Raag Patel / DS 7333 – Quantifying the World / Fall 2023/ Case Study 1

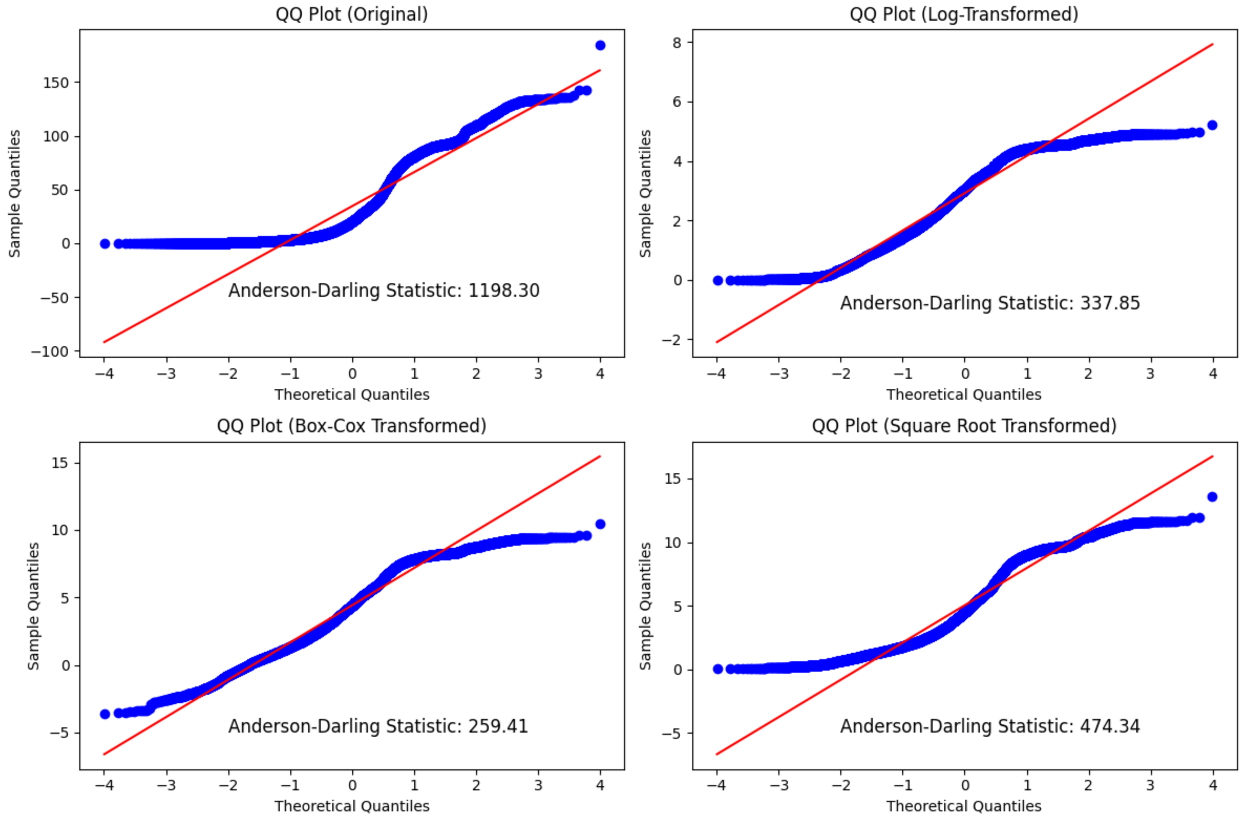
Predicting Critical Temperature of Superconductors Using Linear Regularization

Data Introduction:

Using data sourced from superconductivity research, we have 167 features, with 21,263 entries, that we will use to train models to predict the critical temperature of superconductors. There are a wide range of features, ranging from different measures of atomic mass, first ionization energy (FIE), density, electron affinity, fusion heat, thermal conductivity, and valence. There are also the elemental compounds listed, with the weights of each element for 15,542 unique materials. Though there are multiple similar elemental compounds, evident by having less unique materials than the total entries, there are no duplicates in the data.

Since we are focusing on using linear models to represent the critical temperature, we need to check the distribution of the response. The Anderson-Darling test is a statistical method used to evaluate the normality of a dataset. It assesses the hypothesis that a given sample follows a normal distribution, which is an essential assumption in linear regression. The test yields a statistic known as the Anderson-Darling statistic, which quantifies the degree of deviation from normality. Lower values of this statistic indicate closer adherence to the normal distribution, while higher values suggest significant deviations.

QQ plots are graphical representations used to compare the quantiles of observed data to the quantiles of a theoretical distribution, typically the standard normal distribution. The main objective of a QQ plot is to visually inspect whether the observed data follows a specified theoretical distribution. As mentioned, we are focusing on linear models, so the below QQ plots has a red line showing the standard normal distribution.



Looking at these plots, visually, none of the transformations applied to the original critical temperature data show a perfect normal distribution. The closest, however, is when we apply a box-cox transformation, which is a combination of a power and logarithmic transformation. Since the transformation encompasses a logarithmic transformation, we added 1 to the critical temperature, to make sure we do not have values that would leave a logarithmic transformation undefined. As well as to maintain feature importance, for when we move into feature selection, the 167 features have been scaled, so that their magnitude of

Methods:

To predict critical temperature, we are going to compare using linear regression against applying L1 and/or L2 regularization parameters, better known as Lasso, Ride, and Elastic Net.

Cross-Validation (CV) is a technique used in machine learning and statistics to assess the performance and generalization capability of a predictive model. It helps in estimating how well a model will perform on unseen data and whether it is overfitting or underfitting the training data.

Linear Regression is a fundamental supervised learning algorithm used for predicting a continuous target variable based on one or more independent predictor variables. It models the relationship between the predictors (also called features or input variables) and the target variable as a linear equation.

L1 regularization, also known as Lasso (Least Absolute Shrinkage and Selection Operator), is a technique used in machine learning to add a penalty term to the loss function during model training. This penalty encourages the model's coefficients to be exactly zero for some features, effectively performing feature selection.

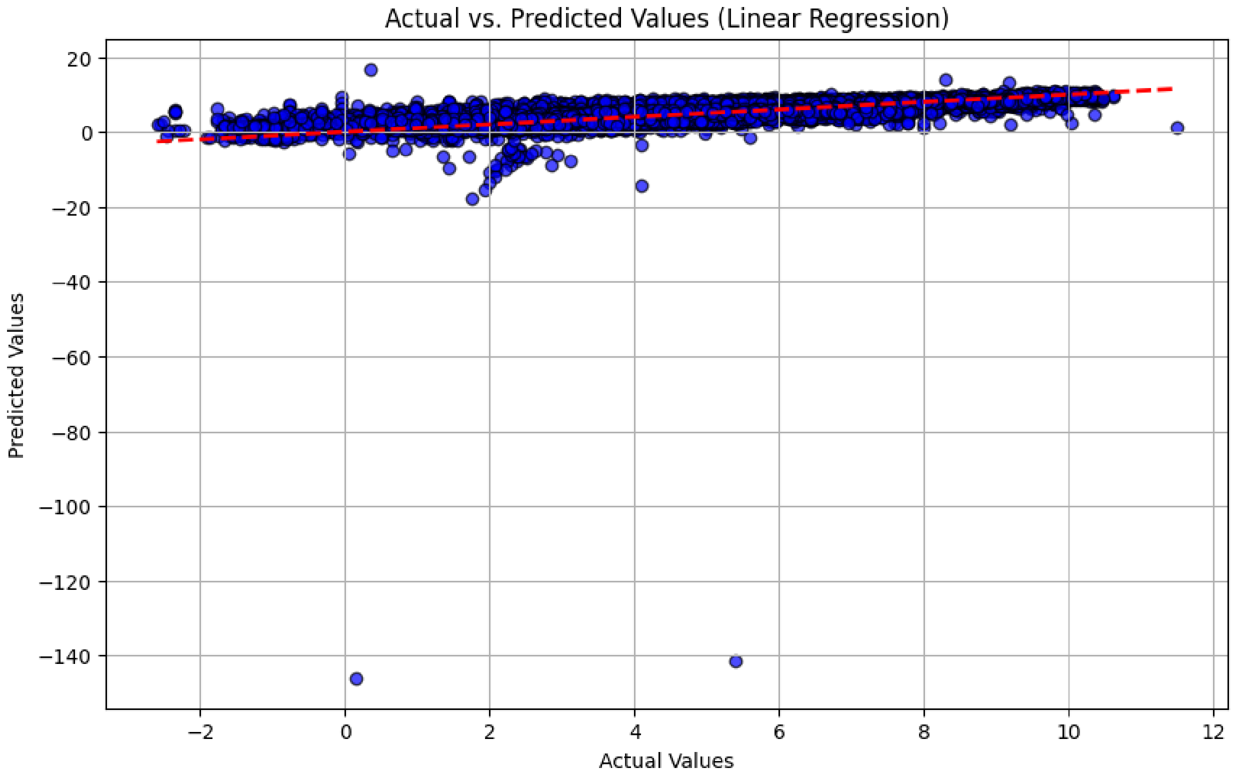
L2 regularization, often referred to as Ridge regularization, is another technique used to add a penalty term to the loss function during training. Unlike L1 regularization, L2 regularization encourages the model's coefficients to be small but not necessarily zero. It prevents overfitting by shrinking the coefficients towards zero, which helps in reducing the impact of multicollinearity and makes the model more stable.

Elastic Net regularization is a combination of L1 and L2 regularization techniques. It adds both L1 and L2 penalty terms to the loss function during training, providing a balanced approach. It combines the feature selection capabilities of Lasso with the coefficient stability of Ridge, by adjusting a hyperparameter called "l1_ratio," practitioners can control the balance between L1 and L2 regularization.

Mean Squared Error (MSE) is a widely used metric in regression analysis and machine learning for evaluating the performance of a predictive model. It measures the average squared difference between the predicted values produced by the model and the actual (observed) values in the dataset.

Multiple Linear Regression Model:

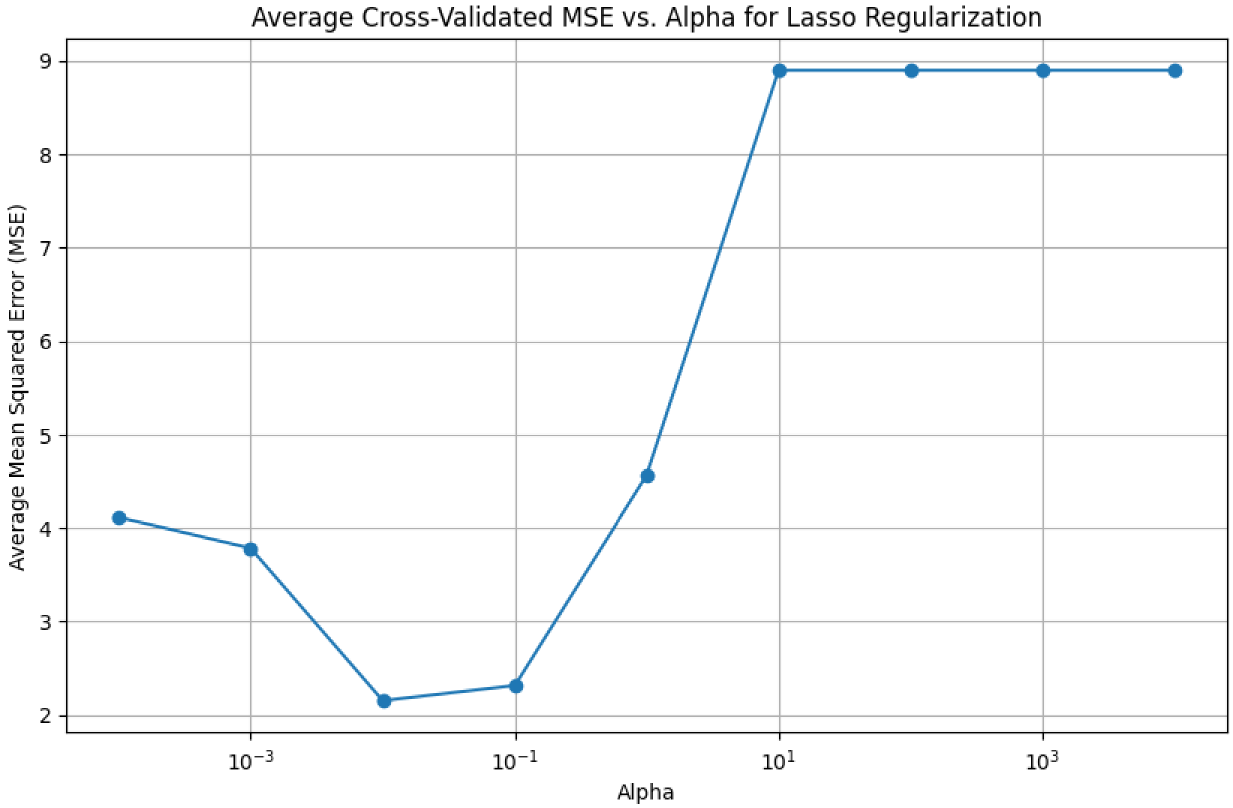
We first started off with using multi linear regression, with a 10-fold CV. All the features were used in predicting ‘critical temperature’. We found a MSE of 4.12. This number seems pretty small, which bolsters our transformation which we conducted above. We don’t have anything to compare it to yet, however, so we will find out if we can predict better.



It’s hard to tell how well the predicted did versus the actual, because of the scale. Somehow, the model severely under calculated the 2 points you see at the bottom. Looking closer at the line however, the model missed a lot of points around the actual 2 degree mark. The red line goes through the origin, which symbolizes that the predicted value matches up perfectly with the actual value.

L1 Regularization:

Using alpha (penalty term) values ranging from 10^{-4} to 10^{4}, incrementing by an order of 10, we applied each penalty term and recorded each MSE.

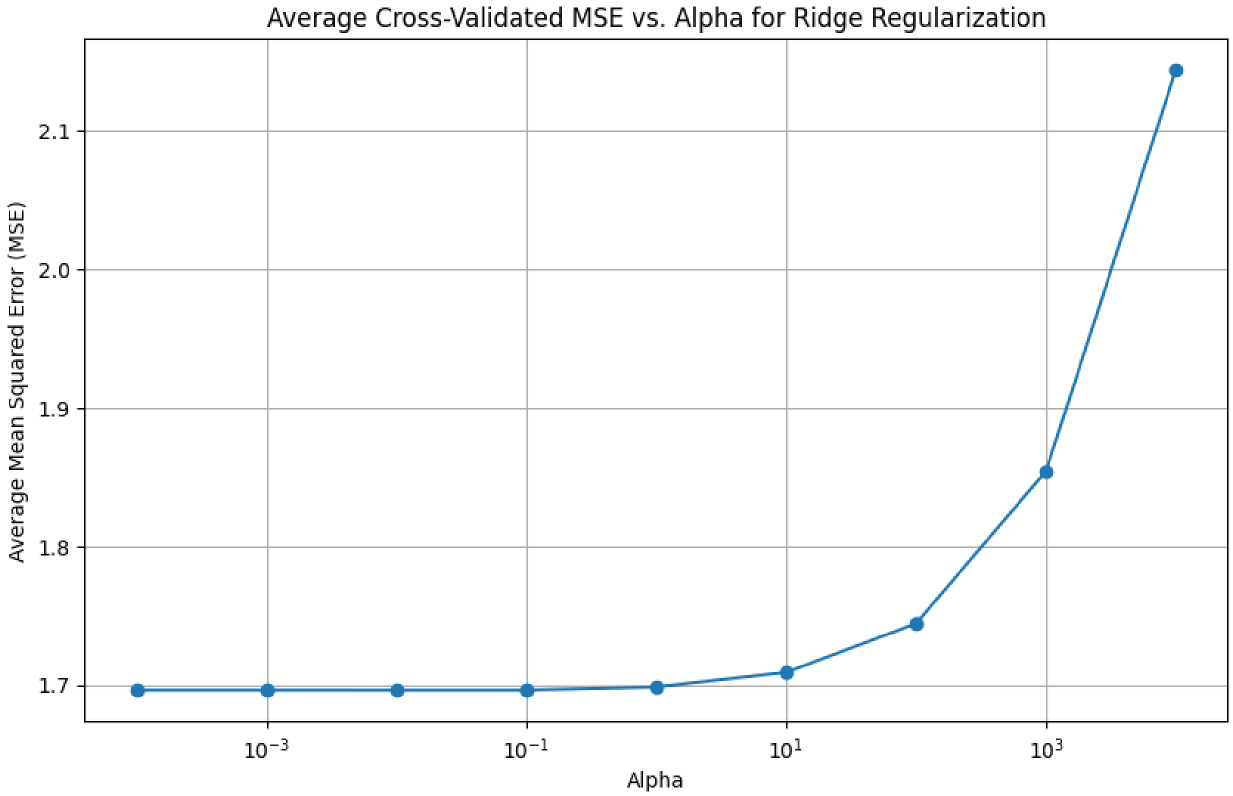


The average MSE generally decreases as alpha increases from 10^{-4} to 10^{-1}. This suggests that lower values of alpha result in higher complexity models (less regularization), which can lead to overfitting and higher MSE. As alpha increases beyond 10^{-1}, the average MSE starts to increase. This is because higher alpha values lead to stronger regularization, which can underfit the data and result in higher MSE. Notice that from alpha values of 10^{1} to 10^{4}, the average MSE remains relatively stable at roughly 9. This indicates that once the regularization becomes very strong (high alpha values), the model essentially becomes too simple and consistently produces higher MSE.

We can see from this plot, which shows the penalty term, against the average MSE across the 10-fold CV. A penalty term of .1 corresponds to a MSE of 2.31. This is already lower than our non-penalized multiple linear regression.

L2 Regularization:

Using the same alpha values, we then sough to look at if L2 regularization can have an impact.

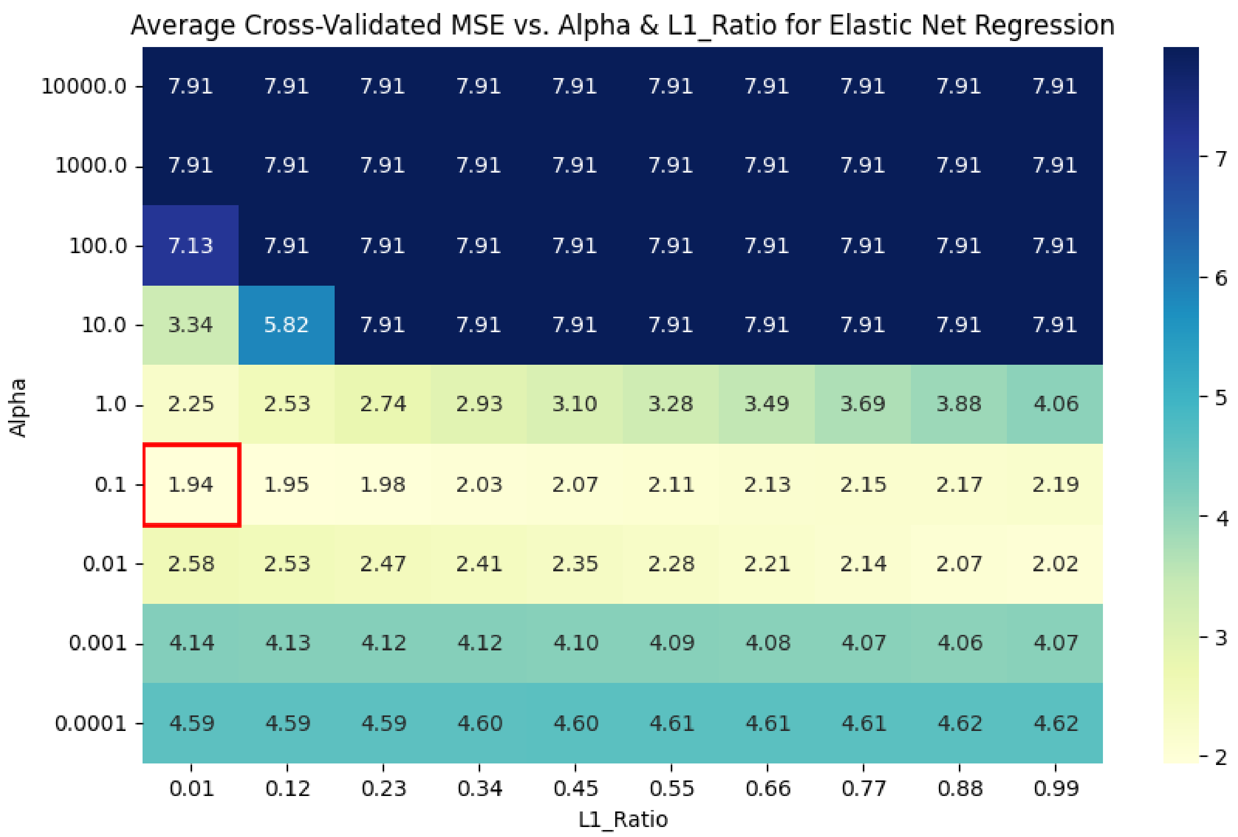


As alpha increased from 10^{-4} to 10^{1}, we observed a relatively stable range of average MSE, with values hovering approximately at 1.7. This suggests that for these alpha values, the level of regularization applied did not significantly affect the model's performance, resulting in consistent MSE. As we further increased alpha, starting from 10^{1} and beyond, a noticeable shift occurred. The average MSE began to rise, indicating stronger regularization. This increase implies that the model, under the influence of higher alpha values, became overly simplified and unable to capture the underlying patterns in the data, leading to higher MSE.

We cannot really tell from the plot, but a penalty term of .1, the same found in L1 regularization, produces the best fit. That penalty term has a corresponding MSE of 1.69, lower than what we found in the multiple linear regression, and lower than the L1 regularization.

Elastic Net Regularization:

Elastic Net regularization incorporates another term, L1_Ratio. As explained above, Elastic Net uses a combination of L1 and L2 regularization, with a controllable variable of L1_Ratio. The same range of alpha's (penalty terms) were tested, and a linearly scaled array ranging from [.01,.99] was used as the L1_Ratio.



Alpha	0.01	0.12	0.23	0.34	0.45	0.55	0.66	0.77	0.88	0.99
10000.0	7.91	7.91	7.91	7.91	7.91	7.91	7.91	7.91	7.91	7.91
1000.0	7.91	7.91	7.91	7.91	7.91	7.91	7.91	7.91	7.91	7.91
100.0	7.13	7.91	7.91	7.91	7.91	7.91	7.91	7.91	7.91	7.91
10.0	3.34	5.82	7.91	7.91	7.91	7.91	7.91	7.91	7.91	7.91
1.0	2.25	2.53	2.74	2.93	3.10	3.28	3.49	3.69	3.88	4.06
0.1	1.94	1.95	1.98	2.03	2.07	2.11	2.13	2.15	2.17	2.19
0.01	2.58	2.53	2.47	2.41	2.35	2.28	2.21	2.14	2.07	2.02
0.001	4.14	4.13	4.12	4.12	4.10	4.09	4.08	4.07	4.06	4.07
0.0001	4.59	4.59	4.59	4.60	4.60	4.61	4.61	4.61	4.62	4.62

The provided heatmap shows the performance of an Elastic Net model under varying combinations of L1_Ratio and alpha parameters, with a focus on the resulting MSE.

As alpha decreases from 10^4 to 10^{-4}, there is a clear downward trend in the value of the MSE. This consistent decrease suggests that lower values of alpha, which correspond to weaker regularization, are associated with lower a lower MSE. In practical terms, this indicates that incorporating more substantial regularization into the model contributes to better predictive performance and helps mitigate overfitting.

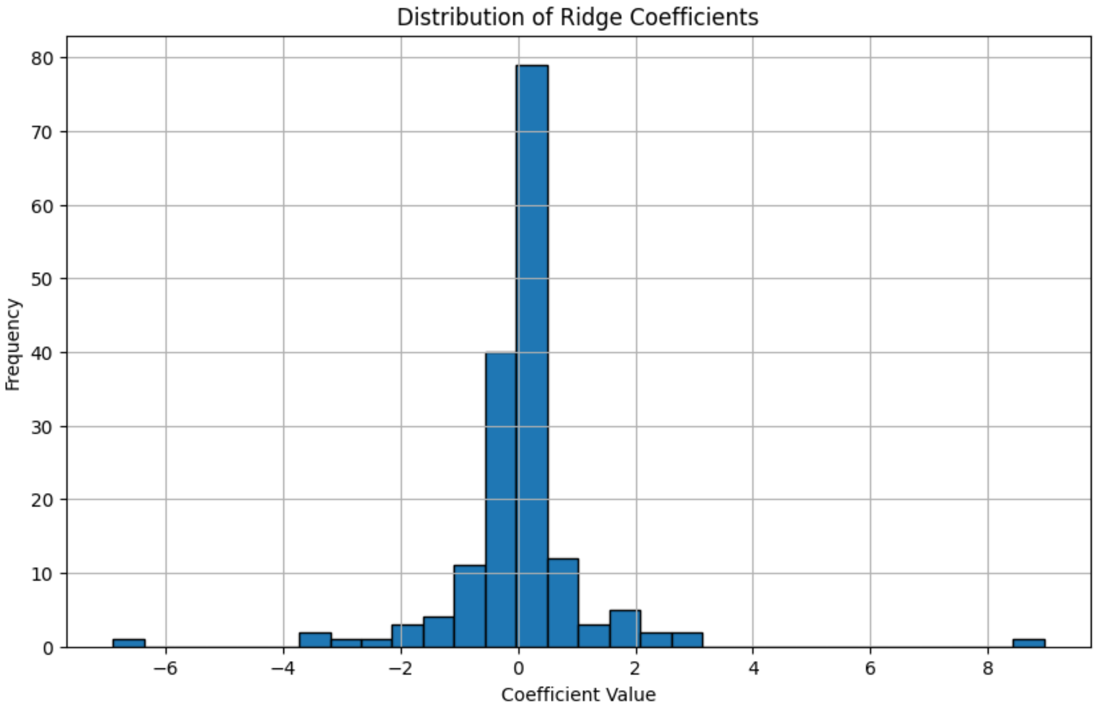
The heatmap also shows different values of the L1_Ratio, representing the mixing proportion of L1 (Lasso) and L2 (Ridge) regularization. Notice that for a given Alpha, there is little effect on the MSE, across different L1_Ratio values. This suggests that the choice of L1_Ratio appears to have a smaller impact on model performance compared to the magnitude of alpha.

It's essential to highlight the presence of an optimal combination of a penalty term of both L1 and L2 regularization, that yields the lowest MSE. The optimal combinations is represented by the red square highlight. An alpha of 10^{-1} as well as an L1 ratio of 0.01 resulted in the a MSE of 1.94. This is higher than Ridge (L2) regularization, but still lower than Lasso, and much lower than the initial linear regression model.

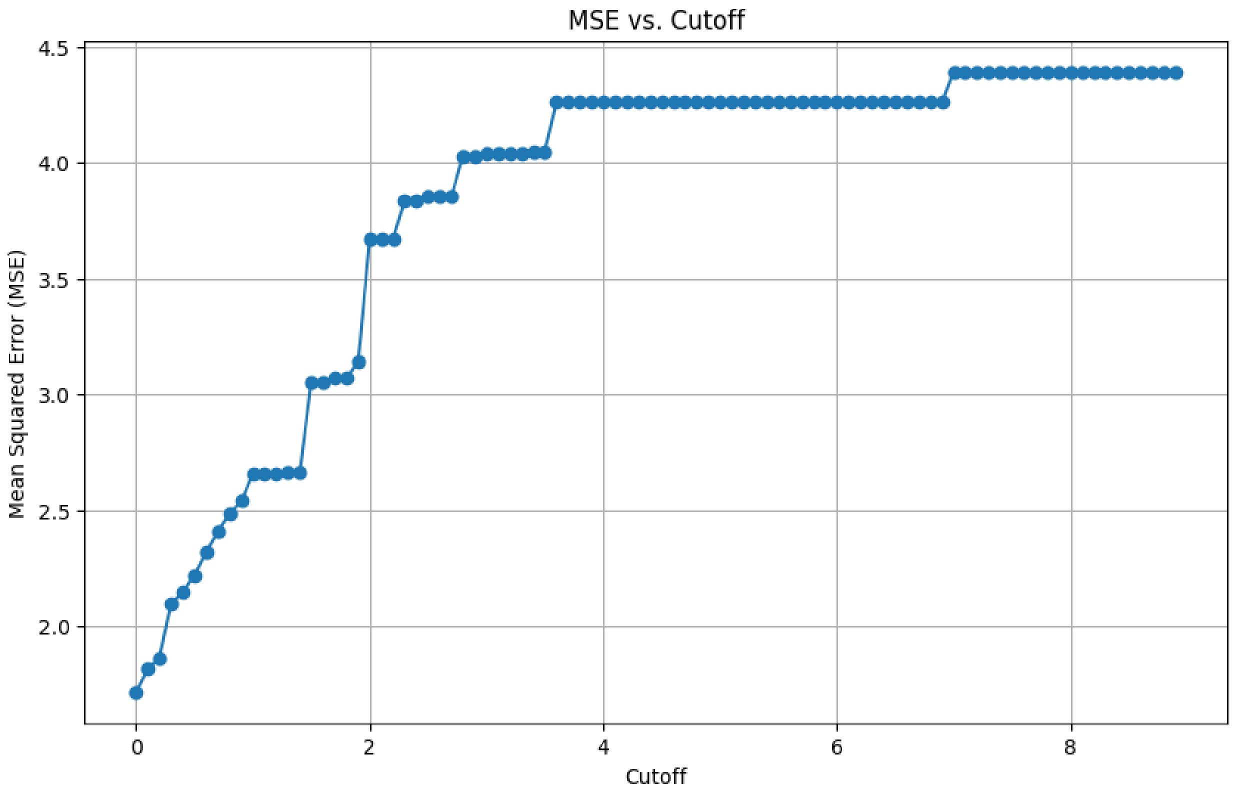
Feature Selection:

Feature selection with a Ridge regression model involves using the regularization properties of Ridge regression to automatically select the most important features. Ridge regression from the scikit-learn's library in Python, holds the attributes from the Ridge regression model stores the coefficients (weights) assigned to each feature in the model. These coefficients represent the contribution of each feature to the linear combination used for making predictions.

First we will filter the features that are weighted as 0, since they have no effect on the model. Next we will look at the distribution of weights, to see if there is a visual way to assess a cutoff metric for features that have little effect on the model.



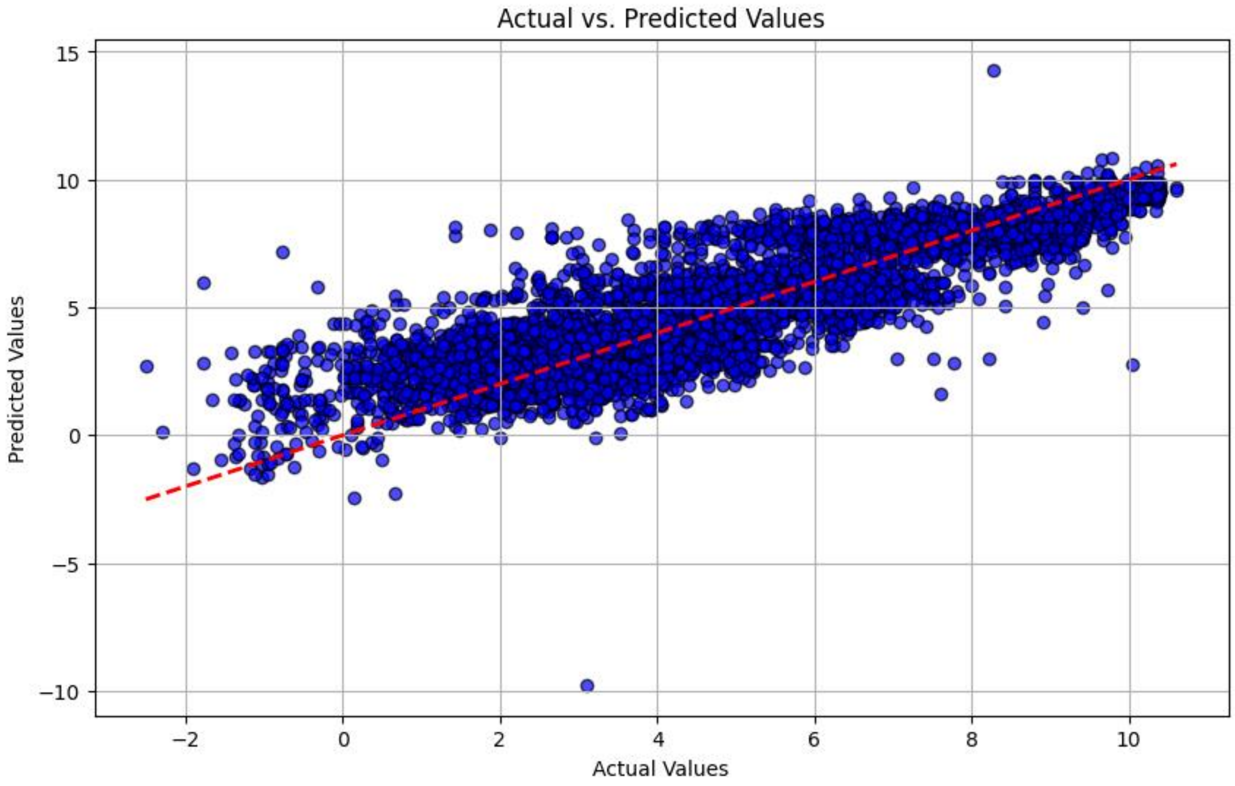
Looking at the magnitude of the absolute value of coefficients, we can see that there are 2 features that have the greatest effect on predicting critical temperature. Those features are, ‘wtd_mean_fie’ and ‘wtd_gmean_fie’ which is the mean and geometric mean of the first ionization energy of the elemental compound. Let’s see how adjusting the cutoff value, which would remove features whose absolute value is below that cutoff value, would affect the MSE.



It is apparent that increasing the cutoff value, or removing features, would have a detrimental effect on the model’s predictability performance. It seems that all features play a significant role in model accuracy.

Conclusion:

After testing these regularization techniques, we can conclude that applying a L2 penalty, of .1 achieves the best performing model with an MSE of 1.69. It must be remembered that there was a transformation applied to the response, so this MSE is relative to the box-cox transformed critical values.



This plot shows the performance of the model. There is still work that can be done, as the model overpredicts for values closer to 0, and underpredicts for large values.