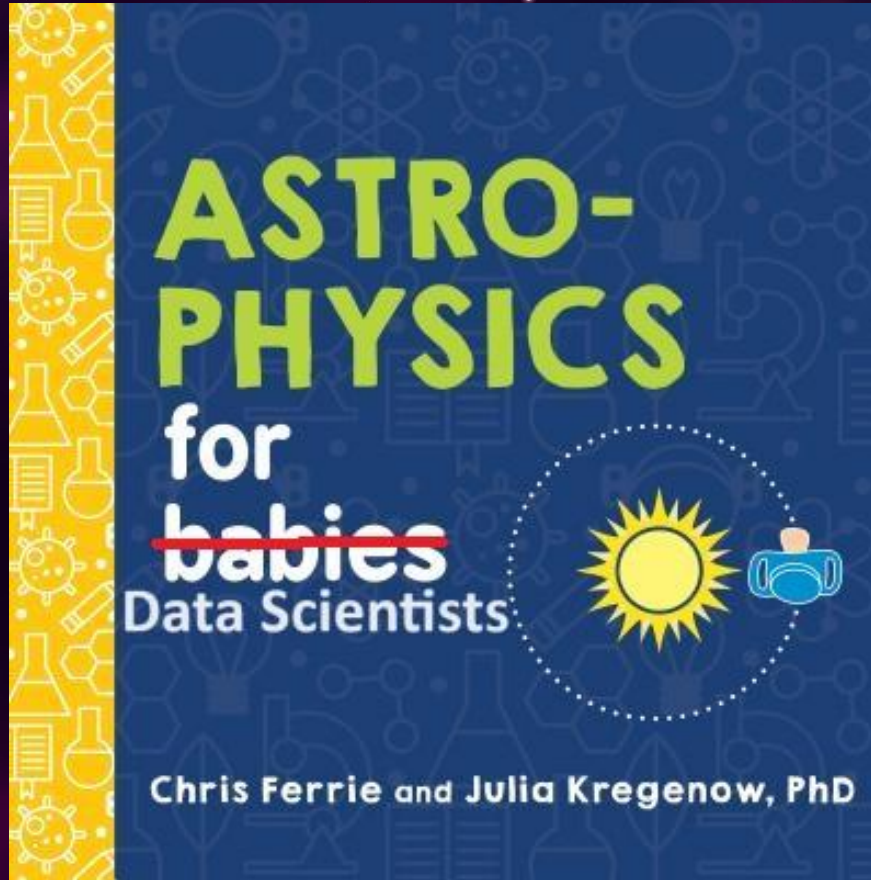# Stellar Object Classification

Sean Hulseman

Problem Statement

- The goal of this project is to create a model that accurately distinguishes between types of stellar objects from Sloan Digital Sky Survey (SDSS) data to help physicists more finely distinguish quasars, galaxies, and stars.
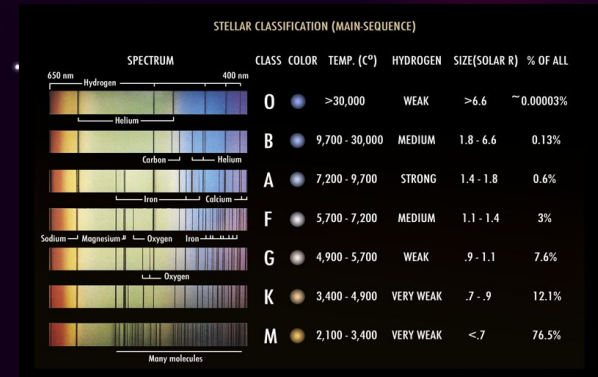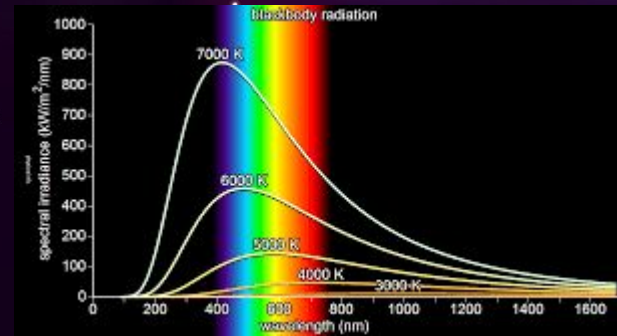
Spacetime, Stars, and Black Holes

Black Body Radiation

Stars - Multiple classes and different life cycles

**Galaxies** - A gravitationally bound system of millions or billions of stars

Quasars - Possibly the most powerful explosions in the universe, occur near supermassive black holes at the center of galaxies

https://en.wikipedia.org/wiki/Stellar_classification

blackbody radiation

spectral irradiance (kW/m²/nm)	wavelength (nm)
7000 K	
6000 K	
5000 K	
4000 K	
3000 K	

STELLAR CLASSIFICATION (MAIN-SEQUENCE)

SPECTRUM	CLASS	COLOR	TEMP. (C°)	HYDROGEN	SIZE(SOLAR R)	% OF ALL
650 nm — Hydrogen — 400 nm	O		>30,000	WEAK	>6.6	~0.00003%
Helium	B		9,700 - 30,000	MEDIUM	1.8 - 6.6	0.13%
Carbon — Helium	A		7,200 - 9,700	STRONG	1.4 - 1.8	0.6%
Iron — Calcium	F		5,700 - 7,200	MEDIUM	1.1 - 1.4	3%
Sodium — Magnesium — Oxygen — Iron	G		4,900 - 5,700	WEAK	.9 - 1.1	7.6%
Oxygen	K		3,400 - 4,900	VERY WEAK	.7 - .9	12.1%
Many molecules	M		2,100 - 3,400	VERY WEAK	<.7	76.5%

"Just a Phase" – Note From an Expert

"Astronomers believe that most, if not all, large galaxies went through a so-called "quasar phase" in their youth, soon after their formation. If so, they subsided in brightness when they ran out of matter to feed the accretion disk surrounding their supermassive black holes."

https://earthsky.org/astronomy-essentials/definition-what-is-a-quasar/
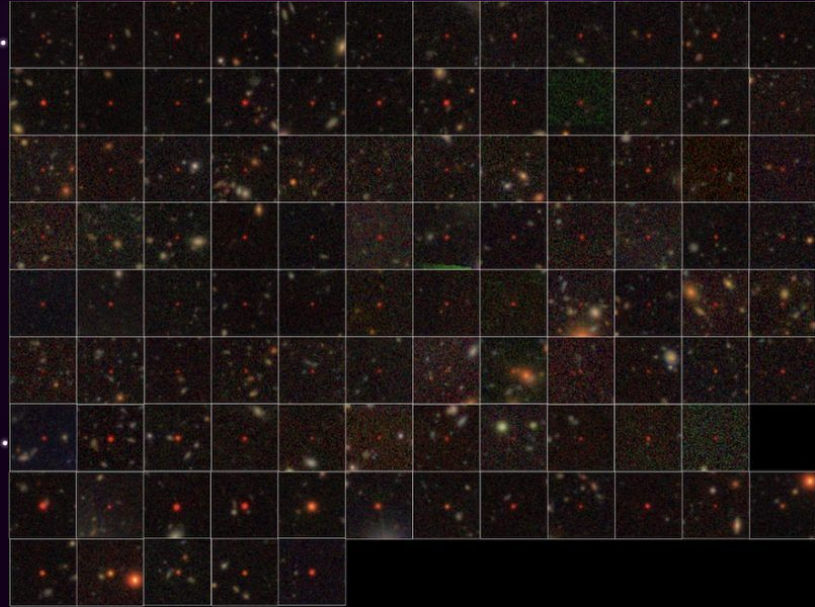
Here are 100 quasars identified via data from the Hyper Suprime-Cam mounted on the Subaru Telescope.
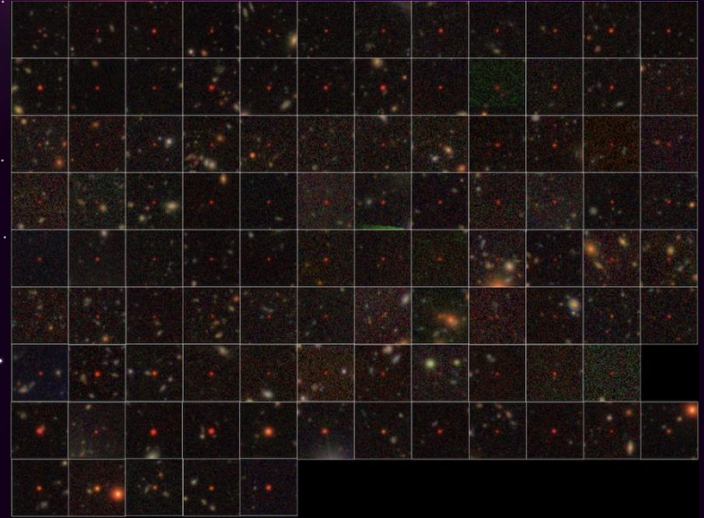
Fibers and Plates – Not Pictures

Taking 'pictures' is often impractical

Exposure time is prohibitively high

Captures more than the object needed

Spectral information is much easier to collect and has many uses!

Data

1st attempt: Kaggle Data – 100K labeled stellar objects

2nd attempt: SQL Query from SDSS– 500K labeled stellar objects

A model comparison – Kaggle vs SDSS

Features:
- 6 specific wavelengths near visible light
- Location of object
- Fiber/plate used to record the observation.

Score to beat: 98.4%
https://www.kaggle.com/code/beyzanks/stellar-classification-98-4-acc-100-auc

	Kaggle Data	SDSS SQL Query
Galaxy	**59,445**	**252,021**
Star	**21,594**	**193,382**
Quasar	**18,961**	**54,590**

Cleaning and EDA

Sensor errors dropped
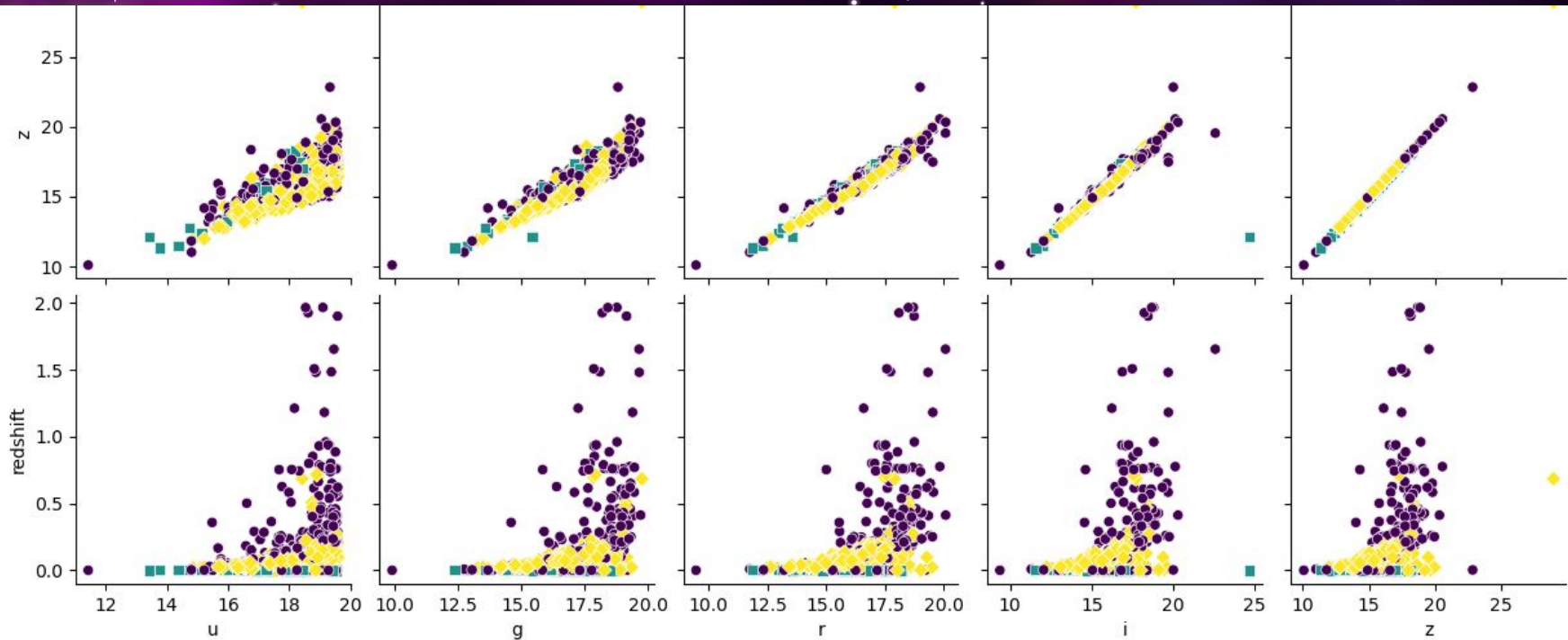
Mapped 'class' variable
    0: Galaxy, 1: Star, 2: Quasar

Imblearn LabelEncoder for categorical variables

Clusters seen visually? Redshift?
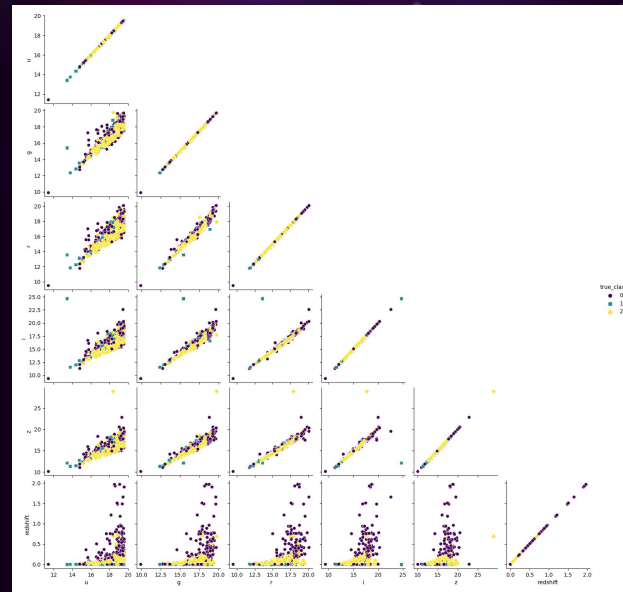
Clustering?

SMOTE – Synthetic Minority Over-sampling Technique

Synthetic samples are generated by interpolating between the selected instance and its k nearest neighbors. SMOTE involves creating new instances along the line segments connecting the chosen instance with its neighbors in the feature space.[1]

Class	Before	After
Galaxy	252,021	252,021
Star	193,382	252,021
Quasar	54,590	252,021

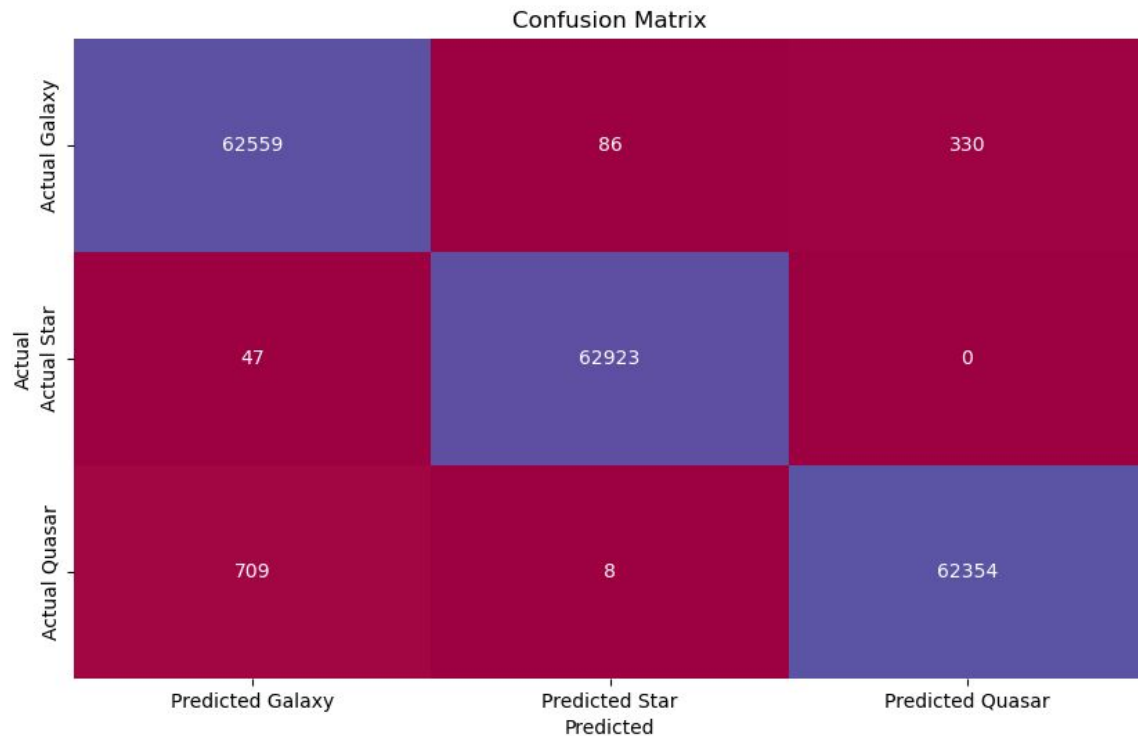[1] https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTE.html

Modeling

1. Grid Search Random Forest model to hyperparameter tune with first data set.
2. SQL Query for 500K data points instead of the meager 100K from kaggle
3. Train single random forest model with best hyperparameters found earlier
4. Analyze incorrect predictions and make inferences from them

Results
   Accuracy: 99.375%
   Score to beat: 98.4%

Analysis of **misclassified** objects

Confusion Matrix

	Predicted Galaxy	Predicted Star	Predicted Quasar
Actual Galaxy	62559	86	330
Actual Star	47	62923	0
Actual Quasar	709	8	62354

Predicted

What are Quasars Again?

Possibly the most violent phenomena in the universe
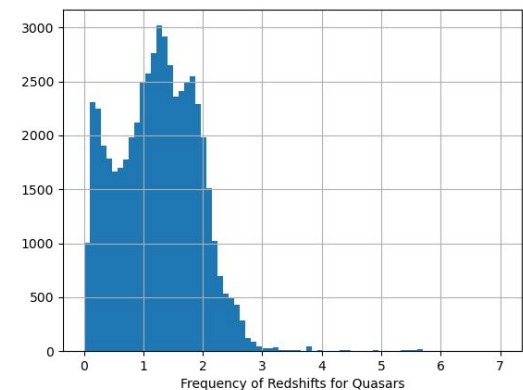
Supermassive black holes actively 'eating'.

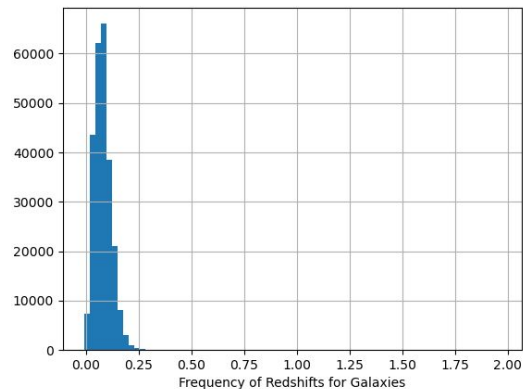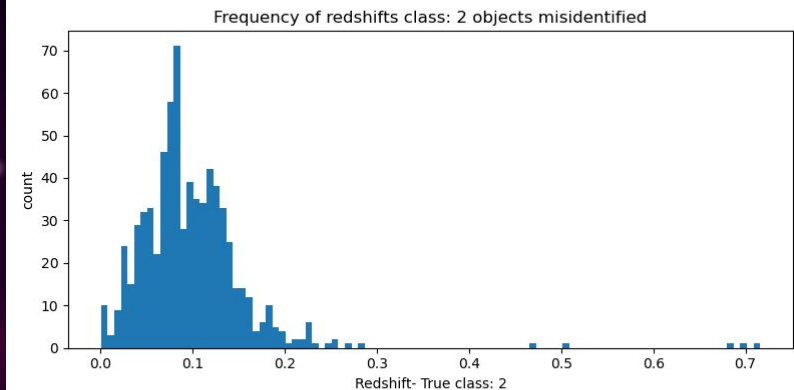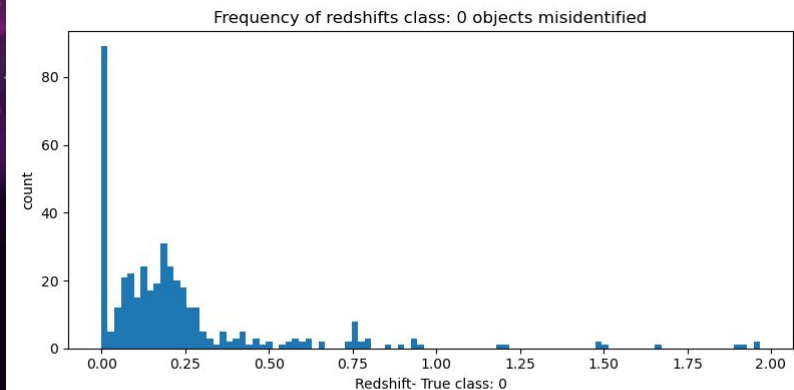Occur earlier in formation of galaxies

Something strange in my data

https://www.space.com/17262-quasar-definition.html

Analysis of Error – REDSHIFT ANOMALY

Analysis of Error – REDSHIFT ANOMALY

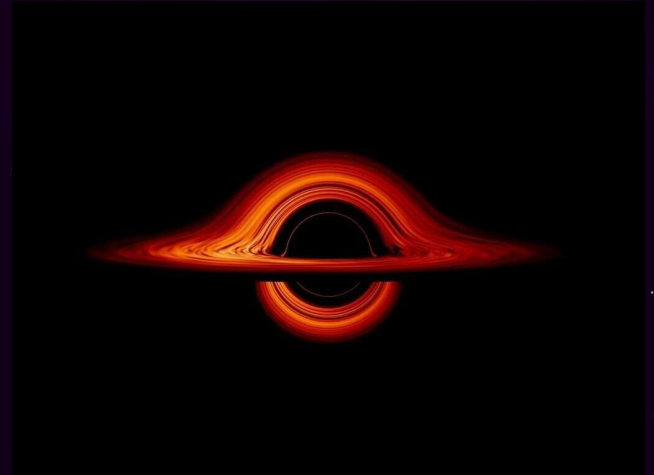|	Mean and Standard deviation of redshift OVERALL	Mean and Standard deviation of redshift for misclassified
Galaxy	Mean = 0.08 Standard dev.= 0.07	Mean = 0.09 Standard dev.= 0.06
Star	Mean = 0 Standard dev.= 0	Mean = 0 Standard dev.=
Quasar	Mean = 1.23 Standard Dev. =0.70	Mean =0.28 Standard dev.=0.311

Redshift Anomaly – Hypotheses

Doppler Effect?

Simultaneousness of emissions of light?

Some unknown physics of quasars?

Is the true distance to quasars further than the measured redshift indicated?

https://www.npr.org/2023/10/05/1203749479/black-hole-best-life-lessons
https://www.scirp.org/journal/paperinformation?paperid=108536

Conclusion

Predicted whether an object detected by the SDSS is a Galaxy, Star, or Quasar with 99.4% accuracy

Stellar objects misidentified provide evidence supporting the theories around the 'Quasar Red Shift Anomaly'

Further collaboration with astrophysicists could also use the positional data to compare the anomalous Quasars with their respective galactic neighborhoods

Thank you!

Slides Resources: slidesgo.com

Research
Youll, D. (2021) Anomalies in the Counts of Low Redshift Quasars. Journal of High Energy Physics, Gravitation and Cosmology, 7, 531-550. doi: 10.4236/jhepgc.2021.72030.

https://earthsky.org/astronomy-essentials/definition-what-is-a-quasar/

Software: Python, Scikit-Learn, Seaborn, Matplotlib, imblearn