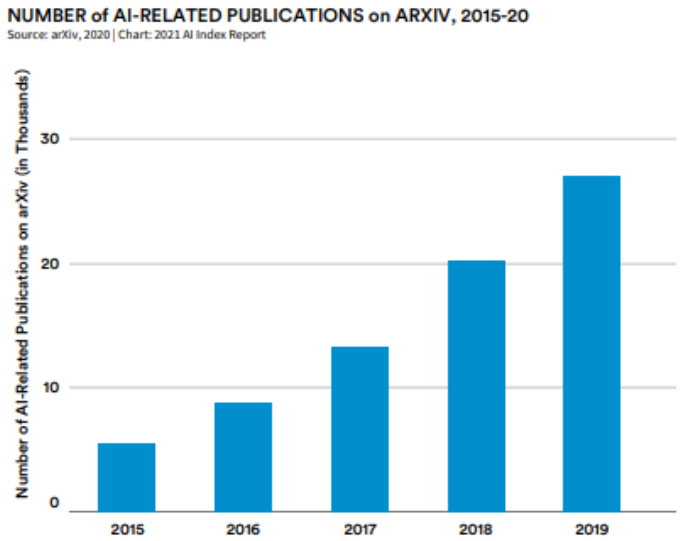
機器學習模型應用於信用違約之預測

一、 前言

硬體的進步與開源程式庫叢立，如 scikit-learn, tensorflow (Abadi, Martín, et al(2016))，成為機器學習快速發展的溫床，重量級文獻揭連發表，啟發許多其他領域的學者和研究人員，躋身機器學習相關的應用，使其漸普及於日常生活。根據 arxiv 的統計資料顯示，AI 相關的論文發表自 2015 至 2020 年，由 5,478 提升到 34,736，年平均成長率達到 106.82%。影像辨識、語音辨識和近期轟動全球的 ChatGPT(由 OpenAI 開發)，AI 在影像和語音的領域上，已獲得巨大的成就，皆有相對成熟的產品；而在金融和醫學上，尚未出現成熟的產品，潛在成長空間相對大。

圖一、arxiv 的 AI 相關論文出版統計資料



Source: Artificial Intelligence Index Report.(2021)。

AI 獲得如此巨大的成功，主要原因來自於機器學習和深度學習模型，汲取特徵的方式和傳統統計方法相差盛鉅，一般線性模型只能捕捉到資料間線性的關係，而機器學習善於捕捉非線性關係。如在二分類問題上，常見的統計方法，可能會使用羅吉斯迴歸來進行預測，但在樣本類別極度不平衡的情況下，羅吉斯迴歸會受到資料佔比較大的一方所影

響，因而計算出失準之權重；在機器學習的模型中，常見用來解決二分類問題的模型，有支持向量機(Support Vector Machine, SVM)和隨機森林(Random Forest, RF)，根據 SVM 的損失函數，便可以了解其運作的核心概念，SVM 運用兩群體間最鄰近的資料，並在資料間找出群體間的邊界，分界線的權重計算，則考慮到兩群體的距離需一致，也正是因為這樣的機制，使得 SVM 不受樣本比例的影響。另一方面，RF 強大的地方在於其運用了集成式學習中 bagging 的概念，透過隨機抽樣組建多組子樣本，再透過不同的子樣本分別去訓練模型，亦即每個模型所使用的訓練資料皆有所差異，其權重也會有所差異。Bagging 的概念與投票機制相同，透過結合上述多個模型的預測結果，使用平均結果作為 RF 模型的最終預測結果，而 SVM 和 RF 模型也被證實在分類上，優於羅吉斯迴歸模型。

在金融產業中，機器學習可應用於投資領域，針對股價和報酬率進行預測、優化資產配置模型等，如 Kolanovic et al. (2017), Raffinot et al. (2017)等人的研究，發現階層式風險平等法(Hierarchical Risk Parity, HRP)，優於傳統財務理論中的配置方法—效率前緣和等權重配置；同時也發現 HRP 能有效解決馬可維茲的詛咒—過度配適樣本內的資料，解決其在樣本外預測效果不佳問題。

而本文旨在研究金融產業的另一項議題—風險分析中信用違約預測，根據 Xolani et al(2020)的回顧性研究表明，集成式的分類模型在整體表現上，優於任何單一分類器，而深度學習又可以透過疊加多個隱藏層，來進一步發掘資料間的隱性特徵，本文重點在於使用深度學習結合元學習的概念，並比較預測結果元學習所訓練出來的模型是否優於集成式和其他指標模型。此處選用之深度學習模型為 MLP、CNN(Bing Zhu et al. (2018))，集成式模型選用 RF，其他標竿模型則選用 SVM。評估模型的指標選用正確率(PCC)、接收者操作特徵曲線(Receiver operating characteristic curve, ROC curve)及線下面積(Area under curve, AUC)。

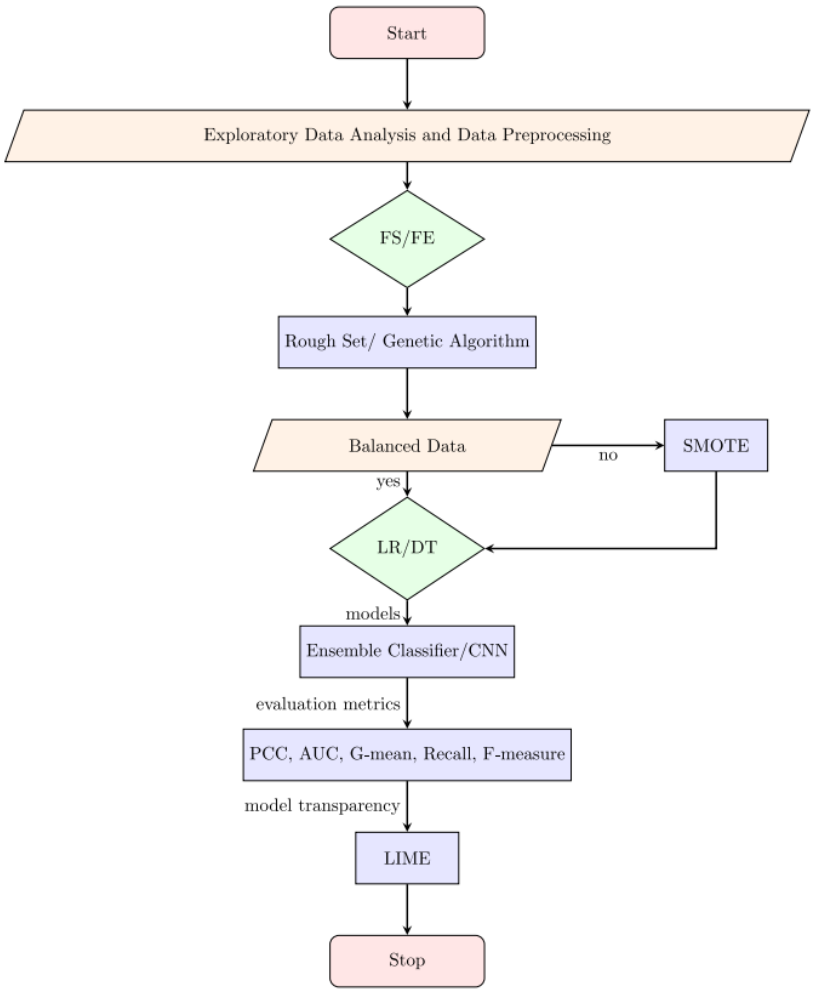
二、 文獻整理與探討

根據 Xolani et al(2020)回顧文獻匯整了 2010 至 2018 年間共 74 篇，關於信用分級模型的預測研究，表明集成式的分類模型在整體表現上，優於任何單一分類器。此外，也為後續信用違約預測方面的研究；訂定了研究架構(圖二)。

依圖二所示，研究流程主要可分為以下多個步驟，第一步資料分析及

資料前處理，第二步進行特徵選取或特徵工程，但現實世界中，實際發生違約的情況很稀少，所以違約類型的資料，常會有樣本不平衡的問題出現。像在影像辨識上，常見調整樣本不平衡的方法 SMOTE、抽樣方法中的過度取樣、低度取樣等方法。(Chawla et al. (2016), Saia et al. (2016))

圖二、研究架構圖



Source: Xolani et al.(2020)。

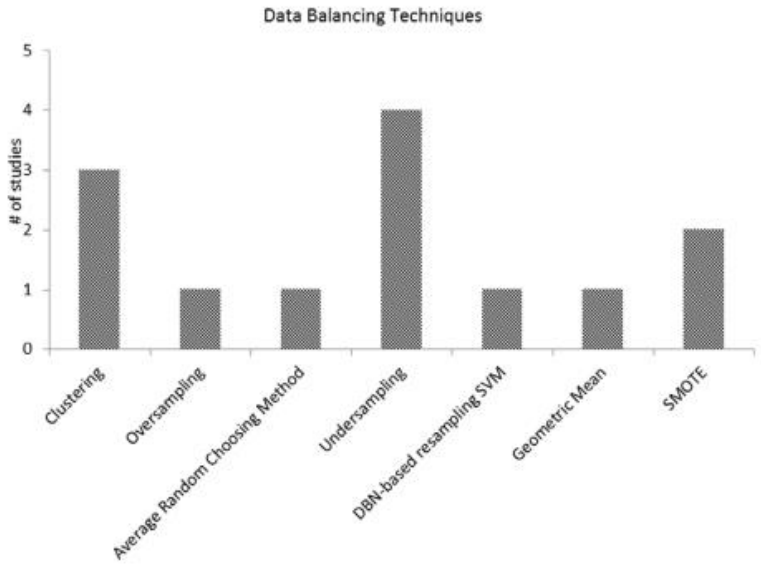
1. 特徵選取/特徵工程

本研究資料選用台灣上市櫃公司，2015 年到 2020 年底的資料，總共 114 個特徵值，主要為財報和董監事資訊。由於特徵數足夠，將不會進行特徵工程。此外，特徵選取也是財經研究的熱門議題，其研究流程又與本文最終目的有些出入，提供其他研究人員作為未來研究方向。

2. 平衡資料

依研究使用的資料集顯示，違約事件(類別=1)占全部資料不到 2%(約 150 筆資料)，若使用傳統統計方法的羅吉斯迴歸，會嚴重受到此樣本不平衡的因素，導致學習成效低落，即模型全部預測非違約(類別=0)，也會有 98%的正確率。針對這種情況，根據圖三指出，最常使用的方式是下取樣(Under sampling)，透過刪除主要類別的樣本數，在此案例中即刪除類別為 0，使兩類別的樣本數盡可能達到平衡，類別 0 和 1 的比例接近 1。然而缺點也相對明顯，使用下取樣會因此失去大量的數據，在整體數據量不夠充足的情況下，使用該方法非明智之舉。

圖三、平衡樣本方法於文獻使用之分布圖



Source: Xolani et al.(2020)。

由於兩類別比例差距過於懸殊，在權衡之下，本研究選擇上採樣(Oversampling)，透過重複抽取次要類別，來調整樣本中類別 1 相較於整體樣本的比例。然而，要從不到 2%拉升至 50%，可能會導致其他問題產生，如:過度配適類別 1 的資料，導致模型訓練出來的權重不具一般性。

在此兩點的考慮下，發現在圖片辨識中，有一套應對此情況的學習方法—小樣本學習(Few-Shot Learning, FSL)，主要目的在於建立多組小樣本(task)，而一個 task，須包含類別數和子樣本數(N-ways K-shot)，依信用風險為例，違約與非違約資料含有兩個類別，故 N=2，而各類別要抽取多少樣本 K 的數量，則由使用者自行決定，本研究將對 K 進行敏感性分析，來尋找最適的 K。

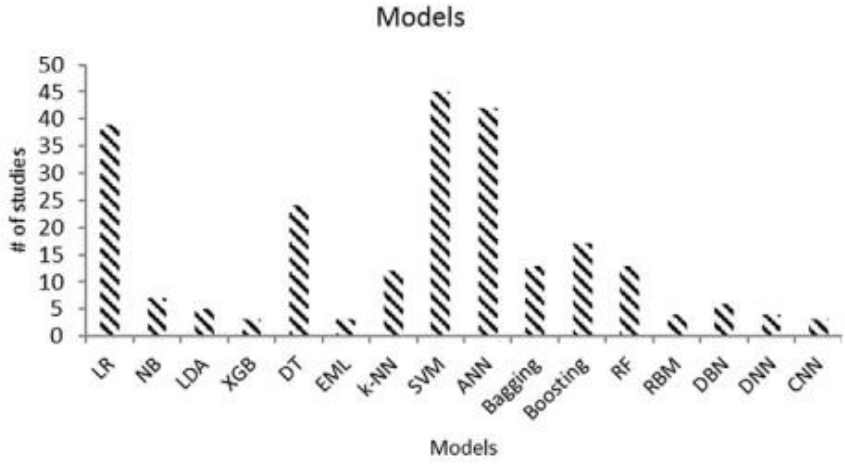
3. 模型選用

承上，本文使用小樣本學習(FSL)的機制，搭配模型不可之元學習(Model-Agnostic Meta-Learning, MAML)的學習架構。

常見的深度學習模型，目的是學習一個用於預測的數學模型。MAML 作為元學習中的一種學習框架，重點在於學習的過程，不是直接用來預測，而是學習"如何更快更好的學習一個數學模型"。MAML 名字中所提到的模型無關 Model-Agnostic，會這麼命名主要是因為 MAML 學習框架，可以套用在任何模型上，絕大多數深度模型皆可無縫嵌入 MAML 中，而本研究這次使用的模型為 MLP，其他模型如 LSTM 和 CNN 的應用，待後續學者進行研究。

根據圖四指出，SVM 作為最常被使用的模型之一，主要在於其核心理念，權重之學習不受樣本不平衡的因素所影響，適合納入本研究的指標模型。此外指標模型選用 RF 模型，儘管 RF 模型使用頻率明顯較低，但 Xolani et al(2020)的研究結論指出，集成式的分類模型在整體表現上，優於任何單一分類器，因此指標模型應將集成式學習的模型(Bagging, Boosting, RF)加入考慮，在此基礎下，本文選擇將 RF 納入並作為指標模型之一。

圖四、預測模型於文獻使用之分布圖

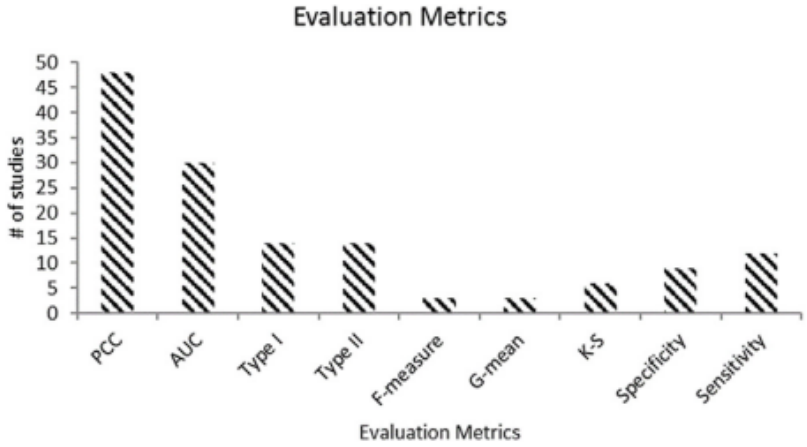


Source: Xolani et al.(2020)。

4. 模型評估

根據圖五，最常使用的前五項模型評估指標，為正確率(PCC)、接收者操作特徵曲線(Receiver operating characteristic curve, ROC curve)及線下面積(Area under curve, AUC)、型一錯誤(Type I Error)、型二錯誤(Type II Error)和敏感性分析，本文遵循此邏輯，會分別評估這五項指標對模型的優劣。

圖五、模型評估方法於文獻使用之分布圖



Source: Xolani et al.(2020)。

三、 模型和變數選用

使用變數由 TEJ 所整理，包含財務資料和屬性資料(產業別)，共計 11 項變數，使用 2015 至 2020 年所有上市公司之資料，資料筆數共計 9000 多筆。指標模型為 SVM、RF，本研究重點在於使用 MAML 對 MLP 做校準學習。

為更細部說明 FSL 和 MAML 的學習概念，以下將舉例說明，以影像辨識為例，FSL 的概念在於從原本的數據中進行抽樣，以組成多個小樣本，但與統計方法中的重複抽樣概念，差在小樣本的維度。假設資料的特徵數量為 100 個變數，每次抽取的資料筆數為 100 筆，則每個子樣本的維度為 100x100，若重複抽取 100 次，則整個子樣本的集合，其維度等於 100x100x100；而在 FSL 中，會分別針對不同類別的資料進行抽樣，以二分類為例，每次抽取 50 筆資料，最後 task 的維度會變成 2x50x100，重複抽取 100 次後，task 集合維度變成 100x2x50x100。兩者有著同樣的資料總筆數，但一次抽取的子樣本數量有所差異，在總資料筆數相同的情況下，單次抽取的資料筆數越多，越容易抽到重複的資料，而使用 FSL 則是透過提高一個維度，使單次抽樣的數量下降，避免

各 task 間的資料重複性太高。

Finn, C. et al. (2017)提出了使用梯度來更新模型，根據圖六可以知道元學習的權重更新可以分為 5 個流程:

1. 設置權重的初始值。

2. 進入第一層迴圈後，對資料進行抽樣並組建 task 集合。

3. 進入第二層迴圈後，逐步將各 task 樣本傳入**內模型**中，並計算每次的梯度，再將梯度用來更新**內模型**的權重。

4. 離開第二層迴圈，回到第一層迴圈後，使用**內模型的平均損失**來計算新的梯度，並用新的梯度來更新**外模型**。

5. 重複迭代直到設定的條件滿足後，即完成整個 MAML 的訓練。

本研究的內模型使用 MLP，隱藏層共 3 層，輸出層 1 層，輸出層之激活函數使用 softmax，會根據輸入資料分別計算出所有類別的機率，優化器使用 Adam 優化器來更新權重，學習率調整為 0.001。外模型同樣使用 Adam 更新權重，此處的學習率設為 0.0001，且外模型更新權重，會使用內模型的平均損失來計算梯度。

圖六、MAML 運作流程

Algorithm 1 Model-Agnostic Meta-Learning

Require: $p(\mathcal{T})$: distribution over tasks
Require: α, β: step size hyperparameters
1: randomly initialize θ
2: **while** not done **do**
3: Sample batch of tasks $\mathcal{T}_i \sim p(\mathcal{T})$
4: **for all** $\mathcal{T}_i$ **do**
5: Evaluate $\nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$ with respect to K examples
6: Compute adapted parameters with gradient descent: $\theta_i' = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$
7: **end for**
8: Update $\theta \leftarrow \theta - \beta \nabla_\theta \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(f_{\theta_i'})$
9: **end while**

Source: Finn, C. et al. (2017)。

四、 實證與模型比較

1. SVM 指標模型

SVM 參考圖四為最常被文獻使用，進行信用違約預測的模型，本文使用該模型進行預測，結果如圖七、八所示。根據圖七的混淆矩陣可觀察到，SVM 預測類別 0 的能力較類別 1 佳。

在真實類別為**非違約**的情況下，預測類別為**違約**的數量，在樣本內和樣本外皆為 0，即無違約模型為違約的機率為 0。

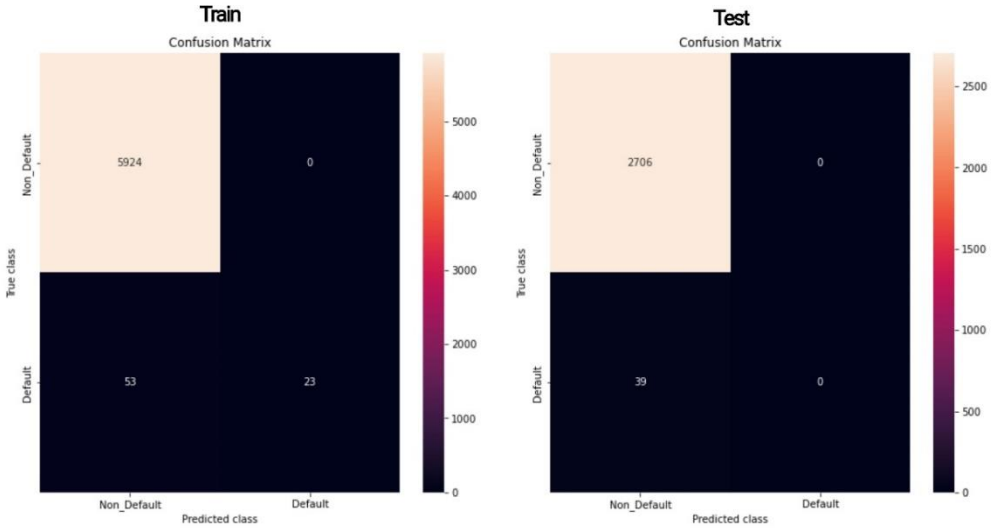
反之，在真實類別為違約的情況下，預測類別為非違約的數量在樣本內為 53，樣本外為 39，型二錯誤為

$$\beta_{train} = \frac{FN}{TP + FN} = \frac{53}{53 + 23} = 0.6974$$

$$\beta_{test} = \frac{FN}{TP + FN} = \frac{39}{39 + 0} = 1$$

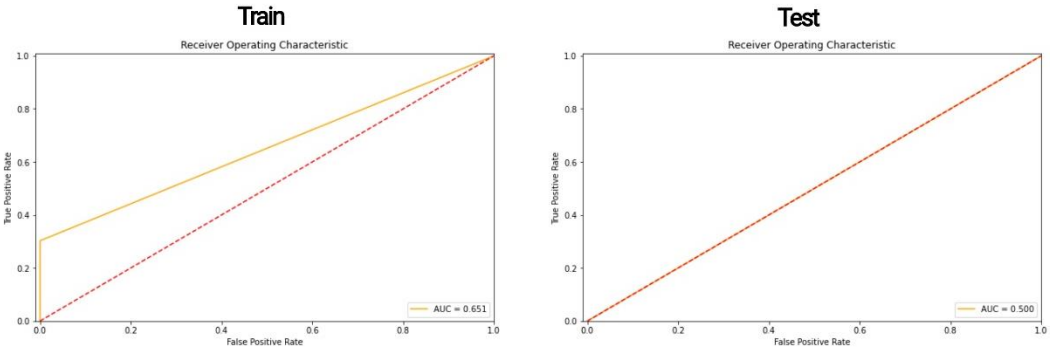
此外，根據圖八可知，在訓練集中 SVM 之 AUC 為 0.651，高於 0.5。然而，在樣本外測試資料則顯示 SVM 之 AUC 為 0.5，AUC 等於 0.5，表示模型的預測力和隨機猜測一樣(二分類的情況下，隨機猜中的機率為 0.5)。

圖七、SVM 之混淆矩陣



Source:本研究自行整理。

圖八、SVM 之 ROC 和 AUC



Source:本研究自行整理。

2. RF 指標模型

儘管 RF 被文獻使用來進行信用違約預測的頻率沒有像 SVM 那般頻繁(如圖四)，但 RF 的集成式學習機制，卻是 Xolani et al(2020)在整合了 74 篇研究所提出的最佳解，在本研究的表現如圖九、十所示。根據圖九的混淆矩陣可觀察到，RF 預測類別 0 的能力較類別 1 來的優秀。

而在真實類別為非違約的情況下，預測類別為違約的數量，在樣本內外皆微為 0，即無違約模型為違約的機率為 0。

反之，在真實類別為違約的情況下，預測類別為非違約的數量在樣本內為 55，樣本外則是 39，型二錯誤為

$$\beta_{train} = \frac{FN}{TP + FN} = \frac{55}{55 + 21} = 0.7237$$

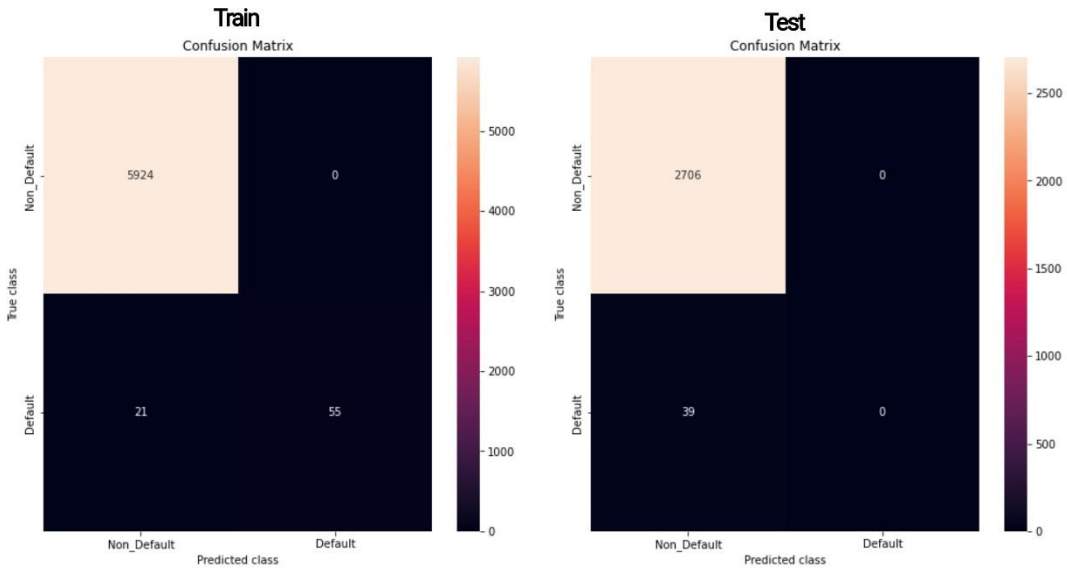
$$\beta_{test} = \frac{FN}{TP + FN} = \frac{39}{39 + 0} = 1$$

此外，根據圖十可知，在訓練集中 RF 之 AUC 為 0.862，高出 0.5 非常多。然而，在樣本外測試資料則顯示 RF 之 AUC 為 0.5，AUC 等於 0.5，表示模型的預測力和隨機猜測一樣(二分類的情況下，隨機猜中的機率為 0.5)。

結合 SVM 和 RF 的結果可以發現，兩模型受類別樣本比例影響甚鉅，儘管 RF 在樣本的結果看似有一定程度預測類別 1 的能力，但這些

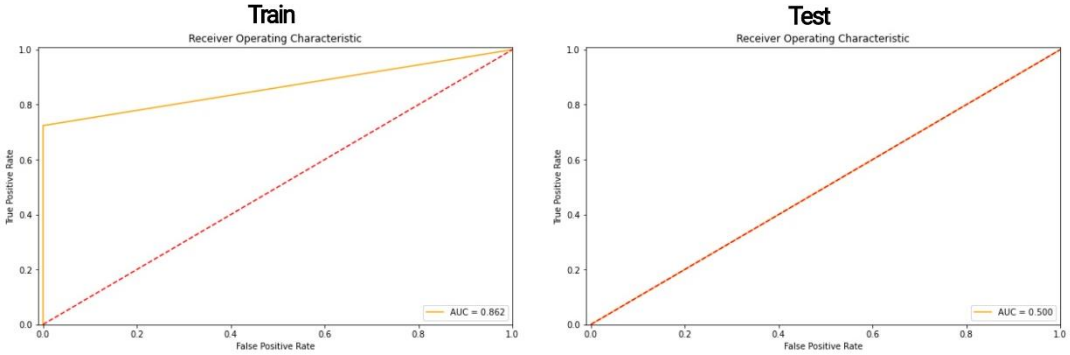
學習成果在樣本外就失去了作用。而這也是為何本研究提出使用 FSL 搭配 MAML 來訓練模型。

圖九、RF 之混淆矩陣



Source:本研究自行整理。

圖十、RF 之 ROC 和 AUC



Source:本研究自行整理。

3. MAML 學習機制

MAML 搭配 FSL 後，可調控的參數又變得更多了，因此在細談 MAML 的訓練結果前，本文先針對 18 組參數組合進行了網格分析(圖十一)，目的在找尋最佳的參數組合。測試的參數有兩個，一是 task 的數

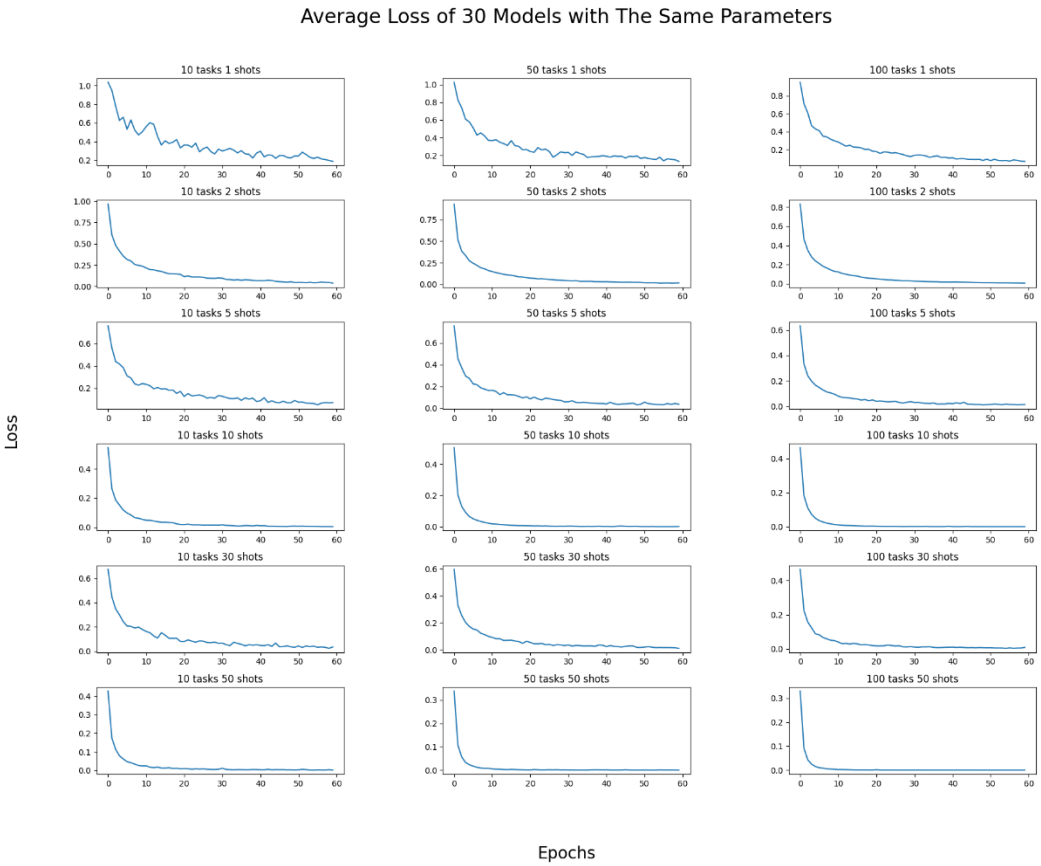
量，task 決定了一次要輸入多少筆資料進入模型中；二是 shot 的數量，shot 則表示一組 task 要抽取多少組樣本。

圖十一為網格分析的結果，本研究針對每組參數皆運行 30 次，並以 30 次的平均損失作為依據來繪製圖表，主要係在訓練深度學習模型時，權重的初始位置也是重要參數之一，使用平均結果較能體現出參數(task, shot)本身所帶來的影響。根據結果可看出，當 shot 數為 1 的時候，代表一組 task 樣本，僅包含類別 0 和 1 的資料各一筆，後續傳入新的 task，其資料容易出現模型沒看過的資料機率較高，因此訓練過程較波動，導致平均損失出現較劇烈的波動。而此現象隨 shot 數量的上升，損失曲線越趨平滑。

觀察 task 數量對模型訓練的影響，可以發現相同 shot 數下，task 數與損失曲線的平滑程度成正比，該現象尤其在 shot 數量等於 1、5、30 時尤為明顯。

最後，要根據訓練資料集來挑選合適的參數組合，評估的要點可參考 task、shot 與外模型更新次數(本研究設為 60 次)的乘積，為總訓練資料的特定倍數來評估，若此倍數等於 10，則表示一筆訓練數據的預期抽樣次數為 10 次。故從圖十一的最左上角到最左下角，倍數的範圍介於 0.1 至 50 倍，被數低於 1 基本上會有抽樣不足的情況發生，高於 30 則有過度採樣的疑慮，故後續將會針對倍數介於區間 5 至 10 的參數，去分析其 PCC、AUC、型一錯誤和型二錯誤。

圖十一、網格分析



Source:本研究自行整理。

根據上述倍數區間所對應的參數組合，有(10, 50), (50, 10), (100, 5), (100, 10)共四組參數。根據圖十二的四組混淆矩陣，可得出四個參數組合相對應的準確率，樣本內之準確率分別是 64.78%、**68.73%**、65.78%和 66.67%，以(50, 10)該參數組合準確率最高。樣本外的結果分別是 61.85%、**65.15%**、62.63%和 63.63%。有樣本內外之準確率可明顯看出 MAML 的訓練結果較具一般性。

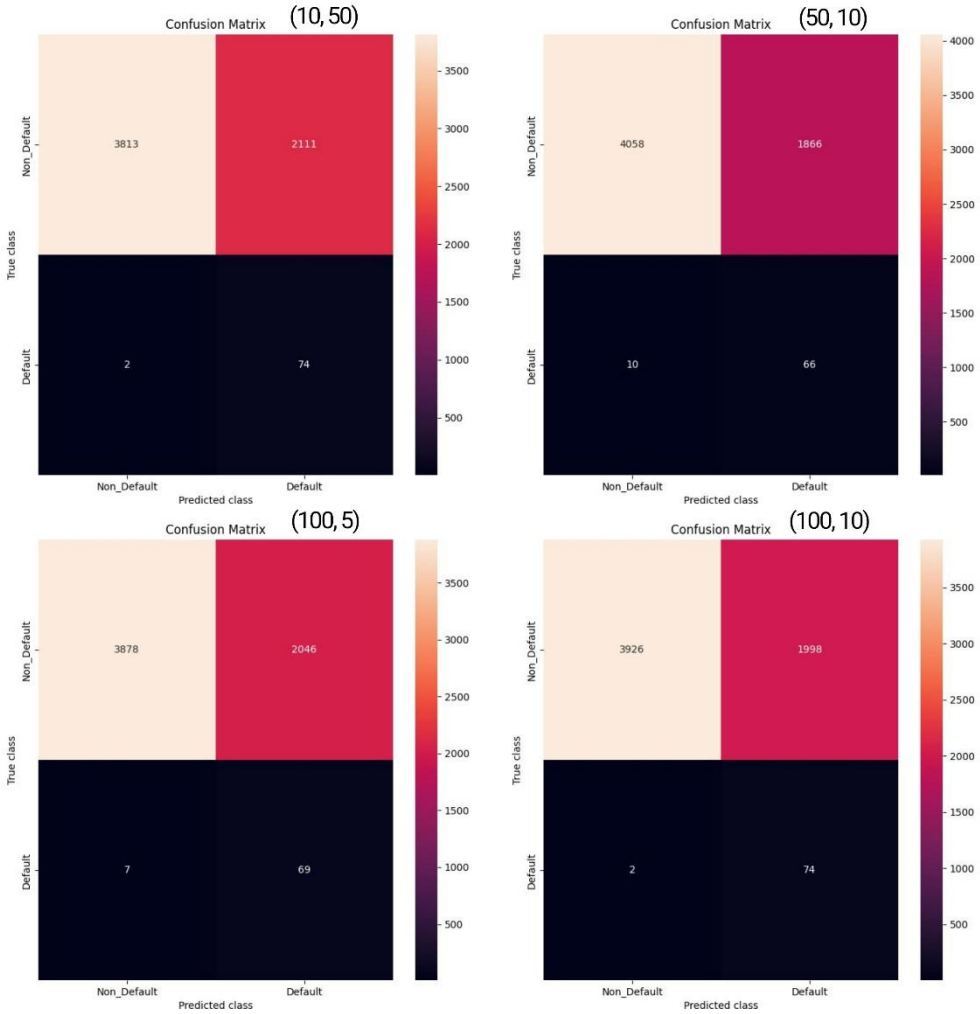
樣本內之型一錯誤分別為 35.63%、**31.50%**、34.54%和 33.73%，以(50, 10)該參數組合型一錯誤最低，(10, 50)最高。樣本外之結果分別為 38.63%、**35.24%**、37.73%和 36.63%。型一錯誤之排序結果樣本外與樣本內之結果相符，顯示 MAML 的訓練結果較具一般性

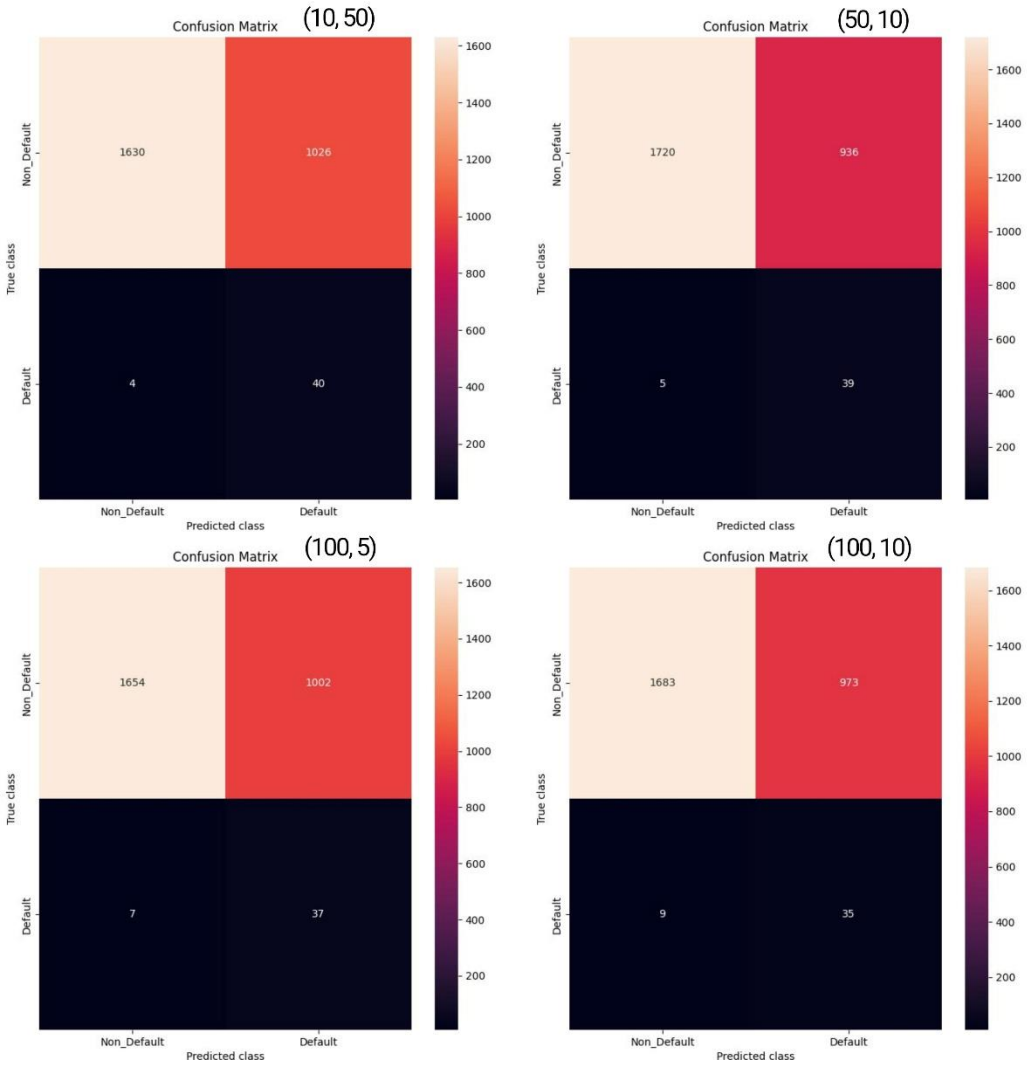
型二錯誤分別為 **2.63%**、13.16%、5.26%和 **2.63%**，以(10, 50)和(100, 10)該參數組合型二錯誤最低，(50, 10)最高。樣本外之結果分別為

9.09%、11.36%、15.91%和 20.45%。型二錯誤之排序結果樣本外與樣本內之結果並非完全相符，但(10, 50)參數組同時在樣本內和樣本外皆為型二誤差最低得參數組，顯示 MAML 的訓練結果有部分一般性。

觀察圖十三可發現 ，在樣本內(100, 10)該參數組合之 AUC 最高，達 81.80%，(50, 10)之 AUC 最低為 77.70%。樣本外則是(50, 10)之 AUC 最高，達 76.70%，(100, 10) 之 AUC 最低為 71.50%。由此結果可觀察出，(100, 10)的訓練結果較其他參數組更配適訓練資料，但(50, 10)參數組所訓練出來的模型在樣本內外的 AUC 較一致，顯示參數組合確實是影響訓練很重要的因素。

圖十二、四組參數組合之混淆矩陣

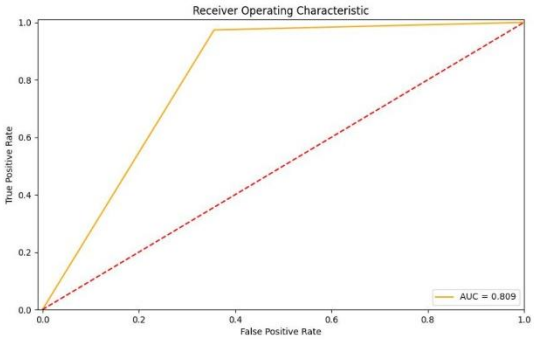




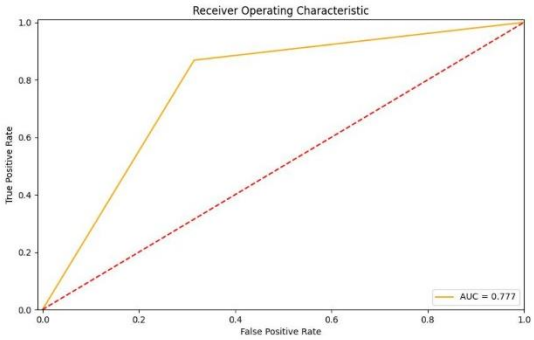
Source:本研究自行整理。

圖十三、四組參數組合之 ROC 和 AUC

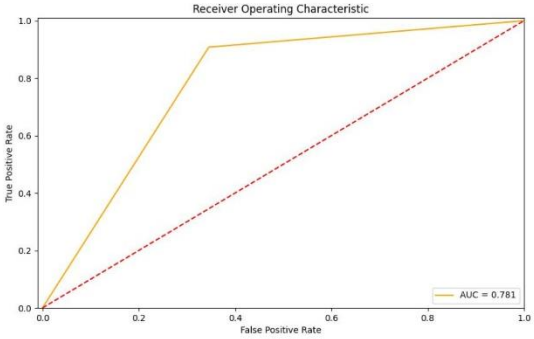
(10, 50)



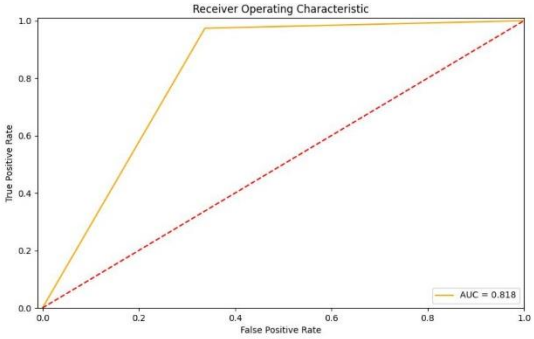
(50, 10)

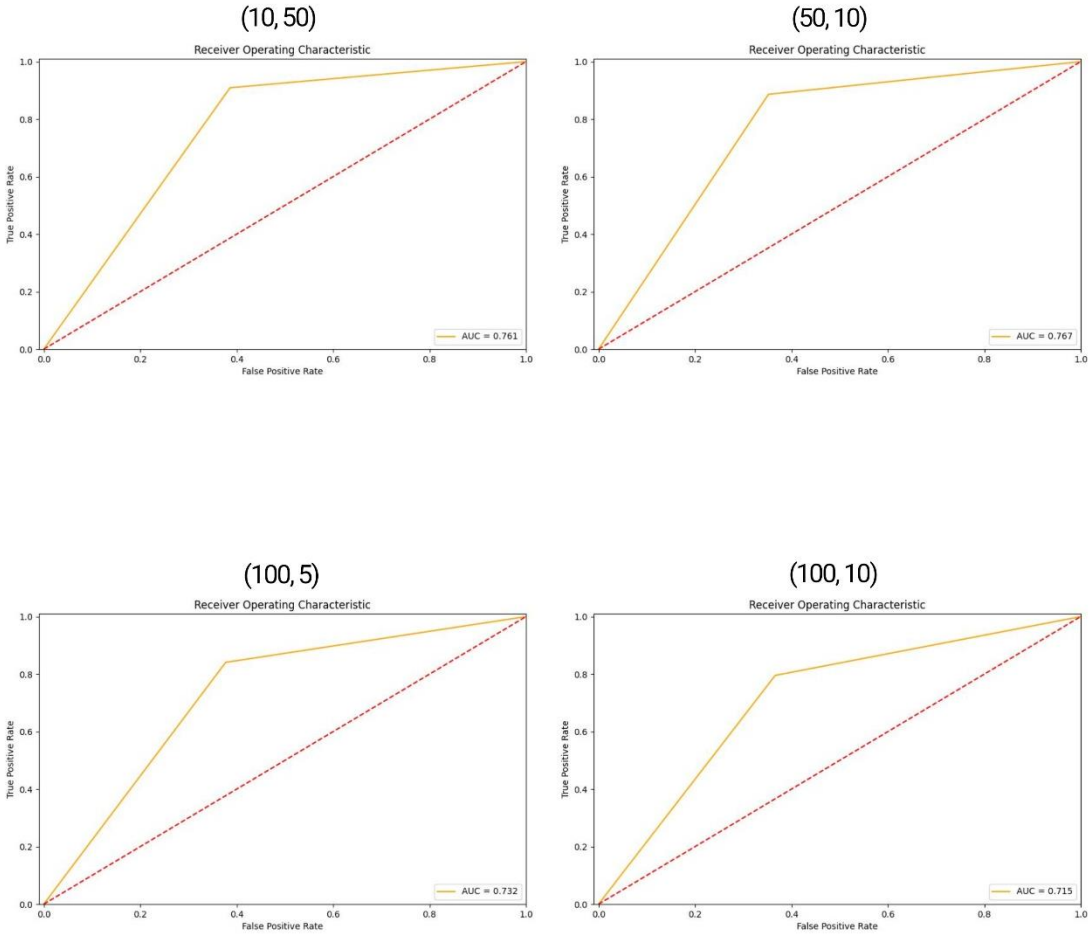


(100, 5)



(100, 10)





Source:本研究自行整理。

五、 結論

整體來說，RF 和 SVM 指標模型有著類似的特色，即對類別 0 之預測能力皆優於類別 1；兩者在型一錯誤皆為 0，以型二錯誤來看，在樣本內 RF(0.7237)較 SVM(0.6974)來的高一些，但兩者在樣本外皆顯示，對於類別 1 毫無預測能力。

比較兩者的 AUC，在樣本內 RF 之 AUC 為 0.862，超出 SVM 的 0.651 許多，也表示 RF 的學習成果較 SVM 為佳；在比較樣本外的結果，卻發現兩指標模型之 AUC 皆為 0.5。樣本內的比較結果比較能說明，Xolani et al(2020)的研究結果屬實，集成式學習模型的預測結果，普遍優於單一模型。

加入 MAML 學習機制比較，其預測類別 0 的準確率雖不及 RF 和 SVM 模型，但觀察三者之型二錯誤，可以發現 MAML 在預測類別 1 上的能力，明顯較任一指標模型突出，甚至在(10, 50)的參數組合上，出現樣本內 2.63%，樣本外 9.09%的型二錯誤。

接著比較三者之樣本外 AUC 可發現，即使 MAML 為 AUC 最小者(0.715)，也比任一指標模型來的高(RF 和 SVM 之樣本外 AUC 皆為 0.5)，亦即 MAML 搭配 FSL 的預測方法，在信用違約的主題上，會優於單純 RF 和 SVM 指標模型。故本研究的結果顯示，MAML 在信用違約這種樣本極端不平均的情況下也適用。

本文最後建議未來研究方向，可考慮加入離群值篩選機制；或是同樣使用 MAML 的架構，但內模型可更換成 CNN 等其他深度學習模型。

六、 參考文獻

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., ... & Zheng, X. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.

2. Wang, Y., Yao, Q., Kwok, J. T., & Ni, L. M. (2020). Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, *53*(3), 1-34.

3. Dastile, X., Celik, T., & Potsane, M. (2020). Statistical and machine learning models in credit scoring: A systematic literature survey. *Applied Soft Computing*, *91*, 106263.

4. Zhu, B., Yang, W., Wang, H., & Yuan, Y. (2018, May). A hybrid deep learning model for consumer credit scoring. In *2018 international conference on artificial intelligence and big data (ICAIBD)* (pp. 205-208). IEEE.

5. Kolanovic, M., A. Lau, T. Lee, and R. Krishnamachari (2017): “Cross asset portfolios of tradable risk premia indices. Hierarchical risk parity: Enhancing returns at target volatility.” White paper, Global Quantitative & Derivatives Strategy. J.P. Morgan, April 26.

6. Raffinot, T. (2017): “Hierarchical clustering based asset allocation.” Journal of Portfolio Management, forthcoming.

7. N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: Synthetic minority over-sampling technique, J. Artificial Intelligence Res. 16 (2002) 321–357.

8. R. Saia, S. Carta, G. Fenu, A Wavelet-Based Data Analysis to Credit Scoring, 2018.

9. Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P. H., & Hospedales, T. M. (2018). Learning to compare: Relation network for few-shot learning. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1199-1208).

10. Finn, C., Abbeel, P., & Levine, S. (2017, July). Model-agnostic meta-learning for fast adaptation of deep networks. In International conference on machine learning (pp. 1126-1135). PMLR.