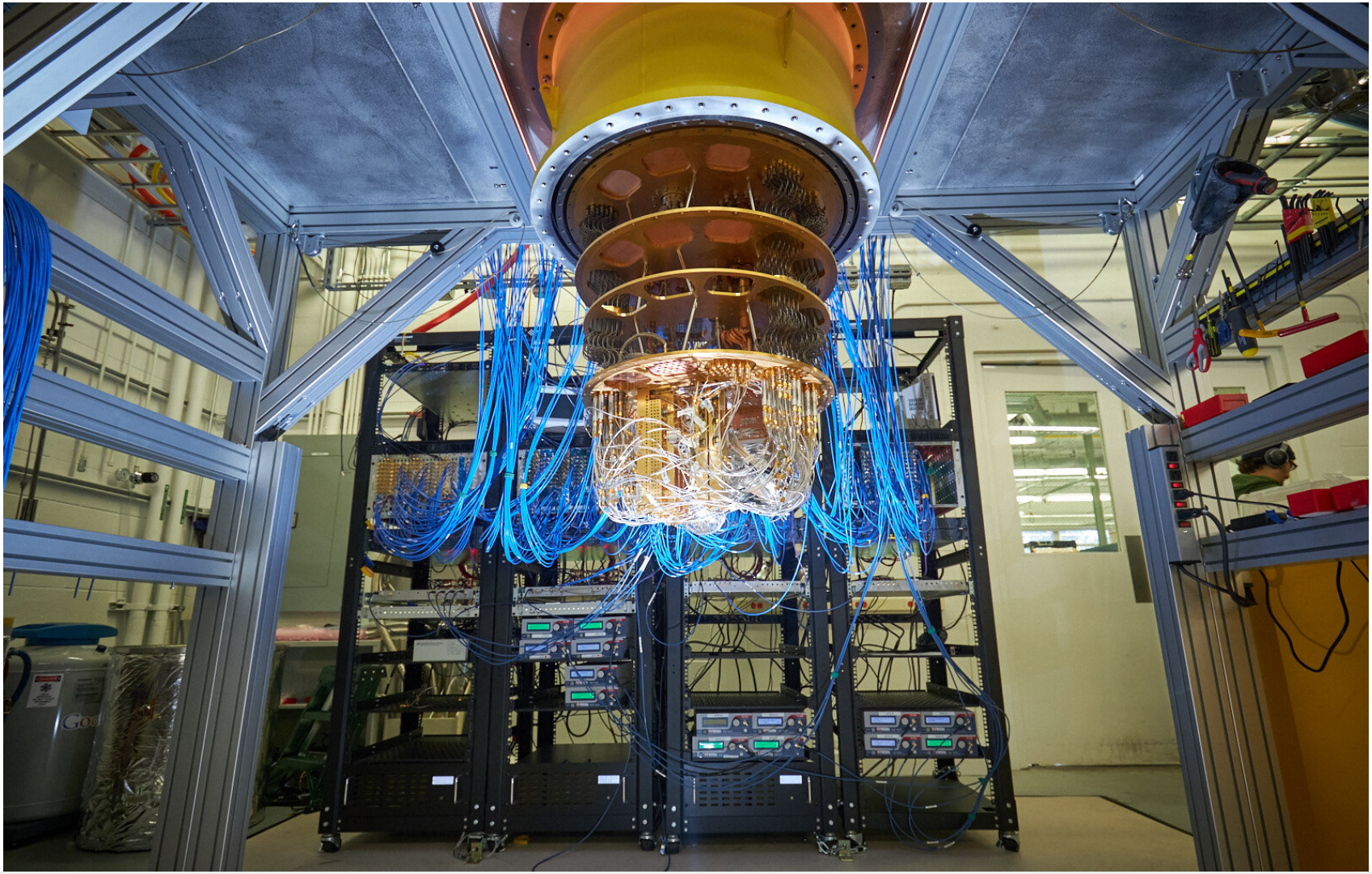
A Fine-Grained Analysis of XGBoost Predicting T_c

Daniel Briseno

January 20, 2021

Introduction



Introduction

Data-driven T_c Prediction

- No good physical model for T_c
- Previous attempts at data-driven T_c prediction focus prediction on small subset of superconductors.(**owolabi˙estimation˙2015**)(**stanev˙machine˙2018**
- XGBoost gradient-boosted decision tree ML algorithm provides T_c prediction for more general class of Superconductors (**hamidieh˙data-driven˙2018**).

Introduction

Principle Questions Motivating Analysis

1. How well does XGBoost perform at predicting T_c across the true T_c quartiles?
2. How well does XGBoost perform at the same task across the predicted T_c quartiles?

Introduction

The Data

Set of 21263 superconductors, with features defined by summary statistics of the atomic properties for the elements they contain.

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/m3)	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: Elemental properties used to define features used by XGboost to predict T_c.

Introduction

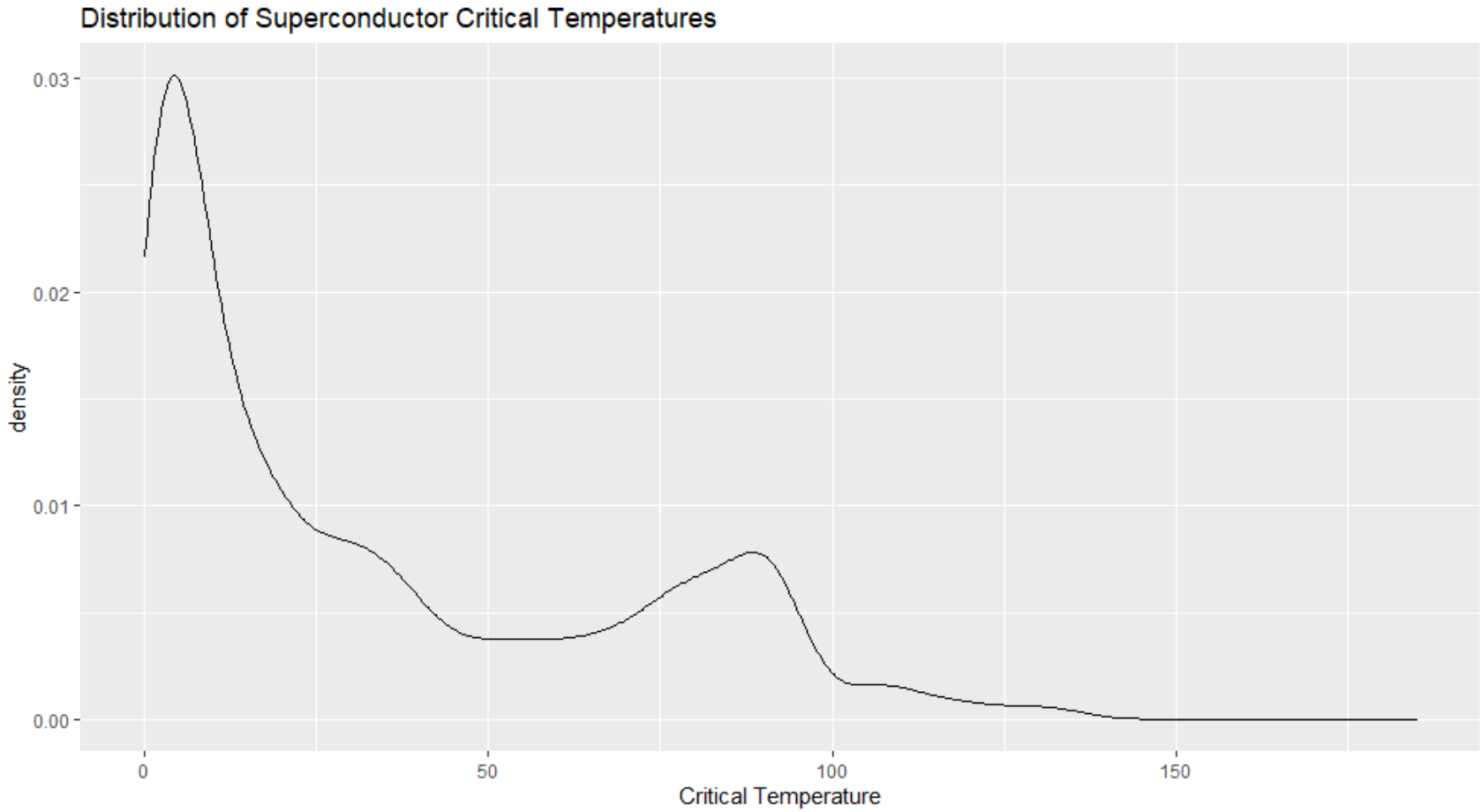


Figure: Density plot of T_c values for all superconductors in dataset.

Methods

Raw Error and Error Vector

For a T_c prediction P and an actual T_c value A, the **raw error** is $P - A$.

An **error vector** is a list of summary raw error statistics listed below.

Statistic	Description
RMSE	The residual mean squared error of predictions
ave_err	Raw average of error
std_err	Standard deviation of raw error values
under_cnt	Number of under-predictions
over_cnt	Number of over-predictions
correct_cnt	Number of exactly correct predictions. Expected to be 0
ave_under	Average value of under-prediction raw error
ave_over	Average value of over-prediction raw error
std_under	Standard deviation of under-prediction raw error
std_over	Standard deviation of over-prediction raw error

Table: Summary error statistics collected for predicted T_c values.

Results – Control Data

	RMSE	ave_err	std_err	under_cnt	over_cnt	ave_under	std_under	ave_over	std_over
1	9.518	-4.998	8.099	5611.04	1469.9	-7.008	7.561	2.673	4.843

Figure: Control Data Error Vector Summary

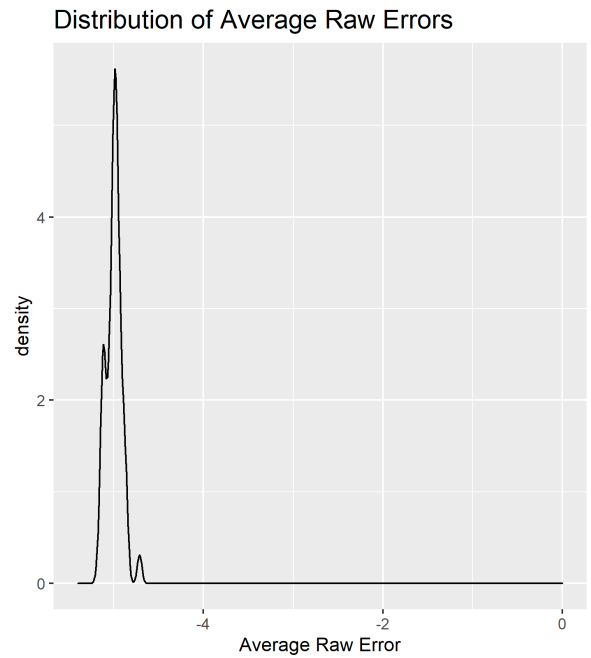


Figure: Density plot of raw error averages.

Results – Control Data

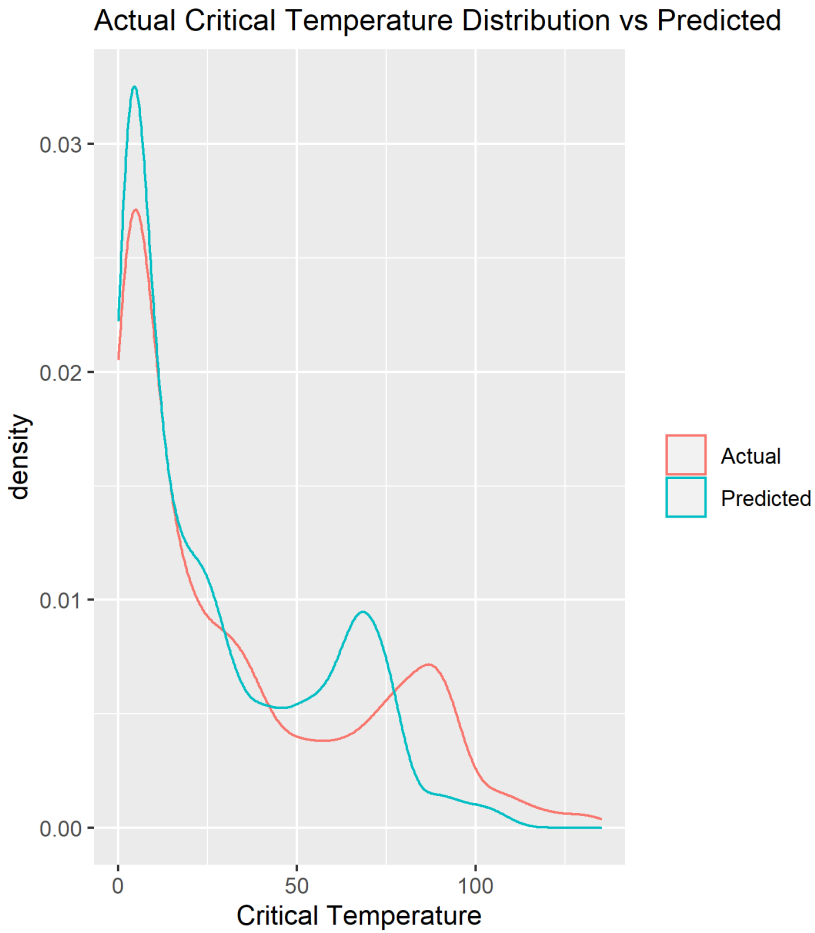


Figure: Actual T_c distribution plotted alongside predicted T_c distribution.

Results – True T_c Quartiles

Figure: RMSE, standard deviation of error, and the average of raw errors plotted alongside each other for each true T_c quartile.

Results – Predicted T_c Quartiles

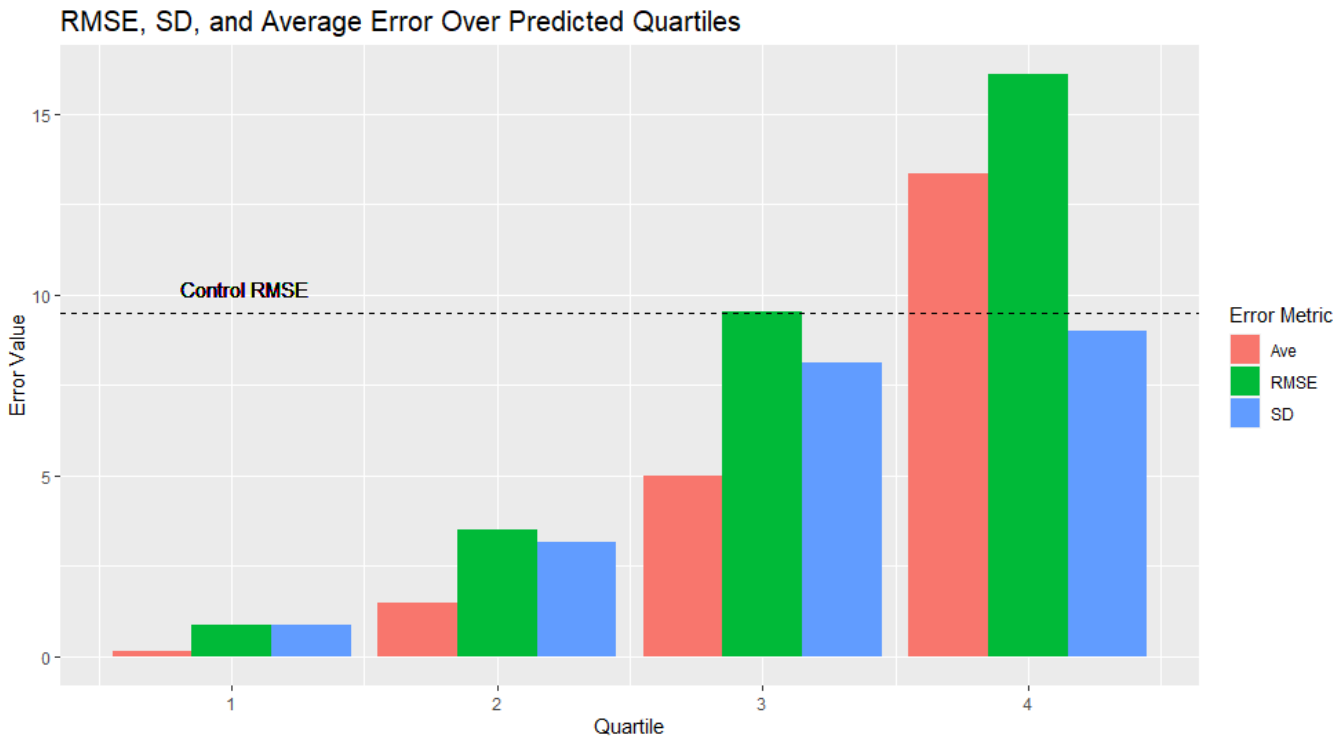


Figure: RMSE, standard deviation of error, and the average of raw error plotted alongside each other for each predicted T_c quartile.

Conclusion

Summary

- XGBoost performs with rapidly decreasing accuracy for increasing true T_c quartiles.
  - Largely due to correctable bias
- XGBoost performs with rapidly decreasing accuracy for increasing predicted T_c quartiles.
  - Largely due to correctable bias

Conclusion

Future Directions

- Implementing corrections to XGBoost
  - Via modified loss function
  - Via additive correction term which depends on predicted quartile
- Analysing distribution of raw errors without taking summary statistics

References