# Predicting Galaxy Masses in SDSS using Emission Data

*Authors : Kondareddy, T.; Miner, J.; Peng Q.; Sharma, K.; Frieden Templeton, W. ; Vasanthawada, S. R. S.*

Carnegie Mellon University

## Introduction

Spectroscopic measurements reveal the energy emitted by a source (galaxy, star, dust, gas, etc.) at specific wavelengths in the electromagnetic spectrum. Equivalent width is a measure of the strength of an emission line relative to the galaxy's brightness near the line. We use different statistical models to predict the mass for galaxies given the observed strengths of ten separate emission lines.

## Data Processing

We use the emission line strength data for galaxies from the Max Planck for Astrophysics - John Hopkins University (MPA-JHU) Group catalog[1,2,3] for all the galaxy spectra available in the Sloan Digital Sky Survey (SDSS) data release 8 (DR8). The dataset contains equivalent width measurements for ten distinct emission lines of 21,046 galaxies and their estimated galaxy masses.

As the Emission line data is right-skewed (Fig. 1), we perform a two-step transformation to scale the predictor variables such that they only have positive values. Post the transformation, outliers were identified through visual inspection and were removed from the dataset. After this preprocessing, the sample size of galaxies is reduced to 19,568.

Emission line data has strong multicollinearity (Fig. 3), meaning that many of the independent measurements are correlated. Multicollinearity reduces the ability to identify which emission spectra is the most valuable for predicting mass, but does not influence the overall predictive ability of trained models.

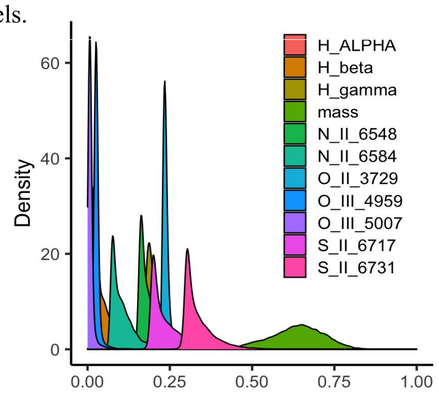


**Figure 1: Standardized dataset showing right-skewed emission spectra**

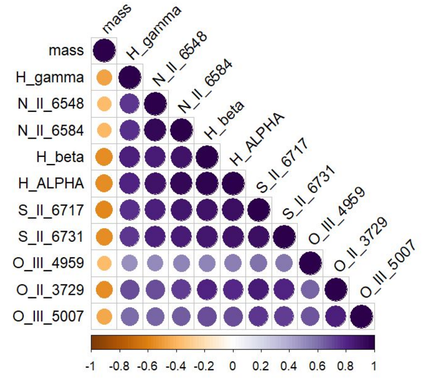


**Figure 3: Correlation plot for the predictor variables. Variables show strong multicollinearity**

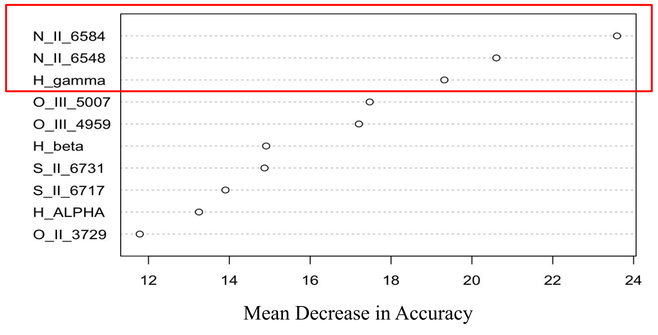


**Figure 2: Some emission spectra explain larger quantities of information more than others despite multicollinearity**

## Methods

We split the data into a training (80%) and test set (20%). We use five statistical models available in RStudio : Linear regression with Akaike Information Criteria, Decision Tree, Random Forest, Gradient Boosting, and K-Nearest Neighbors. For each of these models, the associated Mean-Squared Error is used to determine the goodness of fit.

## Analysis

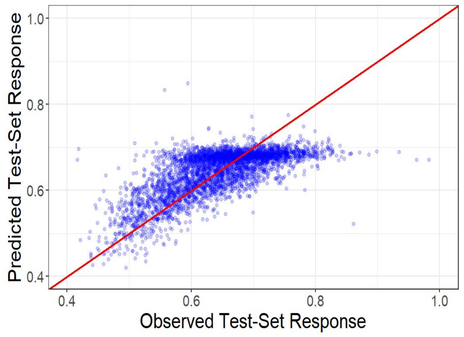


**Figure 4a: Linear regression**

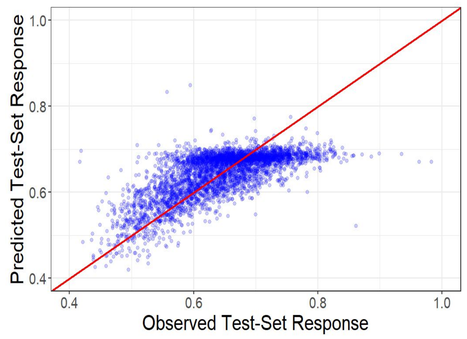


**Figure 4b: Best Subset GLM (AIC)**

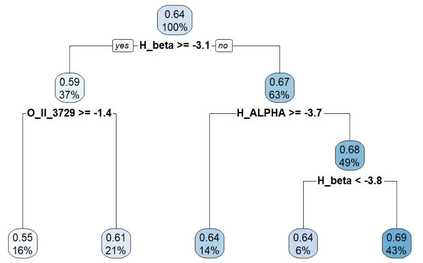


**Figure 4c: Decision Tree**

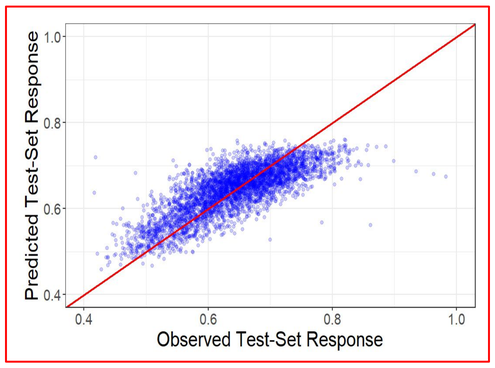


**Figure 4d: Random forest**

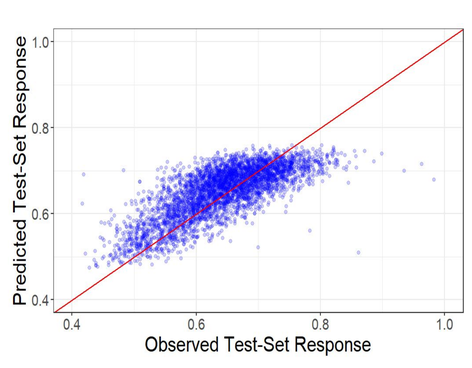


**Figure 4e: K-Nearest Neighbors**

**Table 1 : Mean squared errors (MSE)**

| Model | Mean-Squared Error ($x10^{-3}$) |
|---|---|
| Random Forest | 2.284 |
| Gradient Boosting | 2.369 |
| K-Nearest Neighbors | 2.601 |
| Linear Regression | 2.996 |
| Decision Tree | 3.489 |

## Conclusions

The current project utilizes the emission line data provided by the SDSS spectral signature to predict the galaxy masses.

- Galaxy mass can be approximately modeled using the available data and is generally underestimated across all models
- However, using Random Forest Algorithm, we show that the emission lines are strongly associated with the galaxy mass
- Emission lines such as `N_II_6584`, `N_II_6548`, and *`H_gamma`* are key drivers in predicting the galaxy mass

## References

1. Brinchmann J., Charlot S., White S. D. M., Tremonti C., Kauffmann G., Heckman T., Brinkmann J., 2004, MNRAS, 351, 1151
2. Kauffmann G., et al., 2003, MNRAS, 341, 33
3. Tremonti C. A., et al., 2004, ApJ, 613, 898