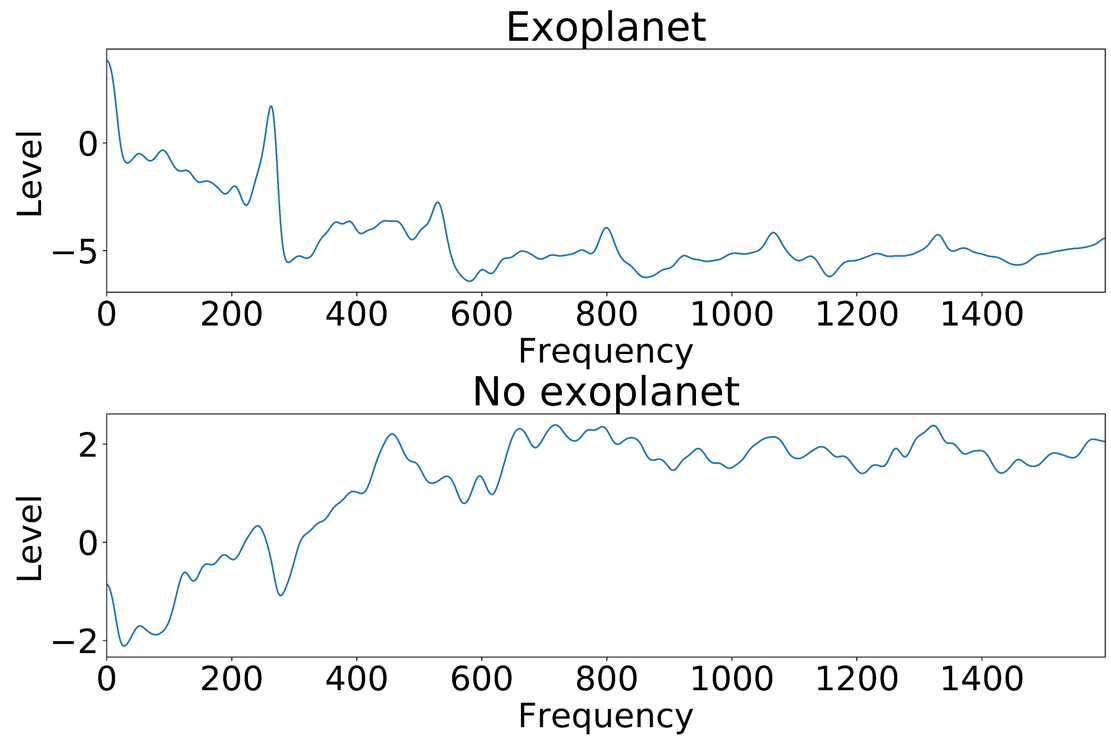
Exoplanet detection using the transit method

Authors: Marko Raidlo, Raido Everest

UNIVERSITY OF TARTU

• Introduction

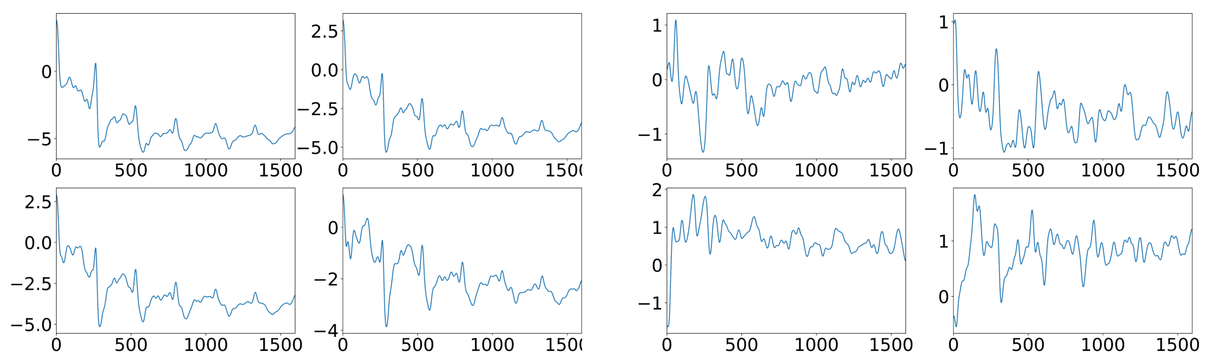
An exoplanet is defined as a planet which orbits a star outside our own solar system. Exoplanets are a field of research for astronomers and astrophysicists to gather more data on different planetary systems and make better theories on system formation and evolution. The transit method (aka. Transit photometry) is the most widely used method of exoplanet detection. The light curve of a distant star is measured for periodic dips in brightness. Usually, these are caused by exoplanets transiting in front of the star.



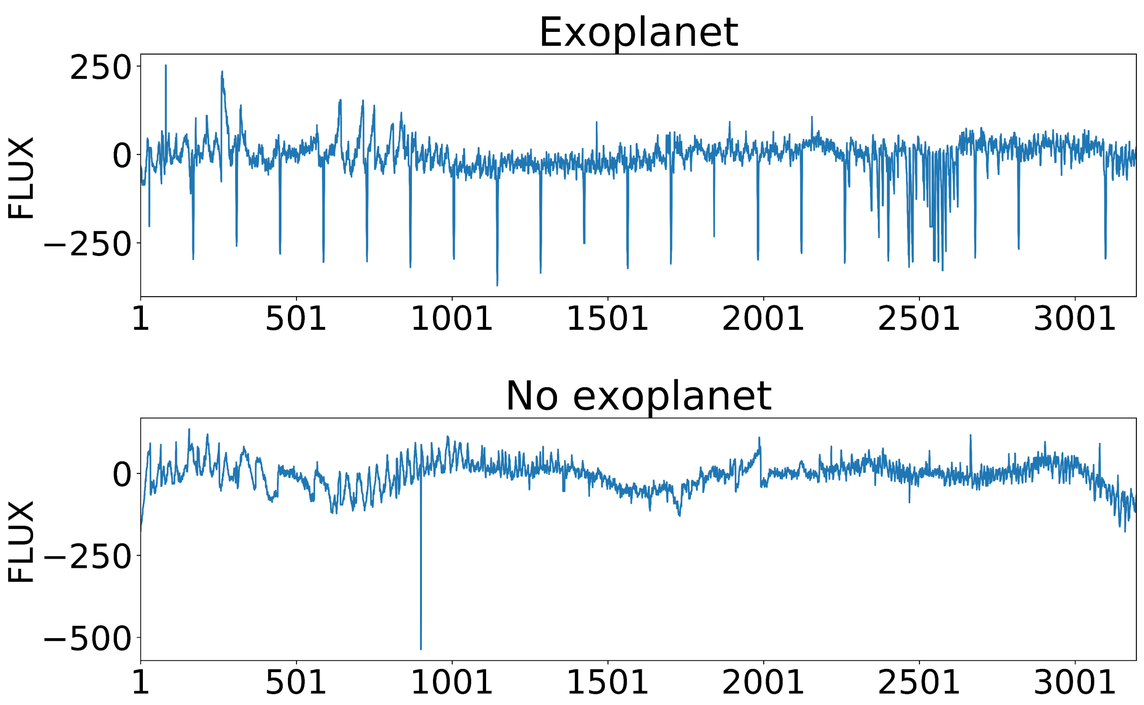
Caption: Post processing data

• Data

The light curve data comes from the Kepler space telescope, publicly available on NASA exoplanet archive. For this project, Kaggle dataset was used, which uses Kepler Campaign 3 data. Kaggle dataset has been made much more usable than raw data from NASA, as it has been made into a .csv file, while as NASA provides raw .fits files. The datasets include indicator of exoplanet existence and 3197 flux measurements. The training set contains 5087 observations with 37 confirmed planets and the test set contains 570 observations with 5 confirmed planets.



Caption: Examples of exoplanets (left) vs non exoplanets (right)



Caption: Raw light curve data

• Preprocessing

1. Fourier transform
2. Normalization
3. Gaussian filter
4. Standardization

• Building models

Following models were tested on the dataset:

- Logistic Regression
- Support Vector Machine
- K-Nearest Neighbours
- Decision tree
- Random forest

• Conclusions

Best results were provided by the random forest models. With a small amount of decision trees in the forest, the model managed to gain 1.0 recall and precision on tests. We find that this model is very efficient way to detect planets with short orbital periods, as they are detected several times during a measurement campaign. Planets with periods longer than measurement campaigns need more work as repeat detections on the light curve are needed for conformation of the exoplanets existence. Below are the confusion matrices for the random forest model.

	Predicted False	Predicted True
Actual False	5050	0
Actual True	0	37

	Predicted False	Predicted True
Actual False	565	0
Actual True	0	5

Confusion matrixes for training and testing datasets.

← GitHub link