



# 國泰人壽『心臟病風險預測』 期末專案

## 第四組

110302058 金融四 邱士展  
109102040 經濟四 周紹璞  
110302013 金融四 易可倫  
110302026 金融四 鄭達嶸  
110308036 風管四 黃以穠



# Agenda

1

商業定位與目標

2

資料探索與預處理

3

特徵工程

4

模型選擇與訓練

5

商業應用

6

結論

# Agenda

1

商業定位與目標

2

資料探索與預處理

3

特徵工程

4

模型選擇與訓練

5

商業應用

6

結論

# 目前現狀

## AI、外溢保單與健康數據生態圈成形

外溢型保單  
成為主流趨勢

AI風險預測模型的  
戰略角色

健康實驗室與跨界  
合作布局

- 導入外溢機制 ( VBI: Value-Based Insurance )
- 透過健康回饋、行為追蹤、動態保費折扣強化保戶參與
- 穿戴裝置數據 ( 如 Fitbit、Apple Watch ) 導入即時健康監測與警示
- 促進保戶行為改變，降低理賠率、提升健康價值

- AI模型可預測高風險族群，導入前端健康介入 ( 健檢、諮詢、行為任務 )
- 結合外溢保單，形成「預測 → 介入 → 風控」完整閉環
- 降低重大理賠發生率，穩定公司損益結構

- AI與內部核保、商品設計、理賠風控深度整合
- 與醫院、學研機構、穿戴設備廠商 ( 如 Fitbit、Apple Watch ) 合作
- 建立完整健康數據生態圈
- 透過 MVP 試驗、A/B測試驗證新模型成效與應用場域可行性

### 挑戰

- 避險成本高企，壓縮利差收益。
- 解約潮持續，影響現金流。
- 人口高齡化加速，醫療與長照需求增加。

### 機會

- 數位科技應用深化，提升營運效率。
- 政策支持與法規鬆綁，鼓勵保險業創新。

# 「預測 → 介入 → 回饋」的健康管理生態圈

## 國泰AI健康保險專案的優劣與定位策略

### 差異化優勢

- 全台首家將AI心血管風險模型應用於實際保戶管理的壽險公司
- 結合健康App、FitBack任務機制與穿戴裝置整合，形成即時監測與介入能力
- 健康行為直接轉換為保費折扣，提升客戶參與與健康意識

### 主要商業痛點

- 心臟病為高額理賠主因，理賠壓力逐年升高
- 目前保戶健康行為管理介入效果有限，缺乏即時預警與個人化追蹤機制
- 產業競爭加劇，需要提升商品創新力與風控能力

### 核心定位目標

- 透過AI風險預測，提前辨識高風險保戶
- 結合健康任務卡、定期健檢與即時介入措施，降低心臟病發病率
- 同步提升保戶續保率、降低理賠支出，強化公司獲利能力與市場競爭力

# 專案目標設定：SMART

透過分析國泰保戶，建構「心臟病風險預測模型」，用於識別高風險族群。

類別	描述
S ( 具體性 )	透過分析內部保戶資料（如體檢報告、既往病史、保單類型、醫療支出）建立心臟病風險預測模型
M ( 可衡量 )	減少保費，以吸菸者（Smokers）若戒菸為例：實質預期支出減少約 6050 元
A ( 可達成 )	國泰擁有健檢、壽險、醫療理賠等綜合資料，具備AI模型訓練所需資料與專業人才
R ( 相關性 )	直接對應公司減損率控管、產品設計創新、與健康服務升級的三大策略方向
T ( 時限性 )	預計於模型建置後半年內導入核保端風險評估流程，並於一年內擴大應用於健康建議通知與產品推薦模組中

# Agenda

1

商業定位與目標

2

資料探索與預處理

3

特徵工程

4

模型選擇與訓練

5

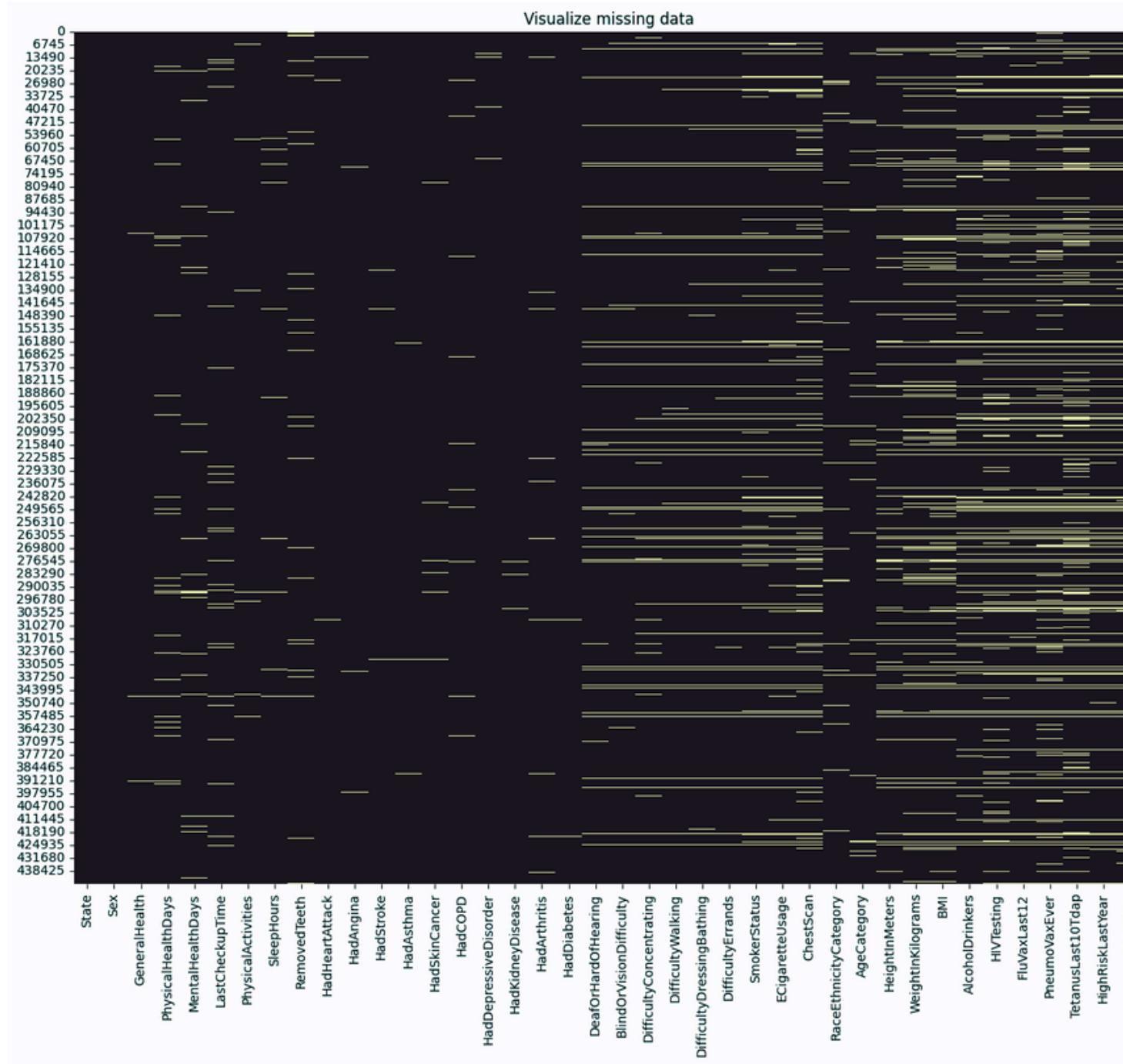
商業應用

6

結論

# 以 Group-based imputation 做 Data cleaning

## 原始資料缺失值



## Step 1

刪除缺失值 30% 以上的樣本

## Step 2

進行分群補值

### 一般補值

### 分群補值

優勢

較快

較能捕捉到同群體  
特徵、解釋力較強

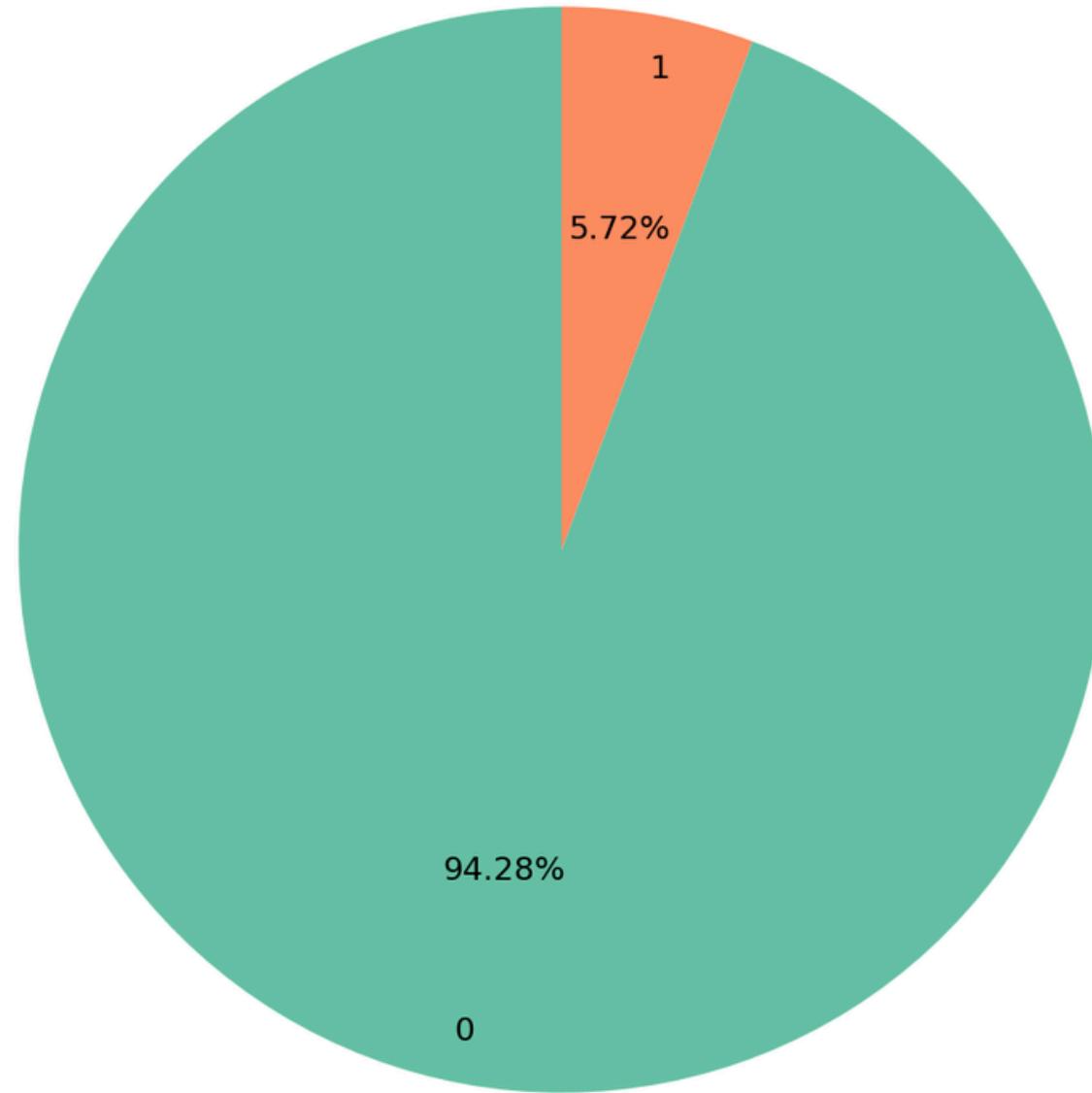
缺點

不同群體間差異大，  
以相同方式補值可能  
會產生嚴重誤差

分群標準難界定

# 患心臟病比例僅 5% , 本組以三種方式進行處理

心臟病患者比例



SMOTE

較為簡單，快速，但是在特徵維度較高的情況下，較容易生成無意義樣本

GAN

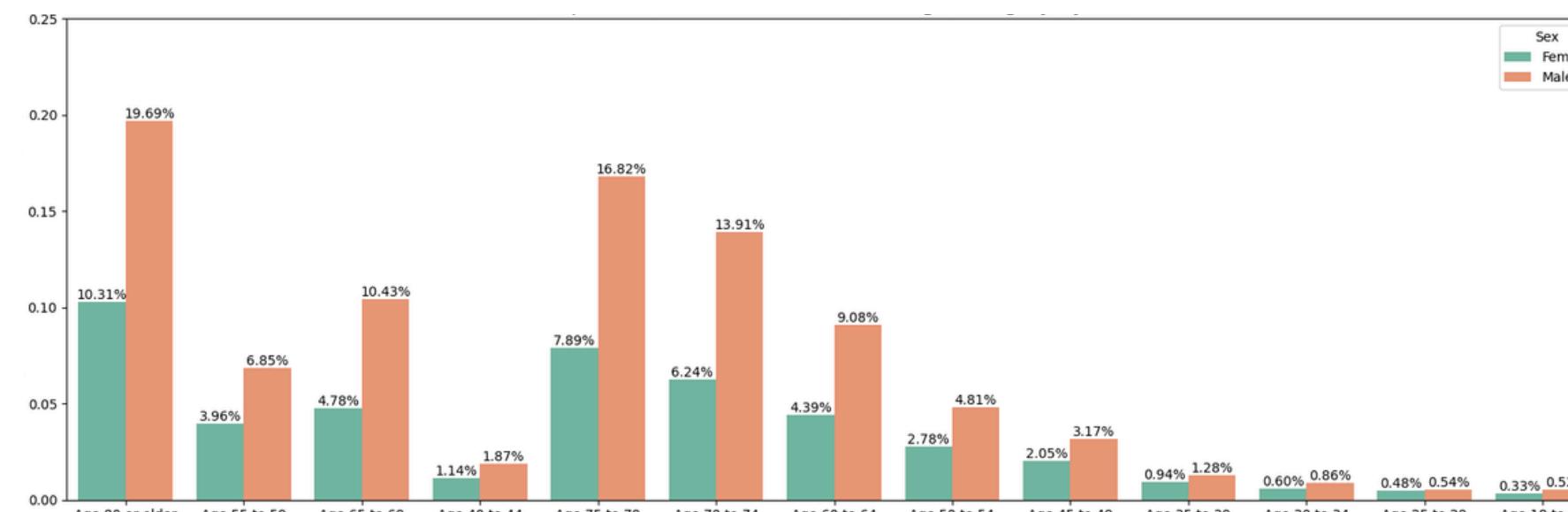
透過生成器跟鑑別器對抗使其生成之樣本質量較 SMOTE 高，但超參數的調優較難界定，且計算成本較高

Tree-based Models  
or  
Ensemble Model

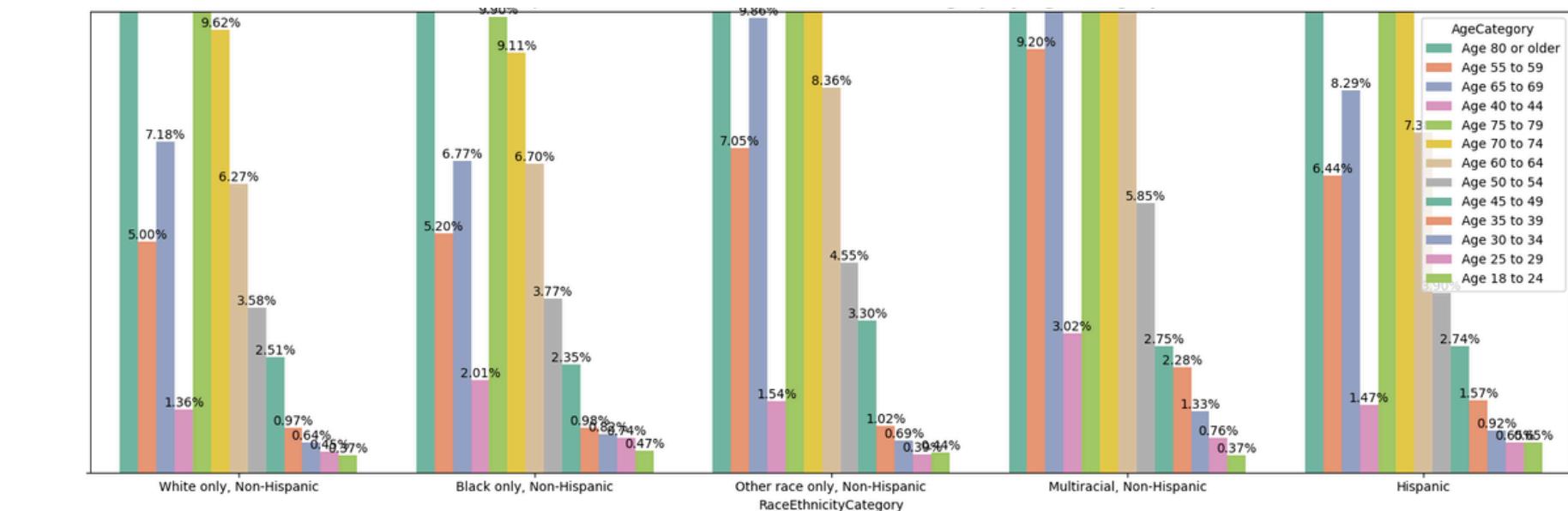
不改變原始資料的分佈，避免生成不合理的樣本，但若是少數樣本比例過低，模型訓練結果可能不佳

# EDA

## 不可控因素



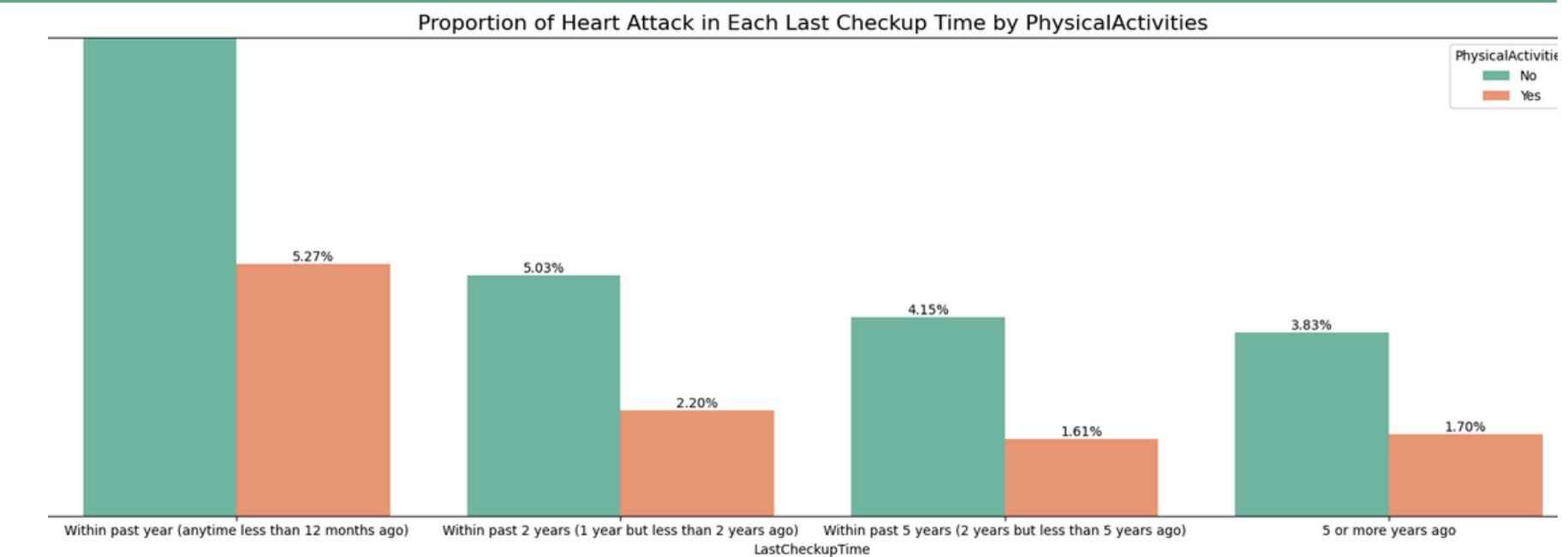
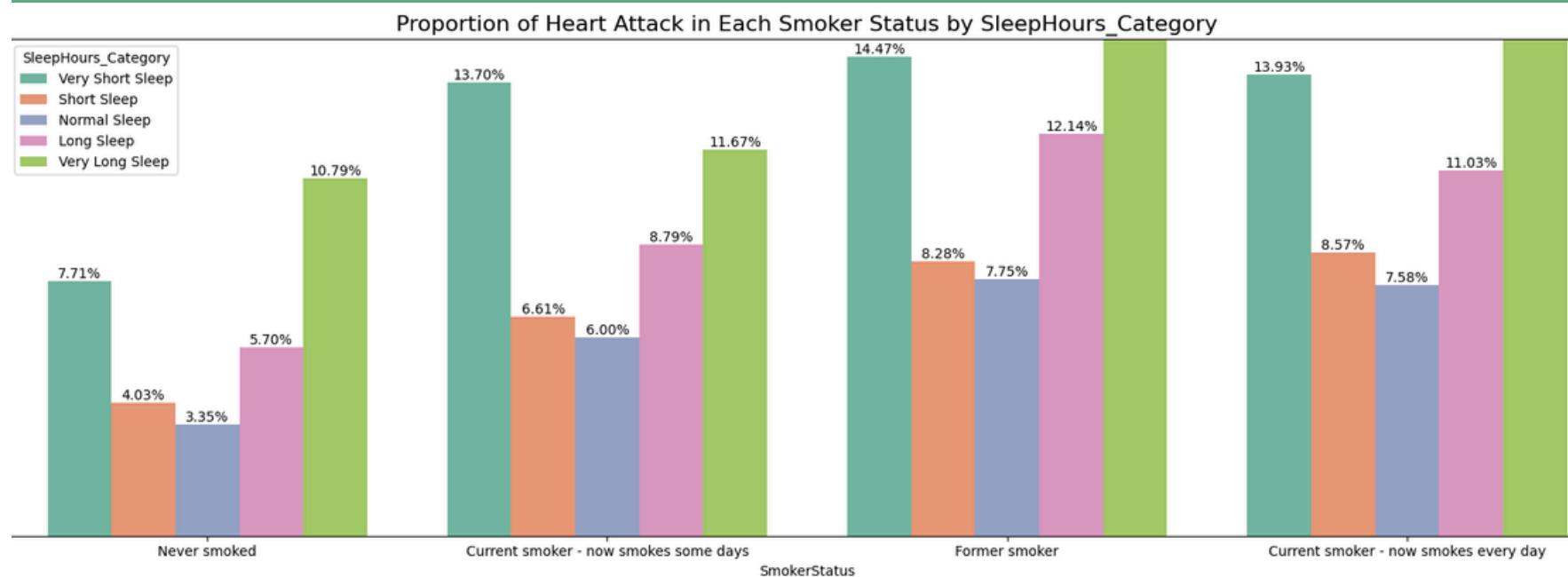
明顯看出在有心臟病的樣本中男性較女性佔比較高；高年齡群比例也顯著較高



種族差異較難看出跟心臟病的關聯性

# EDA

## 可控因素（後續檢測是否為重要特徵，外溢保單設計時納入考慮）

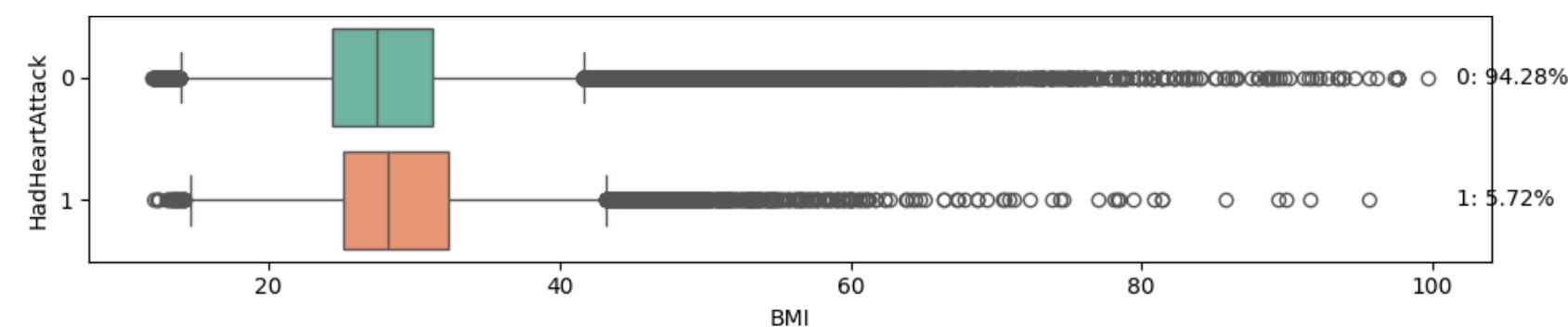


<左圖>由睡眠時長與抽菸可以看出為潛在重要變數，可能為國泰保單設計時須考慮之重要變數

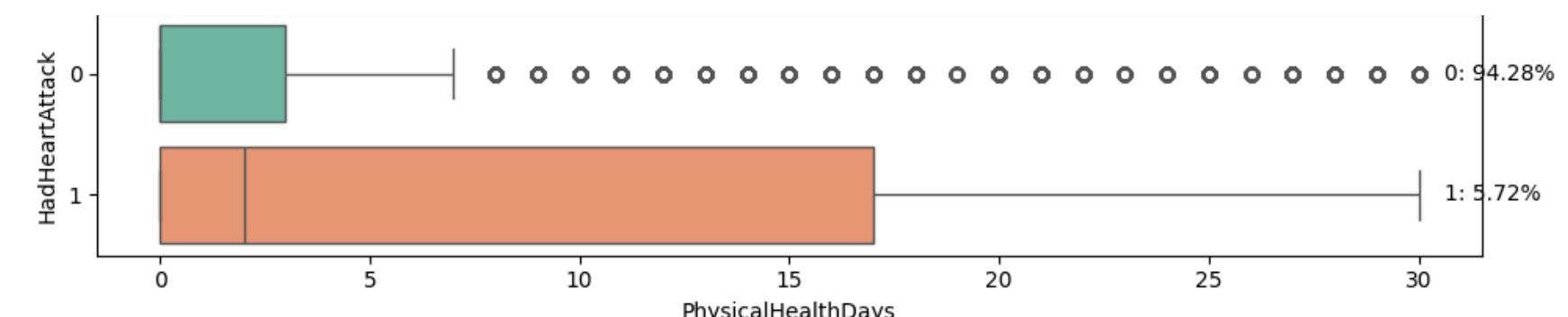
<右圖>有在進行身體活動的類別以及最新一次檢查時間可能也為重要變數，一年之內有進行檢查的心臟病比例較高初步推測原因可能是因為身體有其他問題而去檢查，可能主導原因是其餘變數

# Outliers 視覺化

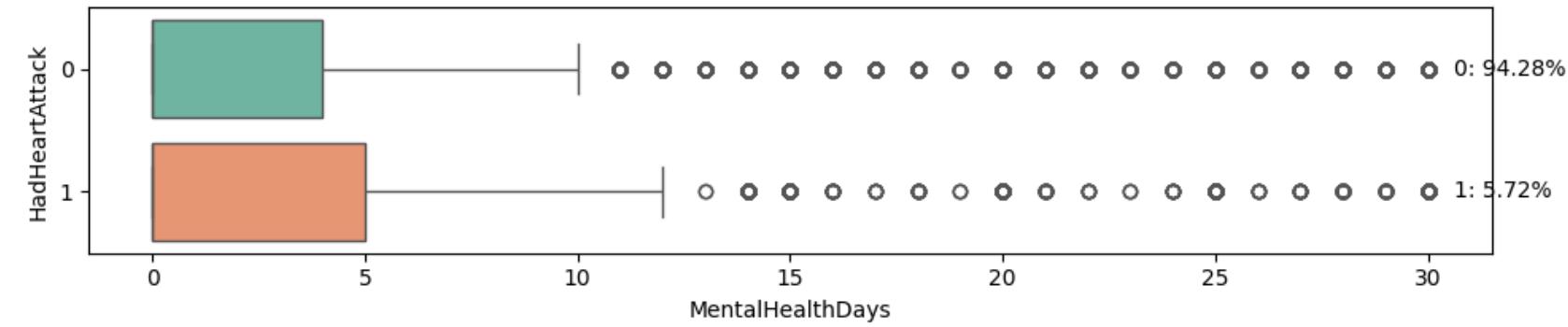
## BMI Outliers



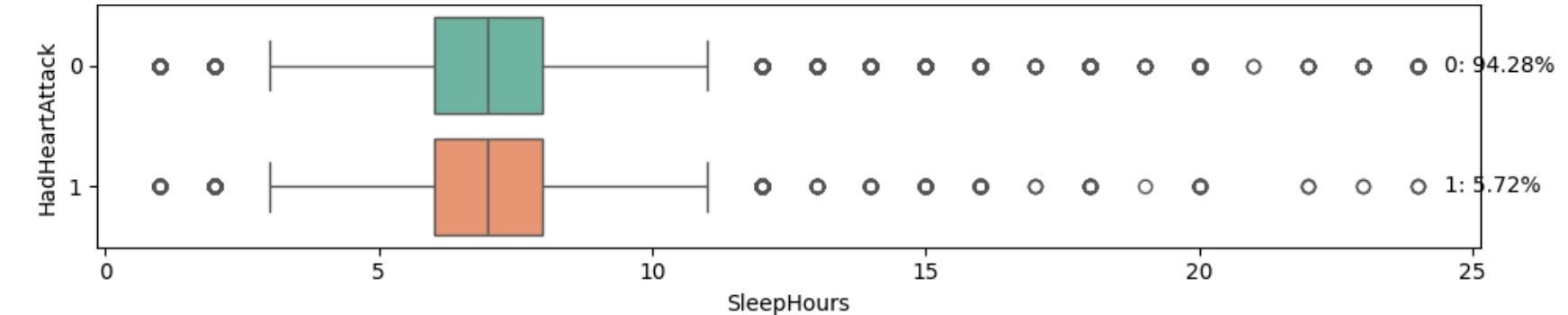
## PhysicalHealthDays Outliers



## MentalHealthDays Outliers

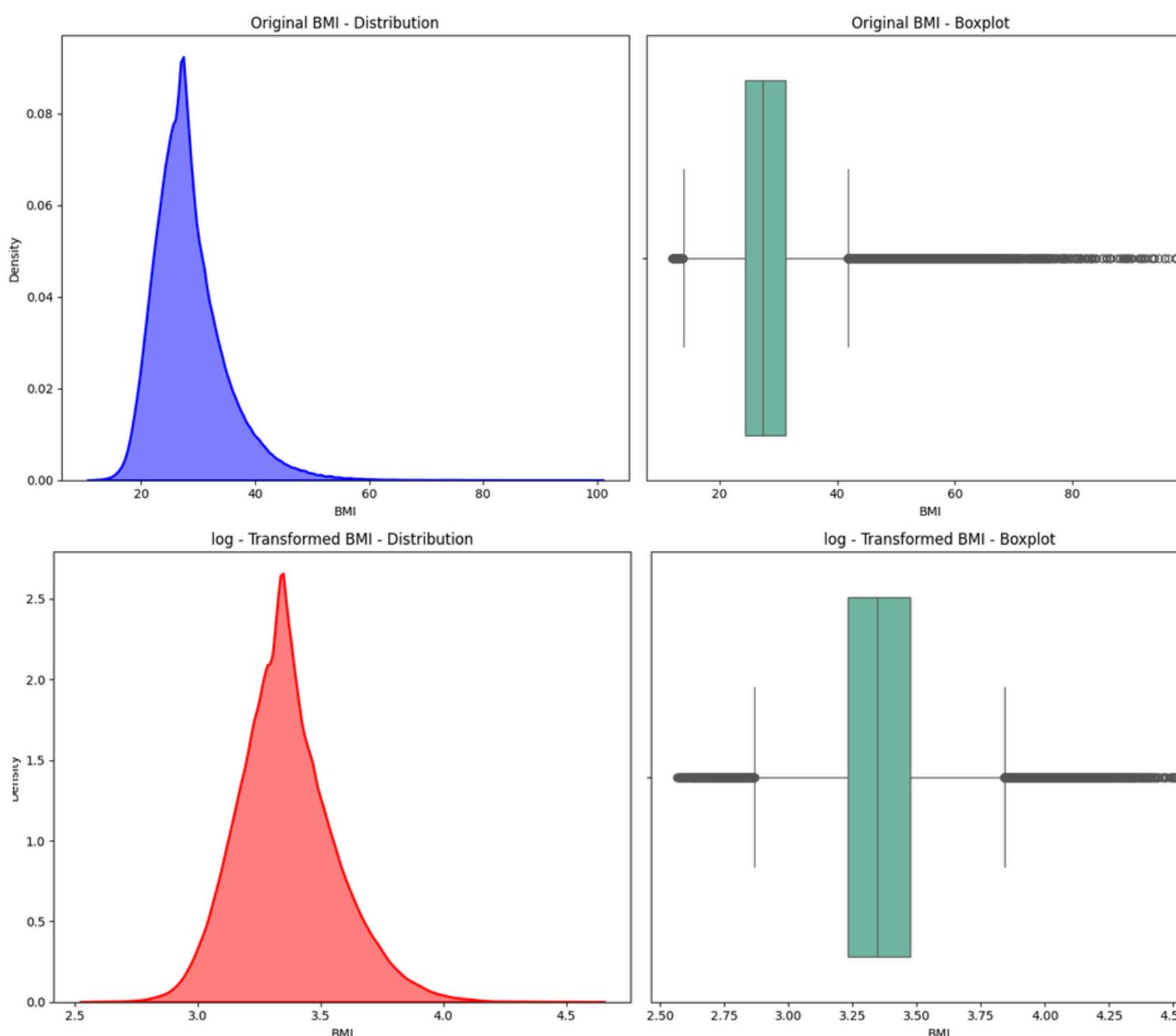


## SleepHours Outliers



# 特徵異常值以 Isolation Forest 進行偵測

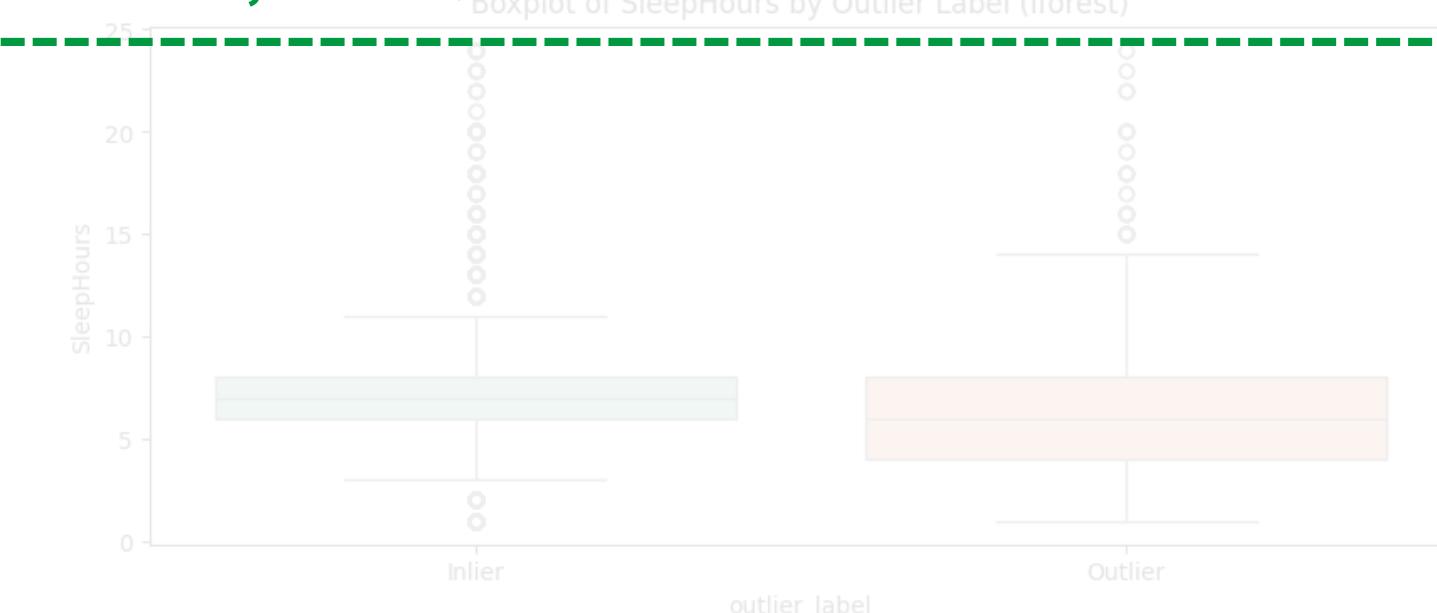
## 傳統 Outliers 處理方法 ( e.g. SleepHours )



## Isolation forest

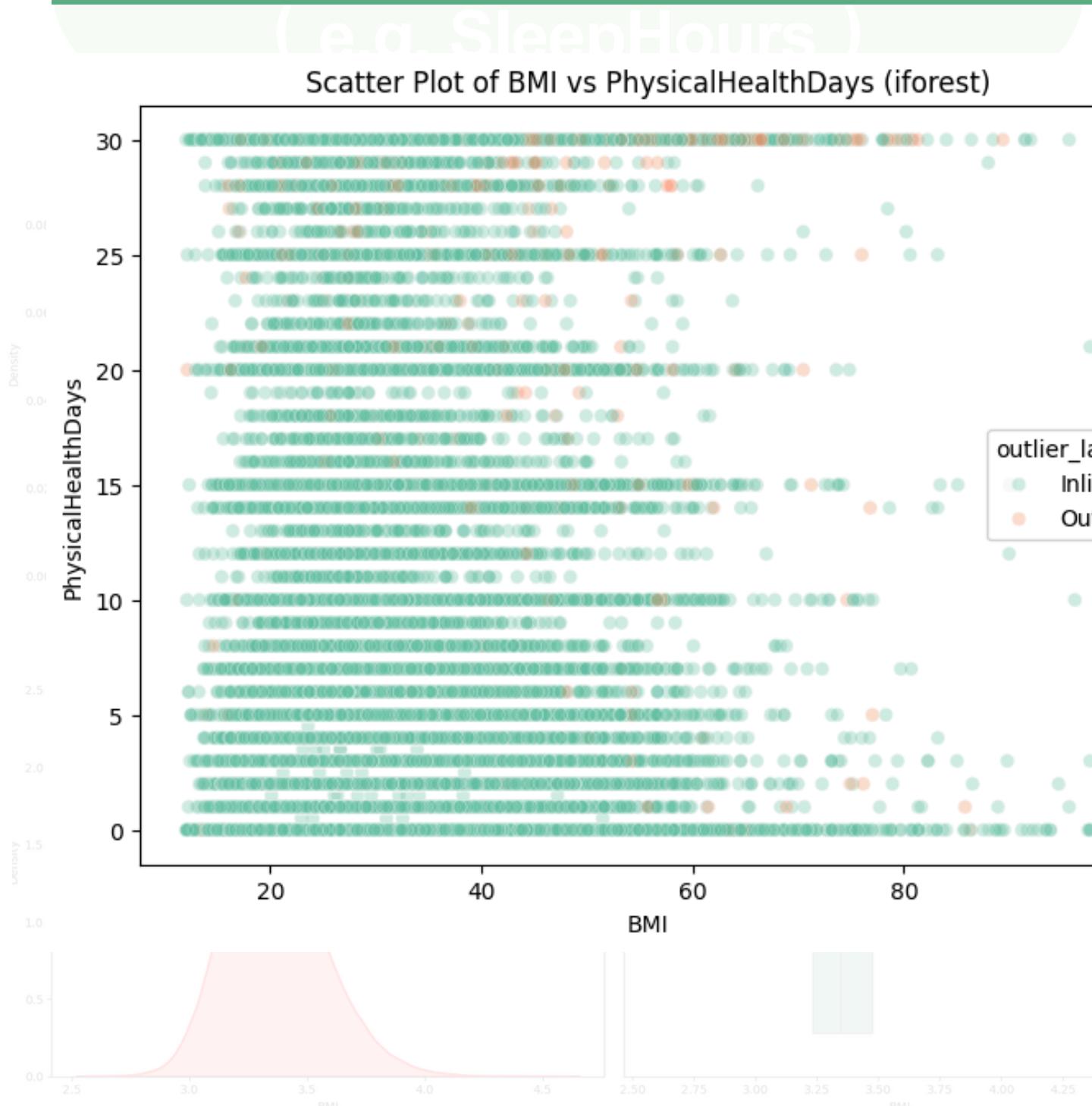
透過單一特徵的觀察來去判斷是否為 Outliers，並且進行轉換或者刪除，可能會有兩種問題

1. 忽略特徵在多維空間的特性，被認為是 Outliers 並且刪除，誤將正常數據點標記為異常，導致數據丟失
2. 有些數據點在單一特徵上正常，但在多維特徵組合下異常，傳統方法無法檢測這種情況

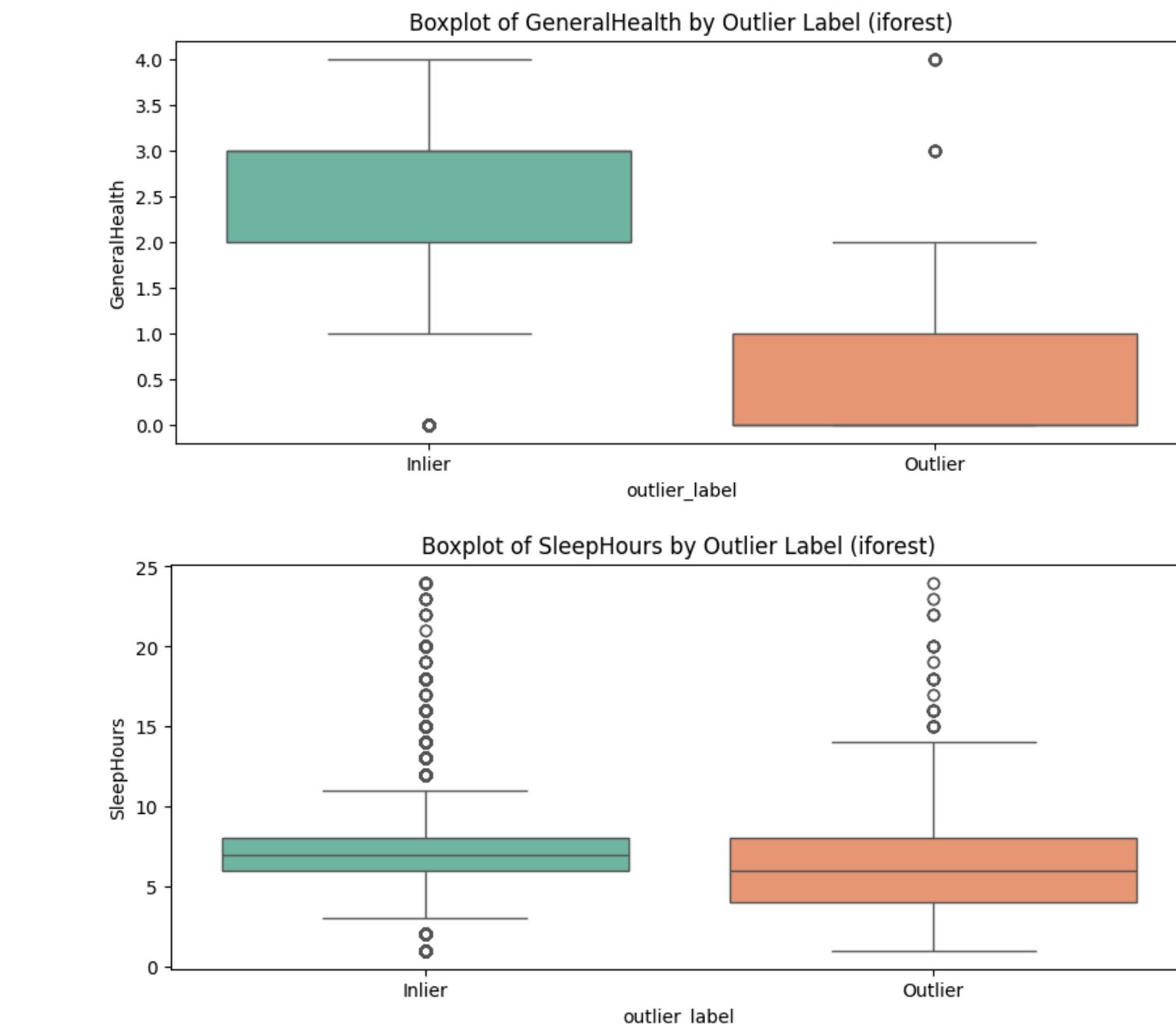


# 特徵異常值以 Isolation Forest 進行偵測

e.g. BMI、PhysicalHealthDays



Isolation forest



# Agenda

1

商業定位與目標

2

資料探索與預處理

3

特徵工程

4

模型選擇與訓練

5

商業應用

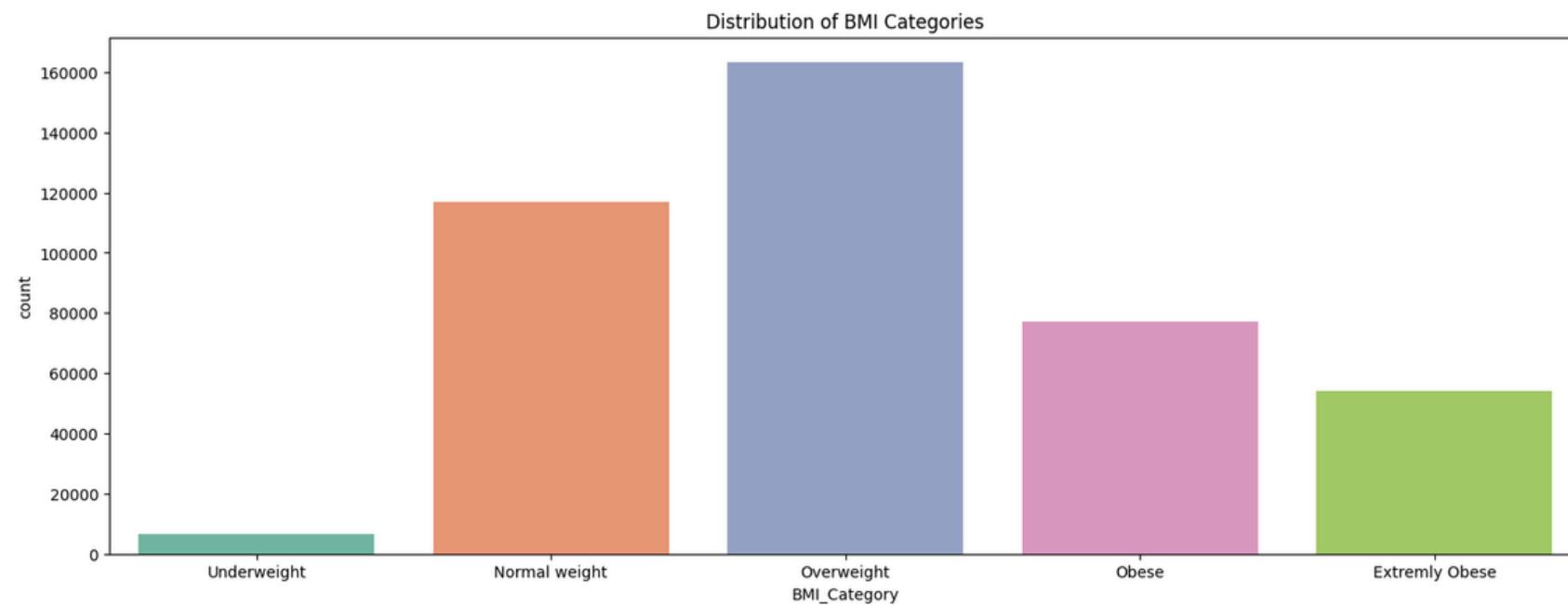
6

結論

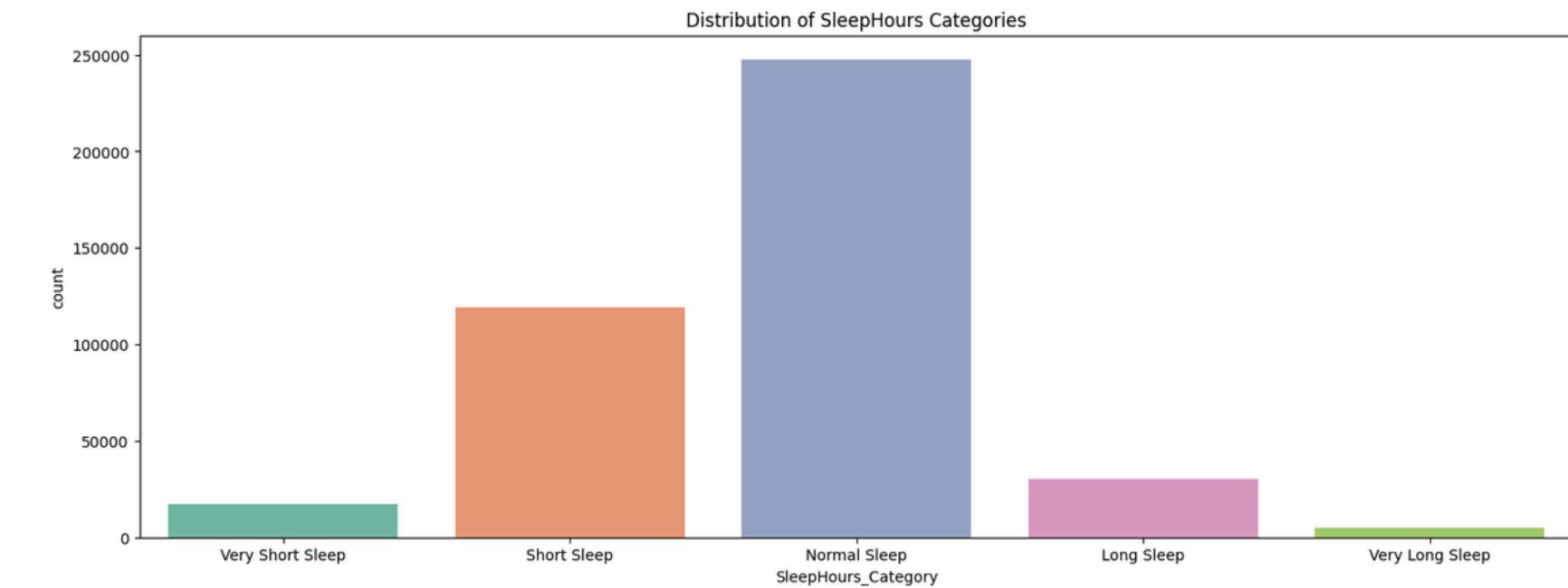
# 將數值型特徵 Binning (分組)，以利於後續分析

- 數值型特徵（如 BMI 和 SleepHours）的分佈可能過於連續或複雜，直接分析可能導致模型過於敏感或難以解釋。透過 Binning，將連續數值轉換為離散的類別（如「正常體重」、「過重」或「短睡眠」、「長睡眠」），讓資料結構更簡單且有意義
- 預期好處：
  - 提升可解釋性
  - 降低噪音干擾

## BMI Binning



## SleepHours Binning



# 將類別型特徵 Encoding

- 將二元分類的類別型特徵進行 Encoding ( 無中風 : 0 , 有中風 : 1 ) , 序數型的類別特徵依照其順序給出數值 ( Age 18 to 24 : 0 , Age 80 or older : 12 )

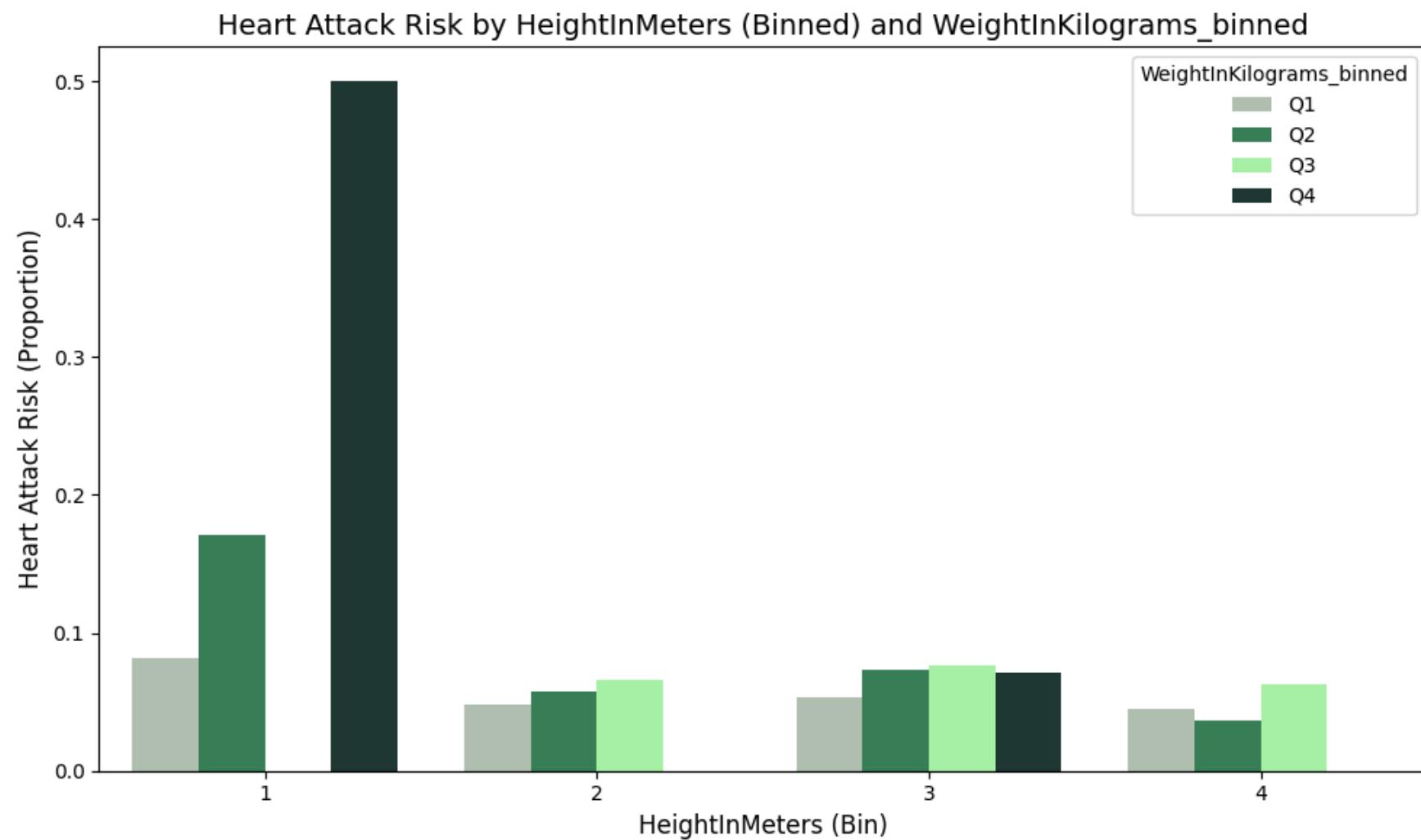
## 部分特徵 Encoding 結果

Sex	GeneralHealth	PhysicalHealthDays	MentalHealthDays	LastCheckupTime
1	3	2.0	0.0	3
0	3	0.0	0.0	3
1	3	0.0	0.0	3
0	3	1.0	6.0	1
1	3	0.0	0.0	3

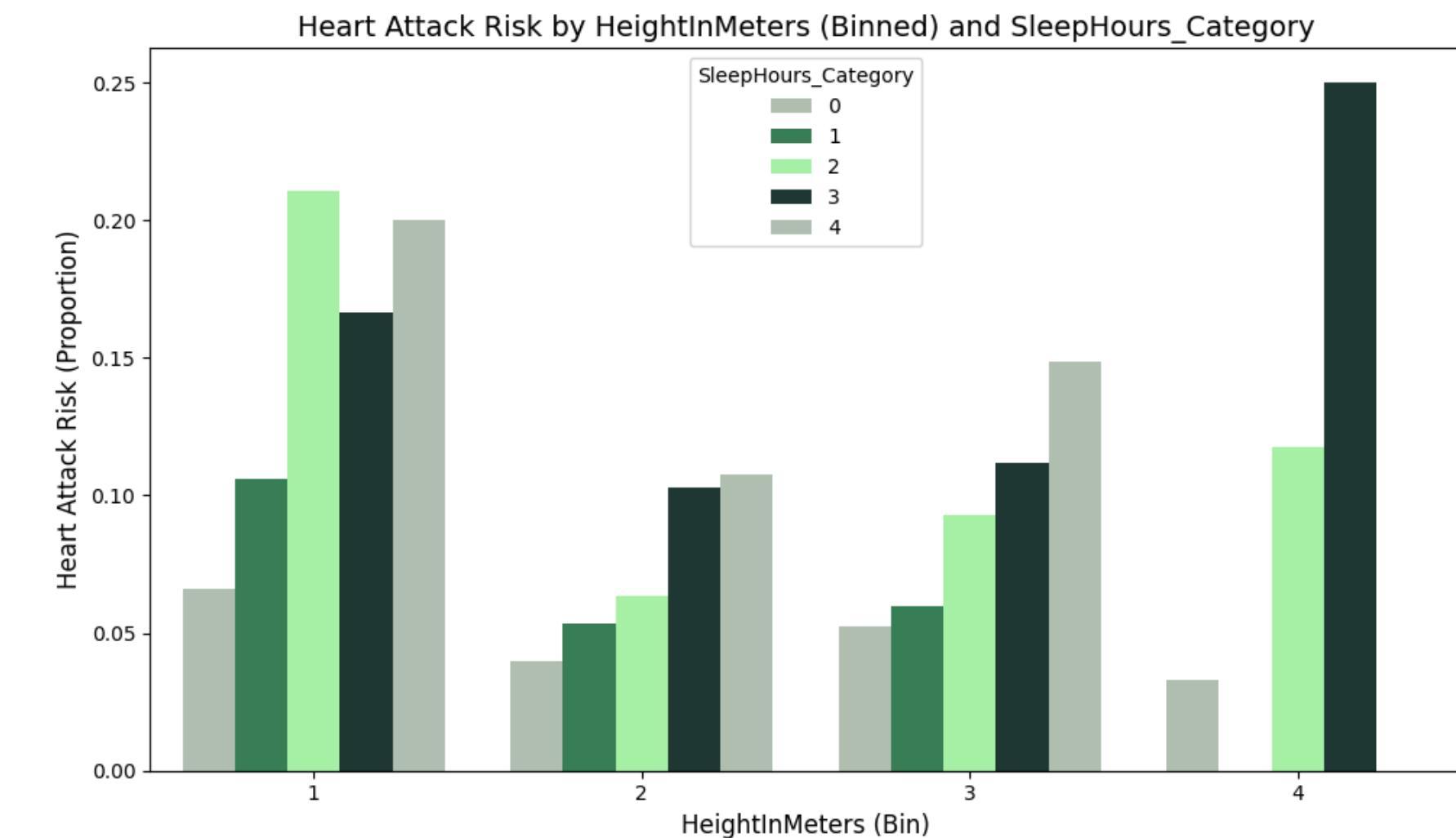
# 特徵工程\_交互作用項設立

- 進行交互作用視覺化，找尋合適的交互作用項建立

## 體重跟身高的交互作用



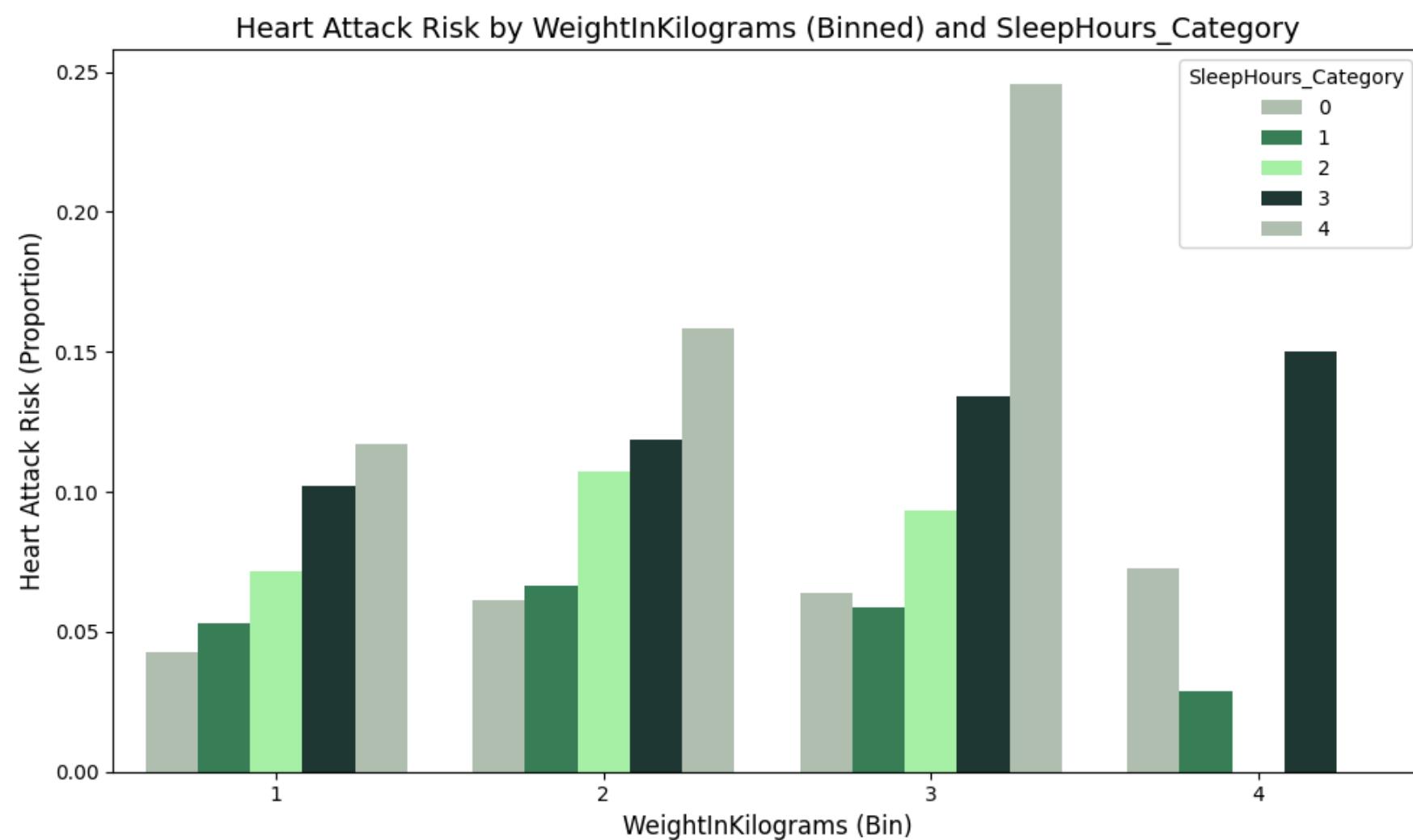
## 睡眠時間跟身高的交互作用



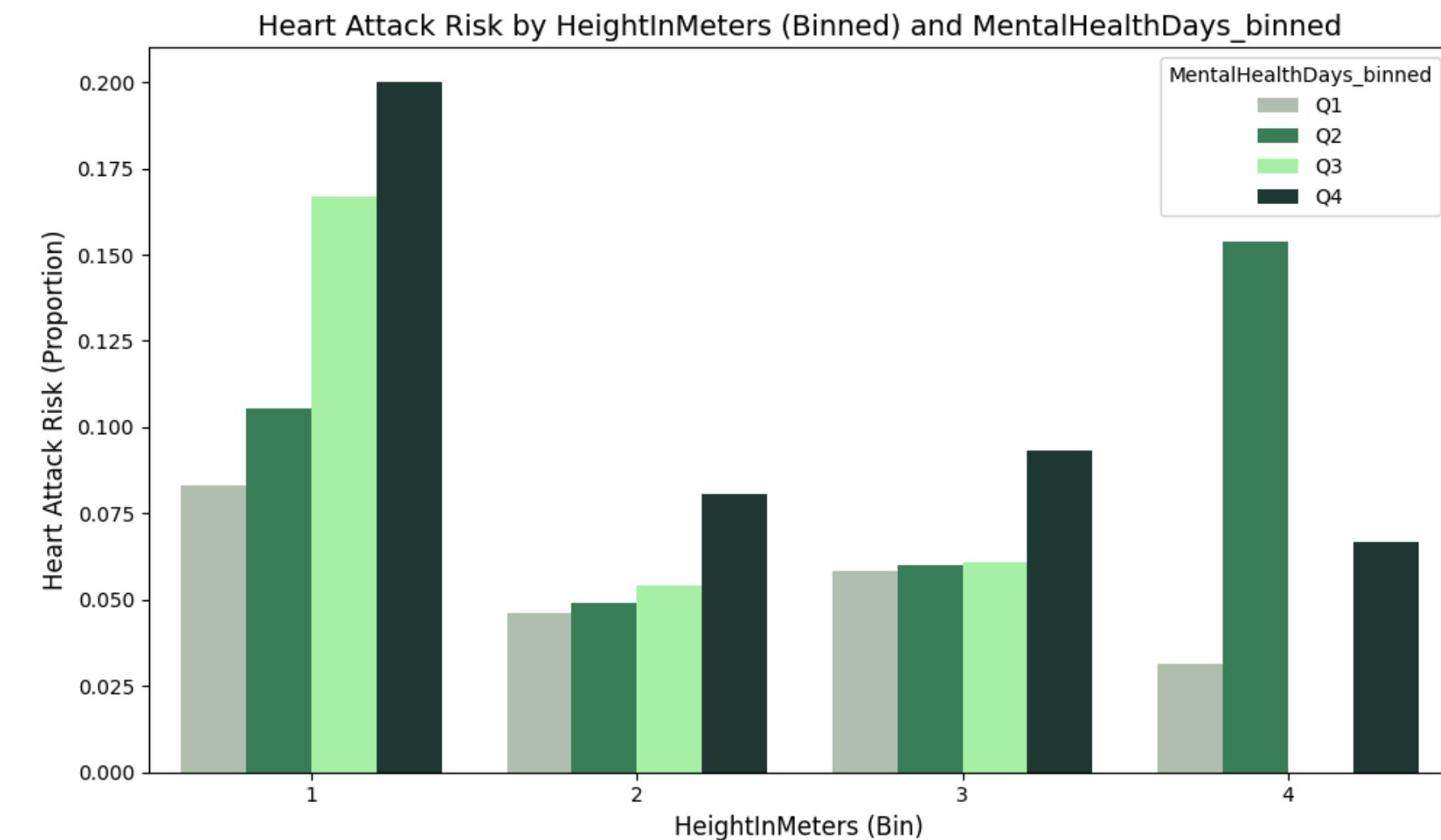
# 特徵工程\_交互作用項設立

- 進行交互作用視覺化，找尋合適的交互作用項建立

## 睡眠時間跟體重的交互作用



## 心靈健康與身高的交互作用



# Agenda

1

商業定位與目標

2

資料探索與預處理

3

特徵工程

4

模型選擇與訓練

5

商業應用

6

結論

# 模型訓練架構

## 定義目標函數

以極大化國泰收  
益為主要目標

## 訓練模型

分出訓練集、測  
試集 ( 80:20 ) ,  
並且驗證找出最  
佳參數

## 收益結果比較

比較有模型訓練  
結果收益跟無模  
型結果收益

## 特徵重要性

列出特徵重要性  
以利後續商業應  
用分析

使用 H2O AutoML, HyperOpt, SMOTE, GAN 等技術選擇模型與找尋最佳參數

# 目標函數設計

以 60 歲女性投保「樂鍾心」為例，保額 100 萬元、繳費 20 年期，每年實繳保險費 16,582 元 (取自國泰新聞稿)

$$\text{Maximize E [Revenue]} = 271,139 \times N(TN) - 401,832 \times N(FN)$$

模型預測為 0

- 如果真的是 0 (TN)  $\rightarrow$  收益 = 271,139
- 如果其實是 1 (FN)  $\rightarrow$  收益 = -401,832

模型預測為 1

- 須請用戶另行提供健康指標報告，暫時不予以保險

保費折現

FV	PMT	I/Y	N	PV
N (# of periods)	20			
I/Y (Interest per year)	2	%		
PMT (Periodic Payment)	\$-16,582			
FV (Future Value)	\$0			
Results				
PV = \$271,139.47				

理賠折現

FV	PMT	I/Y	N	PV
N (# of periods)	20			
I/Y (Interest per year)	2	%		
PMT (Periodic Payment)	\$0			
FV (Future Value)	\$-1,000,000			
Results				
PV = \$672,971.33				

# AutoML 最佳模型

根據目標函數  
設計對應權重

TN reward: +271,139  
FN cost: -401,83

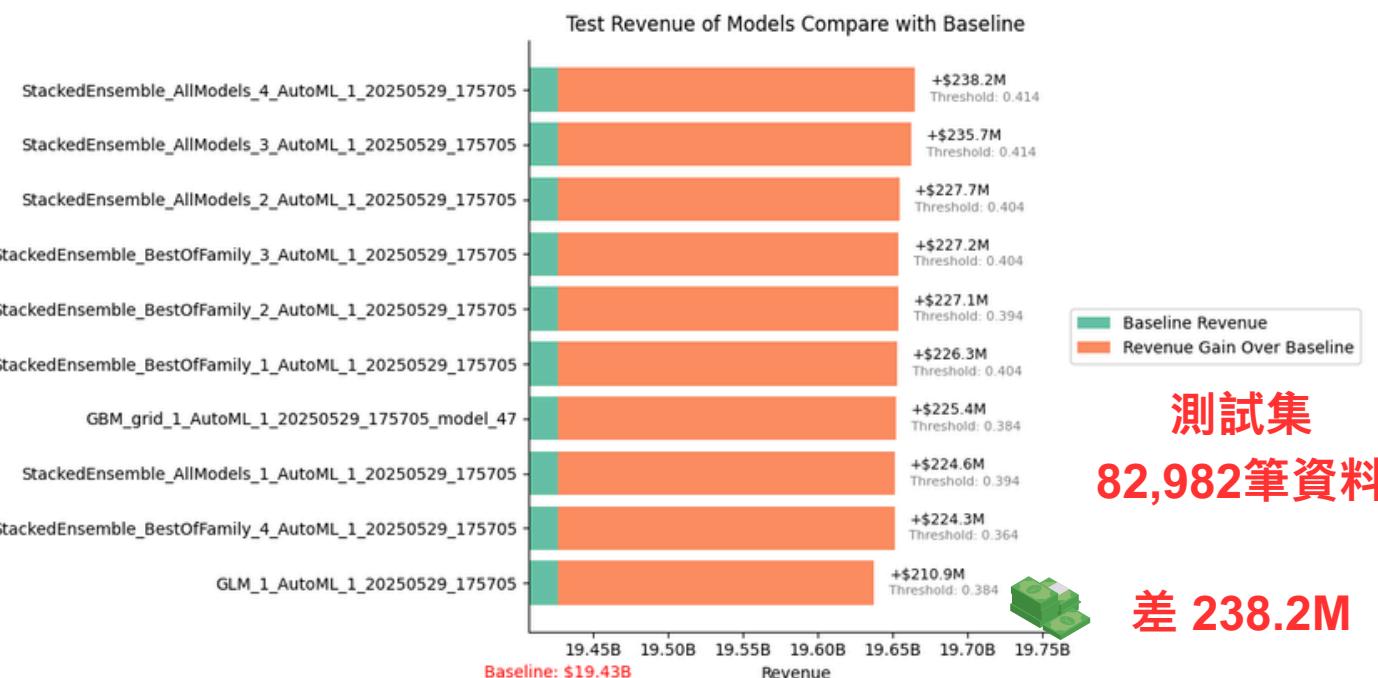
對錯誤分類樣本的懲罰  
class\_0\_weight = 271,139  
class\_1\_weight = 401,832

class 0, class 1  
ratio = [0.5976, 0.4024]

轉換懲罰加權  
penalty\_ratio = [0.4, 0.6]

最後  
對閾值進行最佳化  
以作為判斷[0,1]標準

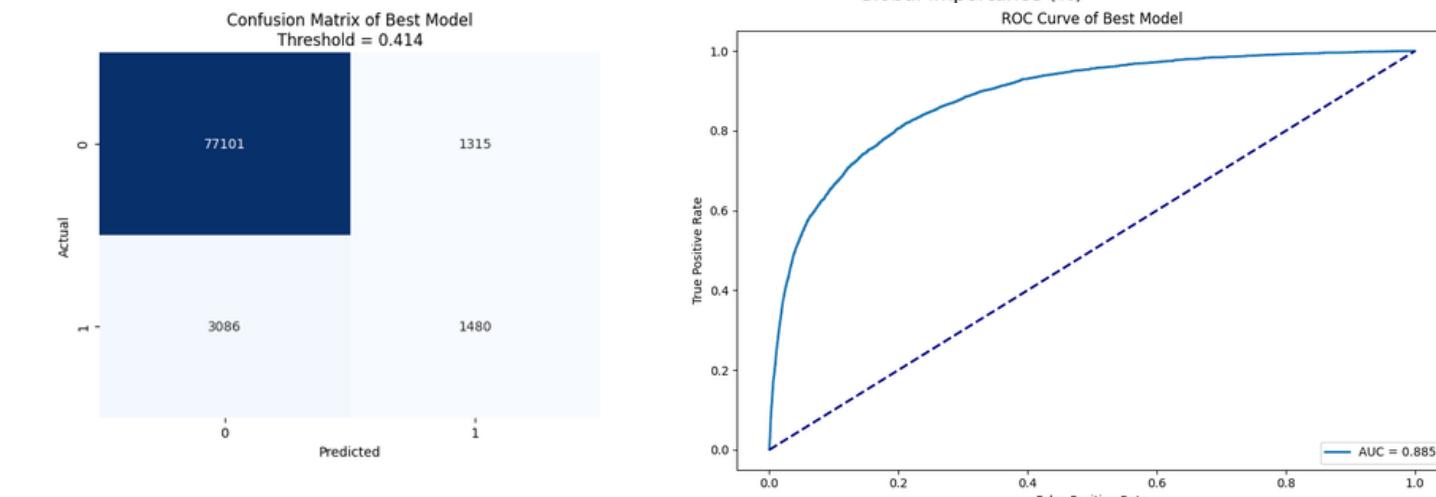
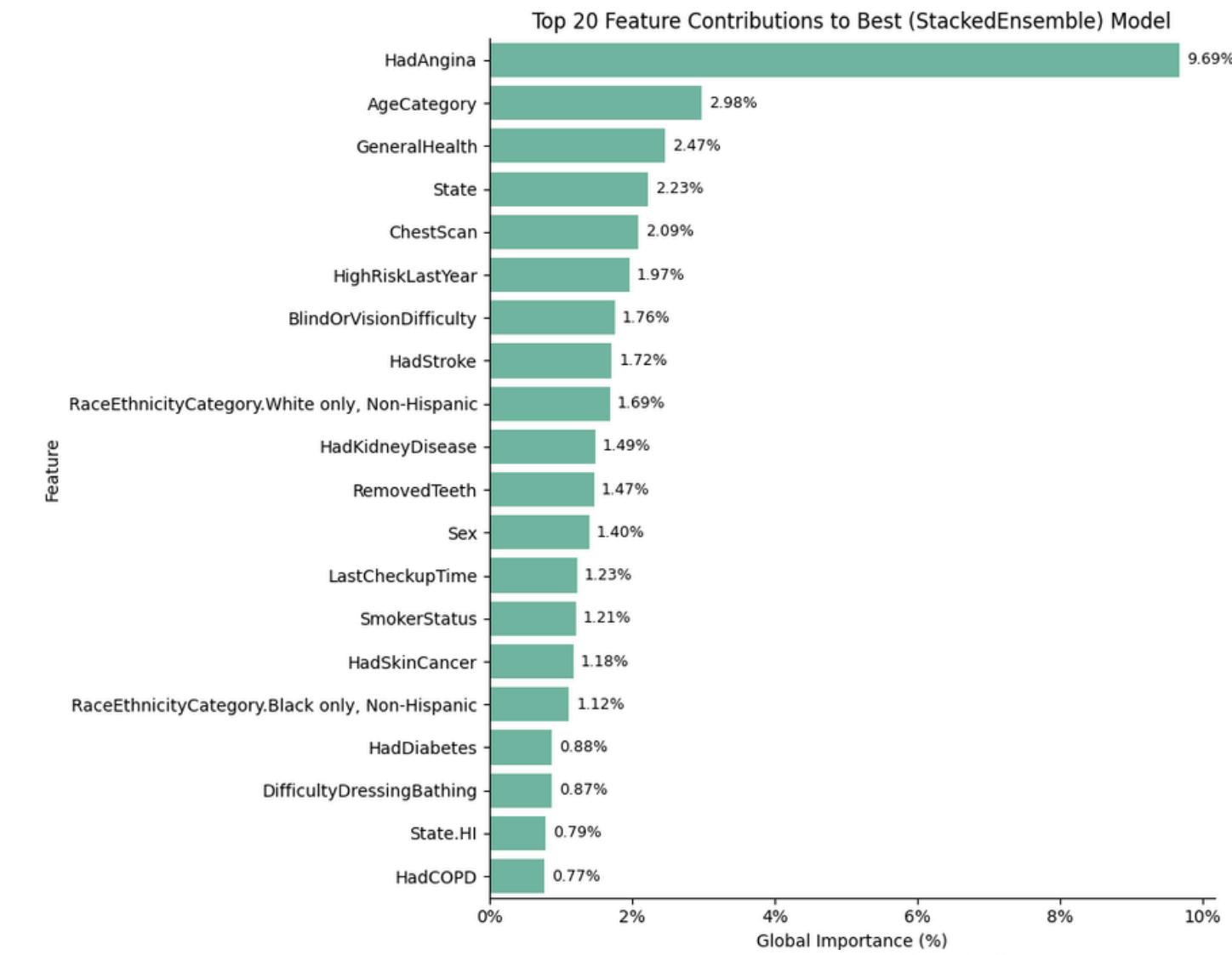
## Top Models vs No model 收益比較



## Best Model Details

Top Ensembled models	global_percentage	Accuracy	94.70%
DeepLearning_grid_3	18.58%	Precision	52.95%
GBM_grid_1_model_6	6.00%	Recall	32.41%
DRF_1_AutoML	5.95%	F1 Score	0.4021
GBM_grid_1model_22	5.74%	AUC	0.8854
GLM_1	5.14%	Threshold	0.414

## 重要特徵 & Performance



# LightGBM 目標函數調參

直接基於極大化收入  
自訂指標

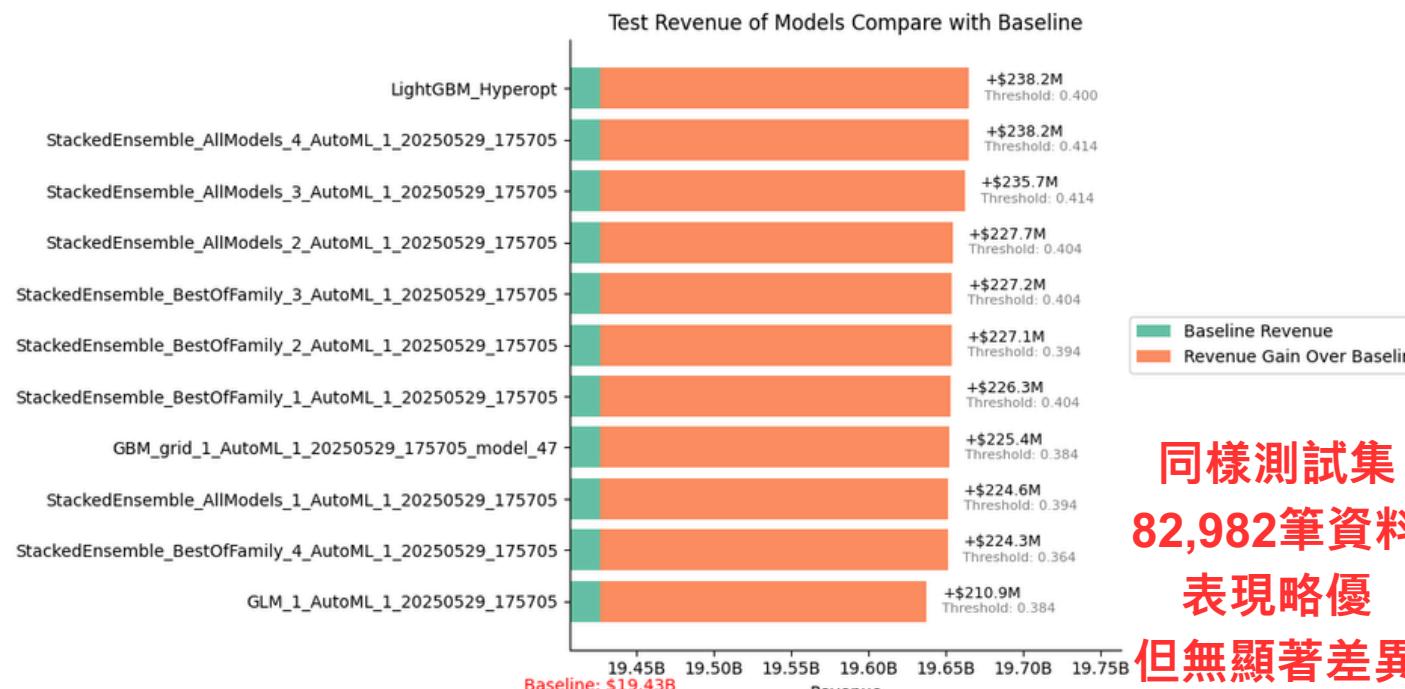
定義了目標函數

```
def total_revenue_score:
    sum(
        271_139 if (yp == 0 and
yt == 0) else
        -401_832 if (yp == 0 and
yt == 1) else
    0)
```

給定的超參數訓練模型

最後  
對閾值進行最優化  
以作為判斷[0,1]標準

## LightGBM\_Hyperopt vs Other models 比較



同樣測試集  
82,982筆資料  
表現略優  
但無顯著差異

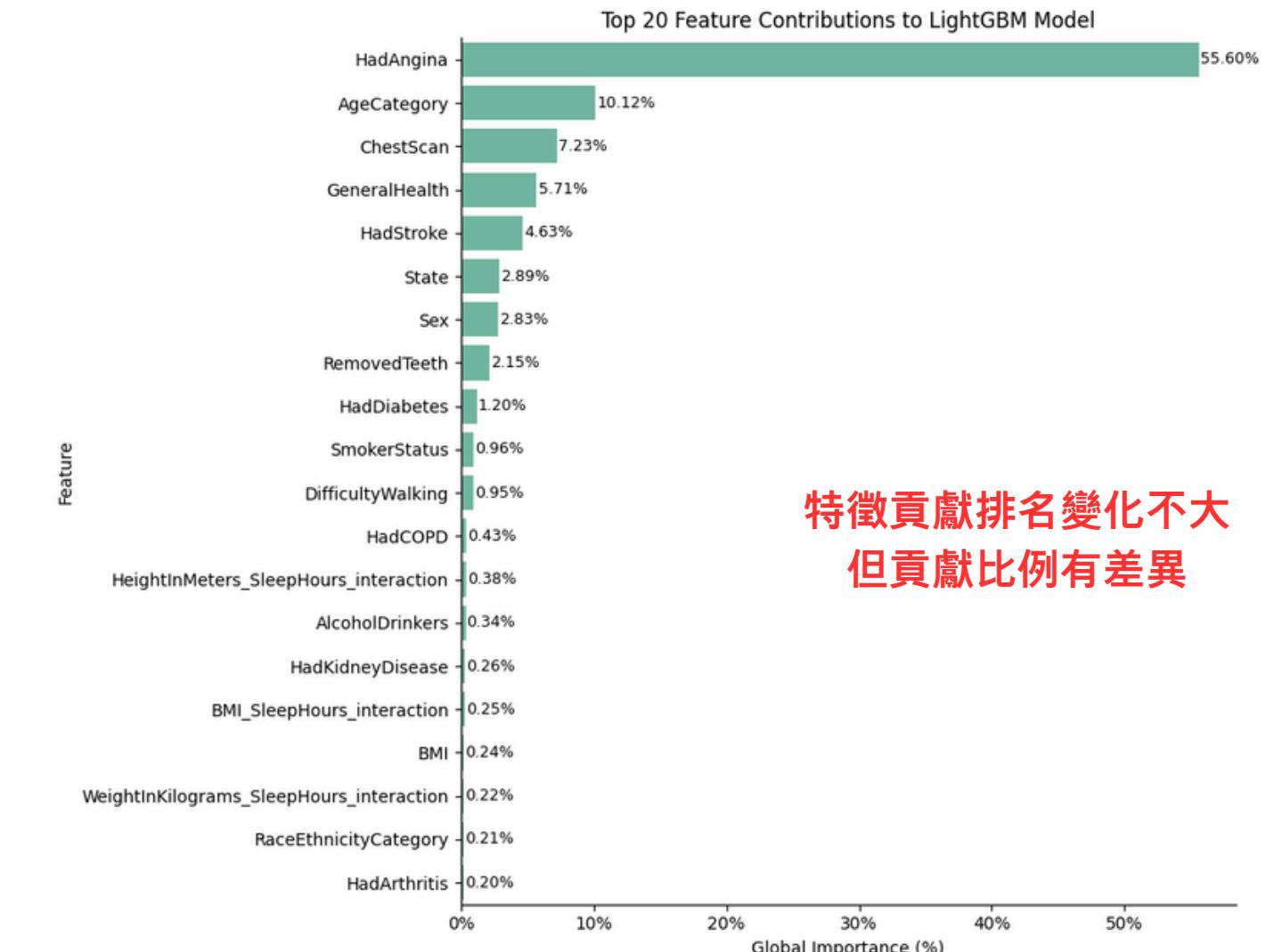
## Best Model Details

### 超參數優化空間

```
space_lgb = {
    'num_leaves': hp.quniform('num_leaves', 20, 150, 1),
    'max_depth': hp.quniform('max_depth', 3, 12, 1),
    'learning_rate': hp.loguniform('learning_rate', -5, 0),
    'min_child_samples': hp.quniform('min_child_samples', 10, 100, 1),
    'subsample': hp.uniform('subsample', 0.6, 1.0),
    'colsample_bytree': hp.uniform('colsample_bytree', 0.6, 1.0),
    'objective': 'binary',
    'metric': 'binary_logloss',
    'verbose': -1
}
```

Accuracy	94.63%
Precision	52.07%
Recall	33.59%
F1 Score	0.4086
AUC	0.885
Threshold	0.4

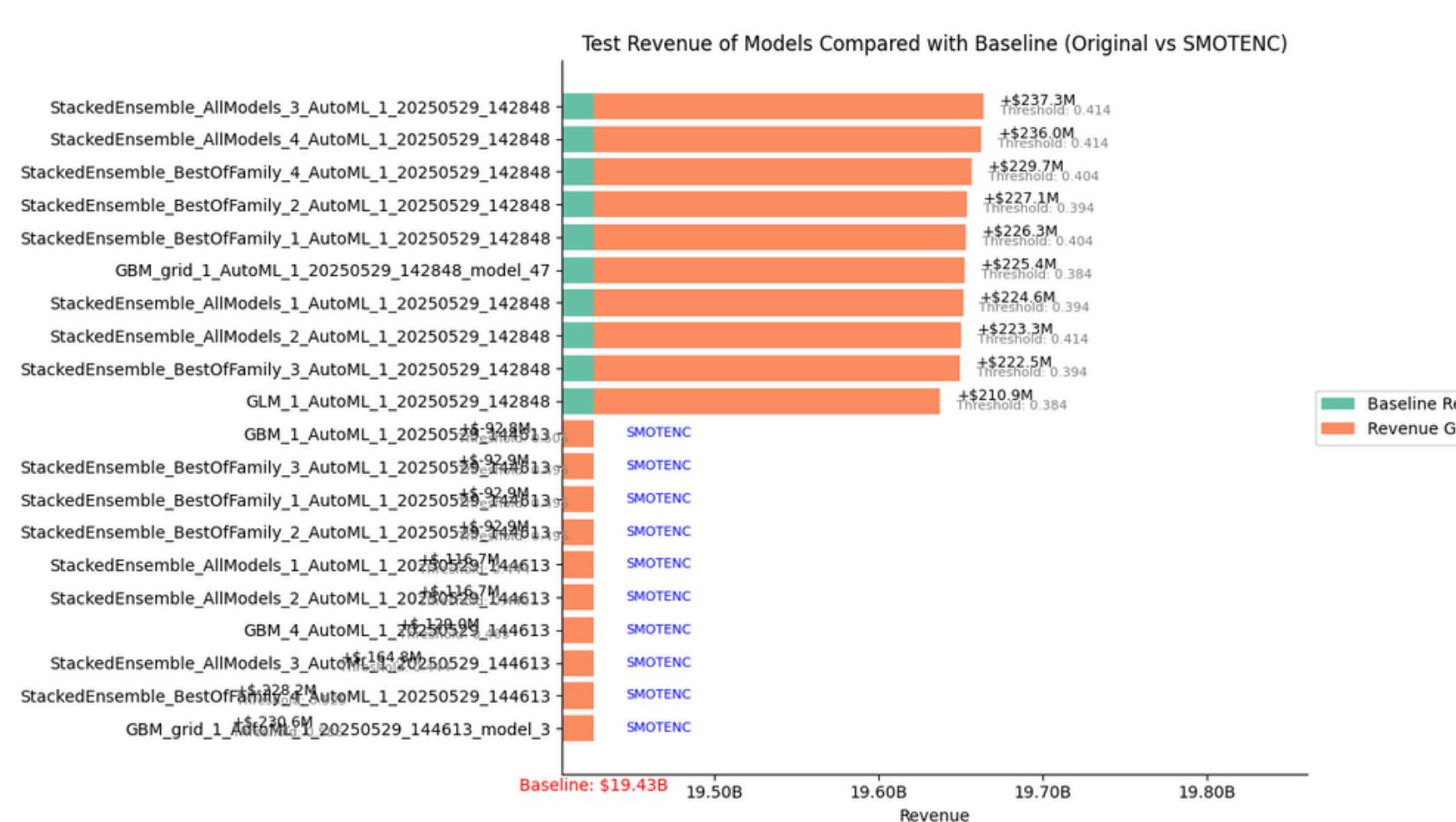
## 重要特徵 & Performance



特徵貢獻排名變化不大  
但貢獻比例有差異

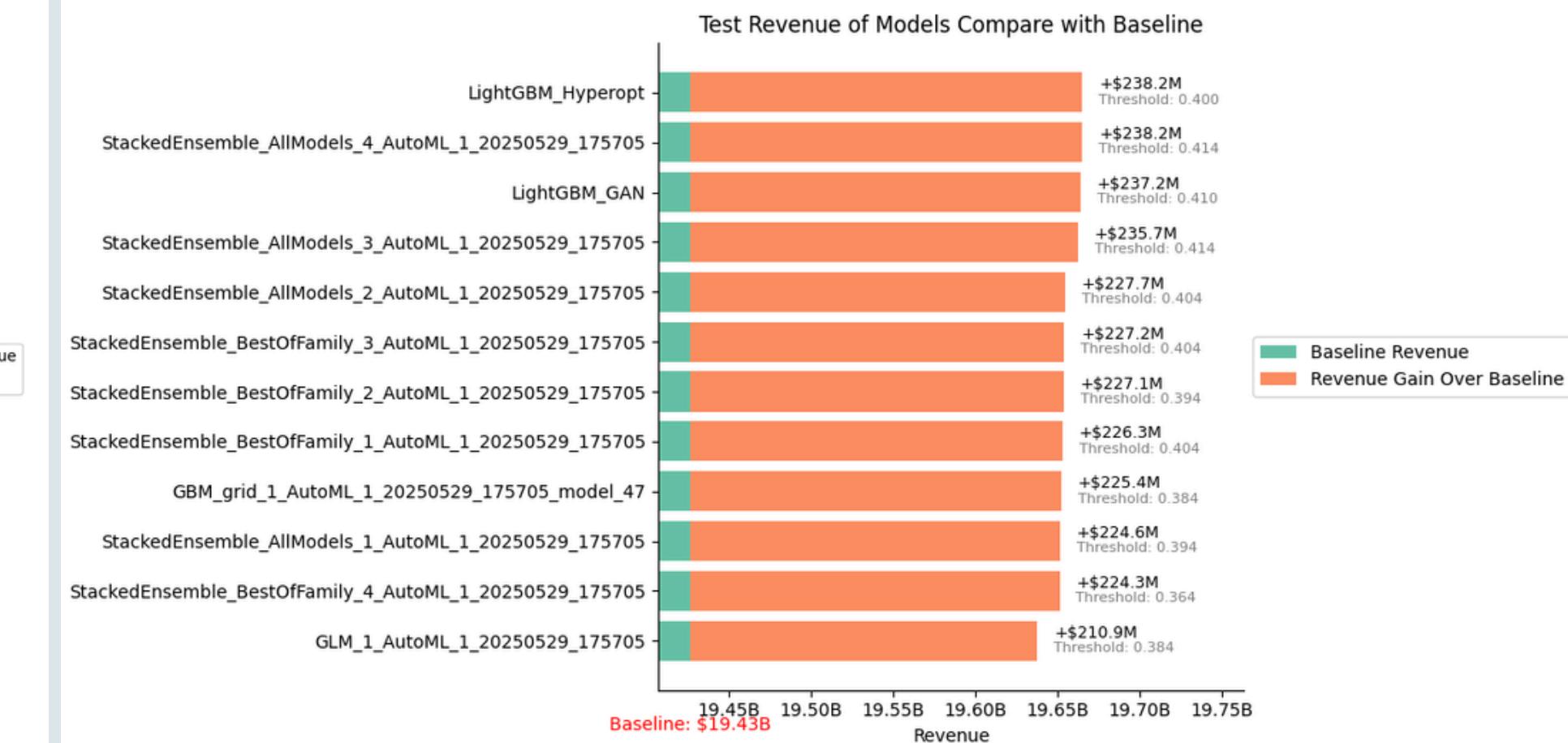
# 為解決資料不平衡問題，搭配 SMOTE 與 GAN 比較收益

## 使用 SMOTE 收益反下降



SMOTE 的過採樣方法可能未考慮學習的適配性  
甚至帶來反效果

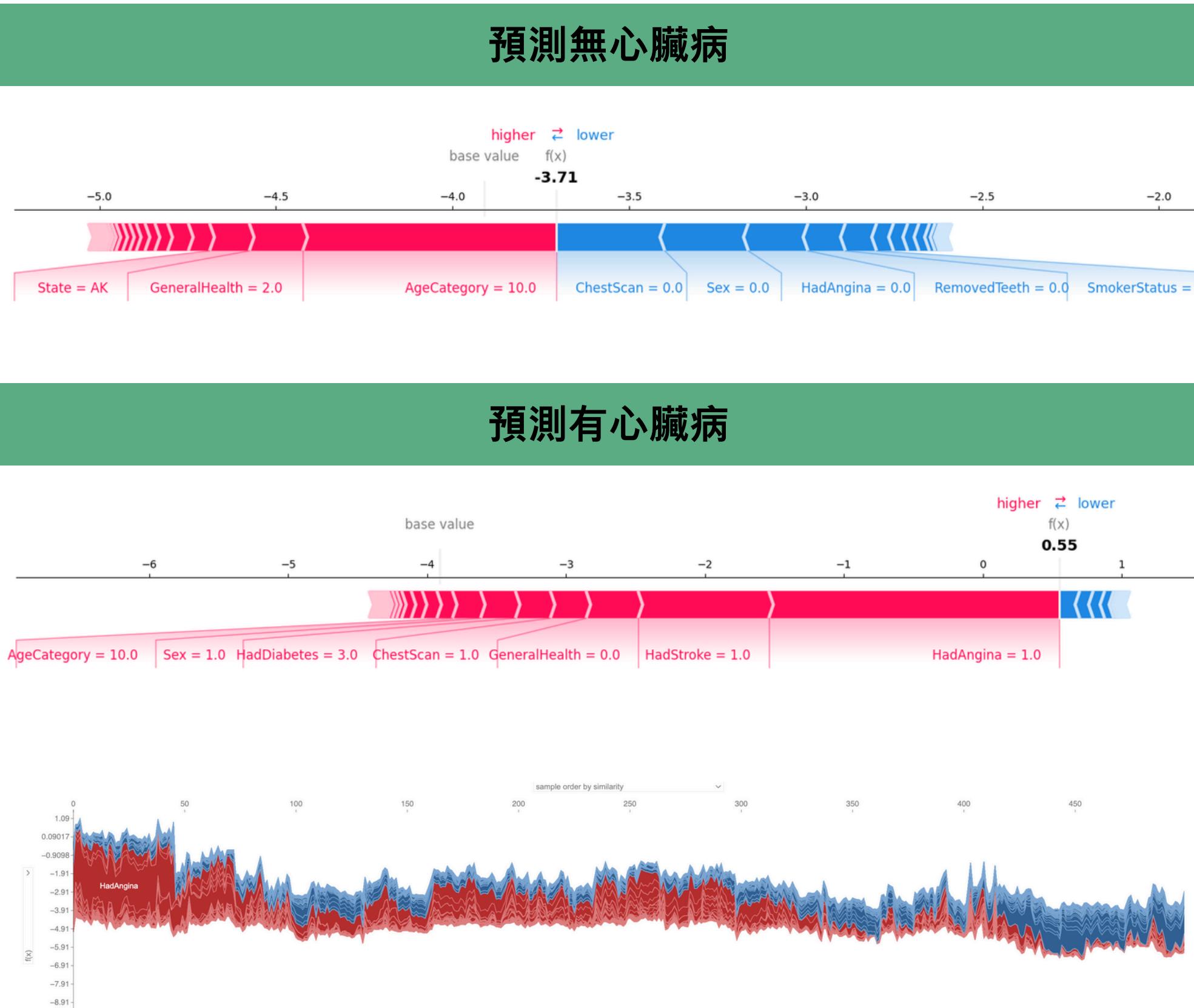
## LightGBM 搭配 GAN 收益也無明顯提升



GAN 透過對抗學習產生的資料能提供更高品質的少數類樣本  
雖然提升有限但未來可考慮優化以提升效果

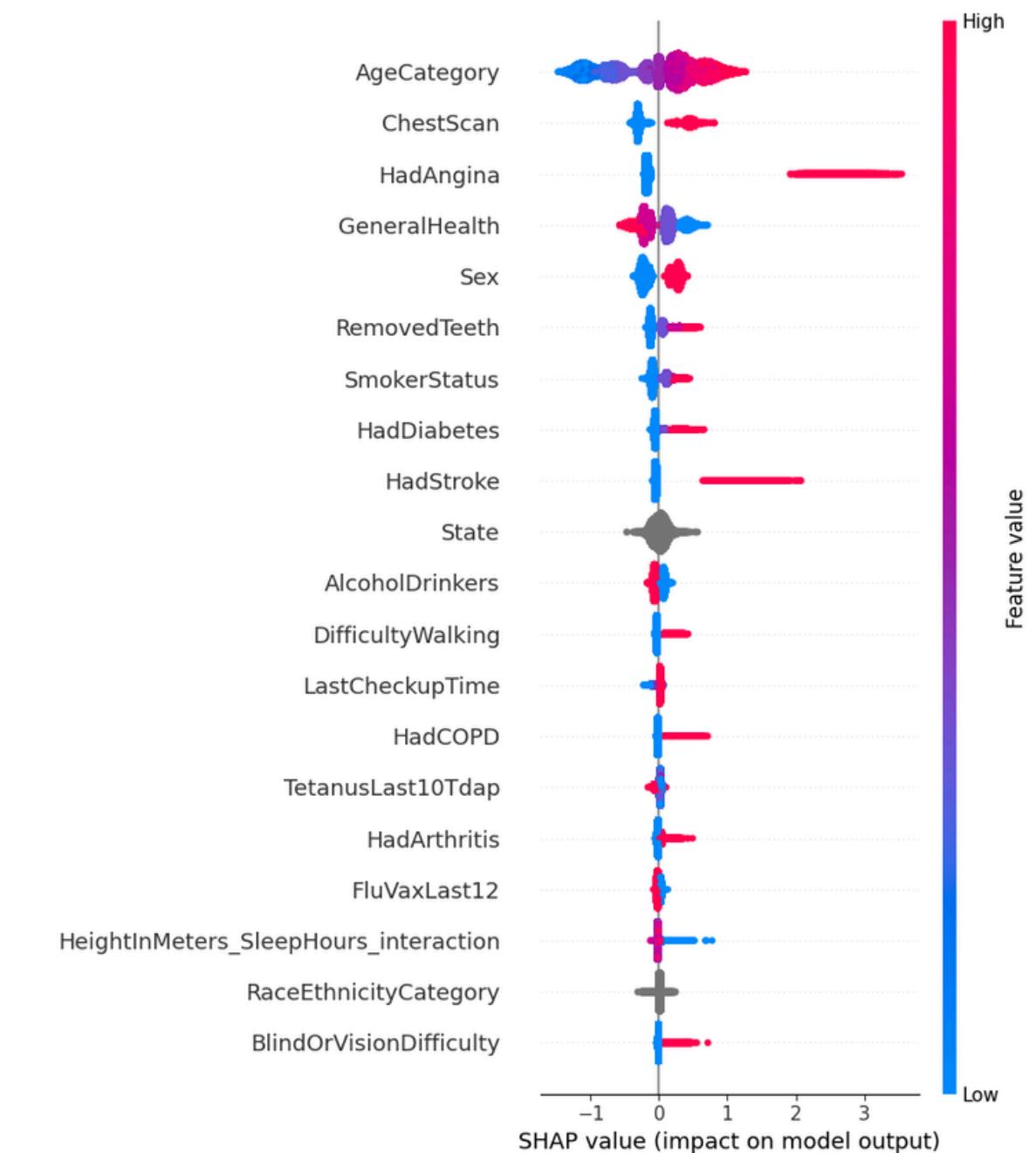
# SHAP 分析

SHAP  
力圖  
可視化



SHAP  
互動圖

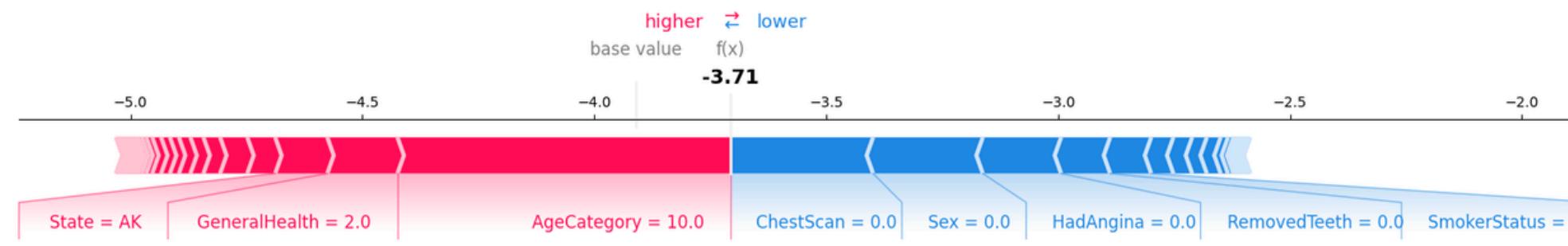
## SHAP 特徵貢獻度



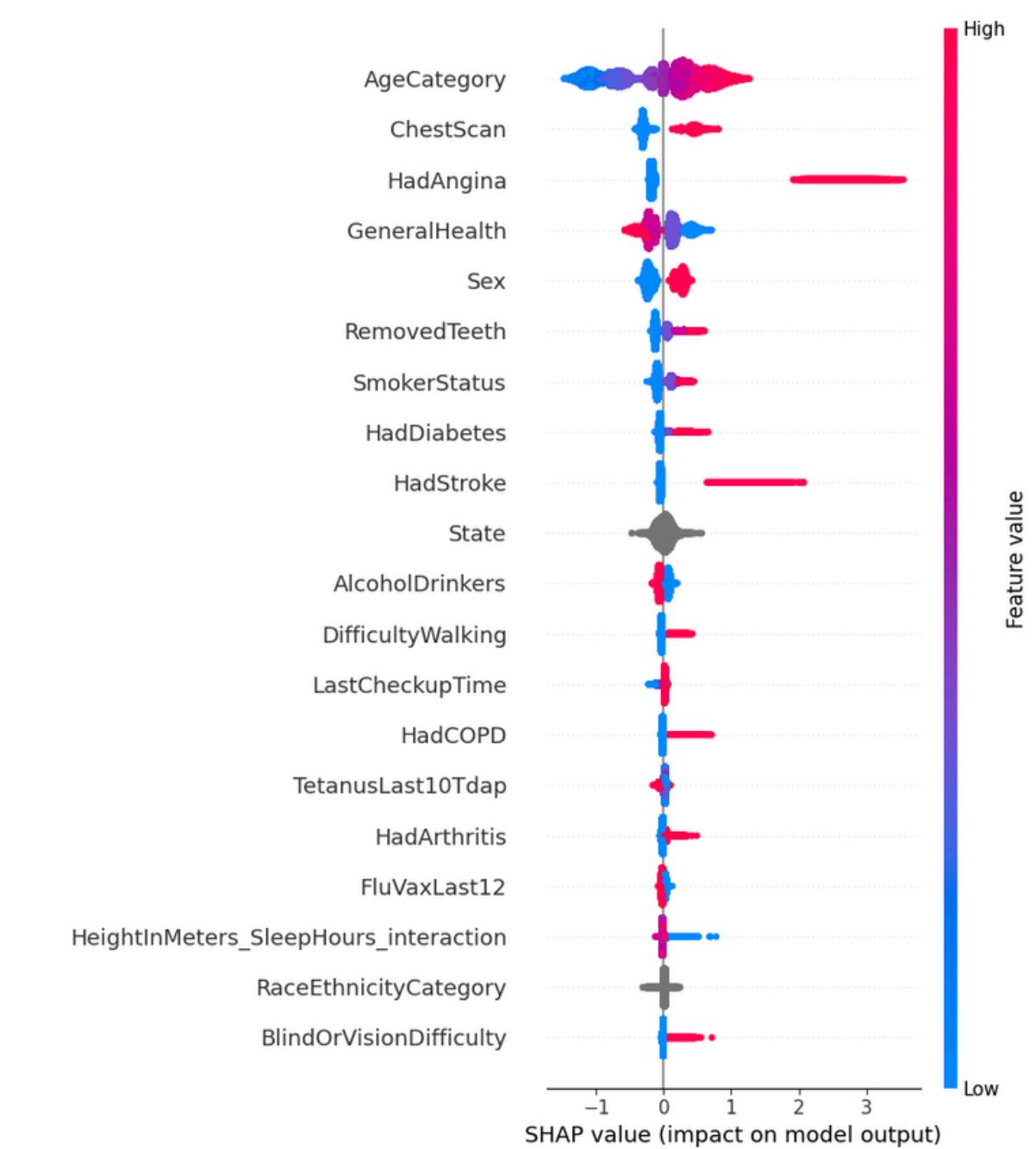
# SHAP 分析

SHAP  
力圖  
可視化

預測無心臟病

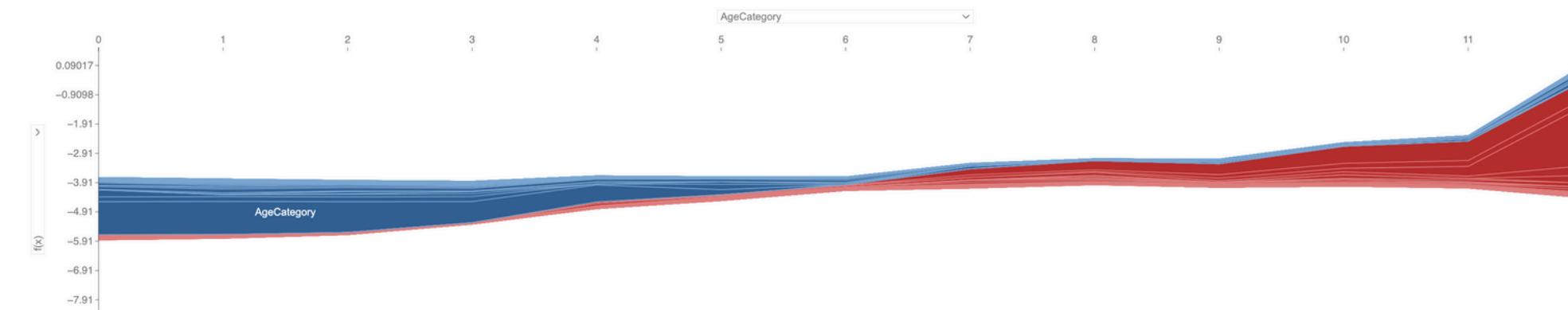


SHAP 特徵貢獻度



SHAP  
互動圖

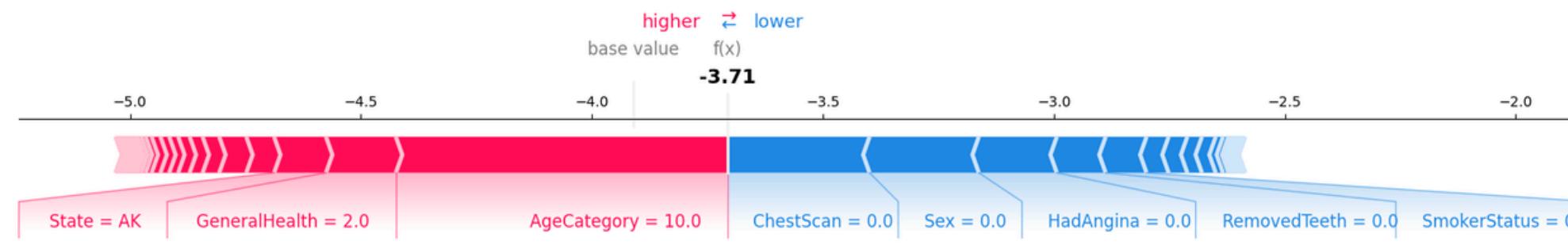
預測有心臟病



# SHAP 分析

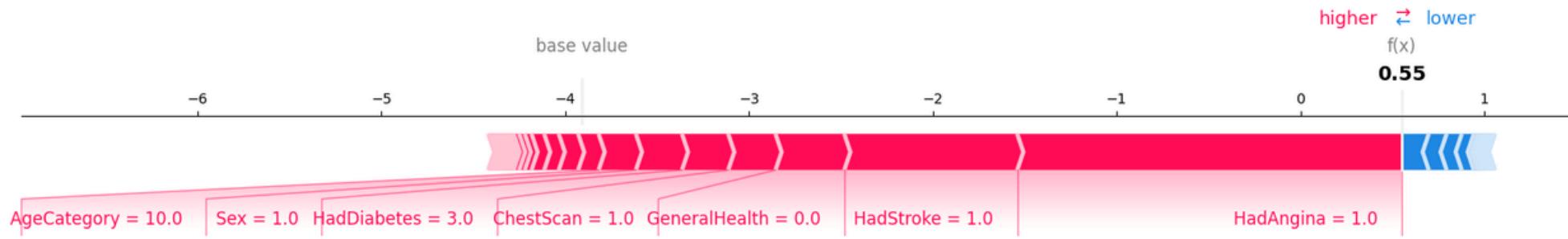
SHAP  
力圖  
可視化

## 預測無心臟病

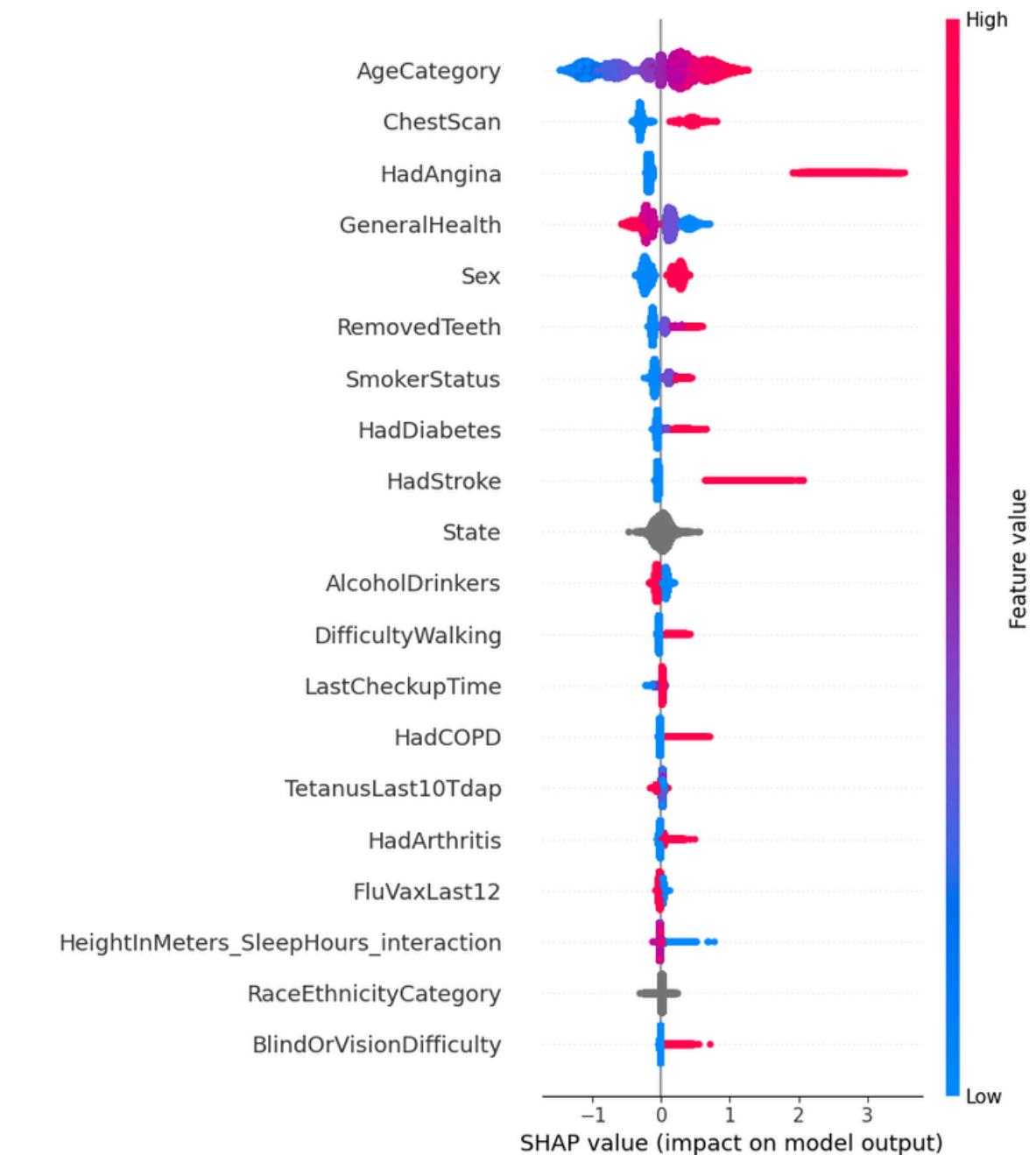


SHAP  
互動圖

## 預測有心臟病



## SHAP 特徵貢獻度



# Agenda

1

商業定位與目標

2

資料探索與預處理

3

特徵工程

4

模型選擇與訓練

5

商業應用

6

結論

透過積極管理這些風險因子，不僅能顯著改善健康狀況，降低罹患相關疾病的風險，亦有可能對未來的保險成本產生正面影響

## 主要可控因子及其管理策略

心絞痛 (Angina)

9.69 %

整體健康

睡眠

1.235 %

運動

1.235 %

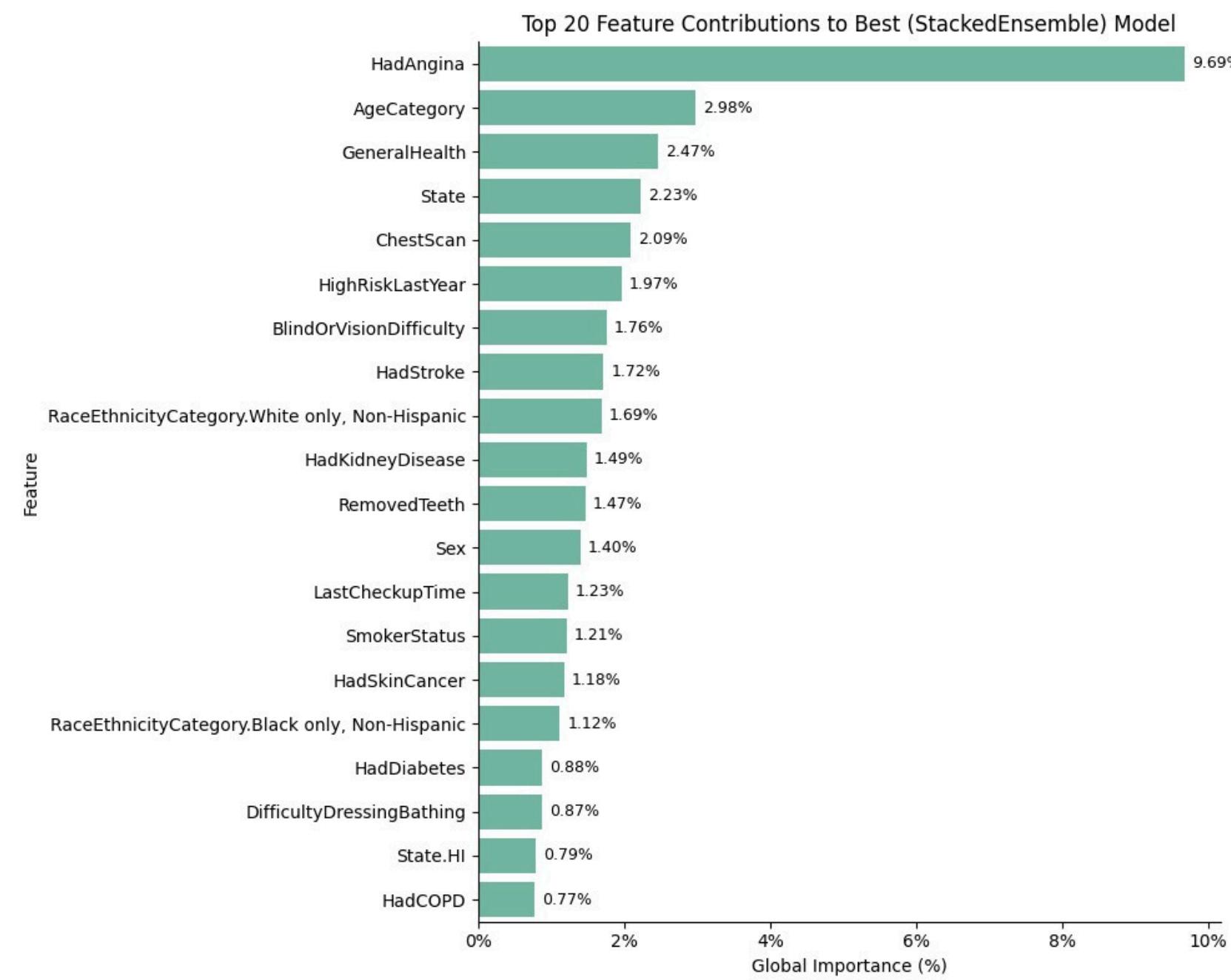
中風 (stroke)

1.72 %

抽菸 (smoke)

1.21 %

## 特徵重要性



心絞痛 (Angina)是冠狀動脈疾病的常見症狀，發生原因是心肌供血不足，通常由動脈狹窄或阻塞引起。若不妥善管理，可能導致心肌梗塞等嚴重後果

主要風險因子：吸菸、糖尿病、肥胖、家族病史、年齡（男性 > 45 歲，女性 > 55 歲）

#### 規律運動

Cleveland Clinic 指出，每週至少150分鐘中等強度運動有益

#### 戒菸

American Heart Association 研究顯示，戒菸後心血管事件風險可降低50%

#### 體重管理

肥胖會增加心臟負荷，控制BMI在18.5-24.9範圍內

#### 外部數據

戒菸、健康飲食已被證實能顯著降低冠狀動脈疾病的進展風險，這些措施可降低 30 % 以上的心絞痛風險

#### 減少保費

$0.0969 * 0.3 * 1,000,000 = 29,070$

# 長期睡眠不足或睡眠品質不佳（如睡眠呼吸中止症）與高血壓、心臟病、中風及糖尿病風險增加有關

主要風險因子：焦慮、不規律作息、年齡、生活方式（如輪班工作）

## 避免刺激物

避免咖啡因、酒精和大量進食。BIDMC - How Does Sleep Help Your Heart? 建議睡前2小時避免咖啡因

## 規律作息

盡可能在每天同一時間上床睡覺和起床，即使在週末也是如此

## 睡眠障礙

若有打鼾嚴重、日間嗜睡等問題，應諮詢醫師是否需要進行睡眠評估

## 外部數據

睡眠不規律的人罹患心臟病的風險幾乎是睡眠規律者的兩倍。ISD Health Solutions 提供相關數據

## 減少保費

$0.01235 * 0.5 * 1,000,000 = 6,175$

# 缺乏體能活動是心血管疾病、糖尿病、肥胖等多種慢性病的主要風險因子

主要風險因子：久坐生活方式、缺乏運動習慣、心理因素（如缺乏動機）

## 設定目標

成人每週至少進行150分鐘中等強度有氧運動（如快走、游泳、騎自行車），或75分鐘高強度有氧運動

## 循序漸進

可從短時間、低強度開始，逐漸增加運動量。Johns Hopkins Medicine 建議從每日10分鐘開始

## 減少久坐

每隔一段時間起身活動，如每小時站立5分鐘，減少久坐可降低全因死亡風險

## 外部數據

僅僅每天多活動10分鐘也能帶來健康益處，增加每日步數有助於降低全因死亡風險。

## 減少保費

$0.01235 * 0.5 * 1,000,000 = 6,175$

# 中風 (Stroke) 是因腦部血管阻塞或破裂，導致腦細胞缺氧受損，可能造成永久性殘疾甚至死亡，平均每 42 分鐘就有 1 人死於中風

主要風險因子：吸菸、高血壓、高膽固醇、年齡（特別是>65歲）、性別（男性風險較高）

## 控制高血壓

高血壓是中風最重要的風險因子，應定期監測並透過藥物治療加以控制。CDC 建議保持血壓  $<120/80 \text{ mmHg}$

## 戒菸

吸菸會損害血管壁，增加血栓風險，戒菸能大幅降低中風機率

## 膽固醇管理

控制低密度脂蛋白水平，預防動脈粥狀硬化，建議LDL $<100 \text{ mg/dL}$

## 外部數據

美國疾病管制與預防中心 (CDC) 指出，高達80%的中風可以透過健康的生活方式和藥物治療來預防

## 減少保費

$$0.0172 * 0.8 * 1,000,000 = 13,760$$

菸草煙霧（Smoke）含有超過 7,000 種化學物質，其中至少 70 種為已知的致癌物。根據 CDC，抽菸每年在美國造成超過 480,000 人死亡

主要風險因子：抽菸量、持續時間、年齡、遺傳因素、環境暴露

尼古丁療法

使用尼古丁貼片、口香糖、錠劑或吸入器來減輕戒斷症狀

藥物治療

醫師可能開立伐尼克靈（Varenicline）或安非他酮（Bupropion）等藥物，幫助減少菸癮

定期健檢

定期進行肺部檢查、癌症篩查（如低劑量CT掃描）和心血管評估，早期發現和治療菸草相關疾病

外部數據

戒菸後，心血管疾病的風險會隨著時間逐漸降低，戒菸 5 年後，心血管疾病的風險會顯著降低，10~15 年後，風險會接近從未抽菸者的水平

減少保費

$0.0121 * 0.5 * 1,000,000 = 6,050$

# 獎勵機制是外溢保單的核心

## 獎勵機制分層進行

### 點數/健康積分系統

- 保戶透過「FIT BACK健康吧」等平台，綁定穿戴式裝置(如Apple Watch, Fitbit, Garmin等)或手機App，自動記錄步數、睡眠等數據
- 達成每日或每週的健康目標(如步數達標、睡眠達標)即可獲得健康點數或積分。參與健康知識問答、完成特定健康任務等也可能獲得點數

### 點數的應用與獎勵

- 小額獎勵/兌換：累積的點數可能用於兌換合作商家(如咖啡、超商商品)的優惠券或商品
- 保費折減/保障增額(外溢效果)：這是外溢保單最主要的獎勵

### 獎勵形式

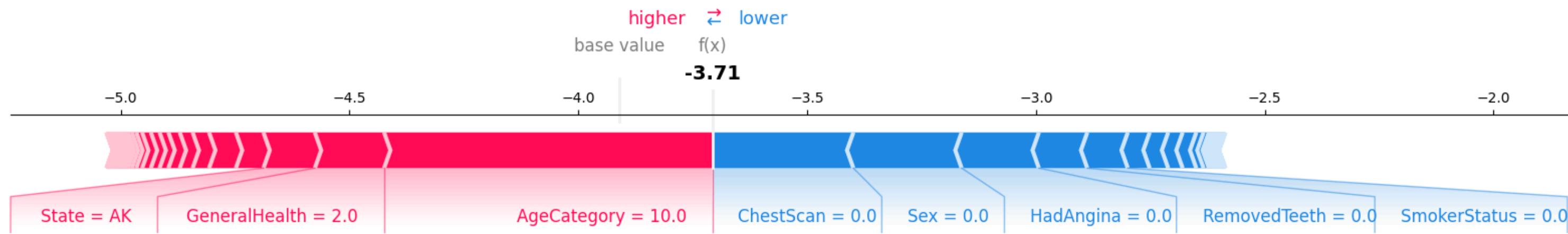
- 保費折減：若達到指定的健康標準，下一年度的續期保費可享有一定比例的折扣(例如2%~10%，視保單條款與達成程度而定)
- 保障增額：保費不變的情況下，提高保險金額(例如增加5%~20%的保障)。
- 現金回饋/獎勵金：直接給予一筆獎勵金

### 機制

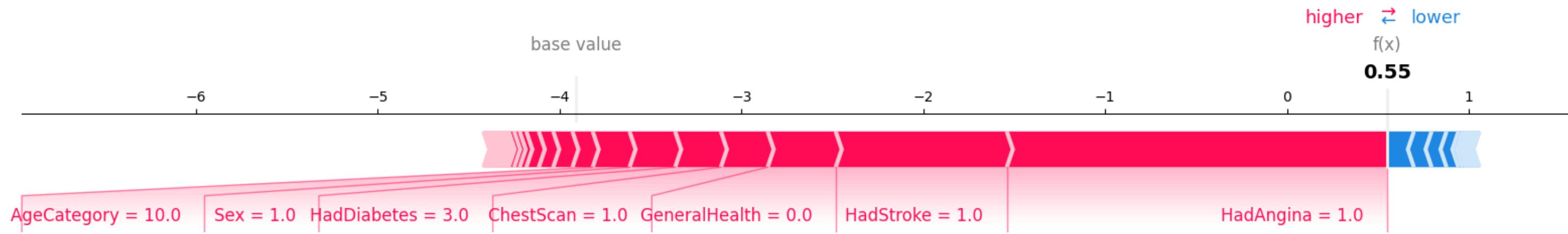
- 保戶在保單年度內累積的健康成果，如年度平均每日步數、健康檢查結果、累積點數等級等會被評估，通常是每年評估一次

# 個案探討

## 預測無心臟病



## 預測有心臟病



General Health (12,350) + Stroke (13,760) + Angina (29,070) = 55,180