Predicting Redshift Using SDSS and GALEX Data

Esha Gupta, Joyce Huang, David Lurie, Mengrou Shou

Project Advisor: Peter Freeman

Introduction

Quasars, or quasi-stellar objects, are ultra-bright cores of galaxies with supermassive black holes at their center. To determine how far away quasars are in the universe, astronomers use a quasar's redshift, or the ratio of the observed wavelength of a photon from an object to its wavelength when it was emitted, minus one. However, determining a redshift precisely is a time-consuming process. In this project, our goal is to determine a model to predict a quasar's redshift from easily obtained brightness measurements made by SDSS (a ground-based telescope) and GALEX (a space-based UV telescope).

Data

The data that we analyze in this project combines information from SDSS and GALEX (Trammell et al. 2006). The dataset contains 5380 observations with one response variable and ten predictor variables.

Predictors: We have six different magnitude variables which are strongly correlated with each other, ascension and declination, and measurements of dust and gas content along the line of sight to the quasars.

Variable Name	Description
ra	right ascension
dec	declination
(u,g,r,i,z) mag	quasar magnitude in the u, g, r, i, and z bands of the optical and near-infrared
nmag	quasar magnitude in the ultraviolet n band
gal_abs_u	dust content
log_nh_gal	gas content

Response: Our response variable is redshift. We find that performing a square root transformation on redshift yields an overall better mean squared error and reduced right-skewness.

Exploratory Data Analysis:
We find that the quasar magnitudes are all heavily linearly correlated with each other (Figure 1). This indicates that using all of the magnitudes in a model may be redundant. However, since the ultimate goal of this study is prediction, we do not remove any of the predictor variables while building models.

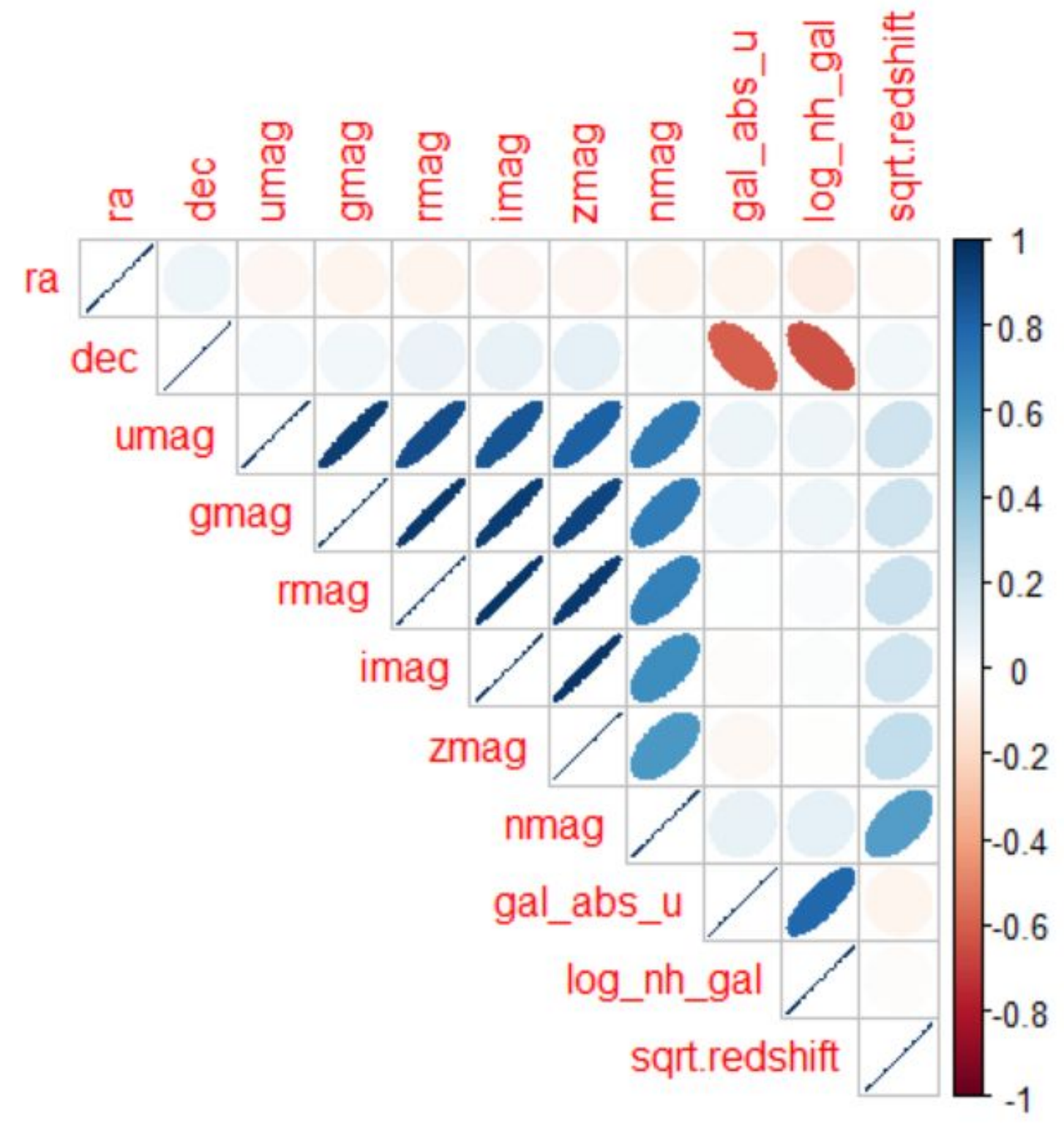

Figure 1. Correlation plot for predictor and response variables

Methods

We train models using 70% of the data and test them on the remaining 30%.
We explored the following models:

- Linear Regression
- Random Forest
- eXtreme Gradient Boosting
- Decision Tree
- Best Subset Selection
- K Nearest Neighbors

Analysis and Results

Out of all the algorithms used, Random Forest produces the lowest mean squared error of 0.030. The Random Forest algorithm aggregates a series of regression trees, each built with a bootstrapped sample of the original data and with a randomly chosen subset of the predictor variables. A redshift prediction for a given quasar is generated by taking the average of the predicted values across all trees in the forest.

MSEs of Models Used

Model	Mean Squared Error
Random Forest	0.030
eXtreme Gradient Boosting	0.033
Linear Regression	0.040
Best Subset Selection	0.040
Decision Tree	0.048
K Nearest Neighbors	0.066

Table 1. Mean-squared error of models

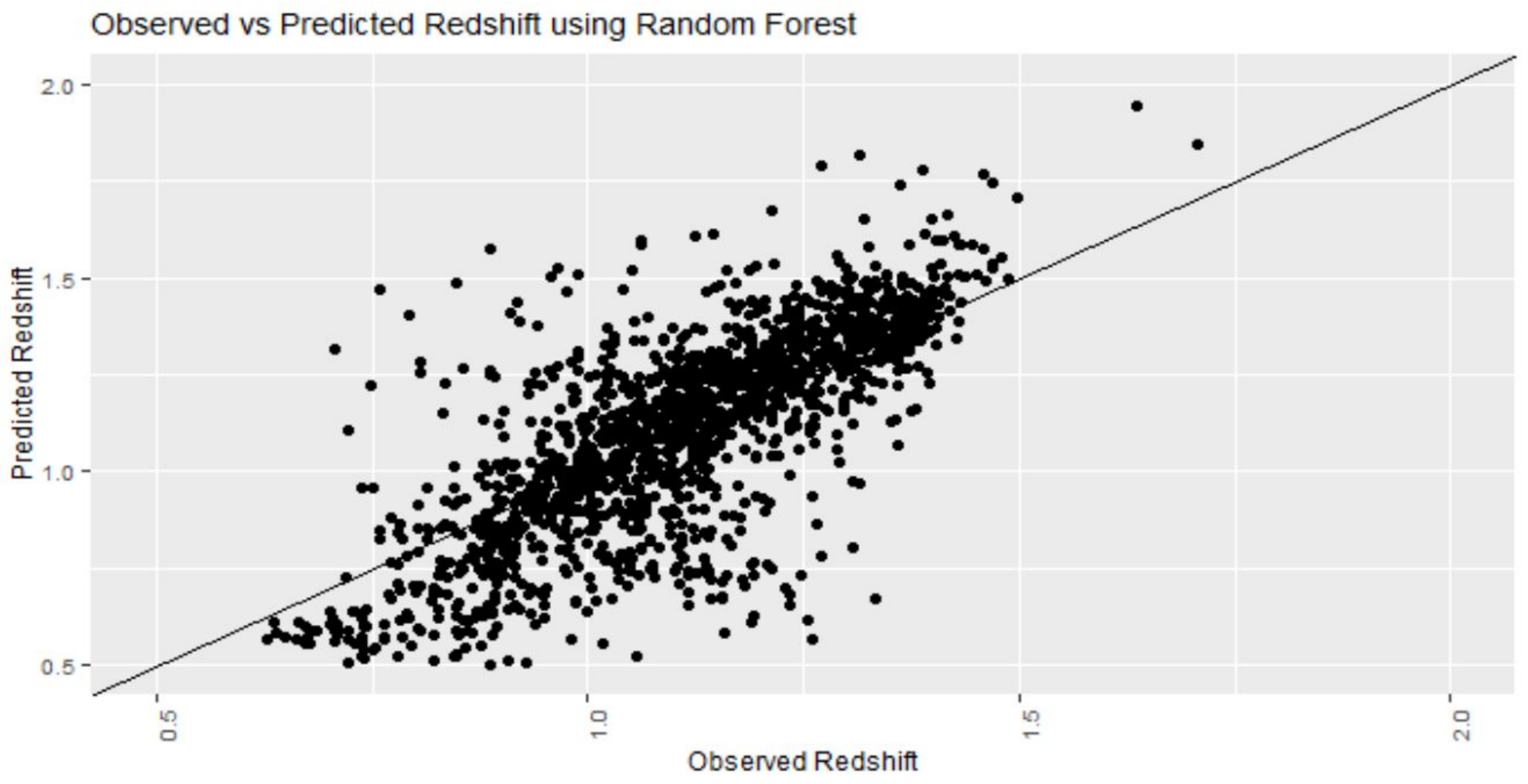

Figure 2. Predicted redshift against observed redshift for the Random Forest model

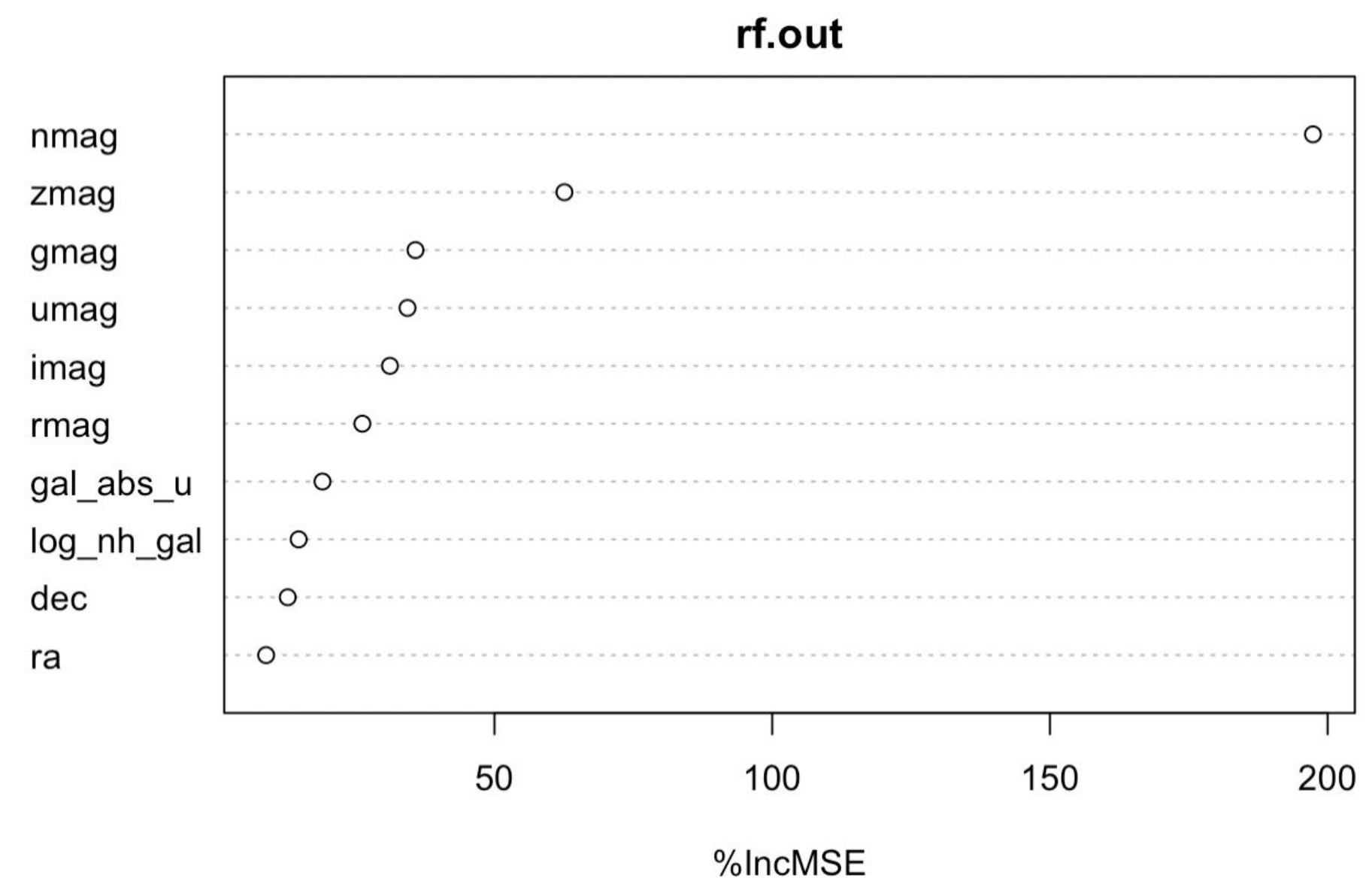

Figure 3. Variable importance plot for the Random Forest model

Conclusion

We conclude that redshift can be most precisely predicted from its predictor variables using the Random Forest model. It performs better than decision trees due to its main advantage of bootstrap aggregation. The two most important predictor variables that predict redshift are *nmag* and *zmag*. We note that *nmag* is a UV magnitude from GALEX. This suggests that including the GALEX brightness information improved the model over what SDSS brightness information would produce by itself.

References

[1] Freeman, P. E. 2021, online at https://github.com/pefreeman/36-290/blob/master/PROJECT_DATASETS/GALEX/README.md

[2] Trammell, G. B., et al. 2006, The UV Properties of SDSS Selected Quasars, online at https://arxiv.org/pdf/astro-ph/0611549.pdf