

STA160 Final Project

Predicting Critical Temperature of Superconductors

Haolin Li | 913838107 | June 12, 2020

# Introduction

In this report, I will conduct data analysis on the superconductivity dataset from UCI Machine Learning Repository, to gain more knowledge about how data are structured and more, as well as to practice data visualization, dimension reduction, regression and more.

### Superconductivity Data Data Set

This dataset is downloaded from the UCI Machine Learning Repository. There are two files contained in the zip file: train.csv and unique\_m.csv, the goal of this project is to predict the critical temperature based on the features in the dataset, therefore, I will only use the train.csv file because it contains the features I needed.

The train.csv file contains 81 features extracted from 21263 superconductors along with the critical temperature in the last column. The unique\_m.csv file contains the chemical formula broken up by elements for all the 21263 superconductors from the train.csv file, as well as the critical temperature and chemical formula in the last two columns.

According to Kam Hamidieh in his paper called “A data-driven statistical model for predicting the critical temperature of a superconductor”, the data within the train.csv file was cleaned up by him. The original dataset was messier and larger, Kam Hamidieh preprocessed and cleaned up the data to obtain the data now in the train.csv file. According to his words, the train.csv file has about 67% of the original data. More about data preprocessing can be found in his paper.

# Data Exploratory

Thanks to Kam Hamideh, as mentioned above, he preprocessed the data already and therefore, the data within the dataset are ready to use. Despite I know about data preprocessing was done, I checked if there are any missing value in the dataset, which there are none. Thus, I can explore the data without doing any preprocessing.

In the train.csv file, there are 81 features from 21263 superconductors, however, they are 10 different features for feature extraction of 8 different properties of an element, as well as the number of elements within the superconductor. For the 8 different properties, there are atomic mass, first ionization energy, atomic radius, density, electron affinity, fusion heat, thermal conductivity and valence. For each of the property, there are 10 different features for feature extraction, they are the mean, the weighted mean, geometric mean, weighted geometric mean, entropy, weighted entropy, range, weighted range, standard deviation and weighted standard deviation. Below I obtained a table to summarize the predictor variables. The value for each feature of a property is the mean of that feature of all the superconductors. For detailed explanations of the features and properties, they can be found in Kam Hamideh’s paper regarding the formula of the features and the unit of the property.

A screenshot of a cell phone

Description automatically generated

Below is the histogram of the distribution of critical temperatures for all the superconductors. The distribution is clearly right skewed with most of the critical temperature are in between 0 and 10 and the unit of measurement is K(Kelvin).

A screenshot of a social media post

Description automatically generated

# Data Correlation

In this section, I will show the correlation between the variables in the data. The reason we want to know about the correlation between the variables is that we want to predict the critical temperature based on the variables, correlation tells us the relationship between the variables so that we can make a decision of whether or not to remove the variables, removing one variable may significantly change the value of other variables.

When we want to look at the correlation between the variables, there are usually two ways: correlation table/map and variable inflation factor. It is impossible to display the correlation table/map of all 81 variables in the dataset because the visualization would be very large to observe. Therefore, below I obtained the correlation heat map for 20 selected variables in the data.

A close up of a colorful background

Description automatically generated

From the correlation heat map shown above, we can see a trend within the variables. We can see that, weighted mean and weight geometric mean of first ionization energy are strongly correlated, and the mean and geometric mean of atomic radius are also strongly correlated. The high correlated pairs are related within the property and as well as the features of the property, such as mean and geometric mean are strongly correlated. To obtain a better look on this trend, I obtained the correlation heat map for each property with their features.

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

From each correlation map, we can see they all share a similar trend is that the features regarding the mean have high correlation, such as there are high correlation between the mean, weighted mean, geometric mean and weighted geometric mean, same trend happens in entropy and standard deviation, while range and weighted range do not have a relatively high correlation as the others.

From the correlation heat maps above, we have an idea of the correlation between the variables in the dataset, however, we only know the correlation of features within the property, we may want to observe the correlation of features of all properties when we want to know whether the one variable is significant or not. Below, I obtained two tables of the variance inflation factor (VIF) of all the variables in the dataset.

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

The VIF table on the left shows the VIF score for each variable in the dataset, while on the right shows the VIF score in a descending order as well as the corresponding variables. A variance inflation factor detects multicollinearity in regression analysis, which is the correlation between predictor variables in a model. One way we can interpret the score of VIF is that if it is equal to 1, then not correlated, if it is between 1 and 5, it is moderately correlated, and if it is greater than 5, then it is highly correlated. The variance inflation factors range from 1 upward. In general, we consider a VIF above 10 has high correlation. For the VIF table shown above, we can see that the maximum of VIF is 8433.872 and the minimum is 17.774. Since they are all greater than 10, we can say that they all have high correlation between the predictor variables, and we should not drop them.

# Principal Component Analysis (PCA)

From the VIF table, we know that all variables have high correlation and we should not drop them, but with 81 predictor variables in the data, it generates a high dimensional space and may cause overfitting in model. Then principal component analysis comes into place to reduce the dimension of the data while retaining as much information as possible. There are two ways to display the PCA, tables of explained variance and plot. Below are the two tables of PCA.

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

Table on the left shows that cumulative explained variance and table on the right shows the explained variance for each component. We can interpret the explained variance as the proportion of the variance in the data. For example, the first component, P1, explains 38.93% of the variance in the data, then p2 explains 10.48% of the variance in the data and so on, and they all add up equal to 100% or very close to 100%. Same idea applies to the cumulative table, p1 explains 38.93% of the variance, p2 is its percentage of the variance combined with p1, so is 49.41%. The other way to interpret cumulative explained variance is that for example p4, means the first 4 components can explain 66.84% of the variance of the data, which is over half of the information of the data. To better understanding the interpretation of the PCA, I obtained the plot for each table.

A picture containing game

Description automatically generated

A close up of a mans face

Description automatically generated

The cumulative plot shows that as more components are included, it explains more proportion of the data which eventually reaches to 100%. The scree plot shows the proportion of the data for each component and therefore components at the end contributes 0% of the variance in the data.

We can see that both from the cumulative and scree plot, the proportion of the variance reaches to maximum with the first 63 components, therefore, we may drop the components after the 63th component since they do not contribute information to the data.

Above we showed the proportion of the components in the data, below I will explain the correlations between the principal components and the original variables in the data. Below are the two tables of the variables contributed in the first component.

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

The way we can interpret the variables in the first component is that the first principal component is strongly correlated with the original variables on the left, and not very correlated with the original variables on the right. Another way we can interpret this is the first principal component increases with increasing the variables on the left and decreasing the variables on the right. When to determine at the level the correlation is important, we generally take correlation above 0.5 as indicating important correlation.

# Regression Models

From the principal component analysis above, we know that the first 63 component can explain nearly 100% of the data and the remaining component are generally useless. However, when using principal components analysis in regression, we generally take number of components explain more about 60% of the data. In this analysis, I will use number of components that explain about 90% of the data. In this case, after splitting the data into train and test data, the number of components that explain about 90% of the data is 12, in other words, the first 12 principal components explain 90% of the data.

Below I will obtain 4 different regression models to predict the critical temperature, they are multiple linear regression, random forest regression, bagging regressor and XGBoost regressor. I will obtain their root mean squared error (RMSE) and r-squared for comparison. I will also generate plots to display the fit of the predicted outputs.

Below is the table showing the rooted mean squared error and r-squared for each regression model.

|  |  |  |
| --- | --- | --- |
|  | RMSE | R^2 |
| Multiple Linear Regression | 21.8776 | 0.5945 |
| Random Forest | 12.4055 | 0.8696 |
| Bagging | 10.4416 | 0.9076 |
| XGBoost | 10.8349 | 0.9005 |

A general idea of these numbers of value is that the lower the root mean square error and the higher the r-squared, lead to the better the fit of the model. RMSE is an absolute measure of fit and r-squared is a relative measure of fit. An r-squared of 90% means that 90% of the data fir the regression model, it is a measure of how well the regression model fits the data.

Below are the plots of predicted and observed data to show the fit of the predicted outputs for each regression model.

A close up of a map

Description automatically generated

A close up of a map

Description automatically generated

A close up of a map

Description automatically generated

A close up of a map

Description automatically generated

We can see that for the multiple linear regression model, the predicted outputs are not close to the observed line, therefore it has the worst fit among all. For the bagging and xgboost, we cannot observe significant difference between the two since their fits are so similar, but in numeric, bagging has a better fit than xgboost.

# Conclusion

Overall, in this report, I displayed two ways to visually and numerically to find the correlation between the variables in the data, they are using correlation heat map and obtaining the variance inflation factor. Then, since the data has a large dimension, meaning having many predictor variables, I used principal component analysis to reduce the dimension of the data while retaining an important proportion of the original data. Finally, I used 4 regression models to predict the critical temperature and compared their root mean square error and r-squared, the results reveal that bagging regressor performed the best fit among all models with a r-squared of 0.9076.