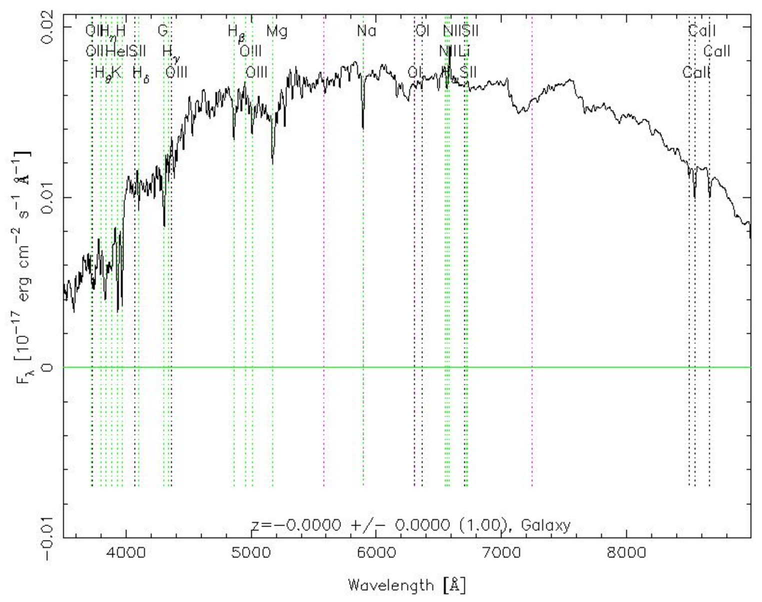
Predicting Galaxy Mass from SDSS Emission Data

By: Aniket Naik, Anthony Vetturini, Arnav Garcha, and Changcheng Ji

Carnegie Mellon University
Statistics & Data Science

Background and Introduction

The Sloan Digital Sky Survey (SDSS) has spent decades collecting data in an effort to make a map of the universe, and it has documented more than 2.86 million galaxies and 960,000 quasars[1]. One portion of its data collection includes atomic emission line observations, as these readings can inform us of astrophysical phenomena, such as the rate of star formation within galaxies[2]. **Herein, we use this atomic emission line data to create a predictive model that looks to predict the mass of an observed galaxy from the SDSS data.**



An example of an emission line reading. The "smooth" black line is the continuum of the measurements. The narrow troughs (or spikes) are specific emission lines cause by electron transitions. Here, the vertical lines name the emission line at a measured wavelength from the SDSS.

https://classic.sdss.org/dr5/algorithms/spectemplates/spDR2-022.gif

Exploratory Data Analysis

The SDSS dataset contains the strengths of 10 emission lines measured across 21,046 galaxies. We have emission data for various Hydrogen, Oxygen, and Nitrogen ionization states of varying wavelengths and are looking to predict the mass of the galaxy

Predictor Variables	Response Variable
H_{Alpha} O_{III_4959} N_{II_6584} S_{II_6717} H_{Beta} O_{III_5007} N_{II_6548} S_{II_6731} H_{Gamma} O_{II_3729}	Estimated Mass of Galaxy ($\log_{10}$ solar mass)

Data Transformation

Why? Predictor variables were heavily skewed leading to high error in prediction

1. Negative emission line values were removed. This removed 9,455 data points
2. The remaining data was first squared then log-transformed
3. The squared values were then $\log_{10}$ transformed

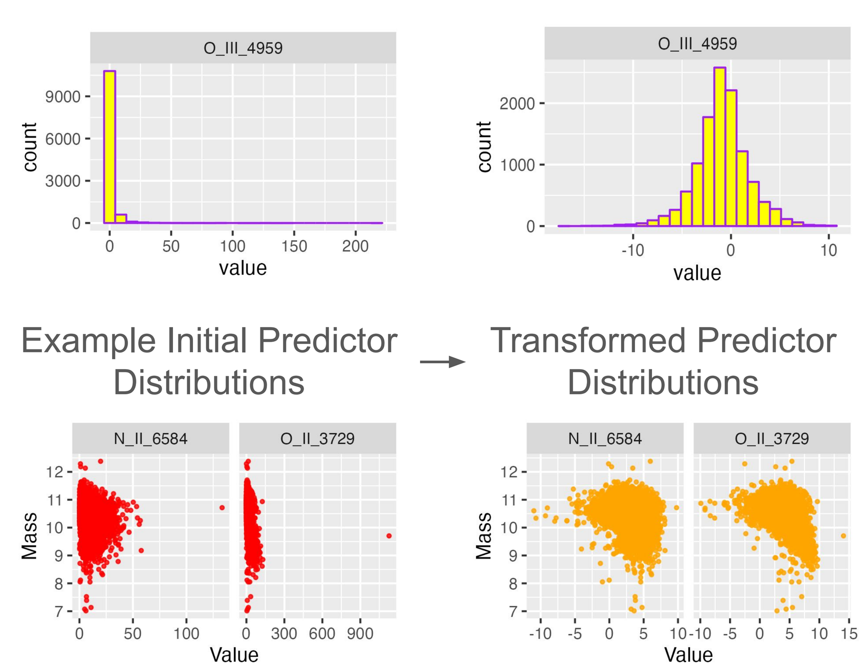


Figure 1. The transformed predictor variables (right) show similar distributions when measured against the response. The original data (left) has different distributions when measured against the response.

Methods

- The same training and test set were used for each model. A 70/30 training and test split were used. A random seed was imposed.
- We used several statistical learning models, including linear regression, random forest, regression trees, and XGBoost to predict galaxy mass.
- To evaluate the models, mean squared error (MSE) of the ground truth mass and the predicted mass from test set were calculated.

Analysis and Results

- In the non-transformed training, best-subset selection with linear regression eliminated 3 predictor variables: H_{gamma}, O_{III_4959}, and N_{II_6548}. In the transformed training, all of the predictors were retained (hence the same MSE as standard linear regression).
- With the transformed data, XGboost produced the smallest MSE (Fig 2, Fig 3). In the non-transformed data, MSE was minimized with XGBoost.

Table 1. Results for the test set MSEs for the non transformed predictor variables (right) vs. the transformed dataset (left) The transformed dataset performed marginally better with XGboost and Regression Trees.

Non Transformed Data	
Model Type	MSE
Linear Regression	0.12305
LR-BSS	0.12304
Regression Tree	0.10577
Random Forest	0.07637
XGBoost	0.05114

Transformed Data	
Model Type	MSE
Linear Regression	0.08549
LR-BSS	0.08549
Regression Tree	0.10233
Random Forest	0.06718
XGBoost	0.02576

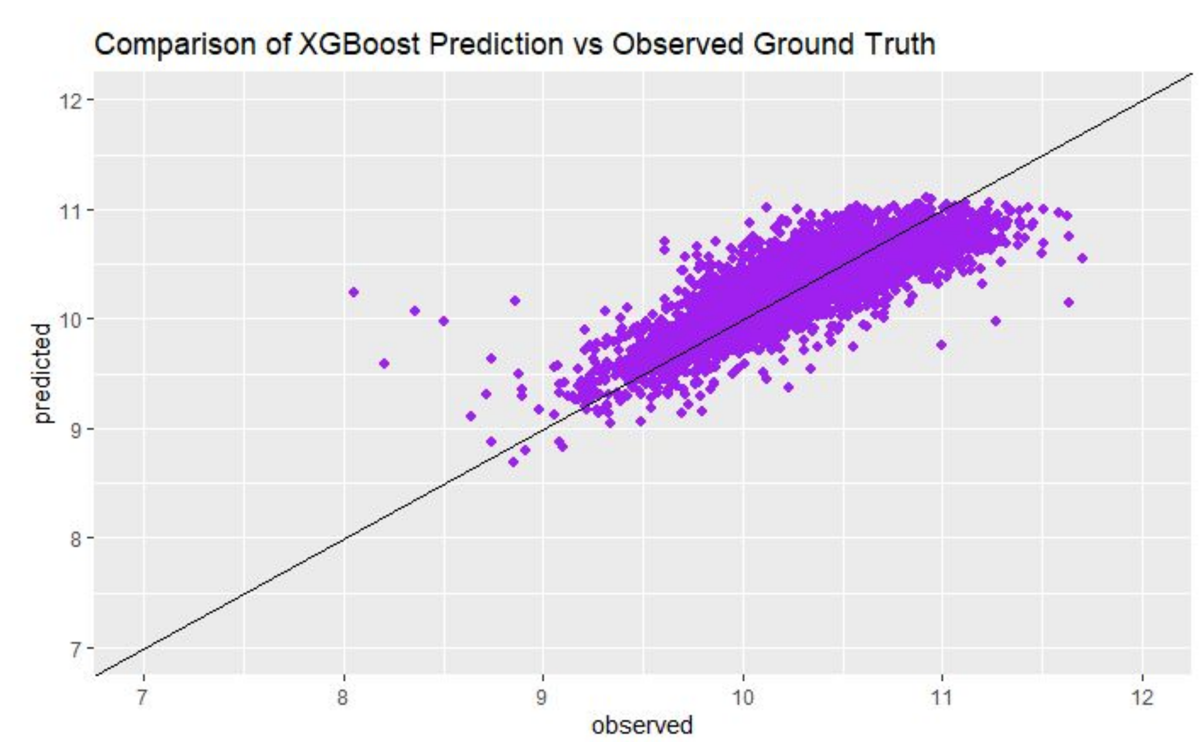


Figure 2. Diagnostic plot of the XGBoost test prediction on transformed data compared to ground truth.

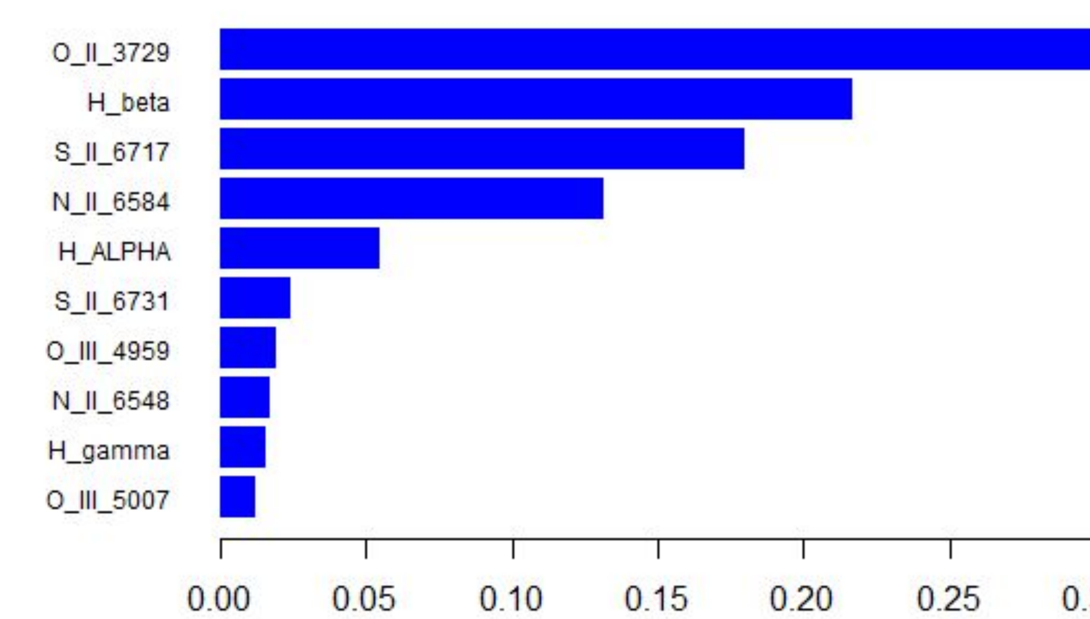


Figure 3. Variable importance plot for the XGBoost model trained on the transformed data.

Conclusions

- Transforming the data as described in the Exploratory Data Analysis led to smaller MSEs despite removing 9,455 data points resulting in a more accurate predictive model.
- We have successfully modelled the relationship between galaxy emission line data and the predicted log solar mass of a galaxy
- We have found that the best predictive model is found using XGBoost with the transformed data with an MSE of 0.02576

References

1) Funding for the Sloan Digital Sky Survey V has been provided by the Alfred P. Sloan Foundation, the Heising-Simons Foundation, the National Science Foundation, and the Participating Institutions. SDSS acknowledges support and resources from the Center for High-Performance Computing at the University of Utah. The SDSS web site is www.sdss.org.

2) von Der Linden, Anja, et al. "Star formation and AGN activity in SDSS cluster galaxies." *Monthly Notices of the Royal Astronomical Society* 404.3 (2010): 1231-1246.