Predicting Quasar Redshifts Given Brightness

By: Michael Chen, Pauline Qin, May Wang
Advisor: Peter Freeman

Background and Introduction

Quasi-stellar objects, better known as quasars, are extremely luminous and lie far from our galaxy. All their light comes from the region around supermassive black holes in the nuclei of galaxies. The positions of quasars in the universe can be measured using redshift, which is the amount by which the wavelength of a photon changes as it travels through the expanding Universe. However, it is rather difficult to measure the redshift of quasars and thus our dataset only includes redshifts determined by "visual inspection," a very laborious process.

Artist's impression of the quasar 3C 279

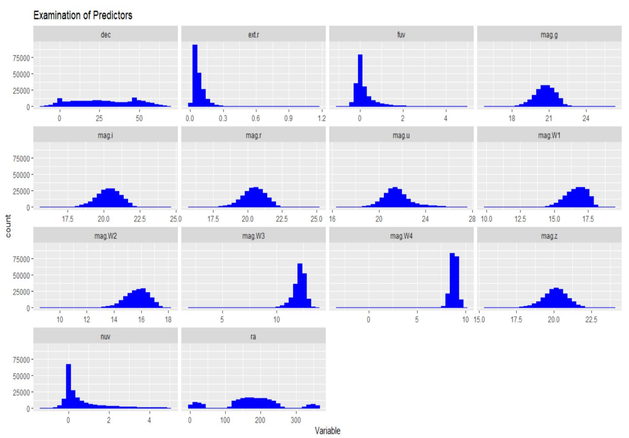
The goal of our study is to learn a statistical model that takes in a quasar's brightness at different wavelengths and produces an accurate and precise redshift estimate.

Data

The original data in this directory includes a predictor data frame with 216,190 quasar observations with 21 variable measurements. The data were collected by the Sloan Digital Sky Survey (SDSS) and is from the SDSS Data Release 14 Quasar Catalog.

Variable Name	Description
ra, dec	right ascension, declination (celestial longitude and latitude)
bal	an index related to the breadth of the C IV quasar absorption line
mag.[ugriz]	SDSS photometric magnitudes
ext.r	Milky Way extinction in the r band (related to dust along the line of sight)
rass	brightness in the ROSAT All-Sky Survey (X-ray)
fuv, nuv	brightness in two GALEX bands (UV)
mag.[W1, W2, W3, W4]	magnitudes in four WISE bans (IR)
flux.[Y, J, H, K]	brightness in four UKIDSS bands (IR)
flux.first	brightness at 21 cm wavelengths (radio)

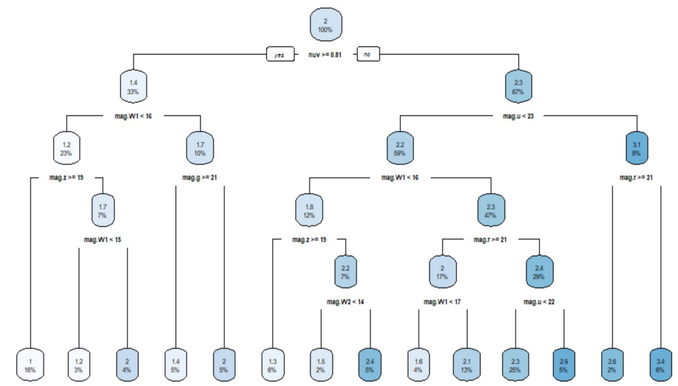
We clean the data by removing predictors that mostly contained missing values and by removing unusual observations that might also indicate missing data. Our final dataset consists of 14 predictors and the response variable, redshift, and 216,100 observations. In addition to cleaning, we also look into transforming the predictors and combining the magnitudes into colors. Ultimately, we found out that our results were better with the data as is and decide to run our models without any additional adjustments to the cleaned data.



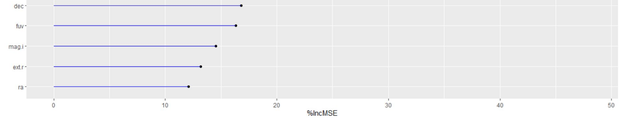
Methods

- We use a number of methods to learn statistical models:
  - Linear Regression, Best Subset Selection (using BIC), Tree Regression, Extreme Gradient Boosting (XGBoost), and Random Forest
- We use mean-squared error (MSE) as our metric to measure the accuracy of our models
- We find that Random Forest produces the best results with an MSE of 0.0898
  - Random Forest is an ensemble learning model that grows multiple decision trees in parallel and takes the mean of each individual tree prediction

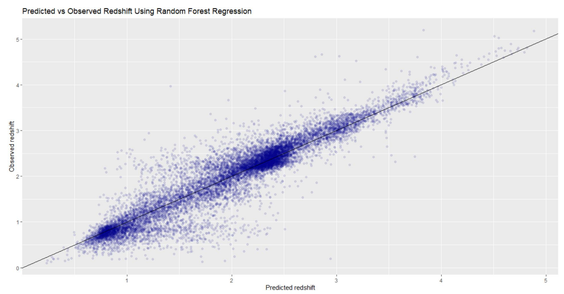
Analysis and Results



Regression tree for our data. For the above tree, we can see that the most important predictor appears to be nuv followed by mag.u and mag.W1.



Variable importance plot generated using random forest regression. The most important predictor appears to be nuv followed by mag.u and mag.W1. This agrees with our inference from the regression tree.



Comparison of our model with our data. Our model reasonably fits the data, and our fit appears to better for higher values of redshift.

Model	Test Set MSE
Linear Regression	0.2114
Best Subset Selection (BIC)	0.2114
Regression Tree	0.2887
Random Forest Regression	0.0898
Extreme Gradient Boosting	0.0980

Table of test set MSEs for all statistical models tried. As seen in the table, random forest regression had the lowest test set MSE followed closely by extreme gradient boosting.

Conclusion

- Random Forest was our best model, with a MSE of 0.0898
- The accuracy is supported by our diagnostic plot which shows that our random forest model fits well
- The most important variables in our model are nuv, mag.u and mag.W1

References

https://www.sdss.org/dr14/algorithms/qso_catalog/
Paris, I., et al. 2018, Astronomy & Astrophysics, v. 613, p. A51