Superconducting Transition Temperature Prediction from Chemical Formula and Elemental Properties

Noravee Kanchanavatee

Findings

- Test root-mean-squared error (rmse) = 9.49 based on XGBoost model using recursive feature elimination to narrow the number of the features down from 81 to 27.
- Most important features are range thermal conductivity, and range atomic radius
- Most important elemental properties are thermal conductivity, atomic radius, and valence

Introduction

Superconductors possess the remarkable ability to conduct electricity with zero resistance below a specific critical temperature T_c. This property opens doors to a multitude of applications, including efficient energy transmission and powerful magnets. However, there are two main obstacles that hinder the widespread adoption of superconductors: the difficulty of finding materials with high Tc and the lack of a comprehensive theory to predict T_c.

In this study an alternate approach was considered. Instead of calculating T_c based on experimental results or physical models, the transition temperature was predicted from chemical formula of superconducting materials and properties of the elements of that material. This method was first utilized

Property	Unit	Description
Atomic Mass	Da	Rest mass of an atom
First Ionization Energy	kJ/mol	Energy required to remove a valence electron
Atomic Radius	pm	Calculated atomic radius
Density	kg/m^3	Density at room temperature and ambient pressure
Electron Affinity	kJ/mol	Energy released on formation of anions
Fusion Heat	kJ/mol	Energy required to change from solid to liquid
Thermal Conductivity	W/m·K	Ability to conduct heat
Valence	No units	Typical number of chemical bonds formed by the element

Table 1: Properties of an element which are used for creating features to predict T_c.

by the following work: *K. Hamidieh, Computational Materials Science 154 (2018) 346–354*, in which the model gives out-of-sample predictions:±9.5 K based on root-mean-squared-error.

Methodology

The raw inputs were a list of chemical formula of superconducting materials and their T_c, and a list of elements and 8 selected properties. A complete list of properties, units, and definitions can be seen in Table 1. The data was then processed by combining the chemical formulas and properties in 10 different ways as described in Table 2.

Since there are 8 properties, the number of features after droping the chemical formula are (8)(10)+1 = 81, where the last feature are the number of elements in the material. As illustrated in Figure 1, Kendall correlation between features and target reveal a few potential important factors.

The processed data used in this work was taken from the aforementioned paper. After an outlier of superconductor, H_2S which only superconducts under extreme pressure, and around 60 duplicates were removed. The final input data contains 21,196 samples.

The data was split at random to 4/5 train and 1/5 test, then the train data was fitted with the following models:

Feature	Formula
Mean (μ)	$\frac{p_A+p_B}{2}$
Weighted mean (ν)	$w_A p_A + w_B p_B$
Geometric mean	$(p_A p_B)^{\frac{1}{2}}$
Weighted geometric mean	$p_A^{w_A} p_B^{w_B}$
Entropy	$-v_A \ln v_A - v_B \ln v_B$
Weighted entropy	$-z_A \ln z_A - z_B \ln z_B$
Range	$\lvert p_A - p_B \rvert$
Weighted range	$\lvert w_A p_A - w_B p_B \rvert$
Standard deviation	$\sqrt{\frac{(p_A-\mu)^2+(p_B-\mu)^2}{2}}$
Weighted standard devieation	$\sqrt{w_A(p_A-\nu)^2 + w_B(p_B-\nu)^2}$

Table 2: Summary of the procedure for feature extraction from material's chemical formula and properties. $p_{A,B}$ and $w_{A,B}$ are properties and weights of element A and B, respectively. $v_{A,B} = \frac{p_{A,B}}{p_A+p_B}$, and $z_{A,B} = \frac{w_{A,B}v_{A,B}}{w_A v_A + w_B v_B}$.

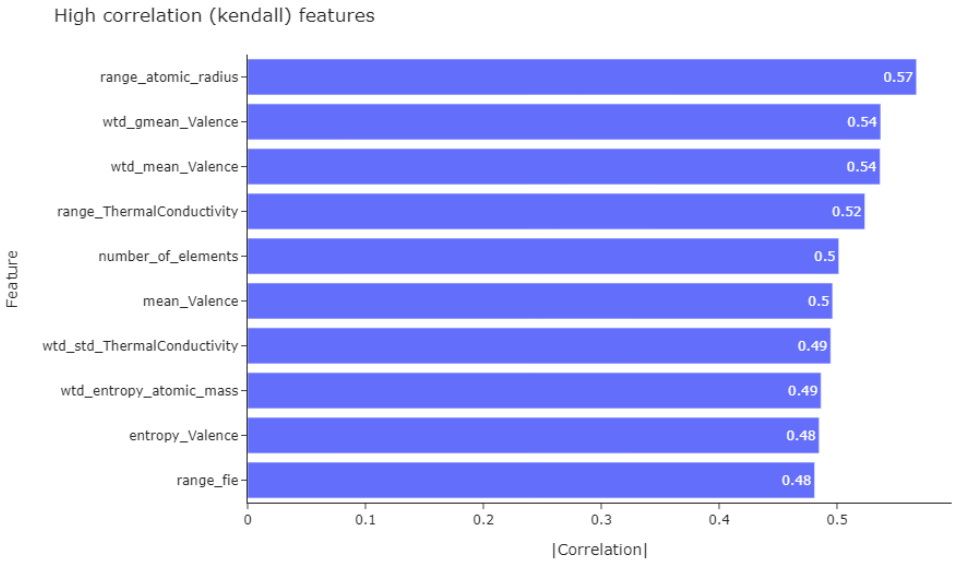

Figure 1: Top ten features with highest Kendall correlation with T_c

- Ordinary least square (ols) as a benchmark
- Regularization (lasso and ridge) to reduce noise
- XGBoost with squred error as the objective function
  - Since XGBoost has multiple hyperparameters, using grid search is too computational expensive. Thus, bayes hyperparameter optimization was utilized.
  - Reduce dimensions by principal component analysis (PCA) and choosing the principal component that has 99% cumulative explained variance.
  - Feature selection using f-regression, mutual info regression (mi), and recursive feature elimination (rfe).

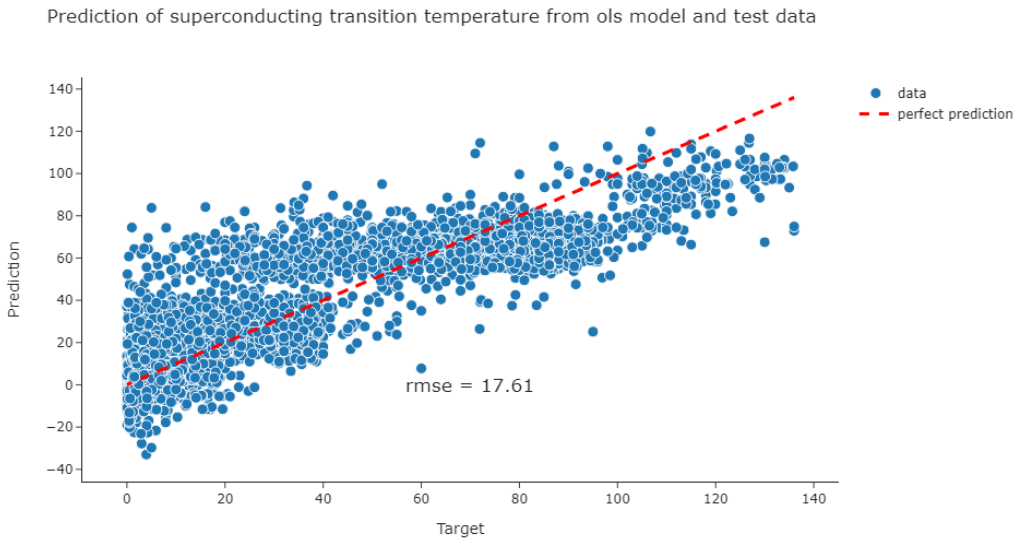

Figure 2: Prediction of T_c from ols model where the red dashed line indicates zero-error predictions

Results

Linear regression

The ols model's test rmse estimated by the procedure above is about 17.6 K while the R^2 is about 0.74. Figure 2 shows the predicted T_c (prediction) versus the observed T_c (target). The plot indicates that the ols model under-predicts T_c of high temperature superconductors and over-predicts low temperature superconductors' T_c. Regularization shows no improvement in cross validation score (CV score), and yield similare test results.

XGBoost

There are a number of tuning parameters that could enhance the model's prediction performance; They are mainly:

- n_estimators: the maximum number of boosting trees
- max_depth: The size of each new decision tree

n_estimators	max_depth	min_child_weight	subsample	colsample_bytree	gamma	lambda	alpha
223	13	1	0.69	0.46	0.57	0.94	0.62

Table 3: XGBoost hyperparameters that yield the highest CV score

- min_child_weight: The minimum number of observations that must be in the children after a split
- subsample: the ratio of data that is randomly selected for growing trees
- colsample_bytree: the subsample ratio of columns when constructing each tree
- gamma: minimum loss reduction required to make a further partition on a leaf node of the tree
- lambda: L2 regularization term on weights
- alpha: L1 regularization term on weights

After tuning these hyperparameters by bayes optimization based on the mean of 5-fold CV rmse scores, the hyperparameters that result in the highest score are shown in Table 3.

Since all 81 features come from only 8 properties, there are high correlations between features. PCA was performed to reduce the dimensionality of the data. As can be seen in Figure 3, 99% of cumulative explained variance ratio corresponds to principal component #30. Hence, the train data was transformed to 31 dimensions and fit with the optimized XGBoost, which yields CV rmse score of 10.3 K.

However, PCA makes it difficult to indicate which features and properties are important in predicting T_{c}. Three alternative feature selection methods were considered.

1. f-regresson: rank f-statisic of each feature and target, but only capture linear relations.
2. mutual info regression: measures the dependency between the variables and target based on entropy estimation from k-nearest neighbors distances, which is also make it sensitive to the chosen number of neighbors.

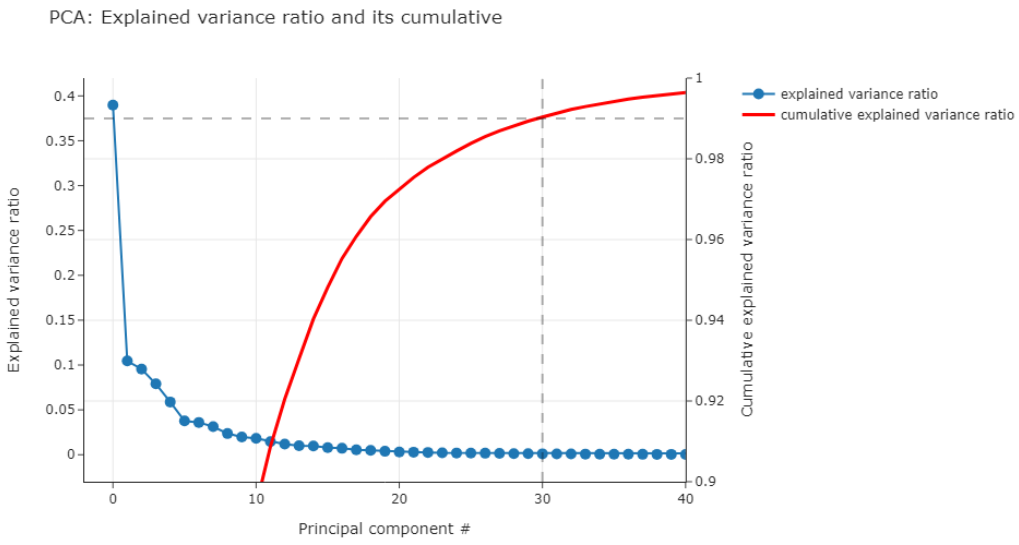

Figure 3: Explained variance ratio, cumulative explained variance ratio, and principal component. The horizon and vertical dashed lines indicate 99% cumulative explained variance ratio and principal component #30, respectively

Method	# of features k	CV rmse score
f-regression	28	9.67
mi regresson	26	9.68
rfe	27	9.46

Table 4: Number of selected features k and CV rmse score for each feature selection method.

3. rfe with XGBoost as estimator: starts with full set of features then recursively eliminate feature with the least imporntance, however using XGBoost for both selection and fitting can introduce bias towards features that are particularly relevant for XGBoost.

Based on CV scores around the elbow when plotted against number of selected features k, all three methods yield similar k and CV scores as shown in Table 4. Moreover, the two most significant features according to XGBoost's feature importance from either of these method are exactly the same, range of thermal conductivity and range of atomic radius. Figure 4 illustrates the average XGBoost's feature importance from 3 feature selection methods. Six of these features coincide with the top 10 features with highest Kendall correlation with the target in Figure 1. Additionally, it can be seen that the most important feature by far is range thermal conductivity, followed by range atomic radius while the most important properties are thermal conductivity, valence, and atomic radius. Note that many high temperature superconductors contain a combination of metal and non-metal, such as copper and oxygen, resulting in a large range thermal conductivity.

Using the model with best CV score (rfe) to predict T_c on test data give rmse = 9.49, which is slighly better than the previous work. The comparison between the predicted and observed values can be seen in Figure 5. It can be seen from the plot that, similar to ols, the XGBoost model still underestimates the critical temperature for high-temperature superconductors, while it overestimates it for low-temperature superconductors.

Thus, one of the potential improvement for the model is to first use classification to divide samples to low and high critical temperatures, then do separate regression for each one. Other possibilities include applying one-hot

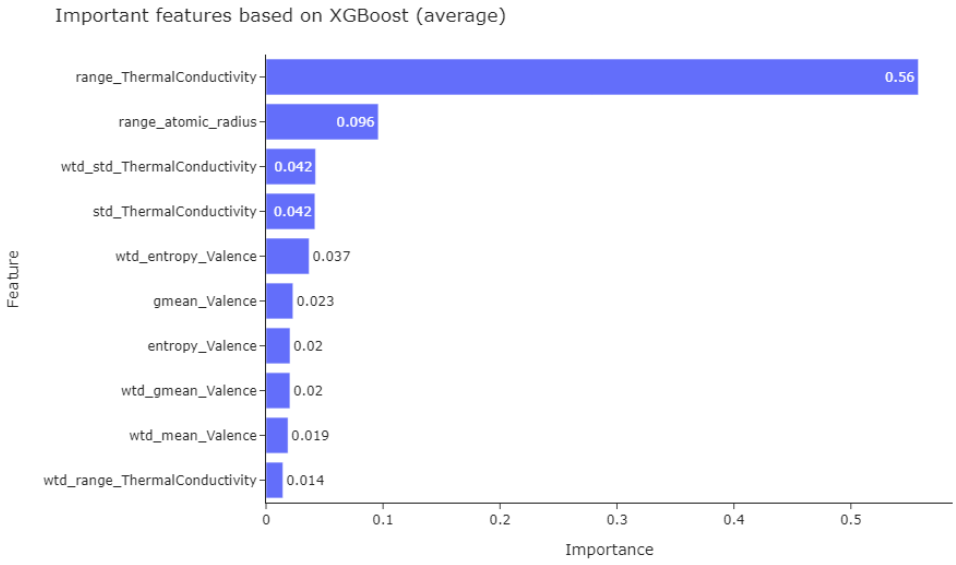

Figure 4: Average values of XGBoost's feature importance from 3 feature selection methods

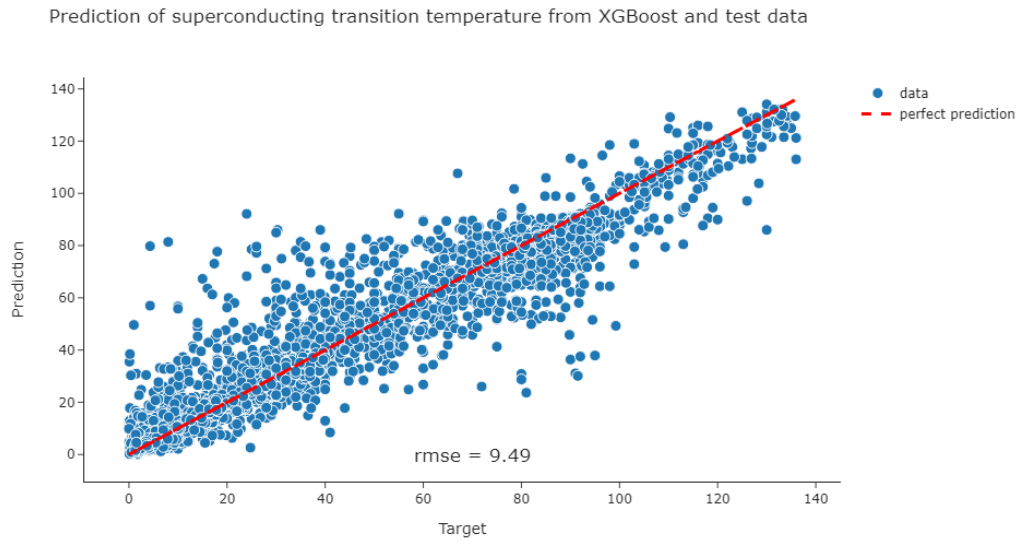

Figure 5: Prediction of T_c from XGBoost model where the red dashed line indicates zero-error predictions.

encoder on the chemical formula, and using other models, such as neural network.

Summary

XGBoost model using chemical formula of superconductors, their superconducting critical temperatures, and elemental properties was utilized to predict T_c. The best model was obtained by selecting 27 features or one third of all columns with rfe, while the best test rmse score is 9.49. The range of thermal conductivity is by far the most significant feature, followed by the range of atomic radius. Meanwhile, the most crucial properties are thermal conductivity, valence, and atomic radius.