Predicting Exoplanets by Model Fitting Kepler Objects of Interest Data

David G. Kinney

Bellevue University

DSC 680 – Professor Catherine Williams

Abstract

Exoplanets are planets that orbit around a star, as in our solar system. These bodies are very hard to see directly with telescopes since they are hidden by the bright glare of the stars they orbit. The Kepler Objects of Interest (KOI) dataset contains various measurements of transit, in addition to many others that aid in identifying exoplanets. In 2009, NASA launched the **Kepler** spacecraft to search for exoplanets. Kepler looked for planets in a wide range of sizes and orbits that circled around stars of varied size and temperature. [2] The Kepler spacecraft detected exoplanets using the **transit** method. As a planet passes (transits) in front of a star, it blocks out a bit of the star's light. By observing the stars' change in brightness astronomers can figure out the size of the orbiting planet as well as how far away it is from the star. Further, this data then aids in calculating the planet's temperature, and the chances that it may contain liquid water—the stuff of life…

Keywords: exoplanet, NASA, RandomForestClassifier, RandomizedSearchCV, PCA

Introduction & Hypothesis

The NASA Exoplanet Archive is an online astronomical exoplanet and stellar catalog and data service that collates and cross-correlates astronomical data and information on exoplanets and their host stars. These data include stellar parameters (such as positions, magnitudes, and temperatures), exoplanet parameters (such as masses and orbital parameters) and discovery/characterization data (such as published radial velocity curves, photometric light curves, images, and spectra). [1] As this data is used by astronomers to arrive at whether an object is an exoplanet, it follows that it is likely a Machine Learning model can be developed to make predictions based on the same data.

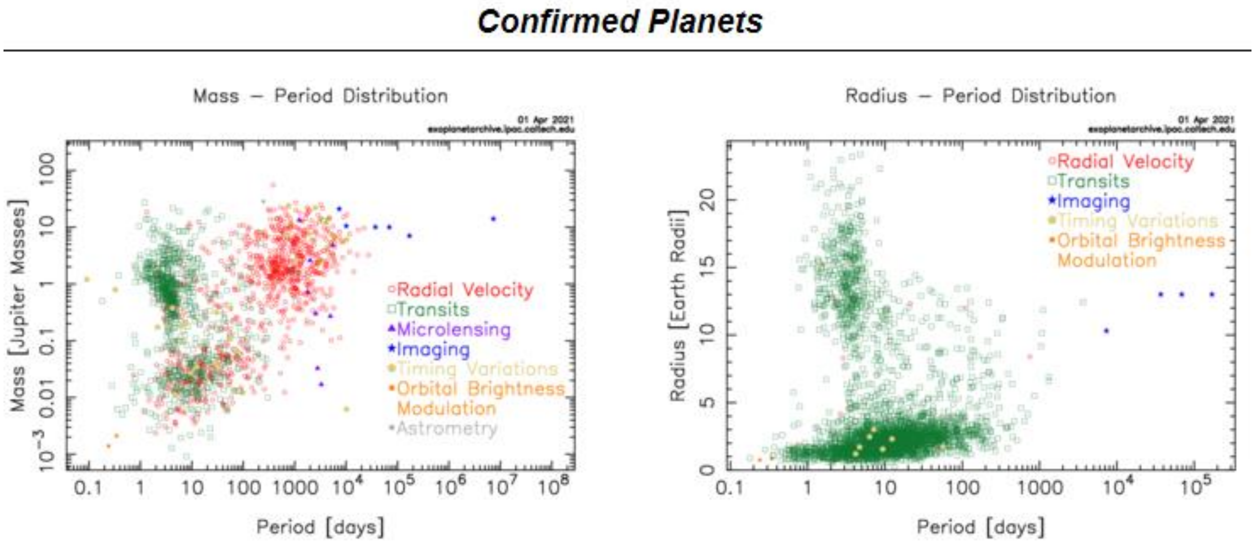
Data

The dataset I am leveraging is the **Kepler Object of Interest (KOI) Cumulative** table. The cumulative KOI table gathers information from the individual KOI activity tables that describe the current results of different searches of the Kepler light curves. The intent of the cumulative table is to provide the most accurate dispositions and stellar and planetary information for all KOIs in one place.

The **Data Columns** documentation can be reviewed here: Data columns in Kepler Objects of Interest Table (caltech.edu). The dataset can be found here: Kepler Objects of Interest (caltech.edu). The dataset consists of 9,565 rows with 2,358 confirmed exoplanets, 2,366 candidates and 4,840 false positives.

The dataset is divided into nine categories, the first three being identification, archive information, and disposition. The remaining six categories are directly related to observational data used to categorize objects as either a confirmed exoplanet, a candidate, or a false positive.

- Transit Properties
- Threshold-Crossing Event (TCE) Information
- Stellar Parameters
- KIC Parameters
- Pixel-Based KOI Vetting Statistics



Source: NASA Exoplanet Archive

Method

Data Preparation

All variables that were completely empty or contained all zeros were dropped. There were also several error factor variables for some of the measurements. The values in these columns were extremely small and not likely to influence the model's performance; they were dropped as well. Lastly, I made the decision to drop the 20 categorical variables. 12 of them were simply descriptive (informational). For the remaining 8, there was very little variation in the values (in at least one case, the value was a veritable constant). I concluded they were not likely to have a large impact on the behavior of the model, and as they would all require encoding, it would only add to the magnitude of the dataset.

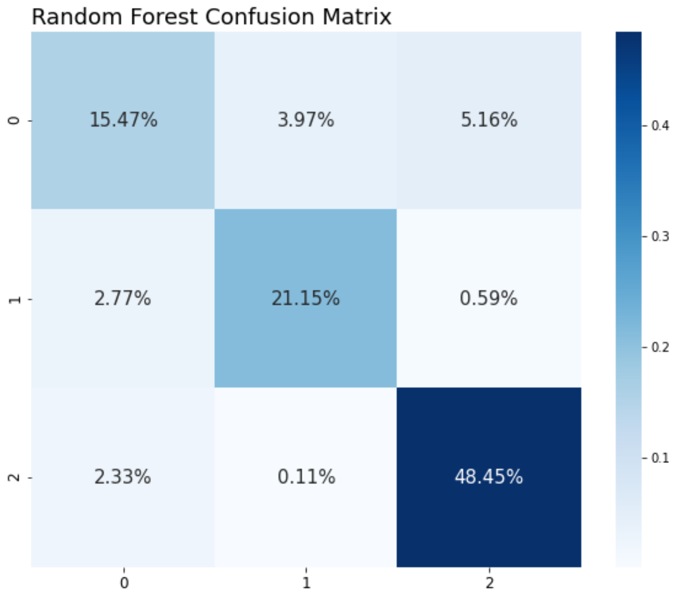
The **koi_disposition** variable was separated from the resulting dataset and saved as the target predictor labels (Confirmed, Candidate, and False Positive). For the remaining features dataset, I employed a **SimpleImputer** to replace missing values with the median of that feature.

Lastly, *dimensionality reduction* was performed by applying **Principal Component Analysis**, reducing the feature set from 53 variables to 29.

Modeling

A baseline model was established by training a **Random Forest Classifier**, randomly assigning 1,000 estimators. The function containing the model also applied a **StandardScaler** to normalize the data prior to model fitting. This resulted in an accuracy score of 86%.

	precision	recall	f1-score	support
CANDIDATE	0.79	0.66	0.72	601
CONFIRMED	0.86	0.87	0.87	600
FALSE POSITIVE	0.9	0.96	0.93	1190
accuracy			0.86	2391
macro avg	0.85	0.83	0.84	2391
weighted avg	0.86	0.86	0.86	2391



This was followed by performing a randomized search by leveraging a **RandomizedSearchCV** model selector. Interestingly, this did not result in an improvement in accuracy.

Research Questions

1. Regarding the decision to exclude the 8 categorical variables, can you elaborate on what they represented?
2. In retrospect, would any of them add any value to the model's accuracy?
3. Why did you decide to employ a standard scaler, as opposed to say, a min/max scaler?
4. In your initial proposal, you were going to train the model on subsets based on the 6 categories. What changed your mind?
5. Do you feel this is still something worth pursuing?
6. Why did you decide to employ a Random Forest Classifier?
7. Are there any other models you feel may give you better results?
8. What other methods might you employ to improve the accuracy of your model?
9. You state in your paper that a randomized grid search did not result in model improvement. Had you considered applying a genetic algorithm to perform hyperparameter tuning?

Conclusions

While 86% accuracy is not exemplary, that was achieved with a simple baseline model. I am confident greater accuracy is possible, either by fine-tuning the hyperparameters for the Random Forest Classifier, or by investigating other models, such as **XGBoost** or Convolutional Neural Networks. Given that Machine Learning algorithms can process data faster than humans by an order of magnitude, deployment of such a model that can classify observations of the trillions of solar objects present in the observable universe with blinding speed will free up astronomers for more creative endeavors; one area where humans still dominate.

> “Whether life exists beyond Earth is one of the most profound questions of all time. The answer will change us forever, whether it reveals a universe rich with life, one in which life is rare and fragile, or even a universe in which we can find no other life at all.” [3]

References

[1] NASA Exoplanet Archive – NASA Exoplanet Science Institute
https://exoplanetarchive.ipac.caltech.edu/index.html

[2] What Is an Exoplanet? | NASA Space Place – NASA Science for Kids

[3] Overview | What is an Exoplanet? – Exoplanet Exploration: Planets Beyond our Solar System (nasa.gov)

[4] https://www.zooniverse.org/projects/nora-dot-eisner/planet-hunters-tess/about/research

Appendix

Figures

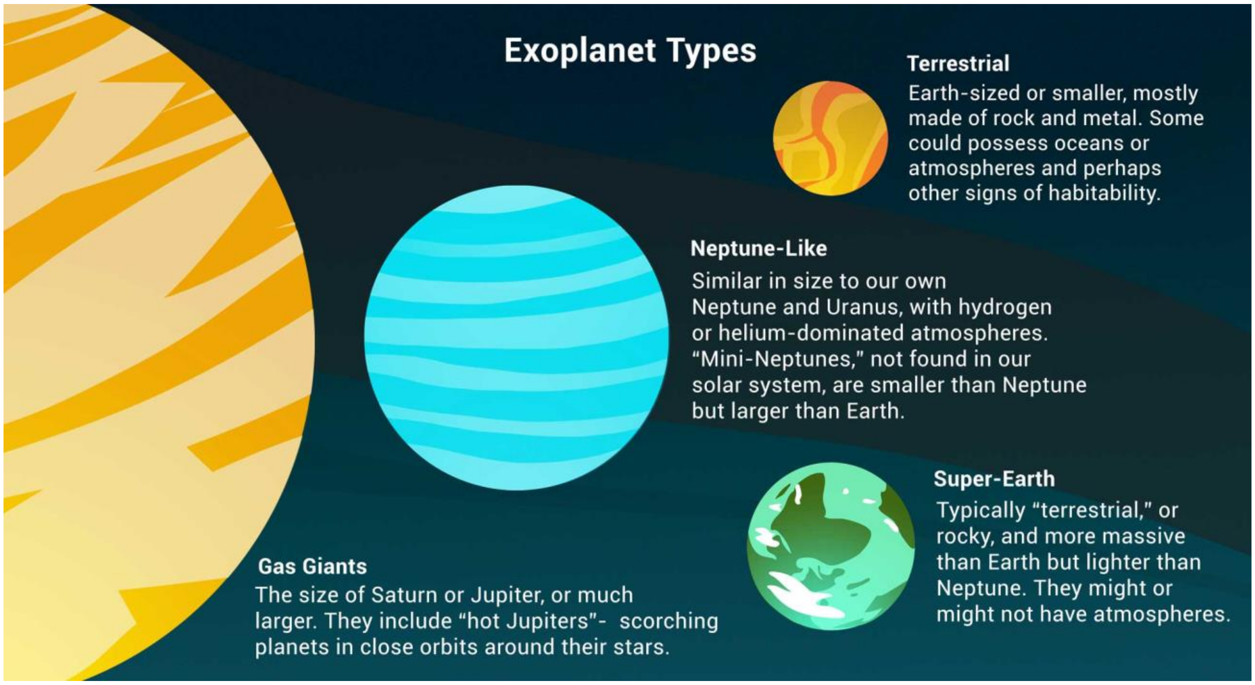

Exoplanet Types
Terrestrial
Earth-sized or smaller, mostly made of rock and metal. Some could possess oceans or atmospheres and perhaps other signs of habitability.
Neptune-Like
Similar in size to our own Neptune and Uranus, with hydrogen or helium-dominated atmospheres. "Mini-Neptunes," not found in our solar system, are smaller than Neptune but larger than Earth.
Super-Earth
Typically "terrestrial," or rocky, and more massive than Earth but lighter than Neptune. They might or might not have atmospheres.
Gas Giants
The size of Saturn or Jupiter, or much larger. They include "hot Jupiters"- scorching planets in close orbits around their stars.

Image credit: NASA/JPL-Caltech/Lizbeth B. De La Torre

A simplified example of what the lightcurve from a transit looks like. It shows that as the planet passes in front of its host star, the light received decreases. [4]

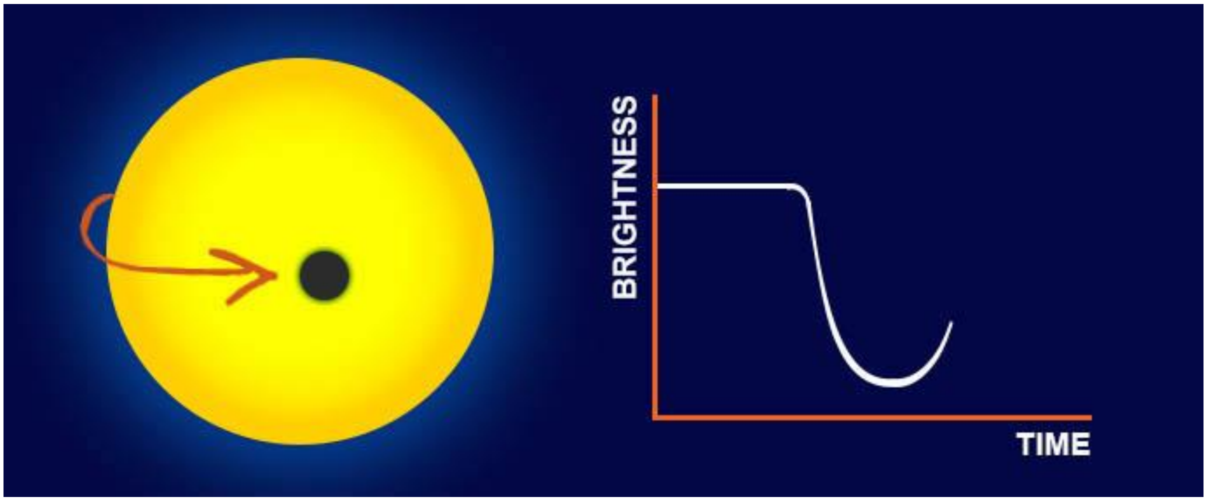


Photo credits: Nora Eisner

In order for us to be able to observe a transit we require the planetary system to be oriented so that the **planet passes between us and the host star** (as shown on the right-hand side of the image below). If this is the case, we will see a dip every time that the planet completes one full orbit around the star. If the planet does not cross our line of sight, we will miss the transit (shown on the left). [4]

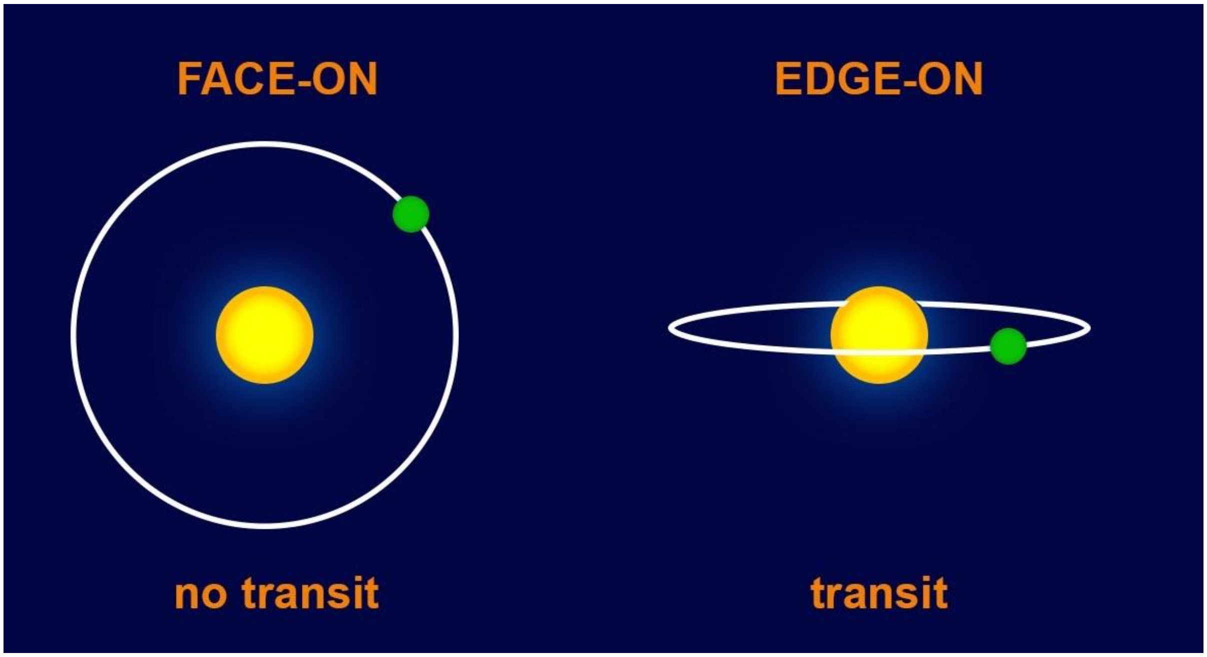


Photo credits: Nora Eisner