Finding Exoplanets with Machine Learning: Developing Classification Models for Kepler Mission Data

Timothy Drexler

August 2019

Background

Over its 9-year mission lifetime, the Kepler space telescope collected time-series measurements of stellar brightness on a sample of more than 150,000 stars (Brown, Latham, Everett, & Esquerdo, 2011). The goal of the Kepler survey was to identify stellar systems in which the transit of extrasolar planetary bodies between the star and the telescope caused periodic decreases of brightness. Preliminary analysis of the light curve data found 9,564 potential planetary systems, known as Kepler Objects of Interest (KOIs). Astronomers then conducted follow-up research on the KOIs using ground-based telescopes. These observations helped determine whether a transiting planet caused the reduction in stellar luminosity or if there was another explanation, such as an eclipsing binary star or instrument noise.

For this machine learning project, I used a data set containing measurements of the nearly 10,000 KOIs, made available on Kaggle by NASA (NASA, 2017). My goal was to explore different machine learning techniques and develop a classification model capable of predicting whether a KOI with a given set of characteristics was more likely to be confirmed as a planetary system or a false positive. Astronomers could use such a model to prioritize the limited telescope time available for follow-up observations and focus their research on KOIs with higher potential to host planetary systems.

Data Set Description & Exploration

The KOI data set originally contained 50 variables. The response variable of interest for this project was koi_pdisposition, a categorical variable with two possible values. The CANDIDATE label indicated a KOI that met the criteria for consideration as a potential planetary system pending further analysis and confirmation. FALSE POSITIVE meant that ensuing observations found that the signal was spurious. NASA uploaded the data set to Kaggle on October 10, 2017; for this project, I assumed the koi_pdisposition value reflected the status of each KOI as of that date, subject to further updates.

My plan for developing the classification model was to use variables representing measured or calculated physical characteristics of a KOI to predict the value of koi_pdisposition. To find the most appropriate variables, I first explored the data set and consulted the data dictionary available from the NASA Exoplanet Science Institute website (2017). I found that many of the included variables, such as measurements of stellar position and object id numbers, were irrelevant for generating predictions. Overall, I reduced the number of predictor variables from 49 available candidates down to 12. All the selected variables consisted of continuous numeric values. The variables fell into roughly two groups, as summarized below (refer to the data dictionary at the NASA Exoplanet Archive for detailed descriptions).

- The first group of variables measured properties of the potential planetary object and the observed transit event:
    - koi_duration, the transit time in hours
    - koi_period, the time between transits, measured in days
    - koi_depth, the fraction of stellar flux (energy per unit area) lost at the transit minimum when the potential planetary object blocked the maximum amount of light from the star
    - koi_prad, the radius of the planet in Earth radii
    - koi_teq, a calculated estimate of the planetary equilibrium temperature in Kelvin
    - koi_insol, the insolation flux, another calculated estimate of planetary temperature, measured in units of Earth flux
    - koi_model_snr, a calculation of the transit depth flux measurement normalized by the mean amount of uncertainty in the measurement
    - koi_impact, a normalized parameter estimating the distance between the center of the planet disc and the center of the stellar disc at conjunction.
- Variables in the second group measured the properties of the star:
    - koi_steff, the stellar effective temperature in Kelvin
    - koi_slogg, the base-10 logarithm of the stellar surface gravity
    - koi_srad, the stellar radius measured in units of solar radii
    - koi_kepmag, the stellar magnitude in the Kepler band

Preliminary Analysis

After selecting the modeling variables, I examined the observations with missing data for at least one predictor variable. For most of the 364 records in this category, no data existed for more than two or three predictor variables, making the imputation of the other missing measurements impractical. Since the incomplete observations represented less than 4% of the overall data set, and since there was no explanation I could find as to why the data was missing, I decided to proceed with the model selection process using only the 9,200 observations with complete records.

Further exploration showed that 11 of the 12 predictor variables had distributions that deviated from normal. Ten variables had distributions skewed to the right, while the koi_slogg variable distribution had a left skew. This suggested variable transformation could be beneficial, both to reduce the variance within each distribution and to decrease the amount of skewI selected the $ln(x+1)$ transformation since multiple variables contained observations with a value of 0. I applied the transformation to all predictor variables with right-skewed distributions, leaving 'koi_slogg' (left-skewed distribution) and 'koi_kepmag' (approximately normal distribution) untransformed.

Figure 1 illustrates the effect of the $ln(x+1)$ transformation on right-skewed distributions, using the variable koi_depth as an example. The boxplot on the left shows the original data distribution: most of the data fall below a value of 1,500 parts per million, but some observations range to above 1,500,000 parts per million. The plot on the right shows how transforming the variable decreases the variance, reducing the range of values to 0 through 14.248. In addition, visual inspection confirms that the data distribution of the transformed variable is closer to the symmetric shape of a normal distribution.

Figure 1

Box Plots of 'koi_depth' Variable Distribution

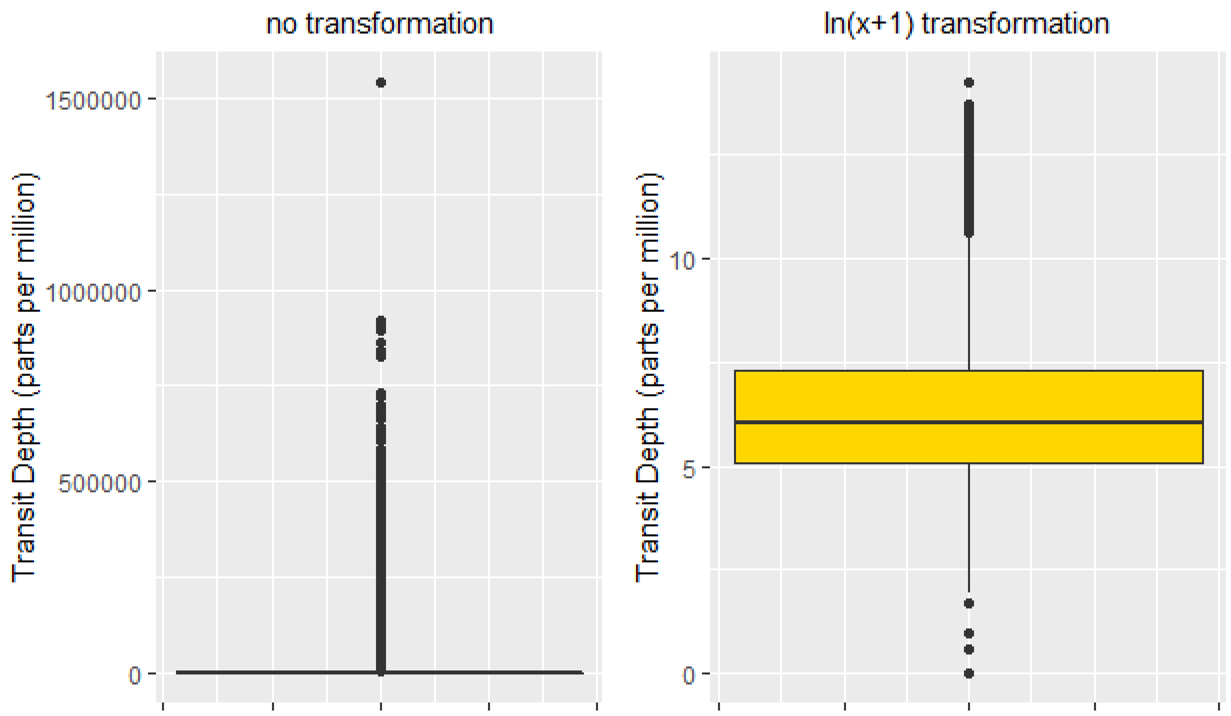
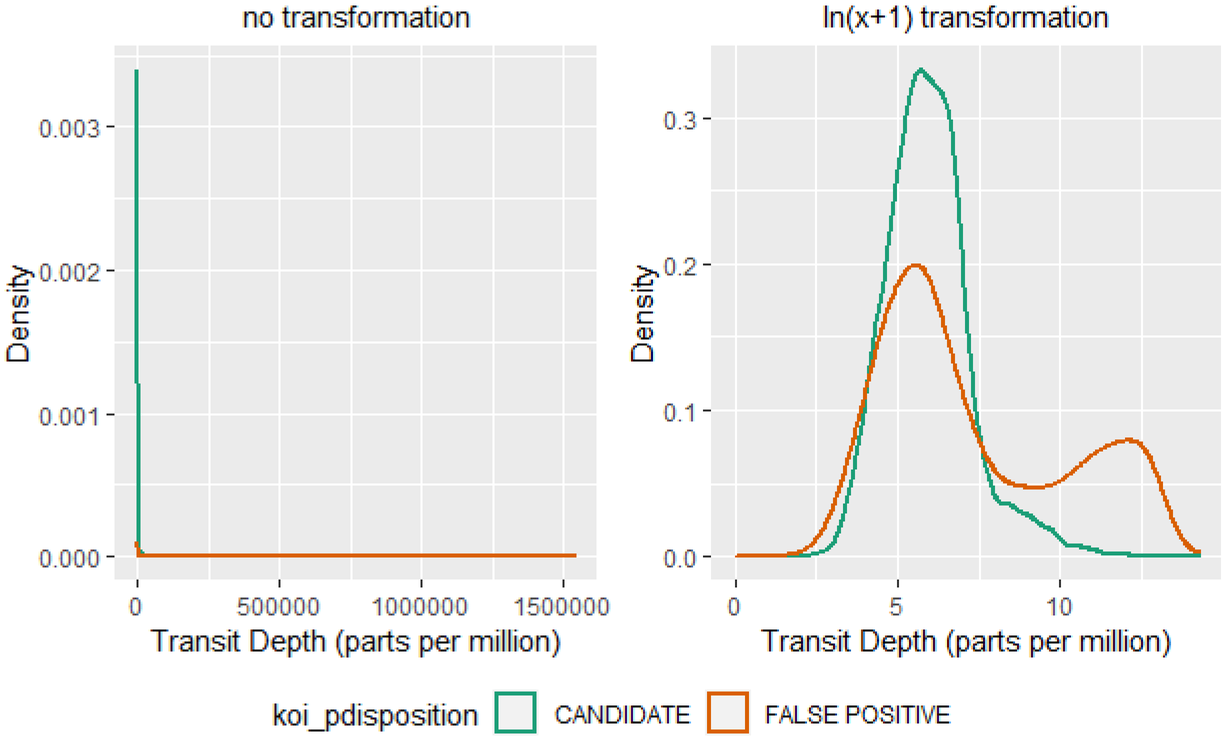


Figure 2 shows a similar comparison between untransformed and transformed distributions of the koi_depth variable using probability density plots the observed values grouped by response category (koi_pdisposition). In the plot on the left, the extreme right skew of the untransformed variable distributions makes it problematic to determine whether the shapes of curves for the two categories are distinct. However, in the plot on the right, the difference between the distribution shapes of the transformed variable categories is apparent. This divergence means that a machine learning model could better distinguish between observations belonging to each classification category when trained using the values of skewed variables after transformation.

Figure 2

Density Plots of 'koi_depth' Variable Distributions by Classification Category



CLASSIFICATION MODEL SELECTION

Several machine learning techniques are suitable for selecting a classification model with a binary response variable and numeric predictor variables. I used the characteristics of the KOI data set to narrow the possibilities and determine which methods to include in the model selection process. For example, statistical tests indicated that the distributions of the predictor variables were not multivariate normal before or after applying the $ln(x+1)$ transformation. Linear discriminate analysis and quadratic discriminate analysis both assume multivariate normality as a condition of model accuracy, so they were both excluded from further consideration. Similarly, the values of the linear correlation coefficients between the response variable koi_pdisposition and each of the predictor variables indicated only weak linear correlations (all $|r| < 0.5$). In the absence of pronounced linear relationships, models such as logistic and penalized regression would likely produce less accurate predictions than models better able to represent non-linear relationships. Therefore, I limited the model selection process to more flexible techniques, including k-nearest neighbors, support vector machines with a radial kernel, decision trees, boosted decision trees, random forests/bagging, and artificial neural nets. I used each of these methods to generate multiple models over a range of model parameter values, as listed in Table 1. I also employed 10-fold cross-validation to allow the models to make predictions on observations not included in model training. Finally, I calculated and compared the classification error rate of each model to determine which parameter values produced the most accurate results for a given technique.

Table 1

Model Selection Tuning Parameters and Ranges

Modeling technique	Tuning parameter	Parameter value range
k nearest neighbors	number of nearest neighbors (k)	integers: 1-30
support vector machine (radial kernel)	cost penalty for points misclassified or within margin	values: 0.01, 0.1, 1, 10, 100, 1000
	radial kernel constant (γ)	values: 0.5, 1, 2, 3, 4
decision tree	number of leaves	integers: 2-20
boosted decision tree	shrinkage parameters (λ)	values: 0.001 to 0.01 by 0.001 intervals
	number of splits in each tree (d)	integers: 1-3
	number of trees (B)	values: 500, 1000
random forest/bagging	number of predictors randomly selected each split	integers: 1-12
artificial neural net	number of ‘hidden’ nodes	integers: 1-20

The model selection process discovered a minimum classification error rate of 15.65% produced by a support vector machine using a radial kernel, cost penalty value of 1, and a radial kernel constant of 0.5. A random forest model had a comparable error rate of 15.68% with a modeling parameter for the number of randomly selected predictor variables available at each split of the trees set to 5. A k-nearest-neighbors model using 18 nearest neighbors (16.51%) and an artificial neural network with 18 hidden nodes (16.17%) had error rates within one standard deviation of the minimum rate produced by the support vector machine model.

Selection Process Assessment

After choosing a modeling method and its associated parameter values, I assessed the model selection process. To do so, I used an outer loop of 10-fold cross-validation to create randomly sampled subsets of the training data used to fit the models and the test data used to measure model accuracy when applied to new observations. I then applied the model selection process to each data fold to find the modeling technique that created the most accurate predictive model within that fold and calculate its classification error rate. To save time when running the script and reduce the number of model fits necessary, I limited the model selection process within each fold to the four techniques that had previously produced the best results on the complete data set: support vector machines, random forests, k-nearest neighbors, and artificial neural nets. These assessment procedures determined that the maximum predictive accuracy of any selected model was 84.61%, equivalent to a minimum possible error rate of 15.39%. The error rates of the support vector machine model and the random forest model chosen by the standalone model selection process were each near this minimum value. Both error rates were likewise an improvement over the null error rate of 48.18% produced by exclusive prediction of the majority classification category (FALSE POSITIVE for this data set). Considering these results, either model would be an acceptable choice for making predictions on truly new data.

Conclusion

The Kepler mission ended in 2018, but the model development process discussed above could be applied to data sets from future exoplanet surveys using whichever measurements of physical characteristics of the stellar systems are available. For example, the variables I included in the model formula of my selection process were limited to what NASA had posted on Kaggle. However, additional Kepler observational data accessible at the NASA Exoplanet Archive (NASA Exoplanet Science Institute, 2017) includes other variables, some of which could potentially improve model fit. Whatever the origin of the data set and the types of measurements available, it is critical to note that the models created and selected by this process are not intended to predict the ultimate disposition of objects of interest. Astronomers should only use the predictions to guide decisions of where to initially apply resources for follow-up observations, not definitively exclude any KOIs from further research.

References

Brown, T. M., Latham, D. W., Everett, M. E., Esquerdo, G. A. (2011). Kepler input catalog: photometric calibration and stellar classification. *The Astronomical Journal, 142*(4), 112. doi:10.1088/0004-6256/142/4/112.

NASA. (2017, October 10). *Kepler exoplanet search results (Version 2)* [Data file]. Retrieved July 27, 2019, from https://www.kaggle.com/nasa/kepler-exoplanet-search-results

NASA Exoplanet Science Institute. (2017, August 3). Data columns in Kepler objects of interest table. Retrieved July 28, 2019, from https://exoplanetarchive.ipac.caltech.edu/docs/API_kepcandidate_columns.html