股價漲跌模型預測

大數據與商業分析期中專案

第8組：王喬、林姝延、李佑婷、陳青妤、張芳瑜、葉家妤

目錄

研究摘要

專案主題

抓取文章關鍵字並以模型進行股價漲跌判斷，決定進出場

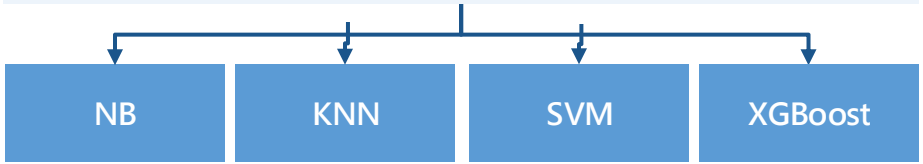
研究主軸

Step 1
資料前處理

- 標記文章漲跌
- 取出漲跌 features
- Mapping features

Step 2
模型訓練、參數調整及漲跌預測

- 取出漲、跌文章
- 80%訓練資料、20%測試資料
- 模型訓練及預測
- 以測試資料評估模型準確率



Step 3
移動回測

- 以前三個月資料建構向量量空間
- 設定進出場閾值
- 進行模型訓練並進行每日股價漲跌預測
-計算準確率

研究產出

- 各模型的最佳模型參數及準確率
- 各模型移動回測混淆矩陣、準確率及出手率
- 各模型成效比較

01

資料前處理

Our goals

Label 文章的漲跌作為 Ground Truth

針對 title 及 content，使用關鍵字挑選出目標股票(元大台灣50)的新聞及論壇文章。漲跌幅經討論及修改後設定為 ± 0.4

取出漲跌的 Features

使用monpa套件將訓練文章斷詞，再使用 chi-square 降低向量稀疏性

Mapping Features 與文章

使用 CountVectorizer 將文章 mapping 到由漲跌 features 組成的向量空間

What we have tried?

取出的文章向量空間過於稀疏

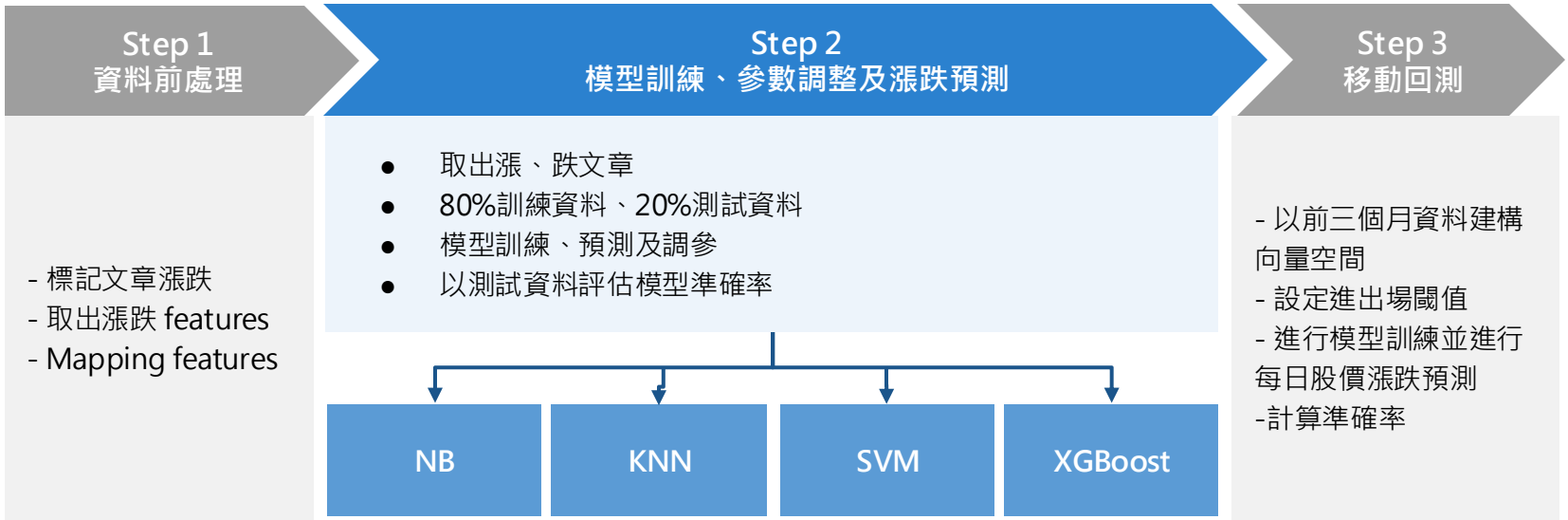
研究後發現：
countVectorize 會先用內建的函式斷詞後再 mapping，與 monpa 斷詞方式不同，故無法比對

改良：
將所有文章都先經過 monpa 斷詞後再丟進 countVectorize 做 mapping

02

模型訓練及漲跌預測

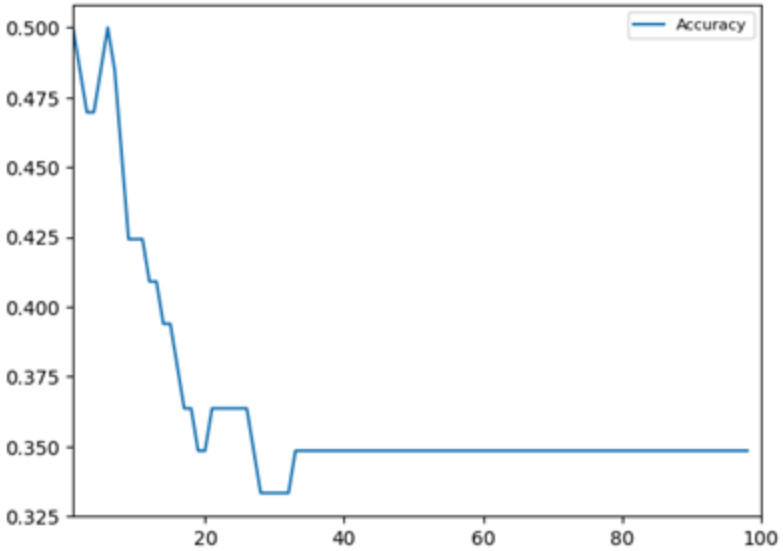
02- 模型訓練及漲跌預測



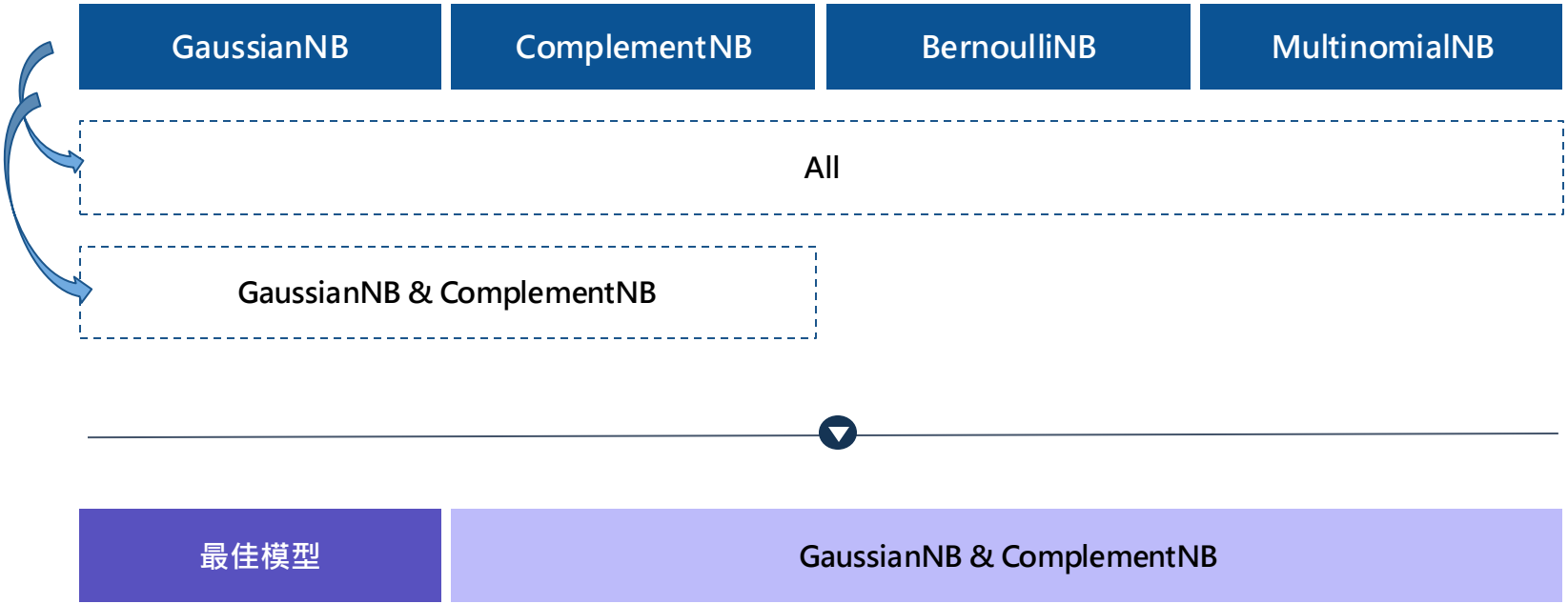
資料前處理 模型訓練參數調整及漲跌預測 移動回測

模型訓練

1. 驗證模型
   - K — folds 交叉驗證
2. 調整超參數
   - 多層 for 迴圈
   - GridsearchCV



參數調整-Naive Bayes



資料前處理 → 模型訓練參數調整及漲跌預測 → 移動回測

參數調整-KNN

N - neighbors	Weights	Leaf Size	p	Resampling Strategies
5	uniform	3	1	normal
7		5	2	under
9		10		over
11		20		
13		30		

最佳模型	
N-neighbors	9
Weights	uniform
Leaf Size	20
p	1
Resampling Strategies	under

參數調整-SVM

kernel	linear	poly	rbf	sigmoid	→ rbf
C	1, 2, ...,19, 20	→ 5			
degree	1, 5, 10, ...,95, 100 & 100, 150, ..., 800, 850	→ 影響不明顯			
decision_function_shape	ovo, ovr	→ 影響不明顯			

最佳模型	kernel	C	degree	decision_function_shape
	rbf	5	3	ovr

參數調整-XGBoost

max_depth	min_child_weight	gamma	learning_rate	reg_alpha
3	1	0	0.1	1e-5
5	3	0.1	0.2	1e-2
7	5		0.3	0.1
9			0.4	1
			0.5	100

最佳模型	
max_depth	7
min_child_weight	1
gamma	0.1
learning_rate	0.4
reg_alpha	1e-5

各模型分類結果

	NB			KNN			SVM			XGBoost		
混淆矩陣		預測漲	預測跌		預測漲	預測跌		預測漲	預測跌		預測漲	預測跌
	實際漲	164	390	實際漲	181	131	實際漲	232	354	實際漲	484	125
	實際跌	8	657	實際跌	446	461	實際跌	165	468	實際跌	432	178
準確率	67%			53%			57%			54%		

資料前處理 → 模型訓練參數調整及漲跌預測 → 移動回測

03
移動回測

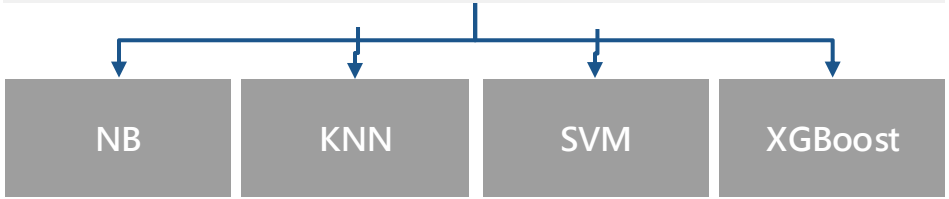
03 - 移動回測

Step 1
資料前處理

- 標記文章漲跌
- 取出漲跌 features
- Mapping features

Step 2
模型訓練、參數調整及漲跌預測

取出漲、跌文章
80%訓練資料、20%測試資料
模型訓練、預測及調參
以測試資料評估模型準確率



Step 3
移動回測

- 以前三個月資料建構向量空間
- 設定進出場閾值
- 進行模型訓練並進行每日股價漲跌預測
-計算準確率

移動回測

- 每次取 3 個月的資料建立文章特徵向量空間，將此 3 個月和第 3+1 個月的文章內容以此轉為向量表示。
- 用前 3 個月的資料訓練模型，來預測第 3+1 個月的漲跌，若漲和跌篇數過於接近則不出手。
- 往後移動 1 個月，重複上述步驟。
- 統計總出手率、準確度。

各模型結果

	NB			KNN			SVM			XGBoost		
混淆矩陣		預測漲	預測跌		預測漲	預測跌		預測漲	預測跌		預測漲	預測跌
	實際漲	73	61	實際漲	20	12	實際漲	27	16	實際漲	49	19
	實際跌	41	71	實際跌	22	30	實際跌	17	28	實際跌	32	34
準確率	59%			60%			62.5%			62%		
出手率	89%			30%			32%			64%		
閾值	0.01			0.8			0.7			0.2		

資料前處理 → 模型訓練參數調整及漲跌預測 → 移動回測

各模型結果

	NB			KNN			SVM			XGBoost		
混淆矩陣		預測漲	預測跌		預測漲	預測跌		預測漲	預測跌		預測漲	預測跌
	實際漲	73	25%	實際漲	20	14%	實際漲	27	18%	實際漲	49	14%
	實際跌	17%	71	實際跌	26%	30	實際跌	19%	28	實際跌	24%	34
準確率	59%			60%			62.5%			62%		
出手率	89%			30%			32%			64%		
閾值	0.01			0.8			0.7			0.2		

資料前處理 → 模型訓練參數調整及漲跌預測 → 移動回測

Demo

Thanks !

影片連結：
https://drive.google.com/file/d/1C4l6czfnI1fE1MOluEKE5jlik sX5zadD/view?usp=sharing