STA 160 Final Project

Xuecheng Zhang
Department of Statistic–STA 160
University of California, Davis
Studnt ID: 915942842
hcczhang@ucdavis.edu

June 12, 2020

Abstract

"There are two files: (1) train.csv contains 81 features extracted from 21263 superconductors along with the critical temperature in the 82nd column, (2) unique-m.csv contains the chemical formula broken up for all the 21263 superconductors from the train.csv file. The last two columns have the critical temperature and chemical formula." I only use the train.csv data to do analysis. My goal is to understand the data set, and cluster the data, visualize the clustering results, and finally find a suitable model to predict the critical temperature.

All definitions come from Wikipedia, and code support from `scikit-learn.org`

`https://archive.ics.uci.edu/ml/datasets/Superconductivty+Data`

1 Exploratory Data Analysis

There are a total of 82 features in the superconductivity data set, and we can divide them into 10 data frames: the number of elements is categorical data, and the rest of them are numerical data. The numerical data include atomic mass, fie, atomic radius, density, electron aff, fusion heat, thermal cond, valance, and critical temperature. In the numerical data, the number of valances electrons a discrete variable, and the rest of them are continuous variables due to their nature. Except for the number of elements and critical temperature; all of the variables are shown by using mean, weighted mean, geometric mean, weighted geometric mean, standard deviation, weighted standard deviation, entropy, weighted entropy, range, and weighted range calculations.

1.1 Distribution of Critical Temperature

From the figure, we can see that the critical temperature of most substances have critical temperature values in 0-10k, and after 100k, only a few substances

have a critical temperature higher than 100k, also the critical temperature is right-skewed. However, there is one outlier which is 200 critical temperature value.

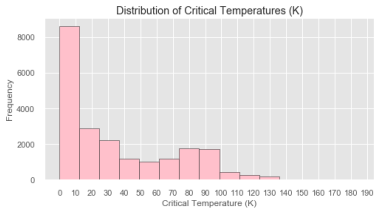

Figure 1:

1.2 Boxplot between Critical Temperature and Number of Elements

First, I found that the substances that have three or less elements have similar critical temperature and it generally low, but many of them may have very high temperature. Secondly, the temperature of substances is increasing linearly from four elements to eight elements, and the increasing rate is small.However, when the substances has 9 elements, its temperature drops compared with the substances which has 8 elements.

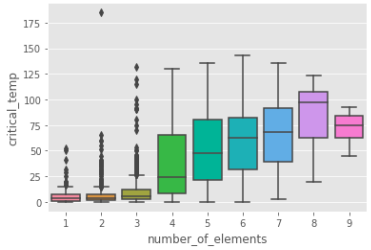

Figure 2:

1.3 Correlation Matrix Heatmaps

I set the correlation by 8 data frames:atomic mass, fie, atomic radius, density, electron aff, fusion heat, thermal cond, valance.

First, we can clear see that on the left top, there are 4 predictors have darker color than others which means they have higher correlation, they are mean, wtd-mean, gmean and wtd- gmean. Second, density and atomic mass have strong positive linear correlation. This make sense because chemical formula for density which is mass / volume. Third, except for thermal conductivity, almost all of

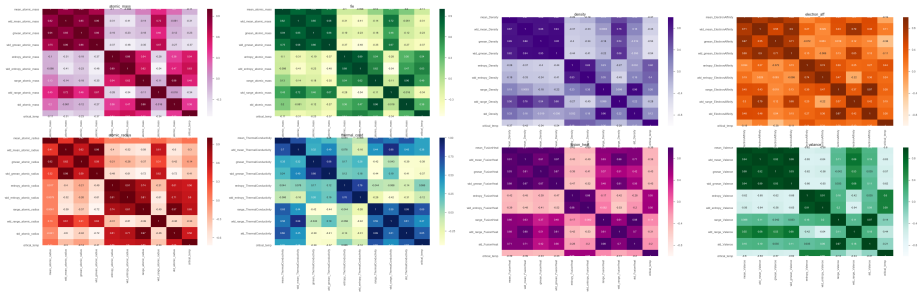

Figure 3:

the features exhibits a strong or significant positive correlation with each others variability measures. In short, it is observed that correlations between variability measures are significantly strong.

In order to see the multicollinearity problem, I find the VIF of each features.

	VIF	features
0	79.644423	number_of_elements
1	414.277383	mean_atomic_mass
2	818.370293	wtd_mean_atomic_mass
3	444.203673	gmean_atomic_mass
4	879.861538	wtd_gmean_atomic_mass
...	...	...
76	307.311308	wtd_entropy_Valence
77	56.759455	range_Valence
78	26.150907	wtd_range_Valence
79	96.823865	std_Valence
80	51.827597	wtd_std_Valence

Figure 4:

As we can see, most of the VIF are higher, this means that our prediction factors have a strong collinearity, and we can see from Figure 2 that there is a complex relationship between the prediction factors. This leads to the fact that linear regression model is not a good model for predicting critical temperature, because the linear regression models assume that all the predictors are independent from each other.

2 Cluster

2.1 Feature Selection

We need to select the features that contribute the most to the predicted critical temperature. From the previous analysis, we know that we have 81 variables, and there are higher correlation between them, so feature selection can reduce over fitting, improve accuracy and reduce training time.

1. I remove all low variance columns that have a very small variance(5 percent). However, all the features have high variance, so we keep all features.

2. I drop any mutually correlated features, then we get 28 features.

3. I get rid of any features that have a very low correlation (abs$<$0.1) with the critical temperature.

Finally, I get 27 features, they are: number of elements, mean atomic mass, range atomic mass, wtd-range-atomic-mass, mean-fie, wtd-mean-fie, wtd-entropy-fie, range-fie, wtd-range-fie, mean-atomic-radius, wtd-range-atomic-radius, mean-Density, range-Density, mean-ElectronAffinity, wtd-mean-ElectronAffinity, range-ElectronAffinity, wtd-range-ElectronAffinity, mean-FusionHeat, range-FusionHeat, mean-ThermalConductivity, wtd-mean-ThermalConductivity, gmean-Thermal, wtd-entropy-ThermalConductivity, range-ThermalConductivity, range-Valence, wtd-range-Valence.

2.2 Hierarchical Cluster

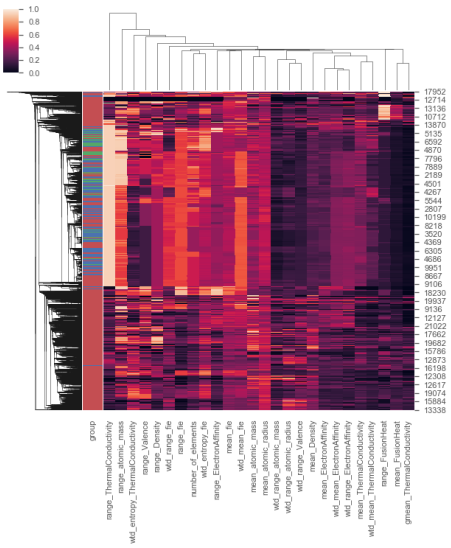

Figure 5: Hierarchical Cluster Heat Map

I use the data set after feature selection. Then I use hierarchical cluster

which is "a general family of clustering algorithms that build nested clusters by merging or splitting them successively"(Wikipedia page). Next, I use single linkage method which minimizes the distance between the closest observations of pairs of clusters. In order to observe the cluster result, I divided critical temperature to 3 groups. The critical temperature lower than 46K is belong to low, and the color is red, the critical temperature between 46K and 92K is mid group which is blue, the rest is green.

From figure we can see that lower group is easy to be clustered, but mid and high group are mix together. For this result, I can say one reason is I created the wrong group for the critical temperature. For instance,the frequency of blue color blocks is high and the area is large, and the number of green color blocks is small, so there may be two groups of data. Other reason is that for this data set, hierarchical is not the best method to cluster. In short, I think this error is normal.

3 Model Selection

Before we do the data analysis. I need to prepare the data first.

1. Split data into training and test sets.By using training set and testing set, I can minimize the effects of data discrepancies.

2. Standardize the data. I need to transform the data onto unit scale which mean equal 0 and variance equal 1.

3.1 Principal Component Analysis

As we known, one of the most important applications of PCA is for speeding up machine learning algorithms. In our original data have 81 features except critical temperature.

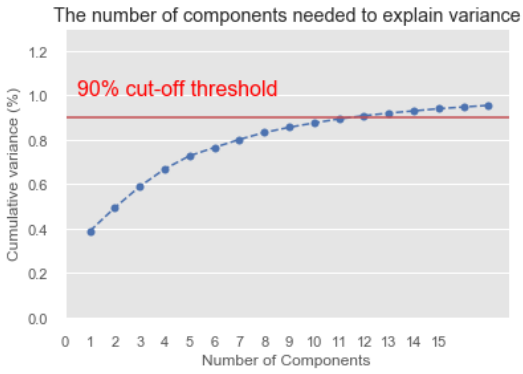

Figure 6: Number of Components Explain Variance

we want that the principal component can explained variance to be around 90 percent, from the figure 6, I choose the first eleven PC. And the explained

variance ratio is 0.39384406, 0.10280271, 0.09414382, 0.07822043, 0.05760655, 0.03903877, 0.03582438, 0.03071241, 0.02338201, 0.01948765, 0.01850802.

3.2 Multiple Linear Regression Model

In the previous analysis, we know that this data is not suitable for linear regression model, because there are collinearity and complex relationship between its features. After we fit the linear regression model, the model coefficients are -3.99219323, - 1.941134, 2.49772055, - 1.16125916, 2.56485398, 3.27119184, - 1.1096712, - 0.40282578, - 1.98117586, 0.109578, - 2.9455536. Further,the mean squared error is 746.50. The residuals for linear regression model and prediction error plots are shown below.

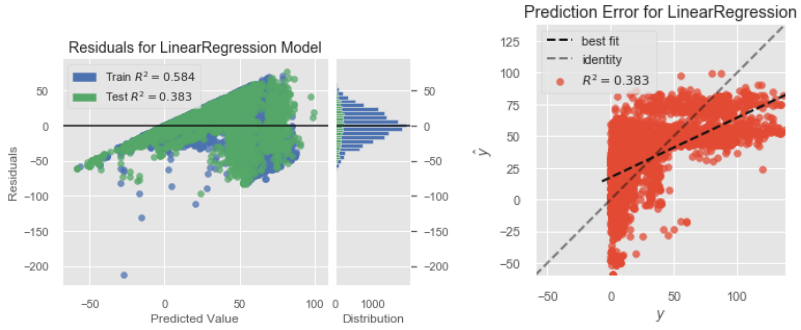

Figure 7: Residuals for Linear Regression Model

Obviously, from the left part of figure 7, we can see the distribution of residuals is follow the normal distribution, and the test R^2 is only 0.383 which means that MLR can not predict the critical temperature.

The right part of the figure 7 is a prediction error for linear regression model. We can see that the predict critical temperature is higher than true critical temperature, after 40K,the predict critical temperature is lower than true critical temperature. The MLR did the bad prediction when critical temperature between 0K and 50K. In short, MLR is not a good model for superconductive data.

3.3 Random Forest Regression

We apply the random forest regression for our data which "is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap Aggregation, commonly known as bagging"(Krishni). The residuals for random forest regression model and prediction error plots are shown below.

Obviously, from the left part of figure 8, Residuals are mostly distributed near 0, and residual tend to rise when the predicted value becomes larger. The train

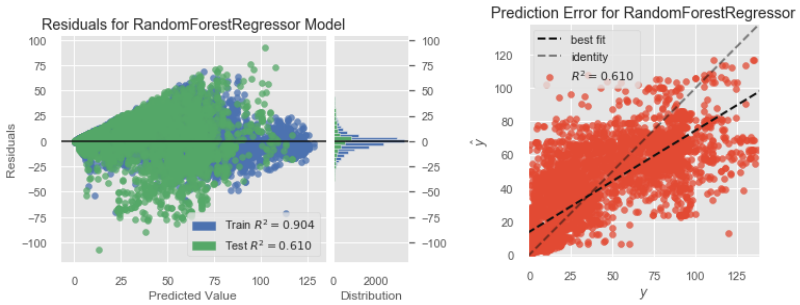

Figure 8: Residuals for Random Forest Regression Model

R^2 is 0.904, which can express 90 percent, but the coefficient of determination is 0.61, its better than MLR model.

The right part of the figure 8 is a prediction error for random forest regression model. We can see that the key y value is 30k, before 30k, the predict value is higher than true value, after 30K, they are the opposite. The distribution of points in the figure is close to the random forest regression model's best fit line, indicating that the prediction has certain error, but the scattered points are in a linear relationship, indicating that the prediction make some sense.

3.4 Support Vector Regression

We apply the support vector regression for our data which "produced by support vector classification depends only on a subset of the training data, because the cost function for building the model does not care about training points that lie beyond the margin"(scikit.learn). The residuals for support vector regression model and prediction error plots are shown below.

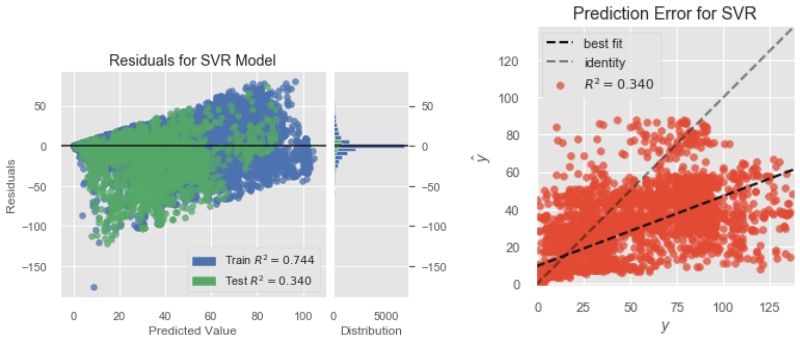

Figure 9: Residuals for Support Vector Regression Model

Obviously, from the left part of figure 9, Residuals are mostly distributed near 0, but the test data a little bit left skew. Also the residual tend to rise when

the predicted value becomes larger. The train R^2 is 0.744, but the coefficient of determination is 0.34, which means the support vector regression is not a good model to predict critical temperature.

The right part of the figure 9 is a prediction error for support vector regression model. We can see that the predict critical temperature is higher than true critical temperature, after 10K,the predict critical temperature is lower than true critical temperature. We can say that the predict value mostly higher than true value. Also the distribution of points in the figure is around bottom of this plot.

3.5 Gradient Boosting Regression

We apply the gradient boosting regression for our data which" builds an additive model in a forward stage-wise fashion; it allows for the optimization of arbitrary differentiable loss functions. In each stage a regression tree is fit on the negative gradient of the given loss function."(scikit.learn). The residuals for gradient boosting regression model and prediction error plots are shown below.

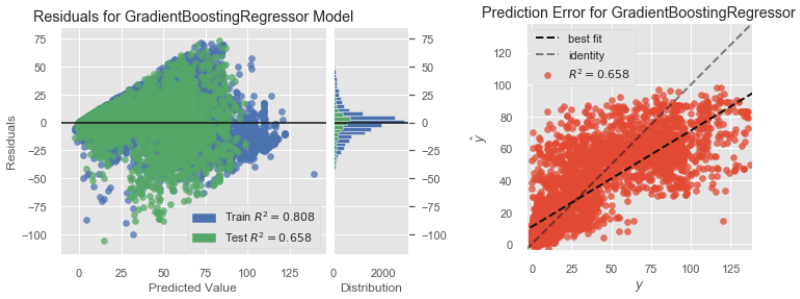

Figure 10: Residuals for Gradient Boosting Regression Model

Obviously, from the left part of figure 10,we can see the distribution of residuals is follow the normal distribution, and the training R^2 is 0.808, and test R^2 is 0.658. Also the residual tend to rise when the predicted value becomes larger.

The right part of the figure 10 is a prediction error for gradient boosting regression model. The key y value is 26k, before 26k, the predict value is higher than true value, after 26K,the predict critical temperature is lower than true critical temperature.

3.6 Comparing Results

"MSE is a risk function, corresponding to the expected value of the squared error loss"(Wikipedia). It tells us how close a regression line is to a set of points. Lower MSE means the expected value close to the true value. Compare to those four models, gradient boosting regression have the lowest MSE.

Table 1: Compare MSE and R^2 for Each Model

Model Name	Mean Squared Error	R^2
Multiple Linear Regression	746.50	0.38
Random Forest Regression	471.58	0.61
Support Vector Regression	798.24	0.34
Gradient Boosting Regression	413.49	0.66

"Coefficient of determination is a statistic that will give some information about the goodness of fit of a model"(Wikipedia). In regression, the coefficient of determination coefficient of determination is a statistical measure of how well the regression predictions approximate the real data points. An coefficient of determination of 1 indicates that the regression predictions perfectly fit the data. From table 1, gradient boosting regression have the lowest R^2, R^2=0.66. If we only look at mean squared error, and coefficient of determination, gradient boosting regression is better for our superconductivity data.

4 Conclusion

Superconductivity data is very complex, it contains 82 features, which makes me have a huge problem in the beginning. I need to understand the meaning of each feature, so I spent a lot of time in exploration data analysis, I split it in 10 data frames: the number of elements, atomic mass, fie, atomic radius, density, electron aff, fusion heat, thermal cond, valance, and critical temperature. Except for the number of elements and critical temperature; all of the variables are shown by using mean, weighted mean, geometric mean, weighted geometric mean, standard deviation, weighted standard deviation, entropy, weighted entropy, range, and weighted range calculations. Except predicted critical temperature, there are complex relationships among other features. Mean,wtd-mean, gmean and wtd-gmean have higher correlation between each other, and strongest correlation lying between 0.7 and 0.9 are observed between atomic mass and density; atomic mass and atomic radius; atomic radius and first ionisation energy; atomic radius and density; density and valance. After exploration data analysis, I did the unsupervised analysis, I did the feature selection before cluster, I remove all low variance features, drop mutually correlated features, drop the low correlation features, then I get 27 features. Then I use hierarchical cluster our data into three group: low, mid and high. Next, I did the supervised analysis, I want to build the model to predict the critical temperature. Before I fit the model, I did the principal component analysis, I use first eleven components to fit the model. I fit the multiple linear regression, random forest regression, support vector regression, and gradient boosting regression, then I compare the MSE and R^2, gradient boosting regression is better to predict the critical temperature.