基於集成學習框架之信用違約預測－以信用卡客戶為例 (post-print)

江彌修*

Mi-Hsiu Chiang

胡聚男**

Chu-Nan Hu

黃立新***

Li-Xin Huang

陳靜怡****

Ching-Yi Chen

* 通訊作者：國立政治大學金融學系，臺北市文山區指南路二段 64 號；電話：(02)29393091 分機:81265；傳真：+886-2-9398004；Email：mhchiang@nccu.edu.tw。
Department of Money and Banking, National Chengchi University

** 國立政治大學金融學系
Department of Money and Banking, National Chengchi University

*** 國立政治大學金融學系
Department of Money and Banking, National Chengchi University

**** 國立政治大學金融學系
Department of Money and Banking, National Chengchi University

摘要

藉由堆疊 (Stacking) 與勻合 (Blending) 學習器 (base estimators) 所產生的異質性集成學習(Heterogeneous Ensemble Learning)框架，本文建構消費金融信用卡客戶之違約風險預警模型。採用 Yeh and Lien (2009) 的資料集，我們的實證結果顯示，堆疊與勻合集成皆能有效降低誤判信用違約客戶為正常的型二誤差。尤其當輔以適當的學習器挑選策略，堆疊集成的綜合辨別能力呈現一定的泛化優越成效，明顯勝出任何非集成的單一學習器。更加地，輔以學習器挑選策略的堆疊集成能夠提高模型識別違約客戶的準確率 (F_1 值)，且在增進識別違約客戶能力的同時有效降低誤判正常客戶為違約的分類誤差 (AUC 值)。

關鍵詞：信用風險、違約風險、信用卡客戶、集成學習、機器學習

Credit Default Prediction Based on Ensemble Learning－The Case of Credit Card Customers (post-print)

Abstract

Based on Heterogeneous Ensemble Learning that allows for the Stacking and Blending of base learners of distinct types, in this study we construct an ensemble-learning assisted credit-risk prediction model in an attempt to prewarn consumer banks of their credit card holders' possibility of default. Using the dataset as in Yeh and Lien (2009), our empirical results show that ensemble learning models that exploit either Stacking or Blending can effectively reduce the Type II error in mis-judging defaulted entities as normal. In particular, when equipped with a learner-selection strategy, heterogeneous ensemble learners that exploit Stacking tend to exhibit superior predictive power over all single base learners. Furthermore, ensemble learners with Stacking are found to be capable of improving the rate of accuracy in nailing down defaulted entities (F_1-score); they demonstrate the ability to identify credit-critical customers while at the same time reduce the possibility of misjudging normal customers as defaulted ones (AUC-value)

Keywords: Credit Risk, Default Risk, Credit Card Clients, Ensemble Learning, Machine Learning

壹、前言

信用評分 (Credit Scoring) 在信用借貸的決策與風險管理中發揮不容忽略的作用，其應用範圍既涵蓋傳統銀行的信用卡與銀行貸款業務，也涉及金融科技借貸 (Fintech Lending) 的 P2P (peer-to-peer) 業務。信用卡與銀行貸款是銀行主要的獲利來源，然而管理不當也會釀成 2005 年底台灣的雙卡風暴。事件後，金管會規定台灣金融機構於 2006 年底實施新巴塞爾資本協定 (Basel II) ，要求各銀行必須具備足夠的資本適足性及風險管理能力，並建立內部評等系統和風險預警機制以評估信用風險。此外，巴塞爾資本協定提倡銀行使用更加精緻的 (sophisticated) 信用評分系統以及追求更高的鑒別客戶的能力。Van Liebergen (2017) 顯示銀行採用機器學習技術來預測信用卡違約已有十多年經驗，並取得一些顯著成效。另一方面，P2P 借貸被認為是零售銀行領域的重大創新，在平台數量與成交量上都穩步增長 (Dorfleitner et al., 2016)。Jagtiani and Lemieux (2019) 指出，使用另類數據 (alternative data sources)、大數據、機器學習以及其他複合人工智慧算法可以減少信用決策以及信用監管的成本，降低貸方操作成本；金融科技的貸方可能潛在地將利益轉移到借方。本文以台灣信用卡數據為例，引入堆疊 (Stacking) 與 勻合 (Blending) 兩種集成學習因應傳統銀行借貸與金融科技借貸之於不同面向的成本（誤判成本與時間成本）考量，為更精緻的信用評分系統提供理論支援，也為後續的信用評分模型之發展提供比較之基準。

傳統信用風險衡量方法以區別分析法 (Discriminant Analysis, DA) 與羅吉斯迴歸 (Logistic Regression Analysis, LRA) 為主，這類方法都屬於線性模型，優點在於模型簡單且通常有不錯的解釋力，其缺點則是在訓練高維度或是特徵複雜的數據時，這些線性模型的表現較差。近年來，機器學習伴隨金融科技的浪潮，已成為財務領域學術研究的熱門研究方法之一，機器學習具有適應力強且處理不同型態資料的優點，適合分析高度複雜數據，應用領域亦相當廣泛。

Desai et al. (1996) 將類神經網路應用在信用評分上，並與傳統統計方法常用的區別分析以及羅吉斯迴歸進行比較，研究結果顯示在以違約借款人之分類正確率為績效指標下，類神經網路模型預測表現最佳；而以整體預測準確率為衡量指標下，類神經網路模型的預測成效與羅吉斯迴歸模型相當。Yeh and Lien (2009)

使用台灣某大型銀行 2005 年的信用卡客戶資料，建構六種模型以預測銀行信用卡客戶之逾期繳款機率，分別有羅吉斯迴歸、線性區別分析、單純貝氏分類法 (naive Bayes classifier) 、K 近鄰分類演算法 (K-Nearest Neighbor, KNN) 、類神經網路模型 (Artificial Neural Networks, ANN) 以及分類樹 (Classification trees) 共六種學習方法。而在以提升圖下方面積 (Area Under Curve, AUC) 來評估模型預測能力下，類神經網路模型亦表現最佳，區別分析模型與羅吉斯迴歸則表現最差。本研究所採用的研究樣本資料由 Yeh and Lien (2009) 提供，並公開放置於 UCI 資料庫中。[1]

然而，多數研究都只僅使用單一演算法進行預測，即便該文採用多種以上的演算法，預測時也僅是採用各別演算模型進行分類預測與互相比較。美國人工智能協會 (Association for the Advancement of Artificial Intelligence, AAAI) 前主席與機器學習領域權威 Thomas G. Dietterich 指出任何一個單一分類模型都會面臨一定的錯誤分類風險，若將多個模型集成起來通常能夠降低整體分類錯誤的風險。由於各個分類演算法皆具不同的特性，也有其適合的應用時機，許多學者們為了更加提升模型之泛化能力 (generalization ability) 與預測表現，而使用集成學習框架 (ensemble learning) 結合多種相異特性之學習器建立預測模型。因此，本研究選擇以堆疊與勻合集成作為主要研究方法，衡量銀行信用卡客戶之違約風險。信用卡客戶違約資料存在類別不平衡 (class imbalance) 問題，即正常客戶數量遠大於違約客戶，本文所使用資料中正常客戶共有 23,364 筆，違約客戶共有 6,636 筆。當以預測準確率為績效指標時，機器學習模型則會產生傾向於預測客戶為正常類別的偏誤；更多關於類別不平衡問題的討論可以參考 Guo et al. (2017) 的回顧文章。在處理訓練資料的過程中，本研究引入合成少數類過取樣技術 (Synthetic Minority Oversampling Technique , SMOTE) 處理違約預測樣本常見的類別不平衡 (class imbalance) 情況。

Xia et al. (2018) 使用結合 Bagging 與模型堆疊的集成框架構 (bstacking) 以支援向量機，隨機森林，GPC (Gaussian process classifier) 與 XGBoost 為第一層

[1] 資料網址：https://archive.ics.uci.edu/ml/datasets/default%2Bof%2Bcredit%2Bcard%2Bclients?fbclid=IwAR05_oGqeUZDNMofvuz4q4O3wH180-JRgMY4d52WLsYmuha4ulu33n11omc。

學習器，建構信用評分系統並研究在信用卡與 P2P 借貸的應用。其資料來自 UCI 資料庫的德國與澳大利亞信用資料以及分別來自中國(人人貸)與美國 (Lending Club) 的 P2P 資料。作者發現 bstacking 在多種衡量指標驗證下在信用卡與 P2P 借貸資料集上都有優異表現。而本研究不同於 Xia et al.有三點：一、本文在集成框架中納入文獻上表現最好的單一算法類神經網路，而 Xia et al.則以類神經網路計算費時拒絕納入類神經網路。本文沒有刻意調整學習器參數，卻發現支援向量機的訓練時間 (275 秒) 長於類神經網路 (46.7 秒)；二、不同於 Xia et al.注重集成學習第二層的集成策略，本文在保持模型簡潔的同時更關注第一層學習器的挑選策略,並不同模型衡量指標下加以檢驗；三、本文在模型堆疊集成框架之外更引入模型勻合，在考量模型表現的同時也兼顧計算效率。

本文的實證結果顯示三點主要發現。第一，當輔以學習器挑選策略，堆疊集成的綜合辨別能力呈現泛化的優越成效，明顯勝出任何非集成的單一學習器。以型二誤差而言，採用堆疊與勻合集成皆能有效降低將違約客戶歸類為正常的錯誤。以 F_1 值而言，輔以學習器挑選策略的集成模型皆高於最佳單一模型(隨機森林)，而不加挑選的集成模型皆低於隨機森林；實證表明缺乏挑選策略的集成學習無法在識別違約客戶的能力穩定地勝過單一模型。以 AUC 值而言，僅剔除支援向量機的模型堆疊勝過最佳的單一模型(類神經網路)，結果表明缺乏學習器挑選策略與框架設計的集成模型不能穩定的在辨別客戶的能力上勝過最佳的單一模型。

第二，我們檢視集成模型的學習器挑選策略與框架設計。不加挑選的集成學習在型二誤差、F_1 值與 AUC 值三個衡量指標下皆差於對應的輔以學習器挑選策略的集成學習，我們推測其原因是支援向量機與隨機森林的模型預測結果高度相關，同時採用則違反了學習器篩選準則的第二條--學習器之間需要具足夠的差異性。進一步比較堆疊與勻合集成，發現模型勻合模型在相同情境下的表現皆弱於模型堆疊，其原因是模型勻合對訓練資料的利用程度不如模型堆疊，從樣本中萃取的資訊有所缺失。本文的研究發現指出不加選擇地使用集成學習，盲目的添加學習器可能導致集成模型表現甚至不如單一模型。

第三，本研究亦計算集成模型的訓練時間，為兼顧模型表現與時間成本之考量提供決策依據。同等情境下，模型堆疊的訓練時間長於模型勻合，這主要因為

模型堆疊在訓練單一模型的過程中採用 K-折交叉驗證的技巧所致。集成模型中剔除支援向量機的模型勻合的訓練成本 (80 秒) 低於單一模型的支援向量機訓練成本 (275 秒)。

本研究架構如下：第貳節為文獻回顧，第一部分說明過去文獻如何探討信用分析，第二部分則為信用分析之評分方式。第參節為研究方法，詳述本研究建構預測模型的過程，主要介紹堆疊與勻合集成兩種集成學習模型之建構方式，以及模型預測能力衡量指標。第肆節為資料分析，包含資料前處理與如何處理原始資料類別不平衡。第伍節為實證結果，包含介紹模型超參數 (hyperparameter) 設定，隨後呈現本研究於單一模型與集成模型的預測表現，並進一步比較兩者之間的差異，與探討集成模型於學習器上的篩選是否會影響其效能。第陸節為研究結論與建議，內容涵蓋本研究的實證結論，而對於信用違約預測領域的學術研究，本研究亦給予可行的建議與未來發展方向。

貳、文獻回顧

一、消費者信用分析

信用風險分析銀行風險管理的重要議題，而銀行組織結構通常依照放款對象分為企業金融與消費者金融部門。本研究主要係以消費者放款的信用分析為主。在 1940 年代，美國的金融機構與郵購公司在貸款的決策上依照經驗法則建構專家系統 (expert systems) 來評估放款與否。而 Durand (1941) 則是單純地利用樣本平均數等敘述統計量區分貸款對象之良莠，其消費者信貸等級的分類因素可分為財務因素與非財務因素兩類；財務因素包含所得、貸款金額、貸款期間、擔保等、非財務則包含貸款者的工作穩定、居住穩定、工作特質與貸款目的等。

在隨後的 1950 年代則出現了以更進階的統計模型作為信貸決策的諮詢公司 (Thomas, 2000)。1960 年代誕生的信用卡產品則催生出了信用評分 (credit scoring) 系統。信用評分模型能幫助金融機構對個人消費者進行評價，從而鑒別出高違約風險群體 (Hand and Henley, 1997)。信用評分模型則主要根據申請人的特徵給予計分，例如 Desai et al. (1996) 根據貸款人年紀、信用卡個數、是否擁有房屋、薪資與其他收入金額、工作年資、現址居住年資、先前借款筆數、過去七個月申請次數等)，將其歸類為「好」的債務人與「壞」的債務人。相較於使用專家系統，信用卡發行機構使用信用評分系統擁有更高的預測準確率，違約機率下降 50% 以上 (Myers and Forgy, 1963; Churchill et al., 1977)。

1980 年代之後，則有更多統計方法被引入到信用評分中。線性區別分析模型是第一個信用評分模型，此模型係由 Reichert et al. (1983) 提出並被沿用至今日。線性區別分析模型雖然簡單且具有不錯的解釋力，但 West (2000) 指出線性區別模型的缺點在於假設是變數服從常態分佈，然而信用相關的數據卻通常不是常態分佈的，且接受的分類與拒絕的分類的相關性矩陣很可能是不一致的，使其實際應用時受到限制。為了克服線性區別模型的缺陷，更多統計方法或機器學習工具被應用於信用評分領域，包含羅吉斯迴歸分析 (Henley, 1995)、K 近鄰 (Henley and Hand, 1996)，支援向量機 (Baesens et al., 2003; Schebesch and Stecking, 2005; Huang et al., 2007; Martens et al., 2007)、決策樹 (decision tree, DT) (Davis et al., 1992)、類神經網路 (Desai et al., 1996; West, 2000) 等。

Desai et al. (1996) 研究三家美國東南部信用合作社 1988-1991 的放款數據，比較類神經網路模型與其他機器學習模型，發現類神經網路在放款違約的分類表現最佳；但假如是以整體的預測能力而言，則羅吉斯迴歸與類神經網路有類似的表現。 Baesens et al. (2003a) 使用來自比盧荷聯盟 (Benelux) 或英國金融機構等 8 組信用評分資料集，比較羅吉斯迴歸、最近鄰居、類神經網路、決策樹、支援向量機與最小平方支援向量機 (least-squares support vector machine, LS-SVM) 等不同學習器的結果。研究結果發現雖然類神經網路與 LS-SVM 在分類準確率與 AUC 指標的表現最佳，但是像羅吉斯迴歸或線性區別分析等簡單的學習器也有不錯的預測表現。

二、集成學習應用於信用評分模型

集成算法是透過訓練多組個體學習器 (individual learner) 以解決同一個問題 (Polikar, 2006)。在集成學習中，組成集成學習的學習器也被稱為基礎學習器 (base learner)；其中，又可分為弱學習器 (weak learner) 與強學習器 (strong learner)。前者是指一個比隨機預測稍微好的學習器，而後者則是能產生相當準確的預測。

相較於單一算法，近年來的研究顯示集成學習算法通常能帶來更佳的分類預測效果 (Yu et al., 2008; Hung and Chen, 2009; Wang et al., 2011) 。Dietterich (1997) 視機器學習為搜尋包含最準確假說的假說空間 (hypothesis space)，並提出三個集成學習優於單一學習器的解釋：(1) 訓練資料未提供足夠的資訊以支持挑選出最優的單一學習器，例如存在多個表現相當的單一學習器，而結合這些單一學習器可能產生更佳的表現；(2) 算法的搜尋程序可能不完美，以至於搜尋程序只能找到次優的假說而非最佳假說，而集成學習可能補償不完備的搜尋進而產生更佳的表現；(3) 所搜尋的假說空間可能未包含目標函數，而集成學習則可能產生良好的近似效果。

Zhou (2009) 認為 Hansen and Salamon (1990) 以及 Schapire (1990) 等兩篇文章帶動了 90 年代集成學習的研究熱潮。Hansen and Salamon (1990) 發現結合使用一系列學習器的預測結果通常比使用單一學習器的預測結果更準確；而 Schapire (1990) 證明弱學習器通過使用 Boosting 算法可以產生強學習器。

Windeatt and Ardeshir (2004) 認為構建一個好的集成學習需要滿足兩個條件：(1) 基礎學習器的預測準確率越高好，(2) 基礎學習器彼此間的差異性 (diversity) 越高越好。

應用於信用評分的集成算法主要有三類，分別為 Bagging，Boosting 以及 Stacked Generalization[2] (Wang et al., 2011)。Bagging 是 **B**ootstrap **agg**regat**ing** 的縮寫，由 Breiman (1996) 提出，是一個符合直覺、運用簡單且具備良好表現的集成算法。Bagging 使用拔靴法 (bootstrap) 的對原始訓練樣本進行抽樣並構建出新的訓練資料子集，基礎學習器使用資料子集進行獨立且並行的訓練，最後以多數決 (majority vote) 的策略集成。隨機森林 (Random Forest) 是 Bagging 的代表算法。

Boosting (Schapire, 1990; Freund and Schapire, 1996) 則是另一大主流集成學習算法。與 Bagging 相似，Boosting 的基礎學習器也是使用不同的訓練集；不同的是 Boosting 的基礎學習器是相關聯且依序進行訓練的。透過重新加權歸類錯誤的樣本的權重，使得基礎學習器在訓練過程中更容易學習到先前的錯誤，進而提升預測準確率。Boosting 的集成策略是根據基礎學習器的表現給予加權 (weighting) 並線性加總。AdaBoost, XGBoost 都是 Boosting 的代表算法。

Bagging 與 Boosting 皆屬於同質集成學習(Homogenous ensemble learning)，其原理為基於相同的學習器產生多個基礎學習器，而基礎學習器的差異性則依賴與抽樣方法。由於 Bagging 對原始樣本進行有放回的隨機抽樣產生多個數據子集，則部分客戶的資料可能在子集中重複出現，而部分客戶則資料則可能被排除在外。基於不同的數據子集，基礎學習器因此產生差異，Bagging 透過基礎學習器的變異 (variance) 降低預測誤差。因此，Bagging 適用於基礎學習器對數據擾動敏感的情境。不同於 Bagging 以並行的方式訓練模型，Boosting 則是透過對原始數據進行加權迭代的方式分階段訓練模型。在每一輪迭代中，提高本輪預測錯誤的客戶資料的權重，從而為下一輪迭代提供更富含信息的資料。Boosting 以避免上一輪錯誤的方式降低預測誤差。所以，Boosting 適用於基礎學習器對數據偏

[2] 此處，我們使用 Stacked Generalization 的術語而非更常見的 Stacking，是為了方便後文的 Blending 算法與 Stacking 算法的區分。

誤 (bias) 敏感的情境。更詳細的討論可以參考 Hastie, et al. (2009)， Kuncheva (2004)，Lessmann et al. (2015)， Lin et al. (2012) 以及 Marqués, et al. (2012)。

有別於 Bagging 和 Boosting，堆疊泛化 (Stacked Generalization) 則屬於異質集成學習，其背後的思想為異質的學習器可能以不同的角度看待相同的趨勢，進而在預測時互補並降低誤差。堆疊泛化使用一個次級基礎學習器 (high-level base learner) 結合較初級基礎學習器 (lower level base learners) 以構建元學習器 (meta-level base learner)，進而達到更高預測準確率 (Wolpert, 1992)。模型堆疊就是堆疊泛化的代表算法。

模型勻合的演算法本質 (Töscher et al., 2009) 與模型堆疊非常相似。模型勻合係由 Netflix 競賽的優勝團隊推出的演算法，雖然該演算法並非準確率最高的算法，但其運算效率較高。與模型的優異表現不相稱的是，集成學習受到的關注較少，而據我們所知，本文是第一篇將模型勻合應用於信用評分的文章。堆疊與勻合集成的不同之處在於兩點，第一，訓練數據的分割。不同於模型堆疊使用的全部的原始訓練集，模型勻合把原始訓練集分為新訓練子集與驗證集 (validation set)，基礎學習器只使用新訓練子集訓練；第二，模型勻合的基礎學習器的預測值數量遠少於模型堆疊。模型勻合僅使用驗證集輸入已經訓練好的基礎學習器，每個基礎學習器產生與驗證集相同個數的預測值，然則模型堆疊在同等情境下產生等同於原始訓練集大小的預測值。最後，堆疊與勻合集成都是使用基礎學習器產生的預測值對元學習器進行訓練。

本研究將過去使用集成學習探討信用評分的文獻綜整於表 1。由表 1 可知，文獻大多使用 Bagging 或 Boosting 等集成學習以提高模型預測能力，較少使用堆疊泛化模型。Wang et al. (2011) 以羅吉斯迴歸、決策樹、類神經網路與支援向量機四種演算法為基礎學習器，比較 Bagging、Boosting 以及堆疊泛化三種集成算法。實證發現堆疊泛化和基於決策樹的 Bagging 在平均準確率，型一誤差及型二誤差三種衡量標準之下具備最佳表現。然而，Lessmann et al. (2015) 則以模型堆疊作為基準模型，發現模型堆疊表現相當差。Xia et al. (2018) 表示模型堆疊令人失望的表現可能是由於使用過於簡單的學習架構以及元學習器性能過弱。

本研究將嘗試使用強學習器，將 Boosting 與 Bagging 的集成模型做為基礎模型，探討堆疊與勻合集成兩種堆疊泛化方法是否能進一步提高預測效果，以符合巴塞爾銀行監管委員會 (2000, 2005a) 的期待，即採用更為精緻的信用評分模型以提高準確率與適合度，藉此增進資金配置的效率、產品定價與獲利能力。

表 1 集成學習應用於信用評分文獻彙整

論文	基礎學習器	集成方法
West et al. (2005)	NN	MV/Bagging/Boosting
Yu et al. (2008)	NN	Bagging
Nanni and Lumini (2009)	NN, SVM, KNN	Bagging
Paleologo et al. (2010)	SVM, KNN, DT, Adaboost	Bagging
Twala (2010)	LR, NB, DT, KNN, NN	Bagging/Boosting
Yu et al. (2010)	SVM	Bagging
Zhang et al. (2010)	DT	Bagging
Heish and Hung (2010)	SVM, NN, BN	Bagging
Finlay (2011)	LR, LDA, DT, NN, KNN	Bagging/Boosting
Wang et al. (2011)	LR, DT, NN, SVM	Bagging、Boosting、Stacked Generalization
Wang et al. (2012)	DT	Bagging
Tsai et al. (2014)	NN, SVM, DT	Bagging/Boosting
Abellán and Mantas (2014)	DT	Bagging/Boosting
Lessmann et al. (2015)	BN, DT, KNN, LR, NB, NN, SVM 等	Bagging、Boosting、Stacking
Xia et al. (2017)	DT	Boosting
Xia et al. (2018)	SVM, RF, XGBoost, GPC	Stacking
Xia et al. (2020)	Five tree-based algorithms: RF, GBDT, XGBoost, LightGBM, CatBoost.	WA

註: 表中的縮寫如下：BN = Bayesian Network、DT = Decision Tree、GPC = Gaussian process classifier (Williams and Rasmussen, 2006)、KNN = K-Nearest Neighbor、LR = Logistic Regression、NB = Naïve Bayes、NN = Neural Network、RF = Random Forest、SVM = Support Vector Machine、MV = Majority Vote、WA = Weighted Average。

叁、研究方法

一、模型堆疊

模型堆疊結合多種機器學習演算法建構兩層學習器，第一層學習器稱為初級學習器，而第二層學習器稱為次級學習器。為了方便比較不同集成學習框架的預測表現，本研究的理想模型第一層包含類神經網路、隨機森林與 XGBoost 三種模型，利用客戶特徵生成違約機率預測值；模型第二層則是使用羅吉斯迴歸模型作為元模型 (meta-model) ，輸入第一層學習器產生的預測值進行最終預測。

值得指出的是，本研究採用分層 K 折交叉驗證 (Stratified K-fold cross validation) 的技巧，每個子集中樣本的類別比例與原始資料相同，因此訓練集和驗證集是不相關的 (uncorrelated) 且基於測試集產生的預測誤差為無偏估計。Rabinowicz and Rosset (2020) 指出，以平方誤差為損失函數的 K 折交叉驗證並非免於分佈假設；透過添加相關性的修正項的方式，作者建構了在相關資料 (correlated data) 情境下預測誤差的無偏估計值 (unbiased estimate)。

步驟 1 資料切割

模型堆疊在訓練第一層學習器時，每一個模型皆會使用 K 折交叉驗證的技巧，將原本的訓練資料切割為 K 個子集資料 (subsets) ，先從中隨機挑出 1 個子集作為驗證集，另外的 K-1 個子集則作為新訓練集。K 折交叉驗證的特點在於每個子集皆會輪流擔任驗證集的角色，較不容易發生模型過擬合問題。新訓練集將被用來訓練模型，而驗證集將會被輸入至已訓練好的模型並獲得預測結果。由於總共有 K 個子集資料，因此第一層的每個學習器總共會訓練 K 次，並產生 K 組預測結果。

步驟 2 模型訓練階段

本研究中理想模型第一層學習器共有 3 個單一模型，其訓練過程皆相同， 因此我們以類神經網路進行詳細說明，隨機森林與 AdaBoost 可以此類推。為了確保模型的訓練效果，本研究採取分層 K 折交叉驗證的技巧，將原始訓練集劃分為 5 個子集 (K=5)，且每個子集中樣本的類別比例與原

始資料相同。由於每個子集 (subset) 皆會輪流作為驗證集，因此類神經網路模型會被訓練 5 次並輸出 5 個預測結果。將 5 個預測結果合併得到與原訓練集大小(N=38,023) 相同的向量。經過訓練，單一模型皆會產生大小為 38,023×1 的向量，集合 3 個單一模型則可得到大小為 38,023×3 的新特徵資料 meta_x:

$$\text{meta\_x} \equiv \begin{bmatrix} \text{Predict\_ANN}_1 & \text{Predict\_RF}_1 & \text{Predict\_XGB}_1 \\ \text{Predict\_ANN}_2 & \text{Predict\_RF}_2 & \text{Predict\_XGB}_2 \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ \text{Predict\_ANN}_{38023} & \text{Predict\_RF}_{38023} & \text{Predict\_XGB}_{38023} \end{bmatrix}$$

其中，3 個向量分別為類神經網路、隨機森林與 XGBoost 學習器以機率形式輸出的預測結果。在模型堆疊集成學習框架中，這些預測結果所集合形成的 meta_x 作為第二層學習器的訓練資料。

簡而言之，上述概念為第一層中第 j 個學習器對第 i 個訓練樣本的預測值將作為新訓練集 meta_x 中第 i 個樣本的第 j 個特徵值 (j=1 to 3 and i=1 to 38,023) 。最後，模型的第二層學習器，也就是元模型－羅吉斯迴歸會對新訓練集 meta_x 進行學習。

步驟 3 模型測試階段

在模型堆疊集成學習框架中，測試集需先經過第一層中每一個模型的 5 個子模型(K=5)產生預測結果，接著依相同模型，將各子模型預測結果取平均。若為分類問題，模型採用簡單多數投票法，若為連續問題，則採用簡單平均法，以形成與原測試集大小相同的新測試集(meta_y) ，再輸入至第二層學習器進行預測。

二、模型勻合

模型勻合集成學習框架與模型堆疊相似，都結合多種單一演算法，建構兩層學習器，並將第一層學習器輸出的預測結果合併為新特徵集，作為第二層模型的輸入資料以訓練完整的模型。模型勻合對測試集進行預測的方式也與模型堆疊相

似，先將測試集輸入至模型的第一層學習器，再將第一層學習器之預測結果輸入至第二層學習器進行預測，以得到最終的預測結果。然而，模型勻合與模型堆疊不同之處在於對資料的切割與使用。模型勻合沒有使用 K 折交叉驗證，而是將原始訓練集分割成新訓練集與驗證集，先使用新訓練集訓練第一層模型，訓練後的模型結合使用驗證集產生第一層的預測結果。具體過程與步驟如下：

步驟 1 原始訓練集的分割

本研究以 6 比 4 的比例將原始訓練集 (38,023 筆) 劃分為訓練集 (22,813 筆) 與驗證集 (15,210 筆) ，此訓練集用於訓練第一層的所有個體學習器，而驗證集會被輸入至已訓練完成的第一層學習器進行預測，接著模型會再將預測結果整合，作為第二層學習器的新輸入特徵 (meta_x) ，以訓練元模型。

步驟 2 模型建構階段

本研究中模型第一層學習器包含支援向量機、類神經網路與隨機森林模型；首先，全部的模型皆會對訓練集進行學習，接著模型訓練完畢後會對驗證集進行預測，而各個模型對驗證集的預測結果會被集合為新特徵集 (meta_x) ：

$$meta\_x \equiv \begin{bmatrix} prediction\_ANN_1 & prediction\_RF_1 & prediction\_XGB_1 \\ prediction\_ANN_2 & prediction\_RF_2 & prediction\_XGB_2 \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ prediction\_ANN_{15210} & prediction\_RF_{15210} & prediction\_XGB_{15210} \end{bmatrix}$$

其中公式第一行的 3 個向量分別為類神經網路、隨機森林與 XGBoost 模型以機率形式輸出的預測結果，而每個結果序列皆可看作一個新特徵，在模型勻合集成學習框架中，這些新特徵會被集合形成 meta_x，作為第二層學習器的訓練資料。最後，模型的第二層學習器，也就是元模型-羅吉斯迴歸會對新訓練集 meta_x 進行學習。

步驟 3 模型測試階段

在模型勻合集成學習框架中，測試集亦需先經過第一層中的每一個模型，而產生預測結果。接著，各模型對測試集的預測值會被集合為新特徵集，再輸入至第二層學習器進行預測。

三、模型預測能力衡量指標

本研究使用型二誤差、 F_1 值與 AUC 值作為模型預測能力的主要衡量指標，前兩項指標用來評估模型對違約客戶的辨識能力，而 AUC 值用於評估模型整體的配適度與分類能力。雖然準確率為模型最常見的評估指標，但並不適用於預測類別不平衡的資料，故僅供參考。

(一) 混淆矩陣與型二誤差

在分類問題中，欲鑑別一個模型的成效，最常見的工具之一便是混淆矩陣。如表 2 所示，混淆矩陣中的每一欄代表各類別的預測值，而每一列代表樣本真實的分類。本研究中，陽性代表違約客戶，而陰性則代表正常客戶。表格中的真陽性代表真實為違約客戶且被正確預測為違約的類別，偽陽性代表真實為正常客戶但被錯誤預測為違約的類別，在統計學上亦稱型一誤差 (Type I error) ；偽陰性代表真實為違約客戶但被錯誤預測為正常的類別，在統計學上亦稱型二誤差 (Type II error) ，真陰性代表真實為正常客戶且被預測正確的類別。實務上，將違約客戶錯判為正常類別的錯誤造成嚴重呆賬損失，因此型二誤差是備受重視的指標。

表 2　混淆矩陣

		預測結果		
		預測為陽性	預測為陰性	總計
實際情況	真實為陽性	真陽性 (True Positive, TP)	偽陰性 / 型二誤差 (False Negative, FN)	TP+FN
	真實為陰性	偽陽性 / 型一誤差 (False Positive, FP)	真陰性 (True Negative, TN)	FP+TN
總計		TP+FP	FN+TN	

基於混淆矩陣，可以方便的定義的衡量指標。以最常見的準確率 (Accuracy) 為例，準確率衡量模型整體正確率，代表所有資料中有多少比例是正確預測的。

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (1)$$

然而在資料類別不平衡的問題中，學習器會傾向將所有資料判斷為正常客戶，使得準確率失真，因此準確率並非評價信用評分模型的良好指標。

(二) F_1 值

F_1 值或稱 F_1-Measure，為精確度 (precision) 與召回率 (recall) 的調和平均數 (harmonic mean) ，用來衡量準確識別出違約類別的能力，是比準確率更適用於類別不平衡的指標。使用調和平均數而非算數平均值，可以更多的懲罰極端值的情境。

$$F_1 - \text{score} = 2 \times \frac{1}{\frac{1}{precision}+\frac{1}{recall}} = 2 \times \frac{precision \times recall}{precision+recall} \quad (2)$$

其中，召回率指標又稱為敏感度 (sensitivity) 或真陽性率 (true positive rate, TPR) ，代表所有真實為違約客戶的資料中多少比例被正確預測，其公式如下：

$$Recall = \frac{TP}{TP+FN} , \quad (3)$$

精確度 (precision) 指標又稱陽性預測值 (positive predictive value，PPV) ，代表預測為違約客戶的結果中有多少比例是真實為違約客戶，其公式為：

$$precision = \frac{TP}{TP+FP} , \quad (4)$$

(三) ROC 曲線與下方面積

ROC 曲線 (Receiver Operating Characteristic curve) 及其衍生的 AUC 值是信用評分行業最常被使用的指標，這可能歸功於其兩個特性。一是，免於樣本類別不均衡造成的扭曲，當測試集中陰陽樣本的分佈改變，甚至在極端比例的情境下，ROC 曲線保持不變 (Fawcett, 2006)。二是，ROC 曲線免於最佳門檻值的選擇，

從而避免人為調參的影響。實務上 ROC 曲線是巴塞爾銀行監管委員會(2005b)建議效力驗證指標的其中一種。

ROC 曲線具體為一個二維圖形，其中 Y 軸為真陽性率 (TPR)，即所有真實為違約的樣本中被正確識別為違約的比例，指標越高越好；X 軸為偽陽性率(False Positive Rate, FPR)，即所有真實為正常的樣本中被錯誤識別為違約的比例，指標越低越好。ROC 曲線上描繪同一模型在給定不同門檻值 (threshold) 下得出的真陽性率與偽陽性率；此門檻值是模型用以區分客戶是否違約的機率值。若設定門檻值為 0，則在 ROC 曲線上的點為 (0,0) ，表示模型不加區分地將所有樣本判別為正常；若設定門檻值為 1，則在 ROC 曲線上的點為 (1,1) ，表示模型不加區分地將所有樣本判別為違約。理論上最完美的點為 (0,1) ，代表正確識別出所有違約客戶並且避免將正常客戶歸類為違約的錯誤，所以，越凸向左上角的 ROC 曲線代表模型的區別能力越強。

然而，從 ROC 曲線並不能方便地比較各分類模型的區別能力，因此需採用其曲線下的面積值 (AUC 值)。其定義為 ROC 曲線與橫軸之間的面積，一般範圍介於 0.5 至 1 之間，此值越大代表模型越靠近左上方，代表模型的區分能力與配適度越好。Iyer et al., (2016) 表示，依據經驗法則，在信息匱乏(豐富)的情境下 AUC 值在 0.6 (0.7) 之上的模型可令人滿意；即使只提升 0.01 的 AUC 值，在信用評分行業也被認為是顯著的增益。

四、模型穩健性檢驗

本研究參考 Xia et al. (2018)，以及該文所引用之 Lessmann et al. (2015)，利用多重資料集 (multiple data sets)，並搭配 Friedman 檢定檢驗模型是否具有顯著差異，並進行模型兩兩比較。然而，本研究對象為台灣信用卡用戶，資料集來源本身之取得即相當有限，因此本研究穩健性檢驗時將原始樣本集隨機分為 4 等份，進行模型績效評估表現檢驗與模型成對比較 (pair-wise test)。

首先，在績效評估表現檢驗分為單一指標檢定與綜合指標檢定，其基本架構都採用 Friedman 無母數檢驗方法。本研究採用準確率、型二誤差、F1 值以及 AUC 等四個分類績效評比指標。首先，我們要將分類績效指標轉換為排序，排序越低

者表現越佳。由於本研究採用之分類器共有九種，含五種基礎分類器與四種集成學習模型，故排序值介於 1 到 9 之間。據此，我們可以取得各模型在不同子資料集的結果。

本研究參考 Demšar (2006) 利用 Friedman 檢定 (Friedman, 1937, 1940) 檢驗九種分類模型 (K=9) 在單一指標與綜合指標是否呈現顯著差異，其檢定統計量如下式 (5):

$$\chi_F^2 = \frac{12D}{K(K+1)}\left[\sum_j R_j^2 - \frac{K(K+1)}{4}\right] \quad (5)$$

其中，$R_j = \frac{1}{D}\sum_j r_j^i$，而$r_j^i$代表第 i 個模型應用於第 j 個資料集的評估指標排序值。本研究資料集共有四組，故 D=4。由於本研究資料集個數與分類模型個數相乘為 4*9 = 36，在大於 30 的情況下，其檢定統計量之近似分配服從 $\chi^2_{df=K-1}$ (Lehmann and D'Abrera, 1973, p265)。

另外在模型預測表現成對比較方面，本研究參考 Demšar (2006) 成對比較各個模型間四種衡量指標是否存在差異，此方法也被 Lessmann et al. (2015) 採用。首先，研究者針對單一衡量指標，先取出欲比較之兩組模型在不同資料集的表現，並計算該指標的差異 (d_i) ，如下式 (6) 所示：

$$d_i = I_{Bi} - I_{Ai} \quad (6)$$

其中，I 代表衡量指標值，下標 i 為某資料集所得之模型評估結果。下標 B 代表基準模型，本研究依照各個指標表現，分別以集成學習模型中表現最佳與最差的作為基準模型，以簡化分析；而下標 A 則代表所欲比較之模型。計算完兩模型指標差異後，再利用差異絕對值進行遞增排序，亦即若 $d_i = 0$ 代表兩模型在該指標的表現一樣，排序值 $rank(d_i)$ 為 1，以此類推。由於差距可能為正或為負，因此我們依照下式 (7) 與式 (8 分別計算正排序和與負排序和。

$$R^{+} = \sum_{d_i>0} rank(d_i) + 0.5 \sum_{d_i=0} rank(d_i) \quad (7)$$

$$R^{-} = \sum_{d_i<0} rank(d_i) + 0.5 \sum_{d_i=0} rank(d_i) \quad (8)$$

依照式 (7) 與式 (8)，計算 Wilcoxon 符號等級檢定統計量 (Wilcoxon signed-ranks test)，如下式 (9)：

$$z = \frac{T - 0.25D(D+1)}{\sqrt{\frac{D(D+1)(2D+1)}{24}}} \quad (9)$$

檢定統計量可利用查表，當樣本數為 4 時，在顯著水準 5%下之雙尾檢定臨界值為 0。[3]因此檢定統計量低於 0 時，拒絕兩模型表現沒有顯著差異之虛無假設。

[3] 其臨界值表可參考 Lyles and Hummer (2012) 或 Mason et al. (1999) 等教科書。

肆、研究資料與處理

一、原始資料來源與變數之定義

由於信用資料難以取得，信用評分的集成學習傾向於使用一些公開的數據庫，例如 UCI 機器學習資料庫 (Xia et al., 2018)。故本研究的資料集取自 Yeh and Lien (2009)，可直接由 UCI 機器學習資料庫取得。該數據描述台灣某大型銀行信用卡持有人資料，資料期間為 2005 年 4 月至 2005 年 9 月，共有 30,000 筆客戶資料與 24 個變數，其中反應變數 *DEFAULT_PAY* 為一虛擬變數，代表客戶在 2005 年 10 月是否會逾期繳款 (違約) ，而其餘 23 個解釋變數包含客戶基本資料、信用卡帳單金額與過去還款紀錄，變數型態共計有 3 個離散型變數與 20 個連續型變數，資料詳細的變數定義如表 3 所示。

表 3 原始資料變數定義表

變數名稱	變數解釋
DEFAULT_PAY	於 2005 年 10 月是否違約：1=違約[4]，0=未違約
LIMIT_BAL	銀行給予的信用額度 (包含個人與家庭的信用額度)
GENDER	性別：1=男性，2=女性
EDUCATION	教育程度：1=研究所以上，2 =大學，3=高中，4=其他
MARRIAGE	婚姻狀況：1=已婚，2=未婚，3=其他
AGE	年齡
PAY_*n*	第前 *n* 個月還款狀況，*n* = 1,…,6。
BILL_AMT*n*	第前 *n* 個月前信用卡帳單金額(信用卡消費金額)，*n* = 1,…,6。
PAY_AMT*n*	第前 *n* 個月支付信用卡的金額，*n* = 1,…,6。

註：1. 原資料集之還款狀況 (PAY_n) 變數的 0=及時付款，1=逾期繳款一個月，2=逾期繳款兩個月，依此類推。此外，原資料集之還款狀況 (PAY_n) 變數亦包含-2,-1,0 三個數值，分別代表的意義為無消費、全額付款與使用循環信貸，意即僅繳交最低應繳金額，而本研究將此 3 種還款狀況皆歸類於數值等於 0，表示未逾期繳款。2. 還款狀況(PAY_n)、信用卡帳單金額(BILL_AMTn)與支付信用卡金額(PAY_AMTn)的第前 *n* 月是以 2009 年 10 月為基準，例如第前 1 個月為 2009 年 9 月還款狀況，以 PAY_1 變數表示。

依據變數 *DEFAULT_PAY* 我們將本資料中的 30000 筆客戶分為正常客戶與違約客戶。經統計後，正常客戶共有 23,364 筆，違約客戶共有 6,636 筆，違約

[4] 本研究違約的定義為當期支付金額低於銀行所規定的最低應繳金額。

客戶佔整體客戶僅約 22%，顯示此份資料的類別呈現類別不平衡問題。然而類別不平衡的資料會使學習器傾向將所有資料判斷為正常客戶，導致對違約客戶的預測結果表現較差，因此我們會使用 SMOTE 過取樣方法來解決此問題，詳細過程將於後文說明。

二、資料處理與特徵工程

(一) 特徵離散化與標準化

對特徵進行離散化處理，會使特徵變化不明顯，降低資料離群值的干擾，通常能夠提高模型的穩定性。首先我們對變數客戶年齡 (*AGE*) 進行離散化，先將客戶年齡劃分 5 個區間：21 歲 (此資料客戶年齡最小值為 21 歲) 至 29 歲、30 歲至 39 歲、40 歲至 49 歲、50 歲至 59 歲，60 歲至 79 歲 (客戶年齡最大值為 79 歲) ，然後再將此五個區間分別指定 1 到 5 的數值，完成變數的離散化處理。

接著我們將還款狀況 *PAY_1 - PAY_6* 這些代表不同月份客戶逾期繳款月數的變數加總，整合成一個新變數 *PAY_RISK*，再將變數劃分為 7 個區間，分別為 0 至 1 個月、2 至 6 個月、7 至 12 個月、13 至 18 個月、19 至 24 個月、25 至 30 個月、31 個月以上，最後指定每個區間為 0 至 6 的數值，共分成 7 組。每一組代表的意義為過去每月平均逾期繳款多少個月，比如當 *PAY_RISK* 等於 0 表示客戶於過去信用狀況良好，幾乎沒有違約，而當 *PAY_RISK* 等於 3 表示客戶過去平均每月遲繳 2 個月以上至多 3 個月。

許多模型背後的演算法是以歐氏距離進行分類，如羅吉斯迴歸、支援向量機或其他使用損失函數 (loss function) 作為目標函數的演算法，若每個特徵的取值範圍差異太大，可能會導致模型傾向於擬合單位尺度 (scale) 較大的特徵。因此為了得到更準確的分類結果，在建構模型前，我們對資料進行特徵縮放 (feature scaling)，除了能提高模型準確性，也能降低離群值的影響，加快模型收斂速度。

本研究採用標準化 (standardization) 對所有連續型變數進行特徵縮放，使得縮放後的每個特徵平均值變為 0，標準差變為 1，其定義如式 (7) 所示：

$$\mathcal{Z}_i = \frac{x_i - \bar{x}}{S_x} \text{ ，} where\ S_x = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}, \tag{7}$$

其中Z_i為特徵x_i進行特徵縮放後之值，$\bar{x}$為某個特徵在樣本空間下之平均值，S_x為某個特定特徵在樣本空間下之標準差。

(二) 變數相關性與主成份分析

接著我們對資料中的連續型變數進行相關性分析，以觀察變數間是否存在多重共線性問題。[5] 表 4 為 13 個連續變數的相關係數矩陣，數字為相關係數的絕對值大小。顯然，*BILL_AMT1 - BILL_AMT6* 這 6 個代表客戶過去信用卡帳單金額的變數彼此具有高度相關性，表示這 6 個變數之間存在較多類似的資訊。

表 4　連續型變數的相關係數矩陣

	LIMIT_BAL	BILL_AMT1	BILL_AMT2	BILL_AMT3	BILL_AMT4	BILL_AMT5	BILL_AMT6	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6
LIMIT_BAL	1												
BILL_AMT1	0.29	1											
BILL_AMT2	0.28	0.95	1										
BILL_AMT3	0.28	0.89	0.93	1									
BILL_AMT4	0.29	0.86	0.89	0.92	1								
BILL_AMT5	0.30	0.83	0.86	0.88	0.95	1							
BILL_AMT6	0.29	0.80	0.83	0.85	0.90	0.95	1						
PAY_AMT1	0.20	0.14	0.28	0.24	0.23	0.22	0.20	1					
PAY_AMT2	0.18	0.01	0.10	0.32	0.21	0.18	0.17	0.29	1				
PAY_AMT3	0.21	0.16	0.15	0.13	0.30	0.25	0.23	0.25	0.24	1			
PAY_AMT4	0.20	0.16	0.15	0.14	0.13	0.29	0.25	0.20	0.18	0.22	1		
PAY_AMT5	0.22	0.17	0.16	0.18	0.16	0.14	0.31	0.15	0.18	0.16	0.15	1	
PAY_AMT6	0.22	0.18	0.17	0.18	0.18	0.16	0.12	0.19	0.16	0.16	0.16	0.15	1

為了解決共線性問題與有效地降低變數的數量，我們採用主成份分析 (principal components analysis，PCA) 對這 6 個變數進行降維處理，以降低模型複雜度。主成份分析之目的為以最少的變數代表原始資料最大的資訊，其方法為

[5] 如同一般線性迴歸模型，羅吉斯迴歸也需避免解釋變數之間共線性的問題。雖然多重共線性問題對大部分機器學習模型的影響較小，如決策樹、隨機森林等，但資料特徵的解釋性會受影響。

將原變數作線性組合，成為另一組新變數—*BILL_AMT*，組合原則為保留原變數間最大的變異，且新變數彼此不相關，具獨立性。

主成份分析結果如表 5 所示，第一主成份的變異量百分比約為 90.5%，代表其貢獻約百分之九十的原始資料訊息，能涵蓋大部分原變數間的變異，因此我們以第一主成份取代 *BILL_AMT1 - BILL_AMT6* 此 6 個變數。

表 5　BILL_AMT 變數之主成份分析結果

主成分	PC1	PC2	PC3	PC4	PC5	PC6
變異量百分比	90.56%	5.099%	1.861%	1.119%	0.693%	0.673%
累積貢獻比率	90.56%	95.66%	97.52%	98.64%	99.33%	100%

三、資料集切割

我們將資料以 9 比 1 的比例隨機劃分為訓練集 (train set) 與測試集 (test set)，其中訓練集包含 27,000 筆資料，測試集包含 3,000 筆資料，且兩個資料集均有 1 個反應變數與 13 個解釋變數。其中，兩資料集中違約客戶占比相當，訓練集中有 5,977 筆為違約客戶、約占訓練集樣本數 22.14%，測試集中則有 659 筆為違約客戶、約占測試集 22.00%。本研究中訓練集資料用來建立模型與確定模型超參數；而測試集資料用來評估模型與預測數據，我們能夠透過模型的測試結果來檢視模型在測試資料上的配適情況。

值得注意的是，在建構模型勻合集成模型時，我們還會再將訓練集依 6 比 4 的比例劃分為新訓練集(包含 16,200 筆資料)與驗證集(包含 10,800 筆資料)。而此驗證集並非用來檢驗模型與優化調整模型參數，我們會將此驗證集放入已訓練完成的第一層學習器做預測，再將所得到的預測結果作為第二層學習器的新輸入特徵，以訓練元模型。

四、類別不平衡問題處理

在前述原始資料分析中提及本研究資料存在著類別不平衡問題，而以往解決此問題常用的做法有 (1) 過取樣法 (oversampling) ，對少數類樣本進行隨機抽樣，再將抽樣得來的樣本新增至資料集中，或 (2) 下取樣法 (undersampling) ，隨機去除一些多數類樣本，以減少多數類規模，達到樣本分類均衡。然而，He and

Garcia (2009) 認為前者最大的風險為若資料特徵少，易導致模型過擬合 (overfitting) 問題，後者最大缺點為丟失多數類樣本中的一些重要資訊。因此本研究採取 He and Garcia (2009) 提出的 SMOTE 技術，以減緩樣本數據呈現類別不平衡的問題。SMOTE 為一種基於改進隨機過取樣法的算法，具有不損失原資料價值的優點，其概念為對少數類樣本間進行插值，以合成新樣本，再將新樣本新增至資料集中，詳細步驟如下：

步驟 1 計算樣本間的歐式距離

找出每個少數類樣本點x_i的 K-近鄰樣本 (K-nearest neighbors) ，而 K-近鄰樣本為與x_i距離最近的 K 個少數類樣本，本研究 K 值取為 5。其距離之定義為樣本點x_i與其他少數類樣本之間的 n 維空間歐式距離，計算公式如下：

$$\text{distance}(x_i, x_j) := \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{in} - x_{jn})^2}, \qquad (8)$$

步驟 2 找出鄰近樣本

從少數類樣本點x_i的 K-近鄰樣本中隨機選取一個樣本點$\hat{x}_i$，再根據以下公式合成新樣本：

$$x_{new} = x_i + (\hat{x}_i - x_i) \times \delta, \qquad (9)$$

其中$\hat{x}_i$屬於樣本點x_i的 K-近鄰樣本，$(\hat{x}_i - x_i)$ 為兩點間的歐式距離，δ為介於 0 與 1 之間的隨機變數。

步驟 3 重複前述步驟

對每個少數類樣本點重複上述過程，直到最後少數類樣本的數量增加至 17,000 筆。

五、最終資料集敘述統計量

本研究由 UCI 機器學習資料庫取得我國某家大型銀行業 30,000 筆信用卡客戶資料集後，經過上述的變數離散化、連續變數取標準化、萃取主成分等資料處理後，先以 9:1 將資料集切割為訓練資料集與測試資料集，再將訓練資料集依 6:4 的比例劃分為新訓練資料集與驗證資料集。此外，由於違約事件通常相對稀少，

故我們會將訓練集中的違約客戶筆數由 5,977 筆增加至 17,000 筆。表 6 為本研究集成學習所使用的樣本與其敘述統計量。其中，經 SMOTE 調整的訓練集樣本有 21,023 筆正常戶加上 17,000 筆違約客戶，共 38,023 筆資料，訓練集違約客戶占比也由 22.14% 增至 44.17%。

此外，本研究使用變數大多與原始資料集提供的相同，惟年齡變數經離散化處理，並自行計算 *PAY_RISK* (平均逾期繳款月份) 與利用主成分分析萃取 *BILL_AMT1 - BILL_AMT6* 等六個變數得到的第一主成分—*BILL_AMT*，代表繳款金額。值得注意的是，經 SMOTE 調整後，由於擴增了違約資料集的筆數，因此 SMOTE 調整前與調整後的平均數具明顯差異，如 *PAY_AMT1 - PAY_AMT6* 與 *BILL_AMT* 等變數在 SMOTE 調整前的平均數皆為 0.001 層級的正值，但經 SMOTE 調整後全變為 0.01 層級的負數；總體而言，違約樣本的賬單金額與繳款金額皆小於正常樣本，調整後特徵值由正轉負也佐證違約樣本的擴增。

表 6　訓練資料集與測試資料集敘述統計量

		訓練資料集 (SMOTE 調整前)		訓練資料集 (SMOTE 調整後)		測試資料集	
變數	變數定義	平均數	標準差	平均數	標準差	平均數	標準差
DEFAULT_PAY	違約與否	0.2214	-	0.4471	-	0.2197	-
LIMIT_BAL	信用額度	0.0018	1.0014	-0.0882	0.9710	-0.0164	0.9872
SEX	性別	1.6032	-	1.5951	-	1.6083	-
EDUCATION	教育程度	1.8406	-	1.8561	-	1.8577	-
MARRIAGE	婚姻狀況	1.5584	-	1.5488	-	1.5470	-
PAY_AMT1	第前 1 個月支付信用卡金額	0.0020	1.0168	-0.0398	0.9139	-0.0176	0.8342
PAY_AMT2	第前 2 個月支付信用卡金額	0.0028	1.0321	-0.0340	0.8949	-0.0255	0.6430
PAY_AMT3	第前 3 個月支付信用卡金額	0.0022	1.0224	-0.0284	0.9544	-0.0196	0.7697
PAY_AMT4	第前 4 個月支付信用卡金額	0.0026	1.0227	-0.0301	0.9368	-0.0234	0.7659
PAY_AMT5	第前 5 個月支付信用卡金額	0.0021	1.0145	-0.0291	0.9547	-0.0190	0.8590
PAY_AMT6	第前 6 個月支付信用卡金額	0.0000	0.9946	-0.0279	0.9490	-0.0002	1.0473

AGE[a]	年齡	2.0849	0.9699	2.0884	0.9727	2.0847	0.9800
PAY_RISK[ab]	平均逾期繳款月份	0.3997	0.7528	0.5544	0.8652	0.3987	0.7589
BILL_AMT[b]	繳款金額(第一主成分)	0.0042	2.3398	-0.0092	2.3313	-0.0378	2.2503
樣本數		27000		38023		3000	

註：1.本表上標 a 代表該變數經過標準化處理、b 代表該變數經過離散化處理，上標 c 代表本研究由原始資料欄位衍生、自行計算變數。2.離散變數不計算標準差，以「-」表示。

伍、實證結果

本章為單一模型與集成模型對測試資料集進行預測的結果，共分為三個小節，第一節說明各個模型的重要超參數設定與最佳超參數值；第二節說明羅吉斯迴歸、支援向量機、類神經網路、隨機森林與 XGBoost 五種單一模型的預測結果，並以型二誤差、F_1 值以及 AUC 值三種指標比較各個模型的預測能力；第三節則探討使用堆疊與勻合集成框架所建構的模型之成效是否優於單一模型，並以單一模型最佳表現作為基準進行比較。

一、模型參數配置

本研究使用 Python 的機器學習套件 Scikit-learn (簡稱 sklearn) 、xgboost、keras 與 vecstack 建立模型，並透過網格搜尋法 (grid search) 與 5 折交叉驗證 (5-Fold Cross Validation) 來設定單一演算法的超參數。鑒於本研究目的旨在結合多種單一機器學習模型後，探討所得到的集成模型表現是否更好，而不是訓練出一個預測能力極佳的模型，因此我們不過於複雜化模型，只針對每個模型的重要超參數進行調整，其他未調整的參數皆為預設值。[6] 每個模型超參數設定請參考附錄。

此外，由於本研究資料存在類別不平衡問題，因此我們使用堆疊與勻合集成，對於元模型—羅吉斯迴歸的訓練則有進行超參數 class_weight 的設定。當 class_weight 設定為「balanced」時，模型會根據訓練集中各類別的樣本數來計算權重，原本樣本數多的類別會被配置較低的權重，而樣本數少的類別所配到的權重較高，能得到模型更大的關注；相比不考慮權重，模型在訓練過程中會將更多的樣本劃分到權重高的類別，能夠減緩類別不平衡的問題。

二、單一模型預測表現

本節檢視單一模型之表現目的有二，其一是比較單一模型在預測集的表現，挑選不同衡量指標下最佳單一模型作為後續集成模型比較之基準；第二，我們以

[6] Blending 與 Stacking 集成模型的第一層學習器中各模型的參數設定皆與原單一模型相同。

隨機森林以及 XGBoost 為例，檢視模型所提取之有效特徵及其重要程度排序，由此觀察單一模型之間的相關程度。

隨機森林與 XGBoost 模型中，影響客戶信用風險的前 8 個重要特徵皆包含 *PAY_RISK* (還款狀況)、*LIMIT_BAL* (銀行給予客戶的信用額度) 與 *PAY_AMT1* - *PAY_AMT4*、*PAY_AMT6* (分別對應 2005 年 9 月至 6 月及 4 月支付信用卡的金額) ；重要程度最低的 5 個特徵皆為 *PAY_AMT5*、*EDUCATION*、*MARRAGE*、*SEX* 與 *AGE*。其中變數 *PAY_RISK* 為本研究自行合併生成的新變數，代表客戶過去每月平均逾期繳款多少個月，能夠評估客戶的還款狀況。以資訊數字化的難易程度與資訊解讀對具體情境的依賴程度為標準，Liberti et al. (2019) 將財經資訊區分為硬資訊 (Hard Information) 與軟資訊 (Soft Information) 。硬資訊以數字形式體現，軟資訊則是以文本形式呈現；將文本訊息數字化的過程也被稱為訊息硬化 (harding of information) ,這一處理過程不可避免地導致資訊丟失 (loss of information) 。銀行給予的信用額度、客戶的還款狀況與支付金額這前 8 個特徵都是硬訊息，排序結果與文獻上 (例如 Thomas, 2000; Iyer et al., 2016) 信用額度、還款記錄等為重要特徵的結論一致。而 *EDUCATION*、*MARRAGE*、*SEX* 與 *AGE* 的信息在轉化的過程中脫離了原有情境，損失了部分資訊。前特徵集 (8 個) 與後特徵集 (5 個) 的一致的區分，說明模型都良好的捕捉了特徵，但也說明模型的相關程度高；特徵集排序上的差異則體現單一算法間的差異性。

本研究採用機率的平均值 0.5 作為各分類模型的最佳門檻值，若模型輸出機率大於 0.5，此資料將被歸類為違約客戶，若輸出機率小於 0.5，則會被歸類為正常客戶。表 7 的實證結果顯示，在根據各個單一模型預測表現所計算出的型二誤差之比較下，類神經網路模型的誤差值最低，隨機森林模型次之，表示類神經網路模型在區分違約客戶資料時，其錯誤分類成本最低。就 F_1 值而言，隨機森林模型的 F_1 值最高，支援向量機次之，表示若將模型精確度與型二誤差兩指標取平衡來看，隨機森林模型的綜合表現最好。在比較各個單一模型的 AUC 值時，類神經網路模型的 AUC 值最高，隨機森林模型次之，代表類神經網路模型不論門檻值的設定為何，其整體分類效果與配適度最好。結合上述指標的綜合比較下，類神經網路模型與隨機森林為表現最好的兩個模型。

除此之外，本研究亦計算各個單一模型的訓練時間，置於表 7 最右方。我們能夠發現羅吉斯迴歸模型的訓練速度最快，訓練時間不到 1 秒鐘；XGBoost 與隨機森林模型的訓練速度次之，均小於 5 秒鐘；類神經網路模型的訓練時間又比前兩個模型慢一些，約 46 秒鐘，而支援向量機模型的訓練速度最久，約 4 分 35 秒，若訓練資料量過於龐大，支援向量機模型的訓練過程與其它模型相比則較無效率。

表 7　單一模型在測試資料集上之表現

模型	準確率	型二誤差	F_1 值	AUC	訓練時間
羅吉斯迴歸	75.73%	42.34%	0.51075	0.7410	**0.18 秒****
支援向量機	**76.27%****	43.10%	**0.51230***	0.7433	275.00 秒
類神經網路	74.07%	**39.45%****	0.50634	**0.7645****	46.70 秒
隨機森林	75.67%	**41.27%***	**0.51463****	**0.7636***	4.97 秒
XGBoost	**76.10%***	45.98%	0.50425	0.7598	**3.08 秒***

註：**代表模型在該指標衡量下的表現最好、*則為次佳。

三、集成模型預測表現之探討

本研究使用堆疊與勻合的集成學習框架建立模型，集成模型中的第一層學習器會使用擬合度高的模型，如 XGBoost、類神經網路與支援向量機等，以充分地對訓練資料進行學習。第一層中的所有學習器會使用不同的演算法進行特徵提取，並從不同的角度訓練模型，得到具有差異性的輸出結果。由於集成模型中的第一層學習器大多使用複雜且非線性的方法提取有效特徵，容易產生過擬合問題，因此為了降低過度擬合的風險，第二層學習器大多使用較簡單的模型，如羅吉斯迴歸、Lasso 迴歸等廣義線性模型。而本研究使用羅吉斯迴歸作為元模型，並配合 L2 正規化方法。

(一) 集成模型的學習器挑選策略

集成模型的預測效果優於單一模型的關鍵在於第一層學習器的選擇。參考 Windeatt and Ardeshir (2004)，本研究集成學習挑選第一層模型的原則為 (1) 各個模型的預測表現不能太差、(2) 各個模型間最好具足夠的差異性。如表 7 所示，

各單一模型準確率皆高於 50%且在 70%以上，AUC 值也高於 0.7 (資訊豐富情境下可接受水平)，表明第一層各個模型表現良好。原始樣本的解釋變數雖然多達 23 個，但是變數所提供資訊集中在客戶基本資料、銀行授信金額、信用卡賬單、支付與還款紀錄，而不包含數位足跡 (digital footprint) (Berg et al., 2020)，友誼與社交網絡 (Hildebrandt et al., 2017) 以及文本資料 (Gao et al., 2018; Netzer et al., 2019) 等具豐富資訊的變數。在前述的單一模型實證結果中，模型所捕捉的重要特徵大致相同。因此，樣本集本身的特性決定了表現良好的模型之間差異度有限。由於本研究所建構的堆疊與勻合模型會使用羅吉斯迴歸模型作為第二層的元模型，因此在挑選第一層學習器時，我們不納入羅吉斯迴歸模型。

我們對 4 個單一模型：支援向量機、類神經網路、隨機森林與 XGBoost 的預測結果進行相關性分析，籍此觀察訓練過後模型的相關性與差異性。Kuncheva and Whitaker (2003) 就學習器的差異性衡量指標進行探討，作者表示指標不存在被公認的正式定義。然而，作者列舉的二分類問題的差異性衡量指標都是基於學習器的預測結果，因此本研究也以學習器預測結果衡量其相關性並作為差異性判斷之依據。表 8 為單一模型輸出值的相關係數矩陣，數字為相關係數的絕對值大小。如表所示，模型輸出結果之相關係數介於 0.79 至 0.91 之間，顯示彼此具有高度相關性，亦代表本研究所訓練的單一模型之間差異較低，其中支援向量機與隨機森林模型的相關性最高。由於隨機森林在不同衡量指標皆優於支援向量機，我們在後續實證研究中比較有無支援向量機的集成學習之表現。

表 8　單一模型預測結果之相關係數矩陣

	支援向量機	類神經網路	隨機森林	XGBoost
支援向量機	1			
類神經網路	0.89	1		
隨機森林	0.91	0.85	1	
XGBoost	0.82	0.79	0.89	1

在選擇第一層學習器時，本研究的理想集成模型將去除支援向量機模型，意即第一層學習器只包含類神經網路、隨機森林與 XGBoost 模型的堆疊與勻合集

成模型。此外，為了檢驗集成模型學習器挑選策略的有效性，也為了全面地驗證集成模型的預測表現是否一定優於單一模型，我們亦建構了全學習器的堆疊與勻合集成模型，即在初級學習器中加入支援向量機。

在模型訓練完成後，我們檢視全學習器的模型堆疊集成所捕捉之重要特徵，並與輔以學習挑選策略的模型堆疊集成進行比較。結果呈現，在全學習器的模型堆疊集的特徵重要程度排序中，支援向量機所生成特徵之重要性最低；而其它 3 個由學習器所生成特徵之重要程度排序由高到低為 XGBoost、隨機森林以及類神經網路，此排序與輔以學習挑選策略的模型堆疊集成相同。此結果亦支持理想的模型堆疊集成學習框架中的第一層學習器應剔除支援向量機模型。

(二) 集成模型預測表現

表 9 為集成模型在測試資料集上的預測表現之實證結果。其中，堆疊與勻合集成模型的第二層學習器 (羅吉斯迴歸模型) 採用 0.5 作為最佳門檻值。經 SMOTE 調整的訓練集樣本含 21,023 筆正常戶與 17,000 筆違約客戶，違約客戶占比為 44.17%。類別不平衡問題有所改善，但傾向於將客戶歸類為正常的模型仍然會具較高準確率。實務中，將違約客戶歸類為正常的錯誤造成嚴重呆賬損失。因此，即使經過 SMOTE 調整，準確率仍非良好的衡量指標。

表 9　集成學習在測試資料集上之表現

模型	準確率	型二誤差	F_1 值	AUC 值	訓練時間
最佳表現之單一模型	支援向量機	類神經網路	隨機森林	類神經網路	羅吉斯迴歸
單一模型最佳數據	**76.27%****	39.45%	0.51463	**0.7645***	**0.18 秒****
Stacking (全學習器)	74.03%	38.09%	0.51160	0.7625	971 秒
Blending (全學習器)	**74.13%***	39.15%	0.50824	0.7630	365 秒
Stacking (剔除 SVM)	73.70%	**35.81%****	**0.51743****	**0.7650****	235 秒
Blending (剔除 SVM)	73.70%	**36.27%***	**0.51565***	0.7634	**80 秒***

註：**代表模型在該指標衡量下的表現最好、*則為次佳。

以準確率而言，單一模型的支援向量機表現最佳 (76.27%) ，次佳為剔除支援向量機的模型勻合 (74.13%)。Yeh and Lien (2009) 在沒有調整數據不平衡 (違約客戶占比為 22.12%) 的情境下，比較六種單一模型在測試集的準確率，類神經網路 (83%) 優於羅吉斯迴歸 (82%)。而本研究在違約客戶占比為 44.17%的情境中，類神經網路 (74.07%) 差於羅吉斯迴歸 (75.73%)。雖然本研究與 Yeh and Lien (2009) 在資料處理上方式不同，但實證結果顯現，以準確率衡量的模型表現受到違約客戶分布的影響，與文獻上的準確率不適配類別不平衡問題的認知一致。

在將違約客戶錯判為正常的錯誤 (型二誤差) 方面，表現最佳模型為輔以學習器挑選策略的模型堆疊 (35.81%)，次佳模型為結合學習器挑選策略的模型勻合 (36.27%)。實證結果呈現，結合學習器挑選策略的集成學習優於不實施挑選策略的集成學習 (型二誤差分別為 38.09%與 39.15%)。比較集成模型設計之表現，則發現同等情境下模型堆疊皆勝過模型勻合。進一步比較集成模型與單一模型，發現集成模型的表現皆優於最佳單一模型 (39.45%)，表明採用堆疊與勻合集成皆能有效降低將違約客戶歸類為正常的錯誤。

在衡量模型準確識別違約客戶的能力 (F_1 值) 方面，表現最佳模型為輔以學習器挑選策略的模型堆疊 (0.51743)，次佳模型為結合學習器挑選策略的模型勻合 (0.51565)。實證顯示，結合學習器挑選策略的集成學習皆優於最佳單一模型 (0.51463)，不實施挑選策略的集成學習 (F_1 值分別為 0.51160 與 0.50824) 皆差於最佳單一模型。結果表明，只有結合學習器挑選策略的集成學習方能提升模型準確識別違約客戶的能力。比較集成模型框架之表現，亦發現同等情境下模型堆疊皆勝過模型勻合。

就模型正確識別違約客戶且避免將正常客戶分類為違約客戶的能力 (AUC 值) 而言，最佳模型為輔以學習器挑選策略的模型堆疊 (0.7650)，次佳模型為單一模型的類神經網路 (0.7645)。AUC 值在 0.7 之上的模型可令人滿意 (Iyer et al., 2016)，而本研究的單一模型與集成模型之 AUC 值皆在此標準之上。實證顯示，具學習器篩選策略的集成模型 (AUC 值分別為 0.7650 與 0.7634) 勝過不具學習器篩選策略的集成模型 (AUC 值分別為 0.7625 與 0.7630)。比較集成模型結構之表現，則發現模型堆疊無法一致地勝過模型勻合。

另外，本研究亦計算集成模型的訓練時間,為兼顧模型表現與訓練成本之考量提供決策依據。同等情境下，模型堆疊的訓練時間皆長於模型勻合，這主要因為模型堆疊在訓練單一模型的過程中採用 K-折交叉驗證的技巧，對於每個模型皆訓練 5 次。集成模型中剔除支援向量機的模型勻合的訓練成本最低，其訓練時間為 80 秒。有趣的是，單一模型中的支援向量機訓練成本最高，耗時 275 秒，遠超剔除支援向量機的集成模型所需訓練時間 (分別為 235 秒與 80 秒)。

最後，比較集成模型有無搭配學習器挑選策略下，實證結果呈現具備挑選策略的集成學習在型二誤差、F_1 值與 AUC 值三個衡量指標下皆優於對應的不加挑選的集成學習。推測其原因，支援向量機與隨機森林的模型預測結果高度相關，同時採用違反了學習器篩選準則的第二條--學習器之間需要具足夠的差異性。比較集成學習框架之表現，則大多數情境下模型堆疊皆優於模型勻合，反映出模型勻合對訓練資料的利用程度不如模型堆疊，從樣本中萃取的資訊也有所缺失所致。但要注意的是，輔以學習器挑選策略的模型勻合在集成學習中仍為次佳模型，且相較於最佳單一模型，模型勻合則在型二誤差與 F_1 值方面皆有更優異表現。綜合上述分析，可知勻合集成在有效減少訓練時間下，亦可使其模型表現接近於堆疊集成。

(三) 穩健性檢定結果

前述內容主要係針對全體樣本估計算得之預測表現結果。然而，此模型彼此間之預測表現是否具有顯著，需要透過進一步檢驗，請參考研究方法第三小節。

首先，針對各個模型在單一指標排序與綜合指標排序平均值與 Friedman 檢定統計量彙整於下表 10。可發現除了由 4 個指標平均排序建構的 avR1 指標未統計顯著外，其餘的單一指標或 avR2 指標檢定結果皆顯著地拒絕模型間預測表現沒有差異的虛無假設。若以型二誤差值為例，該指標用來衡量模型將違約客戶錯判為正常類別的程度，利用剔除 SVM 的堆疊模型在四組資料集的平均排序為 1.25、是 9 種演算法中平均表現最佳者，其次為剔除 SVM 的勻合模型、平均排序為 2.25，此現象與表 9 利用所有資料取得之結果一致。且型二誤差所得之 Friedman 檢定統計量為 26.7，其 P-value 為 0.0008，顯著拒絕 9 種模型所得之型二誤差皆相同的虛無假設。

此外，由於本研究模型績效衡量指標有 4 種，可進一步參考 Lessmann et al. (2015) 與 Xia et al. (2018) 方法，利用各個指標平均排序之平均值 (average of average ranks)，以做整體模型的表現的評估，計算結果可參考表 10 之 avR 欄位，綜合排序指標同樣採用式 (4) 計算 Friedman 檢定統計量，僅須將R_j 改為avR_j 。由表 10 可知，Friedman 檢定結果確實顯現集成學習在準確率的表現較差，但在檢定力、F1 值與 AUC 的表現較優。

職是之故，本研究 avR 值依照是否包含準確率與否，進而分為兩類。表 10 顯示包含準確率的 avR1 指標，其 Friedman 檢定結果顯示各模型間不具顯著差異，但不含準確率的 avR2 指標結果則呈現顯著差異。因此，本研究進行後續兩兩模型之比較分析時，著重在型二誤差、F_1 值與 AUC 等指標的成對比較。

表 10　模型績效指標檢定

	準確率	型二誤差	F_1 值	AUC	avR1	avR1 排序	avR2	avR2 排序
羅吉斯迴歸	**2.75****	7.00	5.75	8.50	6.00	7	7.08	8
支援向量機	2.88	8.25	7.25	6.75	6.28	8	7.42	9
類神經網路	6.00	5.75	7.50	7.50	6.69	9	6.92	7
隨機森林	**2.63***	6.38	4.25	5.00	4.56	5	5.21	5
XGBoost	4.00	7.38	7.50	4.25	5.78	6	6.38	6
Stacking (全學習器)	7.88	**2.25****	**3.13****	**2.63***	**3.97****	**2**	**2.67****	**2**
Blending (全學習器)	5.75	3.88	4.25	**3.13****	4.25	4	3.75	4
Stacking (剔除 SVM)	7.13	**1.25***	**1.38***	3.50	**3.31***	**1**	**2.04***	**1**
Blending (剔除 SVM)	6.00	2.88	4.00	3.75	4.16	3	3.54	3
Friedman 檢定統計量	16.8333	26.7000	19.6833	18.7167	6.1208		17.9593	
P-value	0.0319	0.0008	0.0116	0.0165	0.6337		0.0215	

註： avR1 含準確率; avR2 不含準確率。

在成對比較方面，根據 Wilcoxon 符號等級檢定統計量計算結果如下表 11，此數據越小代表兩模型間的差距越顯著。由該表結果可知，無論是集成學習表現最佳或最差者，在型二誤差與 F_1 值成對比較，幾乎全數與單一模型呈現顯著差

異，代表集成學習確實能夠提高個別模型判斷違約客戶的能力，此模型預測表現對於銀行經營者與政府監管單位甚為重要。而在 AUC 指標部分，可以發現即便是 AUC 表現最差的集成學習-- Blending (剔除 SVM)，仍顯著優於單一學習器之表現。

綜整而言，透過集成學習演算法確實可以提升單一模型的型二誤差與 F1 值的表現，尤其是經過判斷標準篩選的模型堆疊模型；但要注意的是，模型堆疊和模型勻合綜合評比的表現其實相當，但模型勻合可以節省許多運算時間，在資料集龐大或需要快速運算時機下，此一特色將相當受到使用者青睞。

表 11　模型績效指標 Wilcoxon 符號等級檢定

	集成學習表現最佳模型			集成學習表現最差模型		
	型二誤差	F_1 值	AUC	型二誤差	F_1 值	AUC
Stacking (剔除 SVM)	--	--	-0.37*	-1.83*	-1.83*	0.00
Stacking (全學習器)	-1.64*	-1.28*	--	-1.64*	-1.46*	-1.10*
Blending (剔除 SVM)	-1.83*	-1.83*	-1.10*	-1.64*	-0.73*	--
Blending (全學習器)	-1.83*	-1.83*	-0.55*	--	--	0.00
隨機森林	-1.83*	-1.83*	-1.83*	-1.83*	0.00	-1.46*
XGBoost	-1.83*	-1.83*	-1.46*	-1.83*	-1.83*	-0.73*
類神經網路	-1.83*	-1.83*	-1.83*	-1.64*	-1.83*	-1.83*
羅吉斯迴歸	-1.46*	-1.83*	-1.83*	-1.83*	-0.73*	-1.83*
支援向量機	-1.83*	-1.83*	-1.46*	-1.83*	-1.83*	-1.46*

註：1. 當樣本數為 4 時，Wilcoxon 符號等級檢定臨界值為 0，當檢定統計量小於臨界值達 10%統計顯著，以*表示。2. --代表此模型於該項評估指標之基準模型。

陸、結論

本研究利用台灣信用卡使用行為與基本資料，引入堆疊與勻合集成並建構違約風險預警模型，以期滿足傳統銀行借貸與金融科技借貸之於不同面向的成本（誤判成本與時間成本）考量。Finlay (2010) 指出正確辨別借貸客戶是否能如期繳款對於銀行獲利表現產生重大影響，因此如何降低型二錯誤發生的機率與提高 F_1 指標攸關銀行經營績效；此外，Hand and Henley (1997) 認為即便是微幅的提高辨識成功率，對銀行帶來的獲利仍相當可觀。除此之外，當銀行使用之模型的型二錯誤率較低的情況下，意味著其逾放比、所須提列的備抵呆帳費用也較小，這些科目都攸關銀行資產品質 (asset quality) 以及所面臨之信用風險，也是金融監理單位觀察之重點。

本研究實證發現，兼具學習器挑選策略與良好框架設計的集成模型方能穩定地在綜合辨別能力上勝過最佳的單一模型；以衡量辨別客戶能力的不同指標為考量，輔以學習器挑選策略模型堆疊模型誤判成本最低，體現在誤判違約客戶為正常的錯誤最少，正確識別違約客戶能力最強，區別客戶能力最佳，明顯勝出任何非集成的單一學習器。兼顧模型表現與時間成本，結合學習器挑選策略的模型勻合在保有不俗模型表現能力的同時大幅降低訓練時間。

在訓練的單一模型時，本研究所捕捉到的重要特徵大致相同，而位於重要性首位的變數為本研究自行合併生成的 PAY_RISK，該變數代表客戶過去平均逾期繳款月數，即相較於銀行授信額度、客戶過往支付歷史等資料，客戶逾期記錄最有助於識別客戶潛在違約風險。近年來，亦有學者關注軟資訊對信用評分系統的增進。例如數位足跡 (digital footprint) (Berg et al., 2020)，友誼與社交網絡 (Hildebrandt et al., 2017) 以及 文本資料 (Gao et al., 2018; Netzer et al., 2019) 等。然而，這類資料不易從公開渠道獲取，且在使用時也需考量個人資料保護之事宜。Jagtiani and Lemieux (2019) 認為貸方未經借方授權而使用個人資料(諸如社交網絡的詳細情況) 可能造成隱私洩漏的疑慮。本文使用較為質樸的集成學習框架，也沒有刻意地調整與訓練學習器之參數，一方面增補集成學習應用於信用卡評分之文獻，作為後續研究提供比較之基準，一方面也為後續模型結合更多軟資訊留下空間。

參考文獻

Abellán, J., and Mantas, C. J. (2014), "Improving Experimental Studies about Ensembles of Classifiers for Bankruptcy Prediction and Credit Scoring," *Expert Systems with Applications*, *41*(8), 3825-3830.

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., and Vanthienen, J. (2003), "Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring," *Journal of the OperationalRresearch Society*, *54*(6), 627-635.

Berg, T., Burg, V., Gombović, A., and Puri, M. (2020), "On the Rise of FinTechs–Credit Scoring using Digital Footprints," *The Review of Financial Studies*, *33*(7), 2845-2897.

Breiman, L. (1996), "Bagging Predictors," *Machine Learning*, *24*(2), 123-140.

Churchill, G. A., Nevin, J. R., and Watson, R. R. (1977), "The Role of Credit Scoring in the Loan Decision," *Credit World*, 3(3), 6-10.

Desai, V. S., Crook, J. N., and Overstreet Jr, G. A. (1996), "A Comparison of Neural Networks and Linear Scoring Models in the Credit Union Environment," *European Journal of Operational Research*, *95*(1), 24-37.

Demšar, J. (2006), "Statistical Comparisons of Classifiers over Multiple Data Sets," *Journal of Machine Learning Research*, *7*(Jan), 1-30.

Dietterich, T. G. (1997), "Machine-Learning Research," *AI Magazine*, *18*(4), 97-97.

Dorfleitner, G., Priberny, C., Schuster, S., Stoiber, J., Weber, M., de Castro, I., and Kammler, J. (2016), "Description-text related soft information in peer-to-peer lending–Evidence from two leading European platforms," *Journal of Banking & Finance*, *64*, 169-187.

Durand, D. (1941), *Risk Elements in Consumer Installment Financing*. National Bureau of Economic Research, New York.

Fawcett, T. (2006), "An introduction to ROC analysis," *Pattern Recognition Letters*, *27*(8), 861-874.

Finlay, S. (2011), "Multiple Classifier Architectures and their Application to Credit Risk Assessment," *European Journal of Operational Research*, *210*(2), 368-378.

Freund, Y., and Schapire, R. E. (1996), "Experiments with a New Boosting Algorithm," In *icml*, *96*, 148-156.

Friedman, M. (1937), "The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance," *Journal of the American Statistical Association*, *32*(200), 675-701.

Friedman, M. (1940), "A Comparison of Alternative Tests of Significance for the Problem of m Rankings," *The Annals of Mathematical Statistics*, *11*(1), 86-92.

Gao, Q., Lin, M., and Sias, R. W. (2018), "Words Matter: The Role of Texts in Online Credit Markets," *Working paper*.

Guo, H., Li, Y., Shang, J., Gu, M., Huang, Y., and Gong, B. (2017), "Learning from Class-Imbalanced Data: Review of Methods and Applications," *Expert Systems with Applications, 73,* 220-239.

Hand, D. J., and Henley, W. E. (1997), "Statistical Classification Methods in Consumer Credit Scoring: A Review," *Journal of the Royal Statistical Society: Series A (Statistics in Society), 160*(3), 523-541.

Hansen, L. K., and Salamon, P. (1990), "Neural Network Ensembles," *IEEE Transactions on Pattern Analysis and Machine Intelligence, 12*(10), 993–1001.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009), *The Elements of Statistical Learning (2nd ed.)*, New York: Springer.

He, H., and Garcia, E. A. (2009), "Learning from Imbalanced Data," *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263-1284.

Henley, W. E. (1995), "Statistical aspects of credit scoring," *Doctoral dissertation, The Open University*.

Henley, W. E., and Hand, D. J. (1996), "AK - Nearest - Neighbour Classifier for Assessing Consumer Credit Risk," *Journal of the Royal Statistical Society: Series D (The Statistician),45*(1), 77-95.

Hildebrand, T., Puri, M., and Rocholl, J. (2017), "Adverse Incentives in Crowdfunding," *Management Science, 63*(3), 587-608.

Hsieh, N. C., and Hung, L. P. (2010), "A Data Driven Ensemble Classifier for Credit Scoring Analysis," *Expert Systems with Applications, 37*(1), 534-545.

Iyer, R., Khwaja, A. I., Luttmer, E. F., and Shue, K. (2016), "Screening Peers Softly: Inferring the Quality of Small Borrowers," *Management Science, 62*(6), 1554-1577.

Jagtiani, J., and Lemieux, C. (2019), "The Roles of Alternative Data and Machine Learning in Fintech Lending: Evidence from the LendingClub Consumer Platform ," *Financial Management, 48*(4), 1009-1029.

Kuncheva, L. I., and Whitaker, C. J. (2003), "Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy," *Machine Learning, 51*(2), 181-207.

Kuncheva, L. I. (2004), *Combining Pattern Classifiers: Methods and Algorithms*, New York: John Wiley & Sons.

Lehmann, E. L., and D'Abrera, H. J. (1975), *Nonparametrics: Statistical Methods Based on Ranks*, Holden-day.

Lessmann, S., Baesens, B., Seow, H. V., and Thomas, L. C. (2015), "Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring: An Update of Research," *European Journal of Operational Research, 247*(1), 124-136.

Liberti, J. M., and Petersen, M. A. (2019), "Information: Hard and soft," *Review of Corporate Finance Studies, 8*(1), 1-41.

Lin, W. Y., Hu, Y. H., and Tsai, C. F. (2012). Machine Learning in Financial Crisis Prediction: A Survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 42*(4), 421-436.

Lyles, R. W., and Hummer, J. E. (2012), *Effective Experiment Design and Data Analysis in Transportation Research* (Vol. 727), Transportation Research Board.

Marqués, A. I., García, V., and Sánchez, J. S. (2012), "Exploring the Behaviour of Base Classifiers in Credit Scoring Ensembles," *Expert Systems with Applications, 39*, 10244-10250.

Mason, R., Lind, D., and Marchal, W. (1999), "Nonparametric Methods: Analysis of Ranked Data," *Statistical Techniques in Business and Economics*, 541-574.

Myers, J. H., and Forgy, E. W. (1963), "The Development of Numerical Credit Evaluation Systems," *Journal of the American Statistical association*, 58(303), 799-806.

Nanni, L., and Lumini, A. (2009), "An Experimental Comparison of Ensemble of Classifiers for Bankruptcy Prediction and Credit Scoring," *Expert Systems with Applications*, 36(2), 3028-3033.

Netzer, O., Lemaire, A., and Herzenstein, M. (2019), "When Words Sweat: Identifying Signals for Loan Default in the Text of Loan Applications," *Journal of Marketing Research, 56*(6), 960-980.

Paleologo, G., Elisseeff, A., and Antonini, G. (2010), "Subagging for Credit Scoring Models," *European Journal of Operational Research, 201*(2), 490-499.

Polikar, R. (2006), "Ensemble Based Systems in Decision Making," *IEEE Circuits and Systems Magazine, 6*(3), 21-45.

Rabinowicz, A., and Rosset, S. (2020), "Cross-Validation for Correlated Data," *Journal of the American Statistical Association*, 1-14.

Reichert, A. K., Cho, C. C., and Wagner, G. M. (1983), "An Examination of the Conceptual Issues Involved in Developing Credit-Scoring Models," *Journal of Business & Economic Statistics*, *1*(2), 101-114.

Schapire, R. E. (1990), "The Strength of Weak Learnability," *Machine Learning, 5*(2), 197-227.

Thomas, L. C. (2000), "A Survey of Credit and Behavioural Scoring: Forecasting Financial Risk of Lending to Consumers," *International Journal of Forecasting*, *16*(2), 149-172.

Tsai, C. F., Hsu, Y. F., and Yen, D. C. (2014), "A Comparative Study of Classifier Ensembles for Bankruptcy Prediction," *Applied Soft Computing, 24*, 977-984.

Töscher, A., Jahrer, M., and Bell, R. M. (2009), "The BigChaos Solution to the Netflix Grand Prize," *Netflix Prize Documentation*, 1-52.

Van Liebergen, B. (2017), "Machine Learning: A Revolution in Risk Management and Compliance?" *Journal of Financial Transformation, 45*, 60-67.

Wang, G., Hao, J., Ma, J., and Jiang, H. (2011), "A Comparative Assessment of Ensemble Learning for Credit Scoring," *Expert systems with applications, 38*(1), 223-230.

Wang, G., Ma, J., Huang, L., and Xu, K. (2012), "Two Credit Scoring Models Based on Dual Strategy Ensemble Trees," *Knowledge-Based Systems, 26*, 61-68.

West, D. (2000), "Neural Network Credit Scoring Models," *Computers & Operations Research, 27*(11-12), 1131-1152.

West, D., Dellana, S., and Qian, J. (2005), "Neural Network Ensemble Strategies for Financial Decision Applications," *Computers & Operations Research, 32*(10), 2543-2559.

Williams, C. K., and Rasmussen, C. E. (2006), *Gaussian Processes for Machine Learning (Vol. 2, No. 3, p. 4)*, Cambridge, MA: MIT press.

Windeatt, T., and Ardeshir, G. (2004), "Decision Tree Simplification for Classifier Ensembles," *International Journal of Pattern Recognition and Artificial Intelligence, 18*(05), 749-776.

Wolpert, D. H. (1992), "Stacked Generalization," *Neural Networks, 5*(2), 241-259.

Xia, Y., Liu, C., Li, Y., and Liu, N. (2017), "A Boosted Decision Tree Approach using Bayesian Hyper-Parameter Optimization for Credit Scoring," *Expert Systems with Applications*, 78, 225-241.

Xia, Y., Liu, C., Da, B., and Xie, F. (2018), "A Novel Heterogeneous Ensemble Credit Scoring Model Based on Bstacking Approach," *Expert Systems with Applications, 93*, 182-199.

Xia, Y., Zhao, J., He, L., Li, Y., and Niu, M. (2020), "A Novel Tree-Based Dynamic Heterogeneous Ensemble Method for Credit Scoring," *Expert Systems with Applications*, 113615.

Yeh, I. C., and Lien, C. H. (2009), "The Comparisons of Data Mining Techniques for the Predictive Accuracy of Probability of Default of Credit Card Clients," *Expert Systems with Applications, 36*(2), 2473-2480.

Yu, L., Wang, S., and Lai, K. K. (2008), "Credit Risk Assessment with a Multistage Neural Network Ensemble Learning Approach," *Expert Systems with Applications*, *34*(2), 1434-1444.

Zhang, D., Zhou, X., Leung, S. C., and Zheng, J. (2010), "Vertical Bagging Decision Trees Model for Credit Scoring," *Expert Systems with Applications*, 37(12), 7838-7843.

Zhou, Z. H. (2009), "Ensemble Learning," *Encyclopedia of Biometrics, 1*, 270-273.

附錄

附表 1　單一模型超參數之設定

模型	參數	說明	設定	備註
羅吉斯迴歸	penalty	正規化方法	L2	處理過擬合問題
	C	正規化參數	0.008	需為正數
	class_weight	調節樣本類別權重	balanced	
支援向量機	kernel	核函數	rbf	
類神經網路	epochs	訓練次數	150	
	batch_size	批次訓練量	1200	
	metric	評估模型指標	['acc', 'mse',getRecall, getPrecision]	getRecall 與 getPrecsion 為本研究自行定義的函數
隨機森林	n_estimators	決策樹數量	200	
	max_features	節點分割時，參與判斷的最大特徵數	auto	auto 設定為所有特徵數的平方根
	max_depth	樹的最大深度	6	
XGBoost	n_estimators	決策樹數量	90	
	max_depth	節點分割時，參與判斷的最大特徵數	3	
	subsample	子集抽樣比例	0.9	用於訓練模型的子集占整體集合之比例

註：本表的正規化參數為 λ 的倒數。

附表 2　集成學習超參數之設定

模型	參數	說明	設定	備註
Blending (全學習器)	penalty	正規化方法	L2	處理過擬合問題
	C	正規化參數	0.01	需為正數
	class_weight	調節樣本類別權重	balanced	
Blending (剔除 SVM)	penalty	正規化方法	L2	處理過擬合問題
	C	正規化參數	0.002	需為正數
	class_weight	調節樣本類別權重	balanced	
Stacking (全學習器)	n_folds	K 折數	5	
	regression	是否用於迴歸	False	此為分類問題
	penalty	正規化方法	L2	處理過擬合問題
	C	正規化參數	0.009	需為正數
	class_weight	調節樣本類別權重	balanced	
Stacking (剔除 SVM)	n_folds	K 折數	5	
	regression	是否用於迴歸	False	此為分類問題
	penalty	正規化方法	L2	處理過擬合問題
	C	正規化參數	0.001	需為正數
	class_weight	調節樣本類別權重	balanced	

註：本表的正規化參數為 λ 的倒數。