Predicting the Critical Temperature of Superconductors

Yonaton Heit

April 18, 2019

1 Objective

The purpose of this project is to predict the critical point of superconductors using features of properties of the elements that make up the superconductors. Clustering is unitized in hopes to improved accuracy.

2 Introduction

Superconductors have a multitude of applications. The most famous application is the superconducting magnet in magnetic resonance imaging (MRI). Superconductors were used in the superconducting coils in the Large Hadron Collider at CERN to maintain a high magnetic field, which verified the existence of the Higgs boson as predicted by the Standard Model of particle physics. Superconductors could potentially replace components in electronically powered systems to improve power quality and increase system reliability.[1]

The defining quality of all superconductors is their superconductivity. Superconductivity is a phenomenon in which a material has zero resistance, allowing it to maintain an electric current indefinitely. All known superconductors only have this property at low temperatures. The temperature at which a material loses superconductivity is called its critical temperature.

At this time, there is no universal theory on superconductivity.[2] According to Bardeen-Cooper-Schrieffer (BCS) theory, as proposed in 1957, in superconductors, electrons are bound together as Cooper pairs. As they travel along the metal, they distort the lattice structure, allowing electrons to freely flow through the material.[3] As a quantum mechanical explanation, superconductivity cannot be fully understood with Newtonian physics. At higher temperatures, the lattice vibrations decouple the Cooper pairs, which according to the theory, leads to the loss of superconductivity. In 1986, a superconductor

was created with a critical temperature of 35 K.[4] A year later, another superconductor was created with a critical temperature of 93 K. [5] These are higher critical temperatures than were predicted as possible by BCS theory. Other theories that have been proposed are resonating-valence-bond theory [6] and spin fluctuation theory. [7]

In this project, rather than trying to predict the critical point of superconductors from theory, which is difficult because of the lack of a universal theory, we will built a machine learning model based on other features such as atomic mass and first ionization energy. This project is expanding on the work of Hamidieh.[8] While we are using the same data as Hamidieh, our model clusters the superconductors in the hopes of increasing accuracy.

3 Data Set Overview and Wrangling

The data used were the same as used by Hamidieh contain 81 features for 21,263 superconductors.[8] 80 features were derived from the 10 statistical measurements of 8 physical properties for the atoms that make up the material. The atomic physical properties were the atomic mass, the first ionization energy, the atomic radius, density, the electron affinity, the fusion heat, the thermal conductivity, and the valence (the typical number of chemical bonds made by the elements). The 10 statistical measurements of these properties were the mean, weighted mean, geometric mean, weighted geometric mean, entropy, weighted entropy, range, weight range, standard deviation, and weighted standard deviation. The difference between a weighted and unweighted statistical measurement was a consideration of the proportions of the elements in the material. For example B_6Y_1 has a mean atomic mass of 49.86 amu, which is the mean of the atomic mass of boron (10.81 amu) and yttrium (88.91 amu). The weighted atomic mass is 21.97 amu which weights the atomic mass based on the proportion of each element (6/7 for boran and 1/7 for yttium). For more details see Hamidieh's paper.[8] The last feature was the number of elements in the superconductor. The database can be found at UCI Machine Learning Repository.

Wrangling of the data set was already performed by Hamidieh. It contained no missing values or values that had to be removed. Since the values of the features were based on the properties of the elemental components rather than the properties of the superconductor themselves, the values were easy to determine. If the values were based on the superconductors, there would be a concern of missing values for superconductors less examined than the highly studied elements.

4 Data Statistics:

4.1 Breakdown by Element

Table 1 shows the number and percentage of superconductors for the 10 most abundant elements in superconductors. Figure 1 shows the number of superconductors for all elements present. The most abundant elements were oxygen and copper appearing in 56.27% and 50.97% of superconductor in data set respectively. 41.18% of superconductors contain neither. There are 77 elements are present in superconductors data set and 60 were present 5% of the time. Because a high but not overwhelming number of superconductors contain oxygen and copper, there may be some use in including a feature stating whether these 2 elements are present in a predictive model. This was not explored in this project but warrants future examination. Since the properties used by Hamidieh were derived from the elements present rather than the superconductors themselves, the properties are a proxy for the elements. It would be interesting to examine the performance of a model from the elements present rather than the properties. This may be examined in the future.

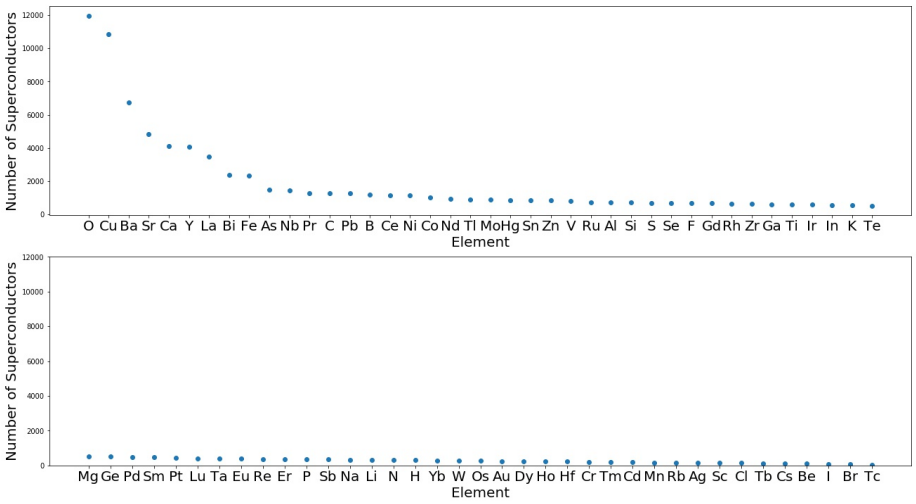

Figure 1: Number of Superconductors with a given element. The elements are ordered greatest to least. The bottom plot is a continuation of the top.

4.2 Property Distribution

Figures 2, 3, and 4 show histograms overlapping with kernel density estimation (kde) distributions of the features and the critical temperature. While there are 81 features, many of them are different statistical measurements of the same properties. In order to reduce the number of features shown to a

	Percent(%)	Number
O	56.27	11964
Cu	50.97	10838
Ba	31.75	6751
Sr	22.82	4852
Ca	19.34	4112
Y	19.16	4075
La	16.29	3463
Bi	11.24	2389
Fe	11.0	2339
As	7.06	1502

Table 1: The percentage and number of superconductors with a given elements. Only the ten most common elements are listed.

reasonable amount, only the weighted means are displayed. For many of the properties, there appears to be a bimodal distribution. This is a particularly important observation for the critical temperature since it is the only property that belongs to the superconductor itself rather than the component elements. Other properties that observed this bimodal pattern were first ionization energy, atomic radius, and density. The superconductors were clustered as shown by Figure 2-4 and the clustering details are in Section 5.1.

4.3 Inferential Statistics

Figures 5 and 6 show the scatterplots of features and the critical temperature of superconductors. Similar to Section 4.2, only the weighted mean features were considered. A linear fit for each feature was created and the Pearson's correlation of these relationships was between 0.2 and 0.5 (positive correlation) or -0.2 and -0.5 (negative correlation). The results were statistically significant. ($p<0.001$) This indicates that these features are correlated with the critical temperature and should provide predictive value to a model.

5 Modeling Predictions

5.1 Clustering

This bimodal pattern, as seen in Figures 2-4, suggest that there were two separate populations of superconductors within the dataset. The superconductors were separated into two clusters with K-means clustering using the 81 features and the critical temperature. Attempts to cluster were also made with both HDBSCAN and affinity propagation. Both methods created thousands of clusters rather than the desired two.

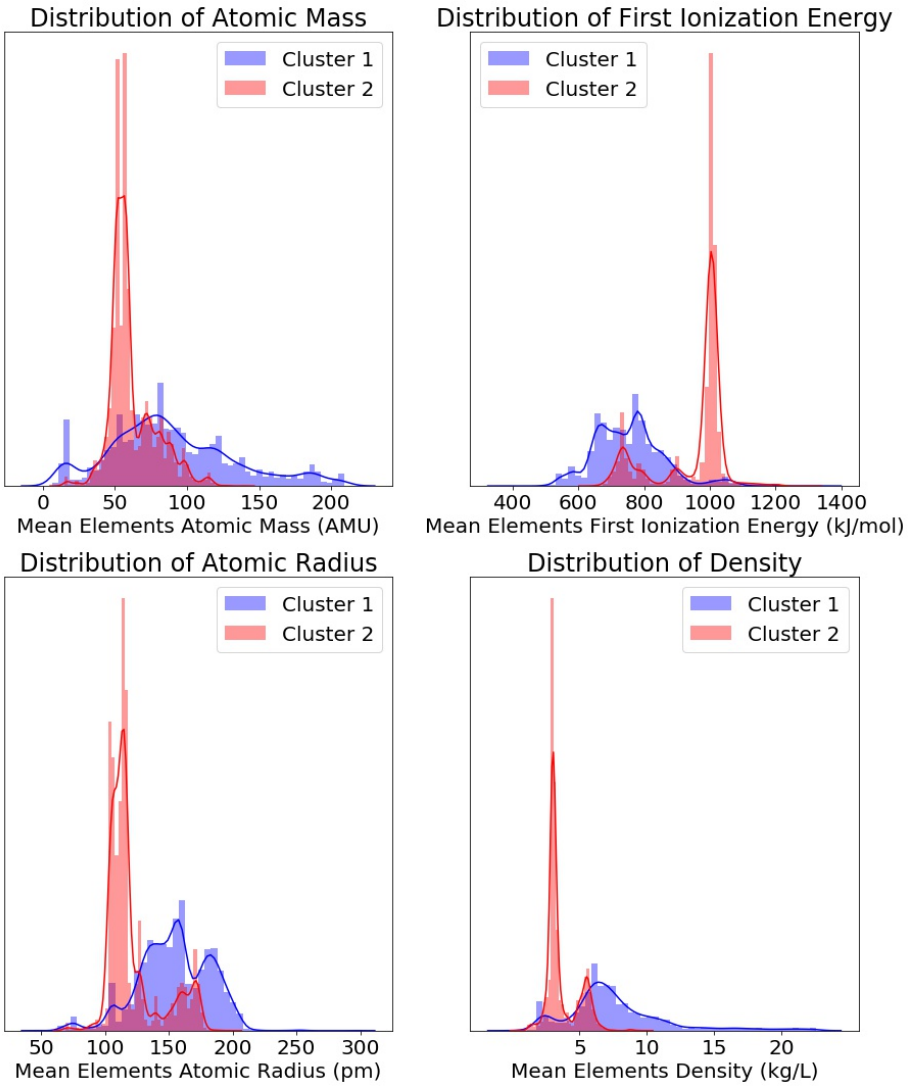

Figure 2: Histogram of the weighted means of the atomic mass, first ionization energy, atomic radius, and density of the elements of the superconductors. The data are color coded by cluster.

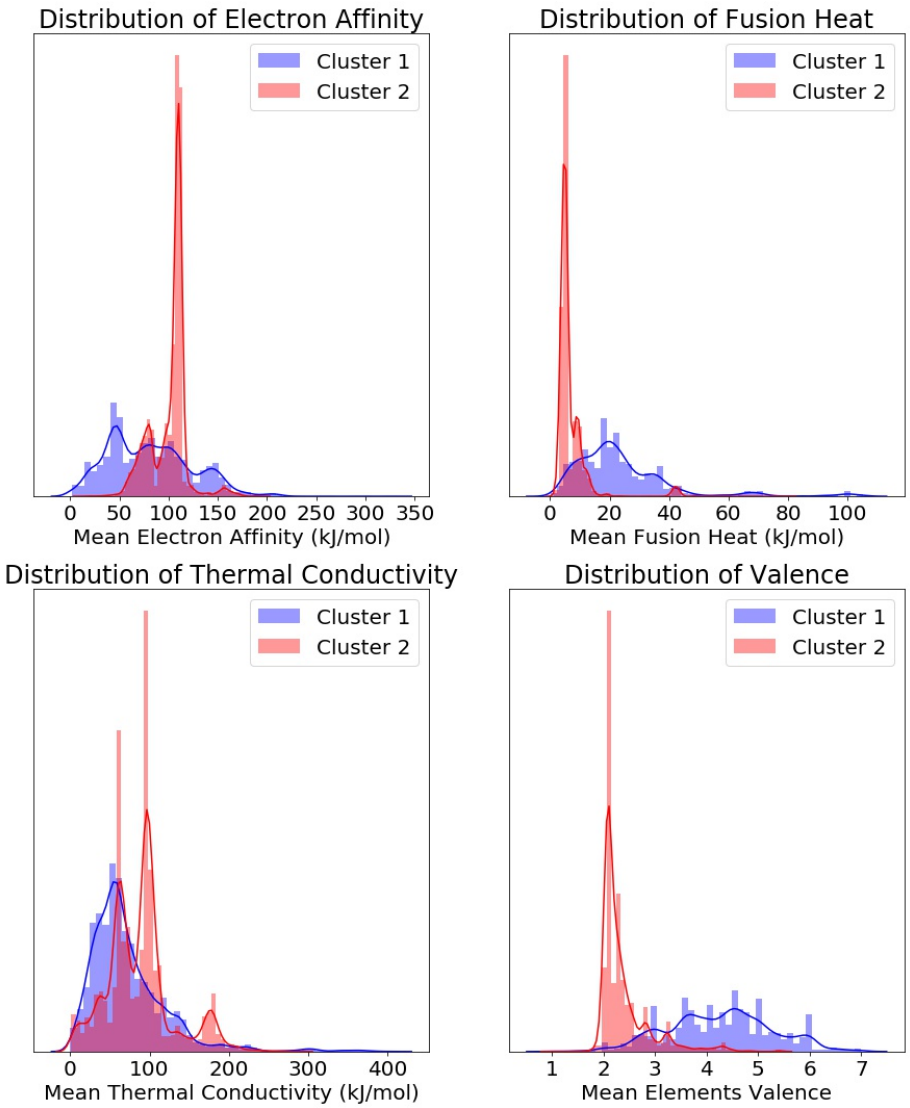

Figure 3: Histogram of the weighted means of the electron affinity, fusion heat, thermal conductivity, and valence of the elements of the superconductors. The data are color coded by cluster.

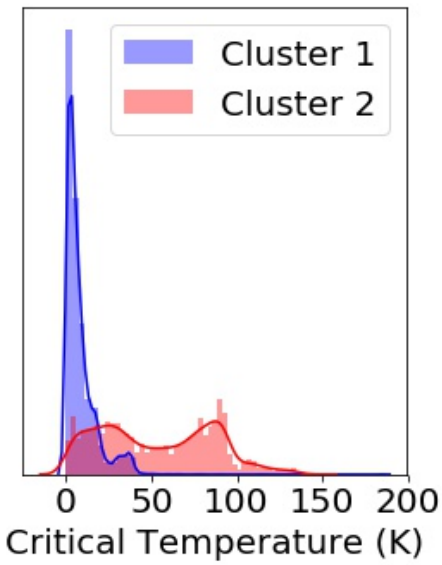

Figure 4: Histogram of the critical temperature of the superconductors. The data are color coded by cluster.

It is well known there are two types of superconductors, type I and type II. There are a few properties that distinguish between the two types such as their interaction with a magnetic field; the most importantly for this project is their critical temperature. Type I superconductors have lower critical points. It is reasonable to assume that the existance of two distributions is a result of the data set containing two types of superconductors. Unexpectedly, it was found that significant numbers of type II superconductors were placed in the lower critical temperature cluster. Therefore, it was concluded that the two distributions were not separated by type. The cause of the distribution being bimodal is unknown and the clusters were simple labeled "Cluster 1" and "Cluster 2". There were 8792 superconductors in Cluster 1 and 12471 in Cluster 2.

5.2 Linear Regression

The first attempted model was a linear least fit square regression. If a simple model is accurate, there is no reason to use a more complex model that may be more time consuming. The data was fitted in two separate methods. The first method both clusters were modeled together in a single linear fit. In total 14000 data point (7000 data points from each cluster) were placed into the training set as part of the training/testing data split. An equal number of data points from each cluster was used in order to avoid having unbalanced data. In the second method, the clusters are modeled separately. The same 7000 data

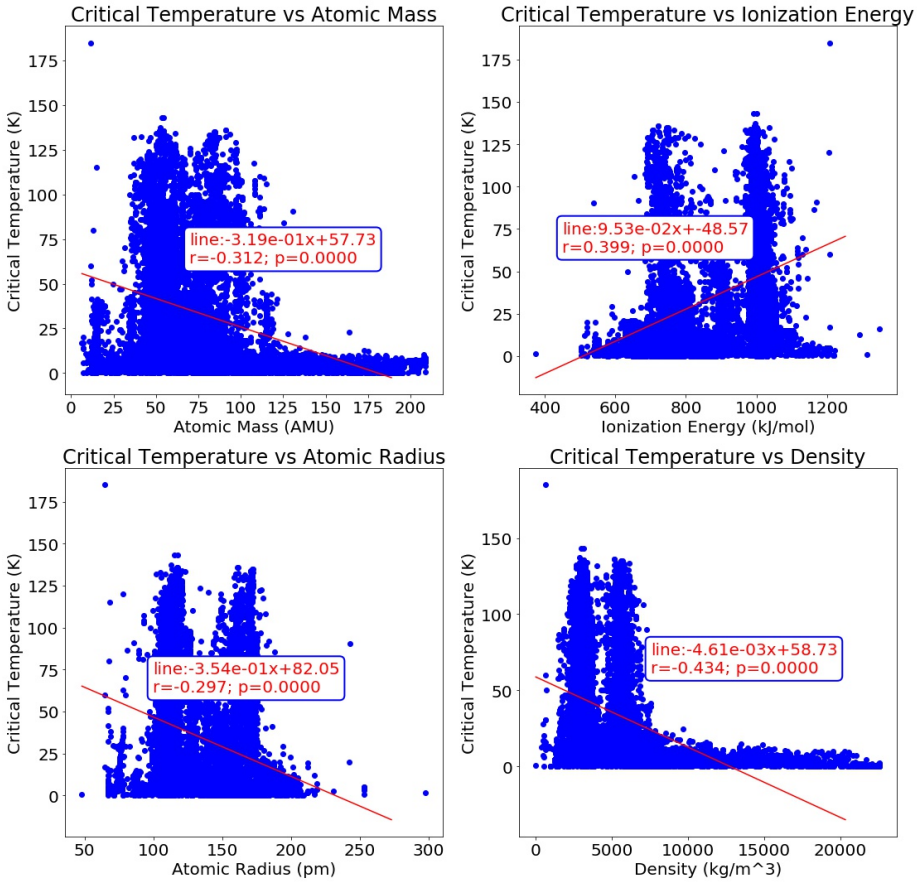

Figure 5: Scatter plot of the critical temperature of superconductors against the weighted mean of the atomic mass, first ionization energy, atomic radius, and density of the elements of the superconductors.

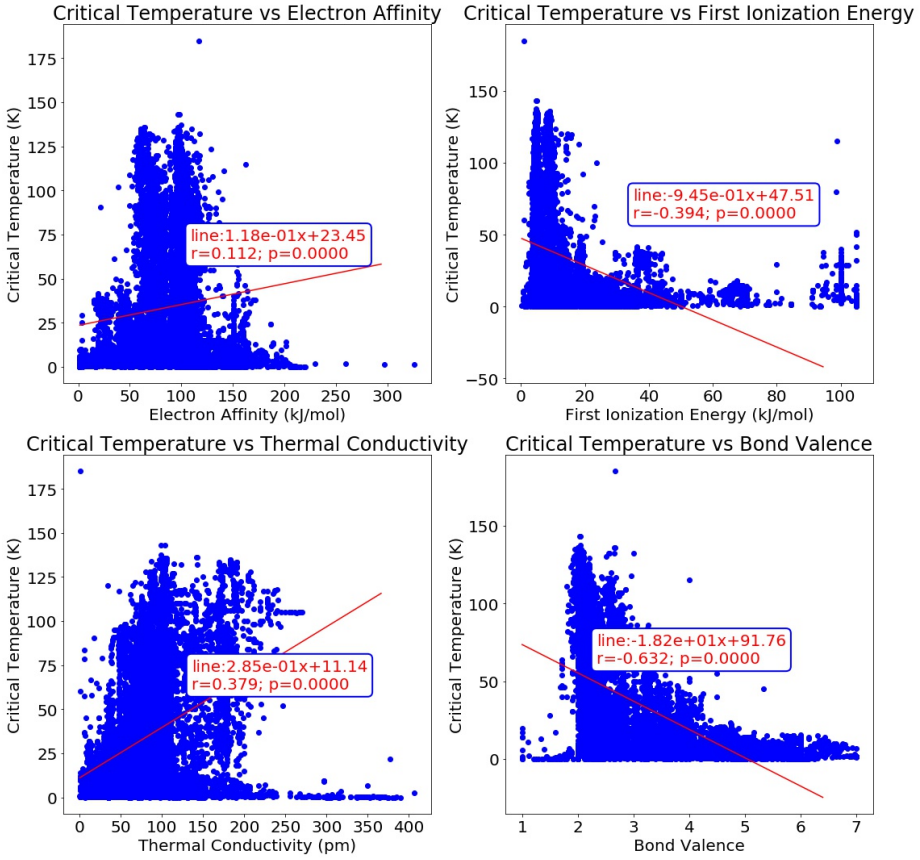

Figure 6: Scatter plot of the critical temperature of superconductors against weighted mean of the electron affinity, fusion heat, thermal conductivity, and valence of the elements of the superconductors

from each cluster were used for the training set.

Before fitting with a linear regression, the data for each of the 81 features were standardized by centering them at the mean and dividing by the standard deviation. Principal component analysis (PCA) was considered but the transformation would make it difficult to determine features importance (particular for the non-linear gradeint boosting method discussed in the next section). The scikit-learn library implemented into python was used for linear regressions. Since the gradient boosting model was more accurate and ultimately chosen, the weights for linear regression was not examined to determine the feature importance. The elastic net regularization was considered, but even after hyperparameterization, linear regression without regularization proved slightly more actuate on both the training and test data.

The results of the linear fit shown Figure 7. The out-of-sample root mean square error (RMSE) of the data when modeling the clusters together was 19.50 K and the out-of-sample R^2 was 0.70. When the clusters were modeled separately, the out-of-sample RMSE and R^2 were 17.36 K and 0.76, respectively. This showed modeling the clusters seperately improved accuracy. This is particularly true for Cluster 1, which RMSE lowered by more than a third (9.58 K to 5.98 K). In addition, the R^2 changes from negative to positive (-0.21 to 0.53). A negative R^2 indicates that the fit performed worst than a horizontal line and that the linear regression method that did not model the clusters poorly fitted Cluster 1.

5.3 Gradient Boosting Regression

In order to improve model accuracy, gradient boosting was next attempted. Gradient boosting is an ensemble method similar to Random Forests that forms an answer from multiple decision trees, called base learners, to make predictions. Gradient boosting differs from Random Foresting that the learners are created sequentially and the errors are corrected from previous learners.

The data were separated by cluster, separated into training/testing data, and standardized exactly as explained in Section 5.2. Two gradient boosting methods were performed, one modeling both clusters together and one where the two clusters were modeled separately.

The gradient boosting was performed with the XGBoost Python library and 4-fold cross-validation was performed using scikit-learn Python library. The XGBoost library contains a scikit-learn API but it was found to be slow so a custom API was created for the scikit-learn Python liberty in order to do cross-validation.

A random search of six separate parameters in 200 possible sets of com-

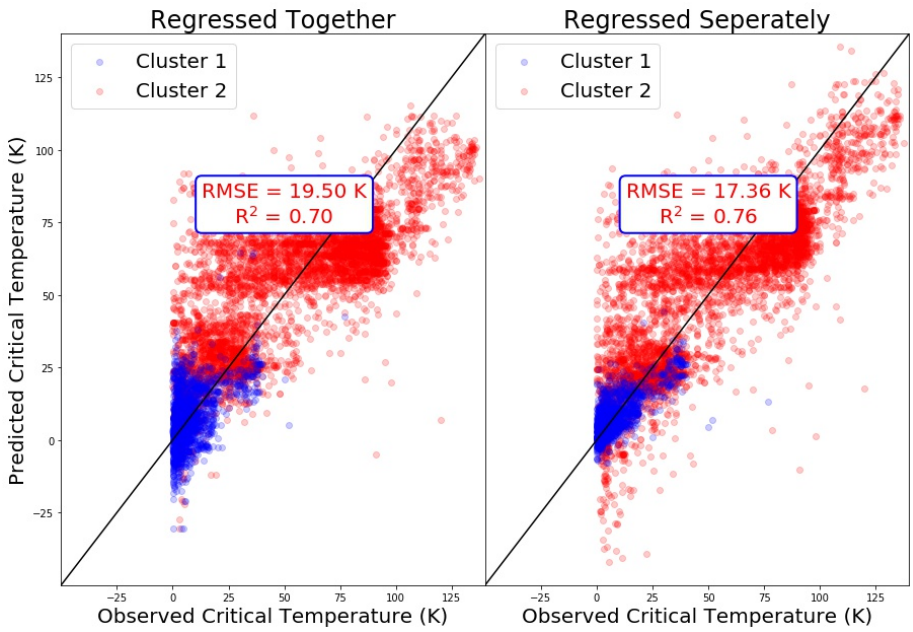

Figure 7: The scatter plot show the predicted vs observed critical temperature as analyzed with linear regression. The closer the observed value is to the predicted value (as shown by the black line) the more accurate the model. The clusters are color coded. On the left, the clusters were modeled together. On the right, the clusters were modeled separately.

binations was performed in order to tune the model. These parameters are the number of learners (n_estimators), the minimum sum of weight of all observations in a child (min_child_weight), the maximum tree depth per learn (max_depth), the learning rate (learning_rate), the ratio of data points used per learners (subsample), and the ratio of features used per learners (colsample_bytree). The optimized parameters for the model using all the data and model for each cluster is shown on Table 2 along with range of values searched for each parameter. The models created from these parameters were evaluated using R^2.

Parameters*	Both clusters	Cluster 1	Cluster 2	values searched
n_estimators	230	130	139	1,2,3..699
min_child_weight	18	3	13	1,2,3...20
max_depth	17	18	12	1,2,3...20
learning_rate	0.40	0.28	0.34	0.01,0.02,0.03...1.00
subsample	1.00	0.75	1.00	0.50, 0.75, 1.00
colsample_bytree	1.00	0.75	1.00	0.50, 0.75, 1.00

Table 2: The optimized value found for model using both clusters, Cluster 1, and Cluster 2. The final column is the values for the random search during parameterization.
*Parameters listed are the internal names within the XGBoost module.

Gradient boosting showed significant increase in accuracy over linear fitting. The out-of-sample RMSE and R^2 for the model using both clusters together was 11.33 K and 0.90, respectively. Modeling the clustering separately did not improve the accuracy. The out-of-sample RMSE and R^2 results when the clusters were modeled separately were 11.39 K and 0.90. The scatter plots comparing the observed critical temperatures to the predicted critical temperatures are shown on Figure 8.

While modeling the clusters separately did not improve accuracy, the different models for the clusters provide insight into what features are significant for predicting the critical temperatures. Table 3 shows the fraction gains of the 20 most important features. The fraction gain is defined by

$$\text{Fraction Gain} = \frac{\text{Total Gains for a feature}}{\text{Sum of Total Gains for all for features}} \tag{1}$$

Our results were consistent with those Hamidieh.[8] When the clusters were modeled together, the most important features were derived from the thermal conductivity. The atomic mass, valence, electron affinity, and density were also found to be important. The model using Cluster 2 found similar results with atomic radius also being significant (Hamidieh also found this to be a significant feature). The model using Cluster 1 has different important features

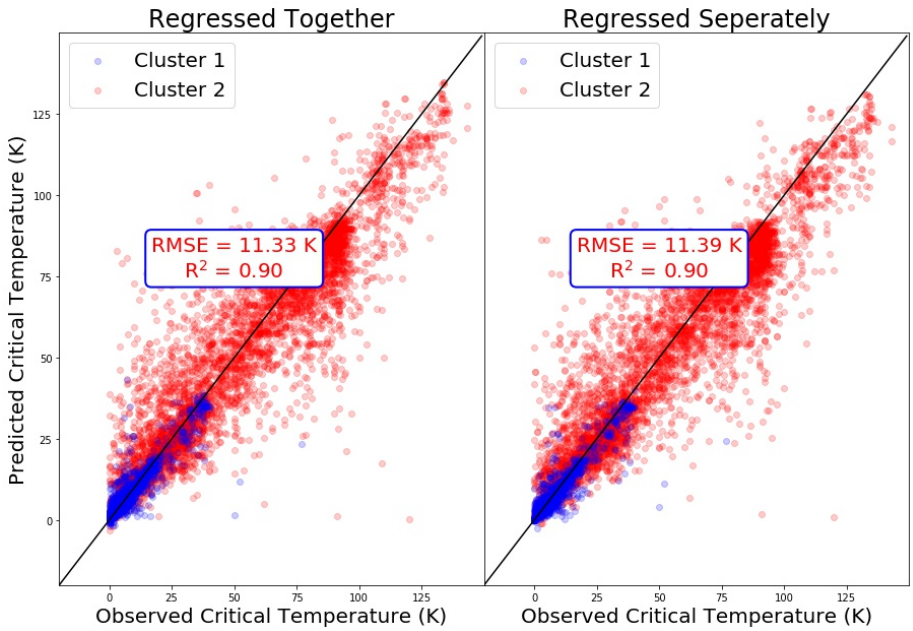

Figure 8: The scatter plot show the predicted vs observed critical temperature as analyzed with gradient boosting regression. The closer the observed value is to the predicted value (as shown by the black line) the more accurate the model. The clusters are color coded. On the left, the clusters were modeled together. On the right, the clusters were modeled separately.

Both Clusters	
Feature	Fraction Gain
Range thermal conductivity	0.59
Weighted geometric mean thermal conductivity	0.13
Standard deviation atomic mass	0.02
Weighted mean valence	0.02
Weighted geometric mean valence	0.02
Weighted standard deviation electron affinity	0.01
Weighted range atomic mass	0.01
Mean density	0.01
Standard deviation density	0.01
Weighted range valence	0.01
Cluster 1	
Feature	Fraction Gain
Weighted mean atomic mass	0.25
Range first ionization energy	0.11
Weighted mean valence	0.05
Weight geometric mean electron affinity	0.05
Mean first ionization energy	0.03
Weighted standard deviation thermal conductivity	0.03
Mean density	0.03
Weighted geometric mean atomic mass	0.03
Weighted entropy valence	0.02
Weighted entropy atomic radius	0.02
Cluster 2	
Feature	Fraction Gain
Weighted mean thermal conductivity	0.46
Weighted mean valence	0.07
Standard deviation atomic mass	0.06
Weighted geometric mean valence	0.06
Weighted standard deviation electron affinity	0.04
Range atomic radius	0.03
Weighted entropy thermal conductivity	0.01
Geometric mean electron affinity	0.01
Weighted mean electron affinity	0.01
Weighted range atomic mass	0.01

Table 3: The most important features when gradient boosting for the model using all the data, the model using only Cluster 1, and the model only using Cluster 2.

from the model using both clusters and the model using Cluster 2. The most important feature for the Cluster 1 model the was derived from the atomic mass. The first ionization energy was found to be significant while it was not by the other models or by Hamidieh. This demonstrates that the model using all clusters closer resembles Cluster 2's model. An equal number of Cluster 1 and Cluster 2 data was included into the training set to avoid unbalancing the model. Superconductors in Cluster 2 have a wider range of critical temperatures than Cluster 1 which may have introduced biases.

6 Conclusion

The models produced from gradient boosting were more accurate than those using linear modeling. This is as expected since linear regression model is a simpler model that would preform poorly for non-linear data. Separating the superconductors into two cluster improved the accuracy of the linear fit model but not the gradient boosting. This is most likely because the clustering introduced flexibility into a linear regression while the gradient boosting method was already flexible. Clustering allow examination of which feature were important for each cluster whereas modeling the clusters together would not.

References

[1] Hassenzahl, W. V.; Hazelton, D. W.; K., J. B.; Komarek, P.; Rei, C. T. *Proceedings of the IEEE* **2004,** *92,* 1655–1674.

[2] Mann, A. *Nature* **2011,** *475,* 280–282.

[3] Bardeen, J.; Cooper, L. N.; Schrieffer, J. R. *Phys. Rev.* **1957,** *108,* 1175–1204.

[4] Bednorz, J. G.; Müller, K. A. *Z. Phys.* **1986,** *64,* 189–193.

[5] Wu, M. K.; Ashburn, J. R.; Torng, C. J.; Hor, P. H.; Meng, R. L.; Gao, L.; Huang, Z. J.; Wang, Y. Q.; Chu, C. W. *Phys. Rev. Lett.* **1987,** *58,* 908–910.

[6] Anderson, P. W. *Science* **1987,** *235,* 1196–1198.

[7] Monthoux, P.; Balatsky, A. V.; Pines, D. *Phys. Rev. Lett.* **1991,** *67,* 3448–3451.

[8] Hamidieh, K. *Comput. Mater. Sci.* **2018,** *154,* 346–354.