Predicting Galaxy Mass from Sky Coordinates and Brightness

Cherie Hua, Eric Huang, Joanna Yao, Neha Choudhari (Advisor: Peter Freeman)

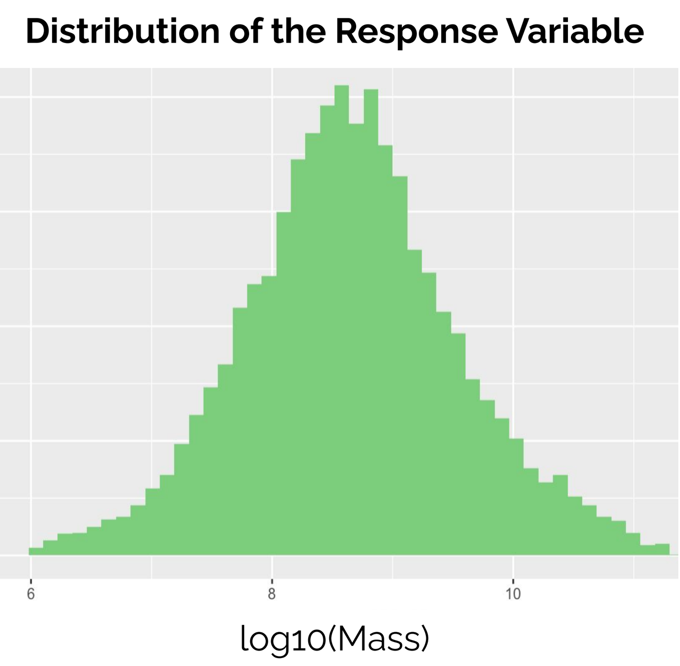
Introduction

The Cosmic Assembly Near-Infrared Deep Extragalactic Legacy Survey (CANDELS) is a program that creates catalogs of distant galaxies observed by the Hubble Space Telescope. Astrophysicists use these catalogs to study galaxy evolution by studying the relationship between galaxy properties and their brightness. The goal of this project is to **predict galaxy mass in the GOODS-North field from galaxy location and brightness data**.

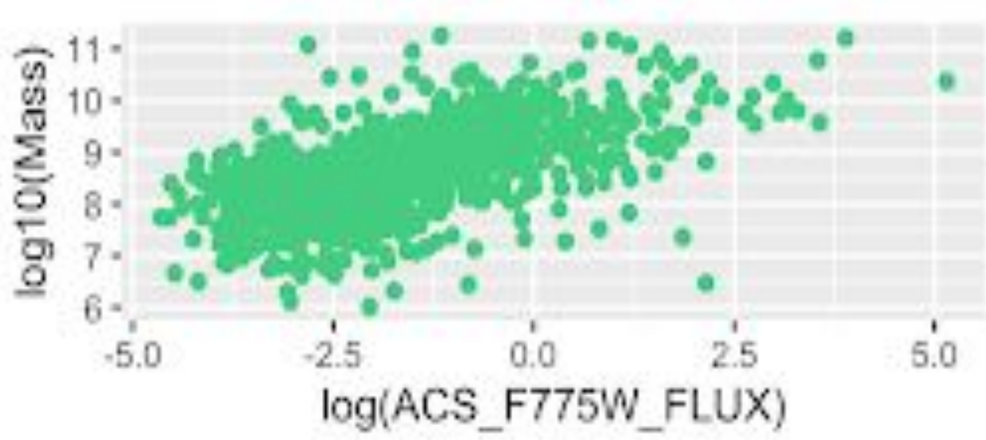
Data Overview

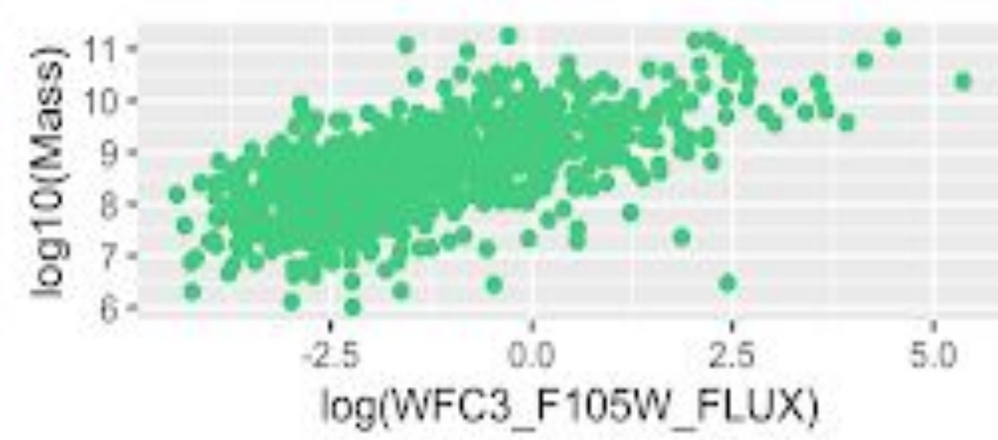
Our data (Barro et al. (2019)) consists of 15 predictors for 13,359 observations,, including 13 measures of brightness and two sky coordinates.

RA, DEC	Celestial longitude and latitude
KPNO_U_FLUX	Brightness of galaxy as observed in the U band at Kitt Peak National Observatory
LBC_U_FLUX	... in the U' band at Large Binocular Telescope
ACS_(F435W, F606W, F775W, F814W, F850LP)_FLUX	... at five wavelengths (0.435 microns, etc.) using Hubble's Advanced Camera for Surveys
WFC3_(F105W, F125W, F140W, F160W)_FLUX	... at four wavelengths (1.05 microns, etc.) using Hubble's Wide-Field Camera
MOIRCS_K_FLUX	... in the K band at the Subaru Telescope
CFHT_Ks_FLUX	... in the Ks band at the Canada-France-Hawaii Telescope

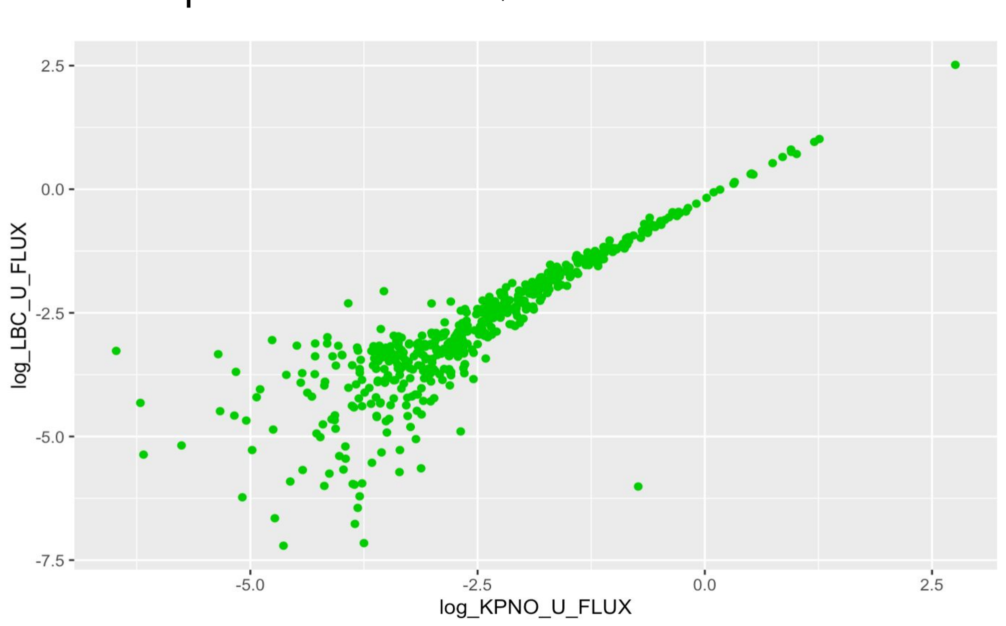


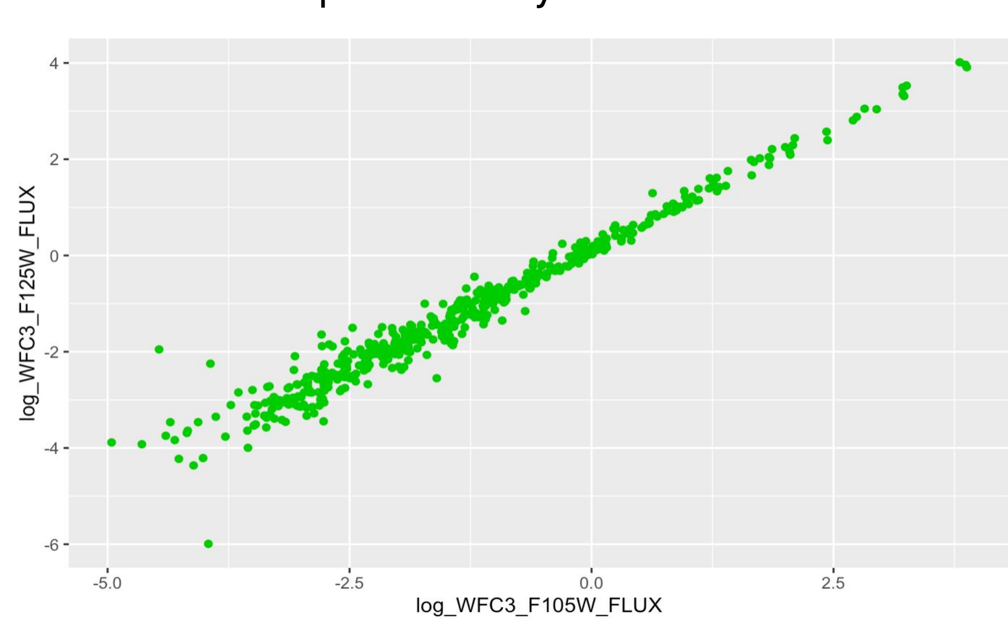
Since all brightness predictors are skewed, we have log transformed them for visualization. They all have a moderate linear relationship with the response.





The brightness predictors are also linearly correlated to one another. Although our goal is prediction, in the next section we will see how multicollinearity affects the model performance, should we seek better model interpretability.





Analysis and Results

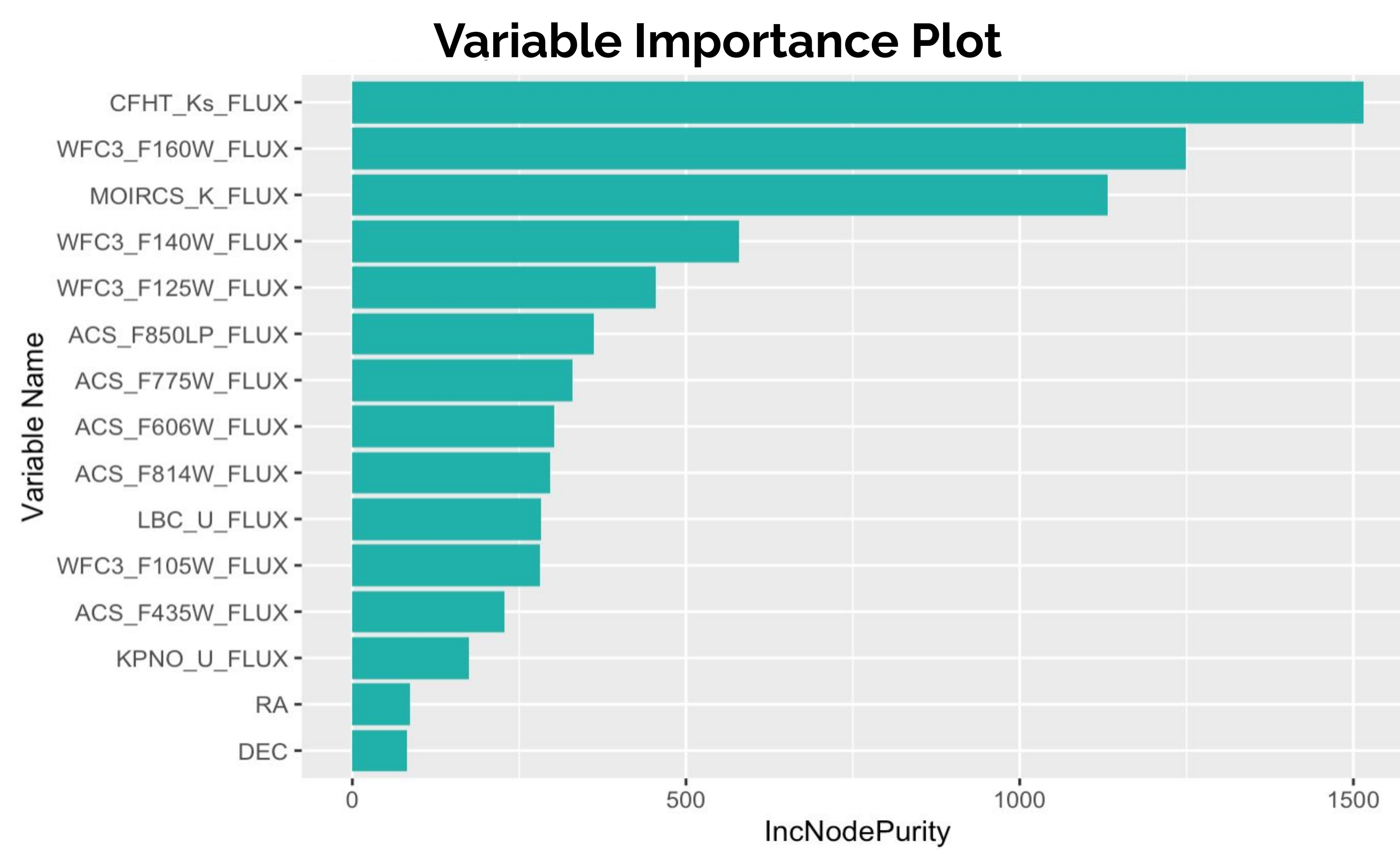
- Among the 15 predictors, all 13 brightness variables are log-transformed, but not the two sky coordinates.
- We split our data set into 70% (9351 galaxies) as the training set and 30% (4008 galaxies) as the test set.
- Following models are included:
  - Four parametric regression models: linear regression, best subset selection, lasso regression, and ridge regression.
    - To deal with collinearity, all parametric regression models are first fit on the full predictor space (15 variables), and then fit on the vif-reduced predictor space (two sky coordinates and four brightness measurements).
  - Four machine learning models: decision tree, random forest, gradient boosting, and k nearest neighbors (KNN).

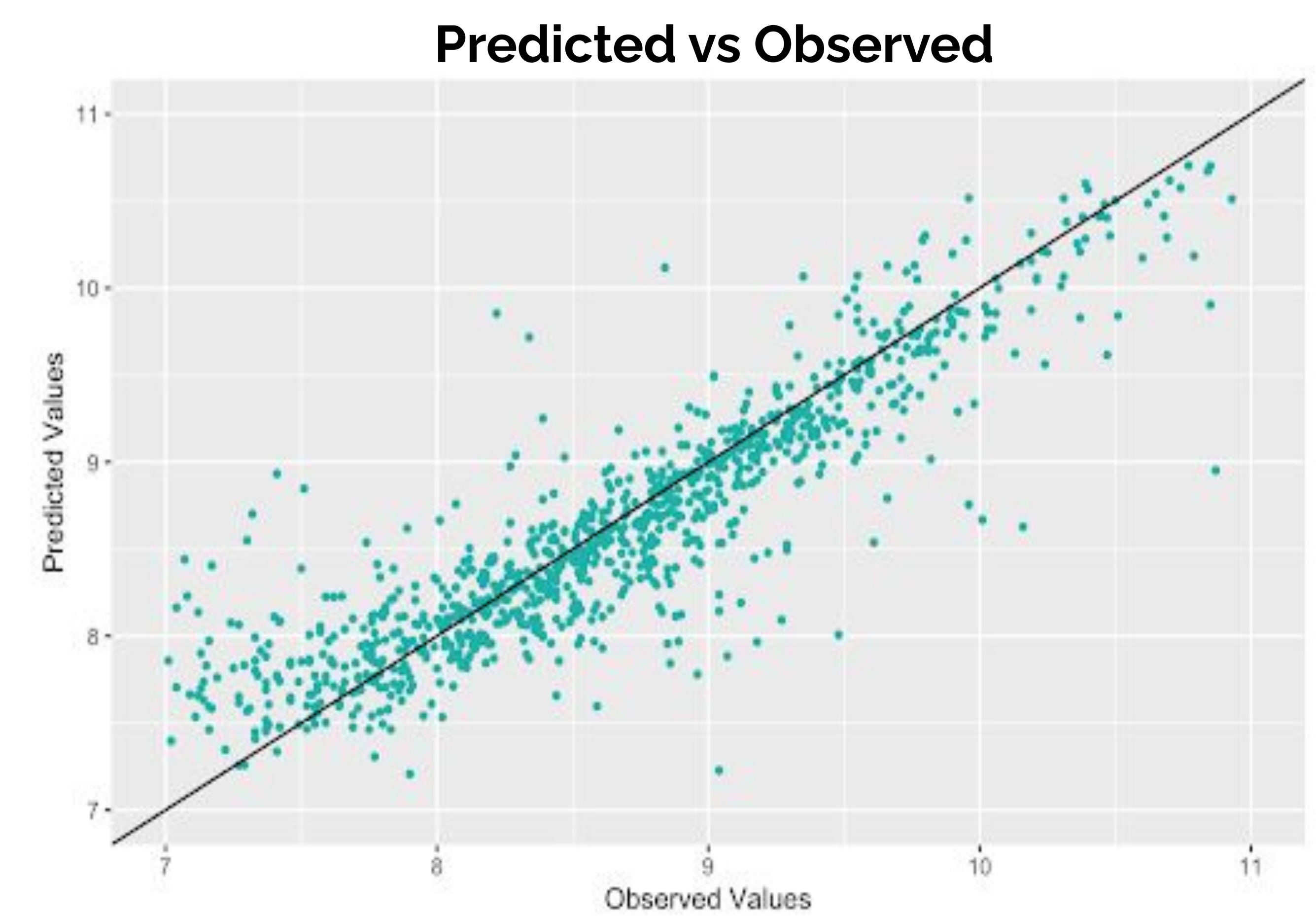
The test-set MSEs are as follows:

Parametric regression models	full	vif-reduced
Best Subset Selection with BIC	0.239	0.360
Lasso Regression	0.239	0.360
Linear Regression	0.239	0.360
Ridge Regression	0.279	0.365

Machine Learning Models	
Random Forest	0.172
K Nearest Neighbors	0.198
Gradient Boosting	0.205
Decision Tree	0.371

As we can see, random forest performs the best. The plot on the left side below shows the variable importance (in decreasing order) in random forest. CFHT_Ks_FLUX, WFC3_F160W_FLUX and MOIRCS_K_FLUX are the most important ones, while the sky coordinates are the least important. On the right is the predicted response from random forest versus observed response.





Conclusion

Random forest performs the best when predicting galaxy mass from brightness measurements, with a test-set MSE of 0.172 and a good predictive ability.

References: Barro, G. et al. 2019, The Astrophysical Journal Supplement Series, vol. 243, id. 22