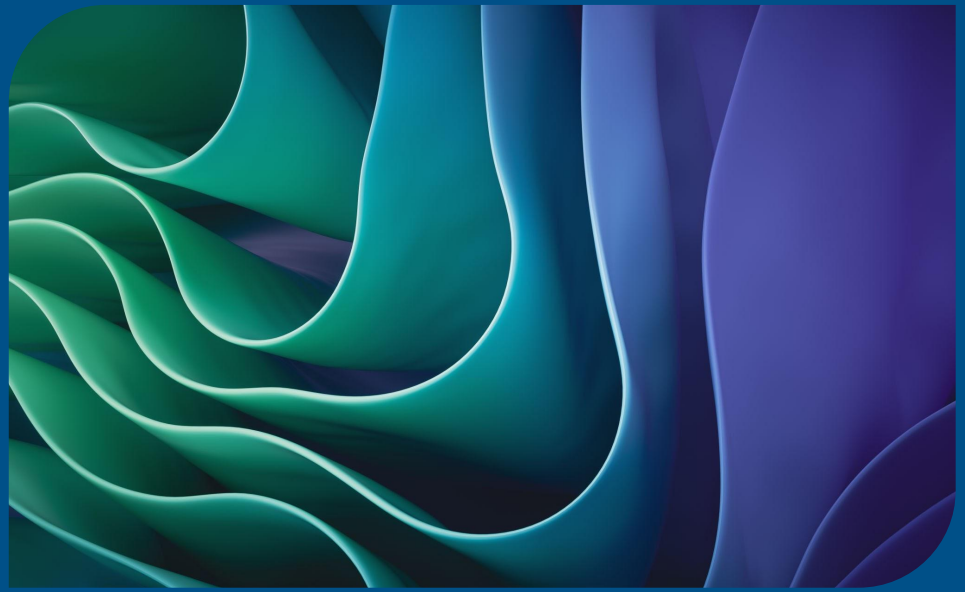
Search for Life

By Brannon Smith



- *Executive Summary: The purpose of this project is to build a model that predicts whether or not an exoplanet is able to support life based on the features provided in the dataset provided by NASA's Exoplanet Archive.*

- *MVP: A supervised or unsupervised model that can correctly predict whether an exoplanet can potentially support life or not.*

- *Data Question: Is it possible to create a model based on top features from the dataset provided by NASA's Exoplanet Archive that can correctly predict whether an exoplanet can potentially support life or not?*

- Data Sources: *Planetary Systems Composite Data*, *PHL @ UPR Arecibo - Habitable Worlds Catalog*

- *Potential Issues and Challenges: The lack of data relative to the sets of data we have usually worked with in this course (100,000 or more rows vs. the little more than 5,000 rows in the exoplanets dataset.*
- *Also within the small amount of data available, any NaN values which further complicates any model making accurate predictions .*

- *Target Variable: Binary variable based on ranges of variables that the exoplanets dataset has in common with the Habitable Worlds Catalog*

Definitions and Determining Features

Exoplanet – A planet that orbits a star outside of our solar system, or a planet that exists in another star system. Exoplanet is short for "extrasolar planet".

Discovery Method - Astronomical method used to discover exoplanet. The two main methods listed in the Habitable Planets Catalog are Radial Velocity and Transit.

Transit Method - Works by observing a star's brightness and detecting the temporary dips in brightness caused by a planet passing in front of it, creating a "mini-eclipse" called a transit.

Radial Velocity Method - Also known as Doppler spectroscopy, the Radial Velocity Method is a technique used to detect exoplanets by measuring the wobble of a star caused by the gravitational pull of an orbiting planet.

Mass - Mass of the planet in Earth masses (Earth = 1.0 ME). Habitable Range = (0.39, 3.19).

Radius - Radius of the planet in Earth radii (Earth = 1.0 RE). Habitable Range = (0.92, 1.60).

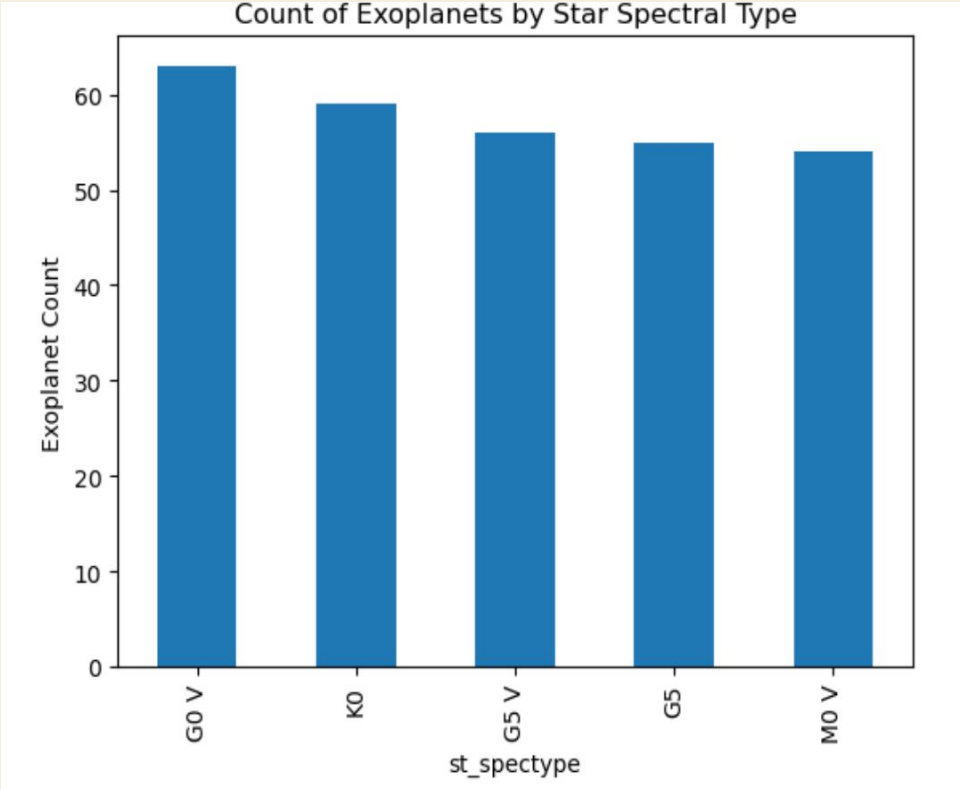
Flux - Average stellar flux of the planet in Earth fluxes (Earth = 1.0 SE). Habitable Range = (0.25, 1.48).

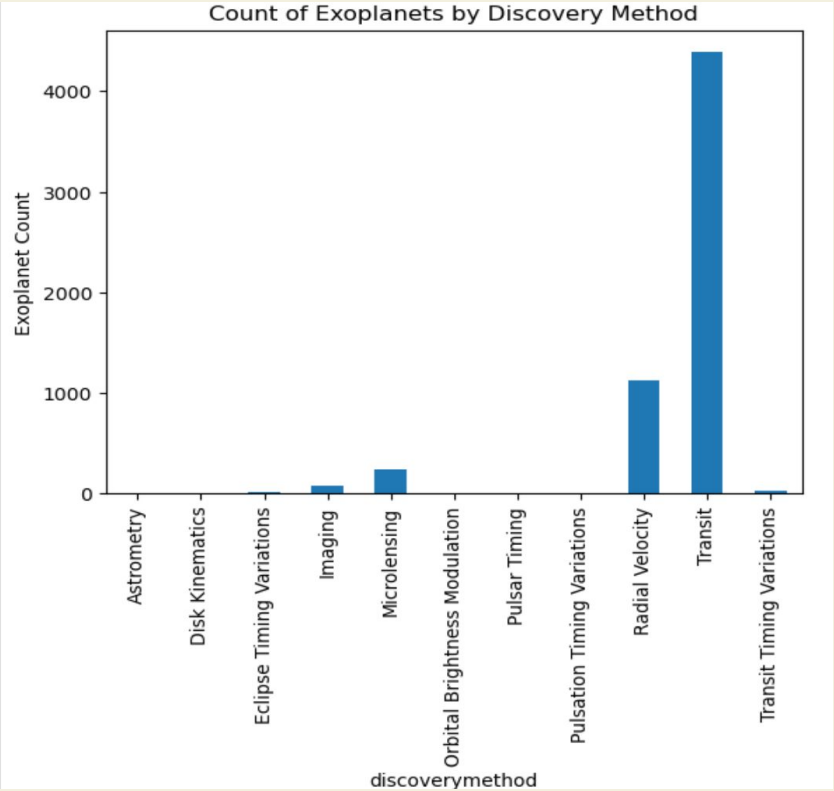
Tsurf - The estimated surface temperature in Kelvins (K) assuming an Earth-like atmosphere. Habitable Range = (203, 316).

Period - Orbital period in days (Earth = 365 days). Habitable Range = (4.05, 267).

Distance - Distance from Earth in light-years (ly). Habitable Range = (4.2, 1193).

EDA





As you can see above, the distribution of exoplanets by Star Spectral Type is fairly even. Star Spectral Type categorizes stars based on temperature (O being the hottest, M being the coolest), further temperature subdivision for each letter (0 being the hottest, 9 the coolest), and size/brightness (e.g. V for main sequence stars, III for giants). This distribution is in sharp contrast to exoplanet count by Discovery Method, with Transit Method Discoveries and Radial Velocity Method Discoveries greatly outnumbering other method discovery counts.

Models Used and Results

Gradient Boosting Classifier Results

```
Accuracy: 0.9977477477477478
MCC: 0.4467094086480023
[[1771    0]
 [   4    1]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      1771
           1       1.00      0.20      0.33         5

    accuracy                           1.00      1776
   macro avg       1.00      0.60      0.67      1776
weighted avg       1.00      1.00      1.00      1776
```

Logistic Regression Classifier Results

```
Accuracy: 0.972972972972973
MCC: 0.1856870506310548
[[1725   46]
 [   2    3]]
              precision    recall  f1-score   support

           0       1.00      0.97      0.99      1771
           1       0.06      0.60      0.11         5

    accuracy                           0.97      1776
   macro avg       0.53      0.79      0.55      1776
weighted avg       1.00      0.97      0.98      1776
```

MLP Classifier Results

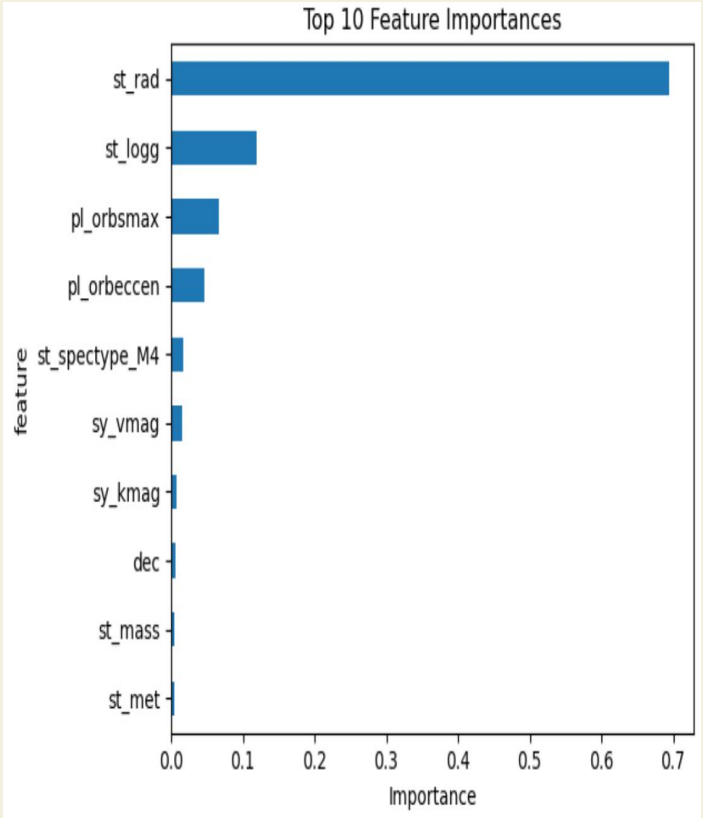
```
Accuracy: 0.9960585585585585
MCC: 0.22165032598869927
[[1768    3]
 [   4    1]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      1771
           1       0.25      0.20      0.22         5

    accuracy                           1.00      1776
   macro avg       0.62      0.60      0.61      1776
weighted avg       1.00      1.00      1.00      1776
```

The models I tried out to predict exoplanet habitability are Logistic Regression, Gradient Boosting Classifier, and MLP Classifier. I also used Randomized Search Cross Validation to get the best possible parameters for each model. Looking at these resulting metrics, even though Logistic Regression predicted the most true positives or habitable planets, it also has the worst f1 score, which is the biggest indicator of a model's effectiveness. Therefore, even though the Gradient Boosting Classifier Model only correctly predicted one potential habitable planet, since it has the best f1 score, it is the best model out of the three, followed by the MLP Classifier Model and then the Logistic Regression Model.

Feature Importance

As you can see in the table and graph to the right, st_rad or stellar radius, or basically the radius of the star at the center of the star system of the exoplanet in question, is many times more powerful as a variable in determining if an exoplanet is habitable than the other top ten variables. Star radius is directly proportional to the size of the star, therefore according to these results, the bigger the central star, the more likely the exoplanet associated with that star is habitable

	feature	importance
5	st_rad	0.693745
8	st_logg	0.118957
2	pl_orbsmax	0.066508
3	pl_orbeccen	0.046616
252	st_spectype_M4	0.016678
11	sy_vmag	0.015066
12	sy_kmag	0.007495
10	dec	0.006246
6	st_mass	0.005228
7	st_met	0.003916



Conclusions

- All three models that I used predicted extremely few exoplanets as habitable correctly, which agrees with the expectation I had about the fact that, out of all the known exoplanets, the conditions required on an exoplanet for life results in very few of those exoplanets being candidates for life, potentially proving that extraterrestrial life, if it exists, is extremely rare in the universe. As a matter of fact, the ranges of values of the variables recorded in the Habitable World's Catalog are extremely narrow compared to the ranges of values of those same variables in the exoplanets dataset.

- The exoplanet correctly predicted by the Gradient Boosting Model is called TOI-700 d. This planet is a rocky, Earth-sized exoplanet that orbits a small, cool M dwarf star in the Dorado constellation, and is about 101.4 light-years from Earth. It's the outermost of four confirmed exoplanets exoplanets in the TOI-700 star system. It is about 1.2 times the size of Earth, with a mass 1.25 times Earth's, and takes 37.4 days to orbit it's central star.

- Having correctly predicted that this planet has the proper conditions to potentially be habitable, this does not conclusively prove that this planet actually does have some form of life on it.

Stretch Goals

- Continue trying to find more effective machine learning models with better f1 scores. The Gradient Boosting Classifier Model had an fl score of 0.33, which leaves much to be desired of a proper predictive model.
- More discovered exoplanets or observations. Even though this factor is more outside of what I am able to do to improve predictions, the more data that is available, the better any machine learning model is able to correctly predict an exoplanet as being potentially habitable.
- Drop any variables with potential correlated coefficients.
- Incorporate any other preprocessing techniques.
- Research and try to find if there is any data available of other variables or features not included in the exoplanets dataset that is purely theoretical for now. What I mean by this is that the data in the dataset is based on features being able to support biological life on Earth, but there could potentially be life out there that requires totally different features from what life on Earth requires.
- Try to find other sets of data of potentially habitable planets like the habitable worlds catalog that includes other features not included in the catalog that the exoplanets dataset includes as well.
- Out of all the models I will try out, see what correctly predicted habitable exoplanets that all or most of the models have in common.

Questions?