SUPERCONDUCTIVITY ANALYSIS

Student IDs:

251238730, <Student ID 2>, <Student ID 3>

MDA9160 - Statistical Modelling

Western University

December 1, 2021

**TABLE OF CONTENTS**

1. **Introduction3**
2. **Summary statistics and data visualization**………..…………………………………………………….**4**
3. **Methods of Analysis**……..………………………………………………………………………………………….**7**
4. **Results**.………..………………………………………………………………………………………………………….**14**
5. **Conclusion..**…………………………………………………………………………………………………………….**17**
6. **References..**…………………………………………………………………………………………………………….**18**
7. **Appendix**…………………………………………………………………………………………………………………**19**
8. **Introduction**

This report is created as part of t

1. **Summary Statistics**

The dataset used for the analysis is the Superconductivity dataset from the UCI machine learning data sets repository (https://archive.ics.uci.edu/ml/datasets/superconductivty+data). The dataset has information of 21263 superconductors and the critical temperature of the superconductor. The dataset has 81 features for the superconductors.

The goal is to create a model to predict the critical temperature of a superconductor. The features include mean atomic mass, geometric mean atomic mass , range atomic mass, mean density, thermal conductivity, electron affinity etc. It can be noticed that the many of the variables are highly correlated as these are variations of the calculations for the properties of the superconductors.

The initially step in the analysis involves looking into the distribution of the variables and finding out the relationships of the variables with the dependent variable.

**Analysis Objective**: Develop model to predict Critical Temperature of Superconductors

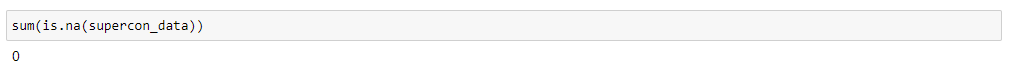
**Dependent Variable**: Critical temperature of a superconductor.

**Independent Variables**: Properties of superconductors such as its mean atomic mass, thermal conductivity etc.

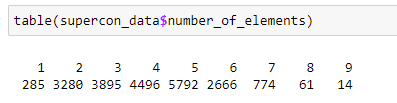
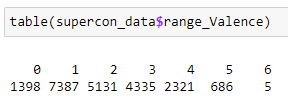
* 1. **Exploratory Data Analysis**

The objective of the exploratory data analysis is to deep dive into the datasets and derive insights on the variables in the datasets such as their distribution and discover patterns or anomalies. The major observations that can be made from the data are as follows:

* All the variables in the dataset are numerical.
* There are no missing values in any of the columns.



* **Number of Elements** and the **Range Valence** are in the ordinal scale and in the analysis we would be considering them without transforming into categorical or dummy variables as the relative ranking of the variable values may have an impact on the critical temperature.

* The dependent variable, critical\_temperature is moderately right skewed and has two peaks, 1 close to a critical temperature of 5 and another peak close to a critical temperature of 80.

Chart, histogram

Description automatically generated

* Then the scatter plot of the independent variables (superconductor properties) against the critical temperature of the superconductor is generated. Some of the interesting relations are as follows:
  + There is a high positive correlation between weighted entropy atomic mass and critical temperature and weighted standard thermal conductivity and the critical temperature.

A picture containing text, tree

Description automatically generated Chart, scatter chart

Description automatically generated

* + There is a high negative correlation between the critical temperature and the weighted mean valence.

Chart, scatter chart

Description automatically generated

1. **Methods of Analysis**
   1. **Outliers and Influential Points.**

* The presence of outliers causes the model to not fit well and removal of influential points which are outliers with a high leverage can cause a noticeable change in the fitted model.
* In the analysis, we cleaned the data by removing the influential points. The influential points were identified using the Cook’s distance. The criteria used was:

***Cook’s Distance > 4/n***

where n is the number of observations.

* After removing the influential points we tried to check the effectiveness of removing the outliers by checking the QQ and residual plotsfor the full model with all the variables before and after removal of the influential points.

Chart, scatter chart

Description automatically generated Chart, line chart

Description automatically generated

Before removal of the influential points

Chart

Description automatically generated Chart, line chart

Description automatically generated

After removal of the influential points

* We could observe improvements in the validation of the assumptions of the model after the removal of the outliers.
  1. **Multicolliearity analysis and creation of baseline model**

In the section for the exploratory data analysis we had seen some of the variables present in the dataset. Many of the variables are related as each of the different property has been transformed into different variables by applying different aggregation techniques on the variables. For instance, mean atomic mass of the elements is a feature that could impact the critical temperature. Among the variables we have the mean arithmetic mean, geometric mean, weighted means, weighted geometric mean etc of the features. This causes the variables to be correlated.

Chart

Description automatically generated with medium confidence

The correlation plot of the data set is given above. It can be observed that there exists high correlation among many variables. The high correlations are also grouped together in the correlation plot which shows that mostly the correlation is between the different aggregation variables created from the same feature.

Since there are a large number of variables to start with, we first focus on removing variables having very high correlation (>=0.75). Here among the correlated groups the variables that are introducing high correlation is removed. In the later sections, the additional techniques used for variable selection are discussed.

After removal of the variables based on correlation plot is as follows:

Chart, scatter chart

Description automatically generated

It can be observed that the correlation between the updated variables has reduced significantly.

After the correlation analysis was done, a first cut baseline model was created to understand the effect of the correlation based feature selection. For this, the data was split into train and test sets

Then a linear regression model was created on the new set of variables.

A screenshot of a computer

Description automatically generated

Text, table

Description automatically generated

The adjusted R squared value is 0.6599 which means 66% of the error in the y-terms is explained by the model.

The linear regression model has been created on the 26 variables. We can see that most of the variables except two of them are still not significant. This could be due to existence of multicollinearity still in the data as we have only removed variables that had a correlation >0.75.

Chart

Description automatically generated Chart, line chart

Description automatically generated

From the plots it can be seen that the normality assumption has been greatly improved by the use of the method. However in the there the homoskedacity assumption is not satisfied and some variable transformations might help with the same.

* 1. **Data Transformations**

Since we notice that the equal variance assumption of the residuals is not valid, we try to apply transformation on the response variable. Two different techniques were tried out:

* + 1. Taylor Expansion
* Since the variance is linearly increasing as seen from the residual plot in the previous section, we transform the response variable as :

y -> squareroot(y)

Chart

Description automatically generated Chart, line chart

Description automatically generated

* From the graph we can see that the Taylor expansion has caused the residuals plot to change slightly but there has been no significant improvemnet hence we adopt the box-cox transformation.
  + 1. Box-Cox Transformation
* In the Box-Cox transformation, we find out the optimal lambda value and use that for the creation of the model. For the data set considered we check the log likelihood values at various lambda levels and chose the value with max log likelihood. Here, the lambda value is set at 0.28
* Then following transformation is applied on y:

Diagram, schematic

Description automatically generated

Diagram

Description automatically generated

Choosing the optimal lambda: Here the optimal value is 0.29

* After transformation , we get the following plots for normality and the residuals.

Chart, scatter chart

Description automatically generated Chart, line chart

Description automatically generated

* From the above graphs, it can be seen that the equal variances assumptions has been met and the normal Q-Q plot is also better than the Taylor series transformation

Table

Description automatically generated

The model with the Box-Cox model has an adjusted R squared value of 0.7089 which is better than the other models

* 1. **L1 Regularization Approach (LASSO)**
* Since the number of variables is quite high, LASSO regression with L1 regularization was also carried out.
* L1 regularization has a penalty term that shrinks the coefficients of the non-important features. L1 shrinks the coefficients to zero and hence helps in variable selection.

Chart

Description automatically generated with low confidence

From the graph, we identify the best lambda parameter value for L1 regression as 0.02

* The LASSO regression model seems to be performing quite well.

A picture containing graphical user interface

Description automatically generated

Chart, scatter chart

Description automatically generated Chart, line chart

Description automatically generated

* The residual plots and the Q-Q plots seems to satisfy the normality and the equal variance assumption. Hence this would be the best model for the estimation of the critical temperature values.

**4. Results**

* In the report we conducted several different analysis methods to create a model that helped in the prediction of the critical temperature of the superconductors.
* There was high correlation in the data which was noticed during the exploratory data analysis phase and it was also noticed that the data is generally clean without needing much cleaning.
* Outliers and influential points were removed prior to the analysis.
* Dropped variables based on correlation to avoid multi collinearity
* Applied Data transformation techniques to improve the statistical validity of the model and boost performance.
* L1 regularization was applied which helped in improving the performance greatly

**4.1. Final Model Evaluation**

* The final model based on the analysis which gave the best performance was the Lasso Regression Model.

|  |  |
| --- | --- |
| **R-squared** | 0.6636 |
| **Adj R-squared** | 0.6632 |
| **RMSE** | 19.732 |

* The model assumptions are also satisfied.

Chart, scatter chart

Description automatically generatedChart, line chart

Description automatically generated

* Linearity: There is no visible trend in the residual plot which means that the linearity assumption is satisfied.
* Independence: There is no correlation between the residuals and seem to be independent.
* Normality: We can see from the Q-Q plot that the residuals line up in a straight line which means that the residuals are normally distributed.
* Equal Variance: There should be equal variance of the residuals across the data set and this assumption is satisfied.
* We have not used the statistical tests for the validation of the assumptions as the statistical tests related to normality and the equal variance assumptions are not quite accurate at very high sample sizes. The maximum suggested size for such tests is 5000 observations beyond which the reliability of such tests reduces.
* It can also be observed that the Rsquared and the adjusted Rsquared values are really close to each other which means that all the variables used in the model are quite significant.

Text

Description automatically generated Chart, scatter chart

Description automatically generated

Final list of variables Predicted vs Actual Plot

**5.Conclusions**

**References**

1. Prediction study on critical temperature (C) of different atomic numbers superconductors (both gaseous/solid elements) using machine learning techniques; G. Revathy, V. Rajendran, P. Sathish Kumar
2. Predictive Modeling of Critical Temperatures in Superconducting Materials; Natalia Sizochenko and Markus Hofmann
3. A data-driven statistical model for predicting the critical temperature of a superconductor; Kam Hamidieh

**Appendix:**

* The R jupyter notebook used for the analyis has been attached below.
* The R codes will be shared as a separate file as well

