The Hunt for Exoplanets

Vetting Kepler Light Curves

Praveen Gowtham

Motivations

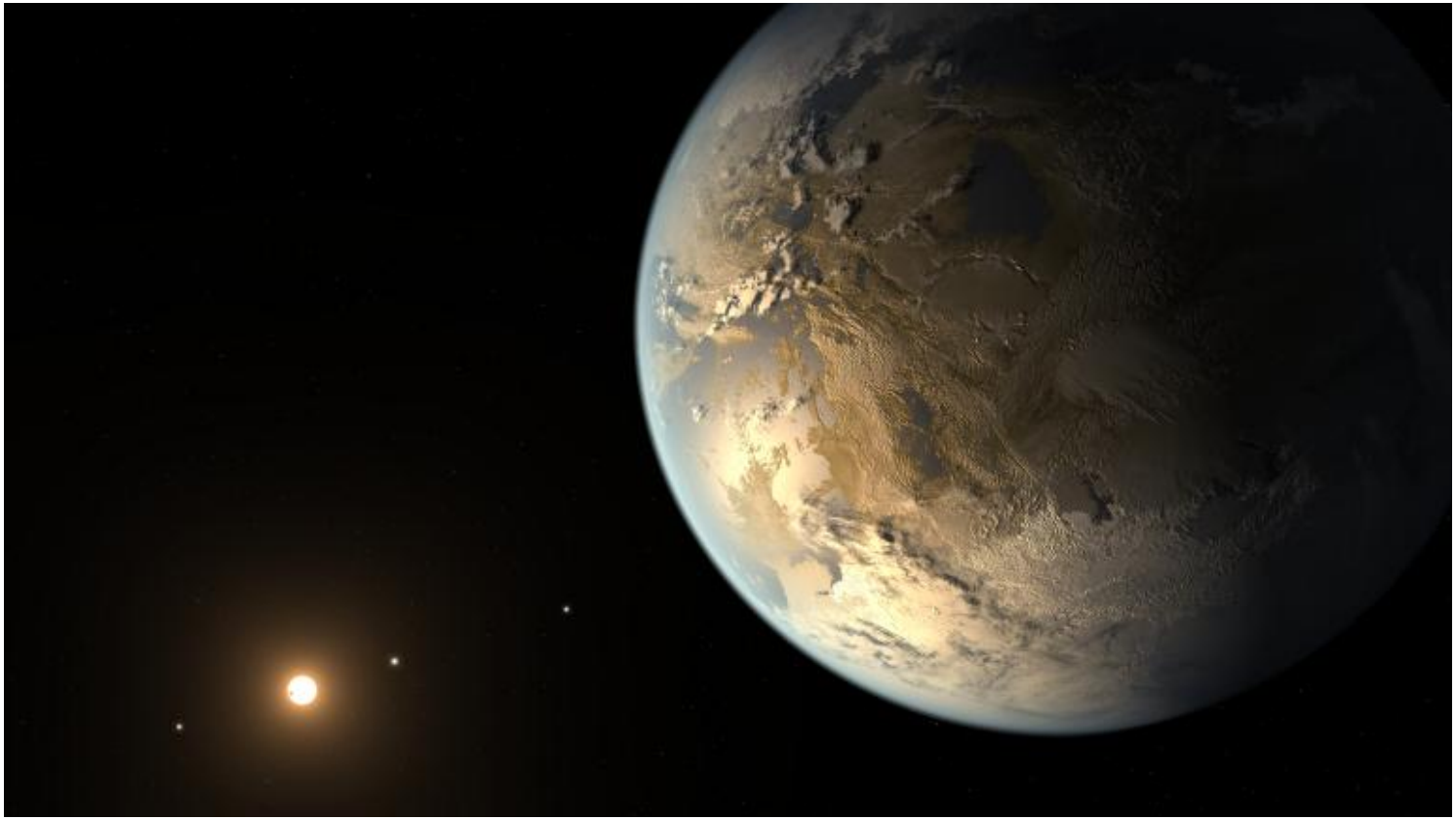


Figure: An artist's concept of Kepler-186f, the first validated Earth-size planet to orbit a distant star in the "habitable zone."

Motivations

- Search for habitable planets.
- The types and distribution of planets and planetary systems in Milky Way.
- Where does our solar system fit in?

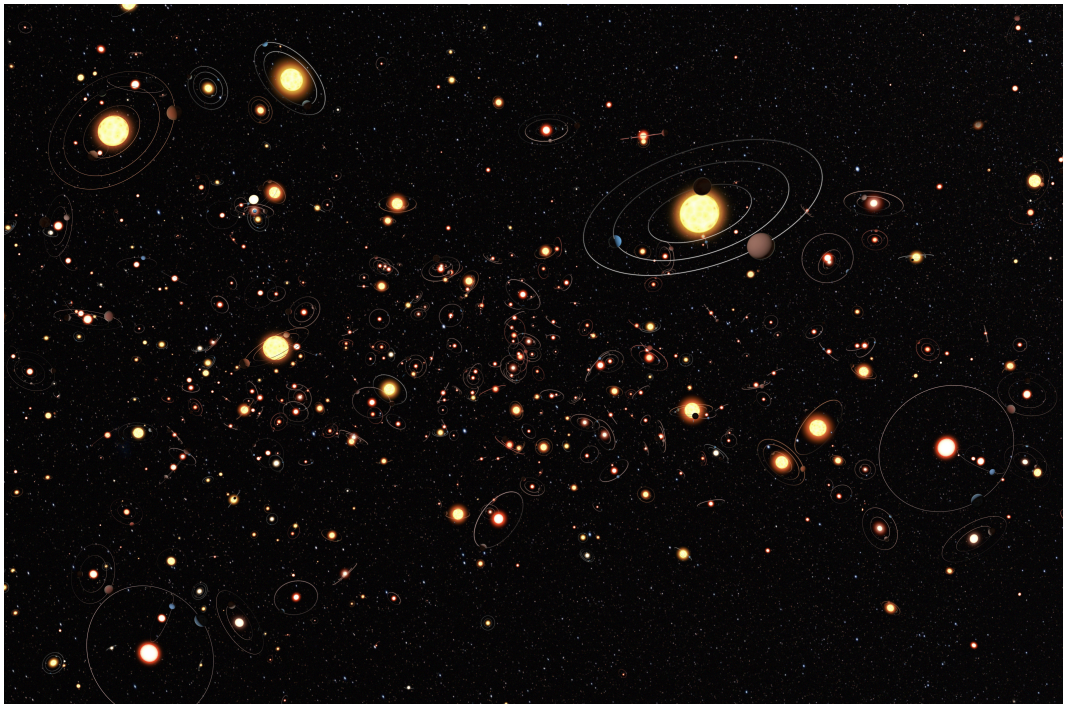


Figure: Artist depiction of planetary system distribution in sector of Milky Way galaxy based off of exoplanetary data.

Detection technique: Light curve transit crossing

Light flux time series from star. Periodic dips in light flux could be transiting exoplanet.

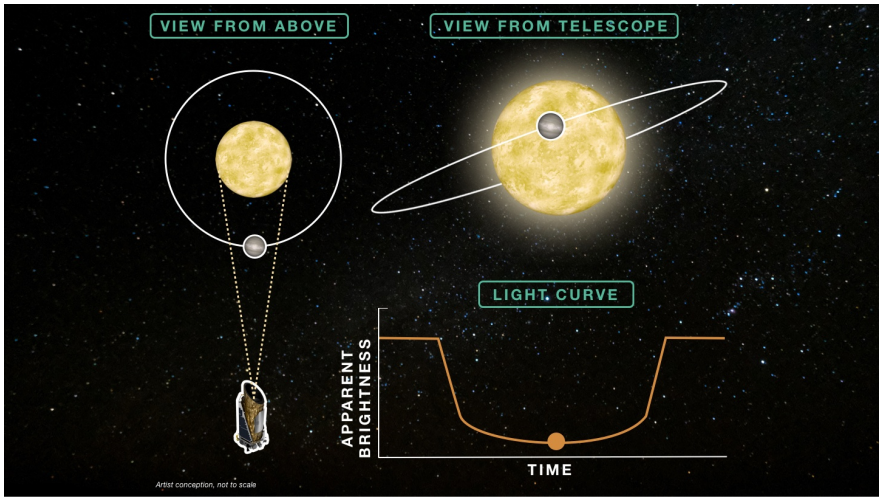


Figure: Schematic of transit detection via space telescope photometry.

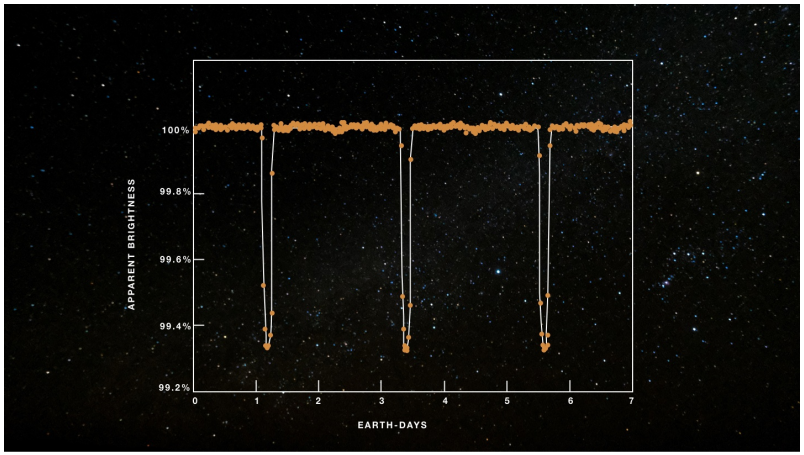


Figure: Light curve with periodic transits from exoplanet HAT-P-7b.

Kepler mission (2009-2018)

- Kepler space telescope: dedicated photometry scanning a section of Milky Way.
- Photometry on $>$ 500,000 stars.
- $\approx$ 2,500 confirmed exoplanets by Kepler via light curve transit method.

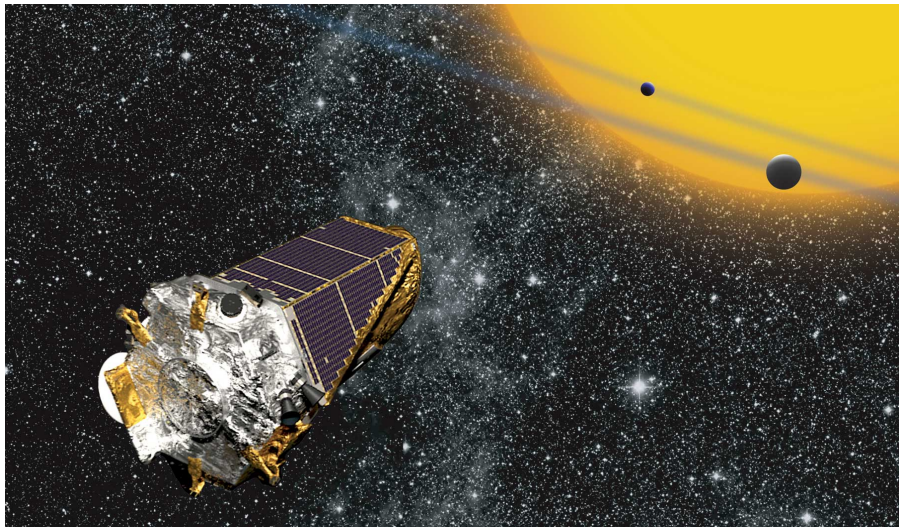
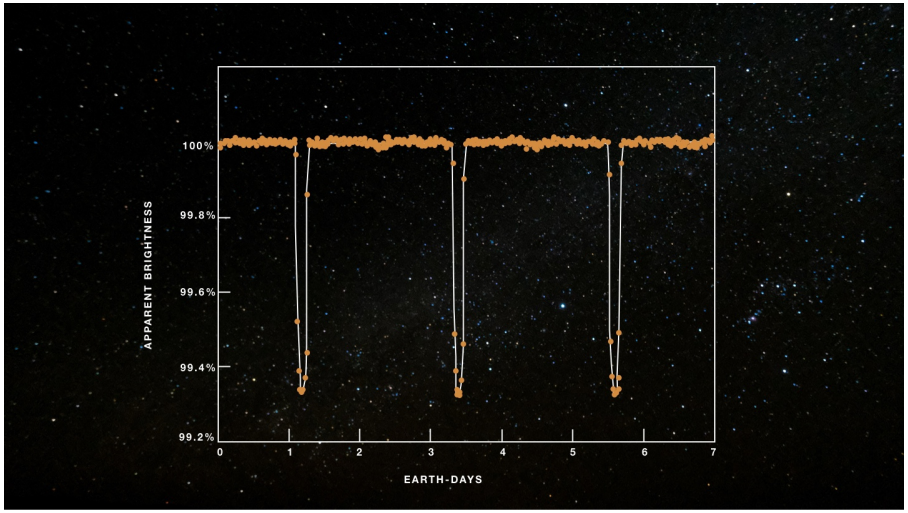


Figure: Kepler Space Telescope.

The problem

- Transit or noise? Some processing/statistics required.



The problem

- Transit or noise? Some processing/statistics required.
- False positives (FPs).

The problem

- Transit or noise? Some processing/statistics required.
- False positives (FPs).
  - Secondary eclipse FPs

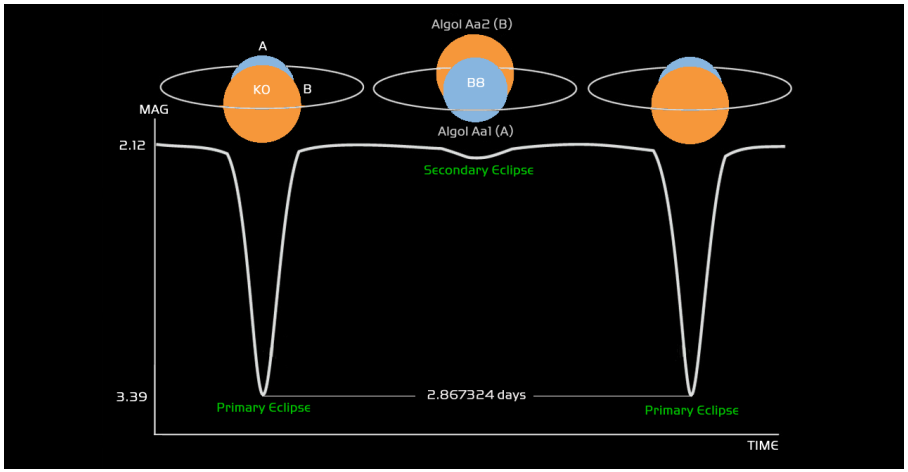


Figure: Light curve from eclipsing binary star system (Algol A/B System)

The problem

- Transit or noise? Some processing/statistics required.
- False positives (FPs).
  - Secondary eclipse FPs
  - Non-transiting Phenomena (NTP FPs)

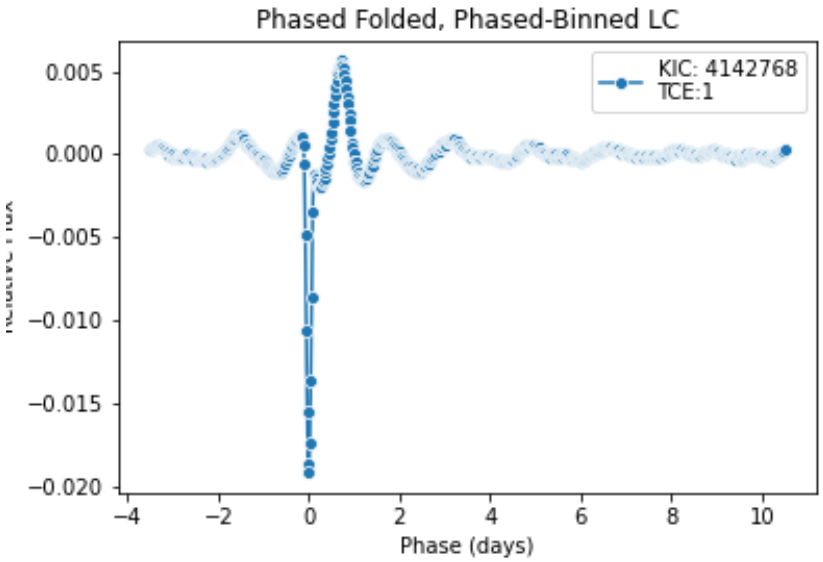


Figure: Single oscillation cycle from pulsating star.

The problem

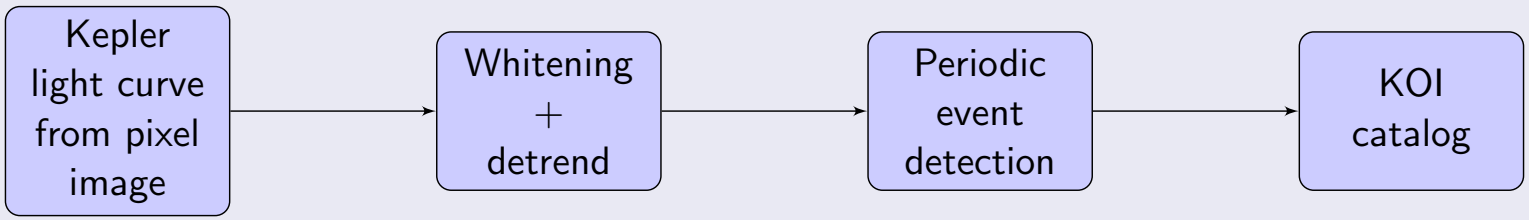
- Transit or noise? Some processing/statistics required.
- False positives (FPs).
  - Secondary eclipse FPs
  - Non-transiting Phenomena (NTP FPs)
- Goal: Real exoplanets vs. different types of FPs

Kepler Data Validation Stream

Kepler mission uses an automated preprocessing and vetting pipeline.

- Detect potentials: periodic statistically significant dip events. These are **Kepler Objects of Interest (KOIs)**.
- Basic preprocessing of KOI light curves.

Kepler pipeline



Focus on classification of KOIs

Problem specification

Label encoding + data breakdown

Class name	Confirmed Planet	Secondary Eclipse FP	NTP FP
Target Label	1	2	3
Count	2333	2140	880

Multi-label classification problem from light curves of KOIs.

- **Key challenge**: Light curve $\rightarrow$ processed light curve $\rightarrow$ relevant features
- Train classifier on constructed features/label mapping

Light curve processing: phase-folding + averaging

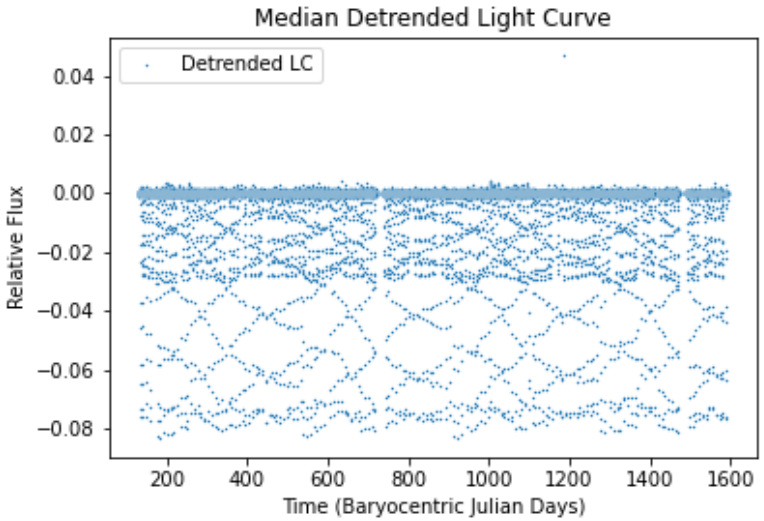


Figure: Whitened, detrended light curve.

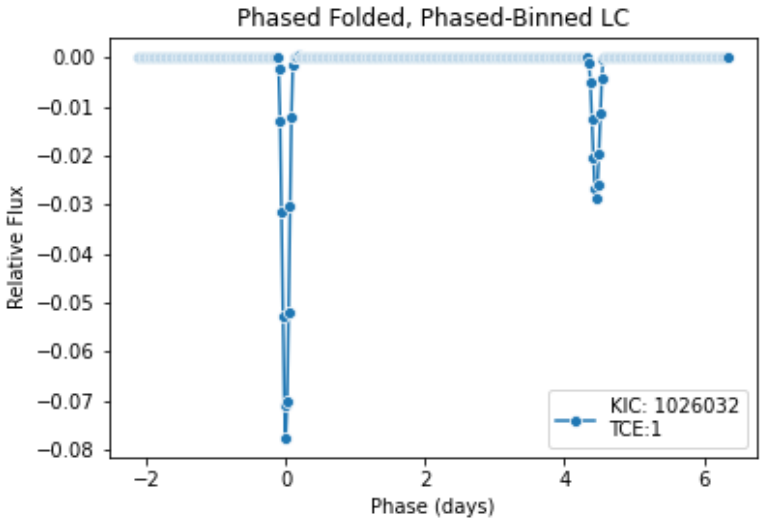


Figure: Phase-folded and bin averaged (secondary eclipse FP).

Constructed Features

All features (except period) constructed from phase-folded, bin-averaged light curves.

Period	TCE period (days).
Duration	TCE duration (hours).
EOS	Even-odd statistic for secondary eclipse detection.
WSS	Weak secondary statistic
min	Almost always corresponds to the depth of the primary transit.
max	Maximum value in phase-folded/bin-averaged light curve.
LCBIN_0 - LCBIN_140	141 points of the xy-normalized primary transit close-ups

Weak Secondary Statistic (WSS)

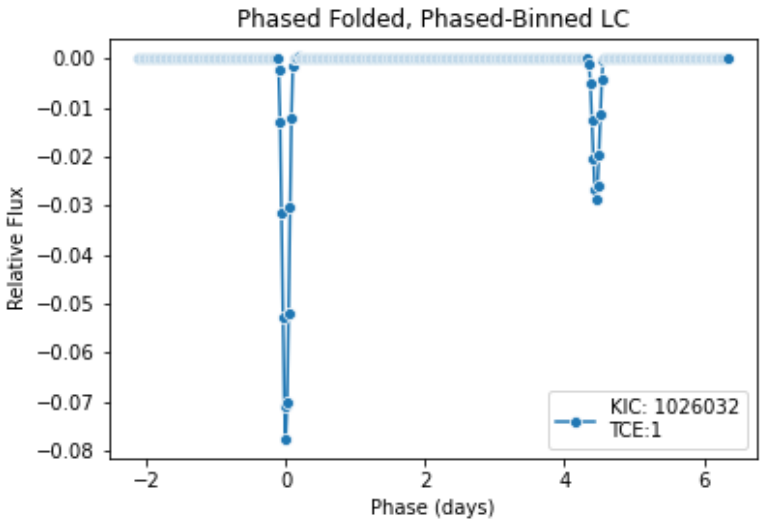


Figure: Whitened, detrended light curve.

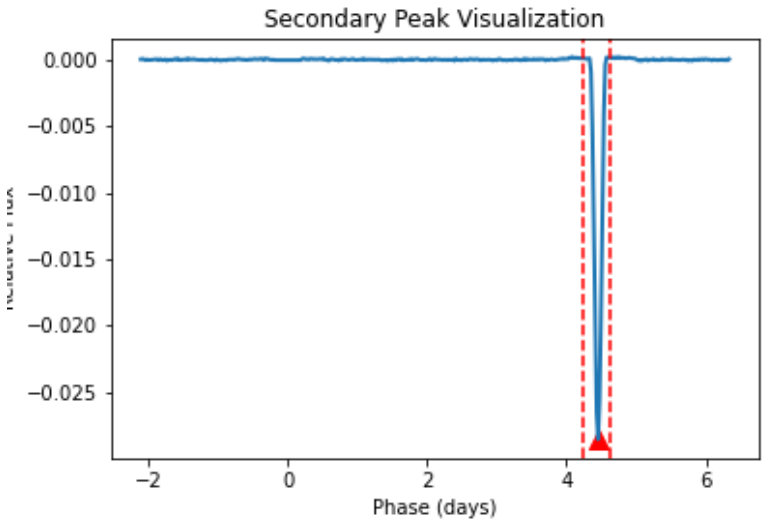


Figure: Primary transit subtracted.

WSS Test

- Subtract primary transit.
- Peak find and extract max amplitude peak (dip) + supports.
- Probability that amplitude (or more extreme) generated by Gaussian noise floor.

Even Odd Statistic (EOS)

Sometimes primary + secondary: similar amplitudes. Secondary at exactly half period.

Problem: primary and secondary registered as same event by Kepler data validation.

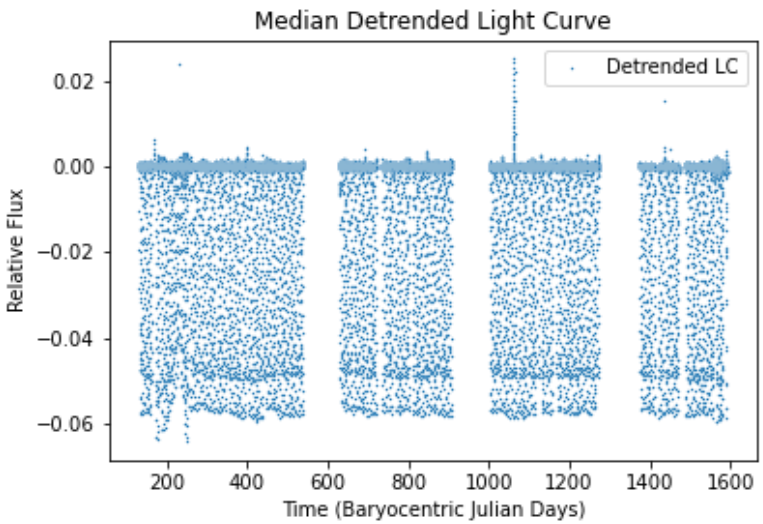


Figure: Primary and secondary transit amplitudes similar.

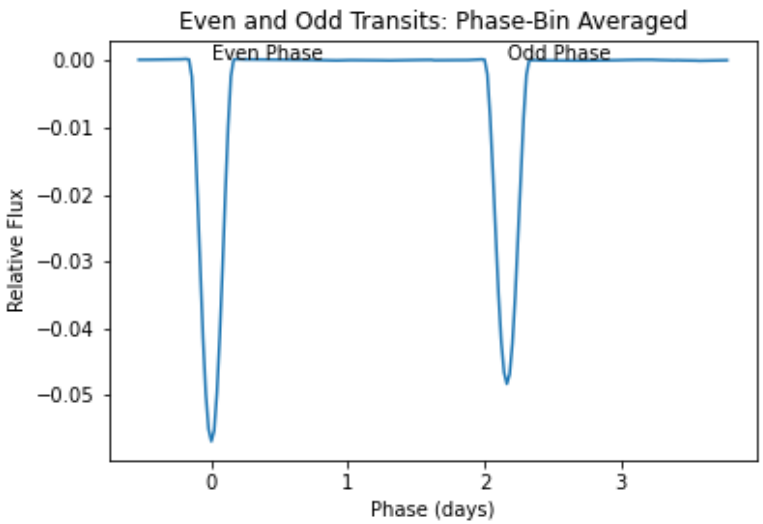


Figure: Alternate cycles binned into separate groups (even/odd). Each phase-folded, averaged.

Even Odd Statistic (EOS)

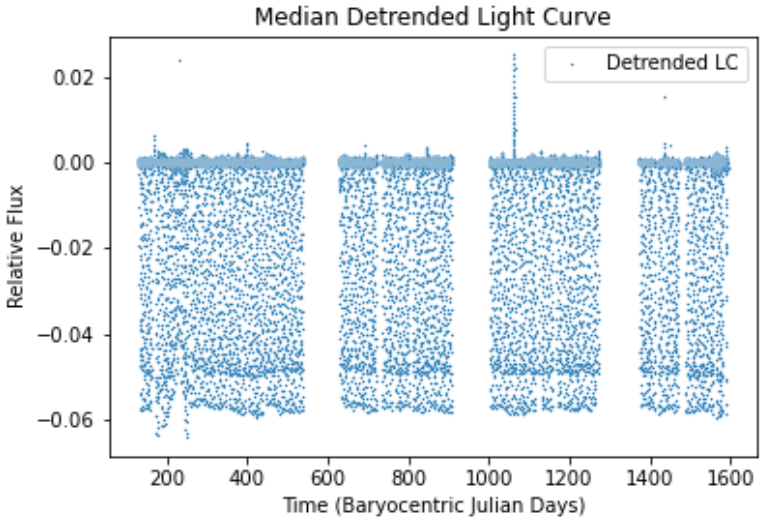


Figure: Primary and secondary transit amplitudes similar.

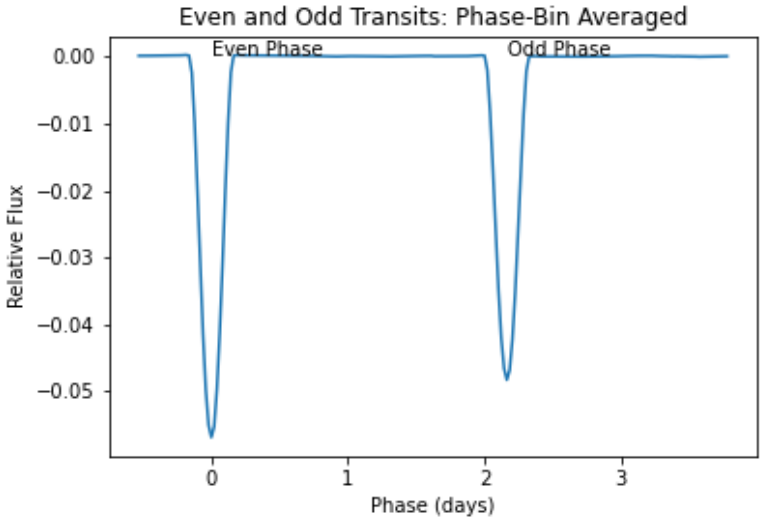


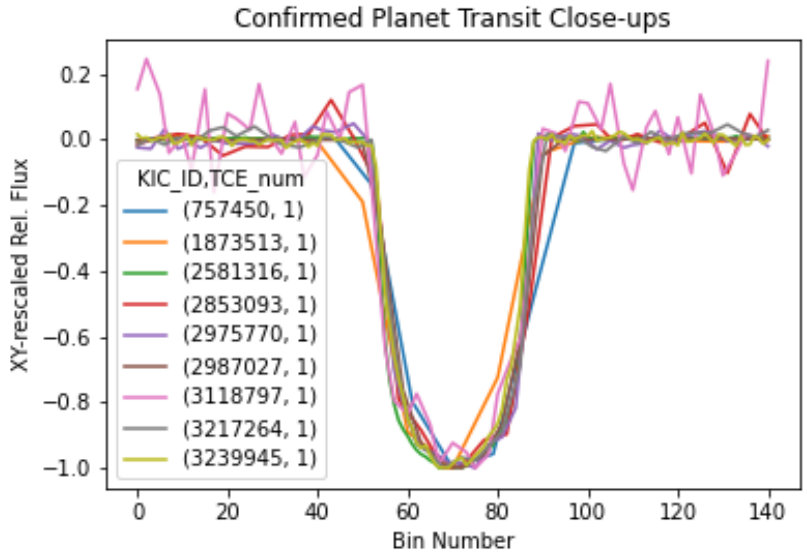
Figure: Alternate cycles binned into separate groups (even/odd). Each phase-folded, averaged.

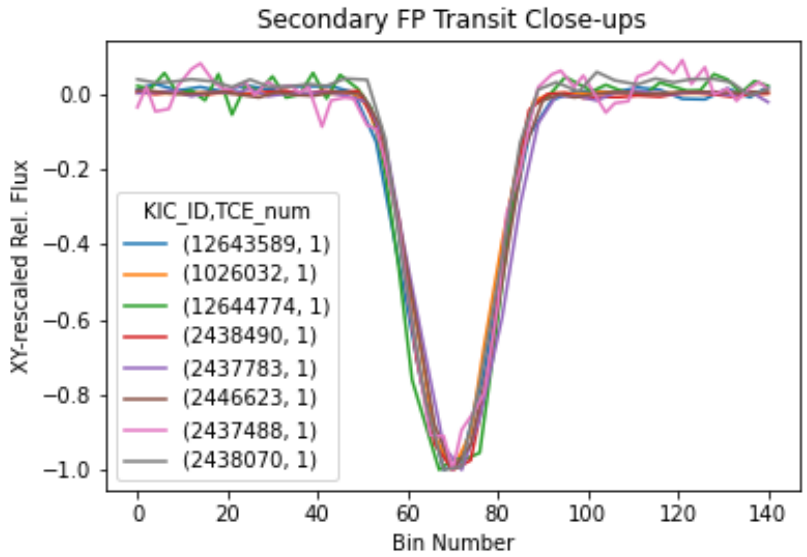
EOS Test

- Two sample t-test on even vs. odd phase transit depths.
- Extract p-value for significance of mean difference.

LCBIN Features

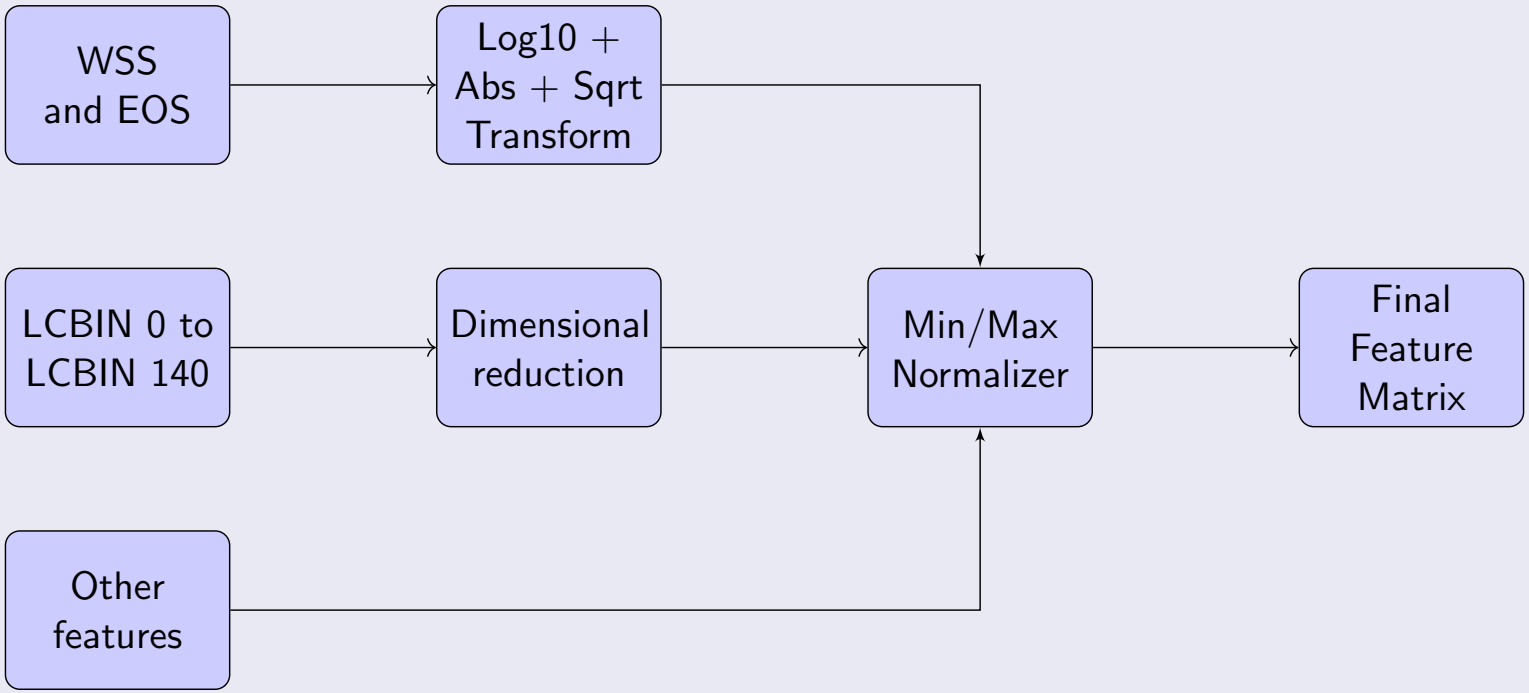
- Primary transit close-up: ± 2 transit durations.
- XY-rescaled, resampled / binned to 141 points fixed length.
- **Shape difference:** U-shape vs. V-shape



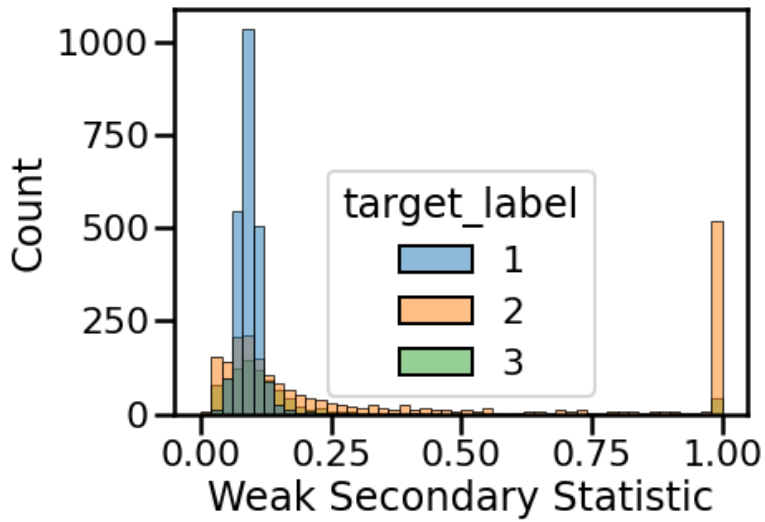


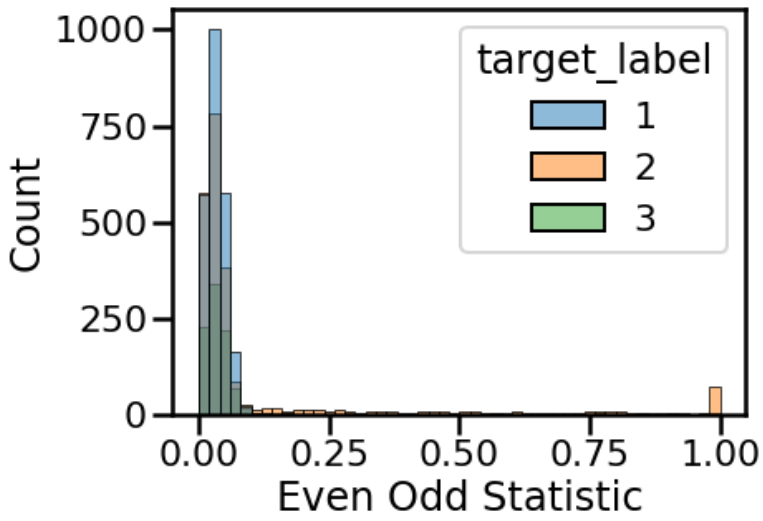
Feature Preprocessing Pipeline

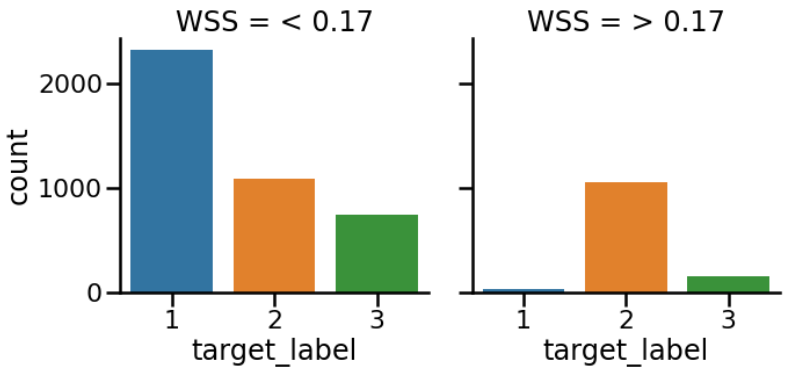
Scikit-learn custom transformers + pipeline integration

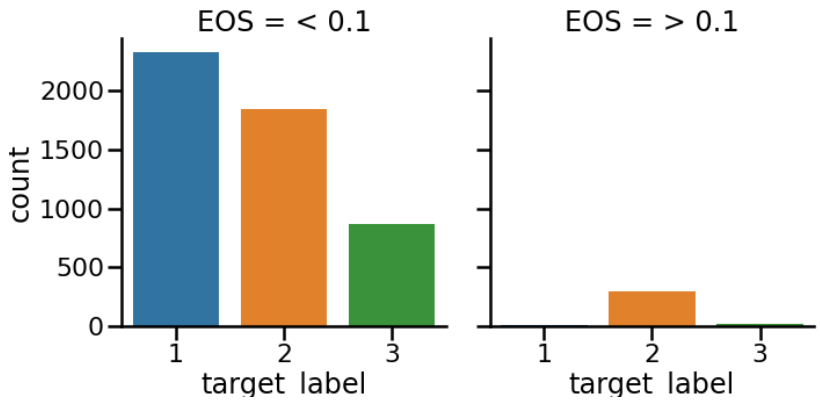


Statistical EDA: Secondary Eclipse Test Features



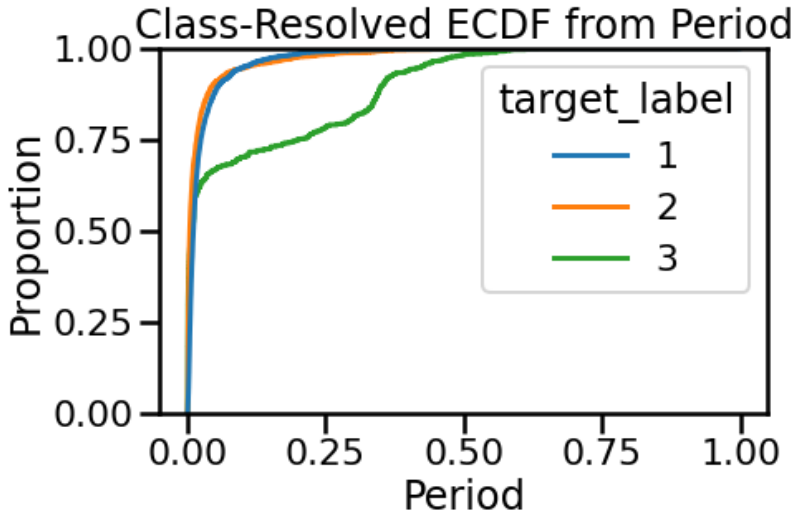


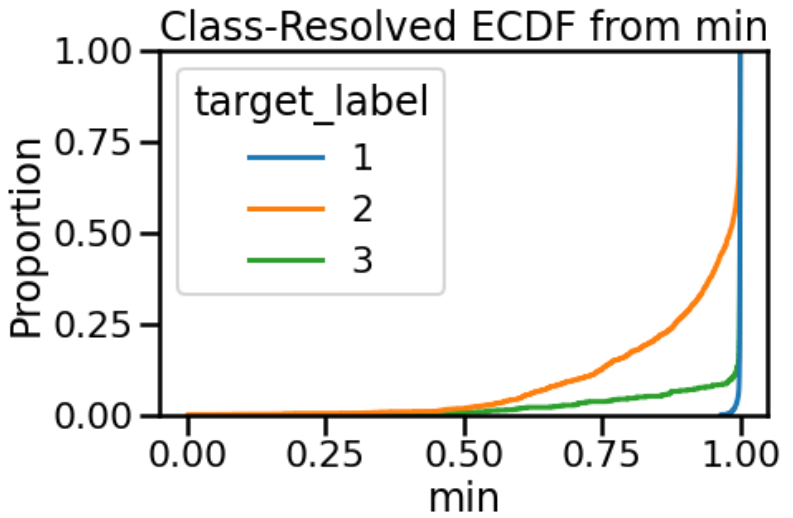




Some other non-LCBIN features

- 'Period' good for class 3 (NTPs) selection.
- 'Min' feature / transit depth: exoplanets vs class 2 (secondary FPs)
- 'Duration', 'max': helps exoplanet vs. FP separation

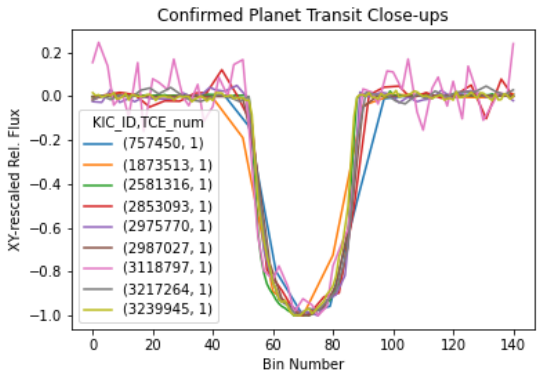


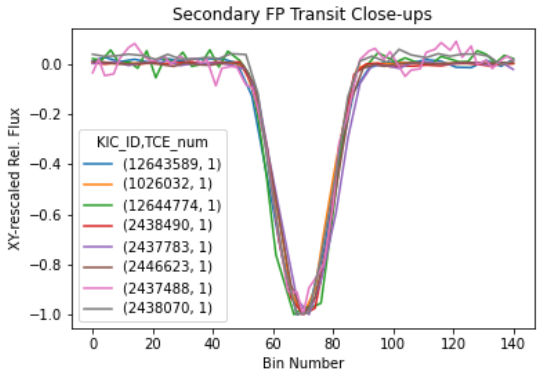


Locality Preserving Projections (LPP)

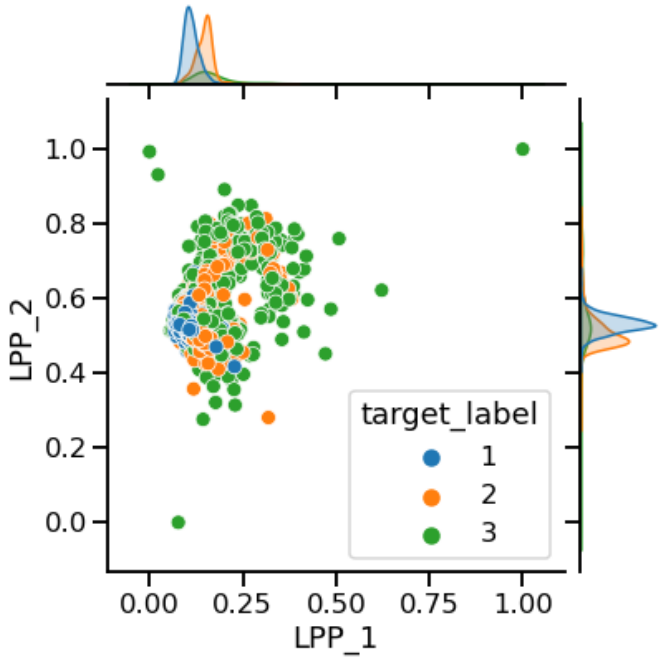
LPP: Dimensional reduction technique

Map to low dimension maintains closeness of objects that are close in high dimensions. Reduce 141-D LCBIN to 2D space.





LPP



Class separation in LPP space

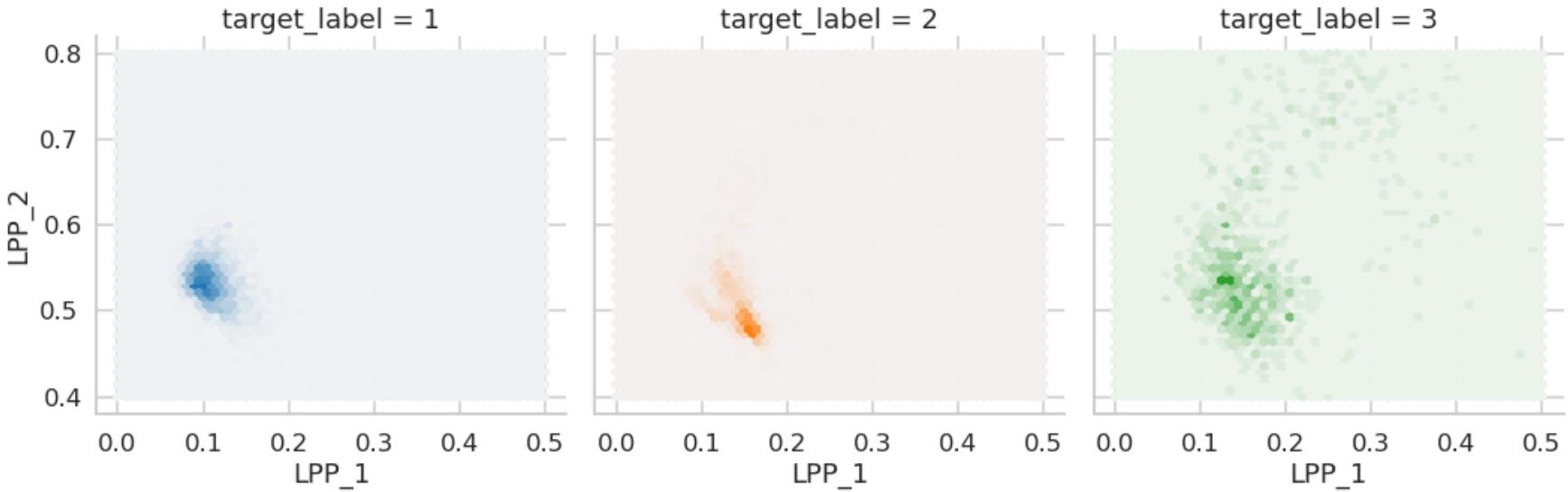


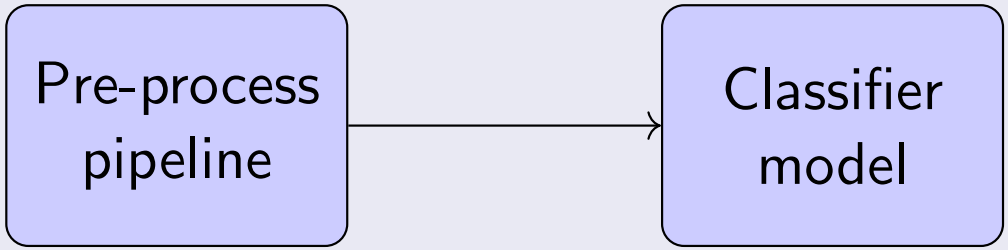
Figure: Hexbin count distribution for each class in 2D LPP space.

Distinction between class 1 (exoplanets) and FPs. Outliers dominated by NTP FPs.

Modeling Pipeline

Hyperparameter tuning: CV grid search

- 75-15 train/hold-out set.
- Scikit-learn full pipeline:



- 5-fold cross validation: 75-15 splits on train set.
- Eval. metric: f1-score micro-averaged across classes.
- Pipeline + CV fit: avoids data leakage.

Classifier Models + Hyperparameters

Random Forest: max_depth, max_features, n_estimators
RBF-Kernelized Soft-Margin SVM: C, gamma
XGBoost (stumps): learning_rate, n_estimators

Best Performing Model

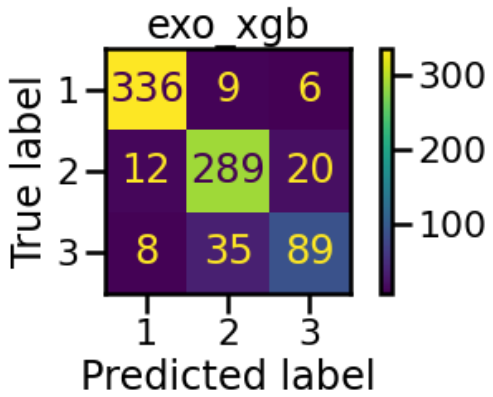
- Final evaluation of tuned models on hold-out set.
- XGBoost using decision stumps performs the best.
- Optimal hyperparameters:
  - Learning rate: 0.1
  - Number of estimators: 1000

XGBoost performance on hold-out

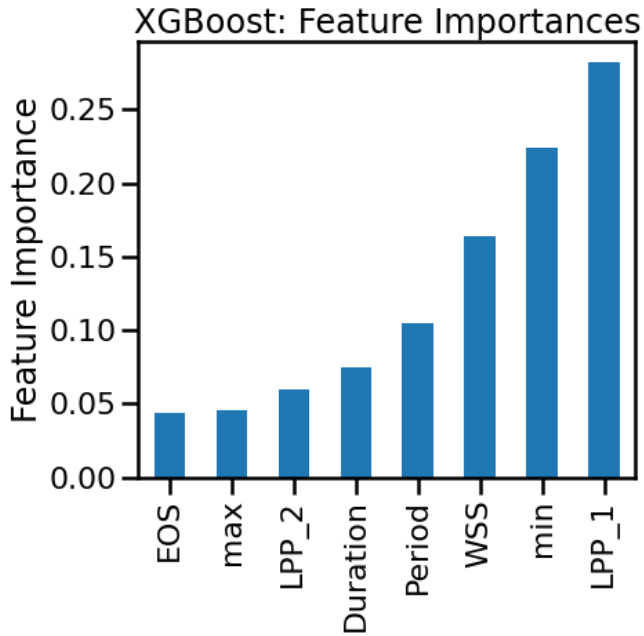
class	precision	recall	f1-score	support
1	0.943820	0.957265	0.950495	351.00000
2	0.867868	0.900312	0.883792	321.00000
3	0.773913	0.674242	0.720648	132.00000

Performance Summary + Feature Evaluation

- Exoplanet vs. FP: excellent.
- Secondary eclipse FP: pretty good.
- NTP FPs: some confusion w/ secondary FPs.

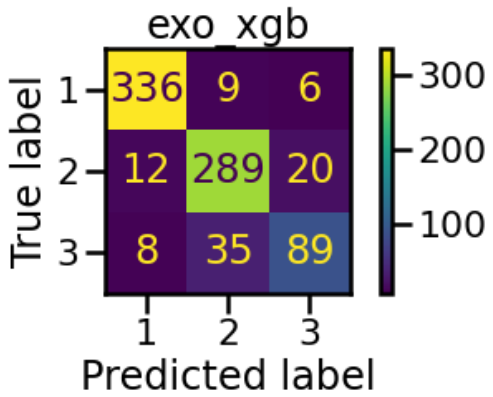


- Feature importances on similar scale.
- Different tree ensemble methods: ordering robust.

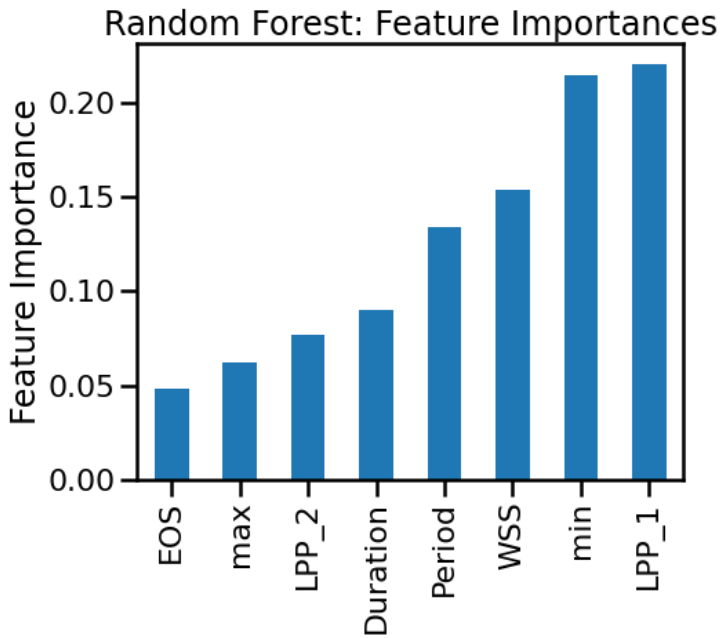


Performance Summary + Feature Evaluation

- Exoplanet vs. FP: excellent.
- Secondary eclipse FP: pretty good.
- NTP FPs: some confusion w/ secondary FPs.



- Feature importances on similar scale.
- Different tree ensemble methods: ordering robust.



Future Directions

- Start vetting from raw data. End-to-end pipeline.
- New missions, similar mission data validation streams.
- Model adaptation to new missions.

Figure: Transiting Exoplanet Survey Satellite (TESS)