MET CS 677 O2 Final Project

Conrad Korzon | 4/21/2024

# Overview

For this project I use the Star Classification data set from Kaggle3, produced by the Sloan Digital Sky Survey, which consists of 100,000 observations of cosmic entities with 8 features each, in which each entity is either a Star, Galaxy, or Quasar. Specifically, the observation rows in the data set consist of 21,594 Stars, 59,445 Galaxies, and 18,961 Quasars, thus more than half of the data set is made up of Galaxy observations while Stars and Quasars are less represented.

The selected features in the data set which will be used for classification models are as follows:

|  |  |
| --- | --- |
| Feature | Description |
| Alpha | Right Ascension angle |
| Delta | Declination angle |
| 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 |

The goal of this project is to find an effective classifier for predicting the type of stellar entity given an observation, thus the following classifiers are trained, tested, and compared using the data set:

1. K-Nearest Neighbors
2. Logistic Regression
3. Gaussian Naïve Bayes
4. Random Forest

The K-Nearest Neighbors and Random Forest classifiers undergo hyperparameter tuning to select the optimal hyperparameter(s) before moving on to predictions analysis and model comparison.

# K-Nearest Neighbors

A graph of blue bars

Description automatically generated with medium confidenceK-NN is chosen as the first simple classification algorithm which can classify the cosmic entities in our data set. To find the optimal K, several values for K are tested, the accuracies of the classifiers for each K are compared, and the K value with the highest overall classifier accuracy is chosen­­1. K value selection findings and confusion matrix observations below:

* As an outcome of K value testing, K=5 was selected as the parameter with the highest overall classifier accuracy.
* A close-up of a color scheme

  Description automatically generatedOverall Accuracy: **94.45%**
* K-Nearest Neighbors performed well, particularly in correctly classifying a high percentage of galaxies and stars correctly.

# Logistic Regression

A close-up of a colorful grid

Description automatically generatedWhile logistic regression is often used as a binary classifier, meaning it predicts whether data belongs to a single class or not1, the SciKit Learn LogisticRegression module can be used for multi-class cases such as the Star Classification data set. It does this by using a one-vs-rest technique (OVR) in which it calculates the probability of each class individually then normalizing the values4, effectively splitting the problem into several layers of binary classification. Testing outcomes and confusion matrix observations below:

* Overall Accuracy: **95.53%**
* Logistic regression performed very well, particularly only misclassified **9** stars out of 5399.
* Performed slightly better than K-NN for Stars and Galaxies, slightly worse for Quasars.

# Naïve Bayes

A purple and white squares

Description automatically generatedSince the Star Classification data is continuous, Gaussian Naïve Bayes is used to predict cosmic entity classes. Testing outcomes and confusion matrix observations below:

* Overall Accuracy: **74.58%**
* Naïve Bayes performed quite poorly.
  + Predicted many Stars as Galaxies
  + Predicted many Galaxies and Stars as Quasars.
* Since Gaussian Naïve Bayes assumes multivariate normal distribution2, poor performance may be accounted for by non-normal distribution in data.
* Upon inspection, data summary hints that distribution may not be Normal, particularly for certain features.
  + Features *u, g, z,* and *redshift* all have their mean much closer to one end of their min-max range than the other.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | alpha | delta | u | g | r | i | z | redshift |
| count | 100000 | 100000 | 100000 | 100000 | 100000 | 100000 | 100000 | 100000 |
| mean | 177.6291 | 24.13531 | 21.98047 | 20.53139 | 19.64576 | 19.08485 | 18.66881 | 0.576661 |
| std | 96.50224 | 19.64467 | 31.76929 | 31.75029 | 1.85476 | 1.757895 | 31.72815 | 0.730707 |
| min | 0.005528 | -18.7853 | -9999 | -9999 | 9.82207 | 9.469903 | -9999 | -0.00997 |
| 25% | 127.5182 | 5.146771 | 20.35235 | 18.96523 | 18.13583 | 17.73229 | 17.46068 | 0.054517 |
| 50% | 180.9007 | 23.64592 | 22.17914 | 21.09984 | 20.12529 | 19.40515 | 19.0046 | 0.424173 |
| 75% | 233.895 | 39.90155 | 23.68744 | 22.12377 | 21.04479 | 20.3965 | 19.92112 | 0.704154 |
| max | 359.9998 | 83.00052 | 32.78139 | 31.60224 | 29.57186 | 32.14147 | 29.38374 | 7.011245 |

# Random Forest

A graph with blue bars

Description automatically generated with medium confidenceRandom Forest is an ensemble learning technique which uses hard voting between a number of decision trees to make predictions2. This classifier was selected as the final classifier algorithm, and first underwent hyperparameter selection for max depth and number of classifiers before fully analyzing test data predictions. Findings presented below:

* As an outcome of hyperparameter selection, the combination of N=12 and D=9 produced the highest overall accuracy and thus were used in the classifier moving forward.
* A close-up of a colorful background

  Description automatically generatedOverall accuracy: **97.52%**
* Random Forest achieved the highest accuracy of all models.
* Only misclassified 5 Stars out of 5,395, and 268 Galaxies out of 14,861.
* Very effective at predicting Stars; only 62 Star predictions were incorrect.

# Comparisons

Confusion Matrix & Accuracy Comparisons by Classifier:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | K-Nearest Neighbors (K=5) | Logistic Regression | Naïve Bayes | Random Forest (N=12, D=9) |
| TS | 5098 | 5390 | 806 | 5394 |
| FS-G | 300 | 3 | 3817 | 5 |
| FS-Q | 1 | 6 | 776 | 0 |
| TG | 14270 | 14347 | 13653 | 14593 |
| FG-S | 415 | 291 | 6 | 60 |
| FG-Q | 176 | 223 | 1202 | 208 |
| TQ | 4244 | 4145 | 4187 | 4392 |
| FQ-S | 10 | 4 | 1 | 2 |
| FQ-G | 486 | 591 | 552 | 346 |
| TSR | 0.944249 | 0.998333 | 0.149287 | 0.999074 |
| TGR | 0.960231 | 0.965413 | 0.918713 | 0.981966 |
| TQR | 0.895359 | 0.874473 | 0.883333 | 0.926582 |
| Accuracy | 0.94448 | 0.95528 | 0.74584 | 0.97516 |

*\*See Terminology section below for row-label definitions.*

Based on overall accuracy, the classifier ranking is:

1. Random Forest
2. Logistic Regression
3. K-Nearest Neighbors
4. Naïve Bayes

The top 3 models all performed very well and were comparable in accuracy, while Naïve Bayes fell behind the pack significantly with an accuracy of only ~74.6% as opposed to the >94% range of the others. Particularly, Naïve Bayes performed terribly for predicting Stars, only achieving a ~15% TSR, as it predicted the majority of Stars to be Galaxies and many other to be Quasars. This poor performance can likely be accounted for by a non-normal distribution in the data, particularly in a subset of features, as discussed in the Naïve Bayes section above.

Of the remaining three classifiers, while they each achieved high accuracy scores, they all performed slightly worse in predicting Quasars compared with Stars and Galaxies. Because each of these three models had relatively high FQ-G and FG-Q scores compared with other misclassification scores, it seems the most challenging aspect of the predictions for each was differentiating between Galaxies and Quasars.

Lastly, an interesting observation is that Random Forest outperforms the other 3 classifiers in *all* categories (Accuracy, TSR, TGR, & TQR), while Logistic Regression mostly outperforms K-NN and Naïve Bayes apart from TQR score. Thus, it seems that Random Forest would be the optimal classifier for this data set regardless of which metric we look to maximize.

# Terminology

|  |  |
| --- | --- |
| Term | Description |
| TS | True Stars. Classifier predicted ‘STAR’ and was correct. |
| TG | True Galaxy. Classifier predicted ‘GALAXY’ and was correct. |
| TQ | True Quasar. Classifier predicted ‘QSO’ and was correct. |
| FS-G | False Star - actual Galaxy. Classifier predicted ‘STAR’, but entity was Galaxy. |
| FS-Q | False Star - actual Quasar. Classifier predicted ‘STAR’, but entity was Quasar. |
| FG-S | False Galaxy - actual Star. Classifier predicted ‘GALAXY’, but entity was Star. |
| FG-Q | False Galaxy - actual Quasar. Classifier predicted ‘GALAXY’, but entity was Quasar. |
| FQ-S | False Quasar - actual Star. Classifier predicted ‘QSO’, but entity was Star. |
| FQ-G | False Quasar - actual Galaxy. Classifier predicted ‘QSO’, but entity was Galaxy. |
| TSR | True Star Rate |
| TGR | True Galaxy Rate |
| TQR | True Quasar Rate |

# Sources

[1] Enxing, J. (2024, March 28). MET CS 677 Module 3 Live Classrooms. Boston, MA, USA.

[2] Enxing, J. (2024, April 10). MET CS 677 Module 5 Live Classrooms. Boston, MA, USA.

[3] fedesoriano. (January 2022). Stellar Classification Dataset - SDSS17. Retrieved April 10, 2024, from <https://www.kaggle.com/fedesoriano/stellar-classification-dataset-sdss17> .

[4] scikit-learn developers. (2024). LogisticRegression. scikit-learn. <https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html>