Jonathan Sheppard

To predict critical temperature in the test data set, I built and fit three models on the training data set.

First, I obtained the training and formula\_training tables and combined them by row number so that my models could consider all possible predictors. I then standardized the X variables in the training set so they were all at the same scale.

Next, I split the training data into training and validation sets. This was done so that I could fit a multiple linear regression model to the training data and evaluate RMSE. I chose a test\_size of 20%, leaving 80% to train the model. From this multiple linear regression, I found a RMSE of 16.83.

Next, I fit a Ridge regression model on the training data to try and minimize many of the predictors and achieve a lower RMSE on the validation set. I tested 100 possible values of lambda, the tuning parameter, ranging from 0 to 1 by 0.01 increments. I used K-Fold cross-validation to compare the different values and determine the best value for lambda. Then, I fit a Ridge regression using that value of lambda, predicted critical temperature on the validation set, and compared it to the actual critical temperatures in the validation set, achieving a RMSE of 16.82.

Finally, I fit a Lasso model on the training data in the hopes that zeroing out predictors would produce a lower RMSE. I tested 100 possible values of lambda, ranging from 1e-5 to 1e5, on a logarithmic scale, as I read that was the preferred lambda range for Lasso. I used repeated K-Fold cross-validation to compare the different lambdas, similar to what I did for Ridge. Then, I fit the model with the best value of lambda, trained it on the training data, and used that to predict critical temperature on the validation set. I achieved a RMSE value of 17.81, higher than both multiple linear regression and Ridge.

Because the lowest RMSE that I found was due to my Ridge model, I used that to predict 1000 values of critical temperature on the actual test data, using the test predictors scaled to the training data.

I did not attempt Principle Components Regression or Partial Least Squares, as Ridge and Lasso seemed to be the preferred methods in the literature generally for dealing with problems such as this one.