Case Study 1 – Critical Temperature

Kevin Boyd
Laura Ahumada
Shikha Pandey

September 5, 2023

Abstract

The first case study for Quantifying the World is to build a linear regression model using L1 and L2 regularization to accurately predict the critical temperature of superconductors and determine which explanatory variables contribute most to the model. Once an initial EDA was performed and our data was normalized, we created linear regression models to test both Lasso (L1) and Ridge (L2) regularizations using negative mean squared error as our score. We found that Lasso did make our data sparser which helped us determine which variables were most important while Ridge was good for not overfitting and giving us the most accurate prediction. We observed that combining both regularization techniques yielded the most accurate and reproducible results for our linear regression model.

Introduction

A group of scientists are looking at multiple superconductors. This area of study is of importance to the scientific community because superconductors have unique characteristics that are very valuable in a wide range of methodologies in modern day science and technology development. Specifically, superconductors are materials that give little or no resistance to electrical current.

The intent of this study is to use the data to produce a model to predict new superconductors based on the properties and the data that they have found. Data points include properties, material composition, and the temperature at which they superconduct. We chose to build a linear regression model using L1 and L2 regularization with the goal to predict the critical temperature of superconductors as closely as possible and to find the variables that carry the most importance in the models.

When considering linear regression models there are a few terms that are important to understand. Specifically in the equation below (left), J is our loss and P is our prediction. Next, we can replace P with our actual formula, where Y is our target and *x* is our data, as shown by the middle equation. Finally, we can introduce our penalty term so we can see it all together (right).

$$J = \frac{1}{n}\sum_{i=1}^{n}(Y_i - P_i)^2 \qquad J = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \sum_{j=0}^{k} m_j x_{ij})^2 \qquad J = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \sum_{j=0}^{k} m_j x_{ij})^2 + \lambda \sum_{j=0}^{k} f(m_j)$$

This is the basic formula for linear regression including our prediction and loss function. The strength of our penalty is denoted by λ. When λ=0 our function will return our original regression values without any penalty. The value of λ is different for each problem and must be tuned experimentally to determine the best value. Also, our penalty term is different when considering L1 (Lasso) or L2 (Ridge) regularization. As seen below, the Lasso term is absolute value of |m_j| while the Ridge term is m_j^2.

$$\lambda \sum_{j=0}^{k} |m_j| \qquad \lambda \sum_{j=0}^{k} {m_j}^2$$

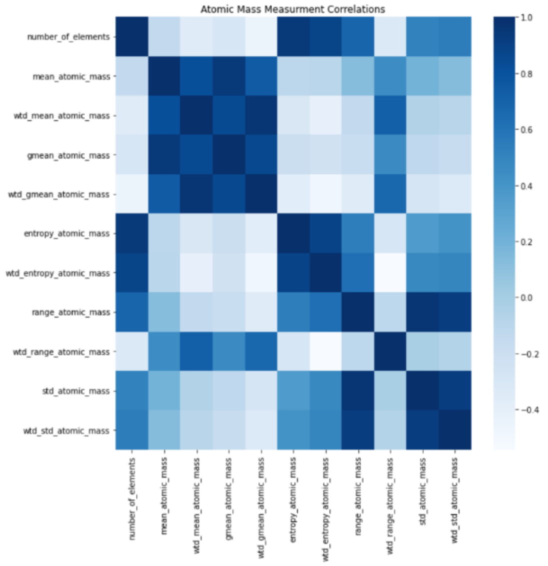
This will lead to Lasso making the data sparser by pushing many of the values to 0 while ridge exponential term will make the important variables contribute even more. Each has it best use case, and when considering our data and objectives, we decided to use both methods.

Methods

Data was obtained from two separate files: ‘train.csv’ and ‘unique_m.csv’. The ‘train.csv’ has 21,263 rows, 82 columns and it contains data for different types of properties of a superconductor and target variable - Critical Temperature. The ‘unique_m.csv’ has 21,263 rows, 88 columns and it contains one hot encoded data for elements present in a superconductor, target variable - Critical Temperature and a ‘material’ column which is a combined notation of elements present in a superconductor.

To predict the Critical Temperature, we first needed to combine these two files. The target variable and the ‘material’ column were removed from ‘unique_m.csv’ in this process. Once combined, we had a dataset of 21,263 rows and 168 columns to further proceed with analysis. We will discuss the target variable later in the EDA.

Some of the elements like 'He', 'Ne', 'Ar', 'Kr', 'Xe', 'Pm', 'Po', 'At', and 'Rn' were not found in any of the material data, so we decided to remove them. Many of the variables relating to Atomic_Mass were highly correlated. This is consistent with Density, Electron_Afffinity, Atomic_Radius, FusionHeat, Entropy_Valance measurements. Specifically shown below are correlation plots for atomic mass and atomic radius. The mean, wtd_mean, gmean, and wtd_gmean measurements are all very corelated as shown by the plots below. We also observed that entropy, wtd_entropy, and range were often highly correlated.



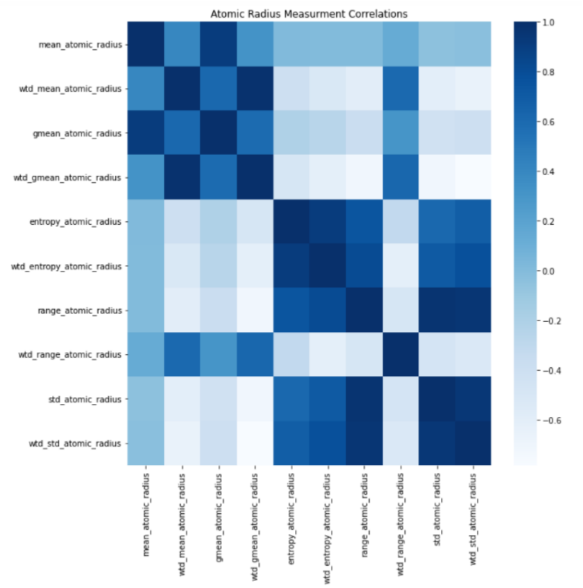


Fig 1 Fig 2

Figure 1: Correlation plot showing atomic mass measurements

Figure 2: Second correlation plot showing atomic radius measurements

After observing the data, we decided to standardize our explanatory variables before running our first model. This is important because many of the variables are on different scales and therefore direct comparisons are very difficult to make with variables with vastly different ranges.

Our target variable is Critical Temperature. We can see that this variable is right skewed, meaning that a majority of the points are low with a few higher points spreading the distribution slightly. We can also see that there is a single point around 175 which is very high but after further investigation does not appear to be an inaccurate measurement.

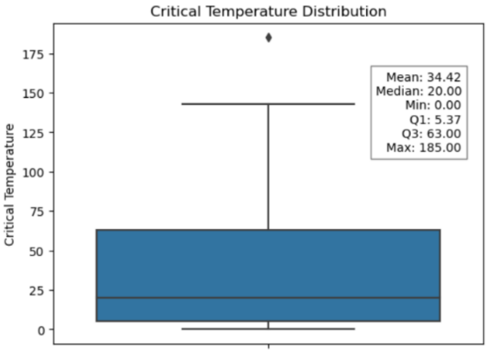


Fig 3

Figure 3: Boxplot showing the distribution of our target variable “critical temperature”.

Results

We opted to use a cross-validation grid-search method to perform an exhaustive search over a wide range of parameters for linear regression regularized with Lasso and Ridge using 10-fold cross validation. Parameter alpha was searched for using 20 values between the range of 10^{-6} and 10^{6}. We decided to use mean squared error scoring to evaluate the performance of the cross-validated model on the test set.

When using Lasso regularization, the best score we observed for mean squared error was 339.322 at an alpha of 0.483. Our Target vs Predictions plot shows a huge difference between the two for a few data points. One with critical temperature around 180 K is predicted with less than 0 K. There are others with critical temperatures around 0 K, 60 K and 80 K are predicted with much higher critical temperatures of around 150 K and 250 K respectively. Many of the coefficients were reduced to zero by this lasso regularization which is what we were expecting. From the list of selected features by lasso, some of the most important features Ba, wtd_mean_ThermalConductivity, wtd_std_ThermalConductivity, Bi, wtd_entropy_atomic_mass, Ca, and range_atomic_mass. These appear to impact the critical temperature positively with change while other most importance features like wtd_gmean_ThermalConductivity, wtd_std_Valence and wtd_gmean_ElectronAffinity impact the critical temperature negatively with change.

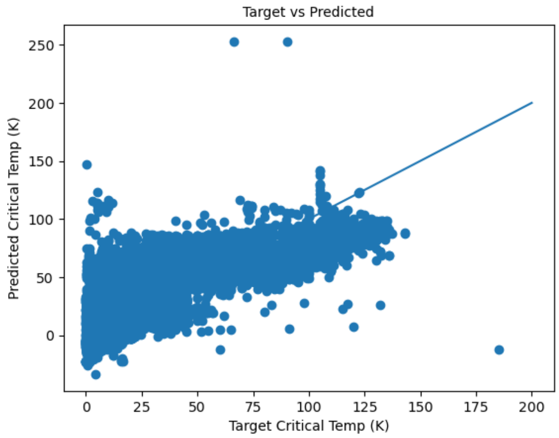


Fig 4

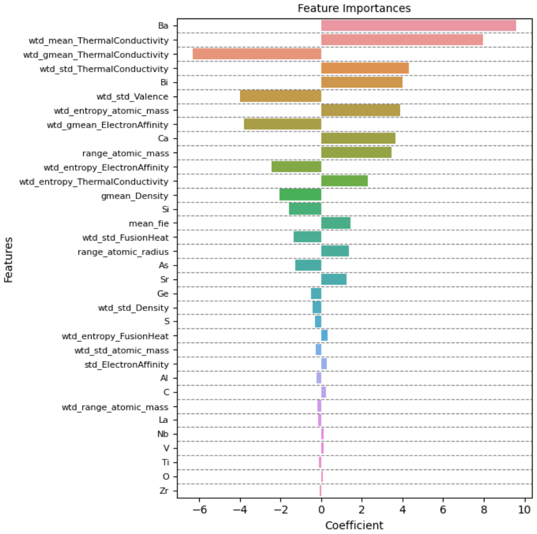


Fig 5

Figure 4: Predicted vs actual values of the target variable 'critical temperature' with regression line.
Figure 5: Values of coefficients from the Lasso regularization.

Next, we tried our Ridge regularization. With Ridge we found the best score when measuring mean squared error was 335.719 at an alpha of 2976.351. Like Lasso, we see the same prediction pattern for data points with target temperatures around 180 K, 0 K, 60 K and 80 K. There is however one additional data point with a huge difference in target temperature around 25 K and predicted temperature less than -300 K. As expected, Ridge didn't reduce any coefficients to zero. Some of the top important features provided by ridge are same as lasso. Ba, wtd_mean_ThermalConductivity, wtd_std_ThermalConductivity, Bi, range_atomic_mass, range_atomic_mass, and Ca impact the critical temperature positively with change while other most importance features like wtd_std_Valence, wtd_gmean_ThermalConductivity, wtd_entropy_ElectronAffinity and wtd_gmean_ElectronAffinity impact the critical temperature negatively with change.

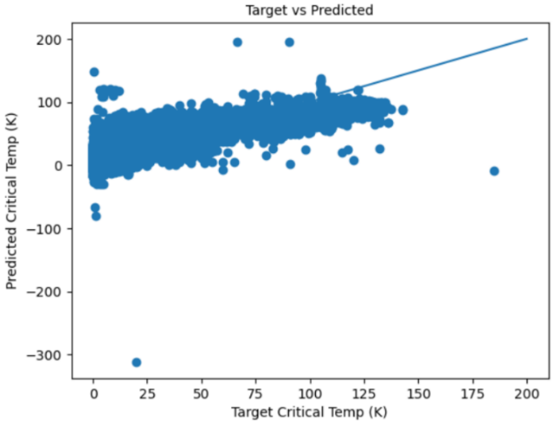


Fig 6

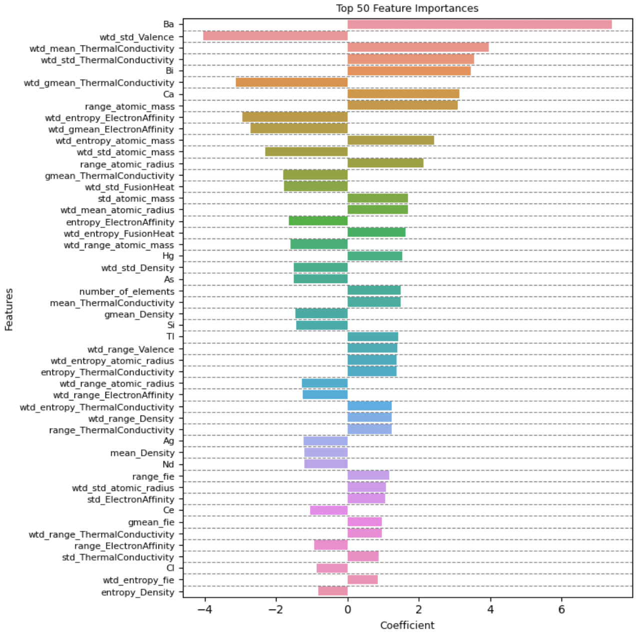


Fig 7

Figure 6: Predicted vs actual values of the target variable 'critical temperature' the regression line.
Figure 7: Values of top 50 coefficients from the Ridge regularization.

We also wanted to compare the different scores at different values of alpha for both models. We did this by plotting both regularization methods together to show all the values between 10^{-6} to 10^{6}. This plot shows a drastic increase in scores for lasso between 10^{-1} and 10^{1}.

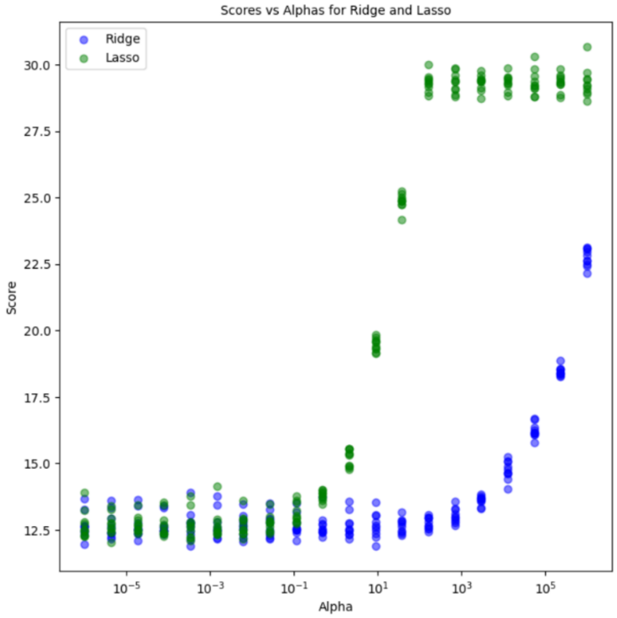


Fig 8

Figure 8: Plotting different values of alpha and their corresponding scores. Negative mean absolute error was used here to score both models.

Conclusion

Ba (Barium) was the most influential variable in both the L1 and L2 regularized linear regression models. Bi (Bismuth) was the second most influential element with Ca (Calcium) as the third most important element. When looking further into this we found that these elements are key components of specific compounds known for their high temperature superconducting abilities. Next, we noticed that the ridge regularization found that thermal conductivity measures; mean, gmean, and standard deviation, were indicative of predicting superconductivity. This also makes sense because energy is converted into heat and substances that can reduce this heat will be better superconductors. Both models' different regularizations performed well while L1 (lasso) made the data more sparce to determine the most influential coefficients, and the L2 (ridge) is more generalizable and better to predict.

Code

Link to the Jupyter notebook with code assisting in this case study is below:

https://github.com/lauraah10/QW-Projects/blob/main/Case%20Study%201%20-%20Linear%20Regression.ipynb