期末報告

主題: IC 之 AI 模型建立與預測

樣本: 元大台灣 50 (0050) +元大高股息（0056）ETF

目標: 提高勝率及穩定性

基於 AI 模型有許多種，重點利用宏觀變量、因子歷史 IC 序列和市場變量等信息來預測每個風格因子未來一期的 IC 值，然後根據預測的 IC 值對風格因子賦予權重。一般的，預測時間間隔越短，機器學習模型的預測能力會越強，但是短時間內股票價格的漲跌幅度一般較小，可能不足以覆蓋交易成本及買賣股票對市場造成的衝擊。綜合考慮到模型預測的準確程度和交易成本，建議選擇以周為再平衡週期，機器學習模型也選擇以周為預測週期，即在每天收盤時進行預測，預測投資組合中成份股 5 個交易日之後的回報率與風格因子當前取值之間的相關係數。在模型輸入特徵的選擇上，主要從因子歷史 IC 序列及外部變量兩個方面來提取特徵，而外部變量又分為市場變量和宏觀變量兩部分。其中，從風格因子歷史 IC 序列出發，我們提取了不同頻率下（日頻、週頻和月頻）每個因子上期的IC值，時間間隔分別為1天、5天和20天，總共得到15個特徵。而外部變量包括4個市場變量和2個宏觀變量。這樣，整個模型共包含 17 個特徵。此外，本報告對每個風格因子都建立了預測模型，因此總共有 7 個模型，模型輸入為相同的 17 維特徵，輸出為下期的 IC 值。本次重點中，多因子策略的選擇股票池為 0050 及 0056 ETF 成份股，以 2007 年 1 月至 2015 年 12 月的市場行情為樣本內訓練數據，以 2016 年 1 月以來市場行情為樣本外回測數據。一般來說，對機器學習模型進行訓練時，希望有更多的樣本，這樣訓練得到的模型泛化能力更強、穩定性更高。

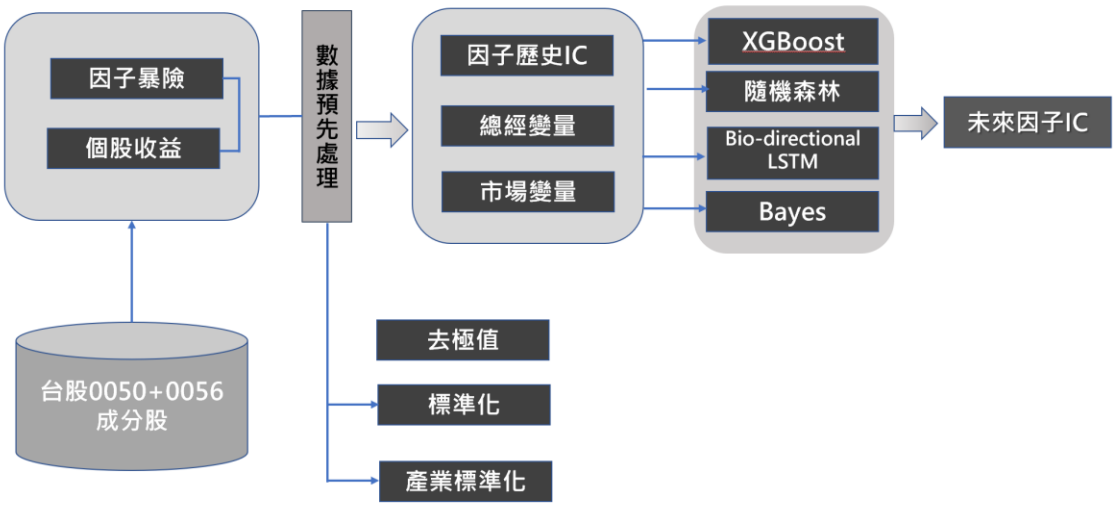
本次工作:

一、資料準備和處理。

二、預測因子的未來 IC： 用四種模型、滾動模型來預測因數的下一期 IC。

三、因子選股比較：利用四種預測的 IC 值，並構建組合比較因子的效果。

投資架構流程圖:



樣本: 在 0050+0056 ETF 股票池中所選因子及其分類:

表格 1

因子類別	因子名稱
Value	EP FCF/P P/NWC
Earning	ROE Gross Margin
Historical Growth	1Y Chg. ROE 1Y Chg. FCF/Assets
Momentum	20day Rolling RS. 1Q Chg. Alpha
Risk	20Day Rolling Vol 1Q Beta

1. 取 5 個因子的值，基於滾動樣本訓練，原始因子數據處理：每個因子採用四分距進行標準化，只保留因子排名信息，對於每一個因子，每支股票先根據因子值進行排序得到 Rank。
2. 每個 60 個交易日建構並更新一次模型並預測未來。

外部變量及其處理方法:

表格 2

變量類別	名稱	處理方法
市場變量	台灣加權月漲跌幅	一階差分
	費半指數月漲跌幅	一階差分
	台灣加權月波動率	不處理
	台灣加權月波動率	不處理
宏觀變量	美金台幣匯率	變化率
	外銷出口訂單	變化率

IC (Information Coefficient，資訊係數)：表示所選股票的因子值與股票下期收益率的截面相關系數，通過 IC 值可以判斷因子值對下期收益率的預測能力。換言之，IC 代表因子選股帶來的超額收益能力。

1. IC 的最大值為 1，表示因子的預測能力 100%準確。
2. IC 最小值為-1，表示因子絕對反向的預測作用。
3. IC 等於 0，表示因子既沒有正向的預測也沒有反向的預測作用。
4. 當 IC 大於 5%或者小於-5%，就會認為因子是比較有效的。
5. 如果 IC 的平均值能夠達到 10%以上，則代表這个因子的預測能力非常強。

常見的 IC 有兩種：一種是 Normal IC，另外一種則是 Rank IC。通常使用的是 Rank IC，此時 IC 表示因子選出股票的排名與股票收益率的相關性。

IC 計算的概念是 *t* 期的因子值於選出的股票中做排序值(Rank)和該股票於 *t+1* 期收益的排序值(Rank)之間的相關系數(Spearman)。而兩者之間的相關程度，就是 IC。因此，如果因子選出排名前面者的股票的收益也很高，則表示因子的預測能力越強。

https://www.joinquant.com/view/community/detail/15290

因子說明與定義

所使用的因子及其分類，整理於表 1。

表 1 因子之定義

變數	定義
EP	益本比。衡量投資者在未來一段時間內利用公司預期收益比率的指標。計算方法是將預計每股收益除以股票的當前市場價格。相對較低的 E / P 比率預計收益增長高於平均水平。 益本比 = 普通股每股收益/股價
FCF/P	自由現金流量對市場價值比。主要將公司的每股自由現金流量與其當前股價進行比較。 自由現金流量對市場價值比 = 自由現金流量/市場價值
P/NWC	股價對淨營運資本比。淨營運資本用於將流動資產減去流動負債來確定公司流動資產的可用性。 股價對淨營運資本比= 股價/(流動資產 − 流動負債)
ROE	權益報酬率。
Gross Margin	毛利率。
1Y Chg. ROE	權益報酬率的年變動。
1Y Chg. FCF/Assets	自由現金流量對資產比的年變動。
20 Day Rolling RS	前 20 天滾動相對強弱勢指標
1Q Chg. Alpha	Alpha 的季變動
20 Day Rolling Vol	前 20 天滾動波動率。衡量證券的風險。20 Day Rolling Volatility =前 20 天滾動波動率 = 股價過去 20 天變動標準差/總報酬股價 $*\sqrt{252}$ 每季的 Beta

外部變量

所使用的外部變量包含市場變量與總體變量。其名稱與處理方法整理於**表 2**，其中，變數 X 之一階差分與變化率的定義如下：

$$一階差分_{t,X} = X_t - X_{t-1}$$

$$變化率_{t,X} = \frac{X_t - X_{t-1}}{X_{t-1}}$$

表 2 外部變數之定義與處理方式

變數名稱	變數定義	處理方式
市場變量		
台灣加權($TAIEX_t$)	全名為台灣加權指數，簡稱 TAIEX，由臺灣證券交易所所編製的股價指數。在本研究中，我們採用一階差分作為主要變數。	一階差分
費半指數(SOX_t)	全名為費城半導體指數，簡稱 SOX，此為全球半導體業景氣主要指標之一。	一階差分
台灣加權波動率 (Vol_t^{TAIEX})	利用台灣加權指數之歷史報酬率計算所得之標準差。	報酬率之標準差
費半指數波動率 (Vol_t^{SOX})	利用費城半導體指數之歷史報酬率計算所得之標準差。	報酬率之標準差
總體變量		
美元兌新台幣匯率 (EX_t)	由外匯市場中獲取其一美元可兌換新台幣數量之成交價格。	變化率
外銷出口訂單(EO_t)	根據外銷廠商承接國外客戶貨品訂單額度資料所編製之統計並計算而得其外銷金額數據。	變化率

靜態方法

研究樣本期間從 2005 年 1 月至 2018 年 12 月的月資料為樣本，其中 2005 年 1 月至 2017 年 12 月為樣本內訓練資料，又以 2018 年 1 月至 2018 年 12 月做為樣本外回測。

動態方法

選取前 n 個月的樣本做模型訓練，以 m 個月當動態移動樣本，動態做未來 k 個月的模型預測。例如：$n = 36, m = 1, k = 1$，第一個樣本時間點是 2007 年 12 月，選取樣本期間從 2005 年 1 月至 2007 年 12 月共 36 個月做模型訓練，去預測下 1 個月時間點為 2008 年 1 月的模型預測，再動態移動 1 個月樣本固定為 36 個月繼續做到 2018 年 12 月。

單因子分析

從台灣經濟新報(TEJ)資料庫中，選取上市之 900 檔股票，樣本期間從 2005 年 1 月至 2018 年 12 月，其中 2005 年 1 月至 2017 年 12 月為樣本內訓練資料，又以 2018 年 1 月以後做為樣本外回測。首先選取會從宏觀面和基本面影響股票收益率的因子，將個股因子區分為五類: Value、Earning、Historical Growth、Momentum 和 Risk，加上外部變量包括 4 個市場變量和 2 個總體變量，整個模型共有 11 個因子。原始因子數據處理：每個因子先去極值後採用四分距進行標準化，只保留因子排名信息，對於每一個因子，每支股票先根據因子值進行排序得到 Rank，將每個因子按分數大小排序並且將股票池按照分數分五等構造組合。

接下來將用下式所有因子和下一個月報酬來計算相關性當作每個因子當期的 IC 值

$$\mathrm{IC_t} = Corr(R_{t+1}, X_t) \tag{1}$$

再來將算出來的〖 IC〗_t 依照 11 個因子分成 11 組，各組中算出的 IC 序列和因子序列分別放入 4 個模型：XGBoost 法、隨機森林法、雙向 LSTM 法及貝氏分類器法，模型分別輸入當期的 11 個因子，輸出預測的下一期 IC 值。

多因子組合

1. 加權組合法

我們根據單因子分析所選取的因子挑選表現最佳的兩個因子，進行權重調整，例如：權重各半調整後得到新的因子，根據新的因子再利用單因子分析法重新評估投資組合。

2. 雙重排序法

假設共有 900 支股票，按照 A 因子排序來分組，高到低分成三組 (A1, A2, A3)，接著再依照 B 因子分組，在 A1 組別中可在拆分出三組 (B1, B2, B3)，依此類推。我們可以用下表作為示意圖，買進(A1, B1)、賣出(A3, B3)

先排 A 再排 B	**A1**	**A2**	**A3**
B1	(A1, B1)*	(A2, B1)	(A3, B1)
B2	(A1, B2)	(A2, B2)	(A3, B2)
B3	(A1, B3)	(A2, B3)	(A3, B3)**

*表示買進，**表示賣出

3. 雙層排序法

假設共有 900 支股票，按照 A 因子排序來分組，高到低分成五組 (A1, A2, A3, A4, A5)，接著再依照 B 因子分組，在 A1 組別中可在拆分出五組 (B1, B2, B3, B4, B5)，依此類推。

下表為示意圖，我們買入 (A1, B1)和(A5, B1)，賣出 (A1, B5)和(A5, B5)

先排 A 再排 B	**A1**	**A2, A3, A4**	**A5**
B1	**(A1, B1)***		**(A5, B1)***
B2, B3, B4			
B5	**(A1, B5)****		**(A5, B5)****

策略評價

從預測出$t+h$期 ($h \in \{1,2,\cdots,12\}$)的 IC 值去挑選較高的因子來做策略評價。例如：因子益本比(EP)預測的 IC 值最高，在t期用 EP 因子將所有公司排序 3 等分(5 或 10 等分)，買最高等分的投資組合賣掉最低等分的投資組合，持有h期來看報酬率。若有兩個因子較高，在 t 期時用這兩個因子將所有公司各排序 3 等分 (5 或 10 等分)共有 9 等分 (25 或 100 等分)，買這 25 等分的最高等分的投資組合賣掉最低等分的投資組合，持有h期來看報酬率。