Case Study 1 – Super Conductors

Balaji Avvaru, Ravi Sivaraman, Apurv Mittal

Abstract

The objective of this study is to predict a composition to create a new superconductor and the critical temperature the superconductor will operate. The superconductors are materials that give little or no resistance to electrical currents at a specific (critical) temperature. This study analyzes the recorded data to predict the most important parameters to create a new superconductor.

1. Introduction

Superconductors are materials that conduct electricity with no resistance. This means a superconductor can carry a current indefinitely without losing any energy. However, the superconductors exhibit this property only at a critical temperature. For example: Mercury if cooled below 4.1 Kelvin becomes a super conductor and offers no resistance to electrical current.

The dataset used for this study provides the list of different superconductors identified with their chemical compositions and the critical temperature. The objective of this study is to predict the critical temperature based on the features extracted for a superconductor.

The dataset provided has two files

a. *train.csv* which contains 81 features extracted from 21,263 superconductors along with the critical temperature.
b. *unique_m.csv* contains the chemical formula broken up for all the 21,263 superconductors from the train.csv file.

The main variables captured in the file include:

- **atomic_mass** - total proton and neutron rest masses, in Atomic Mass Units (AMU).
- **fie** - First Ionization Energy, energy required to remove a valence electron, in kilojoules per mole (kJ/mol).
- **atomic_radius** - calculated atomic radius, in picometer (pm).
- **density** - density at standard temperature and pressure, in kilograms per meters cubed (kg/m3).
- **electron_affinity** - energy required to add an electron to a neutral atom, in kilojoules per mole (kJ/mol).
- **fusion_heat** - energy to change from solid to liquid without temperature change, in kilojoules per mole (kJ/mol).
- **thermal_conductivity** - thermal conductivity coefficient k, in watts per meter-kelvin.

- **valence** - typical number of chemical bonds formed by the element, no units.
- **critical_temp** - superconductor critical temperature, in Kelvin.

Reference: https://www.neuraldesigner.com/learning/examples/superconductivity

Statistics of each variable are included such as: mean, weighted mean, geometric mean, weighted geometric mean, entropy, weighted entropy, standard, weighted standard, range and weighted range.

2. Methods

The data from the files were extracted and joined via index of each row since each row in one file corresponds to the same row in the other file. First step of the data analysis is to look for any missing data. The data was identified to be complete and no missing values, thus no data imputation was needed.

Two methods were primarily used for regression models:

1. Method 1: All features including highly correlated variables
2. Method 2: Reduced feature set with highly correlated features removed

The results from both Method 1 and Method 2 are shared in "Results" section for more analysis.

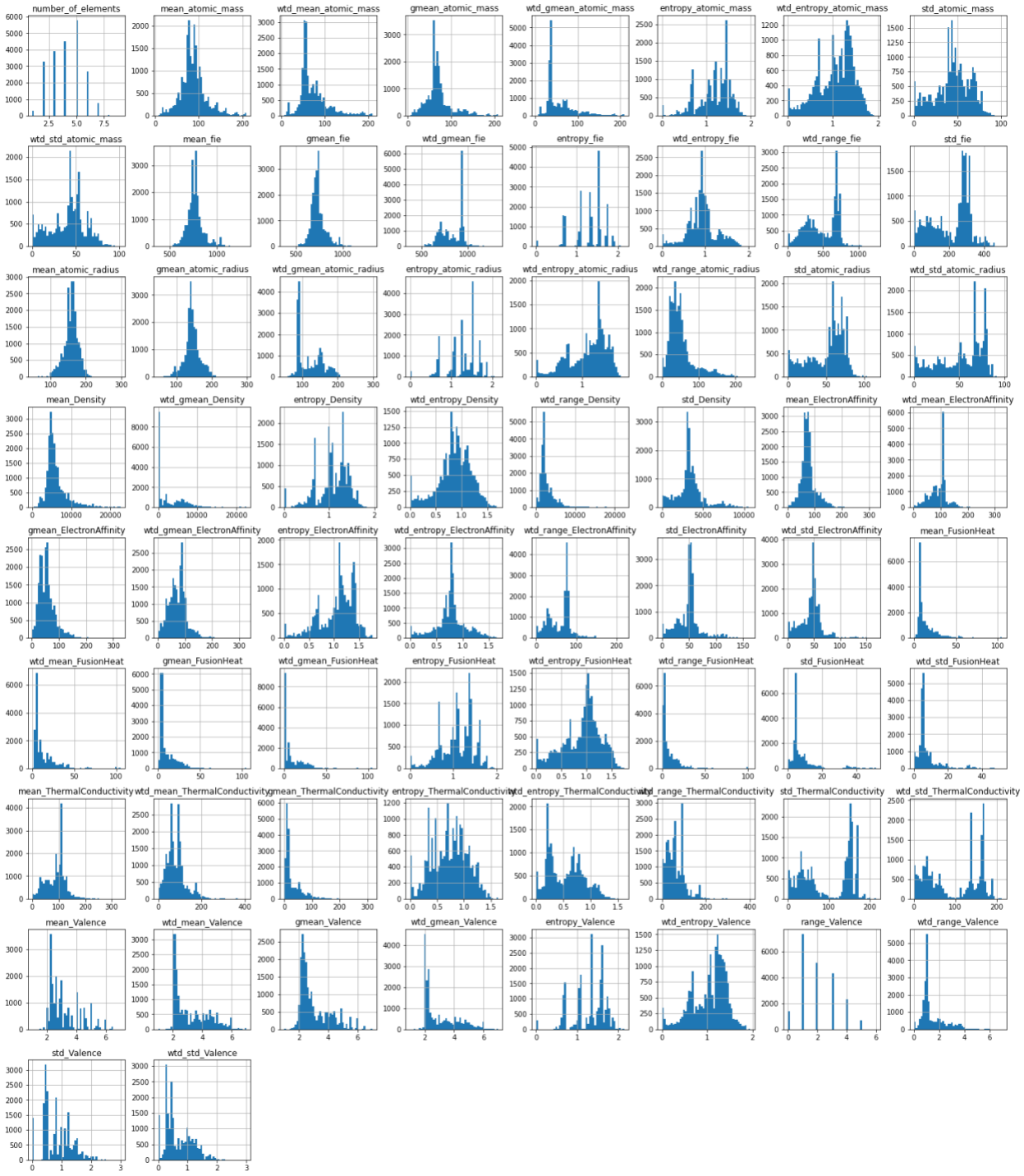
Method 2: Reduced Dataset

Following methodology was used to remove the variables from the original dataset:

The dataset was plotted to check the data distribution. Most of the variables shows normal distribution or close to normal distribution. Some variables were skewed as shown below. The variables which were variant of the original data like: weighted mean, geometric mean or weighted geometric mean etc. were checked for distribution and only the variable with closer to normal distribution was retained in the dataset for further analysis. Since there are several variables capturing similar information of the same variable for example below variable are variants of the same data:

a. mean_atomic_radius
b. wtd_mean_atomic_radius
c. gmean_atomic_radius
d. wtd_gmean_atomic_radius
e. wtd_mean_atomic_radius
f. range_atomic_radius

In above example, ‘wtd_mean_atomic_radius’, ‘range_atomic_radius’ were removed from the dataset as part of the Method 2 reduced dataset.



As explained above, the following variables were removed from the dataset for capturing similar information:

1. range_atomic_mass
2. wtd_range_atomic_mass
3. wtd_mean_fie
4. range_fie
5. wtd_std_fie
6. wtd_mean_atomic_radius
7. range_atomic_radius
8. wtd_mean_Density
9. gmean_Density
10. range_Density
11. wtd_std_Density
12. range_ElectronAffinity
13. range_FusionHeat
14. wtd_gmean_ThermalConductivity
15. range_ThermalConductivity

Several variables in the *train.csv* file have high collinearity and to build a good independent model it's important to reduce the impact of multi-collinearity. The variables with correlation above 95% were removed from the dataset. Total of 12 variables were identified to have greater than 95% correlation with other variables as listed below:

1. 'wtd_gmean_atomic_mass'
2. 'gmean_fie'
3. 'entropy_fie'
4. 'entropy_atomic_radius'
5. 'wtd_entropy_atomic_radius'
6. 'wtd_gmean_FusionHeat'
7. 'wtd_std_ThermalConductivity'
8. 'gmean_Valence'
9. 'wtd_gmean_Valence'
10. 'entropy_Valence'
11. 'wtd_entropy_Valence'
12. 'std_Valence'

In "*Unique_m.csv*" data set removed the variables without variance. There were 9 variables with only "0" values, those variables were dropped. Here is the list of variables dropped:

1. 'He'
2. 'Ne'
3. 'Ar'
4. 'Kr'
5. 'Xe'
6. 'Pm'

7. 'Po'
8. 'At'
9. 'Rn'

Also, the variable "material" which comprises of all elements used as materials is dropped from the data frame. This information is captured in the other variables which covers the combination of variable used in each allow in *"Unique_m.csv"*.

Regression Models:

After the identification of the variables following regression models were executed on the final dataset using cross validation (at k folds = 10):

a. Linear Regression
b. Lasso (at various alpha levels)
c. Ridge Regression (at various alpha levels)
d. ElasticNet

Root Mean Square Error (RMSE) was used to identify the most successful models.

3. Results

Method 1: Models with original data set

Model	RMSE	Standard Deviation of RMSE at CV=10
Multi-Linear Regression	-22.57	12.42
Lasso (at alpha = 0.3)	-17.96	0.28
Ridge (at alpha =1000)	-17.89	1.04
ElasticNet	-19.46	0.27

Method 2: Models with reduced data set as explained in the "Methods" section.

Model	RMSE	Standard Deviation of RMSE at CV=10
Multi-Linear Regression	-22.68	10.94
Lasso (at alpha = 0.3)	-18.35	0.33
Ridge (at alpha =1000)	-18.21	1.09
ElasticNet	-19.67	0.29

Ridge Model showed the lowest RMSE for Method 1, however it has higher standard deviation compared to Lasso at alpha of 0.3. Upon comparing the score for Ridge, its apparent that the scores are not consistent.

Ridge Scores = `-17.20146094, -17.76609165, -17.70643718, -17.55575516, -17.52395924, -18.28908214, -17.33259556, -17.65596932, -17.0831061 , -20.85931023`

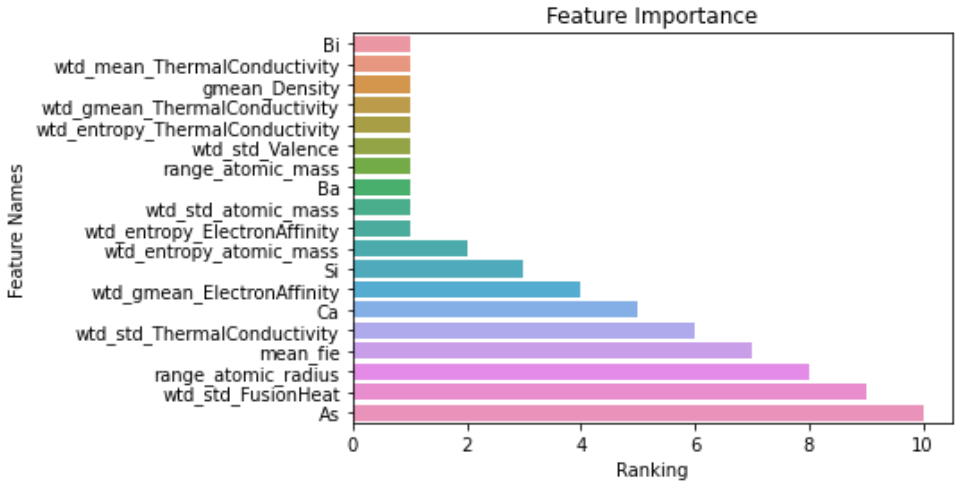
Lasso at $\alpha = 0.3$ shows consistent scores and very low standard deviation. Even with slightly higher RMSE score for Lasso at -17.96 vs -17.89 for Ridge model. The best model with consistent scores is Lasso (α=0.3) for Method 1 (with all data variables considered in the model).

Lasso Scores = `-17.49501274, -18.18279108, -18.04697072, -17.9540785, -17.99042669, -18.24794041, -17.76672689, -18.01578225, -17.48547188, -18.39358418`

The recommended model is **Lasso (α=0.3)** with all data variables in its original form and will be used for further analysis and prediction.

Lasso – Important Features (Top 10 Features)

Based on the Lasso model, the top 10 features to be considered for the regression model are listed below:



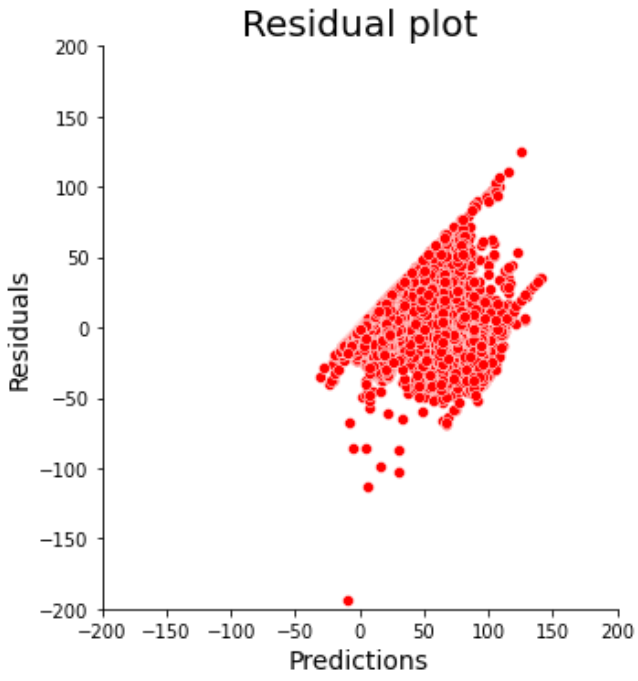
Most important features for the regression model:

1. Bi
2. wtd_mean_ThermalConductivity
3. gmean_Density
4. wtd_gmean_ThermalConductivity
5. wtd_entropy_ThermalConductivity
6. wtd_std_Valence
7. range_atomic_mass
8. Ba
9. wtd_std_atomic_mass

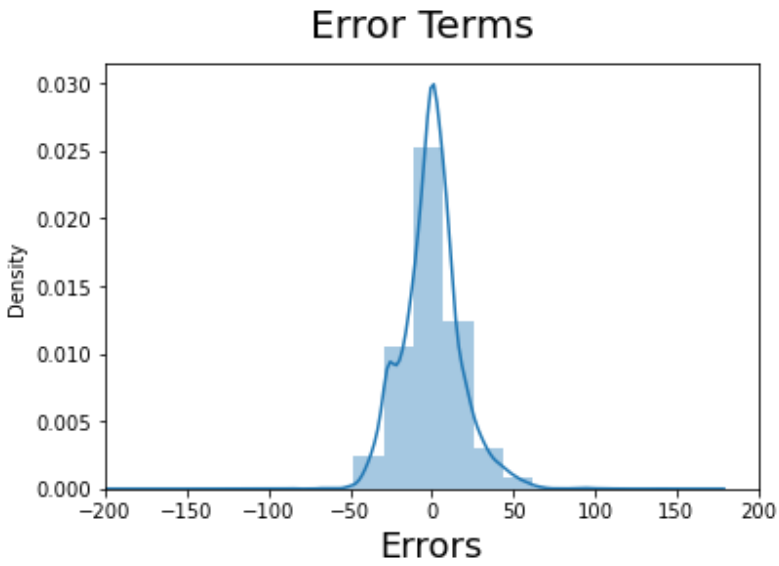
10. wtd_entropy_ElectronAffinity
11. wtd_entropy_atomic_mass
12. Si
13. wtd_gmean_ElectronAffinity
14. Ca
15. wtd_std_ThermalConductivity
16. mean_fie
17. range_atomic_radius
18. wtd_std_FusionHeat
19. As

Lasso Residual Plot

The residual plot of Lasso appears to be well within bounds and together with very few outliers. However, there is slight pattern in the distribution of the residuals. Since majority of the residual values are co-located this provides confidence that the model is successful in predicting the critical temperature with the selected parameters.



Similarly, Error distribution also shows centered at 0 with uniform distribution.



4. Conclusion

Based on the Lasso model, the multiple linear regression model can be interpreted as below:

$$\hat{\mu}\{\text{Critical Temperature}\} = \beta_0 + \beta_1 Bi + \beta_2 \text{wtd\_mean\_ThermalConductivity} + \beta_3 gmean\_Density + \beta_4 wtd\_gmean\_ThermalConductivity + \beta_5 wtd\_entropy\_ThermalConductivity + \beta_6 wtd\_std\_Valence + \beta_7 range\_atomic\_mass + \beta_8 Ba + \beta_9 wtd\_std\_atomic\_mass + \beta_{10} wtd\_entropy\_ElectronAffinity \ldots.$$

$$\hat{\mu}\{\text{Critical Temperature}\} = 34.42 + 3.85\ Bi + 10.60\ \text{wtd\_mean\_ThermalConductivity} - 1.82\ gmean_{Density} - 8.91 wtd_{gmean_{ThermalConductivity}} + 2.65 wtd_{entropy_{ThermalConductivity}} - 4.05 wtd_{std_{Valence}} + 5.43 range_{atomic_{mass}} + 8.75 Ba - 3.34 wtd_{std_{atomic_{mass}}} - 2.75 wtd\_entropy\_ElectronAffinity$$

…..

Interpretation:

1. All features being zero, the critical temperature of the allow will be 34.42 K.
2. Considering all features at constant, with per unit addition of the Bi (Bismuth) the critical temperature increases by 3.85 K.
3. Considering all features at constant, with per unit increase in wtd_gmean_ThermalConductivity the critical temperature decreases by 8.91 K.
4. Similarly, all identified features impact the critical temperature of the superconductor.

Appendix – Code

The Jupyter notebook (code) is uploaded along with the report in 2ds assignments for case study 1.