Gina Krynski
STA 467
Final Report
11 May 2022

Classification of Stellar Objects

Background and Motivation

The Sloan Digital Sky Survey is a project dedicated to creating detailed 3-dimensional maps of the Universe. Data collection began at the University of Chicago in 2000, and is still continued today. The dataset used for this report contains 100,000 observations from space, taken by the SDSS. The observations are classified as stars, galaxies, or quasars. A quasar is a luminous supermassive black hole that is feeding on gas in the center of a galaxy. They are visible to radio telescopes because they emit radiation across the electromagnetic spectrum, such as radio waves (Young 2021).

This data set intrigued me because it gives insight into the process of how astronomers classify newly observed stellar objects. The spectral characteristics in the dataset that are used to classify were new to me and I enjoyed learning about which characteristics are more insightful than others. The cataloging of stellar objects has given astronomers the ability to map the distribution of stars, which gave rise to the understanding that they make up our own Milky-Way galaxy. I am very much interested to continue to learn about the new findings astronomers discover as the telescopes and other technology they use become more and more advanced. My partner for this project is Simon Louisin, together we discussed the initial exploration of the data as well as modeling techniques.

Data Description

The SDSS dataset contains 100,000 observations and has a total of 18 columns. The column “class” identifies the observation as a star, galaxy, or quasar. 8 of the other columns were identification information, such as the run number to identify the specific SDSS scan the observation is from, a field number to identify the field the observation was found in, etc. All 8 identification columns were removed from the dataset because they are not considered stellar characteristics that are used to classify the observation. From my understanding, these identification columns are used by the SDSS when creating 3-d maps of the Universe, so they were not needed for classification. The Modified Julian Date column was also removed from the dataset, as it only states when the SDSS observation was taken. The 8 predictors used in the final dataset to classify the objects are:

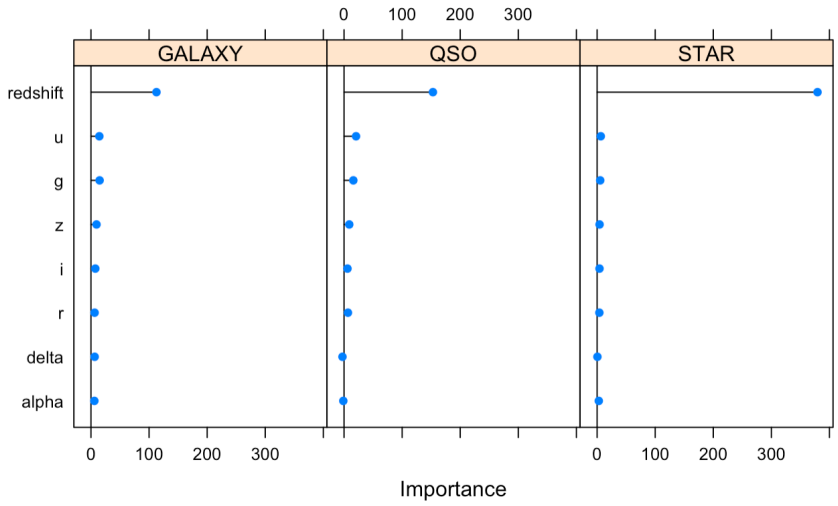
- alpha = Right Ascension angle (at J2000 epoch)
- delta = Declination angle (at J2000 epoch)
- u = Ultraviolet filter in the photometric system
- g = Green filter in the photometric system
- r = Red filter in the photometric system
- i = Near Infrared filter in the photometric system
- z = Infrared filter in the photometric system
- redshift = redshift value based on the increase in wavelength

Methods

Prior to modeling the data, a random sample of 10,000 observations was taken from the dataset in order to reduce the time needed to run each of the models and to knit the final R Markdown file. Also, the sample was split into training and testing sets, which were 8,000 and 2,000 observations respectively. 5-fold cross validation was the validation approach that was used consistently for each of the four models. The seed was set prior to sampling, splitting, and modeling the data to ensure reproducible results.

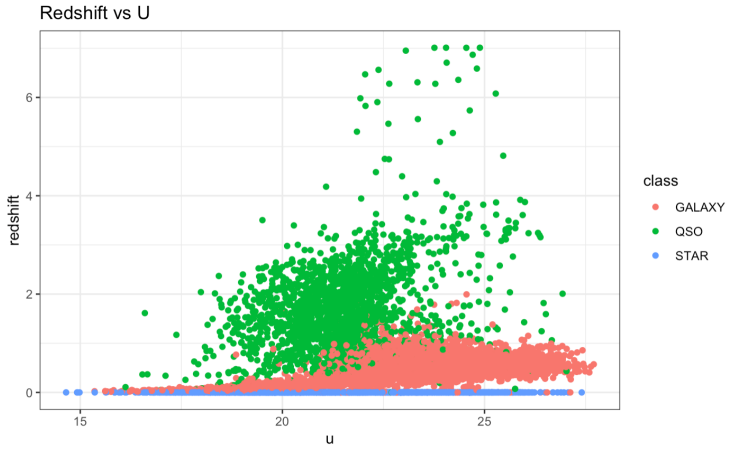
EDA

To begin the exploratory data analysis, a random forest model was created to look at variable importance.



For classification trees, variable importance uses the Gini index to calculate the importance of all predictors in the model. The Gini index measures node purity, which tells us when a node contains observations from mostly one class. The variables that have the largest decrease in the mean Gini Index are considered more important when computing variable importance (James 2021). The plot above compares the importance of the 8 predictors in our model. *Redshift* has been reported to be the most important of the predictors. Predictors *u* and *g* are the second and third most important. They are shown to be much less important than *redshift* but only slightly more important than *z* and the rest of the predictors.

Several exploratory scatterplots were created to get a better understanding of how the predictor related to each other and to the three classes. Below is a scatterplot of *Redshift vs U*



This plot stood out to me from the manu scatterplots that were made during the EDA because it shows how stars almost always have a near 0 redshift. Also, it is clear that quasars typically have the highest redshift values, as galaxies do not often have redshift values greater than 2. The other scatter plots made during the EDA are available in the Appendix.

Model 1

Linear discriminant analysis was used to create the first model. Linear discriminant analysis is a method that approximates the Bayes classifier by plugging in estimated values for the mean and probability that an observation belongs to the *kth* class as well as estimating the variance that is assumed to be shared by all *k* classes (James 2021).

```
Linear Discriminant Analysis

8000 samples
   8 predictor
   3 classes: 'GALAXY', 'QSO', 'STAR'

Pre-processing: centered (8), scaled (8)
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 6400, 6399, 6400, 6401, 6400
Resampling results:

  Accuracy   Kappa
  0.8440019  0.703433
```

The results of training the LDA model with the training data show an 84.4% accuracy rate. The model was then used to predict the classes of the observations in the test set. This resulted in a misclassification rate of 14.6%

Model 2

The next model that was created was a Partial Least Squares Classifier. PLS is a dimension reduction method that identifies a new set of features that are created from linear combinations of the original features (James 2021).

```
Partial Least Squares

8000 samples
   8 predictor
   3 classes: 'GALAXY', 'QSO', 'STAR'

Pre-processing: centered (8), scaled (8)
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 6400, 6401, 6400, 6400, 6399
Resampling results across tuning parameters:

  ncomp  Accuracy   Kappa
  1      0.6924992  0.3308766
  2      0.7532501  0.4775584
  3      0.8122490  0.6370164
  4      0.8167503  0.6478244
  5      0.8237501  0.6635624
  6      0.8311259  0.6772619
  7      0.8257506  0.6673624
  8      0.8268760  0.6693600

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was ncomp = 6.
```

The PLS model was tuned using the number of components from 1 to 8, since there are a total of 8 predictors in our sample dataset. The final value of the ncomp tuning parameter was 6, which produced an accuracy of 83.1%. When this model was tested by predicting the classes of the observations from the test set, the misclassification rate was 16.9%.

Model 3

The third model that was fit was a random forest model. In random forests, a number of decision trees are built and at each split, a random sample of *m* predictors is chosen as split candidates from the full set of predictors. A new sample of *m* predictors is chosen for each split (James 2021).

```
Random Forest

8000 samples
   8 predictor
   3 classes: 'GALAXY', 'QSO', 'STAR'

No pre-processing
Resampling results across tuning parameters:

  mtry  Accuracy  Kappa
   1    0.965375  0.9380453
   2    0.973500  0.9527084
   3    0.974750  0.9548436
   4    0.976750  0.9583624
   5    0.976125  0.9572470
   6    0.975875  0.9567994
   7    0.975125  0.9554524
   8    0.975250  0.9556806
   9    0.974750  0.9547892
  10    0.975250  0.9556887

Accuracy was used to select the optimal model using the
 largest value.
The final value used for the model was mtry = 4.
```

The random forest model was then used to predict the classes of the test observations. This resulted in a misclassification rate of 2.4%, which is the lowest misclassification rate of all of the models thus far.

Model 4

A support vector classifier was created as the fourth and final model. The support vector classifier is based on a hyperplane that purposely does not perfectly separate two or more classes. This is also called a soft margin classifier because some observations are allowed to be on the wrong side of the margin or hyperplane (James 2021).

```
Support Vector Machines with Linear Kernel

8000 samples
   8 predictor
   3 classes: 'GALAXY', 'QSO', 'STAR'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 6400, 6400, 6401, 6399, 6400
Resampling results across tuning parameters:

  C     Accuracy   Kappa
   0.1  0.9368742  0.8867802
   1.0  0.9486247  0.9087554
  10.0  0.9577497  0.9244716

Accuracy was used to select the optimal model using the
 largest value.
The final value used for the model was C = 10.
```

The C tuning parameter was tried at values of 0.1, 1 and 10, and the final value used for the model was C=10. The model produced a misclassification rate of 3.5% when used to predict the classes of the test set observations.

Model Comparison

The four models were compared by their misclassification rate. Below is a table reporting those values for each of the models.

Model	Missclassification
LDA	0.1460
PLS Classification	0.1690
Random Forest	0.0240
SVC	0.0345

The Random Forest model produced the lowest misclassification rate, with the support vector classifier having a similar but slightly larger misclassification rate. The LDA and PLS models performed significantly worse than the other two models.

Closing Remarks

The final model was a random forest model with an mtry tuning parameter value of 4, meaning 4 predictors were tried at each split. As stated in the results of the random forest model below, the quasar class had the highest classification error among all of the classes.

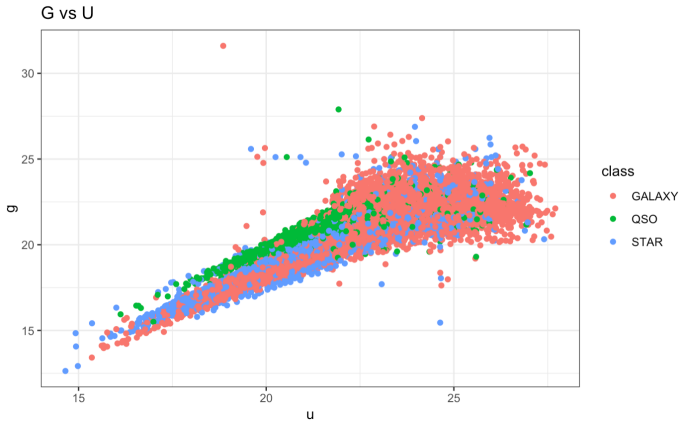
```
Call:
 randomForest(x = x, y = y, ntree = 200, mtry = min(param$mtry,
ncol(x)))
               Type of random forest: classification
                     Number of trees: 200
No. of variables tried at each split: 4

        OOB estimate of  error rate: 2.38%
Confusion matrix:
       GALAXY  QSO STAR class.error
GALAXY   4718   64    7  0.01482564
QSO       115 1429    1  0.07508091
STAR        3    0 1663  0.00180072
```

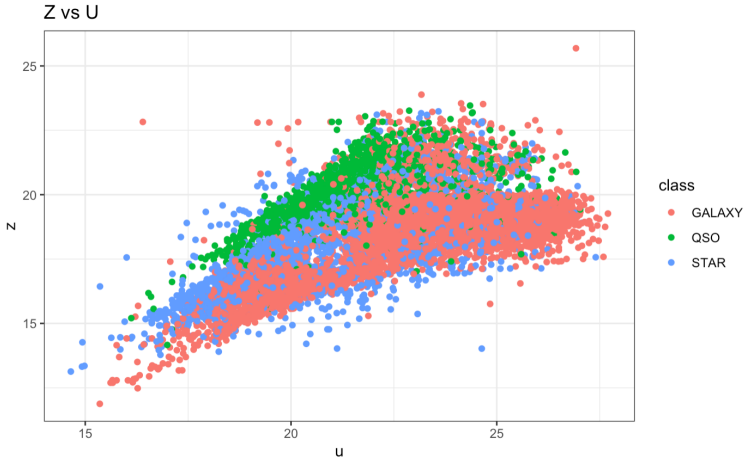
The star class had the lowest classification error from the trained model, with only 3 out of 1,666 star observations being misclassified. Interestingly, when galaxies and quasars were misclassified, they were not misclassified as stars. From these results, it appears that galaxy and quasar observations are more similar to each other than stars, and thus the model has a more difficult time classifying them correctly.

Appendix

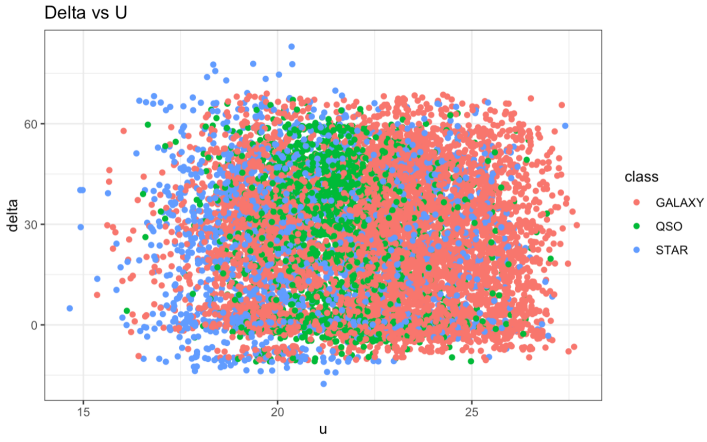
A.



B.



C.



References

Young, Monica. “What Is a Quasar?” *Sky & Telescope*, Sky & Telescope, 28 Sept. 2021, https://skyandtelescope.org/astronomy-resources/what-is-a-quasar/.

James, Gareth, et al. *An Introduction to Statistical Learning: With Applications in R*. Springer, 2021.