Project proposal: exoplanet stars identification

Members

Hadrien Lacroix

Project description

Exoplanets are planets outside of our solar system. Despite a first possible evidence for the existence of exoplanets in 1917, their existence has only been officially confirmed in 1992. Since then, thousands of exoplanets and planetary systems have been identified. In a nutshell, exoplanets are detected by observing the luminosity of stars and seeing whether or not they dim at a regular interval. If so, it's possible that this dim is caused by a planet orbiting around it. Identifying exoplanets is the first step to then potentially finding life outside of our solar system (if you're unfamiliar with the Drake equation, once planets are identified, we can investigate whether they could support life, have actually developed it, potentially an intelligent form, and potentially one that can release signals into space).

The goal of this project is to identify stars that are orbited by exoplanets. This will be a classification project, and our goal will be to test different methods and algorithms to obtain the best accuracy possible for predicting whether a star is orbited by planets or not. This will be a classification problem.

Data

The dataset we will use will, at minimum, be the one that was made available [here](Exoplanet Hunting in Deep Space I Kaggle) by Winter Delta. Each observation is a star, and each of the 3198 columns is a light intensity observation recorded for this star at different times.

However, ideally, we would use the entirety of the Kepler mission dataset for our purpose. The dataset provided by Winter Delta counts 5,657 star observations in total, but the Kepler mission monitored over 100,000 stars over its course. The whole data for the Kepler mission is available on the Mikulski Archive for Space Telescopes website, so we'd like to consider all of the campaigns instead of just one for this project.

Doing so will require some additional data preprocessing effort on the light curves data to match the structure of the Winter Delta dataset, and to convert `.fits` data to `.csv`. We will rely

heavily on the Kepler Data Processing Handbook for this effort.

It's unlikely that we have the bandwidth to do so, but more data could be added on top of the simple light intensity observations to feed more features to our model (full frame images, target pixel files...). However, this would lead us to include image analysis and computer vision techniques, which might lead us too far from the initial goal for this project.

The data might also be enriched with K2 campaign data, K2 being the mission that followed into the Kepler mission footsteps.

Scientific research questions

We're working with data that has already been reviewed by astrophysicists. For the data we're using, the experts have already managed to decide with more or less confidence whether each star observed is orbited by exoplanets or not. The purpose of the project is to leverage this human expertise to build a model that can speed the process in the future.

One of the main disadvantages of such a dataset is how heavily imbalanced it is. Most stars are unfortunately not orbited by a planet, so most of our dataset will consist of non-exoplanet stars. We will need to research and implement methods to deal with this imbalance in order to minimize its impact on our model. Otherwise, the model could systematically predict that a star is not orbited by an exoplanet and still boast an incredibly high accuracy score - while remaining completely useless and failing to be of any help in future identification efforts.

To deal with imbalance, we plan to research the following techniques:

- random over-sampling (over-representing the minority class)
- random under-sampling (under-representing the majority class)
- synthetic Minority Over-sampling Technique (SMOTE, which we need to research further to see if and how it can be applied to categorical variables)
- how to properly leverage K-fold cross-validation when oversampling
- how we can resample the dataset to reuse the same cases or the rare class but randomly sample from the majority class, and how we can fine-tune this approach with different rare/majority class ratios
- how we can use clustering to reduce the number of majority cases to the center of a cluster
- how we can potentially leverage the XGBoost algorithm (provided our data distribution allows it) as it internally takes care of imbalance

- how we can use voting algorithms to potentially leverage the strengths of different algorithms

We will also need:

- to decide which metrics make the most sense when evaluating our model (precision, recall, F1, MCC, AUC)
- to learn more about the `.fits` format
- to evaluate whether using time series techniques on the light intensity observations at different point in time can prove useful at all