1. Une image contenant texte

   Description générée automatiquementPresentation of the challenge

The aim of this project is to come up with a model/algorithm to produce predictions on T1\_RETURN using the available information.  
  
You are expected to use said predictions to form a portfolio that will then be evaluated according to Information Ratio.

**As an output, you are asked to provide:**

* A short report on what you have done together with some results and conclusions, including a performance comparison among the models you propose, using Information Ratio as a metric.
* The source code used for producing your analysis and results.

The project will allow us to evaluate your skill set within programming, data preparation, modeling, and communication.

Une image contenant texte

Description générée automatiquement

# Collect & transform data

First, we inspect data types:

Une image contenant texte

Description générée automatiquement

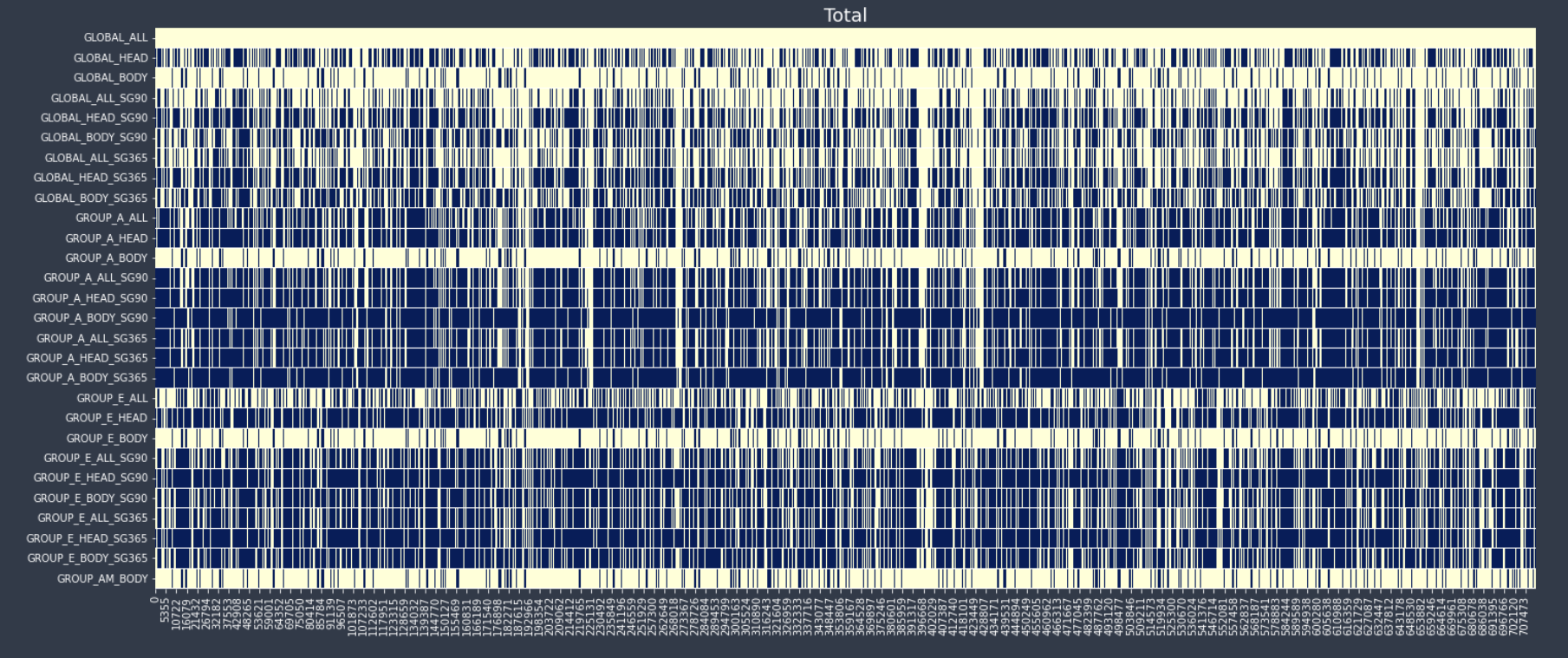
Then, we inspect missing values:

Une image contenant texte

Description générée automatiquementUne image contenant texte

Description générée automatiquement

Missing values don not show any evident pattern (temporal or other depedencies):

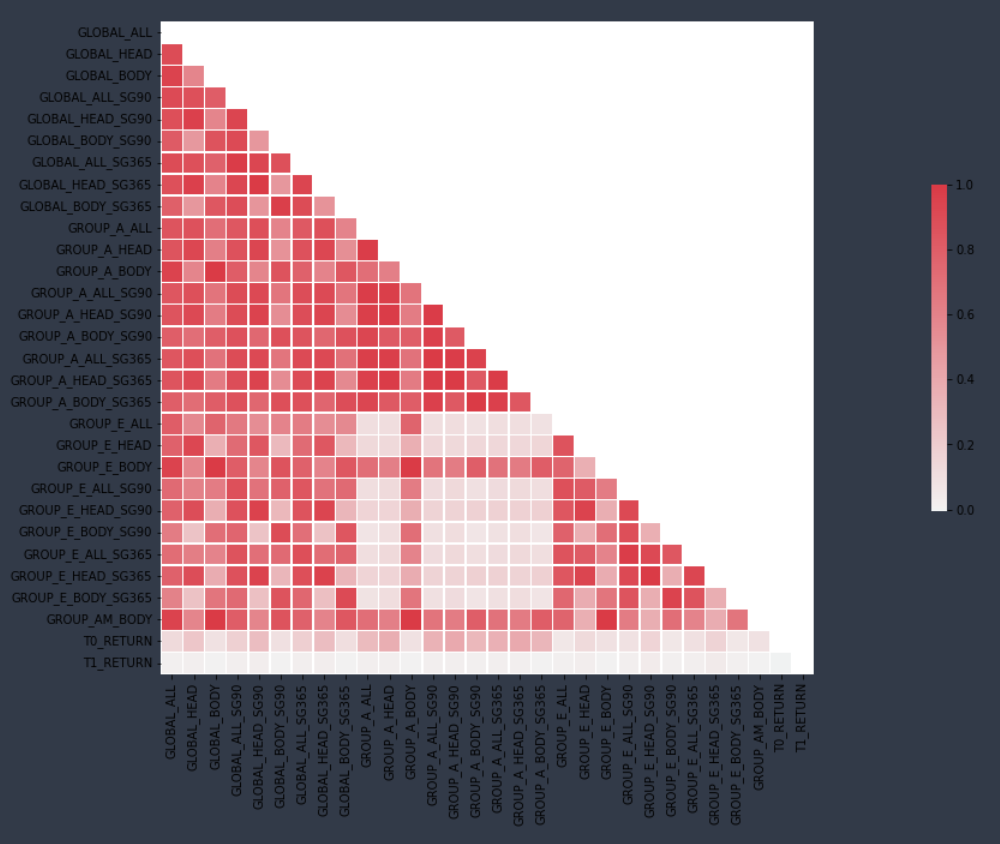


Features value seem to be fairly evenly distributed between -1 and 1

Une image contenant ciel, fenêtre

Description générée automatiquement

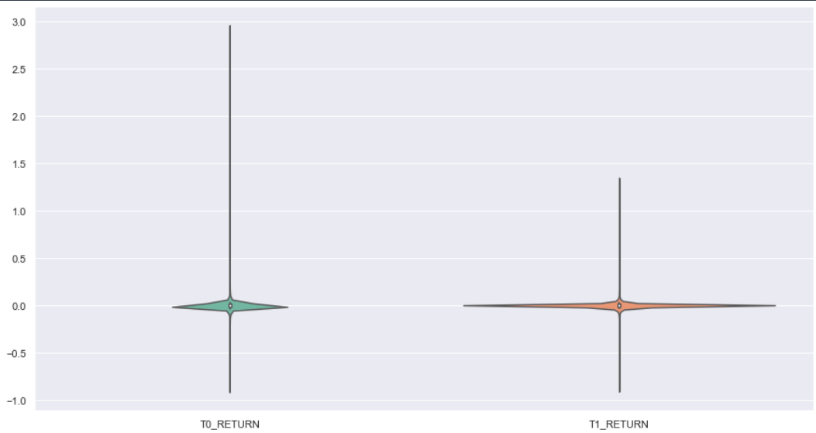
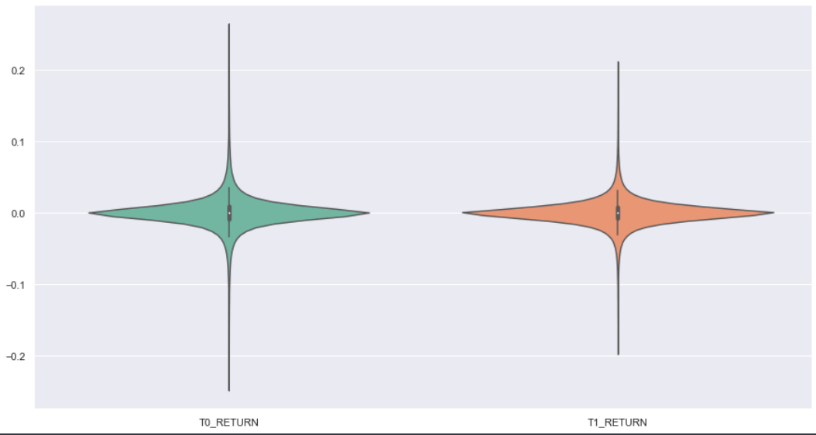
Features seems correlated with each other, and not correlated with T0\_Return nor T1\_Return



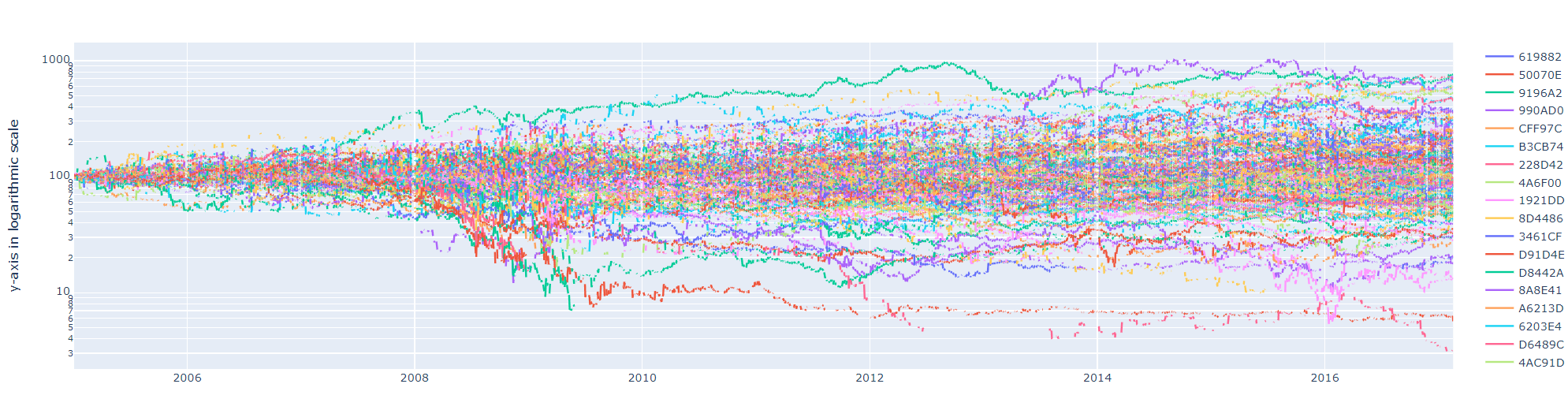
Une image contenant texte, instrument d’écriture, crayon

Description générée automatiquement

Distribution of return are quite normal, take extrem values appart



We plot prices of the assets having more than 756 points in ntime from 2005 to 2016 to have a first look at it:



The range path is wide, and asset behave differently through time.

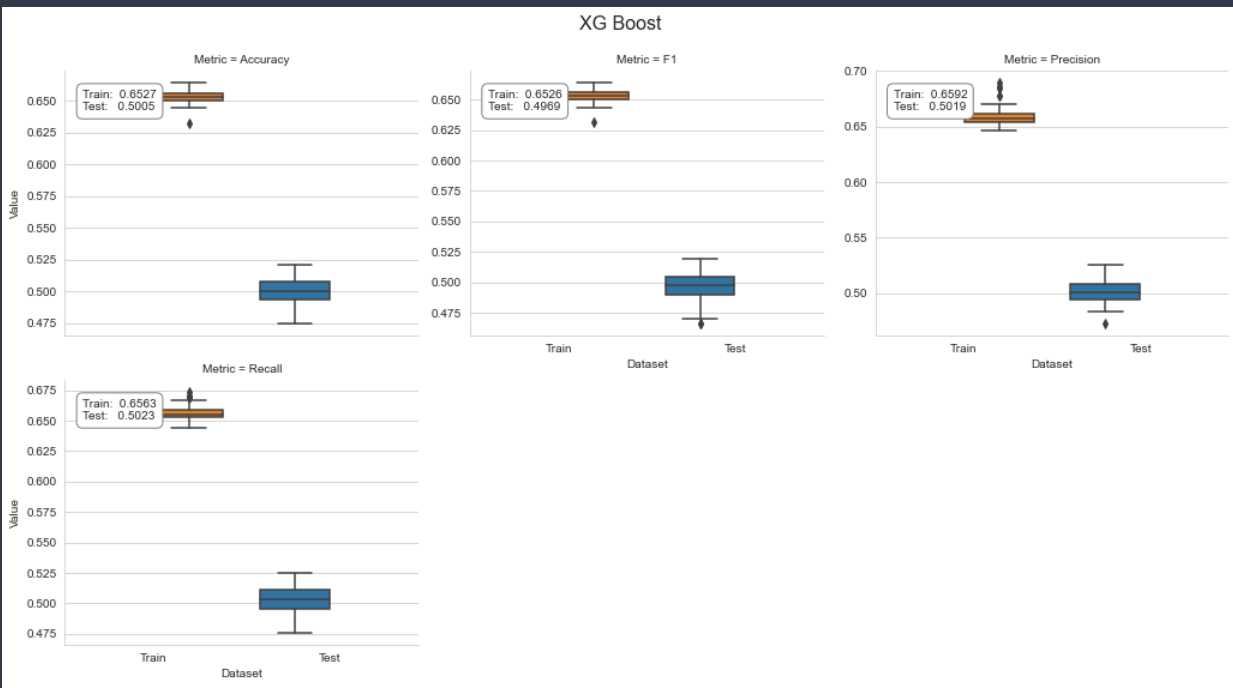
We can take note of the fact that there is a lot of missing date (times series are not complete).

# First model

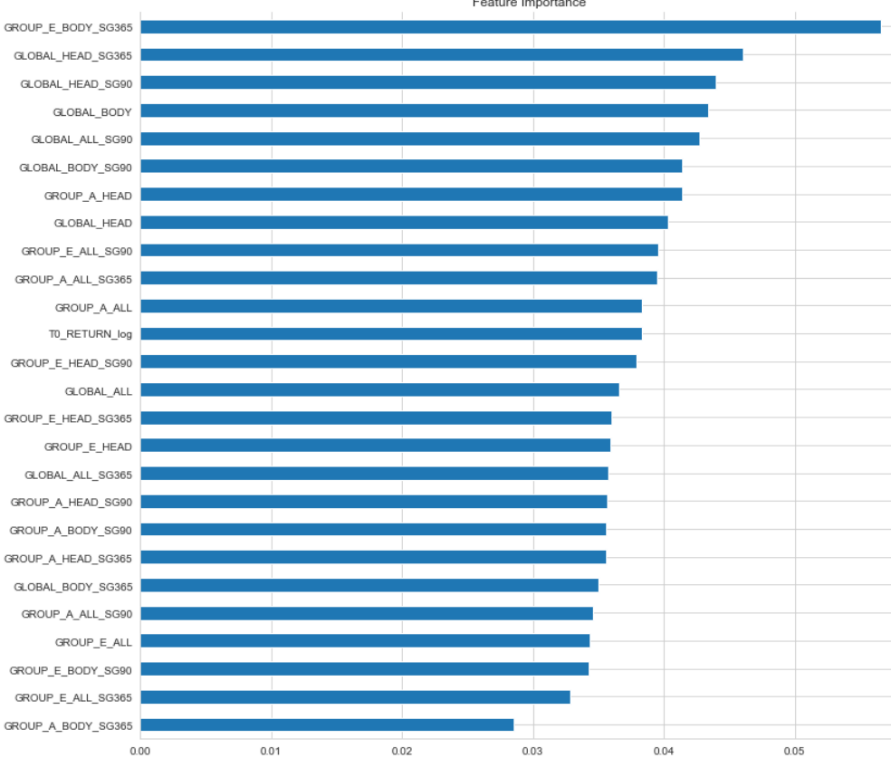
First, we process the following:

* convert the date column to datetime,
* merge duplicated row,
* drop columns with 100% NA,
* keep only assets with a minimum of 3 years of data available on the whole period (2005-2016),
* convert returns to log-returns.

We run the XGBoost algorithm on the whole data. First results are not good as illustrated below:

Une image contenant texte, tableau de points

Description générée automatiquement

Feature importance:

# Second Model