**DATA SCIENCE AND FINANCE 2023**

**Project: Back testing strategies based on firm fundamentals.**

**To do in groups of up to 3 students for March 28**

**The goal of this project is to explore if you can build interesting trading strategies (i.e. with high Sharpe ratios and low correlation to the market) using fundamental data:** In other words, can sorting stocks on some fundamental characteristics generate alpha? and how to find a best combination of features? **Feel free to proceed your own way**, and to use methods you know (e.g. machine learning). Just remember to avoid any “look-ahead bias” and be aware of overfitting problems.

**Prepare a presentation (slides or notebook) of your results that you can show in class and that you can send me together with programs.**

I have constructed two files for this project:

First, the file ‘crsp\_top1000.csv’ contains monthly stock returns. I selected data for the top 1000 US stocks by market cap (as measured at the end of previous calendar year), so these are fairly liquid stocks, and you will construct your back-test on this universe. That file also contains other stock-level information that you might find useful and that you can use in building your signals, All the variables in this file are also available for traders at the end of the corresponding month.

Sic2 is the industry of the stock (First 2 digits of Standard Industrial Classification Code).

vol is Volume (in # shares) during month t

shrout is Shares Outstanding at the end of month t

ivol is idiosyncratic volatility (i.e. volatility after controlling for the risk coming from beta)

b\_mkt is the beta of the stock estimated using rolling windows. (ivol and b\_mkt are

estimated on previous months)

Second, the fundamentals data file: I have saved for you some fundamental characteristics of stocks (financialratios.csv), based on firms’ accounting. The codebook of the variables is Financial\_Ratio\_Manual.pdf. The variable “publicdate” indicates when the variables on that line of the data are publicly available to investors: if the variable is available at month t, it can be used to construct a portfolio at the end of that month, which will result in returns during month t+1. This file can be combined with the previous one using the stock identifier permno and the date (precisely, “publicdate” in this file and “date” in the previous file).

Here are some suggested steps.

First, let’s keep aside an out-of-sample period for final validation: say, all months after January 2017.

1. Start with one signal, say sorting stocks on the variable “roa” (returns on assets). This is a strategy called profitability that is well known.

1. On the training sample, construct the portfolios (for example, long the top quintile, short bottom quintile).
2. Still on the training sample, explore if this signal can be improved by “industry-adjustment”: we change the signal by sorting stocks within their industry. We invest in the top 20% and short the bottom 20% of the ROA distribution *within each industry*.
3. Now check the performance out-of-sample: how does it compare to in-sample?
4. Check with a regression if your signal has alpha vis-à-vis the four factors included in the file FF.csv. (I would expect that the signal has a high beta on “qual”, the quality factor)

2. In this second step, the idea is to generalize the analysis made on one feature (roa) by using all proposed features in a systematic manner. You should propose a process, execute it, and finally test it out-of sample (what is the Sharpe ratio out of sample?) Be careful to test out-of-sample **only once**, i.e. when you have finalized your proposal!

Feel free to propose your own method (I am sure you will find creative approaches and I am very curious to hear about them!), below are some ideas:

-You could for example use rolling-windows to select features at time t that have done well in previous 3 years.

-You can use the change in features over time to build signals.

-You can generate other features by inter-acting features from the data-set (industry-adjustment, multiplications etc.).

-You can put dynamically more weight on signals that did well in the past (on a rolling-window basis, to avoid look-ahead bias).

-You could take into account the correlation between features to avoid putting too much weight on similar signals (one possibility is to use PCA to reduce dimensionality).

-etc.