Towards Discovering a Planet

Background

Our solar system is not the only planet-bearing stellar system in the universe. In recent decades, this has been proven by the emerging field of exoplanetology which seeks to identify and classify planets in systems beyond our own. One technique which has led to major discoveries in this field is the transit method, wherein astronomers measure the drop in brightness as seen from Earth of a star as one of its planets eclipses it.

Of course, to find a single star where a planet eclipses its parent, many stars must be watched for a long period of time and the resulting time-series data must be analyzed to see if there is, in fact, a drop. One of the best observatories which employs this technique is the Transiting Exoplanet Survey Satellite [TESS]. 27 days at a time, TESS looks at a large chunk of the sky and plots out the brightness of the brightest stars in the region.

All objects TESS observes are taken from a list of approximately 2 Billion TESS Input Candidates [TIC], which are filtered by stellar parameters down to a Candidate Target List [CTL] of approximately 9 Million stars which are theoretically likely to exhibit a valuable transit event during TESS observation. A selection of these, approximately 60k at the moment, have then observed by TESS and astronomers have analyzed the resulting light curves to determine which stars exhibited dips in their brightness during observation which were likely to have been caused by a transiting exoplanet. These stars, some of which are likely to contain planets, are placed on a growing list of approximately 1k TESS Objects of Interest [TOI] to be evaluated by follow-up observations.

Objective

The objective of this project is to build a classifier which can take in metadata about the drop in brightness of a star from the CTL along with parameters about its parent star and determine whether or not it belongs as a TESS Object of Interest. If the resulting algorithm performs well enough, the goal is to mine the set of observed CTL objects and find promising, previously unclassified, candidate TOIs which will be submitted to The Exoplanet Follow-up Observing Program for TESS [ExoFOP-TESS] as Community-identified TOIs [CTOI].

Data Collection and Characteristics

The initial set of features for the dataset consists of select stellar parameters which are likely to contribute to the likelihood of a star containing a planet of a size detectable by TESS, metadata recovered from the light curve, and the relevant object class (TOI or not TOI).

Star Selection

To collect the list of CTL objects for the dataset, first, data on all 1,183 official TOIs was downloaded from ExoFOP-TESS[1], which catalogs TICs, CTLs, TOIs, and CTOIs. Then data on the 4,000 objects in the CTL observed by TESS with the highest priority (likelihood of observing a transit event) were selected using the TESS CTL interactive viewer on Filter Graph[2]. A MATLAB script was used to eliminate from this CTL those entries which were missing features or had the same TIC identifier as an entry in the TOC list and remove any entries from the TOI dataset that were missing features. This script then thinned the CTL dataset down to 2,366 objects[3] not currently classified as TOI, combined it with the TOI dataset, and randomized the order of the objects in the table.

[1] https://exofop.ipac.caltech.edu/tess/view_toi.php

[2] Viewer: https://filtergraph.com/tess-ctl , specific source dataset: https://filtergraph.com/2258634

[3] Twice the number of official TOIs. This was chosen to keep the data to a manageable size.

Transit Parameter Calculation

Even though the TOI dataset contains transit parameters, the CTL data does not. Therefore a method must be deployed to capture these parameters from light-curves. To ensure that all that the features from both datasets are comparable, the transit parameters that came with the TOI dataset will be disregarded and the same method will be used to determine the transit values for both objects from both source datasets.

A python script was written to use astroquery to download FITS data of the Pre-search Data Conditioning Simple Aperture Photometry [PDCSAP][4] flux curves for each object from the Barbara A. Mikulski Archive for Space Telescopes [MAST] collection of TESS data and then use astropy to convert all the data to csv files containing just the flux and time data points for the transit observations of each object.

Since the very nature of transit is a sustained drop in brightness, a histogram of the light curve for a start exhibiting such an event would show two distinct peaks. So, MATLAB script was written to comb through all the light curves, and extract the number of peaks in the histogram and their mean separation.

Resulting Features

After the above data processing, the resulting dataset contains 3,454 documents with the following features.

Features		Description
	class	Whether or not the object has been officially identified as a TOI.
Stellar Features	M	Magnitude of star as observed by TESS.
	T_{eff}	Effective surface temperature of star being observed (in Kelvins)
	R_*	Radius of star being observed (in solar radii).
	L_*	Luminosity of star being observed (in solar units).
Transit Features	N_{peaks}	Number of unique peaks in the histogram of the PDCSAP flux (light) curve (N=2 implies transit observed).
	$D_{peaks} = \frac{\Delta F_{peaks}}{\max F_{peaks}} 10^6$	PPM relative difference in flux between peaks in histogram of the PDCSAP flux curve (or 0 if not applicable). Essentially, this is the transit depth which is proportional to the square of the candidate planet's radius.
	TIC ID	TESS Input Catalog Identifier. This will not be used by the algorithm but will be used during error to track down further information about problematic objects (such pulling up their light curves).

Dataset Partitioning

This dataset was then randomly split divided into training dataset (1/2 of full dataset), a testing dataset (2/6 of full dataset), an a final testing dataset (1/6 of full dataset). In addition, a further 50,000 observed objects[5] not classified as TOI were extracted from the CTL using Filter Graph to be used once the model has been trained and refined in order to attempt to find candidate TOIs which have not yet been identified.

Baseline Performance

A basic support vector machine was setup in Weka using LibSVM. Using all default settings and the testing dataset as the supplied test set, this correctly classified 75.16% of instances *but*, given the structure of the training data, this is not impressive and not that is confirmed the abysmal kappa of 0.0574 and by the confusion matrix which showed that the algorithm simply classified the vast majority of the elements as "not TOI". Clearly, not only is a lot of tuning required but also the following improvements to the feature set.

[4] Among other improvements, PDCSAP filters out long-term trends in the SAP flux curves.

[5] https://filtergraph.com/2932740

Plans for Handling the Data

Potential Major Sources of Error

One potential issue with the data is that it's not perfectly segmented. There's a very high likelihood a number of the stars not labelled as TOI should be labelled TOI and just haven't been uncovered yet; in fact, that's the very point of building this classifier.

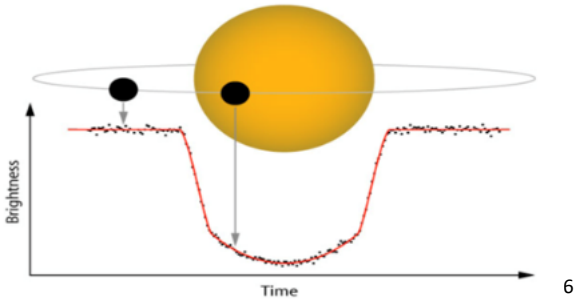
Ideally, the appropriate way to get data to train this classifier would be to be use a list of observed CTL objects which have examined by astronomers and thereby labelled definitively as "TOI" or "non-TOI". Unfortunately, a public dataset meeting these criteria of sufficient volume could not be found. That said, of the ~9M objects in the CTL sample space, only ~1k have been classified as TOI so far; so, for now, this error is going ignored by assuming that the number of misclassified "non-TOI" objects is small within the dataset.

If this does present an issue, one potential remedy would reselected the stars which comprise the "non-TOI" class to include medium or low priority stars from the CTL. The downside of this, however, would be that many of these objects, if they're stars at all, don't have easily available stellar parameters.

Planned Improvements

Currently, the feature space incorporates very few and simple statistics on the light curves. This is currently in the process of being improved but the data processing still has several more days before its finished. The planned improvements are as follows:

Since, very broadly speaking, a transit event appears in light curve as a fast decline in brightness, followed by a period sustained decreased brightness, and then a fast rise in brightness back up to the original value, which all occur at a regular interval as shown below.



[6]

While there certainly are other characteristics to a transit event such as orbital inclination, limb darkening, and a second eclipse when the planet hides behind the star, if one is simply trying to detect the presence, relative depth, and frequency of exoplanet-like transit events and not recover actual properties of a transiting exoplanet, these factors can be ignored and a transit can be modelled as a basic trapezoidal drop in brightness as shown below.

$$F(t) = \begin{cases} F_0 - \dfrac{\Delta F}{t_r}(t - t_0 + iT_t) & \text{if } t_0 + iT_t \le t < t_0 + iT_t + t_r \\ F_0 - \Delta F & \text{if } t_0 + iT_t + t_r \le t < t_0 + iT_t + t_t - t_r \\ F_0 - \dfrac{\Delta F}{t_r}(t_0 + iT_t + t_t - t) & \text{if } t_0 + iT_t + t_t - t_r \le t < t_0 + iT_t + t_t \\ F_0 & \text{else} \end{cases} \quad \text{for } i \in [0, N-1]$$

[6] https://www.apus.edu/academic-community/space-studies/exoplanet-transit-photometry

where F_0 is baseline value of stellar flux, ΔF is the drop in brightness, t_0 is the start time of the first transit event, t_r is the ramp time for the drop/rise in brightness, t_t is the total duration of a transit event, T_t is period of transit events, and *N* is the total transit events observed in the light curve.

For each light curve, this model is fit to the data using MATLAB's *fmincon* on the least squared error between the model and the data. From the fit, the following parameters and statistics will be added as features:

Feature	Description
$\boldsymbol{R^2}$	Root mean-squared error of model's fit to the light curve.
$\boldsymbol{T_t}$	Observed transit period.
$\boldsymbol{N}$	Number of transit events observed.
$\boldsymbol{D = \frac{\Delta F}{F_0} * 10^6}$:	PPM depth of transit.
$\boldsymbol{t_t}$	Transit duration.
$\boldsymbol{t_r}$	Transit rise time.

After these features have been added, a new baseline model will be trained and error analysis be performed to determine which, if any, of these features is not contributing to / hurting the model's performance. Once this is done, different kernels will be tried for the support vector machine and parameter tuning will be performed. It is likely that this process will change which features are significant; so, this whole process will be repeated iteratively until a satisfactory model has been produced. At this point the model will be benchmarked using the final testing dataset.
If these benchmarks reveal significantly better performance than a random classifier, the classifier will be deployed on the set of 50,000 observed CTL objects which have not been identified as TOIs in an attempt to find the small number of objects in that set which should be classified as TOIs.