BDSE18 專題發表

台 股 大 數 據

李子顥 李謀勳 陳姿伶 劉文裕 陳彥伶

報告流程

李子顥

團隊介紹 → 專題介紹 → 平台建構 → 資料蒐集與處理 → 建模應用 → 資料視覺化 → 結論與展望

團隊介紹

李子顥

- **主要:組長**
- 專題架構規劃
- 資料蒐集與處理
- 視覺化

李謀勳

- **主要:Hadoop平台建構**
- 資料蒐集與處理

團隊介紹

陳姿伶

- **主要:資料蒐集與處理**

- 網路爬蟲
- 機器學習與深度學習
- 資料視覺化

Linkedin

Github: Gaspardetlisa

團隊介紹

劉文裕

- 主要:機器學習與深度學習
- 資料處理

陳彥伶

- 主要:資料視覺化與網頁
- 網路爬蟲
- 資料探勘

台灣股市

台股環境介紹-問題點

李子顥

- 投資者是有限理性的
- 資訊不對稱

Why?

台股環境介紹-公司家數

- 上市：948家
- 上櫃：785家
- 總計：1733家

台股環境介紹-成交量

李子顥

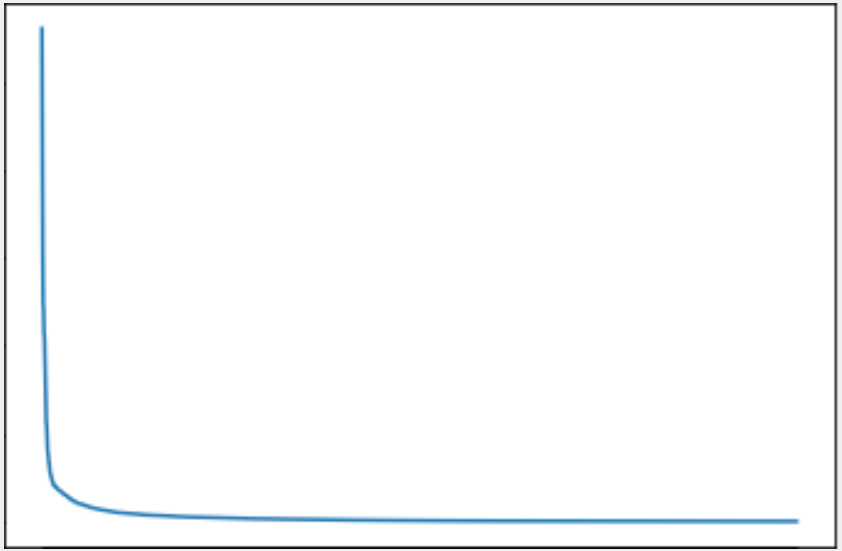
台灣股市個股
四月總成交張數

Stock_id	Volume (張)
3481	11214990
2603	7628701
2409	6266219
6116	4973994
2610	4816802
…	…
2740	93
3629	89
6574	67
2924	64
6236	57

台股環境介紹-成交量圖

李子顥

台灣股市個股
四月總成交張數



公司家數

台股環境介紹-價量意義

- 每天熱門成交的公司數量僅前面的1-200家不到
  - 市場上的熱錢追逐相同標的
    - 價格波動大：短時間收益/虧損可觀
    - 受到資訊影響
    - 容易炒作過頭

台股環境介紹-克服問題點

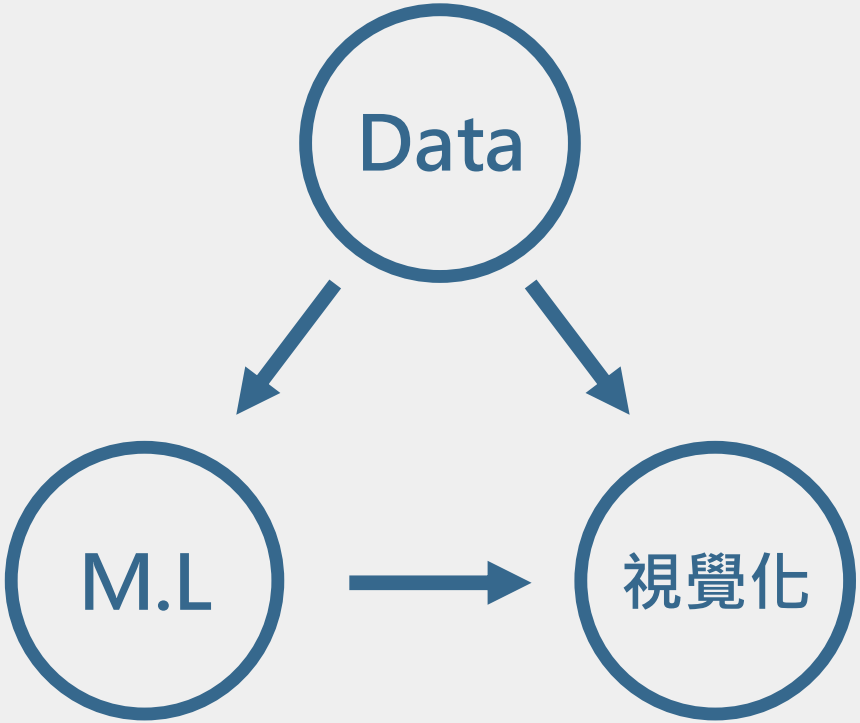
- 投資者是有限理性的
- 資訊不對稱

專題目的

- 程式處理大量的金融資料
- 資料中篩選出與股價關聯度高的欄位
- 建立理性投資觀念的基礎

專題架構

李子顥



專題架構與使用工具

李子顥

專題環境	資料收集、清洗	大數據平台	ML & DL	視覺化
python	Selenium	ubuntu	pandas	HTML JS CSS
jupyter	pandas	hadoop	RANDOM FOREST	D3
	Linux	APACHE Spark	XGBoost	
			LSTM	

專題協作架構

李子顥

GitHub

平台建構

主講人:李謀勳

使用虛擬環境和Hadoop的優點

- HA 高可用性: 容許特定數量虛擬機下線(損毀、維修)
- HDFS 分散式檔案系統: 透過檔案分割、異地存放。讓檔案不易損毀、遺失，確保檔案的完整性
- 檔案壓縮: 透過Pig和Parquet兩種方式，可再度壓縮原本HDFS裡的檔案
- 運算效率: 分散式運算，有效利用各台電腦的運算效能

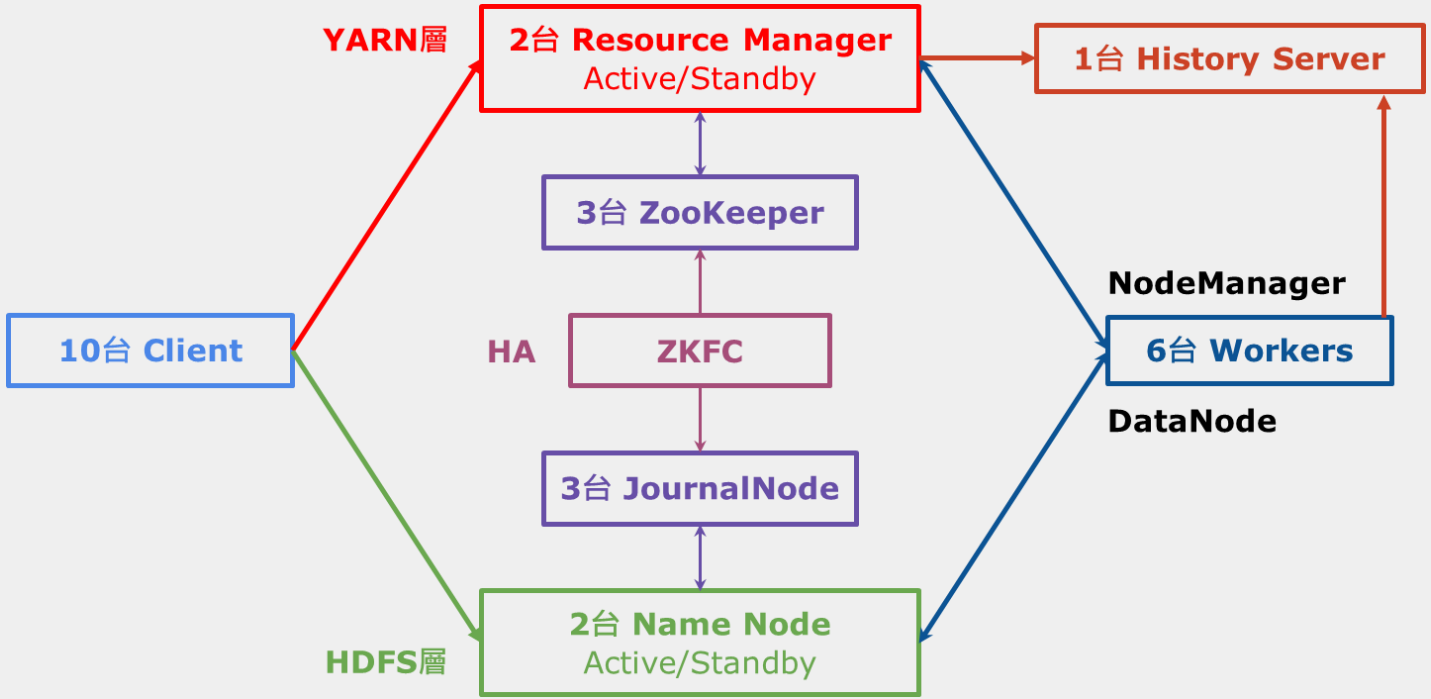
Hadoop 叢集規格：

- 五位組員、五台實體主機：
  - 十台 虛擬主機：
    - 4 核心
    - 24GB RAM
    - 動態硬碟配置，目前最高500GB

- 叢集：
  - 六台 Workers：
    - 24核心、144GB RAM、最高3TB

Hadoop HA叢集工作架構

李謀勳



Hadoop 節點簡介

YARN層

2台 Resource Manager
Active/Standby

負責資源和任務管理

NodeManager

HDFS層

2台 Name Node
Active/Standby

負責分散式儲存

DataNode

Hadoop 節點簡介

李謀勳

Map-Reduce 負責分散式計算

基於 YARN層 + HDFS層 + 1台 History Server

HA

- 3台 ZooKeeper
- ZKFC
- 3台 JournalNode

高可用性，監控狀態
切換Active/StandBy 的RM/NN

Hadoop Demo -單機

台股
每日交易資料(十年)
300mb，400萬筆

```
[1]: # import modules
     import pandas as pd
     import numpy as np

[5]: from pyspark.sql import SparkSession

[6]: spark.sparkContext.appName

[6]: 'sqldemo'

[7]: spark.conf.set("spark.sql.execution.arrow.pyspark.enabled", True)

[8]: import databricks.koalas as ks

[9]: ks.set_option("compute.default_index_type", "distributed")

[10]: %time perfdf = ks.read_csv('daily.csv')

      CPU times: user 33.9 ms, sys: 9.99 ms, total: 43.9 ms
      Wall time: 19.4 s

[12]: %time perfdf.shape

      CPU times: user 2.18 ms, sys: 2.82 ms, total: 5 ms
      Wall time: 4.46 s

[12]: (4078037, 11)
```

李謀勳

Hadoop Demo -Yarn 叢集

台股
每日交易資料(十年)
300mb，400萬筆

```
[1]: # import modules
     import pandas as pd
     import numpy as np

[2]: from pyspark.sql import SparkSession

[3]: spark.sparkContext.appName

[3]: 'PySparkShell'

[4]: spark.conf.set("spark.sql.execution.arrow.pyspark.enabled", True)

[5]: import databricks.koalas as ks

[6]: ks.set_option("compute.default_index_type", "distributed")

[7]: %time perfdf = ks.read_csv("/user/stock/daily.csv")

     CPU times: user 2.28 ms, sys: 23.4 ms, total: 25.6 ms
     Wall time: 7.21 s

[8]: %time perfdf.shape

     CPU times: user 0 ns, sys: 7.84 ms, total: 7.84 ms
     Wall time: 751 ms

[8]: (4078037, 11)
```

李謀勳

資料蒐集與處理

主講人:陳姿伶

資料爬蟲

陳姿伶

使用套件：各網站 API, requests, pandas
抓取上市上櫃台股資料

爬取的網站：

1. 台灣證券交易所
2. GoodInfo! 股市資訊
3. FinMind

資料爬蟲

陳姿伶

拿取資料：

- 每日交易量
- 法人買賣
- 個股融資融卷
- 當日沖銷交易標的及成交量值
- 個股股利、PER、PBR資料表
- 股東結構表

資料清理與整理

使用 Linux 合併表格

使用 Pandas 轉換資料格式，
如日期格式，
format=%Y/%m/%d，
以及 int 轉 float64…等。

每日交易量
法人買賣
個股融資融卷
當日沖銷交易標的及成交量值
個股股利、PER、PBR資料表
股東結構表

58個欄位

資料清理與整理

陳姿伶

個股基本資料

欄位名稱	對照名稱
index	索引
stock_id	股票代碼（primary_key)
stock_name	公司名稱（primary_key)
ISIN_Code	國際證券辨識號碼(ISIN Code)
Listing_date	上市日
Listing_category	市場別
Industry_category	產業別
CFICode	CFICode

無空值

資料清理與整理

陳姿伶

日交易

欄位名稱	對照名稱
date	交易日期
Volume	成交量
Volume_Cash	成交金額
Open	開盤價
High	最高價
Low	最低價
Close	收盤價
Change	漲跌幅（會有正負號）

無空值

資料清理與整理

陳姿伶

個股股利、PER、PBR資料表

法人買賣

欄位名稱	對照名稱
dividend_yield	殖利率
PER	本益比
PBR	股價淨值比
Dealer_buy	自營商_買進
Dealer_Hedging_buy	自營商避險_買進
Dealer_self_buy	自營商_買進
Foreign_Dealer_Self_buy	外資自營商_買進
Foreign_Investor_buy	外資_買進
Investment_Trust_buy	投信_買進
Dealer_sell	自營商_買進
Dealer_Hedging_sell	自營商避險_賣出
Dealer_self_sell	自營商_買進
Foreign_Dealer_Self_sell	外資自營商_賣出
Foreign_Investor_sell	外資_賣出
Investment_Trust_sell	投信_賣出

有空值

資料清理與整理

陳姿伶

股東結構表
（外資持股表）

欄位名稱	對照名稱
ForeignInvestmentRemainingShares	外資尚可投資股數
ForeignInvestmentShares	外資持有股數
ForeignInvestmentRemainRatio	外資尚可投資比例
ForeignInvestmentSharesRatio	外資持股比例
ForeignInvestmentUpperLimitRatio	外資投資上限
ChineseInvestmentUpperLimitRatio	陸資投資上限
NumberOfSharesIssued	發行股數
RecentlyDeclareDate	最近一次異動申報日期
StockShareholding_note	StockShareholding_備註
ForeignInvestmentRemainingShares	外資尚可投資股數
ForeignInvestmentShares	外資持有股數
ForeignInvestmentRemainRatio	外資尚可投資比例
ForeignInvestmentSharesRatio	外資持股比例
ForeignInvestmentUpperLimitRatio	外資投資上限
ChineseInvestmentUpperLimitRatio	陸資投資上限

有空值

資料清理與整理

陳姿伶

個股融資融卷

欄位名稱	對照名稱
MarginPurchaseBuy	融資買進
MarginPurchaseCashRepayment	融資現金償還
MarginPurchaseLimit	融資限額
MarginPurchaseSell	融資賣出
MarginPurchaseTodayBalance	融資今日餘額
MarginPurchaseYesterdayBalance	融資昨日餘額
OffsetLoanAndShort	資券互抵
ShortSaleBuy	融券買進
ShortSaleCashRepayment	融券償還
MarginPurchaseBuy	融資買進
MarginPurchaseCashRepayment	融資現金償還
MarginPurchaseLimit	融資限額
MarginPurchaseSell	融資賣出
MarginPurchaseTodayBalance	融資今日餘額
MarginPurchaseYesterdayBalance	融資昨日餘額

有空值

資料清理與整理

陳姿伶

個股融資融卷

當日沖銷交易標的及成交量值

欄位名稱	對照名稱
OffsetLoanAndShort	資券互抵
ShortSaleBuy	融券買進
ShortSaleCashRepayment	融券償還
ShortSaleLimit	融券限額
ShortSaleSell	融券賣出
ShortSaleTodayBalance	融券今日餘額
ShortSaleYesterdayBalance	融券昨日餘額
Note	註記
BuyAfterSale	可否當沖
Trading_Volume	成交量
BuyAmount	買進金額
SellAmount	賣出金額

有空值

建模應用

特徵擴充、篩選

主講人:劉文裕

投資分析四個面向

劉文裕

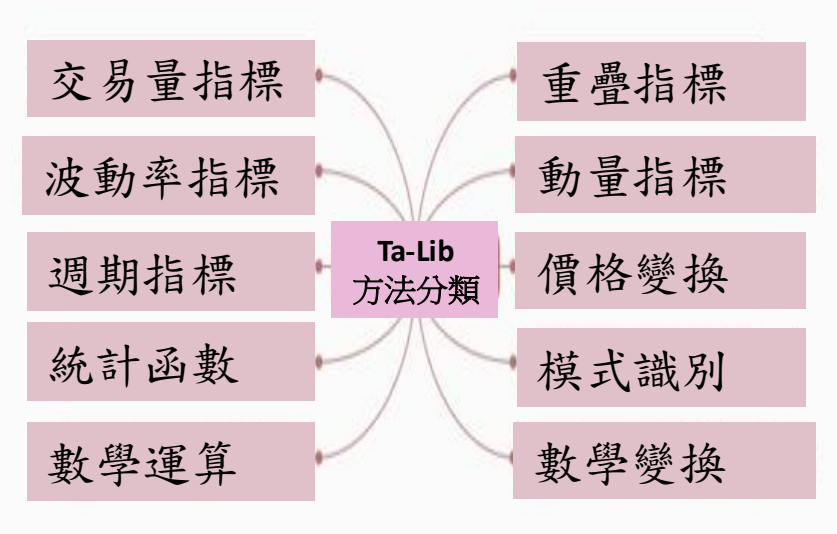
基本面
（公司營收與獲利狀況）

技術面
（技術線圖、均線分析）

籌碼面
（每日三大法人買賣超數量）

消息面
（新聞利多、利空資訊）

TA-Lib



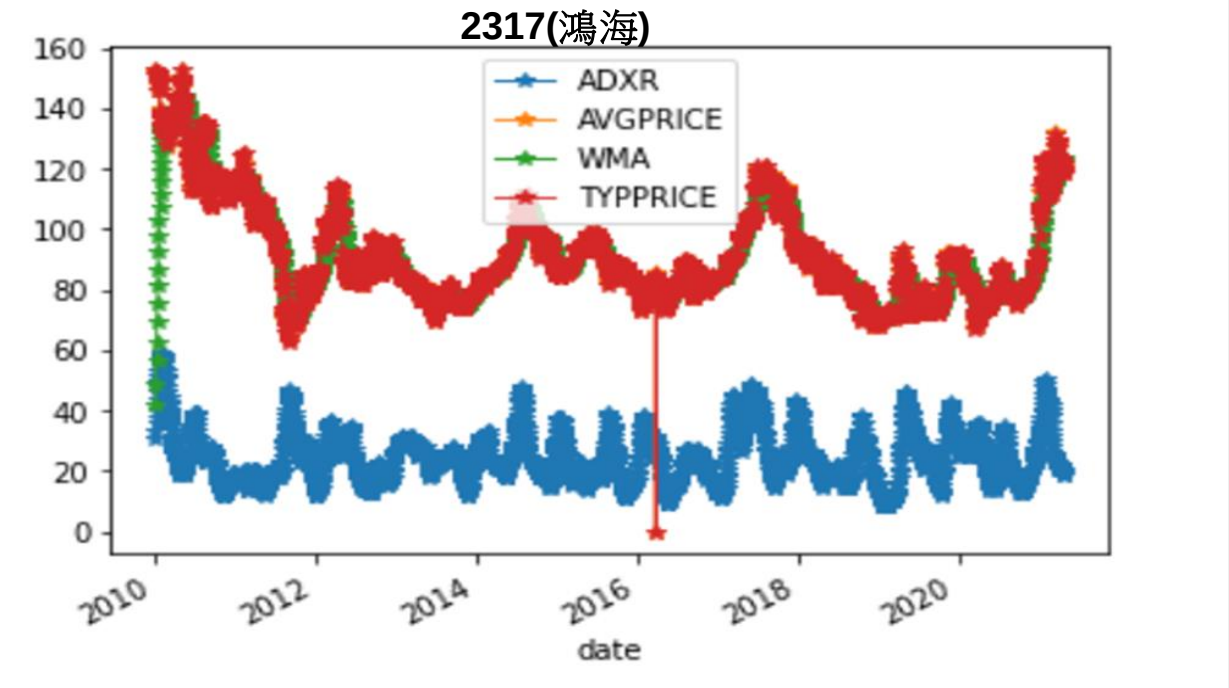
200 種指標

TA-Lib可分為10大主要指標:

- Overlap Studies(重疊指標)
- Momentum Indicators(動量指標)
- Volume Indicators(交易量指標)
- Cycle Indicators(週期指標)
- Price Transform(價格變換)
- Volatility Indicators(波動率指標)
- Pattern Recognition(模式識別)
- Statistic Functions(統計函數)
- Math Transform(數學變換)
- Math Operators(數學運算)

TA-Lib(指標圖)

劉文裕



ADXR:平均趨向指數的趨向指數
AVGPRICE:平均價格函數
WMA:移動加權平均法
TYPPRICE:代表性價格

TA-Lib

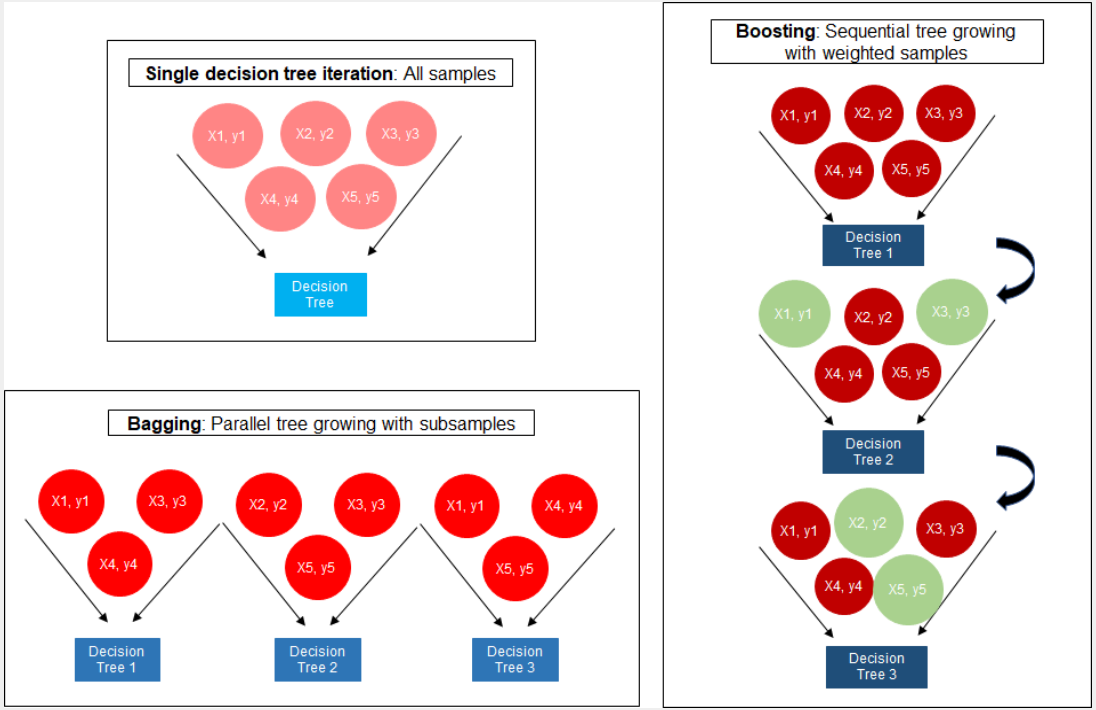

```
(2692, 178)
(2682, 178)
```

Out[281]:

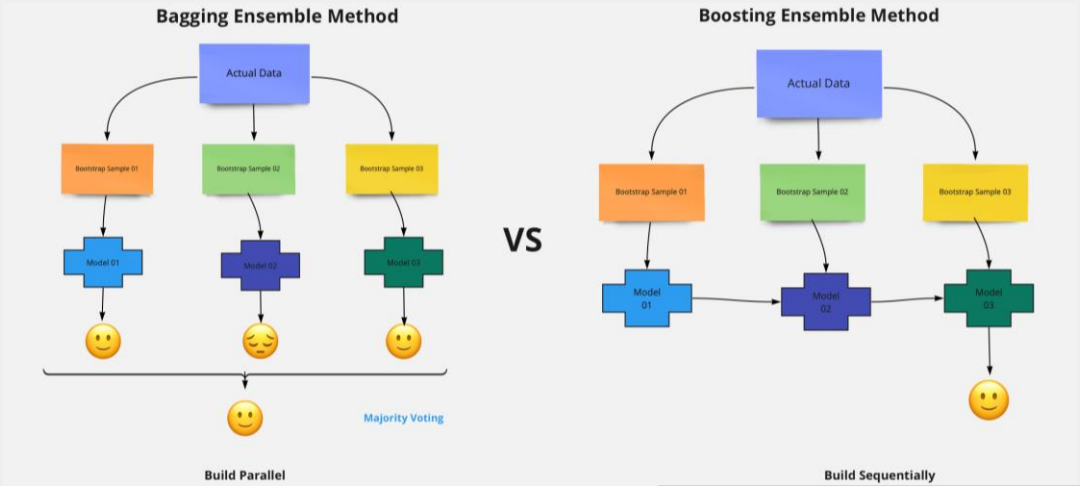
re	sine	...	stddev	tsf	var	atr	natr	trange	ad	adosc	obv	return
00	2682.000000	...	2682.000000	2682.000000	2.682000e+03	2682.000000	2682.000000	2682.000000	2.682000e+03	2682.000000	2.682000e+03	2682.000000
46	0.012239	...	0.186643	25.411893	5.365339e-02	0.324619	1.298594	0.324497	2.163219e+06	6170.262167	1.082137e+04	1.001275
91	0.605760	...	0.137204	3.612736	1.217650e-01	0.098697	0.438140	0.208381	1.472339e+06	19516.785178	6.035243e+05	0.039858
13	-1.000000	...	0.000002	16.599451	3.296918e-12	0.162540	0.658059	0.050000	-1.889946e+05	-137223.749290	-1.038442e+06	0.798283
02	-0.482258	...	0.101980	23.533654	1.040000e-02	0.251908	0.990731	0.200000	8.359461e+05	-5018.631331	-5.657454e+05	0.984382
90	0.001318	...	0.150333	25.161813	2.260000e-02	0.307547	1.171988	0.300000	2.156400e+06	6065.538880	5.278156e+04	1.000000
98	0.525923	...	0.228035	26.457418	5.200000e-02	0.373914	1.496181	0.400000	3.312700e+06	17485.781240	5.009822e+05	1.016319
79	0.999987	...	1.614125	35.567582	2.605400e+00	0.654265	2.916991	3.550000	4.781246e+06	132519.955424	1.897963e+06	1.502890

隨機森林、XGBoost

劉文裕



隨機森林 vs. XGBoost



RF	precision	recall	f1-score	support
False	0.47	0.61	0.53	361
True	0.56	0.41	0.47	433
accuracy			0.48	794
macro avg	0.51	0.51	0.5	794
weighted avg	0.52	0.5	0.5	794

dataaspirant.com

XGB	precision	recall	f1-score	support
False	0.48	0.61	0.53	361
True	0.56	0.41	0.47	433
accuracy			0.5	794
macro avg	0.51	0.51	0.5	794
weighted avg	0.52	0.5	0.5	794

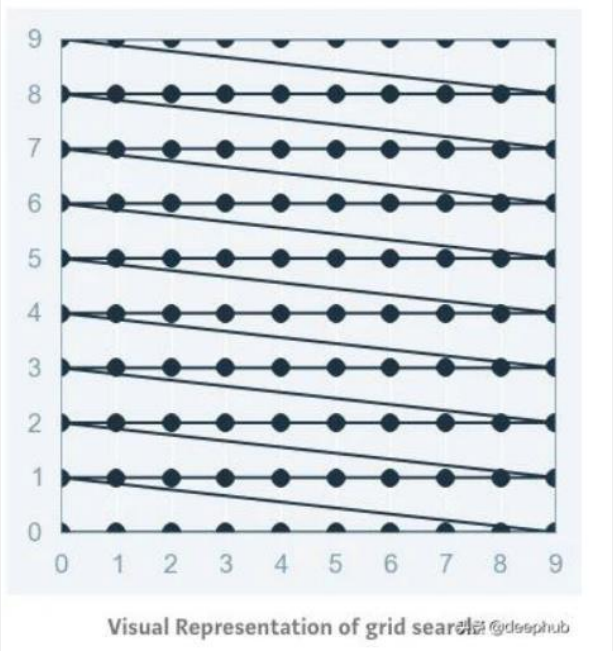
超參數優化問題

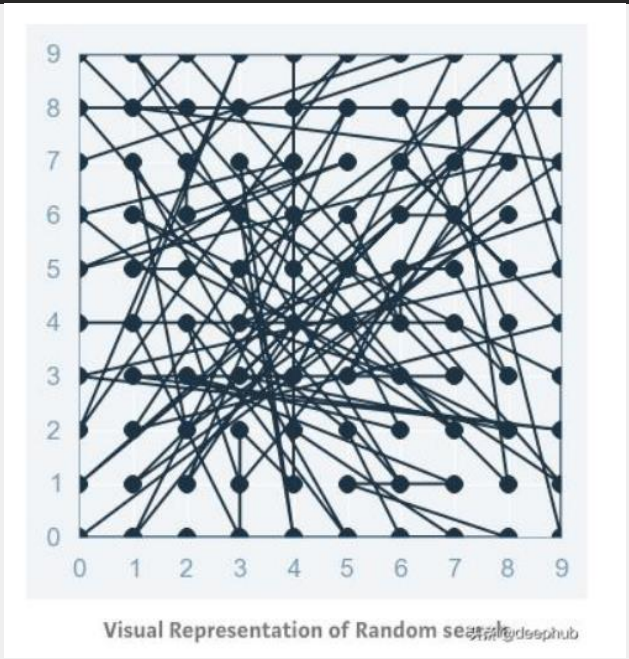
有多種方法可用。 例如 ：

- Grid Search(網格搜索)
- Random Search(隨機搜尋)
- Optuna(奧圖納)
- HyperOpt(超級光電)

網格化尋優（Grid Search） VS 隨機尋優（Random Search）

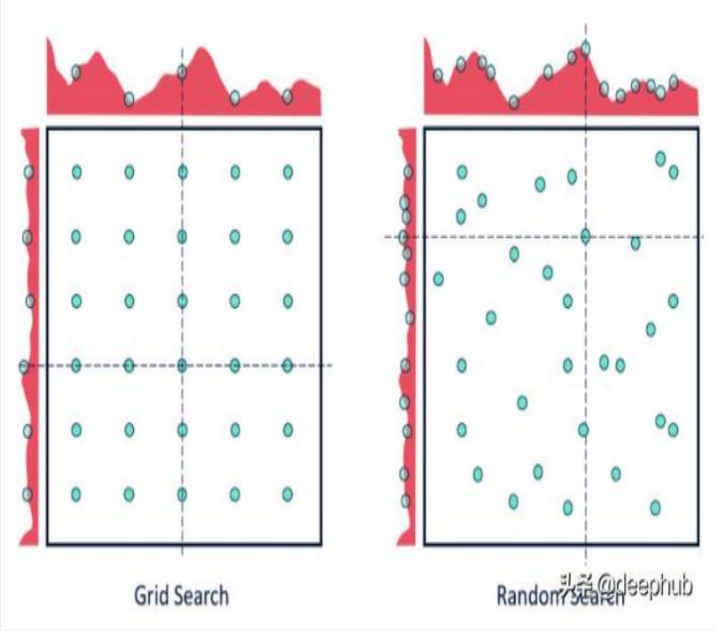
劉文裕





隨機尋優方法(RandomizedSearchCV)

劉文裕

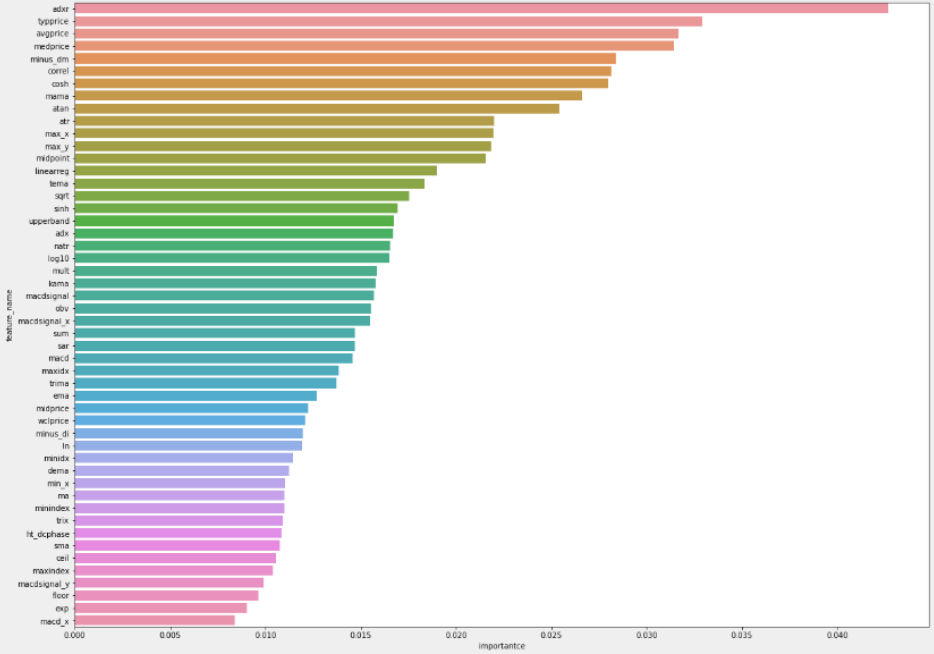


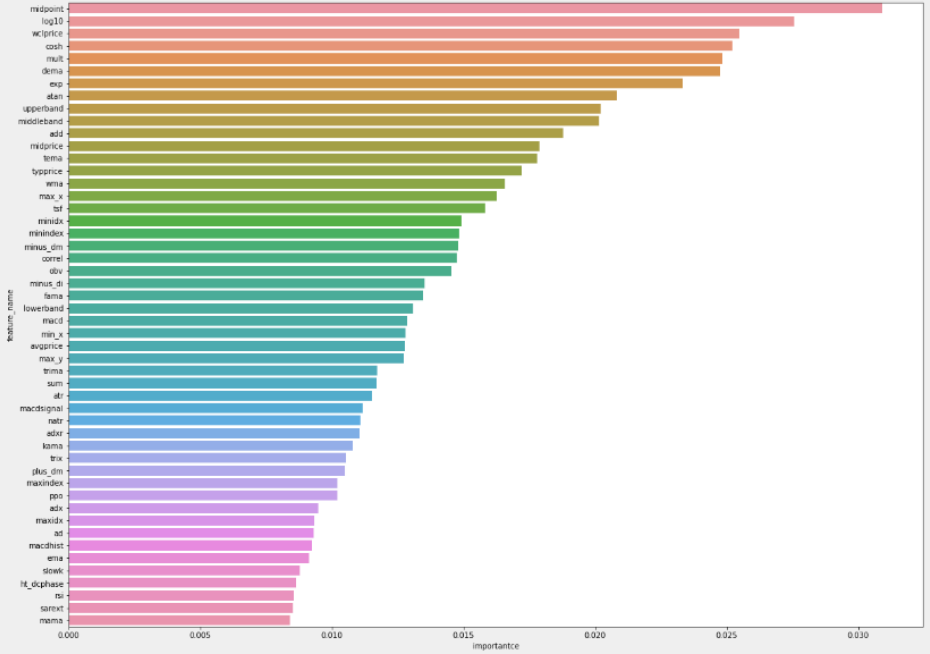
RandomizedSearchCV and estimator

XGB	precision	recall	f1-score	support
False	0.46	0.41	0.43	361
True	0.55	0.59	0.57	33
accuracy			0.51	794
macro avg	0.5	0.5	0.5	794
weighted avg	0.51	0.51	0.51	794
XGB(RCV)	precision	recall	f1-score	support
False	0.47	0.61	0.53	361
True	0.55	1	0.71	433
accuracy			0.55	794
macro avg	0.27	0.5	0.35	794
weighted avg	0.3	0.55	0.38	794

隨機森林和XGB預測出重要的特徵

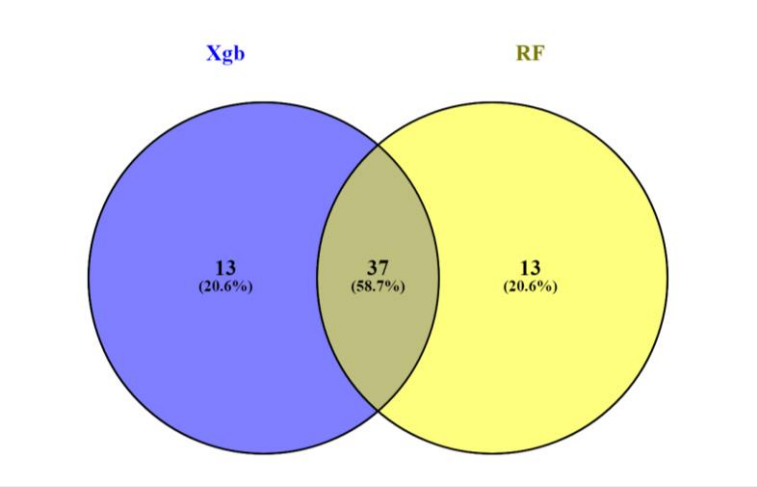
劉文裕



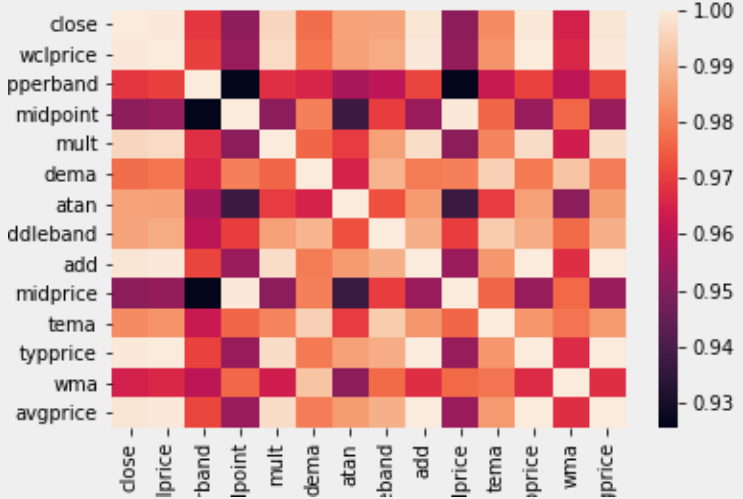


與收盤價相關性最高的特徵

劉文裕



交集



相關性熱圖

建立所有股票日股價與特徵

劉文裕

Out[5]:

stock_id	date	ADXR	AVGPRICE	OBV	MINUS_DM	WMA	TYPPRICE	TEMA	UPPERBBANDS	DEMA	WCLPRICE	return
1101	2010-01-04	NaN	34.1500	8.299290e+03	NaN	NaN	34.200000	NaN	NaN	NaN	34.2250	0.997085
	2010-01-05	NaN	35.1000	5.635866e+04	NaN	NaN	35.266667	NaN	NaN	NaN	35.3250	0.967606
	2010-01-06	NaN	35.8875	9.238540e+04	NaN	NaN	36.016667	NaN	NaN	NaN	36.0750	0.937931
	2010-01-07	NaN	35.9500	7.785438e+04	NaN	NaN	35.816667	NaN	NaN	NaN	35.7500	0.953586
	2010-01-08	NaN	35.4750	6.718324e+04	NaN	NaN	35.450000	NaN	36.653794	NaN	35.4375	0.898305
...	...	...	...	...	...	...	...	...	...	...	...	...
9962	2021-04-19	29.572913	12.5875	8.506548e+08	0.305011	10.093097	12.533333	10.791586	13.124515	10.475719	12.5875	NaN
	2021-04-20	31.326490	12.7000	8.506527e+08	0.283225	10.241678	12.600000	11.107083	13.459858	10.715841	12.5000	NaN
	2021-04-21	32.609524	12.3500	8.506540e+08	0.462994	10.401312	12.400000	11.425286	13.466149	10.963375	12.4125	NaN
	2021-04-22	34.125041	12.6750	8.506491e+08	0.429923	10.545678	12.650000	11.672634	13.017364	11.167456	12.5625	NaN
	2021-04-23	34.599750	12.0625	8.506479e+08	1.099215	10.646001	11.916667	11.775796	12.968767	11.275846	11.8625	NaN

4078037 rows × 11 columns

建模應用

預測比較

主講人:陳姿伶

股票篩選

陳姿伶

從 1733 支股票中
挑出前 54支股票
（0050 成分股）

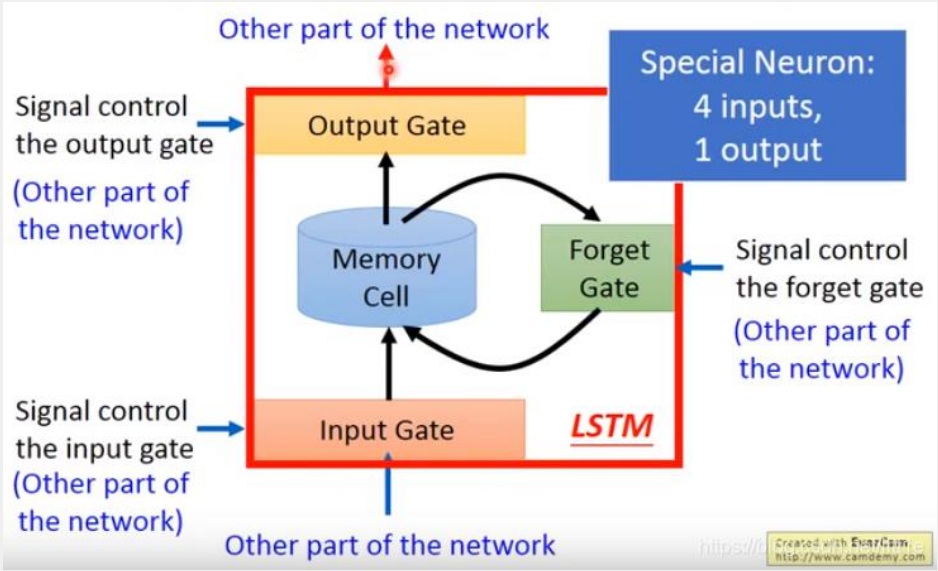
使用LSTM和特徵預測股價

陳姿伶

1	將所有因子計算相關係數	以日交易量因子預測收盤價
2	計算個股 TA-Lib 特徵	將收盤價和 TA-Lib 特徵計算相關係數，取相關係數 >0.9 以上的特徵
3	計算個股 TA-Lib 特徵	將計算出的特徵分別匯入RF和 XGBoost，取出兩者交集的特徵
4	計算 TA-Lib 特徵	將計算出的特徵分別匯入RF 和 XGBoost，取出兩者交集的前十項特徵

使用LSTM預測股價

陳姿伶



LSTM 簡介

Reference: https://reurl.cc/Q9drLO

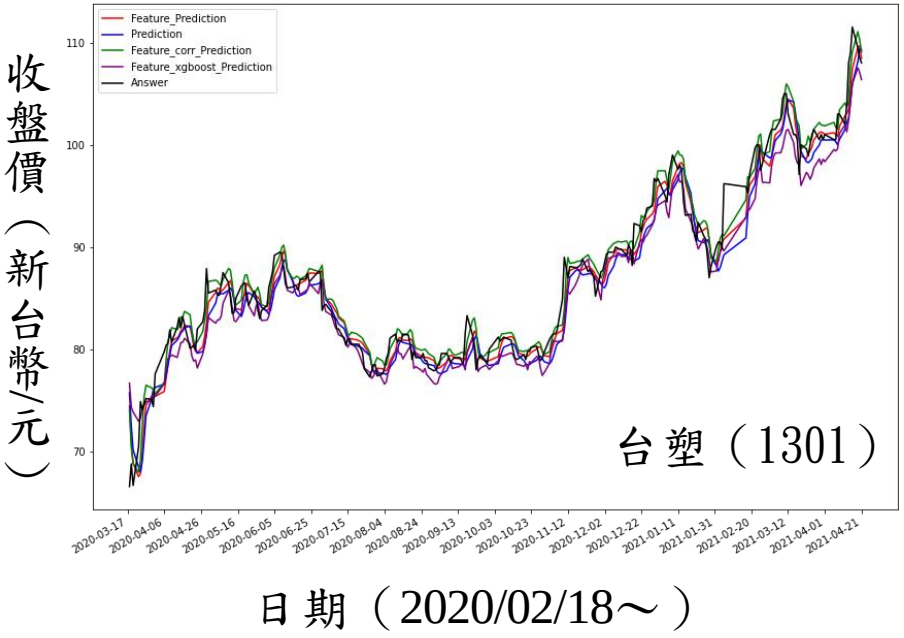
文獻參考：Stock Price Prediction Using Attention-based Multi-Input LSTM, Proceedings of Machine Learning Research 95:454-469, 2018

預定作法：

1. 以 KERAS 做為 LSTM 的實作選擇
2. 兩層 256 個神經元的 LSTM layer
3. 加上了 Dropout 層
4. 兩層有不同數目的全連結層
5. 輸出 1 維數值（收盤價）

使用LSTM和特徵預測股價

陳姿伶



圖片說明

Feature_Prediction: 前十名 TA-Lib 特徵

Prediction: 日交易量因子預測

Feature_corr_Prediction: 與收盤價相關係數 > 0.9 TA-Lib 特徵

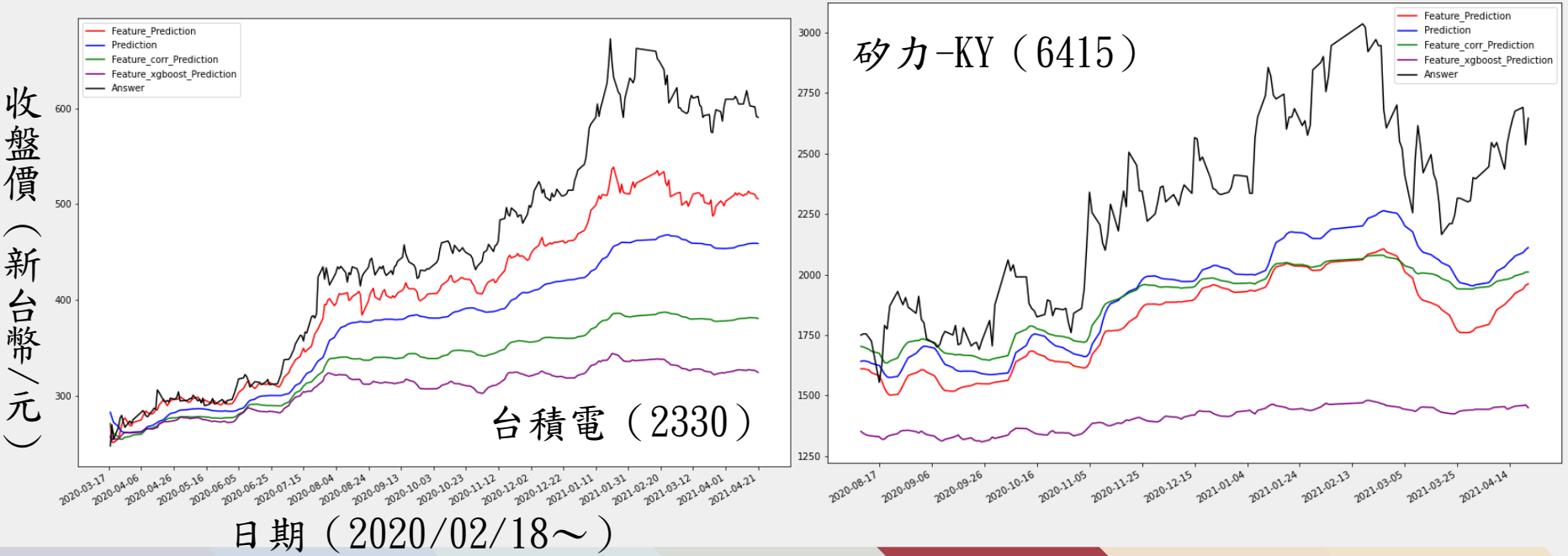
Feature_xgboost_Prediction: TA-Lib 特徵

Answer: 實際收盤價

使用LSTM和特徵預測股價-2

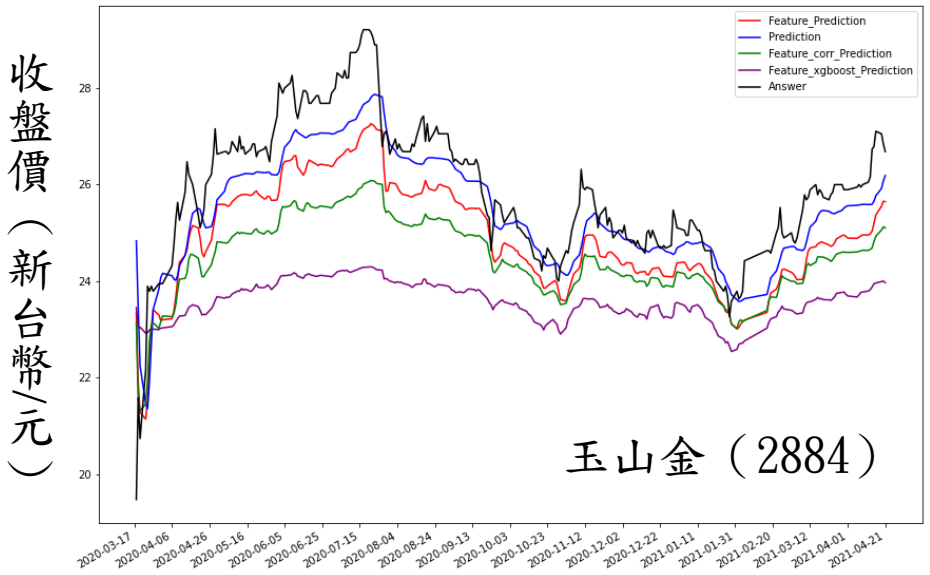
陳姿伶

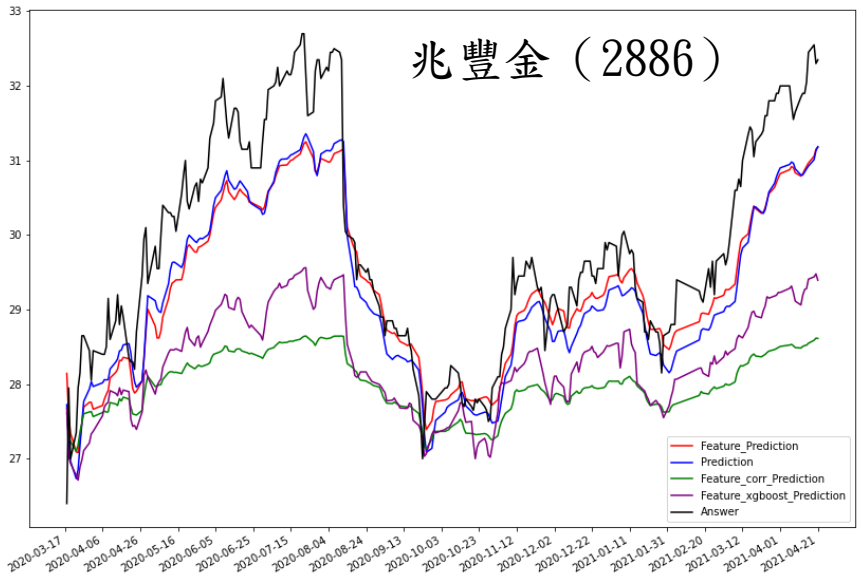
以下這兩支股票，四個模型都無法預測到接近真實值。



使用LSTM和特徵預測股價-3

陳姿伶





RF 和 XGBoost 交集後前十名 TA-Lib 特徵的預測，和用日交易量因子預測，結果和真實的收盤價較為接近。

MSE 一覽表

RF 和 XGBoost 交集後前十名 TA-Lib 特徵預測					
MSE	epoch為50	epoch為100	epoch為150	剔除暫停交易日資料	小計
<10	31	1		1	33
>10	13	5	2	1	21
總計	44	6	2	2	54

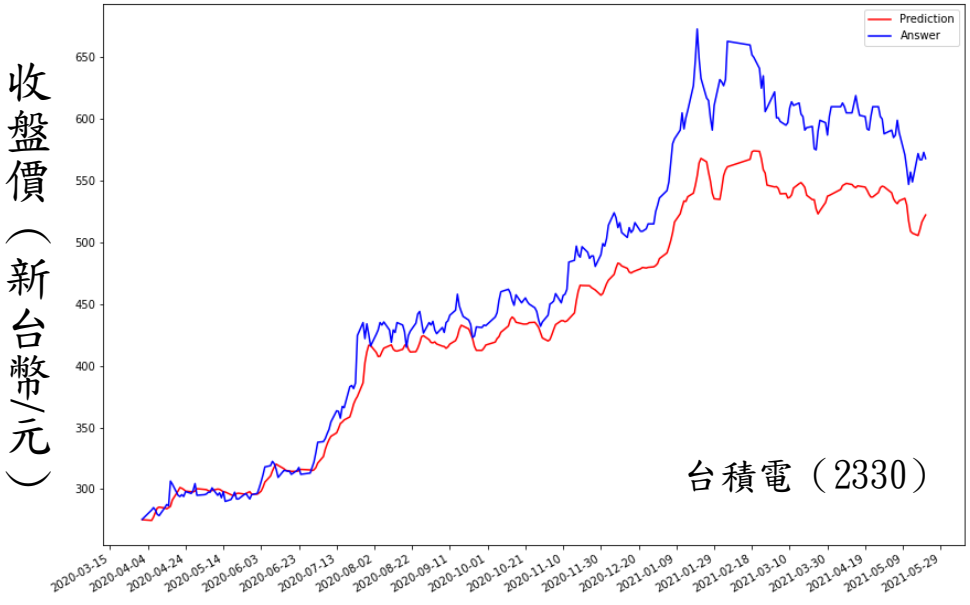
日交易量因子預測					
MSE	epoch為50	epoch為100	epoch為150	剔除暫停交易日資料	小計
<10	30		2	1	33
>10	1	7	12	1	21
總計	31	7	14	2	54

MSE 一覽表-2

RF 和 XGBoost 交集後前十名 TA-Lib 特徵預測 / 日交易量因子預測（結果相同）			
產業別	<10	>10	小計
金融保險業	14		14
航運業 (3), 塑膠工業 (3), 通信網路業 (3)	9		9
油電燃氣業 (1), 鋼鐵工業 (1), 食品工業 (1), 紡織纖維 (1)	4		4
水泥工業	2		2
電腦及週邊設備業	3	3	6
半導體業	1	7	8
光電業 (1), 貿易百貨業 (1), 橡膠工業 (1)		3	3
汽車工業 (2), 電子零組件業 (2), 其他業 (2), 其他電子業 (2)		8	8
總計	33	21	54

驗證是否過擬合

陳姿伶



已知五月台灣疫情升溫，故驗證模型是否可預測到五月股價下跌

結果：模型有預測到股價下跌的趨勢，但無法精準預測數字。

小結

1. 以 RF 和 XGBoost 交集後前十名 TA-Lib 特徵預測和日交易量因子預測，兩個模型對股票預測的趨勢雖接近真實，但無法預測精確數字。

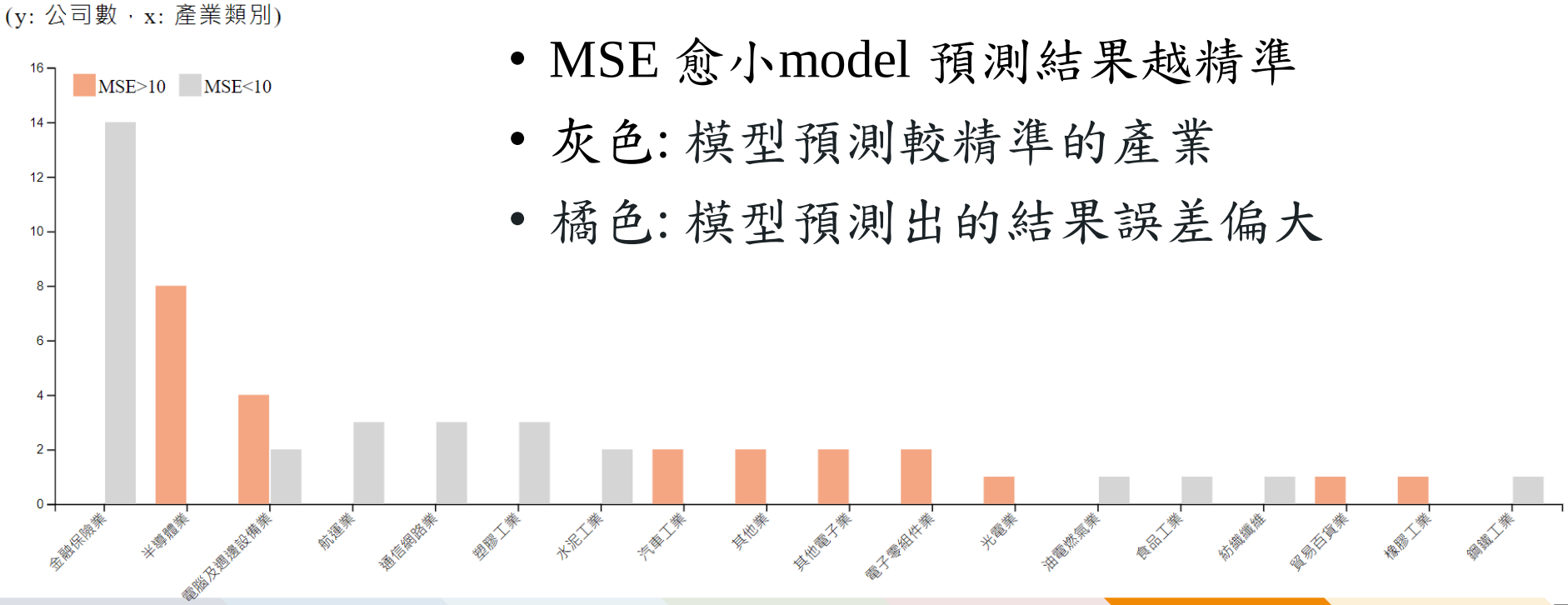
2. 不同的股票有各自適合的預測模型。

資料視覺化

主講人:陳彥伶

MSE與產業的長條圖

陳彥伶

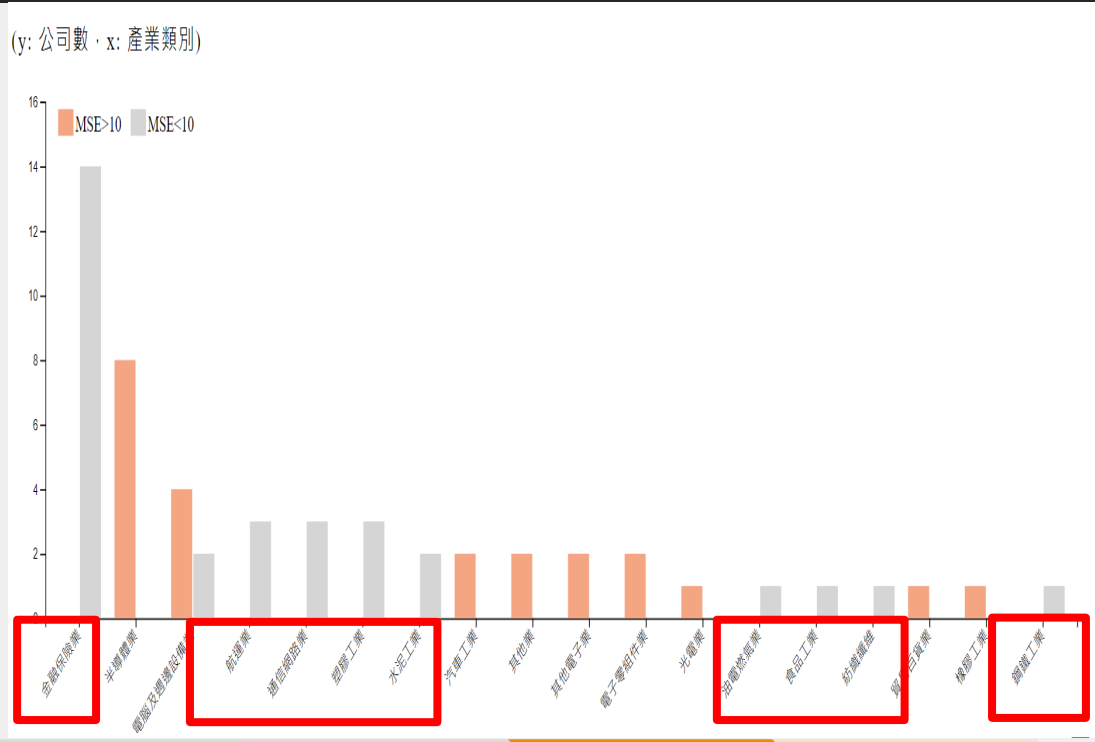


- MSE 愈小model 預測結果越精準
- 灰色: 模型預測較精準的產業
- 橘色: 模型預測出的結果誤差偏大

MSE與產業的長條圖

陳彥伶

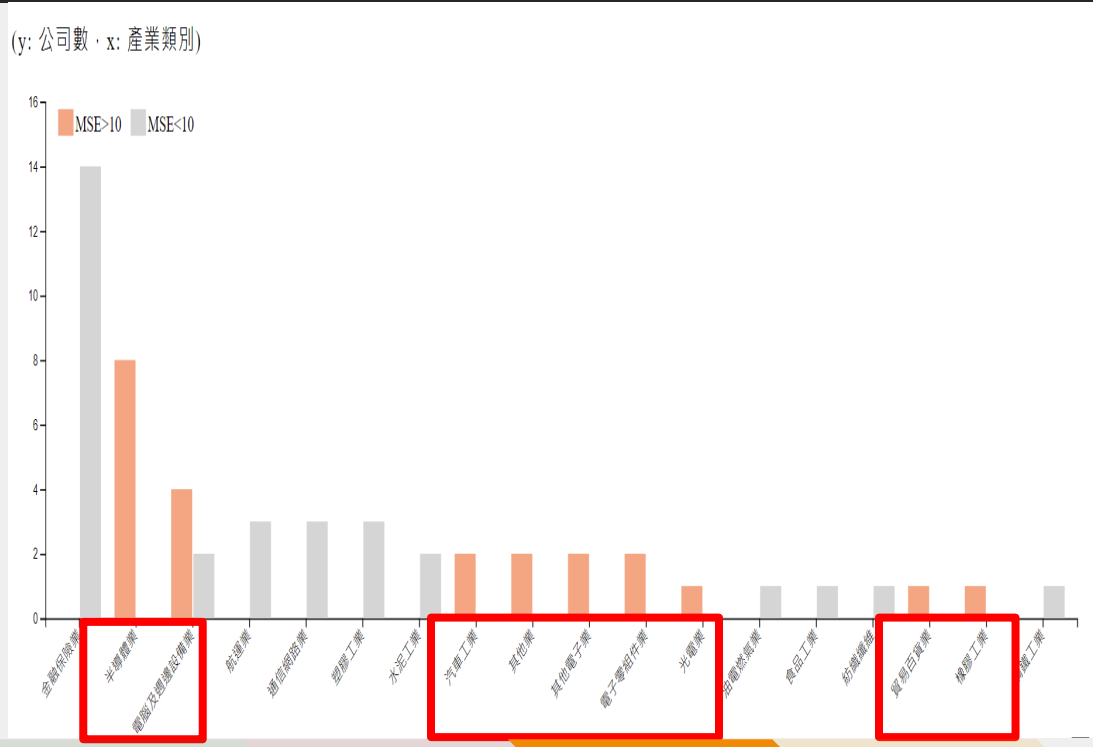
- LSTM 預測較準的產業:
  - 金融保險
  - 通訊



MSE與產業的長條圖

陳彥伶

- LSTM 誤差較大的產業:
  - 半導體
  - 汽車工業
  - 電子



陳彥伶

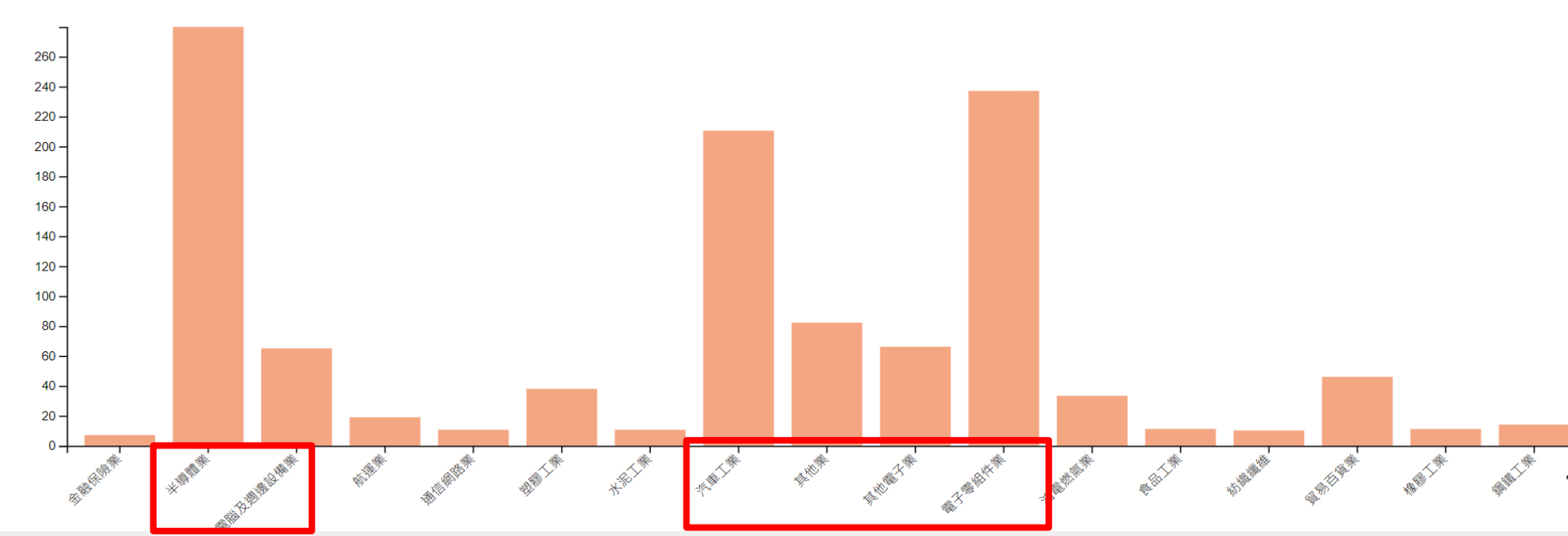
為什麼半導體等產業預測結果誤差會偏大?
原因是什麼?

陳彥伶

股價的變化幅度

股價變化幅度

陳彥伶



註:由於大立光3008(光電業)股價變動幅度大過於其他產業太多，故排除方便分析

造成股價波動的原因與市場需求有關

- 市場需求趨勢可用前面TA-Lib選出的特徵ADXR指數判別

ADXR指標(以鴻海為例)

陳彥伶

25

ADXR>25

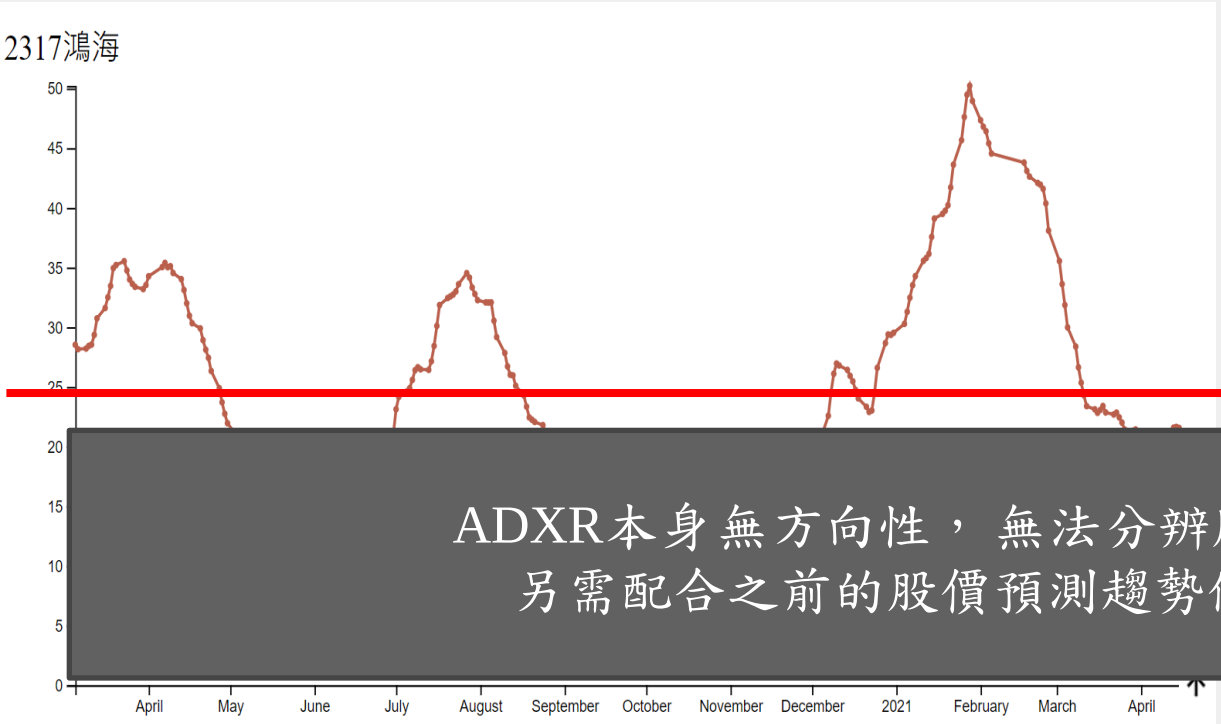
- 股市趨勢動能大
- 股價變化幅度大

ADXR<25

- 股市趨勢動能小
- 盤整中

ADXR指標(以鴻海為例)

陳彥伶



ADXR>25

- 股市趨勢動能大
- 股價變化幅度大

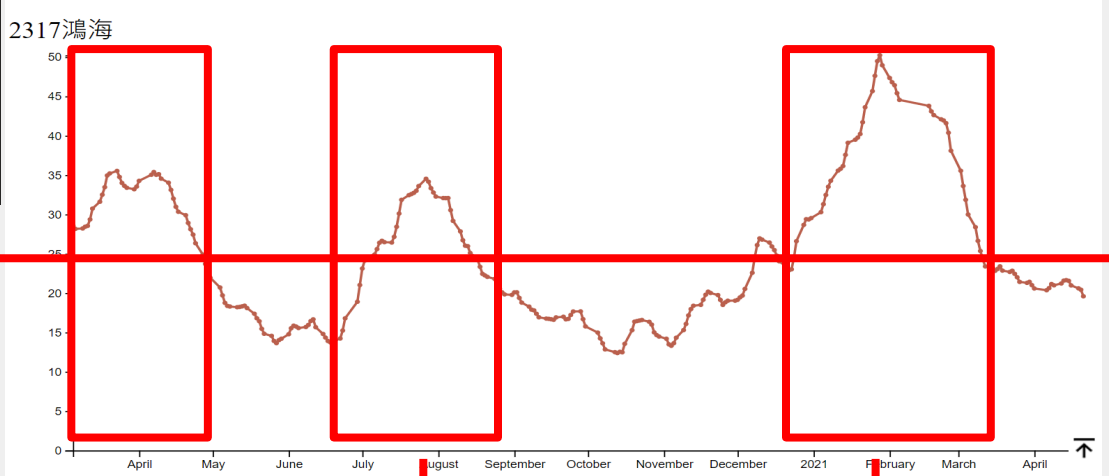
ADXR本身無方向性，無法分辨股價漲跌
另需配合之前的股價預測趨勢做分析

2317鴻海

陳彥伶

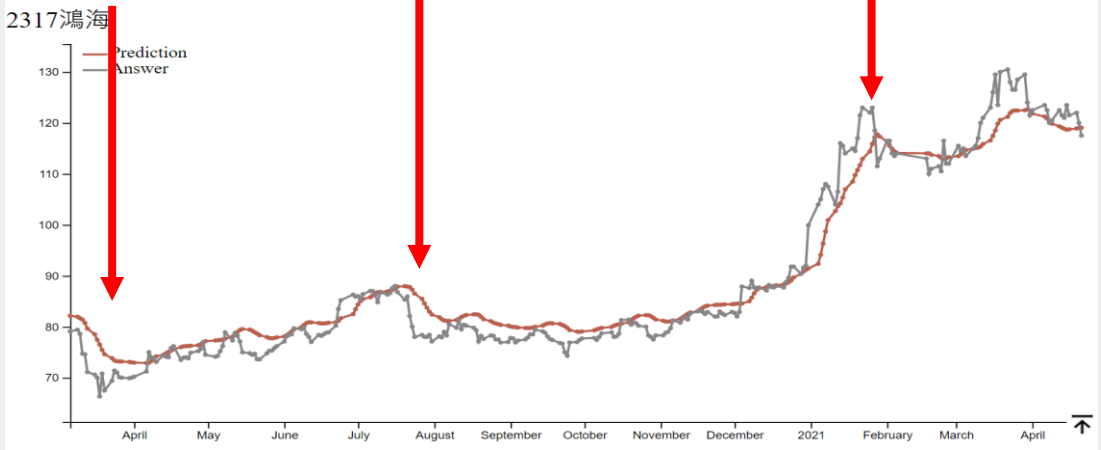
鴻海ADXR>25 有三處

- 左與中間的紅框(下跌)
- 右邊的紅框(上漲)



25

市場趨勢波動

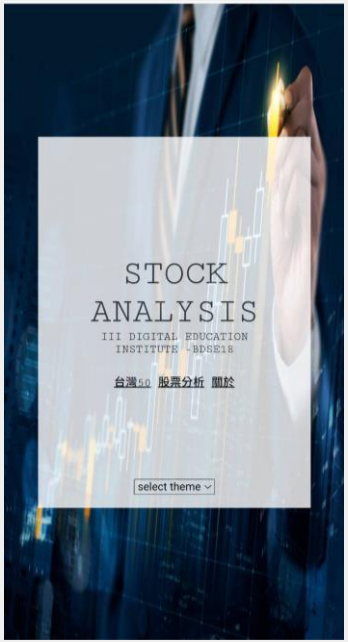


股價預測

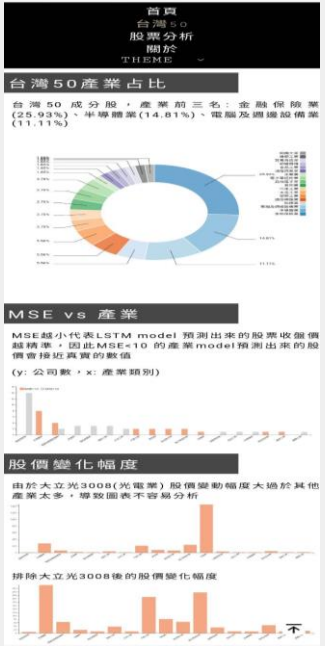
網頁

陳彥伶

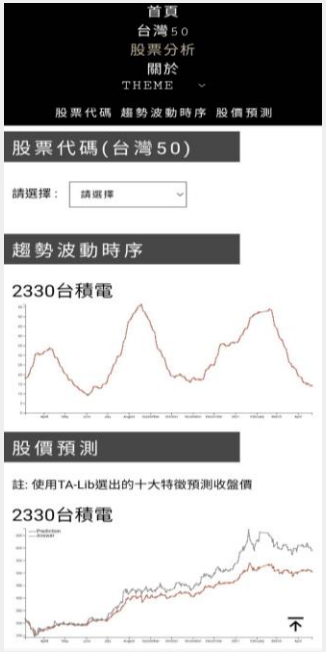
首頁



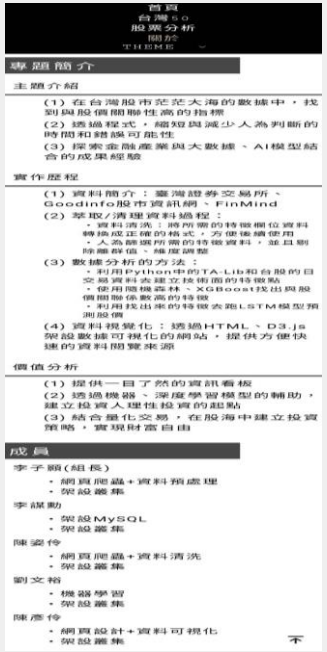
台灣50



股票分析



關於



網頁QR code

台灣50: 股價變化幅度、MSE vs 產業圖

股票分析: 市場趨勢、股價預測圖

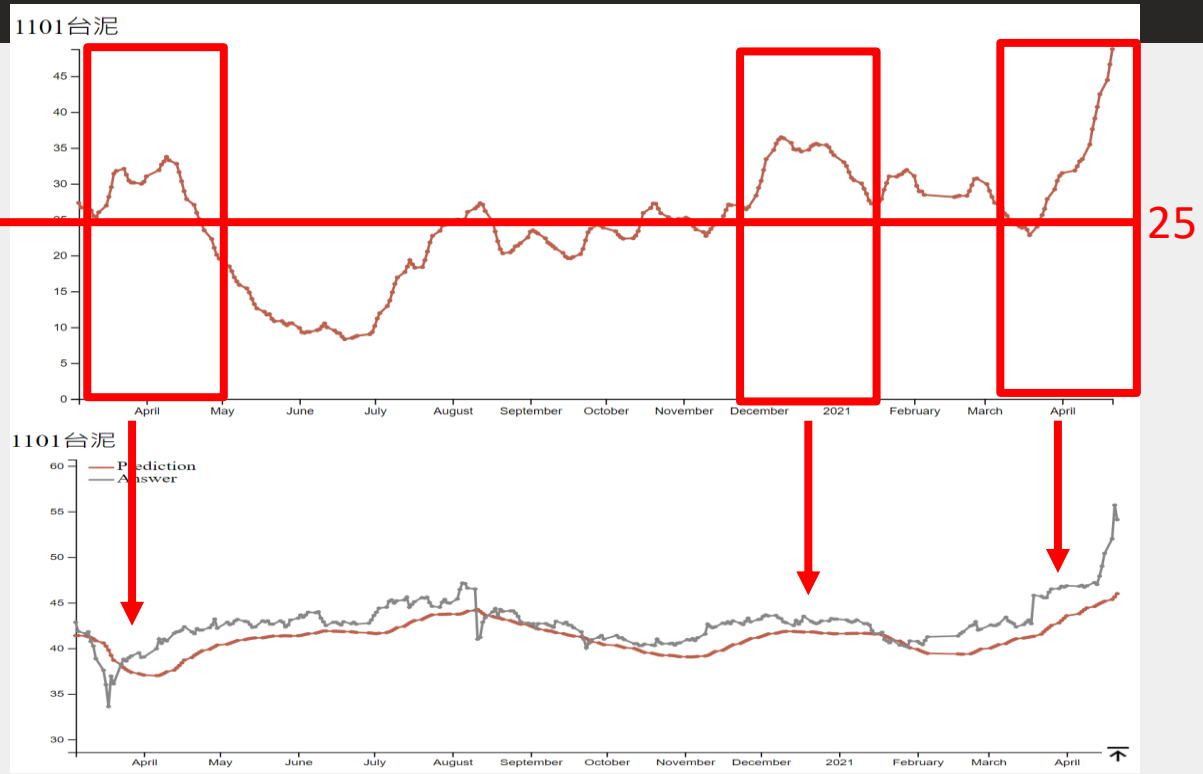
小結

是否可單看ADXR 與股價預測來判斷進場或出場?

小結

台泥ADXR>25 有三處

- 但台泥股價本身就偏穩定，所以就算市場有一定的趨勢，股價也不見得有所起伏



小結

判別是否進出股市，還是需參考其他指標一起分析

網頁QR code

陳彥伶

結論與展望

主講人:李子顥

結論

Data

- 每日交易量
- 法人買賣
- 個股融資融卷
- 當日沖銷交易標的及成交量值
- 個股股利、PER、PBR資料表
- 股東結構表

結論

Machine Learning

- Random Forest, XGBoost 參數最佳化
- 篩選出重要的特徵

Deep Learning

- LSTM 預測未來股價
- 四種不同特徵模型比較

結論

Visualization

- 個股
  - 重要特徵
  - 股價預測
- 台灣50
  - 產業資訊
  - 模型預測、MSE

專題目的

- 程式處理大量的金融資料
- 建立理性投資觀念的基礎

LSTM 想法

- 不同產業類別、特性股票的影響
- 人為因素 ex. 特徵篩選、參數調整
- 擬合問題:過擬合、欠擬合

展望

李子顥

- 探索更多種類的股票指標，ex. 籌碼面的延伸、基本面、期貨
- 根據不同產業類別、不同個股，建立各自的參數挑選、股價預測模型
- 嘗試不同的機器學習、深度學習模型。找出最佳模型和參數
- 以預測模型建立股市交易策略

BDSE18 專題發表

股 市 大 數 據

李子顥 李謀勳 陳姿伶 劉文裕 陳彥伶