Spring 2023

Developing a Data-Driven Statistical Model for Accurately Predicting the Superconducting Critical Temperature of Materials using Multiple Regression and Gradient-Boosted Methods

Emil Agbemade
University of Central Florida, eagbemade@knights.ucf.edu

Developing a Data-Driven Statistical Model for Accurately Predicting the Superconducting Critical Temperature of Materials using Multiple Regression and Gradient-Boosted Methods

Emil Agbemade
dept. Statistics and Data Science
University of Central Florida
Orlando, United States
emil.agbemade@ucf.edu

***Abstract*—This study focuses on developing a statistical model for estimating the superconducting critical temperature (Tc) of materials using a data-driven strategy. The study analyzed 21,263 superconductors and used a combination of multiple regression and gradient-boosted models to make predictions. The analysis included a descriptive analysis of the distribution of Tc, feature selection using the Backwards selection method, and model diagnostics. The results showed that the gradient-boosted method outperformed the multiple linear regression method with an RMSE of 12.01 and an R^2 value of 88.23 after fine-tuning its hyperparameters. The study concludes that the gradient-boosted method is an effective approach for accurately predicting Tc in superconducting materials.**

***Index Terms*—superconducting critical temperature, multiple regression, gradient-boosted model**

I. Introduction

The zero-resistance current-conducting properties of superconducting materials have many real-world applications. Magnetic Resonance Imaging is probably the most well-known example of this technology. Magnetic Resonance Imaging scanners are routinely used by medical professionals to obtain detailed images of the inside of patients' bodies. Power grids may be radically altered by the discovery of new superconductors industries, as frictionless superconducting wires and power grids could theoretically eliminate energy loss during transmission and distribution.

There are two main problems that have slowed the widespread use of superconductors: At or below its superconducting critical temperature, a superconductor conducts current with zero resistance Critical Temperature (Tc). For a superconductor to show its zero-resistance property, it must be cooled to temperatures well below the boiling point of nitrogen (77 K), which is often impractical.

A new method has emerged, however, to probe the underlying factors that set the superconducting critical temperatures (Tc) of materials. Over the years, numerous databases documenting a wide range of material properties, both observed and predicted, have been compiled. In this study, we use data-driven strategy to develop a statistical model for estimating Tc from the available information. The superconductor information is sourced from the UCI machine learning data repository. There were 21,263 superconductors at our disposal, and the data had already been preprocessed. We primarily only tested two distinct statistical frameworks. Due to this, the main prediction model will consist of a combination of a multiple regression model and a gradient-boosted model. We found that the gradient boosted method is superior to the multiple linear regression method with RMSE of 12.01 and R^2 value of 88.23. The gradient boosted method yielded more accurate results after we made some additional adjustments to its parameters.[1]

II. Analysis

There are primarily two parts to this analysis section. To this end, we employ a descriptive analysis of the superconductor and data-driven machine learning techniques to access the model accuracy.

A. Descriptive analysis

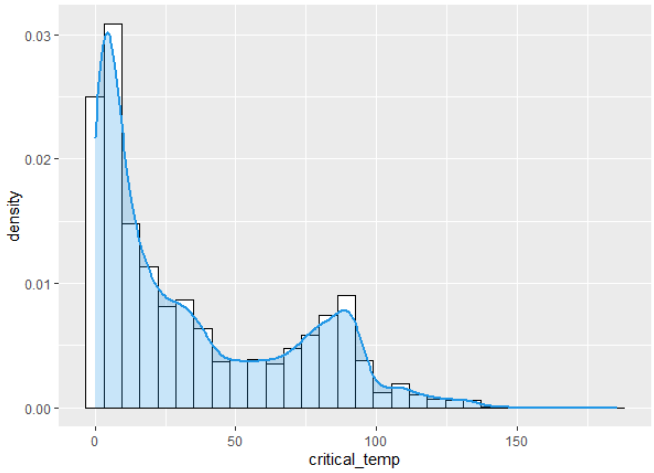

Fig. 1. Distribution of the superconducting critical temperature

The distribution of superconducting critical temperature as depicted in Fig.1 is not normally distributed. We made an effort to perform several transformations, including the log and Box-Cox transformations. These transformations did not yield a favorable outcome. We have discovered that even after the transformation, the critical temperature remained skewed.

TABLE I

Min	1st Qu.	Median	Mean	3rd Qu.	Max	Std
0.00021	5.365	20.00	34.42	63.00	185.00	34.25

Table 1 presents the results of our statistical analysis of the critical temperatures of all 21,263 superconductors. Here we see the distribution of superconducting critical temperatures, with minimum, first quartile, median, third quartile, maximum, and standard deviation indicated in the column headings.

B. Multiple linear regression

The factors may be more easily understood if a regression model is created Modulating the critical temperature of established superconductors and constituting a natural component of a system for locating undiscovered ones. We performed feature selection using the Backwards selection method with a probability value of 0.05 before running the linear regression model. This allowed us to narrow our model down to only the most important features. The training data set consist of 16952 data points and 70 significant features. The residual mean square error (RMSE) of 18.08 and an R^2 of 0.73 was reported from our test set prediction. The multiple linear regression predict quite accurately.

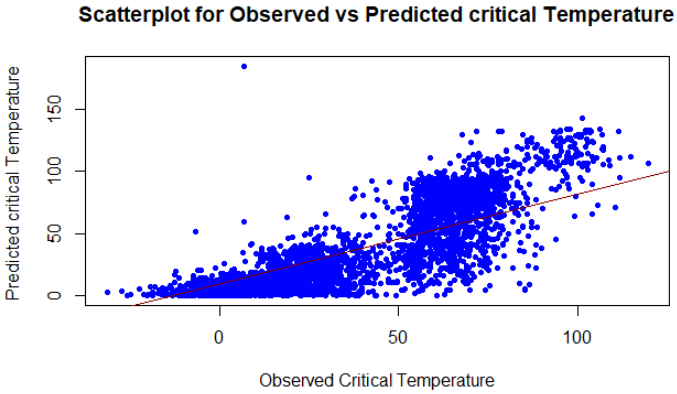

Fig. 2. Predicted vs Observed critical temperature

Based on the results of the multiple regression model, Fig.2 compares the critical temperatures that were predicted to those that were actually observed. The RMSE reported by the model was 18.08, and its R^2 value was 0.73. Additionally, we analyzed all of the model diagnostics, which included the Q-Q plot and the residual plots, which were located in the appendix section of the document.

Training error refers to the average error that results from

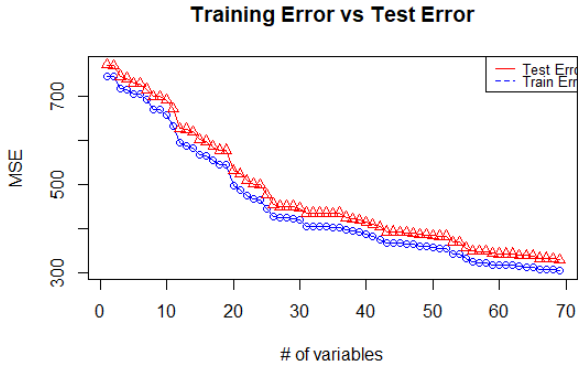

Fig. 3. Training Error vs Test Error

making predictions about responses based on the training data set.[2] The test error is the typical error that results from attempting to predict the response using the testing data set, which was not utilized in the training of the model. Comparing the training error and the testing error from our multiple linear regression shown in Fig.3, we conclude that there is little difference between the two errors. However, as the number of features increased, the testing error decreases.

C. Gradient boosting method

A representation of a decision tree that can handle large amounts of complex data is provided by the Gradient Boosting machine learning boosting system. It is based on the supposition that the next possible model will minimize the gross prediction error if it is combined with the models that came before it. The decision trees are utilized in order to make the most accurate predictions possible. The application of this method helps to simplify the prediction work. To improve the accuracy of this model, we debated whether to split the data in half, 80 percent to 20 percent, before running the analysis. Once again, we stripped away the unnecessary details. We constructed the model and made predictions using the validation data. This model had R^2 of 84.83 and an RMSE of 13.67.

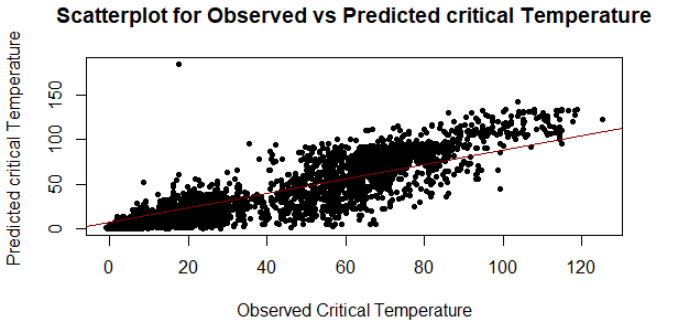

Fig. 4. Predicted vs Observed critical temperature

Fig.4 shows a predicted verses the actual values plot with an RMSE of 13.67 and R^2 of 84.83, it makes more accurate predictions. Indicating the gradient boosting approach is effective.

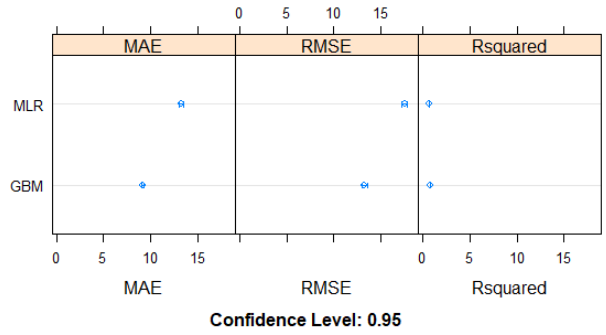

Fig. 5. Comparison of the base models

We compare the two outcomes using their respective MAE, RMSE, and Rsquare values. Observing Fig. 4, each of these variables has better results than the multiple linear regression. We did some fine-tuning of the model's hipper parameters despite the fact that the base gradient boosting already outperforms our multiple linear regression. The tuning led to a further drop in the RMSE to 12.01 and an increase in R^2 to 88.23.

D. Conclusion

In order to make an accurate prediction of the superconductor critical temperature, we made use of a couple of machine learning algorithms. The primary objective is to examine and contrast the two algorithms in order to choose the model that provides the highest level of accuracy. We found that the gradient boosted method is superior to the multiple linear regression method in terms of with RMSE of 12.01 and R^2 value of 88.23. The gradient boosted method yielded more accurate results after we made some additional adjustments to its parameters.

REFERENCES

[1] Kam Hamidieh. "A data-driven statistical model for predicting the critical temperature of a superconductor". In: *Computational Materials Science* 154 (2018), pp. 346–354.

[2] Robert E Schapire and Yoram Singer. "BoosTexter: A boosting-based system for text categorization". In: *Machine learning* 39.2 (2000), pp. 135–168.