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

Christina Liu

ABSTRACT

Are we alone in this universe? Are there any exoplanets other than Earth where humans can live? And what are the stellar or planetary characteristics that make an exoplanet more likely to harbor life? The search and understanding of potentially habitable exoplanets beyond our solar system has been one of the most interesting research fields in astrophysics throughout the past decade.

This research studies the exoplanet habitability and the influential stellar and planetary parameters to habitability through the lens of machine learning. A **Random Forest** classifier and an **XGBoost** classifier were trained with high accuracies (both at **0.95**) and feature important analysis was conducted on the ML models to understand the influential features.

INTRODUCTION

As of January 28, 2025, **5,834** confirmed exoplanets were documented in the **NASA Exoplanet Archive** dataset and the numbers continue to grow. To study exoplanet habitability within this ever-growing dataset, researchers have increasingly adopted machine learning, where lots of work have been focusing on building high-quality machine learning models.

This work distinguishes itself by leveraging the feature importance analysis, specifically, **SHAP** (**SH**apley **A**dditive ex**P**lanations) technique, to understand how stellar and planetary parameters influence the habitability.

RESEARCH OBJECTIVES

The research has the following goals:

- Build high-quality tree-based machine learning models (**Random Forest**, **XGBoost**) to predict exoplanet habitability.
- Conduct model feature important analysis through the ML libraries (**sklearn**, **xgboost**) and **SHAP** to identify influential stellar and planetary parameters to habitability.
- Apply analysis through **SHAP** to understand how different stellar and planetary parameter values lead ML models towards positive (habitable) or negative (non-habitable) prediction outcomes.

METHODOLOGY

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

Joining 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 in **single-host star systems**.

* 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**).

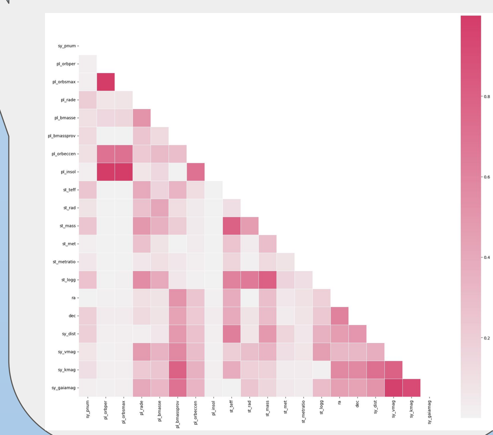
Handling Data Imbalance

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

* We apply a combination of the Synthetic Minority Oversampling Technique (**SMOTE**) and Edited Nearest Neighbors (**ENN**) technique to oversample minorities and undersample majorities.

Data Correlation Analysis

* We 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 training & 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.

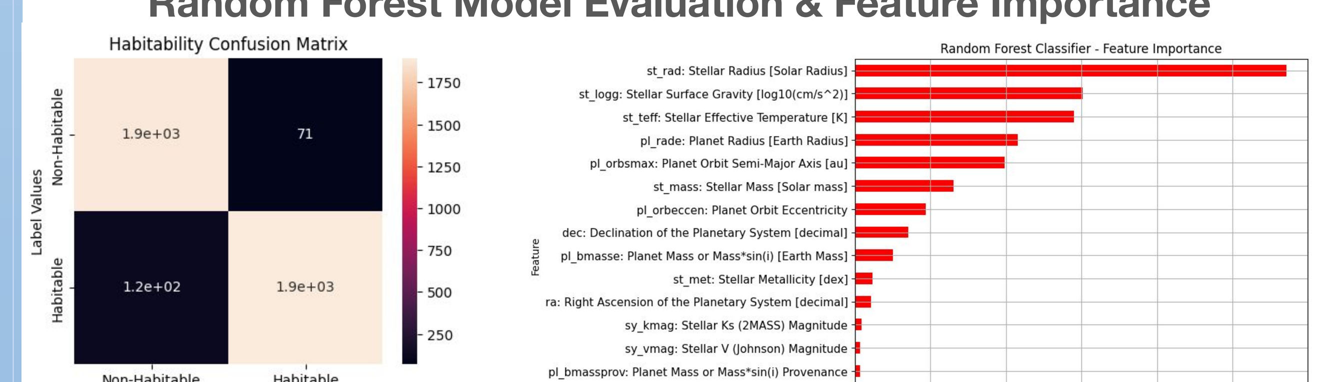
Step 4: Feature Importance Analysis

* Feature importance analysis through **sklearn** & **xgboost** libaries.

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

RESULTS

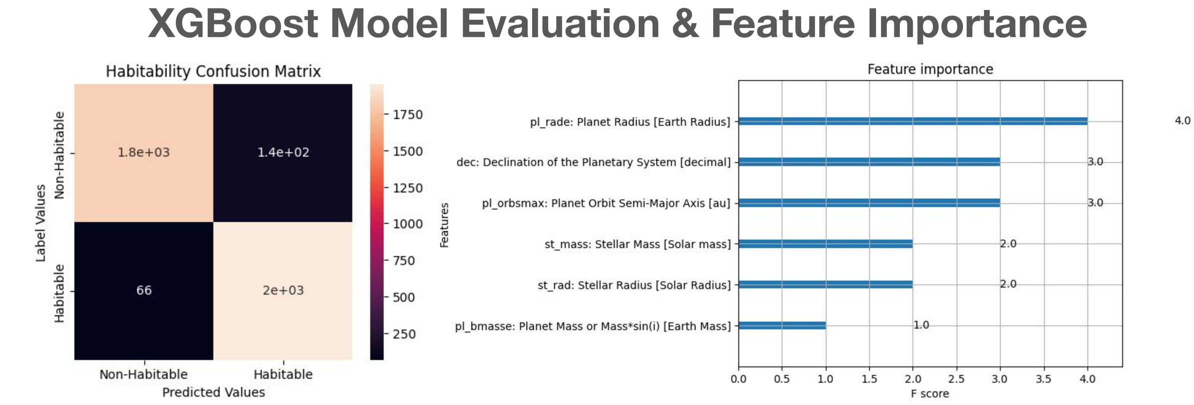
Random Forest Model Evaluation & Feature Importance



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

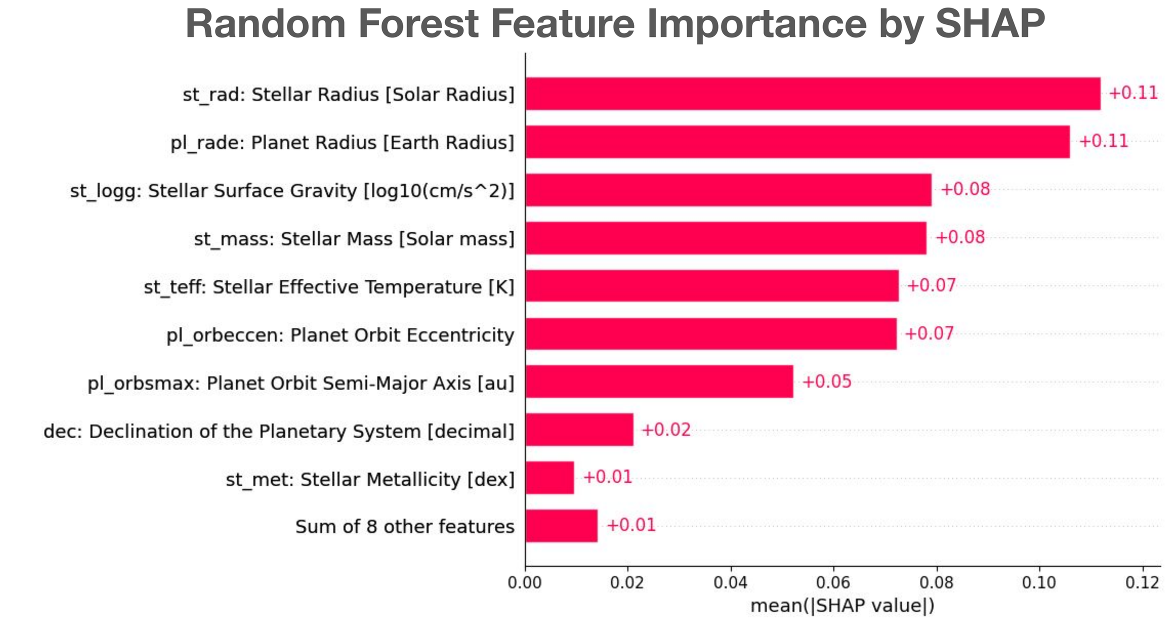
XGBoost Model Evaluation & Feature Importance



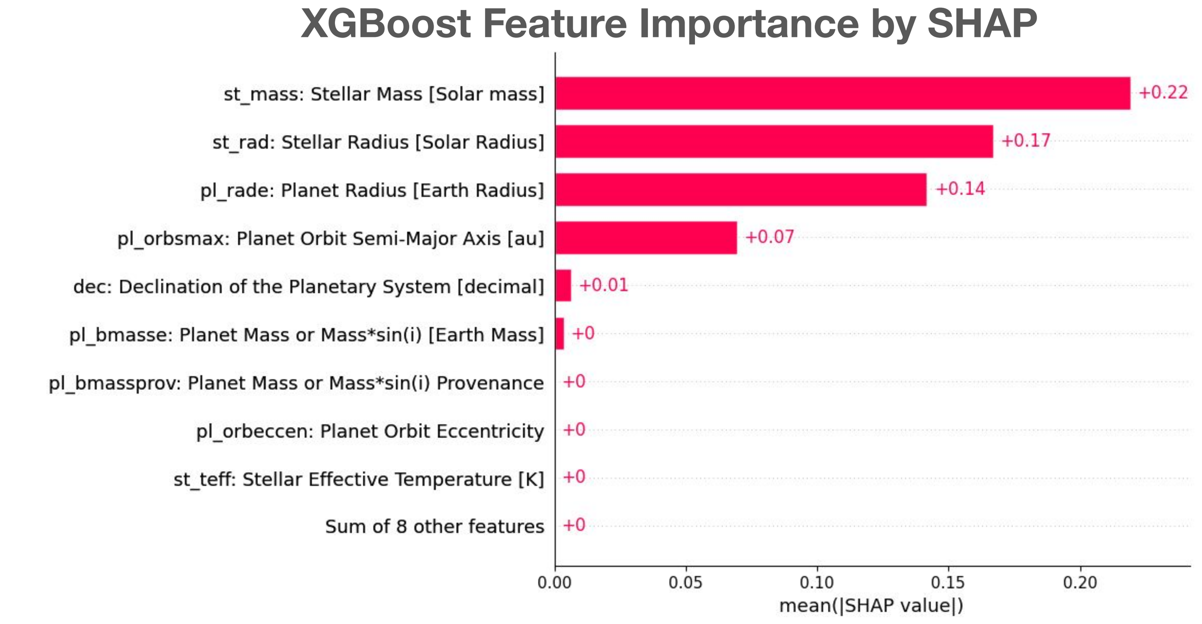
XGBoost Classifier - Classification Report:

	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

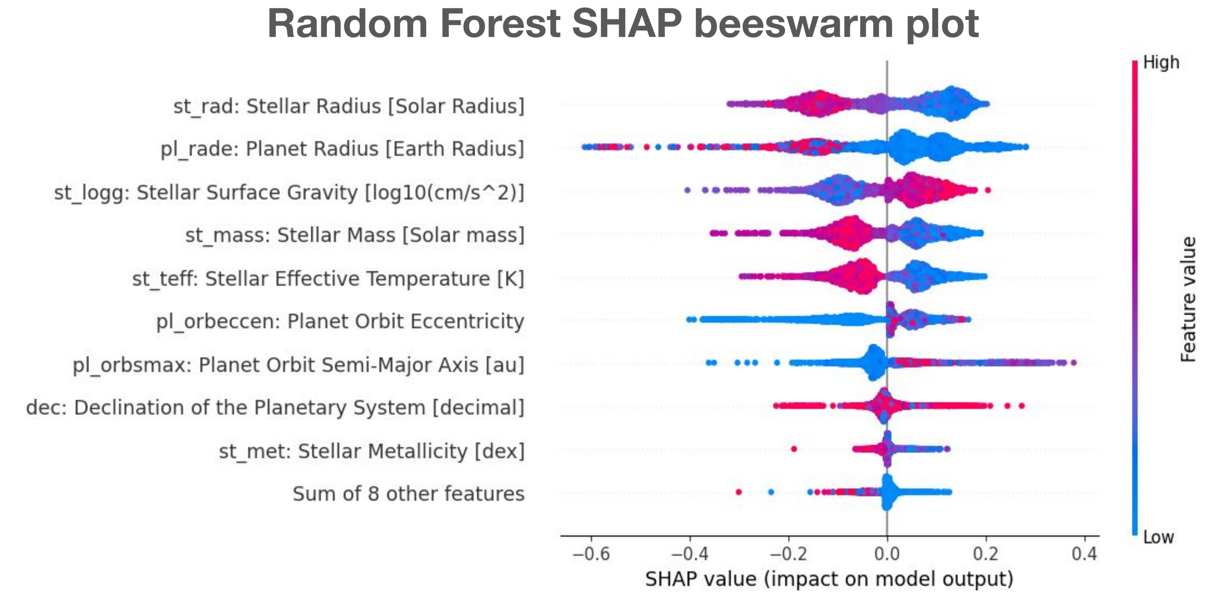
Random Forest Feature Importance by SHAP



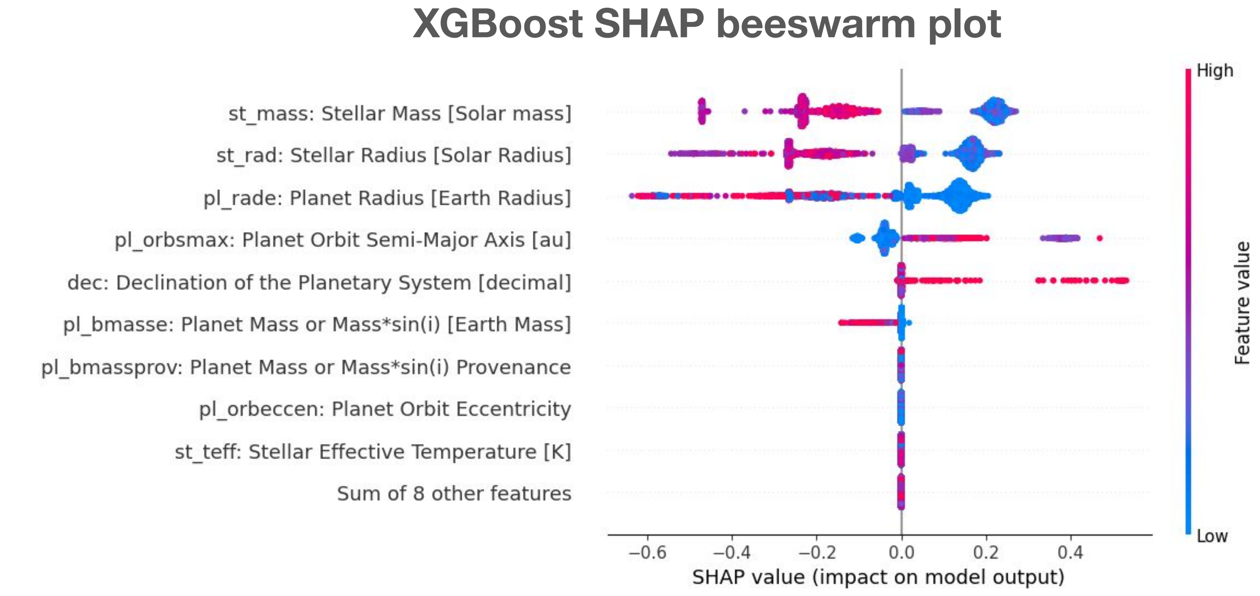
XGBoost Feature Importance by SHAP



Random Forest SHAP beeswarm plot



XGBoost SHAP beeswarm plot

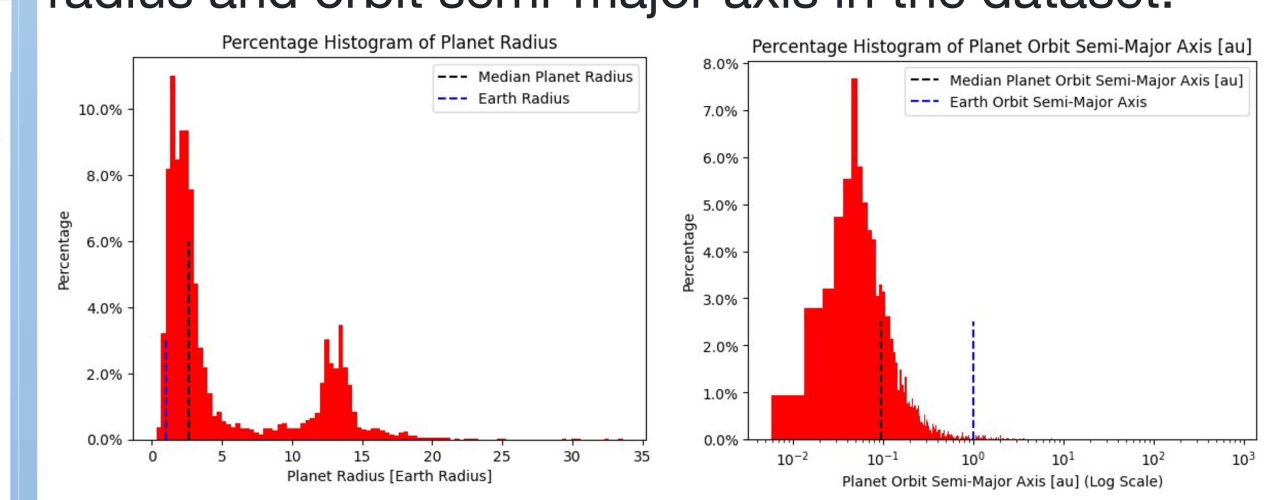


ANALYSIS

Based on the **SHAP beeswarm plot**, one can learn how much each feature contributes to the predictions and towards which outcomes (positive/habitable or negative/non-habitable).

Take the Random Forest's SHAP graph as an example. The beeswarm plot indicates that higher values (relative to other samples in the dataset) in 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.

As a reference, the percentage histogram graphs below show the value distributions of planetary radius and orbit semi-major axis in the dataset.



CONCLUSIONS

A **Random Forest** and **XGBoost** model were trained with high accuracy at **0.95**. Feature important analysis through ML libraries and **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), and how their values impact prediction outcomes. This research sets a good foundation for the further study of exoplanet habitability.

FUTURE WORK

- Understanding the difference in the feature importance analysis results of Random Forest and XGBoost models.
- Train a **Neural Network** model for habitability prediction and feature importance analysis. Research in ML shows that tree-based model in general perform better than deep learning models for tableau dataset.