Exoplanet Ranking

Introduction to Data Science

Goron, Nathan
nathan.goron@helsinki.fi

Motzkus, Franz
franz.motzkus@helsinki.fi

Vopat, Tomáš
tomas.vopat@helsinki.fi

October 22, 2019

1 Introduction

As mankind faces environmental problems nowadays, several approaches come up with the solutions. One way is to modify our habits to act softly with nature, the other way is to find another place to new inception.

In this project, we aim to find an appropriate exoplanet (a planet in another solar system) that might be good enough for life. To deal with this issue, we use data about known exoplanets and AI techniques. The domain of this attempt is to provide an answer whether there is the planet in the universe similar to the Earth or not. To achieve the goal we compute the similarity rank for each exoplanet.

This study might interest companies, that want to raise people's curiosity around the topic of finding Earth-like planets, or simply astronomy enthusiast, that we can inspire with a new view.

2 Preprocessing and Analysis

The dataset [1] we used is from an archive hosted by the NASA exoplanet science institute which is operated by the California Institute of Technology. In the dataset, there is data about 4 056 exoplanets with 47 various attributes (e.g. temperature, star luminosity, distance). These attributes are used to determine the "livability score" since they influence if humanity could survive on this planet.

As a first step, we removed the columns, that contained irrelevant information for our goals such as website links or different names of the exoplanet. Next step was to drop columns with a high ration of missing data. In the column `fpl_eccen` is 61,7 % missing and in the column `fst_spt` 66,3 %. As a result only 20 columns left for the ranking, these attributes are deeply described in Figure 1. Since the domain knowledge of physicists, astrologists and maybe

even biologists are needed, we use all the attributes mentioned above for our experiments. For further experiments, it is supposed to consult with a specialist the livability impact of each attribute in the dataset, as some of those might be more important than the others.

Afterwards, we explored the distributions of each variable. As some of the variables vary a lot, some of them are almost constant across all the dataset. To make it more clear, we plot the mean value of each variable with its standard deviation in Figure 2. As can be seen, attributes like `fst_mass` (stellar mass) or `fst_lum` (stellar luminosity) have small in comparison to others. On the other hand, `fst_teff` (effective temperature) and `fpl_orbper` (orbit period) differ a lot. To be able to plot all the variables, it was necessary to log-scale all the values.

Furthermore, some of the variables might be correlated (mutually influenced). Due to that, we have computed correlation matrix for each pair of variables. As illustrated by Figure 3, `fst_logg` (gravity acceleration at the stellar surface) and `fst_lum` (stellar luminosity) are highly correlated, which means that with the gravity acceleration the luminosity of the star decreases.

To finish the analysis, we try to group exoplanets by its properties, as the planets are situated in the vector space (not in the universe). First of all, we have to normalize all the values to be in the same range (from 0 to 1). After that, the clustering can be applied. We use the K-Means algorithm implemented in *Scikit-learn*[1] library for machine learning. Surprisingly, the optimal number of clusters is 2, since there are 4 thousand exoplanets in the dataset.

We can see in Figure 1 that the Earth is situated in the green cluster where is 77,5 % of the exoplanets. The planets are distinguishable even though space is projected by Singular-Value Decomposition (SVD, *Scikit-learn*[1] implementation is used) into 2 dimensions.

3 Exoplanet Ranking

To compute the ranking we use vector similarity between each exoplanet and the Earth. All those vectors are obtained from the Exoplanet dataset. The vectors are preprocessed, so there is not any missing data and the values are normalized. The ranking output is a value for each exoplanet in the range from 0 to 10 (rescaled by multiplying) where 10 means the most similar planet.

Cosine distance implemented in *Scikit-learn*[1] library is used as a similarity measure between vectors. The presumption is that the most similar exoplanets to the Earth are the most likely sufficient for life. The exact formula is displayed in Formula 1.

$$\text{cosine}(u, v) = \frac{\sum_{i=0}^{n-1} u_i v_i}{\sqrt{\sum_{i=0}^{n-1} u_i^2} \sqrt{\sum_{i=0}^{n-1} v_i^2}} \tag{1}$$

[1] `https://scikit-learn.org`

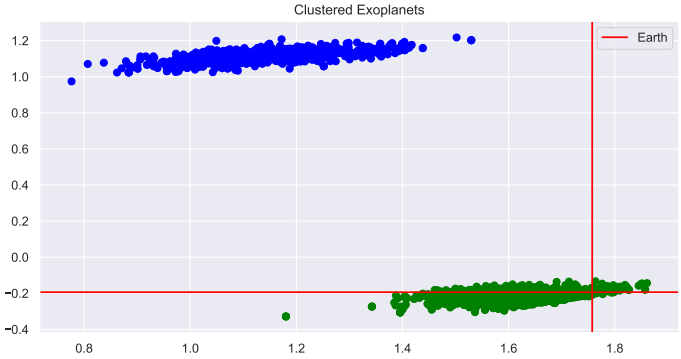

Figure 1: K-Means Clustering the Exoplanets projected by SVD into 2 dimensions

It is supposed that some parameters have a bigger impact than others on how "livable" an exoplanet is. For instance, the metal abundance is probably not as important as surface temperature. Therefore two rankings have been computed – the second one takes into account the weight of the parameters.

We created a weight vector w that can be seen in Table 1 with weight in the range from 1–10 for each attribute of exoplanets. To compute the weighted cosine similarity the library *SciPy*[2] is used. Where it is implemented as a correlation between vectors with weighted averages as describe Formula 2 and Formula 3, where u and v are the vectors of exoplanets.

$$\text{cosine}_{\text{weighted}}(u, v, w) = \frac{\sum_{i=0}^{n-1}(u_i - \overline{u_w})(v_i - \overline{v_w})}{\sqrt{\sum_{i=0}^{n-1}(u_i - \overline{u_w})^2}\sqrt{\sum_{i=0}^{n-1}(v_i - \overline{v_w})^2}} \tag{2}$$

$$\overline{u_w} = \frac{\sum_{i=0}^{n-1} u_i w_i}{\sum_{i=0}^{n-1} w_i} \tag{3}$$

4 Web App

For our product, we created a Node.js[3] web application to display the results of our ranking. The website uses native HTML/CSS/JS with Bootstrap[4] classes, a node.js template engine called EJS[5] to pass data to views as well as some data visualisation frameworks such as chart.js to make it easier to display our data.

[2] `https://www.scipy.org`
[3] `https://nodejs.org`
[4] `https://getbootstrap.com`
[5] `https://ejs.co`

The Web App displays the ranking, so the user can directly compare both variants of the ranking (the standard and the weighted one). Underneath there is a bar chart showing the number of exoplanets discovered in each year from 1991 to 2019, where a remarkable amount of exoplanets was discovered in years 2014 and 2016. In the end, the distance of the top 20 planets is displayed. All the plots and the tables react to the user's mouse hover to show more details about the specific object.

Design is meant to be a little mysterious to catch the eye of potentially interested people since the website is meant to be presented on fairs or other events. The homepage shows features like the caption and an image of a planet on the colourfully modified background to be quickly recognized by astronomy enthusiasts and other people enjoying the wonders of outer space.

Currently, the website can be reached at `http://ec2-52-201-232-146.compute-1.amazonaws.com:3000`.

5 Results

According to our ranking, the most suitable exoplanet is `K2-85 b` with a score of 9,96 for the regular ranking and 9,94 for the weighted ranking. The temperature of 351 K, which is a lot for a human to withstand without proper equipment, but looking at the repartition of the temperature column in the dataset, this is a pretty good result. We can also note that the density is very similar to Earth's, and therefore the gravity wouldn't make feel much different for a human.

On the other hand, the worst planet for the weighted ranking is `BD+20 2457 b`, with a score of 4,21 out of 10. In addition to being 4 762 pc away from Earth (it would take 13 821 years travelling at the speed of light to reach it), it's surface temperature is about 2 370 K, which makes it a really bad choice to be mankind's next home.

Even though providing a reliable ranking would require a deeper knowledge of astronomy, we can safely (but naively) say that our ranking gives a good glimpse of what could be mankind's next home.

References

1. NASA EXOPLANET ARCHIVE. *Composite Planet Data* [online]. 2019 [visited on 2019-10-21]. Available from: `https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=compositepars`.

A Dataset Attributes

Table 1: Attributes Description

Attribute	Description	Unit	Weight
fpl_orbper	orbital period	days	5
fpl_smax	the longest radius of an elliptic orbit	AU	3
fpl_bmasse	mass of the planet	Earth mass	5
fpl_rade	planet radius	Earth radii	5
fpl_dens	density of the planet	g/cm^3	7
fpl_tranflag	wether the lanet transits the star or not	bool	1
fpl_cbflag	does planet orbit a binary solar system	bool	5
fpl_snum	number of stars in the solar system	integer	8
dec	declination of the planetary system	decimal degrees	3
fst_teff	effective temperature	K	10
fst_logg	gravity acceleration at the star surface	$\log_{10}\left(\mathrm{cm}/s^2\right)$	4
fst_lum	star lumonisty	$\log_{10}\left(\textit{Solar luminosty}\right)$	4
fst_mass	stellar mass	Solar mass	6
fst_rad	stellar raidus	Solar radii	6
fst_met	star metallicity	dex	3
fst_metratio	metal abundance	[Fe/H], [M/H]	1
fst_age	stellar age	Gyr	8

B Charts

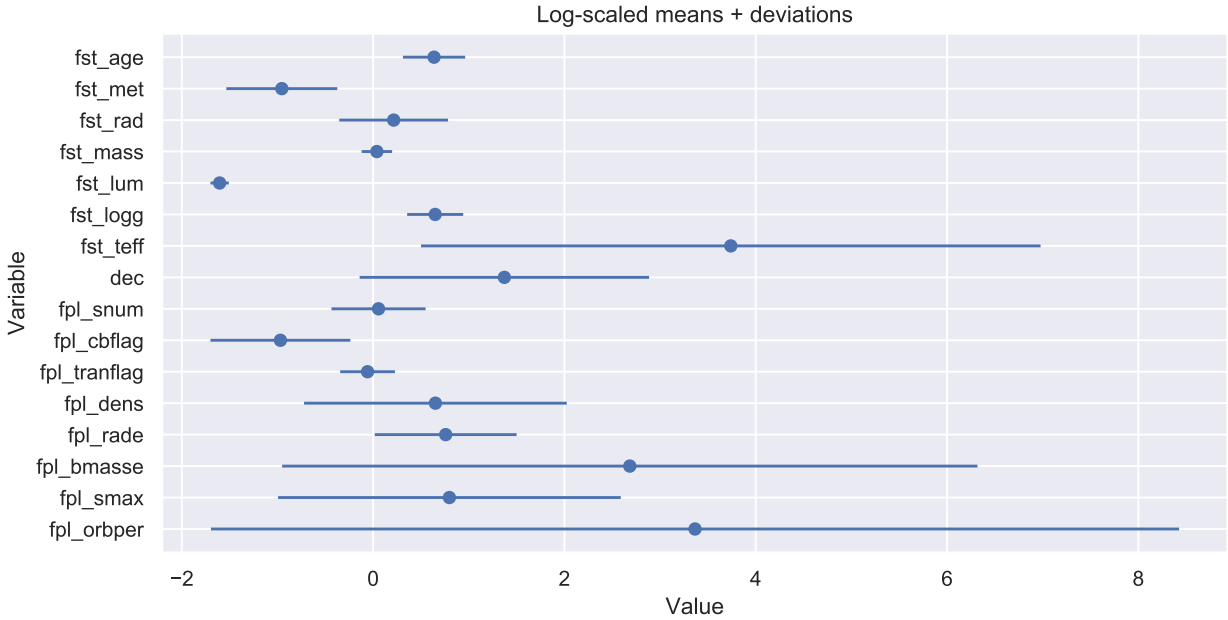

Figure 2: log-scaled mean of each variable with its standard deviation

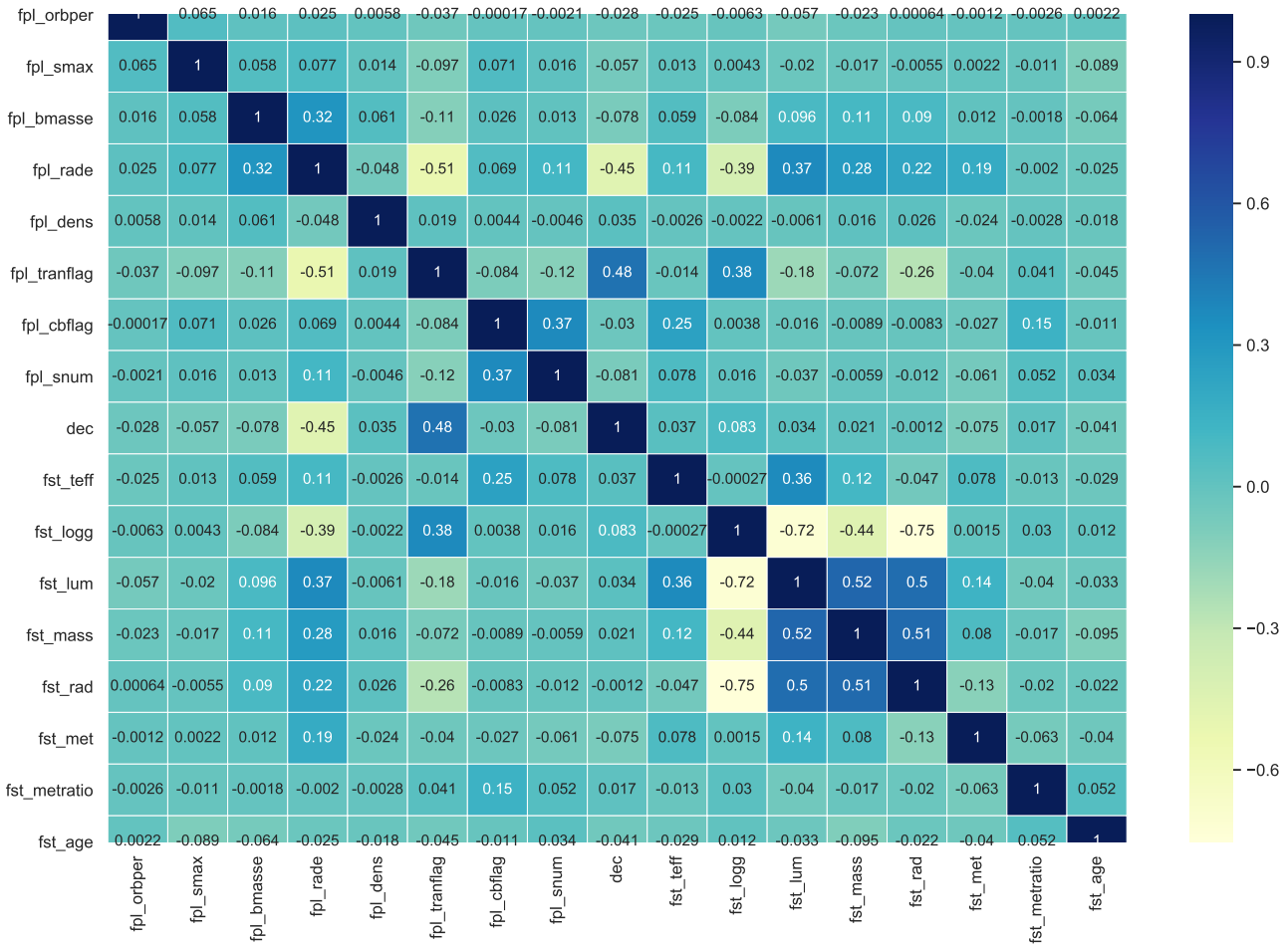

Figure 3: correlation matrix of variables