Exploration on the Superconductivity Dataset From a Statistical Learning Perspective

Hainiu (Johnny) Xu
Undergraduate Student, Department of Statistics
University of California, Davis
Davis, CA 95616
scxu@ucdavis.edu

Abstract

With recent developments in supercomputer, medical technology and the demand for green energy, Superconductivity and superconductor is becoming a increasingly popular research topic in the field of physics. The characteristic of Superconductivity was discovered in the early 20th century Onnes and prospered in theoretical physics. Countless studies about Superconductivity have been conducted and the related studies have also received numerous Nobel prizes in physics. The key feature of superconductor is its superconductivity under a certain critical temperature (in Kelvin). The major difficulty of studying superconductor is the determination of the critical temprature as it is often of extreme values. This raises the interest of statistical learning. Making inferences and construct statistical model that predict the critical value of a given substances based on its other physical features would be exceedingly helpful with studying the superconductivity of substances. Integrating statistical tools with physics study has always been a mainstream of conducting physics research and the growing sophistication of statistical learning techniques has stimulated such trend. Therefore, in this project, we apply modern statistical learning and regression techniques to conduct a comprehensive data analysis and build a predictive model that gives prominent result in predicting the critical value of superconductivity.

1 Introduction

The Superconductivity dataset to be analyzed in this study is obtained from the UCI Machine Learning Repository. The original author of this dataset is Kam Hamidieh, who is a lecturer at the statistics department at the University of Pennsylvania. There are two datasets in this repository, one of which contains information of the physical characteristics of the substances and the other provides information on their chemical features. Both datasets have the potential to be used as the training dataset and the detailed statistical characteristic of both datasets will be examined closely in the exploratory data analysis. Since the prediction target, which is the critical temperature of a specific substance to reveal its superconductivity, are numerical values instead of categorical variables, this would limit the use of statistical models to models that are capable of conducting regression. For instance, the linear regression and its derivatives such as ridge regression and LASSO shall be employed and the corresponding result shall be examined.

Due to the fact that the majority of the models to be fitted in this study are linear models, therefore the assumptions for linear regression is a critical factor in this case. Originated from the assumption that the residuals shall be independently and identically normally distributed with mean 0 and variance $\sigma^2 < \infty$, it follows that the response variable shall also follow a normal distribution. Thus, validating these assumptions are also vital components of this study and the statistical integrity of

A course project for STA 160 (Spring 2020), Davis, CA, USA.

the result is highly depended on it. In addition, the potential possibility of over-fitting is another critical issue with regression as most regression models are variant-greedy and they are easily over-fitted. Thus model selection as well as validating the prediction model using unseen data are both crucial steps to be taken during the study.

The significance of this interdisciplinary study is that it can be constructive to explore the application of modern statistical methods, machine learning in particular, to aid and accelerate physics researches. Superconductivity has been a popular research topic for centuries and its influence extend well beyond the field of physics. Other fields such as computer engineering, biomedical engineering, and civil engineering are all interested in the special characteristic of superconductor– superconductivity. It transports electrons with no resistance thus materials that have superconductivity are capable of generating strong magnetic field [1, 2]. Hence, to systematically explore the possibility of employing modern machine learning techniques to predict substances critical temperature to reveal their superconductivity, the study will be organized in the following way: an exploratory data analysis well be conducted in section 2 to examine the necessary assumptions needed for certain regression models such as multiple linear regression. The *train* data, which contain information of the physical features of the sample substances will be studied in section 3, including fitting and accessing the performance of statistical models. The second dataset *unique_m*, which contains the chemical features of the sample substances will be studied in section 4. Section 5 will summarize the results of the previous sections and make conclusions to this study.

2 Exploratory Data Analysis

The *train* dataset contains a total of 21263 sample substances and 82 of their physical features including information such as the number of elements, mean atomic mass, etc. Thus, the first insight by solely looking at the number of variables available for regression, we already see a potential possibility for overfitting. Due to the limited length of this report, the explicit information on the variables available in this dataset is in the Appendix. Missing value often exists in large datasets like the *train* data and it can produce be hideous to model fitting or other statistical calculation in the later section. Thus, the number of missing value in the *train* dataset is checked and it turns out that no missing value is found, which guarantees that the following calculation and result are representative of the data.

From the correlation matrix of the variables in the *train* dataset shown on the left side of Figure 1, we can see that the majority of the data are not highly correlated. We do see moderate collinearity issue with variables that are very similar to each other. For instance, suppose we the variables of *Density*. Then it would give the correlation matrix on the right side of Figure 1. Thus, the issue with collinearity does exist for some variables that are closely related such as statistics that are directly related to each other. In the case of the *Density* attribute, we see that the mean (*mean_Density*), the weighted mean (*wtd_mean_Density*), the geometric mean (*gmean_Density*), and the weighted geometric mean (*wtd_gmean_Density*) are highly correlated since they all represent a similar statistic. Thus, we will combine the moderate collinearity issue with model selection to see if the correlated variables shall be removed from the model or not.

After looking into the variables, we now shift our attention to the response variable, which is the critical temperature that a substance would turn to a superconductor. First of all, some summary statistics is listed in Table 1

Count	21263.00
Mean Temperature	34.42
Standard Deviation	34.25
Minimum Temperature	0.00021
25% Quantile	5.37
Median Temperature	20.00
75% Quantile	63.00
Maximum Temperature	185.000000

Table 1: Summary Statistics of the response variable, *critical_temp*.

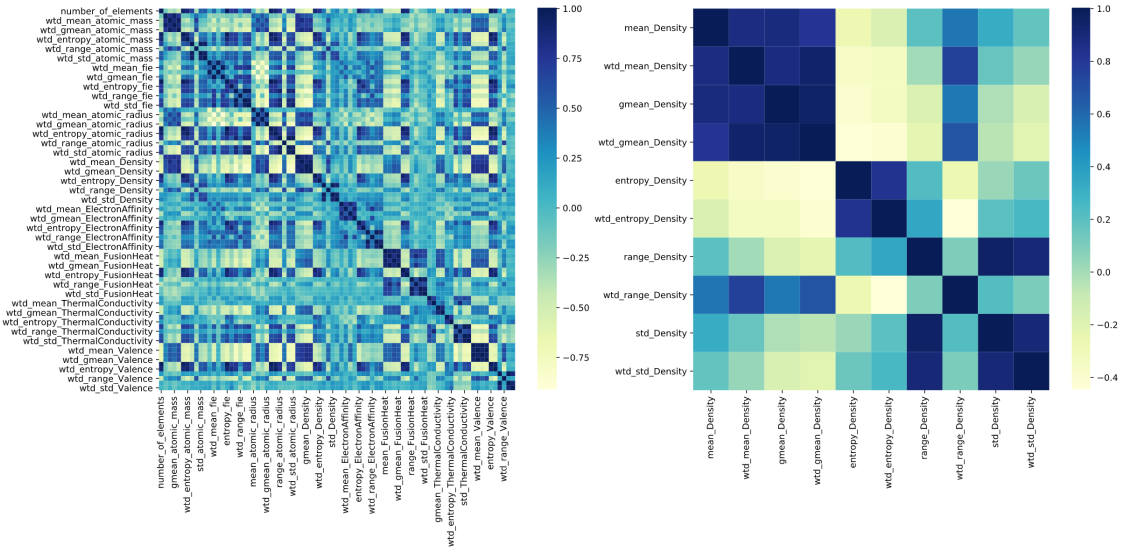

Figure 1: Correlation matrix of variables in *train* dataset. The left figure is the correlation matrix of the full dataset and the right figure represents correlation matrix of the *Density* variables only.

From Table 1, we can see that the critical temperature where a substance would turn to a superconductor is not concentrated as the standard deviation is 34.25 K, which is close to the mean temperature 34.42 K. Thus, we can see that the variation of the critical temperature between substances is quite large, suggesting a potential possibility to divide temperatures into groups, classify the unknown substances into individual groups and then utilize regression model to predict the estimated critical value of the substance within each group. This approach shall be addressed in the modeling section. Further, we have the density plot of the response variable which is shown as the left plot in Figure 2. As we can see, the estimated density based on the original sample data is not normally distributed. In fact, it is heavily skewed to the right suggestion that some substances have a high critical temperature albeit most substances have a relatively low critical temperature. To satisfy the assumption of multiple linear regression: $\vec{Y} \sim \mathcal{N}_p(X\vec{\beta}, \sigma^2)$ where $\vec{Y}$ is the random vector of the response variables, p is the number of variables, X is the random data matrix, and $\vec{\beta}$ is the vector of coefficients. Box-Cox transformation is therefore applied to the original sample data and the resulting density is shown in the middle plot of Figure 2. Further, the right-most plot in Figure 2, which shows that comparison of the density of the Box-Cox transformed data with the normal density, suggests that the density of the transformed data is clearly more alike the normal density but the peak of the density is missing leading to a moderate violation of the multiple linear regression assumption. The Box-Cox transformation, which is defined as

$$\mathcal{B}(Y) = \frac{Y^{\lambda} - 1}{\lambda}$$

where Y is the original sample data and λ is a tuning parameter [3]. Therefore, it is easy to show that the Box-Cox creates a bijection between the original sample space and the space of transform data and it is easy to derive the inverse mapping

$$Y = (\mathcal{B}(Y) \cdot \lambda + 1)^{\frac{1}{\lambda}}$$

Hence, the predicted value obtained in the form of Box-Cox transformed data can be easily inversed back to the standard temperature scale, which allows us to interpret the result as a temperature instead of merely a numerical value.

Excluding the response variable, the *train* dataset provided 81 variables. Running the regression model based on all 81 variables could be time consuming and computationally greedy. Hence, we first attempt to conduct the principal component analysis (PCA) to reduce the dimension of the data and employ regression models on the resulting principal components. In this case, we wish to

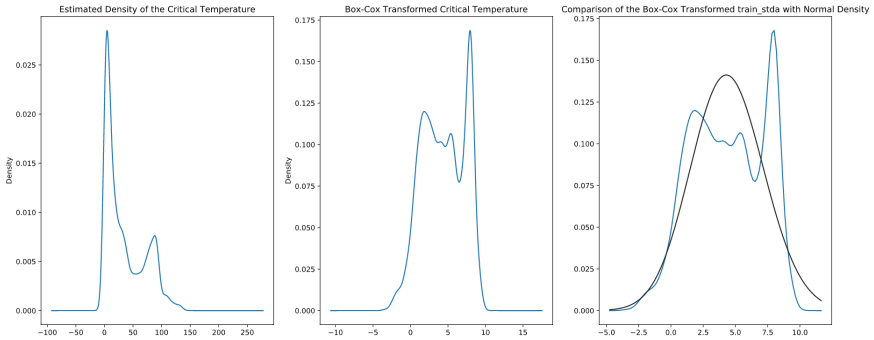

Figure 2: Density plot of the critical temperature variable. Left: the estimated density based on the distribution of the original sample data; Middle: the estimated density based on Box-Cox transformed data; Right: compare the density of the Box-Cox transformed data with the normal density.

calculate the efficient number of principal components such that the fraction of variance explained (FVE) by the leading principal components is larger than 90% of the total variance. The PCA for this study is calculated based on the correlation matrix of the standardized data. To achieve a FVE of above 90%, a total of 12 principal components will need to be used and they explain a total of 90.87% of the total variation. Further, based on the loadings of the leading principal components, the contribution of the determination of the corresponding principal components is studied and the result in shown in Figure 3 along with the biplot of the principal components.

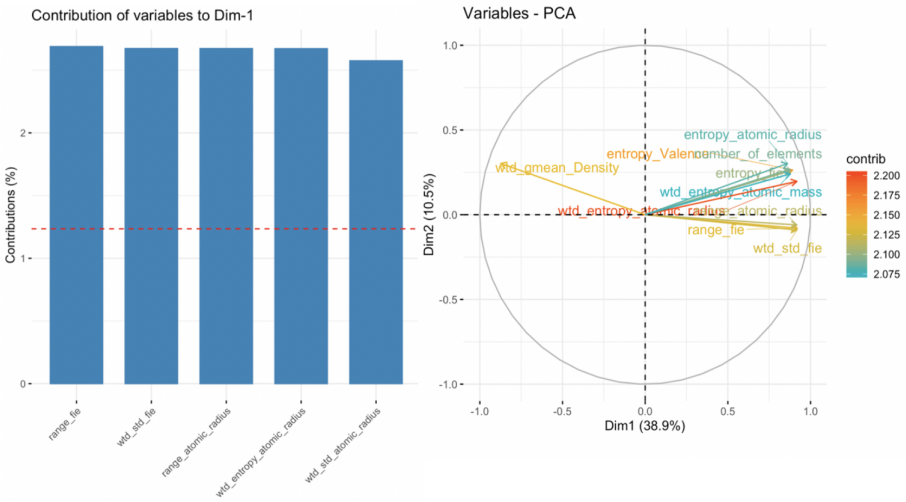

Figure 3: Diagnostic plot of PCA. Left: Variables with largest contribution to the determination of principal components overall. Right: Biplot of the leading 10 principal components.

From the contribution plot, we can see that *FIE* and *Atomic Radius* are the key characteristics of the substances. The fact that they explain the largest amount of variation indicates that it is the easiest to distinguish between the substances based on these features. The same conclusion goes for the variables that contribute the most to the determination of PC2 and PC3. It is reasonable to assume that these variable would also be the main force in the prediction model. From the biplot, we see that the variation direction of the *Density* variable is different from that of the other variables, which indicates that the *Density* variable could also be a good candidate for aiding prediction. Further, we can also see that the direction of variation of some other variables such as *FIE* and *Atomic Mass* it quite similar, which matches the result we had in the correlation matrix– the direction of variation among these variables is similar since they are highly correlated.

Another nice feature of PCA is that some clustering may be observed by plotting the principal components against each other. In this case, we visualized both the 2D plot of PC1 against PC2 as well as the 3D plot where PC1, PC2, and PC3 are used for visualization. The plots are shown in Figure 4. As the plots suggest, there is no obvious clutering among the substances mainly due to the reason that most substances are of continuous values. One insteresting obseration from this visualization is that the 3D plot suggest that the PC scores of these sample substances seem to fall on a hyperplane in the 3D space and the meaning of this phenomenon shall be further studied.

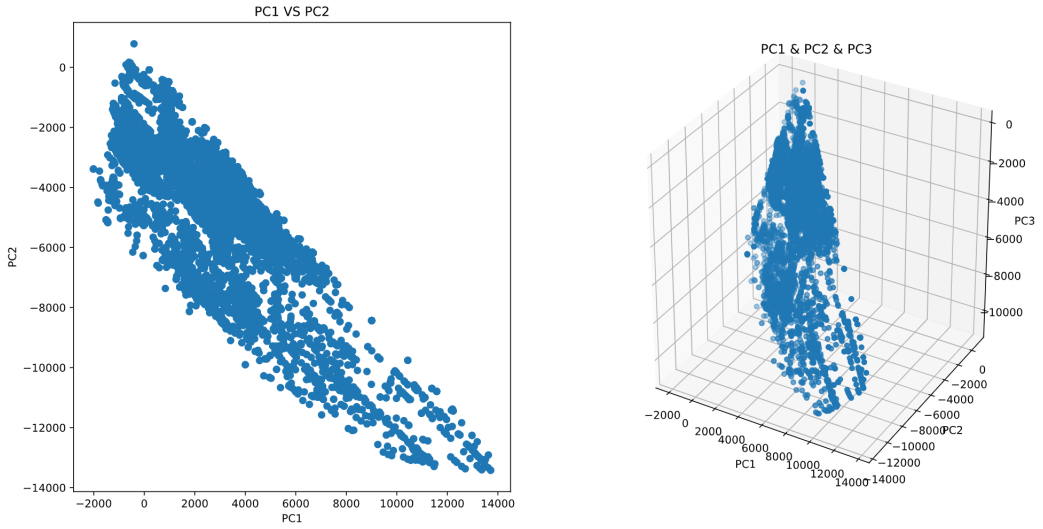

Figure 4: 2D and 3D visualization of the PCA result.

As for the *unique_m* dataset, there is very limited information contained in it. First of all, the data matrix is very sparse. From Figure in the Appdendix, we can see that there is rarely any variable with available information. By further checking the summary statistics of each variable, it shows that the minimum, 25% quantile, median, and 75% quantile value of most variables is 0 with very few exceptions. Therefore, we decided to study the critical temperature mainly based on the *train* dataset and leave the *unique_m* dataset as a backup. We will explore the *unique_m* dataset indepth if the regression result from the *train* dataset do not match our expectation.

3 Model Fitting and Prediction

Numerous regression methods have been proposed and extensively studied and the abundance of regression methods allows us to make comparisons between different approaches and employ the most prominent of them all. Based on the data analysis done in the previous section, we now have two options. The first option is to fit a regression model directly on the full dataset, which utilizes all the variables. Another possible way of doing prediction is to fit a model using only the principal components. Since the leading principal components captures the main mode of variations, the model built based on them shall also contain the majority of the information contained in the dataset. Regression using only the principal components shall be much more efficient compare to the full model as a significantly fewer variable will be used. In this case, the full model would contain 81 variables where as the smaller model using only the principal components only contain 12 variables. Hence, we would examine both approaches mainly based on the prediction accuracy and the computational efficiency shall also be taken into consideration.

3.1 Model Introduction

In this study, we wii mainly examine 5 regression models, which are multiple linear regression model using ordinary least squares (OLS), LASSO, AdaBoosting regression, Gradient Boosting

regression, and XGBoosting regression. The reason for selecting these regression methods is that they are capable of conducting regression and they are the best models that are well represent the performance of models in their category. According to the famous No-Free-Lunch theorem, no model is the universally the best model. Therefore, we do not have any preference among these model albeit some are generally a better option than others. Among these models, linear regression and LASSO are based on the foundation of linear regression where the classical multiple linear regression utilizes the OLS and the LASSO model introduced shrinkage to the coefficients thus focus the computing power to conduct prediction based on the most distinct variables. Explicitly, the use of the "0-1" loss function allows LASSO regression model to completely neglect the nebulous variables. The inherent model selection feature allows it to be used on dataset with massive variables or even high dimensional dataset. The other three methods, AdaBoosting, Gradient Boosting, and XGBoosting regression are derivative of the boosting method. They share similarity with LASSO since boosting also accumulate resources on the most blurry variables. All these boosting method is based on decision tree thus the result may not be as prominent as the linear regression models but it also depends on the data. XGBoosting and Gradient Boosting both uses the classical gradient descent algorithm. The major difference between these two approaches is that the XGBoosting method uses the second-order gradient of the loss function to estimate the direction, which shall provide more information than using the first-order gradient alone.

3.2 Model Fitting and Prediction Result Based on PCA

All 5 models are fitted using the leading 12 principal components. The main reason for fitting using only the principal components is to increase the efficiency of the learning algorithms. The principal component is first determined based on the full dataset and then the dataset is splited into a train dataset and a test dataset. The dimension of the train and testing dataset is then reduced to 12 according to the loadings determined by PCA. In such a way, we have only 12 variables in both the train and testing dataset. The resulting accuracy rate as well as the determination of coefficient $R^2 = \frac{ESS}{TSS}$, which shows the amount of variation explained by regression (ESS: explained sum of squares, TSS: total sum of squares), is displayed in Table 2.

Method	Est. Mean Squared Error	R^2	Box-Cox Inversed
LASSO	2.89	0.65	9.00
AdaBoost	3.54	0.56	13.00
Gradient Boosting	1.97	0.76	5.01
XGBoosting	4.58	0.43	22.00
Linear Regression	2.83	0.66	8.66

Table 2: Regression Results of the 5 models.

Based on the regression results, we can see that the Gradient Boosting model provided the best result. It has the lowest mean squared error. The coefficient of determination of the Gradient Boosting model, $R^2_{GB} = 0.76$, also suggests that the model fits the data relatively well. Based on the Box-Cox inversed mean square error ($\lambda = 0.24$ in this case), we can see that $\sqrt{5.01} = 2.24$ implies that the average error that our model would make when predicting the critical temperature of the substance would be off 2.24 Kelvins on average. Considering the fact that the average critical temperature is 34.42 Kelvins with a 34.52 Kelvins standard deviation and that the original distribution is skewed to the right, suggesting that there are more substances with extremely high critical temperatures, the average error that the Gradient Boosting model would make suggest that it is a prominent model for estimating the critical temperature based on the principal components. The worst model here is the XGBoosting model albeit that it shall perform better than the gradient boosting model in theory. The coefficient of determination of this model is $R^2_{XG} = 0.43$, which indicate that the model did not fit the data very well. The Box-Cox inversed mean squared error indicate that the predicted value would be off 4.69 Kelvins on average, which is also an acceptable result by much worse than that of the Gradient Boosting.

In terms of the computation efficiency, all 5 models are quite efficient. Although not measured explicitly, the run time of all 5 models are similar and the fitting process is completely in several seconds. Hence, with a similar computation efficiency, we conclude that the Gradient Boosting

model is the overall best model to be used with principal components in this dataset. In the next section, we shall examine the result of these models using the full dataset.

3.3 Model Fitting and Prediction Based on the Full Dataset

Even though there are 81 variables in the dataset, which may be computationally greedy, it would still be plausible and worthwhile to compare the PCA based model with the models based on all available variables. The same model statistics are obtained and displayed in Table 3:

Method	Est. Mean Squared Error	R^2	Box-Cox Inversed
LASSO	2.30	0.73	6.24
AdaBoost	2.86	0.65	8.83
Gradient Boosting	1.23	0.87	2.93
XGBoosting	3.89	0.54	15.58
Linear Regression	1.93	0.77	4.90

Table 3: Summary statistics of the regression models based on the full dataset.

From the regression result, we can see that the regression result for all models based on the full dataset gained some improvement compare to the model fitted with only the principal components. Further, the Gradient Boosting method again gives the most prominent result. It fitted the dataset exceedingly well with a coefficient of determination $R^2_{GB} = 0.87$. In addition, if we look at the Box-Cox Inversed mean squared error of the Gradient Boosting model, we can see that the prediction is off by only 1.71 Kelvins on average, which means that the predicted value is already a very precise estimate of the actual critical temperature. Such a accurate estimate can significantly save the research time and effort that physics researchers need to spend to get an idea of the critical temperature of the substance. The researchers are able to obtain a estimated critical temperature based on the Gradient Boosting model and then slightly change the temperature through experiment to get the actual critical temperature of the substance. The worst model again being the XGBoosting regression model, the prediction of which is off 3.94 Kelvins on average and the model still does not fit the dataset very will. The coefficient of determination of this model is only $R^2_{XG} = 0.54$. Another interesting topic is the multiple linear regression model in this case. Theoretically, LASSO is capable of being used on such a multivariate dataset without the need for model selection as the shrinkage feature of the loss function itself is equivalent to conducting model selection while fitting the data. Linear regression, on the other hand, do not have such feature and the resulting model may have overfitting issue.

Therefore, to see if there is any room of improvement with the multiple linear regression model, we conduct model selection using the recursive feature elimination method. We start with a reduced model with only 10 variables and increment 5 variables at a time. Based on the model selection result, the 10 most important variables for the reduced multiple linear regression model are listed in Table 4.

wtd_entropy_atomic_mass	entropy_fie
wtd_entropy_fie	entropy_atomic_radius
wtd_entropy_Density	wtd_entropy_ElectronAffinity
wtd_entropy_FusionHeat	entropy_Valence
wtd_entropy_Valence	wtd_std_Valence

Table 4: The top 10 variables selected by the recursive feature elimination algorithm.

Observe that the 10 most important variables for the multiple linear regression model is utterly different from the top 5 variables that contribute the most to the determination of the leading principal components. This insight tells us that the result of correlation-based PCA is not directly related to the result of linear regression. In the case of linear regression, the variables that are most correlated to the response variable is selected where as the PCA invetigate the overall correlation the each variable with respect to all other variables in the dataset. To better access the result of model selection,

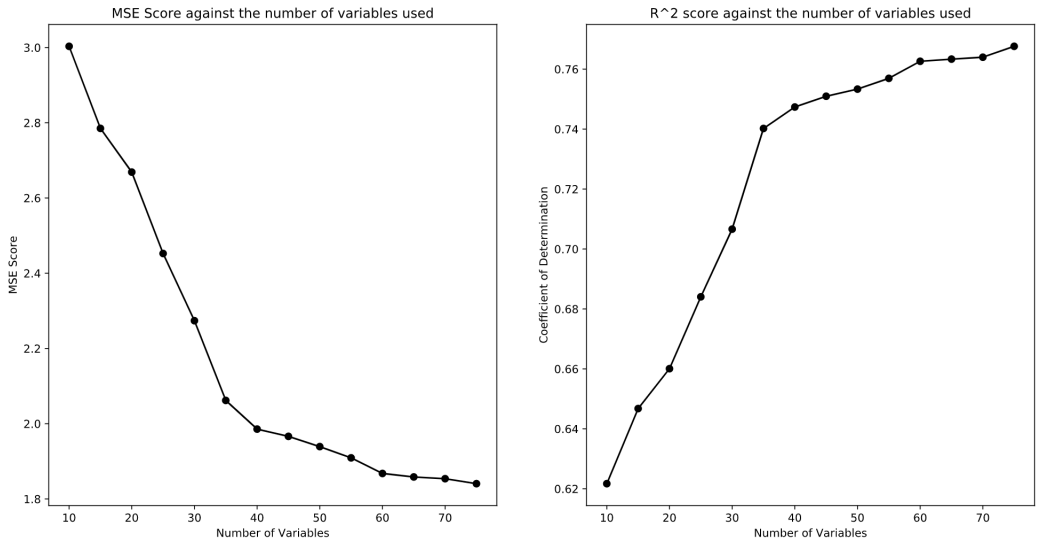

Figure 5: Model selection result for the multiple linear regression model.

the result of the grid search for the best reduced model is visualized in Figure 5. Unlike most model selection result where there is a sweet spot occurred at the model with a certain number of variables, the model selection result shows that the more variable the model has, the better predicting result the regression would give. This is unorthodoxical but it also implies that each individual variable in the data set contains information related to the critical temperature that other variables to not have. Thus, the conclusion drawn from the model selection result is the the full linear regression model is the best model and there is no room for improvement in this sense.

As we have seen from the regression result based on the PCA and the full dataset, the Gradient Boosting regression model gives the best prediction result and it has met our expectation for the precision of predicting the critical temperature. Hence, considering the sparsity of the *unique_m* dataset, we decided not to use the data for regression of only focus on the *train* dataset.

4 Classification and Regression

As mentioned in the exploratory data analysis section, the variance within the response variable, critical temperature, is quite large, making us wondering if there is potential clustering that us to assign each substance to a certain category according to its critical temperature and then fit regression model within each group. Here, we will examine this and make comparisons of the result with the regression-alone approach.

First of all, since the model selection and regression analysis in the previous regression section has informed us that fitting regression model based on the full dataset gives the best result. Therefore, we skip the investigation on principal component analysis and directly make use of all the available variables.

To assign labels according to the values of critical temperature, we first obtain the 33% quantile and the 66% quantile of the data and then consider the substances with critical temperature below the 33% as *Low Temp* (critical temperature between 0K to 61.67K); substances in between the quantiles are considered as *Mid Temp* (critical temperature between 67.67K to 123.33K); substances with critical temperature above the 66% are labeled as *High Temp* (critical temperature between 123.33K to 185K). Since the distribution of the critical temperature variable is skewed to the right, the labeling is also effected by such skeweness. There are a total of 15852 substances being labeled as *Low Temp*, 5235 substances have label *Mid Temp*, and 176 substances assigned label *High Temp*.

To classify the 3 labels, 3 classification models are attempted. The K-Nearest-Neighbors classifier gives an accuracy rate at around 91.5%; the AdaBoosting classifier has accuracy rate at around

82.8%; the Radom Forest Classifier gives an accuracy rate at around 93.9%. The confusion matrix of the Random Forest classification result is shown in Table 5:

	Low Temp	Mid Temp	High Temp
Low Temp	3787	167	0
Mid Temp	135	1184	8
High Temp	1	12	22

Table 5: The confusion matrix of the Random Forest classification result. The rows are true labels and the columns are predicted labels.

The confusion matrix shows that the *High Temp* category is the hardest to be classified where as the other two categories are relatively simple and the accuracy rate is ideal. Hence, based on the overall performance, we decided that the Random Forest classifier is the best classifier in this case.

With an ideal classification algorithm, we now fit the Gradient Boosting as well as the multiple linear regression model to the data of each group. The two regression models considered here are the multiple linear regression model and the Gradient Boosting regression model, both of which give some of the best result in the previous section. In this case, the Gradient Boosting regressor still gives the best regression result.

Before fitting the model, the density of the response variable in each group is double checked to see if there is any need for Box-Cox transformation.

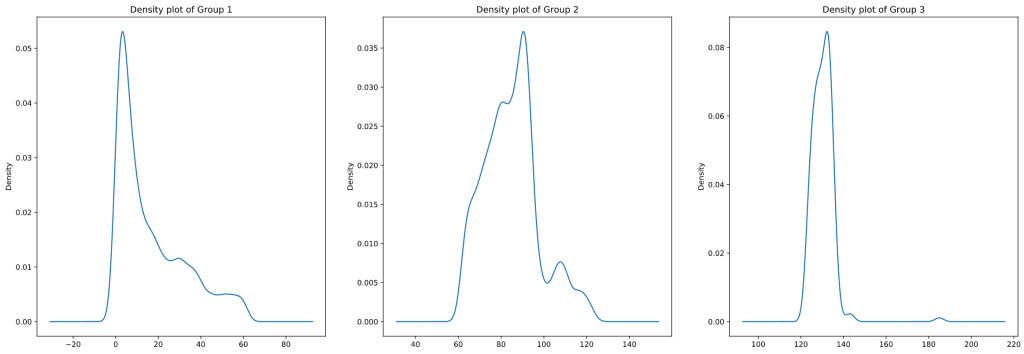

Figure 6: Density plot of the response variable in all 3 Groups. Right: density plot of response variable in Group 1; Middle: density plot of response variable in Group 2; Left: density plot of response variable in Group 3.

From Figure 6, we can see that the density of the critical temperature of each individual group is roughly normally distributed. Hence, Box-Cox transformation is not applied in this case and the regression model is directly fitted.

Group	Est. Mean Squared Error	R^2
Group 1	63.77	0.77
Group 2	71.01	0.65
Group 3	16.81	0.93

From Table 6 we can see that the Gradient Boosting model do not have a good fit on Group 1 and Group 2, which correspond to data with low critical temperature and medium critical temperature. Surprisingly, the model fits Group 3 data exceedingly well with a coefficient of determination $R^2_{Group3} = 0.93$. This information provides us with the insight that the physical characteristics of the substances with high critical temperature is more linearly correlated with the critical temperature. From the mean squared error, we see that the prediction of critical temperature of Group 1 substances is off by 7.98 K on average. The average critical value in Group 1 is 16.67 K. Hence, the

prediction does not provide a very accurate estimation in general. The prediction of critical temperature of Group 2 substances is off by 8.43K on average. Taking into consideration that the average critical temperature for Group 2 substances is 84.93K, the estimation here is much more accurate than that of Group 1. The estimation of critical temperature of substances from Group 3 is the most prominent. The estimation here is off by 4.1 K on average whereas the average critical temperature of substances from Group 3 is 130.32K.

5 Conclusion

To conclude this study. We can see that although the model fitted on the whole dataset seems to provide a good estimation since the estimation is only off by 2.93K on average in the case of Gradient Boosting regressor, the result from the classification-regression approach shows that the estimation result is only adequate for substances with low critical temperature, which is the majority of the substances (74.55% of the dataset). Hence, the first approach misguided our study and provided us with a false confidence with our model. The classification-regression approach here showed that the Gradient Boosting regressor works well only on data with medium to high critical temperature, which is 25.45% of the dataset. The regression model performs especially well on the substances with high critical temperature and the estimation or prediction is exceedingly accurate, suggesting that such approach can be applied to study substances with high critical temperature. Indeed, for low critical temperature substances, although testing the critical temperature can be laborious, it is plausible as the critical temperatures can be easily achieved with modern heating equipments. Heuristically, we assume that the actual complication of studying the critical temperature lies in the case of substances having medium to high critical temperature. These extremely high temperatures could be hard to attain using modern heating equipments, thus the measuring process could be both tedious and costly. Thus, the Gradient Boosting regression model is the felicitous method to aid the prediction of the critical temperature of such substances as it is both computationally efficient and gives very precise and accurate estimation.

For future studying, we suggest that the transformation of the response variable from the original dataset shall be further studied as the Box-Cox transformation did not give a very good result, which in turn led to a moderate violation of the assumptions of multiple linear regression. Further, the unorthodoxical phenomenon occurred during model selection for multiple linear regression where adding more variable to the model will always improve the predictive power shall be further studied. Last but not the least, the Random Forest classifier had a hard time distinguish substances of medium critical temperature and that of high critical temperature. The accurate classification result is the foundation for regression prediction in the classification-regression approach. Thus, alternative methods of classifying medium critical temperature and high critical temperature is also demanded.

References

[1] Wikipedia contributors. Superconductivity — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Superconductivity&oldid=960523366, 2020. [Online; accessed 10-June-2020].

[2] Paul Preuss. Superconductors face the future, Sep 2010.

[3] Wikipedia contributors. Power transform — Wikipedia, the free encyclopedia, 2020. [Online; accessed 11-June-2020].

6 Acknowledgements

This is the last project in my undergraduate study and I hope that it is appropriate to write a little acknowledgements to convey some of my feelings. It is such a mixed feeling to write down the last work and proof read the report. It feels like I am still the freshman who knows nothing about statistics yesterday and I am now submitting my last undergraduate project.

I would like to express my utmost appreciation for Prof. Fushing Hsieh and Xiaodong Wang. The diverse topics covered in the lecture as well as the useful applied knowledges contained in each discussion section has been very helpful with improving my statistical thinking.

Wish we all strive for a better world and prosper in our field in the future.

Appendix

Variables available in the *train* dataset: 'number_of_elements', 'mean_atomic_mass', 'wtd_mean_atomic_mass', 'gmean_atomic_mass', 'wtd_gmean_atomic_mass', 'entropy_atomic_mass', 'wtd_entropy_atomic_mass', 'range_atomic_mass', 'wtd_range_atomic_mass', 'std_atomic_mass', 'wtd_std_atomic_mass', 'mean_fie', 'wtd_mean_fie', 'gmean_fie', 'wtd_gmean_fie', 'entropy_fie', 'wtd_entropy_fie', 'range_fie', 'wtd_range_fie', 'std_fie', 'wtd_std_fie', 'mean_atomic_radius', 'wtd_mean_atomic_radius', 'gmean_atomic_radius', 'wtd_gmean_atomic_radius', 'entropy_atomic_radius', 'wtd_entropy_atomic_radius', 'range_atomic_radius', 'wtd_range_atomic_radius', 'std_atomic_radius', 'wtd_std_atomic_radius', 'mean_Density', 'wtd_mean_Density', 'gmean_Density', 'wtd_gmean_Density', 'entropy_Density', 'wtd_entropy_Density', 'range_Density', 'wtd_range_Density', 'std_Density', 'wtd_std_Density', 'mean_ElectronAffinity', 'wtd_mean_ElectronAffinity', 'gmean_ElectronAffinity', 'wtd_gmean_ElectronAffinity', 'entropy_ElectronAffinity', 'wtd_entropy_ElectronAffinity', 'range_ElectronAffinity', 'wtd_range_ElectronAffinity', 'std_ElectronAffinity', 'wtd_std_ElectronAffinity', 'mean_FusionHeat', 'wtd_mean_FusionHeat', 'gmean_FusionHeat', 'wtd_gmean_FusionHeat', 'entropy_FusionHeat', 'wtd_entropy_FusionHeat', 'range_FusionHeat', 'wtd_range_FusionHeat', 'std_FusionHeat', 'wtd_std_FusionHeat', 'mean_ThermalConductivity', 'wtd_mean_ThermalConductivity', 'gmean_ThermalConductivity', 'wtd_gmean_ThermalConductivity', 'entropy_ThermalConductivity', 'wtd_entropy_ThermalConductivity', 'range_ThermalConductivity', 'wtd_range_ThermalConductivity', 'std_ThermalConductivity', 'wtd_std_ThermalConductivity', 'mean_Valence', 'wtd_mean_Valence', 'gmean_Valence', 'wtd_gmean_Valence', 'entropy_Valence', 'wtd_entropy_Valence', 'range_Valence', 'wtd_range_Valence', 'std_Valence', 'wtd_std_Valence'

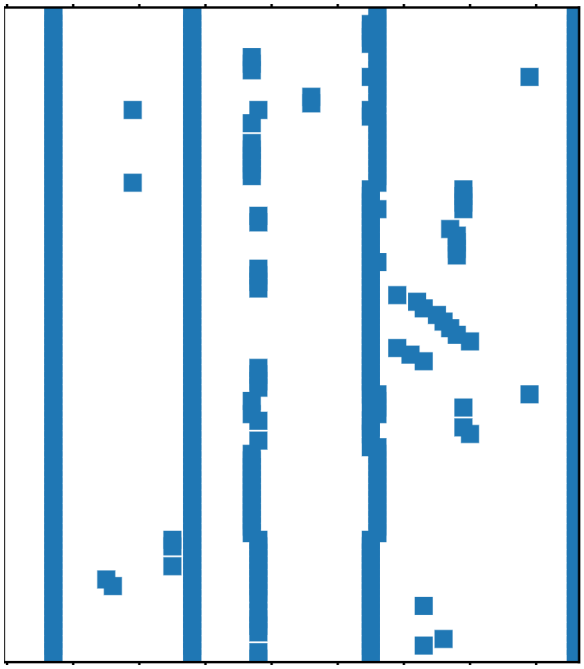

Figure 7: Visualization of the sparsity of the *unique_m* data matrix.