October 21, 2024

exoplanet habitability classification

galaxies paper - ML classifiers

- summary:
  - the following classifiers achieved high accuracy:
    - random forest: **0.95**
    - xgboost: **0.97**
  - feature important analysis on **random forest classifier**:
    - top 4 features in term of feature importance

Feature Name	*Feature Importance*
stellar radius [solar radius]	**0.274373**
stellar effective temperature [k]	**0.212329**
*stellar surface gravity [log10(cm/s**2)]*	0.134561
planet orbit semi-major axis [au]	**0.131208**

    - our formula captures the most important features (**3** out of top 4)

$$T_{surf,ave} = kT_{\odot}(1-A)^{0.25}(R_{\odot}/(2d))^{0.5}$$

galaxies paper - training data processing

- join NASA 03-10-2024 data with HWC data from PHL.
- HWC data has a "*P_HABITABLE*" data field that can be used as label
- training data preprocessing:
  - remove data fields that are not relevant to training
  - drop data fields with too much missing values
  - for categorical data fields:
    - filling missing values with mode
    - encode with LabelEncoder
  - for numeric data fields:
    - filling missing values with MICE imputation
  - use SMOTEENN to oversample and downsample to overcome sample imbalances

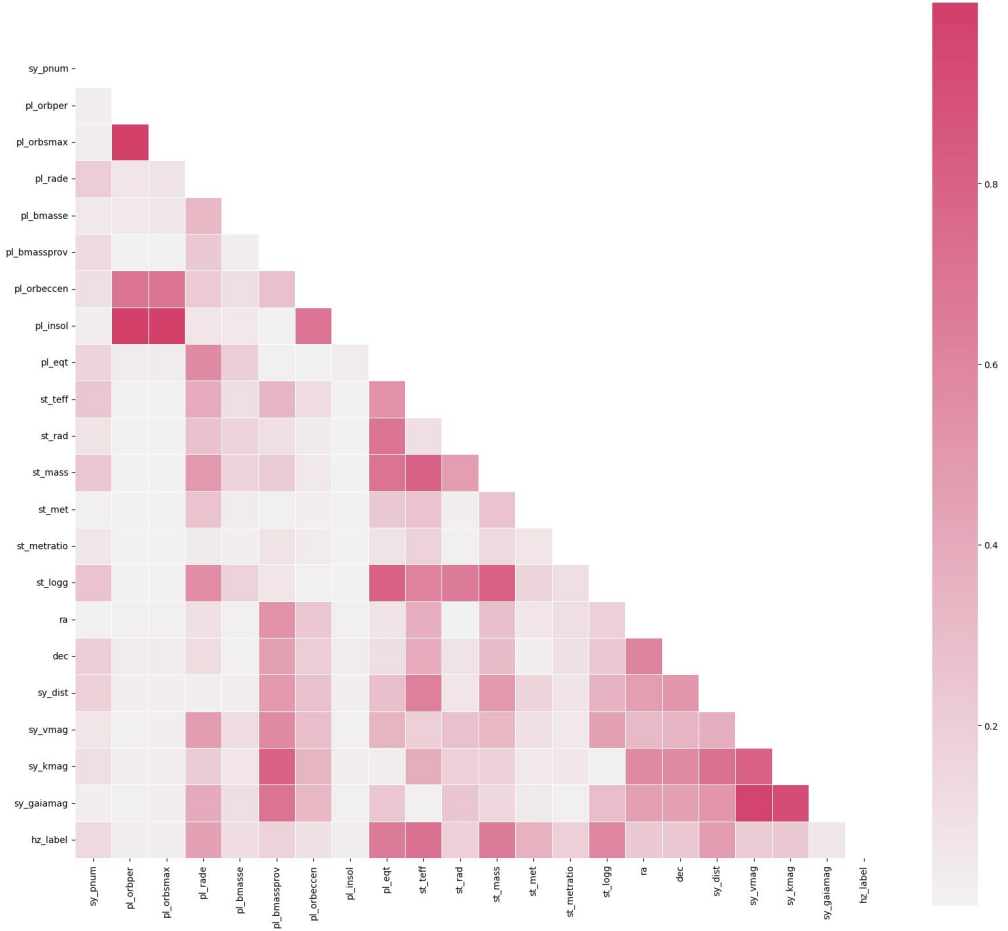
```
hz_label=0, count=4520 (98.798%)
hz_label=1, count=55 (1.202%)
```

galaxies paper - feature correlation analysis

correlation analysis.
remove highly
correlated features:

- pl_orbeccen
- pl_insol
- sy_gaiamag

end up with 17 features in
the training data



galaxies paper - training features

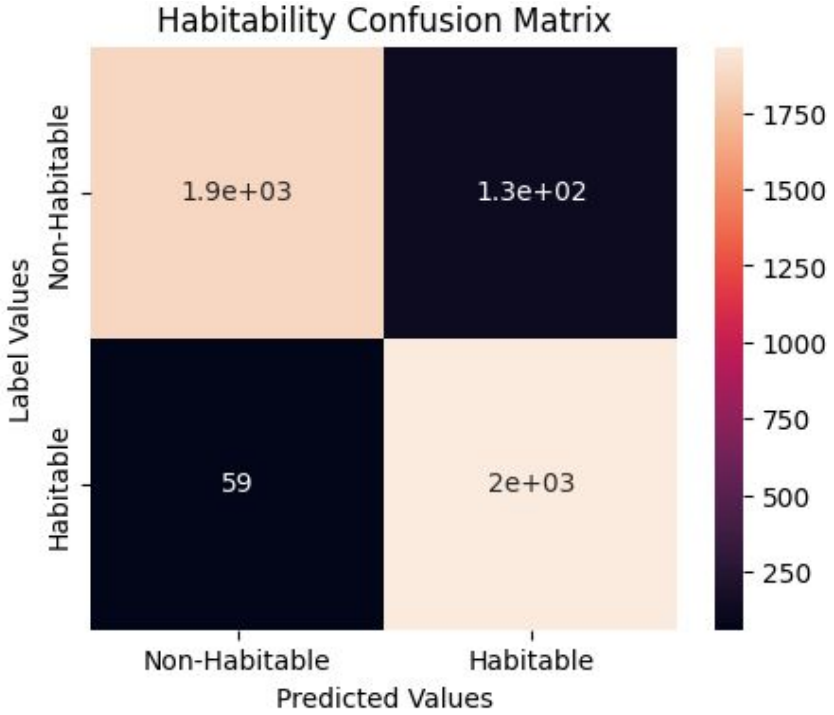
```
Data columns (total 17 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   sy_pnum       8924 non-null   int64
 1   pl_orbper     8924 non-null   float64
 2   pl_orbsmax    8924 non-null   float64
 3   pl_rade       8924 non-null   float64
 4   pl_bmasse     8924 non-null   float64
 5   pl_bmassprov  8924 non-null   int64
 6   st_teff       8924 non-null   float64
 7   st_rad        8924 non-null   float64
 8   st_mass       8924 non-null   float64
 9   st_met        8924 non-null   float64
 10  st_metratio   8924 non-null   int64
 11  st_logg       8924 non-null   float64
 12  ra            8924 non-null   float64
 13  dec           8924 non-null   float64
 14  sy_dist       8924 non-null   float64
 15  sy_vmag       8924 non-null   float64
 16  sy_kmag       8924 non-null   float64
```

galaxies paper - random forest classifier

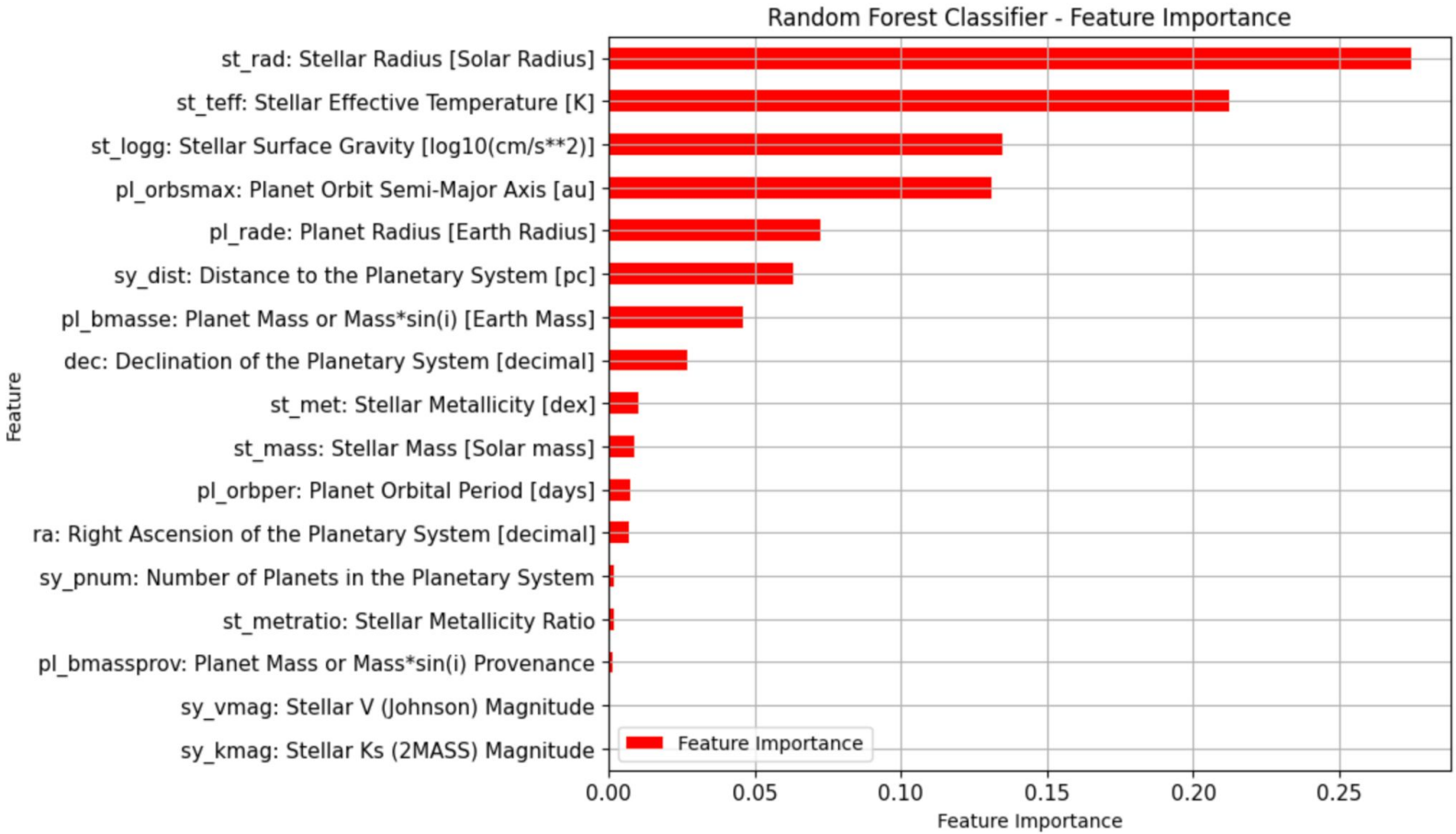
```
RandomForestClassifier(n_estimators = 6,
                       criterion = 'gini',
                       max_depth = 7,
                       max_features = 'log2',
                       max_leaf_nodes = 11,
                       random_state = 0)
```

Random Forest Classifier - Classification

	precision	recall	f1-score
0	0.97	0.94	0.95
1	0.94	0.97	0.96
accuracy			0.95



random forest classifier - feature importance



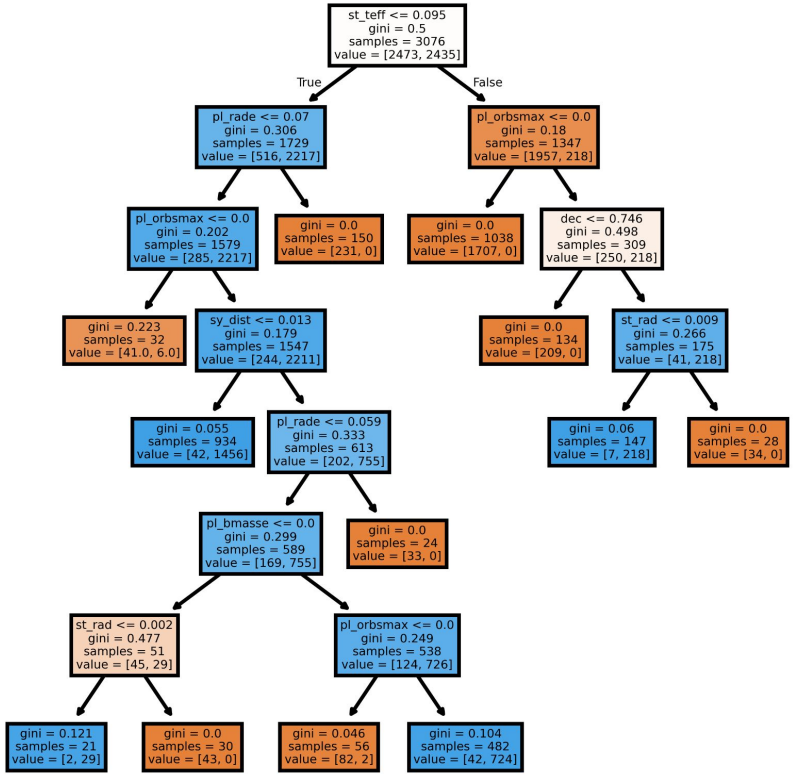
random forest classifier - feature importance

	Feature Importance
st_rad: Stellar Radius [Solar Radius]	0.274462
st_teff: Stellar Effective Temperature [K]	0.212594
st_logg: Stellar Surface Gravity [log10(cm/s**2)]	0.136978
pl_orbsmax: Planet Orbit Semi-Major Axis [au]	0.132699
pl_rade: Planet Radius [Earth Radius]	0.073394
sy_dist: Distance to the Planetary System [pc]	0.062292
pl_bmasse: Planet Mass or Mass*sin(i) [Earth Mass]	0.043432
dec: Declination of the Planetary System [decimal]	0.027246
st_met: Stellar Metallicity [dex]	0.010323
st_mass: Stellar Mass [Solar mass]	0.008836
pl_orbper: Planet Orbital Period [days]	0.007751
ra: Right Ascension of the Planetary System [decimal]	0.004763
sy_pnum: Number of Planets in the Planetary System	0.002106
st_metratio: Stellar Metallicity Ratio	0.001859
pl_bmassprov: Planet Mass or Mass*sin(i) Provenance	0.001267
sy_vmag: Stellar V (Johnson) Magnitude	0.000000
sy_kmag: Stellar Ks (2MASS) Magnitude	0.000000

random forest classifier - one of decision trees

random forest classifier contains 6 decision trees.

the diagram on the right is visualization of one of those 6 decision trees.

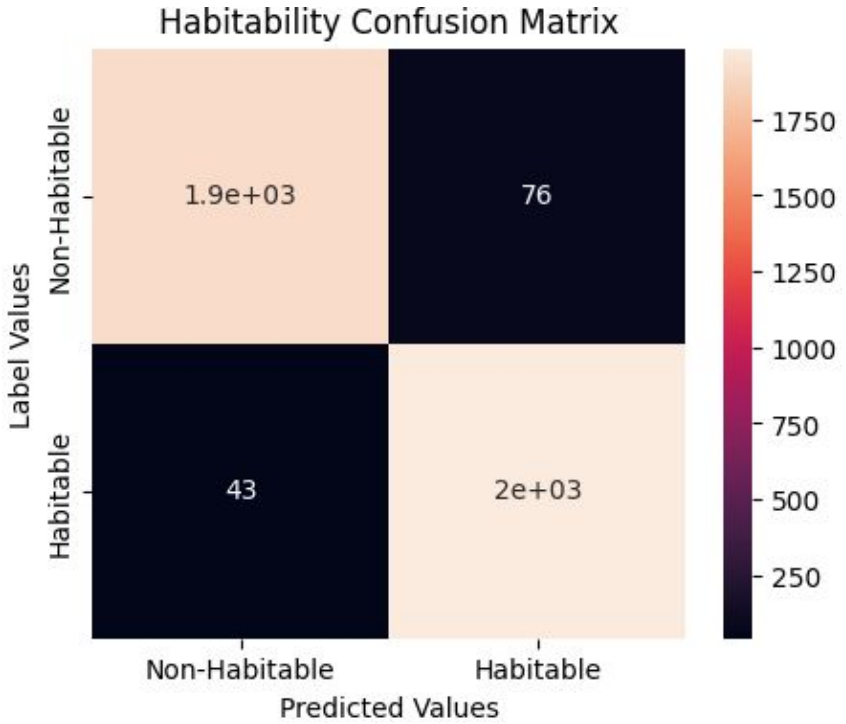


galaxies paper - xgboost classifier

```
XGBClassifier(n_estimators=8,
              max_depth=5,
              learning_rate=1,
              objective='binary:logistic',
              eval_metric='logloss',
              max_leaves=11)
```

XGBoost Classifier - Classification Report:

	precision	recall	f1-score
0	0.98	0.96	0.97
1	0.96	0.98	0.97
accuracy			0.97



galaxies paper - knn classifier (not good enough)

```
KNeighborsClassifier(n_neighbors=5,
                     weights='distance',
                     algorithm='auto',
                     leaf_size=5)
```

KNN Classifier – Classification Report:

	precision	recall	f1-score
0	0.70	0.98	0.82
1	0.98	0.59	0.73
accuracy			0.78

not good enough, suggest not mention it in the paper.