Predicting Galaxy Mass From Sky Location and Brightness

By: Canzhou Qu, Ginny Zhao, Peter Wu, Serena Wang
Project Supervisor: Peter Freeman

Carnegie Mellon University
Statistics & Data Science

INTRODUCTION

Galaxies cannot be weighed on a scale. We can only estimate their masses based on what we can observe in images: how bright they are, their shapes, etc. In this project, we attempt to learn a statistical model that relates galactic observables to estimates of masses made using physics-based software.

DATA

In this project, we analyze the data of 219,812 galaxies observed by the Sloan Digital Sky Survey (SDSS) and provided in a galaxy properties catalog compiled at the University of Portsmouth.

Predictor Variables:

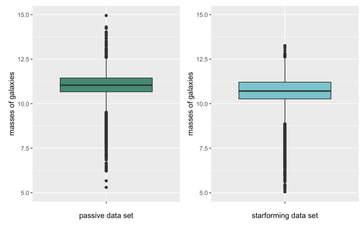
- Sky Location: ra, dec
- Brightness: mag.u, mag.g, mag.r, mag.i, mag.z
- Distance: redshift

The letters u,g,r,i,z represent different wavebands spanning the near-ultraviolet to the near-infrared. In our analyses, we attempt to learn models without the redshift variable, because astronomers do not have redshift measurements for over 99% of catalogued galaxies. We then see how including redshift improves mass prediction.

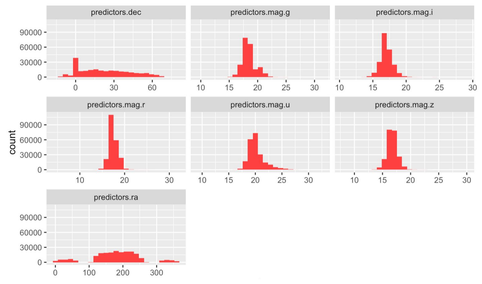
Response Variable:

- Mass

We have two datasets for galaxy mass estimates, made via two types of physics-based algorithms:
Passive: galaxies are not forming new stars now.
Star-Forming: galaxies are in the midst of forming new stars.
Thus, for each galaxy we have two distinct values of mass. We carry out two separate analyses, one for each dataset, and compare the results.



Mass estimates made assuming galaxies passively evolve are larger than if they are assumed to be star-forming.



We observe bimodality in the distribution for mag.g. Mag.u is skewed to the right.

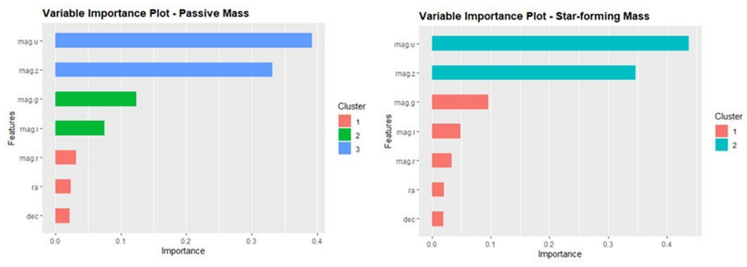
RESULTS AND ANALYSIS

We split our data with 75% used for training and 25% used for testing. The following is our MSE table:

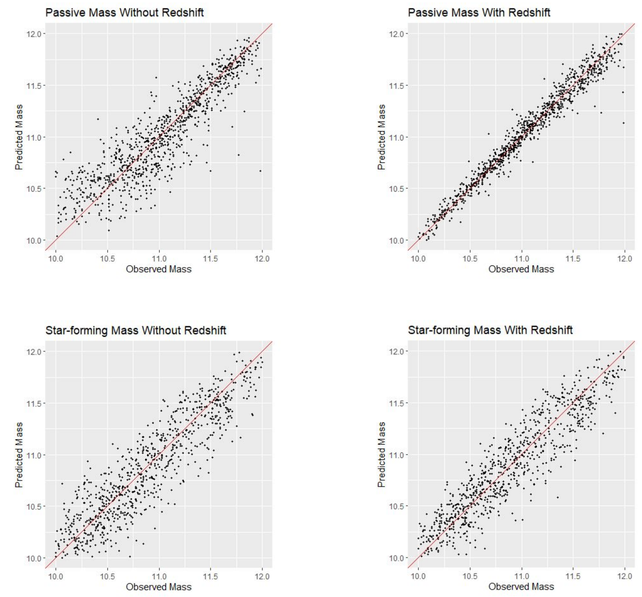
Model	Passive	Star-Forming
Linear Regression	0.097	0.110
Best Subset Selection	0.097	0.110
Regression Tree	0.115	0.173
Random Forest	0.067	0.074
XGBoost	0.064	0.074

If we assume galaxies are not forming stars, our learned models make better predictions.
XGBoost with 10-fold cross validation performed 100 times is the best model for both datasets.

We'd like to see how redshift would influence our models, so we use redshift as a predictor variable and fit our best model and compare the results.



Mag.u and mag.z are two most important predictor variables. Mag.g and mag.i have a stronger influence in the passive dataset compared to the star-forming dataset.



We observe the following two patterns:

- The model with redshift has more accurate predictions, whereas the model without redshift is more scattered around the diagonal line.
- The learned models make better predictions for passive galaxies than for star-forming galaxies.

Passive dataset, XGBoost model

- Mean squared error without redshift: 0.064
- Mean squared error with redshift: 0.013

CONCLUSIONS

We were able to successfully learn statistical models relating sky location and galaxy brightness to galaxy mass, for both when we assume the galaxies are passive and when we assume the galaxies are forming stars. If we are able to add a measurement of distance to the analysis, we find that the mean-squared error of mass predictions improves by a factor of five.

REFERENCES

http://www.sdss3.org/dr10/spectro/galaxy_portsmouth.php
Maraston, C., et al., Monthly Notices of the Royal Astronomical Society, v. 435, p. 2764