# Practice: machine learning in finance

## Part 1: Data problem

1.1 Data nature: The provided data are not the real values but the percentiles for fundamental variables, except forward returns. They are times series from 1998-11-30 to 2019-03-31.

1.2 Data missing: the missing data soars at the end of period, particularly since 2018-12-31 when missing rates jumped from 8.1% to 24.5%. The pct. Missing value edged higher until 54.3%.

1.3 Data outlier: There’s no way to control outlier of fundamental analytics since they are in percentile. As a result, we only apply outlier checks to zero forward returns. We see evidence of extended zero forward return with wider forward-looking period. More specifically

1. No surge of outlier rate for 1M forward return, despite a considerable amount of missing value
2. The outlier rates climb earlier for longer forward-looking period, due to lack of price data.

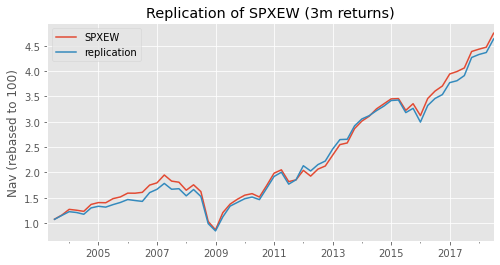
1.4 Data validation: It’s reasonable to assume that DB contains the most liquid stocks in US markets. To test this guess, we will try to replicate S&P500 Equal Weight Total Return Index.

#### Introduction of S&P 500 methodologies.

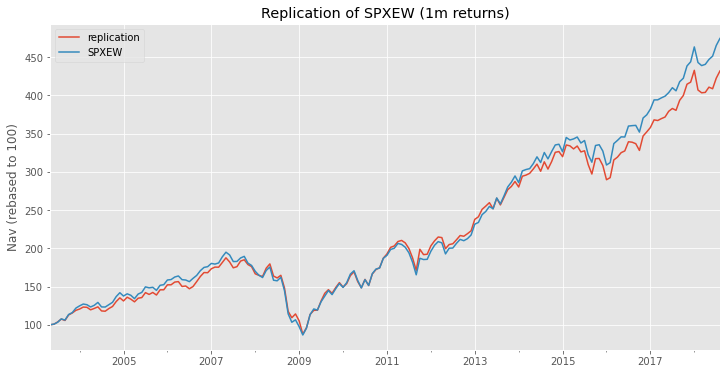
1. Stock inclusion: top ranked stocks.
2. Weighting scheme: “Equal weight”.
3. Rebalance dates: the 3rd Friday of March, June, September, and December.
4. Cut-off dates: the 2nd Friday of March, June, September, and December.
5. Change to index composition: on an as-needed basis, without scheduled reconstitution.

#### Construction of replication portfolio

Replication portfolio will contain the largest 500 values with equal weight just prior to each quarter-ending month (i.g. end-Feb, end-May, end-August, end-Nov). The comparison is still reasonable based on 3M returns.



However, we learnt that the replication is likely to be less perfect if being based on 1m returns.

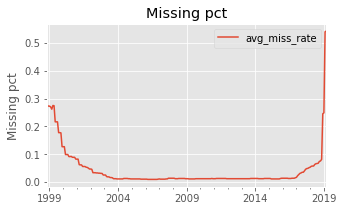
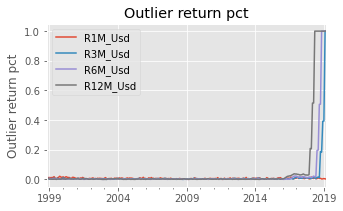


The potential cause for this difference is the inconsistency between ret\_1M\_fwd & ret\_3M\_fwd.

1.5 Rules for backtesting

*Valid periods of forward returns:* Partial period is not eligible for backtesting due to lack of data after Dec. 2018.

1. R1M is applicable until 2018-11
2. R3M is applicable until 2018-09
3. R6M is applicable until 2018-06
4. R12M is applicable until 2017-12

*Selection universe* is limited to the largest 500 active values instead of all the securities

*Rebalancing rules* is to equally weight selected securities just prior to each quarter-ending month.

## Part 2: Factor Screening

## 2.1 what’s information coefficient (IC)?

Information coefficient (IC) gauges the correlation between predicted and realised returns. Unlike information ratio, IC reflects the success rate of bet regardless of return’s amplitude. In a nutshell, there are 3 major measures of PM’s skill:

1. Sharpe Ratio => Measure over a period
2. Information Ratio => Measure over a period
3. Information coefficient => Measure between two rebalancing dates.
   1. IC = Pearson correlation (forecasted return, realized return) => classical correlation
   2. IC = Spearman correlation (forecasted return, realized return) => correlation of ranks
   3. Attention:
      1. It’s worth adjusting returns by risk (volatility) to avoid persistent difference in return dispersion with respect to stock volatility. In other words, the risk-adjusted returns are more likely to be comparable. However, the risk calculation may be delicate.
      2. More specifically, we can simplify the IC expression if we focus only on the return signs, For instance, the sign of stock’s return relative to market (high or low).
      3. In this regard, IC = 2 \* correct decision (%) – 1. The correct decision (%) = Nc / N, where Nc is the number of correct bets.

## 2.2 Fundamental Law of Active management

**Initial FLAM**

According to the **Fundamental Law of Active management**[[1]](#footnote-1), the information ratio can be predicted by

where IR can be decomposed into the

* IC: the success rate of PM’s skill
* N: the breath of PM’s coverage (𝑁 is the number of independent forecasts)

In practice, N may represent for

* **Stock-picking**: the number of cross-sectional bets on different assets at any point in time
* **CTA**: the number of independent bets on the same asset across time.

**Migration of FLAM**

Somehow, there is a migration for FLAM as discussed by Ding and Martin [The Fundamental Law of Active Management: Redux](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2730434) (2017).

Where α and σ are defined below:

is the residual risk that a risk model can not explain. If everything is perfectly explained by risk model, no independent alpha can be obtained. =>

## 2.3 IC of main factors

The main goal of this section is to assess the performance of factors. By reviewing the published results, it turns out to be normal that IC is something less than 5%, regardless of risk-adjusting, or factor purification.

To reach this goal, it requires to carry out following tests:

1. for each period, calculate IC by spearman correlation between factor and forward returns. (due to the fact that Factor data is percentile but not the real value)

* Forward term: 1M or 3M
* Return type: raw return or risk-adjusted return ~~“market neutral return (specific part from CAPM)”~~

=> Attention: Beta need to winsorized to a rational range such as [0.7, 1.3]

2. Evaluation for skewness (it’s not worth using a factor with symmetrical correlation)

* Mean, median: Is mean far away from median? => skewness test
  + Retain the outstanding factors (IC > 2%) for further analysis.
* Consistency of IC sign => is correlation constantly above or below zero?
  + Calculate the pct. positive IC over time
  + Boxplot for each outstanding factors.
* Hypothesis testing: is correlation significantly from zero? => significance test.
* Auto-correlation of same factor between two successive periods
* Cross-correlation of IC between different factors.

Test 1: mean vs. median

Goal: Uncover the factors whose median is very close to zero otherwise average is far from zero. These points do not deserve to study since the average values are just biased by outliers.

Observation:

* Very few factors fall into this trouble. Mean and median are quite linearly correlated most of the time.
* The correlation is higher for 3M forward returns than 1M forward returns, which is nearly doubled. This is a crucial discovery since an absolute correlation beyond 4-5% will deserve to explore (which is the case of 3M forward return). Otherwise, the correlation of 1M forward return sound too little to analyse.

Chart, scatter chart

Description automatically generated

~~From what we know from CAPM, stocks’ returns can be broken into specific part and market (systematic) part. It’s worth checking the correlation between factor and stocks’ specific returns as below:~~

The return dispersion is naturally tied to stock’s volatility regardless of other reasons. To mitigate this bias, we should resort to risk-adjusted processing which turns out useful. As shown below, there are more factors showing correlation greater than 2%.

Chart, scatter chart

Description automatically generated

As suggested by book, it appears more reasonable to focus on 3M return correlation to avoid noise trap.

We retain the outstanding factors (IC > 2%) for further analysis



As explained extensively by academic research, we’re about to retain following descriptors.



Test 2: Consistency of IC sign

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

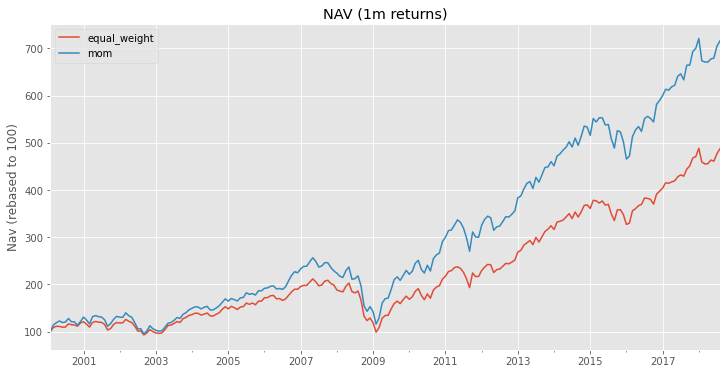
## Part 3: Tree-based method for factor portfolio

This section will start with example of momentum portfolio.

As is widely known, the momentum analytics are deemed useful for Sharpe Ratio improvement. It’s not crazy to expect the Tree-based method find a better return.

### 3.1 Show backtest on momentum factors based on 6m momentum analytics.

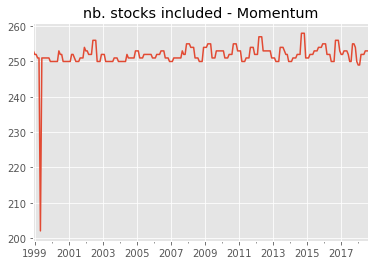
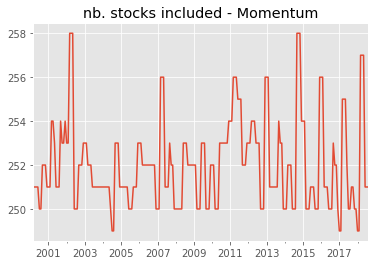
Firstly, the order of momentum score “**Mom\_Sharp\_5M\_Usd**” and “**Mom\_Sharp\_11M\_Usd**” is likely to be opposite to “**Mkt\_Cap\_3M\_Usd**”. Unlike “**Mkt\_Cap\_3M\_Usd**” in descending order for SPX replication, it requires to sort momentum scores in ascending order to obtain outperformance.



Check: what’s the root cause of sudden decline of nb. Stocks included in mom factor before 2000?

Does it arise from significant missing forward returns before 2000?

Answer: yes! It’s caused by period of bad quality before 2000. Just to strip them and it’s fine.

### 3.2 Apply one-layer tree model to momentum strategy

The classical momentum strategy is based on fixed threshold for stock selection. Instead, we expect to use one-layer tree model to learn the best threshold by itself.

Conclusion: the momentum based on Tree models may be worse.

Different parameters which may impact tree-based momentum strategies:

Chart, histogram

Description automatically generated

1. https://quant.stackexchange.com/questions/34189/about-the-number-of-independent-forecasts-in-the-fundamental-law-of-active-manag?rq=1 [↑](#footnote-ref-1)