using an iterative process in which each hyperparameter was treated as an independent module and updated individually until the evaluation criteria converged on optimized results. While this is a naive approach, it was generally straightforward and enabled a quick selection of hyperparameters. However, this method likely resulted in a collection of local optima rather than achieving a true global optimum, as parameters were tuned independently.

Evaluation Criteria

Two metrics were used to evaluate the performance of each model: R^2 and the relative root mean square error, $RRMSE$. The R^2 describes the variability of each model when predicting target values. This can also be understood as the precision of the model. The $RRMSE$ normalizes the $RMSE$ with respect to the target mean values. It describes the accuracy of the model. The $RRMSE$ is not as standard of a choice as the mean absolute error MAE or the mean squared error MSE. However, the normalization provided by $RRMSE$ enables a standardized, unitless analysis across all models. This is not the case for MAE or MSE, as these metrics are based in the unit of the target feature.

5. Experiments

The most performative models across all predictor features were the Random Forest Regression models. The performance of these RFR models are covered below. The metrics provided are the highest achieved after hyperparameter tuning.

Band Gap

For the band gap feature, the RFR model achieved R^2 scores of 0.999, 0.994, and 0.992 with the train, test, and theoretical sets, respectively. This indicates that the data were fitted to the model with high precision and low variance. The corresponding $RRMSE$ scores were 0.030, 0.104 and 0.138. This indicates high accuracy across all data sets and moderately reduced performance with the test and theoretical sets.

After further review, it was determined that two of the predictor features used in the models, VBM and CBM, had a direct correlation with the band gap, which could be trivially calculated using the equation $E_g = CBM - VBM$. Interestingly, the data set contained 43,644 null values for both VBM and CBM, however these null values were not imputed as the RFR and DTR models can handle null values. Despite approximately 28% null values for these two features, the model still produced highly precise and accurate predictions for all molecules.

Dataset	R-squared	Relative RMSE
Train Set	0.999	0.030
Test Set	0.994	0.104
Theoretical Set	0.992	0.138

Figure 1. Predicting Band Gap with Random Forest Regressor

Density

For the density feature, the RFR model achieved R^2 scores of 0.742, 0.668, and 0.534 with the train, test, and theoretical sets, respectively. These figures indicate that the data were fitted with moderate precision with some variance across all data sets. The corresponding $RRMSE$ scores were 0.273, 0.299 and 0.363. This indicates moderately high accuracy across all data sets and slightly reduced performance with the test and theoretical sets.

Dataset	R-squared	Relative RMSE
Train Set	0.742	0.273
Test Set	0.688	0.299
Theoretical Set	0.534	0.363

FIGURE 2. Predicting Density with Random Forest Regressor

Energy Above Hull

For the energy above hull feature, the RFR model model achieved R^2 scores of 0.734, 0.660, and 0.5316 with the train, test, and theoretical sets, respectively. These figures indicate that the data were fitted with moderate precision with some variance across all data sets. The corresponding $RRMSE$ scores were 1.543, 1.763 and 1.643. This indicates low accuracy across all data sets and marginal reductions in performance with the test and theoretical sets.

Dataset	R-squared	Relative RMSE
Train Set	0.734	1.543
Test Set	0.660	1.763
Theoretical Set	0.516	1.646

FIGURE 3. Predicting Energy Above Hull with Random Forest Regressor

Total Magnetization

For the total magnetization feature, the RFR model achieved R^2 scores of 0.853, 0.813, and 0.797 with the train, test, and theoretical sets, respectively. These figures indicate that the data were fitted with moderately high precision with moderately low variance across all data sets. The corresponding $RRMSE$ scores were 1.135, 1.242 and 0.961. This indicates low, but consistent accuracy across all data sets and marginal reductions in performance with the test and theoretical sets.

Dataset	R-squared	Relative RMSE
Train Set	0.853	1.138
Test Set	0.813	1.242
Theoretical Set	0.797	0.961

FIGURE 4. Predicting Total Magnetization with Random Forest Regressor

6. Conclusion

This project illustrates the potential of using machine learning models to predict the properties of inorganic compounds from the Materials Project API. The Random Forest Regression Models consistently produced the best results across all metrics and for all target values. While the predictions for band gap and total magnetization yielded results that may be considered as acceptable, the models trained on density and energy above hull produced over-fitted models.

When compared to the theoretical data set, the RFR model for band gap achieved an R^2 score of 0.979, indicating results acceptable. However, it was determined later on that this model predictor values that were directly correlated with the target feature.

The next best-performing model was the RFR model for total magnetization feature which achieved an R^2 score of 0.797. This indicates that that it is possible in some cases to predict the theoretical data with relatively high precision. However, the best performing models for density and energy above hull features produced low R^2 with scores of 0.534 and 0.516 respectively. The models for these attributes would likely not be considered to be acceptable. These outcomes demonstrate that building highly performative, acceptable models are not guaranteed when predicting the theoretical data in the Materials Project API.

One clear area of improvement in this project lies in the process of hyperparameter tuning, which was performed iteratively without the use of optimization algorithms. Future work could address this with the implementation of an optimization algorithm such as Grid Search. Further advancements could also include the integration of neural networks to find potentially more performative models.

7. Appendix

Dataset	R-squared	Relative RMSE
Train Set	1.000	0.000
Test Set	0.988	0.140
Theoretical Set	0.979	0.215

Figure 5. Band Gap with Decision Tree

Dataset	R-squared	Relative RMSE
Train Set	0.999	0.030
Test Set	0.994	0.104
Theoretical Set	0.992	0.138

Figure 6. Band Gap with Random Forest Regressor

Dataset	R-squared	Relative RMSE
Train Set	0.707	0.290
Test Set	0.624	0.328
Theoretical Set	0.416	0.406

Figure 7. Density with Decision Tree

Dataset	R-squared	Relative RMSE
Train Set	0.742	0.273
Test Set	0.688	0.299
Theoretical Set	0.534	0.363

Figure 8. Density with Random Forest Regressor

Dataset	R-squared	Relative RMSE
Train Set	0.748	1.503
Test Set	0.615	1.875
Theoretical Set	0.647	1.406

FIGURE 9. Energy Above Hull with Decision Tree

Dataset	R-squared	Relative RMSE
Train Set	0.734	1.543
Test Set	0.660	1.763
Theoretical Set	0.516	1.646

FIGURE 10. Energy Above Hull with Random Forest Regressor

Dataset	R-squared	Relative RMSE
Train Set	1.000	0.000
Test Set	0.390	0.181
Theoretical Set	0.112	0.489

FIGURE 11. Energy Above Hull with KNN Regressor

Dataset	R-squared	Relative RMSE
Train Set	0.844	1.170
Test Set	0.774	1.368
Theoretical Set	0.735	1.099

FIGURE 12. Total Magnetization with Decision Tree

Dataset	R-squared	Relative RMSE
Train Set	0.853	1.138
Test Set	0.813	1.242
Theoretical Set	0.797	0.961

FIGURE 13. Total Magnetization with Random Forest Regressor

Dataset	R-squared	Relative RMSE
Train Set	0.825	1.242
Test Set	0.788	1.322
Theoretical Set	0.692	1.185

FIGURE 14. Total Magnetization with Support Vector Regressor

References

[1] Jain, A., Ong, S. P., Hautier, G., Chen, W., Richards, W. D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G., et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL materials 1*, 1 (2013).

[2] LibreTexts. Basic properties of semiconductors: Band gap, 2024. Accessed: 2024-12-11.

[3] Liu, Y., Wang, Y., and Zhang, J. New machine learning algorithm: Random forest. In *Information Computing and Applications* (Berlin, Heidelberg, 2012), B. Liu, M. Ma, and J. Chang, Eds., Springer Berlin Heidelberg, pp. 246–252.

[4] Project, M. Glossary of terms, 2024. Accessed: 2024-12-11.

[5] Project, T. M. Materials project api documentation, 2023. Accessed: 2024-12-10.

[6] Schuch, N., and Verstraete, F. Computational complexity of interacting electrons and fundamental limitations of density functional theory. *Nature physics 5*, 10 (2009), 732–735.

[7] Zhang, H., Ravat, D., Marangoni, Y. R., Chen, G., and Hu, X. Improved total magnetization direction determination by correlation of the normalized source strength derivative and the reduced-to-pole fields. *Geophysics 83*, 6 (2018), J75–J85. Cited: Introduction Section.

Department of Computer Science, University of Southern Maine, Portland, ME