Searching for Planets in the Kepler Data Set

This project intends to predict with confidence whether or not a target star in the Kepler K2 "Second Light" Survey) has an associated exoplanetary system.

As a bonus we will explore the "Inspirational Questions" proposed by NASA:

> - *How often are exoplanets confirmed in the existing literature disconfirmed by measurements from Kepler? How about the other way round?*
> - *What general characteristics about exoplanets (that we can find) can you derive from this dataset?*
> - *What exoplanets get assigned names in the literature? What is the distribution of confidence scores?*

Goals & Success Metrics

This is the ultimate needle-in-a-haystack problem with many risks and limitations that we will elaborate on below. Thus the primary goal is find a method or combination of methods to cluster and predictively model if the target Kepler Object of Interest *(KOI)* is a positive candidate with confidence. The open-source community has yet to find a methodology that outperforms arbitrarily assigning "FALSE POSITIVE" all KOIs. To achieve this would be exemplary, however we hope to find important pieces that moves the open source community closer developing a method that can achieve this.

Risks & Limitations

Data Imbalance: In its 9 1/2 years of operation (2009 to 2018), The Kepler Space Telescope observed 500,000 stars and has confirmed the existence of 2,662 planets. The telescope is designed to to detect exoplanets by way of the *Transit Method* which requires the planetary system to be oriented such that during the period of observation that the planet passes between the telescope and the host star. While the observations and predictions made by astronomers and astrophysicists seem to suggest by now that planets are very common in the universe, this particular orientation is relatively rare. Thus, before looking at the data directly, we can assume the dataset to have extremely unequal class distribution. A model that predicts the majority class across all input observational data of the Kepler Space Telescope Mission would produce a positive prediction rate of ~99.5% (~99.1% accuracy over the *Campaign 3 Test Data Set*). To work around this we have a few tools to our disposal:

1. Design a cost function — e.g. modifying a *Support Vector Machine* — that penalizes wrong classifications of the rare class (transitory planet observed = 1) by a similar ratio of the class populations.
2. Under-sampling the training set.
3. Over-sampling the training set by means of repition or *Synthetic Minority Over-sampling TEchnique (SMOTE)*
4. Cluster the majority class into **n** unique groups of varying size where **n** = The Minority Class Population. Then an unmodified two-group classification model may be trained on the minority class and the mediods of each cluster.
5. Evaluating *precision, recall,* or *F1 score* as opposed to only looking at prediction accuracy.

The above list is not exhaustive, nor is it exclusive. However, it is a solid starting point for tackling the data imbalance inherent in observing exoplanets.

Data Sources

Target Pixel Data: Calibrated flux measurements for the individual pixels of each target star. — *ReadMe* *Mikulski Archive for Space Telescopes*

Light Curve Data: Change in Flux over time across all targets and all campaigns of the K2 Mission. — *ReadMe* *Mikulski Archive for Space Telescopes*

Objects of Interest Data: List and descriptions of all Kepler Objects of Interest *(KOI)* — *Dictionary* *NASA Exoplanetary Archive*

Candidates Table: List and descriptions of all Kepler Objects Candidates for an exoplanetary system — *Dictionary* *NASA Exoplanetary Archive*

Training and Testing Data: Cleaned and grouped data from K2 Campaign 3 — *ReadMe* *Mikulski Archive for Space Telescopes & WΔ*

In [ ]: