Redshift Prediction

Using the COMBO-17 Dataset to investigate redshifts of galaxies from the Chandra Deep Field South Survey

The problem

Redshifts

The redshift of a galaxy is analogous to its distance from earth

Redshifts can be calculated photometrically or spectroscopically

Photometrically is easier but less accurate

Uses

As we approach the launch of JWST, SPHEREX, and other telescope projects, quickly estimating redshifts on large amounts of data is becoming vital

Problem Statement

This investigation will use magnitudes and photon fluxes to estimate photometric redshifts

Galaxies are being used from the COMBO-17 dataset, taken from the Chandra Deep Field South survey

Data Cleaning

Examine Variables

The COMBO-17 dataset contained 63501 objects, 104 variables

Objects classified as galaxies with R < 24 were used for this investigation

Total of 14283 galaxy objects after filterning

Clean NaNs

Cleaning NaNs resulted in 62 variables

Many fields contained NaN as their data did not apply to galaxy type objects

Categorize

Variables were categorized into 7 types:

- Redshifts
- Magnitudes
- Magnitude Errors
- Photon Fluxes
- Flux Errors
- R Magnitudes
- Meta Data

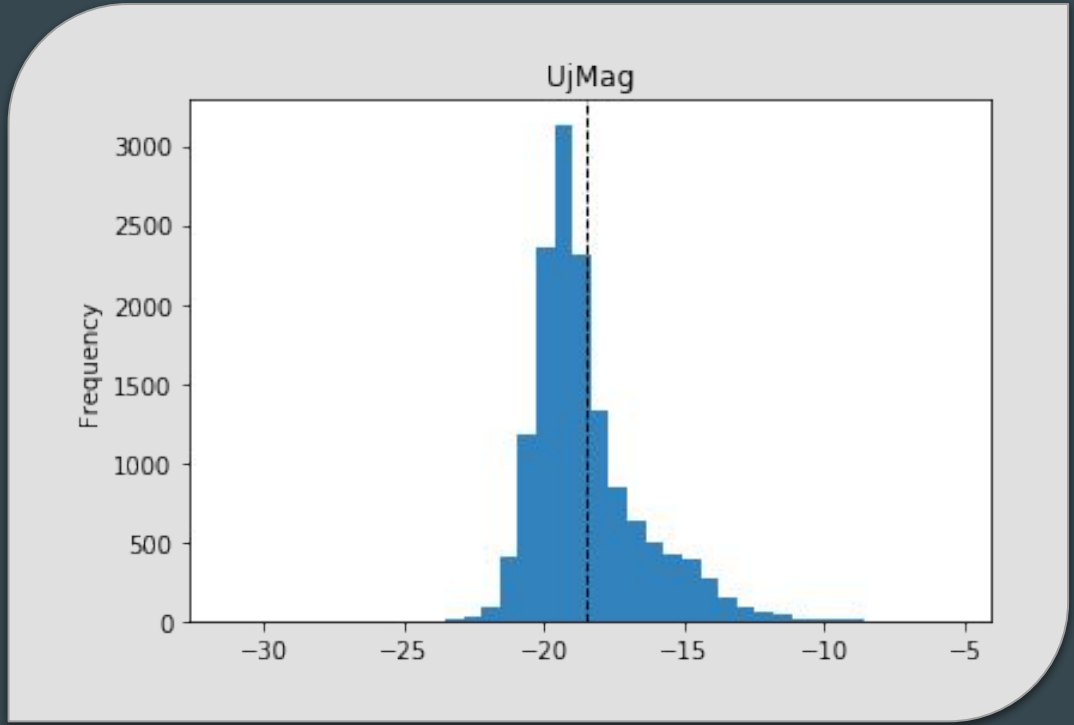
Variables

Magnitudes

Magnitudes were approximately normally distributed

Distribution skewed slightly left of mean

Magnitudes errors almost all less than 1

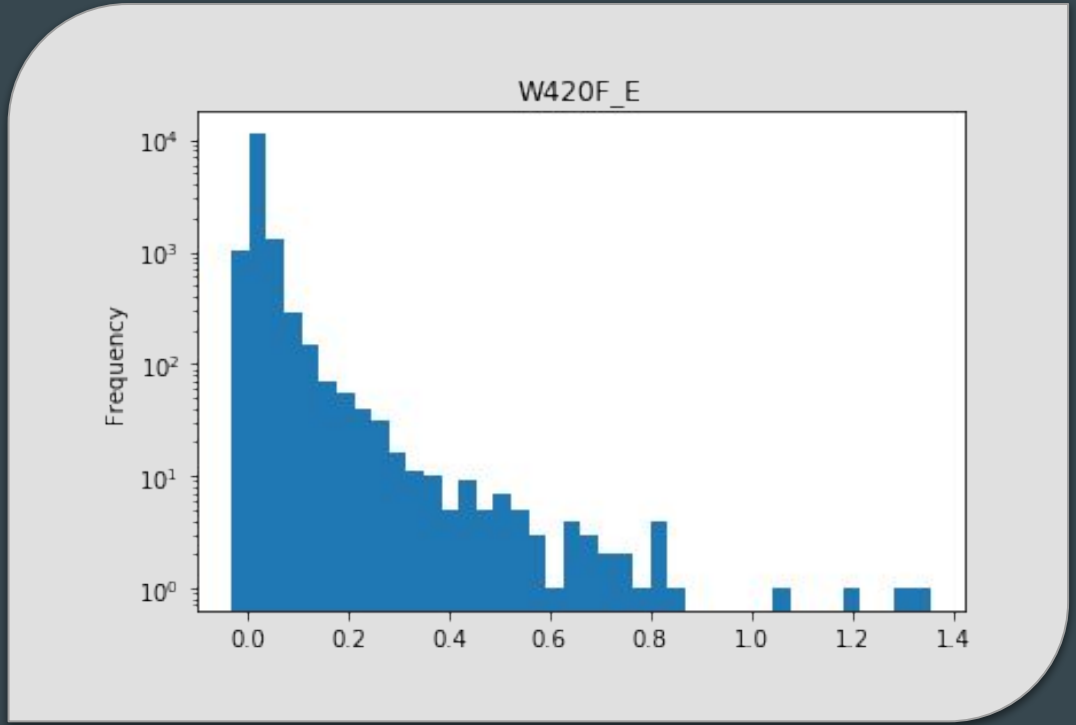


Variables

Photon Fluxes

Majority of photon fluxes less than 0.2 for all wavelengths

Photon flux errors less than 1% percent error



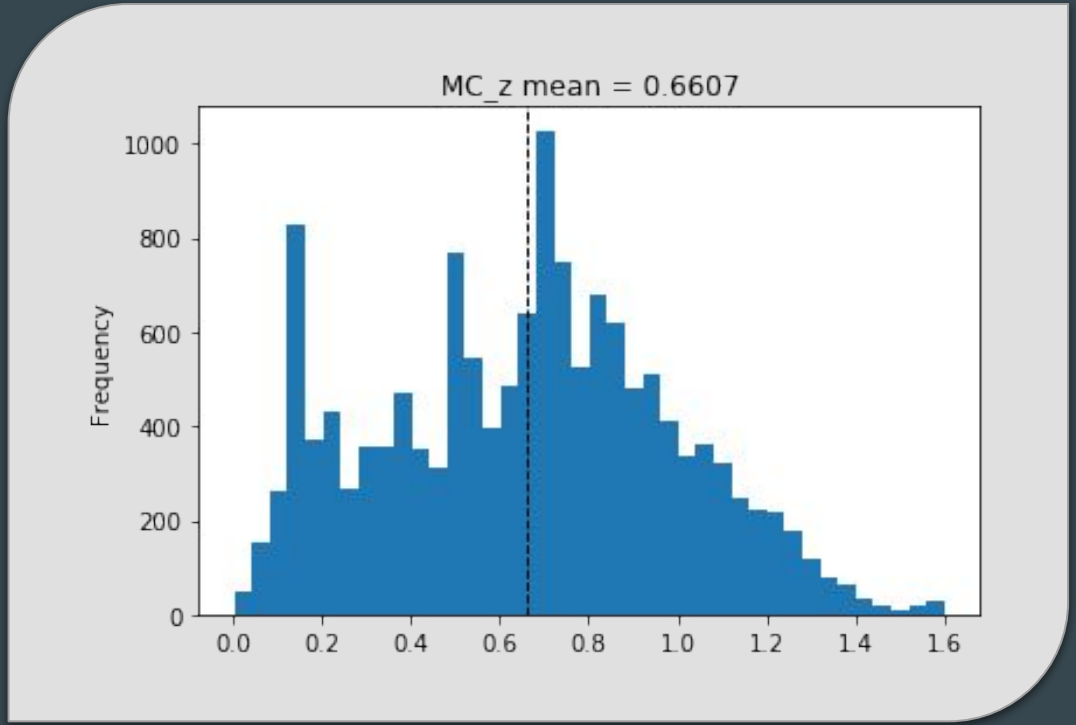
Variables

Redshifts

Redshift values are between 0 and 1.6

Redshifts are not normally distributed

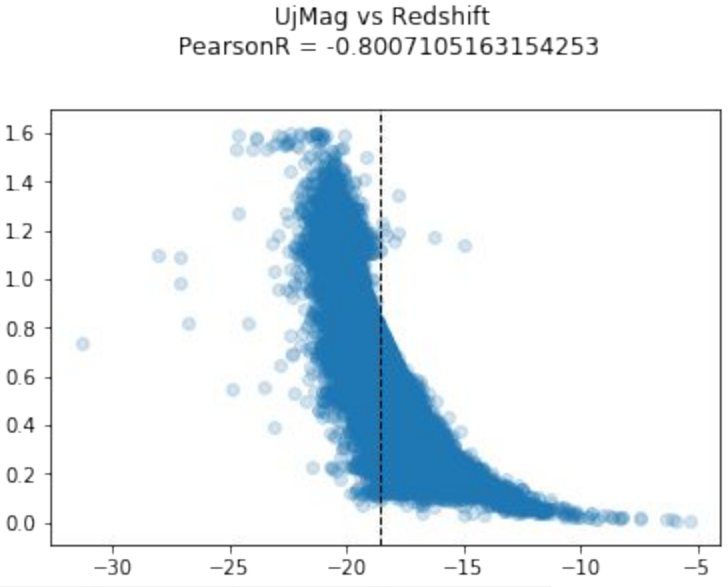
More low redshift objects (<0.8) than high redshift objects

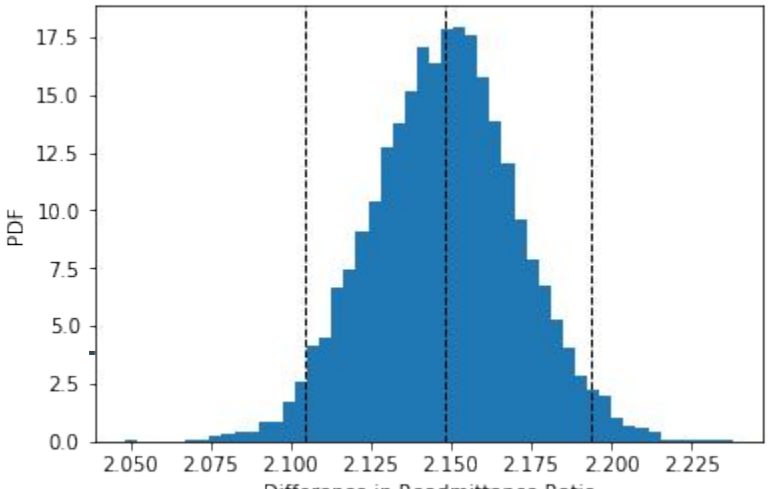


Correlations

- Magnitudes and redshifts all highly correlated
- Absolute value of Pearson R greater than 0.7, most above 0.8

- Bootstrapping shows there is a significant different between magnitudes with $0 < r < 0.8$ and $0.8 < r < 1.6$
- The mean difference in UjMag between small and large redshift is 2.15
- 95% confidence interval 2.10 to 2.19

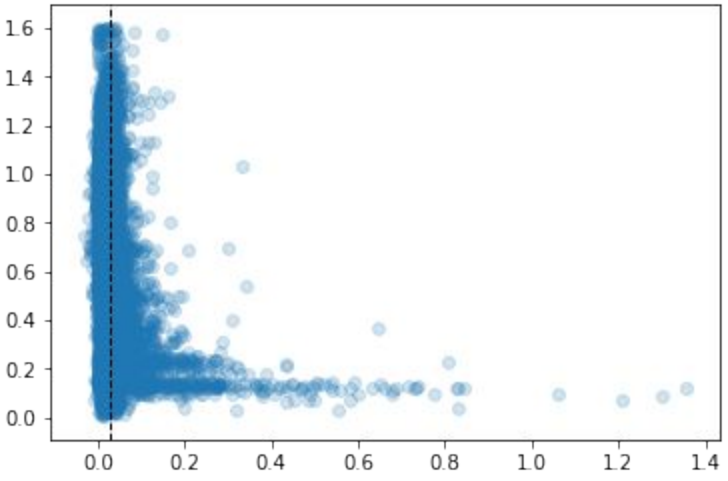


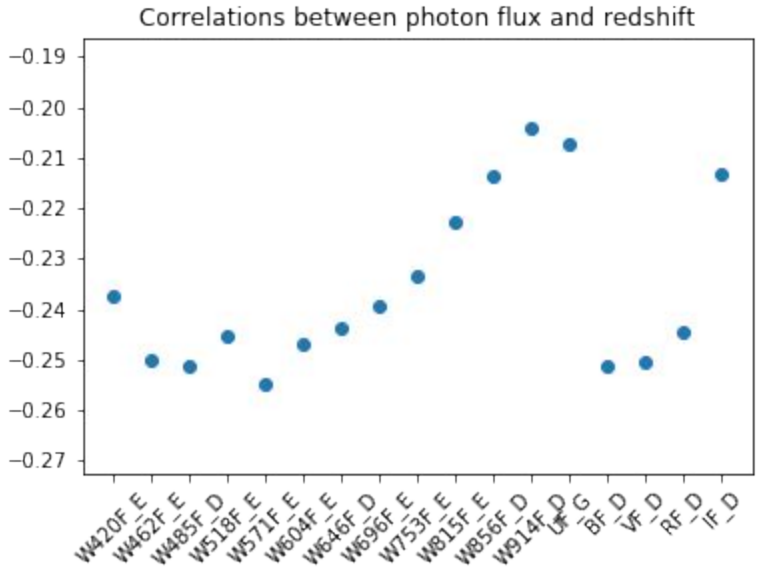


Correlations

- Fluxes and redshifts not highly correlated
- W420 vs Redshift shown in upper right
- Most photon fluxes above 0.2 occur at r < 0.6 for all flux bands

- Absolute values of Pearson R coefficients less than 0.26 for all flux bands





Models

Magnitudes and Fluxes

Models were made using both the full dataset and a subset of the data including the photon flux and flux error only

Photon fluxes are raw data from the telescope

Training and Testing

All models were trained on 70% of the data and tested on the remaining 30%

Training and testing sets were randomized for each model

Optimization

Model variables were optimized using GridSearchCV from Scikit-Learn

- Five models were tested:
  - Random Forest Regression
  - Support Vector Regression (SVR)
  - LASSO
  - Ridge Regression
  - Linear Regression

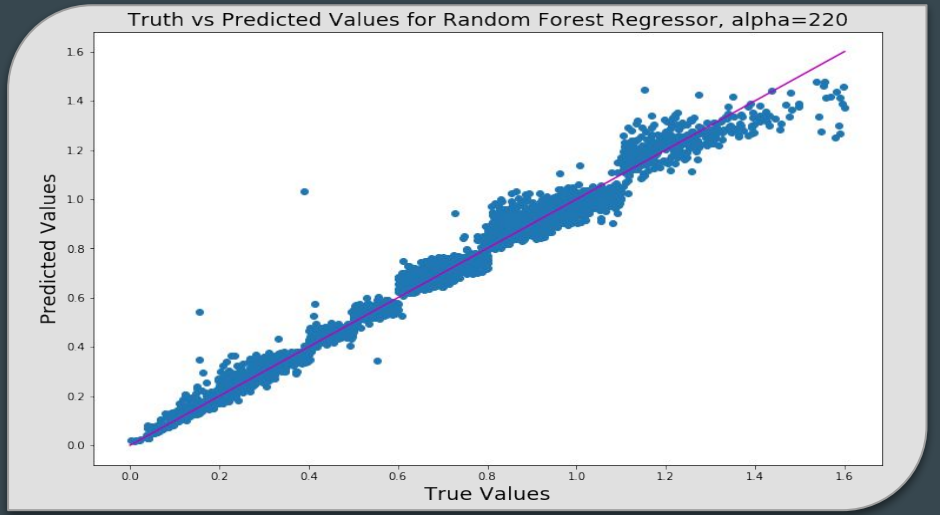
Best:
Random Forest Regression

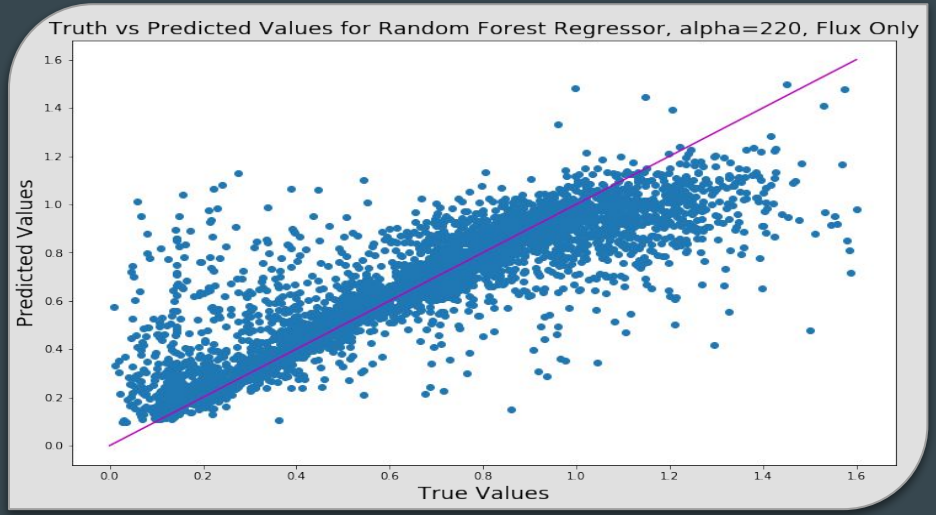
Bad:
Ridge Regression,
Lasso,
Linear Regression

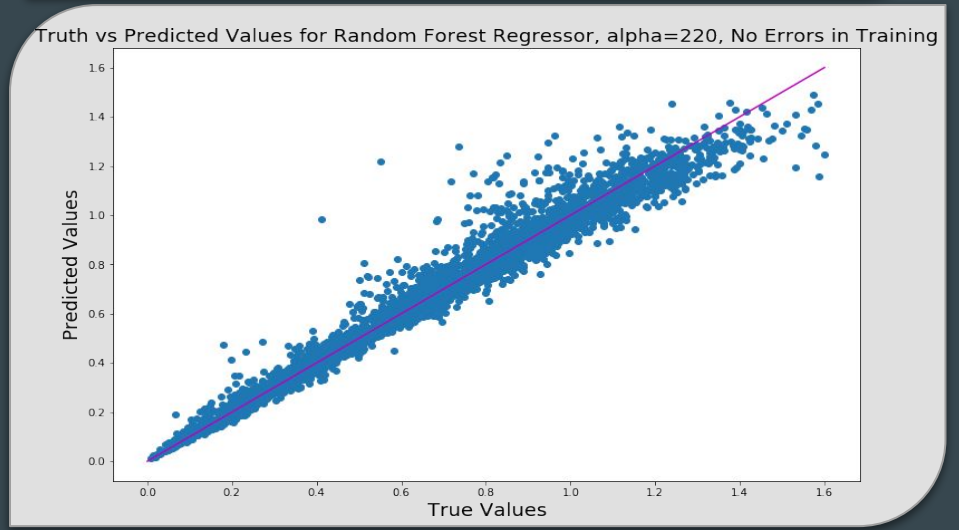
Good:
SVR

Random Forest Regression

- Using full dataset (left), r2 score of 0.983
- Removing errors gave r2 of 0.967
- Fluxes only gave r2 of 0.737





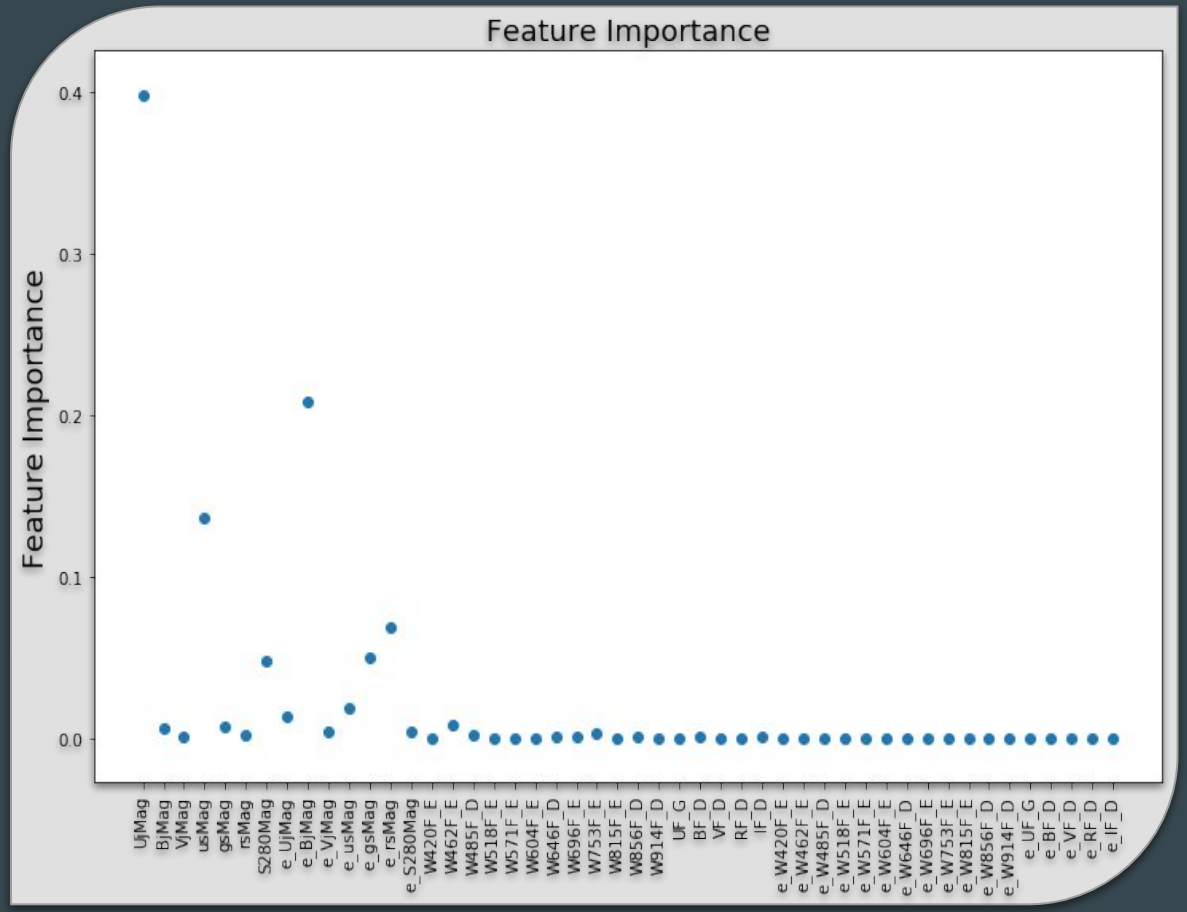


Feature Importance

All Variables

Looking at the feature importance for all variables shows that magnitudes and magnitude errors have the most importance

Random Forest Regressor trained with alpha = 220

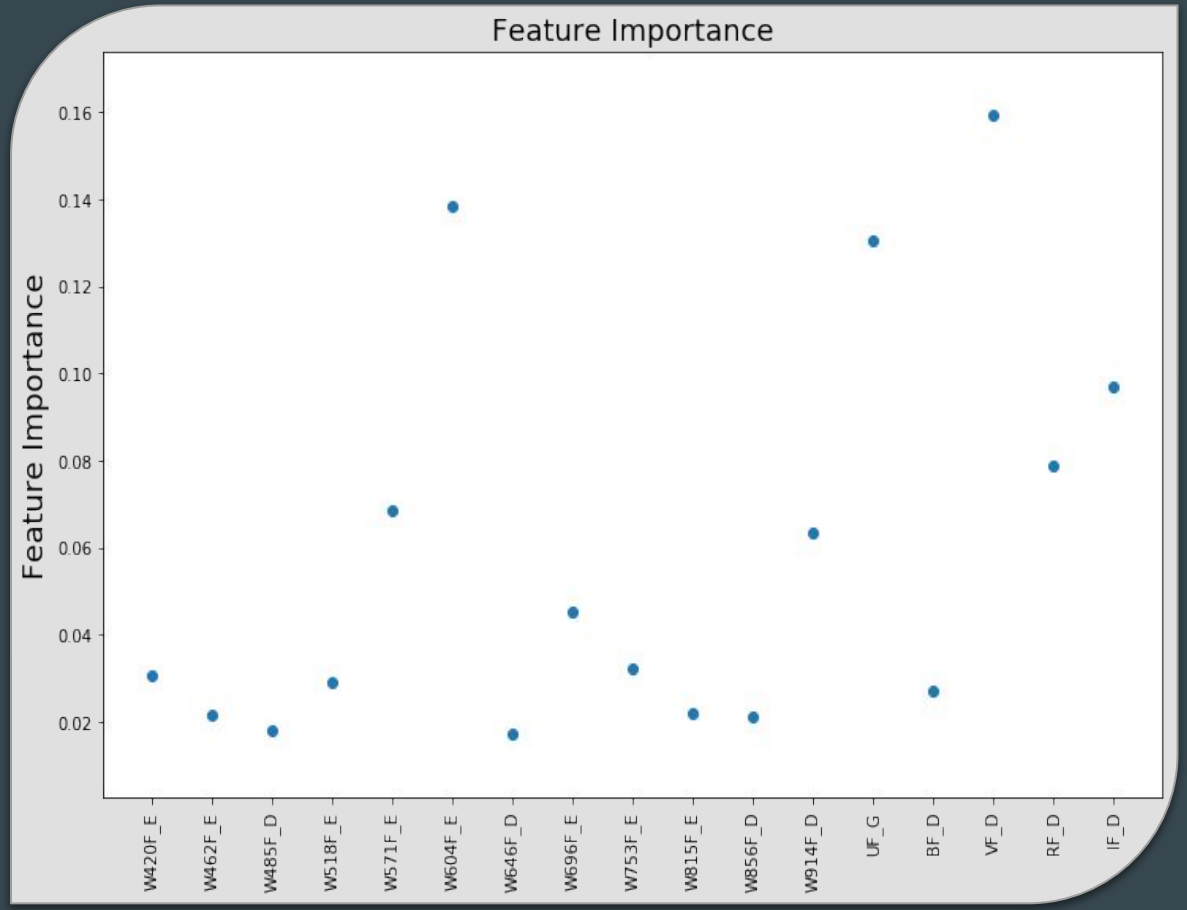


Feature Importance

Fluxes Only

Feature importance for model trained with only photon fluxes and flux errors

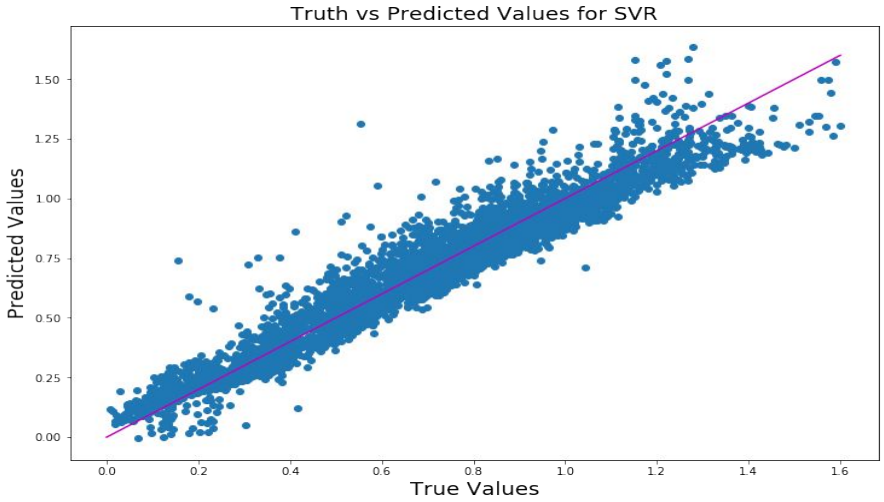
Random Forest Regressor trained with alpha = 220

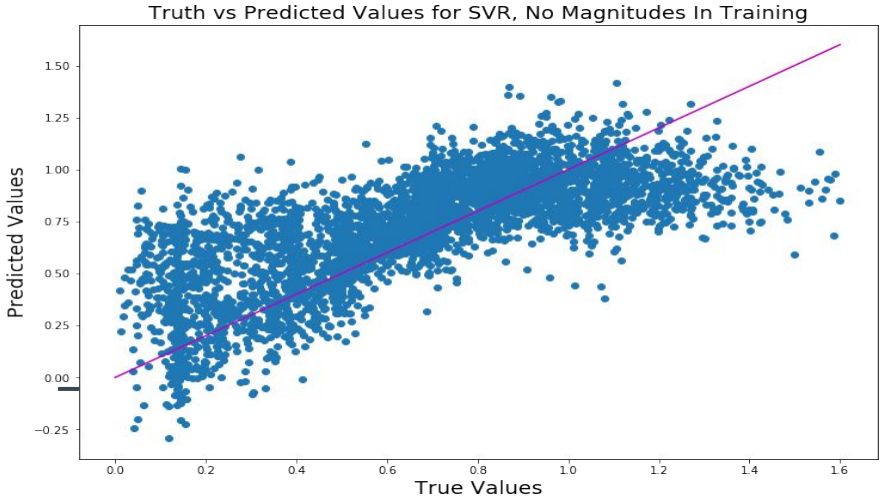


SVR

- SVR with all variables performed almost as well as Random Forest Regressor
- R2 score = 0.935
- No steps in truth vs predicted graph

- Fluxes and flux errors only performed badly
- R2 score = 0.515
- Standard scaling improved performance



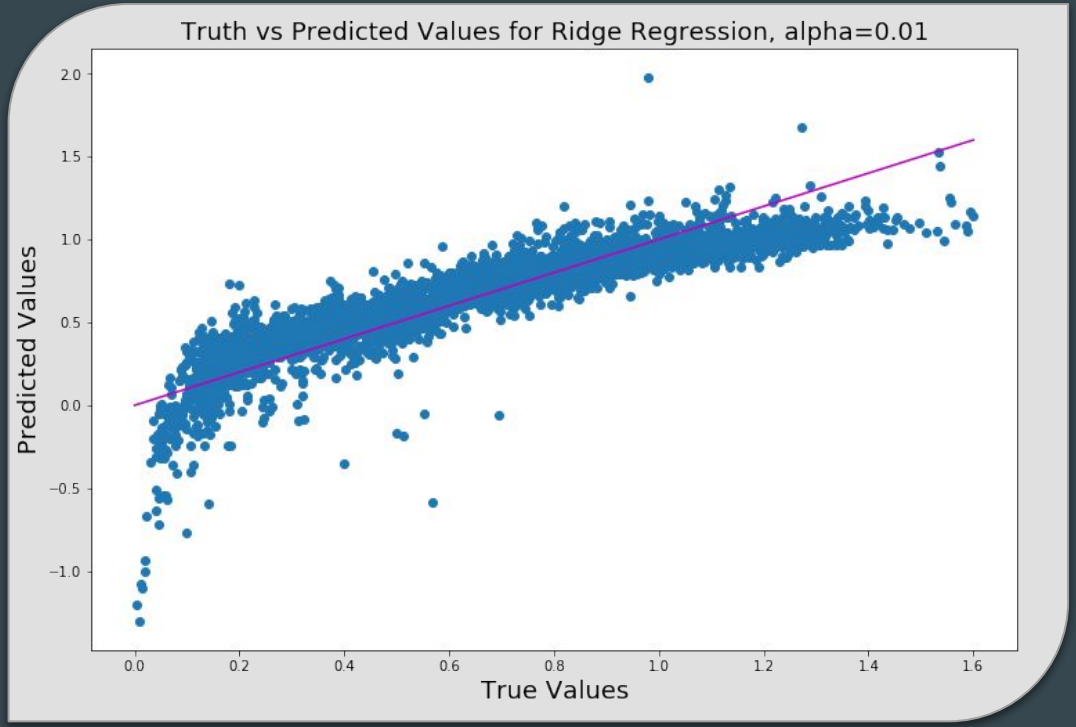


Ridge Regression, Lasso, and Linear Regression

- Ridge regression performed best on all variables
- Test score of 0.811

- Lasso and linear regression rest scores both < 0.66

- All three models predict negative redshift values



Generalizability

High Redshift

This model only valid on data between redshifts of 0 and 1.6

No guarantee that it would generalize to predict galaxies with high redshift

Should not be used to identify rare high redshift objects

Other Telescopes

Different telescopes collect photon flux in different bands

Photon flux not necessarily normalized between instruments

Training on a set of data from each telescope would be necessary to be able to predict redshifts from that telescope's data

Future Steps

Different Telescopes

Repeat the analysis on data from other telescopes

E.g. Hubble, Sloan Digital Sky Survey

Cleaning data would be more difficult than COMBO-17

Use Results

Complete an astrophysics analysis/model using both predicted and true photometric redshifts

Compare results - do predicted redshifts produce statistically significant differences?

Refine Flux Model

Use more advanced models to create a better model to predict redshift from flux only

Deep learning could be useful

Conclusions

Results show that this technique merits further investigation

Could significantly reduce the time of the astrophysics analysis pipeline