Classification of Active Galaxies Observed by SDSS

By: Christina Choi, Sonal Suralikal
Project Advisor: Peter Freeman

Introduction

- Galaxy data typically include images along with measures of brightness from five different bandpasses (denoted u, g, r, i, and z) spanning the optical regime of the electromagnetic spectrum.
- These measures of brightness, or magnitudes, can reveal interesting information about galaxies.
- When a galaxy is *active*—meaning it forms stars at a relatively greater rate or has a supermassive black hole in its center that consumes stars and gas and dust at an enhanced rate—its spectrum will reveal "spikes" called emission lines. Astronomers can use the relative strengths of these emission lines to infer the type of activity occurring in a galaxy and by extension classify it.
- Galaxies can be labeled as having active nucleus or star-forming using the full spectra, but for most galaxies we only have the magnitudes.

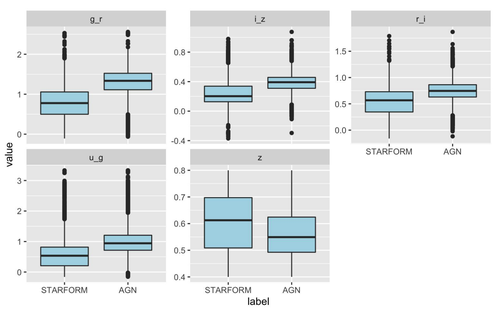
In this study, we attempt to classify galaxies as either star-forming or having an active nucleus using only the magnitude of each galaxy.

Data

- Our data contains information on 28,151 galaxies that have been labeled (through other means) as being star-forming galaxies or galaxies with active nuclei. We see that in this data set, 15,521 galaxies are star-forming galaxies and 13,299 galaxies are galaxies with active nuclei. The dataset is described in Zhang et al. (2019).

Variables	Description
u_g, g_r, r_i, i_z	The four colors of the galaxy. The colors are differences in logarithmic measures of brightness, or magnitudes. Magnitudes are highly correlated with each other as well as galaxy distance.
z	Galaxy redshift. Redshift refers to the ratio of the observed wavelength of a photon from an object to its wavelength when it was emitted, minus 1
label	The factor response variable in this dataset, that indicates whether a given galaxy is observed to be star-forming (denoted STARFORM) or with active nuclei (denoted AGN)

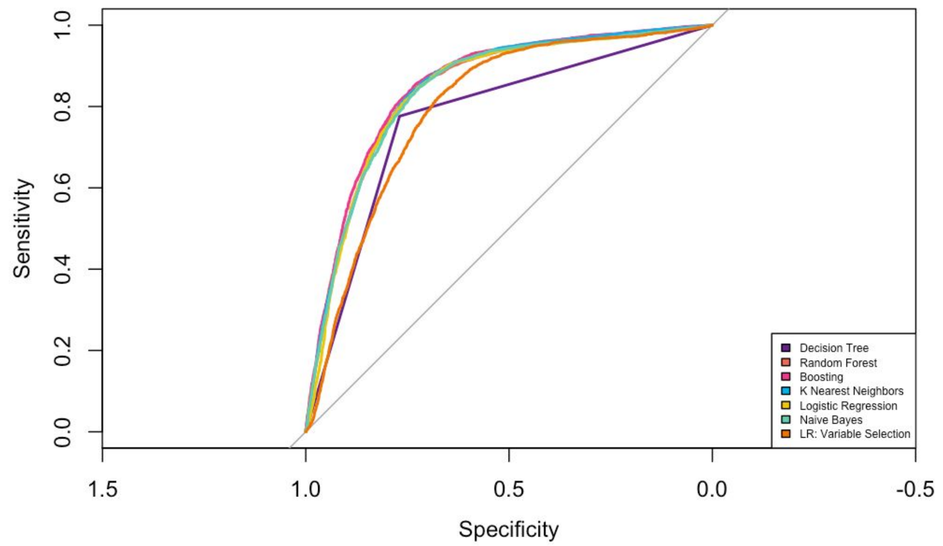
- The boxplots on the right show how the different predictor variables are distributed for each galaxy type (STARFORM and AGN).



Analysis

- We test a variety of classifiers (listed at right) and generate receiver operating characteristics curves for each. ROC curves illustrate the tradeoff between classifying members of each class well. The area under a ROC curve is dubbed AUC; the higher the AUC value, the better the model.
- After comparing the performance of several models, we determine that the models with the highest AUCs are Random Forest, Boosting, and K Nearest Neighbors, with values of 0.844, 0.852, and 0.846 respectively.
  - We determine that Extreme Gradient Boosting is the optimal model in this case.

Comparing ROC Curves for Each Classifier



Model	AUC
Logistic Regression	0.839
LR: Variable Selection	0.797
Decision Tree	0.773
Random Forest	0.844
Boosting	0.852
K Nearest Neighbors	0.846
Naive Bayes	0.841

- We determine optimal class predictions using Youden's J statistic, which balances prediction performance across both classes.
  - We find the optimal threshold for Extreme Gradient Boosting to be 0.404, which is used as the threshold for making class predictions
  - The associated misclassification rate is 0.213.

Confusion Matrix for Boosting Model

		Actual	
		STARFORM	AGN
Predicted	AGN	1452	3822
	STARFORM	3929	650

Conclusion

Overall, we are able to determine the optimal model to classify active nucleus versus star-forming galaxies. In order to find the best classifier, we use the metric of determining which model has the greatest AUC value. Based on this metric, we find that Extreme Gradient Boosting is the optimal model with a misclassification rate of 21.3%.

References

Freeman, P.E. 2021, online at https://github.com/pefreeman/36-290/blob/master/PROJECT_DATASETS/ACTIVE_CLASS/README.md

Zhang, K., et al. 2019, "Machine Learning Classifiers for Intermediate Redshift Emission Line Galaxies", The Astrophysical Journal, online at arxiv.org/pdf/1908.07046.pdf