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 nuclei 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).

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.

Confusion Matrix for Boosting Model

<table>
<thead>
<tr>
<th></th>
<th>AGN</th>
<th>STARFORM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted</td>
<td>1452</td>
<td>3822</td>
</tr>
<tr>
<td>Actual</td>
<td>3929</td>
<td>650</td>
</tr>
</tbody>
</table>

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