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

LIGHT DATA

GALAXY? **QUASAR?** **STAR?**

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

02

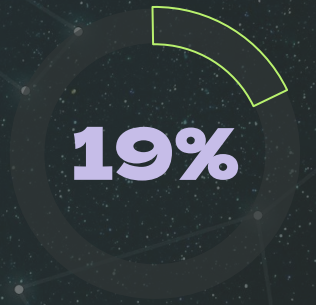
VISUAL INSIGHTS

What The Data Show

CLASS BALANCE

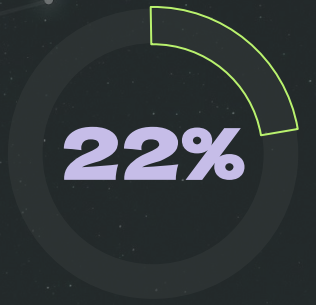
GALAXY

About 59.4% of the data are galaxies.



QUASAR

About 19% of the data are quasars



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.

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-class X are classified as class X)

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

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.

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?