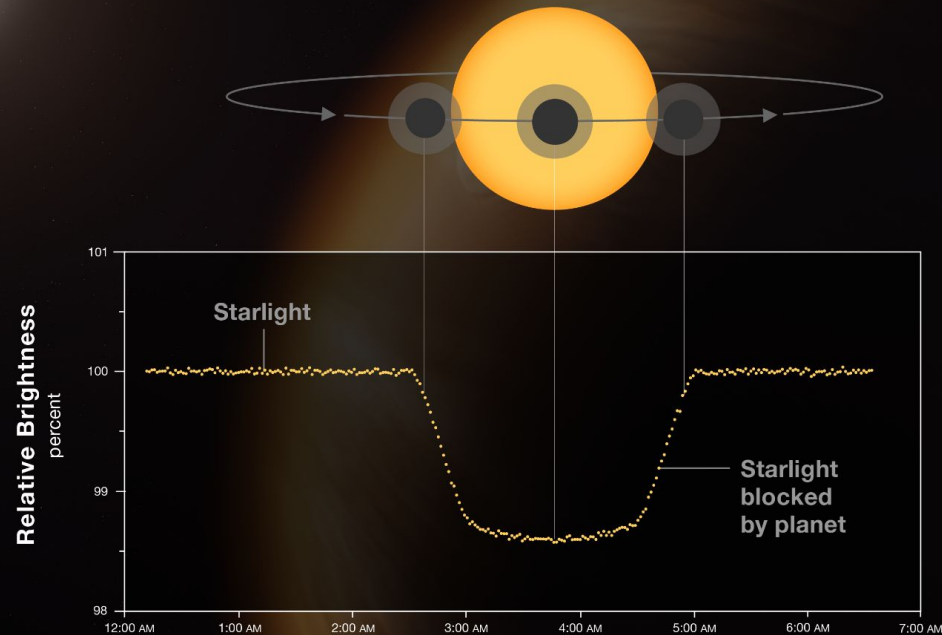# Identifying Exoplanets Using Machine Learning

An Analysis Using Classification Models by Joshua Gottlieb

HOT GAS GIANT EXOPLANET WASP-96 b
TRANSIT LIGHT CURVE

NIRISS | Single-Object Slitless Spectroscopy

Relative Brightness (percent)

Starlight

Starlight blocked by planet

101
100
99
98

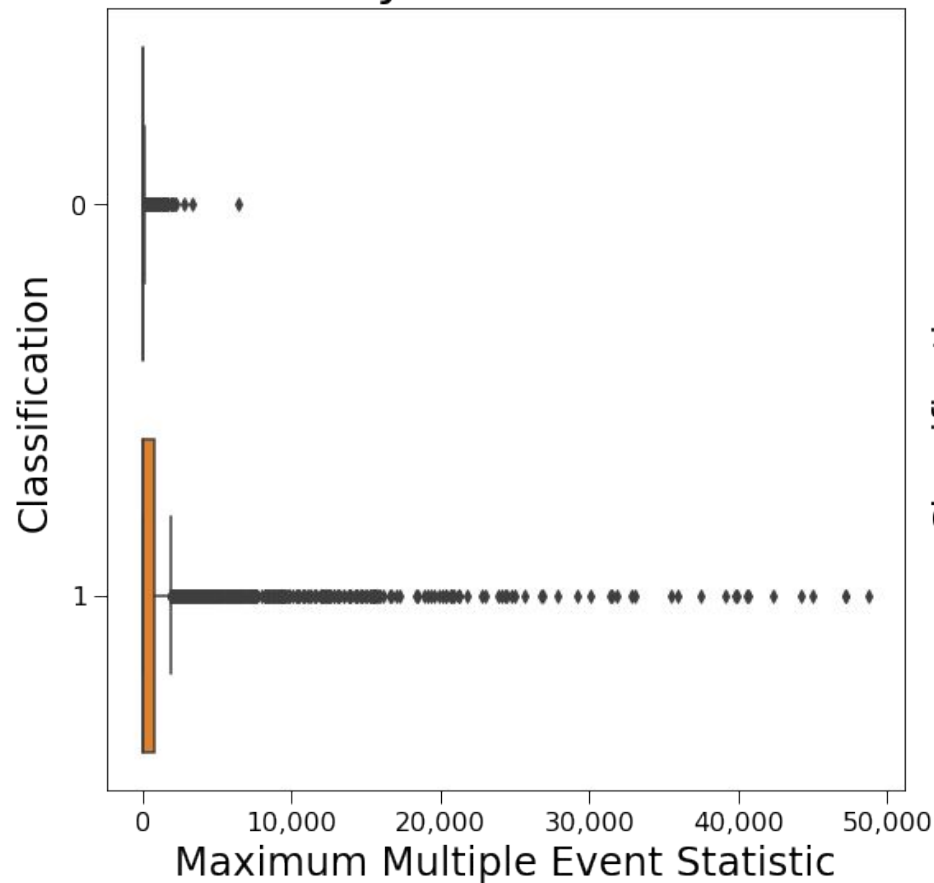12:00 AM   1:00 AM   2:00 AM   3:00 AM   4:00 AM   5:00 AM   6:00 AM   7:00 AM

Time in Baltimore, Maryland
June 21, 2022
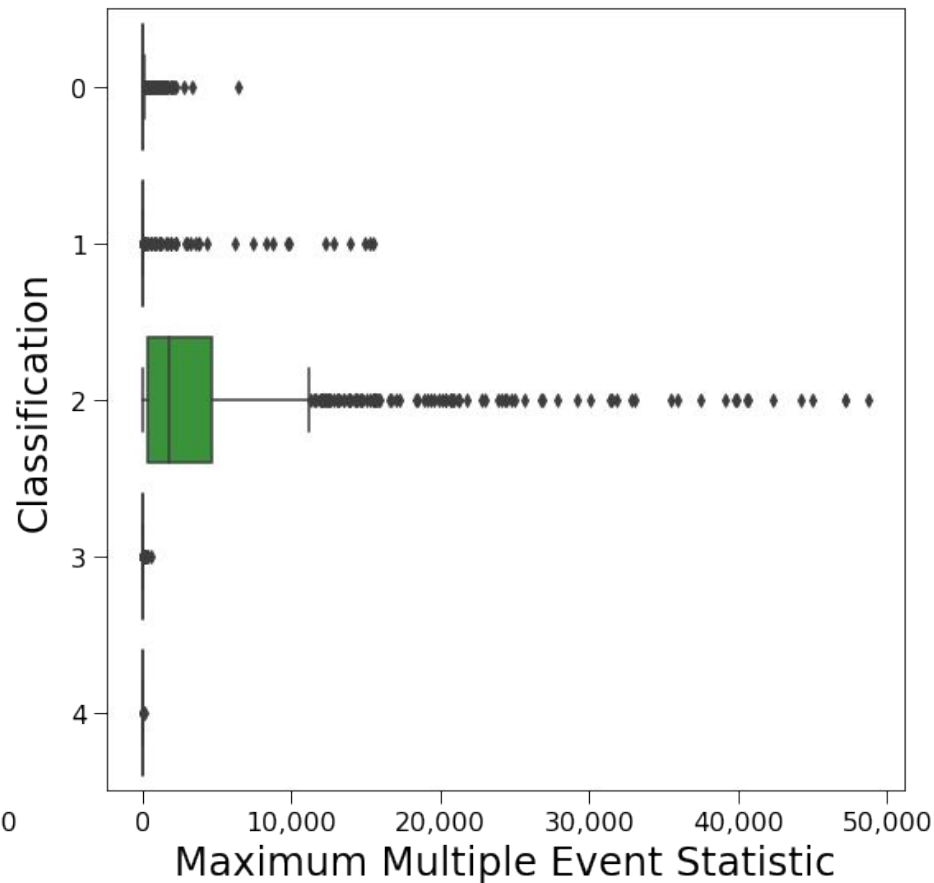
WEBB
SPACE TELESCOPE

Classifications

- Class 0: Confirmed Exoplanets
- Classes 1-4: False Positives
  - Class 1: Not-Transit (Ex: Background Object)
  - Class 2: Eclipsing Binary Star
  - Class 3: Centroid Offset (Ex: Nearby Star)
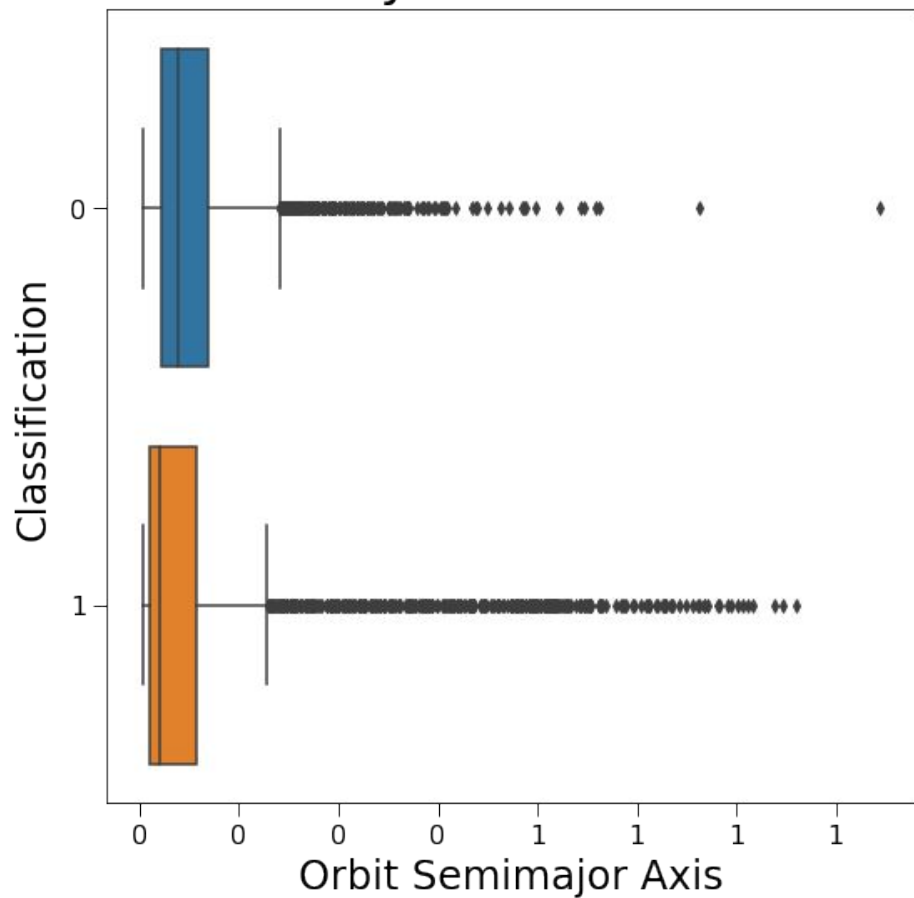 - Class 4: Measurement Contamination
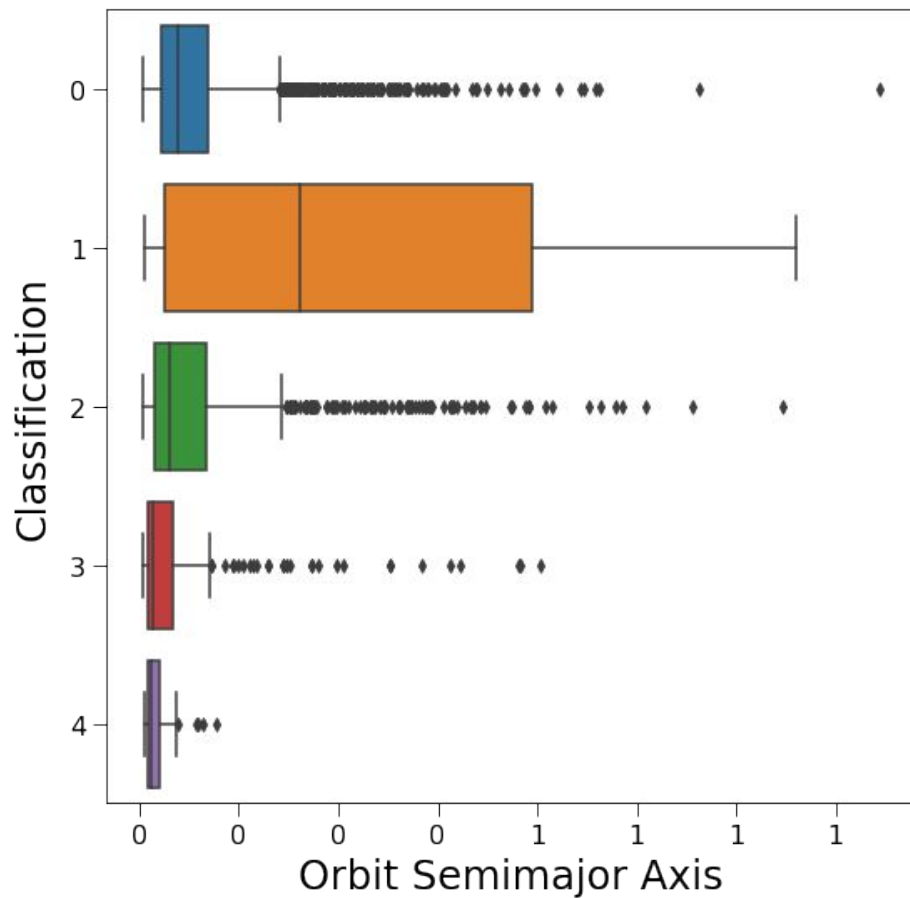
Binary Classification

Classification (y-axis): 0, 1
Maximum Multiple Event Statistic (x-axis): 0, 10,000, 20,000, 30,000, 40,000, 50,000

Multi-class Classification

Classification (y-axis): 0, 1, 2, 3, 4
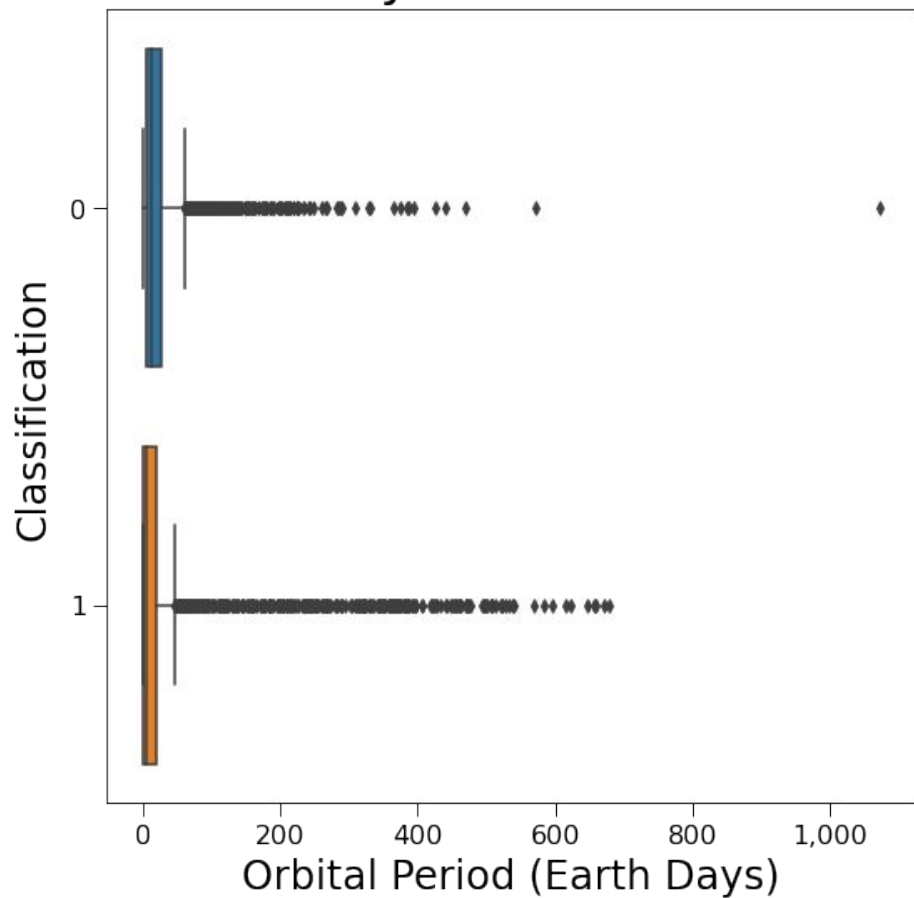Maximum Multiple Event Statistic (x-axis): 0, 10,000, 20,000, 30,000, 40,000, 50,000

Binary Classification

Multi-class Classification

Binary Classification — Multi-class Classification

Binary Classification

Planet-Star Radius Ratio

Multi-class Classification

Planet-Star Radius Ratio

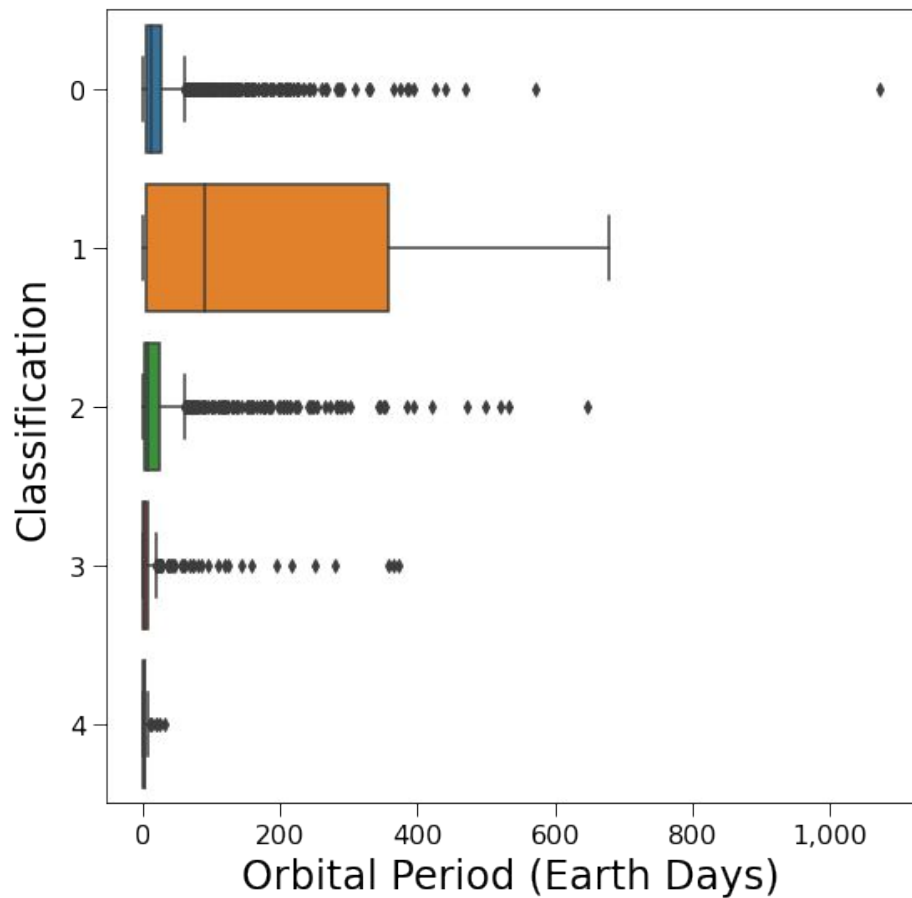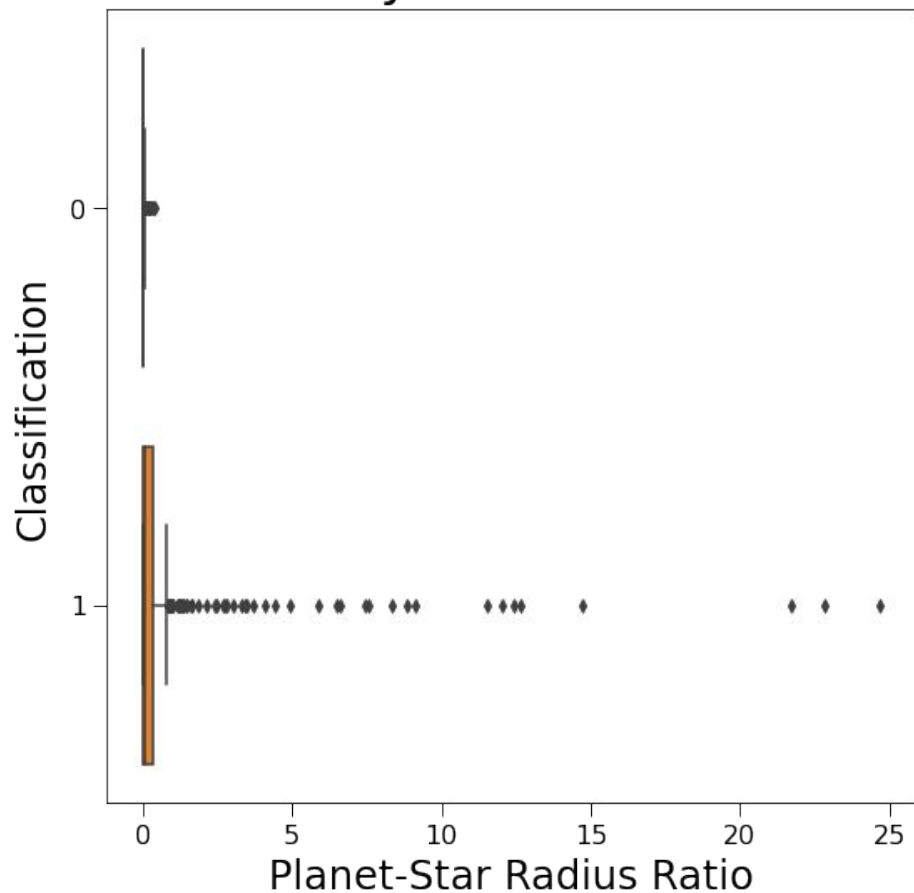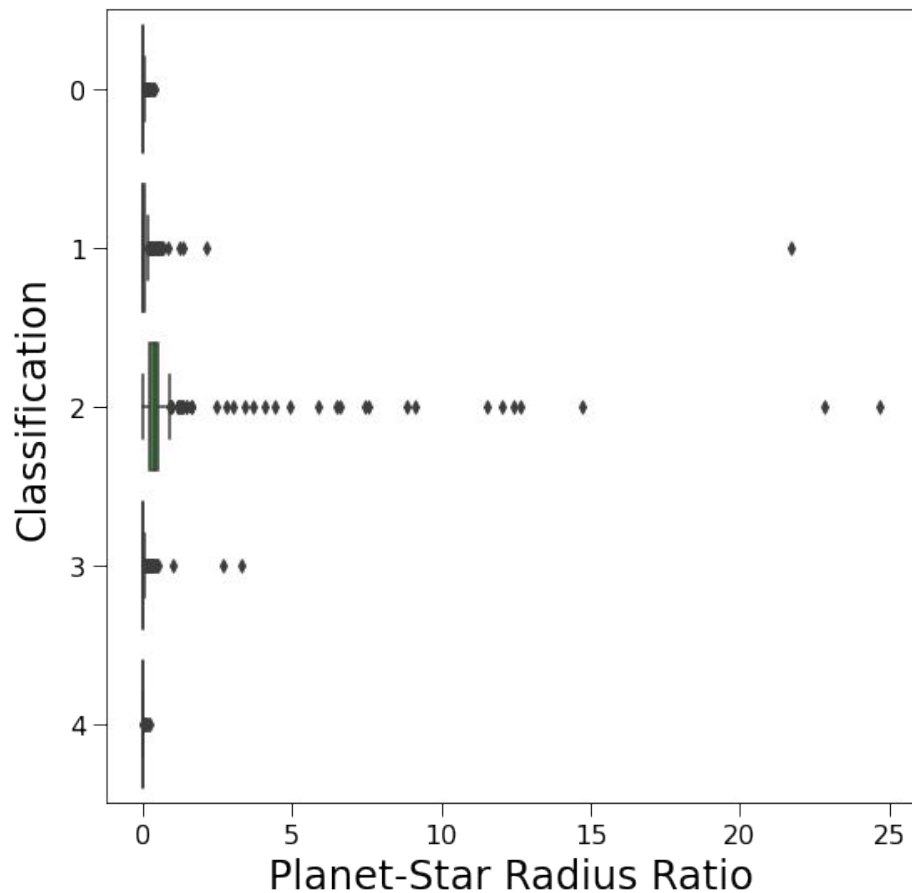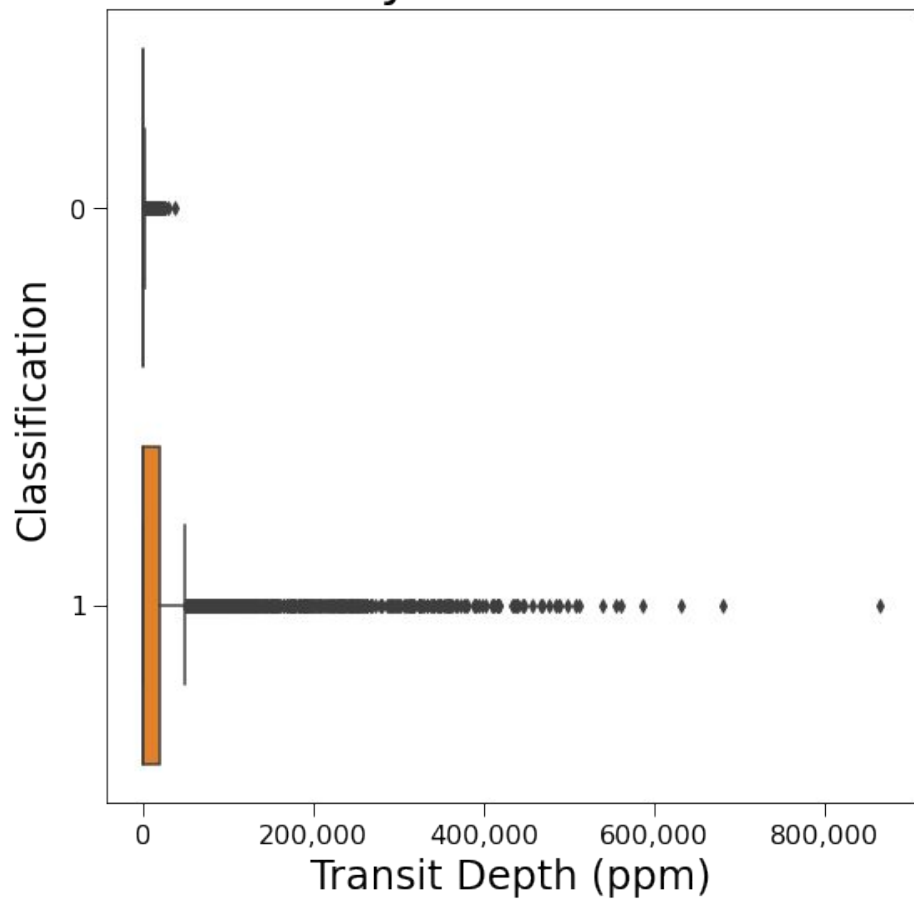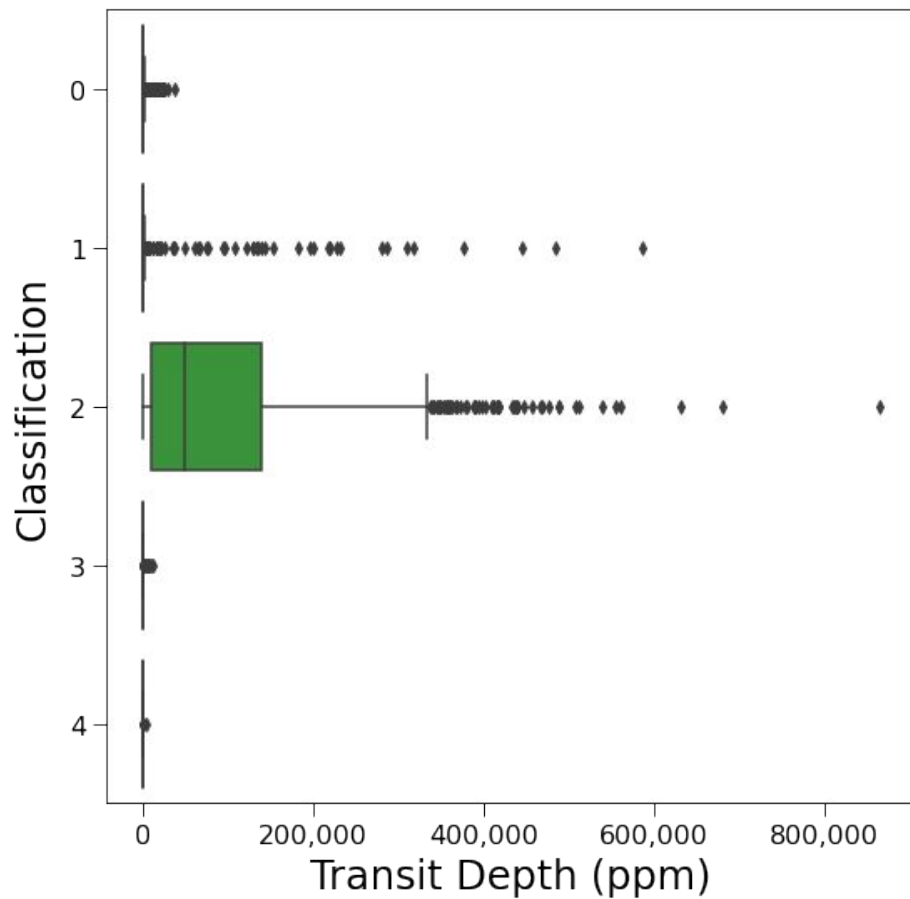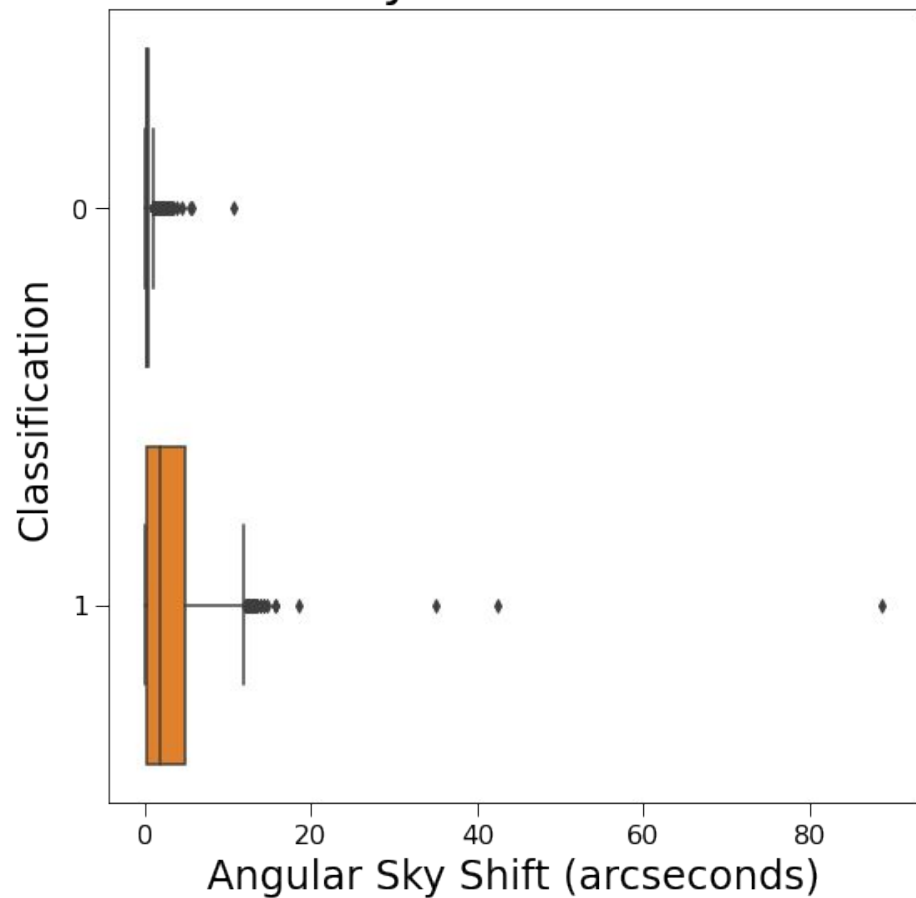Binary Classification ... Multi-class Classification

Binary Classification

Multi-class Classification
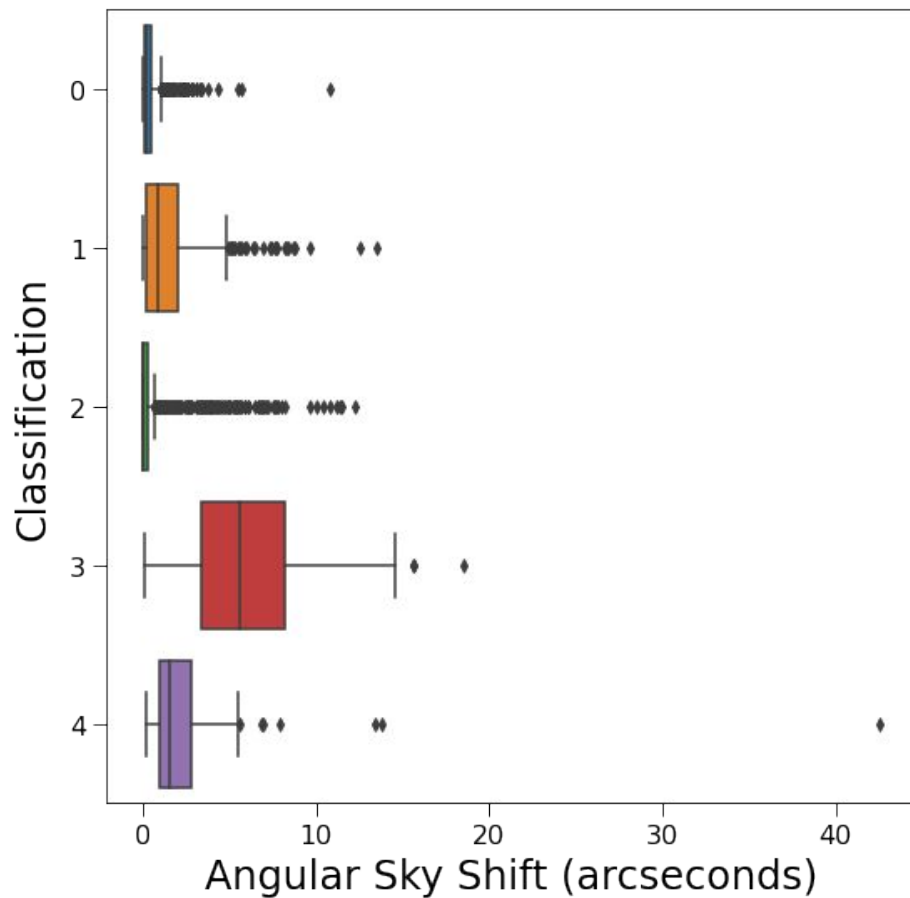
Types of Classifiers Tried

- Logistic Regression
- Decision Tree Classifier
- K-Nearest Neighbors
- Random Forests
- Ada Boosted Trees
- Gradient Boosted Trees
- XG (Extreme Gradient) Boosted Trees

Binary Classifiers - Baseline Logistic

Baseline Logistic Regression
- Default Parameters
- Max Iterations = 1000

	recall	precision	f1	accuracy	auc
0	0.956701	0.929860	0.943089	0.94781	0.986426
1	0.940476	0.963415	0.951807	NaN	NaN

Base Binary Model Confusion Matrix
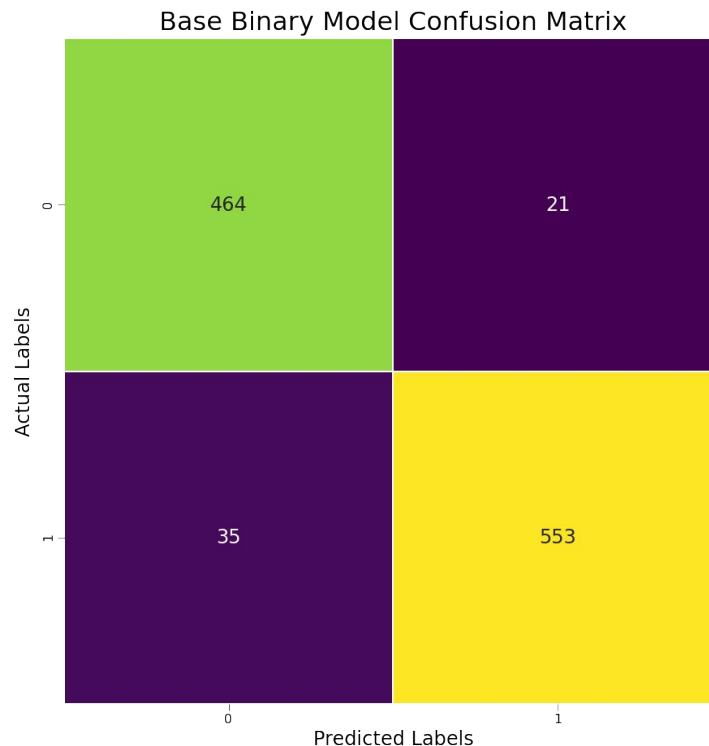
Binary Classifiers - Final XG Boosted

Final XG Boosted Tree
- # of Estimators: 500
- Max Tree Depth: 12

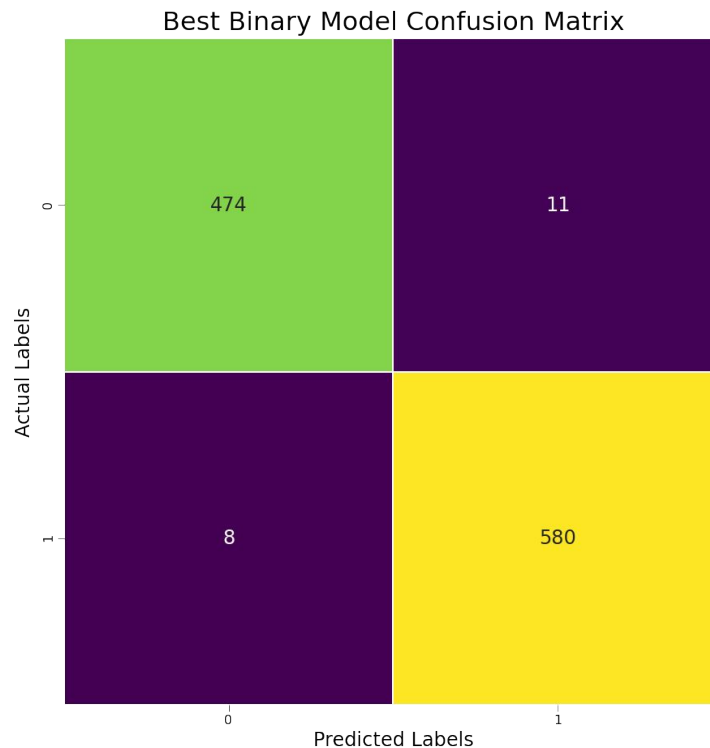|	recall	precision	f1	accuracy	auc
0	0.977320	0.983402	0.980352	0.982293	0.995866
1	0.986395	0.981387	0.983885	NaN	NaN

Best Binary Model Confusion Matrix

Multiclass Classifiers - Baseline Tree

Baseline Decision Tree
- Default Parameters
- Random Oversampling

	recall	precision	f1	accuracy	auc
0	0.925773	0.920082	0.922919	0.856808	0.854234
1	0.732394	0.611765	0.666667	NaN	NaN
2	0.807882	0.906077	0.854167	NaN	NaN
3	0.760000	0.770270	0.765101	NaN	NaN
4	0.444444	0.333333	0.380952	NaN	NaN

Base Multiclass Model Confusion Matrix
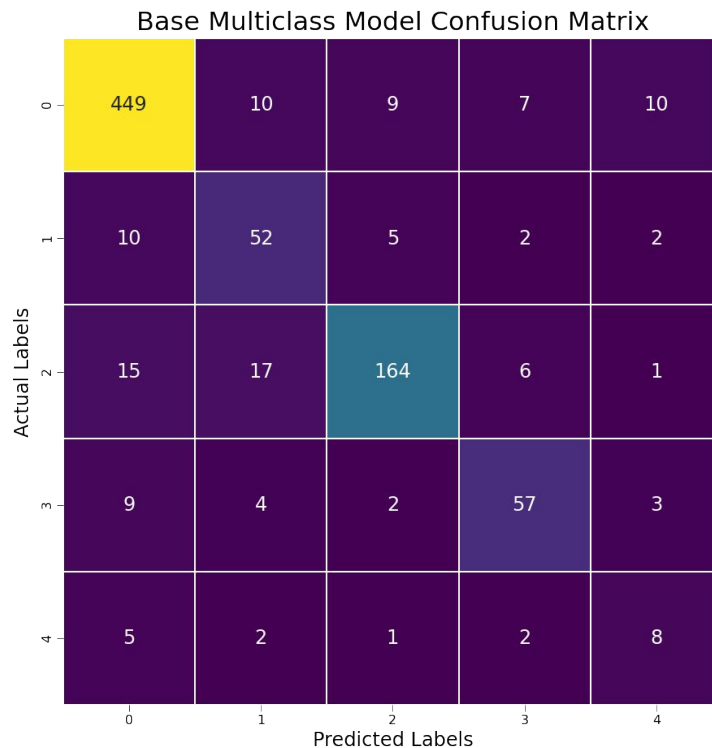
Multiclass Classifiers - Final XG Boosted

Final XG Boosted Tree
- # of Estimators: 1000
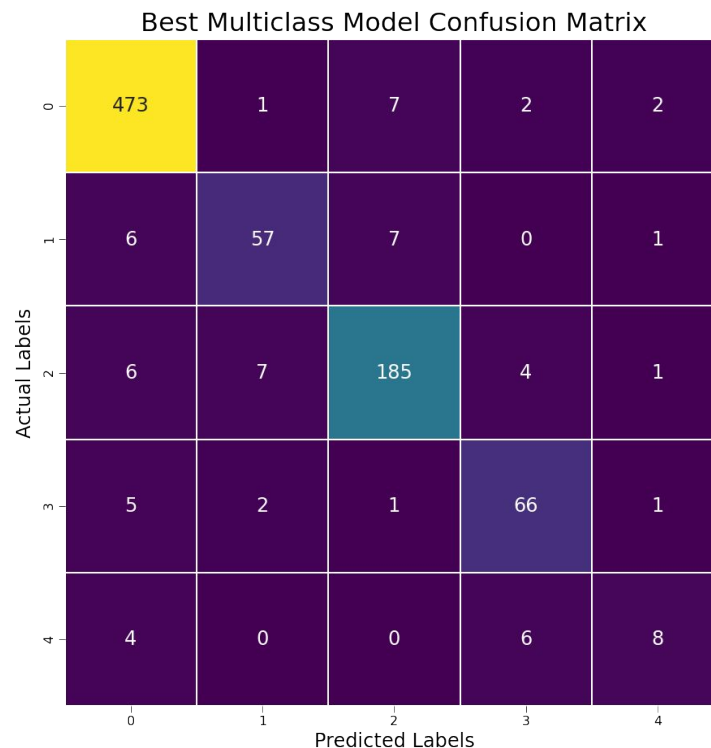- Max Tree Depth: 6
- % of Features by Node: 50%
- Random Oversampling

	recall	precision	f1	accuracy	auc
0	0.975258	0.957490	0.966292	0.926056	0.982354
1	0.802817	0.850746	0.826087	NaN	NaN
2	0.911330	0.925000	0.918114	NaN	NaN
3	0.880000	0.846154	0.862745	NaN	NaN
4	0.444444	0.615385	0.516129	NaN	NaN

Best Multiclass Model Confusion Matrix

Best Binary Classification Feature Importances
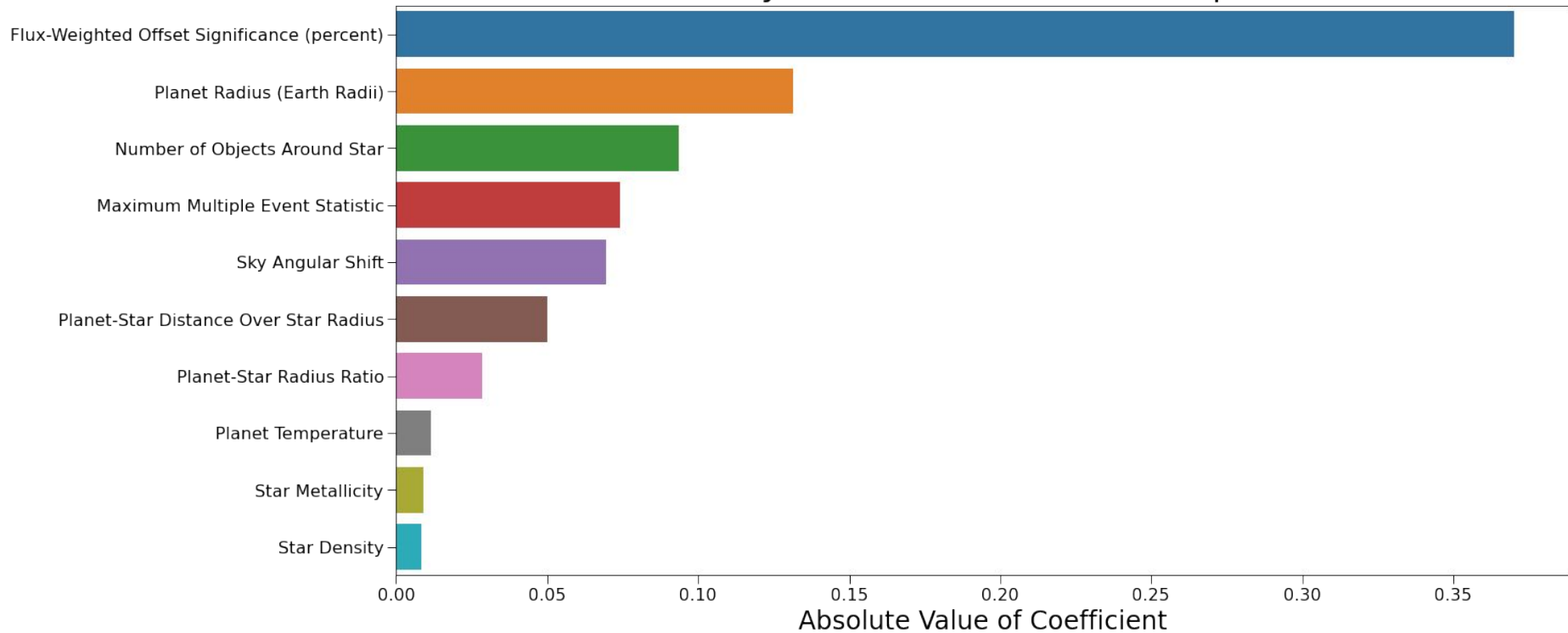
Best Multi-Classification Feature Importances

Feature	
Number of Objects Around Star	
Sky Angular Shift	
Orbit Semimajor Axis Length	
Transit Depth	
Maximum Multiple Event Statistic	
Planet-Star Radius Ratio	
Flux-Weighted Offset Significance (percent)	
Maximum Single Event Statistic	
Orbital Period (Days)	
Kepler-Band Magnitude	

Absolute Value of Coefficient
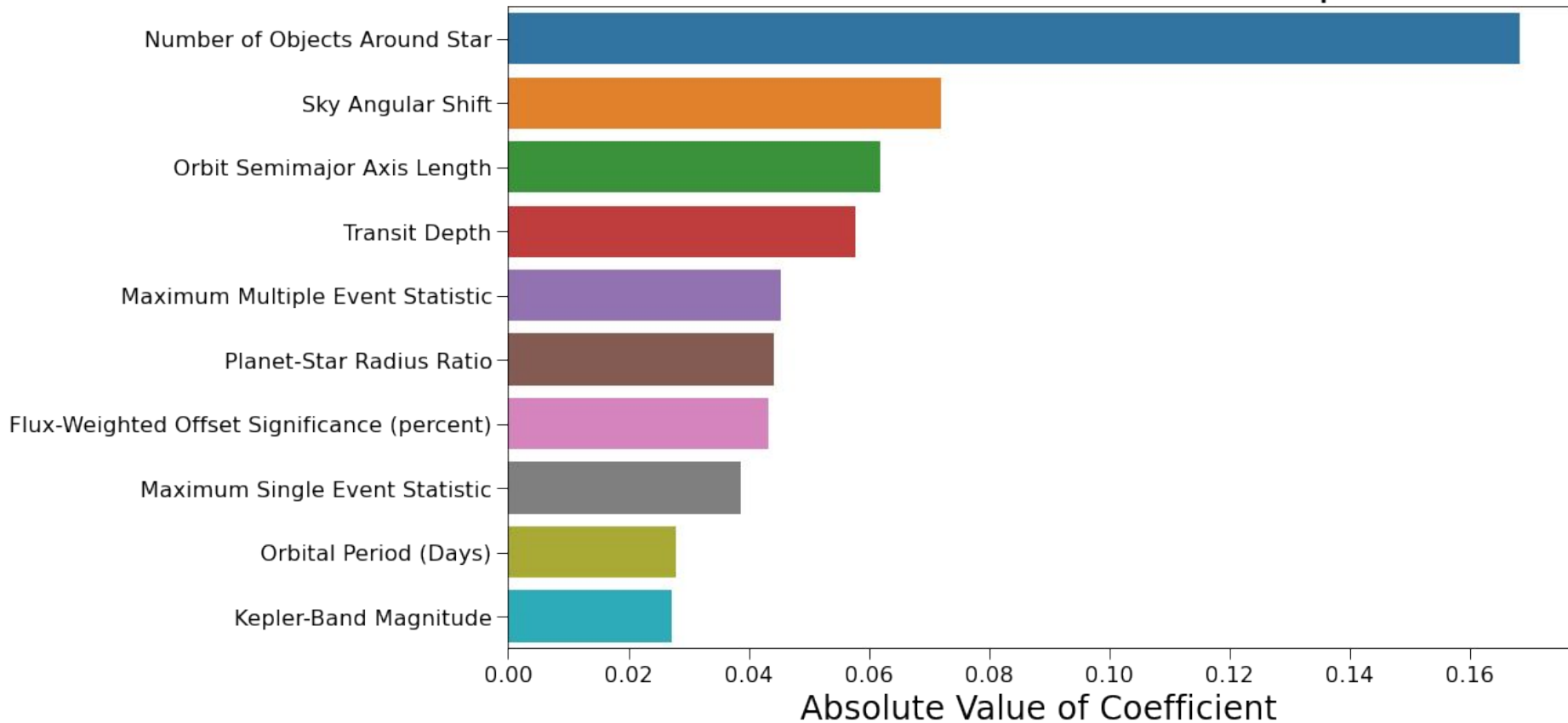
Meaningful Results

- Classifiers are good at separating true Exoplanets from false positives, but struggle to identify class 1, 3, and 4 false positives.
- Features such as Max MES, Orbit Semimajor Axis, Orbital Period, Transit Depth, and Planet-Star Radius Ratio are strong predictors of False Positives

Next Steps

- Gather more data to deal with large class imbalances in a more natural manner.
- Test stacked and/or voting classifiers

Thank you for listening!

Presenter: Joshua Gottlieb

Email: joshuadavidgottlieb@gmail.com

LinkedIn: linkedin.com/in/joshua-gottlieb

GitHub: github.com/JoshuaGottlieb