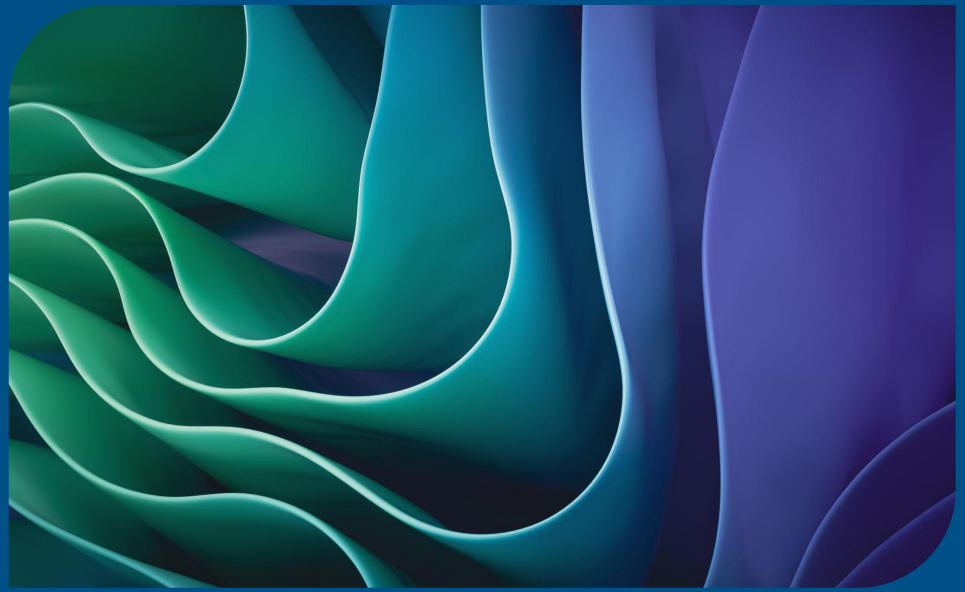
Search for Life

By Brannon Smith



Project Goals

- ***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 by NASA's Exoplanet Archive.***

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

Sources

- **Source of Main Dataframe: Planetary Systems Composite Data**

- **Source of Habitable Worlds Catalog:** *PHL @ UPR Arecibo - Habitable Worlds Catalog*

Features in Common

- **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 - Observes a star's brightness, detects temporary dips in brightness caused by a planet passing in front of it, creates a "mini-eclipse" called a transit.
  Radial Velocity Method - Detects 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).**

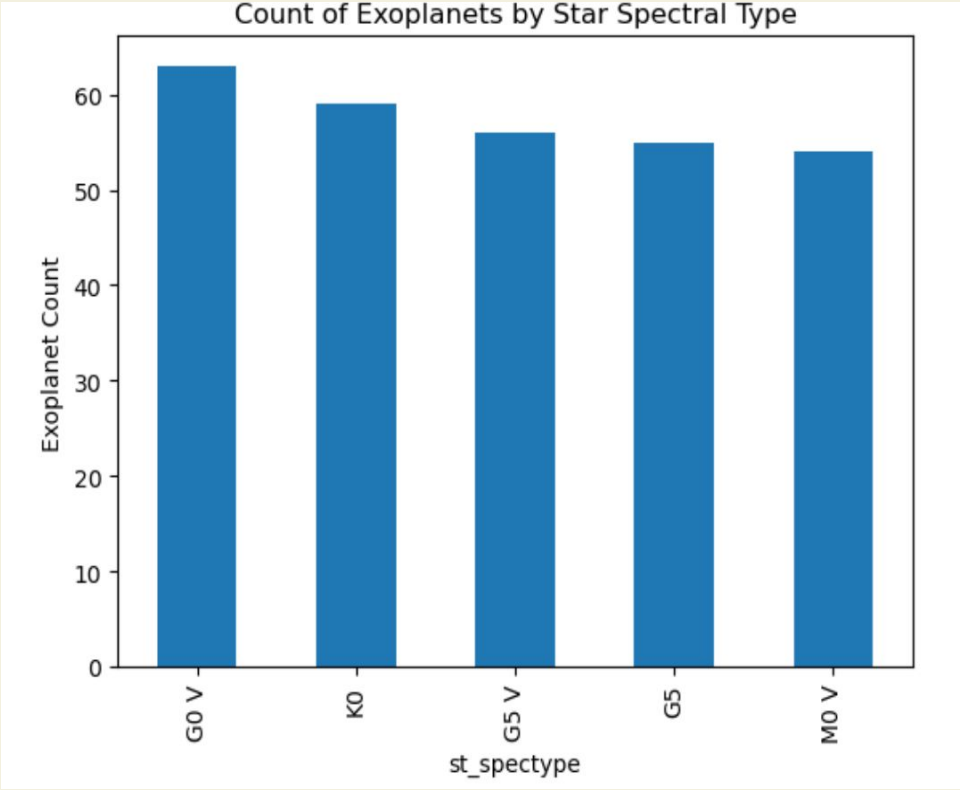
- **Radius - Radius of the planet in Earth radii (Earth = 1.0 RE). Habitable Range = (0.92, 1.60).**

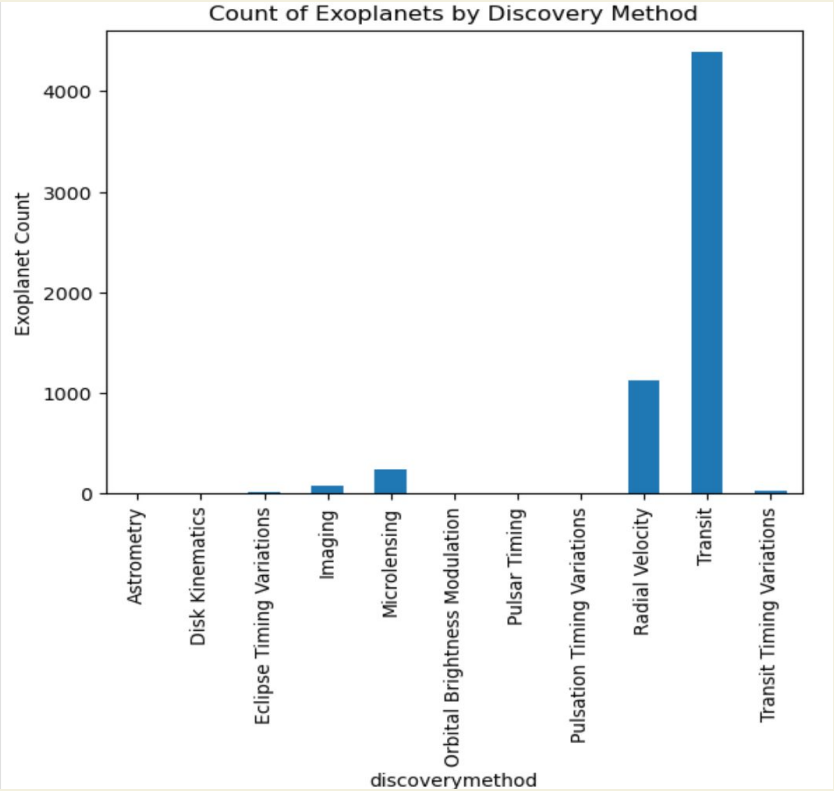
- **Flux - Average stellar flux of the planet in Earth fluxes (Earth = 1.0 SE). Habitable Range = (0.25, 1.48).**

- **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).**

EDA





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.

Target Variable Formation

- **Used the ranges of values of features listed in the previous slide and apply them in a mask function.**
- **Set aside an unlabeled dataframe of rows with NaNs to make predictions on separately.**
- **Within the mask function, conditional code was made that assigns a 1 to rows if the catalog variables within the exoplanets dataframe fall within the value ranges required for a planet to be potentially habitable, and 0 to rows otherwise.**
- **I also dropped the rows from this dataframe with the new is_habitable binary variable that had any NaNs since the rows with NaNs are in a separate dataframe.**

Preprocessing Techniques

- Dropped all variables that had a NaN percentage greater than 70%, which was 4 features
- Dropped the features used to create the Target Variable
- Used Simple Imputer with a strategy parameter of mean for all numeric variables
- Used Simple Imputer with a strategy parameter of most frequent for all categorical and boolean variables
- Created a pipeline that applied Standardized Scaler to all numeric variables and One Hot Encoder to all categorical variables
- Used a train/test split technique with a test size parameter of 0.3 and fit the pipeline to the training data
- Used the fitted pipeline to transform X_train and X_test
- Applied SMOTE to fit and resample X_train_transformed and y_train since the data was very imbalanced (about 0.2% of observations were True in target value)

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
```

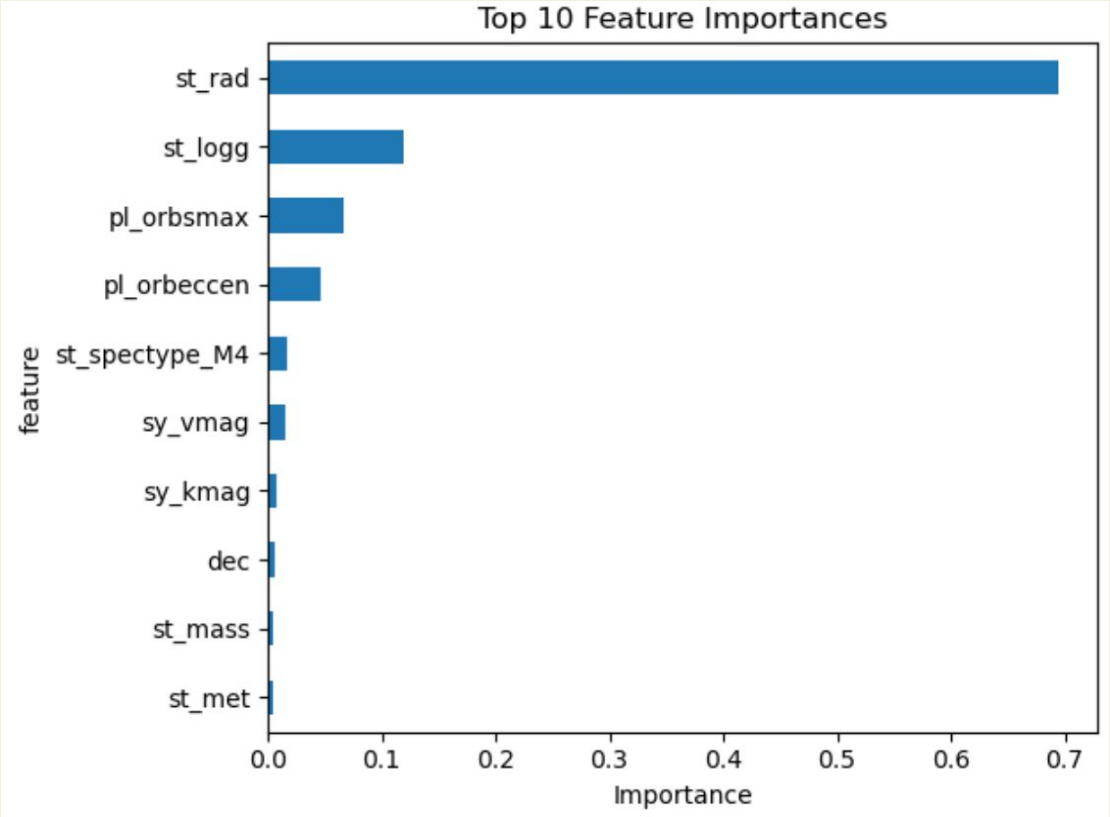
MLP Classifier Results

```
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
```

- **Along with these models, I also used Randomized Search Cross Validation to get the best possible parameters for each model.**
- **Logistic Regression predicted the most true positives or habitable planets, but has the worst f1 score, which is the biggest indicator of a model's effectiveness.**
- **Gradient Boosting Classifier Model only correctly predicted one potential habitable planet, but 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

- **st_rad or stellar radius is many times more powerful as a variable in determining if an exoplanet is habitable than the other top ten variables.**



Prediction Results on Unlabeled Dataframe

- **The same preprocessing methods and model was applied to the unlabeled dataframe**
- **The resulting positive predictions are listed to the right, a total of 9 exoplanets were predicted as being habitable, with 4 having a 100% probability of being habitable based on the model**

	pl_name	predicted	prob
195	GJ 1002 b	1	1.000000
2374	Kepler-1229 b	1	0.999989
2952	Kepler-1652 b	1	1.000000
4048	Kepler-438 b	1	0.999977
4053	Kepler-442 b	1	0.999970
4812	L 363-38 b	1	0.685392
5050	Proxima Cen b	1	1.000000
5577	TOI-715 b	1	1.000000
5617	TRAPPIST-1 e	1	0.999971

Predicted Habitable Planets Info

- **TOI-700 d – A rocky, Earth-sized exoplanet that orbits a small, cool M dwarf star in the Dorado constellation, about 101.4 light-years from Earth. The outermost of four confirmed exoplanets in the TOI-700 star system, it is about 1.2 times the size of Earth, 1.25 times Earth's mass, and takes 37.4 Earth days to orbit it's central star.**

- **GJ-1002 b – 16 ly from Earth, orbiting a cool M dwarf star. It is a rocky planet, 1.08 times mass of Earth, has an orbit of 10 Earth days. Also liquid water could potentially exist.**

- **Kepler-1652 b – super-Earth–class exoplanet (~1.60 × Earth's radius, ~3.2 × Earth's mass). 822 ly away in the constellation Cygnus, with a 38 Earth day orbit, size more similar to a mini-Neptune than a rocky planet.**

- **Proxima Cen b - Earth-sized exoplanet orbiting Proxima Centauri, the closest star to the Sun, just 4.24 light-years away. It may have conditions suitable for liquid water, though its potential habitability is uncertain due to stellar flares and radiation from its host red dwarf.**

- **TOI-715 b – A super-Earth, 1.55 × Earth's radius and 3× Earth's mass, orbiting an M-type red dwarf 137 ly away and orbits every ~19.3 Earth days, cold, but potentially warm enough for liquid water with the right atmosphere.**

Conclusions

- **All three models that I used predicted extremely few exoplanets as habitable correctly, which agrees with the expectation I had that planets having the proper conditions to support life is extremely rare.**

- **This also concurs with the fact that 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.**

Stretch Goals

- **Continue trying to find more effective machine learning models with better f1 scores.**
- **More discovered exoplanets or observations.**
- **Drop any variables with potential correlated coefficients.**
- **Incorporate any other preprocessing techniques.**
- **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.**


DAZ STUDIO
NVIDIA
Iray
GENESIS

Questions?


TS-SCI- S.A.M-422WXY
Report prepared for S.A.A.L.M by XXXXXXXXXXXXX
A-C-T-I-O-N_ACIO PINE GAP
Description
Extra-terrestrial
Species: SAM_Nordics
Aliases: Swedes, Tall Whites, Nordics
Height: 5 - 6.5 feet
Weight: 120 - 240 pounds (estimated)
Eyes: Human
Hair: Blonde
Skin: Pale White
Sex: Male and Female
Communication: Telepathic
Distinguishing Characteristics
Share common physical features with human beings
(especially Scandinavians)
Are taller than the average human
Have more of a muscular build than the average human