Predicting the Critical Temperature of Superconductors

R Tumber

18/03/2021

Introduction

The object of this exercise was to predict the critical temperature of superconductors from data derived from the structure and elemental properties of a list of superconductors obtained at https://supercon.nims.go.jp/en/.

A brief definition of the data used is given below, a full explanation of the method used to calculate these feature values can be found in the original paper from which the data has been taken https://www.arxiv.org/pdf/1803.10260.pdf

Variable	Units	Description
Atomic Mass	atomic mass units (AMU)	total proton and neutron rest masses
First Ionization Energy	kilo-Joules per mole (kJ/mol)	energy required to remove a valence electron
Atomic Radius	picometer (pm)	calculated atomic radius
Density	kilograms per meters cubed (kg/m3)	density at standard temperature and pressure
Electron Affinity	kilo-Joules per mole (kJ/mol)	energy required to add an electron to a neutral atom
Fusion Heat	kilo-Joules per mole (kJ/mol)	energy to change from solid to liquid without temperature change
Thermal Conductivity	watts per meter-Kelvin (W/(m $\times$ K))	thermal conductivity coefficient k
Valence	no units	typical number of chemical bonds formed by the element

The full dataset consisted of two files, one with figures calculated from the elemental properties shown in the above table against superconductor critical temperature, and the other containing details on elemental composition against critical temperature.

For this project, just the first table was used to train a model and make predictions for critical temperature, and once again a brief summary of the calculations used in obtaining the figures is given below, for the superconductor Re_7Zr_1.

Feature & Description	Formula	Sample Value
Mean	= μ = (t1 + t2)/2	35.5
Weighted mean	= v = (p1 t1) + (p2 t2)	44.43
Geometric mean	= (t1 t2)1/2	33.23
Weighted geometric mean	= (t1)p1 (t2)p2	43.21
Entropy	= -w1 ln(w1) - w2 ln(w2)	0.63
Weighted entropy	= -A ln(A) - B ln(B)	0.26
Range	= t1 - t2 (t1 > t2)	25
Weighted range	= p1 t 1 - p2 t 2	37.86
Standard deviation	= [(1/2)((t1 - μ)2 + (t2 - μ)2)]1/2	12.5
Weighted standard deviation	= [p1 (t1 - v)2 + p2 (t2 - v)2)]1/2	8.75

t1 & t2 refer to the thermal conductivities of the two elements. p1 & p2 are the ratios of the two elements. w1 & w2 are fractions of thermal conductivities, t1/(t1+t2) & t2/(t1+t2). A & B are intermediate values calculated from components p & w. A = p1w1/(p1w1+p2w2) & B = p2w2/(p1w1+p2w2).

The dataset contained the figures for 21263 different superconductors, each with 81 features and the critical temperature, and had already been cleaned.

The effect of each feature on critical temperature was checked by first looking at the correlation and then each feature was plotted against critical temperature.

A variety of model types were initially tested using caret against the features that showed greatest correlation and the worst performing model types discarded. In addition to models built using caret, a principal component analysis was performed and a model built using xgBoost A model built using caret and Ranger produced the lowest RMSE for critical temperature predictions on the full dataset for the caret-built models.

When these models were used to make predictions on the test set, the ranger model performed better than the xgBoosted model. On closer examination the ranger model appeared to be more accurate at critical temperatures close to zero.

Method and Analysis

Downloading, Importing and Investigating the Dataset

The zip containing the dataset was downloaded and extracted from the zip file available at the URL given on the UCI Machine Learning Laboratory website. Initially both tables were imported and examined to get a rough of idea of the layout and determine further work that would be required to get the data in a state in which it would be useful.

Tables containing the data definitions, as detailed in the introduction, were scraped from the pdf from which the dataset was derived using Tabulizer and some basic formatting and encoding applied.

Moving forward, further work was restricted to the data table containing the figures calculated from the elemental properties since the data in this is used to build the prediction model.

The table contained 81 features (predictors), the prediction target, critical temperature and 21263 records. Within the data table the features are mostly numeric with three integers and no missing data. (Given the large number of features in this dataset and limited options for pdf output some of the larger tables have been omitted for clarity)

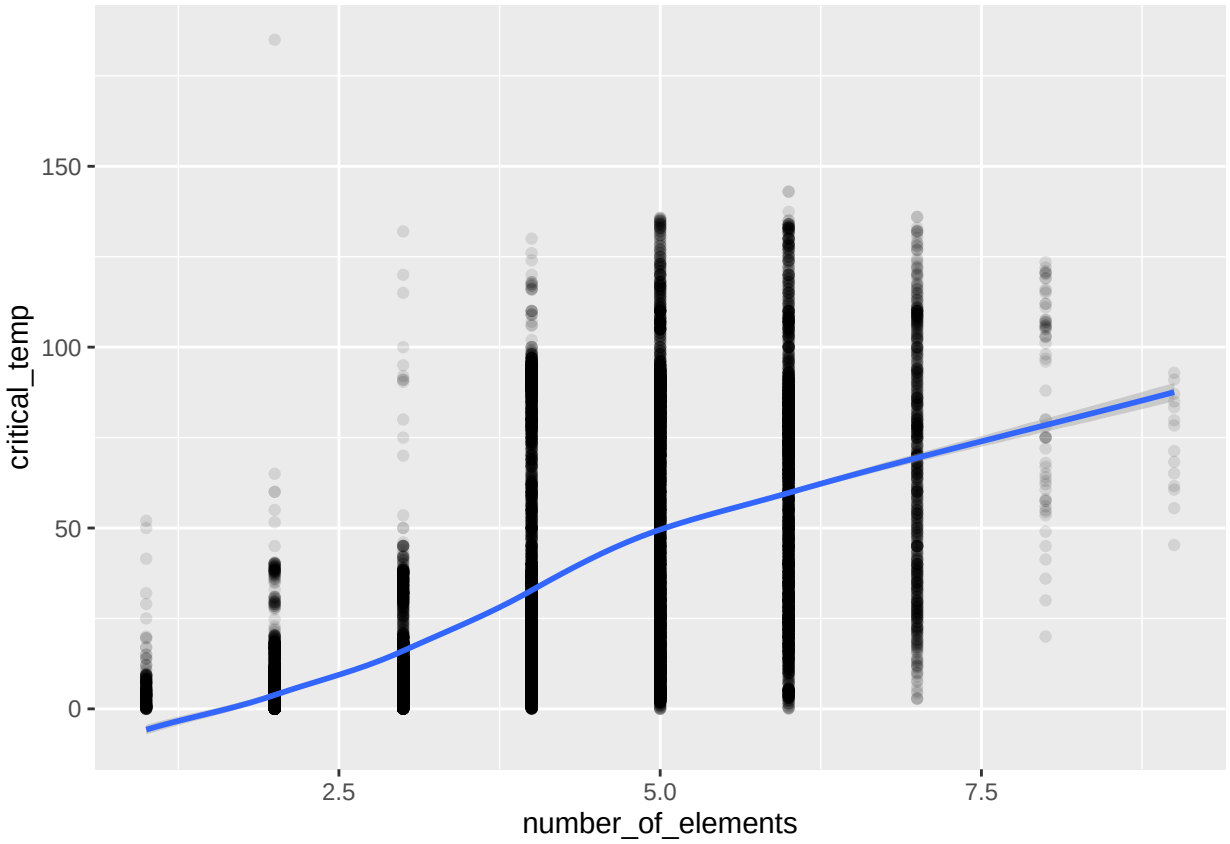
Data Exploration

The correlation coefficient for each feature with critical temperature was recorded which resulted in a range of values between those shown below

	correlation
wtd_std_ThermalConductivity	0.7212711
wtd_mean_Valence	-0.6324010

Following this each feature was plotted against critical temperature to determine if there was anything that help describe the nature of any relationships present.

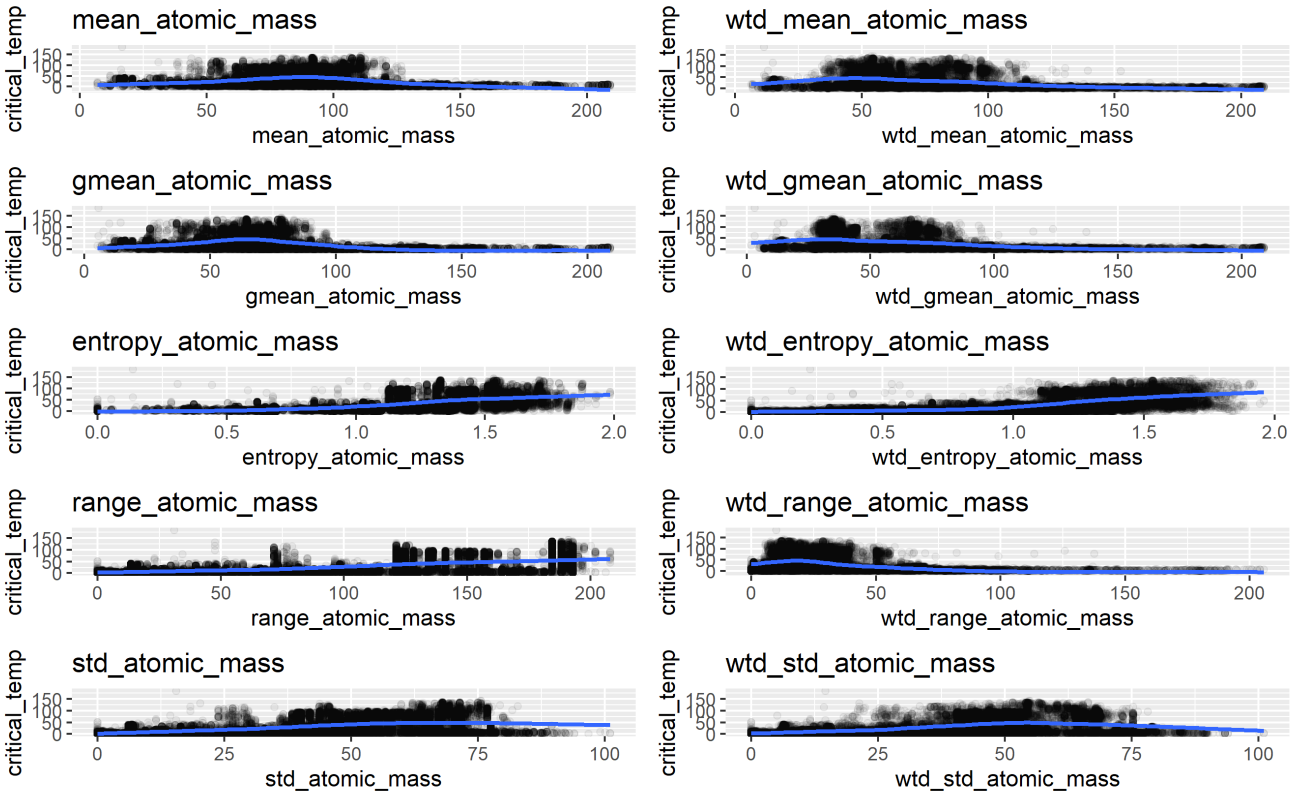
Number of Elements



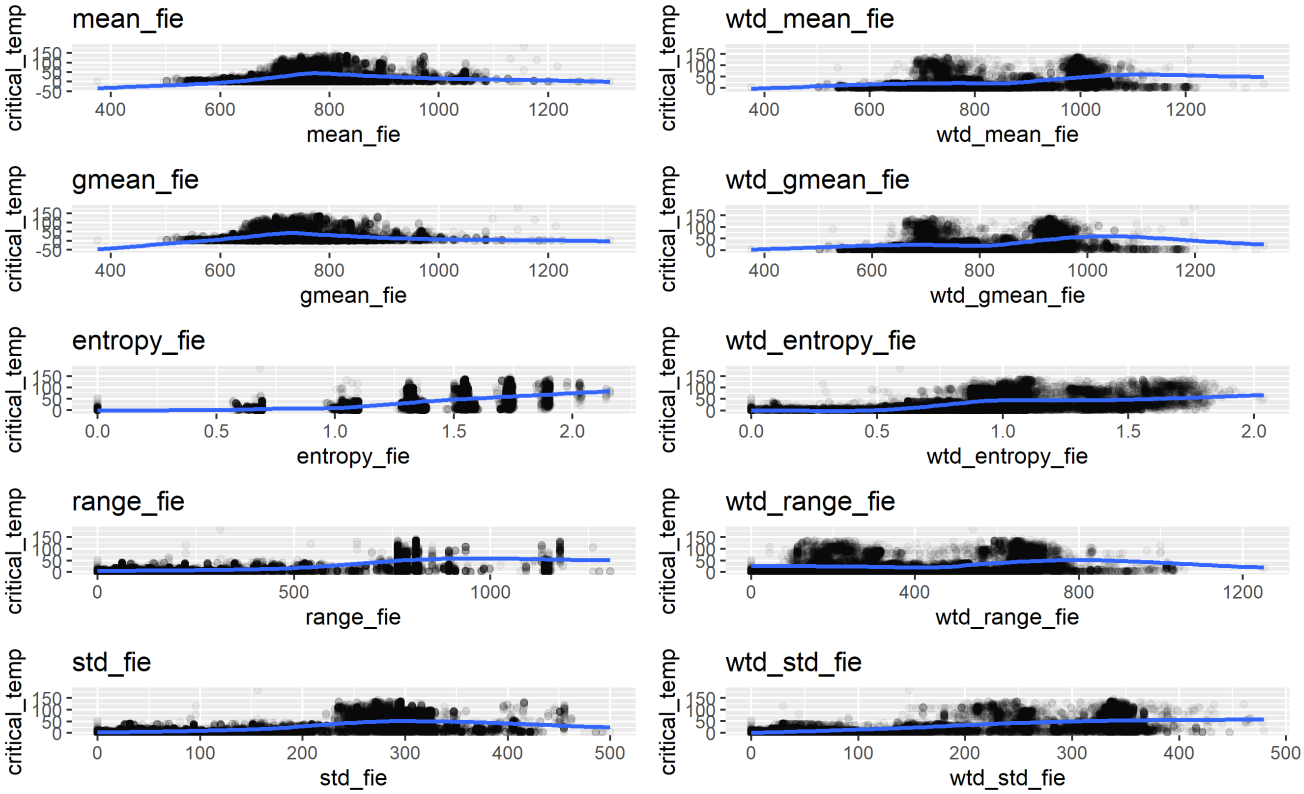
The increase in critical temperature with number of elements may partly be a reflection of the effect of valence and element size on the energy gap between vacant molecular orbitals, however the wide range of critical temperatures for superconductors with any number of elements suggest it may not be appropriate to say there is a definite relationship.

In respect of the features that follow there was little that could be determined that would describe the relationship between them and the critical temperature that would be unexpected when the electronic configuration of the elements that make up the superconductor is taken into consideration.

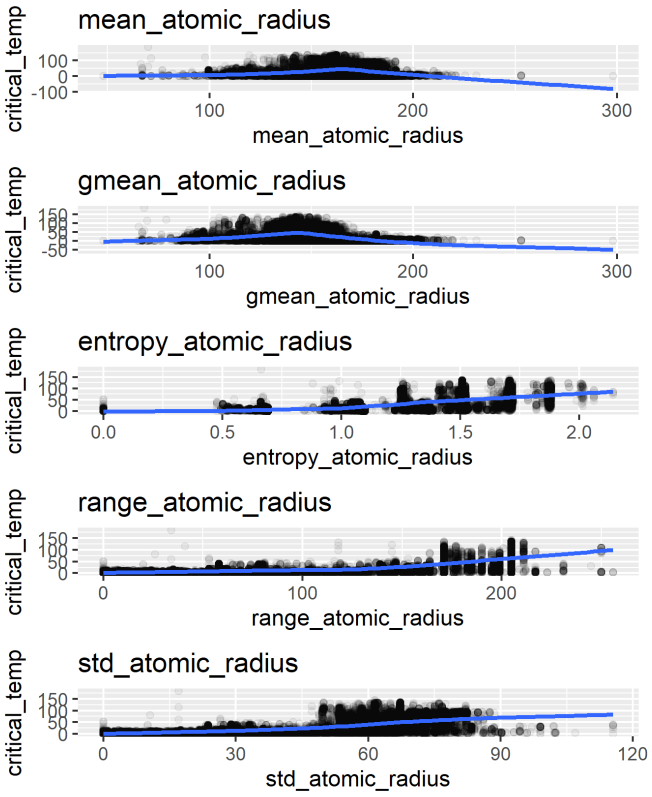
Atomic Mass

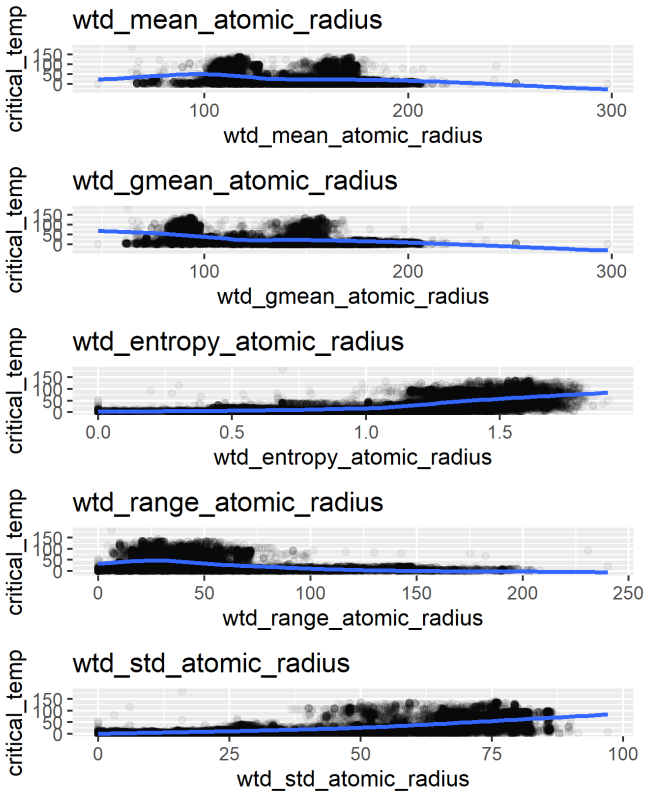


First Ionisation Energy

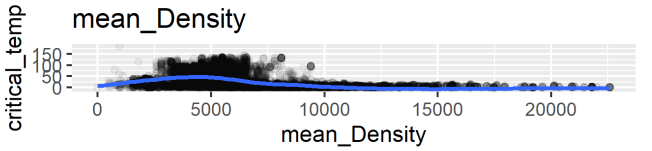


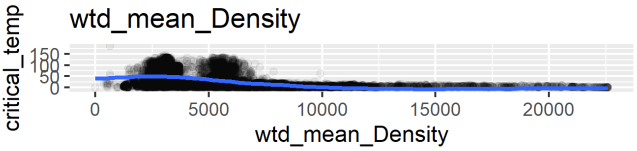
Atomic Radius

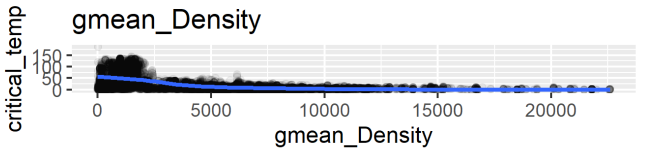


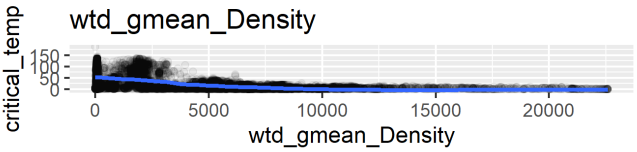


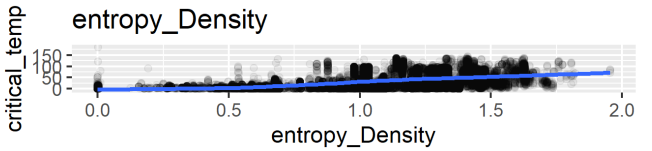
Density

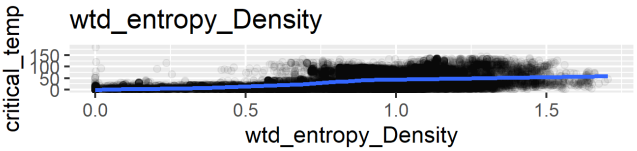


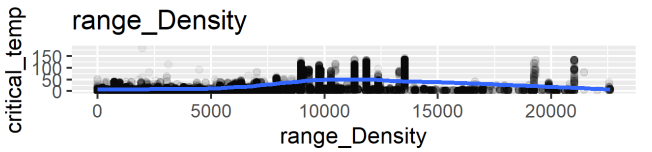


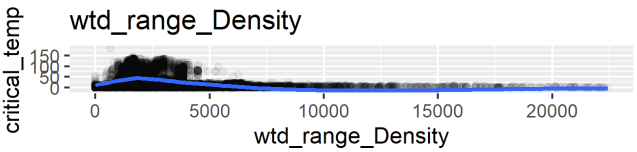


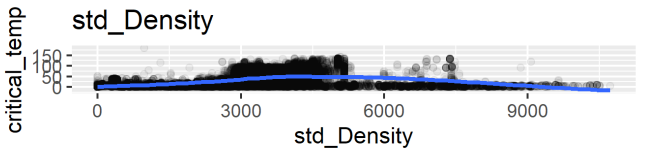


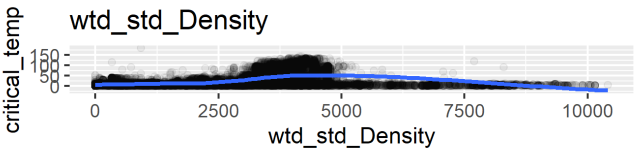




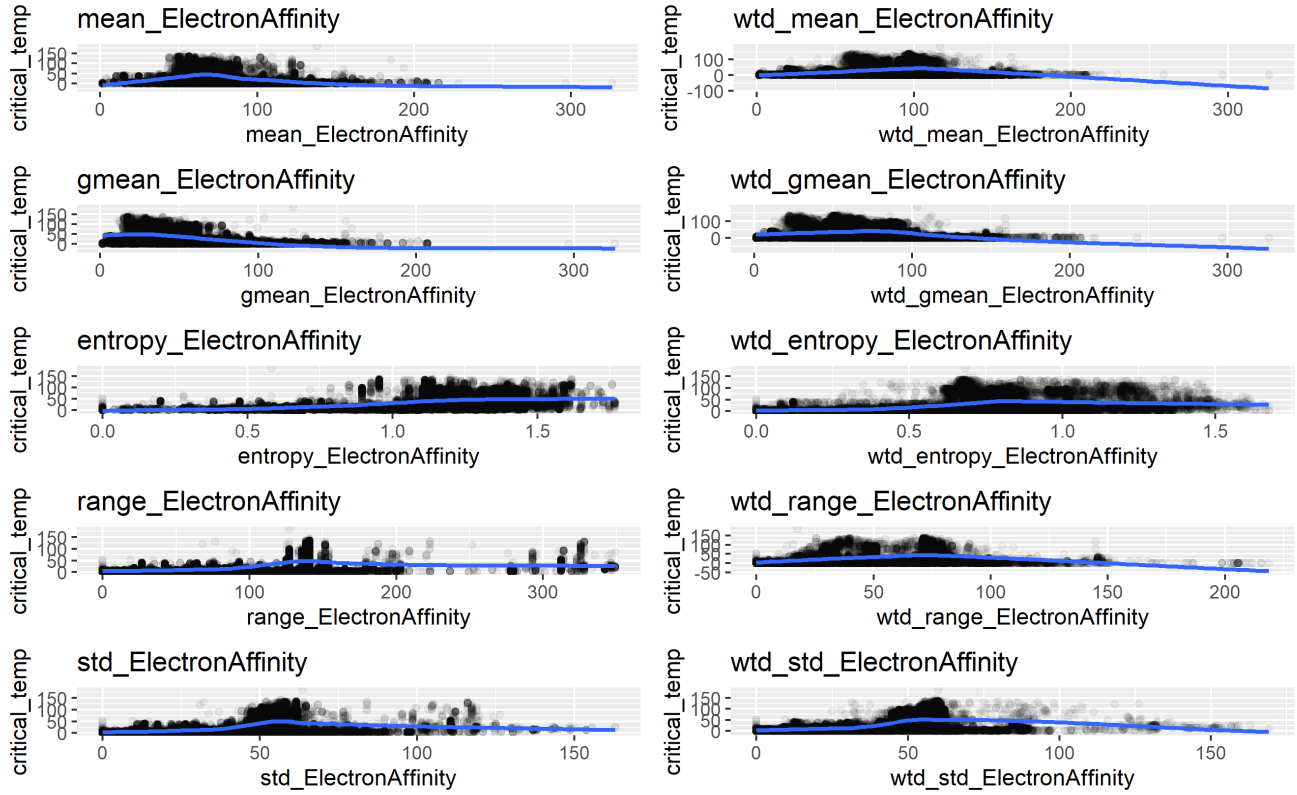




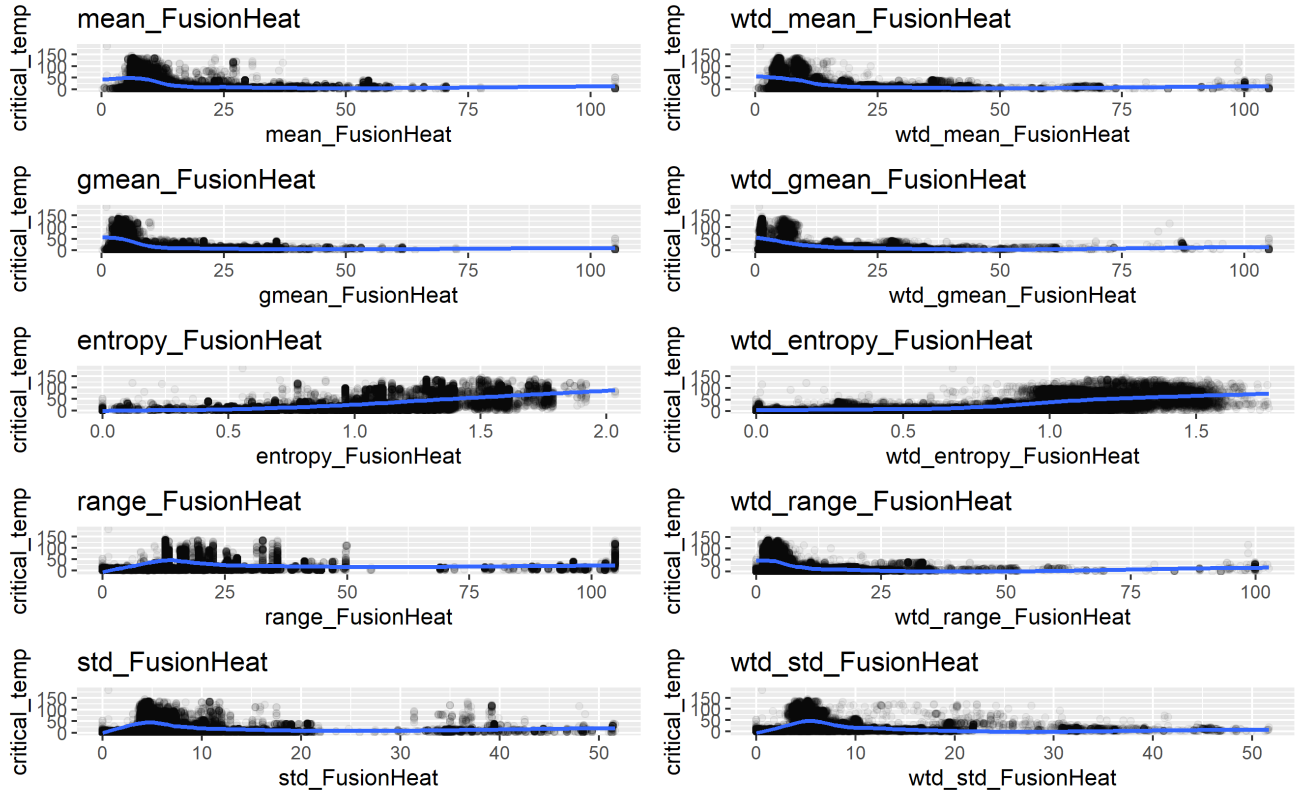




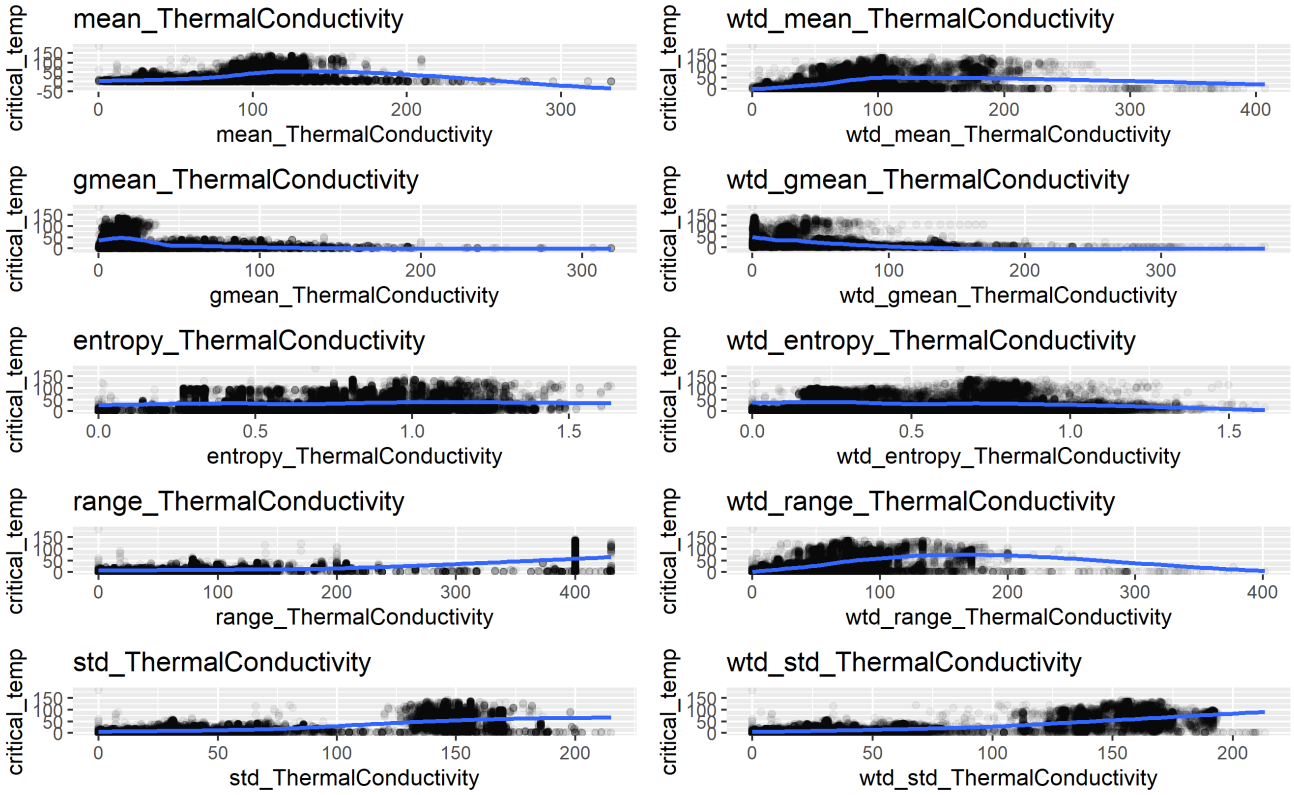
Electron Affinity



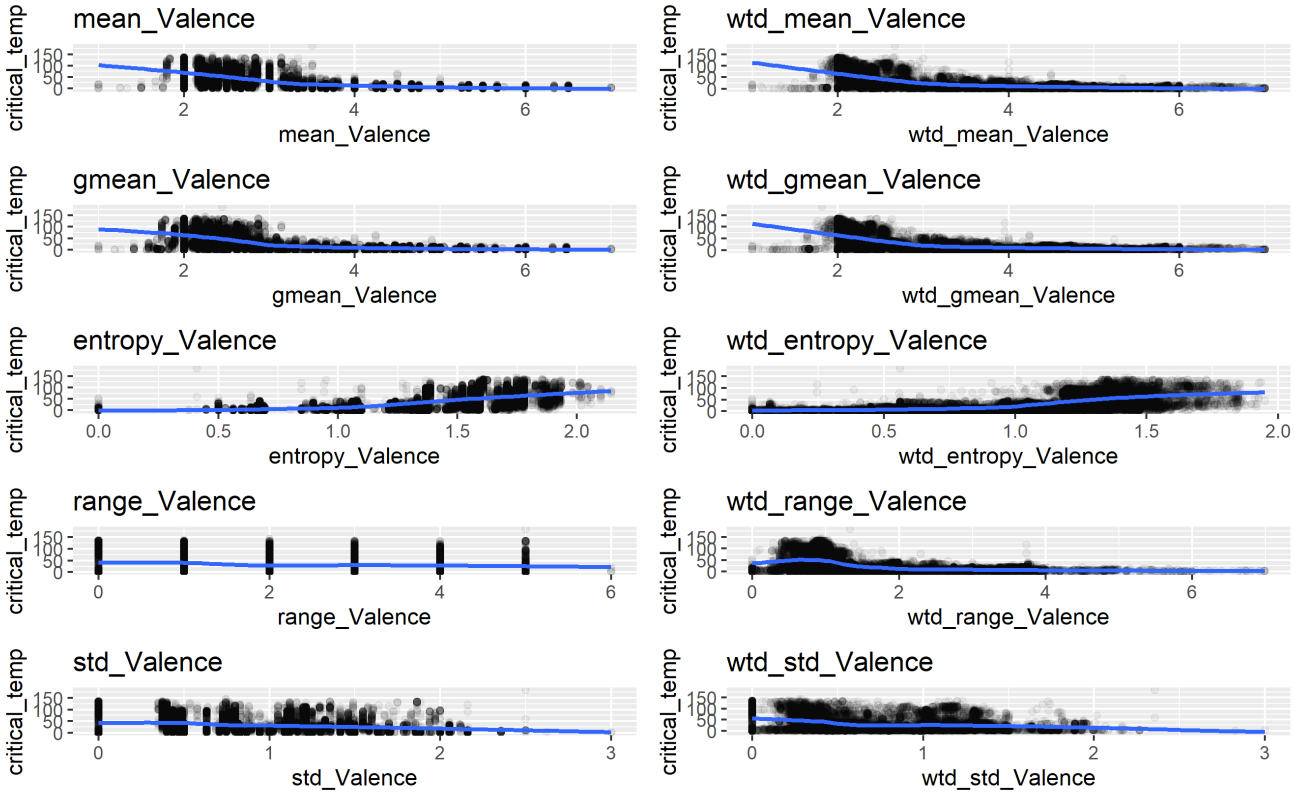
Fusion Heat



Thermal Conductivity



Valence



Since there was little insight to be gained in plotting the features against the critical temperature and no indication

of the nature of the relationships between them was observed, models were constructed using caret. Initially a range of modeling types were employed, linear and non linear, to determine the best approach before focusing on the more successful.

Data Modelling

The feature dataset was split into training and testing sets, with 75% used for training and 25% used for testing the final model.

Caret was used to train a model predicting critical temperature from values of *wtd_std_ThermalConductivity*, this being the feature that showed greatest correlation with critical temperature. The following selection of linear regression model types were first employed: lm; BstLm; gamLoess; glmboost; knn; svmLinear and 10 fold cross validation performed as part of the training RMSE was recorded for these models and the process repeated for non-linear model types: rf; gamboost; bagEarth; brnn; xgbTree.

To ensure the relative accuracy of the models were sound, a second feature, range_ThermalConductivity, was added and the above processes repeated. Initially model types: knn; random forest and tree based models performed best so these model types were retained, with other models of the same type added and training performed again using the three features with the greatest correlation. Parallel processing was employed to expedite the process. The accuracy of all models increased as features were added however eventually it was clear random forest methods were producing the more accurate results so modelling continued using rf, ranger and Rborist models.

While RMSE continued to drop, processing time increased and improvements in accuracy decreased, since boosted models showed some early success construction of dmatrix and modelling using xgboost was performed, resulting in a significant improvement in RMSE. In addition to this, principal component analysis was performed to see if this could produce a more accurate result without success.

Modelling continued using the more accurate of the random forest models, ranger, with further features to be added until RMSE either plateaued or began to increase. Neither of these scenarios occurred and RMSE continued to decrease until all the fgeatures were includeed in the model.

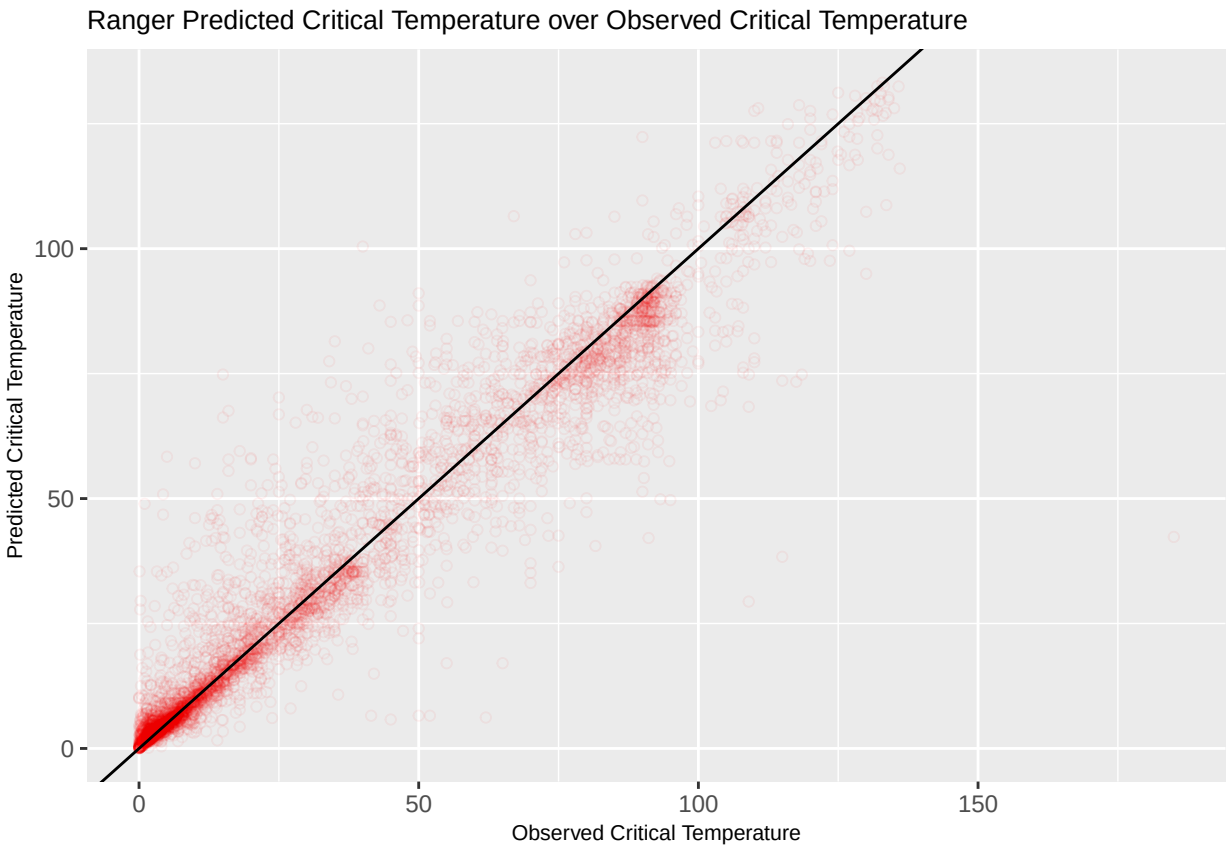
The final RMSE for the ranger model was larger than that of for the xgBoosted model however it was close enough to warrant making predictions on the test set for both models. The result of this was an RMSE for the ranger model lower than that for the xgBoosted model, despite the latter being more accurate in training.

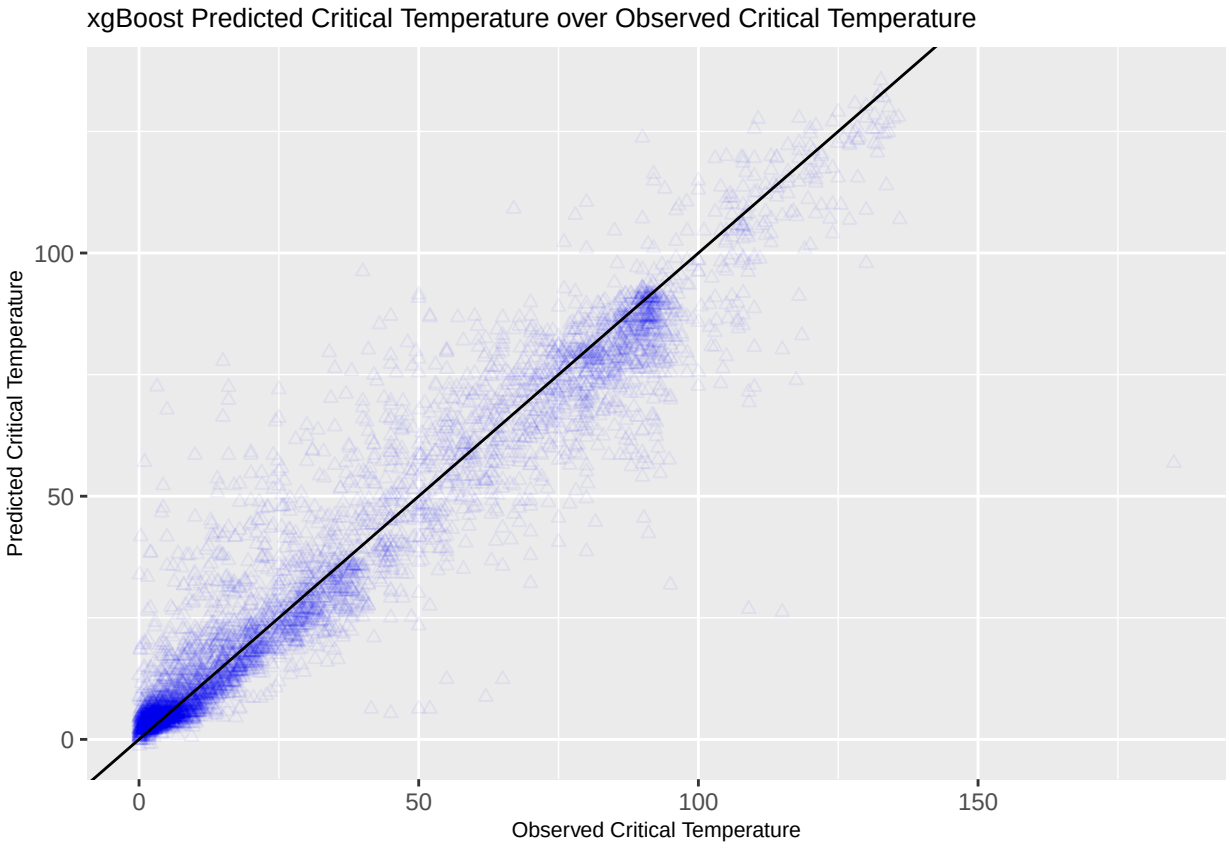
Results and Discussion

The below RMSE values are those calculated after making predictions on the test set with each model. We see the ranger model performed better and the RMSE for the xgBoost model is significantly worse than that obtained during training.

Model Name	RMSE
ranger	9.223133
xgBoost	9.700497

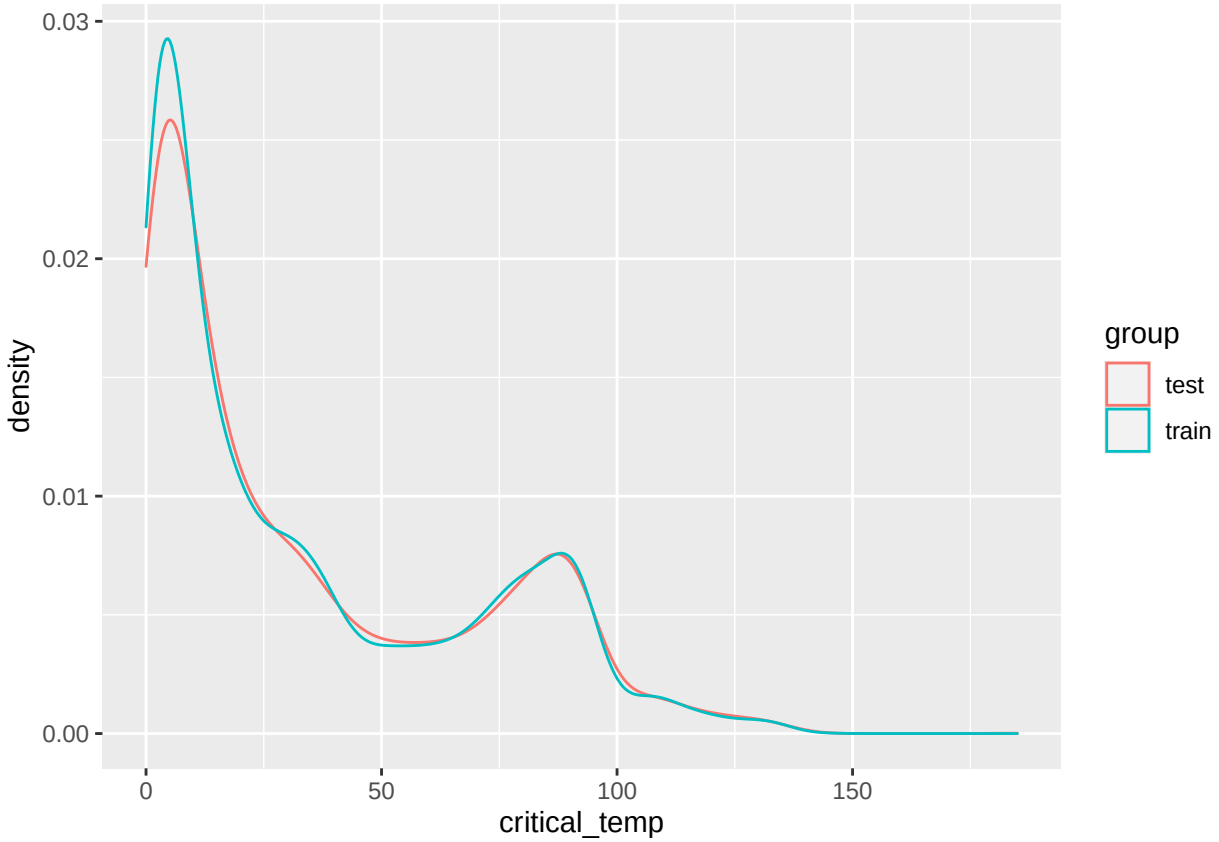
Predicted critical temperature is plotted against observed critical temperature for each model to see if critical temperature magnitude was a factor in model accuracy. The below plots demonstrate the ranger model is more accurate at lower critical temperatures.





This observation could lead to a possible explanation for the difference in RMSE between training and testing sets

for the xgBoost model. If there is a greater proportion of superconductors with a low critical temperature in the testing set than in training the inaccuracy will likely be exaggerated.



Looking at the above density plots for the training and test sets it may be more reasonable to conclude the Boosted model has been overtrained.

Conclusion

A model was created that predicted the critical temperature of a superconductor with an RMSE of 9.22K using ranger and the features provided in the *feature_data* dataset.
However, there is reason to believe there is room to improve on this figure by retraining the xgBoosted model, adjusting the hyperparameters.
In addition to this, there is another set of data that was not touched during this brief analysis, the *superconductor_ref* table describes the elemental composition of superconductors with critical temperature, which may facilitate some form of prediction based solely on the composition or in conjunction with the features used in this model.
A further extension of this would be to look at the crystal structure in relation to the elemental composition and the critical temperature with a view to either determining if certain structures or structure types are energetically more favorable.

References

Dataset URL https://archive.ics.uci.edu/ml/machine-learning-databases/00464/superconduct.zip

Hamidieh, Kam, A data-driven statistical model for predicting the critical temperature of a superconductor, Computational Materials Science, Volume 154, November 2018, Pages 346-354 **Web Link**