



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

2020/7/2



目錄

1. 專案研究方向
2. 再購定義與資料處理
 - a. 再購定義
 - b. 資料處理
3. 模型訓練
 - a. 訓練目標與模型
 - b. 訓練結果
4. 專案結果分析



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

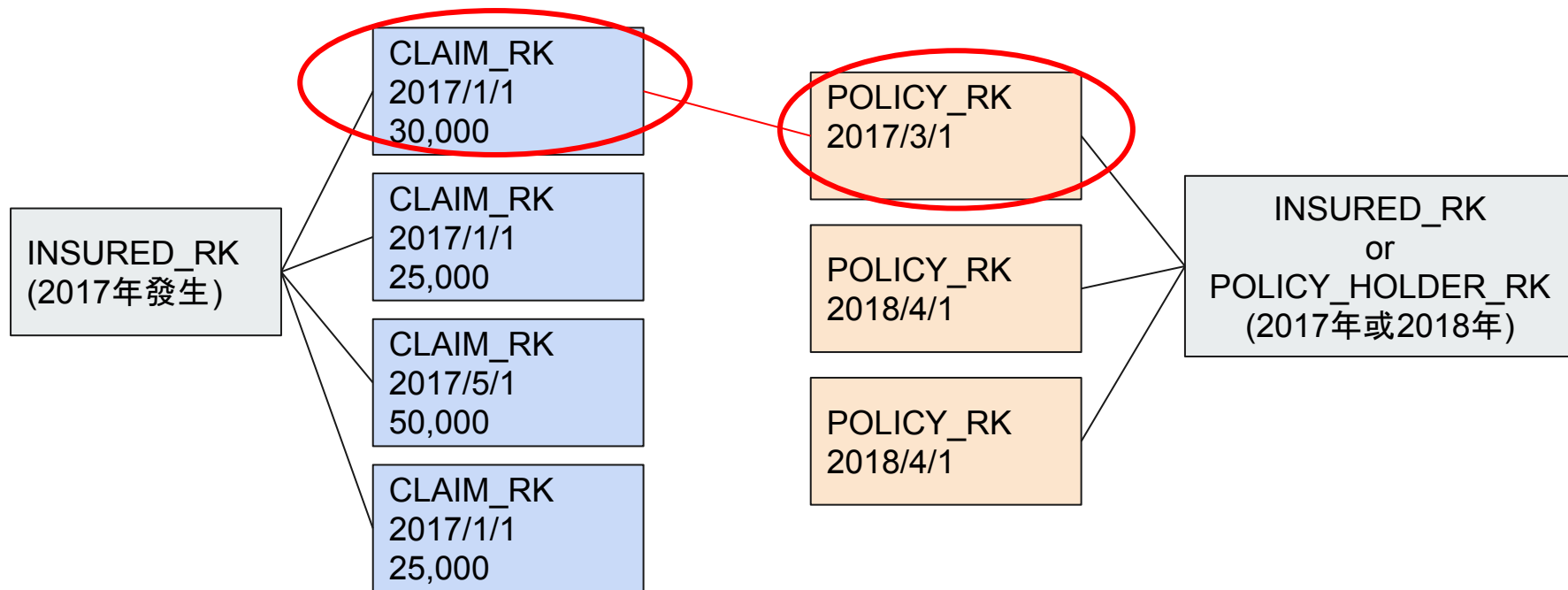
本次專案完成的部分



再購定義(一)

- 理賠檔篩選
 - 2017年有發生理賠紀錄的客戶
- 再購檔篩選
 - **最初篩選**: 僅含有2018年再購客戶
 - **改進篩選**: 包含2017與2018有發生再購的客戶
- 檔案串接
 - 理賠檔中的被保險人欄位與再購檔中的被保險人欄位相同(被對被)
 - 理賠檔中的被保險人藍位與在購檔中的要保人欄位相同(被對要)

再購定義(二)





資料處理

- 最初資料處理
 - 數值資料進行正規化(Z-normalization)
 - Dummy Variable轉成0與1的形式
- 改善後資料處理
 - 將資料分門別類, 共分成三類: **Personal Data , Behavioral Data , All data**
 - Personal Data: 客戶個人資訊 ex: 年齡、性別
 - Behavioral Data: 客戶行為資訊 ex: 過去持有保單、VIP等級
 - All Data: Personal Data + Behavioral Data + 未能分類的欄位

訓練目標與模型

- **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, 提升模型的**Recall Rate** 為**主要目標**並以提升整體預測率 (total accuracy) 為次要目標。
- **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
- **資料輸入**:
 - 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
 - **All data** (behavior data + personal data + 未能分類的欄位)

$$\text{Recall rate} = \frac{\text{模型實際抓到再購人數}}{\text{樣本再購的總人數}}$$



訓練目標與模型

使用模型		
隨機森林(Random Forest)	SVM-支持向量機	深度學習(DNN)模型
<ol style="list-style-type: none">1. 利用隨機抽取 sample 跟 feature 建構許多決策樹2. 能找出每個特徵的重要性 (Feature Importance)	<ol style="list-style-type: none">1. 將資料投影至高維度處理原始空間無法處理的問題	<ol style="list-style-type: none">1. 利用多個非線性回歸方程式捕捉資料特性2. 易解決多維度問題

Training model **without bootstrapping**



	Random Forest	SVM	DNN
Testing set Accuracy	89.85%	89.76%	90.27%
Recall Rate	6.32%	8.56%	22.22%

Bootstrapping to **preprocess the imbalanced data**










	Random Forest	SVM	DNN
Testing set Accuracy	74.23%	82.72%	68%
Recall Rate	65.84%	51.10%	70%

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

Machine Learning Model Training Result



	Behavioral	Personal	All data
<i>Random Forest</i> Accuracy	74.23% 	58.95%	73.77%
<i>Random Forest</i> Recall Rate	65.84% 	59.33%	44.49%
SVM Accuracy	82.72%	85.51% 	82.07%
SVM Recall Rate	51.10% 	1.62% 	49.97%
<i>DNN</i> Accuracy	68% 	55%	61%
<i>DNN</i> Recall Rate	70%	64%	77% 



專案結果分析

- **Bootstrapping** 去平衡原始資料比例, 能更準確的訓練捕獲再購者的模型。
- **Behavior Data** 行為資料對再購預測的影響比 **Personal data** 個人資料 更大。
- 藉由**調整再購定義**(時間區隔調整), 可以抓到更多有效的再購資料進行訓練。
- 從**Random Forest**發現:

客戶年收入、客戶年齡、客戶戶齡和**理賠金額大小**對再購意願有較大的影響。

- 此三模型可以抓出**平均6成以上**的再購客戶, 且表現最好**DNN**能抓到**高達7成**願意再購的客戶。



Thank you for listening