DS7333 Case Study 1

John Girard, Shijo Joseph, Douglas Yip

Modeling Superconductors Using Linear Regression

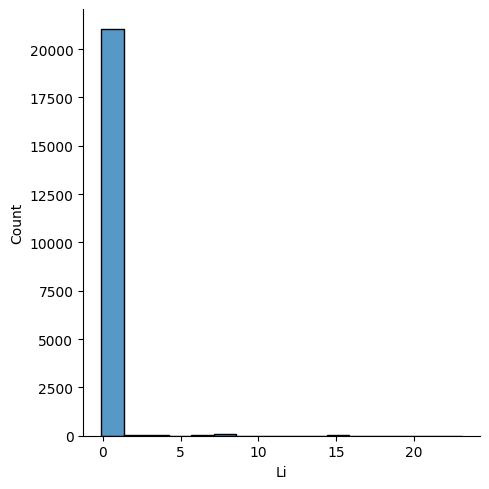
**Introduction**

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 with predicting new superconductors and the critical temperature 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 1.1a:** 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. 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.

Chart, histogram

Description automatically generated

**Figure 1.2b:** A histogram of the mean atomic mass column in the dataset showing an approximately normal distribution.

Chart, histogram

Description automatically generated

**Figure 1.2c:** 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 reduces the number of dimensions needed to create the linear regression model. We ran a VIF for each of the columns in the dataframe to remove columns with high multicollinearity. The next step was to create a correlation matrix to remove any columns with high correlation to another column. In the end we ended up with the following columns:

['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']

We then ran L1 (Lasso) against the dataset. By Utilizing a 5 fold cross validation and loop to iterate through a series of alphas, we ended up with an alpha of .31 with an MSE of -394.17. The top 10 attributes were:

['mean\_atomic\_mass', 'range\_Valence', 'As', 'mean\_Valence',

'entropy\_atomic\_mass', 'range\_atomic\_mass', 'Sr', 'Si', 'Ca', 'Ba']

And the corresponding weights were:

[-1.68307705, -3.01934362, -3.21424955, -4.5381015 , 4.54877242,

4.78508783, 5.48477324, -5.60627905, 6.62230952, 15.09154065]

We also ran a Ridge Regression model. Based on a GridsearchCV with 5 fold cross validation, the best alpha was 2154.434 with a MSE of -481.026 . The top 10 attributes were:

['mean\_atomic\_mass', 'As', 'range\_Valence', 'mean\_Valence', 'Sr',

'entropy\_atomic\_mass', 'Si', 'range\_atomic\_mass', 'Ca', 'Ba']

And the corresponding weights were:

[-2.15028426, -2.22784546, -3.212363 , -3.97111182, 3.98285713,

4.27037221, -4.51545324, 4.62884579, 5.82763934, 12.03356889]

**Results**

As you can see both models had the same Top 10 variables with the only difference being the coefficient values. Because we had a better MSE with the Ridge regression technique, we ended up using that as the final model.

**Conclusion**

The regression model provided great results in predicting critical temperatures for superconductors. We removed over 50% of the attributes when creating the final model while minimizing the mean squared error. Our model states the amount of Barium (Ba), Calcium (Ca), Strontium (Sr) increases, the critical temperature increases. On the other hand if you increase Arsenic (As) or Silicon (Si) the critical temperature decreases. Variables such as the mean atomic mass, the range valence, mean\_valence increase, the critical temperature decrease. If you increase the range atomic mass or the entropy atomic mass the critical temperature increases.