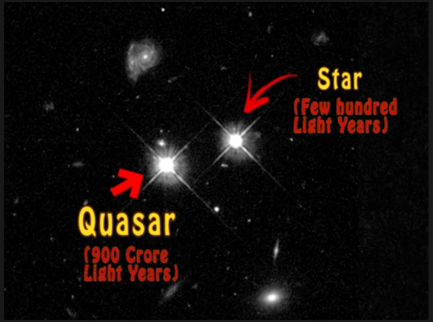
Stellar Classification Dataset SDSS17

Machine Learning Operations (AAI-540-02)
Alyona Kosobokova, Ian Rebmann, Group 7

University of San Diego®

SCIENTIFIC BACKGROUND

Modern sky surveys collect millions of detections, far more than astronomers can classify by hand, so automated star/galaxy/quasar labeling is required to turn raw observations into usable science catalogs. This project builds a production-oriented ML classifier for SDSS DR17-style tabular photometry, sky position, and redshift to support scalable analysis, consistent target selection, and faster follow-up decisions.

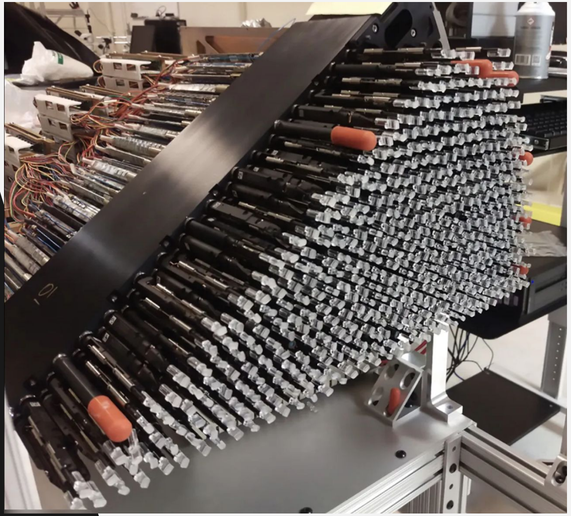


Dark Energy Spectroscopic Instrument (DESI)

Mayall 4-meter telescope in southern Arizona that measures millions of galaxy and quasar spectra to map the 3D structure of the universe.

The instrument is built around a very wide field of view. Light from the sky is focused onto a focal surface that contains roughly 5,000 robotic fiber positioners.

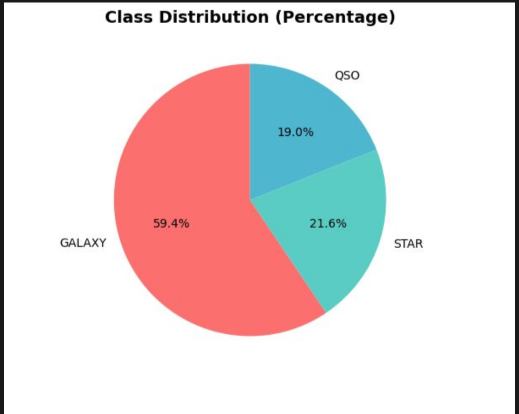
Those fibers send light through long fiber bundles to ten spectrographs. The spectrographs split the light into multiple channels and record wavelengths from roughly 360 to 980 nanometers, covering blue through near-infrared light.

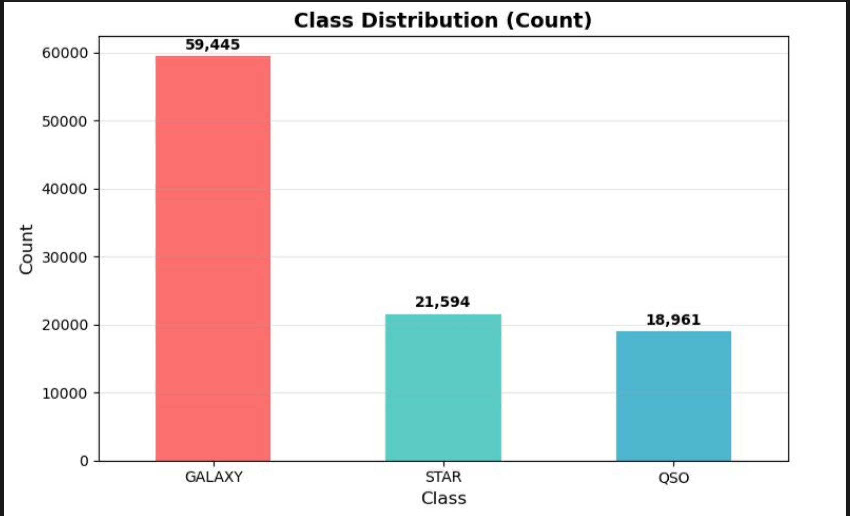


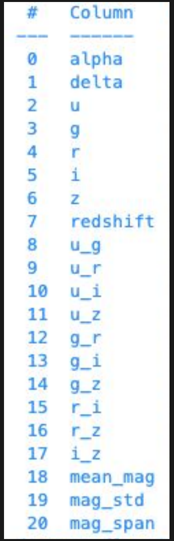
DATA SOURCE AND UNDERSTANDING

Data source contains ~100,000 observations with 17 features and one target class (star, galaxy, quasar).

Features used include sky position, photometric magnitudes (u, g, r, i, z), redshift, and engineered color-index features derived from magnitudes.







```
#    Column
---  ------
0    alpha
1    delta
2    u
3    g
4    r
5    i
6    z
7    redshift
8    u_g
9    u_r
10   u_i
11   u_z
12   g_r
13   g_i
14   g_z
15   r_i
16   r_z
17   i_z
18   mean_mag
19   mag_std
20   mag_span
```

MLOps Architecture

The pipeline loads the SDSS17 dataset into Amazon S3, models and evaluation artifacts are stored in S3.

SageMaker real-time endpoint and CloudWatch metrics/alarms for latency and errors.

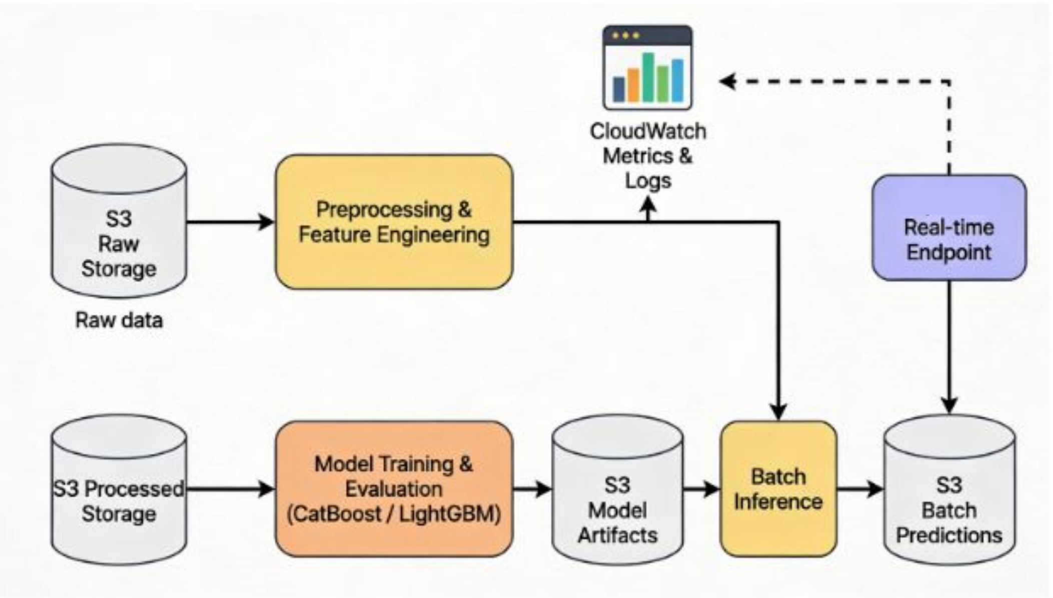


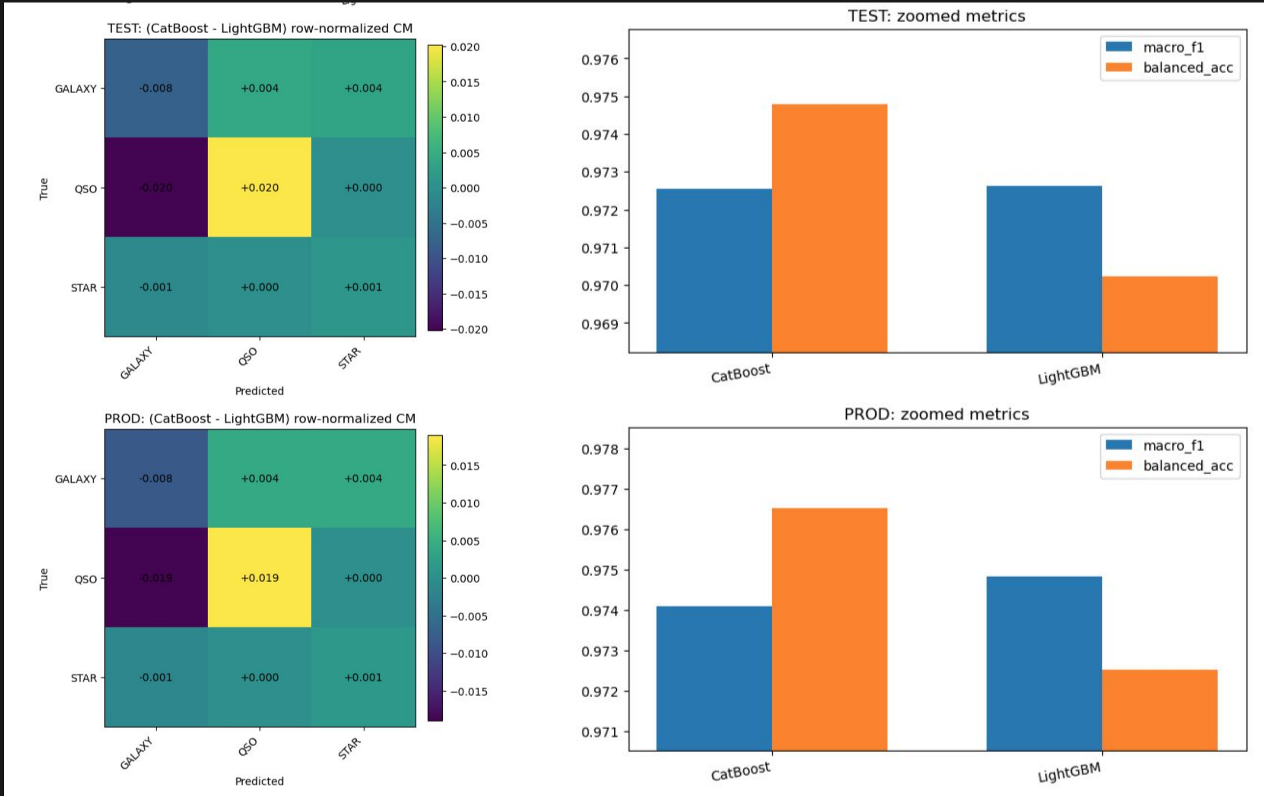
Figure 1: Architecture diagram

Model Comparison

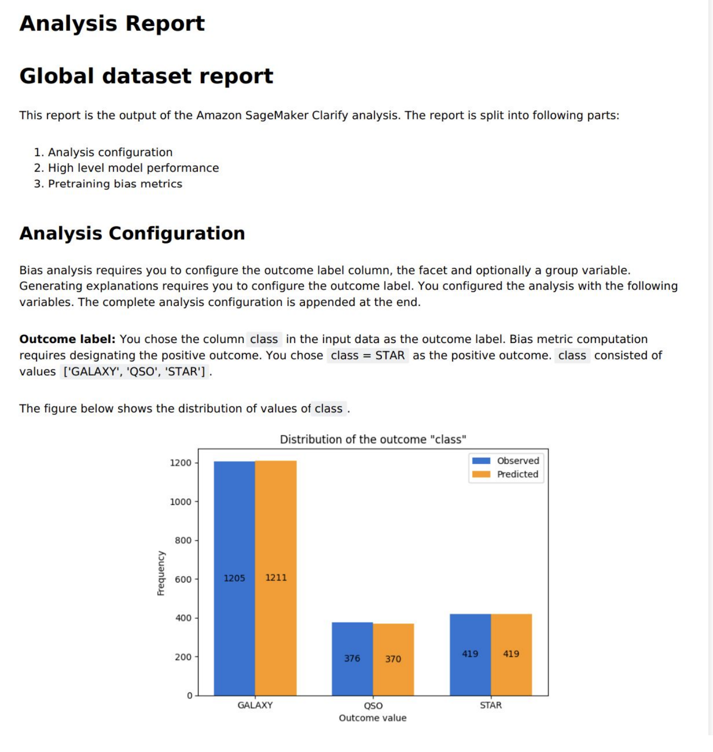
CatBoost
macro_f1: 0.974078
balanced_acc: 0.976524

LightGBM
macro_f1: 0.974829
balanced_acc: 0.972521

Deployed SageMaker built-in XGBoost
macro_f1: 0.975357
balanced_acc: 0.971641



CloudWatch Metrics



Analysis Report

Global dataset report

This report is the output of the Amazon SageMaker Clarify analysis. The report is split into following parts:

1. Analysis configuration
2. High level model performance
3. Pretraining bias metrics

Analysis Configuration

Bias analysis requires you to configure the outcome label column, the facet and optionally a group variable. Generating explanations requires you to configure the outcome label. You configured the analysis with the following variables. The complete analysis configuration is appended at the end.

Outcome label: You chose the column `class` in the input data as the outcome label. Bias metric computation requires designating the positive outcome. You chose `class = STAR` as the positive outcome. `class` consisted of values `['GALAXY', 'QSO', 'STAR']` .

The figure below shows the distribution of values of `class` .

Potential Enhancements

- Upgrading the deployment with automated rollback based on error rate.
- Input contract + schema versioning.
- Uncertainty and “needs review” flags for scientists, providing a “low-confidence” or flag, so astronomers can prioritize manual inspection.
- Explainability for scientific method: adding per-prediction explanations (like which colors/magnitudes drove the decision).

CONCLUSION

- Delivered a production-oriented ML pipeline that classifies SDSS DR17 objects into star, galaxy, or quasar from tabular photometric features, sky position, and redshift.

- Set up monitoring schedule and real time endpoint, alongside batch processing to classify objects based on their photometric features.

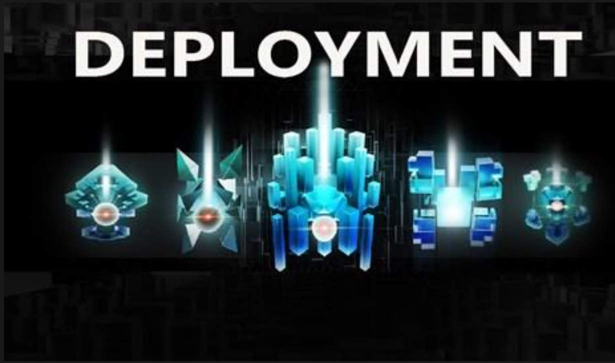
FUTURE DIRECTIONS

Research and Analysis Tool

Provide star/galaxy/quasar predictions as a service for astronomers, with support for custom, reproducible studies such as population statistics by redshift or magnitude bins, quasar-candidate selection, and follow-up prioritization based on model confidence.

Real-Time Data Processing and Object Detection

Integrate into survey/observatory pipelines to classify incoming detections within minutes and flag unusual cases for review (e.g., low-confidence predictions, out-of-distribution photometric/redshift, or abrupt shifts in predicted class proportions), with configurable thresholds and downstream alert outputs.



Thank you!