june 26th, 2025

exoplanet classification + conference prep.

past weeks

- working on updated poster for **AstroAI** + recap over stuff i presented with machine learning
- feature importance analysis through **SHAP** ([**SH**apley **A**dditive ex**P**lanations])
- continuing work on exoplanet classification – literary studies
  - four classes system i briefly covered before → similar, anti-ordered, ordered, mixed
  - now introducing

Data Cleanup and Training Models

- model training data sources: **5,834** exoplanets from NASA Exoplanet Archive joined with **5,599** exoplanets from HWC

Step 1: Data Downloading

- *NASA Exoplanet Archive (5,834 exoplanets)*
- *Habitable Worlds Catalog (HWC), PHL @ UPR Arecibo (5,599 exoplanets)*

Step 2: Data Preprocessing

Data Joining

Join data from NASA and HWC.

HWC dataset has a ***P-HABITABLE*** data field, which indicates exoplanet habitability and is used to label training data.

Data Cleaning

* Filter out exoplanets in questions (with ***pl_controv_flag***=1).

* Limit to exoplanets with **single-host star**.

* Remove data fields that are not relevant to model training.

Handling Missing Values

* Remove data fields with 25+% missing values.

* **Categorical**: fill missing values with **mode**.

* **Numerical**: fill missing values with imputation.

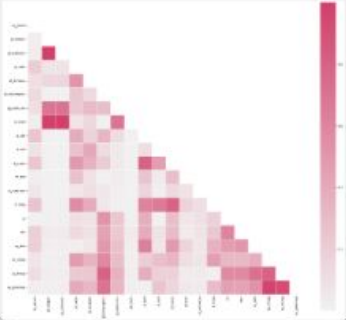
Handling Data Imbalance

* The dataset is highly imbalance: 4,527 negative samples (non-habitable) and 55 positive samples (habitable).

* Apply combination of oversampling and downsampling techniques to oversample minorities and downsample majorities.

Data Correlation Analysis

* Remove redundant data fields based on **data correlation analysis**.



Min-Max Scalar

Standardize value ranges to **[0.1]**

Step 3: Model Training & Evaluation

* Randomly shuffle dataset, then split into train & test datasets for training & evaluation. Repeat. Use avg accuracy as the measure of model quality for the corresponding hyperparameters.

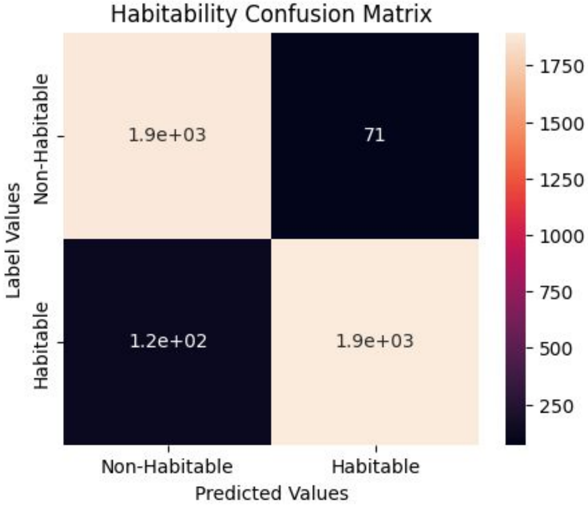
* Hyperparameter tuning through exhaustive search.

Step 4: Feature Importance Analysis

* Feature important analysis through **SHAP** (SHapley Additive exPlanations).

Random Forest and XGBoost model performance

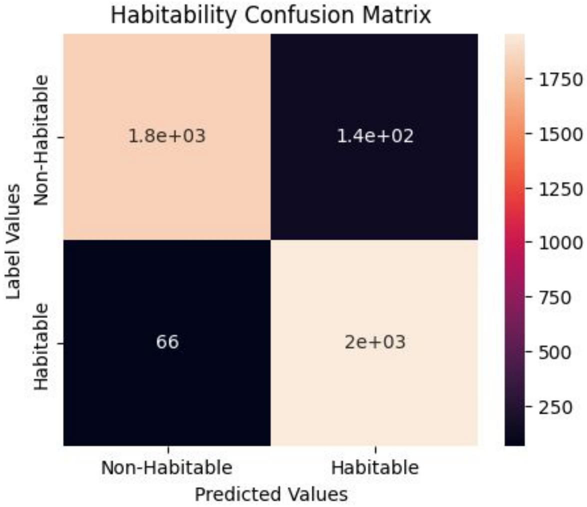
Random Forest classifier



Random Forest Classifier - Classification Report:

	precision	recall	f1-score	support
0.0	0.94	0.96	0.95	1959
1.0	0.96	0.94	0.95	2020
accuracy			0.95	3979

XGBoost classifier



XGBoost Classifier - Classification Report:

	precision	recall	f1-score	support
0.0	0.97	0.93	0.95	1959
1.0	0.94	0.97	0.95	2020
accuracy			0.95	3979

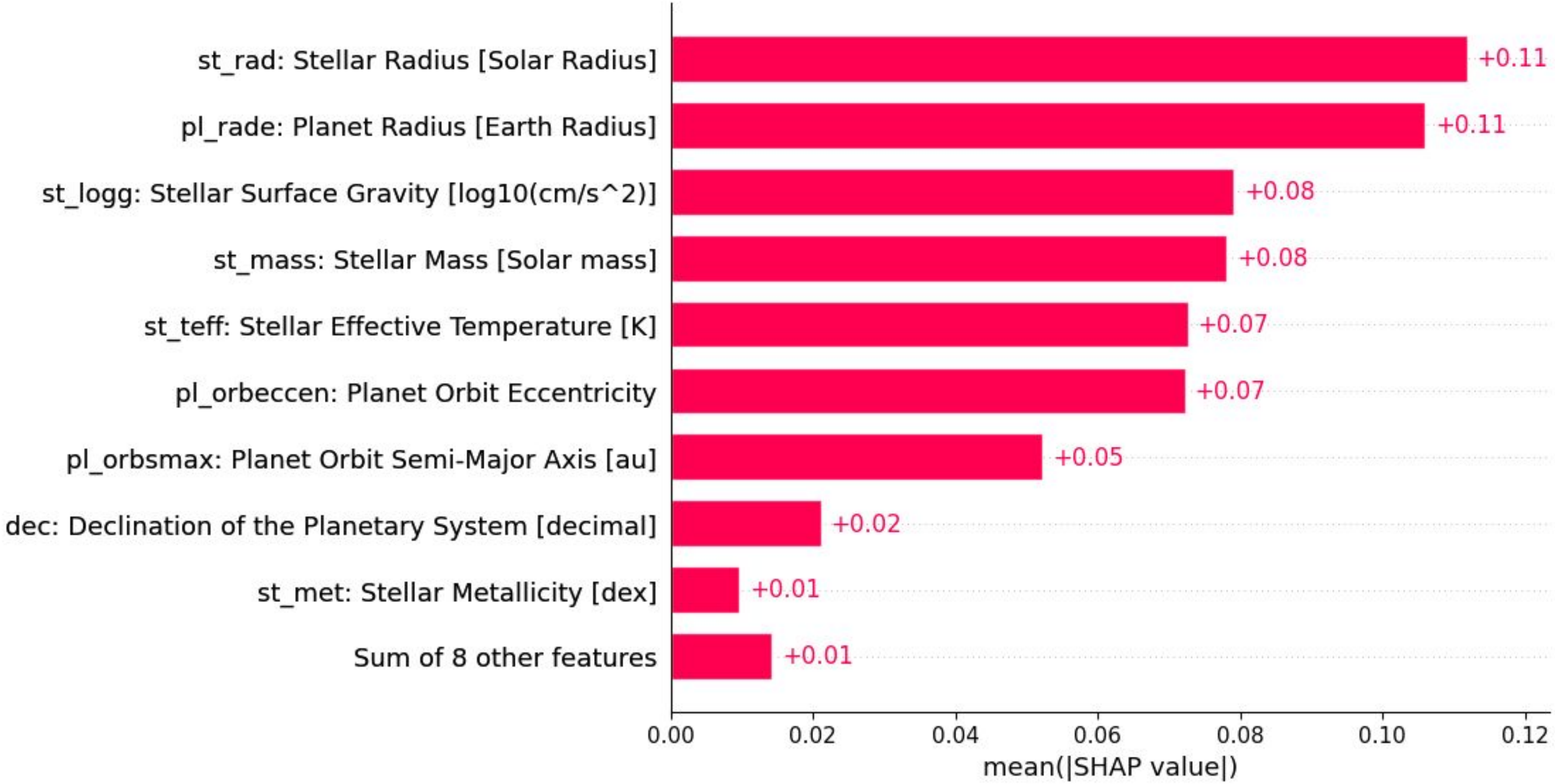
feature importance

- introduce SHAP framework → helps us to recognize the impact of each individual feature positively or negatively affecting our outcome.

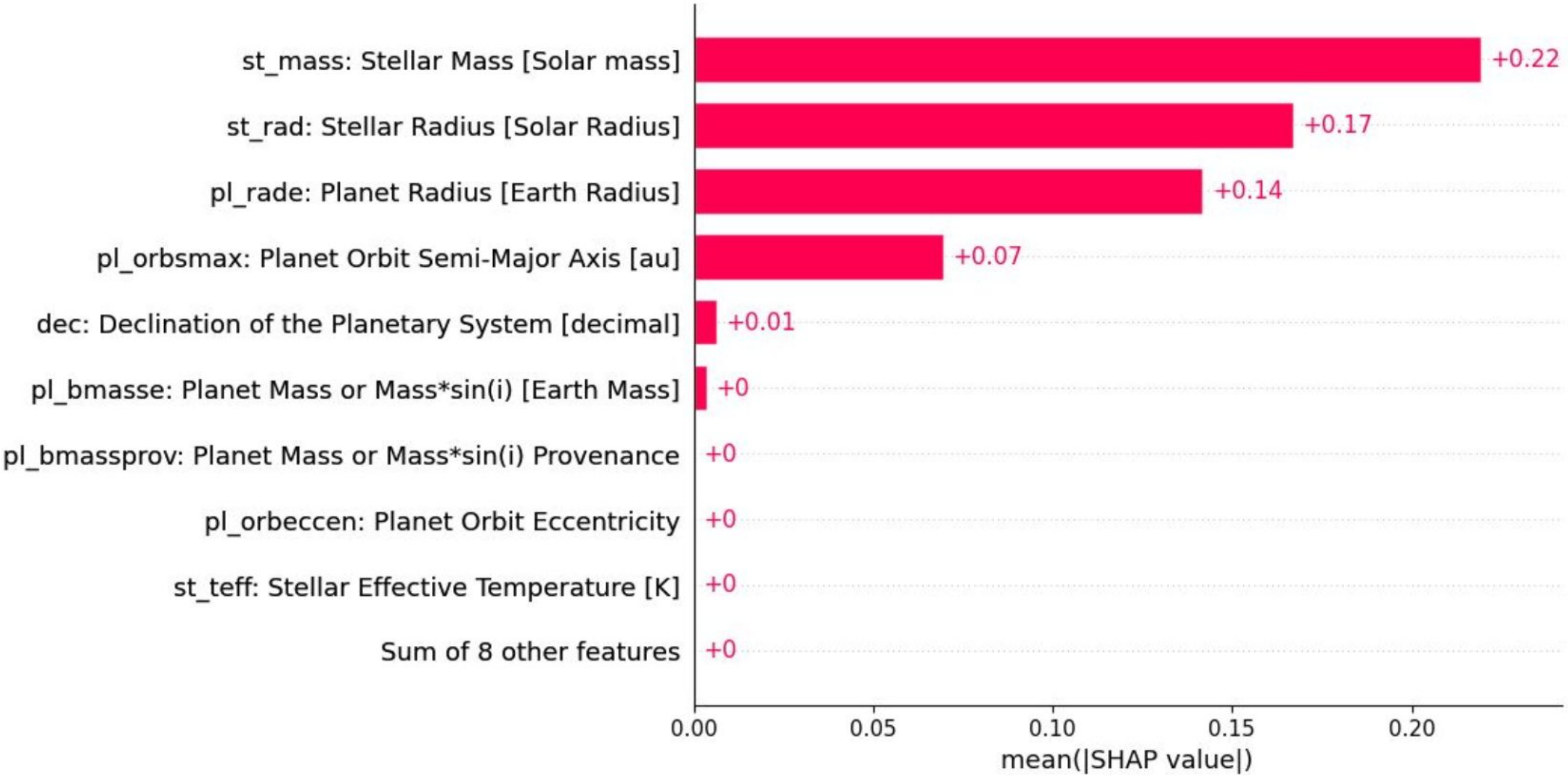
feature importance

- introduce SHAP framework → helps us to recognize the impact of each individual feature positively or negatively affecting our outcome.
- imagine we have a set of features:
  - look at each feature (ex. feature 1) and train the model once with the feature, once without the feature, for each subset of the rest of the features
  - each of these evals. are applied to test dataset; find model predicted diff.
  - find weighted average across all of the diff. subsets → **feature 1’s SHAP value**

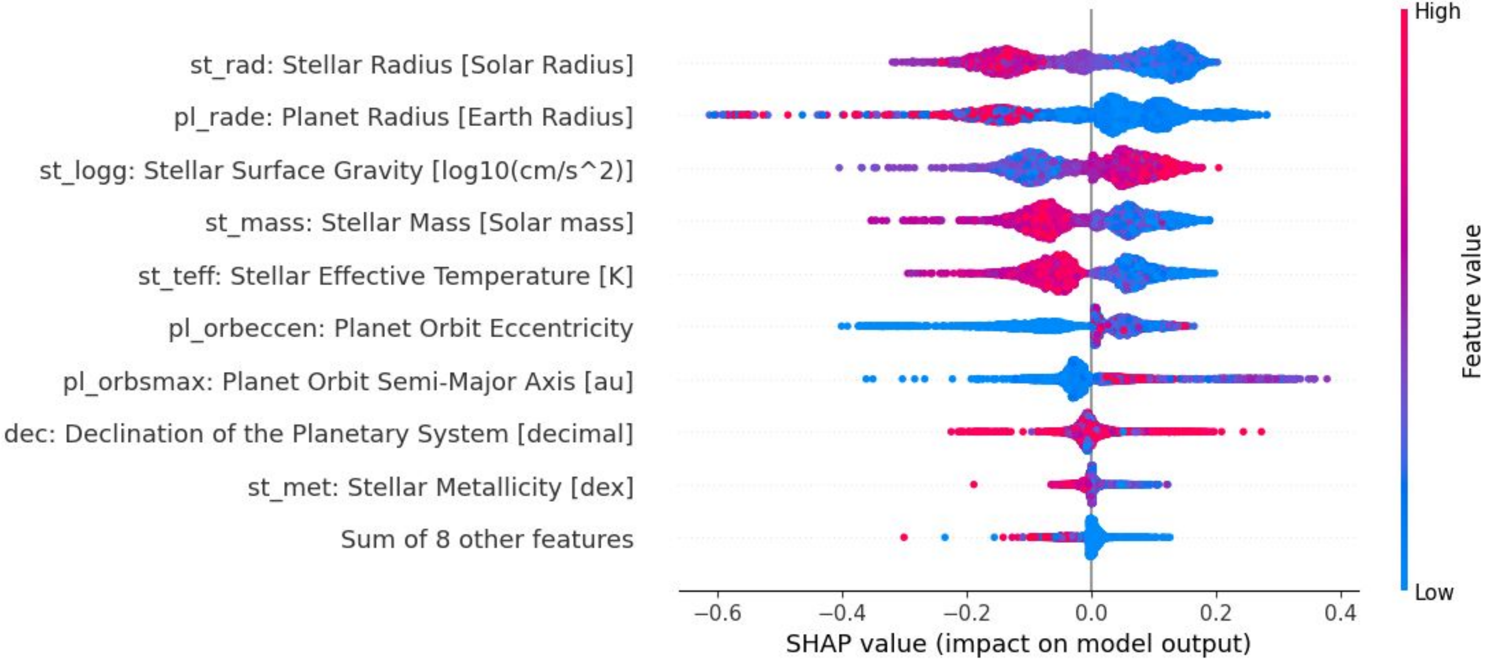
Random Forest feature importance via SHAP



XGBoost feature importance via SHAP

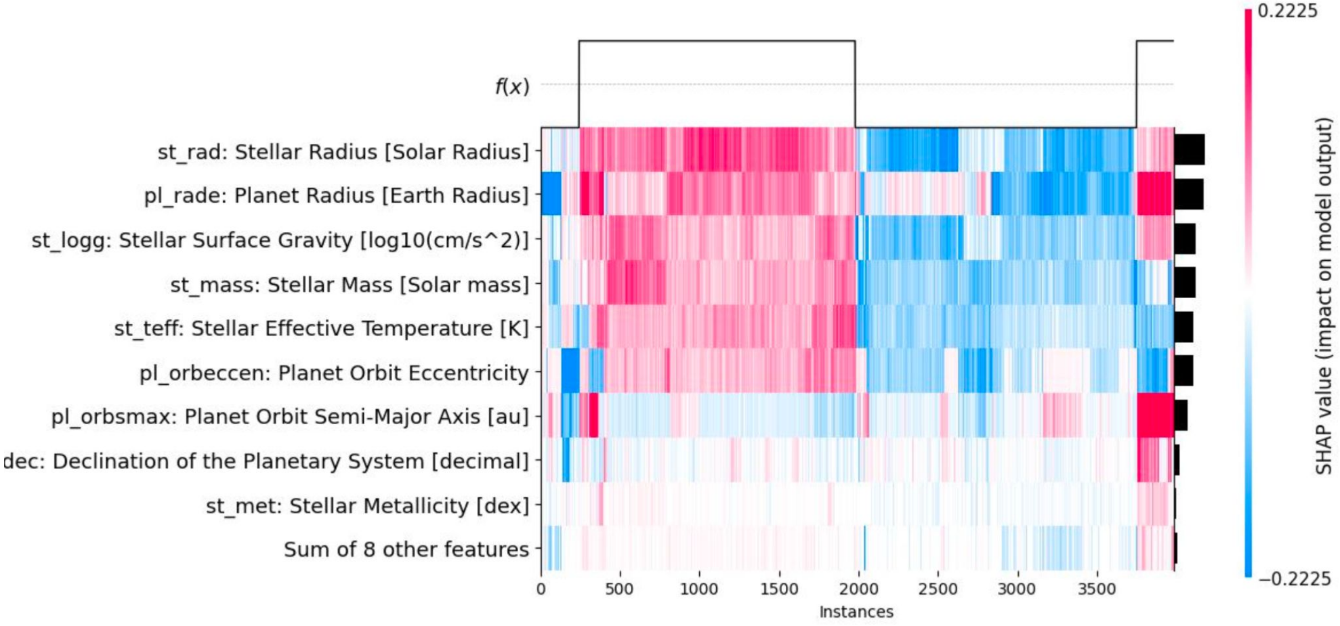


SHAP beeswarm plot - Random Forest



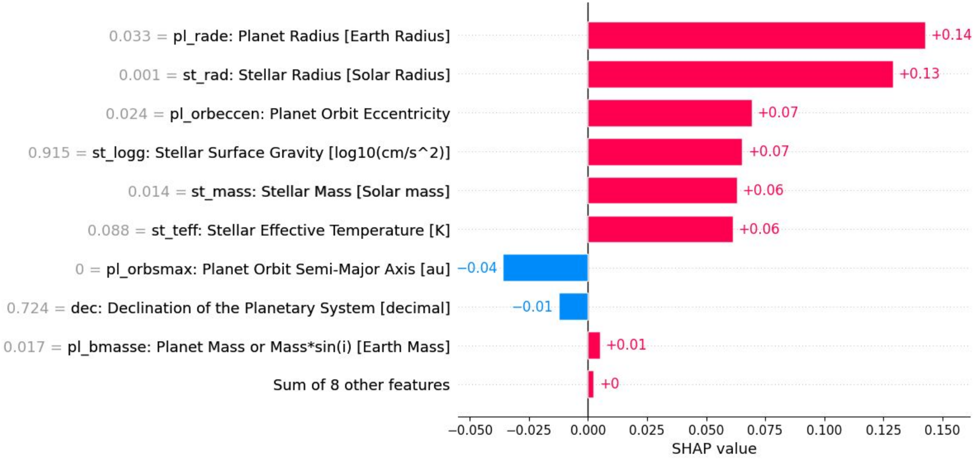
- Higher values (relative to other samples in dataset) of stellar radius, planet radius, stellar mass, and stellar effective temperature lead towards negative predictions, while lower values lead towards positive outcomes.
- Planet orbit semi-major axis, on the other hand, has the opposite impact on prediction outcomes, with higher values leading toward positive predictions while lower values leading towards negative outcomes.

SHAP heatmap - Random Forest

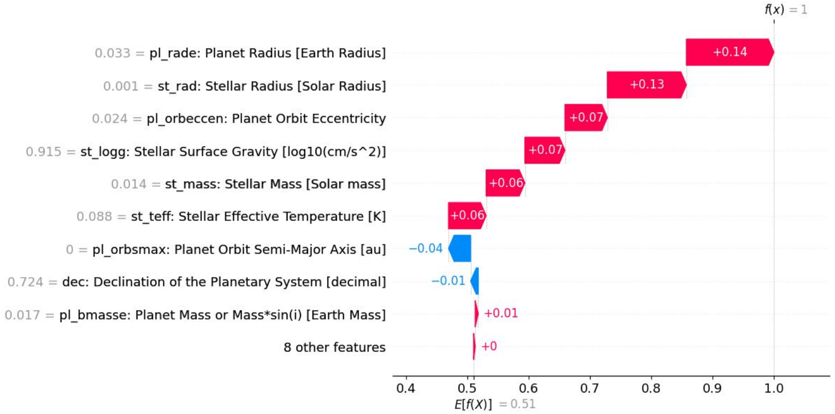


- Higher values (relative to other samples in dataset) of stellar radius, planet radius, stellar mass, and stellar effective temperature lead towards negative predictions, while lower values lead towards positive outcomes.
- Planet orbit semi-major axis, on the other hand, has the opposite impact on prediction outcomes, with higher values leading toward positive predictions while lower values leading towards negative outcomes.

SHAP local bar and waterfall plots - Random Forest

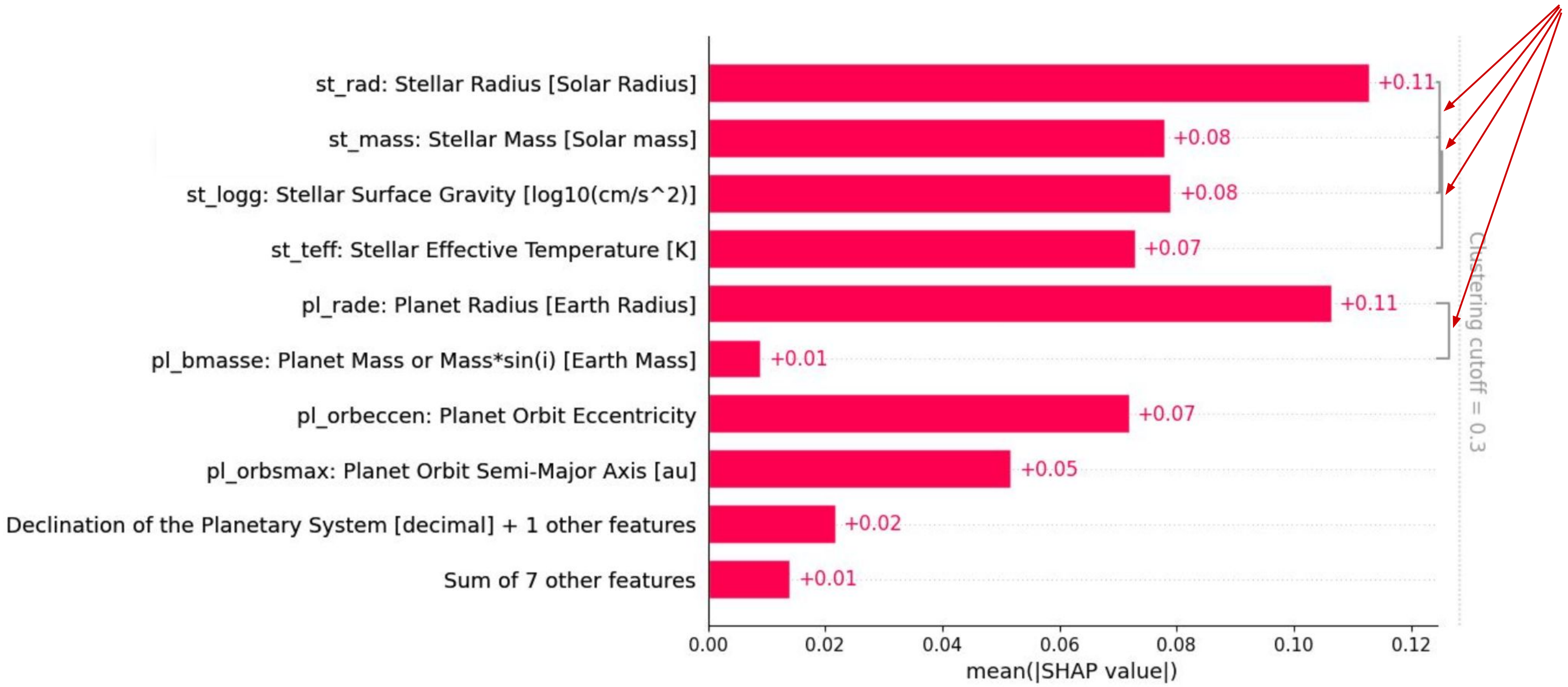


SHAP local bar plot for one sample in the dataset

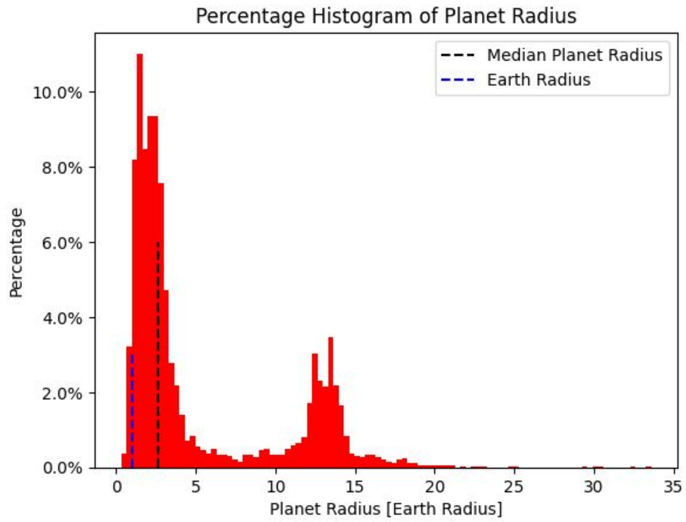


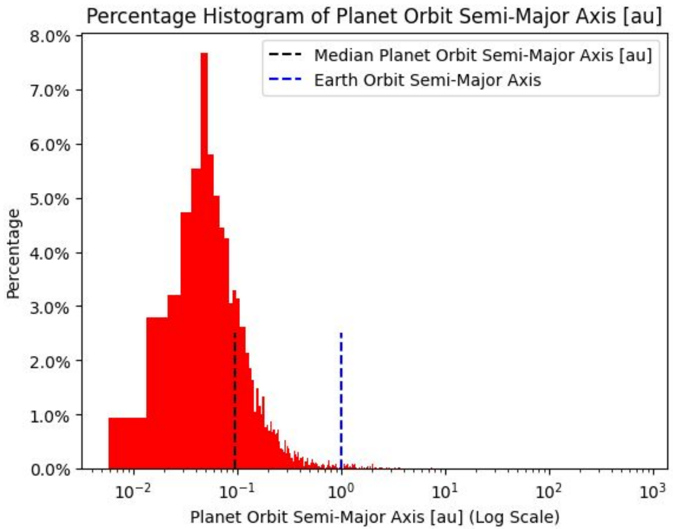
SHAP waterfall plot for one sample in the dataset

SHAP feature correlation analysis



discussion of results and relating to IRL trends





SHAP analysis indicates a higher planet radius leads towards negative predictions, while lower value leads towards positive predictions.

SHAP analysis indicates a higher planet orbit semi-major axis leads towards positive predictions, while lower value leads towards negative predictions.

planetary system classification

exoplanet classification – literary studies

Framework for the architecture of exoplanetary systems (2023)
(DOI: https://doi.org/10.1051/0004-6361/202243751)

Similar – similar regardless of distance
Anti-ordered – as planets further, get smaller
Ordered – as planets further, get larger
Mixed – goes back and forth

Architecture class	Condition
Anti-ordered	$C_S(M) < -0.2$
Ordered	$C_S(M) > +0.2$
Similar	$\lvert C_S(M)\rvert \leq 0.2$ and $C_V(M) \leq \frac{\sqrt{n-1}}{2}$
Mixed	$\lvert C_S(M)\rvert \leq 0.2$ and $C_V(M) > \frac{\sqrt{n-1}}{2}$

(3)

coefficient of similarity – positive for ordered, negative for anti-ordered

$$C_s(q) = \frac{1}{n-1}\sum_{i=1}^{n-1}\left(\log\frac{q_{i+1}}{q_i}\right)$$

where q_i is some planetary quantity q (ex. mass, radius, orbital period, etc.) for the i^th planet in a system.

coefficient of variation – measure magnitude of variation in a set of numbers

$$C_v(q) = \frac{\sigma(q)}{\overline{q}}$$

"while similar systems will have a low value of the coefficient of variation, mixed systems will have a high value of coefficient of variation"

exoplanet classification – literary studies

Framework for the architecture of exoplanetary systems (2023)
(DOI: https://doi.org/10.1051/0004-6361/202243751)

used a model called the GENERATION III BERN MODEL in the process to create synthetic data (under heading *2.1 Theoretical Dataset: Bern Model*)

- system of classification they use requires ≥ 3 planets per system, thus out of their original dataset there were only 41 data points.
- gen iii bern model to generate 1000 such systems

exoplanet classification – literary studies

Architecture Classification for Extrasolar Planetary Systems (2025)
(DOI: https://doi.org/10.1051/0004-6361/202243751)

- uses 6000 exoplanets (only real data!)
- basically just a straight-up split very similar to earlier ones we talked about
- hot Jupiters discussed

exoplanet classification – literary studies

Planetary Population Synthesis and the Emergence of Four Classes of Planetary System Architecture (2023)
(DOI: https://doi.org/10.48550/arXiv.2303.00012)

- This paper also uses synthetic data generated using the **GENERATION III BERN MODEL** which seems to be pretty popular.