Superconductor Critical Temperature Prediction

Sanjay Shanbhag
Dept of Computer Science
University of Central Florida
Oviedo, United States of America
Sanjayshan@knights.ucf.edu

Asst Prof Rui Xie
Dept of Statistic, Data Science
University of Central Florida
Oviedo, United States of America
Rui.Xie@ucf.edu

***Abstract*— The Project is mainly about the semiconductor critical temperature prediction. The critical temperature is a temperature at which the semiconductor conducts with zero resistance. Hence it is very important to predict the if not very correct but at least the nearest temperature for any given semiconductor. The scope of this project is to perform the analysis on the underlying data and finally make predictions on the same data using the statistical methods.**

Keywords—Statistical Analysis, Regression, Random Forest

I. INTRODUCTION

Semiconductor's have always been popular due to their interesting property of partial conductivity. Due to this particular property, it finds its application in variety of domains ranging from automobile industry to power electronics. These semiconductors act as a super conductor and provide zero resistance to the flowing current. This interesting property is only seen at a temperature called as a Critical temperature. This property explains that the current flows in the wire without any loss. Hence to find its applications, it is very critical to know about its critical temperature. In this paper we mainly talk about the prediction of this critical temperature using a combination of two of the datasets consisting of the chemical properties and formula of the Semiconductor material.

II. DATA ANALYSIS

A. Dataset and Descriptive Analysis

As mentioned earlier, to perform the critical temperature prediction of the semiconductor using Superconductor Data Dataset provided by the centre of Machine learning and Intelligent Systems at university of Pennsylvania.The following dataset consists of 2 components namely train.csv and unique.csv. The train.csv purely consists of the properties and statistics that affect the critical temperature. Finally, the unique.csv corresponds to the chemical formula of the semiconductor material. Both the datasets include 21,263 entries but differ by number of predictor variables in them i.e., the train.csv includes 82 features whereas the unique.csv consists of 88 features. Below attached is the snapshot of the descriptive analysis of the given data. The descriptive analysis includes the count, mean, standard deviation, interquartile range of the dataset.

The first dataset fully includes the numerical values and are pertaining to the physical or the chemical properties of the semiconducting material. Hence the descriptive analysis includes slightly higher values as compared to the second dataset, which mostly talks about the count of the chemical elements present in the final semiconductor material. Fig 1. and Fig 2. majorly describe the data in statistical terms.

	count	mean	std	min	25%	50%	75%	max
number_of_elements	21263.0	4.115224	1.439295	1.000000	3.000000	4.000000	5.000000	9.0000
mean_atomic_mass	21263.0	87.557631	29.676497	6.941000	72.458076	84.922750	100.404410	208.9804
wtd_mean_atomic_mass	21263.0	72.988310	33.490406	6.423452	52.143839	60.696571	86.103540	208.9804
gmean_atomic_mass	21263.0	71.290627	31.030272	5.320573	58.041225	66.361592	78.116681	208.9804
wtd_gmean_atomic_mass	21263.0	58.539916	36.651067	1.960849	35.248990	39.918385	73.113234	208.9804
...	...	...	...	...	...	...	...	...
range_Valence	21263.0	2.041010	1.242345	0.000000	1.000000	2.000000	3.000000	6.0000
wtd_range_Valence	21263.0	1.483007	0.978176	0.000000	0.921454	1.063077	1.918400	6.9922
std_Valence	21263.0	0.839342	0.484676	0.000000	0.451754	0.800000	1.200000	3.0000
wtd_std_Valence	21263.0	0.673987	0.455580	0.000000	0.306892	0.500000	1.020436	3.0000
critical_temp	21263.0	34.421219	34.254362	0.000210	5.365000	20.000000	63.000000	185.0000

Fig 1. Description of the First Dataset

B. Visualization and Analysis

In the previous section, a brief overview of the dataset was given. Going ahead in the direction of data modelling it was very important to thoroughly visualize the data and therelation with the respone variable. Various methods were utilised to visualize the data, which include scatter plot shown in the Fig 3. The Fig 3. corresponds to the first dataset and was plotted w.r.t to the predictor variables vs response variables.

	count	mean	std	min	25%	50%	75%	max
H	21263.0	0.017685	0.267220	0.0	0.0	0.0	0.0	14.0
He	21263.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0
Li	21263.0	0.012125	0.129552	0.0	0.0	0.0	0.0	3.0
Be	21263.0	0.034638	0.848541	0.0	0.0	0.0	0.0	40.0
B	21263.0	0.142594	1.044486	0.0	0.0	0.0	0.0	105.0
...	...	...	...	...	...	...	...	...
Pb	21263.0	0.042461	0.274365	0.0	0.0	0.0	0.0	19.0
Bi	21263.0	0.201009	0.655927	0.0	0.0	0.0	0.0	14.0
Po	21263.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0
At	21263.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0
Rn	21263.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0

Fig 2. Description of the second dataset

In the direction of undersatnding the data, it was important to understand the distribution of the critical temperature. From the above Fig 3. It is clear the distribution is more biaed

Identify applicable funding agency here. If none, delete this text b

towards the 0K cases as compred to the varied range of critical temperatures. If it were a classification, it would have meant that the model would be more inclined towards the 0K class and hence would be inefficient in predicting other classes.

Apart the distribution plot, various other visualizations were also made to understand the data well. Few of them include Principle component analysis of the attributes in the dataset. The theory states that the components corresponding to the Eigen vector with larger eigen value are nothing but the principle components. The eigen values obtained were normalized to make all the components visible in the graph in the sorted order.

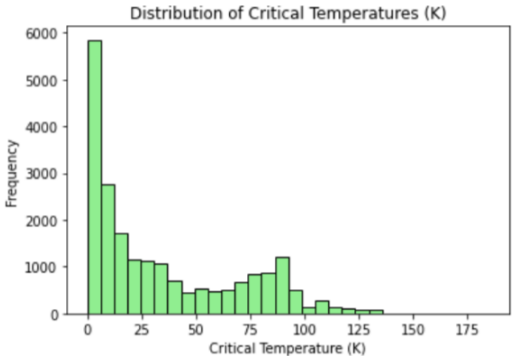

Fig 3. Distribution of Critical temperature (K)

Details of the PCA has not been added in this context due to this constraint, but will be accomodated in the Appendix section of this paper.

Finally as we already understand the importance of the variance and the correlation analysis, these were performed on both the given datasets. It's a common practice to consider all the columns with certain variance, as it it considered to contain more information as compared to the one with the lesser variance.Secondly, the correlation analysis was peformed between the predictor variables and the response variables, to soley understand the relation between them as compred to the inter-predictor variable relations. Fig 4 potrays the variance in the dataset where as the Fig 5.talk about the correlation analysis in terms of the bar plot.

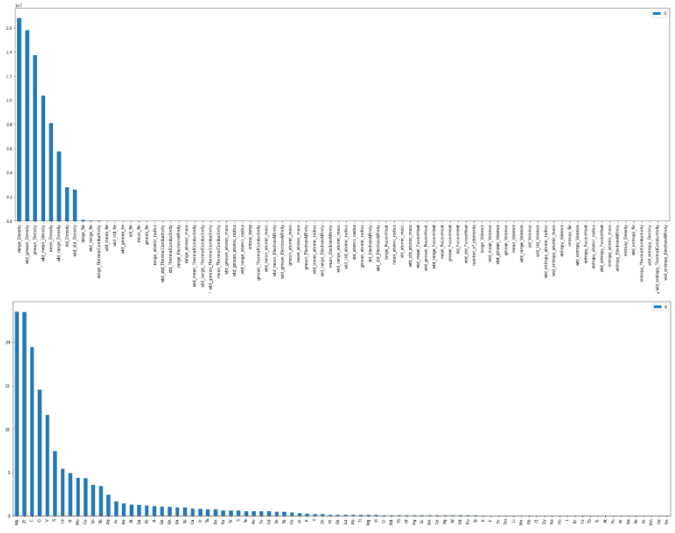

Fig 4. Variance in (Top) first and (Bottom) second dataset

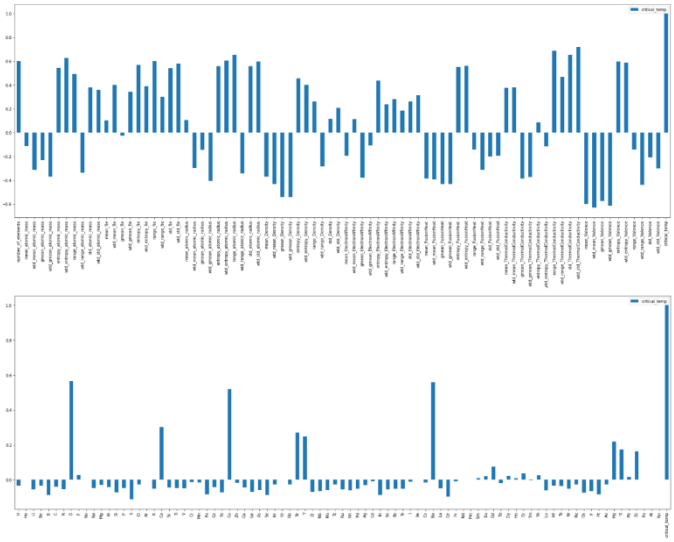

Fig 5. Correlation Analysis on (Top) first and (Bottom) second Dataset

III. Data Modelling Methodologies

In this section we will mainly talk about the modelling methodologies required to predict the accurate critical temperature. In the previous section we have purely discussed about the relation between the various predictor variables and the response and also most visualized the data. Now that we understand the data thoroughly, we will focus on training the models using Linear Regression and Random Forest Regressor.

A. Linear Regression

Coming to Linear regression, steps were performed while running Linear Regression on the data

1. It was noticed that for few of the features, the values are sufficiently apart, hence to get them all in the same range, Standardization was performed.
2. Next step was to understand if the principal components found in the data analysis are representative of the dataset. Hence PCA was performed with varied number of components.
3. Once the features and the labels were made available, the following step was to train the model using Linear regression using a 5-fold cross validation approach.
4. Finally, the model was evaluated using the R2 statistic and the Mean-squared-error. And a line was fit to understand the prediction of the model.

Linear Regression was purely used as a baseline model. A detailed report was generated and included the following

1. MSE for cross validated method
2. MSE for the full data
3. R2 statistic for the cross validated method
4. R2 statistic for the full data
5. Number of features used in training the model.
6. Intercept of the line
7. Regression coefficients

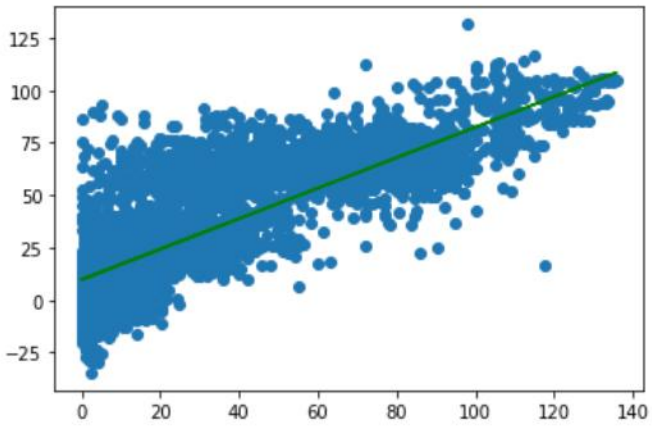

Fig 6. Line fitting on the Linear Regression model

The generated report for the regression analysis is not attached in this due to the constraint. But a data table Table I. and Table II. has been added to visualize the outcome w.r.t the first and the second dataset.

Table I. Outcome of Linear regression for the first dataset

PCA Features	R^2 Statistic	MSE
20	0.5876	483.81
40	0.6529	407.20
80	0.6922	361.12

Table II. outcome of Linear Regression for the second dataset

PCA Features	R^2 Statistic	MSE
20	0.429	669.82
40	0.444	651.49
80	0.3866	718.44

Looking at the Table I and II, it is evident that the R^2 Statistic is gradually increasing and MSE is decreasing as a result of increase in the number of features in the training model. And if we compare the performance of the model on both the datasets, it seems like the critical temperature is more related to the First dataset as compared to the second one. Also looking into the Fig 6. it is evident that the model is underestimating the high temperature as compared to the low temperature. Hence we will explore some slightly more flexible model like Random Forest.

B. Random Forest Regressor

Random Forest has always been raised as the winner in various Kaggle competitions. Hence the Radom Forest regressor was chosen to check its performance on the underlying dataset. In here considering the results from the Linear Regression, PCA was dropped and also the modelling is performed on the first dataset, as the second dataset is less correlated with the response variable.

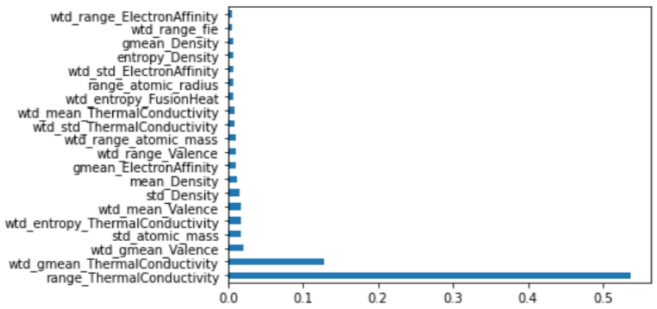

Fig 7. Feature Importance derived from the RFR

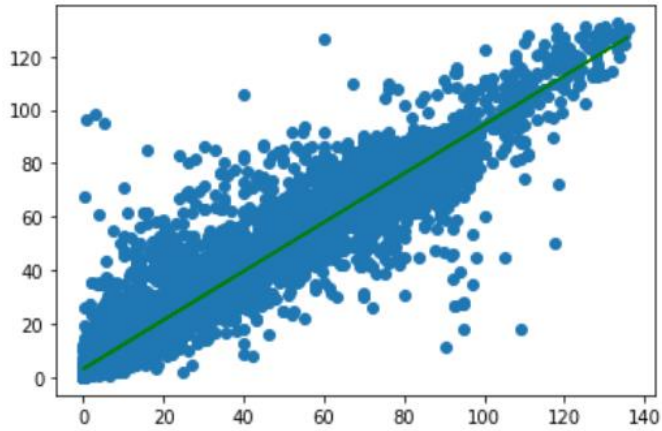

Fig 8. Line fitting on the Random Forest Regressor

Due to the large computation required for the training of Random Forest, the model was trained on a train split and tested on a test split, and not using the cross-validation method. Finally, once the model was trained, the feature importance was derived from the random forest and depicted in the figure Fig 7. below. It was noted that the following predictor variables were allotted high importance by the Random Forest algorithm

1. Range_ThermalConductivity
2. Wtd_gmean_ThermalConductivity
3. Wtd_gmean_valence

And coming to the outcome of the random forest regressor, it made a significant difference in terms of the MSE and R^2 Statistic. The outcome is depicted in the Table III below. Also the line fitting presented in Fig 8. Is much more promising than the Fig 6.

Table III. outcome of Random Forest Regressor

Features	R^2 Statistic	MSE
80	0.8471	179.35

C. Conclusion

A thorough data analysis was performed on the underlying dataset in terms of the correlation, variance, Principe component analysis, Critical temperature distribution. Following which the data modelling was done, which involved

the analysis of the effect of PCA features on the regression. It was concluded that entire set of predictors were required to get the best results on the data. Alongside, the second dataset consisting of the chemical formula was declared as not very well correlated with the Critical Temperature. The outcome from the Random Forest revealed that it performed exceptionally well on the underlying dataset in terms of the R^2 Statistic and MSE, paving the path towards better prediction. Also, on referring it was found out that line fitting is the best way to see the residual errors of the model, hence the line fitting was done and it verified the performance of the Random forest over the Linear Regression model.

D. Future Scope

As mentioned in the reference paper attached below, XGBoost can be used to make better predictions.

REFERENCES.

[1] 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].

APPENDIX

1. The code that was utilized perform the various analysis and the modelling is attached in the form of link below.

   https://www.kaggle.com/code/sanjayshan/semiconductor-critical-temperature-prediction

2. Additional Visualizations and Analysis performed during the course of this project has been attached for the better understanding of the dataset.

   Appendix.docx