Write-Up:

To include the chemical formulas in the training, I concatenated this dataset with the training data, and dropped the categorical features (material). Since this was only one field, I decided to drop it, but an alternative method would be to one-hot encode it to make it numeric.

Based on the examples given in lecture, I chose to evaluate the data on a Linear Regression Model and on the LASSO model. In addition to these, I also ran the data on scikitlearns RanfomForestRegressor and SVM (Support Vector Machine) models to see their performance.

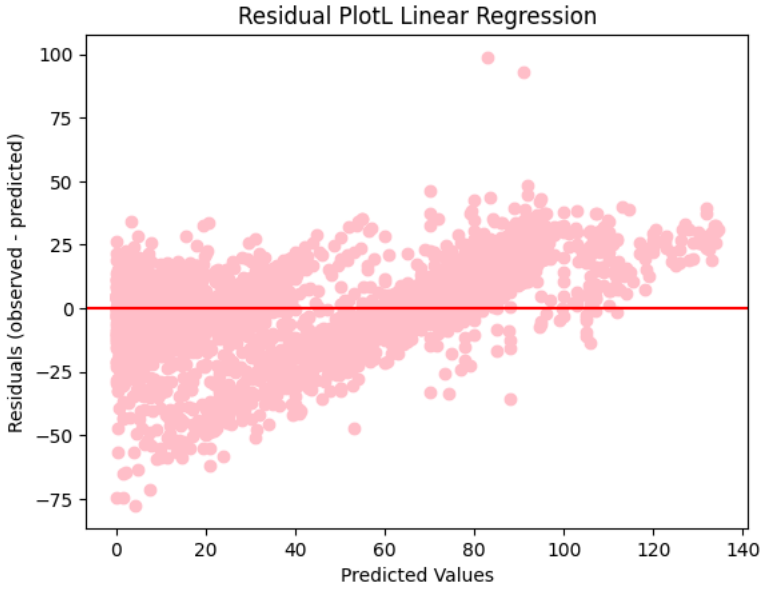
Motivation for models and their analysis:

Linear Regression:
Motivation:

- Serves well as a baseline model and I chose to have it act as a benchmark for the other models.

Performance:

- The RMSE for Linear Regression was 16.658, which I used as a benchmark for the other models.



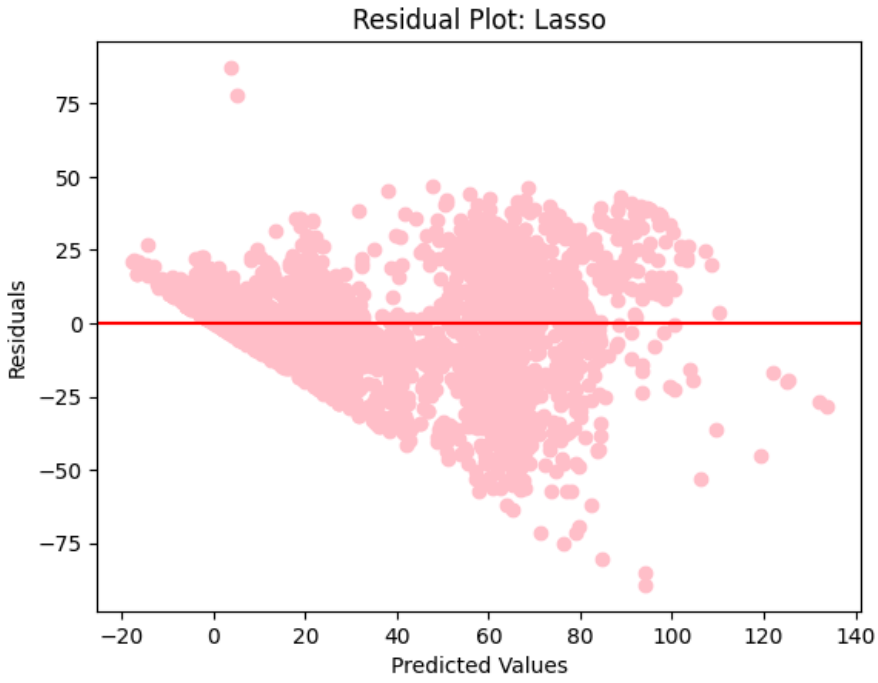
LASSO (Least Absolute Shrinkage and Selection Operator):

Motivation:

- LASSO can handle multicollinearity which may be present in this data. If only a few predictive features matter, LASSO is expected to perform well.

Performance:

- LASSO had an RMSE of 17.4656, which was surprisingly worse than that of Linear Regressions.



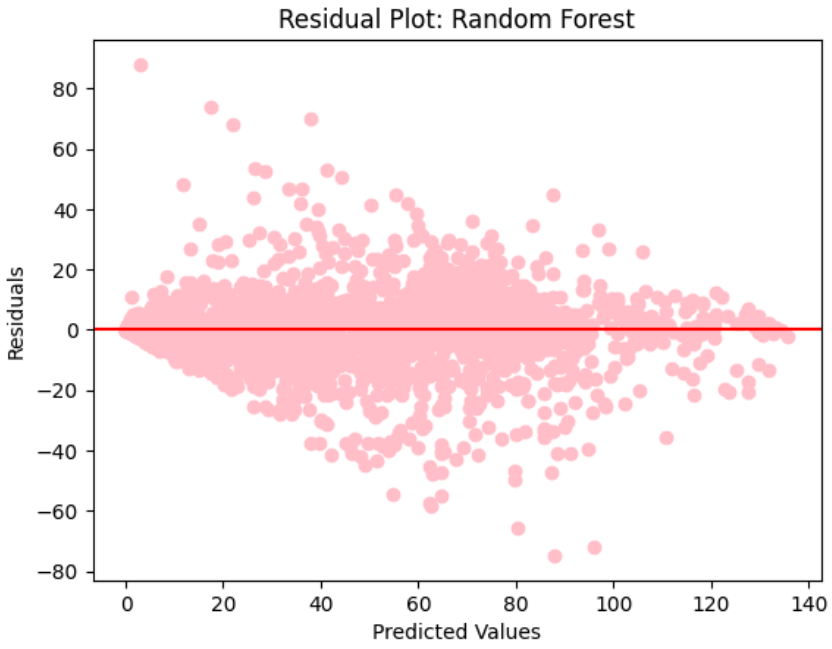
Random Forest

Motivation:

- I chose to test the Random Forest Regression model because of its ability to handle large datasets with nonlinear relationships. RF is also robust to overfitting, which is beneficial since we ultimately want to run the model on the unlabeled dataset.

Performance:

- The RMSE for Random Forest was 9.6452, which made it my best performing model.



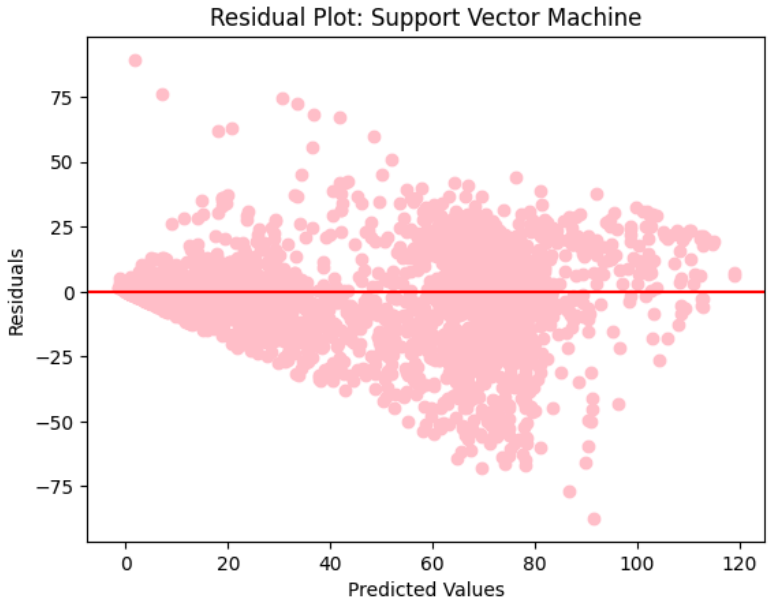
SVM

Motivation:

- Since the Random Forest performed well, I wanted to observe how the SVM model would handle the data. SVM models are known to perform well on high-dimensional spaces, so I thought it would be a fitting model for our data, though the scale and noise in the data may have hindered its performance.

Performance:

- The RMSE of SVM was 15.602 – better than linear regression and LASSO, but not as good as Random Forest.



Conclusion:

The Random forest ended up having the best RMSE score, so these predicted values (run on the original, unlabeled test set) were converted to a CSV.