Author: Yves D'hondt

Domain background

Stock markets are the second largest public market, valued at roughly $70 trillion (cmc markets, n.d.). With a market that large and a lot of money to be gained, it is no surprise that stock price prediction is a big topic. Traditionally, stock price prediction has been analysed through *factors*. In essence, factors are common sources of risks (& returns) among different stocks. Fama & French defined the de facto baseline factor model by identifying 5 different factors (Fama & French, 2015). They identified the following 5 factors: market returns, size, value, profitability & investment patterns. Looking for new sources of stock predictions, researchers have turned to analysing stock chart images using machine learning. Jin & Kwon investigate the impact of chart image characteristics on stock price returns using CNNs (Jin & Kwon, 2021).

Problem statement

This project investigates whether CNNs can be successfully designed to predict stock returns. Concretely, the following research question is addressed:

Can a CNN be designed that accurately predicts stock returns into one of the following three classes?

1. Positive returns (defined as returns > 1% over the next calendar month)
2. Neutral returns (defined as returns <= 1% and >= -1% over the next calendar month)
3. Negative returns (defined as returns < -1% over the next calendar month)

Dataset

The raw data used for this project will consist of adjusted prices for (almost) all US stocks & ETFs between 1984 and 2017 (Marjanovic, 2017). The dataset consists of adjusted prices, meaning that stock splits and dividends have been taken into account. This dataset was released to the public domain under CC0.

The raw data will be cleaned and transformed into a collection of (labelled) images according to the three predictive classes from the problem statement.

Solution statement

This problem will be tackled by setting up a number of different CNNs that classify the stock return images. To start a simple CNN trained on stock chart images of the past month will be trained. Depending on the performance of this model, the model could be extended by feeding it multiple input images (for instance 1 month, 3 month & 6 month return images) and/or adding numeric input features such as 1 month return and standard deviation.

Benchmark model

A simple momentum model will be used as the benchmark model. Momentum is one of the most predictive stock factors. The simple momentum model will be constructed as an OLS regression model that takes the past 1 month, 3 month & 6 month return of a stock and predicts the return over the next calendar month. Afterwards, these predicted returns can be classified into the 3 classes defined by the problem statement.

Evaluation metrics

The model will be evaluated based on its accuracy. This is simply defined as the % of samples that is classified correctly. While it is a basic metric and has some limitations (especially for unbalanced classes), it still gives a quick and easy to understand idea of the quality of the model.

To offer some additional insights, the model will be evaluated based on its precision and recall as well. However, since this is a multi-class problem, the precision and recall will have to be calculated separately for each class. Ultimately, with stock return predictions, the aim is to catch most opportunities (high recall) and avoid (costly) wrong investments (high precision).

Project design

The project will be completed by completing the following roadmap:

1. Clean and process the raw dataset into a dataset that contains:
    a. Numerical information on past stock returns (volatility, returns)
    b. Stock chart images of past stock returns (1 month, 3 month, 6 month)
2. Build an OLS benchmark model
3. Build a baseline CNN model (based on 1 image)
4. Build extended CNN models (based on multiple images and/or inclusion of numerical inputs)
5. Evaluate the benchmark model
6. Evaluate the CNN models
7. Draw insights and conclusions

The project will be constructed using the PyTorch framework. Beyond from creating the images and calculating some numerical features, minimal analysis of the data prior to constructing the models will be needed.

Sources

cmc markets. (n.d.). *Bonds vs stocks*. cmc markets. Retrieved from https://www.cmcmarkets.com/en/trading-guides/bonds-vs-stocks

Fama, E. F., & French, K. R. (2015). A five-factor asset pricing model. *Journal of financial economics*, *116*(1), 1-22.

Jin, G., & Kwon, O. (2021). *Impact of chart image characteristics on stock price prediction with a convolutional neural network*. PLOS ONE. Retrieved from https://doi.org/10.1371/journal.pone.0253121

Marjanovic, B. (2017). *Huge Stock Market Dataset*. Kaggle. Retrieved from https://www.kaggle.com/datasets/borismarjanovic/price-volume-data-for-all-us-stocks-etfs