

股票預測分析

判斷買多還是賣空、預判漲跌，使利潤最大化

Stock prediction and analysis

黃秉茂

Ping-Mao Huang

國立交通大學資訊學院資訊工程學系

Department of Computer Science

National Chiao-Tung University

中華民國 110 年 11 月

November, 2020

Abstract

AI 發展時至今日，在分群和辨識系統上已經做得很好了，人臉辨識、遊戲甚至是遠超人類，然而，在金融上的成就遠遠不及其他領域，也許是因為財金這塊領域的 **domain knowledge** 太深，導致模型在基礎設定上本來就考慮的不夠周全，因此我希望能運用機器學習看看 ML 是否能幫助股票分析。

Motivation

個人利益方面，當然是希望所學能幫自己賺錢。社會方面，因為 AI 在有些方面已經超越人類了，像是圍棋，而運用在財金方面是否有顯著的提供協助？財金相較於其他領域是更貼近人們的生活的，因此可能對人類的生活產生巨變，而且投資要考量的變數實在很多，所以我很好奇運用 AI 技術的投資能否打敗人類。

AI 已經有發展到一定的程度了，然而運用 AI 的領域實在不算多，目前可能只有遊戲、視覺、跟自然語言處理有比較傑出的成果，AI 勢必還需要在某些能大大改變人們生活的領域發展，更能顯的 AI 會如何影響人類。如果 AI 加上財金將會大大地改變人們的生活，而且投資並不是只有考量數學，還是一門跟人文、心理相關的學問，而 AI 是否能連這些都考慮到是我所好奇的。

Introduction

在 AI 在其他領域有不凡的成就時，很多資料分析師都有個疑問，像是財金那麼複雜且和人們息息相關的領域，ML 有辦法去輔助人們做出決策嗎？以下是很多國外學者在股票分析上的成果，而我想看看當分析的資料不是全世界，而是僅僅是台灣時，會發生什麼事情？畢竟一般的台灣人通常只在乎國內的情形，很少對國外的金融商品下手。如果下列的成果都那麼好，那台灣的資料也會一樣好嗎？所以我會透過技術面去預測及分析股價。

Related Work

- ["Global stock market investment strategies based on financial network indicators using machine learning techniques.](#) Lee, Tae

- Kyun, et al. "Global stock market investment strategies based on financial network indicators using machine learning techniques." Expert Systems with Applications 117 (2019): 228-242.
- ["Supporting Investment Management Processes with Machine Learning Techniques."](#) Groth, Sven S., and Jan Muntermann. "Supporting Investment Management Processes with Machine Learning Techniques." Wirtschaftsinformatik (2). 2009.
 - ["A machine learning model for stock market prediction."](#) Hegazy, Osman, Omar S. Soliman, and Mustafa Abdul Salam. "A machine learning model for stock market prediction." arXiv preprint arXiv:1402.7351 (2014).
 - ["Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques."](#) Patel, Jigar, et al. "Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques." Expert systems with applications 42.1 (2015): 259-268.
 - 希望能用台灣的股票資料，做出和上述差不多好的結果

Methodology

Problem statement - Technical analysis

1. 起始金額為現金 110,000 元，希望最後能賺最多錢
2. 每天都要判斷買多還是賣多這固定的十檔股票(0050、0056、鴻海、台積電、聯發科、大立光、富邦金、國泰金、玉山金、元大金)
3. 自變數會用先前的開盤價、最高價、最低價、收盤價、成交量、5 日均價、20 日均價、網路聲量(如果有學會爬蟲的話)等等，而變數就是這十檔分別該買還是賣
4. 假設想買一定買得到，想脫手一定賣得掉
5. 假設股票能以前一天的收盤價買進，融券也是以前一天的收盤價為代價
6. 假設手續費為交易金額的 1%
7. 先預測哪些會漲那些會跌之後，先將現金 10,000 元保留以備不時之需。在認為漲超過手續費的股票中，覺得投資報酬率最高的投入剩餘現有現金的 50%，第二高的再投入剩餘現有現金的 50%，以此類推，並將剩餘所有的現有現金投入投

資報酬率最低但還是認為能漲超過手續費的，然後隔天要將所有持有的股票都賣掉。認為會跌超過手續費的也是按照買股的方式，在認為跌超過手續費的股票中，覺得賣空的投資報酬率最高的就借價值約為現有現金的 50% 的券，賣空的投資報酬率第二高的再借價值約為剩餘現有現金的 25% 的券，以此類推，並在最後借與前一個同價值但還是認為能跌超過手續費的券，然後隔天也要將所有持有的股票都賣掉。隔天都會平倉。

8. 可以全部都不買也不賣

Baseline Method

- 將所有的變數拿去跑回歸，並預測報酬率
- 預測股價與目前股價做比較
- 每一次都是研究 30 天的資料決定第 31 天的買賣
- Testing data 也是 30 天

Data

data source: TEJ, yahoo finance, apple stock

會先將.csv 下載下來後，再使用 pandas, numpy 等進行 EDA

stock:

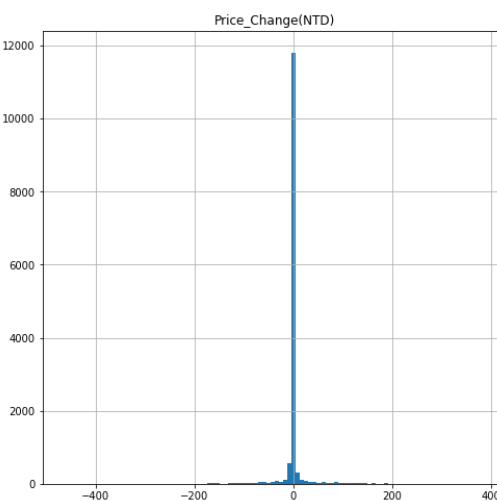
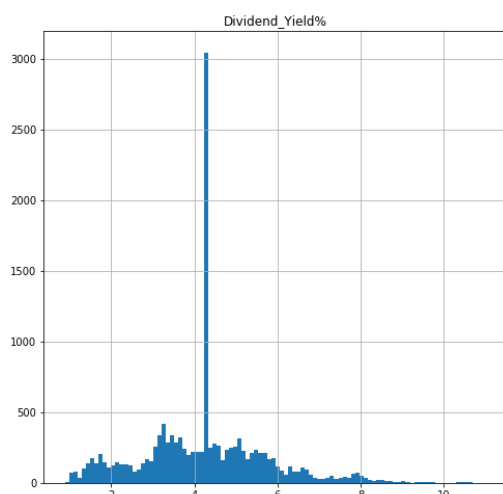
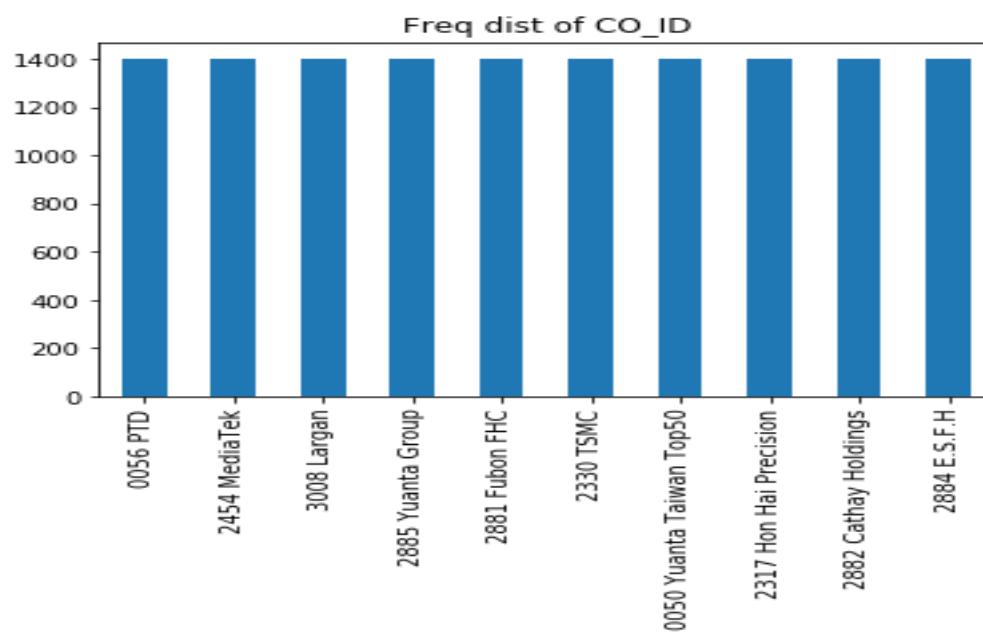
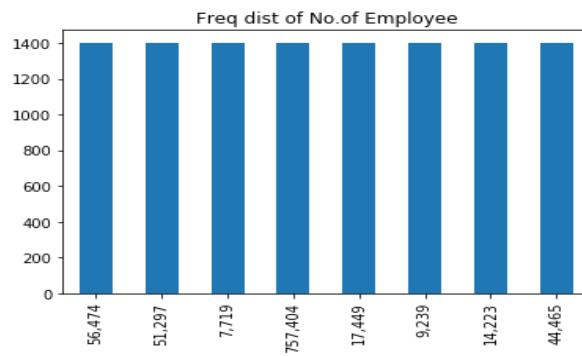
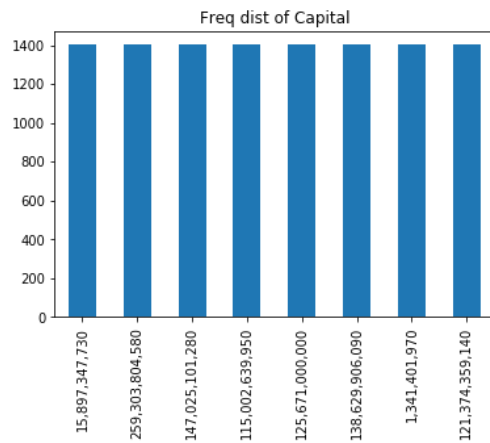
- | | |
|--|-------------------------------------|
| ▪ 0050 Yuanta Taiwan
Top5: 0050 元大台灣 50 | ▪ 3008 Largan: 3008 大立光 |
| ▪ 0056 PTD: 0056 元大高股息 | ▪ 2881 Fubon FHC: 2881 富邦金 |
| ▪ 2317 Hon Hai Precision:
2317 鴻海 | ▪ 2882 Cathay Holdings:
2882 國泰金 |
| ▪ 2330 TSMC: 2330 台積電 | ▪ 2884 E.S.F.H: 2884 玉山金 |
| ▪ 2454 MediaTek: 2454 聯發科 | ▪ 2885 Yuanta Group:
2885 元大金 |

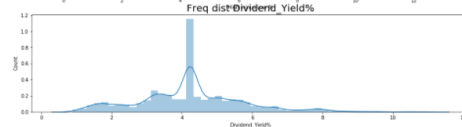
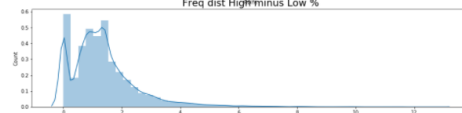
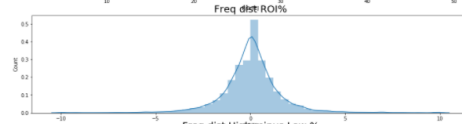
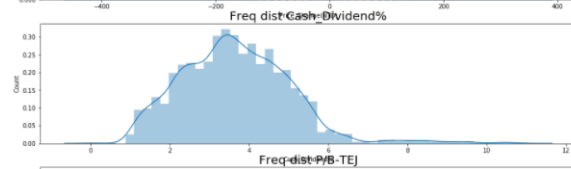
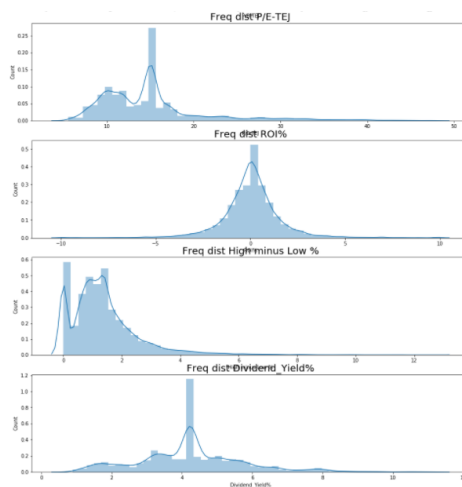
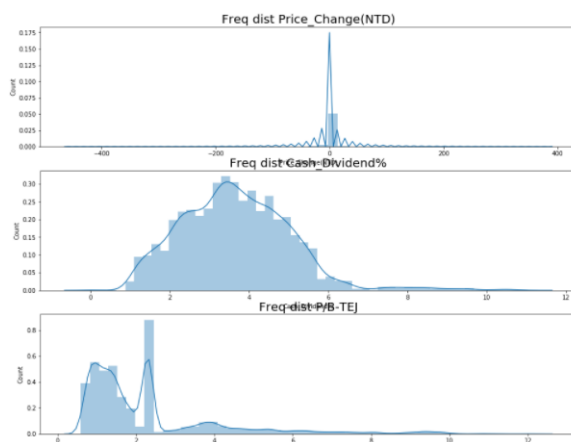
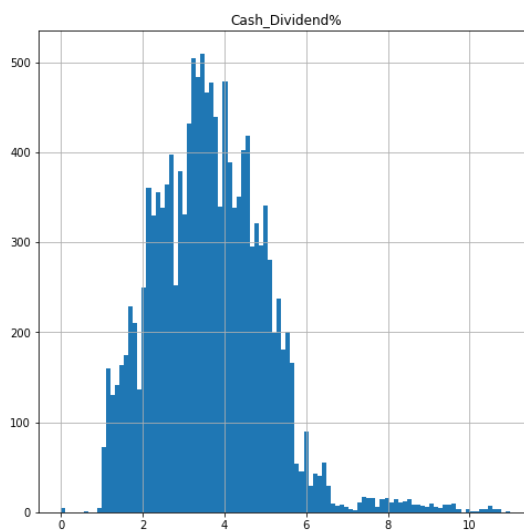
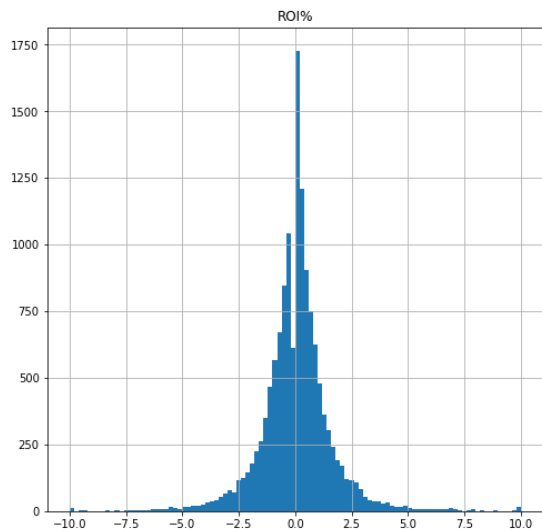
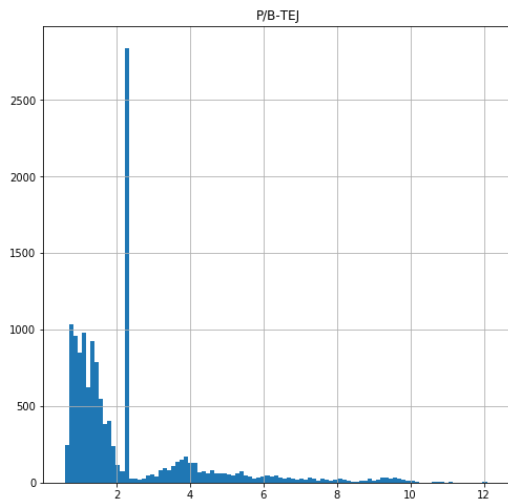
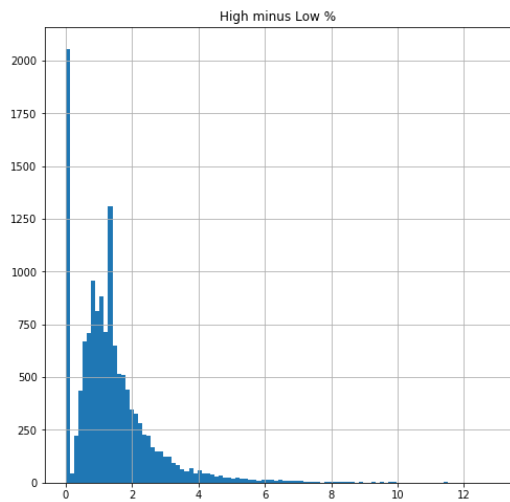
attribute:

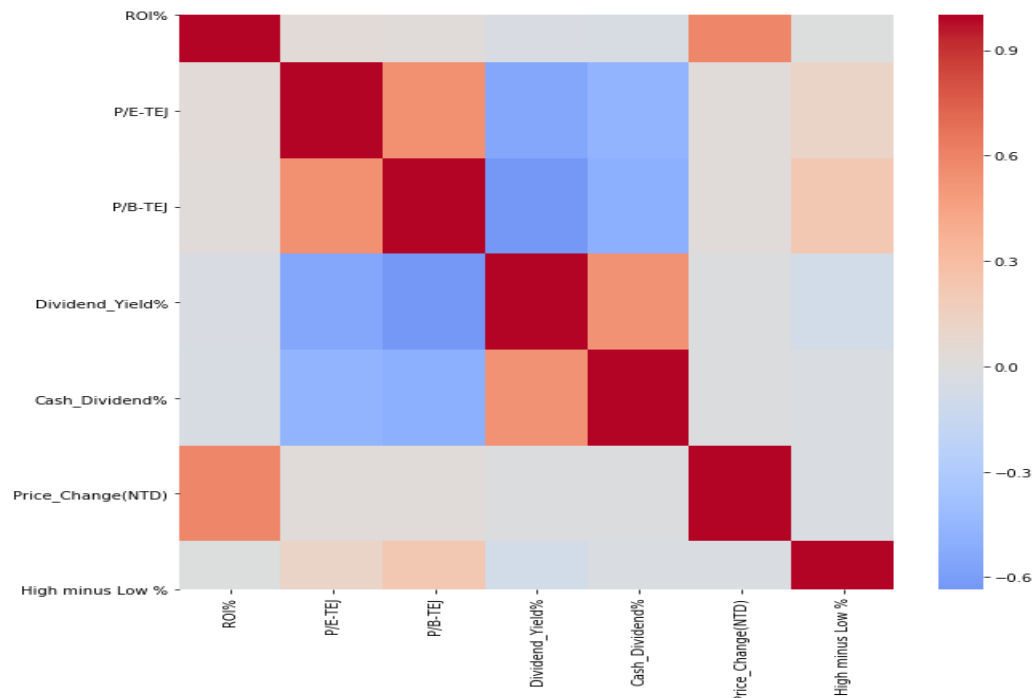
- CO_ID: 公司代碼
- Date: 年月日
- Open(NTD): 開盤價(元)
- High(NTD): 最高價(元)
- Low(NTD): 最低價(元)
- Close(NTD): 收盤價(元)
- Volume(1000S): 成交量(千股)
- Amount(NTD1000): 成交值(千元)
- AVG CLOSE: 當日均價(元)
- AVG CLOSE 5D: 5 日均價(元)
- AVG CLOSE 10D: 10 日均價(元)
- AVG CLOSE 20D: 20 日均價(元)
- AVG Vol 5D: 5 日均量
- AVG Vol 10D: 10 日均量
- AVG Vol 20D: 20 日均量
- ROI%: 報酬率%
- Shares(1000S): 流通在外股數(千股)
- Market Cap.(NTD MN): 市值(百萬元)
- P/E-TEJ: 本益比-TEJ
- P/B-TEJ: 股價淨值比-TEJ
- Dividend_Yield%: 股利殖利率
- Cash_Dividend%: 現金股利率
- Price_Change(NTD): 股價漲跌(元)
- High minus Low %: 高低價差%
- Market: 上市別
- Capital: 資本
- No.of Employee: 員工人數

EDA(Exploratory Data Analysis)

	ROI%	P/E-TEJ	P/B-TEJ	Dividend_Yield%	Cash_Dividend%	Price_Change(NTD)	High minus Low %
count	14020.000000	14020.000000	14020.000000	14020.000000	14020.000000	14020.000000	14020.000000
mean	0.046360	15.044964	2.296988	4.227955	3.663234	0.139675	1.391597
std	1.617399	6.357659	1.845026	1.513258	1.391331	32.966458	1.234087
min	-10.000000	5.530000	0.590000	0.890000	0.000000	-465.000000	0.000000
25%	-0.662300	10.760000	1.110000	3.320000	2.654900	-0.350000	0.659900
50%	0.000000	14.820000	1.660000	4.227955	3.590000	0.000000	1.183400
75%	0.732825	16.240000	2.296988	5.020000	4.540000	0.400000	1.800150
max	10.000000	47.550000	12.200000	11.000000	11.000000	390.000000	12.797600







Current Results

```

money since long: 28049.625
money since short: 22443.550225
money since long: 31160.625132389996
money since short: 128715.97324999167
money since long: 155839.64671309618
money since short: 242675.4458572173
money since long: 242675.4458572173
money since short: 917941.1080564369
money since long: 3001839.537302309
money since short: 3619167.838148529
money since long: 5601250.344308548
money since short: 11381250.590229843
money since long: 27109996.64029511
money since short: 38254916.259128435
money since long: 57785007.382307895
money since short: 93937474.93845655
money since long: 109253980.2271719
money since short: 249623494.02304232
money since long: 455026186.0799027
money since short: 801665134.6355726
money since long: 1016671723.7448331
money since short: 225524344.3175526
money since long: 6129923092.714725
money since short: 7409697786.396241
money since long: 18045763231.434063
money since short: 28285606005.070927
money since long: 45589325478.673065
money since short: 48716753206.51004
money since long: 48716753206.51004
money since short: 80751063236.8658
money since long: 94707875129.06758
money since short: 172923557652.84714
money since long: 255421063920.07916
money since short: 312654538817.9709
money since long: 391771769865.8584
money since short: 625708487947.0092
money since long: 1127863013592.782
money since short: 1202696724544.663
money since long: 1572616169596.4878
money since short: 1951144881618.3623
money since long: 4963029678128.548
money since short: 5717037961978.228
money since long: 7487104377981.212
money since short: 8863982873091.957
money since long: 17981808055712.086
money since short: 31191693798639.582
money since long: 31191693798639.582
money since short: 66003183662611.28
money since long: 131785256659952.81
money since short: 190485704607712.3
money since long: 234373610949329.22
money since short: 396519134344348.94
money since long: 396519134344348.94
money since short: 754233914903923.9
money since long: 1238961196164003.0
money since short: 1575865719431917.2
money since long: 1575865719431917.2
money since short: 3861737738753885.0
money since long: 6014656528109176.0
money since short: 1.1921876253985e+16

```


`final money: 1.1921876253985e+16`

從 10000 賺到 `1.1921876253985e+16`，很誇張的數字但覺得 `code` 的邏輯沒錯，因此認為有些變數可能影含未來的資訊或是 `model` 放入了一些未來的資訊，或是因為每次都是把錢花光，因此本金充足的情況下容易錢滾錢。

Conclusion

雖然結果有點太誇張，但每一次的買賣幾乎都在賺錢，所以認為起碼在預測漲跌方面 regression 已經算是很準了。而因為完全沒有人為因素，且流通性假設為很高，所以才能賺到不可思議的數字，又或是少了很多人為影響決策的因素，導致只憑 `model` 會大賺。

Future Use

希望能透過更多 ML 的方法，甚至是 DL 的方法，讓成果變得更好，未來也會考慮考量更多變數並篩掉不太有解釋效用的變數，也希望未來能不僅僅從技術面分析，也考量到基本面分析，這樣不僅能做的更全面，也更能分辨哪些因素才是重要因子，也期許未來能考量更多市場機制，例如流動性等等。