**Stock Predictive Model – On Equity Fundamentals**

By Hsin

# GOAL – predict if a stock is likely to go up given the fundamentals data

It has been a long-time discussion which people are always interested in: Can we use a company’s equity fundamentals data to predict stock price, or to know if the stock price is likely to go to go up next year? I

# DATASET – NYSE S&P 500 companies historical prices with fundamental data

I would like to use NYSE Fundamentals dataset for my final project (or my first data science project).

There are 4 csv files for this dataset:

1. Fundamentals - 5 year worth annual Fundamentals data for S&P 500 members. This file has 79 columns and 1781 rows. The columns essentially represent fundamental measures and therefore mostly are numeric values.
2. Securities – company information including company name, sector, and company address etc.
3. Prices - 5 year worth *daily* stock prices for S&P 500 members
4. Split-adjusted-prices – Same as Prices, but are adjusted for split already which is what I will be using this for the model.

# PROBLEM STATEMENTS

It has been a long-time discussion which people have been always interested in: Can we use a company’s fundamentals data to predict stock price, or, at least to know if the price is likely to go to go up next year? Today a lot of trades are already executed based on the machine algorithm however to satisfy my own curiosity I would like to experiment it myself and try to answer the following questions:

* What are the most important fundamentals variables that could drive the price?
  + There are 78 fields in the dataset. Some may influence the price and some not may not. Understand the important variables should not only help me build the model, more importantly, from business perspective, corporates and investors want to know what impact their business. I will examine the correlation for all the fields with the price and see if the result matches the convention.
* Can we predict which securities are likely to go up next year? (recommender)
  + Given the historical fundamentals and prices, what securities would be recommended as “BUY”. I.e., if we can predict certain companies with high probability of higher price (annually averaged) by certain threshold, we can recommend those securities.
* Is it possible to use the company’s fundamentals to predict the price? (regression)
  + I would like to see if we can try to predict the price within certain range for a given stock. We could then also potentially calculate the net profit (by taking the probability into consideration)
* Given the fundamentals variables, is a security’s price likely to go up? E.g., Average price
  + A simple yes/no question, should a security be a buy if I just give you all the fundamentals data - for a security that we have not seen (i.e., members outside of S&P 500)

# CHALLENGES – digging a bit into the dataset

* The dataset Fundamentals has 78 fields – this is quite a lot of variables so the challenge would be finding the useful variables for the model. In fact, many fields are highly dependent on other fields (such as Revenue and Expense determine the Net Income) so I have to be careful with using fields that are highly correlated with each other. As we make progress throughout the teaching course hopefully we will go through some techniques that could address such issue. Some domain expertise might also be needed when it comes to making assumptions for the model.
* The Fundamentals dataset only has annual data – which means for a stock, there are only 5 records. This makes things very challenging because the general rule of thumb is, you can’t use other company’s fundamentals model to predict a different company’s price. They are just simply not relevant. Normally we would perform analysis on an individual ticker level – if we have sufficient amount of the data for that ticker. Here, because the amount of data per ticker is very small, it would not make much sense to conduct analysis on ticker level. I would have to come up with some way to cluster the companies, e.g., by sector, in order to make the prediction. My assumption here is that the companies that belong to the same group should have similar pattern in terms of the price movement. For the purpose of standardization, I would probably not use the price but use the PE ratio (price /EPS) and assume that the companies in the same cluster to share the same range of PE ratio.
* The Prices data is provided for daily. Given that the fundamentals are only available in annual we would have to convert the daily price to annual basis. I will simply use the average price for the year for this purpose. Fortunately, there is a price file that already has split-adjusted numbers so this will take care of the spikes due to stock splits so I will just use this file rather than the raw prices.
* In order to predict if a stock is likely to grow, I also need to get the growth rate – which could also be another way to standardize the values across tickers. Imagine the values of all these fundamentals variables are only relevant within a ticker, if we translate them into the growth rate (year over year), then we are technically standardizing them across tickers – assuming at the end of the day, we only care if a stock is likely to grow, no matter the size of the company. We could potentially use the growth rate in our model rather than actual amount to get away with the comparability issue.
* At this moment, it is hard for me to determine the metrics of success given the challenges mentioned above and the time restraints for the final presentation. However, for the very minimum, building a model using as much knowledge/techniques learned from the course as possible and understand the pros and cons for each application would be my primary goal for this project. The knowledge area of the dataset I chose here is in line with my work and hence it would be beneficial for my career path as well.