國泰金控
Cathay Financial Holdings

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 資料描述

2 預計使用模型

3 預期解決方法

一、 資料描述

資料介紹

Main Objective
預測變數 HadHeartAttack 表示受訪者是否曾經罹患心臟病，屬於 Yes/No 的二元分類問題。

資料來源	來自美國 CDC 的行為風險因子監測系統 Behavioral Risk Factor Surveillance System, BRFSS 為全球最大規模的健康電話調查系統。
原始資料集	包含 400,000 多筆成年受訪者健康相關資料。
此版本	由原始近 300 個變數精簡 至 40 個與心臟病最相關的變數， 提供有缺失值與無缺失值的兩個版本。

缺失值處理

數值型資料：填補為該欄位的中位數

類別型資料：填補為該欄位的眾數

Key Variables

根據 CDC，以下為與心臟病顯著相關的主要變數類別：

慢性病指標

Diabetes,
Asthma,
Kidney Disease,
Skin Cancer,
Other Cancer

身體狀況

BMI,
General Health,
Mental Health,
Physical Health

生活習慣與行為

Smoking,
Alcohol Drinking,
Physical Activity

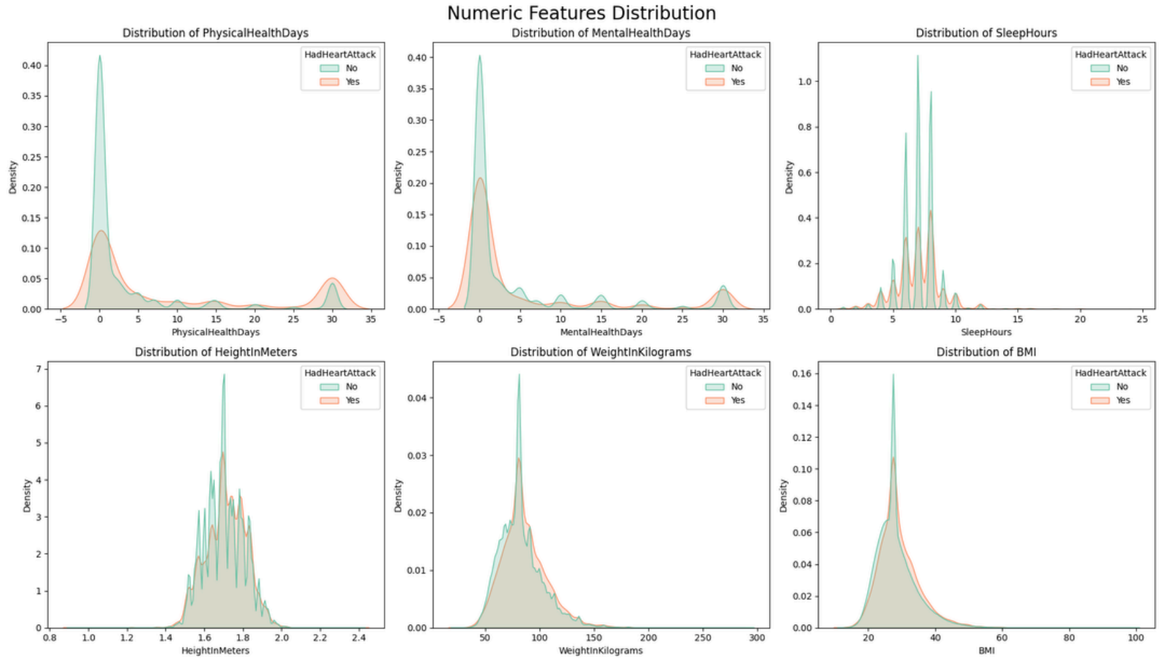
社經背景

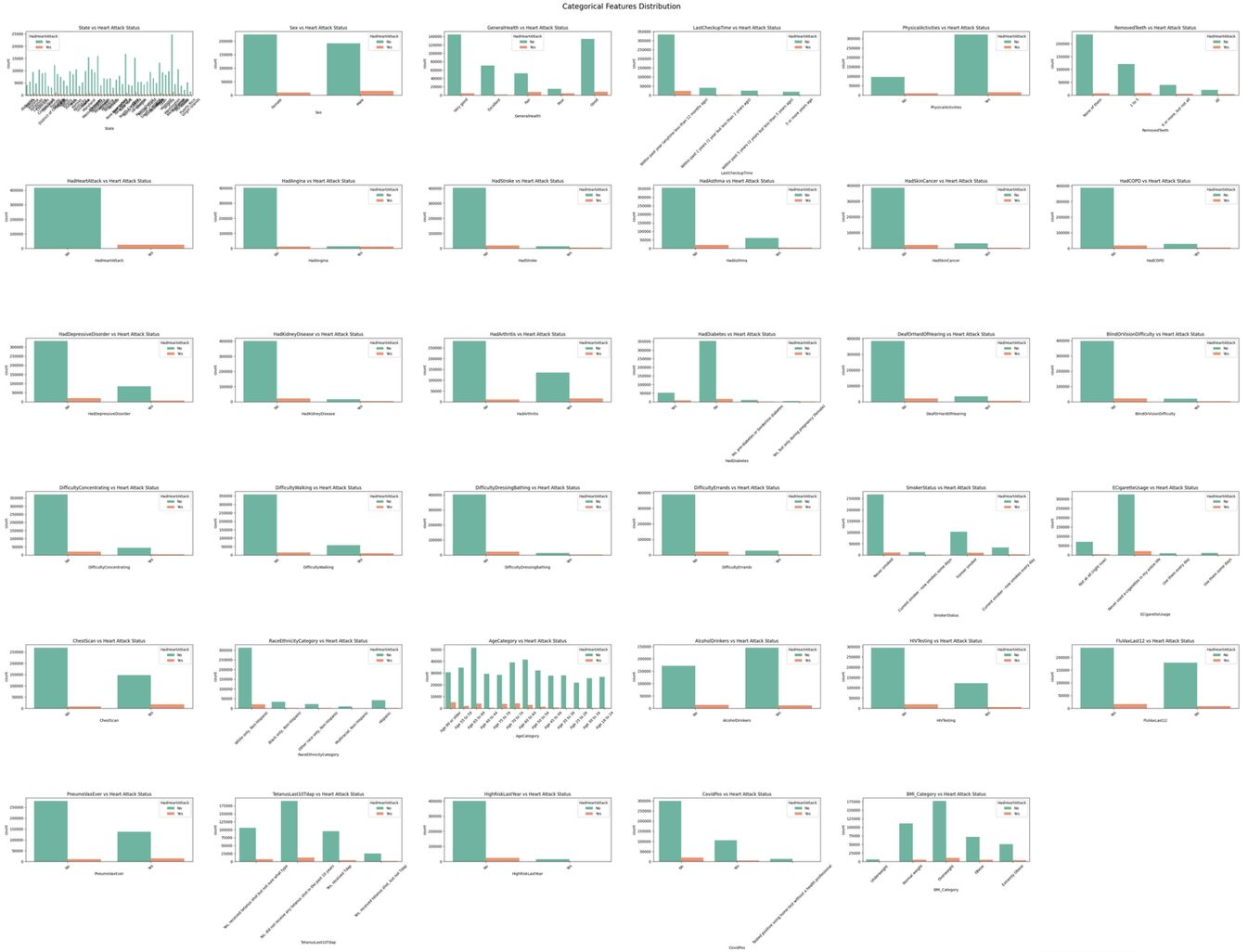
Education Level,
Income Level,
Race,
Gender,
Age Category

二、　預計使用模型

EDA

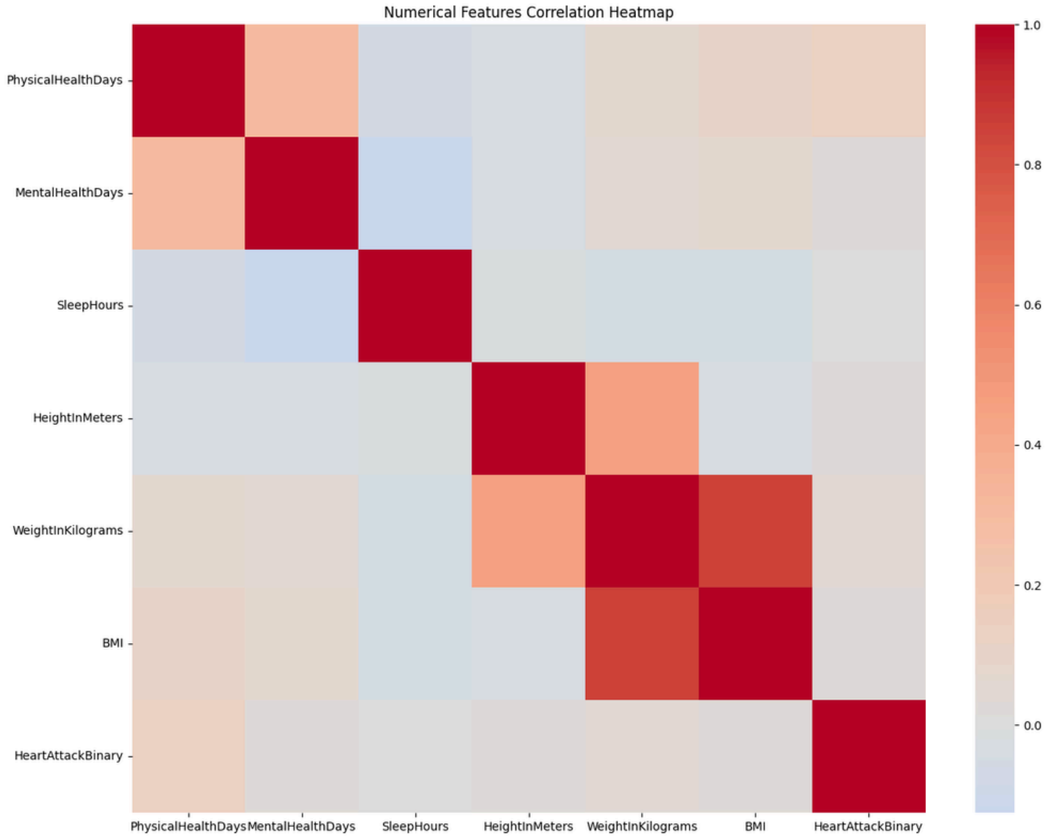
各特徵（數值型、類別型）的分佈
分別以 HadHeartAttack 分開。

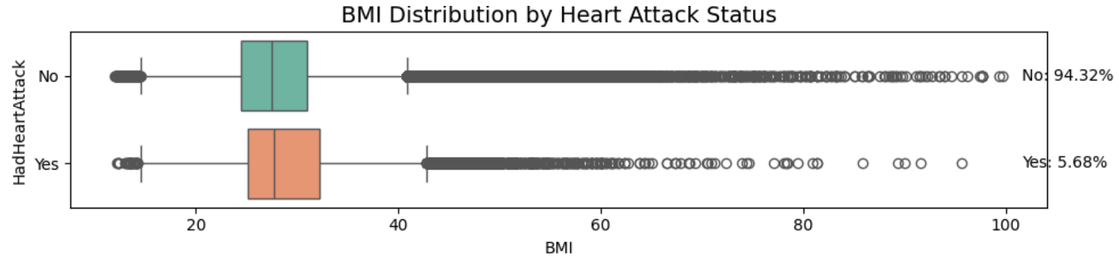


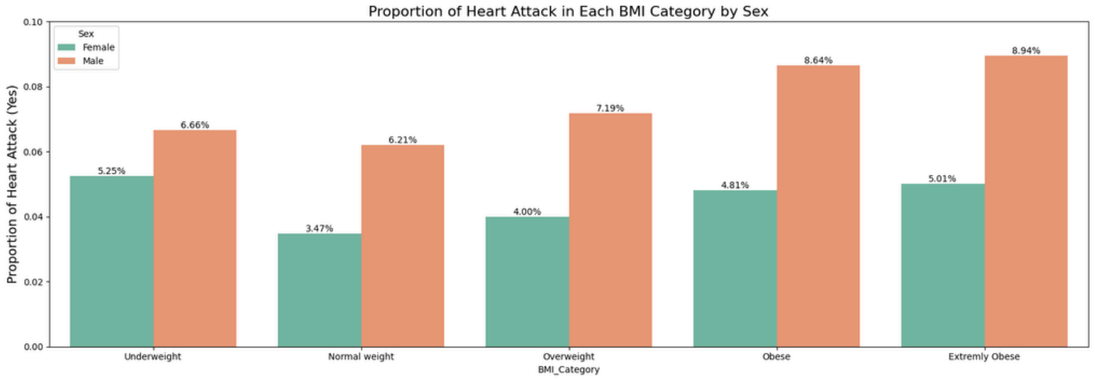


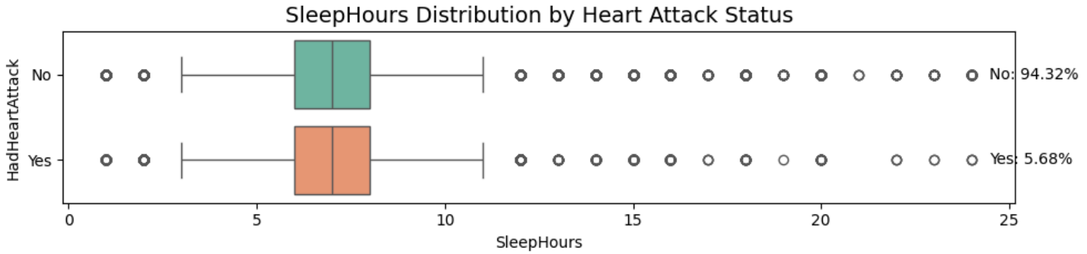
EDA

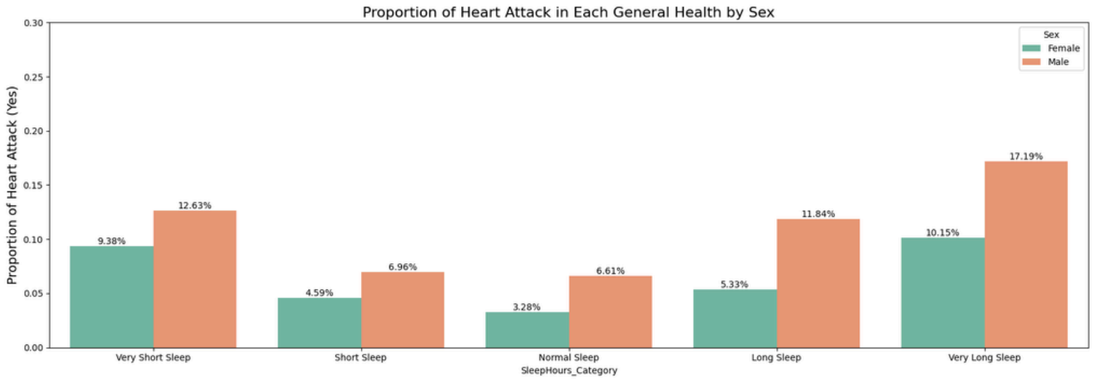
單一變數解釋性可能不足，但對變數進行區間分類 (binning) 並結合其他變數可以獲得更好的差異性。





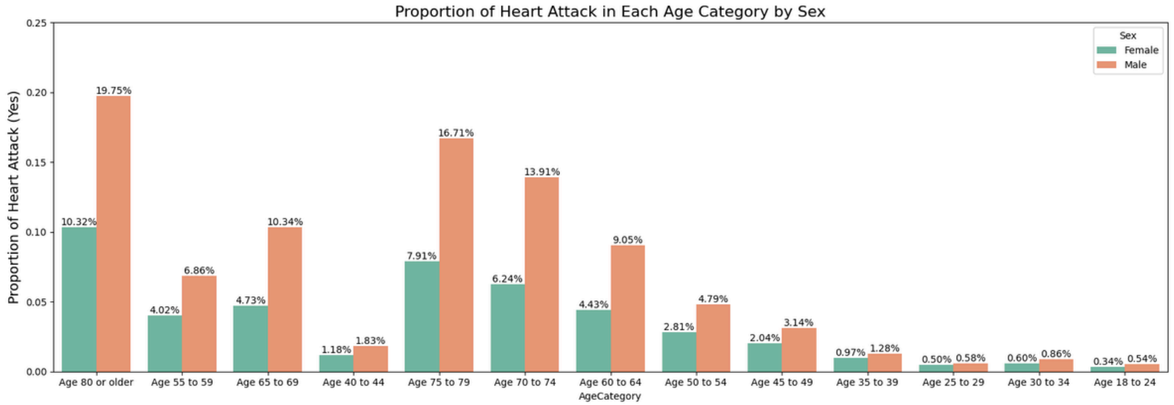


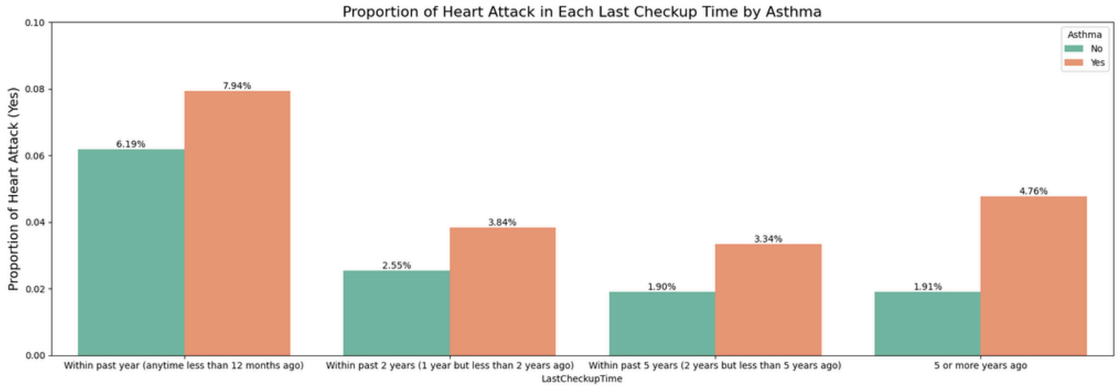


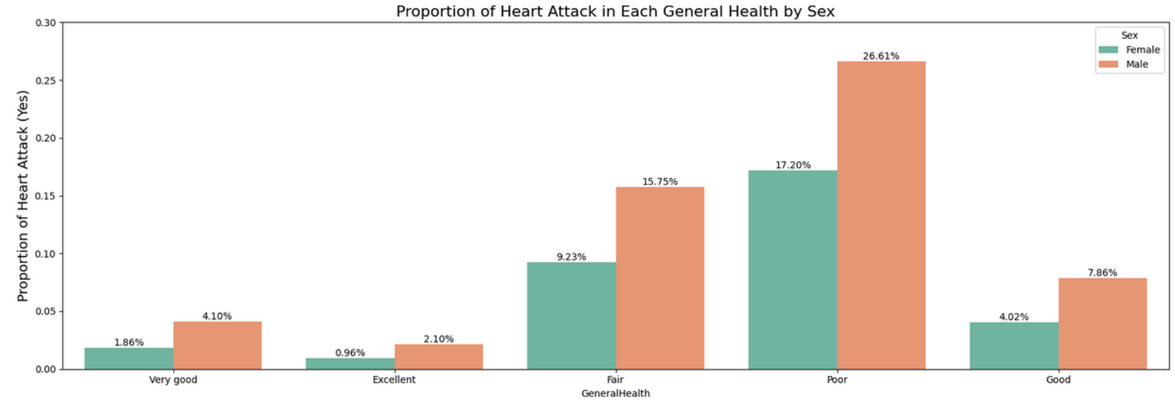


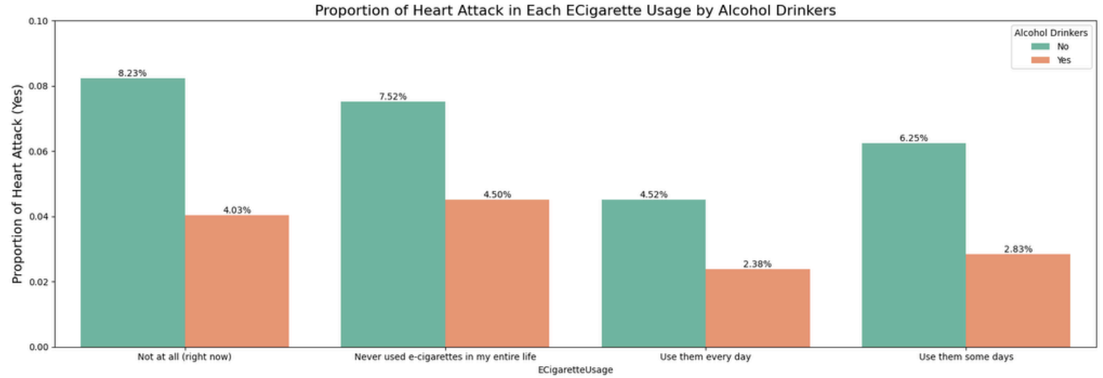
EDA

其他結合變數有更好的差異性例子。







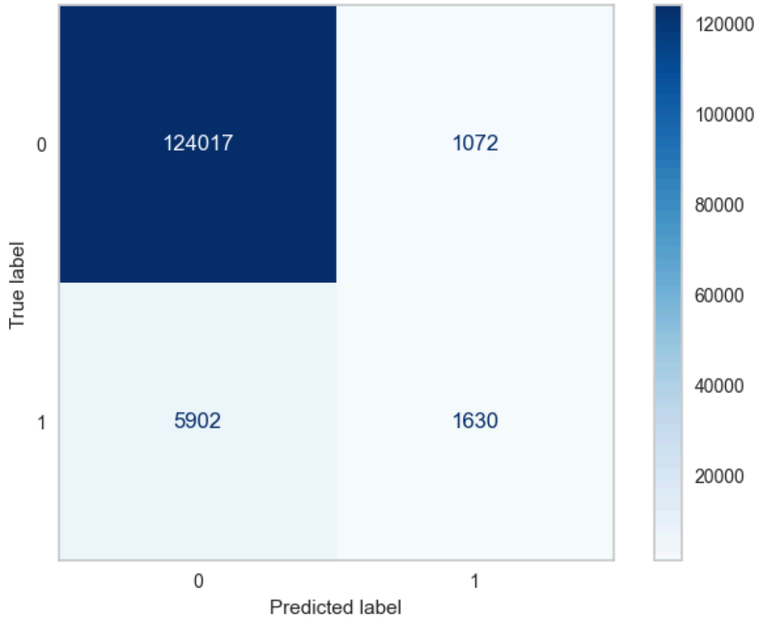


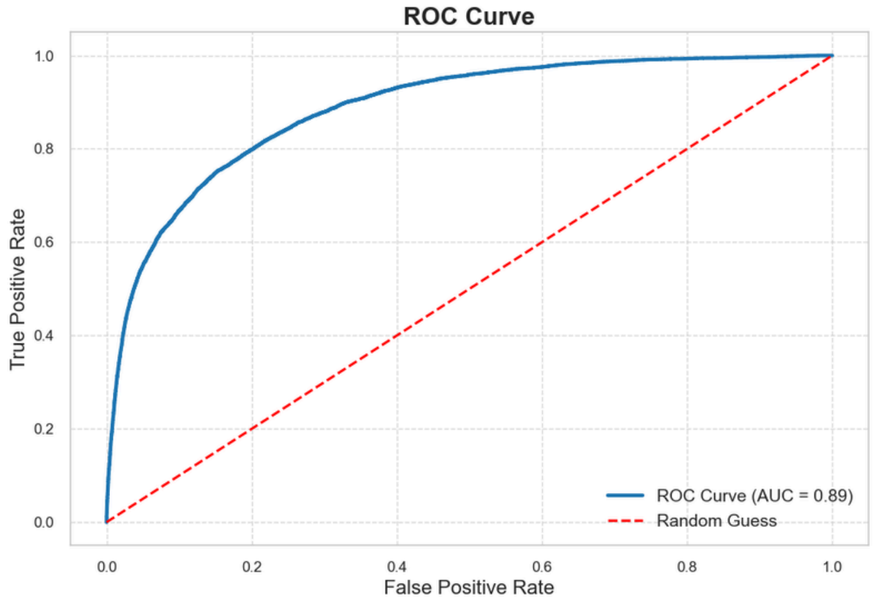
機器學習試跑結果

使用AutoML 對原始清理資料跑模型，
會自動跑數種模型並給出最好的模型與超參數。

Best model: XGBoostLimitDepthEstimator
Best model config: {'n_estimators': 1742, 'max_depth': 11, 'min_child_weight': 0.0596970726114001,...}
Best loss: 0.10975

	precision	recall	f1-score	support
0	0.95	0.99	0.97	125089
1	0.60	0.22	0.32	7532
macro avg	0.78	0.60	0.65	132621
weighted avg	0.93	0.95	0.94	132621





三、　預期解決方法

問題：資料不平衡

模型偏向多數類別，預測時容易只猜「0」，
導致整體 accuracy 看似高，
但實際效果不佳，須以不同的指標共同觀察

即使少數類別有部分被預測出來，
仍可能發生 Precision-Recall 不平衡的狀態

高 Recall, 低 Precision

模型猜了很多 1，但大部分猜錯，
導致假陽性（False Positive）過多

高 Precision, 低 Recall

模型僅在最有把握時才猜 1，
但漏掉大多數真正的 Positive 樣本

模型學不到少數類別的特徵 ，少數類別樣本太少，模型無法學習其代表性的模式

預期解決方法：Undersampling&不平衡設計的演算法

Undersampling

做法：刪除多數類別樣本數

優點

1.資料極容易達到平衡
2.模型可以更好的學習到少數類別

缺點

1.可能遺失有價值的多數類別資訊
2.難以界定刪除樣本的標準

不平衡設計的演算法

方法：Balanced Random Forest、XGBoost、LightGBM

優點

1.內建針對不平衡的處理邏輯。
2.通常表現佳，容易整合進 pipeline。

缺點

1.需理解演算法本身的參數調整與效果。

預期解決方法：Oversampling

做法：增加少數類別樣本數（如 SMOTE、Boarderline-SMOTE、ADASYN）

方法	SMOTE	Borderline-SMOTE	ADASYN
重點	提升少數類別的樣本數， 平均分佈合成資料	加強模型學習決策邊界， 減少分類模糊地帶錯誤	根據分類難度生成新樣本， 重點加強難分類的地方
風險	容易生成 落在多數類別區域的無效樣本	若邊界定義不準， 可能仍會有噪音 或overfitting 風險	如果資料本身分佈混亂， 可能過度擴增錯誤樣本， 造成 overfitting

優點

1.增加少數類別樣本，有助模型學習其特徵
2.保留全部的原始多數類別資料

缺點

1.易導致 overfitting (尤其是簡單重複樣本的情況)

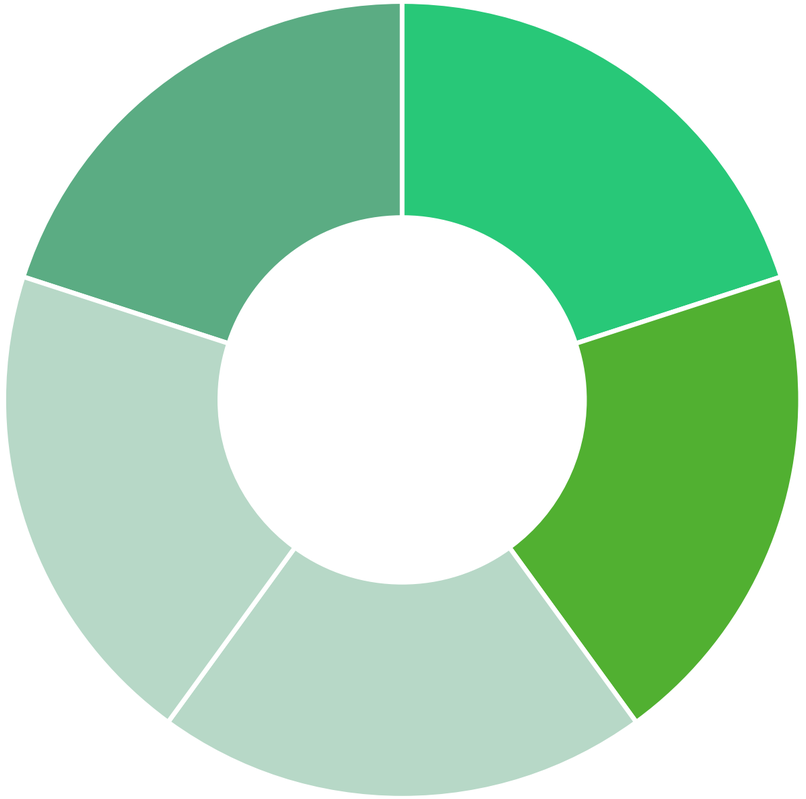
特徵重要性分析

PCA

所有變數中變異最大的前 K 個主成分
保留主要資訊
缺點為難以解釋個別變數的貢獻為何

Filter Methods

相關係數（Correlation Coefficient）
卡方檢定（Chi-square test）
ANOVA F-test



Embedded Methods

效率較高
LASSO (L1 Regularization)
Ridge (L2 Regularization)
Random Forest

Permutation importance

測量特徵的重要性
（打亂某特徵後看模型表現掉多少）

SHAP value 分析

更精細地解釋每個特徵對單一預測的影響