DS7333 Case Study 1

John Girard, Shijo Joseph, Douglas Yip

Modeling Superconductors Using Linear Regression

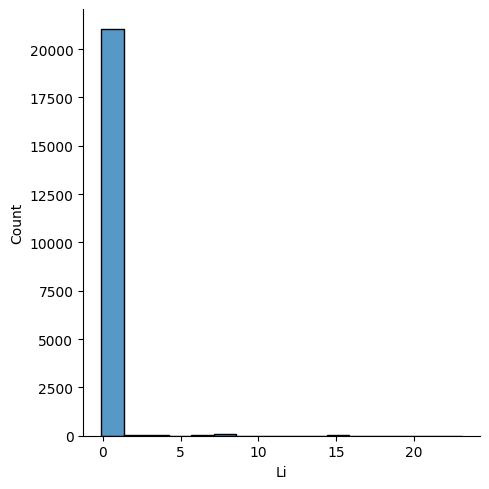
**Introduction**

A group of scientists have been looking at superconductors, which is of pretty big importance to the scientific community. Superconductors are materials that give little or no resistance to electrical current. They are looking for us to produce a model that predicts new superconductors based on the properties in the data they found so far. The data provided shows multiple characteristics of various researched superconductors. Some of these characteristics are number of elements, mean atomic mass, entropy, material composition, temperature at which they superconduct (critical temperature) and other important categories that are useful in defining each superconductor.

Our team is tasked to build a linear regression model using L1 or L2 regularization to predict the critical temperature of new superconductors in which they operate based on the inputs from the dataset. We must also define the important features that influence the critical temperature.

**Methods**

Our initial step was to do EDA on the files. Once we joined the files together in a dataframe, we looked for the total records in the dataset. The data contained 21263 rows and 169 columns. The target variable for the linear regression was found to be the critical temperature. We then looked to see if there was any missing data and found that there was no missing data. The next step was to scale the data using the StandardScaler function from scikit learn. The team looked at the histogram charts of attributes that were not sparse, to show the standard deviations and distributions.



**Figure 1.** A histogram of lithium. The element is mostly sparse as few materials have lithium.

Most elements were sparse based on the composition of the material as shown in Figure 1. We then looked at other attributes and noticed that they were for the most part normally distributed with slight skewness to the left or right as show in Figure 2 and 3.

Chart, histogram

Description automatically generated

**Figure 2:** A histogram of the mean atomic mass column in the dataset showing an approximately normal distribution.

Chart, histogram

Description automatically generated

**Figure 3:** A histogram of the entropy atomic mass column in the dataset showing an approximately normal distribution with left skewness.

With so many rows in our dataset we decided to invoke the central limit theorem.

The next step is to reduce the number of dimensions prior to creating the linear regression model. We used two methods to identify columns that are high correlated 1) Variance Inflation Factor (VIF) identifying high multicollinearity; 2) A correlation matrix to identify columns with high R2. The removal of correlated variables resulted in the reduction of variables from 169 to 94 with the important columns listed in Figure 4.

['H', 'He', 'Li', 'Be', 'B', 'C', 'N', 'O', 'F', 'Ne', 'Na', 'Mg', 'Al', 'Si', 'P', 'S', 'Cl', 'Ar', 'K', 'Ca', 'Sc', 'Ti', 'V', 'Cr', 'Mn', 'Fe', 'Co', 'Ni', 'Cu', 'Zn', 'Ga', 'Ge', 'As', 'Se', 'Br', 'Kr', 'Rb', 'Sr', 'Y', 'Zr', 'Nb', 'Mo', 'Tc', 'Ru', 'Rh', 'Pd', 'Ag', 'Cd', 'In', 'Sn', 'Sb', 'Te', 'I', 'Xe', 'Cs', 'Ba', 'La', 'Ce', 'Pr', 'Nd', 'Pm', 'Sm', 'Eu', 'Gd', 'Tb', 'Dy', 'Ho', 'Er', 'Tm', 'Yb', 'Lu', 'Hf', 'Ta', 'W', 'Re', 'Os', 'Ir', 'Pt', 'Au', 'Hg', 'Tl', 'Pb', 'Bi', 'Po', 'At', 'Rn', 'mean\_atomic\_mass', 'entropy\_atomic\_mass', 'range\_atomic\_mass', 'mean\_ElectronAffinity', 'range\_ElectronAffinity', 'mean\_FusionHeat', 'mean\_Valence', 'range\_Valence']

**Figure 4:** List of 94 variables remaining post removal of high correlated explanatory variables

Two Linear Models, L1 (LASSO) and L2 (Rigid), using a 5 fold cross validation will be performed for 100 different strengths of regularization from 10-10 to 10, with the best regularization strength, alpha, selected based on the lowest average mean squared error.

**Results**

In Table 1, the results of the two models yielded the following RMSEs with the best alpha for the dataset. Table 2 and Table 3 are the top 10 attributes and strength of each model.

**Table 1**. Linear Regression Results

|  |  |  |
| --- | --- | --- |
| Model | Best regularization α strength after 5 Fold CV | RMSE |
| L1 Lasso | 0.31 | **-394.17** |
| L2 Rigid | 2154 | **-481.026** |
|  |  |  |

**Table 2**. L1 (α = 0.31) Top 10 attributes and weights

|  |  |
| --- | --- |
| Attribute | Weights |
| mean\_atomic\_mass | -1.68307705 |
| range\_Valence | -3.01934362 |
| As (Arsenic) | -3.21424955 |
| mean\_Valence | -4.5381015 |
| entropy\_atomic\_mass | 4.54877242 |
| Range\_atomic\_mass | 4.78508783 |
| Sr (Strontium) | 5.48477324 |
| Si (Silicon) | -5.60627905 |
| Ca (Calcium) | 6.62230952 |
| Ba (Barium) | 15.09154065 |

**Table 2**. L1 (α = 2154) Top 10 attributes and weights

|  |  |
| --- | --- |
| Attribute | Weights |
| mean\_atomic\_mass | -2.15028426 |
| As (Arsenic) | -2.22784546 |
| range\_Valence | -3.21424955 |
| mean\_Valence | -3.97111182 |
| Sr (Strontium) | 3.98285713 |
| entropy\_atomic\_mass | 4.27037221 |
| Si (Silicon) | -4.51545324 |
| Range\_atomic\_mass | 4.62884579 |
| Ca (Calcium) | 5.82763934 |
| Ba (Barium) | 12.03356889 |

**Conclusion**

The L1 regression model was the best model (RMSE = -394.17) for predicting the critical temperatures for superconductors. We removed over 50% of the attributes when creating the final model while minimizing the mean squared error. Our model suggest the following elements Barium (Ba), Calcium (Ca), Strontium (Sr) increases the critical temperature while Arsenic (As) or Silicon (Si) reduces critical temperature. Variables such as the mean atomic mass, the range valence, mean valence increase also increase critical temperature decrease. The top attribute that has significant impact to a super conductors critical temperature is Barium.

**Appendix A.**

*Linted python code (utilizing flake8)*

* Cut code here