

Online News Popularity

以網路文章分享數 預測潛在行動者數量

陳幸君 Chen, Hsing-Chun
國立清華大學 服務科學所
專長：資料探勘應用於商業分析
2018.11

原始資料簡述

- Online News Popularity Data Set
- 網路媒體 Mashable 文章資料
- 資料筆數: 39797; 變數數量: 61; 目標變數 : 文章分享量
- A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News
- 原始分析題目:
網路新聞是否會受歡迎(預測分類)
提供文章優化方針(局部搜尋)



Kelwin Fernandes, Pedro Vinagre, and Paulo Cortez

INESC TEC Porto/Universidade do Porto, Portugal

ALGORITMI Research Centre, Universidade do Minho, Portugal

'CRISIS IN OUR CLASSROOMS'

What kids need to be taught about mental health in schools

BUG BOUNTIES

The hackers getting paid to keep the internet safe

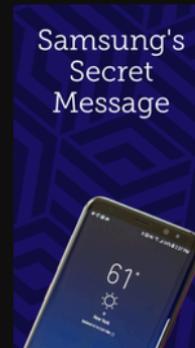
MEET MINKA

How a Tamagotchi made me realise how I really feel about technology

MAGICAL

J.K. Rowling adores cute babies in 'Harry Potter' Halloween costumes

Mashable REELS



Samsung's
Secret
Message



This iPhone
XS Costs
More Than
\$7,000



The History
of Warcraft



We met the mind
behind the Rubik's Cube



PlayStation Classic and
the return of the retro
gaming consoles



These super-realistic
robots are creepy as hell

Watch

What's New

What's Rising

What's Hot

變數說明

字數

超連結

多媒體

發佈日

關鍵字

文字相關

分享數

變數說明

文章字數
標題字數
平均字長
不重複單詞
...等

超連結數量
網超內連結數量
超連結分享數
...等

圖片數
影片數
頻道

週末
星期幾

關鍵字數
Top5關鍵字之
分享數

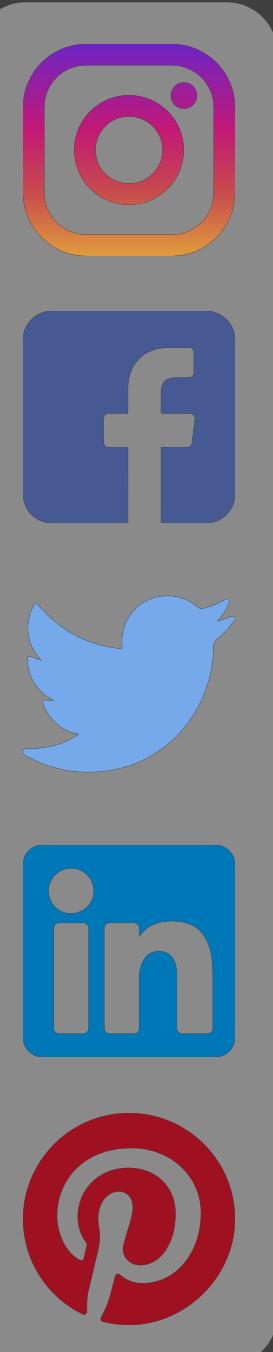
語意分析
主客觀分析
Top5相關文章

分享數

網路文章生態圖



商業題目：提升廣告訂價能力



商業題目：？



商業題目：預知議題潛在行動者的數量

行動者

觸及者

閱覽者

分享者

行動者

- 購買相關商品
- 參加活動/講座 ...
- 鎖定潛在行動者範圍
- 提升廣告、合作的提案效率及訂價能力

未來：

- 分享者群體樣貌
- 市場定位
- 社群影響者
- 時間及管道

商業題目：預知議題潛在行動者的數量

分析題目：預測網路新聞分享量

需求探索

- 資料初探
- 生態圖
- 應用發想
- 定義題目

資料前處理

- 統計數值
- 視覺化
- 檢視清理
- 標準化

預測模型

- 模型選擇
- 變數篩選
- 各群分享數預測結果

效益評估

- 洽談合作
數量
- 合作案
收入成長比

資料前處理

資料筆數：39793 -> 38463

- 統計數值
 - 不正常零值
 - 極端值
- 視覺化

	timedelta	n_tokens_title	n_tokens_content	n_unique_tokens	n_non_stop_words
count	39644.000000	39644.000000	39644.000000	39644.000000	39644.000000
mean	354.530471	10.398749	546.514731	0.548216	0.996469
std	214.163767	2.114037	471.107508	3.520708	5.231231
min	8.000000	2.000000	0.000000	0.000000	0.000000
25%	164.000000	9.000000	246.000000	0.470870	1.000000
50%	339.000000	10.000000	409.000000	0.539226	1.000000
75%	542.000000	12.000000	716.000000	0.608696	1.000000
max	731.000000	23.000000	8474.000000	701.000000	1042.000000

資料前處理

資料筆數：39793 -> 38463

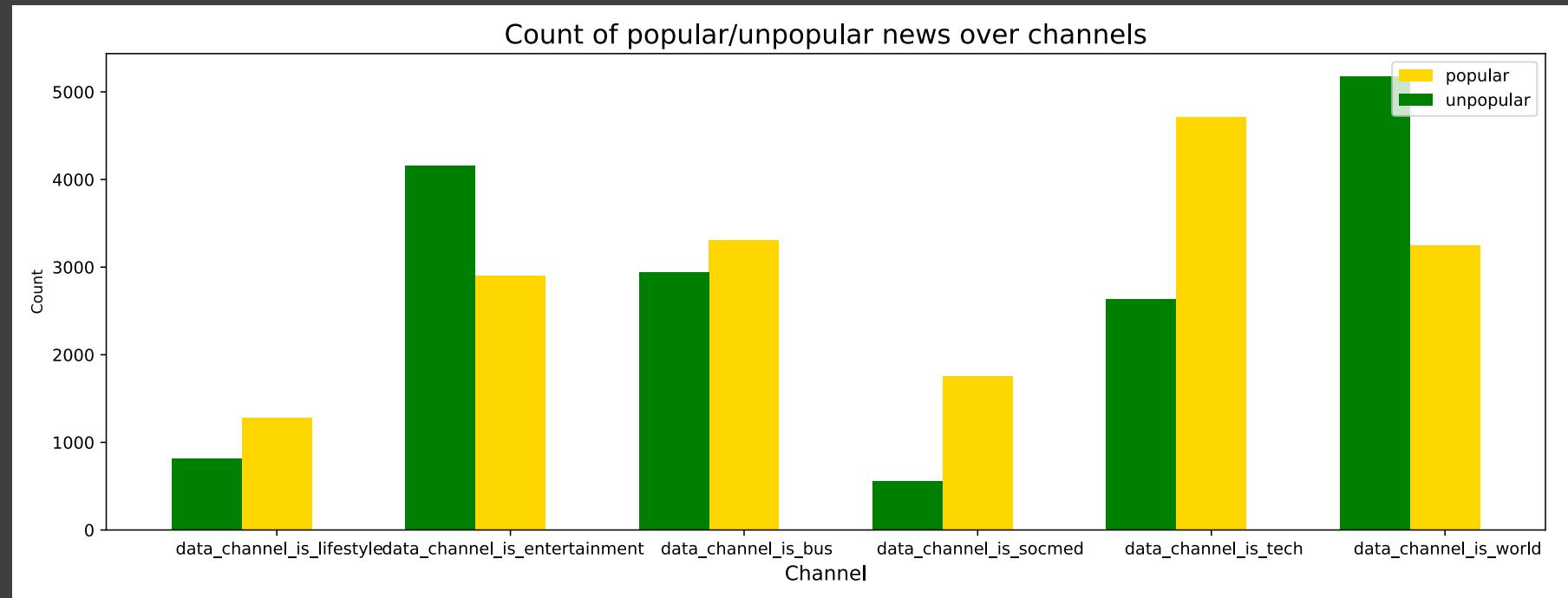
- 統計數值
 - 不正常零值
 - 極端值
- 視覺化

	timedelta	n_tokens_title	n_tokens_content	n_unique_tokens	n_non_stop_words
count	38463.000000	38463.000000	38463.000000	38463.000000	38463.000000
mean	360.385747	10.382419	563.295375	0.565049	1.027065
std	212.773031	2.113800	468.299538	3.573022	5.307978
min	8.000000	2.000000	18.000000	0.114964	1.000000
25%	174.000000	9.000000	259.000000	0.477419	1.000000
50%	347.000000	10.000000	423.000000	0.542986	1.000000
75%	547.000000	12.000000	729.000000	0.611111	1.000000
max	731.000000	23.000000	8474.000000	701.000000	1042.000000

資料前處理

資料筆數：38463 -> 33873

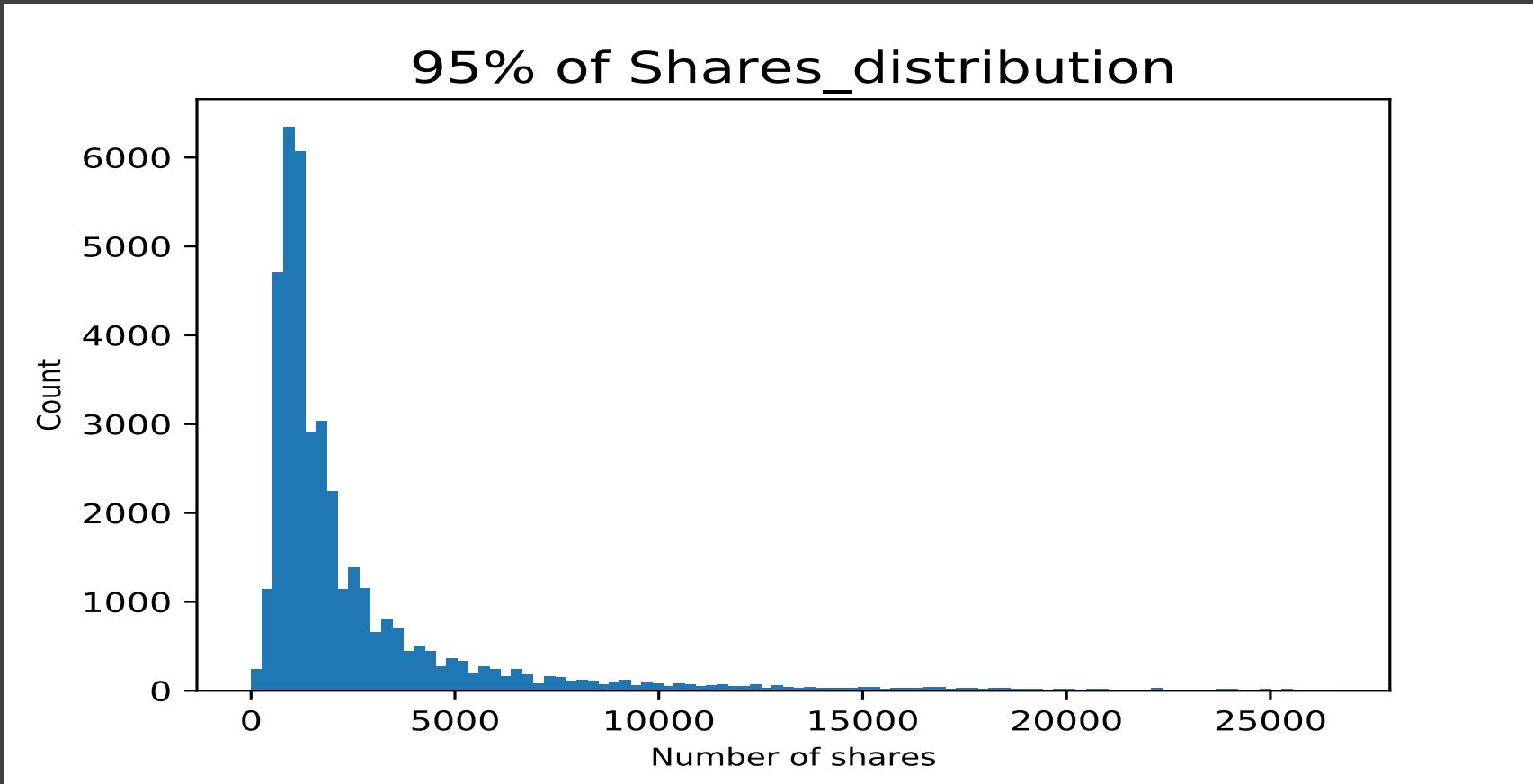
- 統計數值
 - 不正常零值
 - 極端值
- 視覺化
- 標準化



- 時間：週五、週六、週日
- 議題：生活、商業、科技、社會

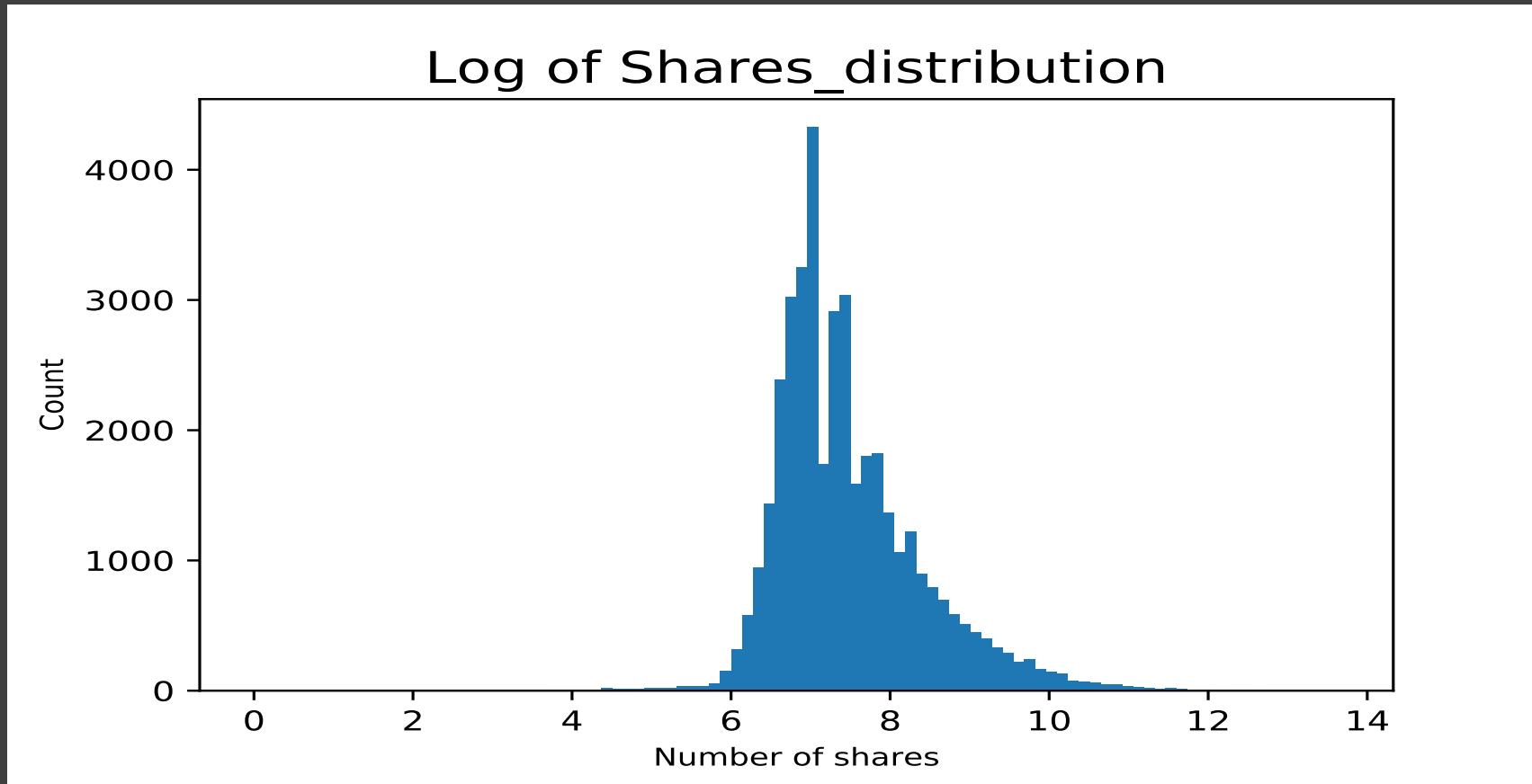
預測模型

- 目標變數 – 分享數：取Log



預測模型

- 目標變數 – 分享數：取Log



預測模型

- 目標變數 – 分享數：取Log，極端值

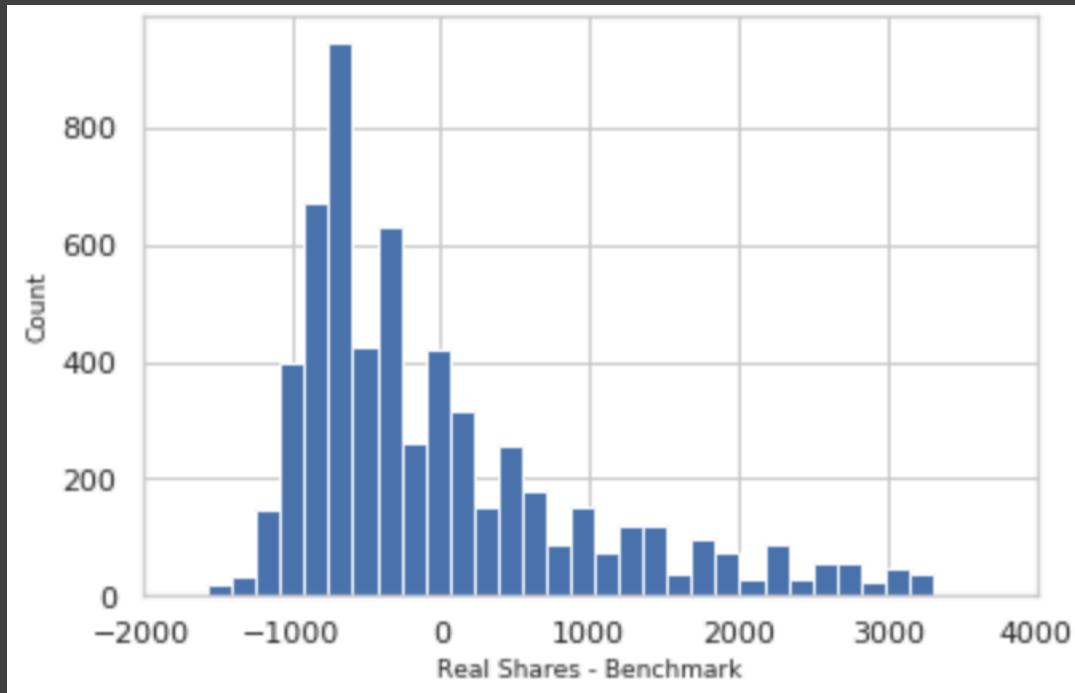
取分享數上限值	無	5,000	8,000	10,000
剩下總資料筆數	38,463	29,991	31,822	3,2397

- 模型選擇 : Adaboost , Regression , Regression Tree
- 變數篩選 : 共線性測驗
- 資料切分 : 80% 訓練集、 20% 測驗集
- 評分方法 : MAE (平均誤差絕對值) 、 殘差圖
- 評分標準 : 平均分享數

預測模型

分享數 < 5000

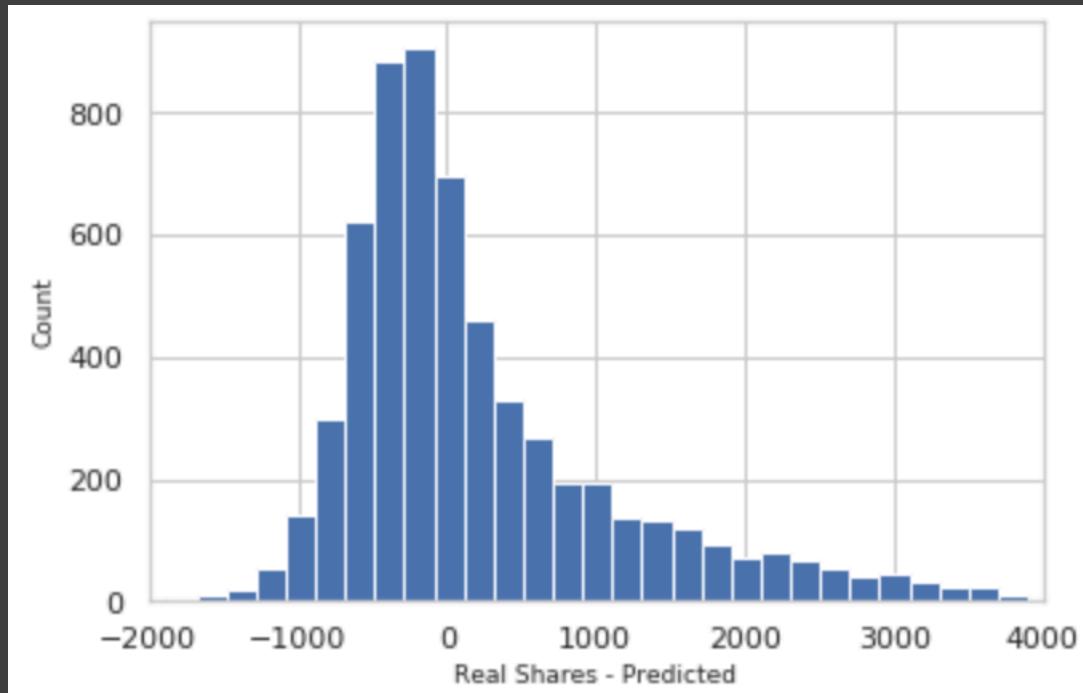
平均分享數 : 1591



MAE = 769

正負200筆 = 846

預測結果



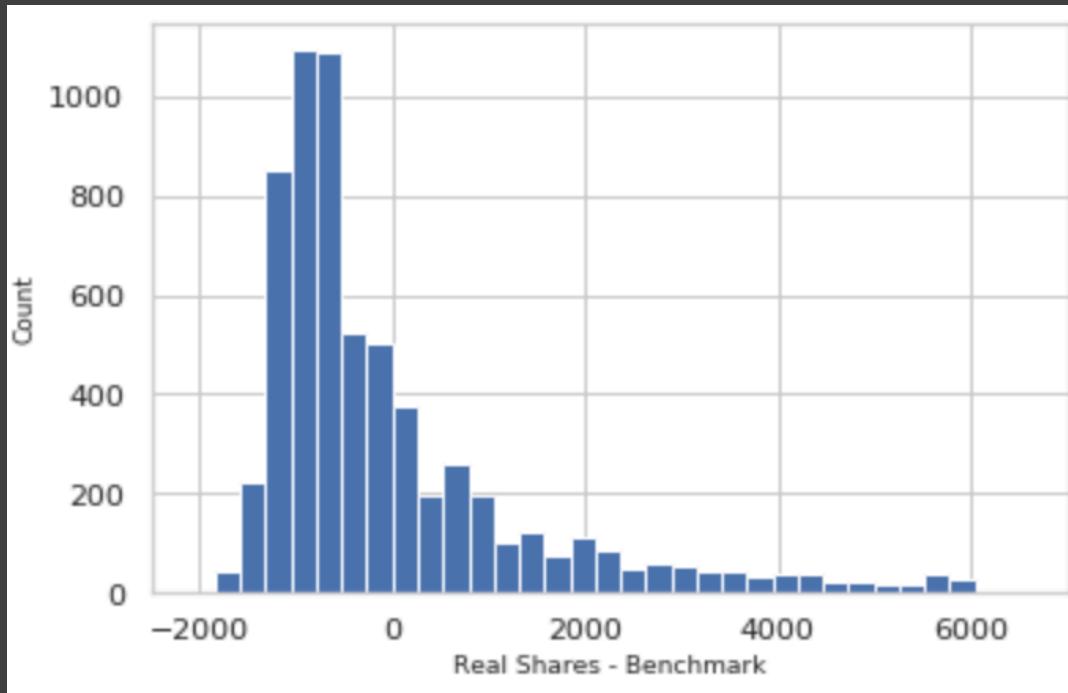
MAE = 633

正負200筆 = 1456

預測模型

分享數 < 8000

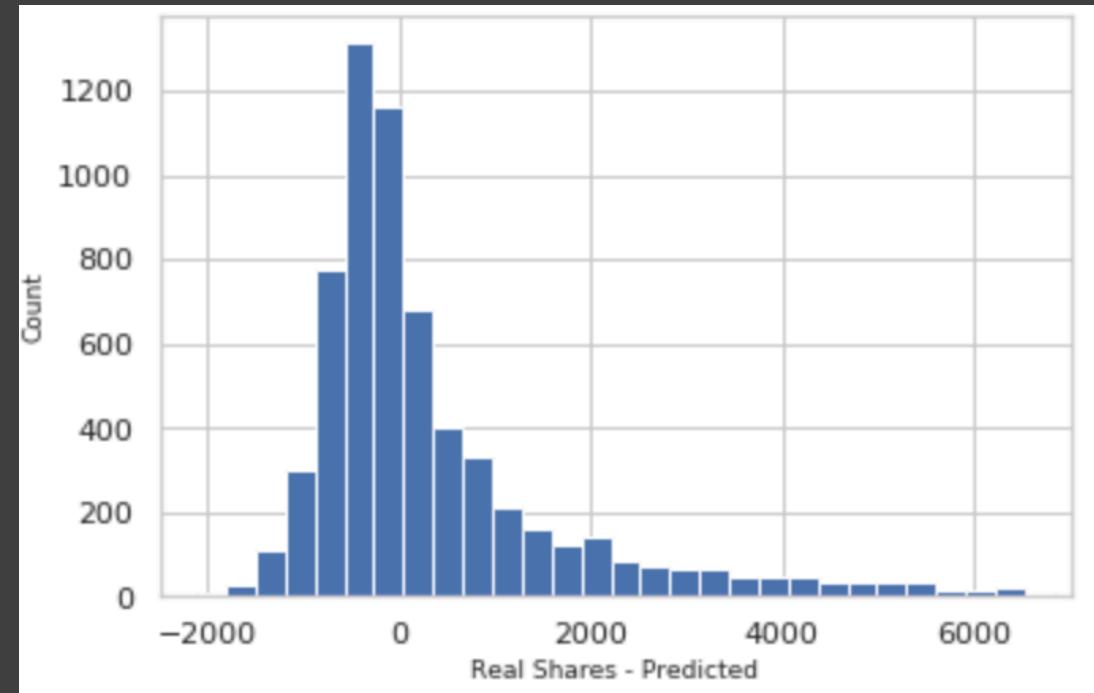
平均分享數 : 1861



$$\text{MAE} = 1061$$

$$\text{正負200筆} = 563$$

預測結果



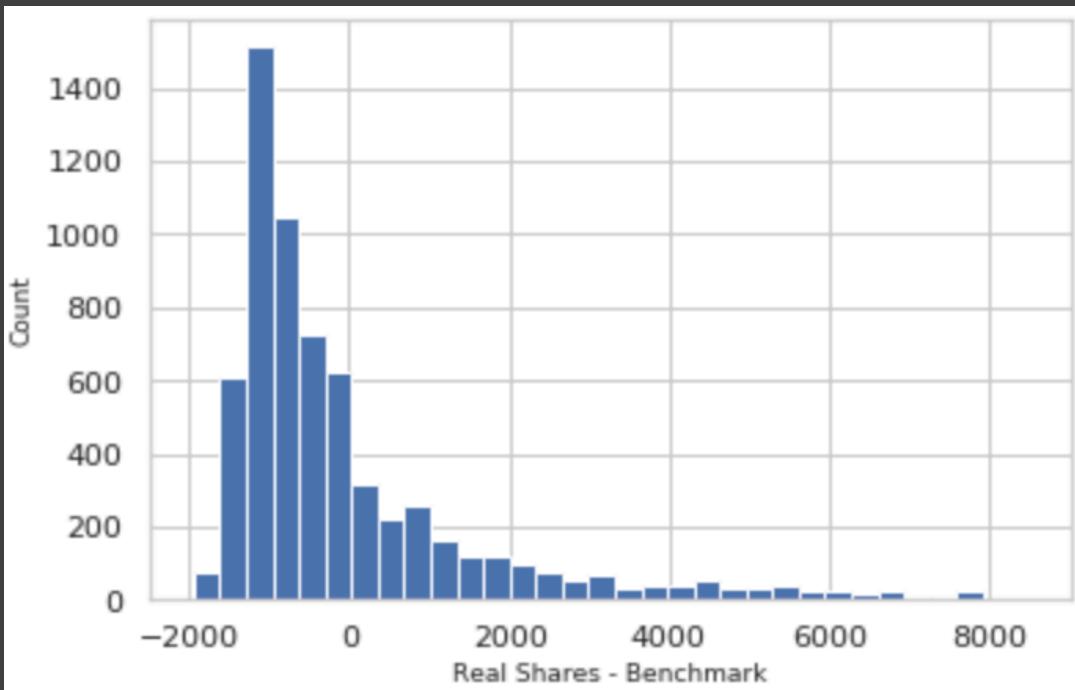
$$\text{MAE} = 888$$

$$\text{正負200筆} = 1267$$

預測模型

分享數 < 10000

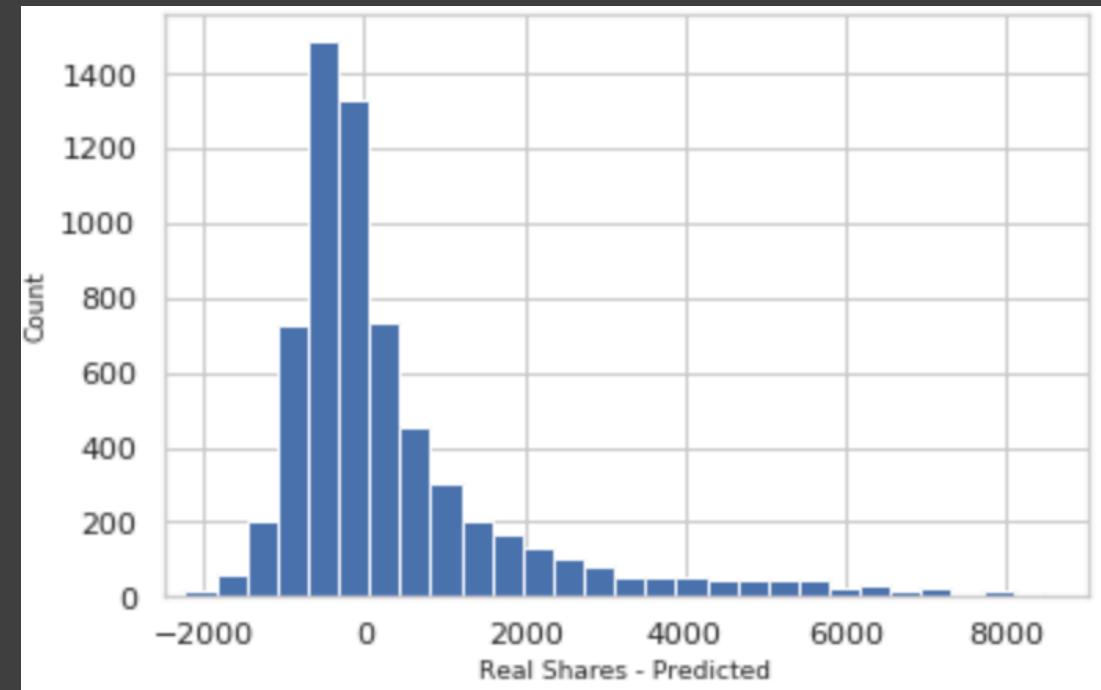
平均分享數 : 1986



$$MAE = 1208$$

$$\text{正負200筆} = 534$$

預測結果



$$MAE = 1007$$

$$\text{正負200筆} = 1162$$

效益評估及後續發展

- 洽談合作數量
- 合作案收入成長比
- 調整模型參數
- 分群的方式：數值、比例、頻道...

效益評估及後續發展

商業題目：預知議題潛在行動者的數量

分析題目：預測網路新聞分享量



- 資料初探
- 生態圖
- 應用發想
- 定義題目

- 統計數值
- 視覺化
- 檢視清理

- 分群界線
- 模型選擇
- 變數篩選
- 分群結果

- 模型選擇
- 變數篩選
- 各群分享數
預測結果

- 洽談合作
數量
- 合作案
收入成長比

效益評估及後續發展

- 洽談合作數量
- 合作案收入成長比
- 調整模型參數
- 分群的方式：數值、比例、頻道...
 - 其他可能變數
 - 加強模型預測力
 - 分享者群體樣貌
 - 市場定位
 - 社群影響者
 - 時間及管道

Thank you.

陳幸君 Chen, Hsing-Chun
hsingchun.c@gmail.com