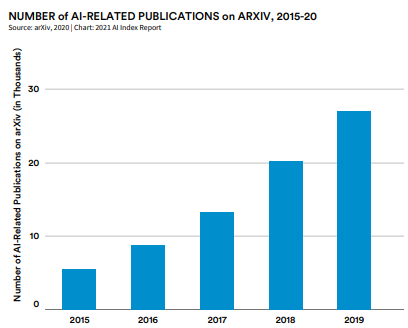
機器學習模型應用於信用違約之預測

1. 前言

硬體的進步與開源程式庫叢立(scikitlearn, tensorflow)，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，使得機器學習和深度學習的應用日漸普及於日常生活。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1、arxiv的AI相關論文出版統計資料



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與資料間線性的關係，而機器學習善於捕捉資料間的非線性關係。如在二分類問題上，常見的統計方法可能會使用羅吉斯回歸來進行預測，但在樣本類別極度不平衡的情況下，羅吉斯回歸會受到資料占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積(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模型解決問題的核心概念雖不同於SVM，但其成效絲毫不弱於SVM，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，訓練出來的權重也會有所差異，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回歸模型。

在金融產業中，機器學習除可應用於投資領域，針對股價、報酬率進行預測、優化資產配置模型等，如Kolanovic et al. [2017], Raffinot [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、等權重配置，同時也發現HRP能有效解決馬可維茲的詛咒(過度配飾樣本內的資料)，解決其再樣本外預測效果很差的問題。

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，本文研究重點在於使用深度學習結合集成式學習的概念，並預期結合後的預測結果會優於集成式模型和其他標竿模型。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)，最後由於巴賽爾委員會(Basel committee)針對信用風險模型，要求需保持模型之透明，為符合Basel之規範，本研究同時加入局部可解釋模型(Local Interpretable Model-Agnostic Explanations, LIME)和Shapley Value來解析深度學習模型。

1. 文獻整理與探討🡪先參考其他文獻如何研究設計和選模型變數

根據Xolani et al(2020)的研究，該回顧文獻匯整了2010至2018年間共74篇關於信用分級模型的預測研究，而該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，而現實世界中，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。像是在影像辨識上常見的調整樣本不平衡的方法SMOTE，和抽樣方法中的過度取樣(Chawla et al.)、低度取樣等方法(Saia et al.)。

**特徵選取/特徵工程**

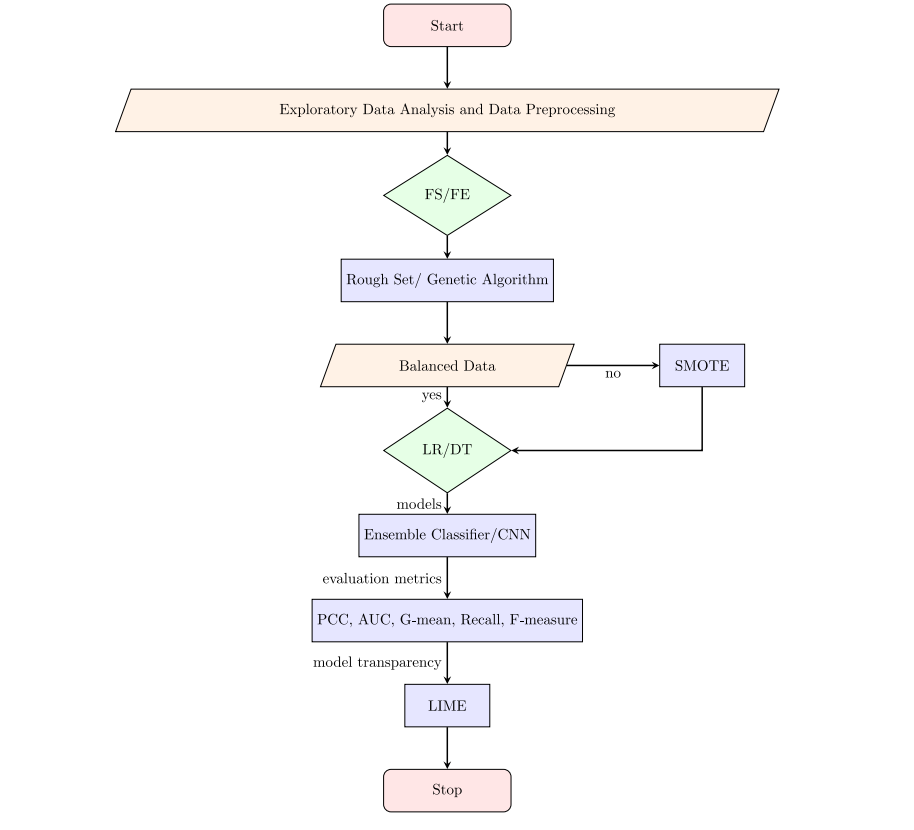
資料選用台灣上市櫃公司，2015年到2020年底的資料，總共160個特徵值，主要為財報、董監事資訊。由於特徵數足夠，本研究將不會進行特徵工程和特徵選取。

**平衡資料**

**模型選用**

**評估模型**

圖二、研究架構圖



1. 模型和變數選用🡪依上章文獻和預期研究結果挑選出適合模型和變數

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115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。選用模型為SVM、RF和MLP模型和加入bagging概念之MLP模型。

1. 實證與模型比較🡪導入資料比較各模型
2. 結論
3. 參考文獻