



機器學習與人工智慧個案實作

第一組：預測客戶是否會發生心臟疾病

組長 | 110303052 會計四 陳芊穎

組員 | 110304020 統計四 張采婕

110304005 統計四 卓姿潔

110508005 風管四 謝柔樺

110102061 風管四 謝宜蓁

AGENDA

01

定義商業問題

02

資料清洗

03

EDA

04

模型訓練與篩選

05

實際商業應用

06

附錄



01

定義商業問題

以如何建立識別心臟病高風險族群與健康分級系統為本次商業問題

專案題目

希望利用現有的顧客健康資料，預測顧客是否罹患心臟病，為公司提供風險評估工具

專案背景

壽險業接軌IFRS 17和 ICS後，國泰人壽將透過**搶攻健康險市場達成累積CSM**，為壽險公司穩定獲利的關鍵指標

壽險業新制上路

全面改為外溢保單

心臟病為主要理賠原因之一

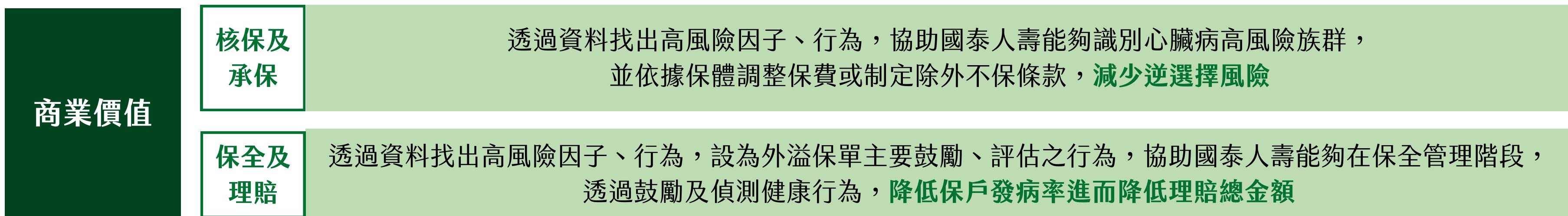
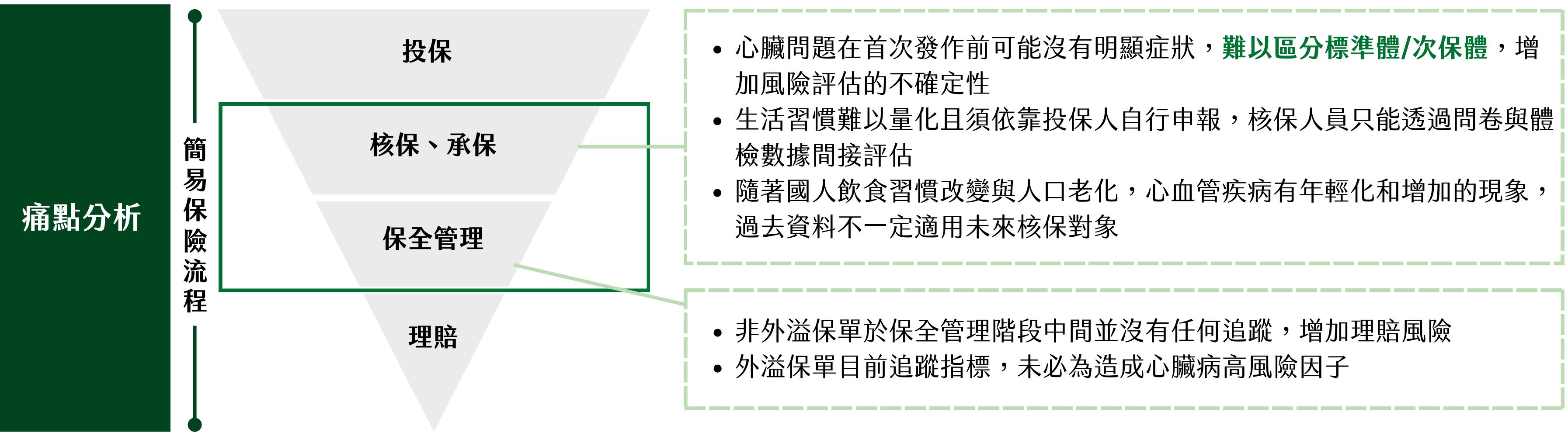
2025 年第二季後國泰人壽全面將健康險納入「外溢機制」

- 健康險中，**心臟病為主要理賠原因之一**，且為國人十大死因排名第二
- 2023年台灣壽險業中，**理賠人次心臟病佔14.2%居二**，且其理賠佔比近年有上升趨勢

小組定義之商業問題

如何運用健康資料與預測模型，建立心臟病高風險族群之識別與健康分級系統，協助解決國泰人壽核保、保全管理流程之痛點，並優化外溢保單設計？

協助國泰人壽減少核保時逆選擇風險，以及強化外溢保單優點降低理賠總額



透過選定之模型將進行分級，降低核保泥選擇及設計分級外溢保單內容

分析架構





02

資料清洗

資料背景

基本資料與生活習慣

State
Sex
AgeCategory
RaceEthnicityCategory
HeightInMeters
WeightInKilograms
BMI
SmokerStatus
ECigaretteUsage
AlcoholDrinkers
PhysicalActivities
SleepHours
LastCheckupTime

健康狀況與疾病病史

GeneralHealth
PhysicalHealthDays
MentalHealthDays
HadHeartAttack
HadAngina
HadStroke
HadAsthma
HadSkinCancer
HadCOPD
HadDepressiveDisorder
HadKidneyDisease
HadArthritis
HadDiabetes
RemovedTeeth

功能障礙與身體限制

DeafOrHardOfHearing
BlindOrVisionDifficulty
DifficultyConcentrating
DifficultyWalking
DifficultyDressingBathing
DifficultyErrands

疫苗接種與篩檢紀錄

HIVTesting
ChestScan
FluVaxLast12
PneumoVaxEver
TetanusLast10Tdap
HighRiskLastYear
CovidPos

資料清洗

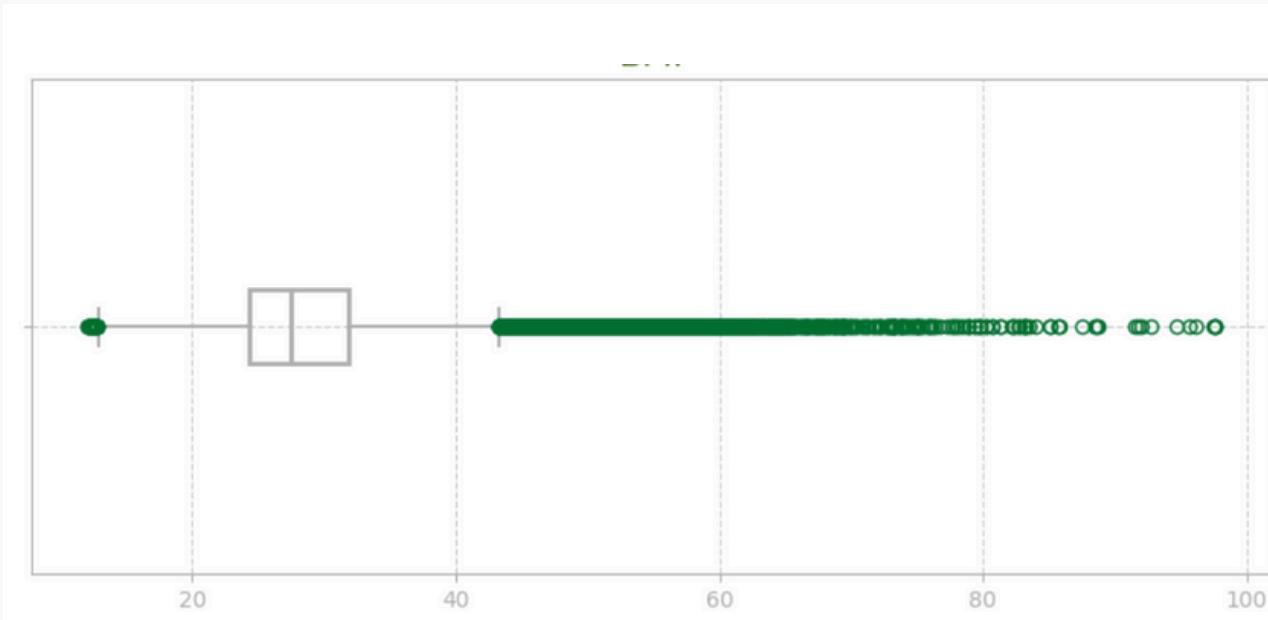
資料清洗

刪除空值

- 原資料筆數：445132
- 具有空值筆數：199110，約佔原資料筆數之45%
- 僅6欄位為數值型，其餘34欄位為類別型，補值困難
- 推斷刪除具空值之所有筆數後資料筆數仍足夠進行建模

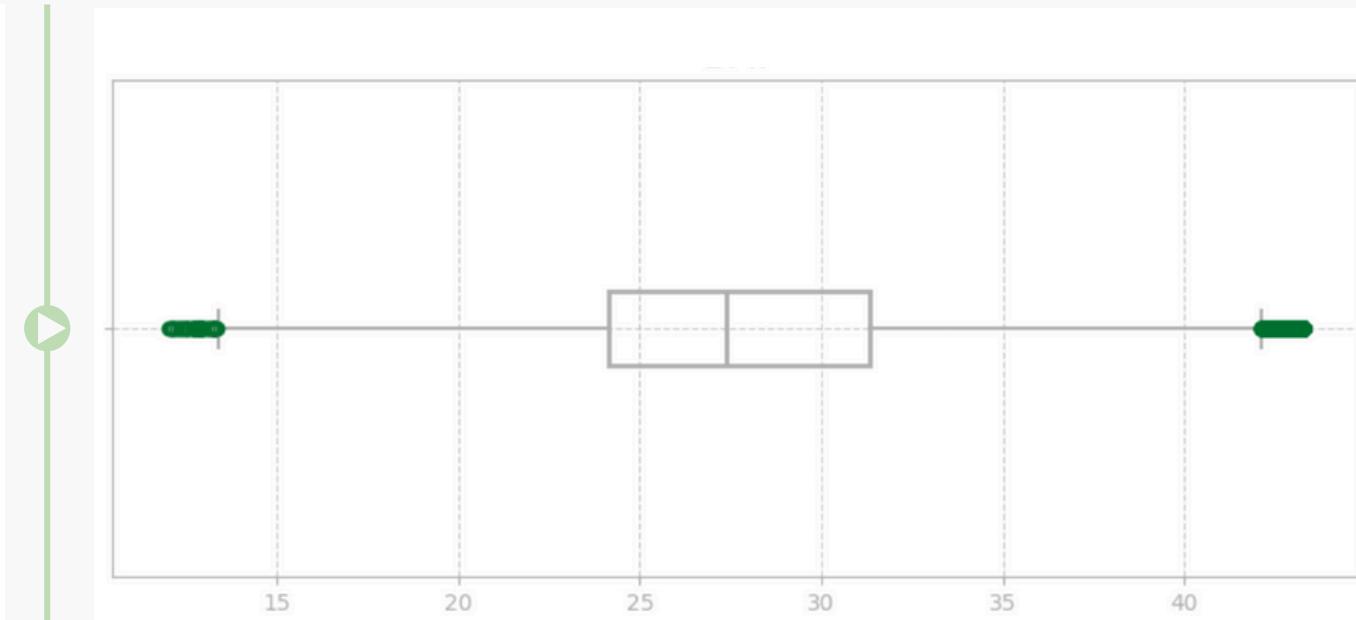
異常值處理

原 BMI 分佈



Q1:24.27 Q2:27.46 Q3:31.89 max=97.65
刪除大於Q3+1.5IQR之BMI

刪除異常值後之 BMI 分佈



Q1:24.14 Q2:27.37 Q3:31.32 max:43.28

資料清洗

資料清洗

連續型
&
類別型

- 針對連續型變數進行標準化，避免因單位差異影響模型權重學習（平均數為 0，標準差為 1）
- 對具順序意義的類別採用 Ordinal Encoding；對無順序關係的類別則使用 One-Hot Encoding

標準化

數值型變數
PhysicalHealthDays
MentalHealthDays
⋮

Ordinal Encoding

GeneralHealth
AgeCategory

One-hot Encoding

其他類別型變數
LastCheckupTime
SmokerStatus
⋮

GeneralHealth	PhysicalHealthDays	MentalHealthDays	SleepHours	HadHeartAttack	AgeCategory	HeightInMeters	WeightInKilograms	BMI	State_Alabama	...	PneumoVaxEver_Yes
0	3.142932	0.743778	0.676470	0	11	-1.003140	0.464867	1.369731	0	...	1
2	-0.481146	-0.507245	0.676470	0	2	-1.475038	0.176685	1.386395	0	...	1
3	-0.360343	-0.507245	0.676470	0	4	1.167590	0.464867	-0.168907	0	...	0
3	-0.481146	-0.257041	-0.021244	0	12	-1.286279	-1.217077	-0.787324	0	...	1
2	-0.239541	-0.507245	0.676470	0	9	0.412553	0.704842	0.551346	0	...	1

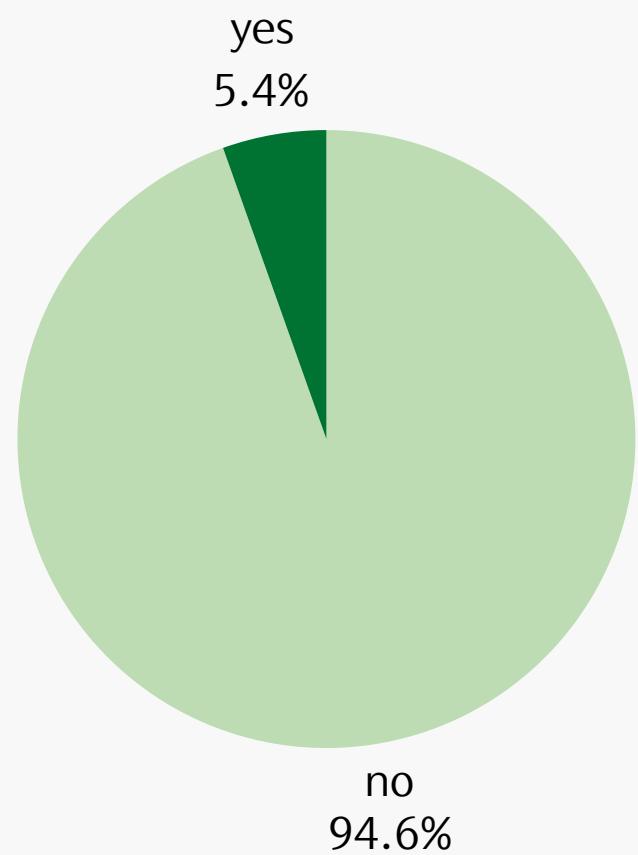


03

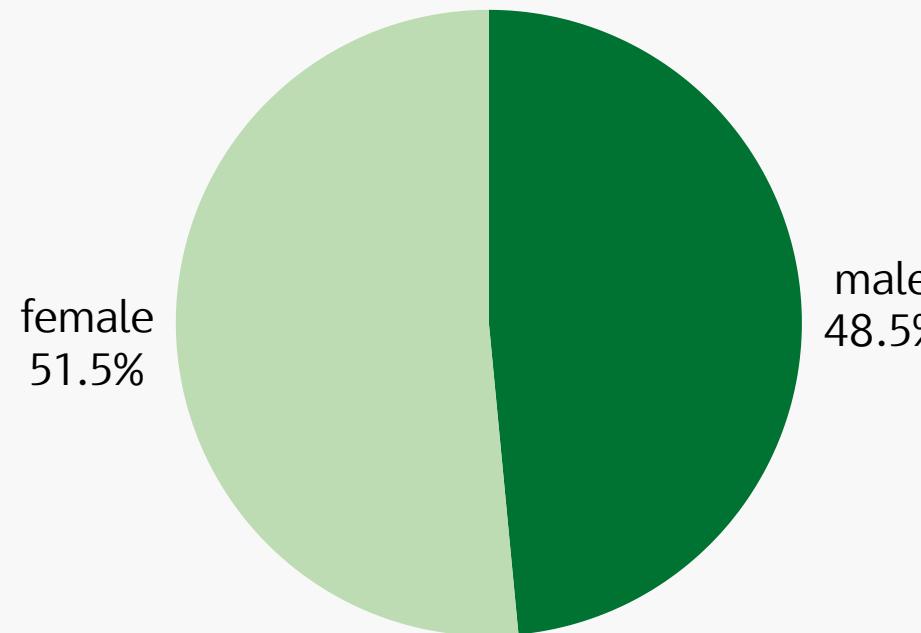
EDA

資料背景

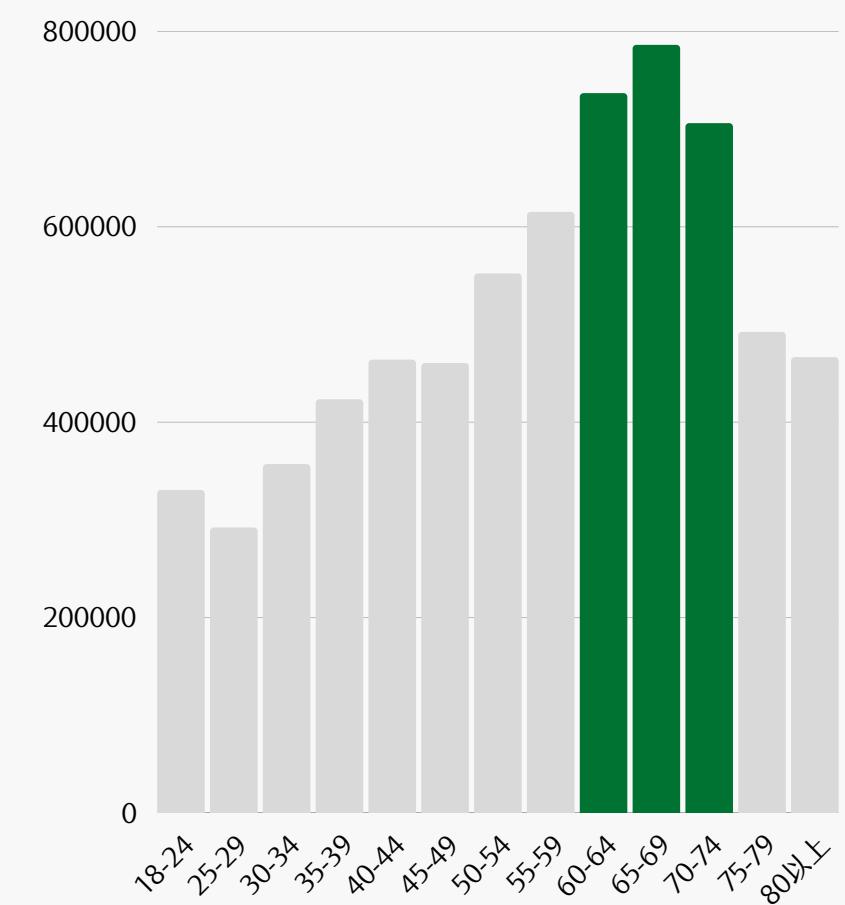
是否有心臟病



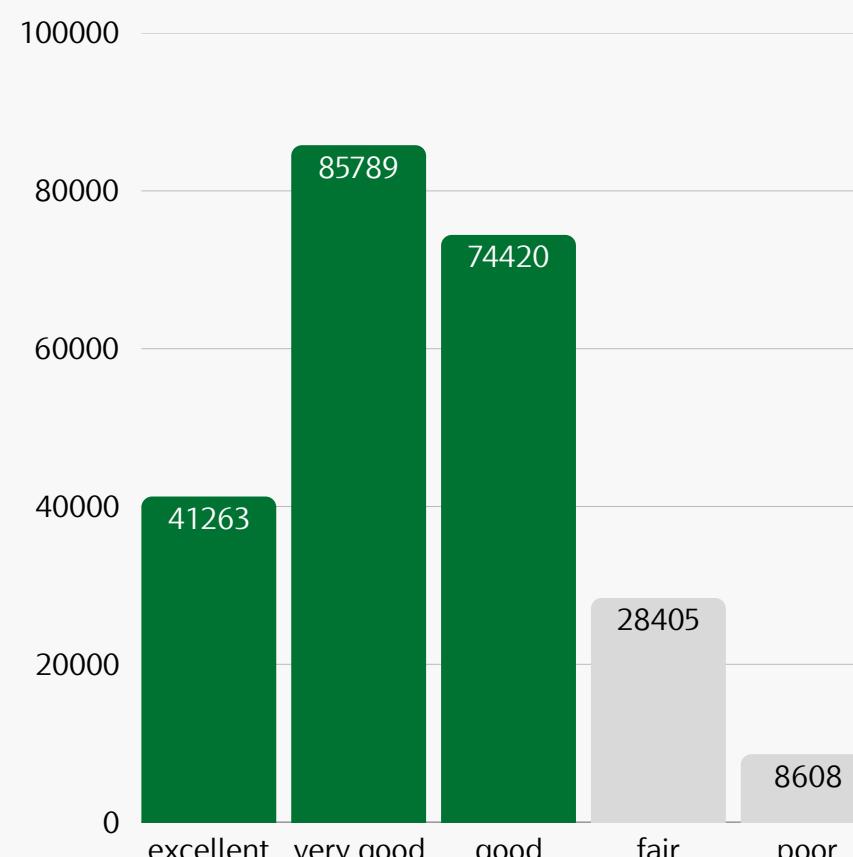
性別



年齡



自述健康狀況



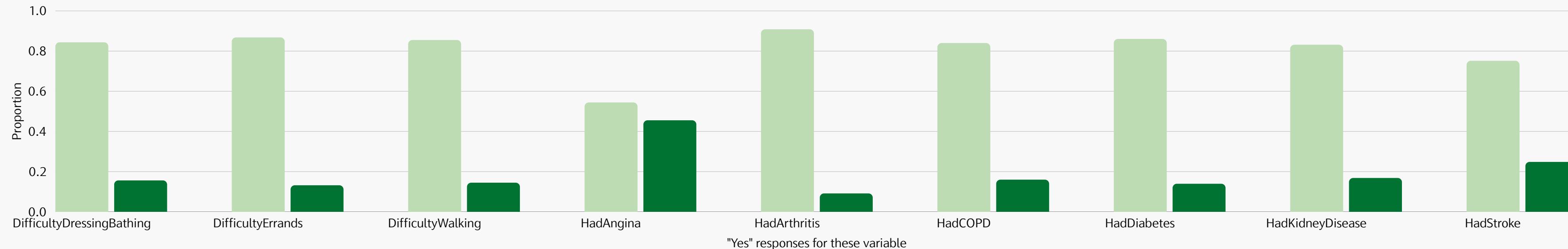
沒有心臟病人數佔94.6%，
後續需處理資料不平衡之問題

男女比相近

60-74歲人數最多

八成以上認為自身健康
狀況良好

資料背景



"Yes" responses for these variable



"No" responses for these variable

上圖為具有這些疾病或功能障礙的人群，下圖則為未具備這些病史或障礙的人群。

淺綠為未曾心臟病發之比例，深綠為有心臟病發之比例。

可以明顯觀察到，在有這些健康問題的族群中，心臟病的發生比例明顯高於沒有的人群，凸顯了這些變數的重要性，後續會用特徵篩選再做確認。



04

模型訓練與篩選

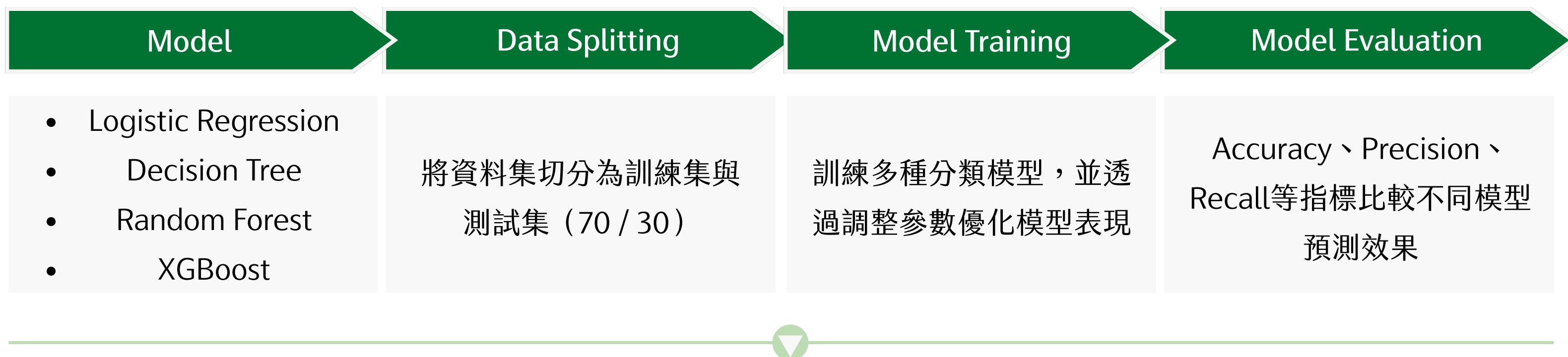
模型選擇與訓練過程

選擇羅吉斯回歸/決策樹/隨機森林/XGBoost作為我們訓練與評估的模型

模型選擇

我們的目標是希望能夠協助國泰「預測顧客是否罹患心臟病」，屬於 **分類（Classification）問題**

建模流程



選擇Recall分數為主要判斷標準

保險預測場景中，錯判一個高風險族群為低風險（FN）是高代價，因此為避免 **False Negative（漏判）**，
我們傾向使用漏判低的模型，並透過最後一層「人工核保」方式過濾

模型篩選

根據最終的模型指標，我們選擇XGBoost作為主要模型，後續將詳細說明過程

	處理	ROC-AUC	Precision	Recall	F1-score	Accuracy
Logistic Regression	<ul style="list-style-type: none">刪除 HeightInMetersWeightInKilograms• GridSearch• Lasso	0.89	0.21	0.77	0.33	0.89
Decision Tree	<ul style="list-style-type: none">• Random oversample• RandomizedSearchCV• class_weight='balanced'	0.79	0.19	0.75	0.31	0.81
Random Forest	<ul style="list-style-type: none">• Class Weight	0.81	0.4	0.55	0.46	0.93
XGBoost	<ul style="list-style-type: none">• Class Weight• GridSearch• Feature Importance+SHAP• Voting (soft)	0.89	0.3	0.67	0.42	0.90

Step1：Baseline與不平衡處理

進行Baseline後，我們發現模型效果不佳待優化

Classification_Report

✖ Baseline Classification Report				
	precision	recall	f1-score	support
0	0.9575	0.9894	0.9732	67661
1	0.5613	0.2355	0.3318	3885
accuracy			0.9485	71546
macro avg	0.7594	0.6125	0.6525	71546
weighted avg	0.9360	0.9485	0.9384	71546

Confusion Matrix

66,946	715
2,970	915

Train / Test 指標

Train	Test
ROC-AUC:0.93	ROC-AUC:0.88
Recall:0.36	Recall:0.23
Precision:0.79	Precision:0.56
Accuracy:0.96	Accuracy:0.94

模型效果不佳 可能有overfit問題

Over Sampling

處理方法 : Random Over Sampling

precision recall f1-score support

0	0.9818	0.8644	0.9194	67661
1	0.2339	0.7210	0.3532	3885

accuracy			0.8566	71546
macro avg	0.6079	0.7927	0.6363	71546
weighted avg	0.9412	0.8566	0.8886	71546

F1 score: 0.3532
ROC-AUC: 0.8790

Under Sampling

處理方法 : Random Under Sampling

precision recall f1-score support

0	0.9854	0.7995	0.8828	67661
1	0.1853	0.7941	0.3004	3885

accuracy			0.7992	71546
macro avg	0.5853	0.7968	0.5916	71546
weighted avg	0.9420	0.7992	0.8511	71546

F1 score: 0.3004
ROC-AUC: 0.8817

SMOTE

處理方法 : SMOTE

precision recall f1-score support

0	0.9604	0.9773	0.9688	67661
1	0.4294	0.2973	0.3513	3885

accuracy			0.9404	71546
macro avg	0.6949	0.6373	0.6600	71546
weighted avg	0.9315	0.9404	0.9352	71546

F1 score: 0.3513
ROC-AUC: 0.8557

Class Weight

處理方法 : Cost-sensitive (scale_pos_weight)

precision recall f1-score support

0	0.9821	0.8628	0.9186	67661
1	0.2332	0.7266	0.3531	3885

accuracy			0.8554	71546
macro avg	0.6077	0.7947	0.6359	71546
weighted avg	0.9415	0.8554	0.8879	71546

F1 score: 0.3531
ROC-AUC: 0.8782

0.2332 0.7266

Precision&Recall差距過大
且Precision明顯過低

Step1 : Baseline與不平衡處理

在嘗試多種不平衡處理方式後，最終選擇使用 Class Weight

Classification_Report					Confusion Matrix			Train / Test 指標			
								Train		Test	
✗ Baseline Classification Report					66,946	715					
	precision	recall	f1-score	support				ROC-AUC:0.93	ROC-AUC:0.88	Recall:0.36	Recall:0.23
0	0.9575	0.9894	0.9732	67661				Precision:0.79	Precision:0.56	Accuracy:0.96	Accuracy:0.94
1	0.5613	0.2355	0.3318	3885							
accuracy			0.9485	71546							
macro avg	0.7594	0.6125	0.6525	71546	2,970		915				
weighted avg	0.9360	0.9485	0.9384	71546							
模型效果不佳 可能有overfit問題											

Over Sampling					Under Sampling					SMOTE					Class Weight				
處理方法: Random Over Sampling					處理方法: Random Under Sampling					處理方法: SMOTE					處理方法: Cost-sensitive (scale_pos_weight)				
	precision	recall	f1-score	support		precision	recall	f1-score	support		precision	recall	f1-score	support		precision	recall	f1-score	support
0	0.9818	0.8644	0.9194	67661	0	0.9854	0.7995	0.8828	67661	0	0.9604	0.9773	0.9688	67661	0	0.9821	0.8628	0.9186	67661
1	0.2339	0.7210	0.3532	3885	1	0.1853	0.7941	0.3004	3885	1	0.4294	0.2973	0.3513	3885	1	0.2332	0.7266	0.3531	3885
accuracy			0.8566	71546	accuracy			0.7992	71546	accuracy			0.9404	71546	accuracy			0.8554	71546
macro avg	0.6079	0.7927	0.6363	71546	macro avg	0.5853	0.7968	0.5916	71546	macro avg	0.6949	0.6373	0.6600	71546	macro avg	0.6077	0.7947	0.6359	71546
weighted avg	0.9412	0.8566	0.8886	71546	weighted avg	0.9420	0.7992	0.8511	71546	weighted avg	0.9315	0.9404	0.9352	71546	weighted avg	0.9415	0.8554	0.8879	71546
F1 score: 0.3532					F1 score: 0.3004					F1 score: 0.3513					F1 score: 0.3531				
ROC-AUC: 0.8790					ROC-AUC: 0.8817					ROC-AUC: 0.8557					ROC-AUC: 0.8782				

Precision&Recall差距過大
且Precision明顯過低

Step2 : GridSearch 參數選擇

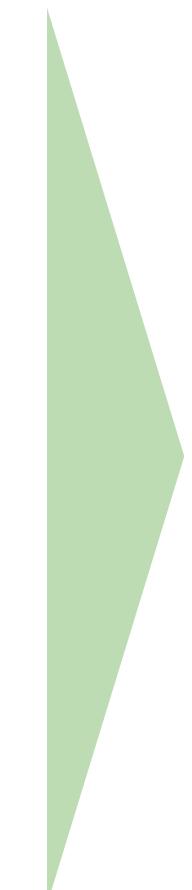
利用 GridSearch 選出最佳參數，包括不平衡的處理權重

GridSearch 參數

```
param_grid = {  
    'clf__max_depth': [3, 5, 7],  
    'clf__learning_rate': [0.01, 0.1],  
    'clf__n_estimators': [100, 200],  
    'clf__min_child_weight': [1, 5],  
    'clf__subsample': [0.8],  
    'clf__colsample_bytree': [0.8],  
    'clf__scale_pos_weight':  
        [round(scale_pos_weight),  
         int(scale_pos_weight) + 5,  
         int(scale_pos_weight) + 10],  
}
```

GridSearch 最佳參數

```
grid.best_params_ = {  
    'clf__colsample_bytree': 0.8,  
    'clf__learning_rate': 0.1,  
    'clf__max_depth': 7,  
    'clf__min_child_weight': 1,  
    'clf__n_estimators': 200,  
    'clf__scale_pos_weight': 17,  
    'clf__subsample': 0.8  
}
```



Step3 : VotingClassifier - Soft Voting 模型融合

以 VotingClassifier 融合模型，縮小 Precision 與 Recall 的差距並降低 FN

Precision 過低，Recall 明顯較高

預測結果誤判率高，實務應用效果受限

融合 XGBoost、Logistic Regression、
Random Forest、LightGBM

- LogReg & RF：穩定、可解釋，但對複雜關係與不平衡資料處理有限
- XGB & LGBM：補足非線性結構學習、提升對少數類別的偵測能力（希望達到 Recall ↑）

透過 Soft Voting，不同模型彼此互補，組合成可以彌補單一模型弱點、提升穩定性與整體效能的模型

Classification_Report

📌 VotingClassifier (XGB + Logistic + RF + LGBM, soft voting)

	precision	recall	f1-score	support
--	-----------	--------	----------	---------

0	0.9795	0.9118	0.9445	67661
1	0.3032	0.6685	0.4172	3885

F1 與 Precision 已上升同時不免犧牲 Recall

accuracy			0.8986	71546
macro avg	0.6414	0.7901	0.6808	71546
weighted avg	0.9428	0.8986	0.9158	71546

Confusion Matrix

61,693

5,968

1,288

2,597

Step4：特徵篩選

將 Feature Importance 與 SHAP 前 20 名聯集，選出 29 個重要特徵

Feature Importance		SHAP	
		feature	shap_value
0	HadAngina_Yes	AgeCategory	0.805083
1	ChestScan_Yes	HadAngina_Yes	0.803296
2	HadStroke_Yes	ChestScan_Yes	0.498051
3	GeneralHealth	GeneralHealth	0.343719
4	AgeCategory	Sex_Male	0.302372
5	Sex_Male	SmokerStatus_Never smoked	0.210527
6	HadDiabetes_Yes	RemovedTeeth_None of them	0.137815
7	RemovedTeeth_All	HadStroke_Yes	0.127398
8	SmokerStatus_Never smoked	AlcoholDrinkers_Yes	0.126683
9	DifficultyWalking_Yes	HadDiabetes_Yes	0.101443
10	RemovedTeeth_None of them	FluVaxLast12_Yes	0.101252
11	State_South Dakota	SmokerStatus_Former smoker	0.085977
12	State_Colorado	TetanusLast10Tdap_Yes, received Tdap	0.066766
13	DeafOrHardOfHearing_Yes	RaceEthnicityCategory_White only, Non-Hispanic	0.063395
14	State_Kansas	HIVTesting_Yes	0.061972
15	State_Oklahoma	PneumoVaxEver_Yes	0.060443
16	State_Nebraska	HadArthritis_Yes	0.053796
17	AlcoholDrinkers_Yes	DifficultyWalking_Yes	0.050894
18	HadKidneyDisease_Yes	HadDepressiveDisorder_Yes	0.048000
19	HadCOPD_Yes	PhysicalActivities_Yes	0.045686

聯集後找出
29個特徵

Step4：特徵篩選

將 29 個重要特徵帶入模型，最終結果 precision 與 F1 皆有提升且 FN 下降

聯集後特徵數量：29
(原特徵數：138)

列舉分數前 10 高的特徵

- 'ChestScan_Yes'
- 'Sex_Male'
- 'HadStroke_Yes'
- 'GeneralHealth'
- 'RemovedTeeth_None of them'
- 'AgeCategory'
- 'HadDiabetes_Yes'
- 'RemovedTeeth_All'
- 'SmokerStatus_Never smoked'
- 'HadAngina_Yes'
-

Classification_Report

📌 VotingClassifier (XGB + Logistic + RF + LGBM, soft voting)

	precision	recall	f1-score	support
0	0.9796	0.9119	0.9446	67661
1	0.3040	0.6700	0.4182	3885
accuracy			0.8988	71546
macro avg	0.6418	0.7910	0.6814	71546
weighted avg	0.9430	0.8988	0.9160	71546

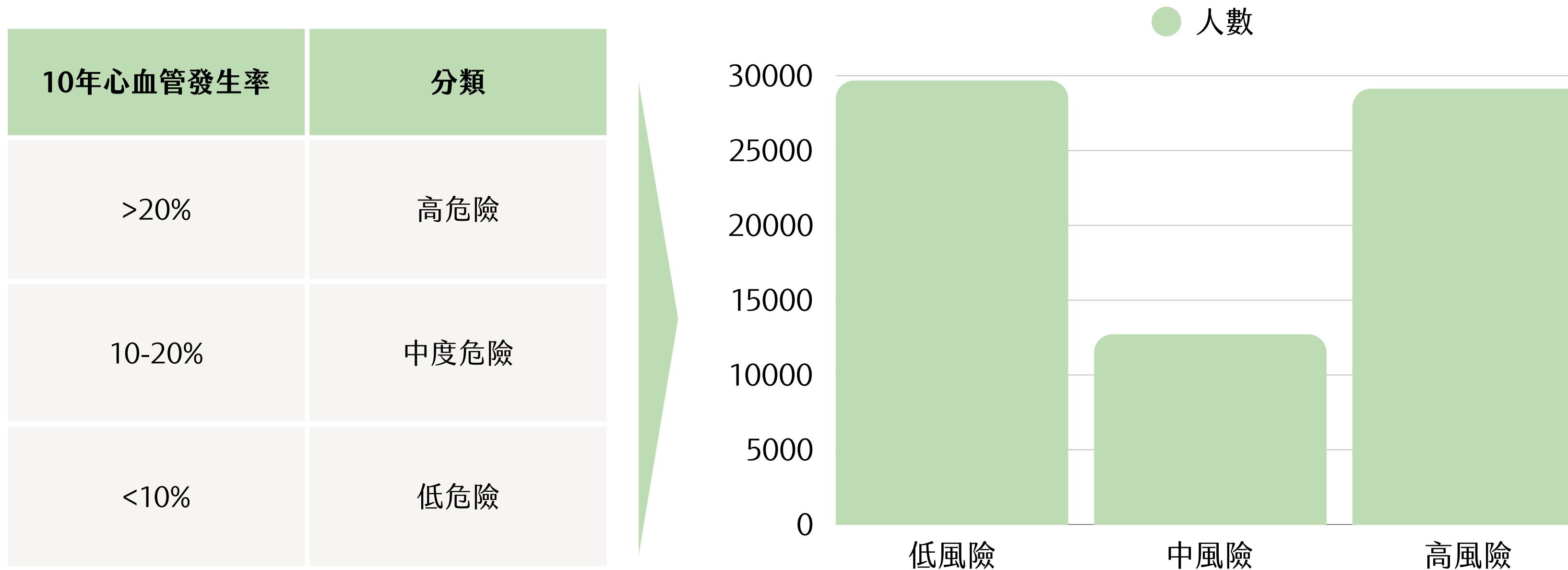
Confusion Matrix

61,701	5,960
1,282	2,603

同時提升
Precision Recall 與 F1

Step5：應用風險分級

依照輸出機率設定門檻做用戶的心臟病風險分類，以利後續策略設計





05

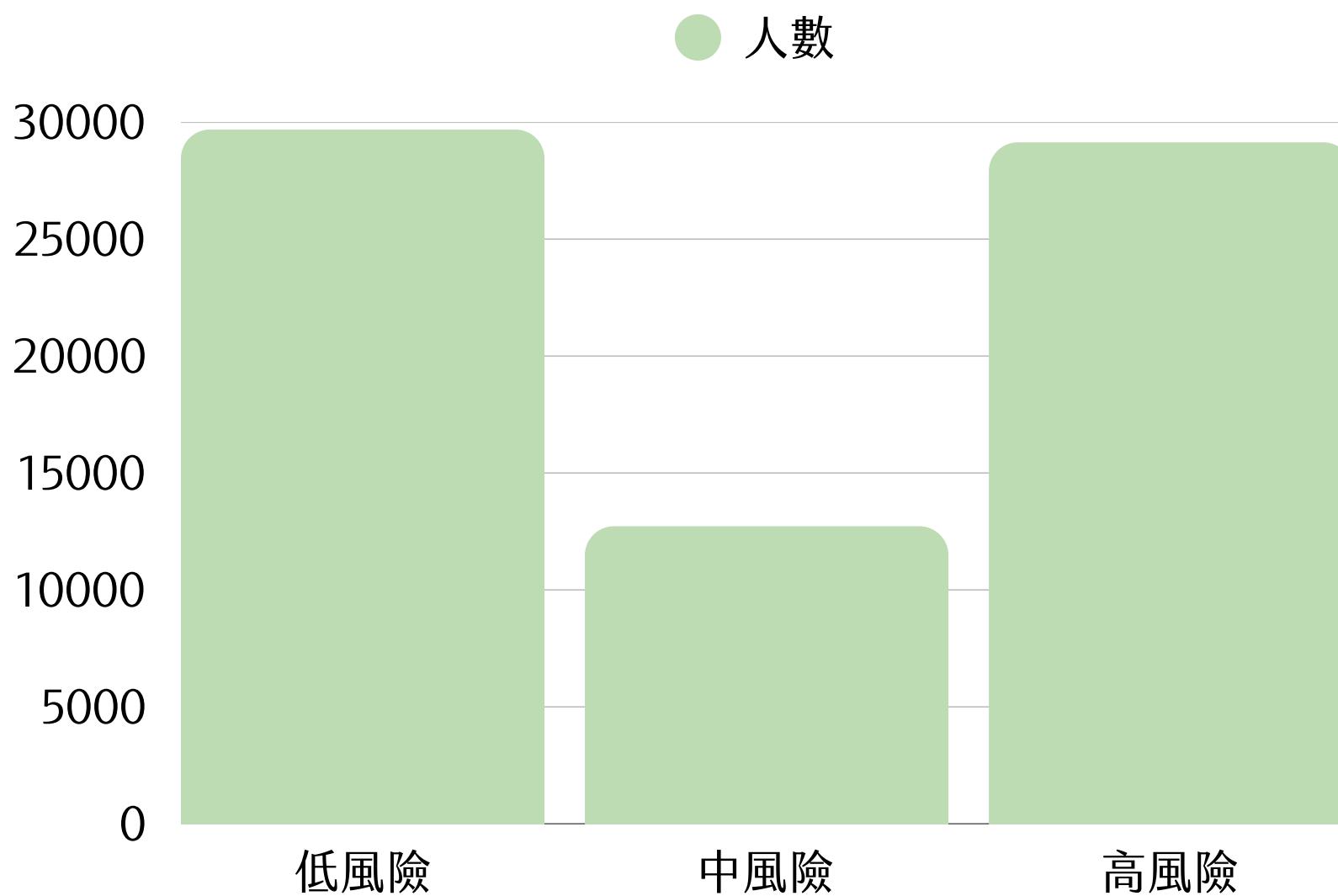
實際商業應用

依據模型找出連集特徵中前 11 個重要特徵以及創建風險分級機制

29 個重要特徵之中 11 個權重較高特徵

特徵名稱	說明
HadAngina_Yes	是否曾患心絞痛
ChestScan_Yes	是否曾做過胸部掃描
HadStroke_Yes	是否曾中風
GeneralHealth	自評健康狀況
AgeCategory	年齡區間
Sex_Male	性別為男性
HadDiabetes_Yes	是否有糖尿病
SmokerStatus_Never smoked	從未吸菸
RemovedTeeth_None of them	沒有拔除過牙齒
AlcoholDrinkers_Yes	是否有飲酒習慣
DifficultyWalking_Yes	是否行走困難

依照模型創建風險分級



篩選權重高之特徵，依據特徵「生活習慣與否」納入保險流程

依「生活習慣」與否區分特徵

特徵名稱	說明	生活習慣
HadAngina_Yes	是否曾患心絞痛	
ChestScan_Yes	是否曾做過胸部掃描	
HadStroke_Yes	是否曾中風	
GeneralHealth	自評健康狀況	V
AgeCategory	年齡區間	
Sex_Male	性別為男性	
HadDiabetes_Yes	是否有糖尿病	
SmokerStatus_Never smoked	從未吸菸	V
RemovedTeeth_None of them	沒有拔除過牙齒	
AlcoholDrinkers_Yes	是否有飲酒習慣	V
DifficultyWalking_Yes	是否行走困難	

商業價值

核保及承保

依據模型初步分出風險高、中、低三級作為核保參考，
並將「非生活習慣特徵」等納入核保標準，
綜合評估下完成核保、承保流程
減少逆選擇風險

保全及理賠

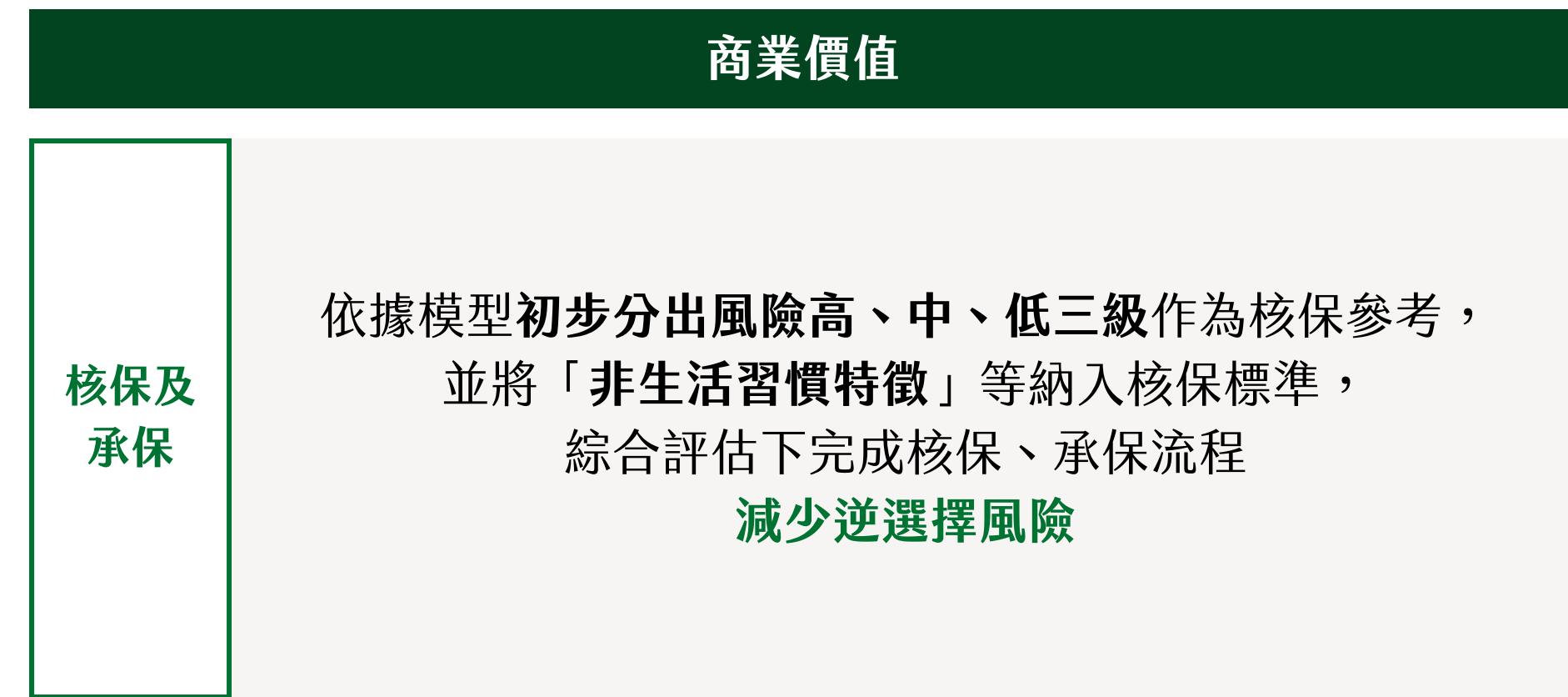
將「生活習慣特徵」納入外溢保單主要鼓勵、評估之行為，
協助國泰人壽能夠在保全管理階段，
透過鼓勵及偵測健康行為，
降低保戶發病率進而降低理賠總金額

篩選權重高之特徵，依據特徵「生活習慣與否」納入保險流程

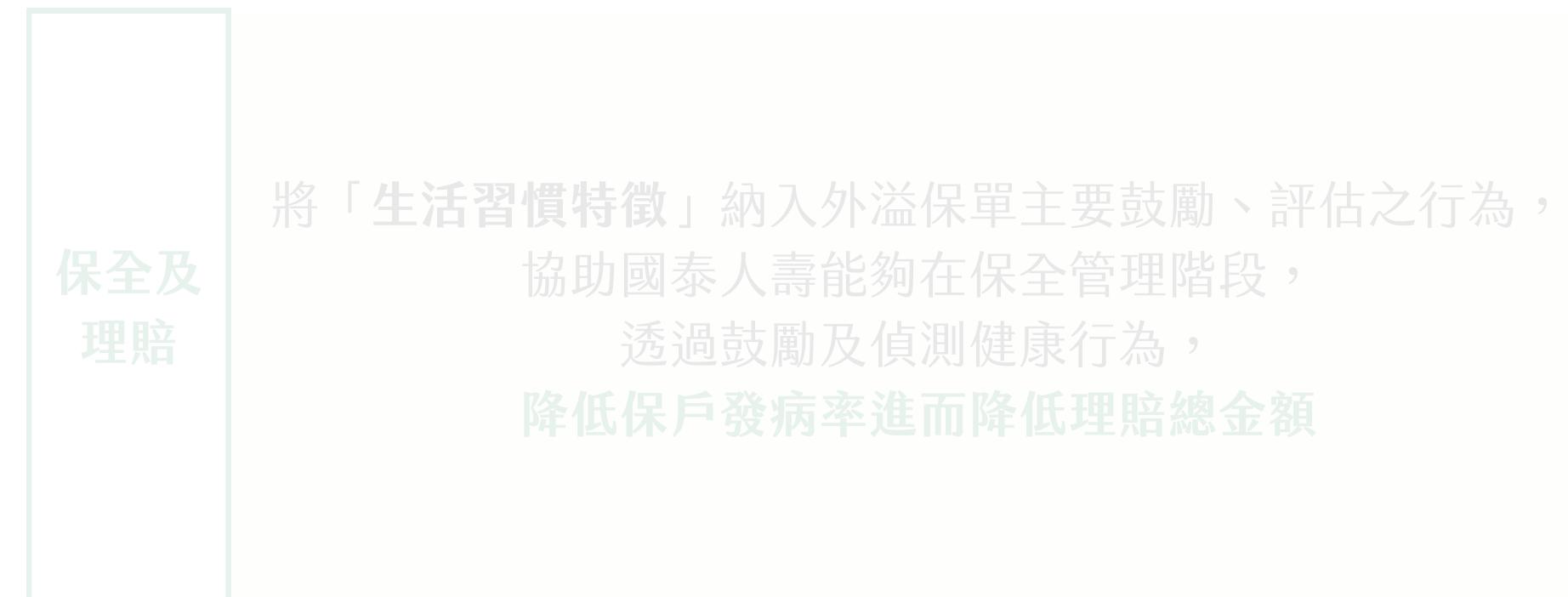
依「生活習慣」與否區分特徵

特徵名稱	說明	生活習慣
HadAngina_Yes	是否曾患心絞痛	
ChestScan_Yes	是否曾做過胸部掃描	
HadStroke_Yes	是否曾中風	
GeneralHealth	自評健康狀況	V
AgeCategory	年齡區間	
Sex_Male	性別為男性	
HadDiabetes_Yes	是否有糖尿病	
SmokerStatus_Never smoked	從未吸菸	V
RemovedTeeth_None of them	沒有拔除過牙齒	
AlcoholDrinkers_Yes	是否有飲酒習慣	V
DifficultyWalking_Yes	是否行走困難	

商業價值



依據模型初步分出風險高、中、低三級作為核保參考，
並將「非生活習慣特徵」等納入核保標準，
綜合評估下完成核保、承保流程
減少逆選擇風險

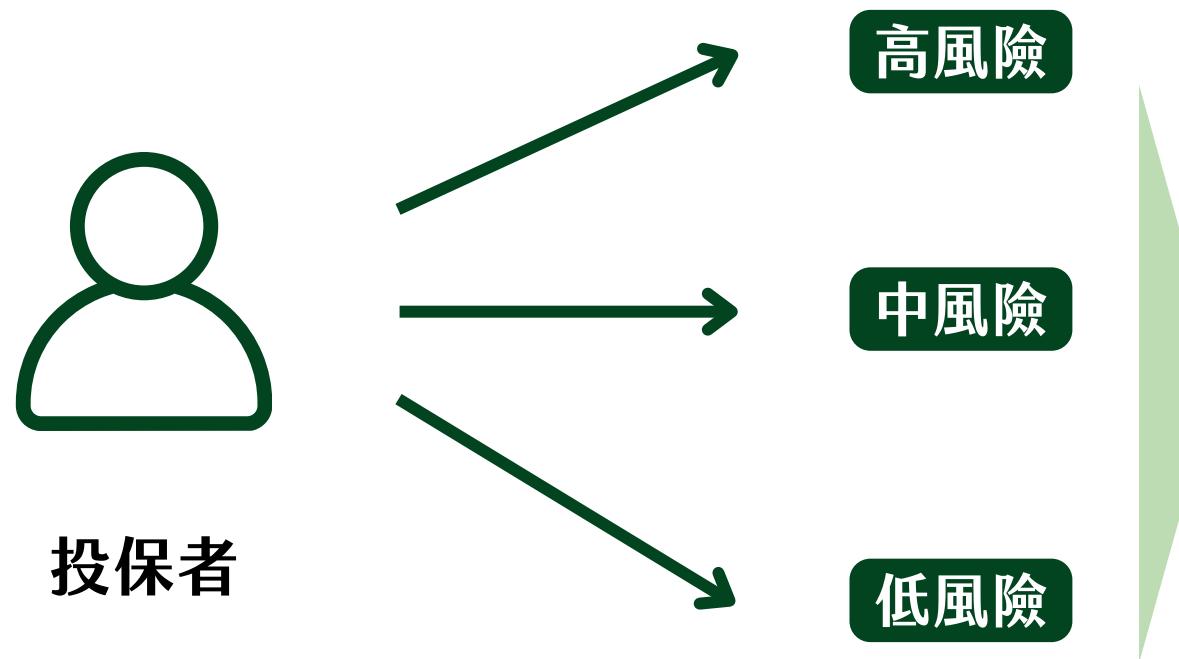


將「生活習慣特徵」納入外溢保單主要鼓勵、評估之行為，
協助國泰人壽能夠在保全管理階段，
透過鼓勵及偵測健康行為，
降低保戶發病率進而降低理賠總金額

流程：核保及承保

依據模型建立風險分級機制以區分投保者，並以重要特徵進一步核保

Step1: 初步風險分級機制



Step2: 依據重要特徵進一步核保

再依據前11個重要特徵，視該投保者真實情況以及補件情況「最後人為評估」是否加費或拒保或維持。

Step3: 定期監控&優化模型

定期監控模型效果，確認實際理賠率是否與預測一致，進行模型優化。

並提供模型結果可解釋性紀錄、公平性檢查報告（bias check）

目前台灣保險業者使用 AI，原則上僅能輔助核保，不能當作唯一拒保或訂價依據，否則可能違反公平對待原則

流程：核保及承保

分級後依據重要特徵人工核保，依真實情形判斷正常保費、加費或拒保

預先將重要特徵「胸部掃描」、「心絞痛」
列入拒保或優先加費之指標

將「健康自評」、「中風」、「吸菸」、「
拔牙」、「糖尿病」、「飲酒」、「
行走困難」
列入需要觀察之項目，請投保人提供更詳細
說明或補件條件，依情形加費

國泰人壽公開資料已知，目前健康醫療險已
針對年齡、性別做保費差異

重要特徵	提供之核保建議	風險提示
是否曾做過胸部掃描	要求補件說明掃描原因與結果，考慮是否例外處理	可能因疑似心肺疾病做過胸部掃描
是否曾患心絞痛	依據病史，考慮是否例外處理	心臟病前兆，極具指標性
自評整體健康狀況	依據病史，考慮是否加費	自評健康狀況可揭示生活習慣或慢性病徵兆
是否曾中風	依據病史，考慮是否加費	可能表示血管問題嚴重
是否拔牙	無拔牙紀錄可視為正向特徵	良好口腔健康與心臟病風險負相關 (慢性發炎理論)
是否有吸菸	依據頻率，考慮是否加費	吸菸直接提升心血管風險
是否有糖尿病	若控制不佳，加費或延期	糖尿病為心血管病重大因子
是否有飲酒習慣	要求說明頻率高低，視為加費條件	飲酒與高血壓/中風等有關
是否行走困難	要求說明原因與檢查結果	需注意項目，可能為心臟衰竭表徵
年齡區間	年齡越高，保費級距提升	已用於現行定價模型
性別為男性	與心血管疾病有關聯	性別風險差異已納入現行模型

篩選權重高之特徵，依據特徵「生活習慣與否」納入保險流程

依「生活習慣」與否區分特徵

特徵名稱	說明	生活習慣
HadAngina_Yes	是否曾患心絞痛	
ChestScan_Yes	是否曾做過胸部掃描	
HadStroke_Yes	是否曾中風	
GeneralHealth	自評健康狀況	V
AgeCategory	年齡區間	
Sex_Male	性別為男性	
HadDiabetes_Yes	是否有糖尿病	
SmokerStatus_Never smoked	從未吸菸	V
RemovedTeeth_None of them	沒有拔除過牙齒	
AlcoholDrinkers_Yes	是否有飲酒習慣	V
DifficultyWalking_Yes	是否行走困難	

商業價值

將「非生活習慣特徵」納入核保標準，並依據模型分出風險高、中、低三級作為核保參考，以及年齡等調整保費或拒保，減少逆選擇風險

核保及承保

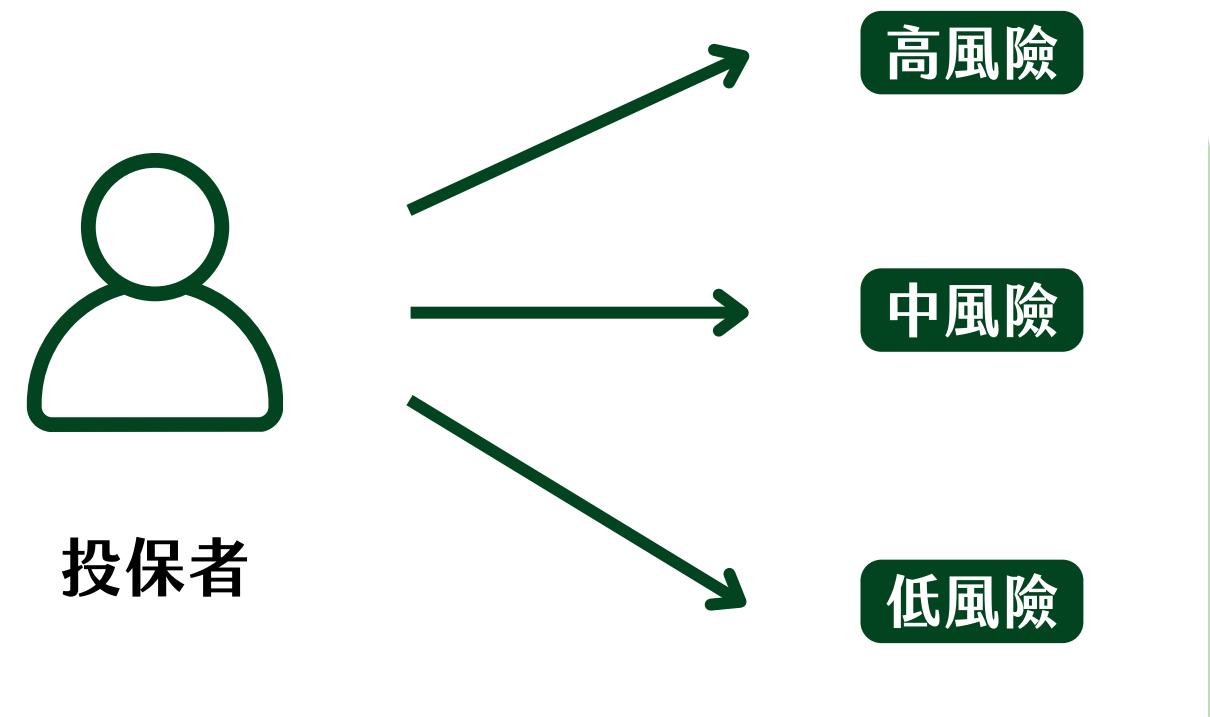
保全及理賠

將「生活習慣特徵」納入外溢保單主要鼓勵、評估之行為，協助國泰人壽能夠在保全管理階段，透過鼓勵及偵測健康行為，降低保戶發病率進而降低理賠總金額

流程：保全及理賠

依據風險分級機制制定差異化外溢以及回饋機制，降低理賠風險與金額

Step1: 核保先依照模型創建風險分級



Step2: 外溢保單差異化機制

- 干預強度因風險等級而異
- 高風險 → 設定較高運動目標
(如每日需走一萬五～兩萬步、
健身場館一週打卡3~4次)
 - 中、低風險 → 設定較基本運動目標
(如每日需走五千～一萬步、
健身場館一週打卡1~2次)

Step3: 外溢保單回饋機制

- 回饋設計因風險等級而異
- 高風險 → 完成可獲得國泰小樹點
 - 中、低風險 → 完成可獲得任務值
- 根據國泰FitBack平台，
約100~200任務值等同於一個國泰小樹
點，而小樹點1點=現金1元

利用保戶的罹患心臟病的預測機率進行風險分級，設計差異化的健康管理介入與回饋機制，
提升健康行為參與度，同時降低理賠成本。



謝謝大家！

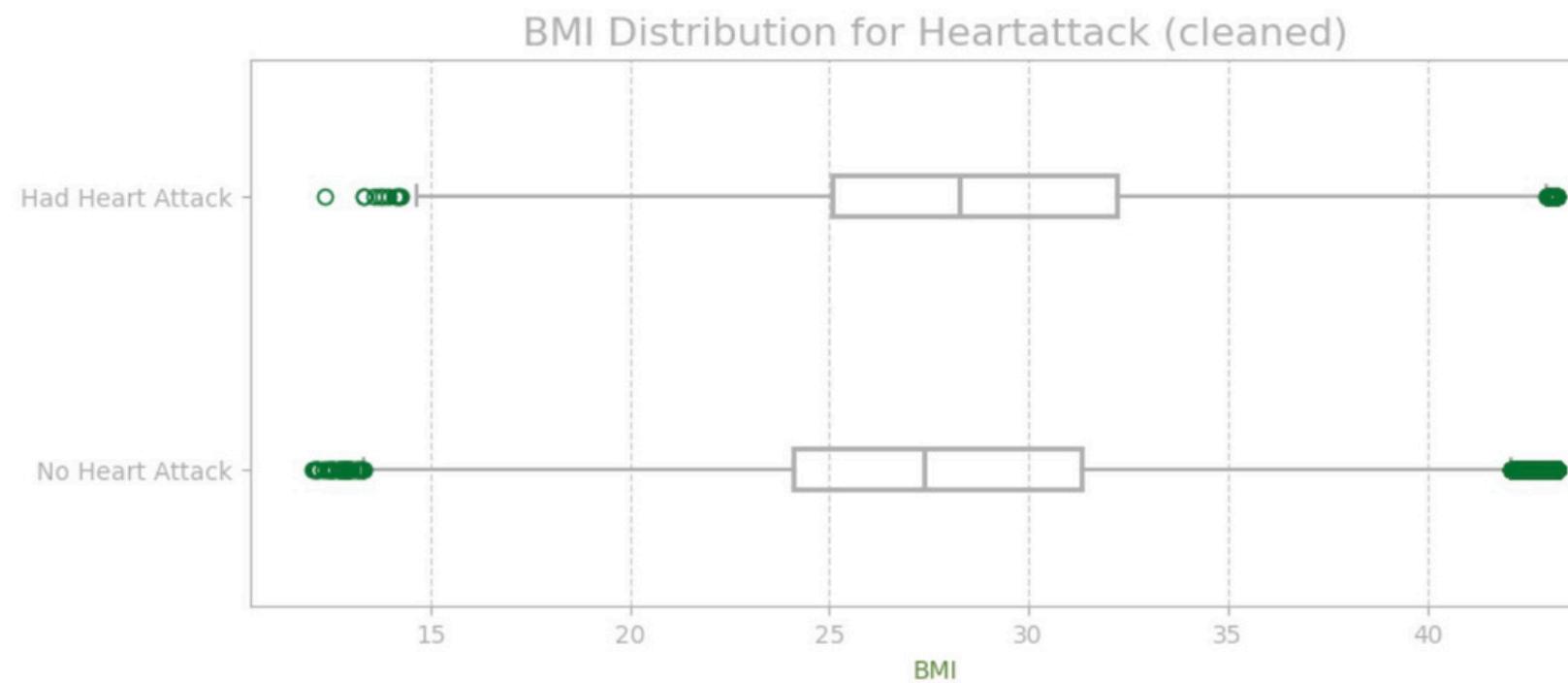
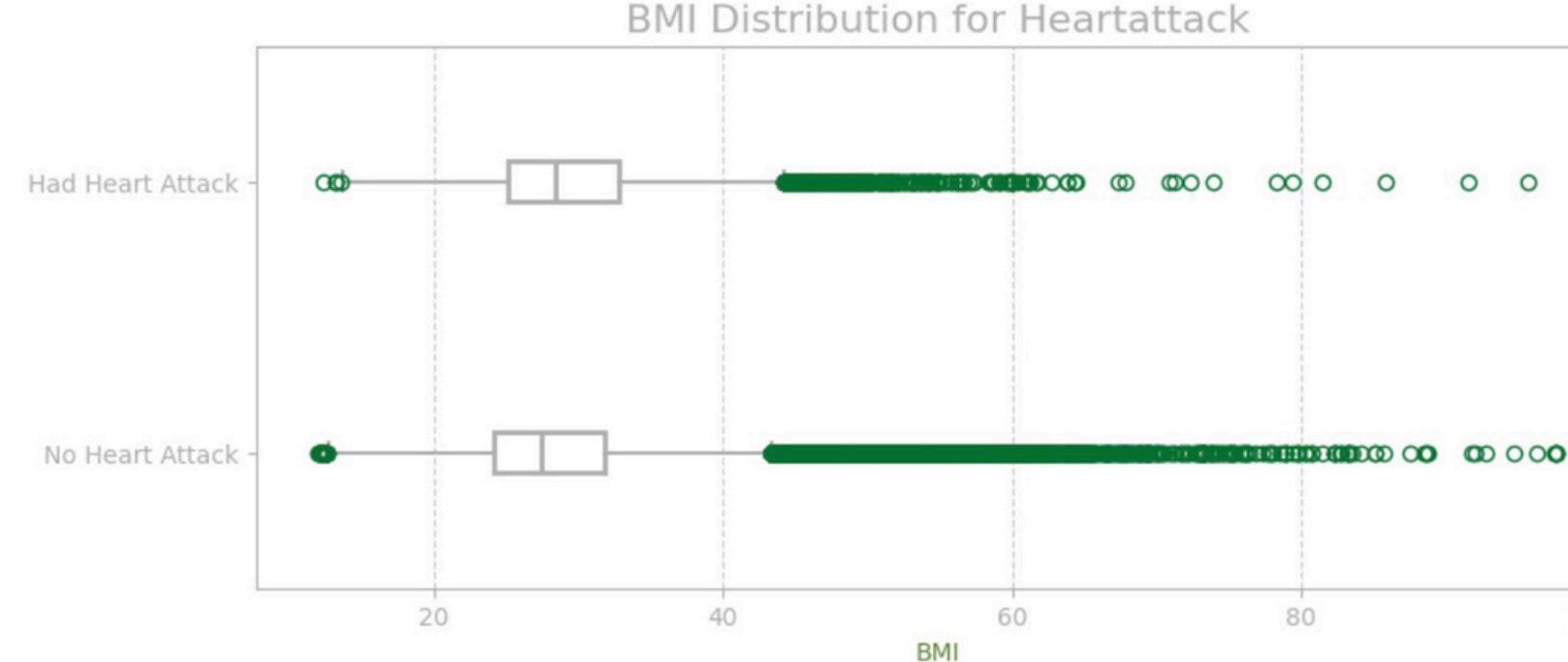


06

附錄

BMI 處理

刪除BMI異常值前後有心臟病人數皆大約佔5.7%，統計量相近，標準差減少



	count	mean	std	min	25%	50%	75%	max
HadHeartAttack								
No	225535.0	27.983758	5.396466	12.02	24.11	27.37	31.31	43.28
Yes	12950.0	28.760106	5.327608	12.34	25.06	28.25	32.23	43.28
df.groupby(['HadHeartAttack'])['BMI'].describe()								
HadHeartAttack								
No	232587.0	28.620521	6.507084	12.02	24.21	27.44	31.87	97.65
Yes	13435.0	29.492435	6.577941	12.34	25.10	28.48	32.76	95.66
[] 12950/225535								
[] 0.0574190258718159								
[] 13435/232587								
[] 0.05776333157055209								

資料不平衡之處理

調整方法

模型層面

資料層面

Under-sampling

刪減多數類資料，讓兩類平衡

RandomUnderSampler、NearMiss、ENN、Tomek Links...

Over-sampling

增加少數類樣本數

RandomOverSampler、SMOTE、ADASYN...

Hybrid-sampling

同時進行under and over-sampling

SMOTE + Tomek Links...

Class Weight

對少數類給更高權重

Cost-sensitive Learning

在模型損失函數中對不同錯誤類型給不同懲罰成本

評斷方式

Accuracy

整體預測中正確的比例

在資料不平衡時，結果可能具誤導性。

Precision

模型判定為「病患」者中，實際為病患的比例

常用於誤判代價高的情境，例如司法審判

Recall

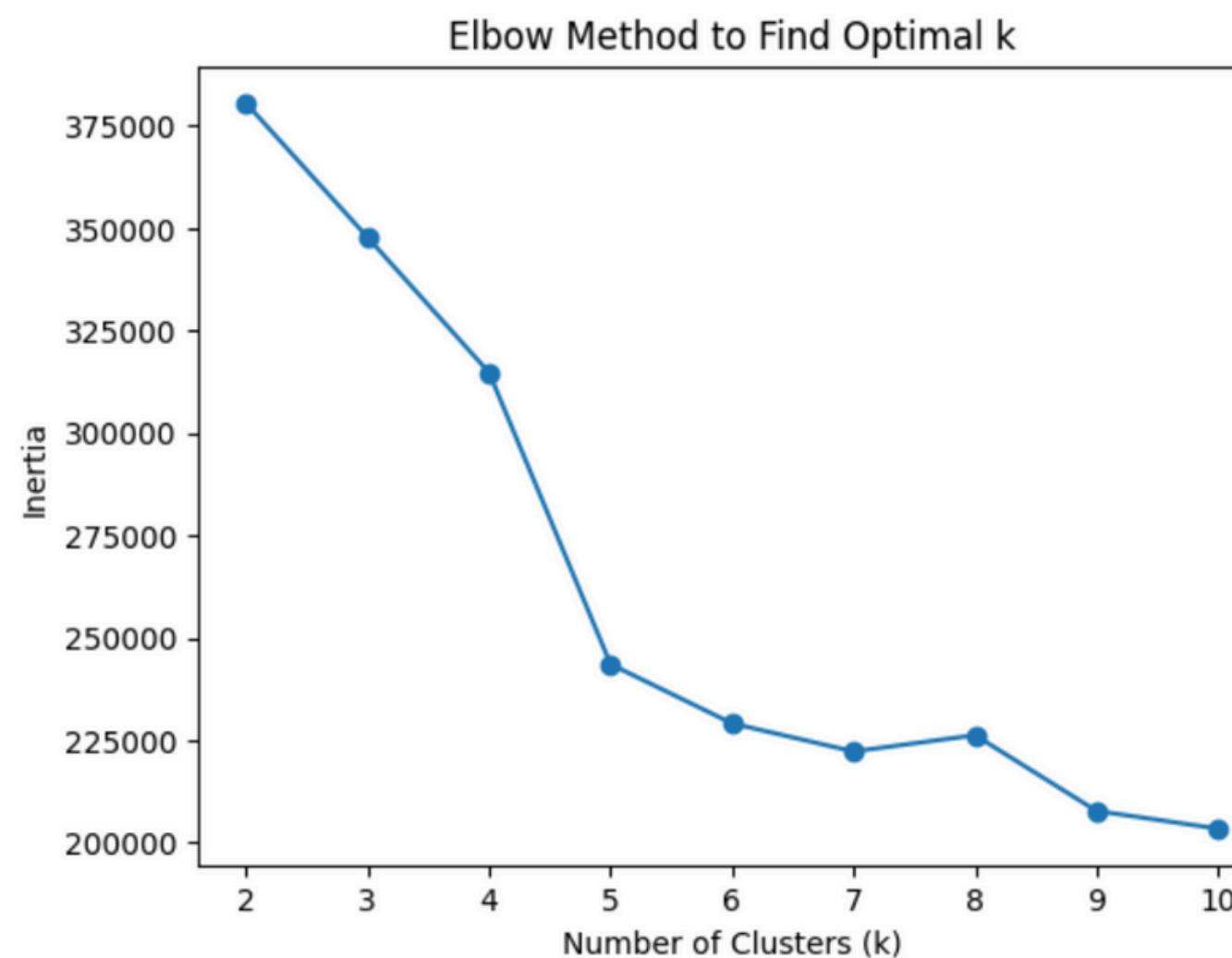
實際病患中，被正確辨識出的比例

常用於漏判代價高的情境，例如疾病篩檢

F1-score

Precision 與 Recall 的加權平均

KMeans分群



根據年齡性別吸菸分為5群

	0	1	2	3	4
年齡	60-69 18-24	65-74 75-79	55-69	60-74	60-74
性別	男最多	男第二	男女比4:6	女第二	女最多
吸菸程度	無	曾抽過/有時	頻繁	曾抽過	無
心臟病	少	最多	次多	普通	最少
統整	低風險 男性	高風險 高齡 男性	高風險 中年 重度吸菸	中風險 戒菸 女性	健康 女性

羅吉斯回歸

encoding (age/health 做 ordinal 剩下類別型 one-hot encoding 欄位數 109)

→ VIF 刪除 ['BMI', 'WeightInKilograms', 'HeightInMeters'] 欄位數 106

→ 跑 baseline

→ lasso / stepwise (p-value=0.05) / selectbest

📌 Baseline Logistic Regression (no feature selection, default C=1)

- Test Set -

	precision	recall	f1-score	support	method	AUC_train	AUC_test	Precision	Recall	F1	F2	\
0	0.9848	0.8363	0.9045	67661	Lasso	0.890917	0.891398	0.213890	0.775290	0.335281	0.508406	
1	0.2137	0.7748	0.3350	3885	BackwardElim	0.889980	0.891285	0.213391	0.776062	0.334740	0.508106	
accuracy			0.8330	71546	SelectKBest	0.889436	0.892478	0.214458	0.777349	0.336172	0.509756	
macro avg	0.5992	0.8055	0.6197	71546								
weighted avg	0.9429	0.8330	0.8736	71546	n_features							

Confusion Matrix:

```
[[56586 11075]
```

```
[ 875 3010]]
```

F2 (Test): 0.5080

ROC-AUC (Test): 0.8914

ROC-AUC (Train): 0.8909

method	
Lasso	103
BackwardElim	37
SelectKBest	30

決策樹

oversample+RandomizedSearchCV+class_weight='balanced'

Fitting 5 folds for each of 10 candidates, totalling 50 fits

最佳參數：{'min_samples_split': 10, 'min_samples_leaf': 10, 'max_depth': 10}

測試集分類報告：

	precision	recall	f1-score	support
0	0.9829	0.8210	0.8947	67668
1	0.1938	0.7509	0.3081	3878
accuracy			0.8172	71546
macro avg	0.5884	0.7860	0.6014	71546
weighted avg	0.9401	0.8172	0.8629	71546

Accuracy: 0.8172

Precision: 0.1938

Recall : 0.7509

F1-score: 0.3081

隨機森林

encoding (age/health 做 ordinal 剩下類別型 one-hot encoding)
→ class_weight={0: 1, 1: 12} 處理不平衡

混淆矩陣：

```
[[64373 3275]
 [ 1757 2141]]
```

分類報告：

	precision	recall	f1-score	support
0	0.97	0.95	0.96	67648
1	0.40	0.55	0.46	3898
accuracy			0.93	71546
macro avg	0.68	0.75	0.71	71546
weighted avg	0.94	0.93	0.93	71546