Case Study 1: Superconductor Dataset - Prediction of Critical Temperature and Interpretation of Superconducting Material Properties

Kristin Henderson

May 12, 2025

1 Introduction

This study aims to identify important properties of superconductors, materials that give little or no resistance to electrical current, and to predict the critical temperature at which their superconducting properties operate.

The Superconductivity Dataset is a University of Califonia at Irvine Machine Learning Repository dataset. It contains 21,263 entries consisting of 82 features. The dataset contains two files, one containing material properties, which are numeric features, and the other containing the chemical formula encoded by a vectorized format, with each element represented as a feature and the value being the number of atoms. The target variable is the *superconducting critical temperature* in Kelvin. There is no missing data.

Table 1 contains the material properties, i.e. the category of features, and their descriptions.

Table 1: Category of Features in the Superconductivity Dataset

Variable	**Description**
number_of_elements	unique elements in the material
atomic_mass	atomic mass of the compositional elements
fie	first ionization energy
atomic_radius	atom size
Density	mass per unit volume
ElectronAffinity	energy change when electron is added to neutral atom
FusionHeat	heat required to change solid to liquid
ThermalConductivity	ability to conduct heat
Valence	combining capacity of elements
material	chemical formula

Table 2 lists the statistical metrics for each property of the materials. Each of the properties (except `number_of_elements` and `material`) is represented by all ten metrics listed in Table 2, resulting in multiple features per property.

Table 2: Statistical Metrics of Each Material's Property

Metric	**Description**
mean	average value
wtd_mean	weighted mean
gmean	geometric mean
wtd_gmean	weighted geometric mean
entropy	distribution of values
wtd_entropy	weighted entropy
range	difference between max and min values
wtd_range	weighted range
std	standard deviation
wtd_std	weighted standard deviation

A linear regression model is chosen for several reasons. Primarily, it is an appropriate choice with the goal of predicting a continuous variable and having the ability to identify and explain the most important features in determining the critical temperature. Secondarily, a linear model is a good choice for computational efficiency.

The target variable in a linear regression model is a linear combination of the input features.

$$y = m_0x_0 + m_1x_1 + \cdots + m_nx_n \text{ where } x_0 = 1 \text{ and } m_0 \text{ is the intercept.}$$

Penalized linear regression models with different types of regularization are used as a tool to prevent overfitting. Penalty terms are added to the base loss function:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \sum_{j=0}^{k} m_j x_{ij}\right)^2$$

Here I compare LASSO, which uses a first-order L1 penalty term $\lambda\sum_{j=0}^{k}|m_j|$ to Ridge regression with its second-order L2 penalty term $\lambda\sum_{j=0}^{k}m_j^2$. For interpretability of error in the model and to get a sense of its usefulness and practicality, I will also compute the root mean squared error (RMSE) from the chosen loss metric, mean squared error (MSE).

2 Data

Many of the features in this dataset are different statistical metrics of the same general property. For example, `FusionHeat` the heat required to change a solid to a liquid, is represented by 10 features: mean, weighted mean, geometric mean, weighted geometric mean, entropy, weighted entropy, range, weighted range, standard deviation and weighted standard deviation. These likely are necessary to describe the material based on the different atoms in its composition. Generally, the middle fifty percent of all the features are roughly symmetric. Some categories of features have outliers, more high than low, and some are skewed, especially in their outer range. Boxplots of fusion heat (see Figure 1) and valence (see Figure 2) are representative of variables with, respectively, smaller and larger ranges, greater and lesser skewness, and more and fewer outliers.

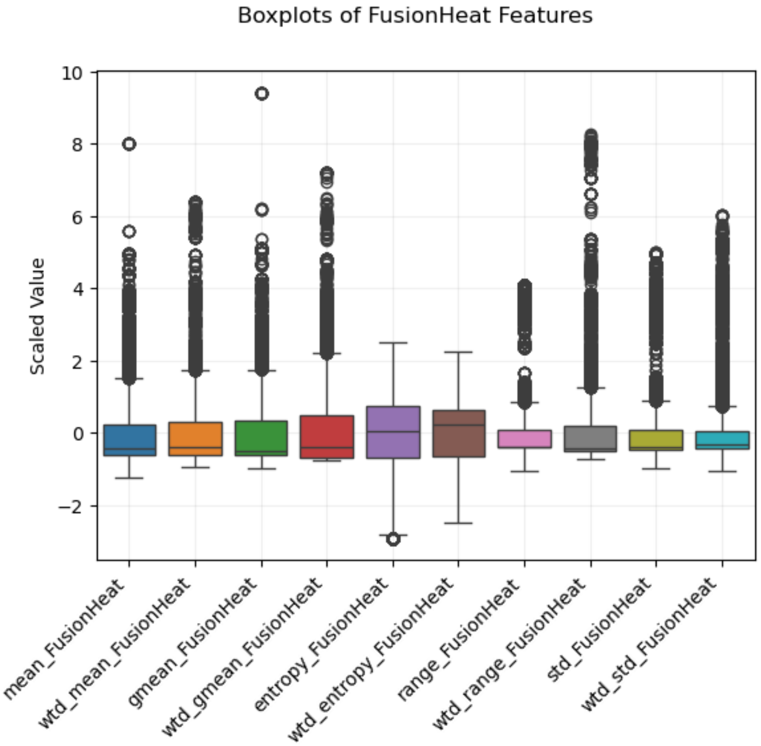

Figure 1: Boxplots of `FusionHeat` features. Notice relatively small ranges for the middle 50% of the data, some right skewness, and high outliers. Values were scaled (z-scored) using `StandardScaler` a function within the scikit-learn package, to compare on the same scale and axis while preserving the distributions.

Despite many of the features being different metrics of the same property, no features were perfectly correlated. Without sufficient domain knowledge to determine that some could be excluded, all features were included in the modeling process.

The distribution of the target variable, critical temperature, is right-skewed with a range from 0 to 185 Kelvin, shown in Figure 3. Over one third of the

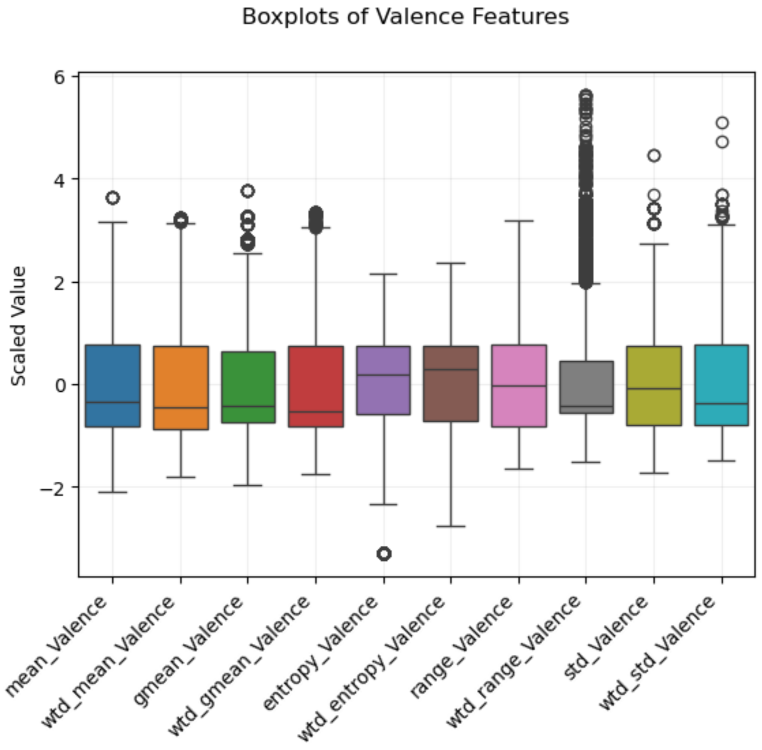

Figure 2: Boxplots of `Valence` features. Notice broader ranges for the middle 50% of the data and few outliers. Values were scaled (z-scored) using `StandardScaler` to compare on the same scale and axis while preserving the distributions.

entries fall within the 0 to 10 Kelvin range. I also plotted a histogram of the log-transformed critical temperature, shown in Figure 4, but the transformation did not fully rectify the skewness. I chose to continue the modeling process without log-transforming the target variable to make interpretation of errors and the coefficients more straightforward.

Though some of the feature have outliers, more commonly high ones, none seem to be unusually abnormal or suggest an error in the data. Given that, and the fairly normal distribution of the data, particularly the reasonable symmetry within the interquartile ranges (IQR), I choose to stick with `StandardScaler`. I choose to scale only the numerical material property features and not to scale the vector encoded chemical formula features.

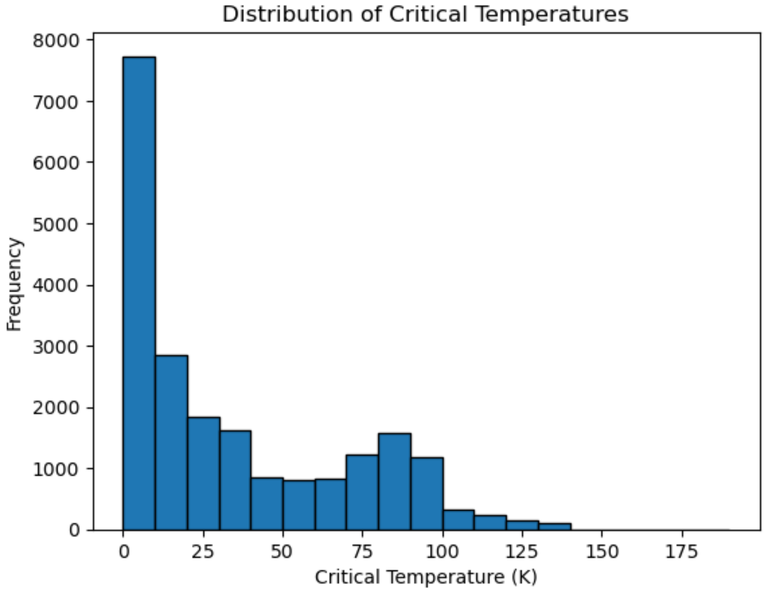

Figure 3: Histogram of the target data, critical temperature. Notice that over 35% of the values fall within the 0-10 K bin.

3 Modeling

Five-fold cross validation with a random shuffling of the data was used for hyperparameter tuning. The out-of-fold predictions were used to calculate the MSE. A manual grid search of the penalty term λ was performed for both LASSO and Ridge models, starting with a broad range of values to identify a region likely to optimize the loss. Using the lowest MSE, the best λ for each model was chosen from a second, narrower search. For LASSO, the λ values searched ranged from 0.5 to 10, focusing on smaller increments under 1, while for Ridge the values ranged from 10^{-5} to 10^{4} increasing by orders of magnitude with finer increments toward the upper end of that range. I also chose to tune an ElasticNet model as a comparison to the pure L1 or L2 regularization of the LASSO or Ridge models respectively. For ElasticNet a grid search was conducted for both the penalty term, λ, with values in the range from 10^{-3} to 10^{3}, and for the L1 mixing term from 0.01 to 0.9, focusing for both hyperparameters particularly on the values between 0.01 and 0.1.

4 Results

Tuning curves for the LASSO and Ridge models are shown in Figure 5 and Figure 6, respectively. The estimated MSE on out-of-fold predictions for LASSO with an optimized λ of 0.34 was 332.03 ± 16.00 (mean ± standard deviation)

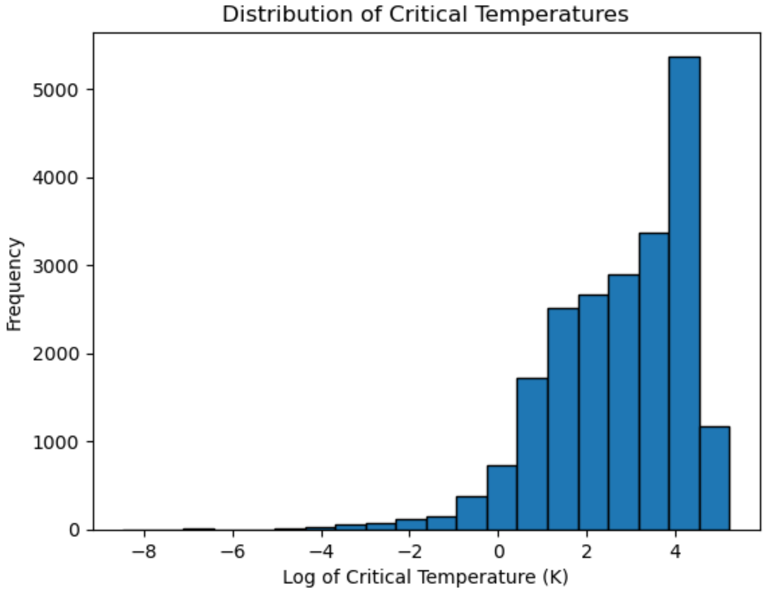

Figure 4: Histogram of the target data, critical temperature after log transformation. Notice that the distribution is not fully normal.

and for Ridge regression with an optimized λ of 2000 was 333.28 ± 25.46.

The results of a narrowed range of the ElasticNet grid search can be seen in Figure 7. The estimated MSE on out-of-fold predictions for ElasticNet with an optimized λ of 0.1 and mixing term of 0.09 was 332.27 ± 30.31. For comparison to all the regularized models, a ordinary least squares regression model yielded an MSE of 687.55 ± 584.20.

A plot of residuals versus the observed critical temperature values for both the LASSO and Ridge models display a slight upward trend and fan shape with a few extreme points, shown in Figure 8 and Figure 9.

The most important properties and chemical components of superconductors identified by the LASSO model, which achieved the lowest MSE and smallest standard deviation, are listed in Table 3 along with their coefficients, indicators of the importance and direction of their relationship with the target variable. Interestingly, the most important elements in the chemical makeup are Barium, Bismuth, and Calcium. Unsurprisingly, several of the metrics of thermal conductivity are important in predicting the conductive temperature of a material. The range of the atomic mass, and metrics of valence and electron affinity are also important features.

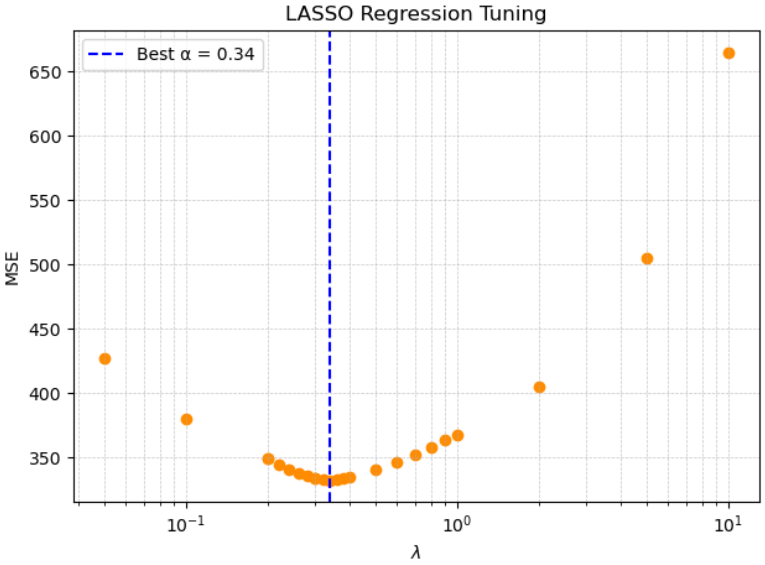

Figure 5: Tuning Curve of λ for the LASSO model. The optimum value of λ is indicated by the dotted blue line.

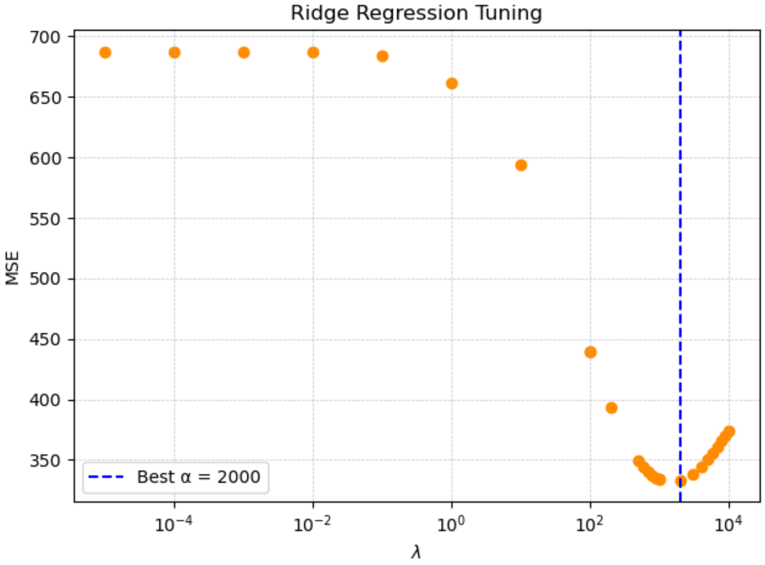

Figure 6: Tuning Curve of λ for the Ridge regression model. Again, the optimum value of λ is indicated by the dotted blue line.

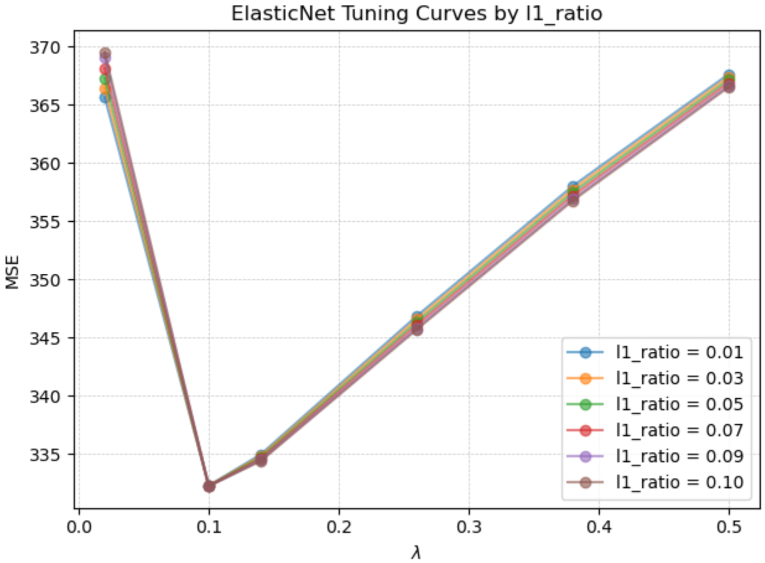

Figure 7: Tuning Curve of λ using different mixing terms for the ElasticNet regression model.

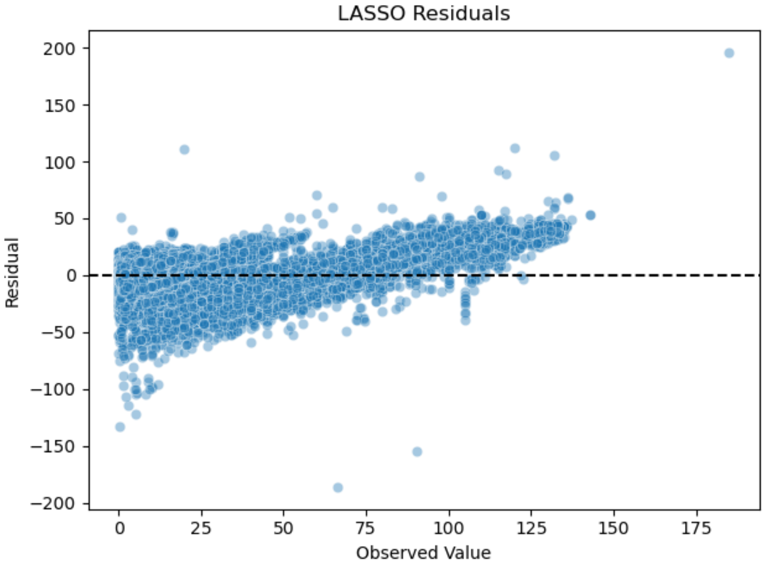

Figure 8: Residuals of the predictions plotted versus the observed values for the LASSO model show a slight upward trend and fan shape and a few extreme values.

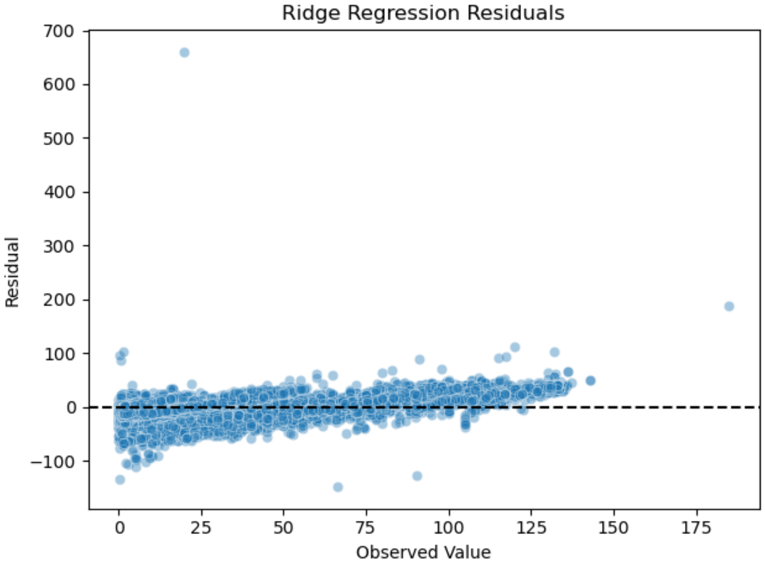

Figure 9: Residuals of the predictions ploted versus the observed values show a slight upward trend and fan shape and a few extreme values.

Table 3 contains the 10 most important features and their coefficients from the LASSO model.

Table 3: Most Important Features from the LASSO Model

Variable	**Coefficient**
Ba	9.66
wtd_mean_ThermalConductivity	9.36
wtd_gmean_ThermalConductivity	-7.82
range_atomic_mass	5.69
Bi	4.70
wtd_std_Valence	-4.14
wtd_std_ThermalConductivity	3.52
wtd_gmean_ElectronAffinity	-3.49
Ca	3.44
wtd_entropy_ElectronAffinity	-3.09

The most important properties and chemical components of superconductors found using the Ridge regression model are listed in Table 4 along with their coefficients. These ten features are the same as the top ten features from the LASSO model. Additionally, the signs of the relationship with the target variable, positive or negative are all also the same between the models. However, the variables rank slightly differently, and the coefficients are generally smaller in the Ridge model.

Table 4 contains the 10 most important features and their coefficients from the Ridge regression model.

Table 4: Most Important Features from the Ridge Regression Model

Variable	**Coefficient**
Ba	8.00
wtd_mean_ThermalConductivity	4.64
wtd_std_Valence	-4.57
wtd_std_ThermalConductivity	4.09
Bi	4.00
wtd_gmean_ThermalConductivity	-3.68
range_atomic_mass	3.63
wtd_entropy_ElectronAffinity	-3.32
Ca	3.23
wtd_gmean_ElectronAffinity	-2.98

5 Conclusion

All three models which used regularization, Lasso, Ridge, and ElasticNet, achieved statistically equivalent results with similar cross-validated MSEs and overlapping standard deviations. There was little difference in computational efficiency with the exception of ElasticNet, which required tuning an additional parameter. Additionally, the optimum L1 ratio for ElasticNet was 0.09, meaning it is primarily using L1 rather than an balanced combination of L1 and L2 penalties. The LASSO model has a very slightly lower MSE and standard deviation than the others. The RMSEs of each regularization model, reflecting error in the original units, were all 18 Kelvin. In contrast, the unregularized linear regression model had a much higher error, an RMSE of 26 K, and larger variability than the other models.

The trend and fan shape in the residual plots indicate potential heteroscedasticity in the errors. Although the log transformation did not eliminate the skewness in the target variable distribution, it may be worth revisiting the modeling process using the log-transformed variable to try to improve the variance in the residuals and better meet the assumptions of the model.

The LASSO and Ridge models were consistent in their variable importance, with Barium and weighted mean of thermal conductivity being the most important features in determining the critical temperature.