An Analysis of Exoplanet Habitability and Most Influential Stellar and Planetary Parameters to Habitability through the Lens of Machine Learning

Christina Liu

INTRODUCTION

Are we alone in this universe? Are there any exoplanets other than Earth where humans are able to thrive? The search of potentially habitable exoplanets has been an active research field in astrophysics throughout the past decade.

As of January 28, 2025, there were **5,834** confirmed exoplanets documented in the **NASA Exoplanet Archive** dataset, each associated with hundreds of parameters. With advancements in the observational capabilities of satellite and telescope based techniques, the number of discovered exoplanets continues to grow.

To identify potential habitable exoplanets among such a large and ever-growing set of candidates, machine learning (ML) model has been increasingly adopted to predict habitability. Furthermore, ML model feature importance analysis techniques such as **SHAP** (**SH**apley **A**dditive ex**P**lanations) provide unique opportunities for studying and identifying stellar and planetary parameters that impact the habitability and how they impact it, which is the focus of this research work.

RESEARCH OBJECTIVES

This research aims to study the influential stellar and planetary parameters to habitability through the lens of machine learning, with the following goals:

- Build high-quality ML models (**Random Forest**, **XGBoost**) to predict exoplanet habitability.
- Conduct feature important analysis via the **SHAP** technique to identify influential stellar and planetary parameters to habitability.
- Perform analysis through **SHAP** to understand how different stellar and planetary parameter values positively or negatively affect the exoplanet habitability.

DATA SOURCES

The primary data sources for this study:

- Planetary Systems Composite Data @ **NASA Exoplanet Archive**: **5,834** confirmed exoplanets as of January 28, 2025.
- Habitable World Catalog (**HWC**), PHL @ UPR Arecibo: **5,599** exoplanets as of January 28, 2025.

The experiment joined data from NASA Exoplanet Archive and HWC. The HWC dataset has a ***P_HABITABLE*** data field, which indicates exoplanet habitability and is used to label training data.

pl_name	hostname	sy_snum	sy_pnum	discoverymethod	disc_year	disc_facility	pl_controv_flag	pl_orbper	pl_orbpererr1	...
11 Com b	11 Com	2	1	Radial Velocity	2007	Xinglong Station	0	323.21000	0.06000	...
11 UMi b	11 UMi	1	1	Radial Velocity	2009	Thueringer Landessternwarte Tautenburg	0	516.21997	3.20000	...
14 And b	14 And	1	1	Radial Velocity	2008	Okayama Astrophysical Observatory	0	186.76000	0.11000	...
14 Her b	14 Her	1	2	Radial Velocity	2002	W. M. Keck Observatory	0	1765.03890	1.67709	...
16 Cyg B b	16 Cyg B	3	1	Radial Velocity	1996	Multiple Observatories	0	798.50000	1.00000	...
17 Sco b	17 Sco	1	1	Radial Velocity	2020	Lick Observatory	0	578.38000	2.01000	...
18 Del b	18 Del	2	1	Radial Velocity	2008	Okayama Astrophysical Observatory	0	982.85000	1.06000	...
1RXS J160929.1-210524 b	1RXS J160929.1-210524	1	1	Imaging	2008	Gemini Observatory	0	NaN	NaN	...
24 Boo b	24 Boo	1	1	Radial Velocity	2018	Okayama Astrophysical Observatory	0	30.33000	0.00000	...

Table extracted from Google Colab notebook showcasing just some of the data fields from the combined NASA Exoplanet Archive + HWC dataset.

MODEL SELECTION

Tree-based machine learning models – specifically **Random Forest** and **XGBoost** – were chosen to build classifiers for predicting exoplanet habitability.

Research in machine learning shows that tree-based models still outperform deep learning models on the tabular dataset. The dataset for this study is entirely tabular based data and therefore tree-based models were well suited for this classification problem.

DATA PROCESSING & MODEL TRAINING

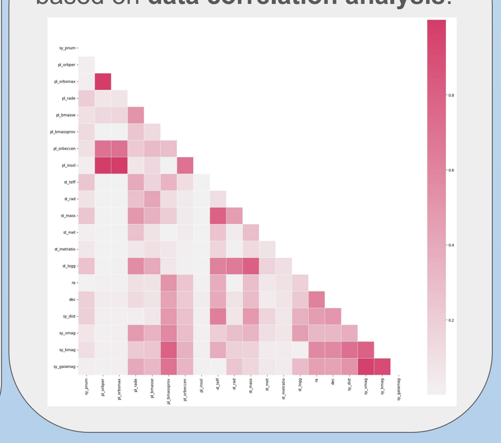
Step 1: Data Downloading

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

Step 2: Data Preprocessing

Data Joining → Data Cleaning → Handling Missing Values → Converting Categorical Features → Handling Data Imbalance → Data Correlation Analysis → Min-Max Scaler

- 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 with questionable status (i.e. where *pl_controv_flag*=1). * Limit to exoplanets with a **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 Multivariate Imputation by Chained Equation (**MICE**).
- Converting Categorical Features: * **Label Encoder** to convert categorical features into numerical values.
- Handling Data Imbalance: * The dataset is highly imbalanced: 4,527 negative samples (non-habitable) and 55 positive samples (habitable). * Apply Synthetic Minority Oversampling Technique (**SMOTE**) and Edited Nearest Neighbors (**ENN**) technique to oversample minorities and undersample majorities.
- Data Correlation Analysis: * Remove highly-correlated data fields based on **data correlation analysis**.



- Min-Max Scaler: Standardize value ranges to **[0,1]**

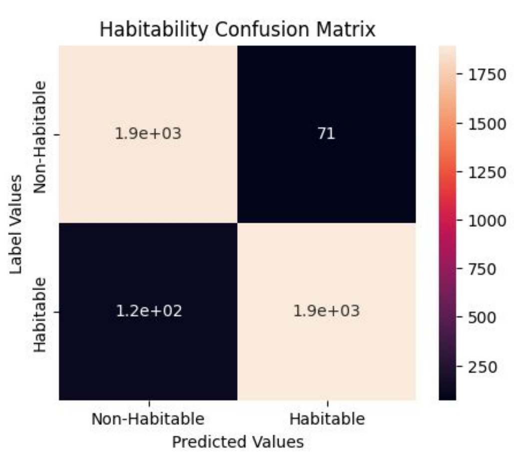
Step 3: Model Training

* Randomly shuffle data, then split into train & test datasets for training & evaluation. Repeat 100 times for each hyperparameter combination. Use average accuracy as the measure of model quality.

* Hyperparameter tuning through exhaustive search.

MODEL EVALUATION

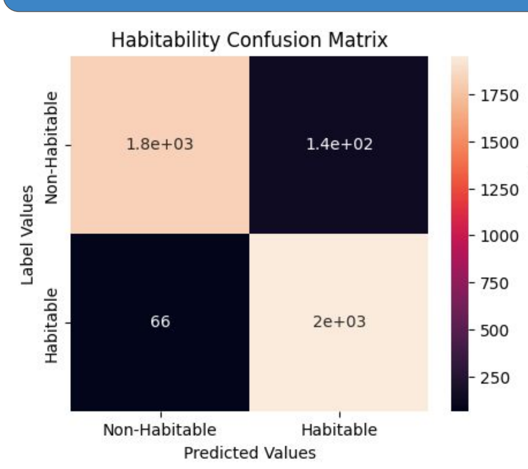
Random Forest Evaluation



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 Evaluation

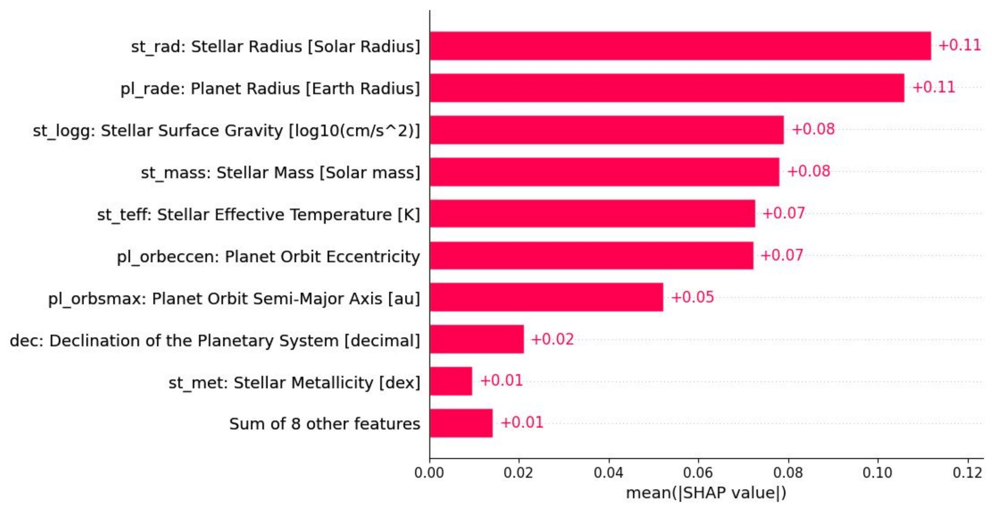


XGBoost Classifier – Classification Report:

	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

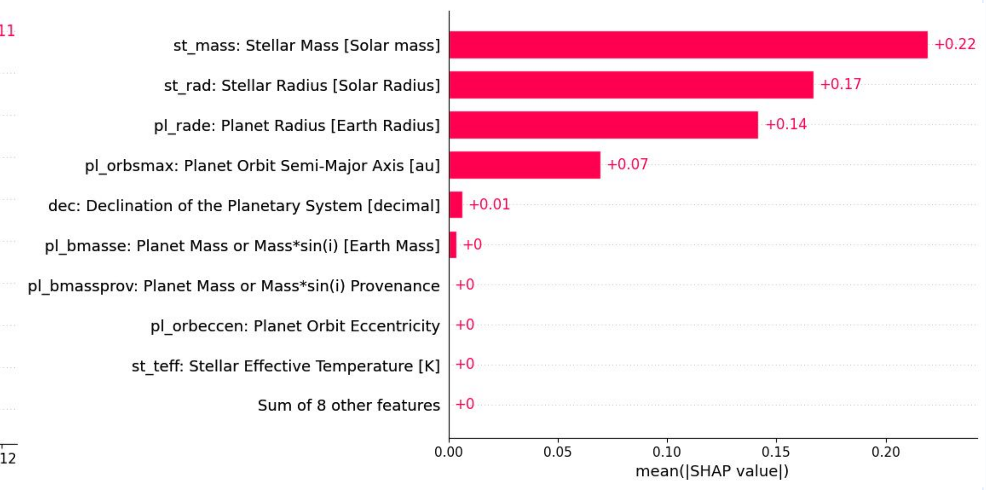
FEATURE IMPORTANCE ANALYSIS

Random Forest Feature Importance by SHAP



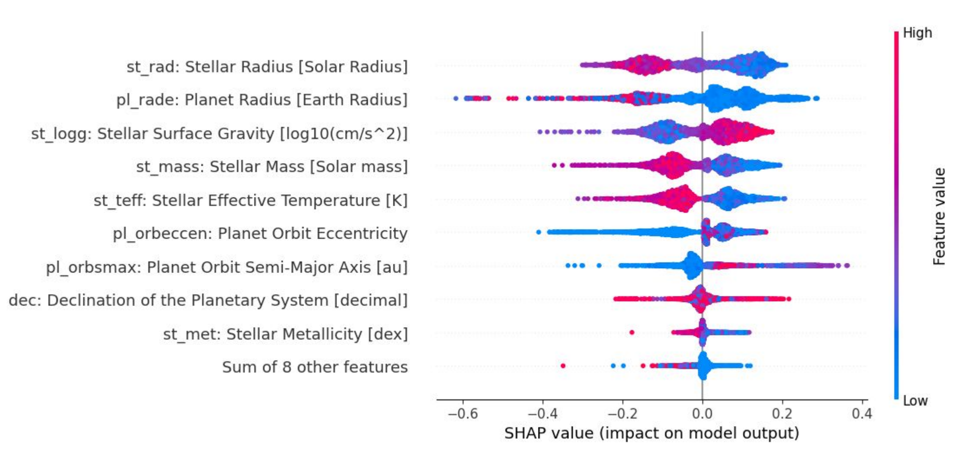
The most influential features of Random Forest classifier: (1) Stellar Radius, (2) Planet Radius, (3) Stellar Surface Gravity, (4) Stellar Mass, (5) Stellar Effective Temperature, (6) Planet Orbit Eccentricity, (7) Planet Orbit Semi-Major Axis.

XGBoost Feature Importance by SHAP



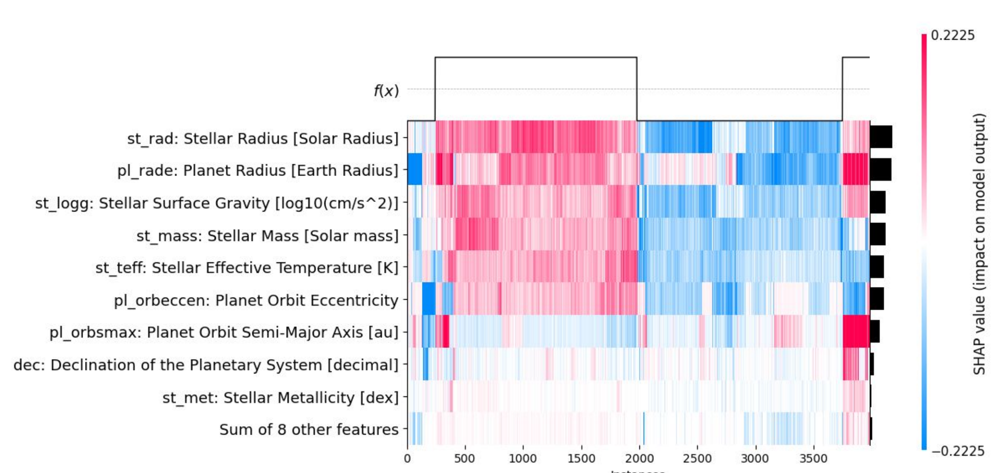
The most influential features of XGBoost classifier: (1) Stellar Mass, (2) Stellar Radius, (3) Planet Radius, (4) Planet Orbit Semi-Major Axis, (5) Declination of the Planetary System.

Deep Dive on Random Forest through SHAP

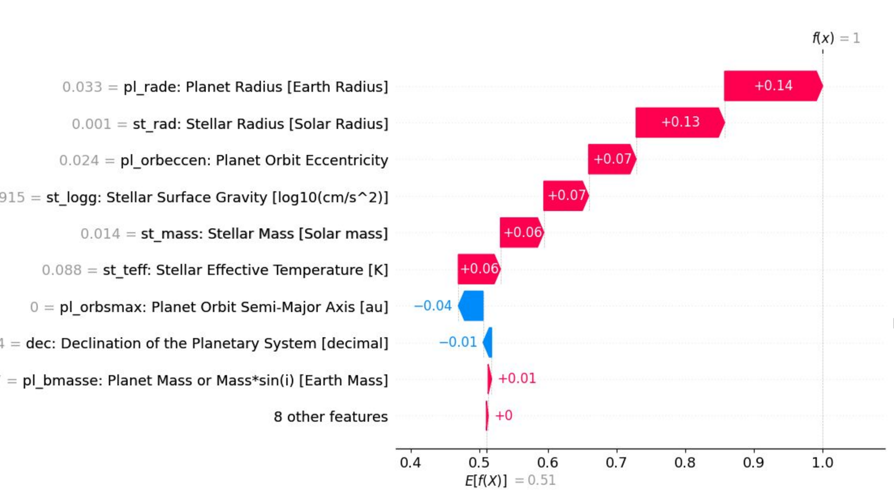


SHAP **beeswarm** plot shows that:

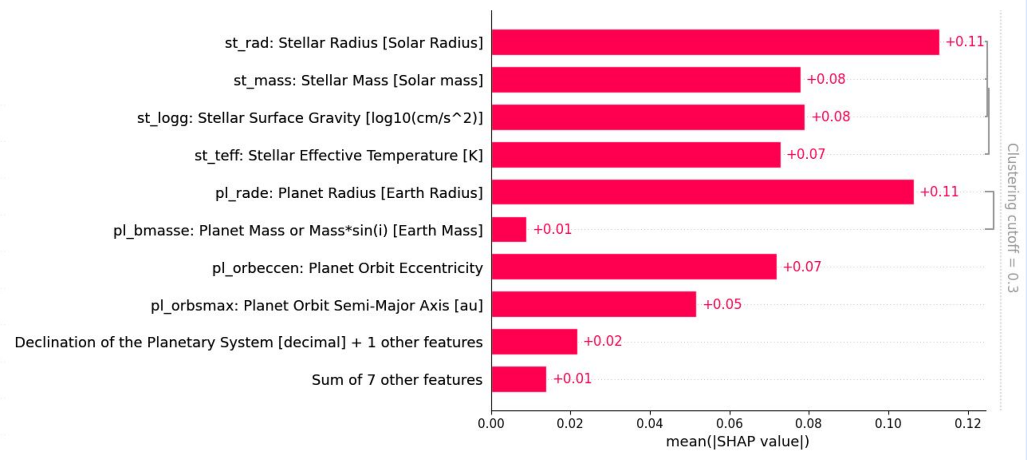
- Relatively higher values in stellar radius, planet radius, stellar mass, and stellar effective temperature lead towards non-habitability, while relatively lower values in those parameters lead towards habitability.
- Relatively higher planet orbit semi-major axis leads towards habitability, while relatively lower value leads towards non-habitability.



SHAP **heatmap** plot groups samples (exoplanets) that have the same model outputs for the same reasons together (e.g., the exoplanets that are predicted to be habitable due to stellar radius, etc.).



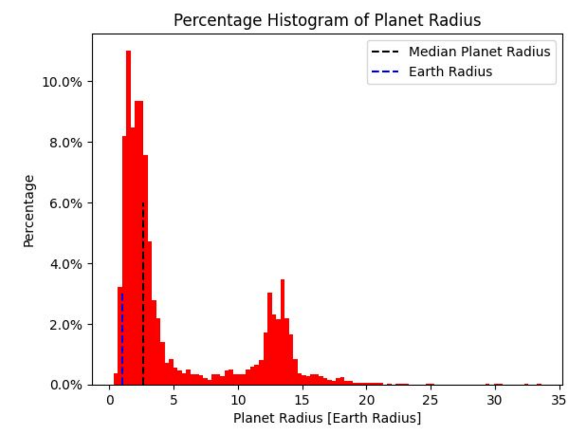
SHAP **waterfall** plot shows how the stellar and planetary parameter values influence habitability for a specific sample (exoplanet).



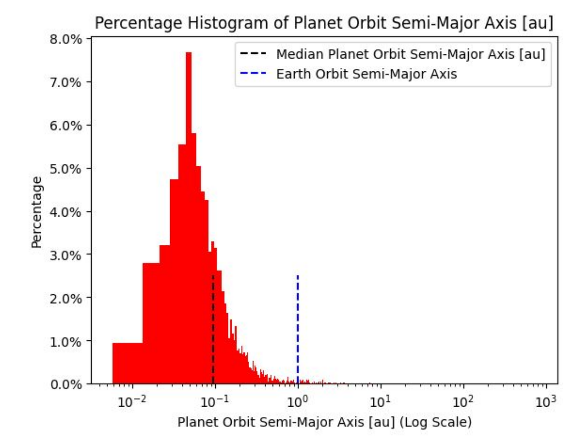
SHAP plot shows correlated features. Stellar radius is correlated with stellar mass. Same goes with planet radius and planet mass.

DISCUSSION

SHAP analysis indicates that relatively larger planet radius leads towards non-habitability. This indeed matches the reality. The below figure shows the percentage histogram of planet radius, with a black vertical line indicating the median and a blue line indicating Earth. As shown in the figure, Earth (habitable) is on the far left of the median, while the exoplanets on the far right are gas-giants and non-habitable.



Similarly, SHAP analysis indicates relatively higher planet orbit semi-major axis leads towards habitability. As shown in the figure, Earth (habitable) is on the far right of the median, while the exoplanets on the far left are too close to their host stars and thus non-habitable.



CONCLUSIONS

A **Random Forest** and **XGBoost** model were trained to predict exoplanet habitability with high accuracy at **0.95**.

Feature importance analysis through **SHAP** identified several influential stellar and planetary parameters to habitability, including stellar radius, stellar mass, stellar effective temperature, planet radius, and planet orbit semi-major axis.

Further analysis by **SHAP** showed that stellar radius, stellar mass, stellar effective temperature, and planet radius have different impacts on habitability than planet orbit semi-major axis. The relatively higher values in stellar radius, stellar mass, stellar effective temperature, and planet radius lead towards habitability while the relatively higher value in planet orbit semi-major axis leads towards non-habitability.

FUTURE WORK

- Train a **Neural Network** model for habitability prediction and feature importance analysis, and compare with tree-based models.
- Study the **Planetary System** as a whole to understand what planetary systems might be more likely to host habitable planets.

KEY REFERENCE

- Seager, Sara. "Exoplanet Habitability." *Science* 340.6132 (2013): 577-581.
- Saha, Snehanshu, et al. "Theoretical validation of potential habitability via analytical and boosted tree methods: An optimistic study on recently discovered exoplanets." *Astronomy and computing* 23 (2018): 141-150.
- Basak, Suryoday, et al. "Habitability classification of exoplanets: a machine learning insight." *The European Physical Journal Special Topics* 230 (2021): 2221-2251.
- Lundberg, Scott. "A unified approach to interpreting model predictions." *arXiv preprint arXiv:1705.07874* (2017).
- Grinsztajn, Léo, Edouard Oyallon, and Gaël Varoquaux. "Why do tree-based models still outperform deep learning on typical tabular data?." *Advances in neural information processing systems* 35 (2022): 507-520.