About Dataset

**Context**

The algorithmic trading space is buzzing with new strategies. Companies have spent billions in infrastructures and R&D to be able to jump ahead of the competition and beat the market. Still, it is well acknowledged that the buy & hold strategy is able to outperform many of the algorithmic strategies, especially in the long-run. However, finding value in stocks is an art that very few mastered, can a computer do that?

**Content**

This Data repo contains the following datasets (in .csv format):

* 2014\_Financial\_Data.csv
* 2015\_Financial\_Data.csv
* 2016\_Financial\_Data.csv
* 2017\_Financial\_Data.csv
* 2018\_Financial\_Data.csv

Each dataset contains 200+ financial indicators, that are commonly found in the 10-K filings each publicly traded company releases yearly, for a plethora of US stocks (on average, 4k stocks are listed in each dataset). I built this dataset leveraging [Financial Modeling Prep API](https://financialmodelingprep.com/developer/docs/) and [pandas\_datareader](https://pandas-datareader.readthedocs.io/en/latest/" \t "_blank).

Important remarks regarding the datasets:

1. Some financial indicator values are missing (nan cells), so the user can select the best technique to clean each dataset (dropna, fillna, etc.).
2. There are outliers, meaning extreme values that are probably caused by mistypings. Also in this case, the user can choose how to clean each dataset (have a look at the 1% - 99% percentile values).
3. The third-to-last column, Sector, lists the sector of each stock. Indeed, in the US stock market each company is part of a sector that classifies it in a macro-area. Since all the sectors have been collected (Basic Materials, Communication Services, Consumer Cyclical, Consumer Defensive, Energy, Financial Services, Healthcare, Industrial, Real Estate, Technology and Utilities), the user has the option to perform per-sector analyses and comparisons.
4. The second-to-last column, PRICE VAR [%], lists the percent price variation of each stock for the year. For example, if we consider the dataset 2015\_Financial\_Data.csv, we will have:
   * **200+ financial indicators for the year 2015**;
   * **percent price variation for the year 2016** (meaning from the first trading day on Jan 2016 to the last trading day on Dec 2016).
5. The last column, class, lists a binary classification for each stock, where
   * for each stock, if the PRICE VAR [%] value is positive, class = 1. From a trading perspective, the 1 identifies those stocks that an hypothetical trader should **BUY** at the start of the year and sell at the end of the year for a profit.
   * for each stock, if the PRICE VAR [%] value is negative, class = 0. From a trading perspective, the 0 identifies those stocks that an hypothetical trader should **NOT BUY**, since their value will decrease, meaning a loss of capital.

The columns PRICE VAR [%] and class make possible to use the datasets for both classification and regression tasks:

* If the user wishes to train a machine learning model so that it learns to *classify* those stocks that in buy-worthy and not buy-worthy, it is possible to get the targets from the class column;
* If the user wishes to train a machine learning model so that it learns to *predict* the future value of a stock, it is possible to get the targets from the PRICE VAR [%] column.

**Inspiration**

I built this dataset during the 2019 winter holidays period, because I wanted to answer a simple question: is it possible to have a machine learning model learn the differences between stocks that perform well and those that don't, and then leverage this knowledge in order to predict which stock will be worth buying? Moreover, is it possible to achieve this simply by looking at financial indicators found in the 10-K filings?