ETS DATA SCIENCE CHALLENGE

Introduction

ETS designs quantitative investment algorithms. We test our methodologies with available historical data. But historical data is scarce and the reality unpredictable.

To avoid overfitting and be able to design more robust algorithms, we have created a Financial Series Generator capable of producing endless synthetic series.

In order to test our models fairly and introduce the synthetic series in our work, we need them to be as similar as possible to the real financial series. Help us test our series generator. From a set of financial series, are you able to distinguish the real ones from the synthetic ones?

- Real financial series (**class 1**).
- Synthetic financial series (**class 0**).

We challenge you to predict the class of the series based on the provided features.

Data

The data set contains 3 files:

train.csv
test.csv
sample_submission.csv

In the train.csv for each of the 12.000 series you will find 260 features and its class label.

- **Feature1 to Feature 260**: are consecutive daily returns during one year.
- **Class**: 1 if the series is real and 0 if it is synthetic.

In **test.csv** you will find the same structure with another 12.000 series, but **without** the class label. Your task is to predict the class for each of these series.

Also, you have an example of the output format in **sample_submission.csv**.

Evaluation

The performance is measured with the AUC. We need you to send the probability (between 0 and 1) of each series belonging to the positive class, class 1 (real).

You are allowed to send 2 .csv files in the same format as in **sample_submission.csv** with your first and last names. For example: "AntonioCortijo1.csv" and "AntonioCortijo2.csv".

Please **send us the code you have written** to distinguish real series from synthetic ones.

You are allowed to use the programming language of your choice, but the use of **Python** will be highly valued.