Case Study 1

Joey Hernandez

Daniel Chang

DS 7333 | Quantifying the World

2023

Table of Contents

**Introduction1**

**Data1**

Data Source1

Data Analysis1

Target Variable Inspection2

**Method3**

**Results6**

**Conclusion7**

**Recommendation8**

**Appendix9**

**Introduction**

**Background**

Superconductivity is a fascinating quantum mechanical phenomenon where a material, under certain conditions, can exhibit zero electrical resistance, allowing electrical current to flow without loss of energy. The temperature at which a material transitions into a superconducting state is termed its "Critical Temperature". This parameter is crucial for applications involving superconductors, including energy storage, transportation, and advanced computing systems. Accurate prediction of the critical temperature for a given superconducting material can offer invaluable insights into material design, facilitate optimization processes, and accelerate technological innovations. However, predicting Critical Temperature​ has remained a significant challenge, given the complex interplay of variables such as material composition, structural features, and external conditions.

## Objective and Scope

The objective of this case study is to develop a robust predictive model for the critical temperature (Tc​) of superconducting materials. Specifically, we aim to construct a Linear Regression model optimized with L1 or L2 regularization techniques—or a combination of both, known as Elastic Net regularization—to predict Tc​ as accurately as possible. Regularization methods like L1 (Lasso) and L2 (Ridge) can prevent overfitting, thereby making the model generalizable to unseen data. Additionally, these techniques can assist in feature selection, an essential aspect when dealing with high-dimensional data sets commonly encountered in materials science.

## Data Source

This case study will utilize two datasets, “train” and “unique\_m”. The data was given to us in the Case 1 Study Module and is in the form of two separate csv files. When combined the data contains 21263 observations and 168 features.

## Data Inspection

Before creating any models or analysis with the data our first step was to inspect our data to better understand data types (such as int, cat, object, etc.), distributions of values, identification of missing values, duplicated data, and outliers. This step is vital in understanding how we should approach any types of transformations or adjustments to the modeling and analysis process of our data.

## Target Variable Inspection

To better understand the target variable distribution a visual inspection was performed using subplots containing various transformations of our target variable.

By plotting the histogram of the target variable in its original form, we gained insight into its inherent characteristics. However, a prominent right skew prompted the exploration of alternative transformations. The logarithmic transformation, Box-Cox transformation, and square root transformation were applied in attempt to get a more normally distributed target variable. Ultimately while none of the results were a textbook “normal” distribution, we proceeded with both normal data, and a Box Cox Transformation so that we can see performance discrepancies between various scenarios.

***Figure 1:*** *Four-Quadrant Bar Plot Illustrating the Distribution of the Target Variable*

A group of blue and black bars

Description automatically generated with medium confidence

## Correlation Plot (Original Target Variable)

Prioritizing the preprocessing steps for the target variable in the step prior to this was important because it allowed us to assess integrity and accuracy of subsequent analyses. By addressing the target variable's distribution and potential transformation needs, the resulting correlation values can be trusted to either accurately reflect the underlying relationships between variables or understand what limitations may arise from the less than desirable target variable distribution. Failure to preprocess and identify data discrepancies in the target variable could lead to misinterpretations, as correlations might be influenced by skewedness, outliers, or nonlinearities within the target data.

Next, performing a correlation heatmap provides a visually informative representation of the relationships between variables within a dataset. By illustrating the strength and direction of linear associations, the heatmap becomes an indispensable tool for uncovering patterns and dependencies that might not be immediately apparent from individual variable analyses. Each cell in the heatmap corresponds to a pair of variables, with the color gradient indicating the magnitude of correlation. This enables the rapid identification of high and low correlation values, highlighting potential areas of interest for further investigation.

***Table 1:*** *Original Target vs. Explanatory Variables: Smallest Correlations*

|  |  |
| --- | --- |
| **Feature** | **Correlation Coefficient** |
| Cs | -0.076822 |
| Tc | -0.075295 |
| std\_atomic\_radius | -0.071642 |
| S | -0.071229 |
| Er | -0.070134 |

***Table 2:*** *Original Target vs. Explanatory Variables: Largest Correlations*

|  |  |
| --- | --- |
| **Feature** | **Correlation Coefficient** |
| Pd | 0.090037 |
| Sb | 0.072646 |
| Ga | 0.058372 |
| Be | 0.057971 |
| Mg | 0.055814 |

***Table 3:*** *Transformed Target vs. Explanatory Variables: Smallest Correlations*

|  |  |
| --- | --- |
| **Feature** | **Correlation Coefficient** |
| Er | -0.113125 |
| Tc | -0.085807 |
| B | -0.080060 |
| Cs | -0.077309 |
| S | -0.076124 |

***Table 3:*** *Transformed Target vs. Explanatory Variables: Largest Correlations*

|  |  |
| --- | --- |
| **Feature** | **Correlation Coefficient** |
| Pd | 0.064658 |
| Ga | 0.050046 |
| Sb | 0.049051 |
| wtd\_range\_ThermalConductivity | 0.046485 |
| Mg | 0.045164 |

***Figure 2:*** *Example - Correlation Heatmap of Variables against original target*

A blue and green striped chart

Description automatically generated with medium confidence

***Figure 3:*** *Example - Correlation Heatmap of Variables against Transformed Target*

A blue and green striped chart

Description automatically generated with medium confidence

## Modeling

## Lasso | L1 Regularization

Lasso Regression, also known as L1 regularization, is an extension of linear regression that not only seeks the best-fitting line through the data but also constrains the size of the coefficients.

In our mission to create the most accurate model to predict the Critical Temperature, we will be using the cross\_val\_score and cross\_val\_predict functions. These functions perform cross-validation, a technique that helps us understand how well the model will perform on unseen data. It does this by splitting the dataset into training and testing sets multiple times and averaging the performance across all splits. Additionally, we will utilize LassoCV, a tool designed to find the best "alpha" value, which is the tuning parameter that balances between fitting the data well and keeping the model simple. By identifying the optimal alpha, we aim to make our Lasso model as accurate and generalizable as possible.