



南山人壽：理賠客戶再購與商品推薦

指導業師 陳仕龍

指導老師 石百達、張智星

工管系大四

胡進揚

財金所碩一

張芮綺

財金系大四

馮啟倫

生醫電資所碩一

曾煒翔

2020/6/18



研究方向

1. 何謂理賠客戶再購
2. 理賠客戶再購預測模型
3. 理賠客戶商品推薦模型
4. 家庭關係與再購

本次專案完成的目標並著重於第一個研究方向



目錄

再購定義

- 資料基本分析
- 理賠後再購定義詳細說明

再購預測模型

- 模型簡介 — Random Forest, SVM, NN
- 階段一:模型預測結果
- 階段二:模型預測結果與修改

結果分析與修改方向

- 6/25 階段三 :模型預測結果與修改
- 本次專案結果分析

分工與資料連結

- 分工表
- GitHub連結



再購定義

- 資料基本分析
 - 理賠檔、再購檔資料說明
 - 關鍵發現
- 理賠後再購定義詳細說明
 - 理賠檔資料篩選
 - 再購檔資料篩選
 - 再購定義篩選

資料基本分析



資料表分析-理賠檔

- 客戶ID相關欄位(與人有關)
 - INJURED_RK
 - **INSURED_RK**
 - POLICY_HOLDER_RK
 - MATURITY_BENEFICIARY_RK
 - DEATH_BENEFICIARY_RK
- 客戶理賠資訊欄位(與事件有關)
 - Claim_RK
 - Policy_RK
 - BundleSubtype2
 - illness_code
 - illness_desc
 - DiagnosisCode_DESC
 - **claim_settle_dt**
 - REIMBURSED_YR_TW



資料表分析-再購檔

- 客戶ID相關欄位(與人有關)
 - INSURED_RK
 - POLICY_HOLDER_RK
 - MATURITY_BENEFICIARY_RK
 - DEATH_BENEFICIARY_RK
- 客戶再購資訊欄位(與產品有關)
 - RRKER_CD
 - Policy_RK
 - payment_period
 - AFYP_NT
 - SHORT_NAME
 - EFFECTIVE_DT



資料表分析-重要資訊

- 理賠檔重要資訊
 - 受理理賠期間為 2014/12/31至2017/12/31
 - 一位客戶可以具有多次理賠紀錄
 - 一位客戶可能在一天中有多筆理賠紀錄
 - 在理賠檔中的客戶為INSURED_RK(被保險人)
- 再購檔重要資訊
 - 保險在購期間為 2017/12/31至2018/12/31
 - 一個客戶可以有多筆再購紀錄
 - 一位各課可以在一天中有多筆再購紀錄
 - 在再購檔中的客戶為INSURED_RK(被保險人)與POLICY HOLDER_RK(要保人)

理賠後再購定義詳細說明



再購定義(一)

- 理賠檔資料篩選準則
 1. 挑選於2017/1/1至2017/12/31間有發生理賠事件的顧客
 2. 若一個人有多筆理賠紀錄, 則挑選理賠時間最晚發生者
 3. 若仍有多筆理賠紀錄時, 則挑選具有最高理賠金額者
 4. 若理賠金額依然相同, 則挑選第一筆被觀察之理賠事件



再購定義(二)

- 再購檔資料篩選準則
 1. 挑選於2018/1/1至2018/12/31發生再購行為的客戶
 2. 若有一位客戶具有比再購紀錄, 則挑選時間最早的再購紀錄
 3. 若最早的在購紀錄有多筆, 則以最早觀察到的在購紀錄為準

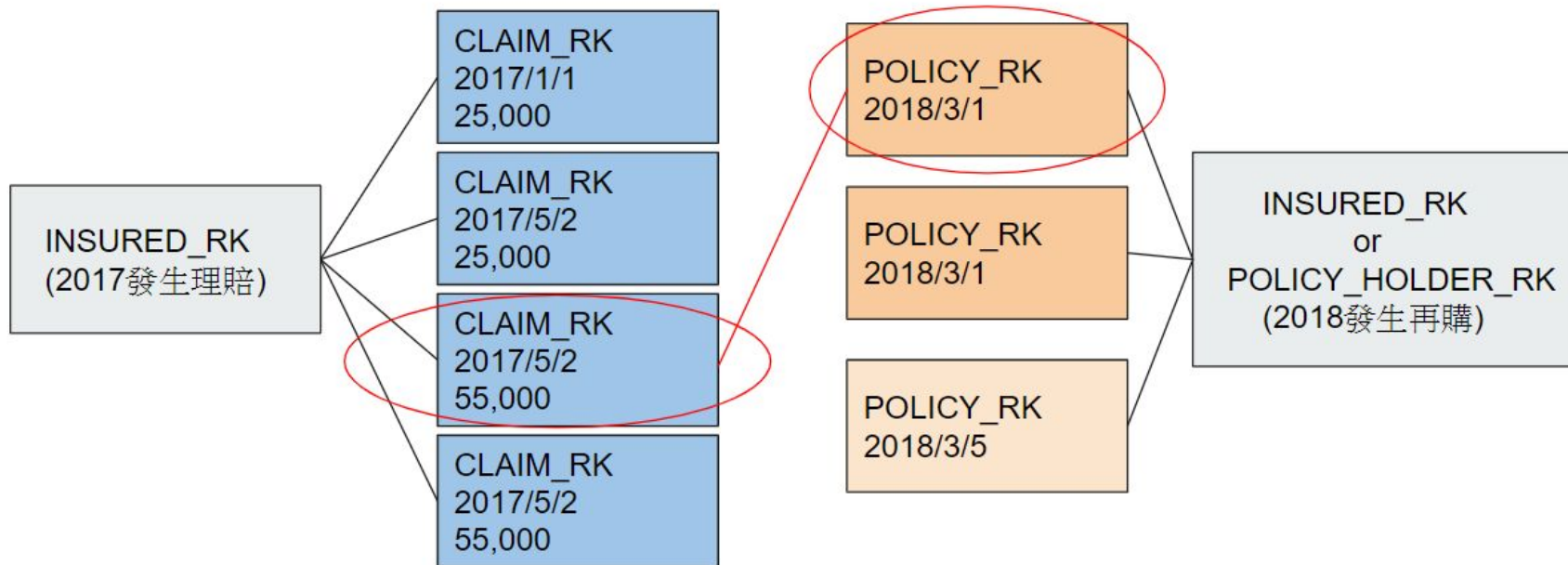


再購定義(三)

- 再購檔與理賠檔連接準則

- 理賠檔中的INSURED_RK與再購檔中的INUSURED_RK相同 (被對被)
- 兩陪檔中的INSURED_RK與再購檔中的POLICY_HOLDER_RK相同 (被對要)

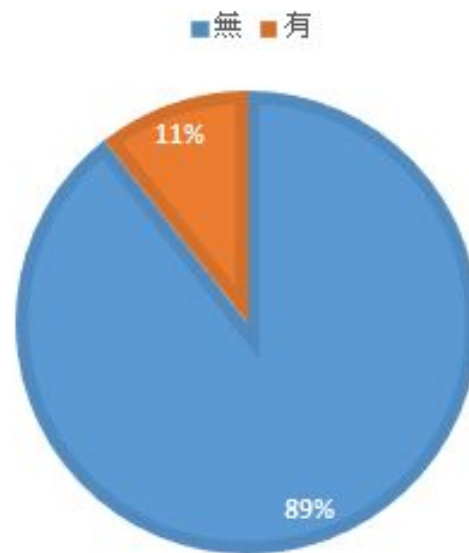
再購定義(四)



定義篩選結果

- 2017發生理賠總人數: **95,408**
- 2017理賠後再購人數: **10,097**

2017發生理賠之顧客資訊





再購預測模型

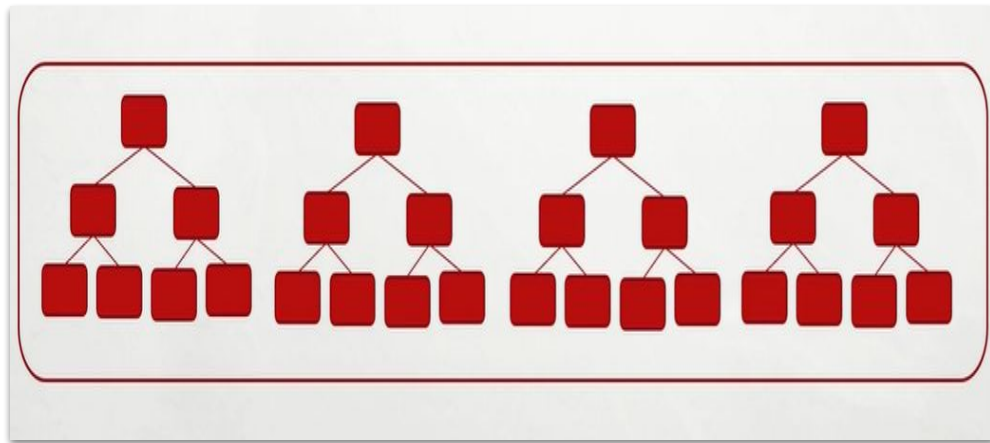
- 模型簡介 — Random Forest, SVM, NN
- 階段一:模型預測結果
- 階段二:模型預測結果與目標調整並修正

模型簡介 — Random Forest, SVM, NN

Random Forest — 隨機森林模型

特點：

- 利用隨機抽取sample跟feature建構許多決策樹
- 離散跟連續型資料都可以使用
- 結果可視化程度高



SVM — 支持向量機

特點：

- 將資料投影至高維度
- 非線性投影方式(kernel)有多種選擇
- 可在高維度空間處理原始空間無法處理的問題



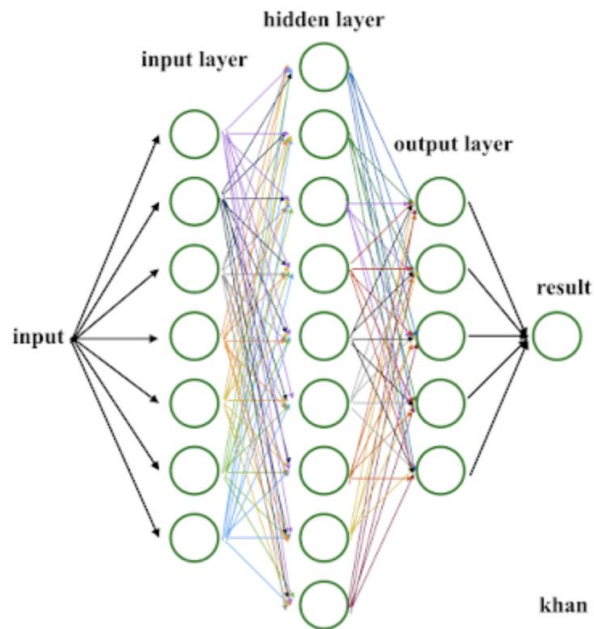
NN — 深度學習模型

特點：


- 利用多個非線性回歸方程式捕捉資料特性
- 善於解決多維度的資料
- 可藉由梯度下降的方式找出解答

應用在再購模型優勢：

- 客戶資料為多維度(70維)資料
- 目標為分類問題



模型階段一



資料型態與目標

- **訓練目標:** 根據客戶理賠資訊, 預測客戶未來是否有再購行為, 並以追求高**整體預測率(total accuracy)為目標**
- **預測任務:** 為二元分類問題, 預測未來是否有再購行為發生
 - 若預測值為1: 未來**有**再購需求
 - 若預測值為0: 未來**無**再購需求
- **解釋因子:** 客戶理賠檔資訊欄位、客戶屬性檔資訊欄位
 - 數值型態資料: 進行Z-Normalization
 - 多類別型態資料: 轉換成Dummy Variable

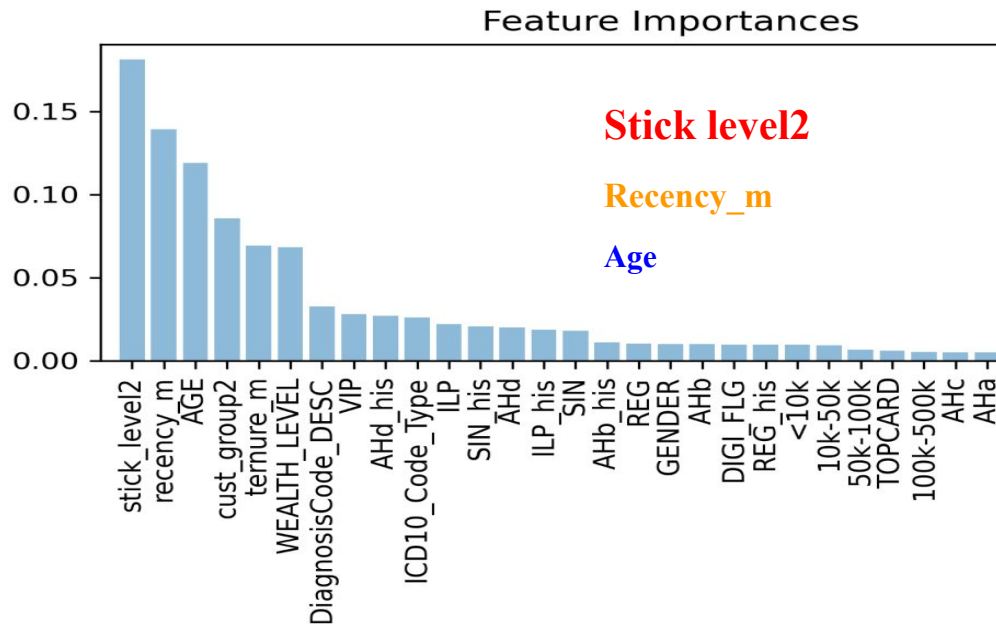
Random Forest 預測結果

訓練結果：

-train set : 90.33%

-test set : 89.85%

```
repurchase customer : 3146  
predict repurchase customer : 293  
actually repurchase customer : 199
```





SVM 預測結果

【linear kernel】, C=1 預設值

-training set: 89.66% 正確, $P(\hat{y}=1 \mid Y=1) = 763/(763+6321) \sim 10.77\%$

-testing set : 89.67% 正確, $P(\hat{y}=1 \mid Y=1) = 317/(317+2696) \sim 10.52\%$

【rbf kernel】, C=1 預設值

-training set: 90.26% 正確, $P(\hat{y}=1 \mid Y=1) = 811/(811+6273) \sim 11.45\%$

-testing set : 89.76% 正確, $P(\hat{y}=1 \mid Y=1) = 258/(258+2755) \sim 8.56\%$



NN 預測結果

訓練結果:

-training set: 90.27% 正確

-testing set : 89.36% 正確

-測試集 19000 筆資料有 2058 筆再購的實際案件, 模型預測其中有 456 筆為再購案件


*model
code: <https://reurl.cc/xZK384>



模型預測結果

	Random Forest	SVM		NN
kernel		linear kernel	rbf kernel	
Training set Accuracy	90.33%	89.66%	90.26%	90.27%
Testing set Accuracy	89.85%	89.67%	89.76%	89.36%
Precision Rate	67.91%	10.52%	8.56%	
Recall Rate	6.32%			22.2%

模型階段二



階段二 目標調整與修正

調整後訓練目標:除了追求高整體預測率(Total accuracy), 提高成功預測再購需求用戶的比率(Precision rate)更為重要。

(註: Precision rate = 模型實際抓到再購人數 / 樣本再購的總人數)

修正:

1. bootstrap: 樣本資料的比例嚴重失衡, 不再購的資料較多, 可能影響到模型預測的結果
2. 進行理賠檔欄位、客戶屬性檔欄位分類, 並分成三種資料進行模型訓練:
 - Behavior data, ex: 過去持有保單紀錄、VIP等級
 - Personal data, ex: 年齡、性別、理賠原因
 - Original data (behavior data + personal data + 未能分類的欄位)

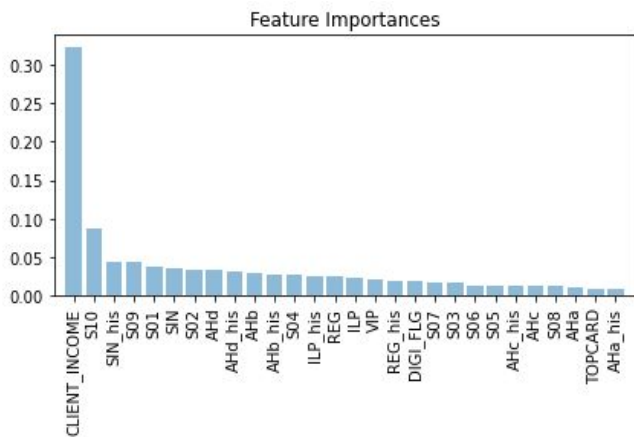
bootstap sampling, 不再購 : 再購 = 1 : 1

Random Forest & Decision Tree 預測結果

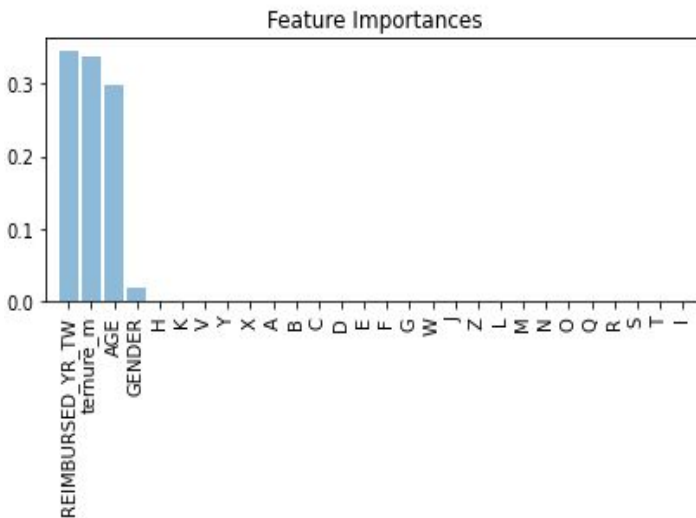
	Behavioral	Personal	All data
<i>Random Forest</i> Testing set Accuracy	65.87%	58.22%	67.43%
<i>Decision tree</i> Testing set Accuracy	64.90%	56.31%	64.29%
<i>Random Forest</i> Precision Rate	69.60%	58.89%	64.96%
<i>Decision tree</i> Precision Rate	60.31%	59.47%	62.28%

Random Forest — Feature Importance

Behavioral data



Personal data

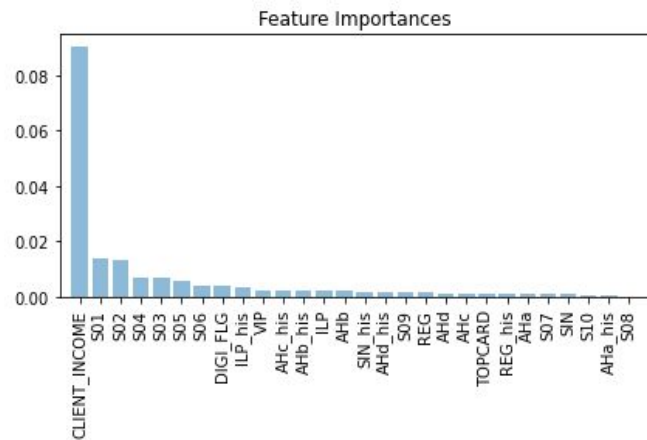


Original data

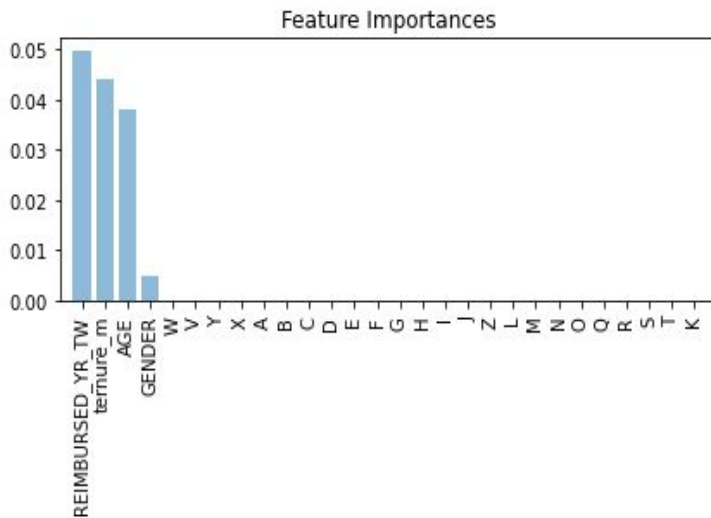
1) recency_m	0.167158
2) CLIENT_INCOME	0.150162
3) AGE	0.089024
4) ternure_m	0.070677
5) REIMBURSED_YR_TW	0.059798
6) G4	0.055268
7) S10	0.035363
8) G1	0.030353
9) SIN_his	0.025488
10) G0	0.019207

Decision Tree — Feature Importance

Behavioral data



Personal data



Original data

1) CLIENT_INCOME	0.088308
2) recency_m	0.055944
3) REIMBURSED_YR_TW	0.036472
4) AGE	0.026460
5) ternure_m	0.025363
6) S01	0.007556
7) S02	0.007348
8) S05	0.004310
9) W5	0.003975
10) GENDER	0.003618

bootstap sampling, 不再購 : 再購 = 2 : 1

SVC 預測結果

【linear kernel】, C=1 預設值

- Original : 82.10%正確, $\text{num}(\hat{y}=1)=15050$, $P(\hat{y}=1 | Y=1) = 4036/10097 \sim 39.97\%$
- Behavioral: 82.10%正確, $\text{num}(\hat{y}=1)=15050$, $P(\hat{y}=1 | Y=1) = 4036/10097 \sim 39.97\%$
- Personal : 89.42%正確, $\text{num}(\hat{y}=1)=2$, $P(\hat{y}=1 | Y=1) = 1/10097 \sim 0\%$

【rbf kernel】, C=1 預設值

- Original : 86.16%正確, $\text{num}(\hat{y}=1)=10042$, $P(\hat{y}=1 | Y=1) = 3468/10097 \sim 34.35\%$
- Behavioral: 84.73%正確, $\text{num}(\hat{y}=1)=11528$, $P(\hat{y}=1 | Y=1) = 3526/10097 \sim 34.92\%$
- Personal : 89.37%正確, $\text{num}(\hat{y}=1)=83$, $P(\hat{y}=1 | Y=1) = 19/10097 \sim 0.19\%$



Neural Network 實驗設置

- 利用bootstrap調整label 0與 1的資料比例, 並且從label 0的資料抽取十份降低整體bias
- Model setting: 三層NN, CE loss, Adam, ReLU, lr = 0.01, epoch = 30(但是會train10份資料)
- testing樣本為19082筆資料(佔總資料20%), 其中 label=1的資料為2033筆

Neural Network 預測結果

	Behavioral + Personal			Behavioral			Personal		
不再購：再購	1:1	1.5:1	2:1	1:1	1.5:1	2:1	1:1	1.5:1	2:1
Predict number	7800	4674	2053	9998	4075	2214	9591	2441	0
Total acc	59%	76%	86%	54%	78%	85%	54%	81%	89%
Repurchase acc	74%	53%	33%	78%	46%	34%	61%	19%	0%



階段二 結論

- Behavior data 對再購預測的影響較 Personal data 大, precision rate 明顯較大
- Total accuracy 與 Precision rate 存在 trade-off
- Neural Network :
 1. All data較有價值 (dimension較多)
 2. 當資料比例在1 : 1至1 : 5時, 效果最佳



結果分析與修改方向

- 6/25 階段三：模型預測結果與修改
- 本次專案結果分析



6/25 階段三

目標: 整體預測率 total accuracy & 成功預測到再購需求用戶的比率 precision rate

修正:

1. 再購資料重新定義並抓取 (原本為2017理賠並於2018再購)
2. 透過 Logistic Regression 找出再購資料的特色
3. 呈現 ROC curve or DET curve
4. 模型 k-fold Cross Validation



本次專案結果分析

- bootstrap sampling 調整再購與不再購的樣本比例後，能更準確捕獲再購者的特徵
- Behavior data 對再購預測的影響較 Personal data 大，應著重探討該行為(Behavior)類別資料
- Total accuracy 與 Precision rate 存在 trade-off
(precision rate = 模型實際抓到再購人數 / 樣本再購的總人數)
- 從Random Forest和Decision Tree可發現：
客戶年收入、客戶年齡、客戶戶齡和理賠金額對再購與否的預測有較大的影響



分工與資料連結

- 分工表
- GitHub連結



分工表

名字	分工內容
胡進揚	再購定義以及資料前處理和 EDA
張芮綺	模型建構與設計實驗及結果分析
馮啟倫	模型建構與設計實驗及結果分析
曾煒翔	模型建構與設計實驗及結果分析



GitHub連結

GitHub:

https://github.com/chiluen/Fintech_NanShan