Willamette University

Exoplanet Prediction using Machine Learning Techniques

Classifying Kepler Objects of Interest from the Kepler Mission dataset using supervised and unsupervised methods.

Nina Hernandez and Megan Eichelberger
Data Science in the Natural Sciences
Professor Jed Rembold
July 1st, 2022

1. Abstract

It is no longer effective to use direct visual detection or radio pulses of a pulsar as the sole means of detecting exoplanets when the computational power and quantitative techniques to analyze data are readily available. This paper utilizes metadata from the Kepler Object of Interest dataset from the Kepler Mission space observatory launched by NASA in 2009 [1], which according to Bugueño et al. [2] accounted for 65% of all exoplanet discoveries. The goal was to use unsupervised and supervised machine learning techniques to predict or describe exoplanets given a list of unknown candidates. Additionally, the authors sought to find the top features that contrast false positives from confirmed Kepler Objects of Interest. Using the techniques described to analyze the metadata of transit lightwave data, the model identified 850 new exoplanets in the Kepler candidates list that shared similar qualities to previously confirmed exoplanets by building powerful models to tease out hidden false positives.

2. Introduction

Transit photometry is a method of detecting exoplanets based on changes in light as an object passes in between a star and the viewer. Advancements in this method over the past 23 years allowed projects like the Kepler Mission to detect over 2,951 exoplanets and counting [2]. The Kepler dataset referenced in this paper contains information not only about the characteristics of the object - e.g. transit and orbital period - but also characteristics of the star - e.g. radius and gravity. This allows researchers to examine a broader range of characteristics and is prime for building machine learning models.

Utilization of machine learning in exoplanet classification is relatively nascent. As space exploration progresses and the depth of the data grows, there is great opportunity to identify machine learning techniques that will accurately predict exoplanets with minimal human interaction. This paper will explore various models -K Nearest Neighbors, Logistic Regression, and Random Forest Classification- to determine the most effective prediction method.

3. **Exploratory Analysis**

a. Data

In this paper the authors utilized the Kepler Object of Interest (KOI) dataset [1] accessed through CalTech. This dataset originally had 9564 observations with 44 features total. Each record represents a single KOI, labeled as a candidate (2,667 objects), confirmed (4,562 objects), or false positive (1,949 objects)[1].. While Yucheng et. al.[3] used a correlation matrix to identify relevant features, the authors of this paper utilized Principal Component Analysis. This yielded the working dataset - 9,178 rows with 9 features. *fig 1* The working data also included the object's classification along with right ascension and declension for mapping.

Column Name	Table Label	Description
koi_slogg	Stellar Surface Gravity	The photospheric temperature of the star
koi_srad	Stellar Radius	The photospheric radius of the star
koi_teq	Equilibrium Temperature	Approximation of planet's temperature
koi_kepmag	Kepler Band Magnitude	Estimated magnitude as seen by the Kepler telescope
koi_prad	Planet Radius	Radius of planet
koi_insol	Insolation Flux	Solar irradiance
koi_period	Orbital Period	Interval between consecutive transits
koi-duration	Transit Duration	Duration of observed transit
koi_time0bk	Transit Epoch	Time of first detected transit

Figure 1. Features chosen from Kepler dataset for supervised and unsupervised Machine Learning.

b. PCA

In order to identify the salient features in a systematic way, the authors chose principal component analysis. PCA gives a visual *fig 2* and numeric representation of the most important features in predicting future data. Further, PCA can reduce overfitting by removing variables which do not significantly contribute to the accuracy of the model. Using the elbow method in conjunction with a scree plot to determine appropriate dimensionality, it appears that two dimensions are ideal to create predictive models. Figure 2 shows the top features across the first two dimensions. This then informed the list of features chosen for the final models. *Fig 2*

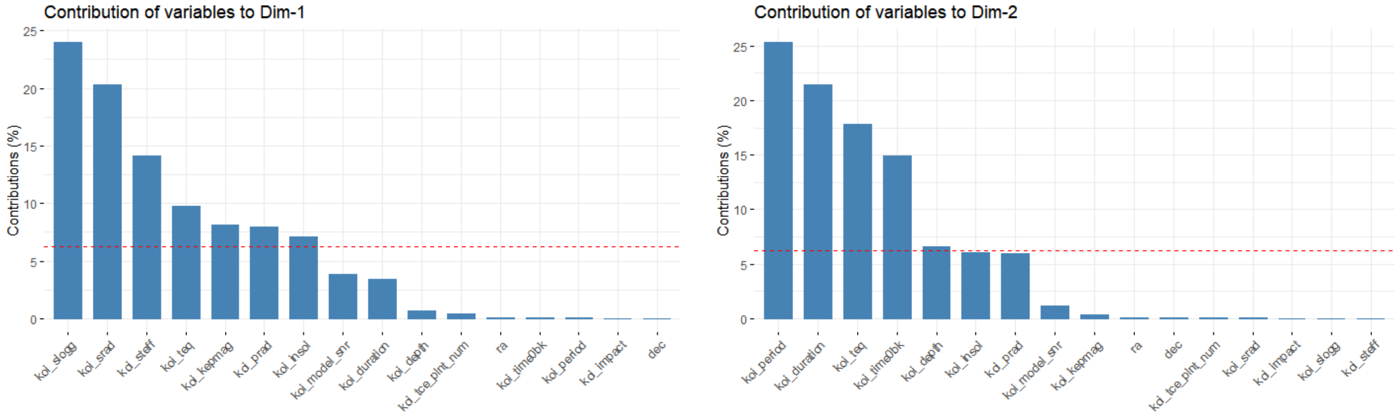

Figure 2. PCA conducted across two dimensions.

4. Methodology

In order to create the models the data was split into training, validation, and test sets. The training and validation sets contained no “candidate” planets - creating a model completely trained on objects that were classified with relative certainty. Three models were chosen to compare: K nearest neighbors - using *k* samples to classify data, logistic regression - classification on a binary model, and random forest - classification through randomly sampled decision trees. These models are laid out in Bugueño, et.al. [2].

In order to reduce model bias, it is necessary to compare the data across a common scale. The data was preprocessed using BoxCox for centering and scaling. After training models on known KOI dispositions with hyper tuning and cross validation using the bootstrap method with three repeats, we could then validate the model and obtain metrics of performance like Kappa and Accuracy. While Accuracy is the de facto metric for summarizing the performance of classification models, accuracy can fail on classification problems with a skewed class distribution. As the working dataset contained 22% candidates, 29% confirmed, and 49% false positives, the authors also use Kappa as a secondary metric for model validation. Kappa measures the expected accuracy, or how well a classifier would do by chance. Generally a kappa between 0.4 and 0.75 is acceptable, and above 0.75 is excellent.

5. Results

A random forest model had the highest kappa and accuracy metrics against the other two models. *Fig 3* While the other models did have a good to excellent kappa, random forest's 0.7805 was the most robust. Although this model was the most accurate, it also predicted more false positives than the others exemplifying its sensitivity to the class imbalance. However, the imbalance of false positives or non-exoplanets to exoplanets in classification problems extends beyond the dataset [2]. Access to additional data was limited, thus random forest was the best option.

	K-Nearest Neighbors	Logistic Regression	random forest
Kappa	0.7532	0.7365	0.7805
Accuracy	0.8827	0.8768	0.8972
# Predicted Exoplanets	1071	1204	875
# Predicted False Positives	880	747	1076

Figure 3. The reported Kappa and Accuracy metrics of each model as well as the number of predicted exoplanets and false positives after testing the models on the unknown candidates set.

After finding the optimal model, the caret package in R identified variable importance of the features in the model, the top five of which are visualized in figure 4. The variable importance is a measure of how great of an effect a particular coefficient has on the model.The features in our random forest model with the greatest importance were the object's inferred radius (koi_prad), the object's orbital period (koi_period), the transit duration (koi_duration), the insolation flux (koi_insol), and the equilibrium temperature (koi_teq). We then used these features to uncover predicted versus actual objects of interest.

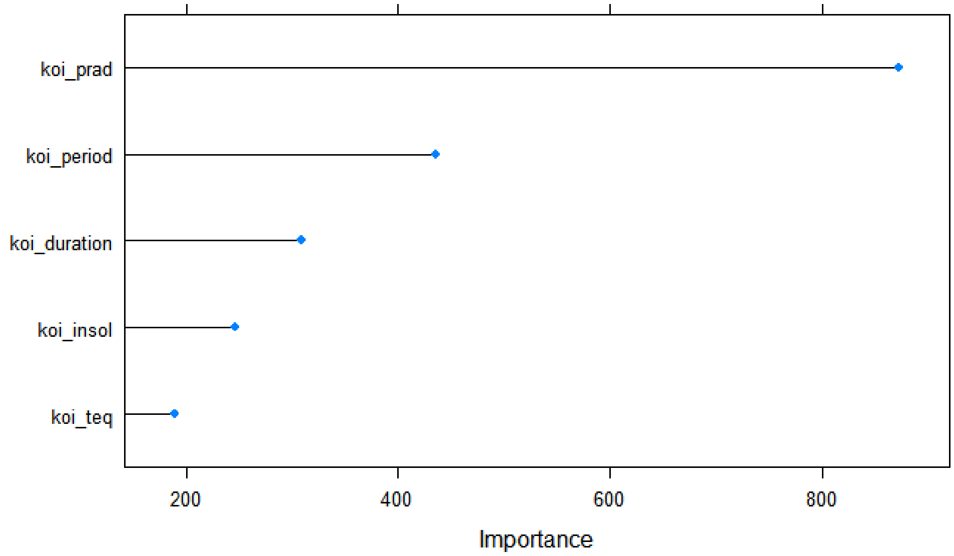

Figure 4. Features in the optimal model (random forest) with greatest importance were the object's inferred radius, the object's orbital period, the transit duration, the insolation flux, and the equilibrium temperature.

The authors discovered that the features with the highest variable importance in the model gave insight about the characteristics of false positives and confirmed exoplanets. Looking at the actual true vs false KOIs for the median metrics of those features in *fig 5* on lines 1 and 3, there are distinct contrasts across all 5 metrics. The findings show that intervals between consecutive planetary transits that are exceptionally low or high are likely to be false positives [3] - demonstrated on lines 3 and 4. For actual true versus predicted values on lines 1 and 2, there are comparable metrics across the object's median radius, the median number of planets in the object's system, the median transit period, and the median surface temperature. The model was especially powerful in predicting eclipsing binary star-like objects [1] with smaller object radii, much longer orbital periods, lower equilibrium temperature, and much less insolation flux than previously confirmed false positives (see *fig 5*).

	Median Object's Radius (Earth radii)	Median Transit Duration (hours)	Median Orbital Period (Days)	Median Equilibrium Temperature (Kelvin)	Median Insolation Flux (Earth Flux)
Actual True	2.16	3.499	11.35011	777	86.18
Predicted True	1.63	3.4769	13.05609	750	74.65
Actual False	8.96	4.1178	4.939131	1152	416.75
Predicted False	1.89	3.809	34.8639	629.5	37.135

Figure 5. Median metrics of the features with the highest variable importance from our model.

It was initially anticipated based on reference research [2][3] that the location of candidates and their system could play a role in their disposition. However, mapping the Right Ascension and Declination across dispositions *fig 6* illustrates that a relationship between the disposition of an object and its location data alone was nontrivial.

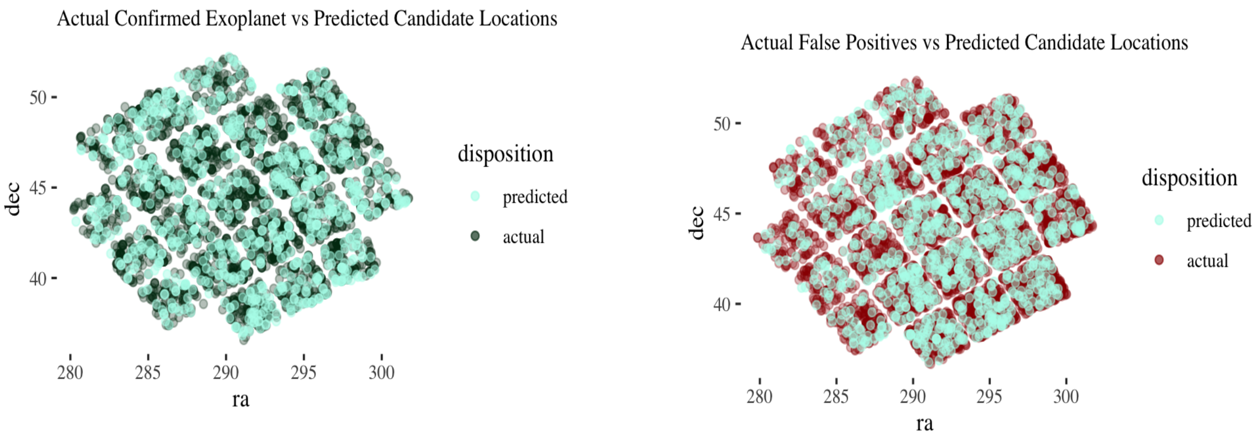

Figure 6. There is no found distinction between the predicted and actual candidates for both confirmed and false positives by the location of their systems.

7. Conclusions

In addition to the more than 2,951 exoplanets already detected by the Kepler mission using the transit method, both supervised and unsupervised machine learning methods predicted an additional 850 potential exoplanet candidates. While locations of the systems were not indicative of the dispositions of objects of interest, several other features were integral to accurate modeling such as the object's inferred radius or the object's orbital period. Furthermore it was found that the orbital period was a critical feature in both predicted and actual false positives. Orbital periods observed as exceptionally low were a hallmark of actual false positives whereas exceptionally high orbital periods were common in the predicted false positives. One limitation to analysis was the class imbalances in the dataset. A similar approach to modeling the data in [2] found downsampling to be ineffective in fixing model bias. However, future work may involve using techniques such as regularization or upsampling, engaging in further analysis around feature correlates, or utilizing other transformations. As the influx of object data and methods to gather data quantify and advance, so will our knowledge and understanding of our universe and the classification of objects that exist therein.

Works Cited

[1] "Links to Data in The Exoplanet Archive." *Data Resources in the Exoplanet Archive*, Feb. 2011, https://exoplanetarchive.ipac.caltech.edu/docs/data.html.

[2] Bugueño, Margarita & Mena, Francisco & Araya, Mauricio. (2018). "Refining Exoplanet Detection Using Supervised Learning and Feature Engineering." 278-287. 10.1109/CLEI.2018.00041.

[3] Yucheng Jin, et al. (2022). "Identifying Exoplanets with Machine Learning Methods: A Preliminary Study". *International Journal on Cybernetics; Informatics,* 11. 31–42.