What Kinds of Stars can host Aliens?

May 15, 2023

Abstract

The first step to finding alien life is knowing where to look. Finding planets around other stars was the goal of NASA's Kepler Mission and the Kepler Exoplanet Archive contains observations of potentially habitable exoplanets. Based on the confirmed cases of planets in the archive, we created a multiple logistic regression model to predict if a planet is in the habitable zone of its star, where there is potential for liquid water, using stellar qualities. With our model, we are only confident associating an increase in the transformed planet-stellar radius ratio with an increase in the multiplicative odds of a planet being within a stars' habitable zone. Our analysis has one key limitation: the full dataset has 3983 planets outside their star's habitable zone and 56 inside. Due to this imbalance, we created a reduced model based on a more balanced dataset. Stellar qualities are significant predictors with the full data, but with low accuracy. With the subset, significance is lower while prediction ability increases. Ultimately, we find it is challenging to predict if planets are in their star's habitable zone. Yet, even if some qualities are slightly more useful, they could be used to sift through future telescope data to narrow down candidates for the presence of life.

1. Background information and Introduction

Could there be life on other planets in our universe? In 2009, NASA launched the Kepler space telescope from Cape Canaveral Florida with the goal of identifying potentially habitable exoplanets outside of our solar system (NASA, 2018). The first Kepler mission lasted from March 2009 until November 2012. During the mission, the telescope collected revolutionary, exploratory data on planets with potential for life. According to NASA, a “habitable zone” can be defined as “the distance from a star at which liquid water could exist on orbiting planets’ surfaces" (NASA Exoplanet Exploration, 2021). From 2014 until 2018 Kepler entered its second mission, the K2 phase, during which it continued the search for habitable planets outside of the Milky Way until it ran out of fuel. After the missions concluded, scientific discovery continued. In 2020 NASA scientists reviewing Kepler’s data from its second mission discovered that there was an exoplanet 300 million light years away from our solar system with conditions similar to Earth and potentially able to support life - Kepler 1649c (Chou & Hawkes, 2020). Our group set out to use the data collected by the Kepler Telescope to explore the relationship between the qualities of an exoplanet and its habitability.

2. Data and Exploratory Analysis

a. Data and Variables

Our group used the Kepler Exoplanet Archive in order to conduct our statistical analysis. The Exoplanet Archive contains over eight thousand data points on exoplanets. The data set contained a lot of falsely identified exoplanets, so our first step was reducing the data to only include confirmed exoplanets. This left us with 4,029 data points of confirmed Exoplanets, of which only 51 were in their star’s habitable zone. This data set is clearly only a fraction of all of the exoplanets in the universe, so we are limited by that. Furthermore we had to work with a data set where only a diminutive fraction of the data points were in the habitable zone. The habitable zone is a categorical response variable, indicating whether or not a planet is within the habitable zone of its star. The habitable zone is calculated by $\sqrt{\frac{luminosity}{1.1}}$ for the inner bound and $\sqrt{\frac{luminosity}{0.53}}$ for the outer bound around a star, where luminosity is the energy output of the star per second (in watts). For our predictors, the Kepler database includes information about the Stellar Effective Temperature, Stellar Surface Gravity, Stellar Metallicity,, Stellar Radius, Stellar Mass, and Stellar Age of an Exoplanet’s Star (Table 1). The data set also includes the disposition score of a planet, which is the amount of confidence between 0 and 1 that an exoplanet is a true planetary candidate (Table 1).

Table 1 *Variable codebook outlining all predictor and response variables explored*

Variable Name	Variable Label	Valid Range or Variable Code	Value Indicating Missing Data
koi_score	Disposition Score	1 = 100% disposition score 0 = not 100% disposition score	NA
koi_ror	Planet-Star Radius Ratio	Between 0 and 1	NA
koi_steff	Stellar Effective Temperature [K]	~3,000 to ~100,000 K	NA
koi_slogg	Stellar Surface Gravity [log10(cm/s**2)]	Greater than 0	NA
koi_smet	Stellar Metallicity [dex]	-4.5 to 1	NA
koi_srad	Stellar Radius [Solar radii]	0.01 to 1,000	NA
koi_smass	Stellar Mass [Solar mass]	0.1 to 100	NA
koi_sage	Stellar Age [Gyr]	Greater than 0 (no theoretical upper bound)	NA
hab_zone	Habitable Zone [is the planet within the bounds of a star's habitable zone?)	TRUE = Yes, it is within the habitable zone FALSE = No, it is not within the habitable zone	NA

b. Exploratory Data Analysis

For our exploratory Data analysis, we ran model diagnostics for the Disposition Score and Stellar Effective temperature of Exoplanets (Appendix Fig. 2 & 3). From the visual and numerical summaries of **disposition score**, we learned that 2257 or 55.95% of planetary

candidates had a disposition score that is not equal to 1 and that 1777 or 44.05% of planetary candidates have a disposition score that is equal to 1. This means that less than half of the candidates in our dataset are high confidence planetary candidates, while the majority of candidates are not high confidence. Looking at the two-way contingency tables and mosaic plot of **disposition score vs. habitable zone**, the conditional distributions of the two variables appear to be different enough to suggest the variables are not independent of one another. There appears to be a larger proportion of planets within a habitable zone if the disposition score is not 1. From the visual and numerical summaries of the **Stellar Effective temperature,** the temperatures of the stars are distributed relatively normally with 95% of the temperatures are between 2661 and 5543 Kelvin. By analyzing the relationship between **habitable zone** and s**tellar effective temperature**, the median and mean values are similar for planets in and outside the habitable zone, but there is a larger spread of temperature for those outside the habitable zone and many more outliers. This makes sense, since the presence of liquid water for habitable zones requires a specific distribution of temperatures. Thus, extreme temperatures might indicate a planet is not in the habitable zone and therefore is key for our model.

3. Model Results

a. Analytic Methods

Since our response variable is a TRUE or FALSE categorical variable– whether a planet is in the habitable zone can only be true or false– we decided to use a multiple logistic regression. Our predictor variables in the initial model included 7 quantitative variables–stellar effective temperature, gravity, metallicity, radius, mass, age, and planet to star radius ratio, –and 1 categorical variable– disposition score (1/0). We saw strong multicollinearity in the pair plots but performed stepwise variable selection before removing them to first understand what the most useful variables were to include in our model. We considered creating interaction terms but none made sense in terms of interpretation of our model in its context, as we were primarily concerned with assessing the effect of stellar qualities individually. In the initial model, the linearity assumption was violated for the predictors stellar radius ratio and planet to star radius ratio so a $\log_{10}$ transformation was applied to solve it. Using a two-directional stepwise selection, and removed metallicity because the AIC value with it included in the model was greater than without it. We used VIF to assess multicollinearity and found severe multicollinearity in multiple variables. We removed variables one at a time with the highest VIFs until the VIFs of all variables were < 5 and we were left with temperature, radius, radius ratio and disposition score.

b. Final Model and Results

Our final model was left with 4 predictors presented below: stellar effective temperature, surface gravity, log(planet to star radius ratio) and disposition score.

$$\widehat{log(odds)_{HabitableZone}} = -36.36 + 0.000689 StellarEffectiveTemperature + 6.74 StellarRadius + 0.977 log_{10}(RadiusRatio) - 2.027 I_{DispositionScore1}$$

Our model was more effective as compared to an intercept only model shown via likelihood ratio tests. The results from our finalized model are summarized in Table 2. All variables except for stellar effective temperature are found to be individually significant given other variables in the model as a result of a Wald test. However, based on the either very wide, very narrow or unrealistic nature of three of the 95% confidence intervals (koi_steff, koi_slogg, and koi_score), we were only confident moving forward with interpretations and conclusions surrounding one predictor variable, the log transformation of the planet to star radius ratio. We are 95% confident that every 1 unit increase in the log of the planet-star radius ratio is

associated with between a 0.58 and 0.83-fold increase in the odds of a confirmed planetary candidate being within its star's habitable zone.

In the end, we are only partially able to answer our research question by considering, realistically, the impacts of one transformed stellar quality on the likelihood of a planet being within its star's habitable zone. This is less than we were expecting to achieve with our model, as we were hoping to be able to learn useful information about multiple stellar qualities.

Due to the severe imbalance in the amount of planets in the habitable zone and the poor confusion matrix that we received with the full sample, we wanted to try using a subset of the planets in the habitable zone to attempt a new model. We ran the same tests with the smaller dataset of 100 planets in the habitable zone. With the smaller dataset, the significance of the variables is reduced (Table 4, Appendix) so that only the surface gravity and radius ratio are significant variables in our model, but the confusion matrix (Table 5, Appendix) is much better than our model ran on the full dataset.

Table 2 *Significance, z-test results, and 95% confidence intervals of coefficients associated with predictor variables in the final multiple logistic regression model. Full hab_zone sample size.*

Variable	Coefficient	P-Value	Wald test Significance (Y/N)	95% C.I. of Coefficient (π/1-π)
koi_steff	0.000689	0.02888	Y	0.50002 - 0.5003
koi_slogg	6.74	4.78e-06	Y	0.981 - 0.99
$\log_{10}$(koi_ror)	0.977	0.00251	Y	0.575 - 0.829
koi_score(1)	-2.0271	4.19e-06	Y	0.047 - 0.2242

5. Discussion and Conclusions.

The main purpose of our project was to identify whether there was a relationship between the qualities of an exoplanet, and whether it was in the habitable zone. Our final model was one that included the stellar effective temperature, surface gravity, and a logarithmic transformation of a star's radius ratio as well as the disposition score of the exoplanet. Our model diagnostics show us that there seems to be a statistically significant positive relationship between our predictor variables and the odds of an exoplanet being in the habitable zone. However, when we used a confusion matrix to assess our model, we found that the model was not very accurate at predicting whether a planet was in the habitable zone or not. Because of this we created another confusion matrix in which we randomly sampled 100 planets outside of the habitable zone. With this tweak, our model was better at predicting whether a planet was within the habitable zone, but the model itself became less statistically significant. What this shows us is that with the available data, there is a tradeoff between having a statistically significant model and having a model that is good at predicting whether a planet is in the habitable zone. In part, this might be because we need more data from planets within the habitable zone (we only had 51 data points in this data set) so that we can better understand what qualities of exoplanets affect their status as being a planet within the habitable zone.

References

Chou, F. & Hawkes, A. (2020, April 15). *Earth-size Habitable Zone Planet Found Hidden in Early NASA Kepler Data*. NASA Exoplanet Exploration. https://exoplanets.nasa.gov/news/1637/earth-size-habitable-zone-planet-found-hidden-in-early-nasa-kepler-data/.

NASA. (2018, Oct. 30). *Kepler and K2: Mission Overview*. NASA. https://www.nasa.gov/mission_pages/kepler/overview/index.html.

NASA Exoplanet Archive. *Kepler Objects of Interest DR25 Table* [Data Set]. https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=q1_q17_dr25_koi.

NASA Exoplanet Exploration. (2021, April 2). *The Search for Life*. NASA. https://exoplanets.nasa.gov/search-for-life/habitable-zone/.

Appendix

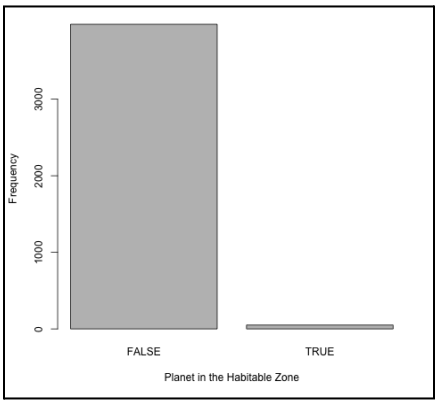

Figure 1 Bar chart displaying distribution of planetary candidates in the data set across TRUE and FALSe categories for the response variable habitable zone. Note the large visual discrepancy between the two categories.

a) b)

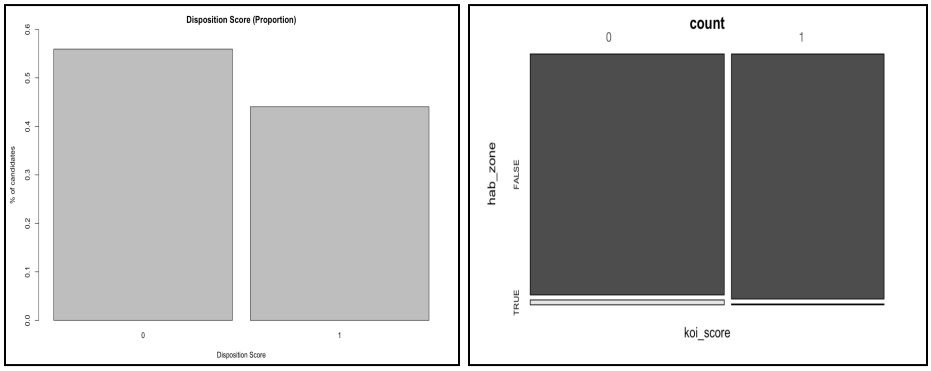

Figure 2 a) Univariate EDA: Bar chart displaying distribution for the distribution score variable b) Bivariate EDA: Mosaic plot representing the relationship between habitable zone response variable and disposition score predictor. We have categorized the disposition score to be 0 for any score < 1 and 1 otherwise.

a) b)

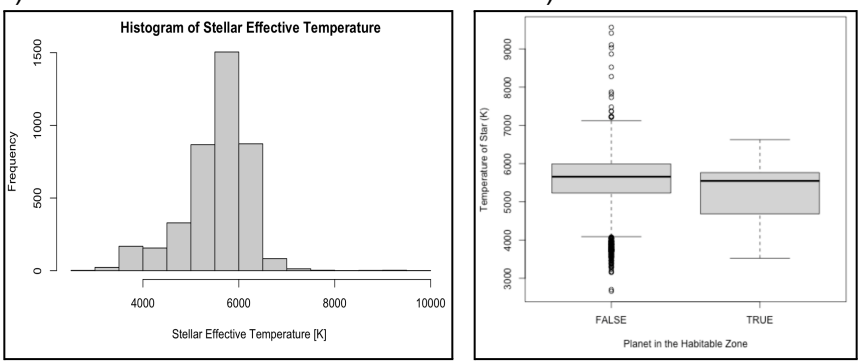

Figure 3 a) Univariate EDA: Histogram of distribution of stellar effective temperature b) Bivariate EDA: Side by side boxplot representing distribution of stellar effective temperature across response variable, habitable zone, categories.

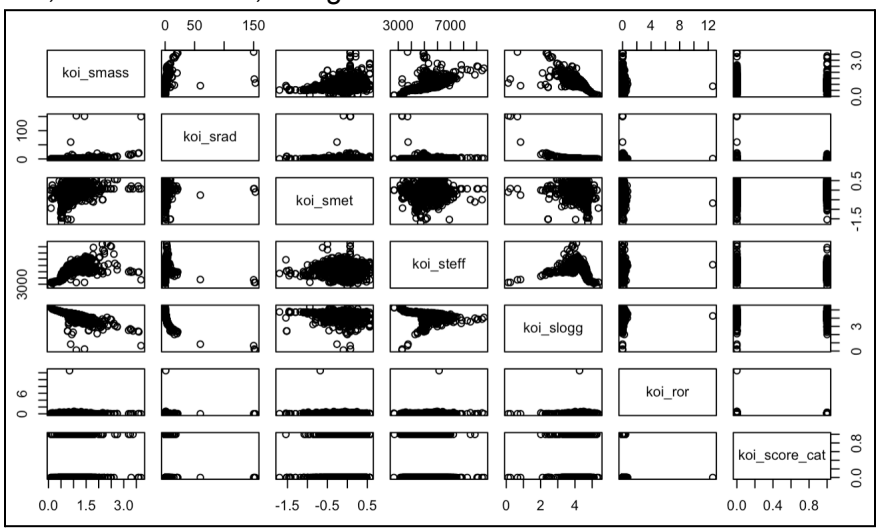

Figure 4 Pairs plot demonstrating relationships between all predictor variables included in the original model.

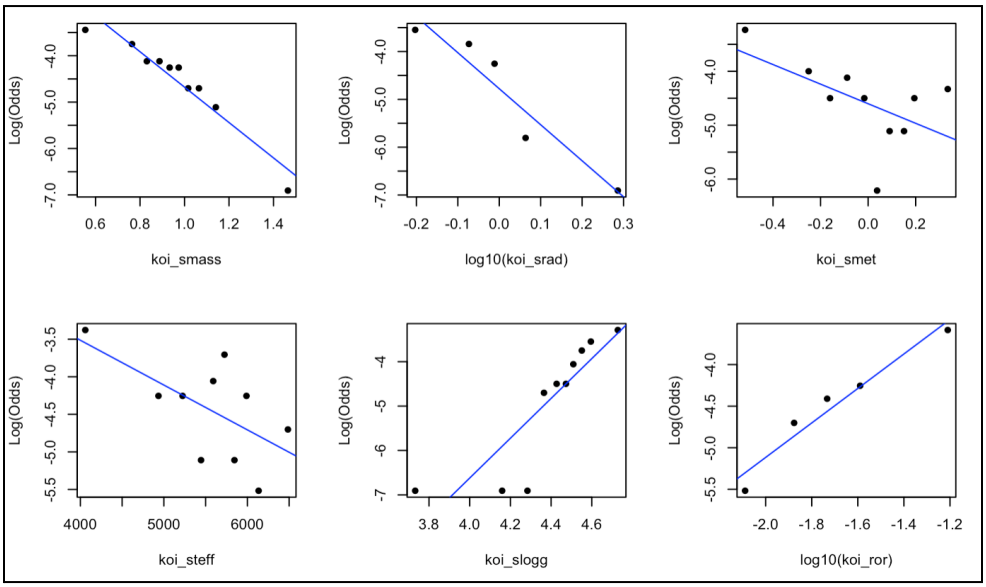

Figure 5 Model diagnostic empirical logit plots for all quantitative predictor variables included in original regression. The plots display the variables post- log transformation for koi_srad and koi_ror. Note all model assumptions are followed.

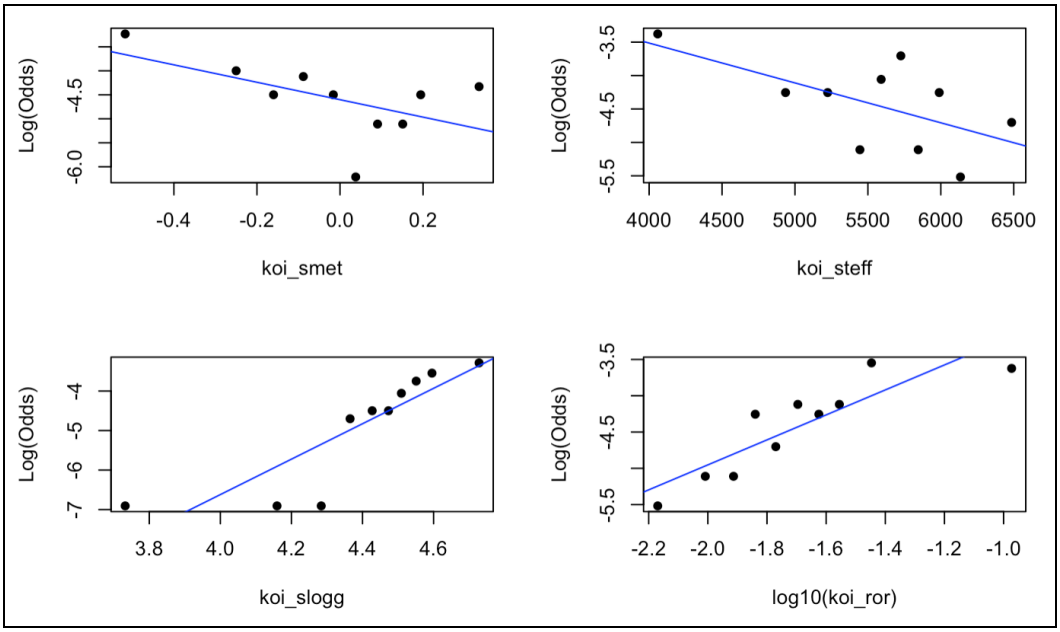

Figure 6 Re-assessed model diagnostic empirical logit plots for all quantitative predictor variables included in the final model.

Table 3 *Significance, z-test results, and 95% confidence intervals of coefficients associated with predictor variables in the final multiple logistic regression model for reduced hab_zone sample size.*

Variable	Coefficient	P- Value	Z-test Significance (Y/N)	95% CI of Coefficient ($\pi/1-\pi$)
koi_steff	0.0005748	0.3799	N	0.998 - 1.001
koi_slogg	6.225	0.0814	Y	$1.442 - 2.055*10^{6}$
$\log_{10}$(koi_ror)	2.892	0.000282	Y	4.178 - 100.6
koi_score(1)	-0.5673	0.3320	N	0.1690 - 1.728

Table 4 *Confusion Matrix where each of the rows signify the total amount of planets predicted to be in the habitable zone based on our model. The columns show us the true count of planets in the habitable zone based on our data. A higher fraction of planets in the diagonal of this matrix signify a model better at predicting.*

	In Hab Zone	Not in Hab Zone
Predicted in Hab Zone	**3975**	**51**
Predicted not in Hab Zone	**2**	**0**

Table 5 *Confusion Matrix where each of the rows signify the total number of planets predicted to be in the habitable zone based on our model, for reduced hab_zone sample size. Same as Table 3 for our reduced sample. Notice how there are now more planets in the diagonal.*

	In Hab Zone	Not in Hab Zone
Predicted in Hab Zone	**116**	**14**
Predicted not in Hab Zone	**8**	**13**