Case Study 1:

Dylan Scott

In this assignment I was tasked with investigating critical temp of a list of chemical compounds. The goal would be to build a model that could predict the critical temperature of any new material presented to the model given its features provided. The data was provided in the form of two sheets. ‘unique\_m’ was a one hot encoded sheet noting all the elements that could be found within the compound. ‘train’ listed all the other features for the compound. After some initial investigation I found it best to not use ‘unique\_m’ and solely focus on the features provided in ‘train’.

Upon investigating the data I found that there was no null vales. In ‘train’ there was 21,263 number of values with 81 number of features. No categorical data found outside the ‘unique\_m’ document. As expected I went ahead and isolated my target variable to ‘crit temp’ and removed it from data set. Next, I standard scaled the data to help with normalizing my features. I did check for multicollinearity since they could lead to bias in our models. I used a threshold of .9 which I would consider to be high but in chemistry many features influence each other, and I didn’t want to remove all my columns. Starting with 81 columns I found 38 features to be colinear and removed them from the dataset:

Removed features:

Text

Description automatically generated

Correlation plot:

A picture containing chart

Description automatically generated

Keep in mind for all models K fold cross validation was used with 10 splits. For this case I ran a simple linear regression to give a baseline of MSE to compare my L1 and L2 to. The simple linear regression model got an MSE score of 357.67. I used this as a baseline to compare L1 and L2 with their different alpha values.

Alpha is a very important metric in L1 and L2 as picking the right alpha will help improve the model while also preventing many outcomes that would lead us to the wrong conclusion. For L1 a smaller alpha is expected as it suppresses the features that have a low coefficient while boosting the features that have a high coefficient. This is their absolute values that we are talking about because highly negative coefficients are also considered “high impact”.