Classification of White Dwarfs Observed by SDSS

By: Bo Xia, Kaylin Li, Roochi Shah, Yixuan Wu
Advisor: Peter Freeman

The Sloan Digital Sky Survey (SDSS) has observed high-resolution spectra, in addition to brightness over five different bandpasses, for many objects known as white dwarfs (WDs). WDs are remnants of low-mass stars like the Sun, and their spectra are historically classified into a number of types. Our goal is to see if we can identify WDs of spectral type DA given easily obtained, non-spectroscopic information.

DATA

The dataset based on the catalog of Kepler et al. (2014) contains information for 9,112 white dwarfs labeled as being either DA (spectral type A) or NOT-DA. There are eight predictor variables:

<table>
<thead>
<tr>
<th>Variables</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ra, dec</td>
<td>celestial longitude and latitude</td>
</tr>
<tr>
<td>umag, gmag, rmag, imag, zmag</td>
<td>magnitude in each SDSS bandpass</td>
</tr>
<tr>
<td>pm_tot</td>
<td>yearly motion relative to background stars</td>
</tr>
</tbody>
</table>

The response classes are imbalanced, with 79.5% DA and 20.5% NOT-DA. We remove one row with an outlier value for zmag. The summaries for the five magnitudes are similar. As for pm_tot, its values are 0 or >2, and to make the positive values less skew, we performed a logarithmic transformation.

The figure to the right shows that there are significant differences in the magnitude variables between the two classes, which might indicate possible associations. Meanwhile, ra and dec are relatively similar between the two classes.

ANALYSIS

We split the dataset, retaining 75% for training and 25% for testing, and learn multiple classification models, as listed in the table below. We compute a receiver operating characteristics (ROC) curve for each model, which illustrates the tradeoff between making accurate predictions in each class. We select the model with the largest area under the ROC curve (AUC) and utilize Youden's J statistic to generate class predictions. In the figure at right, the estimated probability of being NOT-DA is shown for objects of each class. The dotted line corresponds to the optimized Youden's J value; WDs with probabilities below the line, for instance, are predicted to be of class DA. The metric used to evaluate classification success was AUC, or area under the ROC curve. Out of the models tested, Random Forest had the highest AUC of approximately 0.826.

CONCLUSION

Given our dataset, we found that the best model for predicting whether a white dwarf was of spectral type A was a Random Forest model with an AUC of 0.826 and a misclassification rate of 0.233. Therefore, we can conclude that we can determine the spectral type of a white dwarf given information about its brightness, location on the sky, and apparent movement on the sky. A possible next step to improve prediction accuracy is collecting more data on white dwarfs that are not of spectral type A.

References:
Introduction: Freeman, P. E. 2021, online at https://github.com/pefreeman/36-290/blob/master/PROJECT DATASETS/WD CLASS/README.md