Seminar Report (MM 694)

On

DATA DRIVEN APPROACH TO PREDICT THE CRITICAL TEMPERATURE OF A SUPERCONDUCTOR

Submitted by

RAJRISHI SARKAR
(203110061)

In partial fulfillment for the award of the degree

Of

MASTER OF TECHNOLOGY

In

METALLURGICAL ENGINEERING AND MATERIALS SCIENCE

Guided by

Prof. HINA AMOL GOKHALE

INDIAN INSTITUTE OF TECHNOLOGY, BOMBAY

Powai, Mumbai- 400076

APRIL 2021

CONTENTS:

ABSTRACT:

We know that superconductivity has been the focus of research for quite some time now. But the features responsible for this particular phenomenon remains quite mysterious. In this study we review some past papers which tries to estimate the superconducting critical temperature on the basis of a statistical model. Features were extracted based on atomic radius, atomic mass, thermal conductivity etc. To improve the model accuracy even new features were incorporated using the materials data from the AFLOW online repositories. Based on RMSE the models showed reasonable predictions with a deviation of 9.5K to 10K. The models don't say much about superconductivity. It only predicts the critical temperature of the superconductors.

LIST OF FIGURES:

INTRODUCTION:

Materials that conduct current without offering any resistance are called superconductors, they have several applications. One of the applications would be Magnetic Resonance Imaging (MRI) used vastly in the field of medical science. Among other applications, would be Large Hadron Collider (LHC) and SQUIDS. Although superconductivity is a very interesting area to explore but there are two major drawbacks:

a) It only behaves like a superconductor below a certain critical temperature (Tc) which is impractical and difficult to achieve.

b) The model that is responsible for predicting Tc is an open problem which has become quite unfathomable in the scientific community.

Application of statistical methods in order to predict the superconductivity started in the early 1980s with simple clustering methods. But due to absence of any theoretical solid models just simple empirical rules which are based experimental results have helped the scientists for several years. Say, the physicist Matthias concluded that Tc is related to the number of valence electrons per atom, is one of the empirical rules. Many of these empirical rules are violated.

In this report we look into the papers which takes entirely Data-driven approach to create a statistical model that predicts the critical temperature Tc. The required data is extracted from the superconductivity material Database maintained by NIMS in Japan. Features are derived based on the elemental properties of the superconductor. Several Features are derived from the existing features creating a total of 81 features for each superconductor. Off course all the properties are not required for the model. Various statistical models were tried out in different papers but the most efficient are: A multiple regression model and a gradient boosting model. These are the models that generated maximum accuracy. Unlike regression based models the Gradient bosting models creates an ensemble of trees to predict Tc. After fitting a gradient boosting model the weighted average of all the trees are taken into consideration that finally gives the required prediction. It is found that in maximum cases the gradient boosting model outperformed the regression model as they were able to account for the complex correlations among the features. The multiple regression model shows an root mean squared error (R-squared) of 0.75where as the XG boost model delivered an R-squared value of about 0.95.

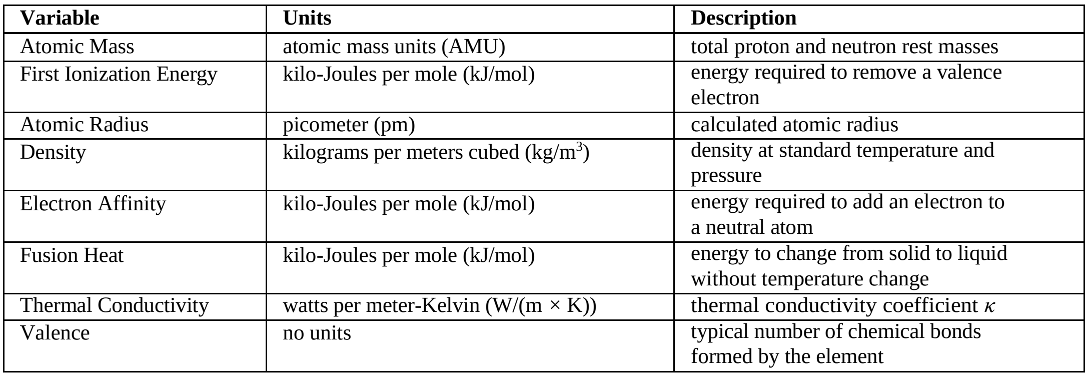
LITERATURE SURVEY:

From the above discussion it is evident that statistical modelling has been a keen interest of many researchers to predict the critical temperature of the superconductors. Although we discussed about the best two models for the same but the main issue with these approaches is feature selection, after extracting them. Because out of all those features many will show correlation that will affect the accuracy of the models. This means various data pre-processing steps are required before fitting the required model into the data, followed by model tuning to improve the accuracy. We look into various literatures in order to figure out the best features and several data engineering techniques adopted to generate the best output i.e a model of maximum accuracy. All of those approaches are summarized below:

Data Pre-processing:

Stage 1: Element Data Preparation

The data consists of total 86 features i.e that many rows. Conder (2016) is a big help for choosing the important properties out of all those features but different paper uses different properties as per they see fit. Variables like “boiling point” are dropped and the corresponding feature which is correlated to it is used, like “Fusion heat” variable. After several trial and errors and gaining various insights literature survey suggests that there are 8 important variables that are the most important ones and hence used in the statistical model.



Variable	**Units**	**Description**
Atomic Mass	atomic mass units (AMU)	total proton and neutron rest masses
First Ionization Energy	kilo-Joules per mole (kJ/mol)	energy required to remove a valence electron
Atomic Radius	picometer (pm)	calculated atomic radius
Density	kilograms per meters cubed (kg/m^3)	density at standard temperature and pressure
Electron Affinity	kilo-Joules per mole (kJ/mol)	energy required to add an electron to a neutral atom
Fusion Heat	kilo-Joules per mole (kJ/mol)	energy to change from solid to liquid without temperature change
Thermal Conductivity	watts per meter-Kelvin (W/(m × K))	thermal conductivity coefficient κ
Valence	no units	typical number of chemical bonds formed by the element

Table 1: The table is taken from Kam Hamidieh’s paper on Data driven approach to predict Tc of a superconductor. It shows the important features selected for predicting the critical temperature Tc[1].

Stage 2: Pre-processing of the Superconducting Material data:

As mentioned before, the Superconducting Material data is maintained by NIMS, a Japan based institution. This database is the ultimate source of the superconductors along with their critical temperatures and elemental formulas. The elemental formulas are used to generate the required feature. This is considered as the most authentic database. The features that were extracted from the database are mentioned in the table:

Feature & Description	**Formula**	**Sample Value**
Mean	$= \mu = (t_1 + t_2)/2$	35.5
Weighted mean	$= \nu = (p_1 t_1) + (p_2 t_2)$	44.43
Geometric mean	$= (t_1 t_2)^{1/2}$	33.23
Weighted geometric mean	$= (t_1)^{p1} (t_2)^{p2}$	43.21
Entropy	$= -w_1 \ln(w_1) - w_2 \ln(w_2)$	0.63
Weighted entropy	$= -A \ln(A) - B \ln(B)$	0.26
Range	$= t_1 - t_2\ (t_1 > t_2)$	25
Weighted range	$= p_1 t_1 - p_2 t_2$	37.86
Standard deviation	$= [(1/2)((t_1 - \mu)^2 + (t_2 - \mu)^2)]^{1/2}$	12.5
Weighted standard deviation	$= [p_1(t_1 - \nu)^2 + p_2(t_2 - \nu)^2)]^{1/2}$	8.75

Table 2: The table shows the process of feature selection from the material's chemical formulae[1]

Data Analysis:

This is the most important step towards model building as it helps to gain insight on how the features of the dataset are distributed, the correlation among the variables and the dependency of the target variable on the extracted features.

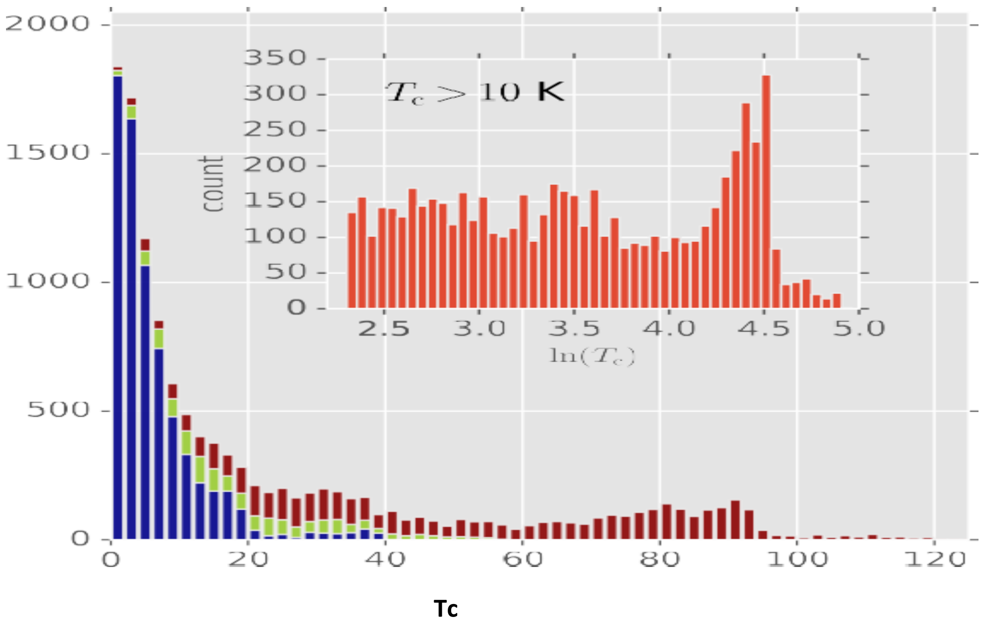


Fig 1: The Histogram of materials categorized by critical temperatures Tc. Blue, green, red denotes “low-Tc”, iron based and cuprate semiconductors[2].

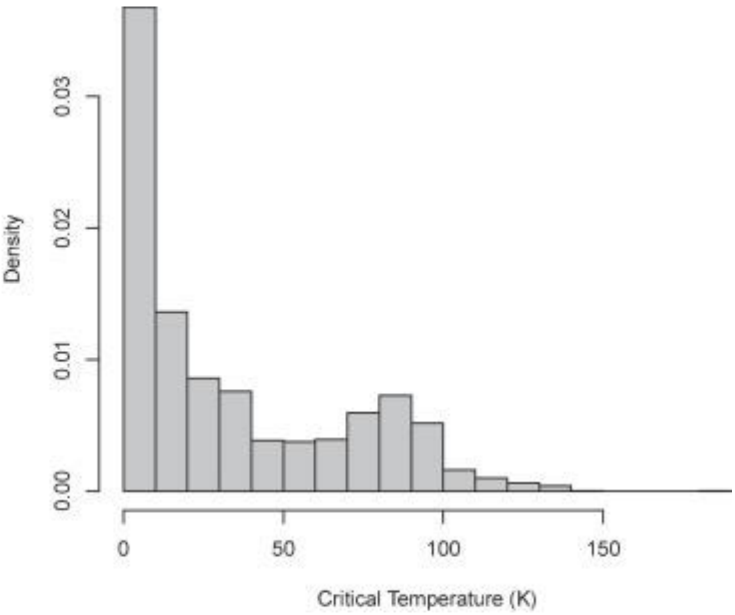


Fig 2: The above figure shows the distribution of critical temperatures Tc[3]

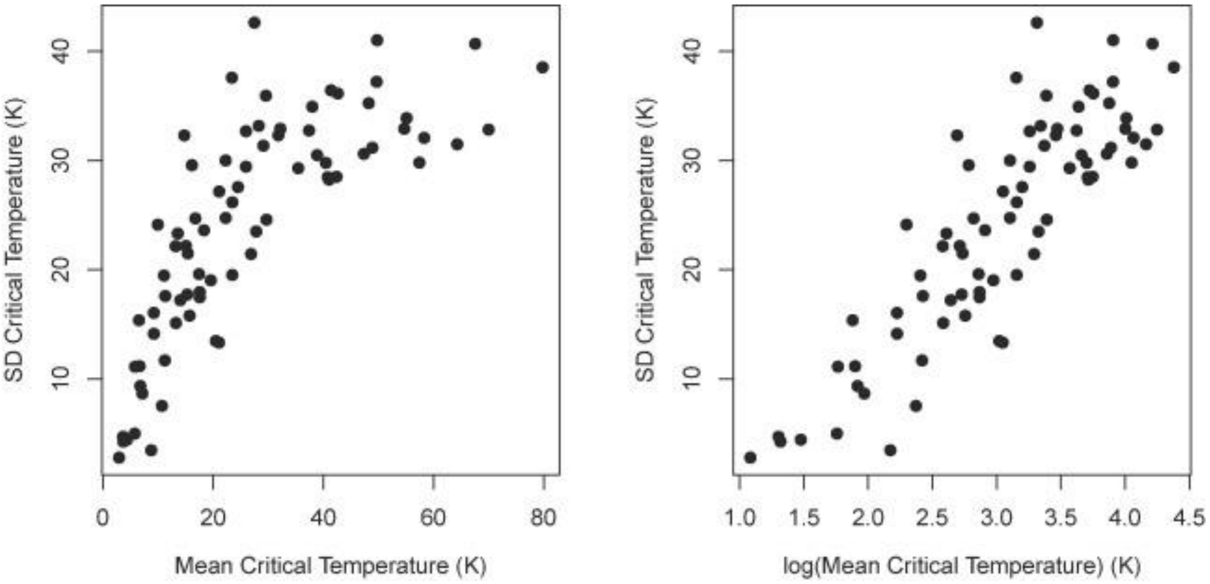


Fig 3: The first figure shows the relationship between the standard deviation per element and the mean critical temperature. The right panel shows the relation of log Tc with SD of the elements. It is evident that with the increase in mean critical temperature, value the standard deviation of critical temperature per element increases[4]

Model analysis:

Multiple Regression Model:

The R-squared value for multiple regression models lies between 0.70-0.76 approximately. The figure shown below shows the observed Tc versus the Predicted values of Tc for the superconductors Actually the multiple regression model under-predicts critical temperature of high temperature superconductors, not only that, it also overpredicts the critical temperature values for the low temperature superconductors. Although it is serves as the basis of other models it should not be used for predicting the critical temperature values.

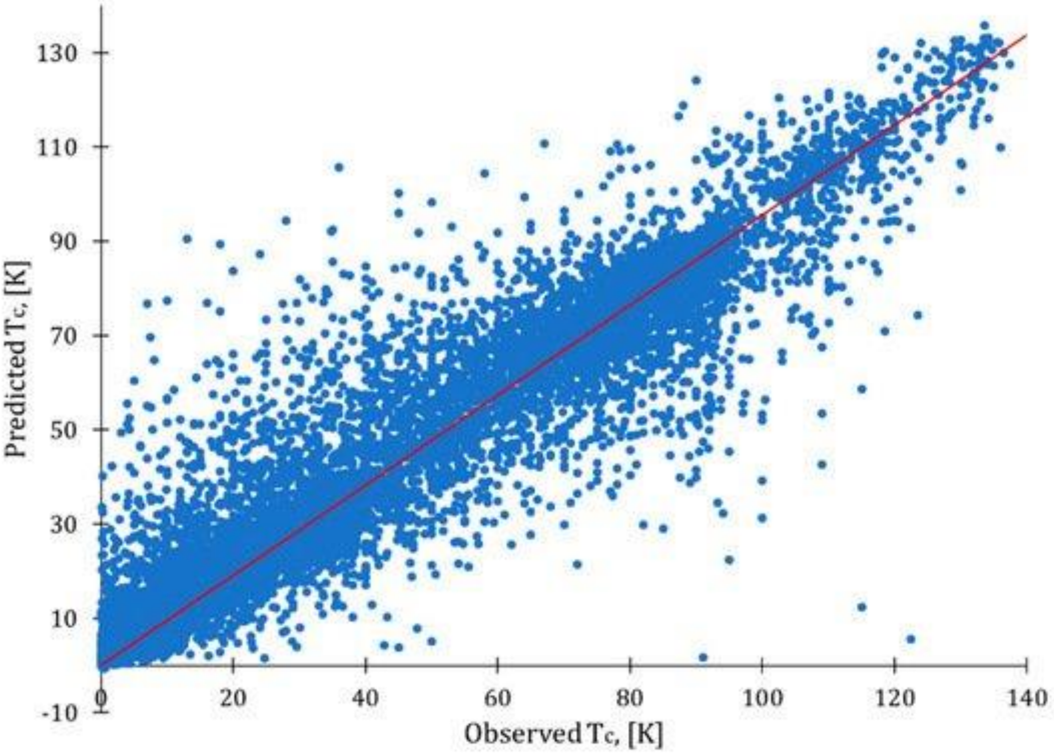


Fig 4: This is the plot of observed critical temperature vs the predicted critical temperature values based on the regression model[4]

XG-boost model:

The XG-boost model performs on[5]:

$$\begin{aligned}\mathcal{L}(\phi) &= \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k)\\ \text{where } \Omega(f) &= \gamma T + \frac{1}{2}\lambda\|w\|^2\end{aligned} \tag{2}$$

Now x_i is the input feature, y_i is the feature that we need to predict and $f_1 \ldots f_k$ is basically sequence of trees Here L is the expected loss function for the ith predicted response and Ω is the penalty function. T is the no. of leaves per tree. The penalty function in the XG-boost model is beneficial for increasing the accuracy of the model. We can also use parameter tuning to achieve higher level of accuracy and increase the efficiency of the model. This includes tuning the learning parameter η and also feature sub sampling, which means to choose a fraction of features randomly while adding a new tree at each stage. The figure below shows the observed critical temperature vs the predicted critical temperatures for XG-boost model with an R-squared value of 0.90 to 0.97 approximately.

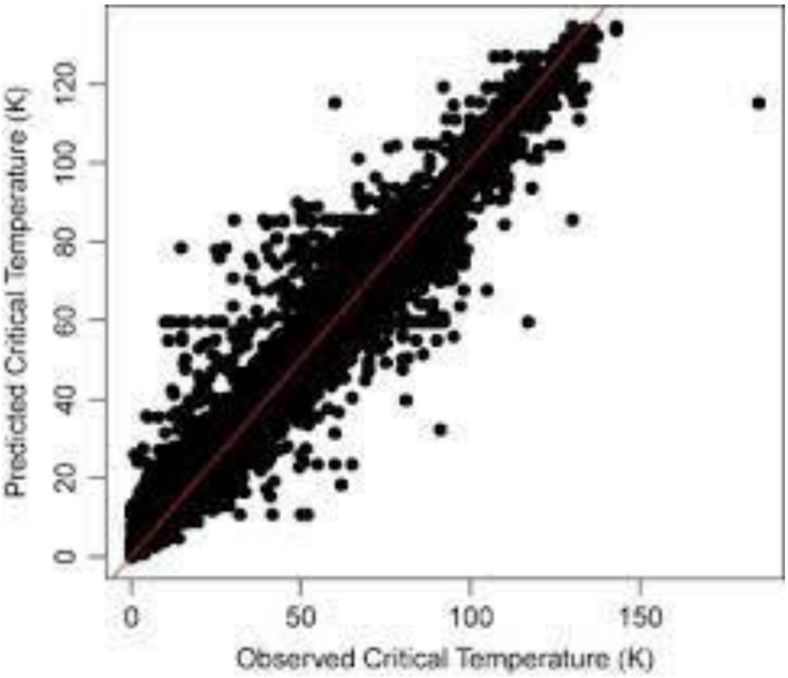


Fig 6: The plot between the predicted Tc versus observed value of Tc for XG-boost model[6]

Well certainly this model performs far better judging by the plots. The R squared value of regression model do not measure up-to the standards of the XG-boost model. Now we look into the most important features for the model. The features with maximum importance with respect to the given model can be calculated by the following formulae:

*Feature Gain = (Sum of gains for the feature) * (sum of gains for all the features)*$^{-1}$

Features having higher gain are more important. The features with most importance are selected for next model evaluation, dropping the ones with least gains. This in turn increases the model accuracy and the model XG boost model improvises.

Feature	**Gain**
range ThermalConductivity	0.295
wtd std ThermalConductivity	0.084
range atomic radius	0.072
wtd gmean ThermalConductivity	0.047
std ThermalConductivity	0.042
wtd entropy Valence	0.038
wtd std ElectronAffinity	0.036
wtd entropy atomic mass	0.025
wtd mean Valence	0.022
wtd gmean ElectronAffinity	0.021
wtd range ElectronAffinity	0.016
wtd mean ThermalConductivity	0.015
wtd gmean Valence	0.014
std atomic mass	0.013
std Density	0.010
wtd entropy ThermalConductivity	0.010
wtd range ThermalConductivity	0.010
wtd mean atomic mass	0.009
wtd std atomic mass	0.009
gmean Density	0.009

Table 3: The table shows the top features with maximum gain. Note that wtd=Weighted,std= standard deviation and gmean= geometric mean[6].

CONCLUSION:

In this report many studies on statistical modelling in order to predict the critical temperature of superconductors are reviewed. This work basically demonstrates the significant role statistical models play in superconductivity research. Several another databases may be used. Out of all the different models we narrowed down our focus to Multiple Regression and XG-boost models. Data preparation and Data analysis also plays a significant role to visualize the data and increase the efficiency of the models. Application of these models has the potential to enhance the search for high temperature superconductors.

REFERENCES:

1.Chen, T. and Guestrin, CXgboost: A scalable tree boosting system. https://arxiv. org/abs/1603.02754.(2016),135(4)

2.Chen, T., He, T., Benesty, M., Khotilovich, V., and Tang, Y. xgboost: Extreme Gradient Boosting. R package version 0.6.4.1.,(2018),34(8)

3.Chen, X., Huang, L., Xie, D., and Zhao, Q,Egbmmda: Extreme gradient boosting machine for mirna-disease association prediction. Cell Death and Disease,(2019), 9(3).

4.Conder, K. A second life of the matthiass rules. Superconductor Science and Technology,(2016),29(8).

5.Dick, J. M. Calculation of the relative metastabilities of proteins using the chnosz software package. Geochemical Transactions,(2018), 9(10).

6.Freund, Y. Boosting a weak learning algorithm by majority. Information and Computation, (1995),121(2), 256 { 285.