Kepler Space Telescope Exoplanet Search Data Machine Learning Classification Project

Timothy Drexler, August 2019

Executive Summary

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 survey was to identify stellar systems in which periodic decreases of brightness were caused by the transit of extrasolar planetary bodies between the star and the telescope. Preliminary analysis of the light curve data found 9,564 potential planetary systems, known as “Kepler objects of interest” (KOIs). Follow-up observations made by ground-based telescopes are then used to determine whether the observed KOI signal is produced by a transiting planet or if there is another explanation, such as an eclipsing binary star or instrument noise. Using a data set containing information on the full list of KOIs, this machine learning project attempted to discover a classification model capable of predicting which KOIs were more likely to be confirmed as planetary systems and which were more likely to be false positives. Such a model could be used by astronomers to prioritize the limited telescope time available for follow-up observations and focus initial research on KOIs with higher potential to contain planets.

The KOI data set originally contained 50 variables, including the response variable of interest for this project: ‘koi_pdisposition’, a categorical variable with “CANDIDATE” and “FALSE POSITIVE” as possible values. The “CANDIDATE” label indicated that the KOI met criteria to be considered a potential planetary system pending further analysis, while the label “FALSE POSITIVE” meant that subsequent observations had confirmed the KOI signal was spurious. The status of each KOI was assumed to be current as of October 10, 2017, the date the data set was posted online (NASA, 2017).

The planned approach to developing the classification model was to use variables representing measured or calculated physical characteristics of the KOI to predict the value of ‘koi_pdisposition’. Preliminary data set exploration and consultation of the data dictionary indicated that many of the included variables would not be useful in this regard, so the number of predictor variables was reduced to 12. All the selected predictor variables contained continuous, numeric values, and fell into roughly two groups as summarized below (complete variable descriptions are available from the data dictionary at the NASA Exoplanet Archive (NASA Exoplanet Science Institute, 2017)).

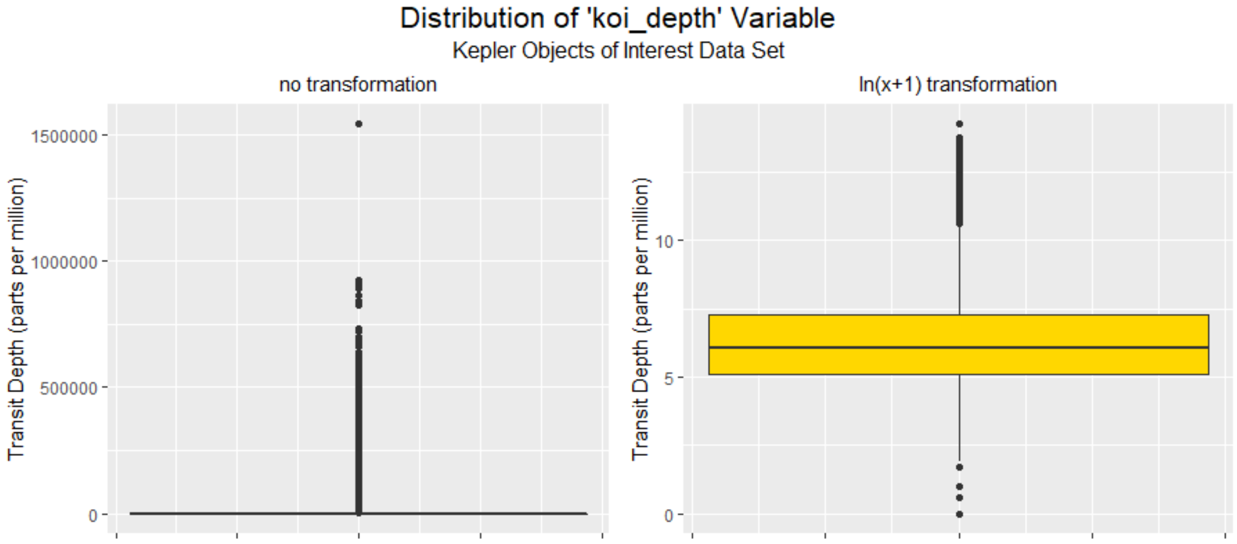
- The first group of variables measured properties of the potential planetary object and the observed transit event:
    - ‘koi_duration’, the transit time in hours;
    - ‘koi_period’, the time between transits, measured in days;
    - ‘koi_depth’, the fraction of stellar flux (energy per unit area) lost at the transit minimum (i.e., when the maximum amount of flux was lost);
    - ‘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;

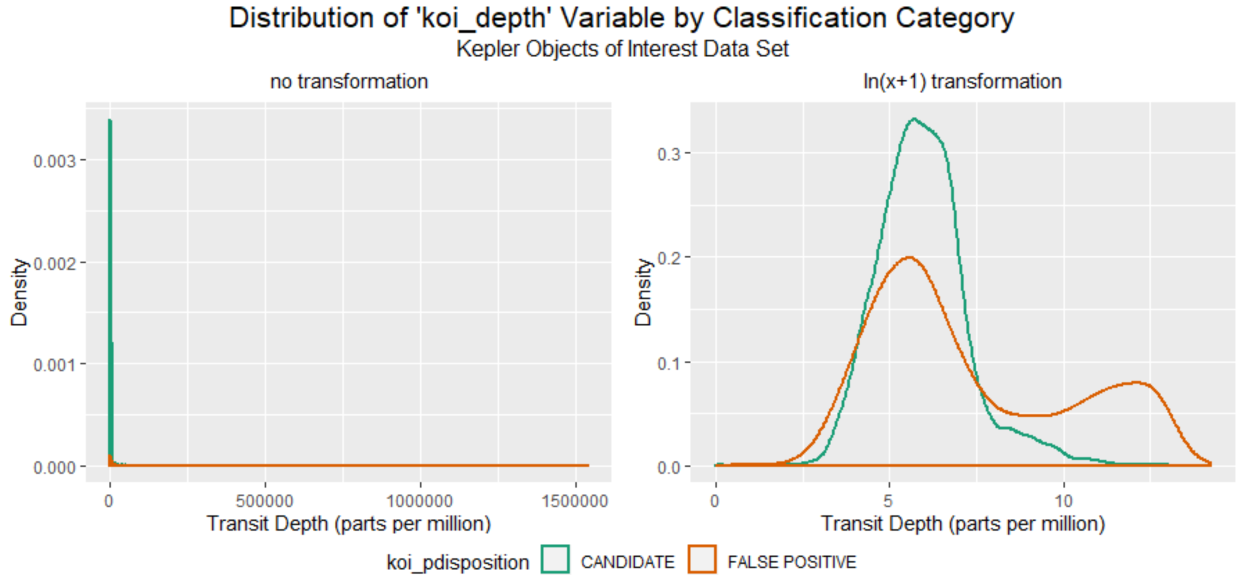
    - ‘koi_model_snr’, a calculation of the transit depth flux measurement normalized by the mean amount of uncertainty in the measurement; and
    - ‘koi_impact’, a normalized parameter estimating the distance between the center of the planet disc and the center of the stellar disc at conjunction.
- Variables 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; and
    - ‘koi_kepmag’, the stellar magnitude in the Kepler band.

After selecting the modeling variables, an exploratory analysis was made of the set of observations with data missing for at least one predictor variable. A majority of the 364 observations in this category were missing values for all but two or three of the predictor variables, making imputation of the missing data impractical. Since the incomplete observations represented less than 4% of the overall data set, and since there was no explanation available as to why the data was missing, it was decided to continue with the model selection process using only the 9,200 observations with complete data.

Further exploration of data set characteristics showed that the distributions of 11 of the 12 selected predictor variables deviated from normal: 10 variables had distributions skewed to the right, while the ‘koi_slogg’ variable distribution was skewed to the left. This suggested that transforming the predictor variable values would be beneficial, both to reduce the variance within each distribution and to decrease the amount of skew, rendering the distribution shapes closer to normal. The $ln(x + 1)$ transformation was chosen since multiple variables contained observations with a value of 0, and the transformation was applied to all predictor variables with right-skewed distributions, leaving ‘koi_slogg’ (left-skewed distribution) and ‘koi_kepmag’ (approximately normal distribution) untransformed.

Figure 1 below 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 distribution of the untransformed variable: most observations have values under 1,500 parts per million, but the observation values range as high as 1,500,000 parts per million. In the plot of the transformed variable distribution on the right, the range of values has been reduced to 0 to 14.248, a notable decrease in variance. Visual inspection also confirms that the distribution





shape of the transformed variable is closer to the symmetric shape of a normal distribution when compared to the right-skewed distribution of the untransformed variable.

Figure 2 (above) makes a similar comparison between untransformed and transformed distributions of the ‘koi_depth’ variable using density plots to show the probability density of the observed values grouped by response category. In the plot on the left, the right skew of the untransformed variable distributions makes it difficult to determine whether the shapes of the two categories are different. However, in the plot on the right, a distinct difference is visible between the shapes of the categorical distributions of the transformed variable. This suggests that a model created using the transformed values of ‘koi_depth’ would be better able to distinguish between observations belonging to each classification category, potentially increasing the model prediction accuracy compared to a model fit using the untransformed values.

There are several machine learning techniques suitable for selecting a classification model with a binary response variable and numeric predictor variables, so the characteristics of the KOI data set were used to determine which techniques 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 the application of 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 $|r| < 0.5$). As a result, the predictions produced by models which rely on linear relationships, such as logistic regression and penalized regression, were likely to be less accurate than predictions made by models better able to represent non-linear relationships. Therefore, the model selection process was limited 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. Each of these methods was used to generate multiple models with different values of model parameters as listed in Table 1 (page 4). The use of 10-fold cross-validation allowed the models to make classification predictions on observations not included in model training, and the resulting error rates were calculated and compared.

Modeling Technique	Parameter Tuned	Parameter Tuning 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

Table 1 – Tuning parameters used in the model selection process

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 produced a comparable error rate of 15.68% with the modeling parameter value of five for the number of randomly selected predictor variables available at each split of the trees. A k-nearest-neighbors model using 18 nearest neighbors (16.51%) and an artificial neural network with 18 hidden nodes (16.17%) both had error rates within one standard deviation of the minimum rate produced by the support vector machine. The model selection process was then assessed using 10-fold cross-validation to create randomly sampled subsets of both the training data used to fit the models and the test data used to measure model accuracy when applied to new observations. The assessment determined that the maximum predictive accuracy of any model selected by this process was 84.61%, equivalent to a minimum possible error rate of 15.39%. The error rates of the selected support vector machine model and the selected random forest model were each near this minimum value, and 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.

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 model formula used in the selection process was limited to variables included in the data set posted by NASA; however, additional Kepler observational data accessible at the NASA Exoplanet Archive (NASA Exoplanet Science Institute, 2017) may include other variables that could be added to the formula to improve model fit. Regardless of the origin of the data set, it is important to note that the models created and selected by this process are not intended to predict the ultimate disposition of objects of interest. The model predictions should be used only to guide decisions of where to initially apply resources for follow-up observations, not to definitively exclude any objects 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