Nickolas Mertens

COMP 4334: Parallel & Distributed Computing

June 9, 2022

Machine Learning Models for Classification of Stellar Objects

# Motivation

The motivation for the current analysis stems from a personal interest in space and a long-held desire to work for NASA. When I was a kid, I remember watching Star Wars, Star Trek, and other space-related shows and wishing for nothing else but to travel through space. This fascination with the universe carried through to my applications to undergraduate degree programs where I specified my desired major as “Astrophysics”. However, at the time my enjoyment of mathematics was not where it is today, and I decided not to pursue a career in Astrophysics because I thought all I would be doing is mathematical equations. Still, I never lost my fascination with “the final frontier”. For me, looking up at the night sky and comprehending that the tiny lights I was seeing was light from a distant star that has traveled millions upon millions of lightyears to reach Earth always takes my breath away. Now that I have begun to pursue a career in Data Science and a newfound respect for the role of calculus and mathematics in our daily lives has sparked my interest in space once more. The present analysis is the first analysis I have done on space data, but I doubt it will be the last.

# Data

The dataset utilized for the current analysis is Stellar Classification data from the Sloan Digital Sky Survey (SDSS) collected from Kaggle (<https://www.kaggle.com/datasets/fedesoriano/stellar-classification-dataset-sdss17?resource=download>). The SDSS is an open-source data repository that includes data from scientist and volunteer telescopes to help map out the night sky. The dataset is made up of 17 different feature columns and one column of Classification data.

|  |  |
| --- | --- |
| *Table 1.* Variable Descriptions | |
| Variable Name | Variable Description |
| objID | Object Identifier |
| alpha | Right Ascension angle (at J2000 epoch) |
| delta | Declination angle (at J2000 epoch) |
| ultra | Ultraviolet filter in photometric system |
| green | Green filter in photometric system |
| red | Red filter in photometric system |
| nearInfra | Near Infrared filter in photometric system |
| infra | Infrared filter in photometric system |
| runID | Run ID for specific scan |
| rerunID | Rerun ID to specify how image was processed |
| camCol | Camera column to identify scanline with run |
| fieldID | Field number |
| specObjID | ID used for optical spectroscopic objects |
| objClass | Object class (galaxy, star, quasar) |
| redshift | Redshift value based on increase in wavelength |
| plate | Plate ID |
| MJD | Modified Julian Data (when data was collected) |
| fiberID | Fiber ID for fiber that pointed the late at the focal plane |

This data consists of 100,000 observations collected by the SDSS. For the present analysis only the following variables were used as features: alpha, delta, ultra, green, red, nearInfra, infra, and redshift. The other columns were not included because they only had information pertaining to how the data was collected and not features of the images themselves.

The first two variables, alpha and delta, refer to the positions of the celestial bodies being recorded at the SDSS. The alpha parameter is the to the east-west coordinate of the object and the delta parameter is the north-south coordinate. Together these variables represent the location of the celestial body in the night sky. The next five parameters, namely ultra, green, red, nearInfra, and infra refer to which wavelengths along the electromagnetic spectrum the object is emitting. The wavelength measurements can tell a lot about the objects we view in the night sky. Primarily by measuring which wavelengths an object emits astronomers and scientists can assess things like the luminosity of a given object, the chemical composition of that object, and the density of those chemicals (Belleville, 2019). The final parameter, redshift, refers to the spectral displacement of a celestial object that is caused by the Doppler Effect, which describes the change in the wavelengths of objects when either one or both objects are moving in relation to one-another (Britannica, 2022). Within the context of space observation this redshift is critical since both Earth and these stellar objects are moving at such high speeds through space. The final column used in this analysis is the classification column, which denotes whether an object is a star, galaxy, or quasar object. Quasars are extremely luminous objects in space that are caused by black holes in space.

For the analysis the data are split into a training and testing set. The train-test split was set to 70-30 given the large amount of data available.

# Machine Learning Problem

The primary problem to be solved with this dataset is to classify celestial objects into one of three categories: star, galaxy, quasar. Three separate machine learning (ML) models were trained and then fitted to the training data to see if there were any differences between model performance. The three classification ML models that were ran are Logistic Regression, Decision Tree, and Random Forest. The Accuracy, Precision, Recall, and F1 scores for each of the models is reported in Table 2 below.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| *Table 2.* Evaluation Metrics for Machine Learning Classification Models | | | | | | | |
|  | Evaluation Metrics | | | | | | |
|  | Accuracy |  | Precision |  | Recall |  | F1 |
| Models |  |  |  |  |  |  |  |
| Logistic Regression | .963 |  | .999 |  | .990 |  | .995 |
| Decision Tree | .949 |  | .992 |  | .992 |  | .992 |
| Random Forest | .958 |  | .988 |  | .990 |  | .989 |

Clearly, all the classification models performed extremely well, which is to be expected given that the training data included 70,000 data points. However, only one model is needed to run on the test data. The Logistic Regression was the highest performing model but was not selected as the final model to run against the test data as this model had already been used in the course prior to this project. Therefore, the Random Forest model was selected to run against the training data.

# Streaming

To setup the stream the testing data was repartitioned to exist on 100 RDDs. From there, the testing data was saved to the DataBricks File System stored in a folder named “starsTestData”. After the models were trained on the training dataset and the Random Forest model was selected to run against the model the stream was setup to initiate the streaming of the test data. Critically, instead of using the .writeStream() function used throughout the course, the .transform() method on the trained Random Forest model was ran on the sourceStream variable and the label, probability, and prediction columns were selected. Finally, using the display() command the stream was started and the results were updated and displayed with each new stream occurrence. A sample image of the output is shown below.

Text

Description automatically generated

# Issues

## Issue 1 – Including the “redshift” variable

When I ran the analysis originally, I did not include the redshift variable as a parameter. Consequently, the models performed poorer compared to the models that did include this variable. The models that did not include the “redshift” parameter were displaying accuracy rates around 75%. Meanwhile, the accuracy of the models that did include the parameter was roughly 95%. Clearly, these results illustrate the importance of including the “redshift” values when classifying celestial bodies.

## Issue 2 – Model Errors

When I originally thought of this problem, I wanted to run a series of classification machine learning models beginning with a Decision Tree, Random Forest, and Gradient Boost classifiers. However, the current GBTClassifier class in pyspark only works with binary classifications and the current analysis had three classifications. Therefore, I used the Logistic Regression classifier in place of the GBTClassifier to ensure that model performance would be consistently assessed across the models included in the analysis.

# References

Belleville, M. (2019, September 27). *Explore—Light* [Text]. NASA. <http://www.nasa.gov/content/explore-light>

Britannica. (2022). *Redshift | Definition & Facts | Britannica*. <https://www.britannica.com/science/redshift>