july 3rd, 2025

christina liu

stuff i did

- working on poster for astroAI on july 7th (so this will be a short presentation)
- some new SHAP updates (force plots)
- literary studies on more of the astrophysics side of classification instead of just using machine learning
- some ideas on how to approach planetary systems analysis

Analyzing the Impacts of Stellar and Planetary Parameters to Exoplanet Habitability through Machine Learning

Christina X. Liu[1], Jonathan H. Jiang[2]
[1]Lakeside School, Seattle, WA, [2]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA

Introduction

Are there any planets other than Earth potentially habitable by humans? What factors affect habitability and how do they impact habitability? Seeking answers to these questions has been an active research field in astrophysics in the past decade.

In this study, we identified the most influential stellar and planetary parameters to exoplanet habitability and analyzed how they impact habitability through the lens of machine learning.

Objectives

Build high-quality ML models (**Random Forest**, **XGBoost**) for exoplanet habitability prediction.

Perform analysis on the ML models through **SHAP** (**SH**apley **A**dditive ex**P**lanations) techniques to understand how stellar and planetary parameters impact exoplanet habitability.

Dataset

The primary data sources for this study:

- Planetary Systems Composite Data @ **NASA Exoplanet Archive**: **5,834** confirmed exoplanets.
- Habitable World Catalog (**HWC**), PHL @ UPR Arecibo: **5,599** exoplanets.

The NASA dataset was joined with the HWC dataset. Then the `P_HABITABLE` data field from the HWC dataset to label the training data.

Model Selection

Research in machine learning shows that tree-based models could be very effective for tabular dataset (on par with deep learning models). In this study, we chose to build **Random Forest** and **XGBoost** classifiers for exoplanet habitability prediction.

Feature Engineering & Model Training

Step 1: Data Downloading

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

Step 2: Feature Engineering

- **Data Joining**: Join data from NASA and HWC. Use `P-HABITABLE` data field from HWC to label training data.
- **Data Cleaning**: * Filter out exoplanets with questionable status. (`pl_controv_flag`=1) * Limit dataset to exoplanets. with **single-host stars**. * Remove irrelevant features.
- **Handling Missing Values**: * Remove features with ≥25% missing values. * **Categorical**: fill missing values with **mode**. * **Numerical**: fill missing values with Multivariate Imputation by Chained Equation (**MICE**).
- **Converting Categorical Values**: * **One-Hot Encoding** to convert categorical features into numerical values.
- **Handling Data Imbalance**: * The dataset is highly imbalanced. Apply a combination of Synthetic Minority OVersampling Technique (**SMOTE**) and Edited Nearest Neighbors (**ENN**) techniques to oversample minorities and downsample majorities.
- **Data Correlation Analysis**: Remove redundant features based on **data correlation analysis**.

- **Min-Max Scalar**: Standardize value ranges to **[0.1]**

Step 3: Model Training

* Randomly shuffle dataset, then split into train & test datasets for training & evaluation. Repeat 100 times. Use avg accuracy as the measure of model quality for the corresponding hyperparameters.

* Hyperparameter tuning through exhaustive search.

Step 4: SHAP Analysis

* Feature important analysis through **SHAP**.

* Perform analysis through **SHAP** to understand how the influential features impact the exoplanet habitability.

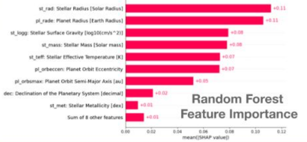
Random Forest Model Evaluation

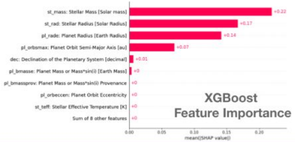
Class	Random Forest Precision	Recall	F1 score
1 (Habitable)	0.96	0.94	0.95
0 (Non-Habitable)	0.94	0.96	0.95

XGBoost Model Evaluation

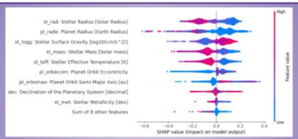
Class	XGBoost Precision	Recall	F1 score
1 (Habitable)	0.94	0.97	0.95
0 (Non-Habitable)	0.97	0.93	0.95

SHAP Analysis

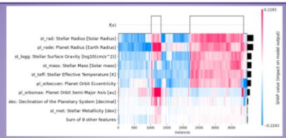




Deep Dive on Random Forest through SHAP



This SHAP **beeswarm** plot shows that relatively higher stellar radius, planet radius, stellar mass, and stellar effective temperatures contribute to habitability negatively, while relatively higher planet orbit semi-major axis contributes to habitability positively.



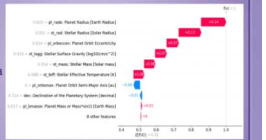
This SHAP **heatmap** plot groups samples (exoplanets) that have the same model predictions and the similar feature impacts together.

SHAP **force** plot for a positive prediction sample (exoplanet)

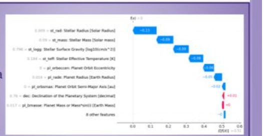
SHAP **force** plot for a negative prediction sample (exoplanet)

More SHAP Analysis

SHAP **waterfall** plot (Random Forest) for a positive prediction sample (exoplanet)



SHAP **waterfall** plot (Random Forest) for a negative prediction sample (exoplanet)



Conclusions

Random Forest and **XGBoost** models were trained to predict exoplanet habitability with high F1 scores at **0.95**.

Feature importance analysis through **SHAP** identified influential stellar and planetary parameters to habitability. Further **SHAP** analysis on those parameters showed that relatively higher stellar radius, stellar mass, stellar effective temperature, and planet radius have negative impacts on habitability, while relatively higher planet orbit semi-major has a positive impact.

This study demonstrated that machine learning techniques could be effective for both predicting exoplanet habitability and understanding how stellar and planetary parameters impact exoplanet habitability.

Future Work

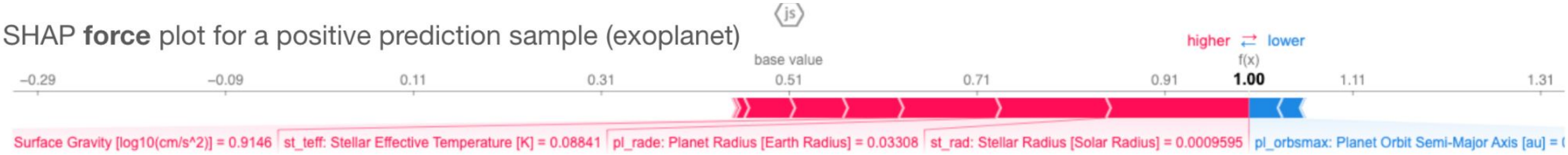
Train a **Neural Network** model for habitability prediction and SHAP analysis. Compare with tree-based models.

Study **Planetary Systems** as a whole to understand what planetary systems might be more likely to host habitable planets.

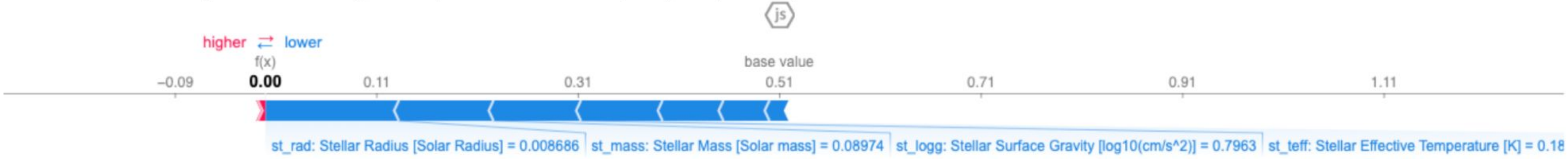
POSTER

new SHAP graphs: force plots

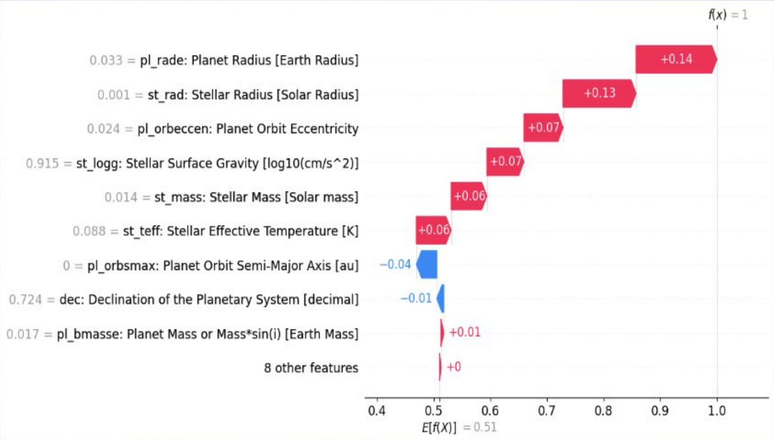
SHAP **force** plot for a positive prediction sample (exoplanet)



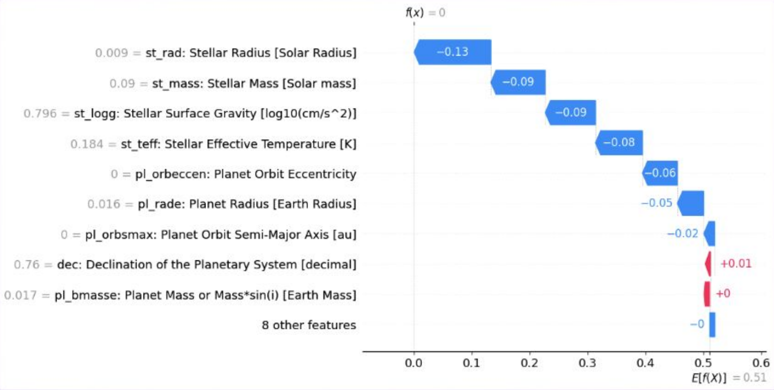
SHAP **force** plot for a negative prediction sample (exoplanet)



SHAP **waterfall** plot (Random Forest) for a positive prediction sample (exoplanet)



SHAP **waterfall** plot (Random Forest) for a negative prediction sample (exoplanet)



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

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.

next week

- learn more about the astrophysical side of how classifications are created (less ML, lots of literary studies)

that's all for this week. :)