Predicting the Critical Temperatures

with Superconductivity Dataset

Kazoua Vang

06/12/20

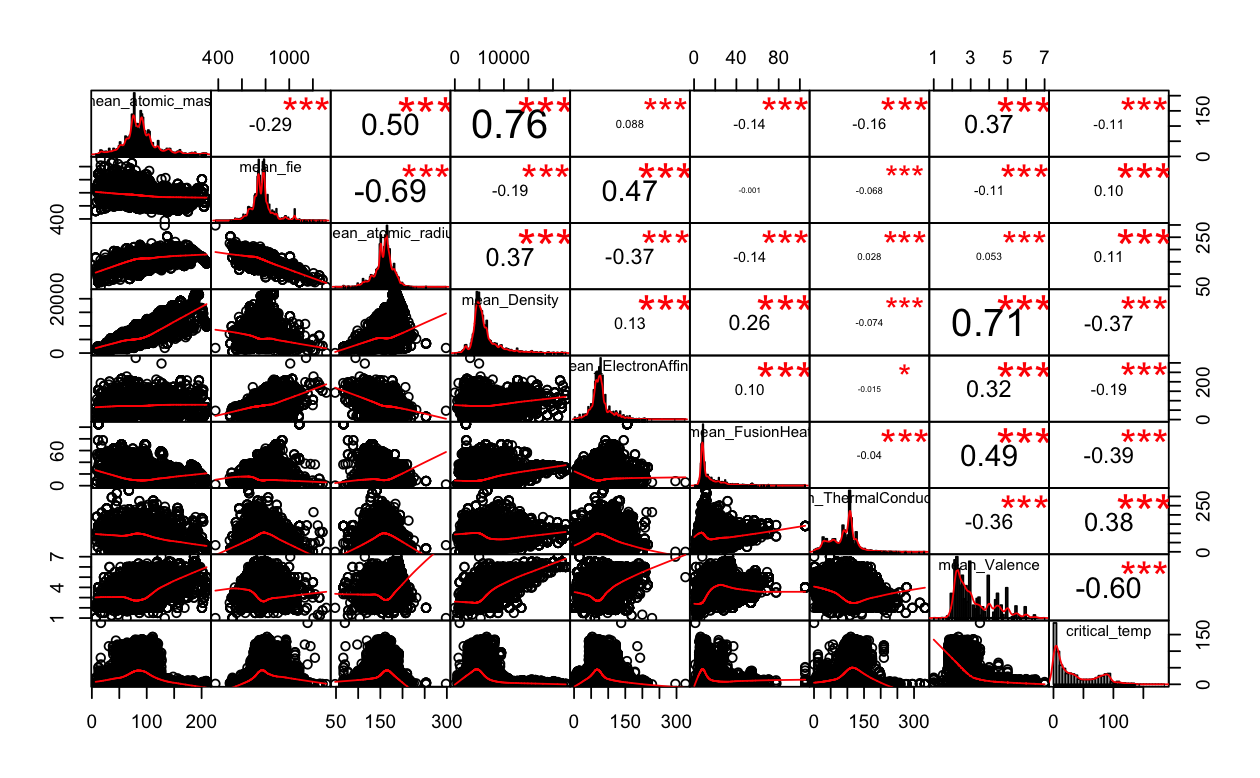
**Introduction**

The Superconductivity Data set which was obtained from the UC Irvine Machine Learning Repository website, contains 81 features extracted from 21,263 superconductors along with the critical temperature. There are eight main variables which are atomic mass, entropy fie, atomic radius, density, electron affinity, fusion, heat, thermal conductivity, and valence. Each of these variables has a total of 10 variations of which includes the mean, weighted mean, geometric mean, weighted geometric mean, entropy, weighted entropy, range, weighted range, standard deviation, and weighted standard deviation for a total of eighty variables. The last feature is the number of elements present in the superconductor. A superconductor conducts electricity without resistance and is only able to do so when the temperature is at or below its superconducting critical temperature. The goal of this analysis is to predict the critical temperature based on the features extracted. For this dataset, I have decided to focus on the variable mean and geometric mean for each of the main features. I want to know of the two, which feature is the better predictor of critical temperature. I will analyze this using the following methods – multiple linear regression, linear discriminant analysis, principal component analysis, and extreme gradient boosting.

**Summary of Data**

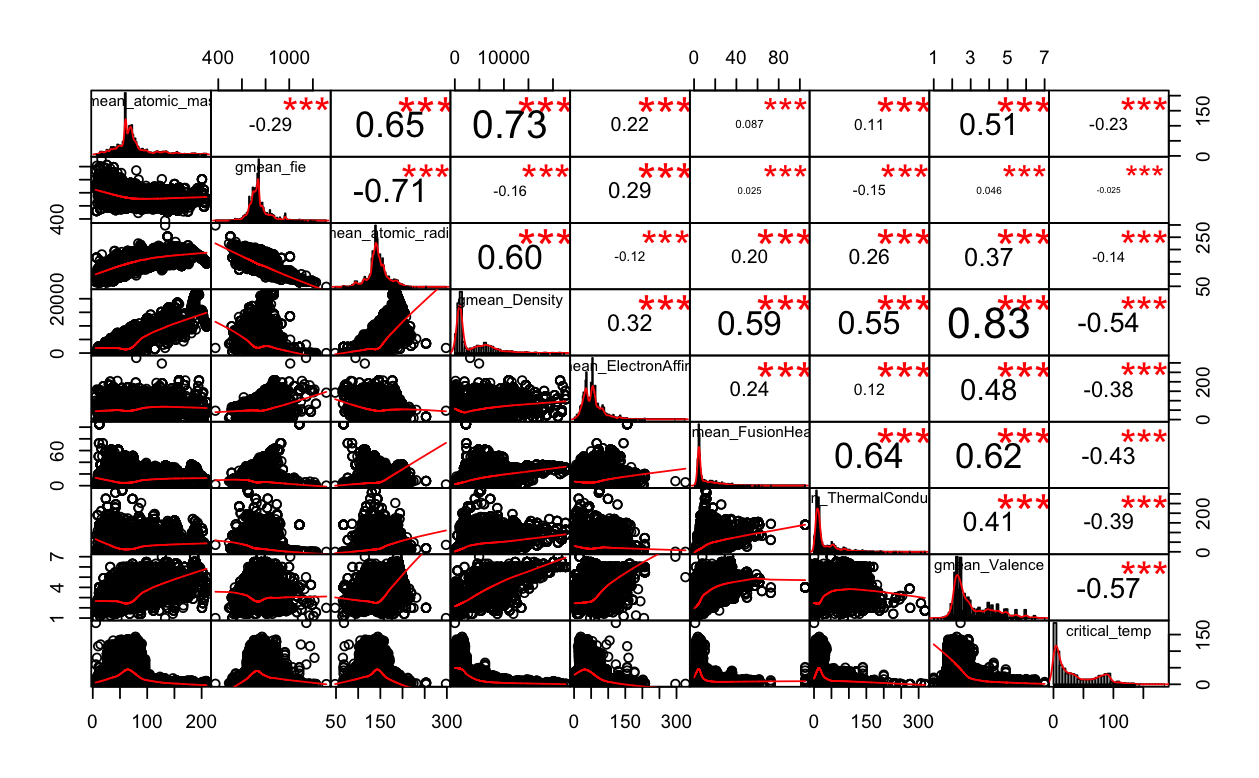
To better visualize the data and see how each feature correlates with one another as well as with the response variable critical temperature, a matrix scatter plot was created for each of the main variables – mean and geometric mean, which is shown in **figure 1** and **figure 2**, respectively. The scatterplots also show the correlation coefficient between each of the nine variables and whether or not it is significant.

**Matrix Scatter Plot of Variable Mean**



**Figure 1** Matrix Scatter plot of the variable mean along with their correlation coefficients as well as significance score

**Matrix Scatter Plot of Variable Geometric Mean**



**Figure 2** Matrix Scatter plot of the variable geometric mean along with their correlation coefficients as well as significance score

The scatterplots tell us that the variable geometric mean has more features correlated to the critical temperature compared to the “mean” variable. The higher the absolute value of the correlation coefficient, the more the two features are correlated with each other. If we consider the cutoff to be 0.5 for there to be a good correlation, then there are two geometric mean features that correlates well with critical temperature. These two variables are geometric mean density with a correlation coefficient of -0.54 and geometric mean valence with a correlation coefficient of -0.57. This is in comparison to the mean variable which only has one feature that is well correlated with critical temperature. This variable is mean valence with a correlation coefficient of -0.60.

Furthermore, the matrix scatterplot can be broken down visually into a correlation plot where we only consider the correlation and whether this correlation of the variables is significant or not. A correlation plot is plotted for each of the variable – mean and geometric mean, which is shown in **figure 3** and **figure 4**, respectively. A dark blue circle signifies that the relation is really strong with a coefficient of one. Likewise, a dark red circle signifies a strong negative correlation with a coefficient of negative one. The lighter the color and the smaller the circle, the less correlated the two variables are. Also, if the correlation is not significant then, there will be a blank.

Comparing the two correlation plots, we see that there seems to be more multicollinearity amongst features in the geometric mean variable compared to the features in the mean variable. We can also easily see that there are more features in the geometric mean variable correlated with critical temperature compared to the mean variable. We see that these correlations are mostly negative correlations.

**Correlation Plot of Mean Variable** **Correlation Plot of Geometric Mean Variable**

A close up of a map

Description automatically generatedA close up of a map

Description automatically generated

**Figure 3** Corplot of Mean Variable and **Figure 4** Corplot of geometric mean

significance of correlation variable and significance of correlation

Another visualization tool to compare the two variables is via heat maps. Shown in **figure 5** and **figure 6** are heat maps for the mean and geometric variable, respectively. Dark red squares represent a strong positive correlation, with dark blue representing a strong negative correlation. The lighter the squares, the less the features are correlated. In comparison, we see more negative correlations between the features in variable mean compared the features in geometric mean where there seems to be more positive correlation which also signifies multicollinearity between the features. Here, we clearly see that critical temperature correlates negatively with most of these features in both mean and geometric mean variable.

A close up of a logo

Description automatically generatedA screenshot of a cell phone

Description automatically generated

**Figure 5** Heatmap for Mean Variable **Figure 6** Heatmap for Geometric Mean

Variable

**Analysis**

To determine whether the variable “mean” or “geometric mean” is the better predictor for critical temperature, a multiple linear regression model will be performed on both sets of variable data – mean and geometric mean. This will be followed by classification methods such as principal component analysis, linear discriminant analysis, and Extreme Gradient Boost (XGBoost). Looking at the matrix scatter plots in **figure 1** and **figure 2**, I predict that the regression method will be less accurate compared to the classification methods due to high multicollinearity between the features. I also predict that Xtreme Gradient Boosting will be the most accurate method for classification.

**Multiple Linear Regression**

When performing a linear model using multiple linear regression, the beta coefficients for each feature is shown below in **Table 1** for variable ‘mean’ and in **Table 2** for variable ‘geometric mean’. According to the correlation coefficients observed from the matrix scatterplots in figure 1 and 2, we should expect the beta coefficients to be negative or near zero for both the mean variables and geometric mean variables, which we do as shown in table 3 and table 4. Furthermore, the calculated multiple R-squared for the mean variables is 0.4966 with a p-value of 2.2x10-16 compared to the geometric mean variables which is 0.4231 with the same p-value of 2.2x10-16. With the higher R-squared value, the mean variable does a better job at predicting the critical temperature than the geometric variable. However, both of these R-squared values are very low and suggests that the multiple linear regression may not be the best model for predicting the critical temperatures.

**Table 1** Beta Coefficients for **Table 2** Beta Coefficients for

Mean Variables Geometric Mean Variables

|  |  |
| --- | --- |
|  | Estimate |
| (Intercept) | -1.877e+02 |
| mean\_atomic\_mass | 4.040e-01 |
| mean\_fie | 1.715e-01 |
| mean\_atomic\_radius | 5.989e-01 |
| mean\_Density | -6.615e-03 |
| mean\_ElectronAffinity | -2.299e-01 |
| mean\_FusionHeat | -1.854e-01 |
| mean\_ThermalConductivity | 3.325e-01 |
| mean\_Valence | -2.794e+00 |

|  |  |
| --- | --- |
|  | Estimate |
| (Intercept) | -5.830e+01 |
| gmean\_atomic\_mass | 3.295e-01 |
| gmean\_fie | 8.474e-02 |
| gmean\_atomic\_radius | 4.132e-01 |
| gmean\_Density | -6.313e-03 |
| gmean\_ElectronAffinity | -1.850e-01 |
| gmean\_FusionHeat | 2.715e-01 |
| gmean\_ThermalConductivity | -1.653e-02 |
| gmean\_Valence | -7.671e+00 |

Furthermore, if we look at the variance inflation factor (VIF) scores of each mean and geometric mean variable which is show in **table 3** and **table 4,** respectively, we see that all of them are greater than one suggesting some correlation. For both variables, it seems that the features mean density and geometric mean density seem to have high correlation with the other features having a VIF score of 7.69 and 10.95, respectively. Overall, the mean variables have lower VIF scores relating to lower multicollinearity when compared to geometric mean variables.

**Table 3** VIF for Mean Variables **Table 4** VIF for Geometric Mean Variables

|  |  |
| --- | --- |
|  | VIF |
| mean\_atomic\_mass | 4.5656 |
| mean\_fie | 2.8822 |
| mean\_atomic\_radius | 2.7692 |
| mean\_Density | 7.6886 |
| mean\_ElectronAffinity | 1.9815 |
| mean\_FusionHeat | 1.6956 |
| mean\_ThermalConductivity | 1.7473 |
| mean\_Valence | 5.2110 |

|  |  |
| --- | --- |
|  | VIF |
| gmean\_atomic\_mass | 4.4862 |
| gmean\_fie | 2.9137 |
| gmean\_atomic\_radius | 4.9128 |
| gmean\_Density | 10.9478 |
| gmean\_ElectronAffinity | 1.6013 |
| gmean\_FusionHeat | 2.7518 |
| gmean\_ThermalConductivity | 2.4449 |
| gmean\_Valence | 4.7210 |

When checking to see how well this multiple linear regression model predicts the critical temperature, the accuracy was not good at all. For the mean variable the accuracy was -22.1859 whereas for the geometric variable, it was worse with an accuracy of -36.4434. The inaccuracy of this model may be due to high multicollinearity which violates the assumptions of a multiple linear regression model. To improve the model, features that are well correlated should be removed or combined together in order for better accuracy.

**Principal Component Analysis (PCA)**

Another method to predict critical temperature is by classifying where the data is split into three groups. Critical temperatures were assigned a “0” for low temperature which is any temperature less than 20, a “1” for medium temperature which is any temperature between 20 and 100, and a “2” for high temperature for any temperatures above 100.

The goal of PCA is to simplify the dataset down by turning the eight variables of the mean variable data and the geometric mean variable data into a smaller number of principal components. Principal components give the direction where there is most variance and using PCA helps to visualize According to the scree plot shown in **figure 7** and **figure 8** for mean and geometric mean variables, respectively, the first two principal components explain most of the variability in the dataset. **Table 5** list the proportion of variance in the components of the mean variables where the first component explains 34% of the variance and the second component explains 26% of the proportion. Similarly, **table 6** shows that for the geometric variables data, the first component explains about 46% of the variance and the second component explains 24% of the variance. In comparison, the geometric variable data is able to detect more variability in the data. **Table 7** and **table 8** tells us that the features that loaded heavily on the first principal components is mean density with a loading of 0.534 for the mean variable data and for the geometric mean variable data, it is also geometric mean density with a loading of 0.501. **Figure 9** and **figure 10** which are the principal components plots of the two data, respectively, also confirms this.

**Figure 7** Scree plot for Mean Variables **Figure 8** Scree Plot for Geometric Variables

A close up of a map

Description automatically generatedA close up of a map

Description automatically generated

|  |  |  |
| --- | --- | --- |
| **Table 5** Importance of Components in Mean Variable | | |
|  | Comp.1 | Comp.2 |
| Standard deviation | 1.6508981 | 1.4461109 |
| Proportion of Variance | 0.3406831 | 0.2614046 |
| Cumulative Proportion | 0.3406831. | 0.6020877 |

|  |  |  |
| --- | --- | --- |
| **Table 6** Importance of components in Geometric Mean Variables | | |
|  | Comp.1 | Comp.2 |
| Standard deviation | 1.9080678 | 1.3727872 |
| Proportion of Variance | 0.4550904 | 0.2355681 |
| Cumulative Proportion | 0.4550904 | 0.6906584 |

|  |  |  |
| --- | --- | --- |
| Table 7 Loadings for Mean Variables | | |
|  | Comp.1 | Comp.2 |
| mean\_atomic\_mass | 0.493 | 0.0292 |
| mean\_fie | -0.323 | -0.443 |
| mean\_atomic\_radius | 0.388 | 0.437 |
| mean\_Density | 0.534 | -0.174 |
| mean\_ElectronAffinity | - 0.016 | -0.510 |
| mean\_FusionHeat | 0.131 | -0.340 |
| mean\_ThermalConductivity | - 0.138 | 0.198 |
| mean\_Valence | -0.422 | -0.404 |

|  |
| --- |
| **Table 8** Loadings on Geometric Mean Variables |
| Comp.1 Comp.2 |
| gmean\_atomic\_mass 0.369 0.228 |
| gmean\_fie -0.153 -0.600 |
| gmean\_atomic\_radius 0.359 0.483 |
| gmean\_Density 0.501 -0.351 |
| gmean\_ElectronAffinity 0.185 -0.444 |
| gmean\_FusionHeat 0.350 -0.290 |
| gmean\_ThermalConductivity 0.327 -0.131 |
| gmean\_Valence 0.444 -0.236 |

PC1 vs PC2 in Mean Variables PC1 vs PC2 in Geometric Mean Variables

A close up of a map

Description automatically generated A close up of a map

Description automatically generated

**Figure 9** PCA plot of loadings pc1 vs pc2 **Figure 10** PCA plot of loadings pc1 vs pc2

in mean variables in geometric mean variables

According to these PCA plots, we see that there may not be a need for three clusters as we do not really see much of the high critical temperatures being in a distinct cluster. Instead, we see mainly low critical temperatures and then a small cluster of medium critical temperatures that overlaps with some low critical temperatures. Also, it seems that medium critical temperatures tend to have higher mean thermal conductivity and higher geometric mean fie as well as geometric electron affinity compared to most low critical temperatures (figure 9 and 10).

**Linear Discriminant Analysis (LDA)**

To best categorize the critical temperatures based on the eight features of each mean and geometric mean variables, the classification method of linear discriminant analysis will be used. First, fit the LDA model and then use the model to predict and perform classification. After fitting the LDA model, we obtain the following results shown in **table 9** and **table 10**. The proportion of trace for first and second linear discriminant in the mean variable data is 0.9699 and 0.0301, respectively. For the geometric mean variable data, the proportion of trace for the first and second linear discriminant is 0.9811 and 0.0189, respectively. This tells us that for both data sets, the first linear discriminant has a far more significant effect in categorizing the critical temperatures. Table 9 and 10 shows us that for both data, the valence feature for both mean and geometric mean has the highest absolute coefficient of -0.39 and -0.54, respectively, signifying that it is the main contributor in specifying the first linear discriminant and determining the boundary between the different temperature types.

When performing the prediction using LDA, looking at the confusion matrices in **figure 11** and **figure 12**, we see that the prediction accuracy for the mean variable data is 81.30% and for the geometric mean variable data, it is 80.63%. We see that the prediction is a little bit better for the mean variables compared to the geometric mean variables.

|  |
| --- |
| **Table 9** Coefficients of Linear Discriminants in Mean Variable Data |
| LD1 LD2 |
| mean\_atomic\_mass 0.016194391 0.0438767657 |
| mean\_fie 0.008973945 0.0040767268 |
| mean\_atomic\_radius 0.030922806 -0.0362097653 |
| mean\_Density -0.000307154 -0.0002199038 |
| mean\_ElectronAffinity -0.011017674 -0.0312528945 |
| mean\_FusionHeat -0.018145733 0.0077788694 |
| mean\_ThermalConductivity 0.015709498 0.0211966282 |
| mean\_Valence -0.391884878 0.6861232227 |

**Confusion Matrix for Mean**

Predicted

|  |  |  |  |
| --- | --- | --- | --- |
| Actual | 0 | 1 | 2 |
| 0 | 8266 | 2250 | 71 |
| 1 | 888 | 9020 | 0 |
| 2 | 4 | 764 | 0 |

**Figure 11** Confusion Matrix for

Mean Variable Data with an

Accuracy of 0.8129615.

|  |
| --- |
| **Table 10** Coefficients of Linear Discriminants in Geometric Variables Data |
| LD1 LD2 |
| gmean\_atomic\_mass 0.0186003440 2.675233e-02 |
| gmean\_fie 0.0042682810 4.918867e-03 |
| gmean\_atomic\_radius 0.0199709403 -2.616960e-02 |
| gmean\_Density -0.0003602409 -3.449732e-05 |
| gmean\_ElectronAffinity -0.0099133382 -3.797208e-02 |
| gmean\_FusionHeat 0.0108601494 -7.729738e-03 |
| gmean\_ThermalConductivity -0.0009959898 1.318347e-02 |
| gmean\_Valence -0.5367561092 3.530764e-01 |

**Confusion Matrix for**

**Geometric Mean**

Predicted

|  |  |  |  |
| --- | --- | --- | --- |
| Actual | 0 | 1 | 2 |
| 0 | 7869 | 2695 | 23 |
| 1 | 632 | 9276 | 0 |
| 2 | 3 | 765 | 0 |

**Figure 12** Confusion Matrix for

Mean Variable Data with an

accuracy of 0.806330

**Conclusion**

Overall, when comparing the all the eight features in the mean variable data with the geometric mean variable data to see which variable is the better predictor for critical temperature, according to the multiple linear regression method and the linear discriminant analysis, the mean variables performed better than the geometric mean variables. As to which method performed better, it is clear that the linear discriminant analysis outperformed the multiple linear regression, predicting the correct classes about 80% of the time. For future analysis, to increase the accuracy of the multiple linear regression, the data must be cleaned of multicollinearity as too much exists in the data violating assumptions of the linear regression model. Also, according to the principal component analysis, there seemed to be two clusters instead of three, so the linear discriminant analysis may perform better if the critical temperatures were split into two categories instead of three.