Not-Project1

Kirsten Bauck

2025-02-07

Introduction

I chose to explore this stellar classification dataset because of my fascination with astronomy and the challenges it presents in physics and data science. The dataset, derived from the Sloan Digital Sky Survey (SDSS), tackles one of astronomy's fundamental tasks: classifying celestial objects. In this case, classifying them as stars, galaxies, or quasars. This can be a challenge when it comes to identifying minority classes like quasars, which despite being less common, play a much needed role in helping us understand our universe.

This data is significant as it plays a foundation role in astronomical research. Stellar classification is key to understanding the fundamental properties of celestial objects, including their temperature, composition, and evolutionary stage. The more scientists can accurately classify stellar objects, the better they can understand how such objects interact with each other. Classifications such as these have historically led to discoveries such as the distinction between the Andromeda galaxy and out own Milky Way. From this, I hope to learn how data science techniques can be used to analyze and explore a dataset in such a way that makes future modeling more efficient.

Regarding the data source, I have a low level of confidence. While this dataset is available on Kaggle and was curated by a data scientist with a double masters credentials, it represents a processed version of SDSS data. This curated process, while probably making the datasource more accessible for analysis does add a layer of separation from a primary source and should not be used in a real analysis

Data Cleaning

Since this data was curated, the data cleaning process was relatively simple and few challenges were encountered. The data cleaning process worked as follows. First, I calculated and removed any extreme outliers from the dataset. Then I changed the modified Julian Date (MJD) column to a simple date column with year, month, day. Finally, I removed any unnecessary columns within the dataset, such as extraneous ID fields and renamed columns to be more easily interpretable.

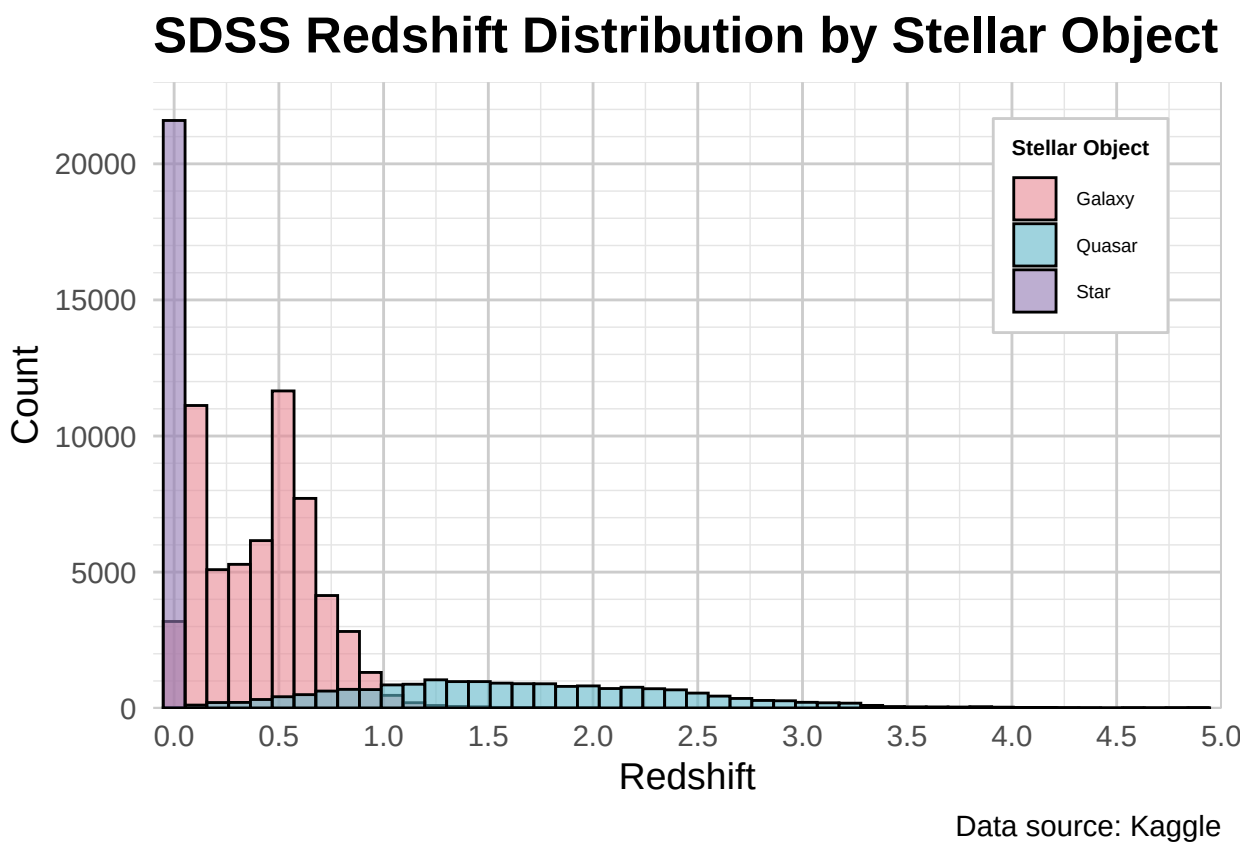
The dataset consists of 99,999 observations collected by SDSS, describing celestial objects on 10 features. The key variables include:

1. Spectral measurements:
    - Ultraviolet, green, red, and infrared filter readings
    - Redshift measurements (indicating the Doppler shift of light)
2. Positional data in the form of Right Ascension angle and Declination angle
3. Classification label
    - Three categories GALAXY (majority class), QSO (quasars), and STAR
    - Galaxies represent over 50% of the observations, with quasars and stars making up the remainder

Class	Right Ascension Angle	Declination Angle	Ultraviolet Filter	Green filter	Red Filter	Near Infrared Filter	Infrared Filter	Redshift	Date
GALAXY	135.689107	32.4946318	23.87882	22.27530	20.39501	19.16573	18.79371	0.6347936	2013-03-03
GALAXY	144.826101	31.2741849	24.77759	22.83188	22.58444	21.16812	21.61427	0.7791360	2018-02-09
GALAXY	142.188790	35.5824442	25.26307	22.66389	20.60976	19.34857	18.94827	0.6441945	2011-01-31
GALAXY	338.741038	-0.4028276	22.13682	23.77656	21.61162	20.50454	19.25010	0.9323456	2017-10-13
GALAXY	345.282593	21.1838656	19.43718	17.58028	16.49747	15.97711	15.54461	0.1161227	2012-09-17
QSO	340.995120	20.5894763	23.48827	23.33776	21.32195	20.25615	19.54544	1.4246590	2011-10-21
QSO	23.234926	11.4181876	21.46973	21.17624	20.92829	20.60826	20.42573	0.5864546	2018-12-04
GALAXY	5.433176	12.0651860	22.24979	22.02172	20.34126	19.48794	18.84999	0.4770090	2012-10-10
GALAXY	200.290475	47.1994023	24.40286	22.35669	20.61032	19.46490	18.95852	0.6600120	2013-04-04
STAR	39.149691	28.1028416	21.74669	20.03493	19.17553	18.81823	18.65422	-0.0000079	2006-12-13
GALAXY	328.092076	18.2203105	25.77163	22.52042	20.63884	19.78071	19.05765	0.4595958	2011-10-18
GALAXY	243.986638	25.7382804	23.76761	23.79969	20.98318	19.80745	19.45579	0.5914091	2011-05-12
STAR	345.801874	32.6728679	23.17274	20.14496	19.41948	19.22034	18.89359	0.0000718	2013-09-05
GALAXY	331.502030	10.0358020	20.82940	18.75091	17.51118	17.01631	16.62772	0.1521936	2011-06-27
GALAXY	344.984770	-0.3526158	23.20911	22.79291	22.08589	21.86282	21.85120	0.8181597	2016-10-21
GALAXY	244.824523	25.1545640	24.88680	22.13311	20.44728	19.49171	18.97470	0.4849288	2011-05-12
STAR	353.201522	3.0807959	24.54890	21.44267	20.95315	20.79360	20.48442	-0.0004286	2011-10-30
QSO	1.494389	3.2917463	20.38562	20.40514	20.29996	20.05918	19.89044	2.0315280	2016-10-02
STAR	14.383135	3.2143262	21.82154	20.55730	19.94918	19.76057	19.55514	-0.0004403	2015-12-17
GALAXY	167.131669	67.3399356	20.48292	18.67807	17.61680	17.11936	16.73351	0.1115879	2001-01-20

Above is a glimpse into the tidied dataset. Using this dataset we can predict a stellar object's class.

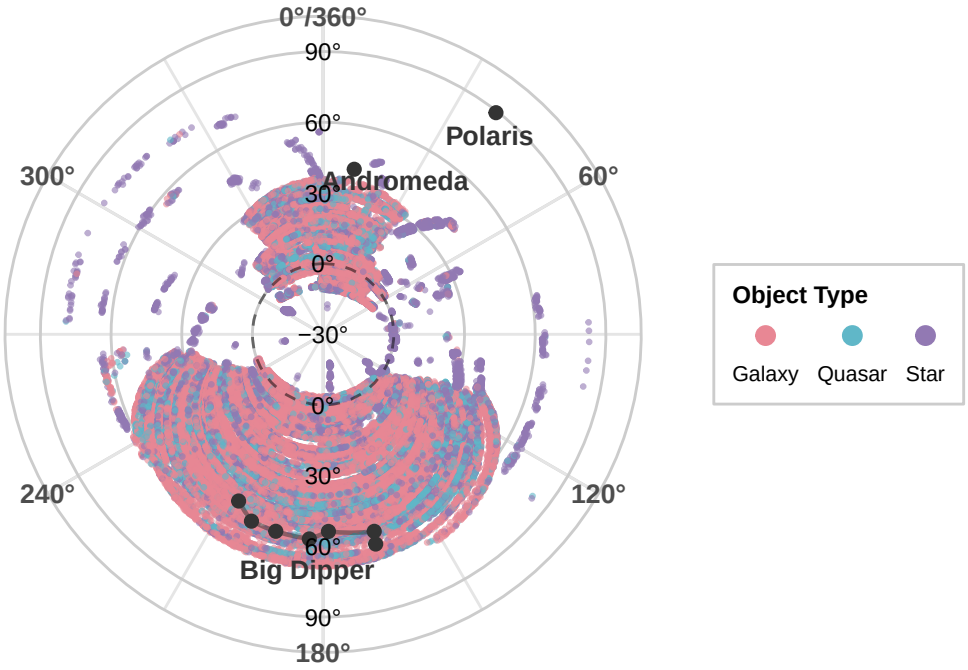
Data Visualizations



This distribution of redshifts shows a clear structure in our observable universe. Stars, with their redshift values clustered at zero, are primarily local objects within the Milky Way galaxy. This is because redshift corresponds to the ratio between the observed wavelength of light and its original wavelength. As such, the higher the redshift, the further away the object is and the further back in time we are seeing it. Galaxies show a broader distribution of redshifts between 0 and 1, representing the galaxies in our cosmic neighborhood and into the medium-distant universe. Quasars, showing the highest and most widely distributed redshifts (up to 3.5), are among the most distant objects we can observe. These high redshifts are a consequence of them being extremely luminous, as they are powered by super massive black holes. This makes them visible even at great distances where normal galaxies or stars would be too faint to detect. This distribution pattern demonstrates how different classes of astronomical objects occupy different time periods in the cosmos. Quasars serve as a way to let us peer deeper into the early universe then most other stellar objects.

Map of SDSS Stellar Objects in Northern Sky

Featuring the Big Dipper, Andromeda Galaxy, and Polaris (North Star)



This SDSS sky map reveals the systematic way in which this subset of data was chosen from the SDSS survey, with two primary survey regions clearly visible. These two regions are the larger northern galactic cap region (120-240 degrees RA) and the southern galactic cap region, which is divided into two (0-40 and 320-360 degrees RA). The choice in these two regions show how this dataset avoids the dusty galactic plane, which is filled with dust clouds and dark lanes. By avoiding this region, one can maximize the surveys ability to accurately detect, measure, and classify extragalactic objects. In this map we can see that galaxies dominate numerically and appear to be uniformally distributed within survey lines across the two survey regions. Quasars are more sparsely scattered but follow a similar pattern to galaxies. Stars, while also within the survey regions, also notably form an arc-like structure between the two main survey regions. I included familiar reference points like the Big Dipper and Polaris to help orient viewers and provide scale context. Through this visualization, we can analyze the two distinct survey regions in order to get a more fundamental understanding of our universe.

Conclusions

Given more time, there are a couple more analysis that I would have done. First of all, I would want to do a full feature importance analysis, where I could quantify the contribution of each feature to a model. In addition, I would also like to use techniques such as PCA to visualize the class separability in the high dimensional space of all the variables (filters, redshift, and position) interacting. Finally, I would also want to go more in depth into an error analysis, where I examine misclassified objects and try to identify any systemic biases or challenging regions in the space.

In a previous analysis of this SDSS data, various machine learning models were applied to classify the stellar objects. A summary of the models. their performance, advantages, and limitations is provided below:

Table 1: Model Summary and Comparison

Model	Test.Error	Advantages	Limitations
Decision Tree	4.17	Interpretable, Visualizable	Prone to overfitting, Unstable
Random Forest	2.06	Handles non-linear relationships, Robust to outliers	Computationally expensive, Less interpretable
AdaBoost	2.41	Reduces bias and variance, Focuses on misclassified examples	Sensitive to outliers, Computationally intensive

The Random Forest model, having an ensemble apprpoach, demonstrated the lowest test error, indicating its effectiveness at capturing some of the complex patters in the dataset while also reducing noise. AdaBoost also performed well, as it was able to leverage its ability to refine predictions by focusing on misclassified examples but it was computationaly expensive. Throughout my analysis, it seemed that the redshift variable was the most important in classifying the stellar objects. From this previous analysis, I learned that balancing interpretability (Decision Tree) and accuracy (Random Forest) is a key challenge, especially as a datasets' dimensions increase and modeling becomes more computationaly expensive. Before my analysis I did not fully understand the point of an intepretable model that was known to perform less well. But after conducing my analysis, I now understand that the more interpretable models are necessary to try and understand what less interpretable models are doing in the background.

When it comes to analyzing this dataset further, there are some potential challenges. First of all, the class imbalance of the dataset consisting mostly of galaxies can easily bias any model created. In addition, as one tries to incorporate additional spectral features, analysis could become more computationally intensive and harder to understand. The fact that most of the filters are highly correlated could lead to redundancy in models, which then complicates feature importance interpretation and could even affect model stability. Most of this data is photometric as well, which is less precise compared to spectroscopic data, so the reliance on such data to classify objects could limit the classification accuracy, especially when distinguish between quasars and stars.

Link to GitHub Repository

```
devtools::install_github('KirstenBauck/StellarClassification')
```

Citations

fedesoriano. (January 2022). Stellar Classification Dataset - SDSS17. Retrieved November 30th, 2024 from https://www.kaggle.com/fedesoriano/stellar-classification-dataset-sdss17.