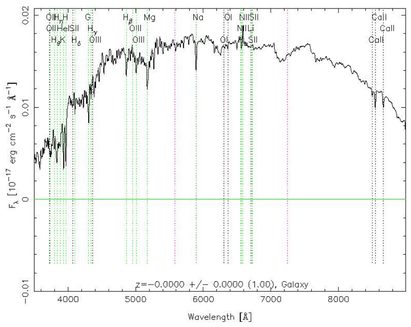
Carnegie Mellon University

36600-A Overview of Statistical Learning and Modeling-Fall 2024

Galaxy Mass Prediction from Emission Lines

Haipei Bie, Dominic Dimambro, Xiaoyu Huang, Christa Michel & Huawei Yu

Context & Introduction



Example spectrum from an “early-type” galaxy

Galaxies emit light at specific wavelengths, forming distinct emission lines in their spectra. These lines result from atomic transitions, where electrons move from higher to lower energy states. The strength of these lines, represented by their equivalent widths, is a measure relative to the surrounding continuum. This normalization ensures that equivalent widths remain consistent regardless of the galaxy’s distance from Earth. Understanding the relationship between emission line strength and galaxy mass can help improve models of galaxy formation and evolution.

This study aims to investigate the relationship between the strengths of various emission lines in galaxy spectra and their estimated masses. Specifically, it seeks to determine which emission lines are most strongly associated with galaxy mass and how well these emission lines can predict mass. Modeling this relationship will provide insights into how emission line strengths are connected to galaxy mass, helping us understand the processes involved in their formation and evolution.

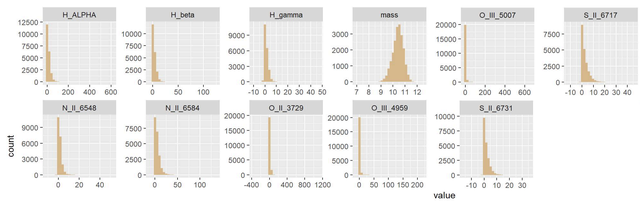
The primary goal of this study is to build a statistical model that accurately predicts galaxy mass and identifies which specific emission lines contribute most significantly to mass prediction based upon their respective strengths.

Data

Data were collected from a comprehensive galaxy survey, including measurements of emission line intensities and associated physical properties of galaxies. The original data consists of ten emission-line-strength measurements for each of 21,046 separate galaxies.

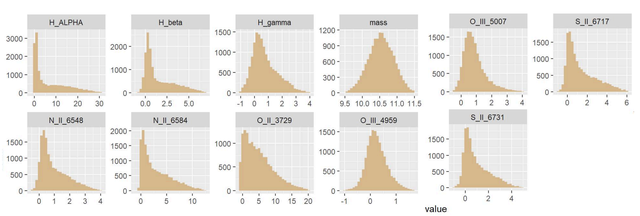
Variables:

- Emission Line Features: H-alpha, OIII, NII, and other prominent spectral lines.
- Target Variable: Galaxy stellar mass (in solar masses).

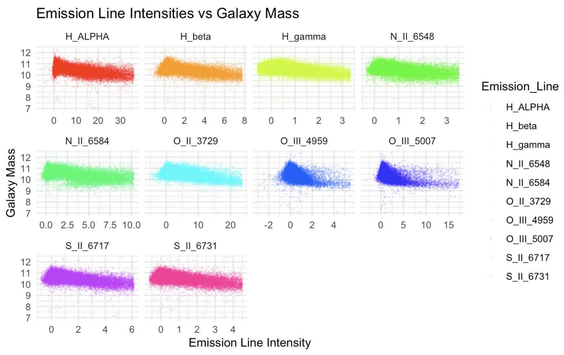


Preprocessing:

- Data cleaning to remove outliers and handle missing values.
- The cleaned dataset now includes 15,055 rows of data.



- Compare ‘mass’ with the predictor columns
- From the graph below, it shows that there is no clear evidence proving a robust relationship between the ‘mass’ and any single predictor. Therefore, the following sections we will try multiple models to try to find the relationship.



- Split the data into training and test data.
- A seed must be set to ensure the model learns the same data set each time. The training data set will account for 70% of the total, while the test data set will account for the rest.

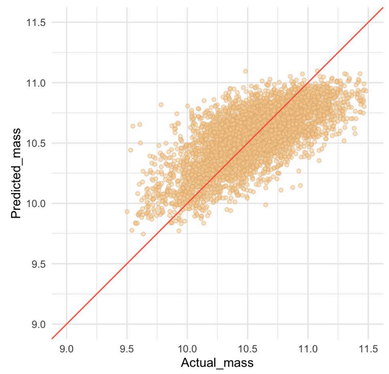
Methods

We learned the following statistical models:

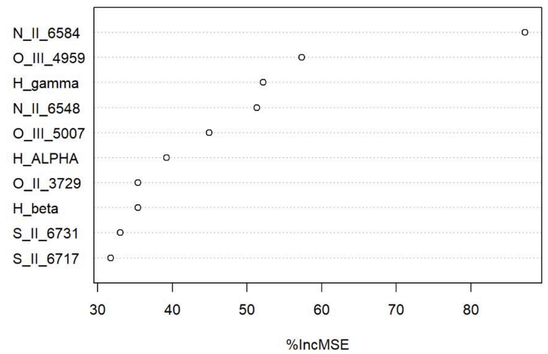
1. Linear Regression Model Creation: A baseline model to establish relationships between emission lines and galaxy mass.
2. Best Subset Selection: Used to select the most effective subset of predictors for the regression model.
3. Regression Tree: A decision tree-based model capturing non-linear relationships.
4. Random Forest: Ensemble learning technique used for feature importance analysis and non-linear modeling.
5. Extreme Gradient Boosting (XGBoost): A highly efficient implementation of gradient boosting.
6. K-Nearest Neighbors (KNN): A non-parametric model based on proximity to predict galaxy mass.

Analysis

Random Forest



Feature Importance from Random Forest Model



	% IncMSE
N_II_6584	87.25
O_III_4959	57.35
H_gamma	52.15
N_II_6548	51.31
O_III_5007	44.97
H_ALPHA	39.23
O_II_3729	35.44
H_beta	35.41
S_II_6731	33.05
S_II6717	31.74

Model Ranking:
In the following table the models are ordered from lowest to highest Test set MSE, showing Random Forest as the best model followed by XGBoost and KNN. The a plot of the predicted test-set response values versus observed test-set response values of the random forest model is shown above. According to the result, the R-squared of the learned random forest model is 0.51.

Order	Model Name	Test Set MSE
1	Random Forest	0.0634
2	XGBoost	0.0672
3	KNN	0.0752
4	Best Subset Selection	0.08
5	Linear Regression	0.0801
6	Regression Tree	0.0883

Conclusions

- The random forest model provided the most accurate estimated mass predictions based upon emission line strength with a MSE of 0.0634 and an R-squared of 0.51.
- Emission lines O_III_4959, H_gamma, and N_II_6548 are among the most important for predicting a galaxy’s estimated mass, with respective increases in MSE of 57.3%, 52.2%, and 51.3% when excluded.
- N_II_6584 is the single most important emission line in galaxy spectra for predicting a galaxy’s estimated mass, with its exclusion resulting in an 87.25% increase in MSE for the random forest model.

Reference

1. Sloan Digital Sky Survey (SDSS): Spectra templates and MPA-JHU catalog. Available at SDSS Spectemplates and MPA-JHU Stellar Masses.