東吳大學巨量資料管理學院學士專題成果報告書

Bachelor Degree Program

School of Big Data Management

Soochow University

Report of practices

錢進陷阱？

用 AI 預測詐欺的蛛絲馬跡

11173103 資科四 A 杜鈺柔

11173120 資科四 A 許芷羚

11173122 資科四 A 武昀臻

11173124 資科四 A 鄧羣霖

指導教授：鄭宏文老師

中華民國 114 年 12 月

摘要

隨著數位金融交易量快速成長，信用卡詐欺偵測已成為金融機構風險管理的重要議題。然而，實際交易資料通常具有極端類別失衡、特徵高度抽象化以及詐欺手法不斷演變等特性，使得傳統機器學習方法雖可能呈現高準確率，卻常在召回率不足、偵測能力有限與決策不透明等面向受到限制，降低其在真實場景中的實用性。

本研究旨在建構一套兼具效能、穩定性與可解釋性的詐欺偵測系統。我們整合資料前處理、基於 SHAP 的特徵工程策略、SMOTE 與 classweight 等不平衡處理方法、Optuna 超參數搜尋以及多模型集成方法，以提升模型對多變詐欺模式的適應能力。同時，本研究引入 SHAP 與 LIME 等 XAI 工具，使模型決策具備透明度與可追溯性。

最終模型在不同資料集上皆展現良好的 F1-score 與穩定性能，並能提供明確的風險因素解釋，兼具偵測效果與實務部署可行性，展現其於高度不平衡信用卡詐欺偵測領域的應用潛力。

目錄

圖目錄

第一章、研究背景與動機

第一節、研究背景

2024 年台灣信用卡詐欺金額已突破 32 億元，並連續兩年維持在 30 億元以上。其中，網路盜刷佔所有詐欺損失的 97%以上，常見形式包括「零元註冊」、「飛航中竊取」與跨境網購詐欺，雖然即時通知門檻下調使單筆平均損失有所降低，但詐欺案件筆數卻持續增加，反映詐欺手法日益自動化與多樣化。

在此環境下，金融機構仰賴交易監控模型辨識異常行為，但模型若過度敏感，可能將正常交易誤判為詐欺，造成客訴並影響客戶信任；若模型偵測不足，又可能放過真正的詐欺交易，使銀行與持卡人承受實質金流損失。如何在降低誤報與避免漏報之間取得平衡，已成為提升信用卡詐欺偵測效能的迫切課題，也是本研究的重要背景。

第二節、研究動機

信用卡詐欺在近年快速上升，使金融機構面臨「如何在第一時間準確偵測出真正的詐欺交易」的重大挑戰。現有的偵測系統常因類別極度不平衡與詐欺手法快速演化，而無法有效捕捉少數異常模式，導致真正的詐欺交易未被識別，進而造成銀行與持卡人承受損失。

另一方面，若模型過度敏感，又可能將正常交易誤判為可疑，造成客戶不便、提高人工覆核成本，甚至影響客戶對金融機構的信任。如何在提升偵測能力的同時，降低誤殺正常客戶的比例，已成為風險管理中的核心痛點。

因此，本研究的主要動機在於發展一套能夠準確辨識真正詐欺行為、改善銀行在偵測詐欺上的痛點、並有效降低誤報情形的信用卡詐欺偵測模型，使其兼具實務可行性與高可靠度，並為金融業的智能風控提供可持續運作的技術基礎。

第三節、研究目的

1. 比較多種機器學習模型(Logistic、LightGBM、XGBoost、Ensemble)在高度不平衡之信用卡詐欺資料上的表現
2. 評估不同不平衡處理方式(SMOTE vs class_weight/scale_pos_weight)
3. 結合 SHAP 與 LIME，說明模型在全局與個案層級的可解釋性

第四節、問題情境與挑戰

信用卡詐欺偵測在實務上面臨多重挑戰，包含：

1. **資料高度不平衡(Class Imbalance)**

在真實世界的信用卡交易資料中，正常交易佔絕大多數，詐欺交易通常只佔極少數，這種極度不平衡的類別分布會導致一般標準分類器傾向預測全部為正常，在整體準確率看似很高的情況下，卻完全忽略少數的詐欺樣本，對風控而言是不可接受的。

2. **高維度與龐大筆數(High-dimensional & Large-scale)**

信用卡交易資料常包含數十個以上的特徵，同時交易筆數可能動輒數十萬甚至上百萬，在這樣的規模之下，模型必須兼顧計算效率與預測性能，並避免過度擬合。

3. **模型可解釋性(Model Interpretability)**

在金融領域，模型的預測結果不僅要準確，更需要可解釋，監管機關與內部稽核單位會關切為何這筆交易被判斷為詐欺、哪些特徵在決策中扮演關鍵角色，若模型完全黑箱，將不利於實務導入，因此如何兼顧高預測效能與適度的可解釋性，是我們研究必須面對的另一項挑戰。

第二章、相關研究探討

第一節、信用卡詐欺偵測與資料特性

信用卡詐欺偵測的核心挑戰,在於必須在高度不平衡且資訊部分被匿名化的交易資料中,從大量正常交易裡找出極少數可疑樣本。實務上的信用卡交易資料通常包含多種不同來源的特徵,例如交易時間、交易金額與幣別、商店類型與交易通路、持卡人的歷史行為統計等。

部分公開資料集為了保護個資會先對原始欄位做匿名化與降維處理,事先對原始變數進行**主成分分析**(PCA),這些經過 PCA 轉換的特徵已經是線性組合與標準化的結果,使得資料在高維空間中較具分離性,也利於後續模型訓練。

第二節、類別不平衡處理方法:SMOTE 與替代方案

為了應對類別不平衡問題,外界提出了多種處理方法,大致可分為以下幾類:

1. **欠採樣**(Under-sampling)

透過隨機刪除部分多數類別樣本,使各類別數量接近平衡。其優點是計算效率高、實作簡單,可快速縮小資料量;缺點是可能刪除含有重要資訊的多數類別樣本,導致決策邊界學習不完整,尤其在少數類別本來就極度稀少時,風險較大。

2. **過採樣**(Over-sampling)

透過複製少數類別樣本或合成新樣本來擴充少數類別。其優點是保留所有原始樣本,讓少數類別在訓練過程中有足夠權重;缺點是若在人造樣本密集填滿特徵空間,模型可能在某些區域產生過度自信,甚至學到不符合真實世界的人造規律。SMOTE 是在少數類別樣本之間插值產生人造樣本,避免重複複製同一筆資料。

3. **混合方式**(Hybrid Methods)

結合欠採樣與過採樣,先刪除部分多數類別樣本,再對少數類別進行 SMOTE,希望在資訊保留與平衡程度之間找到折衷。

本研究在相關研究回顧中，整理上述方法的優缺點，並在實作上選擇只在訓練集上使用 SMOTE 做過採樣(避免測試資料受到污染)，同時與不產生人造樣本的權重調整策略(class_weight、scale_pos_weight)進行比較。

第三節、權重調整策略：class_weight 與 scale_pos_weight

相較於直接改變資料分布的欠採樣、過採樣不同，另一類常見做法是在模型的損失函數中調整類別權重，讓模型在訓練時自動更重視少數類別的錯誤，包括：

1. class_weight

對於支援 class_weight 參數的模型，可以指定各類別在損失函數中的權重，例如設定少數類別權重大於多數類別，讓模型在訓練過程中更在意少數類別被誤判的情況。常見設定之一為 class_weight='balanced'，由模型根據各類別樣本數自動計算權重，這種作法不需生成任何合成樣本，因此避免了 SMOTE 在特徵空間捏出人造點的風險。

2. scale_pos_weight

在 XGBoost 等演算法中，常使用 scale_pos_weight 參數來加強正類的影響力。常見估計方式是將其設為多數類別樣本數/少數類別樣本數，讓模型在每一次樹的分裂與損失計算中，都更重視正類篇的錯誤。

在相關研究中，權重調整策略常被視為比過度依賴過採樣更穩健的替代方案，特別是在少數類別樣本本身已具有一定代表性時。

第四節、Boosting 模型在詐欺偵測上的應用：LightGBM/XGBoost

Boosting 是一種將多個弱分類器反覆疊加，逐步改善錯誤的集成學習方法，每一次迭代都針對前一輪模型做錯的樣本加強學習，最終形成一個強分類器，代表性演算法含：

1. XGBoost(Extreme Gradient Boosting)

以梯度提升決策樹為基礎，加入正則化項與多種剪枝策略，提升模型的泛化能力與訓練效率。XGBoost 支援 scale_pos_weight 等不平衡處理參數，並提供豐富的超參數可調整模型的深度、樹數、學習率、子採樣比例等。

2. LightGBM(Light Gradient Boosting Machine)

相較於 XGBoost，LightGBM 針對大規模與高維特徵的情境做了大量效率優化，包括使用 histogram-based 的分裂方式以及 leaf-wise 的樹成長策略，能在維持預測性能的前提下大幅降低訓練時間與記憶體使用量。LightGBM 亦支援 class_weight 等機制，以配合不平衡資料的需求。

相關研究顯示，在信用卡詐欺偵測與其他金融風險管理議題中，LightGBM 與 XGBoost 常被證實在多數評估指標上優於傳統的 Logistic Regression 或單一決策樹，這類梯度提升樹模型擅長處理高維且經過 PCA 轉換的特徵、特徵之間潛在的非線性關係與交互作用、混合連續型與離散型特徵。

第五節、模型可解釋性：SHAP 與 LIME 之相關研究

在金融領域，模型的預測結果往往會直接影響到用戶權益與風險管理決策，因此**可解釋性**成為模型實務導入的重要前提。近年來，**可解釋** AI 領域提出多種方法，其中 SHAP 與 LIME 是兩種常被應用於表格型資料的技術。

1. SHAP

基於合作博弈論中的 Shapley value 概念，將每一個特徵視為參與預測的玩家，透過考慮不同特徵組合下模型輸出的變化，計算各特徵對預測結果的邊際貢獻。在全局層級，SHAP 可以提供每個特徵的平均重要度排序，協助研究者了解模型整體主要倚賴哪些變數；在個別樣本層級，SHAP 可以說明某一筆交易中，哪些特徵值推高或壓低了該筆交易被判為詐欺的機率。

2. LIME

核心概念是在某一筆特定樣本附近，建立一個簡單且可解釋的近似模型，用來近似原本複雜模型在該區域的決策邏輯，透過對特徵進行擾動與重新預測，能

提供單筆預測結果的局部特徵重要性排序。特別適合用於個案分析，例如解釋某一筆被判為詐欺的交易，究竟是因為金額偏高、時間點異常，還是某些 PCA 特徵呈現極端值。

本研究結合 SHAP 與 LIME，先以 SHAP 分析模型的整體特徵重要性與影響方向，再用 LIME 針對個別交易案例進行解釋。透過這樣的可解釋性分析，本研究不僅關注模型在各種評估指標上的表現，也同時檢視模型決策背後的邏輯是否符合實務直覺，作為未來實際應用與調整門檻的重要參考。

第三章、研究方法

第一節、資料集介紹與欄位說明

1. 資料來源與概述

本研究採用公開的信用卡詐欺資料集 creditcard.csv 作為實驗資料，該資料集收錄了某歐洲發卡機構在特定期間內的信用卡交易紀錄，為信用卡詐欺偵測領域中最常被用作基準的公開資料之一。資料已經過 PCA 處理，因此 V1~V28 本身已是經過中心化與縮放後的線性組合，各主成分在數值尺度上已大致可比較，分布也接近常態，有利於後續模型在高維空間中學習分界。

相較之下，Time 與 Amount 則保留較接近原始的數值意義，因此本研究會在後續特徵工程中特別針對這兩個欄位設計額外的轉換與衍生變數。

2. 變數說明：Time、Amount、V1~V28、Class

資料集中主要欄位可分為以下幾類：

(1) Time：每筆交易自資料收集開始時點起算的秒數，數值範圍代表交易發生的時間先後，而非實際日期與時刻。

(2) Amount：該筆交易的金額，單位為通用數值(具體幣別未公開)，可作為衡量交易風險大小的重要指標之一。

(3) V1~V28：由原始特徵經 PCA 轉換後所產生的 28 個匿名化變數，這些特徵多已完成標準化與降維處理，包含了與持卡人行為、交易模式等相關的資訊，但其原始意義未被揭露。

(4) Class：目標變數，用於標記該筆交易是否為詐欺交易，通常以 0 表示正常、以 1 表示詐欺。

3. 資料量、詐欺比例與基本統計

本研究使用的資料集共包含 284,807 筆交易紀錄，原始欄位為 Time、Amount、V1~V28 以及 Class，共 31 個欄位。在不刪除任何原始欄位的前提下，本研究額外針對時間與金額設計 5 個基礎特徵，以及 8 個基於 SHAP 分析的策略性特徵，最終形成共 44 個欄位(含目標變數)的特徵集合。類別分布方面，標記為詐欺的交易(Class=1)僅有 492 筆，約佔全部交易的 0.17%，詐欺交易與正常交易的比例約為 1:578，屬於典型的極度不平衡資料。

第二節、特徵前處理與工程設計

在不刪除原始欄位的前提下，針對 Time 與 Amount 設計一系列額外特徵，並根據後續 SHAP 分析結果，進一步構造交互特徵與極端值標記，目的是在不破壞原有 PCA 特徵的情況下，提高模型捕捉異常模式的能力。

1. 時間相關特徵：Time 轉換為週期特徵

首先對 Time 欄位進行初步的**探索性資料分析**，如圖 1 及圖 2 所示，可以觀察到 Time 並非均勻分布，而是呈現明顯的多峰結構，顯示交易量在資料期間的某些時間區段較為集中。

由於 Time 代表的是自資料收集開始後經過的秒數，而非實際的日期與時刻，其與詐欺行為之間未必存在簡單的線性關係，若直接將其視為一般連續變數，模型不一定能有效捕捉「在資料期間不同階段，詐欺出現頻率是否不同」這類較複雜的模式，因此，本研究將 Time 做以下轉換：

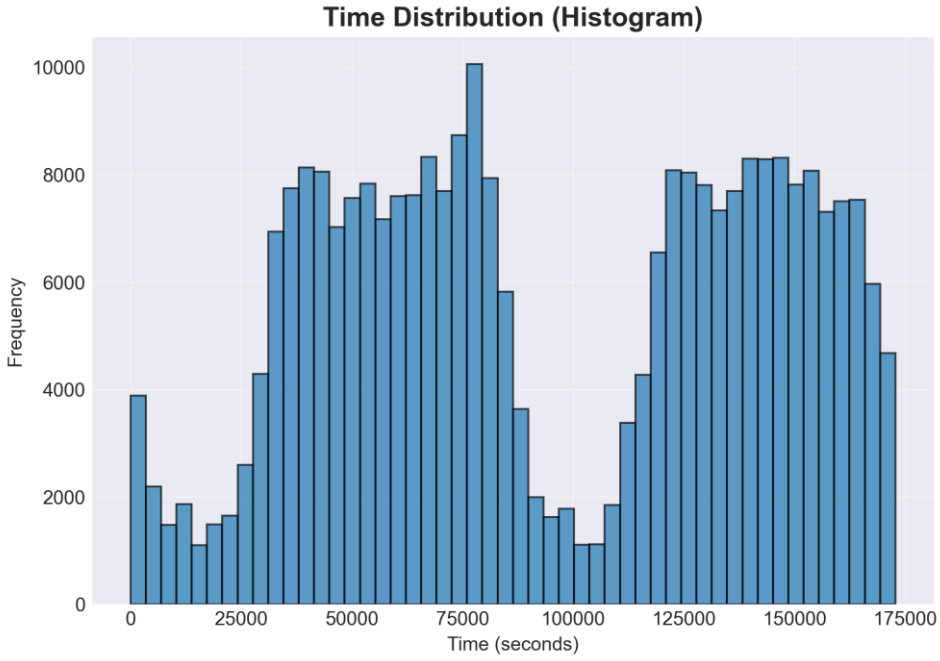

圖 1、信用卡交易時間分布圖(直方圖)

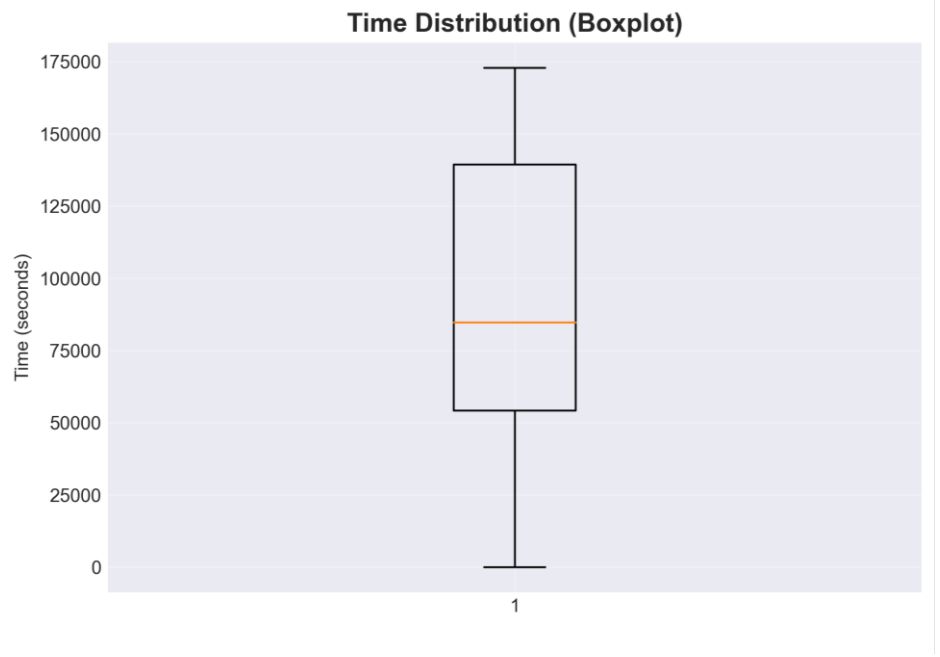

圖 2、信用卡交易時間分布圖(箱型圖)

- **時間相對位置(time_norm)：**

將 Time 縮放至 0~1 範圍，表示每筆交易在整體觀測期間中的相對位置，接近 0 代表發生在資料期間的前段，接近 1 則代表接近資料尾端。

$$\text{time\_norm} = \frac{\text{Time}}{\max(\text{Time})}$$

- **週期編碼(time_sin、time_cos)：**

進一步將 time_norm 映射到單位圓上，得到兩個週期特徵，這樣的轉換可以讓模型以較平滑的方式學習在觀測期間不同階段，詐欺發生機率是否有系統性差異，而不是把 Time 當成一般線性數值。

$$\text{time\_sin} = \sin(2\pi \cdot \text{time\_norm})$$

$$\text{time\_cos} = \cos(2\pi \cdot \text{time\_norm})$$

- **相鄰交易時間差(time_diff)：**

此外，本研究也計算相鄰兩筆交易的時間差，並將第一筆交易的 time_diff 設為 0，該特徵用來描述這筆交易距離前一筆交易隔了多久，若 time_diff 極小，代表短時間內連續交易；若 time_diff 很大，則代表很久才出現一次交易。詐欺行為有時會出現在突然一串密集交易或長時間沒有交易後突然一筆可疑交易的情境中，time_diff 能協助模型捕捉這類時間密度異常的模式。

$$\text{time\_diff}_t = \text{Time}_t - \text{Time}_{t-1}$$

2. 金額相關特徵：Amount 取對數與交互特徵

針對交易金額，本研究同樣先檢視其分布情形，如圖 3 及圖 4 所示，Amount 呈現典型的右偏長尾分布，多數交易為小額消費，少數交易金額極大，且存在明顯極端值。若直接將原始 Amount 輸入模型，可能導致模型過度受少數極大金額影響，而忽略在中小額區間中也可能存在的詐欺模式，因此，本研究對 Amount 進行自然對數轉換。

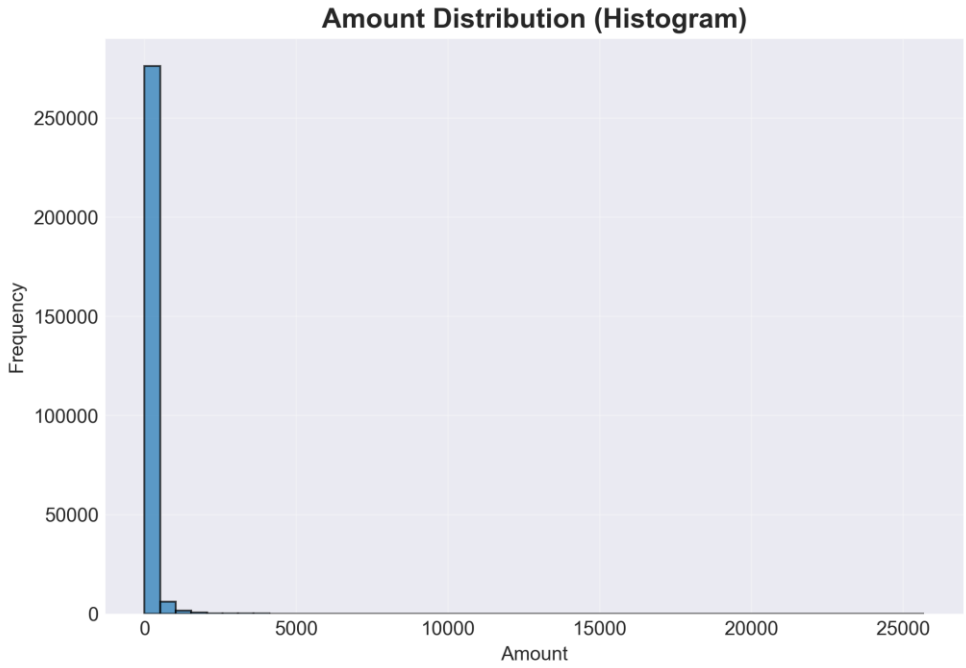

圖 3、信用卡交易金額分布圖(直方圖)

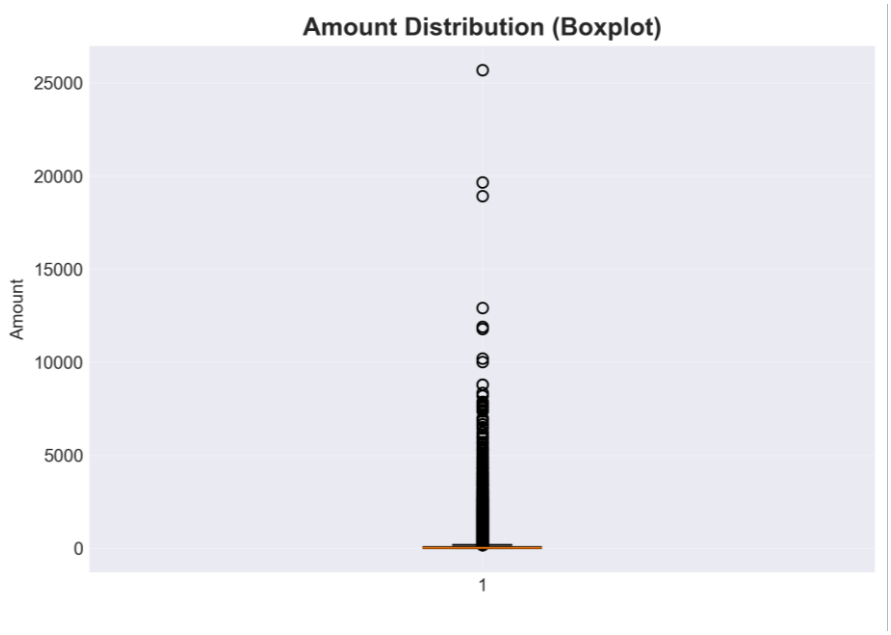

圖 4、信用卡交易金額分布圖(箱型圖)

- **自然對數轉換：**

此轉換在保留金額大小相對關係的前提下，壓縮了極端大額交易的數值範圍，並讓小額至中額區間的差異更明顯，有助於模型在不同金額區間上學習較為平衡的決策邏輯。

$$\text{log\_amount} = \log(1 + \text{Amount})$$

完成基本轉換後，先以樹模型訓練基礎模型，並透過 SHAP 分析各特徵的重要度，結果顯示，**V14、V4、V12、V10** 以及與金額相關的特徵在偵測詐欺交易上扮演關鍵角色。基於此觀察，本研究並非隨機新增大量特徵，而是針對這些關鍵特徵設計少量但有針對性的衍生變數，主要分為兩類：

- **交互特徵(Interaction Features)：**

這些特徵用來捕捉當隱含行為特徵出現異常變化，且金額同時偏高時，詐欺風險是否特別顯著等非線性關係。

$$\text{V14\_x\_logA} = \text{V14} * \text{log\_amount}$$

$$\text{V4\_x\_logA} = \text{V4} * \text{log\_amount}$$

$$\text{V12\_x\_logA} = \text{V12} * \text{log\_amount}$$

$$\text{V14\_x\_V4} = \text{V14} * \text{V4}$$

- **離散與極端值標記特徵(Binning & Flags)：**

(1) **amount_bin**：將 **log_amount** 依四分位數分為四個等級，約略對應很低/中等/偏高/極高金額區間，方便樹模型學習不同金額段落的風險差異。

(2) **V14_extreme、V4_extreme**：分別以 **V14、V4** 的 1%與 99%分位數作為門檻，標記該筆交易是否落在分布的極端區域，用來明確標示非常異常的隱含行為特徵。

(3) **high_amount**：標記 **log_amount** 是否高於其 95%分位數，用於捕捉極高金額交易。在信用卡詐欺情境下，單一特徵異常未必足以構成詐欺風險，但高金額交易搭配異常行為模式往往更具警示意義。上述交互特徵與極端值標記，旨在讓模型能更細緻地學習這類複合異常模式。

3. PCA 特徵 V1~V28 的性質與對模型的影響

對於模型訓練而言，PCA 特徵對本研究有兩個主要影響：

(1) 有利於模型穩定訓練

由於 **V1~V28** 已在相同尺度上，且彼此相關性較低，無論是 **Logistic Regression** 或樹狀模型，都較不會受到多重共線性或尺度差異的影響。因此，本

研究在特徵前處理時，並未再對這組 PCA 特徵額外進行標準化，而是將資源集中在對 Time 與 Amount 做適當轉換與衍生。

(2) **降低特徵可解釋性**

V1~V28 為多個原始欄位的線性組合，無法直接對應具體的金融行為或交易類型，單看數值難以說明「為何這筆交易被判定為可疑」。為了部分彌補其可解釋性不足的問題，本研究在後續章節中結合 SHAP 與 LIME，分析這組 PCA 特徵在模型中的重要性與作用方向，從全局與個案兩個層次解讀模型的決策邏輯。

第三節、基礎模型設計

1. Logistic Regression 基準模型(Baseline)

Logistic Regression 是一種廣義線性模型，透過線性組合特徵後套用 sigmoid 函數，將輸出映射為屬於詐欺類別的機率。其模型具備可解釋性高、訓練速度快的優點，適合作為比較用的基準模型。然而，由於其決策邊界為線性形式，面對信用卡詐欺資料中高度非線性的行為模式時，模型較難捕捉複雜特徵間的交互關係，因此通常作為 Baseline 參考，而非最終最佳模型。

為了提供清楚的比較基準，本研究首先建立 Logistic Regression 作為 Baseline，相較於複雜的樹模型與集成模型，其結構較為簡單、參數具一定可解釋性，可用來觀察各特徵與詐欺風險之間的線性關係方向。假設輸出機率為輸入特徵的線性組合經由 sigmoid 函數轉換

$$P(y = 1 \mid \mathrm{x}) = \sigma(\mathrm{w}^{\top}\mathrm{x} + b)$$

在本研究中，Logistic Regression 的角色主要有二：

(1) **作為效能比較的基準點：**

先在經 SMOTE 過採樣後的訓練資料上訓練 Logistic Regression，評估線性模型在類別平衡情境下的預測能力，提供與 LightGBM、XGBoost 等進階模型比較的參考基準。

(2) **作為 Stacking Ensemble 的 meta-learner：**

在後續的 Stacking Ensemble 中，本研究亦採用 Logistic Regression 作為第二層的整合模型，利用其穩定且可解釋的特性，整合 LightGBM 與 XGBoost 的預測機率，形成最終的集成預測。

本研究並未對 Logistic Regression 啟用 class_weight 等權重調整機制，而是搭配 SMOTE 產生的平衡訓練資料進行訓練，以避免與其他不平衡處理策略混淆。

2. LightGBM/XGBoost 模型架構與輸入特徵

LightGBM 是基於梯度提升的樹狀模型，以葉節點優先的成長策略與直方圖演算法提升訓練效率，特別適合大規模與高維度資料。此外，LightGBM 能自然處理特徵間的非線性與交互作用，其可利用 class_weight 強化少數類別的影響，使其在不平衡資料上有良好表現，因此常被應用於金融詐欺偵測領域。

XGBoost 採用梯度提升樹，透過更完善的正則化項與先進的樹剪枝策略有效避免過擬合，並以高效率計算與分散式運算架構廣泛應用於實務競賽。對於不平衡資料問題，XGBoost 能透過 scale_pos_weight 調整少數類別（詐欺交易）的權重，使模型在 Recall 與 F1-score 上取得更佳效果，是詐欺偵測任務中表現穩定的主流模型之一。

在進階模型部分，本研究選用兩種主流 Boosting 演算法：LightGBM 與 XGBoost。兩者皆屬於梯度提升樹家族，透過反覆疊加弱學習器，逐步修正前一輪模型的錯誤，最終形成一個具高表現力的強分類器。

第四節、類別不平衡處理策略

本研究的資料集中，詐欺交易(Class=1)僅約佔全部交易的 0.17%，屬於極度不平衡資料，若直接在原始資料上訓練模型，模型容易傾向預測所有交易為正常交易，導致對少數類別的 Recall 極低。為了比較不同不平衡處理方法在信用卡詐欺情境中的優缺點，本研究採取雙軌實驗設計。

1. SMOTE 過採樣版本(Logistic/LightGBM/XGBoost with SMOTE)

在第一組實驗中，本研究採用 SMOTE 作為代表性的過採樣方法，其核心概念是利用少數類別樣本之間的鄰近關係，在特徵空間中插值產生新樣本，避免單純複製少數類別導致的過度擬合，具體作法如下：

(1) 僅對訓練集進行 SMOTE：

只在訓練資料上對詐欺類別(Class=1)進行過採樣，使少數類別樣本數量與正常類別相當，驗證集與測試集則維持原始不平衡分布，以更貼近實務應用情境，並避免在評估階段引入人造樣本。

(2) 在過採樣後的訓練集上訓練模型：

Logistic Regression(SMOTE 版本)、LightGBM(SMOTE 版本，不再額外設定 class_weight)、XGBoost(SMOTE 版本，不再額外設定 scale_pos_weight)，透過這樣的設計，可以觀察在資料被平衡後，不同模型對少數類別的識別能力是否提升，以及是否出現對人造樣本過度擬合的風險。

2. 權重調整版本：class_weight 與 scale_pos_weight

在第二組實驗中，不修改訓練資料的分布，而是直接在模型損失函數中調整不同類別的權重，讓模型在訓練過程中更重視少數類別的錯誤，具體設計如下：

(1) LightGBM with class_weight = 'balanced'

使用原始不經 SMOTE 的訓練資料，並將 class_weight 設為'balanced'，讓模型根據實際類別比例自動計算權重，使詐欺類別在損失函數中的影響力大於正常類別，藉此提升對少數類別的敏感度。

(2) XGBoost with scale_pos_weight

同樣使用原始不經 SMOTE 的訓練資料，將 scale_pos_weight 設為負類樣本數/正類樣本數，以反映類別比例的極度不平衡，進一步加強正類(詐欺類別)在梯度更新過程中的作用。

透過將 SMOTE 與權重調整兩種策略套用到相同模型，我們得以系統性比較這兩類方法在詐欺偵測情境中的實際優缺點。

第五節、超參數最佳化與 Ensemble 設計

1. Optuna 超參數搜尋流程設計

為了公平比較不同模型版本，也避免僅憑經驗手動調參，本研究採用 Optuna 進行系統性的超參數搜尋。整體設計如下：

(1) 目標指標

由於研究重點在於在詐欺類別上同時兼顧 Precision 與 Recall，本研究選擇 F1-Score(以 Class=1 為主)作為超參數搜尋的主要目標指標，讓 Optuna 在搜尋過程中偏好能在少數類別上取得較好平衡的參數組合。

(2) 交叉驗證設定

為降低單一資料切分造成的偶然性，本研究在每一次 trial 中皆採用 Stratified K-Fold 交叉驗證，維持各折中的類別比例與原始分布相近。實務上，本研究設定 K=3，並在每一折計算詐欺類別的 F1-Score，再取平均作為該組超參數的評估指標。

(3) 搜尋空間概念

對 LightGBM/XGBoost 設定的超參數含樹數量、單棵樹的最大深度、學習率、觀測值子抽樣比例、特徵子樣本比例、正則化參數等，Optuna 透過貝氏優化與試驗結果的回饋，在本研究設定的 20 次 trials 中逐步收斂至較佳的超參數組合。

透過上述設計，每一個模型版本皆對應一組由 Optuna 搜尋而得的最佳參數，提升模型間比較的公平性與可信度。

2. 單一模型與 Ensemble 策略規劃

在完成各模型版本的超參數搜尋後，本研究分別評估單一模型與多種 Ensemble 策略，整體規劃如下：

(1) 單一模型

- Logistic Regression(SMOTE 版本為主作為 baseline)
- LightGBM(SMOTE 版本與 class_weight 版本)

- XGBoost(SMOTE 版本與 scale_pos_weight 版本)

(2) Ensemble 策略

- Simple Average Ensemble

選取驗證集表現最佳的 LightGBM 與 XGBoost 模型，對同一筆資料分別輸出詐欺機率，將兩者取簡單平均作為最終預測機率。

- Weighted Average Ensemble

以各模型在驗證集上的 F1-Score 作為權重，對 LightGBM 與 XGBoost 的預測機率進行加權平均，讓表現較好的模型在最終預測中占較高比重。

- Stacking Ensemble

將 LightGBM 與 XGBoost 的預測機率視為第二層的輸入特徵，再以 Logistic Regression 作為 meta-learner，訓練一個整合模型，學習如何在不同情況下加權兩個基礎模型的輸出。

3. 最終模型選擇原則

在眾多單一模型與 Ensemble 組合中，本研究採用以下原則選擇最終模型，用於後續在測試集上的最終評估與模型解釋：

(1) 以驗證集 F1-Score 為主要依據

優先選擇在詐欺類別(Class=1)上 F1-Score 較高的模型，以兼顧 Precision 與 Recall 的平衡。

(2) 輔以 Precision、Recall 與 ROC-AUC

若 F1-Score 相近，則偏好在 Recall 表現較佳者，並觀察 ROC-AUC 是否合理，在實務場景中漏抓詐欺往往比誤判正常更具風險，因此在相同 F1-Score 的情況下，本研究傾向選擇 Recall 較高的模型。

(3) 考慮模型可解釋性與部署難易度

若性能差距不大，而某一模型在 SHAP/LIME 解釋上更清楚，或計算成本較低，也會納入實務選擇考量。

在本研究的實驗結果中，Stacking Ensemble 在 F1-Score 上略優於最佳單一模型(LightGBM with class_weight)，且可望結合兩者優點，因此最終選擇 Stacking 作為最佳模型，並在後續章節中進一步就其在測試集上的表現與可解釋性進行分析。

第四章、實驗設計與結果分析

第一節、資料切分策略(Data Splitting Strategy)

本研究為了在維持資料代表性的同時，避免評估結果過於樂觀，採用三階段資料切分策略，分別建構**訓練集**、**驗證集**與**測試集**，並搭配分層取樣及固定隨機種子，以確保結果的穩定性與可重現性。

1. 資料切分比例與方式

(1) 先切出測試集(Test set, 20%)：

先從原始資料中保留約 20%作為最終測試集，在切分過程中採用**分層取樣**，以 Class 作為分層依據，確保測試集中詐欺交易的比例與原始資料相近。測試集在整個建模流程中被保留到最後，不參與任何訓練、調參或模型選擇。

(2) 再將剩餘 80%切分為訓練集與驗證集：

剩餘的 80%資料再切分，約 75%作為訓練集(相當於原始資料的 60%)、約 25%作為驗證集(相當於原始資料的 20%)，同樣使用分層取樣，確保訓練集與驗證集中的詐欺比例與整體資料分布一致。

整體資料比例為訓練集約 60%、驗證集約 20%、測試集約 20%。此種三分法有助於在模型開發的不同階段，分別扮演「學習」、「調整」與「最終驗證」的角色，避免混用資料造成評估偏差。

2. 僅於訓練資料上進行 SMOTE 的理由

由於詐欺交易數量極少，本研究在處理類別不平衡時，採用 SMOTE 方法，於訓練階段合成額外的少數類別樣本，刻意僅針對訓練集進行 SMOTE，而不對驗證集與測試集做任何重採樣處理，原因如下：

(1) 避免資料洩漏(Data Leakage)：

若在資料切分之前，直接對整份資料進行 SMOTE，則合成出來的少數類別樣本，會同時依賴未來預留給驗證集與測試集的資訊，這會造成模型在驗證與測試時間接看過部分樣本的特徵結構，導致評估指標過於樂觀，無法真實反映模型在未來新資料上的表現。

相較之下，僅在訓練集上執行 SMOTE，可以確保驗證集與測試集完全由真實觀測資料構成，不受合成樣本干擾，有效降低資料洩漏風險。

(2) 讓驗證與測試集維持真實世界分布：

實務中，金融機構在接收到線上交易時，資料本身不會事先被重抽樣，詐欺交易仍然是極少數事件，因此，用來評估模型的驗證集與測試集，理論上應該維持原始的不平衡比例，這樣所計算出的 Precision、Recall、F1-Score 與 ROC-AUC 才具有實務意義。若在驗證或測試資料上使用 SMOTE，將人為生成大量詐欺樣本，會使得評估環境偏離實務場景，可能導致模型在不自然的類別比例下表現良好，但實際部署時效果不如預期。

(3) 將 SMOTE 的角色限制在協助模型學習：

SMOTE 的目的，是在訓練過程中提供模型更多的少數類別樣本，使其能學到較完整的詐欺行為模式，而非改變評估資料的分布，因此，本研究將 SMOTE 嚴格限制在訓練集上使用，讓模型在學習階段獲得足夠的詐欺樣本，而在評估階段仍回到原始不平衡的驗證與測試集上進行性能衡量。

綜合上述考量，訓練集可以被人為強化，驗證與測試集必須保持乾淨，這是處理不平衡資料時，兼顧模型效能與評估嚴謹性的關鍵原則。

3. 符合嚴謹機器學習流程之設計

在整體流程設計上，本研究刻意區分用來學習的資料與用來做決策與驗證的資料，以符合較嚴謹的機器學習實務規範：

(1) **訓練集**(Training set, 60%)

- 作為模型參數學習的主要來源。
- 在處理不平衡問題時，僅於此階段引入 SMOTE，以增加詐欺樣本數量。
- Logistic Regression、LightGBM、XGBoost 以及後續的 Ensemble 模型，皆基於訓練集進行學習。

(2) **驗證集**(Validation set, 20%)

- 不進行 SMOTE，保留原始不平衡比例，用來模擬真實場景。
- 用於模型選擇與超參數調整，比較使用 SMOTE 的版本 vs 使用 class_weight/scale_pos_weight 的版本、單一模型 vs Ensemble 模型。
- 亦用於後續**閾值**調整，在固定模型權重後，掃描不同閾值以平衡 Precision 與 Recall，選出最能提升 F1-Score 的切割點。
- 所有模型架構、超參數、是否採用 Ensemble、最終使用的判斷閾值皆僅根據訓練與驗證資料決定。

(3) **測試集**(Test set, 20%)

- 在模型與閾值皆確定後，才將測試集用於最終評估。
- 在此階段，不再進行任何調參或調整 threshold，而是直接套用前一階段決定的最佳模型與固定閾值。
- 測試集的評估結果(包括混淆矩陣、Precision、Recall、F1-Score、ROC-AUC 等)被視為本研究對模型泛化能力的客觀估計。

本研究的流程為先用訓練集學會模型，再用驗證集做選擇與調整，最後才用測試集一次性檢驗模型的真實實力，此種設計不僅降低資料洩漏與過度適配的風險，也使得本研究之評估結果更具可信度與實務參考價值。

第二節、評估指標設計與實作

在高度不平衡的信用卡詐欺偵測問題中，單純以整體準確率作為評估基準，往往會產生嚴重誤導，因此，本研究採用 Precision、Recall、F1-Score 與 ROC-AUC 四項指標，從不同面向評估模型在少數類別(Class=1，詐欺交易)上的辨識能力，並作為模型選擇、超參數調整與閾值設計的主要依據。

1. 混淆矩陣與基本概念

為了說明各指標的意義，首先以二元分類的混淆矩陣為基礎，以本研究為例，將 Class=1 定義為「詐欺交易」、Class=0 為「正常交易」，則混淆矩陣可分為四種情況：

(1) **真正類(True Positive, TP)**：實際為詐欺，模型也預測為詐欺。

(2) **假負類(False Negative, FN)**：實際為詐欺，但模型預測為正常。

(3) **假正類(False Positive, FP)**：實際為正常，但模型預測為詐欺。

(4) **真負類(True Negative, TN)**：實際為正常，模型也預測為正常。

後續所有評估指標都可以從 TP、FN、FP、TN 這四個數量推導而來。

2. Precision 與 Recall 偵測品質的兩個面向

(1) **Precision**：在所有被模型判定為詐欺的交易當中，實際上真的為詐欺的比例。

$$\text{Precision} = \frac{TP}{TP + FP}$$

Precision 越高，代表誤報 FP 越少，風控人員在接收警報時的信任度越高，亦可降低人工覆核的成本與客戶被錯誤風控的機率。

(2) **Recall**：在所有實際為詐欺的交易當中，有多少比例被模型正確偵測出來

$$\text{Recall} = \frac{TP}{TP + FN}$$

Recall 越低，代表 FN 越多，也就是大量詐欺交易被當成正常放過，在金融風控實務上，這類錯誤通常帶來實質金錢損失與風險累積，因此 FN 的成本往往遠高於 FP。

3. F1-Score：平衡 Precision 與 Recall 的綜合指標

在實務上，Precision 與 Recall 通常存在取捨關係，為避免只偏重其中一個指標，本研究採用 F1-Score 作為主要模型比較依據之一，F1-Score 定義為 Precision 與 Recall 的調和平均數：

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

本研究在**模型選擇、SMOTE/class_weight** 策略比較以及 Ensemble 設計時，均以 F1-Score 作為核心比較指標，並搭配 Precision 與 Recall 進一步分析模型行為。

4. ROC-AUC：整體排序能力與區分度評估

除了以上三項在固定閾值下計算的指標之外，本研究亦使用 ROC-AUC 評估模型在不同閾值設定下的整體表現。

(1) ROC 曲線(Receiver Operating Characteristic Curve)：

- **X 軸：**假陽性率(False Positive Rate, FPR = FP/(FP+TN))
- **Y 軸：**真陽性率(True Positive Rate, TPR = Recall = TP/(TP+FN))

為座標，描繪當分類閾值從 0 到 1 變動時，模型在抓到詐欺與誤報之間的折衷關係，ROC-AUC 則為 ROC 曲線下的面積，其數值介於 0.5(隨機猜測)至 1.0(完美分類)之間。

(2) 在高度不平衡的資料集中，ROC-AUC 有兩項重要意義：

- 它不依賴單一閾值，而是度量模型對正負類別的整體排序能力
- 即使在不同場景下需要不同的操作閾值，ROC-AUC 仍能提供一個相對穩定的比較基準

本研究因此將 ROC-AUC 作為輔助指標，用於檢查模型是否具備良好的區分能力，並與 F1-Score 一同作為各模型版本比較的依據。

5. 為何在詐欺偵測情境中特別重視 F1-Score 與 Recall

綜合以上討論，在信用卡詐欺偵測的實務場景中，FN(漏偵詐欺)與 FP(誤報詐欺)的成本並不對稱，FN 直接造成金錢損失、客戶權益受損，甚至衍生法律與信任風險，FP 雖然會增加人工覆核成本，或短暫造成客戶不便，但通常可透過後續人工審查與客戶確認進行補救。

在此情境下，完全追求少誤報而忽略漏報是不可接受的，因此，本研究在模型設計與比較時，刻意將焦點放在 Recall 與 F1-Score。

第三節、模型效能比較與結果分析

本節綜整各模型在驗證集與測試集上的表現，依序從單一模型、Ensemble 方法、閾值調整到最終測試結果，分析不同設計選項對詐欺偵測效能的影響。

1. 單一模型與權重版本比較

首先比較五個單一模型版本在驗證集上的效能：

- Logistic Regression(SMOTE)
- LightGBM(SMOTE)
- XGBoost(SMOTE)
- LightGBM(class_weight)
- XGBoost(scale_pos_weight)

對每個模型，皆在驗證集上計算 Precision、Recall、F1-Score 與 ROC-AUC，並彙整成比較表，同時以長條圖視覺化，如**圖 5**：

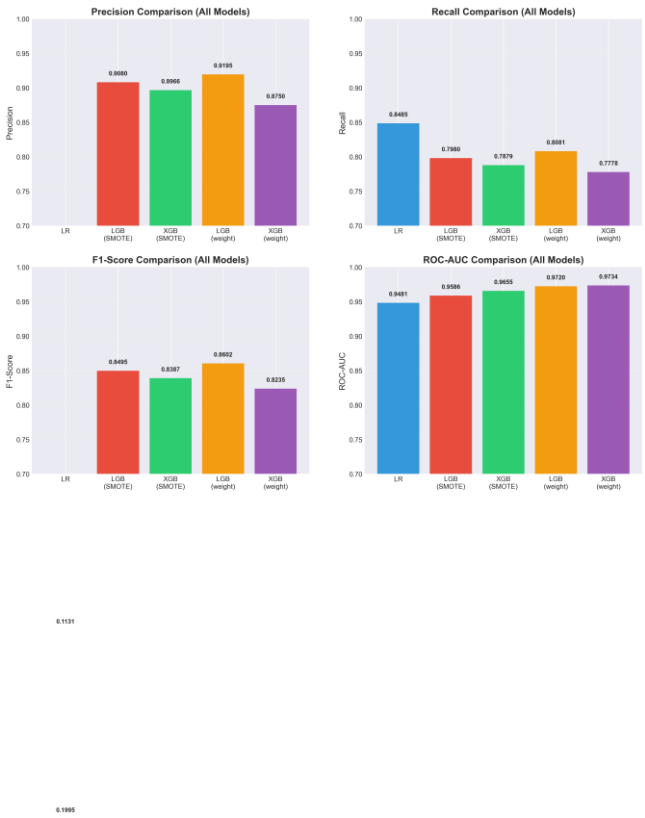

圖 5、各單一模型於驗證集之效能指標比較圖

從結果可以觀察到幾個重點：

(1) **梯度提升樹模型整體優於** Logistic Regression：即使在使用 SMOTE 之後 Logistic Regression 的 Recall 表現尚可，但由於僅能刻畫線性關係，較難捕捉特徵之間的非線性與交互作用，其 F1-Score 明顯低於兩種梯度提升樹模型(LightGBM、XGBoost)，這顯示在信用卡詐欺偵測這類結構複雜的問題上，樹模型更適合作為主要架構。

(2) **LightGBM 的** class_weight **版本略優於** SMOTE **版本**：在 LightGBM 中，SMOTE 版本於驗證集的 F1-Score 約為 0.8495，採用 class_weight='balanced' 的版本則可提升至約 0.8602，Precision 也略為提高。這說明在 LightGBM 上，直接在原始不平衡資料上調整類別權重，就能取得不遜於 SMOTE 的效果，甚至更穩定

(3) **XGBoost 的 SMOTE 版本略優於** scale_pos_weight **版本**：對 XGBoost 而言 SMOTE 版本的 F1-Score 約為 0.8387，稍高於使用 scale_pos_weight 版本的

0.8235。也就是說，對 XGBoost 這個模型來說，適度增加少數類別的人造樣本，反而有助於學習較清晰的決策邊界。不過，scale_pos_weight 版本在 ROC-AUC 上略高，代表其整體排序能力仍然不錯。

(4) **最佳單一模型為 LightGBM(class_weight)**：綜合 F1-Score、Precision 與 ROC-AUC **考量**，LightGBM(class_weight) 在五個單一模型中整體表現最佳，其 F1-Score 約 0.8602、ROC-AUC 約 0.972，同時維持不錯的 Precision。因此，在進一步討論 Ensemble 方法之前，本研究將 LightGBM(class_weight) 視為最佳單一模型，作為後續集成設計與比較的重要基準。

2. **Ensemble 方法比較**

在確認 LightGBM(class_weight)為最佳單一模型之後，本研究進一步評估集成學習(Ensemble)是否能在此詐欺偵測任務中帶來額外提升。我們以 LightGBM(class_weight)、XGBoost(SMOTE)作為兩個基礎模型，分別使用其在驗證集上表現較佳的設定，並實作三種 Ensemble，策略各方法於驗證集上的 F1-Score 與 ROC-AUC，如圖 6 所示：

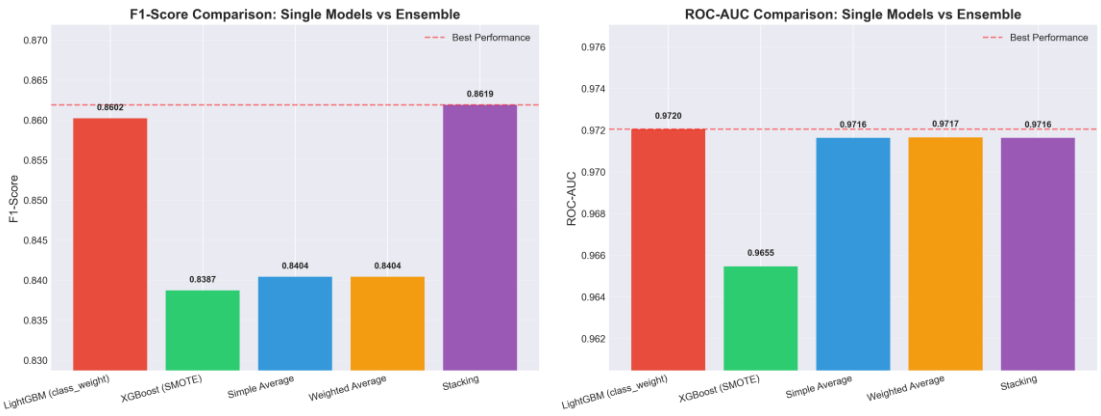

圖 6、單一模型與各 Ensemble 方法之 F1-Score 與 ROC-AUC 比較圖

從圖中可以觀察到：

(1) **Simple/Weighted Average 與最佳單一模型相近，但未明顯超越**：

Simple Average 與 Weighted Average 的 F1-Score 約為 0.8404，略高於 XGBoost(SMOTE)的 0.8387，但仍低於 LightGBM(class_weight)的 0.8602；在 ROC-AUC 上，兩者約為 0.9716~0.9717，與 LightGBM(class_weight)的 0.9720 相

當，顯示簡單平均與加權平均可以維持不錯的整體區分能力，但在 F1-Score 上沒有帶來顯著增益。

(2) Stacking 在 F1-Score 上取得整體最佳表現：

以 Logistic Regression 作為第二層 meta-learner 的 Stacking 方法，其 F1-Score 約為 0.8619，略高於最佳單一模型 LightGBM(class_weight)的 0.8602，為所有模型組合之中最高；雖然其 ROC-AUC 約為 0.9716，略低於 LightGBM(class_weight)的 0.9720，但差距非常小，在統計上可視為相近。

整體結論而言，Stacking 作為最終模型，LightGBM(class_weight)為關鍵基底，綜合 F1-Score 與 ROC-AUC 的結果，Stacking Ensemble 在驗證集上呈現出略優於所有單一模型的整體效能，因此本研究後續以 Stacking 作為「最終預測模型」，同時，由於 Stacking 的第一層仍以 LightGBM(class_weight)為核心基底模型，本研究亦以其作為代表模型進行可解釋性探討，以兼顧實務部署與模型解釋的需求。

3. 閾值調整與最終模型選擇

如前節所示，本研究在驗證集中表現最佳的模型為 Stacking Ensemble，因此在進行分類閾值調整時，同樣以 Stacking 模型在驗證集上的預測機率為基礎，系統性掃描多個不同的閾值(約 0.10~0.85)，對每一個閾值計算對應的 Precision、Recall 與 F1-Score，結果如圖 7 所示：

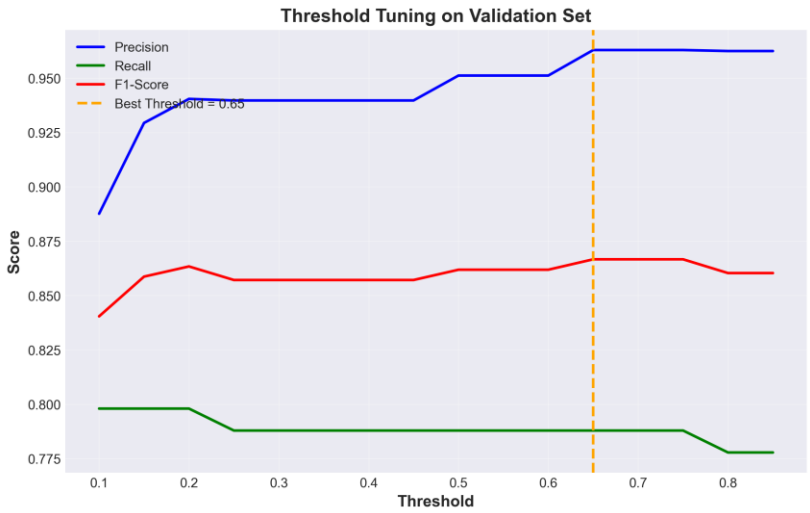

圖 7、驗證集中不同分類閾值下 Precision、Recall 與 F1-Score 之變化圖

從圖中可知：

(1) 閾值愈低，Recall 愈高、Precision 愈低

在閾值約 0.10 時，模型幾乎把大多數有可能的交易都抓出來，Recall 接近 0.80，但同時產生較多誤報，Precision 約落在 0.89 左右，此時雖然漏抓風險較低，但人工覆核成本會明顯提高。

(2) 閾值拉高時，Precision 持續上升、Recall 緩步下降

隨著閾值從 0.3、0.4 逐步提高，Precision 明顯提升，而 Recall 只略微下降，整體 F1-Score 在中間區段逐漸上升，代表在這個範圍內可以同時兼顧少誤報與不過度漏報。

(3) F1-Score 在閾值 0.65 左右達到最高

當閾值調整為 0.65 時，驗證集上的 Precision 約為 0.9630、Recall 約為 0.7879、F1-Score 約為 0.8667，為所有測試閾值中最高點。

綜合以上結果，本研究最終採用 Stacking Ensemble + **閾值 0.65**，作為在詐欺偵測情境下的最佳設定。

4. 測試集最終結果與混淆矩陣分析

在確定最終模型 Stacking Ensemble + **閾值 0.65** 後，我們將此固定模型與閾值套用到事先保留、完全未參與訓練與調參的測試集上，並計算相關指標與混淆矩陣，如**圖 8** 所示：

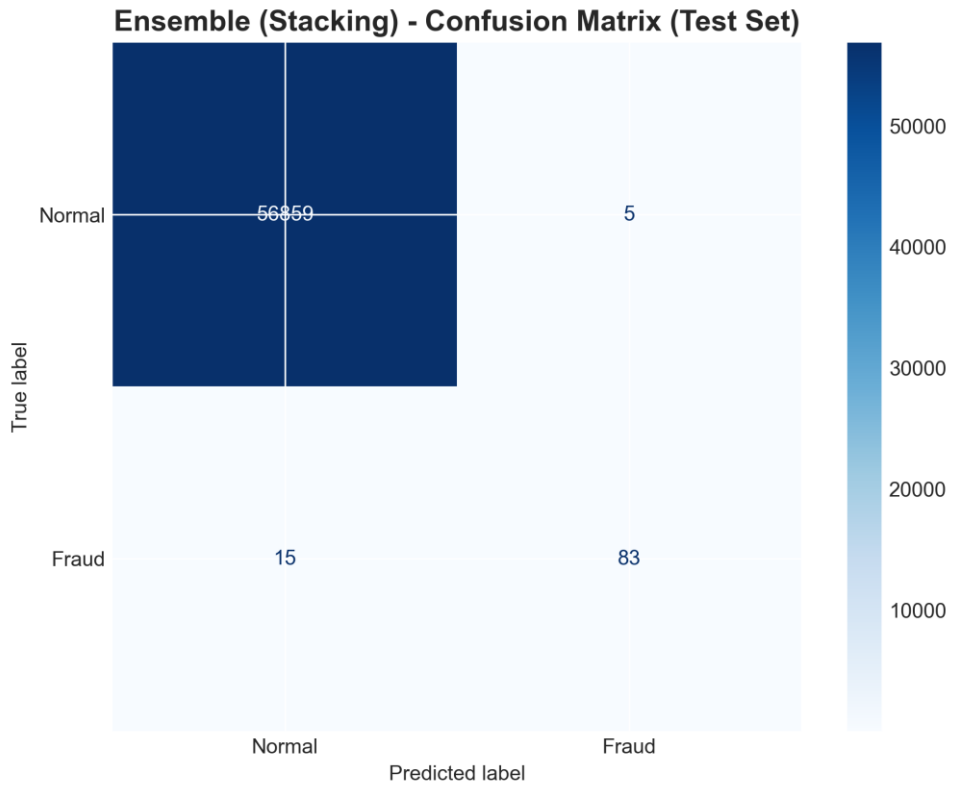

圖 8、Stacking Ensemble 測試集混淆矩陣圖

測試集共有 56,962 筆交易，其中 98 筆為詐欺交易(Class=1)。在此設定下，混淆矩陣為真正常(TN)56,859 筆、假正類(FP)5 筆、假負類(FN)15 筆、真詐欺(TP)83 筆，由此可得測試集上的主要指標為：

- Precision(Fraud) ≈ 0.9432
- Recall(Fraud) ≈ 0.8469
- F1-Score(Fraud) ≈ 0.8925
- ROC-AUC ≈ 0.9805
- Accuracy ≈ 99.96%

從結果可以得到幾點觀察：

(1) 誤報率極低：

在 56,864 筆實際正常交易中，僅有 5 筆被錯判為詐欺，約占正常交易的 0.009%。對實務上的風控團隊而言，代表模型發出的警示大多是有意義的，可減少人工覆核負擔與對客戶的干擾。

(2) 能抓到約 8.5 成的詐欺交易：

在 98 筆實際詐欺中，模型成功偵測出 83 筆(Recall ≈ 0.8469)，仍有 15 筆被誤判為正常。這說明即使在經過 SMOTE、權重調整與閾值優化之後，模型仍無法完全消除漏偵情形，但已能在極度不平衡的情境下維持相當水準的偵測能力。

(3) Precision 與 Recall 之間取得合理折衷：

相較於較低的閾值設定，0.65 在測試集上適度犧牲少量 Recall，換取更高的 Precision，使得每一筆被標記為「詐欺」的交易中，有超過九成是真正可疑個案。這種折衷較符合實務上「同時重視漏報與誤報」的需求。

此外，本研究亦將測試集中每一筆交易的實際標籤、預測標籤與模型輸出機率整理為 prediction_results.csv 檔案，作為後續進一步分析的基礎。

綜合而言，Stacking Ensemble 搭配閾值 0.65 在測試集上展現了良好的整體區分能力與詐欺偵測效能，在極度不平衡的資料條件下，能在 Recall 與 Precision 之間取得兼顧實務需求的平衡，為後續部署與延伸研究提供可靠的基準模型。

第五章、SHAP&LIME 模型可解釋性分析

在信用卡詐欺偵測任務中，即便模型能在實驗中達到良好的 F1-score 與召回率，若其決策過程缺乏透明度，仍難以通過金融監管的審查，也無法被實務風控單位採納。為此，本研究結合 SHAP 與 LIME 兩種主流可解釋人工智慧方法，從全局與局部層次分析最終模型的行為。本章將呈現 SHAP 如何揭示模型整體依賴的關鍵特徵，以及 LIME 如何針對個別交易提供可追溯的決策理由，形成一套能支援實務部署的完整可解釋性分析框架。

第一節、SHAP 全局特徵重要性分析

SHAP 基於 Shapley value，能衡量每個特徵在模型所有預測中的平均貢獻度。本研究使用最終選用之 LightGBM（class_weight）模型於測試集計算 SHAP 值，並繪製前二十名特徵的重要性，如**圖 9** 所示。

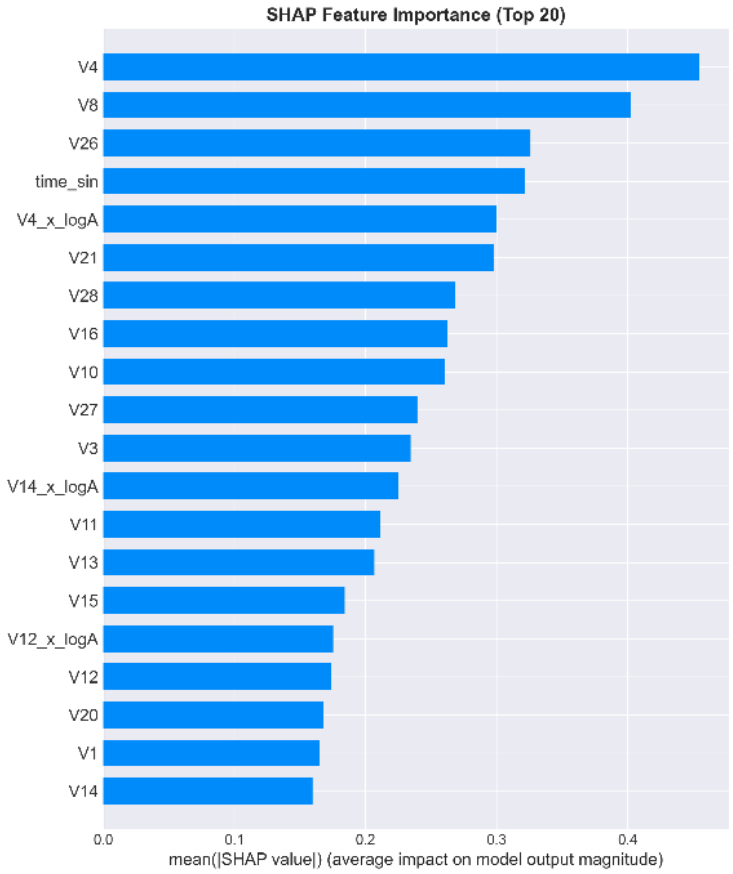

圖 9、SHAP Feature Importance（Top 20）

從圖中可以觀察到三個主要結論：

1. PCA 特徵是模型的主要訊號來源

例如 V4、V8、V26 具有最高平均 SHAP 值。這些 PCA 特徵雖缺乏直觀語意，但 SHAP 顯示模型大量依賴它們來判斷異常行為，顯示詐欺交易在高維空間中的行為模式確實與一般交易不同。由於 PCA 已將多個原始行為變數整合成具變異度代表性的軸向，其 SHAP 重要性提高代表模型成功捕捉了詐欺特徵在特徵空間的偏離程度。

2. 本研究的特徵工程明顯提升模型性能

衍生特徵如 V4×log_amount、V14×log_amount 在排名中名列前茅。這說明模型並非只依賴單純的「高金額」，而是辨識出「高金額結合特定高風險行為模式」的更複雜風險結構。此結果也驗證了本研究透過 SHAP 指引後設計的交互特徵，確實提升模型對詐欺行為的敏感度。

3. 時間週期特徵具有高度辨識力

time_sin 的 SHAP 重要性高於多數 PCA 特徵，代表時間週期性與詐欺風險具有關聯。許多詐欺交易並非在典型消費時間出現，而是於非高峰時段大量發生。透過將 Time 轉換為週期座標，本研究成功讓模型辨識時間段異常行為，使得時間特徵成為重要訊號來源。

第二節、SHAP 特徵影響方向與模型行為解讀

為了解特徵值在不同取值下如何影響模型判斷，本研究使用 SHAP Summary Plot（如圖 10 所示），呈現每個特徵的取值（高/低）對該筆交易詐欺機率的推動方向。

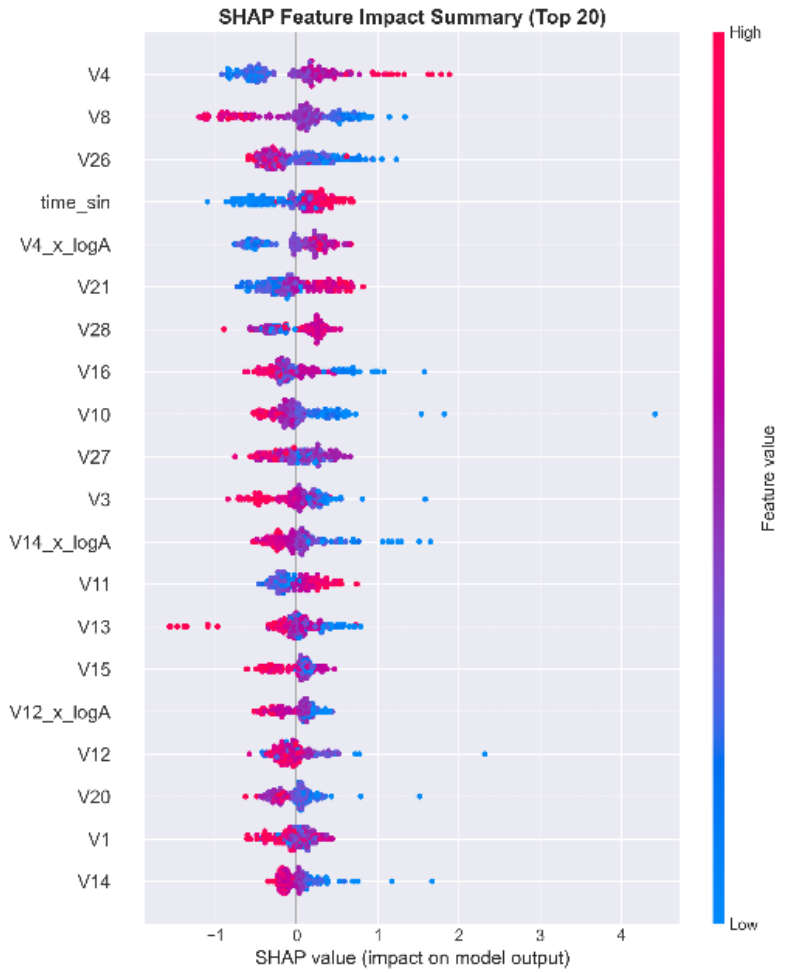

圖 10、SHAP Feature Impact Summary（Top 20）

SHAP Summary Plot 帶來三項重要觀察：

1. 高 PCA 特徵值與詐欺高度相關

例如 V4、V8、V26 高值（深紅點）普遍對應正 SHAP 值，代表模型認為此類特徵的「異常高值」具有明顯詐欺訊號。這反映詐欺行為在 PCA 空間中具有更突出的極端模式。

2. 金額與行為模式的交互異常具有強烈風險訊號

交互特徵 V4×log_amount、V14×log_amount 在高取值時 SHAP 貢獻急遽上升。這說明模型偵測的是「高金額 × 不尋常行為模式」的組合，而非單純的高金額消費。此類訊號通常出現在跨境盜刷、機器人自動詐欺行為中。

3. 時間週期特徵能辨識非典型時段的異常交易

time_sin 的紅色分布集中於正 SHAP 區域，顯示詐欺交易集中於特定「非正常交易時間」。這意味著詐欺者常利用深夜、非營業時段進行測試交易，使時間成為模型重要的判斷依據。

第三節、LIME 個案解釋：詐欺交易案例

在完成 SHAP 的全局分析後，本研究進一步以 LIME 針對個別交易進行局部可解釋性分析，以了解模型在單筆交易尺度上的判斷邏輯。由於 LIME 能在樣本鄰域內建立可解釋的線性模型，因此非常適合用於金融風控中「少量高風險案例的人工覆核」情境。

為說明模型如何識別詐欺，本研究選取一筆模型預測為高風險的交易，並將其 LIME 結果如圖 11 融入以下分析內容。圖中可見模型針對該筆交易給出 Fraud = 1.00 的預測機率，顯示模型對其詐欺性具有高度信心。特徵貢獻面向中，V14_extreme、V12、V3、V17、V16 等特徵皆呈現明顯正向貢獻，使模型的局部線性分類器強烈偏向「詐欺」方向。尤其是交互特徵 V14 × V4，其極端取值對模型預測造成顯著推力，形成典型「高金額 × 異常行為模式」的高風險組合，反映出常見於盜刷、跨境測試交易等詐欺手法的行為軌跡。

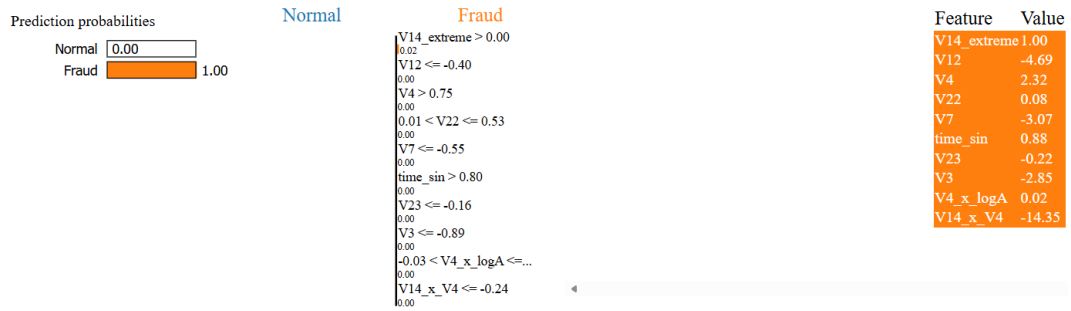

圖 11、詐欺交易案例之 LIME 局部特徵解釋圖

從圖中可以觀察到，模型並非依賴單一特徵，而是由多項特徵共同形成一致的高風險訊號。這種「多軸向同方向異常」的特徵結構，驗證模型的決策具有高度一致性與合理性，同時也使人工審查人員能在實務場景中清楚理解模型標記該交易為高風險的原因。

第四節、LIME 個案解釋：正常交易案例

為檢驗模型是否具備穩定且合理的低風險判斷能力，本研究同樣選取一筆模型預測為正常之交易進行 LIME 解析。此案例對應的 LIME 結果如圖 12 所示，可見模型將該筆交易的 Normal 預測機率提升至接近 1.00，代表該交易整體行為模式落於模型認定的安全區間。

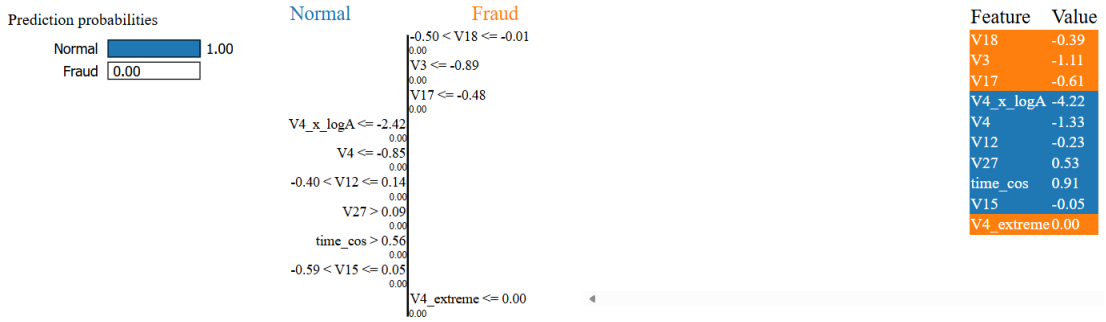

圖 12、正常交易案例之 LIME 局部特徵解釋圖

如圖中所示，模型在此案例中的主要特徵貢獻呈現大量負向或弱正向的權重，包括 V13、V17、V12、V15、time_cos 等，這些特徵的取值均與一般用戶之交易行為高度一致。交互特徵 V14xlogA 在此交易中亦呈現低值，使模型認定金額與行為間不存在異常組合，進一步降低該筆交易的詐欺可能性。

從特徵貢獻方向可以清楚看見與詐欺案例的對比：詐欺交易的特徵呈一致正向推動，而正常交易的特徵則呈現一致負向拉回。這種鏡像式的風險結構差異，不僅強化模型的解釋性，也顯示其具備有效區辨正常與詐欺行為的能力，有助於降低誤報率並避免干擾顧客的正常交易。

第五節、模型可解釋性綜合討論

綜合 SHAP 與 LIME 的分析結果，本研究在模型可解釋性方面得到以下幾點結論：

1. 全局層級(Global)：關鍵特徵清晰可見

SHAP 特徵重要性圖顯示，模型最依賴的特徵包括多個 PCA 成分與金額相關的特徵，以及時間週期特徵。這些特徵的組合與實務上對詐欺行為的理解相吻合，高金額、異常時間、以及某些隱含的交易模式，往往與詐欺行為高度相關。

2. 方向與行為(Directionality)：高值/低值如何影響預測

SHAP 摘要圖說明，高金額與某些 PCA 成分的高值，普遍會增加詐欺機率，反之，落在一般範圍內的特徵則傾向支持正常交易分類，這種「高值→詐欺」、「一般值→正常」的關係，在多數關鍵特徵上具備一致性，有助於確保模型行為不違反基本風控直覺。

3. 個案層級(Local)：LIME 提供交易級別的理由說明

對於被判為詐欺的個案，LIME 能列出具體特徵與其正負貢獻，告訴風控人員這筆交易被標記為高風險，是因為哪些條件組合，對於一般交易，LIME 也能反向說明為何模型認為其風險較低，例如金額不大、時間正常、關鍵 PCA 特徵落在安全區間等，這使得模型的決策可以在單筆交易的尺度上被檢視與質疑，而不是一個黑箱。

4. 實務部署價值與監管友善度

在金融產業中，模型不僅需要有高偵測率，也必須能對監管機關與內部稽核提供合理解釋，本研究的架構結合 SHAP 與 LIME，可將「整體風險因子」與「個案解釋報告」一併輸出：

- SHAP：用於定期檢視模型是否仍依賴合理的特徵組合。
- LIME：用於個別高風險交易的人工審核與決策佐證。

綜合而言，本研究不僅在偵測效能上達到良好表現，更透過 SHAP 與 LIME 建立起一套可解釋、可審核、可溝通的模型說明框架，為未來在實務風控系統中部署此類信用卡詐欺偵測模型提供有力支撐。

第六章、結論與未來展望

第一節、研究成果總結

本研究以公開之信用卡詐欺資料為基礎，針對金融情境中高度類別不平衡與行為特徵抽象化的問題，設計並驗證了一套具備高效能與高可解釋性的詐欺偵測模型。研究首先透過資料前處理與特徵工程強化交易行為之描述，並以 SMOTE、class_weight 與 scale_pos_weight 等方法改善不平衡資料問題。接著，模型採用 LightGBM、XGBoost 與 Logistic Regression 作為主要分類器，並搭配 Optuna 進行超參數搜尋，以提升模型泛化能力。

在效能評估方面，本研究採取三階段資料切分（訓練、驗證、測試）並嚴格避免資料洩漏，僅於訓練階段使用 SMOTE。最終模型於測試集上的結果顯示，調整後的 LightGBM(class_weight)搭配最佳化分類閾值，能有效提升少數類別的 F1-score 與 Recall，使模型在降低誤報率與提升偵測能力兩個面向均取得平衡。此外，本研究使用 SHAP 進行全局特徵重要性分析，並利用 LIME 對代表性樣本進行個案解釋，成功釐清模型判定詐欺與正常交易的邏輯差異，展現模型良好的決策透明度。

綜合而言，本研究不僅在效能層面達成提升詐欺偵測能力的目標，更透過可解釋性分析確保模型具備金融監管環境中所需之可稽核性，使本研究成果兼具實務價值與部署可行性。

第二節、未來研究方向

未來的研究可從多個面向深化本研究成果。首先，建議擴充資料來源，導入來自不同國家、不同期間或多家金融機構的交易資料，以提升模型在多元市場中的穩健性與泛化能力；同時若能引入更多時間序列與使用者行為軌跡資料，將有助於發展具備長期行為辨識能力的時序型模型。此外，成本敏感學習亦為重要方向，可使模型在訓練過程中直接反映「漏報詐欺」與「誤報正常交易」所造成的

風險差異，更符合金融機構實務需求。另一方面，考量詐欺案件常具有網路化與群體行為特徵，未來可結合圖神經網路（GNN）偵測不同交易或帳戶之間的潛在關聯；同時導入線上學習或增量式模型，使偵測系統能隨詐欺手法快速變化而持續更新。綜合而言，若能整合多來源資料、進階模型架構與動態學習策略，將有助於建立更具彈性、可擴充且符合金融情境的智慧詐欺偵測系統。

參考文獻

1. Iqbal, A., & Amin, R. (2025). An efficient mechanism for time series forecasting and anomaly detection using explainable artificial intelligence.
   The Journal of Supercomputing, 81(4), 523.
2. Cui, Y., Han, X., Chen, J., Zhang, X., Yang, J., & Zhang, X. (2025). FraudGNN-RL: a graph neural network with reinforcement learning for adaptive financial fraud detection. IEEE Open Journal of the Computer Society.
3. Zeng, Q., Lin, L., Jiang, R., Huang, W., & Lin, D.(2025). NNEnsLeG: A novel approach for e-commerce payment fraud detection using ensemble learning and neural networks.Information Processing & Management, 62(1),103916.
4. Bahnsen, A. C., Stojanovic, A., Aouada, D., & Ottersten, B. (2013). Cost sensitive credit card fraud detection using Bayes minimum risk. 2013 12th International Conference on Machine Learning and Applications (ICMLA), 333–338.
5. Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. Statistical Science, 17(3), 235–255.
6. Carcillo, F., Le Borgne, Y.-A., Caelen, O., Kessaci, Y., Oblé, F., & Bontempi, G. (2019). Combining unsupervised and supervised learning in credit card fraud detection. Information Fusion, 52, 128–142.

7. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357.
8. Chen, W., Yang, K., Yu, Z., Shi, Y., & Chen, C. L. P. (2024). A survey on imbalanced learning: Latest research, applications and future directions. Artificial Intelligence Review, 57, Article 137.
9. Dal Pozzolo, A., Caelen, O., Le Borgne, Y.-A., Waterschoot, S., & Bontempi, G. (2014). Learned lessons in credit card fraud detection from a practitioner perspective. Expert Systems with Applications, 41(10), 4915–4928.
10. Dal Pozzolo, A., Caelen, O., Le Borgne, Y.-A., Waterschoot, S., & Bontempi, G. (2015). Learned lessons in credit card fraud detection from a practitioner perspective (extended work). In Learning under extremely imbalanced data distributions (technical report).
11. Jurgovsky, J., Granitzer, M., Ziegler, K., Calabretto, S., Portier, P.-E., He-Guelton, L., & Caelen, O. (2018). Sequence classification for credit-card fraud detection. Expert Systems with Applications, 100, 234–245.
12. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems (NIPS 2017).
13. Whitrow, C., Hand, D. J., Juszczak, P., Weston, D., & Adams, N. M. (2009). Transaction aggregation as a strategy for credit card fraud detection. Data Mining and Knowledge Discovery, 18, 30–55.