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

2020/6/25

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

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
  - 保險再購期間為2016/12/31至2018/12/31
  - 一個客戶可以有多筆再購紀錄
  - 一位客戶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)

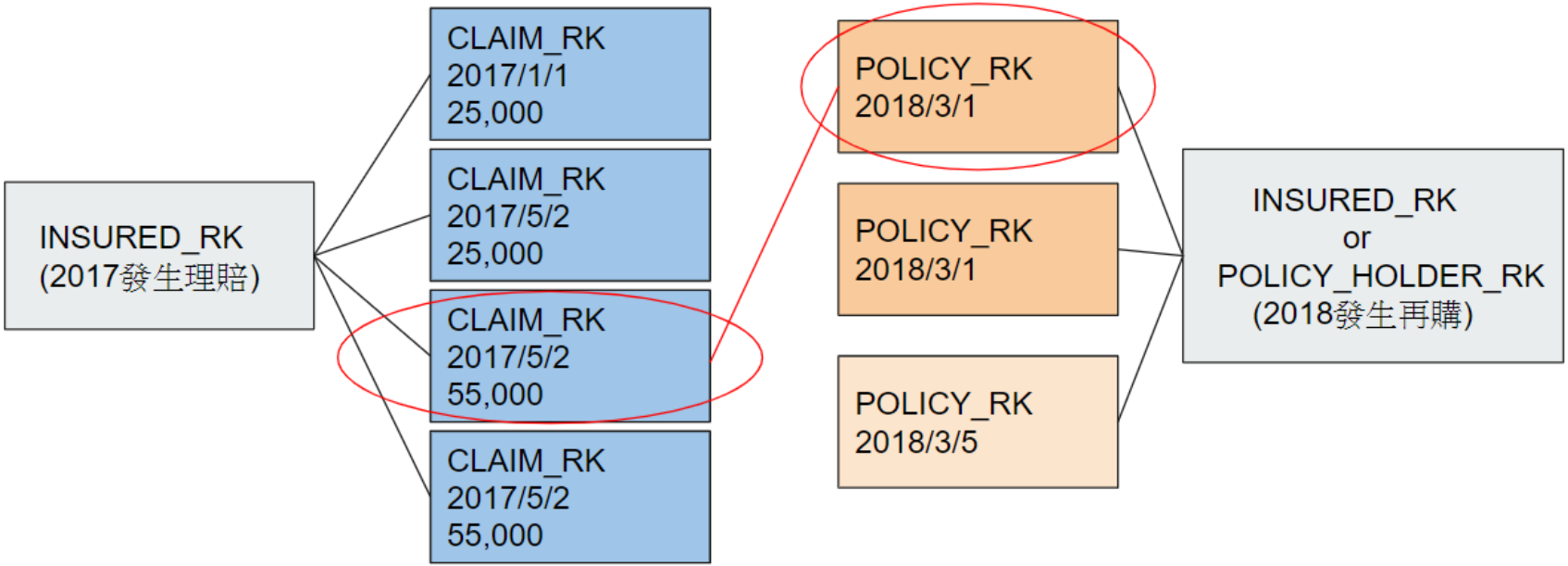
- 再購檔資料篩選準則

1. 挑選於2018/1/1至2018/12/31發生再購行為的客戶
2. 若有一位客戶具有多比再購紀錄，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)

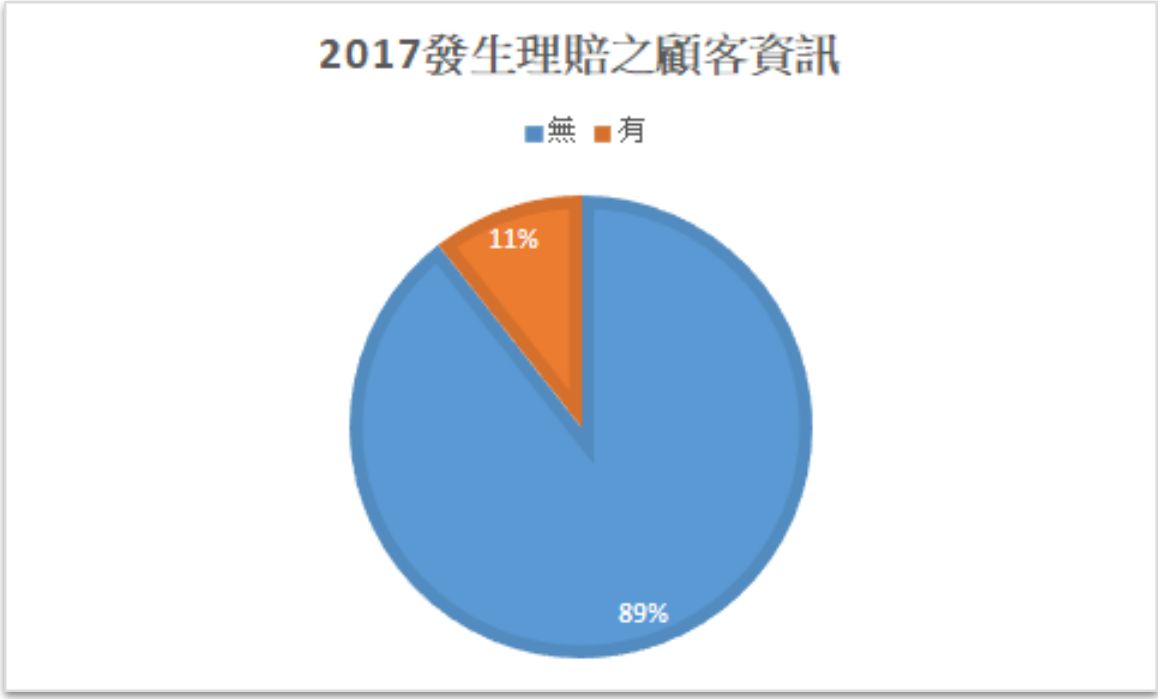
- **再購檔與理賠檔連接準則**
  - 理賠檔中的INSURED_RK與再購檔中的INUSURED_RK相同 **（被對被）**
  - 理賠檔中的INSURED_RK與再購檔中的POLICY_HOLDER_RK相同 **（被對要）**

再購定義(四)


INSURED_RK
(2017發生理賠)
CLAIM_RK
2017/1/1
25,000
CLAIM_RK
2017/5/2
25,000
CLAIM_RK
2017/5/2
55,000
CLAIM_RK
2017/5/2
55,000
POLICY_RK
2018/3/1
POLICY_RK
2018/3/1
POLICY_RK
2018/3/5
INSURED_RK
or
POLICY_HOLDER_RK
(2018發生再購)

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



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
- 階段三：模型預測結果與目標調整並修正

模型簡介 － Random Forest, SVM, NN

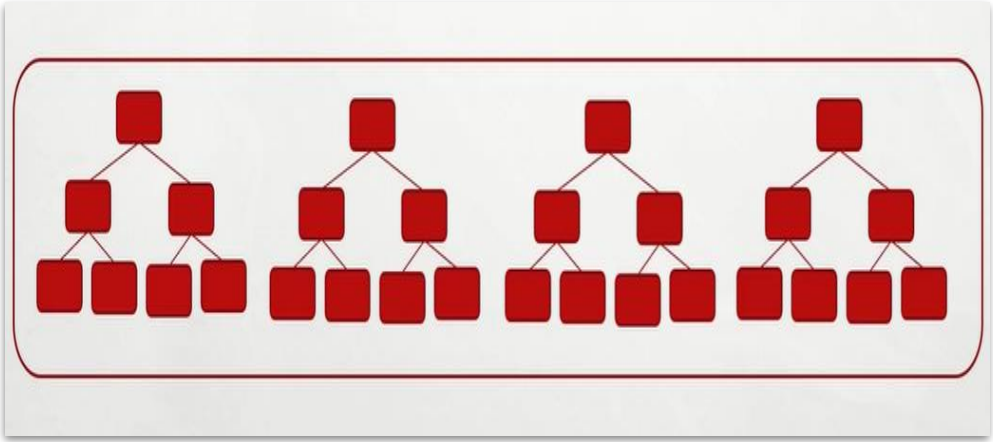
Random Forest – 隨機森林模型

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



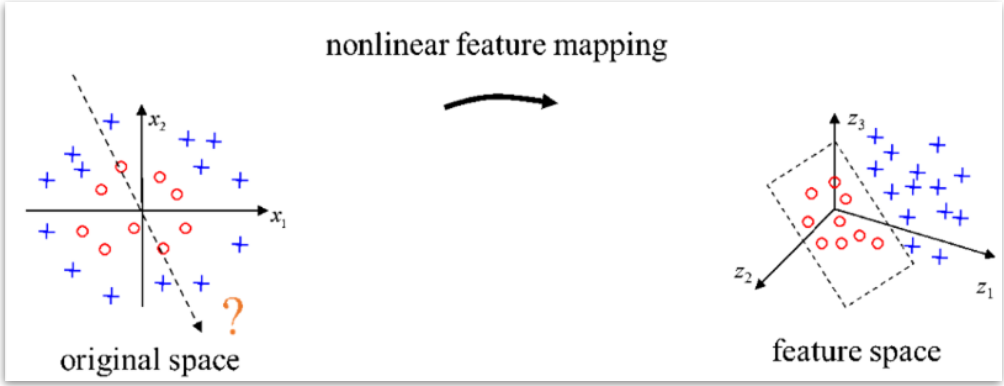
SVM － 支持向量機

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

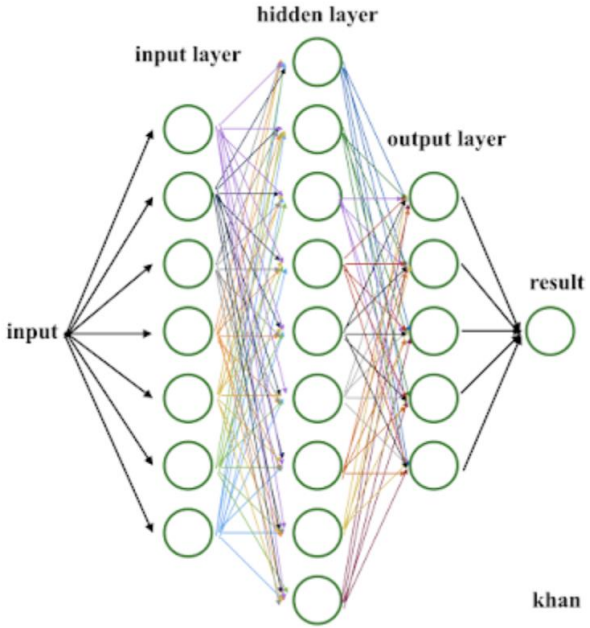
-善於解決多維度的資料

-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

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%
Recall Rate	6.32%	10.52%	8.56%	22.2%

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

模型階段二

階段二 目標調整與修正

調整後訓練目標：

提高成功預測再購需求用戶的(Recall rate)

比追求高整體預測率(Total accuracy)更為重要。

（註: Recall 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
- Oringinal 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
Random Forest Testing set Accuracy	65.87%	58.22%	67.43%
Decision tree Testing set Accuracy	64.90%	56.31%	64.29%
Random Forest Recall Rate	69.60%	58.89%	64.96%
Decision tree Recall Rate	60.31%	59.47%	62.28%

bootstap sampling，不再購：再購 = 2 : 1

SVC 預測結果

【linear kernel】	all	personal	behavioral
accuracy	82.10%	89.42%	89.30%
recall rate	34.37%	0.00%	39.97%

【rbf kernel】	all	personal	behavioral
accuracy	86.16%	84.73%	89.37%
recall rate	34.35%	0.19%	34.92%

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%
Recall rate	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 大非常多。

  (Behavior data預測的Recall Rate較好)

  Total accuracy 與 Recall rate存在 trade-off的情形

- Neural Network :

1. All data train 的 model較有價值(dimension較多)
2. 當資料比例在1 : 1至1 : 1.5時，效果最佳

模型階段三

6/25 第三階段目標修正與調整

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

修正：

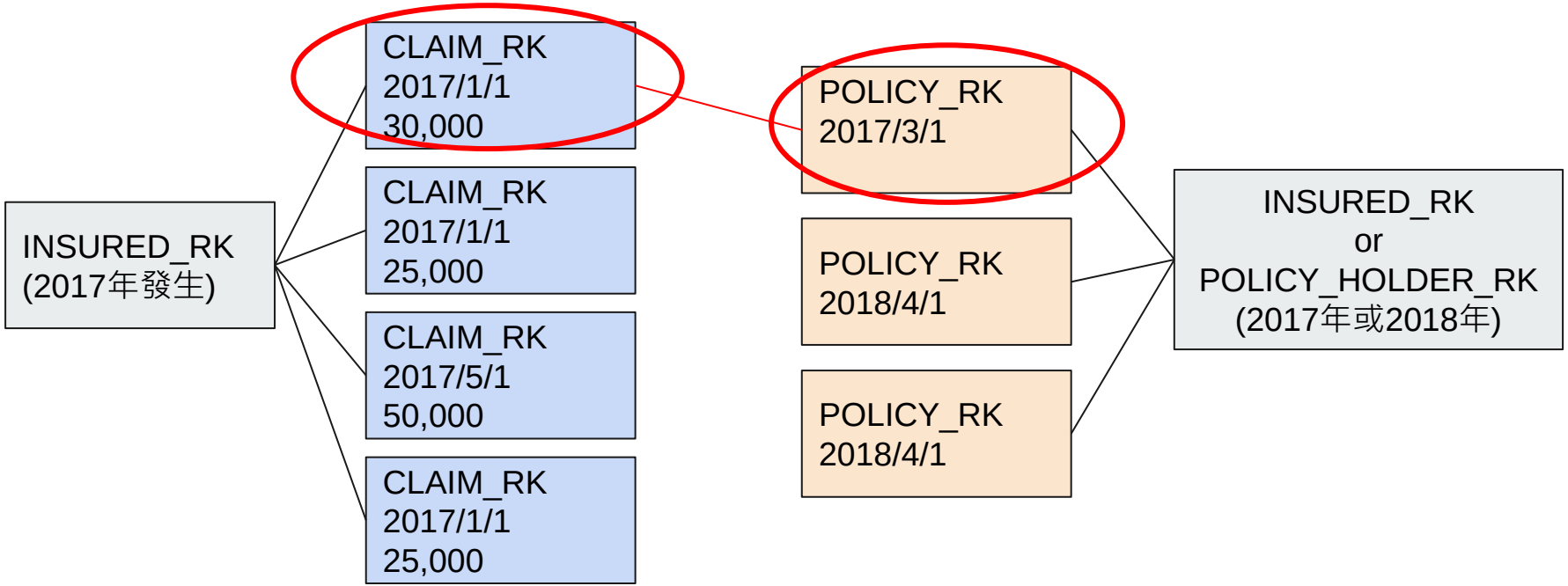
1. 再購資料重新定義並抓取（原本為2017理賠並於2018再購）
2. 呈現 ROC curve or DET curve
3. 模型 k-fold Cross Validation

6/25 階段三-再購定義調整(一)

再購定義調整:

1. 若一位客戶有多筆理賠紀錄，則選擇最早的理賠紀錄
2. 若理賠時間相同，則選擇理賠金額最大者作為紀錄
3. 再購檔的資訊更改為包含2017年與2018年的所有再購紀錄
4. 其餘規則皆與最原始定義相同

6/25 階段三-再購定義調整(二)



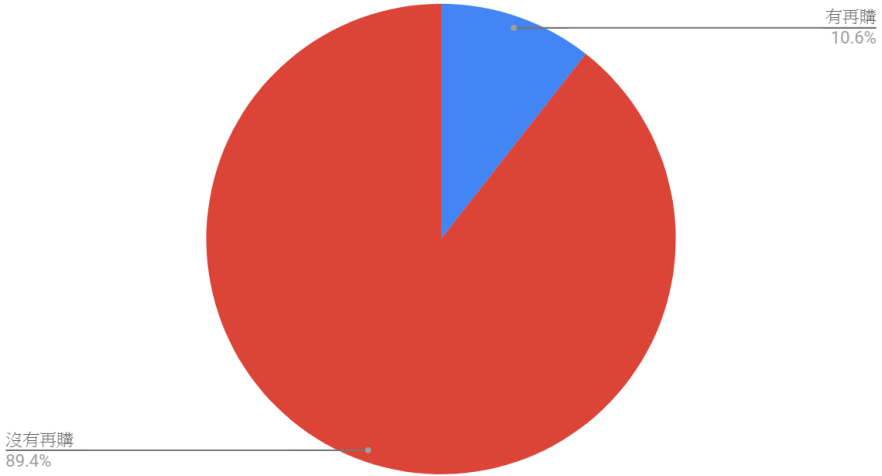
6/25 階段三-再購定義調整(三)

- 定義調整後變化

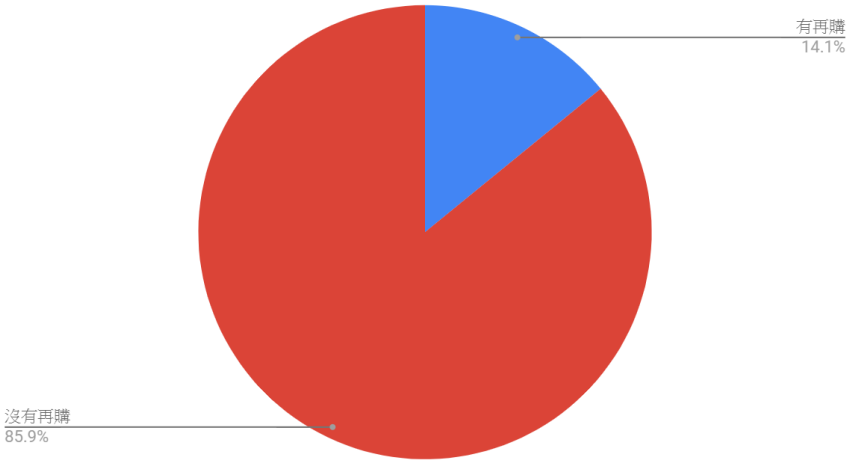
	原始定義	新定義
再購人數	10,097人	13,465人
無再購人數	85,311人	81,943人

6/25 階段三-再購定義調整(四)

理賠後在購人數比例分布(原始定義)



理賠後再購人數比例分布(新定義)



6/25 階段三-模型調整

模型調整:

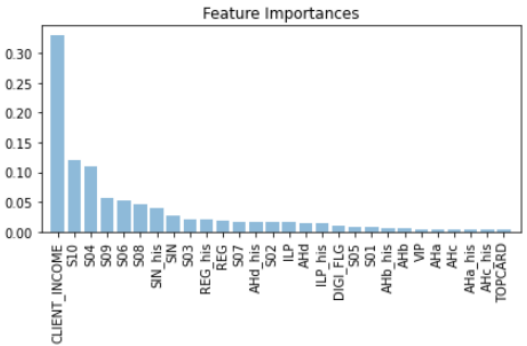
1. 利用K-fold做多次的testing，解決單次testing可能導致的bias
2. 利用Bootstrap算平均結果，增加模型可信度
3. 採用ROC curve，比較整體模型優劣

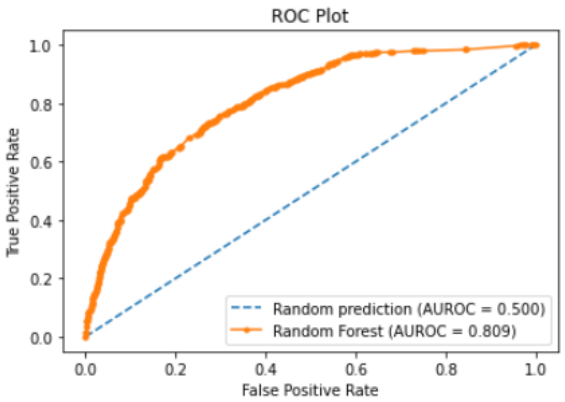
bootstap sampling，不再購：再購 = 1 : 1

Random Forest 預測結果

	Behavioral	Personal	All data
Random Forest **Accuracy**	74.23%	58.95%	73.77%
Random Forest **Recall Rate**	71.84%	59.33%	44.49%

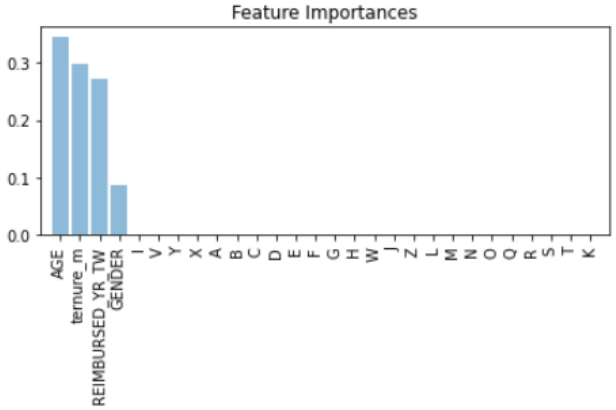
Behavior

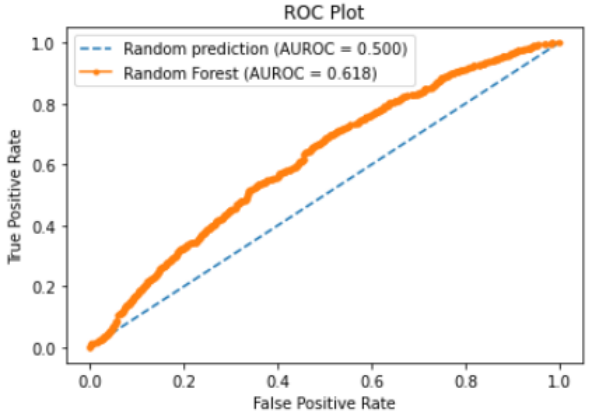




80.9%

Personal

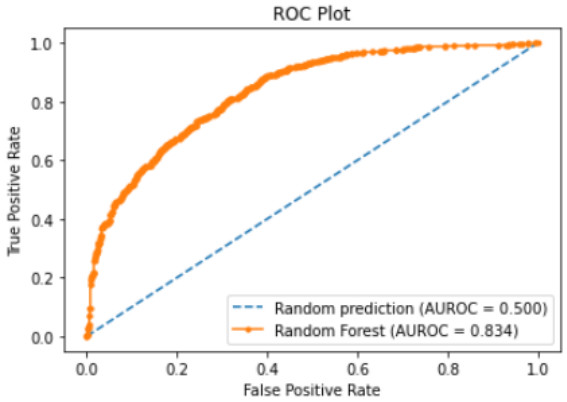




61.8%

All data

1)	CLIENT_INCOME	0.212517
2)	recency_m	0.165515
3)	G4	0.093525
4)	AGE	0.053052
5)	tenure_m	0.052784
6)	G2	0.040291
7)	S10	0.039327
8)	G1	0.031943
9)	REIMBURSED_YR_TW	0.031701



83.4%

註: bootstap sampling比例:不再購 : 再購 = 2 : 1

SVC 預測結果

【linear kernel】	all	personal	behavioral
accuracy	0.7989	0.8588	0.7989
recall rate	0.5328	0.0000	0.5328

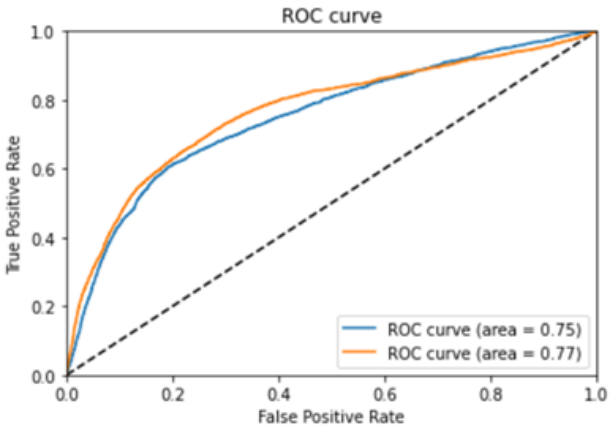
【rbf kernel】	all	personal	behavioral
accuracy	0.8272	0.8551	0.8207
recall rate	0.5110	0.0162	0.4997

藍線 : linear kernel

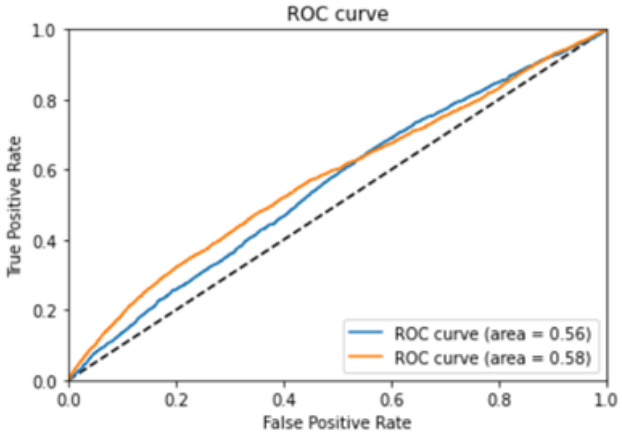
橘線 : rbf kernel

SVC ROC curve

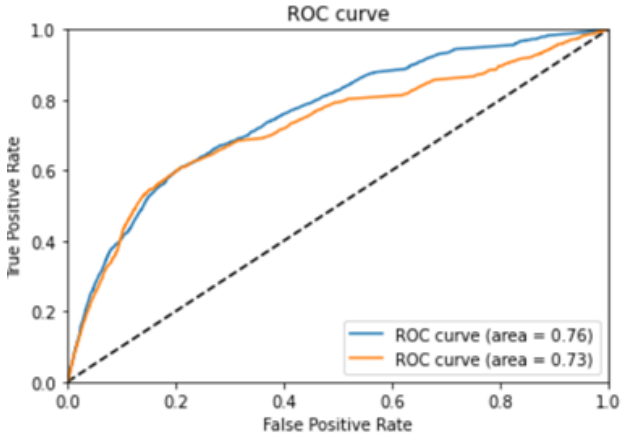
All



Personal



Behavior

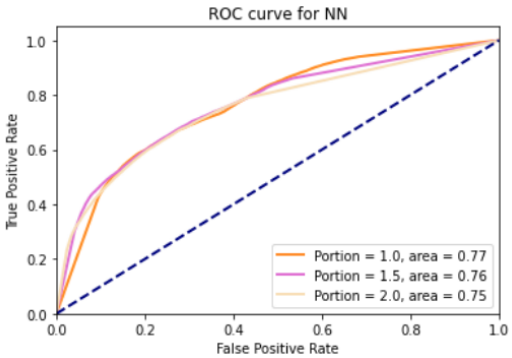


NN 結果

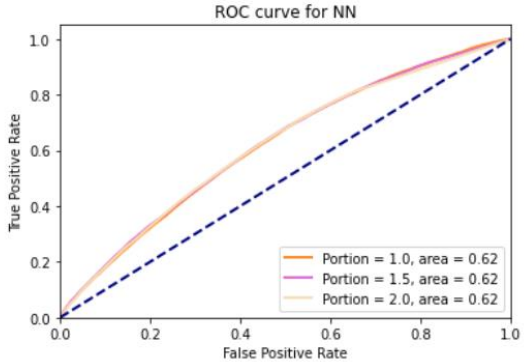
	Behavioral + Personal			Behavioral			Personal		
不再購 ： 再購	1:1	1.5:1	2:1	1:1	1.5:1	2:1	1:1	1.5:1	2:1
Total acc	61%	71%	81%	68%	80%	82%	55%	79%	84%
Recall rate	77%	60%	50%	70%	55%	50%	64%	21%	6%

NN 結果

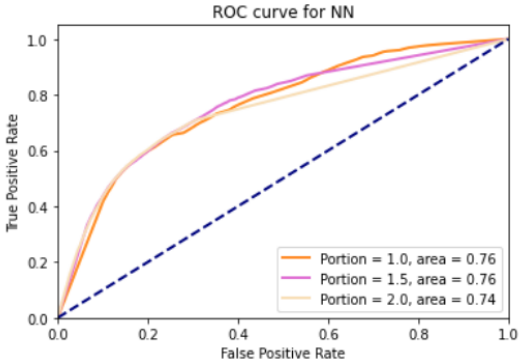
All_data



Personal_data



Behavior_data



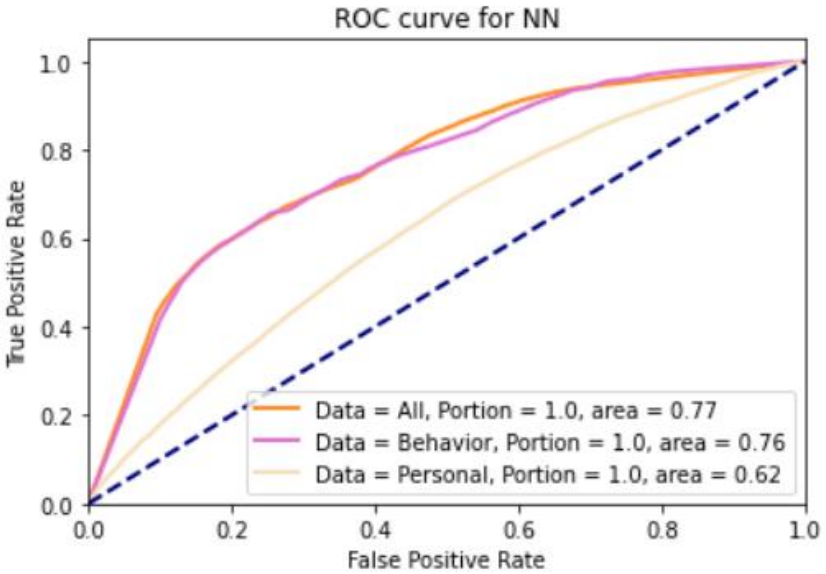
NN 結果

比較結果:

1. 使用全部data效果最好

1. 單用Behavior幾乎可以貢獻全部

1. 相較於階段二，Recall rate提升很多



階段三 結論

1. 再購定義時間間隔調整使得Recall rate顯著增加
2. 從三個模型的ROC curve可以看到 behavior data幾乎可以跟 All data(註一)表現一樣
3. 在Neural Network模型的實驗中，調整Bootstrap的比例:

Boostrap比例(註二)越高 ⇧	Accuracy越高 ⇧	Recall Rate越低 ⇩

註一: All data為personal data +beahavioral data + 不能區分類別的資訊

結果分析與建議

- 本次專案結果分析
- 建議與延伸

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 Recall 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

報告總結論

1. Behavior data可以有效的代表整份資料，未來在訓練模型時可以減少大量運算時間
2. 藉由調整再購定義(時間區隔調整)，可以使Recall rate有效提升
3. Recall rate在最終階段的各模型都大於60%，代表利用這些模型可以抓出6成以上的潛在客戶
4. ROC curve底下的面積都大於60%，代表模型效力遠比隨機來的好

未來方向

1. 類別資料除了one-hot分類，可以嘗試其他分類方式
2. 可以增加K-fold與bootstrap的次數，提供更穩定且較無偏差的模型
3. 再購資料定義的細項調整，驗證哪些部分也會影響Recall rate