Predicting Critical Temperatures for Superconductors

Satvik Ajmera

May 10, 2022

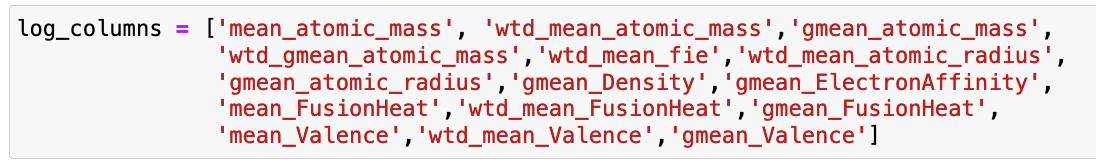
1 Introduction

The features in the data consist of material composition, properties, and temperatures for superconductors. The goal is to predict at what critical temperatures can new superconductors become superconductors. The task is to build a linear regression model using L1 or L2 regularization (or both) the task to predict the Critical Temperature as closely as possible. In addition, to look at the variable importance of the linear models. The entire dataset consists of 21,263 observations. Additionally, the input features, *X*, consists of 81 regressors and response variable, *y* (critical temperature) is also a regressor.

2 Methods

Because there is no missing data, there is no need for data imputation or excessive data cleansing. Looking at a kernel density plot for each input feature, I wanted to see if a log transformation could be applied to any overly skewed data. I determined 14 input features that needed a log transformation (below). Upon applying a log transformation, the kernel density plots for these features looked more normal.

**Figure 1 - Logarithmic Transformed Input Variables:**



For preprocessing, I standardized and scaled the input features using sklearn’s StandardScaler, so *X* has a mean of 0 and standard deviation of 1. From sklearn’s model selection, I used the Repeated K-Fold cross validation for the train-test split. Specifically, I used 10 splits (n\_splits) and 5 repeats (n\_repeats). The performance metric used for testing is mean squared error.

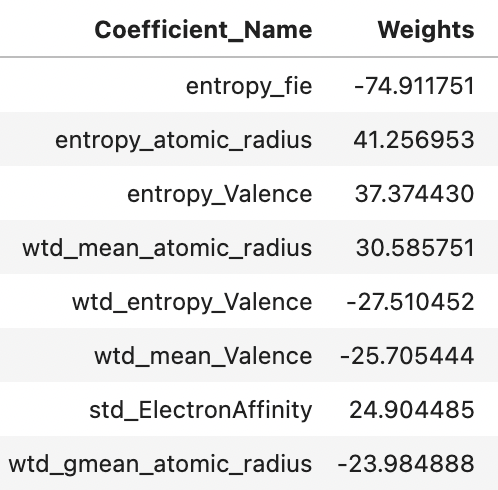
3 Results

**Figure 2 - Model Comparison and Performance:**

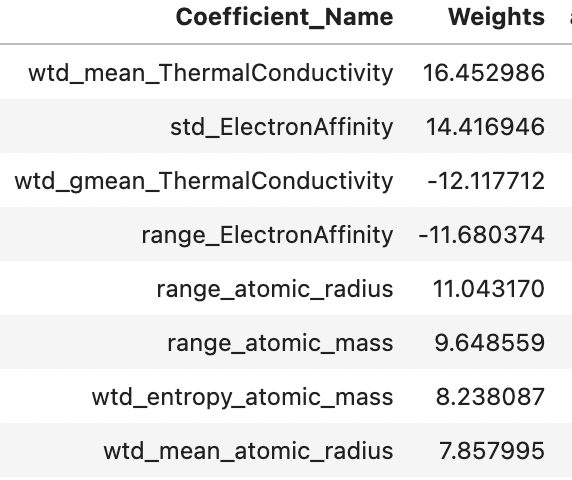
| **Feature Transformation** | **Regularization** | **Alpha** | **Average MSE Over All Folds** |
| --- | --- | --- | --- |
| None | None | N/A | 315.82 |
| Logarithmic Transformation | None | N/A | 312.35 |
| Logarithmic Transformation | L1 | 0.08 | 334.2 |
| Logarithmic Transformation | L1 | 0.2 | 357.36 |
| Logarithmic Transformation | L1 | 0.5 | 381.32 |
| Logarithmic Transformation | L1 | 0.7 | 391.86 |
| Logarithmic Transformation | L2 | 0.9 | 312.41 |
| Logarithmic Transformation | L2 | 0.7 | 312.392 |
| Logarithmic Transformation | L2 | 0.5 | 312.372 |
| Logarithmic Transformation | L2 | 0.3 | 312.357 |
| Logarithmic Transformation | L2 | 0.1 | 312.352 |

There are several models that were built and tested. After log transforming the 14 features above, the average mean squared error over all the folds decreased from 315.82 to 312.35. Moving forward, I used the log transformed features with L1 and L2 regularization. L1 regularization with alpha = 0.08 resulted in the lowest average MSE of 334.2. For L2 regularization and alpha = 0.1 resulted in the lowest average MSE of 312.352. L1 regularization performed the worst of the models and did not minimize the average MSE as well. L2 regularization models with different alphas resulted in approximately the same average MSE as linear regression with the log transformed features.

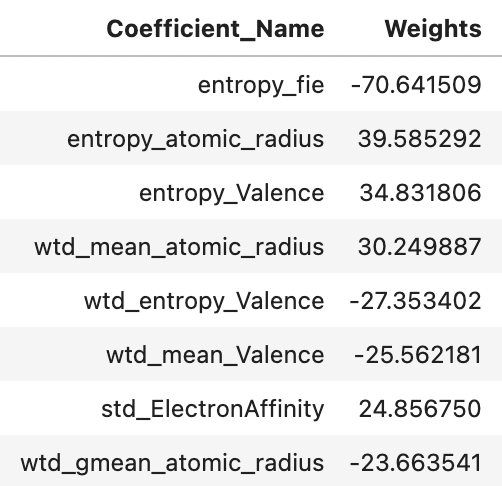
**Figure 3 - Linear Regression Top 8 Weights:**



**Figure 4 - L1 Regularization Top 8 Weights (Alpha = 0.08):**



**Figure 5 - L2 Regularization Top 8 Weights (Alpha = 0.1):**



Looking at the top 8 weights for linear regression, entropy\_fie is the most negatively associated with critical temperature (-74.91), followed by entropic\_atomic\_radius, and entropy\_Valence. We can see that L1 regularization significantly penalized the top 3 weights for linear regression. Additionally, the overall magnitude of the top 8 weights for L1 regularization are much lower than linear regression and L2 regularization. This is expected, because L1 regularization penalizes weights that grow too quickly and have high magnitude. On the other hand, L2 regularization penalizes the weights without removing them entirely. L2 regularization kept all the same top 8 weights, but reduced the magnitudes of the top 3 weights compared to linear regression without regularization.

4 Conclusion

Overall, L1 regularization is used for feature selection, while L2 regularization can be used for final model selection. In this case study, entropy features were highly correlated with the critical temperature. However, we may not want to remove features that are high collinear with the response variable, critical temperature. Log transforming some of the skewed features did result in a slight improvement in the average MSE.

5 Code

Submitted separately in Jupyter Notebook file.