march 8th, 2025

exoplanet classification

past weeks

- retrained the **Random Forest** and **XGBoost** classifiers based on the latest datasets
- feature importance analysis through **SHAP** (**SH**apley **A**dditive ex**P**lanations)

retrain Random Forest and XGBoost classifiers

- 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 Multivariate Imputation by Chained Equation (**MICE**).

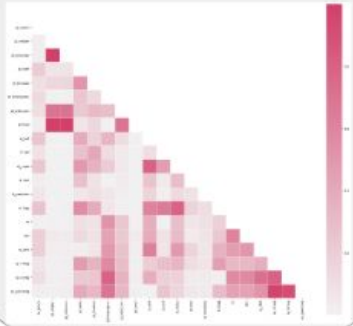
Handling Data Imbalance

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

* Apply combination of Synthetic Minority OVersampling Technique (**SMOTE**) and Edited Nearest Neighbors (**ENN**) technique to oversample minorities and undersample 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 100 times. Use avg accuracy as the measure of model quality for the corresponding hyperparameters.

* Hyperparameter tuning through exhaustive search.

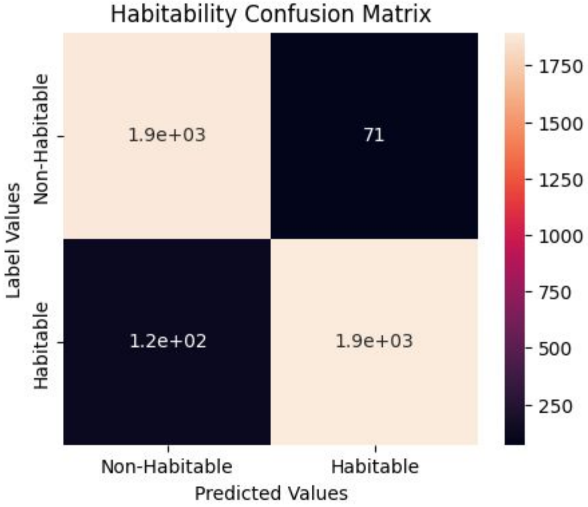
Step 4: Feature Importance Analysis

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

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

Random Forest and XGBoost model performance

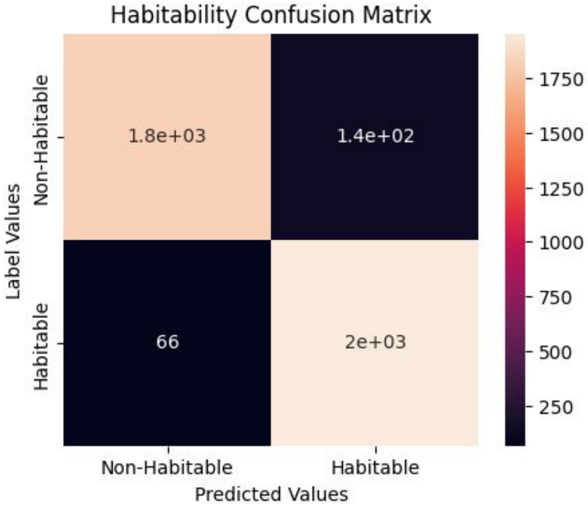
Random Forest classifier



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

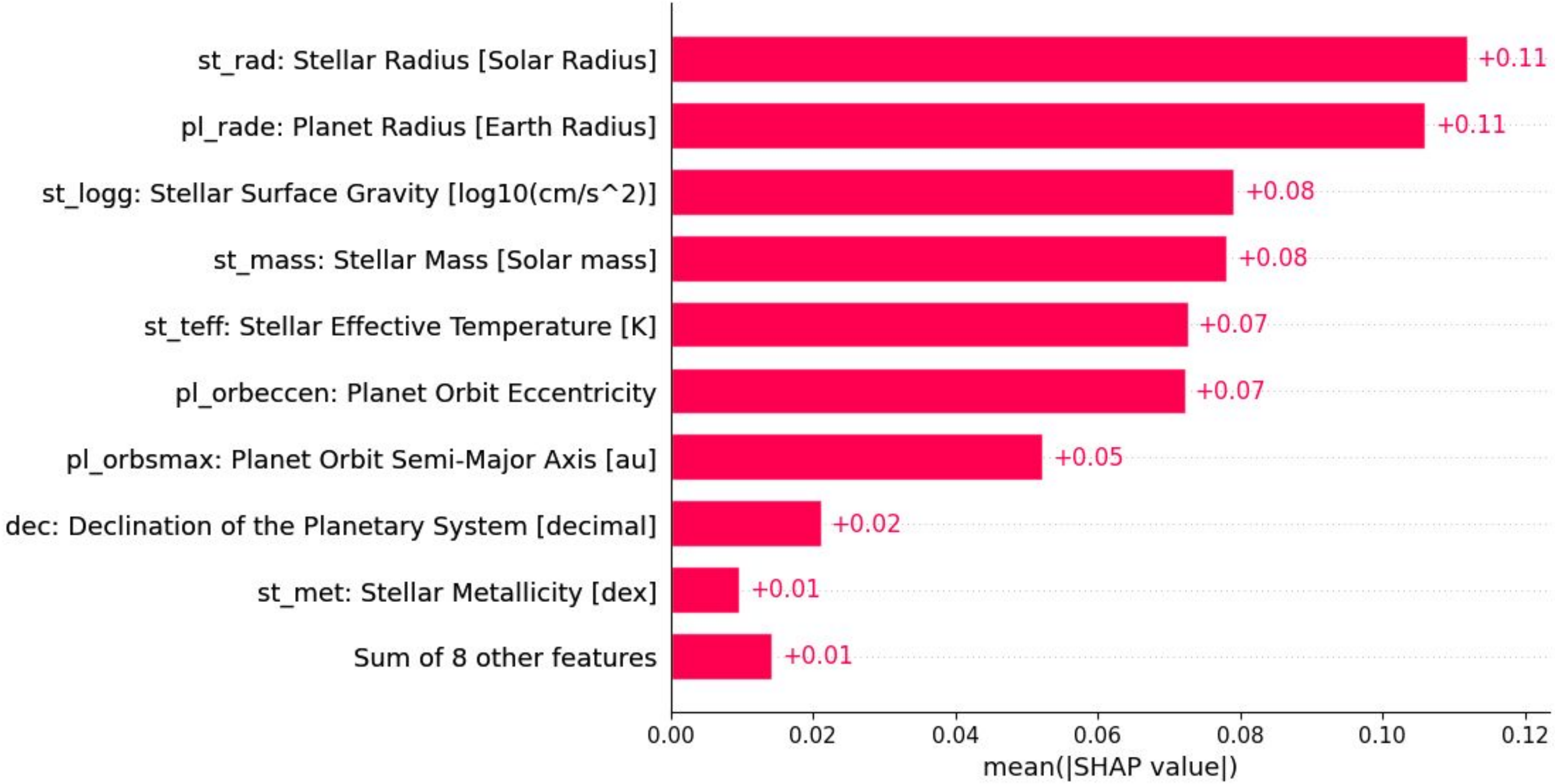
XGBoost classifier



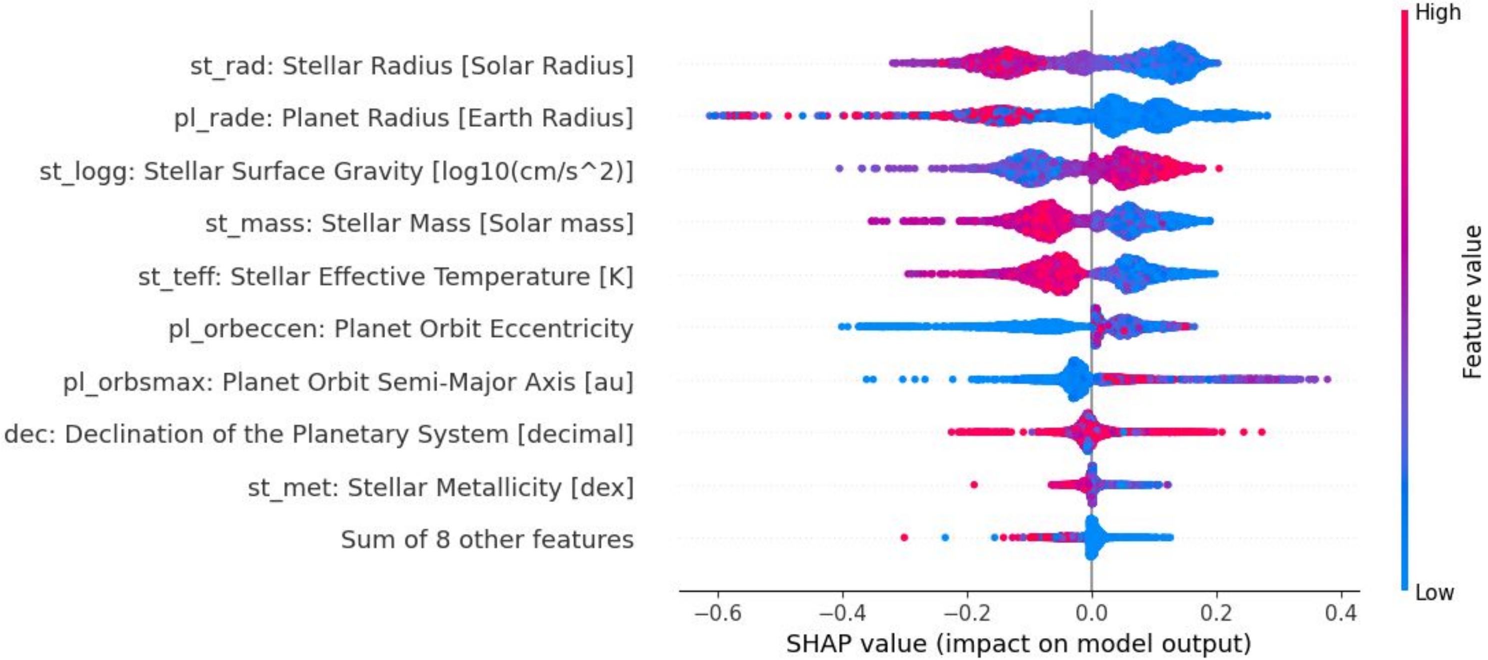
XGBoost Classifier - Classification Report:

	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

Random Forest feature importance via SHAP

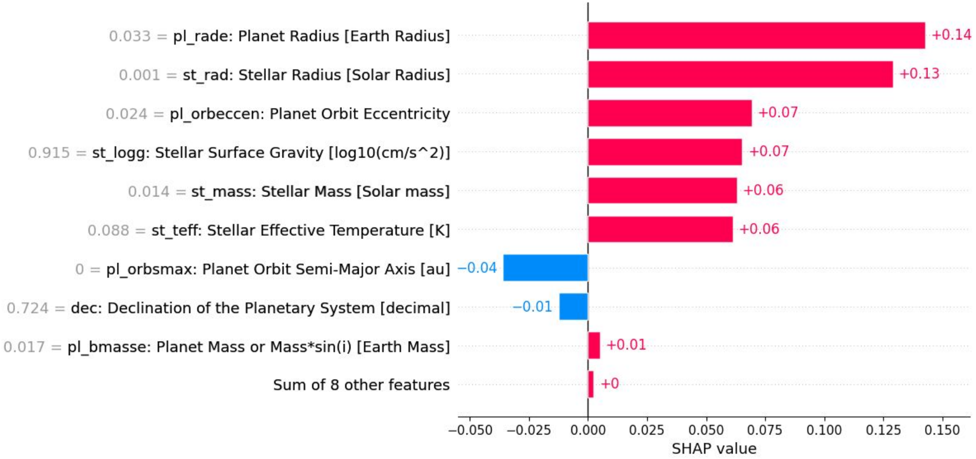


SHAP beeswarm plot - Random Forest

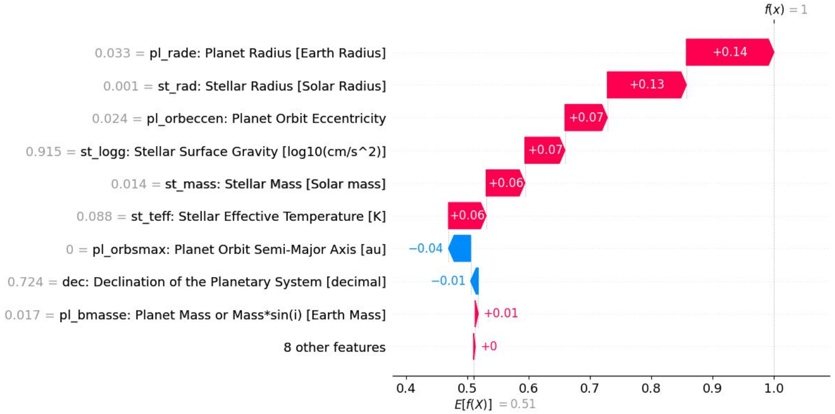


- Higher values (relative to other samples in dataset) of stellar radius, planet radius, stellar mass, and stellar effective temperature lead towards negative predictions, while lower values lead towards positive outcomes.
- Planet orbit semi-major axis, on the other hand, has the opposite impact on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

SHAP local bar and waterfall plots - Random Forest

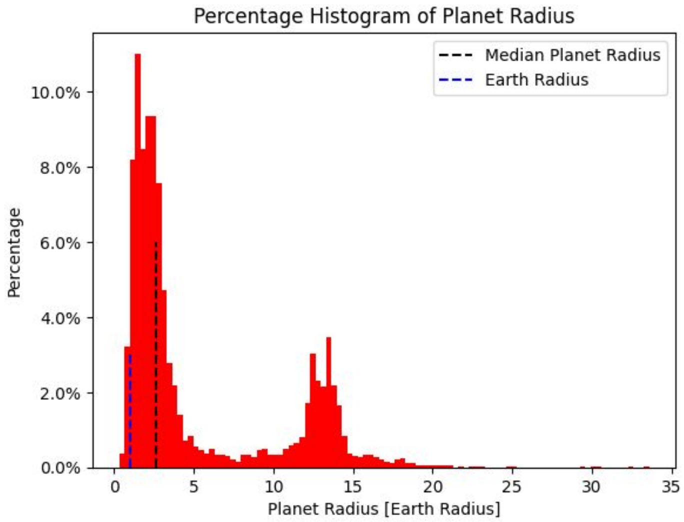


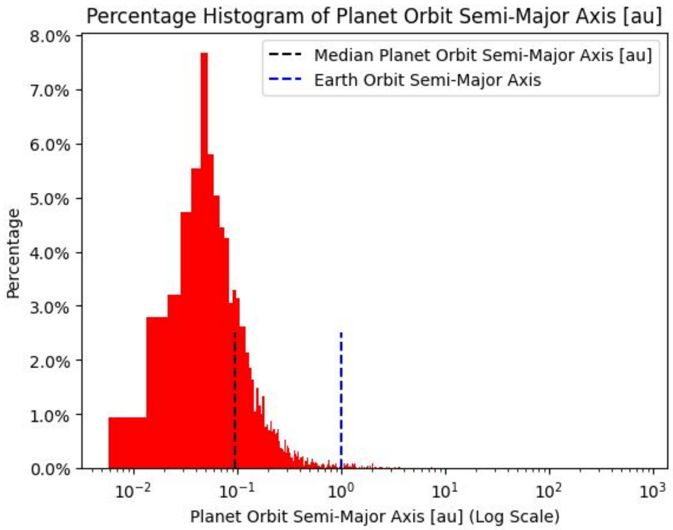
SHAP local bar plot for one sample in the dataset



SHAP waterfall plot for one sample in the dataset

feature influence on predictions - Random Forest





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.