Final Project

Tyler Chang

STA 160

Professor Fushing Hsieh

8 June 2020

Tyler Chang
914163826

Final Project

STA 160
June 8, 2020

Introduction

Superconductivity is a phenomenon where a charge moves through a material without any resistance. Thus, a superconductor is a material that can produce infinite conductivity. In practice, thermal conductors play a significant role in everyday live. One of the largest applications of superconductivity is magnetic levitation. Although the public is generally not aware, magnetic levitation is used with MRI scanners, high-speed rail systems, and even with advanced military technology.

In order for a superconductor to perform magnetic levitation, it must reach its critical temperature first. A material's critical temperature is the exact temperature at which the electrical resistance drops to zero. That is, a material's critical temperature is the exact point where it is able to produce infinite conductivity. Thus, countless physicists and researchers across the globe are quite interested in factors that play into a material's critical temperature as well as developing models to predict a given material's critical temperature.

The UC Irvine Machine Learning repository provides public data on 21,263 different superconductors. Not only does the data provide the material's respective critical temperature, but other information such as atomic mass, density, elements, and several other features are provided. The goal of this analysis is two sided and focused on critical temperature. First, we shall explore certain features and their relationship to critical temperature. Then, we will train several machine learning models in order to accurately predict a material's critical temperature

Exploratory Data Analysis

In this section of the analysis, we will begin with an exploratory data analysis of the data set. In particular, we wish to determine if any patterns can be drawn from certain features. The data set contains 81 different features, ranging from density to atomic mass, along with the elements a given material is made of.

Chemical Elements

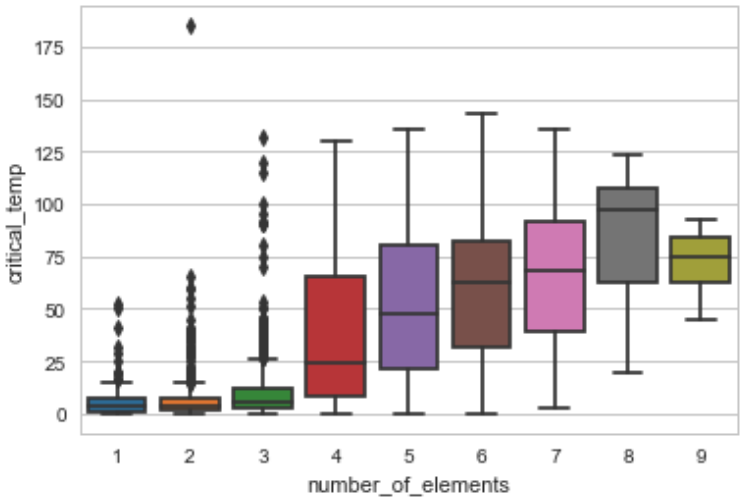

Figure 1: Number of Elements vs. Critical Temperature

Figure 1 displayed above shows the relationship between the number of elements a material has and its critical temperature. Based on the plot, we can see a general increasing trend in critical temperature as a material has more elements. Furthermore, it appears that materials with fewer elements (1-3) vary much less in temperature than materials holding more elements (4-9). Thus, it may be easier to classify materials with less elements than materials with more. The elements that were prevalent for low numbers of elements (1-3) appeared to be specific for that grouping. However, after investigating materials with a medium (4-6) and large (7-9) number of elements, it seems that there were three elements that were consistently prevalent- Oxygen (O), Copper (Cu), and Strontium (Sr).

To further investigate the relationship between elements and critical temperature, we were interested in observing which elements are needed at lower temperatures, and which elements were prevalent at different levels of temperatures. One preliminary pre-processing step that needed to be taken before determining the different temperature levels was transforming the target variable, critical temperature.

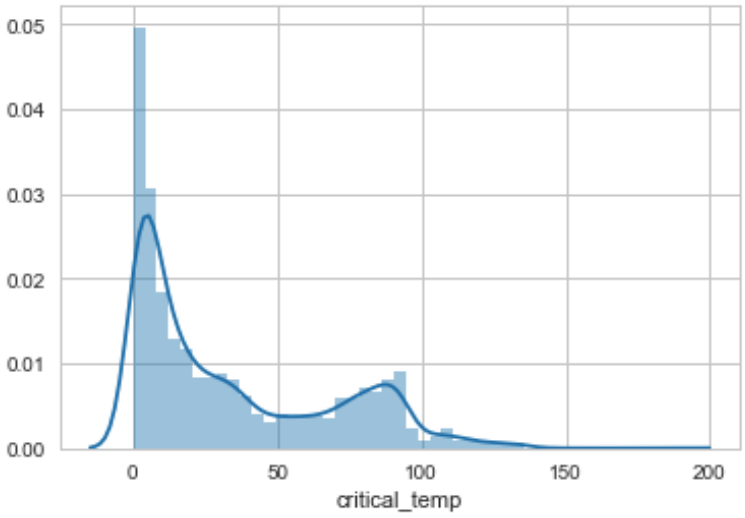

Figure 2: Distribution of Critical Temperatures

As *Figure 2* shown above highlights, the distribution of critical temperature is heavily skewed right. Thus, problems would surely arise if simple cuts were made to group the temperatures. To mitigate this issue, we take a simple cube-root transformation of the data to transform critical temperatures into an approximate Gaussian distribution.

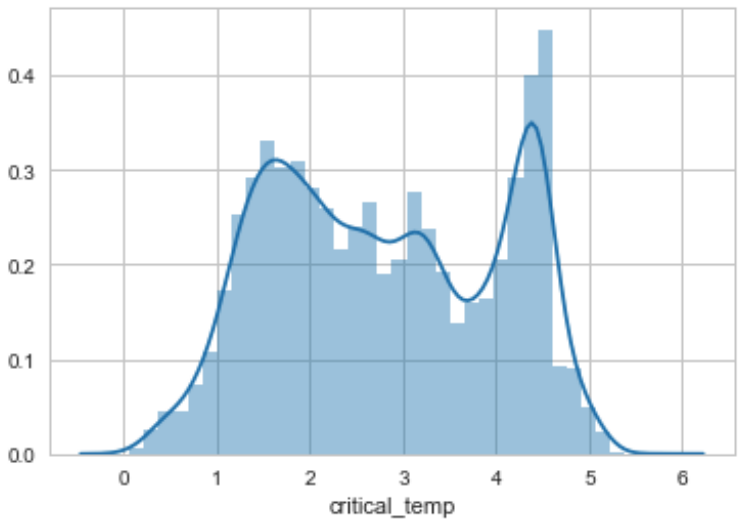

Figure 3: Distribution of Transformed Critical Temperatures

After transforming the critical temperatures, we split the superconductors into three equal groups- low, medium, and high. For lower temperature superconductors, the most prevalent elements were Zirconium (Zr), Niobium (Nb), and Silicon (Si). For both medium and high temperature materials, Oxygen (O) was a clear staple and Copper (Cu) was prevalent as well. To further confirm these results, we were interested in determining which elements are most correlated with critical temperature. In doing so, we determined that approximately 10% of the elements (9 total) were not used in any of the materials.

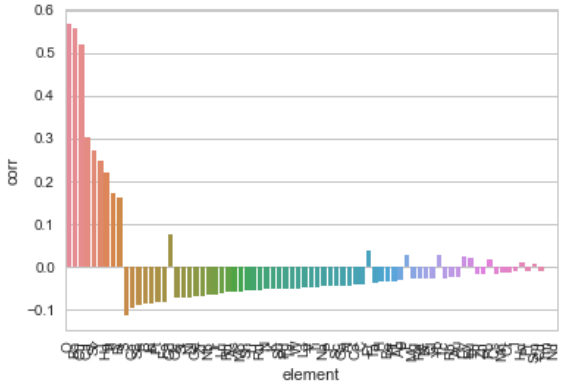

Figure 4: Correlation of Elements

Displayed in *Figure 4* above, most of the elements used did not show significantly high correlation with critical temperature. In fact, only 10 elements had a correlation of $\pm$ 0.1. Notably, Oxygen (O), Barium (Ba), and Copper (Cu) were most correlated with critical temperature, with each element having correlation about 0.5. In fact, all three elements were the most prevalent for the higher temperature materials.

Other Features

Irrespective of the chemical elements, the data set also provides more than 80 different features to describe an individual superconductor. Though there are a large number of features, most are just a different measurement of the same feature. This can be seen through *Figure 5* plotted below.

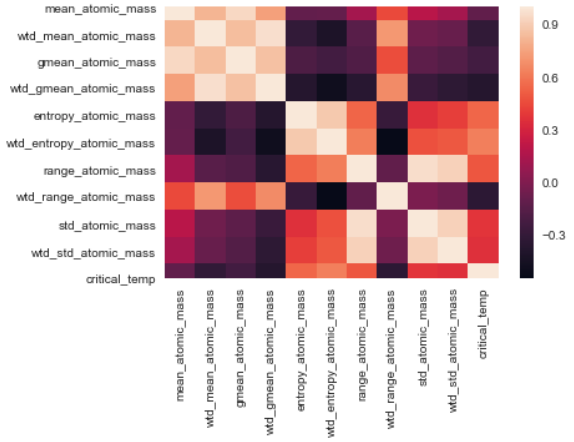

Figure 5: Heat Map of Different Measurements of Atomic Mass

Figure 5 shows a heat map of the different measurements regarding atomic mass. As expected, this subset of features are highly correlated with one another since each feature is describing one specific characteristic of the material. Also, some measurements (as indicated by the last row of the heat map) appear to be more correlated with critical temperature than others. This observation is the same for the other 7 characteristics of the superconductor.

Feature Selection

Combining the elements with all of the features the data set contains 167 different features to describe a single superconductor. The next natural pre-processing step is to determine if we can reduce the feature space to improve run time and accuracy.

Principal Component Analysis

With a large number of features in the data set, we will utilize a dimensionality reduction technique, Principal Component Analysis (PCA), to explain the data with a far less number of features. Essentially, the goal of PCA is to capture the story told by all of the variates by a smaller number of principal components.

The exact values corresponding to each principal component are not particularly important, but rather the amount of information that is maintained after reducing the feature size. To quantify the amount of information maintained, we observe the porportion of variance each principal component contributes to the overall variance. Since $Var(Y_k) = \lambda_k$, the proportion of total variance the first k principal components is defined as:

$$\frac{\lambda_1 + ... + \lambda_k}{\lambda_1 + ... + \lambda_p}$$

	PC 1	PC 2	. . .	PC 66
Proportion of Variance	0.20844765	0.0580478	. . .	0.00976635
Cumulative Proportion	0.20844765	0.26649545	. . .	0.90280403

Notice that that the cumulative proportion of variance for the first sixty six principal components is approximately 90%. That is, only about 10% of information was lost after reducing the number of features to about one-third of its original amount.

Correlated Features

As mentioned previously, many of the features that described a single characteristic were highly correlated with one another. This will create issues when running regression models,

such as model overfitting and multicolinearity. Of the 8 different characteristics, we first selected the one measurement that was most correlated with critical temperature since this likely will be the measurement that best describes the superconductor. Since there was high correlation between the measurements of a single characteristic, we chose only one feature to avoid potential multicolinearity. With respect to elements, the preliminary step was to first remove the 9 elements that were not used in any of the superconductors. Since most of the elements showed little to no correlation with critical temperature, we used the 10 elements with a correlation above $\pm$ 0.1. Thus, this new, greatly reduced matrix now only had 18 features to be used for prediction, as shown in *Figure 6*.

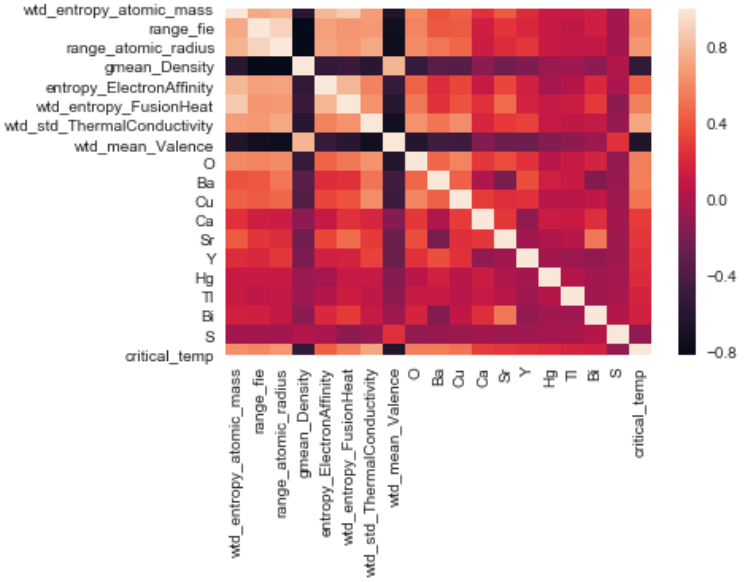

Figure 6: Heat Map of Subset Features

Machine Learning: Regression

As previously stated, there are a total of 167 different features used to explain a single superconductor. After performing a principal component analysis, we determined that we could instead use the first 66 principal components to describe a material without losing too much information. Following that, we determined 18 correlated features that could potentially perform well in describing a superconductor. As a result from both feature selection methods, we now have two different feature matrices that can be compared to the full feature matrix when making predictions.

Model	Full	Subset	PC 66
Linear Regression	1059.885	946.213	1114.266
Ridge Regression	908.449	945.536	907.7289
Lasso Regression	910.901	944.006	913.416

The table above shows the Mean Squared Error rates under each regression model with the different feature matrices utilizing 5-fold cross validation. First, it is worth mentioning that linear regression does considerably worse than both ridge regression and lasso regression, which perform quite similarly. More importantly, for all three models there is no particular feature matrix that unilaterally dominates the other two. For example, even though the subset matrix of 18 features significantly outperforms the full feature matrix in linear regression, the full feature matrix outperforms the subset matrix. In conclusion, this shows that not all features are necessarily needed to make accurate predictions of a superconductor's critical temperature.