新光人壽 預測潛在風險業務員

組員:郭許謙、洪翊婕、曾煒哲、張肇峰

新光金控 | 新光人壽 Shin Kong Life

報告大綱

新光金控 | 新光人壽 Shin Kong Life

01 研究目的

詐欺業務員帶給公司的負面影響

2016到2019 年, 有爭議的保費多達2663萬

金管會和處罰鍰 300 萬元(風險業務)

海外投資上限從 43% 降至 39%

公司聲譽(市占率) = **巨大損失**

估約 4000 萬元

模型開發目標

提升
公司對風險業務
的掌握

提升
風險業務審查工
作的效率

提升
風險業務預測的
精准度

降低成本

4 資料來源: Anue 鉅亨新聞

新光金控 | 新光人壽 Shin Kong Life

模型發展目標

Confusion Matrix

	預測 正常業務	預測 風險業務
真 正常業務	3981 人 TN	0 人 FN
真 風險業務	0 人 FP	38 人 TP

公司對風險業務的掌握

EDA 資料探索/分析

風險業務審查工作的效率

Precision TP / (FN + TP)

風險業務預測的精准度

Recall TP / (TP + FP)

新光金控 | 新光人壽 Shin Kong Life

02 問題與發現

資料型態、規模

資料總覽

- 4019 x 66 筆資料
- 類別型變數 : 12 項
- 連續型變數 : 54 項

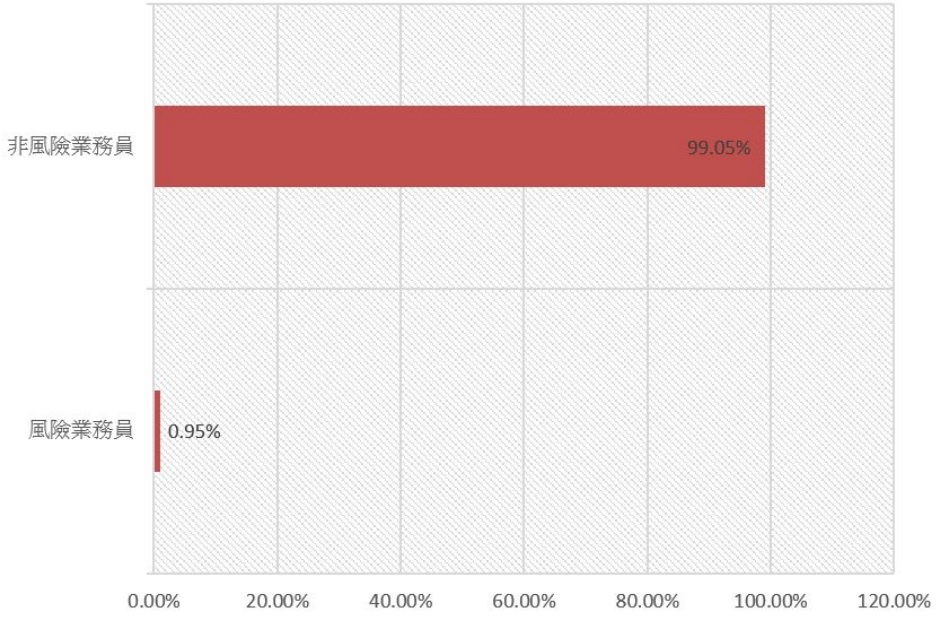
風險業務員 : 38 個

風險比例僅佔0.95%

非風險業務員 : 3981 個

Abnormal Target

資料不平衡



新光金控 | 新光人壽 Shin Kong Life

資料缺失

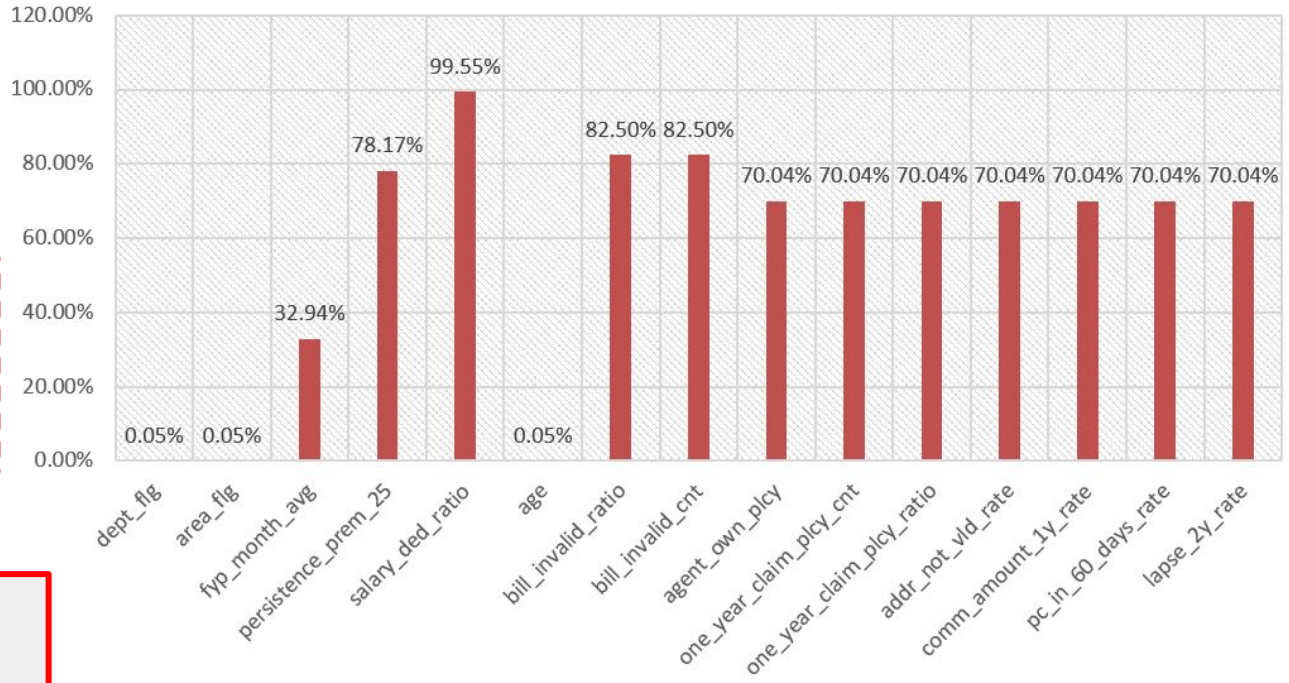
缺失型態

類別型缺失 : 2 筆
連續型缺失: 13 筆

缺失率

<30% : 2筆
30%~50% : 1 筆
>50% : 11筆

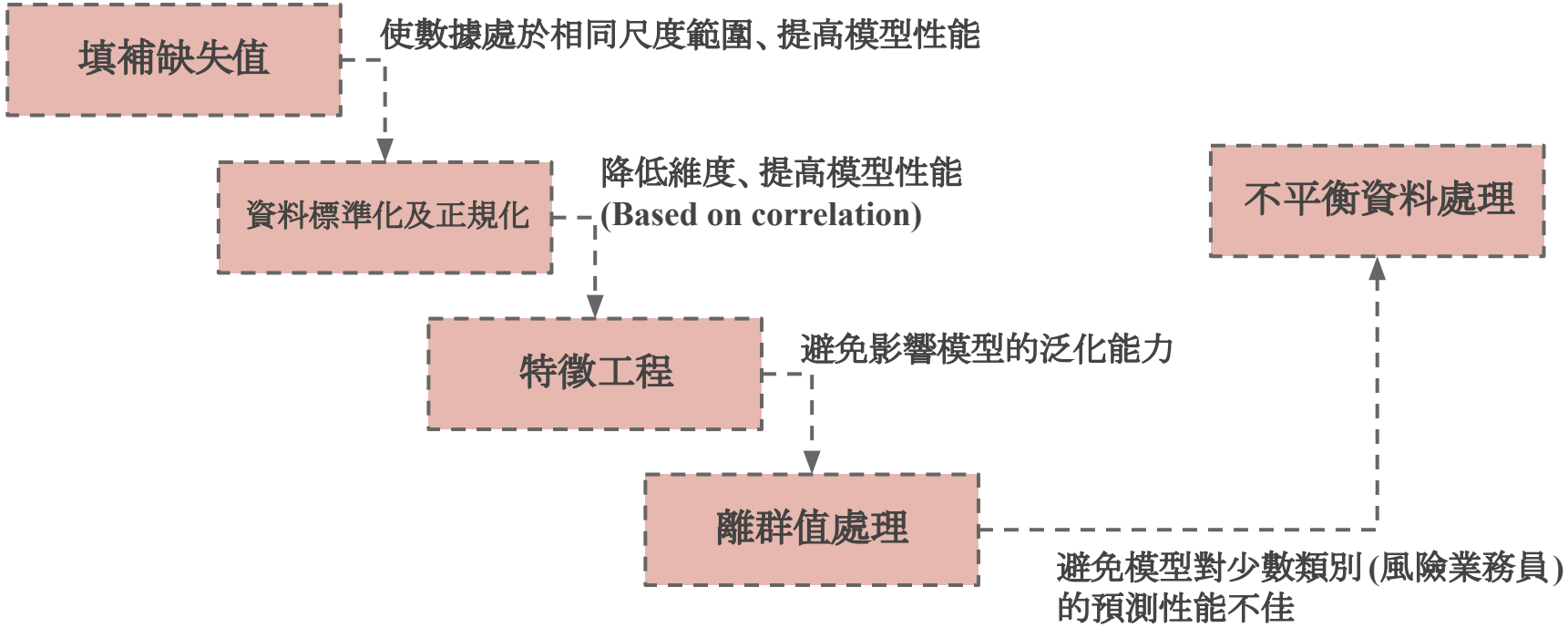
不完整資料



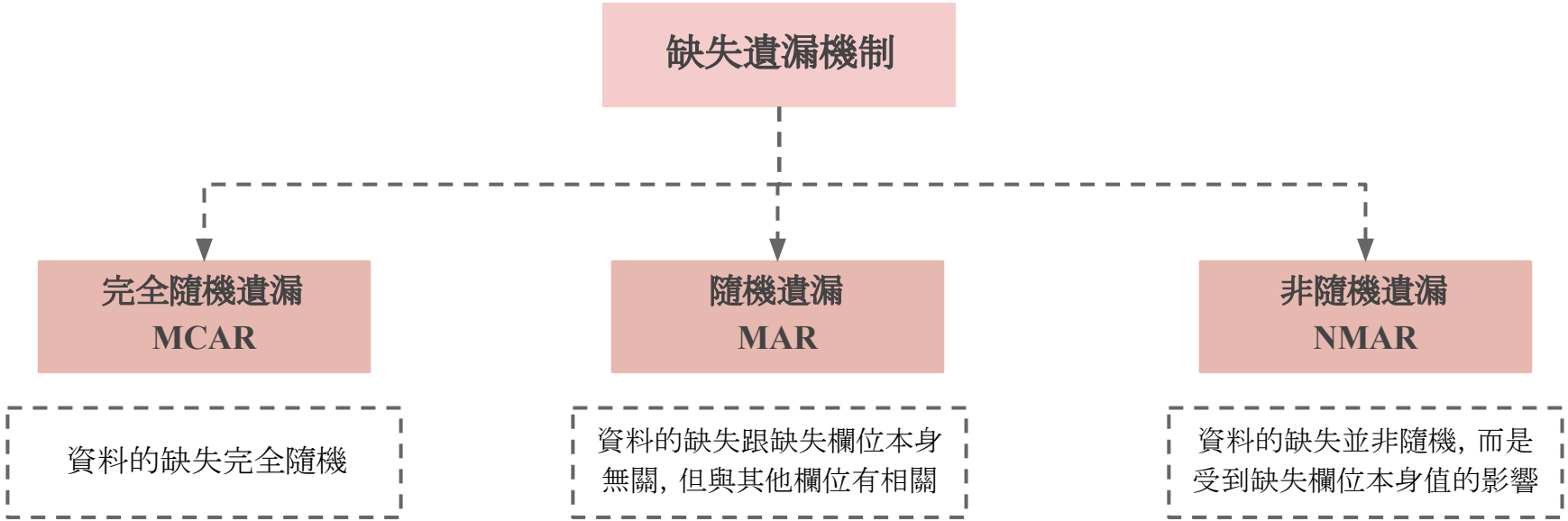
新光金控 | 新光人壽 Shin Kong Life

03 資料處理

資料處理流程



缺失值探討



Filling NA | Standardization | Feature Engineering | Outliers

 資料來源：蔡志豐教授 資料分析之遺漏 值處理技術探討

缺失值填補

填補流程

1.缺失>80%: 直接刪除

2.缺失<80%: 填補缺失值

3. 決定缺失資料型態

月平均業績/fyp_month_avg	MAR
25個月繼續率/PERSISTENCE_PREM_25	MAR
保單相關	MCAR

AGENT_STATUS —Random Forest→ FYP_MONTH_AVG

業務員自有保單數 —隨機抽樣→ 25個月繼續率

保單相關

業務員自有保單數
保單生效後1年內申請賠償請求的數量
保單生效後1年內申請賠償數量佔比
郵寄地址與居住地址不相符張數佔比
近一年佣金收入與近三年佣金收入比
保單生效日後60天內保單變更數佔比
於理賠後兩年內保障停效之佔率

業務員透漏意願不高 —利用現有資料→ 業務員均值填補

Filling NA → Standardization → Feature Engineering → Outliers

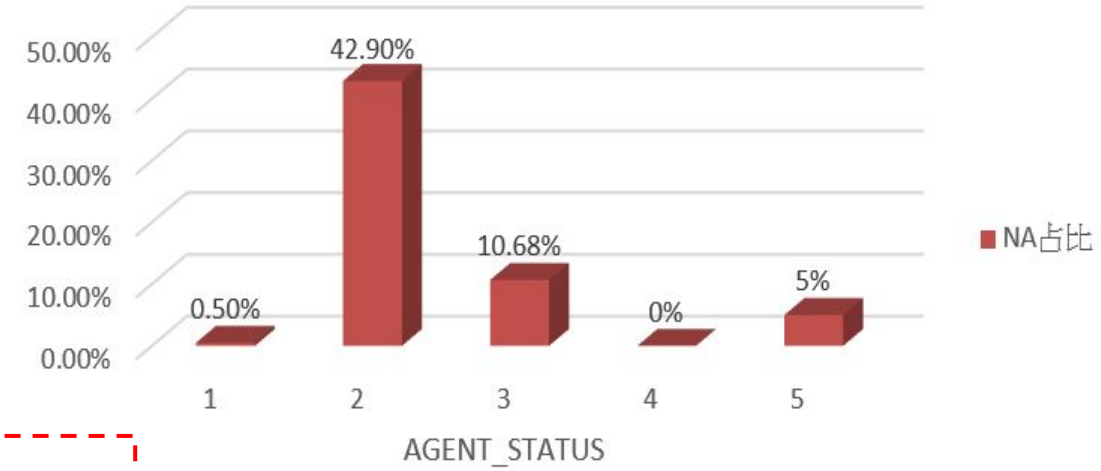
新光金控 | 新光人壽 Shin Kong Life

12 資料來源: Ryan Lu (Medium)

缺失值填補

月平均業績/fyp_month_avg

發現	月平均薪資為缺漏值的比例中，有42.9%的業務員agent_status=2
推論	月平均薪資的缺失與agent_status有一定關係，故判斷其為MAR



根據agent_status預測月平均薪資的缺失值

Filling NA | Standardization | Feature Engineering | Outliers

新光金控 | 新光人壽 Shin Kong Life

缺失值一覽

變數名稱	dept_flg	area_flg	fyp_month_avg	persistence_prem_25	salary_ded_ratio	age	bill_invalid_ratio
變數型態	類別型	類別型	連續型	連續型	連續型	連續型	連續型
缺失種類	MCAR	MCAR	MAR	MAR	MCAR	MCAR	MCAR
缺失率	0.05%	0.05%	32.94%	78.17%	99.55%	0.05%	82.50%
填補方法	眾數填補	眾數填補	random forest	隨機抽樣	刪除變數	平均值填補	直接刪除

缺失值一覽

bill_invalid_cnt	agent_own_plcy	one_year_claim_plcy_cnt	one_year_claim_plcy_ratio	addr_not_vld_rate	comm_amount_1y_rate	pc_in_60_days_rate	lapse_2y_rate
連續型	連續型	連續型	連續型	連續型	連續型	連續型	連續型
MCAR	MCAR	MCAR	MCAR	MCAR	MCAR	MCAR	MCAR
82.50%	70.04%	70.04%	70.04%	70.04%	70.04%	70.04%	70.04%
直接刪除	平均值填補	平均值填補	平均值填補	平均值填補	平均值填補	平均值填補	平均值填補

正規化及標準化之方法

Why Normalization ?

1.減少類別數量、模型複雜度

2.增加模型的表達能力

Label Encoding

1.二元變數轉為0/1
2.單位代號等去除文字, 僅留下數字部分
3.agent_level依據字母順序作編號

Agent_level	Agent_level_encoded
Z1	33
KD	29

Why Standardization ?

1.避免尺度範圍差異

2.提高模型收斂速度

3.避免梯度消失或爆炸

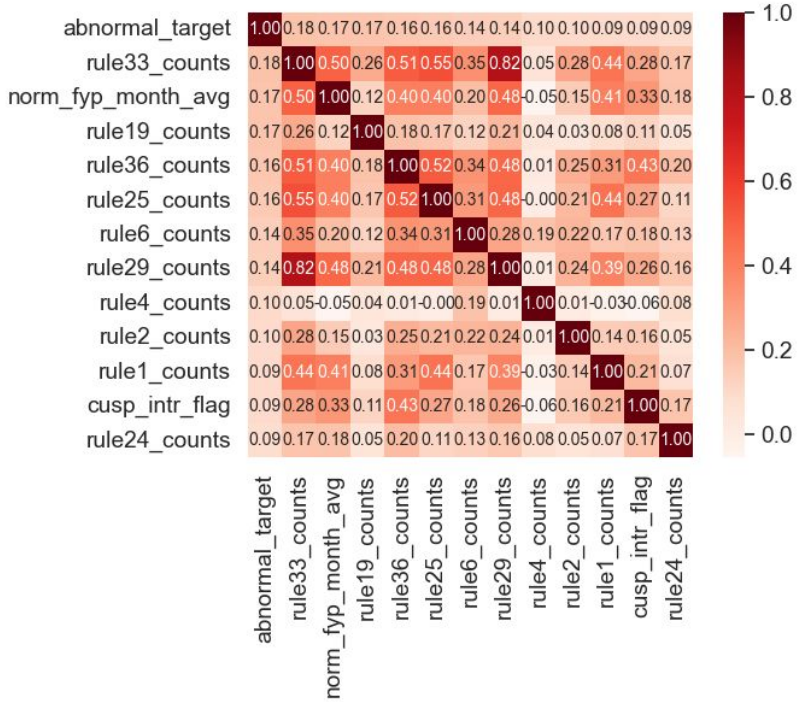
Stardand Scaler

$$(x - mean) / std$$

平均值=0
標準差=1

fyp_month_avg	fyp_month_avg_norm
46969.37	0.191016
32539.06	-0.220258

相關係數分析



針對RULES_COUNT

相關係數 >0.09

因為變數本身與目標變數的相關性夠高，我們認為將他們留在模型裡對其有正向影響

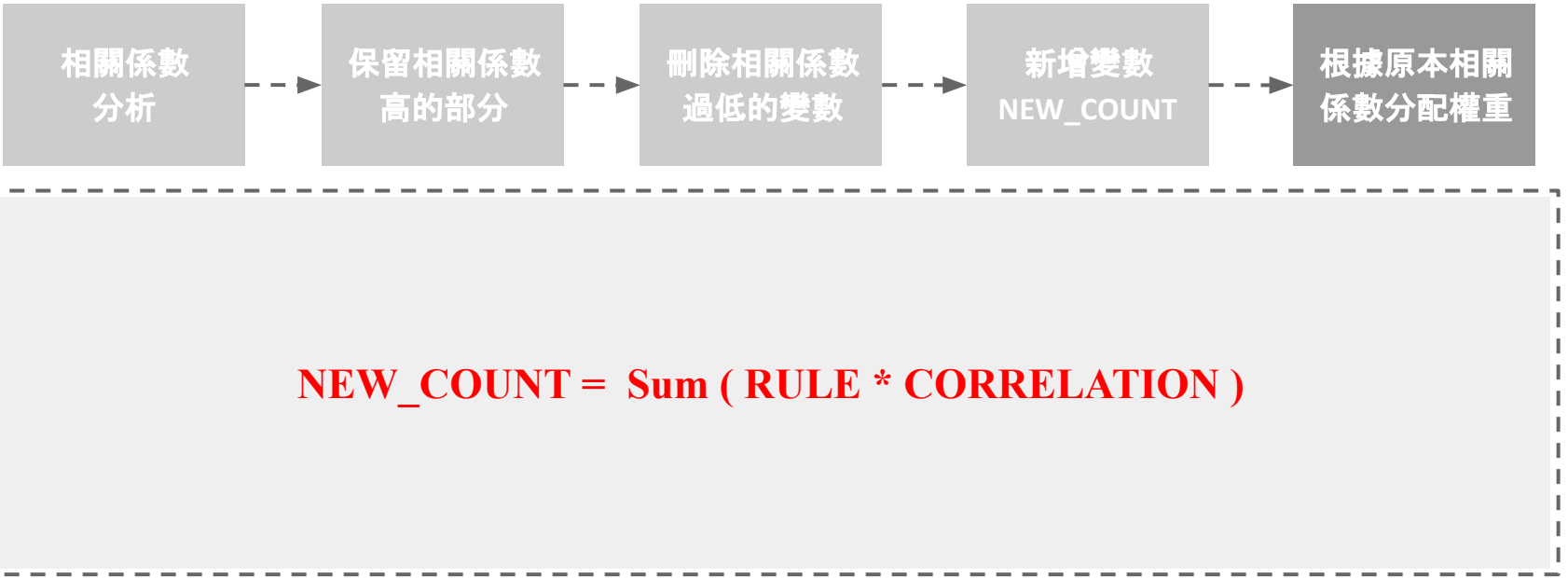
相關係數 <0.09

變數與目標變數相關性低，因此新增一個變數把他們放在一起，並根據相關係數分配權重

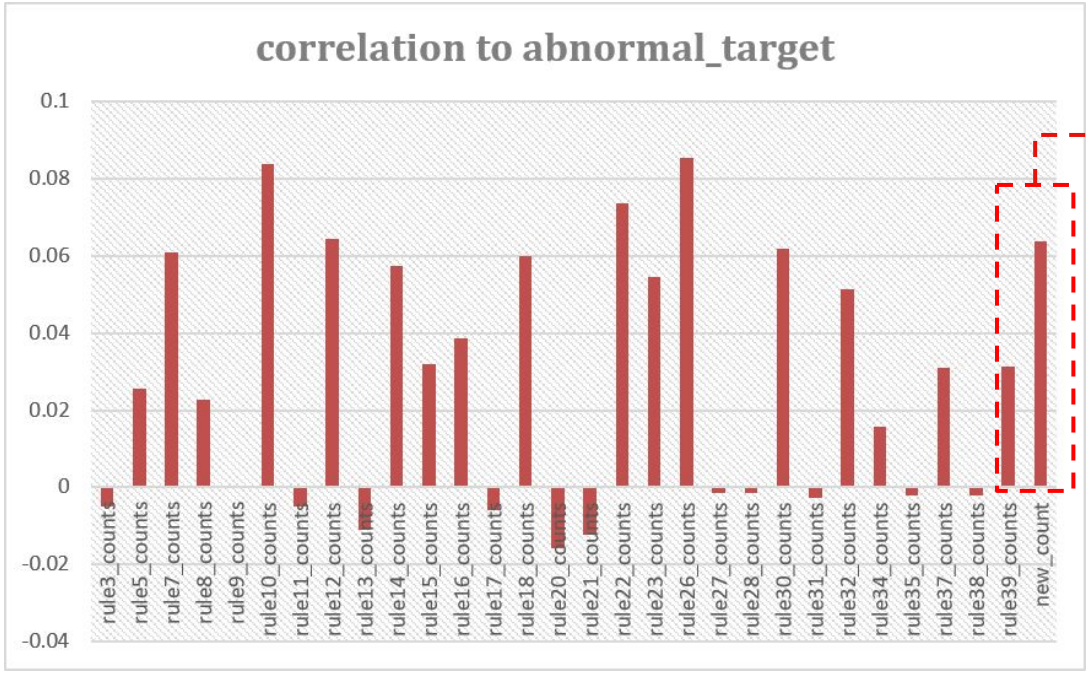
Filling NA → Standardization → Feature Engineering → Outliers

新光金控 | 新光人壽 Shin Kong Life

新增變數 NEW_COUNT



新增變數效果評估



新增的new_count相關係數達0.06以上，已高於多數rules的相關係數，因此我們判斷新增此變數將為模型帶來正面影響。

Filling NA | Standardization | Feature Engineering | Outliers

特徵縮減

細節探討

1. 變數之間相關性過高，兩者擇一

xagfd_flag & rule15_counts

cusp_intr_flag & rule26_counts

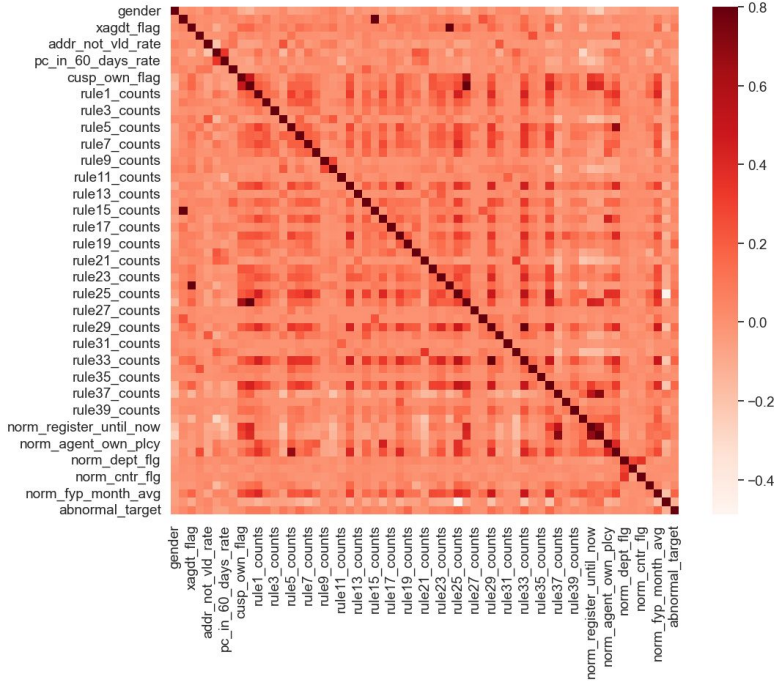
2.對目標變數相關係數過低，刪除變數

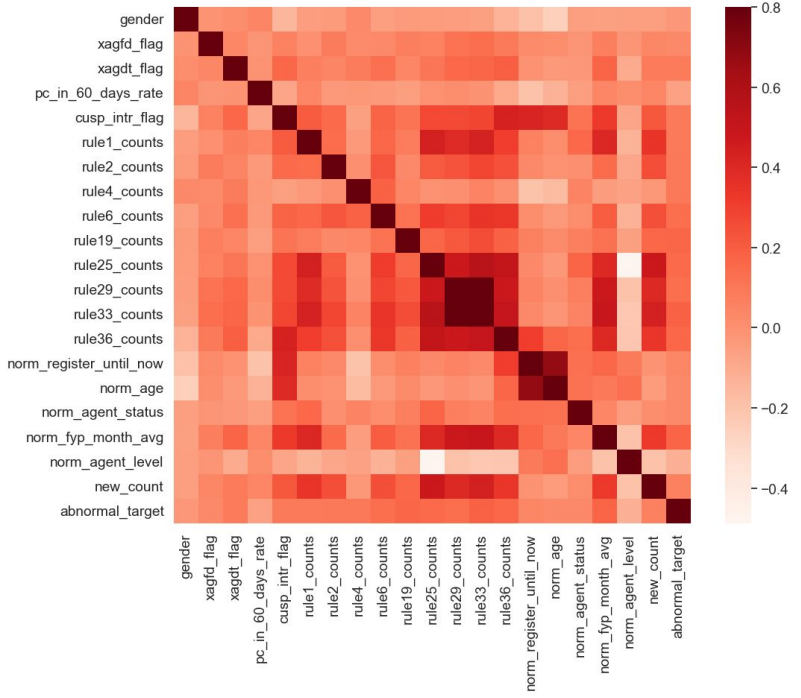
"one_year_claim_plcy_ratio","addr_not_vld_rate","cusp_own_flag","comm_amount_1y_rate等

Filling NA | Standardization | **Feature Engineering** | Outliers

新光金控 | 新光人壽 Shin Kong Life

特徵工程前後對比





Filling NA → Standardization → **Feature Engineering** → Outliers

新光金控 | 新光人壽 Shin Kong Life

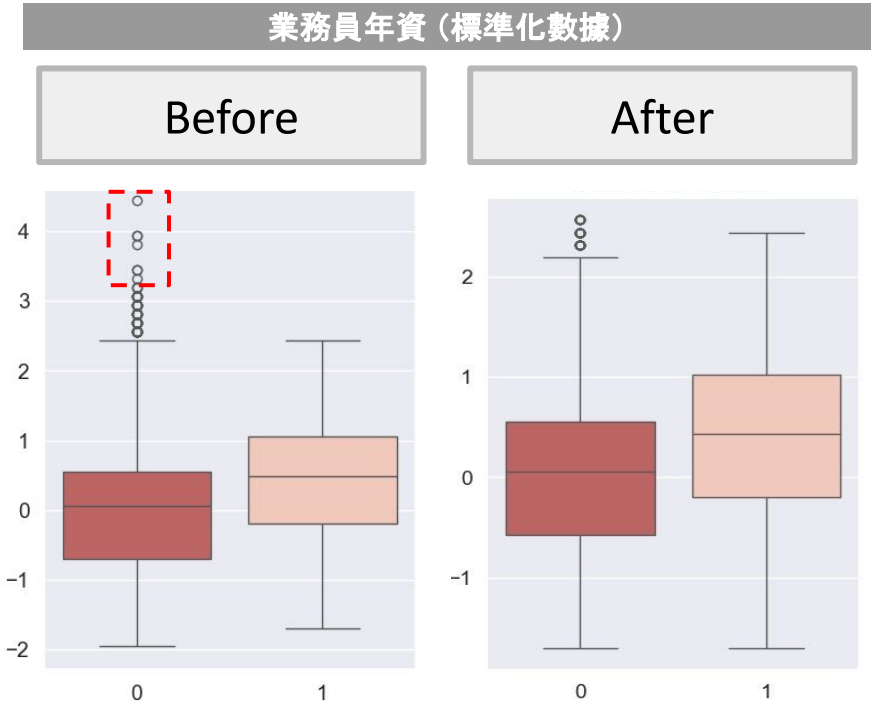
處理Outliers (IQR)

- norm_fyp_month
- norm_register_until_now
- new_count
- norm_age

針對非風險業務員進行Outliers移除

參數調整及目標

目標	使模型更加準確地捕捉到數據的特徵和規律，從而提高模型的準確性和可靠性。
IQR計算	提高模型對於風險業務員樣本的檢測能力，利用其資料計算變數的Q1~Q4以及IQR



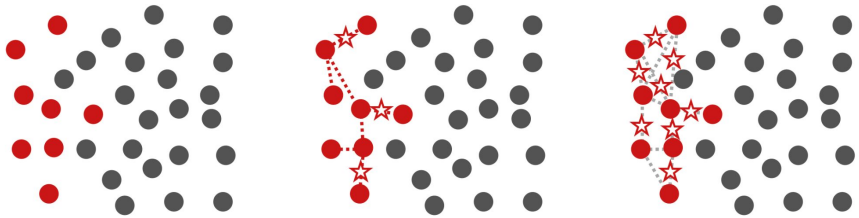
Filling NA → Standardization → Feature Engineering → Outliers

新光金控 | 新光人壽 Shin Kong Life

04 模型介紹

SMOTE 合成少數過採樣法

原理



$$x_{new} = x_{chosen} + (x_{nearest} - x_{chosen}) * \delta\,;\ \delta \in [0,1]$$

1. 設定採樣倍率 N, 也就是對每個樣本需要生成幾個合成樣本
2. 設定鄰近值 K, 從 K 個最近樣本中挑一個
3. 根據以上公式, 創建 N 個樣本

- **解決不平衡資料**
- **獲得大量資料有助於發展神經網路**

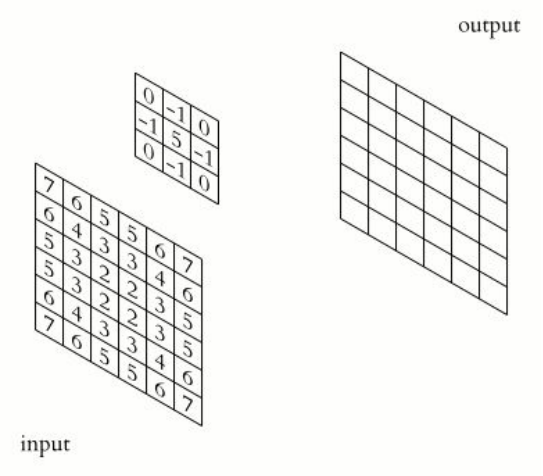
風險評估

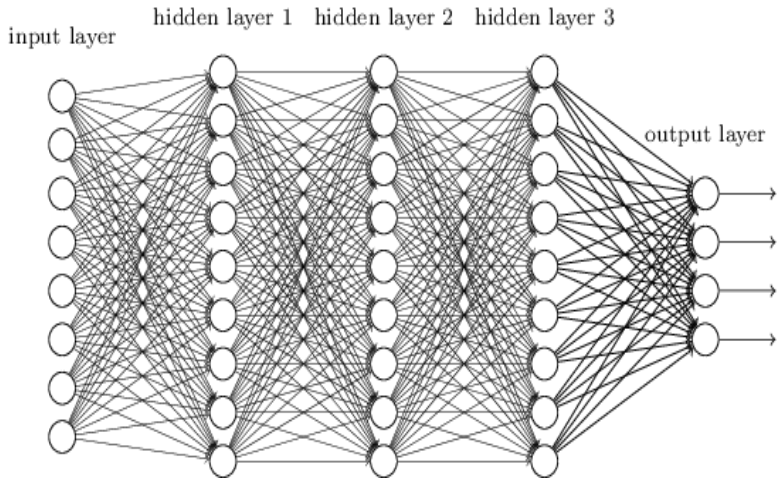
Over Fitting	SMOTE會引入一定程度的噪音, 過度依賴合成樣本可能會導致過度擬合。
降低解釋性	特徵重要性失真, 分類模型的解釋性大幅降低。

資料來源: Edward Tung (Medium) 、David's Perspective

CNN 卷積神經網路

原理





Why CNN ?

相比 XGB/SVM/LR/Stacking/ Blending/DNN, 是我們這組試過表象相對突出、相對好掌握的演算法。

雖然常見的 CNN 模型都是在處理次序行的資料較多也較傑出, 不過一維卷積其實也有特徵抓取的功效。

SMOTE | CNN | MODEL TUNING

 資料來源: Cinnamon AI Taiwan (Medium) 、Cupoy AI 共學社群、Analog Devices

Model Tuning 模型參數調整

```
for batch_size in range(600, 2800, 200):
    for epoch in range(5, 35, 5):
        CNN_MODEL(
            Dropout(0.2)
CNN_MODEL.fit(validation_split = 0.3))
```

參數調整目標

1. 避免 Over fitting
2. 取捨 Recall 和 Precision
3. 穩定 Model

參數說明

小 ⟷ 大

只篩到局部特徵	**batch_size**	梯度下降不太穩定
預測不精準	**epoch**	過度配適

選擇方法

Recall_mean	Precision_mean	Recall_std	batch_size	epoch
0.8285	0.0730	0.0857	1200	10
0.7857	0.1208	0.1317	2000	35

SMOTE → CNN → MODEL TUNING

 資料來源: Cynthia (HackMD)

Model A 和 Model B 間的取捨

提升風險業務審查的工作效率

		Model A	Model B
Model A / Model B	Precision	7.3 %	12 %
Model A / Model B	Recall	82.9 %	78.6 %

資料解讀（假設實際有 40 個風險業務）

Model A 預測 454 個風險業務，其中有33個是真的

Model B 預測 262 個風險業務，其中有31個是真的

模型穩定性 (假設模型Output 是常態分佈)

		Model A	Model B
Model A / Model B	Recall_std	8.6 %	13.2 %

Model A 預測 407 ~ 501，有30 ~ 37 個是真的

有 70 % 的機率

Model B 預測 218 ~ 306，有26 ~ 37 個是真的

結論

風險業務員的罰款**遠高於**增加的工作效益

A

模型的本質是打造穩定又有效率的工具

 資料來源：自行分析

05 結論

檢視模型成效及研究目的

降低成本

增加對風險業務員的掌握

1. 建議在相關係數高的變特徵上少一點缺值。
2. 因使用 CNN 所以對特徵解釋能力較有限。

提升風險業務審查工作的效率

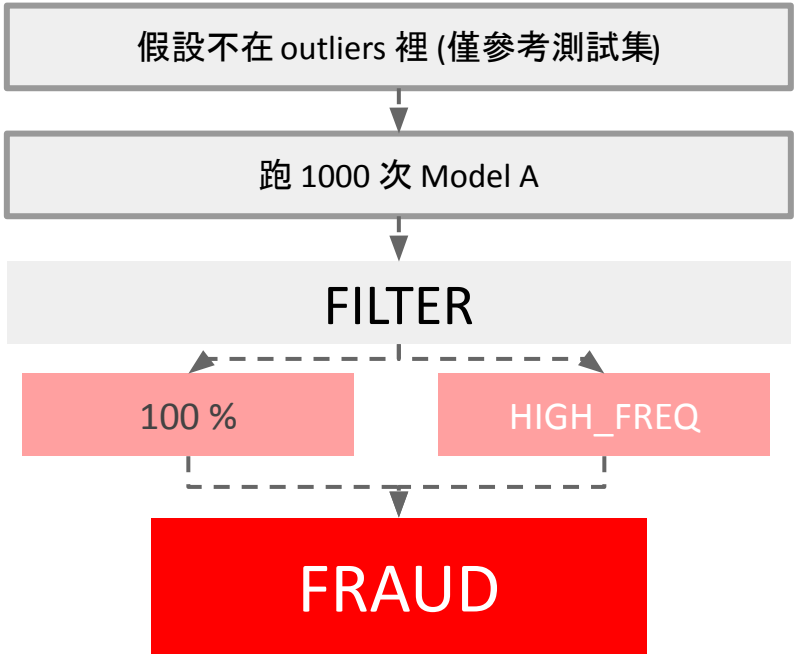
1. 雖然 Precision的表現不算太好，不過我們認為把 Precision作為 Recall 是非常值得的。

提升風險業務預測的精准度

1. 基本上這就是我們模型的主軸，想盡辦法將所有的風險業務員貼上標籤。

預測潛在風險業務員

預測流程



預測結果

agnt_0233	agnt_0289	agnt_0322	agnt_0337
agnt_0440	agnt_1167	agnt_1230	agnt_1432
agnt_1559	agnt_1820	agnt_1892	agnt_2023
agnt_2171	agnt_2219	agnt_2316	agnt_2620
agnt_2641	agnt_2672	agnt_2769	agnt_2829
agnt_2905	agnt_2966	agnt_3036	agnt_3251
agnt_3295	agnt_3321	agnt_3350	agnt_3371
agnt_3394	agnt_3504	agnt_3614	agnt_3625
agnt_3821	agnt_1710		

新光金控 | 新光人壽 Shin Kong Life

06 QA

新光金控 新光人壽 Shin Kong Life