Machine Learning for Stock price prediction

王茗冠、王子軒、黃三騰

National Dong Hwa University

2019/06/13

大綱

研究目的

- 利用統計與機器學習的方法，預測未來股票漲或跌的可能性，提升投資的報酬率。

資料蒐集

- 利用資料爬蟲: 台灣證卷交易所、看盤軟體的個股交易資訊.
  - 資料來源的可靠性.
  - 三大法人的買賣資訊.

原始資料

	open	high	low	close	k_value	SMA_10	WMA_10	fi	de	d_value	it	rsi
2019-04-11 08:00:00	253.0	254.0	251.5	252.0	85.16878	248.00	250.2182	-2963431	-2963431	78.72874	-65000	67.52751
2019-04-12 08:00:00	251.5	253.0	251.0	252.0	84.16014	249.05	250.9455	2036947	2036947	80.53921	-31000	67.52751
2019-04-15 08:00:00	255.0	256.0	254.0	255.5	87.92494	250.40	252.1182	6786182	6786182	83.00112	-22000	73.82995
2019-04-16 08:00:00	257.0	257.0	255.5	257.0	91.94996	251.55	253.3182	3936165	3936165	85.98407	-247000	76.06932
2019-04-17 08:00:00	260.0	263.0	259.5	261.5	91.60300	253.15	255.1273	19234055	19234055	87.85705	-323000	81.43177
2019-04-18 08:00:00	264.0	266.0	263.5	264.5	91.17620	255.00	257.1909	17986624	17986624	88.96343	-75000	84.10339
2019-04-19 08:00:00	269.0	269.5	263.5	264.5	85.10846	256.80	258.9182	15656372	15656372	87.67844	2000	84.10339
2019-04-22 08:00:00	266.5	267.5	265.0	266.0	83.76600	258.10	260.5909	9604000	9604000	86.37429	-250197	85.42999
2019-04-23 08:00:00	266.5	268.0	266.0	268.0	86.47463	259.50	262.3909	9807070	9807070	86.40774	-42000	87.05092
2019-04-24 08:00:00	270.0	270.0	267.5	269.0	89.22870	261.00	264.1182	5748093	5748093	87.34806	114000	87.81353
2019-04-25 08:00:00	268.5	269.0	267.0	267.5	87.61080	262.55	265.3000	4535974	4535974	87.43564	-276000	79.87532
2019-04-26 08:00:00	262.0	263.0	257.5	260.0	68.75203	263.35	264.8364	-9696228	-9696228	81.20777	-10000	52.95047
2019-04-29 08:00:00	260.0	262.0	258.5	259.5	51.16802	263.75	264.1364	6386064	6386064	71.19452	-66000	51.64481
2019-04-30 08:00:00	260.0	260.5	258.0	259.0	38.11201	263.95	263.2727	6366172	6366172	60.16702	109000	50.25084
2019-05-02 08:00:00	261.5	262.5	258.5	259.0	29.40801	263.70	262.3727	11474951	11474951	49.91401	-18000	50.25084
2019-05-03 08:00:00	262.0	265.0	260.5	265.0	39.60534	263.75	262.6091	11473494	11473494	46.47779	9075	64.71526
2019-05-06 08:00:00	260.0	260.0	258.0	259.0	30.40356	263.20	261.7455	-7570053	-7570053	41.11971	85000	48.76477
2019-05-07 08:00:00	259.5	263.0	259.0	262.5	33.60237	262.85	261.6182	-3941242	-3941242	38.61393	817956	55.89812
2019-05-08 08:00:00	260.0	261.5	259.5	260.0	29.64796	262.05	261.1000	-8064778	-8064778	35.62527	-88023	50.27356
2019-05-09 08:00:00	259.5	259.5	256.0	256.5	21.61716	260.80	260.0909	-4699965	-4699965	30.95590	10000	43.39616
2019-05-10 08:00:00	257.0	259.0	255.0	256.0	17.74477	259.65	259.2182	-4181858	-4181858	26.55219	-318000	42.46260

反應變數與解釋變數

1. 收盤價：為個股當天收盤的價格。

2. 移動平均線 (SMA)：N 日收盤價總和/N。這裡我們用五日、十日及二十日，若股價向上突破 SMA 則代表股價走強，為買進訊號，反之亦然。

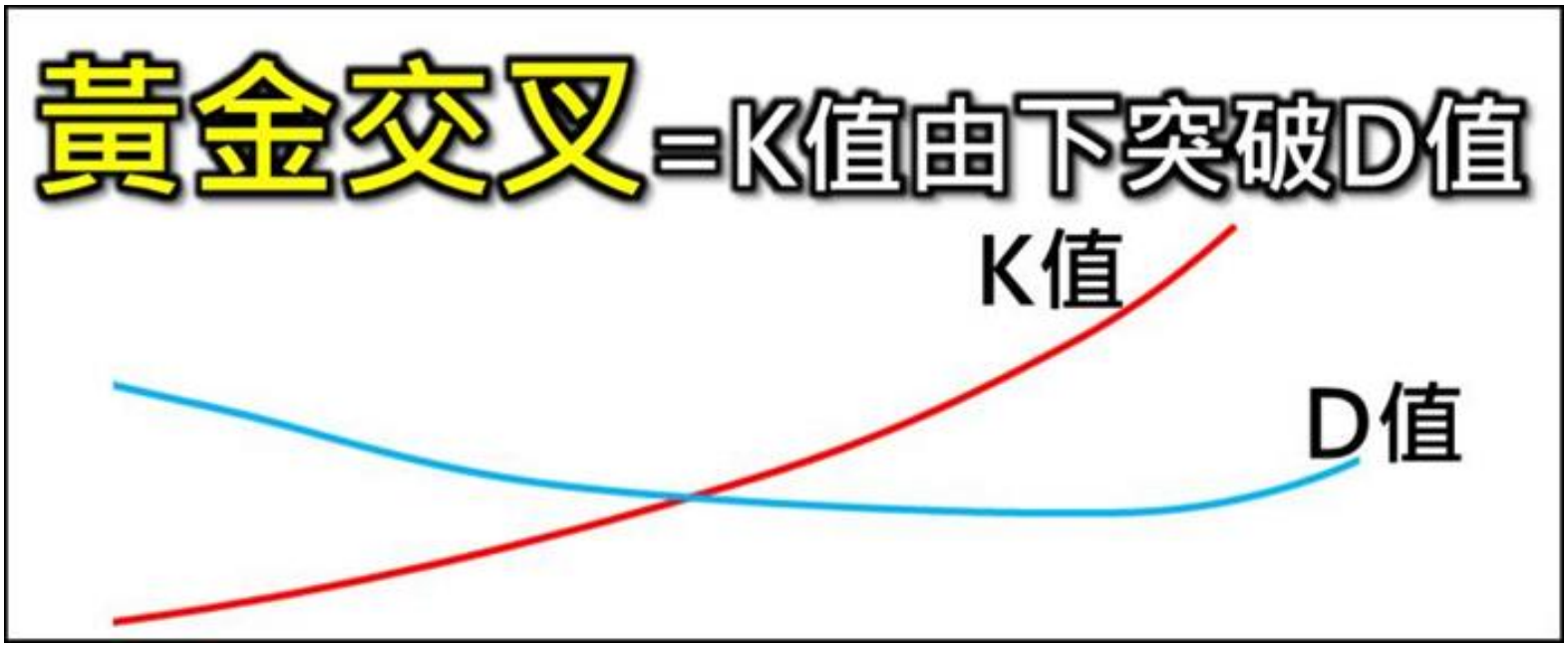
3. 加權移動平均線 (WMA)：(N 日前收盤價 *1+(N-1) 日前收盤價 *2+⋯+ 今日收盤價 *N) / (1+2+⋯+N)。我們用十日的股價加權平均。

反應變數與解釋變數

4. 相對強弱指標 (RSV)：(今日收盤價-九日內最低價)/(九日內最高價-九日內最低價)*100，介於 0 到 100 之間。

5. 隨機指標 (KD)：可分為 K 值與 D 值。K 值 = 前日 K 值 $\times\frac{2}{3}$ + 當日 RSV$\times\frac{1}{3}$；D 值 = 前日 D 值 $\times\frac{2}{3}$ + 當日 K 值 $\times\frac{1}{3}$。介於 0 到 100 間。
當 K 值由下而上穿越 D 值，為黃金交叉，行情看好；當 K 值由上而下跌破 D 值，為死亡交叉，行情看壞。

KD 指標圖形



反應變數與解釋變數

6. 相對強弱指標 (RSI)：(九日股價上漲幅度的加總/ (九日股價上漲幅度的加總 + 九日股價下跌幅度的加總)) *100。越高時代表市場越熱絡，越低時越冷清。其中 RSI 值大於 80 時，股價下跌的機率大；RSI 值小於 20 時，上漲的機率高。

7. W%R：(十日內最高價 - 今日收盤價)/ (十日內最高價 - 十日內最低價)* -100%，一般來說高於-20% 為超買; 反之，低於-80% 以下為超賣，值為 0 時表示今日收盤價為十日內最高價。

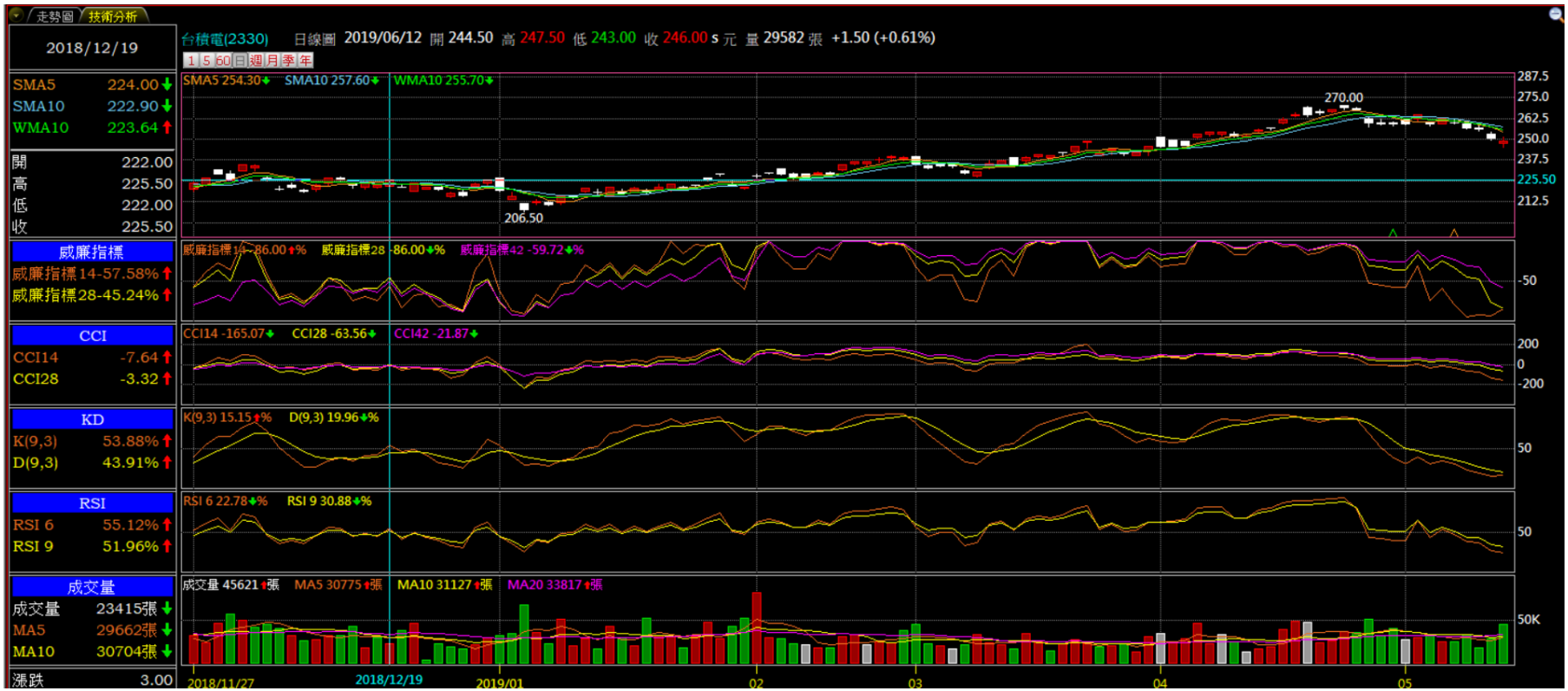
反應變數與解釋變數

8. 動量指標 (Mom)：當日股價 - 前九天的股價，由此可以看出股價在其中波段漲跌幅度。

9. CCI: 當典型價格 (TP) 等於其平均值時，CCI 值會等於零，只有當最後股價在極短期內作劇烈的向上或向下運動時，CCI 值才會出現突然向上或向下大幅擺盪的極端值。

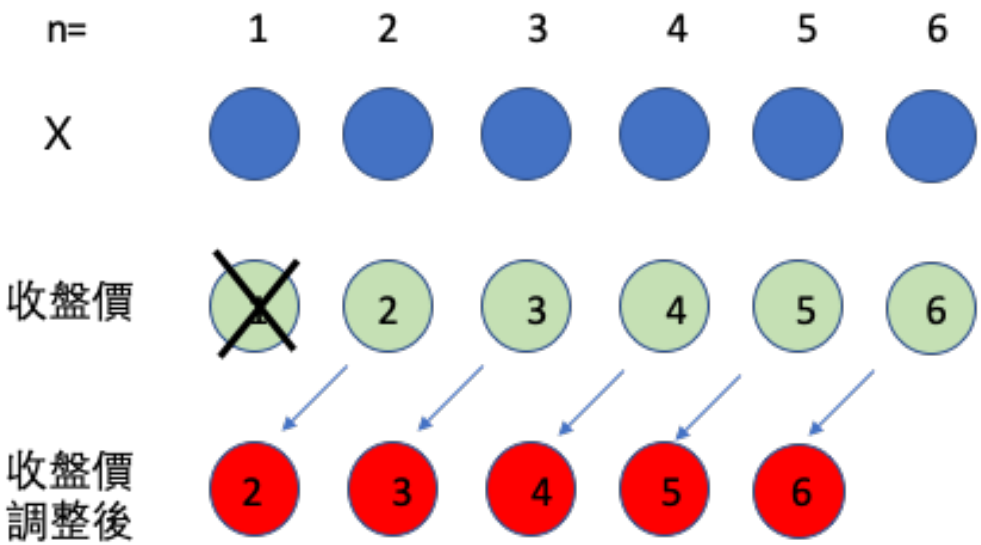
10. 三大法人：分別為外資、投信、自營商，所佔的持股比例。一般而言，若外資持股比例在股票市場中佔有較大，則股價的漲跌會容易受到外資的買賣而影響。

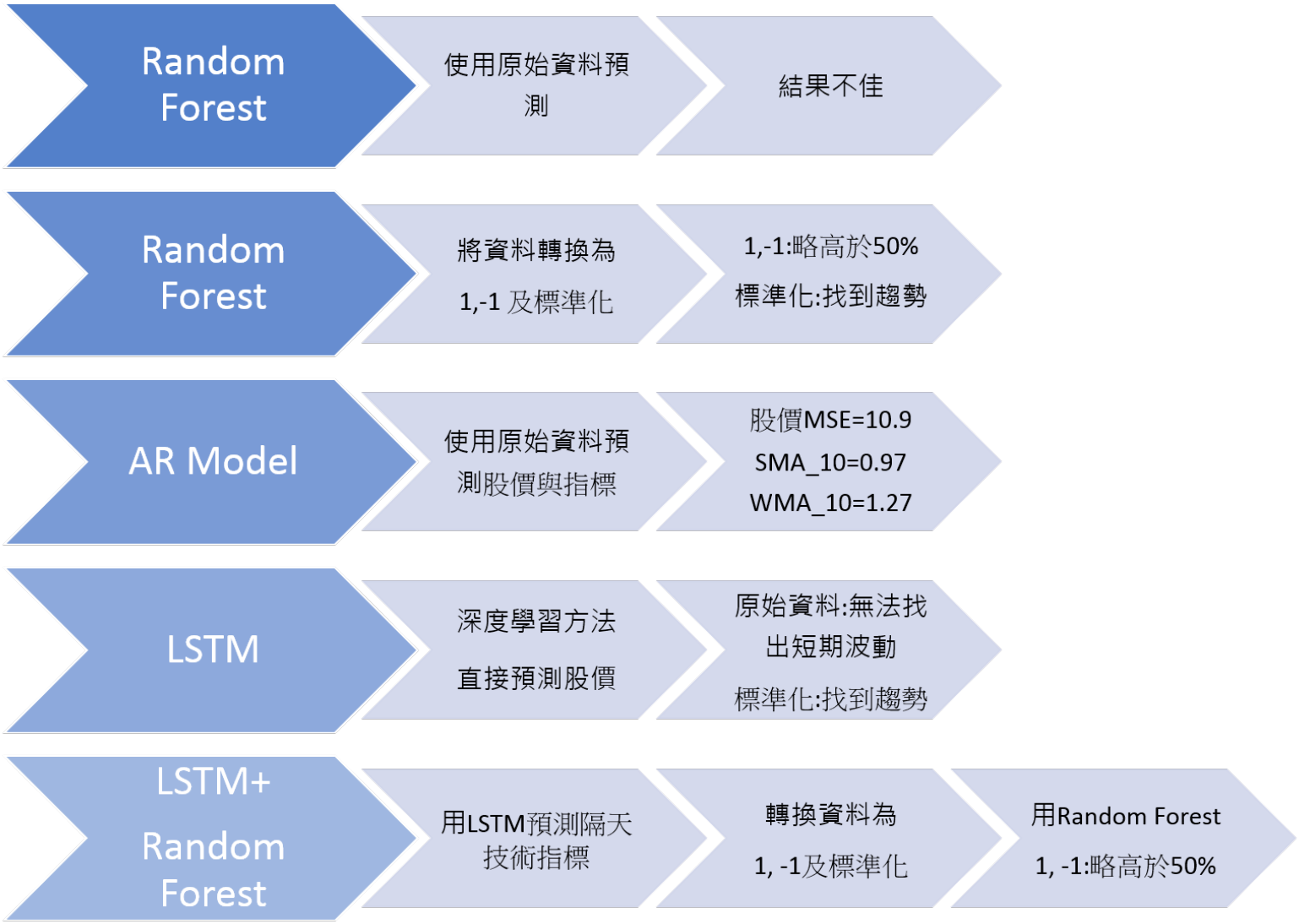
RSI,RSV,KD,W%R 指標



預測明日股價

- Y: 明日收盤價；$X_1, X_2, \ldots, X_p$: 當日各種技術指標值.




Random Forest
使用原始資料預測
結果不佳
Random Forest
將資料轉換為 1,-1 及標準化
1,-1:略高於50%
標準化:找到趨勢
AR Model
使用原始資料預測股價與指標
股價MSE=10.9
SMA_10=0.97
WMA_10=1.27
LSTM
深度學習方法
直接預測股價
原始資料:無法找出短期波動
標準化:找到趨勢
LSTM+ Random Forest
用LSTM預測隔天技術指標
轉換資料為 1, -1及標準化
用Random Forest
1, -1:略高於50%

Random Forest-原始資料

- train error (MSE): 0.5560 ; test error (MSE): 1983.96

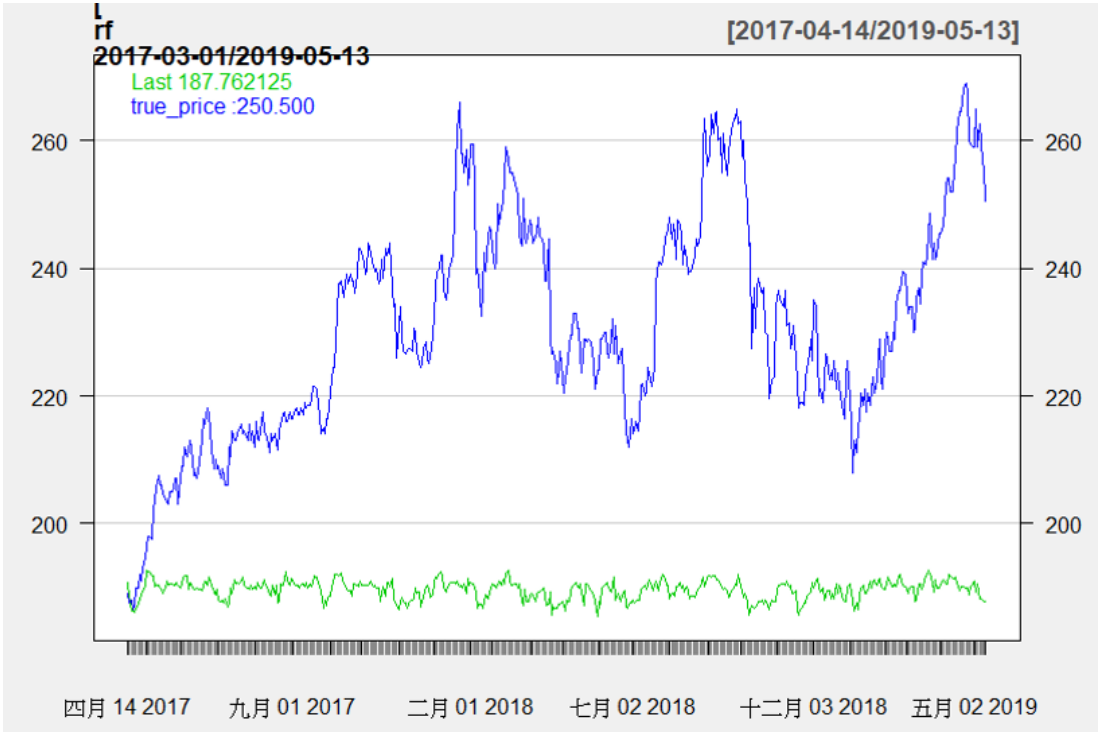


Figure: 籃色: 真實值；綠色: 預測值

Random Forest-資料標準化 (1)

- Y 不標準化, X 標準化 (train & test 分別做).
- train error (MSE): 0.5664 ; test error (MSE): 16383.32

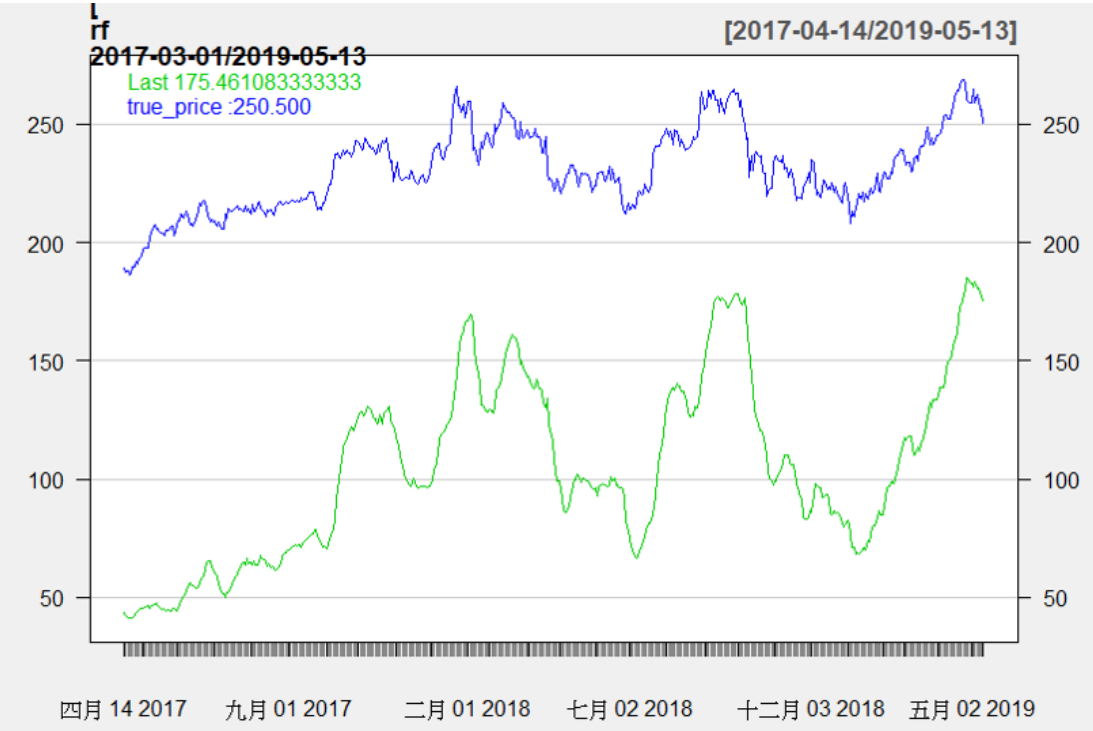


Figure: 籃色: 真實值；綠色: 預測值

Random Forest-資料標準化 (2)

- Y、X 皆標準化 (train & test 分別做).
- train error (MSE): 0.5623 ; test error (MSE): 33.40811

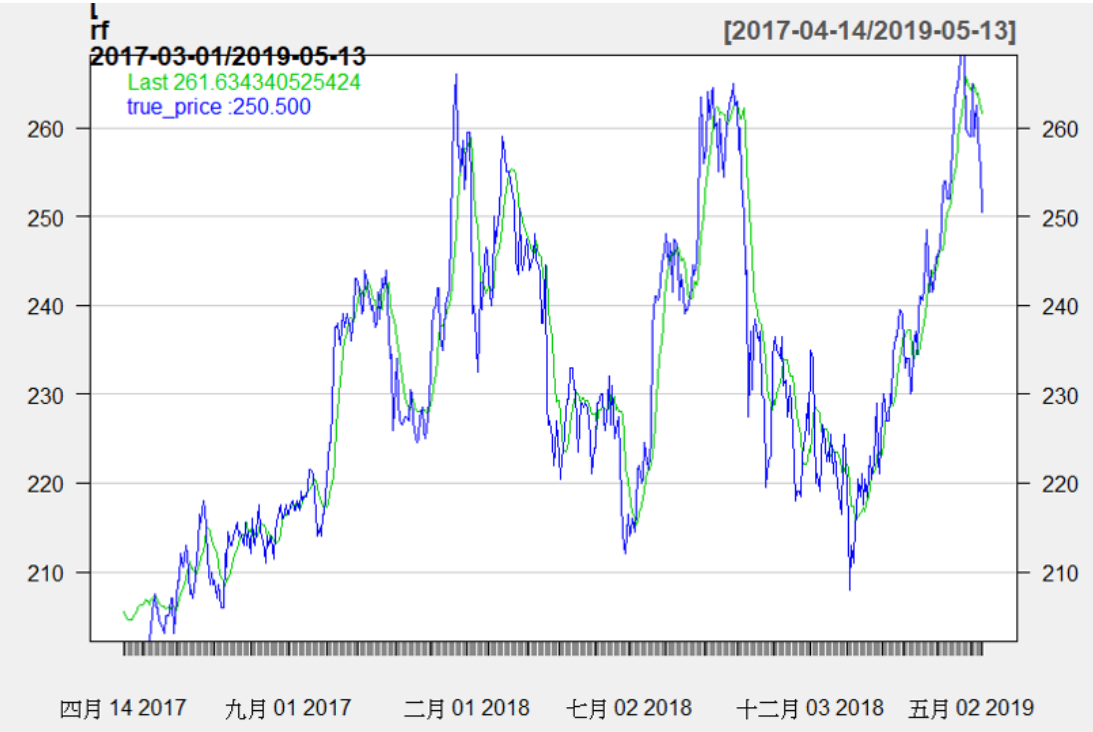


Figure: 籃色: 真實值；綠色: 預測值

Random Forest-資料轉換

- 參考 Patel, Jigar, et al (2015) [1],

- 將明日收盤價與當日做比較得到股價的漲跌: $Y = +1$ 代表漲; $Y = -1$ 代表跌.
- 當日為止技術指標及三大法人資訊 $(X_1, ..., X_p)$ 值轉換為 ± 1: $+1$ 代表看漲; -1 代表看跌.

predict \ true	-1	1
-1	163	149
1	95	104

Train_error(錯誤率): 27.25%
Test_error(錯誤率): 47.74%

Accurancy(正確率): 52.26%

AR model

Autoregressive model (AR model):

$$Y_t = \beta_0 + \sum_{j=1}^{p} \beta_j Y_{t-j} + \epsilon_t,$$

where ϵ_t is distributed with mean 0, variance σ^2.

- Y_t 當期值等於一個或數個過去期 $Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p}$ 的線性組合, 加常數項, 加隨機誤差.
- AR(1): $p = 1$, 利用今天的股價去預測明天的股價.

AR model

- 同樣地, 今天的技術指標去預測明天的技術指標.

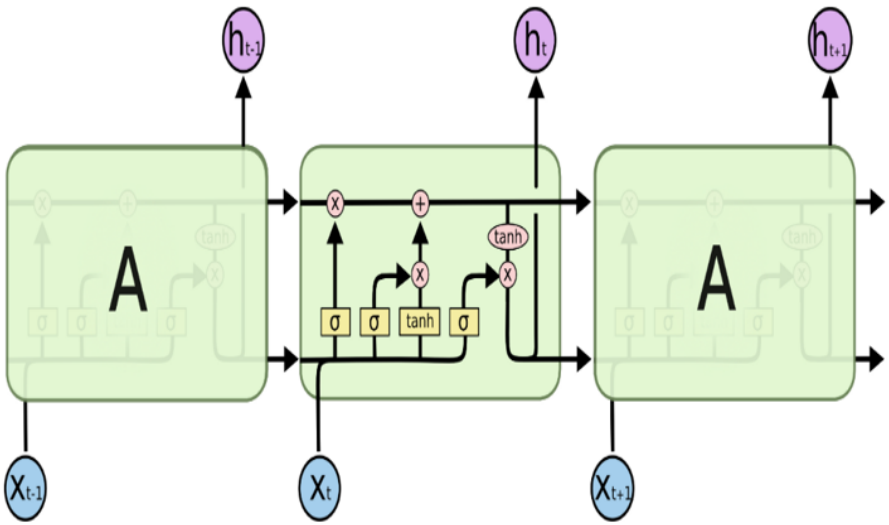
MSE	AR(1)	AR(2)	AR(3)	AR(4)	AR(5)
Y(股價)	10.9944	10.9740	11.1530	11.1813	11.2540
SMA_10	0.9659	0.2002	0.2007	0.2013	0.2010
WMA_10	1.2684	0.3748	0.3758	0.3759	0.3752
RSI_9	48.5738	48.7603	49.1203	49.2162	49.3017
K_value	72.2447	55.7346	55.2089	55.5410	55.4482
D_value	24.4506	6.3513	6.0907	6.1434	6.1638
CCI_14	1993.6600	1941.0450	1944.4980	1945.9440	1949.0350

AR model

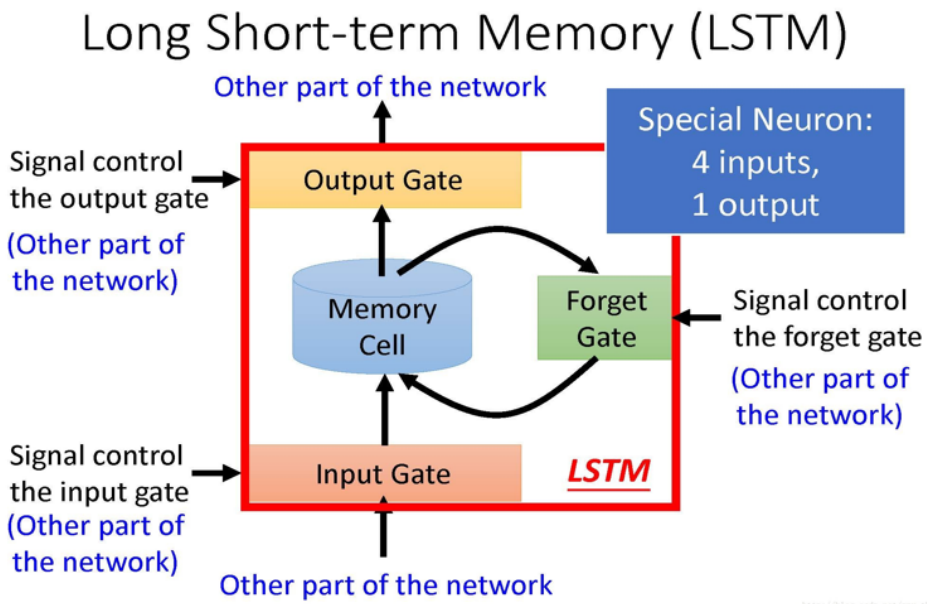
- 在 AR(1), Y: 明天的股價; 解釋變數: 今天的股價.
- train error (MSE): 2.4788 ; test error (MSE): 10.9944

Figure: 籃色: 真實值；綠色: 預測值

LSTM 模型講解



The repeating module in an LSTM contains four interacting layers.



LSTM-直接預測股價

- 將 X 和 Y 皆標準化, 直接丟入 LSTM 進行預測.
- train error (MSE): 2.5295 ; test error (MSE): 19.0717

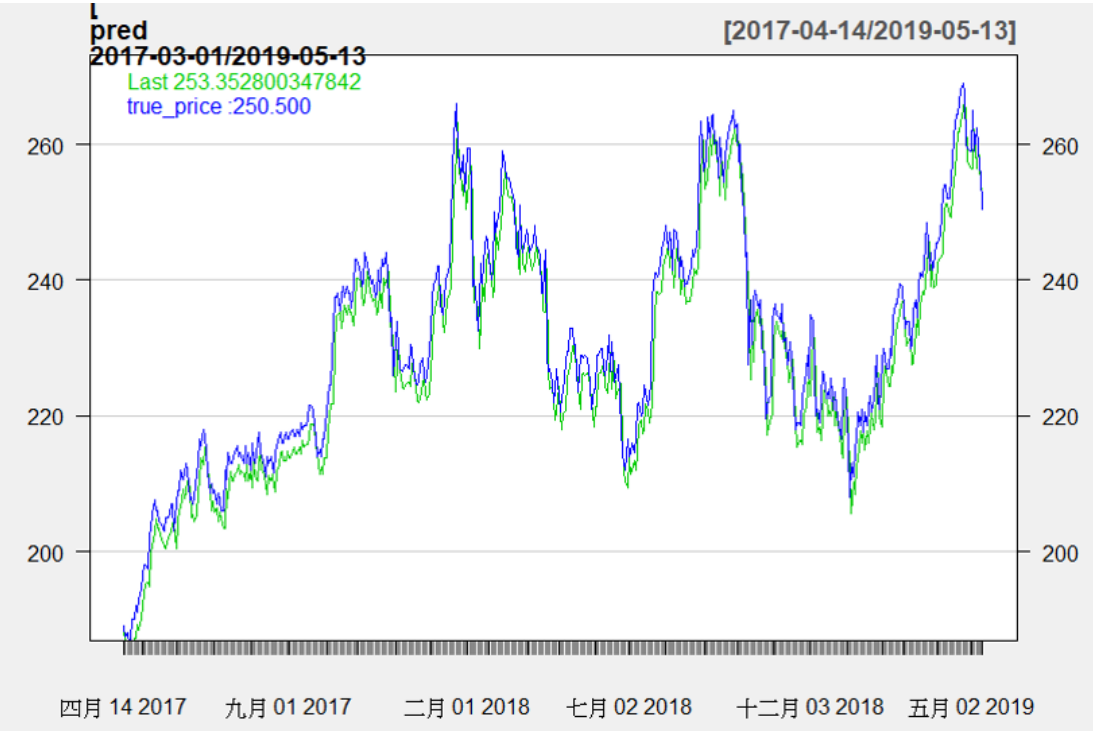


Figure: 籃色: 真實值；綠色: 預測值

LSTM+Random Forest(1)

- 先用 LSTM 估計出明日的 $X_1, ..., X_p \rightarrow$ 轉換成 $\pm 1 \rightarrow$ 由 Random Forest 預測明日的股價漲跌 $Y = \pm 1$.

(i) LSTM 估計技術指標 $X_1, ..., X_p$.

LSTM	SMA10	WMA10	RSI	Mom	k_value	d_value	CCI	%R
MSE	0.2053	0.0863	7.0251	0.2827	1.0709	1.671	458.5	53.77

LSTM+Random Forest(1)

(ii) Random Forest 預測明日的股價漲跌 $Y = \pm 1$.

predict \ true	-1	1
-1	175	150
1	83	103

Test_error(錯誤率): 45.59%
Train_error(錯誤率): 27.25%

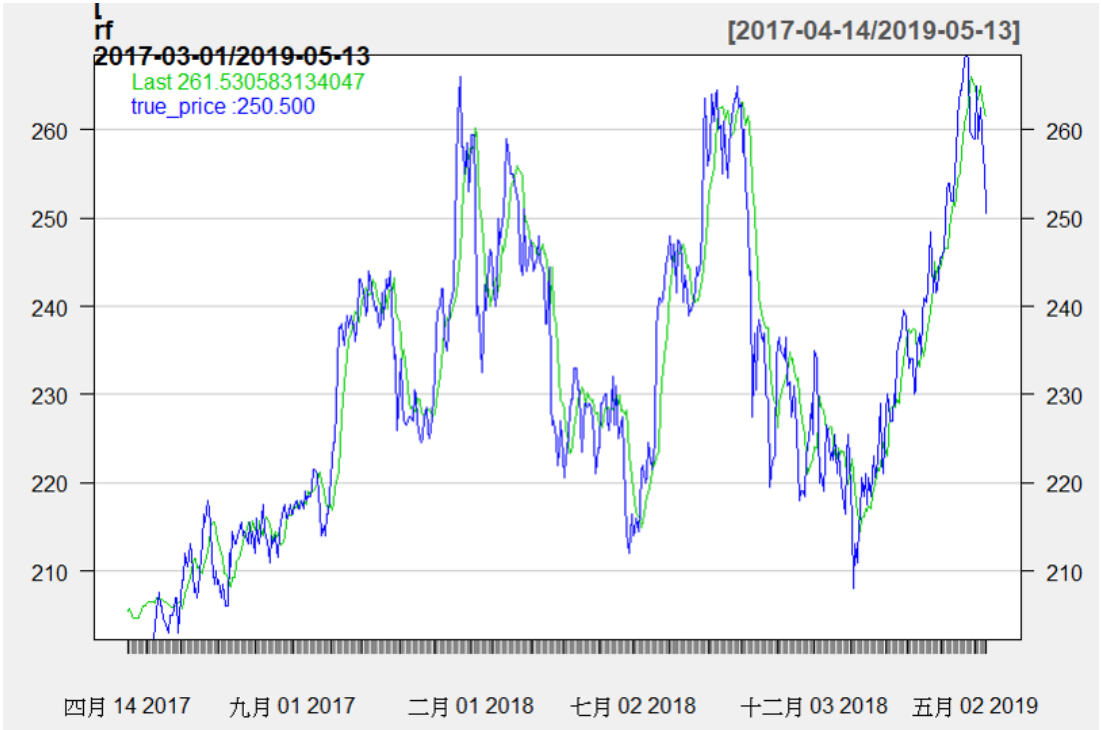
Accurancy(正確率): 54.41%

LSTM+Random Forest(2)

- 先用 LSTM 估計出明日的 $X_1, ..., X_p \rightarrow$ 由 Random Forest 預測明日的股價 Y.

其中 X 和 Y 皆標準化 (train & test 分別做).

train error (MSE): 0.4942 ; test error (MSE): 38.2950



模型比較

	Random Forest	AR	LSTM+ Random Forest
變數資料	多	少	多
訓練速度	快	快	最慢
準確率	普通	高	普通

問題討論 (1)

- 資料是否有標準化，對預測的結果差異極大。
- 將技術指標轉換成 1,-1 的方法並不可行。
- Random Forest 和 Random Forest + LSTM 方法預測差異不大。

問題討論 (2)

- AR 模型預測結果最佳，可能原因在於資料包含時間性，因此機器學習方法預測誤差較大，利用時間序列模型會得到較好結果，未來也可以試著利用時間序列模型結合機器學習方法，或許會得到更好的結果。
- 由於我們目前是只預測隔天的股價，期望未來可以做出預測長期趨勢，或是預測出最佳投資策略。

- 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.
- LSTM 作法: http://rwanjohi.rbind.io/2018/04/05/time-series-forecasting-using-lstm-in-r/
- Random Forest: https://www.finlab.tw/Machine-learning
- Ar model: http://homepage.ntu.edu.tw/ sschen/Book/Slides/Ch3AR.pdf