資料科學與人工智慧競技

Final Project Presentation 1

JPX Tokyo Stock Exchange Prediction

組員 : 製造所 P96101148 巫清賢
製造所 P96104112 蘇冠瑜

比賽簡介

目標：預測2000支JPX提供股票未來的Change Rate(Target)，以建立投資組合。

Change Rate(Target)：利用明天及後天的Close值計算，公式如右：

$$r_{(k,t)} = \frac{C_{(k,t+2)} - C_{(k,t+1)}}{C_{(k,t+1)}}$$

Target越大表示有利的投資(投資100$，得到120$)，越小則表示不利的投資(投資100$，得到80$)。

評分計算方法為計算Top / Bottom 200的排名分數，公式如右：

$$S_{up} = \frac{\sum_{i=1}^{200}(r_{(up_i,t)} * linearfunction(2,1)_i))}{Average(linearfunction(2,1))}$$

簡單而言，就是利用現有資料，找出未來一天的2000支股票的Target值並排序即可完成預測。

資料簡介 - 輸入

stock_prices.csv：主要訓練資料，紀錄2000種最常交易的股票，包含日期、股票編號、開高收低價、交易量、Target等資訊。

secondary_stock_prices.csv：包含較為冷門的股票數據，紀錄的資料與stock_price.csv一樣。

option.csv：紀錄各種股票細節狀態的數據，包含整天、白天、夜間的開高收低價、結算價、理論價、基本波動率、利率等等。

trades.csv：上一個營業周的交易量匯總摘要。包含自營交易、證券商交易、個人交易、外匯等的售出額、購買額、總額(購買額+售出額)、差額(購買額-售出額)。

資料簡介 - 輸入

共2000支股票資料。

時間2017-01-04～2021-12-03。

	RowId	Date	SecuritiesCode	Open	High	Low	Close	Volume	AdjustmentFactor	ExpectedDividend	SupervisionFlag	Target
0	20170104_1301	2017-01-04	1301	2734.0	2755.0	2730.0	2742.0	31400	1.0	NaN	False	0.000730
1	20170104_1332	2017-01-04	1332	568.0	576.0	563.0	571.0	2798500	1.0	NaN	False	0.012324
2	20170104_1333	2017-01-04	1333	3150.0	3210.0	3140.0	3210.0	270800	1.0	NaN	False	0.006154
3	20170104_1376	2017-01-04	1376	1510.0	1550.0	1510.0	1550.0	11300	1.0	NaN	False	0.011053
4	20170104_1377	2017-01-04	1377	3270.0	3350.0	3270.0	3330.0	150800	1.0	NaN	False	0.003026
...	...	...	...	...	...	...	...	...	...	...	...	...
2332526	20211203_9990	2021-12-03	9990	514.0	528.0	513.0	528.0	44200	1.0	NaN	False	0.034816
2332527	20211203_9991	2021-12-03	9991	782.0	794.0	782.0	794.0	35900	1.0	NaN	False	0.025478
2332528	20211203_9993	2021-12-03	9993	1690.0	1690.0	1645.0	1645.0	7200	1.0	NaN	False	-0.004302
2332529	20211203_9994	2021-12-03	9994	2388.0	2396.0	2380.0	2389.0	6500	1.0	NaN	False	0.009098
2332530	20211203_9997	2021-12-03	9997	690.0	711.0	686.0	696.0	381100	1.0	NaN	False	0.018414

2332531 rows × 12 columns

資料簡介 - 輸出

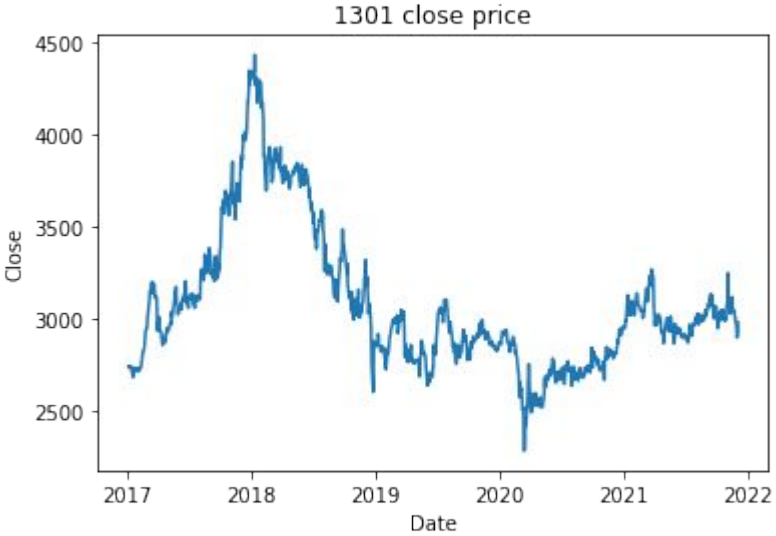
共3個欄位

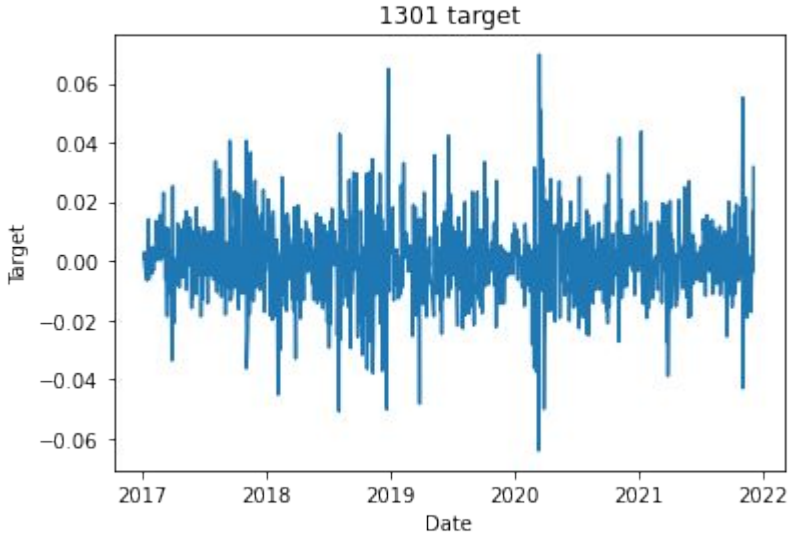
- Date：時間，年-月-日
- SecuritiesCode：股票編號
- Rank：排名，數字越小，排名越高。

	Date	SecuritiesCode	Rank
0	2021-12-06	1301	0
1	2021-12-06	1332	1
2	2021-12-06	1333	2
3	2021-12-06	1375	3
4	2021-12-06	1376	4
...	...	...	...
1995	2021-12-06	9990	1995
1996	2021-12-06	9991	1996
1997	2021-12-06	9993	1997
1998	2021-12-06	9994	1998
1999	2021-12-06	9997	1999

2000 rows × 3 columns

資料觀察


1301 close price
Close
Date
4500
4000
3500
3000
2500
2017
2018
2019
2020
2021
2022


1301 target
Target
Date
0.06
0.04
0.02
0.00
−0.02
−0.04
−0.06
2017
2018
2019
2020
2021
2022

資料觀察

	Date	SecuritiesCode	Open	High	Low	Close	Volume	Target
0	2017-01-04	1301	2734.0	2755.0	2730.0	2742.0	31400	0.000730
1	2017-01-04	1332	568.0	576.0	563.0	571.0	2798500	0.012324
2	2017-01-04	1333	3150.0	3210.0	3140.0	3210.0	270800	0.006154
3	2017-01-04	1376	1510.0	1550.0	1510.0	1550.0	11300	0.011053
4	2017-01-04	1377	3270.0	3350.0	3270.0	3330.0	150800	0.003026
...	...	...	...	...	...	...	...	...
2332526	2021-12-03	9990	514.0	528.0	513.0	528.0	44200	0.034816
2332527	2021-12-03	9991	782.0	794.0	782.0	794.0	35900	0.025478
2332528	2021-12-03	9993	1690.0	1690.0	1645.0	1645.0	7200	-0.004302
2332529	2021-12-03	9994	2388.0	2396.0	2380.0	2389.0	6500	0.009098
2332530	2021-12-03	9997	690.0	711.0	686.0	696.0	381100	0.018414

2332531 rows × 8 columns

```
RowId                0
Date                 0
SecuritiesCode       0
Open              7608
High              7608
Low               7608
Close             7608
Volume               0
AdjustmentFactor     0
ExpectedDividend  2313666
SupervisionFlag      0
Target             238
dtype: int64
```

將資料依照SecurtiesCode切成2000個檔案，再對每個檔案中的空值使用平均數(mean)補值。

最後合併2000個檔案依照Date/照SecurtiesCode排序還原成原始資料。

預計使用方法

最直接的做法是訓練2000支LSTM或時間序列模型預測各股票未來3天的Close值再執行使用公式計算Target值，但實際上不太可能這樣做..

根據Discussion平台之討論，我們預計使用Light-GBM進行訓練。

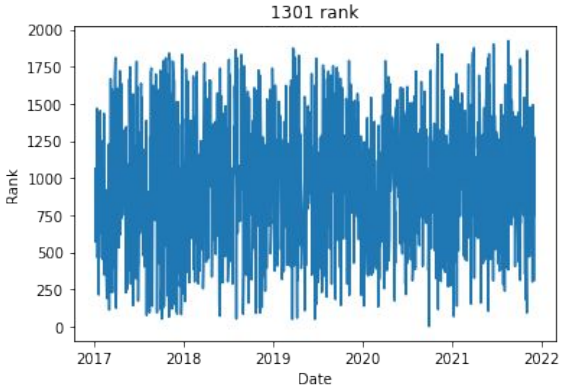
針對stock_prices.csv中的不同Feature，嘗試最佳的Feature組合。

Baseline ("Open", "High", "Low", "Close", "Volume") → ("Target")

直接預測Rank ("Open", "High", "Low", "Close", "Volume") → ("Rank")

將原始Feature經過換算/抽取 ("Close", "???"...) → ("Target")

加入產業類別(stock_list.csv/33SectorName)特徵(0~32)。



預計使用方法

	RowId	Date	SecuritiesCode	Open	High	Low	Close	Volume	AdjustmentFactor	ExpectedDividend	SupervisionFlag	Target	Rank
0	20170104_1301	2017-01-04	1301	2734.0	2755.0	2730.0	2742.0	31400	1.0	NaN	False	0.000730	898.0
1	20170104_1332	2017-01-04	1332	568.0	576.0	563.0	571.0	2798500	1.0	NaN	False	0.012324	325.0
2	20170104_1333	2017-01-04	1333	3150.0	3210.0	3140.0	3210.0	270800	1.0	NaN	False	0.006154	567.0
3	20170104_1376	2017-01-04	1376	1510.0	1550.0	1510.0	1550.0	11300	1.0	NaN	False	0.011053	364.0
4	20170104_1377	2017-01-04	1377	3270.0	3350.0	3270.0	3330.0	150800	1.0	NaN	False	0.003026	750.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
2332526	20211203_9990	2021-12-03	9990	514.0	528.0	513.0	528.0	44200	1.0	NaN	False	0.034816	580.0
2332527	20211203_9991	2021-12-03	9991	782.0	794.0	782.0	794.0	35900	1.0	NaN	False	0.025478	1119.0
											False	-0.004302	1941.0
											False	0.009098	1768.0
											False	0.018414	1472.0

Sector