So You're a Star— But How Hot Are You?

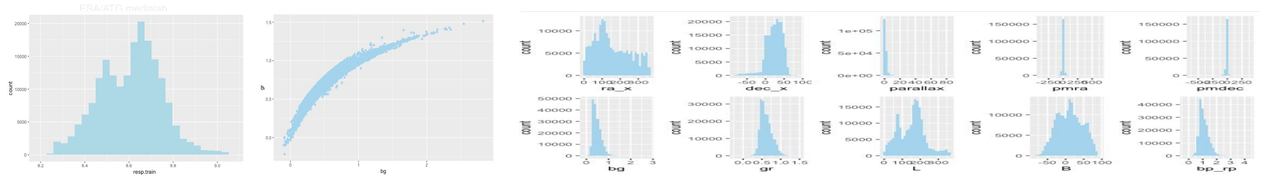
Leon Lu, Kat Phelps, Tara Prakash, Aramya Trivedi (Advisor: Peter Freeman)

Introduction

A space-based telescope, *Gaia*, launched by the European Space Agency, is currently gathering data on over one billion stars "in order to further understand the structure, formation, and evolution of our Galaxy." One of the propertis of stars that we can study is their effective temperature. A standard method of predicting stellar temperature is to fit physics models to the spectra of individual stars. The goal of our project is to build a model to effectively predict stellar temperature given other stellar properties measured by *Gaia*.

Data

Our data consisted of 22 predictor variables and a quantitative response variable, *stellar temperature*. Our total sample size is 250,000 data points of which we train our model on 75% and test on 25%. We cleaned the data down to 10 predictor variables after eliminating error columns and combining columns for interpretation purposes. After examining the distributions of the data we log transformed the response to make the distribution closer to normal.. We observed a strong correlation (0.98) between bg and gr, which is to be expected due to the nature of the relationship between colors. This did not affect our results as we focus on prediction as opposed to inference.



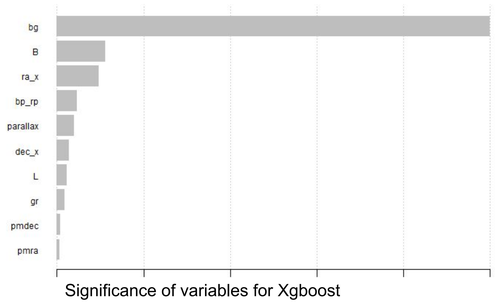
The response variables distribution (left),
bg and rg correlation (middle),
10 predictor variables (right)

Analysis and Results

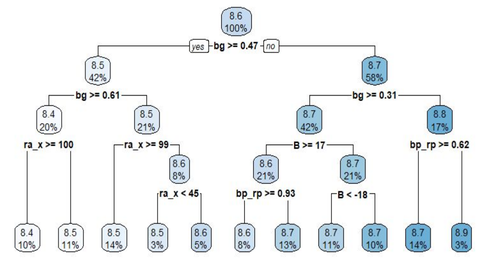
We use various techniques to build models to predict stellar temperature. The main models we work with are multiple linear regression, best subset selection, principal component analysis, boosting, decision trees, and random forest.

In terms of identifying important variables we observed that bg was the most important variable by far with B and ra_x being mildly important. This was consistent across models which supports our assertion that these variables are important in order to determine stellar temperature.

We determined that xgboost was generated the most successful and interpretable results. Xgboost is known for being exceptional at classification and regression predictive models which we saw within our own results; of the models we explored, xgboost had the lowest mean squared error and highest area under the curve which demonstrates its successful performance.



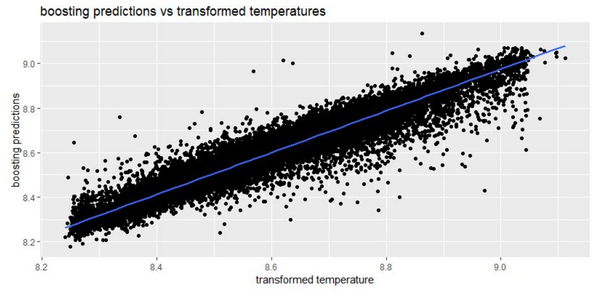
Significance of variables for Xgboost



Significance of variables for Decision Tree

Model	MSE
Linear Regression	0.00650
Decision Tree	0.00554
Random Forest	0.00138
Boosting	0.00123

MSE for each model run



Accuracy of Xgboost

Conclusion

- Though our model ended up performing very well with an MSE of 0.00123 it could still be improved upon. Looking at the graph of predicted temperature vs actual temperatures we see that there are some predictions that are very far off which could either be outliers or our model having a terrible prediction.
- For future analyses we should we focus on more nonlinear types of models as many of the linear models we used performed far worse than the nonlinear ones.
- Relating back to the original question of how to predict stellar temperature Xgboost is clearly the best model and bg, B, and ra_x being the most significant predictors. This allows us to understand which factors are most significant and gives us insight for the future as to which variables should be given the most weight.

References

Bai, Y., et al. 2019, The Astronomical Journal, v. 158, id. 93
ESA/ATG medialab

Carnegie Mellon University
Statistics & Data Science