The background is a dark space scene. In the lower half, the curved horizon of the Earth is visible, showing blue oceans and dark landmasses. Several large, dark, irregularly shaped asteroids are scattered throughout the scene, some in the foreground and others further away. In the upper left corner, there are thin white lines representing orbital paths, with a small purple dot marking a point on one of these orbits. The overall lighting is dim, with some light reflecting off the surfaces of the Earth and asteroids.

CLASSIFYING SPACE OBSERVATIONS

John Andrew Dixon



01

PROJECT OVERVIEW

The Data & The Stakeholders

THE DATA

My dataset contains 100,000 observations taken by the Sloan Digital Sky Survey telescope (SDSS).

A space observation is something the SDSS sees in space.

Source:

<https://www.kaggle.com/datasets/fedesoriano/stellar-classification-dataset-sdss17>



THE CLASSES



GALAXY

Lots of stars, dust, and gas grouped together by gravity.



QUASAR

A bright, far away space object that's pretty much a young galaxy or black hole.



STAR

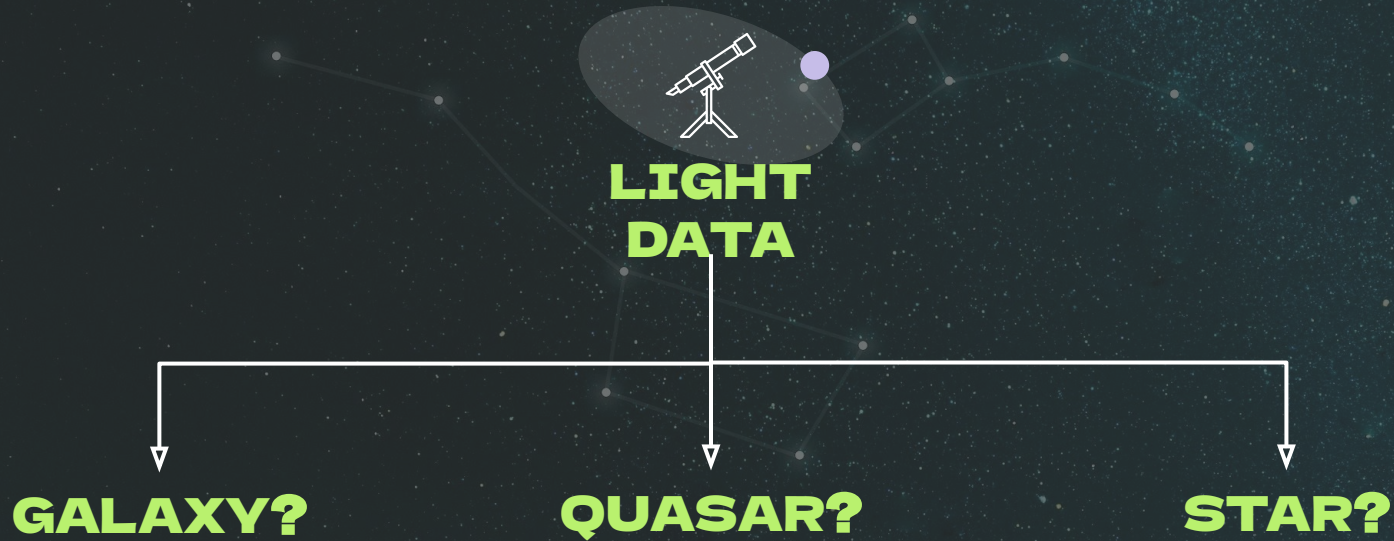
Giant ball of bright, burning gas; it's self-sustaining.

THE STAKE HOLDERS

The stakeholders are a small astronomy company performing work like the SDSS team but without the proper means to verify the nature of their observation. They only have the ability to gather light data but not the ability to classify it. So, they need to rely on machine learning as a means of classification.



THE PROBLEM



How well can my stakeholders classify space observations based on their incoming light?



02

VISUAL
INSIGHTS

What The Data Show

CLASS BALANCE



59%

A donut chart with a dark grey background and a light green outline. The chart is mostly filled with a light green color, representing 59% of the total. The percentage '59%' is written in white in the center.

GALAXY

About 59.4% of the data are galaxies.



19%

A donut chart with a dark grey background and a light green outline. The chart is mostly empty, with only a small segment filled with light green, representing 19% of the total. The percentage '19%' is written in white in the center.

QUASAR

About 19% of the data are quasars



22%

A donut chart with a dark grey background and a light green outline. The chart is mostly empty, with only a small segment filled with light green, representing 22% of the total. The percentage '22%' is written in white in the center.

STAR

About 21.6% of the data are quasars

INCOMING INFRARED LIGHT



A stellar observation that has a higher than average amount of infrared light is more likely a quasar. Stars and galaxies, on average, have less infrared light.

INCOMING RED SHIFT



A stellar observation that has a higher than average red shift in its light is highly likely a quasar. Stars, on average, don't have much observable red shift while galaxies do, but not as much as quasars.



03

MODEL METRICS

How Good Is The Model?

BRIEF INTRO



FALSE POSITIVES AND NEGATIVES

Aside from slight embarrassment, false positives and false negatives for any class won't affect the stakeholders much. Money won't be lost and people won't die. Because of this, the most important metric is accuracy.

The background is a dark, starry space scene. In the top left, a portion of a planet with a ring system is visible. In the top right, a thin white orbital line curves around a small purple planet. The bottom left features faint, overlapping white orbital lines. The central text is contained within a dark gray rectangular box.

TUNED XGBOOST

THE METRICS

	GALAXY	QUASAR	STAR
ACCURACY	98%		
PRECISION	97.8%	96.8%	99.7%
RECALL	98.9%	93.1%	100%

Higher precision = less false positives (i.e. non-X classes are classified as X class)

Higher recall = less false negatives (i.e. X classes aren't classified as X class)

THE GOOD, THE BAD

STRENGTHS

- High accuracy
- Great at classifying all three classes

WEAKNESSES

- Slower training time
- May be *slightly* weak at classifying if something is or isn't a quasar, possibly due to it being trained on an unbalanced dataset.

The background of the slide is a deep space scene. A large, curved horizon of the Earth is visible in the upper half, showing a blue and white atmosphere. Several dark, rocky asteroids are scattered in the foreground and midground. The sky is a dark, starry void.

04

FINAL SUGGESTIONS

A Path For The Stakeholders



WHAT TO DO?



IMBALANCED DATA



MORE DATA

Try to find a more balanced dataset. It doesn't need to be perfectly balanced but should be more balanced than the current data. Then, retrain the model on this data.

BALANCE DATA

If balanced data can't be found, use class balancing techniques to balance the current data and then retrain the model on this new balanced data.



THANKS!

Do you have any questions?

CREDITS: This presentation template was created by **Slidesgo**, and includes icons by **Flaticon**, and infographics & images by **Freepik**