CASE STUDY 1: SUPERCONDUCTORS

BY: ANDREEA CAROLINA CRAUS

1 | INTRODUCTION

Superconductors are materials that give little or no resistance to electrical currents without an onset or buildup of heat, creating a magnetic field that generates a constant flow of electricity. The goal of this case study is to produce a model using linear regression with both L1 and L2 regularization to predict the critical temperature that would identify new superconductors based on different properties. The critical temperature is the temperature at which a material acts as a superconductor.

2 | METHODS

DATA PREPROCESSING

The original data is comprised of two separate CSV files: train.csv, with 21,263 rows and 82 columns, and unique_m.csv, with 21,263 rows and 88 columns. The former file has 81 features for 21,263 superconducting materials with the superconducting critical temperature included. The latter file has chemical formulas comprising each of the 21,263 superconducting materials, i.e. the specific ratios of elements, given in a format that is not entirely dissimilar to one-hot encoding. These two files were merged, resulting in a merged dataset consisting of 21, 263 rows and 168 columns. Given the "material" feature, which is a categorical feature, is essentially one-hot encoded with the specific ratios of the elements, the "material" column was dropped. In addition, the following nine columns were removed as they contained no unique columns: He, Ne, Ar, Kr, Xe, Pm, Po, At, Rn. The target column, critical temperature, was dropped, resulting in a dataset comprised purely of numerical features used for modeling. This resulted in 157 features in the dataset. No further preprocessing was needed.

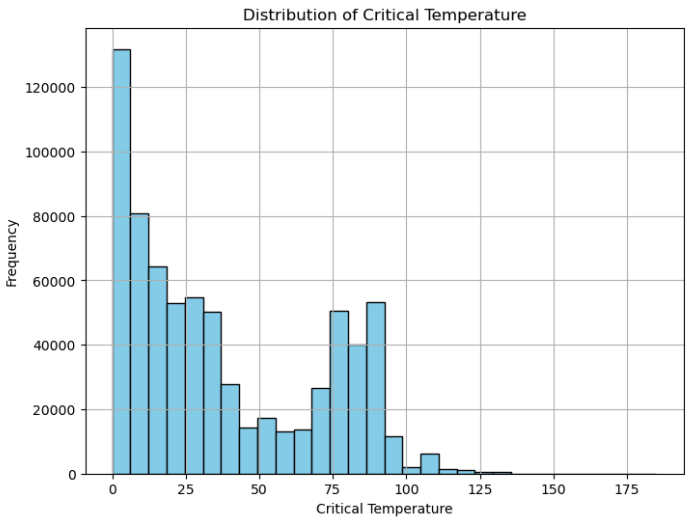

Figure 1

EXPLORATORY DATA ANALYSIS

Given such a large amount of features and not a lot of additional information about the data, an exploratory data analysis was conducted to determine if the data required scaling. The first task was to inspect the distribution of the response variable, critical temperature, which can be seen in Figure 1. The histogram shows the response variable is quite skewed to the right. In addition, a correlation matrix of the response variables against feature variables was created from which some of the histograms for some of the most correlated features are shown in Figure 2. Most of these also show evidence of skewness which is evidence against the assumption of normality required for linear regression.

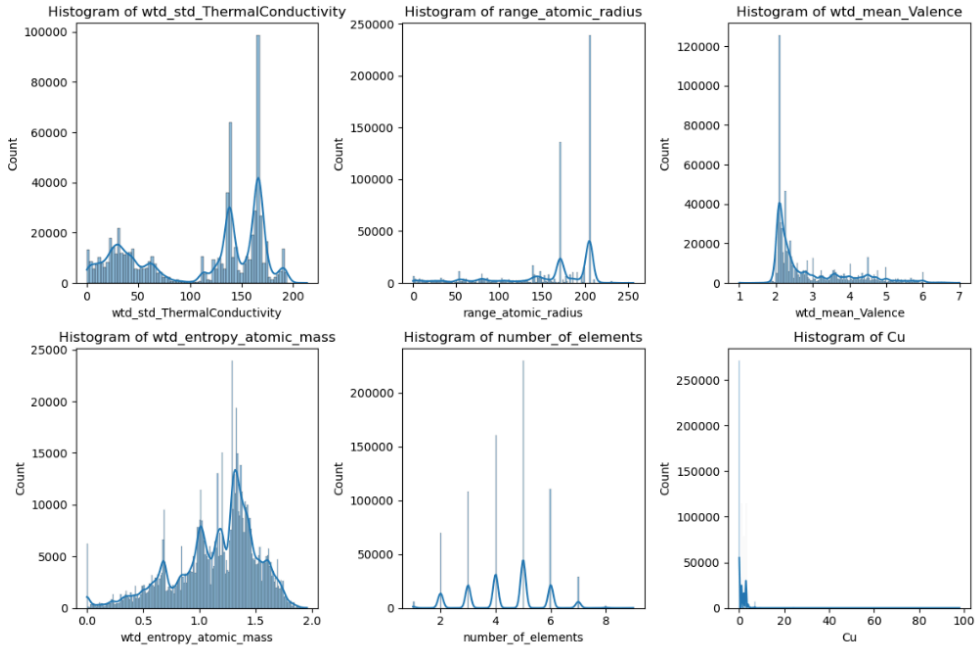

Figure 2

DATA PREPERATION

In order to meet the requirements for linear regression, the explanatory variables only (excluding the response variable, critical temperature) were scaled using the ScandardScalar provided by Sklearn which standardizes features by removing the mean and scaling to unit variance.

3 | RESULTS

To assess the performance of the model, a resampling technique called cross-validation was used which helps in estimating how well a model will generalize to an independent dataset. This technique splits the dataset into multiple subsets, train the model on some of these subsets, and then evaluate its performance on the remaining subsets. This ensures a more reliable estimate of a model's performance and reduces the risk of overfitting. The performance metric used

While linear regression is a powerful tool, it is sometimes not sufficient by itself in certain situations for various reasons including if non-linearity if present in data, as it is in our case. In addition, linear regression by itself is susceptible to both overfitting and underfitting. When features are highly correlated, as in our case, that can also lead to unstable and unreliable coefficient estimates as linear regression can be sensitive to multicollinearity as well as to outliers. When assumption violates occur, including that the residuals are normally distributed and have constant variance, which ours did not from a visual inspection, this can lead to biased parameter estimates and inaccurate confidence intervals. For this reason, we perform regularization which plays a crucial role in addressing some of the limitations of linear regression by adding penalty terms to the linear regression objective function which discourage the model from fitting the training data too closely, preventing overfitting. To determine the best alpha value for the regularization models, we used grid search to loop over alpha values 0.1, 1, 10, 100, and 1000 and calculated the average cross-validation score for each alpha value.

L1 REGULARIZATION (LASSO)

L1 regularization modifies the objective function by adding a penalty term that is proportional to the absolute value of the coefficients. The results can be seen in Figure 3. In the end, an alpha value of 0.1 yielded the highest cross-validation score of 0.563. A scatter plot of the predicted values against the actual values, represented by the red line, is seen in Figure 4. The predicted values are fairly clustered around the actual values with a few outliers but there is definitely room for improvement.

```
Alpha: 0.1
Cross-validation scores: [0.59706383 0.59358701 0.58475395 0.68168831 0.35960713]
Mean score: 0.5633400431425108
------
Alpha: 1
Cross-validation scores: [0.5282424  0.53643114 0.51549462 0.61434319 0.24303855]
Mean score: 0.4875099806355664
------
Alpha: 10
Cross-validation scores: [ 0.22410157  0.26607961  0.03751175  0.32916063 -1.28812447]
Mean score: -0.08625418115892658
------
Alpha: 100
Cross-validation scores: [-0.1863932  -0.15014428 -0.6505522  -0.13231696 -3.30956441]
Mean score: -0.8857942098694774
------
Alpha: 1000
Cross-validation scores: [-0.1863932  -0.15014428 -0.6505522  -0.13231696 -3.30956441]
Mean score: -0.8857942098694774
------
Best Alpha: 0.1
Best Mean Cross-validation Score: 0.5633400431425108
```

Figure 3

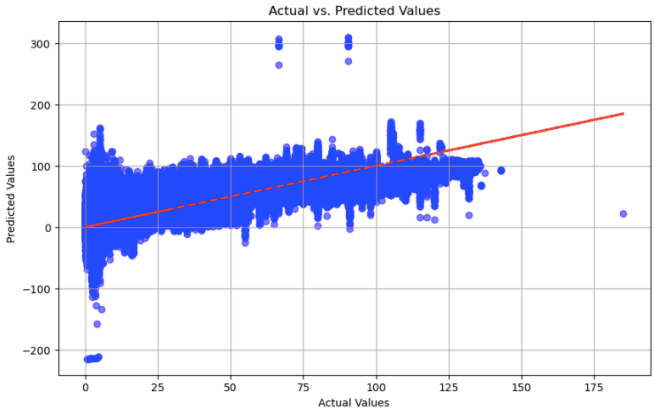

Figure 4

L2 REGULARIZATION (RIDGE)

L2 regularization adds a penalty term to the linear regression objective function based on the L2 norm of the coefficients, which is the square root of the sum of the squared values of the coefficients. While L2 regularization is frequently helpful in dealing with data that has multicollinearity, such as in this case, it seems that this regularization method did not perform as well as L1 regularization with the best cross-validation score being 0.428 at an alpha of 1000, which was also surprising. These results can be seen in Figure 5. A scatter plot of the predicted values against the actual values, depicted as the red line, can be seen in Figure 4. The distribution looks light the predicted points are actually more clustered around the actual values than the L1 regularization plot above, which should be investigated further.

```
Alpha: 0.1
Cross-validation scores: [-0.33391872  0.61901511  0.61515148  0.69691439  0.19692612]
Mean score: 0.3588176753779258
-------
Alpha: 1
Cross-validation scores: [-0.33352533  0.61903244  0.61516971  0.69691977  0.19719741]
Mean score: 0.3589587996511264
-------
Alpha: 10
Cross-validation scores: [-0.32953717  0.61916028  0.61528387  0.69694919  0.19910484]
Mean score: 0.3601922031755075
-------
Alpha: 100
Cross-validation scores: [-0.29015753  0.61939686  0.61507395  0.69677709  0.20492961]
Mean score: 0.3692039941644468
-------
Alpha: 1000
Cross-validation scores: [0.00200935 0.6165977  0.61125781 0.6944564  0.2178354 ]
Mean score: 0.42843133055138916
------
Best Alpha: 1000
Best Mean Cross-validation Score: 0.42843133055138916
```

Figure 5

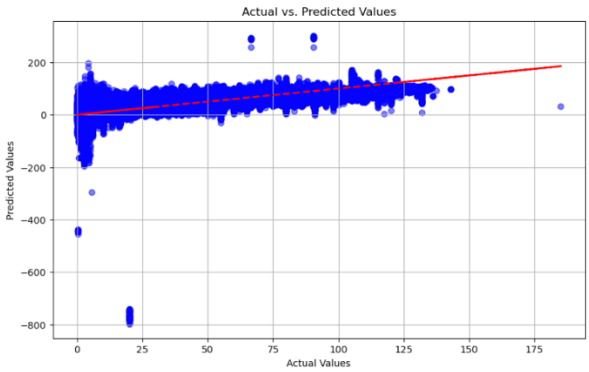

Figure 6

ELASTIC NET

Elastic Net is a regularization technique that combines both L1 (Lasso) and L2 (Ridge) Regulariation. It includes both penalty terms in the objective function allowing for feature selection (like Lasso) and handling correlated features (like Ridge). The same method for determining the best alpha was used as for the previous models. However, Elastic Net uses an additional hyperparameter L1_ratio, which determines the mix between L1 and L2 regularization penalties in the objective function to control the trade-off between the two regularization terms. The L1_ratio takes a value between 0 and 1, where a value of 0 corresponds to pure L2 (Ridge) regression, a value of 1 corresponds to pure L1 (Lasso) regression, and values in between 0 and 1 allow for a flexible combination of both penalties. To determine the best L1_ratio to use in the

Elastic Net model, a grid search was performed to loop over the L1_ratio values of 0.1, 0.5, 0.7, and 1. The highest cross-validation score was 0.563 with the hyperparameters, alpha equal to 0.1, and an L1_ratio equal to 1. This model performed the same as the L1 regularization model, which is expected with an L1_ratio of 1 which mimics pure L1 regularization. The plot of the predicted values against the actual values, represented by the red line, is seen in Figure 7.

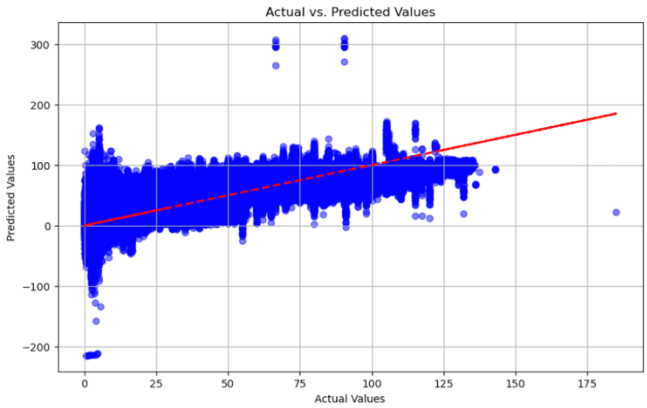

Figure 7

4 | CONCLUSION

While the results for the L1 Regularization model and the Elastic Net model were the same, the Elastic Net Model was used to determine the most important features in predicting critical temperature. The top ten most important features can be seen in Figure 8 along with the value of their coefficient. A bar graph is shown in Figure 9 that visualizes the top ten features and their coefficients.

	Feature	Coefficient
136	Ba	11.035727
62	wtd_mean_ThermalConductivity	8.561095
7	range_atomic_mass	6.845027
64	wtd_gmean_ThermalConductivity	-6.648629
49	std_ElectronAffinity	6.271378
47	range_ElectronAffinity	-4.878852
94	Si	-4.272386
80	wtd_std_Valence	-4.082574
118	Sr	3.921893
100	Ca	3.689896

Figure 8

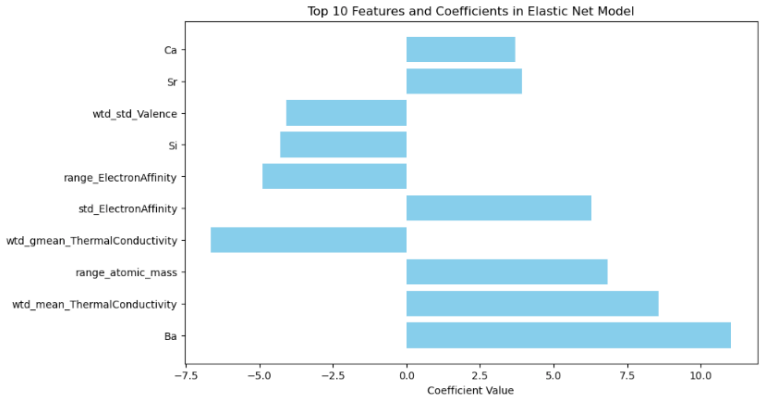

Figure 9

5 | CODE

Relevant code is attached in CarolinaCraus_CS1.ipynb