國立政治大學金融學系

碩士學位論文

集成學習框架下 BERTopic 主題學習之於企業違約預測

Frame of Ensemble Learning Using the Latent Dirichlet Allocation Model and BERTopic for Corporate Default Prediction

指導教授:江彌修 博士

研究生:李宗曄 撰

中 華 民 國 110 年 06 月

DOI:10.6814/NCCU202100708

摘要

因應資產證券化(Asset Securitization)後，擔保債務憑證(Collateralized Debt Obligation,CDO)，信用違約交換(Credit Default Swap,CDS)等與信用風險相關之衍生性商品蓬勃發展，直至 2007 年時因房地產價格下跌導致貸款違約率上升，而被誤判之 CDO 違約率大幅提升，導致 2008 年金融海嘯發生。本研究著重於研究公司違約預測之模型，結合 Duan,Sun and Wang(2012)之財務數據模型加上 Lopatta,Gloger and Jaeschke(2017)運用文字資訊於公司違約模型之方式建立訓練資料及測試資料，而後運用 N Peinelt(2020)之深度學習結合 Blei,Ng and Jordan (2003)之 LDA 主題模型，及機器學習方式對公司財務模型進行預測，並比較傳統 Logit model 和機器學習模型之 Random Forest 和 XGBoost 準確度，其結果也顯示出當加入主題模型之文字資訊時，LDA 模型在各主題參數下其效果較單純運用財務數據之準確度要來得好，而 Bertopic 模型在主題參數少的情況下也有相同效果，而兩文字探勘模型運用機器學習訓練出的公司違約預測模型準確度也較傳統 Logit model 效果要來得好，且當面臨不平衡之樣本資料集時使用 NV Chawla(2002)之 SMOTE 演算法，可透過過採樣之方式，將其違約特徵值樣本放大，並解決違約樣本財務數據不足之問題，其實證結果顯示在加入 SMOTE 演算法後，在各演算法及各主題參數組合的預測中 AUC 分數皆有顯著提升，也顯示在樣本不平衡資料集的訓練中，採用合成資料演算法有其必要性。

關鍵字 ： 公司違約預測、機器學習、主題模型

DOI:10.6814/NCCU202100708

Abstract

Until 2007, due to the fall in real estate prices, the default rate of loans increased, and the default rate of misjudged collateralized debt obligation increased significantly leading to the 2008 financial crisis. This research focuses on the research of the company's default prediction model, combined with the financial data model of Duan, Sun and Wang (2012) plus Lopatta, Gloger and Jaeschke (2017) uses textual information to create training data and test data in the company's default model, then use the deep learning of N Peinelt (2020) combined with the LDA topic model of Blei, Ng and Jordan (2003), and machine learning to predict the company's financial model and compare the accuracy of Random Forest and XGBoost of traditional Logit model and machine learning model.

The results also show that when the text information of the topic model is added, the effect of the LDA model under each topic parameter is better than the accuracy of the pure use of financial data, and the Bertopic model has the same effect when the topic parameters are few. The accuracy of the company default prediction model trained by the two-text exploration model using machine learning is also better than that of the traditional Logit model. When faced with an imbalanced sample data set, the SMOTE algorithm of NV Chawla (2002) can be used. The empirical results show that after adding the SMOTE algorithm, the AUC scores in the prediction of each algorithm and each subject parameter combination are all significant The improvement also shows that in the training of sample imbalanced data sets, it is necessary to use synthetic data algorithms

Keywords: Company default prediction, Machine learning, Topic model

DOI:10.6814/NCCU202100708

目次

DOI:10.6814/NCCU202100708

表次

DOI:10.6814/NCCU202100708

圖次

DOI:10.6814/NCCU202100708

DOI:10.6814/NCCU202100708

第一章 緒論

投資人習慣參考資本市場上的信用評估機構如穆迪(Moody's)、惠譽國際(Fitch Rating)、晨星公司(Morningstar DBRS)、標準普爾(Standard & Poor's)等機構評價系統來評估各公司發行之金融商品違約風險之大小以及預測企業破產之概率。自 1970 年代美國發行之房貸擔保證券(Mortagage-Backed Securities,MBS)開始帶動資產證券化(Asset Securitization)，許多金融機構將信用良好但低流動性之房屋抵押貸款組合透過證券化之方式將其銷售給投資人，但自 1998 年後聯準會開始逐步調降利率，低利環境下使得結構化商品蓬勃發展，擔保債務憑證(Collateralized Debt Obligation,CDO)，信用違約交換(Credit Default Swap,CDS)等產品相繼推出，因金融機構當時對信用評級的誤判將許多 CDO 評為 AAA 等級，讓投資（承銷）機構能易於進行銷售，2007 年時因房地產價格下跌導致貸款違約率上升，而被誤判之 CDO 違約率大幅提升，導致 2008 年金融海嘯發生，因金融海嘯之發生導致多家銀行宣布倒閉及被美國銀行併購，各國股市也因此全面崩盤下跌，並且造成市場流動性壓力遽增，信評公司調降評等房貸抵押擔保證券價格暴跌，景氣衰退貿易成長減緩等巨大的負面影響，也因此自從金融海嘯過後，各金融監管單位也開始注重擔保債務憑證(Collateralized Debt Obligation,CDO)，信用違約交換(Credit Default Swap,CDS)等各式資產證券化(Asset Securitization)之商品設計方式及信用評等評估方法，也開始探討各模型於商品評等預測之準確度，商品價格定值之合理性等，避免金融海嘯再次發生。

而訓練出具預測性之企業違約預測模型，可更精準地評估金融商品的信用評級是否被高估或低估，進而降低投資之信用違約風險和減少不必要之損失，除此之外也可降低金融危機再次發生之概率，本論文著重於優化此預測模型，並探討除了一般財務資訊外加入文字資訊是否能進一步地改善預測模型。

在以往企業違約的研究當中，Black and Scholes(1973)和 Merton(1974)

DOI:10.6814/NCCU202100708

將公司股東的權益當作公司資產，執行價格為公司負債價值的買權，並將公司債券視作無風險債券，加上一個標的物為公司資產且執行價格為公司負債價值之賣權，設立一個動態模型以探討公司價值與破產之關係，而 Altman(1968)提出 Z-Score 模型、Altman, et al.(1977)的 Zeta 模型還有 Ohlson(1980)的 Logit 模型以及 Duan,Sun and Wang(2012)則是利用財務報表中的財務數據來做為關鍵變數。

然而不同於前述提到的研究著重於財務指標之資料，在量化文字情緒近年來的發展中有十分快速的變動，從最開始的網頁數據探勘(Web Data Mining)、文字探勘(Text Mining)再到情感分析(Sentiment Analysis)，為了能將文字量化成有用之因子，主題模型是近期發展的重點之一，主題模型藉由文字探勘的技術，嘗試從文本中萃取有用之信息，用於分析文本潛在之風險主題以利於後續分析，主題模型包括 Deerwester, Dumais, Furnas,Landauer and Harshman(1990)的潛在語意分析(Latent Semantic Analysis, LSA)、Hofmann(1999)的機率潛在語意分析(Probabilistic Latent Semantic Analysis, PLSA)、Biei, Ng and Jordan(2003)的主題模型(Latent Dirichlet Allocation, LDA)以及 Lin and He(2009)結合正負面情感之模型（Joint sentiment/topic model, JST），還有 Peinelt(2020)運用深度學習結合 LDA 之主題模型(Topic Models and BERT Joining Forces for Semantic Similarity Detection, tBERT)，這些經由主題模型萃取出的數據可以用在許多層面上，例如 Lopatta,Gloger and Jaeschke(2017)探討 10-k 財報中的文字資料對於企業違約的影響，證實在企業違約當中，除了財報資訊外，違約公司的文字資訊也會明顯透露出負面情緒及風險字詞。

考量到從 U.S. Securities and Exchange Commission(SEC)網站的 EDGAR System 抓取企業的年度 10-K 報告，並參考 UCLA LoPucki 公司倒閉研究資料庫獲得各年度倒閉公司的參考資料中其符合抓取資料條件之違約公司數相對於整體樣本數較為稀少，因此需使用針對不平衡資料集之演算法來進行處理，

DOI:10.6814/NCCU202100708

Chawla(2002)提出 SMOTE 演算法，利用過採樣方式加重不平衡資料集之權重，使其稀少資料集之特徵能更為突出，而後 Han(2005)提出了 Borderline SMOTE 還有 He(2008)之後又提出了 ADASYN 演算法皆是利用過採樣之方式來對不平衡資料集進行改善，而另外一種模式則是透過欠採樣方式對多數資料集進行權重上之降低，如 Wilson(1972)提出 Edited Nearest Neighbor (ENN)和 Tomek(1976)則提出 TomekLinks 演算法，藉由欠採樣及過採樣之演算法，即使是在樣本稀少之情況下，依然有辦法從稀少樣本中萃取出特徵值並帶入模型使模型依舊保持預測準確度。

本研究將透過 LDA 主題模型及結合深度學習 BERT 處理之 LDA 主題模型量化 10-K 報表之文字訊息，藉由風險字詞在主題出現之頻率，量化成主題風險參數，並結合財務數據，解析企業違約的潛在因素，再透過機器學習之預測方法，提高預測準確率，優化違約預測模型，本研究著重於(1)文本之文字訊息是否能優化預測模型。(2)是否有演算法能處理違約樣本不平衡之問題。(3)機器學習演算法是否能優化預測模型。(4)分析影響主題違約模型的因素。根據實證分析結果，主題模型 LDA 以及結合深度學習處理之主題模型生成之文本參數能改善無主題變數之預測模型績效，使用深度學習的 BERTopic 模型在 XGBoost 和 Random Forest 模型中預測績效與原 LDA 模型的績效相比可發現，BERTopic 在大多數主題參數中 AUC 分數及非違約樣本的預測績效較 LDA 模型好，但在 Logit model 中普遍較差，但若使用 SMOTE 演算法，則不管在各模型中之 AUC 分數、Precision、Recall、F1-Score 都有所提升，而從 10-k 報表的風險字詞當中，也可看出當主題中的風險字詞在文本出現的頻率越高，與當前企業經營的營運狀況也有相關。

本研究之後續安排如下:第二章介紹破產預測模型與主題模型和機器學習模型還有數據處理演算法相關之文獻，第三章介紹研究方法，包括 LDA、BERTopic、XGBoost、Random Forest、SMOTE、羅吉斯迴歸預測模型與衡量模

DOI:10.6814/NCCU202100708

型預測績效之指標如 AUC 分數、Precision 值、Recall 值、F1-Score 值及混淆矩陣，第五章說明資料的蒐集來源與處理，蒐集而來之財務資訊及其樣本分布及相關矩陣還有違約公司樣本及非違約公司樣本之平均值，以及各爬取文本資訊後該如何處理文字資訊例如斷詞(tokens)、詞形還原(Lemmatization)、刪除停用詞(stop words)、TF-IDF 演算法（Term Frequency - Inverse Document Frequency）等使文字資訊可以代入後續之 LDA 主題模型及 BERTopic 模型，還有如何呈現模型之預測績效及文字主題，並透過各分數及混淆矩陣，交叉比對財務資訊及財務資訊加入主題模型萃取各 10-K 文本之風險字詞資訊產生之文本主題參數，比較是否能進一步改善違約樣本及非違約樣本之預測效率，第六章總結研究結果和各模型之比較結論及未來可改進之方向如可其他可參考之財務數據研究方法，不平衡資料集處理方法以及可採用之機器學習演算法，以及是否有其餘因素可改進使得預測準確度更加提升。

DOI:10.6814/NCCU202100708

第二章 文獻探討

第一節 破產預測研究

Merton(1974)依據 Black and Scholes(1973)所建構推導選擇權之公式，將其模型運用至公司債務層面並藉由公式計算公司債價值及公司負債到期時違約之機率，作者假設資產市場為完美市場且無交易成本及稅收，利用公式判斷債務到期時，若資產總額小於負債面額且無法再融資的情況下，則稱作公司到期時債務違約。

Altman(1968)為最早進行公司違約預測研究的作者，其提出 Z-Score 模型，分別採用多項財務數據並將其合併成五項財務比率變數，以多項式迴歸分析並建模出公司違約預測模型，Altman, et al.(1977)提出 Zeta 模型，為 Z-Score 的改良模型，將原本的五項財務比率變數拓展為七項，並作出適當的調整以適用於較為脆弱敏感的零售業，Ohlson(1980)首度利用羅吉斯迴歸模型進行公司違約預測，以 1970 年至 1976 年之 105 家破產公司與 2058 家正常公司為樣本，並採用 9 項財務比率變數來預測財務危機，其研究結果顯示有四項財務比率顯著影響公司違約的機率，包括流動負債、財務結構、公司規模、營運績效，且其模型在前三年預測公司違約準確率皆高於 93%。

Duan,Sun and Wang(2012)以遠期強度模型預測 1999 年至 2011 年美國上市公司未來多期之公司違約機率，除了運用公司規模、現金流量等財務變數外，還加入了三個月美國國庫券利率(treasury rate)、公司違約距離(distance-to-default)、標普 500 指數一年報酬率(S&P500)、一年期異質波動率(idiosyncratic volatility)作為應變數，其結果顯示在短期間內其預測能力相當好，雖然在兩到三年期的預測能力會下降，但其模型依然能捕捉公司違約之模式，無論是三個月美國國庫券利率，亦或是一年期異質波動率皆對公司違約預測有良好的預測能力。

DOI:10.6814/NCCU202100708

Lopatta, Gloger and Jaeschke(2017)除了財務比率變數 Profitability、Liquidity、Size、Cash、Market 外，還使用了 Loughran and McDonald(2011)的情緒字典來對 10-K 報表進行文字情緒分析，作者利用負面字詞出現的頻率數與正面字詞出現的頻率數之差對文本總字數的比率作為衡量財務報表風險程度的參數，研究結果也顯示破產公司之 10-K 報告使用負面字詞的頻率更顯著提升。

第二節 主題模型

在資料探索的研究當中，大多數將文本視為 bag-of-word 的表示法，嘗試使用統計方法擷取文本特徵以建構資料探索的模式，此方法稱為向量空間模型(Vector Space Model，VSM)，但此模型的缺點為不考慮一詞多義(polysemy)和一義多詞(synonymy)的情況，例如使用者在模型中搜尋「airplane」即飛機，傳統向量空間模型只會返回包含「airplane」的文章但不會連帶返回「aircraft」的文章，且其算法之空間維度相當於字典個數的大小，表示當字典字數擴充時，易造成運算上之困難和記憶體之不足，而近年來已有一些文獻被提出以解決向量空間模型高維度的問題，在 1990 年時 Deerwester，Dumais，Furnas，Landauer and Harshman (1990) 提出了潛在語意分析 (Latent Semantic Analysis)，也被稱為潛在語意索引(Latent Semantic Index，LSI)，將文本和字詞表示成矩陣的方法，透過奇異值分解 (Singular Value Decomposition，SVD) 將原本維度過高之矩陣投射到一個低維度的空間，並假設每一個奇異值及其對應的奇異向量(singular vector)代表其潛在主題或概念，以提高精準度。但 LSA 模型由於每一列的值我們無法去推斷說模型矩陣與真實世界之解釋關係，也無法從機率或統計的角度去理解此模型。

因此後續出現了許多機率主題模型，其中機率潛在語意分析(Probabilistic Latent Semantic Analysis，PLSA) 和潛在狄立克雷分布 (Latent Dirichlet Allocation，LDA) 模型被後續學者廣泛使用。PLSA 模型由 Hofmann (1999) 所提出，做法是擷取與文本關聯的意向模型 (Aspect model)，其利用期望最大化

DOI:10.6814/NCCU202100708

算法(EM) 來訓練出隱含的模型參數，但儘管 PLSA 是一個生成模型，其模型卻無法生成新的未知之文件，且隨著文本數量的增加，訓練參數也會隨著線性增加，這就導致無論有多少訓練資料，都容易導致模型的過擬合問題。為了解決這個問題，LDA 模型在 (2003) 由 Blei, Ng and Jordan (2003) 提出了 LDA 模型，其認為每一個文檔與各主題之間都是有關聯性的，且單詞與主題之間亦然，用機率分布去描述「單詞-文本-主題」之間的關係，並克服了 PLSA 模型之缺點使 LDA 模型擁有描述新文章的生成現象，更減輕了過擬合的情形被廣泛運用於資料探索。

但在 LDA 模型中依然存在這無法擷取情緒且在擷取文本與主題的關係時有時會因人為的參數微調呈現不精準的情況，因此後續有相當多的研究者分別針對 LDA 做出了不同的改良及加強，其中 Peinelt(2020)結合了 BERT 與 LDA 主題模型，利用 BERT 能更好地處理上下文之間的關係，並更精準地找出關鍵字詞，可提升 LDA 模型捕捉文本訊息的精度。

第三節 機器學習演算法

由於在預測模型數據中常會出現數據不平衡之例子，因此在有許多學者提出了解決方法，主要分為過採樣和欠採樣方法，過採樣分法主要是用於將少量樣本變多，欠採樣方法則是將多數樣本進行 scale down 使得模型的加權權重改變，降低一點多數樣本的權重重要性。

Chawla(2002)提出 smote 演算法，此研究從少數樣本中選取樣本並按照採樣倍率參數，向量空間中按倍率參數從少數樣本的鄰近空間隨機合成新樣本，使得多數樣本及少數樣本的比率接近平衡，Han(2005)提出了 Borderline SMOTE，其演算法是在 SMOTE 上改善過採樣算法，該算法僅採用向量空間邊界上之少數樣本合成生成新樣本，其算法將少數樣本分成 Safe、Danger、Noise 三類樣本，並僅對 Danger 類的少數樣本進行過採樣，減少產生樣本混淆現象，導致分類效果不佳，He(2008)之後又提出了 ADASYN 演算法，其演算法採用了

DOI:10.6814/NCCU202100708

自適應式算法，對不同的少數樣本賦予不同的權重，從而產生不同數量之合成樣本，進一步改善少數樣本生成情形。

Wilson(1972)提出 Edited Nearest Neighbor (ENN)演算法，為欠採樣演算法，其透過對多數類樣本尋找 K 個鄰近點，如果有一半以上的點都不屬於多數樣本，則將該樣本從訓練樣本中剃除，Tomek(2010)則提出 TomekLinks 演算法，其架構將所有樣本遍歷一次，若少數樣本及多數樣本的距離過大，中間並沒有任何樣本點符合距離條件，則代表此多數樣本點與多數樣本距離過大，被歸類為雜訊樣本必須被刪除。

Quinlan(1986) 提出 Decision Tree，其演算法在判斷過程中將各數據點當作節點並運用熵(Entropy)評斷各節點與分類類別之相似程度，運用樹狀圖的方式判斷後找出各類別所對應之判斷條件，Breiman(2001)提出 Random Forest 演算法，其用隨機的方式建立一個森林，森林由許多決策樹(Decision Tree)組成，每一顆決策樹(Decision Tree)都是沒有關聯的，其隨機採樣的過程降低了過擬合(overfitting)現象的發生，並且其較決策樹(Decision Tree)演算法準確率更加提升，Chen(2016)提出 XGBoost 演算法，為集成學習算法的一種，將許多的弱分類器結合在一起形成一個強分類器，XGBoost 使用 CART 迴歸樹，利用貪婪算法遍歷所有特徵劃分點，並找出各組合於損失函數下之最小值，且運用了許多策略如正則化項、Shrinkage and Column、Subsampling 等防止過擬合(overfitting)的產生，由於 XGBoost 是建立在眾多 Decision Tree 下擬合之預測算法，有別於 Decision Tree 只是單一決策樹決定，XGBoost 不只是多顆決策樹且為共同影響投票決定，也因此 XGBoost 準確度通常較一般決策樹高。

DOI:10.6814/NCCU202100708

第三章 研究方法

LDA

Blei, Ng and Jordan (2003)提出的 LDA 主題模型，是一種典型的詞袋模型(Bag-of-words model)，其假設一篇文檔由多個主題隨機混合生成，且文檔中的潛在主題則由相關字詞所組成，在原 LDA 模型中字詞與字詞間沒有順序關係，而下述文章則會詳述 LDA 模型的演算法及相關術語。

假設 LDA 模型設定有 z 個主題，i 個文檔以及每個文檔分別有不同的字詞數i_j以下是 LDA 生成文本的過程:

1. 從主題的先驗參數 α(α 為 Dirichlet distribution)取樣生成文檔 i 的主題分布 θ_i。
2. 從主題分布θ_i(θ_i為 Multinomial Distribution)取樣生成文檔 i 第 j 個詞的主題 $z_{i,j}$。
3. 從字詞分布 β(β 為 Dirichlet distribution)取樣生成主題$z_{i,j}$的字詞分布$\varphi_{z_{i,j}}$
4. 從字詞的多項式分布$\varphi_{z_{i,j}}$中採樣最終生成詞語$\omega_{i,j}$

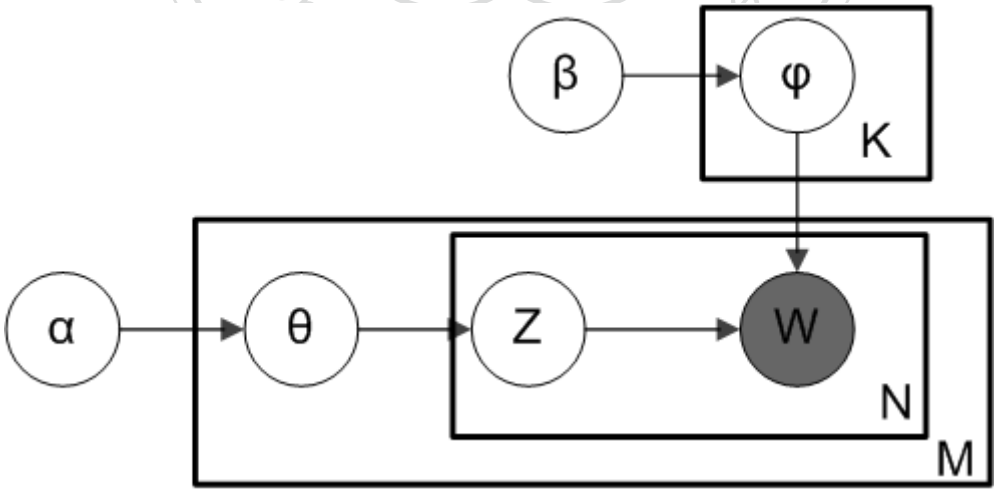

圖 3-1 LDA 主題模型的示意圖

圖片來源 : Blei，Ng and Jordan (2003)

DOI:10.6814/NCCU202100708

在 LDA 模型中由於字詞與字詞之間沒有順序關係，LDA 模型假設字詞由主題生成，而主題在文本中是可無限交換的，De Finetti's theorem 對於一組滿足了可交換性(exchangeability)的隨機樣本，其聯合分配不因隨機變數排列順序變動而改變，且無限可交換的隨機樣本數列等價於某先驗分配抽樣隨機參數，也因此可再抽樣獨立同分布(independent and identically distributed)的隨機變數，故我們可推得給定先驗參數 α 和 β，主題分布機率θ，主題之集合 z，和字詞之集合ω的聯合分配為:

$$P(\theta_i,\ \omega_i,\ z_i,\ \varphi\ |\alpha,\beta) = P(\theta|\alpha)\Pi_{n=1}^{N}P(z_i|\theta)P(\omega_i|\ z_i,\beta) \quad (1)$$

而為了運行 LDA 主題模型，還需要計算潛在變數的後驗分配:

$$P(\theta_i,\ z_i\ |\ \omega_i,\alpha,\beta) = \frac{P(\theta_i,\omega_i,z_i\ |\ \alpha,\beta)}{P(\ \omega_i\ |\ \alpha,\beta)} \quad (2)$$

其中分母項可推導為:

$$P(\ \omega_i|\ \alpha,\beta)=\iint P(\varphi\ |\beta)\ P(\theta|\alpha)(\ \Pi_{n=1}^{N}P(z_i|\theta)P(\omega_i|\ z_i,\ \varphi))d\theta d\beta \quad (3)$$

$$=C\iint\Pi_{k=1}^{K}\Pi_{i=1}^{V}\varphi_{ki}^{\beta_i-1}\Pi_{k=1}^{K}\varphi_{k}^{\alpha k-1}\Pi_{n=1}^{N}\Pi_{k=1}^{K}\Pi_{i=1}^{V}\theta_k\beta_{ki}{}^{\omega_n^i}\ d\theta d\beta \quad (4)$$

$$C=\frac{\gamma(\sum_{k=1}^{K}\alpha_k)}{\Pi_{k=1}^{K}\gamma(\alpha_k)}\ \frac{\gamma(\sum_{i=1}^{V}\beta_i)^K}{\Pi_{i=1}^{V}\gamma(\beta_i))} \quad (5)$$

但因後驗分配的θ和φ有耦合情況，因此 Blei，Ng and Jordan (2003)使用變分推論的方式求出近似於原後驗分配之變分分配，其假設所有隱變量都是透過各自獨立的分布生成，如此便可去掉耦合情況並得到變分分布(variational distribution)

$$q(\varphi_{1:K},z_{1:M},\theta_{1:M}|\ \gamma,\psi,\lambda) =\ \Pi_{k=1}^{K}q(\varphi_k|\lambda_k)\ \Pi_{d=1}^{M}\ q(\theta_d,z_d|\psi_d,\gamma_d) \quad (6)$$

$$=\Pi_{k=1}^{K}q(\varphi_k|\lambda_k)\ \Pi_{d=1}^{M}(q(\theta_d|\gamma_d)\ \Pi_{d=1}^{M}q(\theta_d|\gamma_d)\Pi_{n=1}^{Nd}q(z_{dn}|\psi_{dn})) \quad (7)$$

因變分分布 q 可用來近似真實的後驗機率分布因此我們的最終目標就是最小化兩個分布之間的 KL 距離:

$$(\gamma^*,\ \psi^*,\ \lambda^*) = \frac{argmin}{(\gamma,\psi,\lambda)}D(q(\varphi,z,\theta|\ \gamma,\psi,\lambda)||p(\varphi,z,\theta|w,\alpha,\beta)) \quad (8)$$

DOI:10.6814/NCCU202100708

此距離無法直接計算與優化，因此採用將此最佳化問題等價於極大化對數函數的一個下界，因此我們最佳化對數函數的下界 L，即可解出變分參數，從而得到後驗機率分布的近似。

$$L=E_q[\log p(\varphi|\beta)]+E_q[\log p(z|\theta)]+E_q[\log p(\theta|\alpha)]+E_q[\log p(\omega|z,\beta)]- E_q[\log q(\varphi|\lambda)]-E_q[\log p(z|\psi)]- E_q[\log p(\theta|\gamma)] \quad (9)$$

得到函數下界後透過(E-step)極大化此下界並得出近似分布 q 參數(γ,ψ,λ)再固定變分參數，極大化下界(α,β)(M-step)，在 M-step 步驟中為求解最優的模型參數一般會使用牛頓法對 α,β展開一階和二階導數的逼近，然後迭代求解 α,β在 M 步的最佳解。

而除了透過變分推斷(Variational Inference)和 EM 算法來得到 LDA 模型的文檔分布和主題分布還有字詞分布外，Gibbs sampling 算法也是常用的一種求解方法，Gibbs sampling 利用馬可夫鏈蒙地卡羅(Markov chain Monte Carlo，MCMC)的原理，假設在機率分布 P(x,y)的二維空間中，固定其中一維度 X 坐標的情況下，可以求出其條件機率

$$p(x1,y1)p(y2|x1)=p(x1)p(y1|x1)p(y2|x1) \quad (10)$$

$$p(x1,y2)p(y1|x1)=p(x1)p(y2|x1)p(y1|x1) \quad (11)$$

推得

$$p(x1,y1)p(y2|x1)=p(x1,y2)p(y1|x1) \quad (12)$$

若運用到高維度空間中，即可透過此方法對 LDA 模型後驗分布進行迭代求解，若要求解 p(w|α,β)，則可以在每次迭代中固定所有其他變數，並對單一潛在變數進行抽樣，以此方式得到後驗分布。

DOI:10.6814/NCCU202100708

整個變分流程的演算法如下:

Algorithm: Variational EM Procedure
1 initialize α_k and β_i for all k and i
2 while not covered do:
3 //E-step
4 initialize Φ_{dnk}^0:= 1/K for all d,n and k
5 initialize γ_{dk}^0:= α_k + N_d/ K for all d and k
6 initialize λ_{ki} for all k and i
7 repeat
8 for d=1 to M:
9 for n=1 to N_d:
10 for k=1 to K:
11 Φ_{dnk}^0:= eq.25 according to $\lambda^t$and γ_d^t
12 end for
13 normalize Φ_{dn}^{t+1} to sum to 1
14 end for
15 γ_{dk}^{t+1}:= eq.28 according to Φ_d^{t+1}
16 end for
17 for k=1 to K:
18 for i=1 to V:
19 λ_{ki}^{t+1}:= eq.32 according to Φ^{t+1}
20 end for
21 end for
22 until covergence
23 //M-step
24 update α and βusing Newton-Raphson method with the newly Estimated variational parameters fixed
25 end while

DOI:10.6814/NCCU202100708

BERTopic

BERTopic 參考論文 Top2Vec: Distributed Representations of Topics. (Angelov, D, 2020)，BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding，BERTopic 選擇 BERT 模型來做文字預處理及打印向量判斷文字與文字之間的相似度，再運用 Topic2vec 模型進行各主題中詞彙的分類，BERT 與傳統 Word2vec 打印詞向量不同的要點在於，BERT 模型中的 Contextual word repr.讓同 word type 的 word token 在不同語境下有不同的表達方式，因同詞彙在不同的上下文中所表達的意思也會有所差異，而傳統的詞向量無論上下文，都會讓同 type 的 word token 的 repr.相同，Topic2vec 和傳統的 LDA 模型不同的地方則在於，LDA 模型需要定義已知的主題數量，且 LDA 模型忽略了詞彙於上下文中的順序和語意，也因此先用 BERT 做 fine-tuning 的任務打印詞向量，再利用 Topic2vec 模型，可以比傳統 LDA 模型更精準的抓出主題與詞彙的關係和精準的主題數目。

XGBoost

XGBoost 是基於 Gradient Boosted Decision Tree(GBDT)改良與延伸，其透過增量訓練的方式(additive training)的方式，每一輪訓練中，皆會保留原訓練模型不變並加入新函數至模型中，因此每一輪訓練皆會補足上一顆樹的不足以提升目標函數使損失值為最小。

除了採用增量訓練，XGBoost 還使用劃分點查找算法(split finding algorithms)，此演算法利用貪婪演算法(exact greedy algorithm)列舉所有可能的情形和近似算法(approximate algorithm)將所有分割點的特徵信息映射到相對應的 buckets 當中還有加權分位數算法(weighted quantile sketch)解決貪婪演算法效率低落的問題，最後再使用稀疏分割搜尋法(sparsity-aware Spilt Finding)於特徵分裂生成時避開缺失值提升計算效率。

DOI:10.6814/NCCU202100708

SMOTE

SMOTE 過採樣通過線性插值的方法在兩個少數樣本之間合成新的樣本，避免過擬合(overfitting)的影響，首先從少數樣本中依次選取每個樣本x_i作為合成新樣本之原樣本，其次根據向上採樣倍率 n, 從x_i的同類別的 k(k 一般為奇數, 如 k=5)個鄰近樣本中隨機選擇一個樣本作為合成新樣本的輔助樣本，重複 n 次，然後在樣本x_i與每個輔助樣本間通過公式進行線性插值，創造新樣本之公式如下:

$$x_{new} = x_{chosen} + (x_{nearest} - x_{chosen}) * \zeta ; \zeta \in [0,1] \quad (13)$$

過採樣演算法與欠採樣演算法的差異在於，過採樣演算法用於需要合成少樣樣本之情況，將少數樣本數量增加以平衡分類問題中樣本不平衡之問題，欠採樣演算法則是將多數樣本之權重降低並進行縮小(scale down)，使得模型在訓練時多數樣本之特徵不被過度放大，但無論使用過採樣或欠採樣演算法，若訓練模型之樣本集其樣本特徵不明顯時，就算增加多數樣本之權重或是減少少數樣本之權重，其演算法改善模型訓練之效益也相當有限，此情境在後面實證結果部分也會藉由 BERTopic 演算法的實際分數來顯示出此問題。

。

DOI:10.6814/NCCU202100708

SMOTE 演算法(T,N,K)

輸入: T 為少數類樣本個數，N 為演算法中每個少數樣本需要生成多少合成樣本，K 值則為少數樣本會從 K 個鄰近樣本中隨機抽取並生成合成樣本

輸出: 會有(N/100)*T 個少數樣本合成而成的樣本

演算法過程:

若 N<100%，隨機抽取少樣樣本並隨機化為更少樣本集

若 N=(int)(N/100)，則會代表需生成(T*N)-T 整數個合成樣本

K =需從多少個鄰近樣本中抽取

Numattrs = 有多少分類集

Sample[][] = 原本少數分類集之樣本，將其樣本屬性轉換為陣列

Newindex = 計算已經生成了多少合成樣本，初始係數為 0

Synthetic = 合成樣本，將其屬性轉換為陣列

(並計算每個少數樣本最近之 K 個樣本)

對 T 個樣本計算 K 個鄰近樣本並填入陣列並開始生成合成樣本

計算公式為 $T_{new} = T_{chosen} + (T_{nearest} - T_{chosen}) * \zeta ; \zeta \in [0,1]$

開始生成樣本直到符合樣本集個數符合設定集

DOI:10.6814/NCCU202100708

混淆矩陣

混淆矩陣在統計分類問題和機器學習的分群問題中，可使得模型架構者清楚地知道模型測試集分類的狀況，混淆矩陣的每一列表示一個分類之標籤，而每一行則代表每一個分類的實際預測結果，從混淆矩陣可以檢查每個類別分類的實際效果及情況，並藉此算出 AUC 分數，Precision，Recall，F1-score 等評估指標，並藉此判斷是否應該更換模型或標準化處理方式甚至增減樣本集。

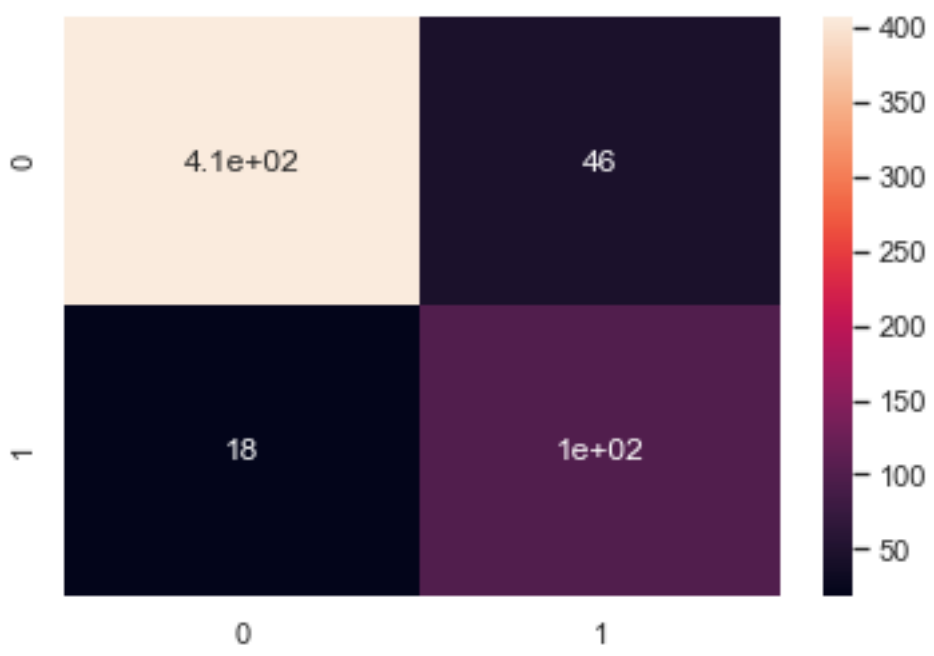

圖 3-2 混淆矩陣示意圖

DOI:10.6814/NCCU202100708

AUC

AUC 是 ROC 曲線中其曲線下方面積，ROC 曲線在兩分類問題中，其 Y 軸代表的是將正樣本預測為正的機率，X 軸代表的是將負樣本預測為正的機率，因此我們可知道當曲線越靠近左上方時，其模型的預測效果越好，而 ROC 曲線下方面積越大至少代表其正樣本預測準確度越高，因此 AUC 分數越高，代表模型預測能力越好，若 AUC 分數>0.5 時，則表示模型預測能力優於隨機預測。

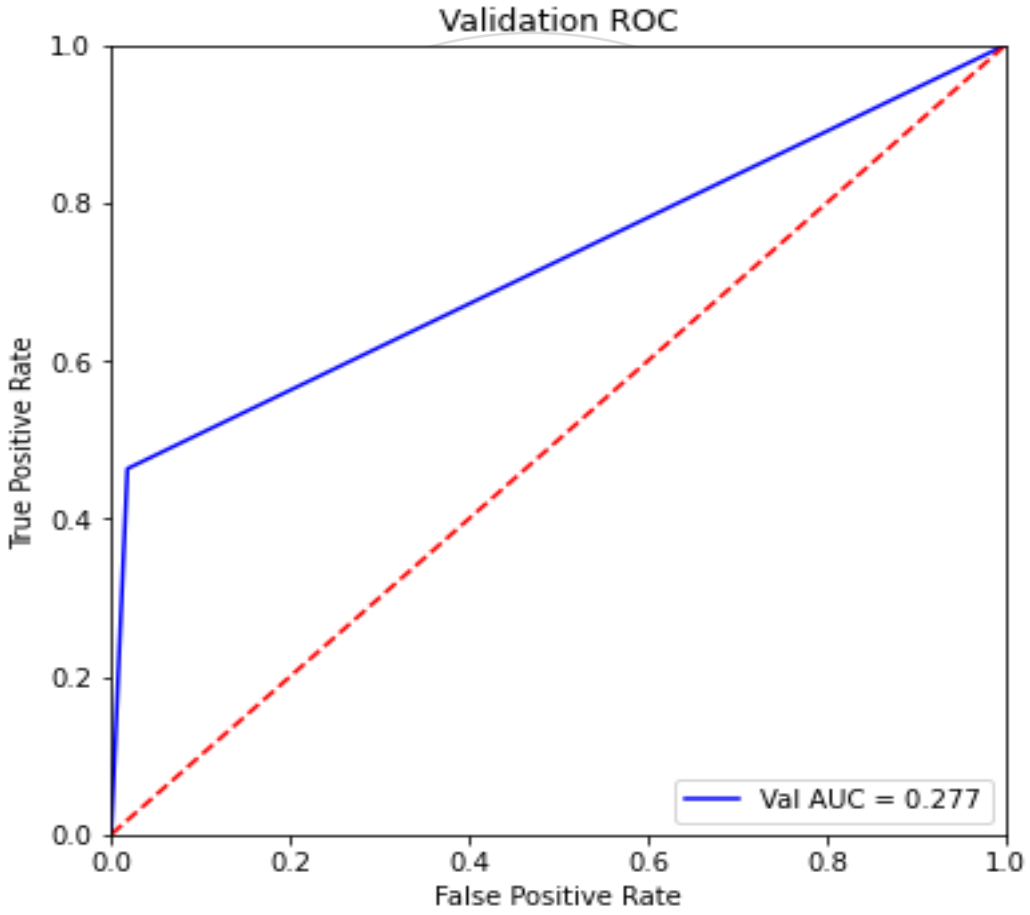

圖 2-3　AUC 曲線圖

精確度

精確度(Precision)代表當模型將測試集中的樣本識別為正樣本，並且其樣本真的為正樣本之機率，若 Precision 越高，則表示模型預測正樣本之正確率越高，其公式為:Precision = TP/(TP+FP)

DOI:10.6814/NCCU202100708

召回率

召回率(Recall)表示在所有測試集的正樣本中，被模型正確判定為正樣本之機率，通常提高精確率會降低召回率值，因此必須在各模型中找出最適的精確度以及召回率，其公式為:Recakk = TP/(TP+FN)

F1-Score

F1-Score 為精確率及召回率之綜合評價指標，其計算公式為:F1-Score = (Recall + Precision) / 2，因在特殊極端情形下，無論是 Recall 值或是 Precision 其中一個指標特別高，另一個特別低的情況下，此模型都是具備妥善辨識能力的模型，因此設計出了 F1-Score 來觀察兩者指標的綜合分數

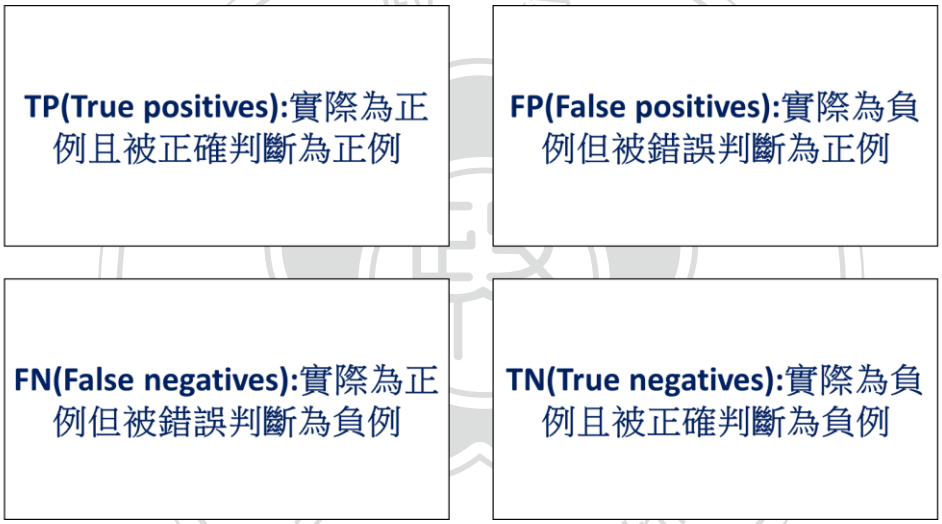

圖 3-3 準確度解釋圖

```
Report :              precision    recall  f1-score   support

           0       0.92      0.96      0.94       425
           1       0.87      0.76      0.81       147

    accuracy                           0.91       572
   macro avg       0.89      0.86      0.88       572
weighted avg       0.91      0.91      0.91       572

Report : 0.8609523809523809
```

圖 3-4 示意圖預測結果圖

DOI:10.6814/NCCU202100708

文字探勘文本處理技巧

Lemmatization

詞形還原(Lemmatization)是指將各文本裡句子裡的單字在不同詞性下還原成相同字根之處理技巧，例如將 “drove“還原成“drive“，將“driving“還原成“drive“，還原後可以讓後續無論是停用字刪除(stop words)或是詞彙出現頻率統計以及 LDA 模型的建立都變得更加有效率且直觀。

停用字刪除(stop words)

停用字刪除主要是刪除例如像 “he“、“she“、“it“等詞彙在整體主題模型中對主題呈現並無影響之字詞，或是參考論文或研究單位提供之辭典，更進一步地刪除非必要之單詞，在 Python 的 NLTK 或是 spaCy 套件庫中皆有提供停用字刪除之詞庫。

TF-IDF

跟傳統的數詞頻率不同，傳統數詞頻率並無針對各自文本總字數去做統計，而是純粹統計各單詞出現的總頻率，TF-IDF 在資料檢索中採取的是一種加權的計算方式，在加權方法中，如果是重要的字詞，則字詞在各文本隨著總字詞增多其出現頻率也會增加，並不會只是在單一文本中的頻率特別高，其保留重要的詞語之加權公式為:

$$tf_{i,j} = \frac{n_{i,j}}{\sum k n_{k,j}}$$

$n_{i,j}$是該詞在檔案d_j中的出現次數，而分母則是在檔案d_j中所有字詞的出現次數之和。

逆向檔案頻率(inverse document frequency, idf)是一個探測字詞在總文本重要性的衡量標準，某一特定詞語的 idf，可以由總檔案數目除上包含該詞語之檔案的數目，再將其商取以 10 為底之對數得:

$$idf_i = \ln \frac{|D|}{|\{j: t_i \in d_j\}|}$$

$|D|$:語料庫中的檔案總數，$|\{j: t_i \in d_j\}|$:包含詞語t_i的檔案數目(即$n_{i,j}$不為 0 的檔案數目)，若詞語不在資料中，則導致分母為零，因此一般情況下使用 $1+|\{j: t_i \in d_j\}|$最終得到 $\text{tfidf}_{i,j} = tf_{i,j} \times idf_i$，此公式可較為妥善地讓使用者知道字詞在文本分布之情形及決定是否要更進一步刪除字詞。

DOI:10.6814/NCCU202100708

第四章 實證結果

第一節資料來源與處理

本研究使用 U.S. Securities and Exchange Commission(SEC)網站的 EDGAR System 抓取企業的年度 10-K 報告，並參考 UCLA LoPucki 公司倒閉研究資料庫獲得各年度倒閉公司的參考資料，再透過 Compustat 資料庫中搜尋得到本章第二節提及之財務資料，最終整理成 1999 年至 2018 年間 476 筆倒閉公司資料及 5462 筆非倒閉公司資料，共計 5938 筆 10-K 報告以及其財務資訊。

在生成模型所需之主題參數前，本研究對 10-K 報告之文字資料進行資料預處理，首先刪除文本中的標點符號，數字及特殊符號，接著進行斷詞(tokens)並進行進行詞形還原(Lemmatization)，還原詞形後進行刪除停用詞(stop words)，本研究使用 Loughran and McDonald 所提供各類別的停用詞字詞庫做為本研究 10-K 文本字詞刪減的停用詞，刪除完後利用 TF-IDF 演算法（Term Frequency - Inverse Document Frequency）對剩餘進行刪除停用詞完後字詞進一步分析，再度刪除不必要之字詞。

```
'compound': 9.932517730102835e-05,
'rate': 4.390273272021182e-05,
'mid s': 0.000296594809591732,
'experience': 6.821501826331049e-06,
'significant': 1.9809758281493558e-05,
'change': 6.833510219088956e-06,
'distribute': 0.00012969553052385907,
'emergence': 0.00018182877942408786,
'national': 0.00013698703792541688,
'superstores': 0.005273662929554386,
'contract': 2.7600565450444277e-05,
'stationer': 0.006513797075547116,
'mass': 0.0016858448960313912,
'merchandiser': 0.0015693152829959526,
'consolidation': 6.006745948917755e-05,
'end user': 0.00017460296175218805,
'warehouse': 0.0006555325736427353,
'club': 0.001125541729243513,
'commercial': 2.616525541929285e-05,
```

圖 4-1 (TF-IDF)

DOI:10.6814/NCCU202100708

本研究使用之財務數據方面，Lopatta, Gloger and Jaeschke (2017)文章中提到之五項控制變數 Profitability，Liquidity，Size，Cash 及 Market 在文中證明可有效進行破產公司與非破產公司之分類，因此本論文也使用此五項控制變數來建構破產預測模型，變數資料主要來自 Compustat 資料庫中 1999 年至 2018 年之年度財務數據，以下為此五項控制變數之定義:

Profitability : 稅前息前獲利(earnings before interest and tax, EBIT)除以總資產(total assets)

Liquidity : 流動比率(current ratio)，其公式為流動資產(current assets)除以流動負債(current liabilities)

Size : 權益市值(market value of equity)取自然對數

Cash : 現金(cash)與其他短期投資(short-term investment)總和除以總資產

Market : 市價淨值比(market-to-book ratio)

利用此五項控制變數加入主題變數並代入 XGBoost 和 logit 還有 Randomforest 模型進行公司違約預測之判斷並比較。

表 4-1 樣本控制變數平均值

種類	Profitability	Liquidity	Size	Cash	Market
違約樣本	-0.065	1.469	4.336	0.0845	1.007
非違約樣本	-0.035	2.559	5.218	0.2187	2.383

DOI:10.6814/NCCU202100708

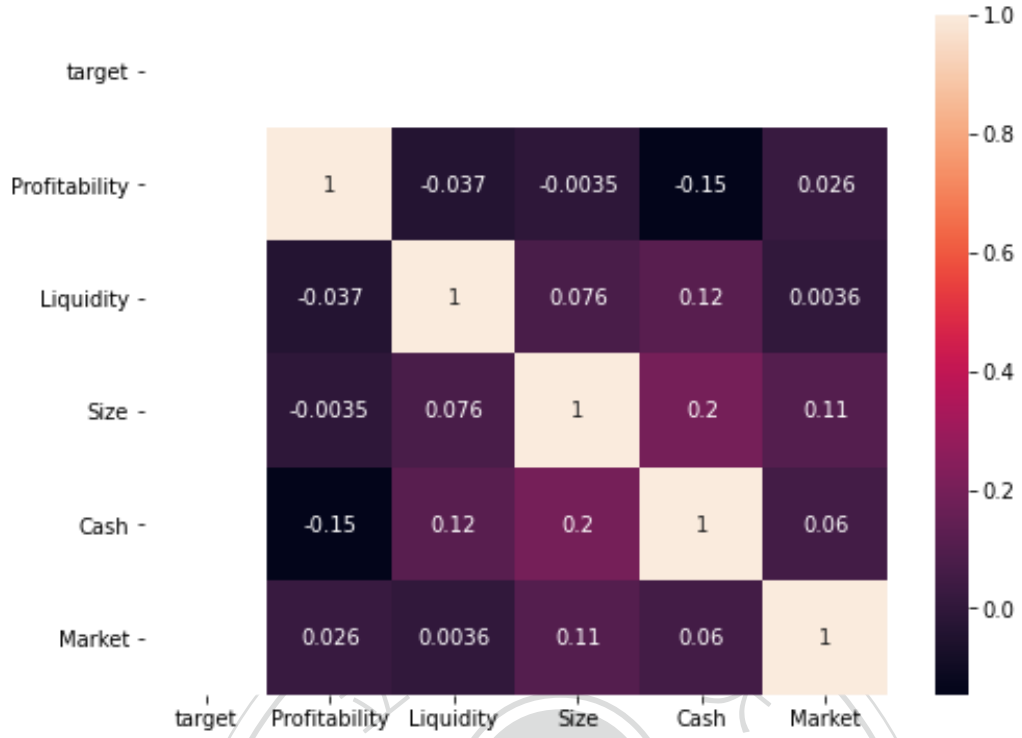

圖 4-2 控制變數相關矩陣

最終共取得樣本數為 5938 筆，共有未違約筆數 5462 筆，違約筆數 476 筆，從 1999 年至 2018 年，為了準確評斷出模型預測的能力，因此本研究將樣本用不同的切割方式，而為了解決樣本不平衡來進行違約公司的預測

預測違約方式採下列四種方式:

(1) 違約筆數 476 筆，未違約筆數 5462 筆

(2) 違約筆數 476 筆，未違約筆數 5462 筆，並採用 SMOTE 演算法

(3) 違約筆數 476 筆，未違約筆數 5462 筆，並加入主題模型參數

(4) 違約筆數 476 筆，未違約筆數 5462 筆，並採用 SMOTE 演算法和主題模型

訓練集和測試集比例為 7:3，並使用 Logit model，Random Forest，XGBoost 演算法來預測公司違約的資料集，以混淆矩陣，AUC 曲線，精確度，召回率，F1 值，來顯示預測模型的好壞。

DOI:10.6814/NCCU202100708

第二節 結果呈現

本研究將資料集依前頁所呈述之四種方式，進行訓練及測試，從下方圖表中可以看到在還沒加入主題模型，純粹使用財務數據之模型 AUC 曲線分數、precision、recall、F1-score。

從下方圖表中可看到財務數據分別使用 Logit model、Random Forest、XGBoost 模型訓練之 AUC 曲線、precision、recall、F1-score 之分數表格。

表 4-2 原始財務數據模型準確度

模型	Precision	Recall	F1-Score	AUC 分數
Logit model	0.46	0.5	0.48	0.5
Random Forest	0.79	0.69	0.72	0.6869
XGBoost	0.85	0.73	0.77	0.7259

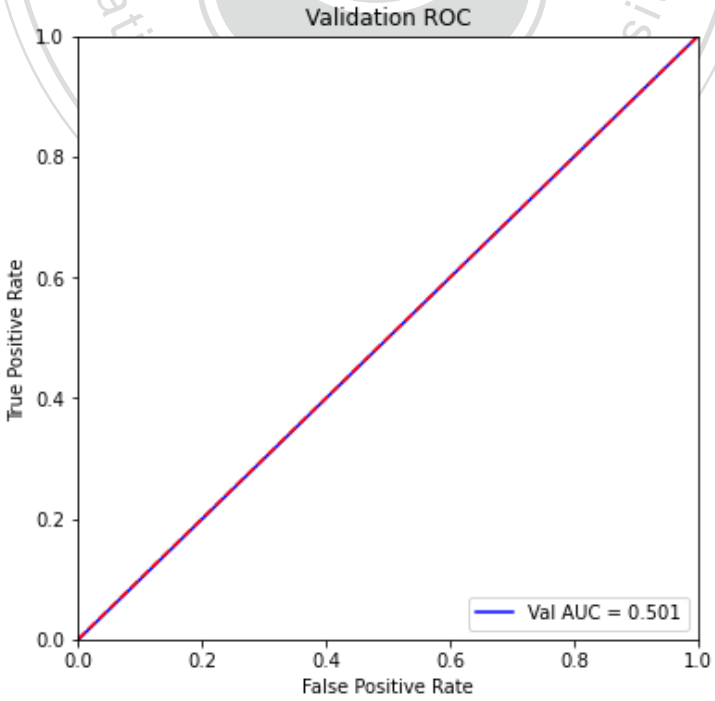

圖 4-3 Logit model 不同門檻值準確度

DOI:10.6814/NCCU202100708

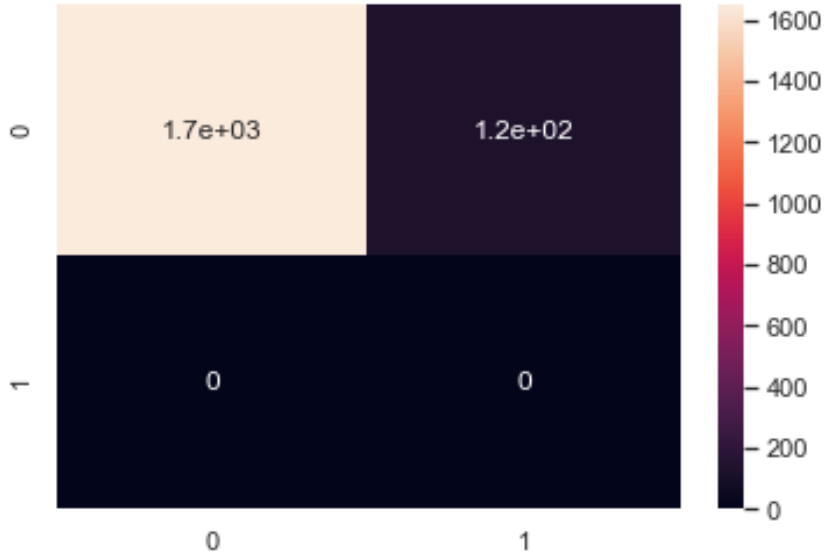

圖 4-4　Logit model 混淆矩陣

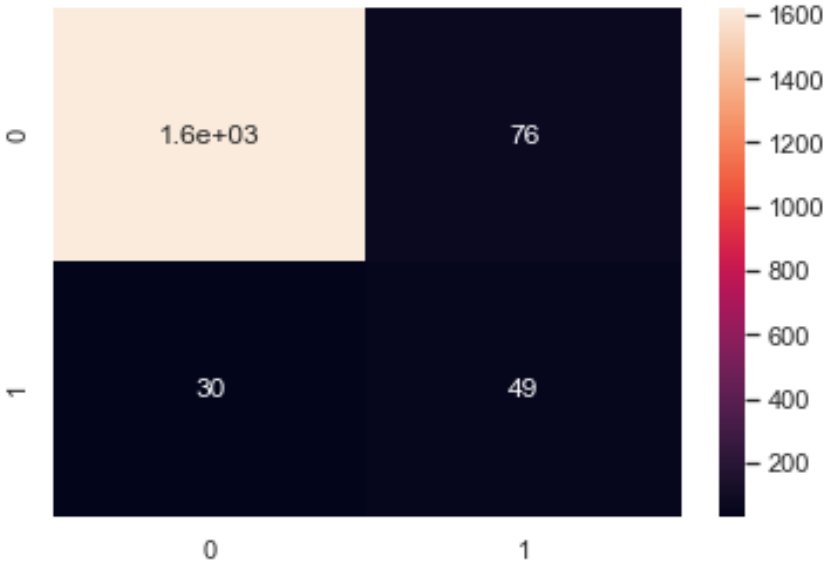

圖 4-5　Random Forest 混淆矩陣

DOI:10.6814/NCCU202100708

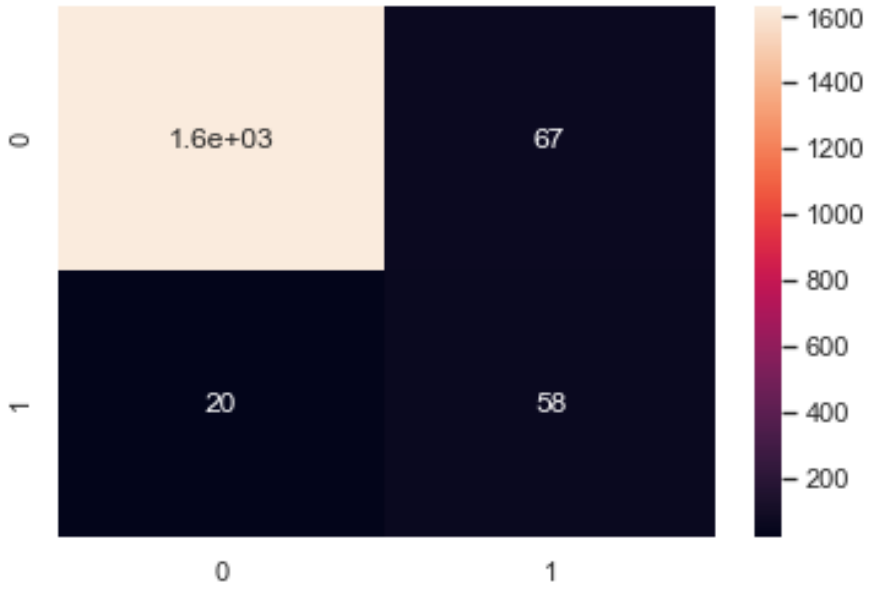

圖 4-6　XGBoost 混淆矩陣

表 4-3 原始財務數據加入 SMOTE 演算法後模型準確度

模型	Precision	Recall	F1-Score	AUC 分數
Logit model	0.54	0.64	0.46	0.6398
Random Forest	0.69	0.78	0.72	0.7787
XGBoost	0.71	0.83	0.75	0.8261

從各圖表中可看出，使用 XGBoost 演算法和 Random Forest 在使用 SMOTE 演算法的情況下，其 AUC、Recall、F1-Score 分數皆高於未使用，但 Precision 分數會微下降，代表將違約樣本標記為非違約樣本的能力下降、找到所有非違約樣本以及模型整體效能上升。

而在 Logit model 中，各項數值皆有明顯上升，在使用 SMOTE 模型之前，其 AUC 分數只有 0.5，代表此模型幾乎無預測能力，表示 SMOTE 演算法能初步改善 Logit model 預測能力。

DOI:10.6814/NCCU202100708

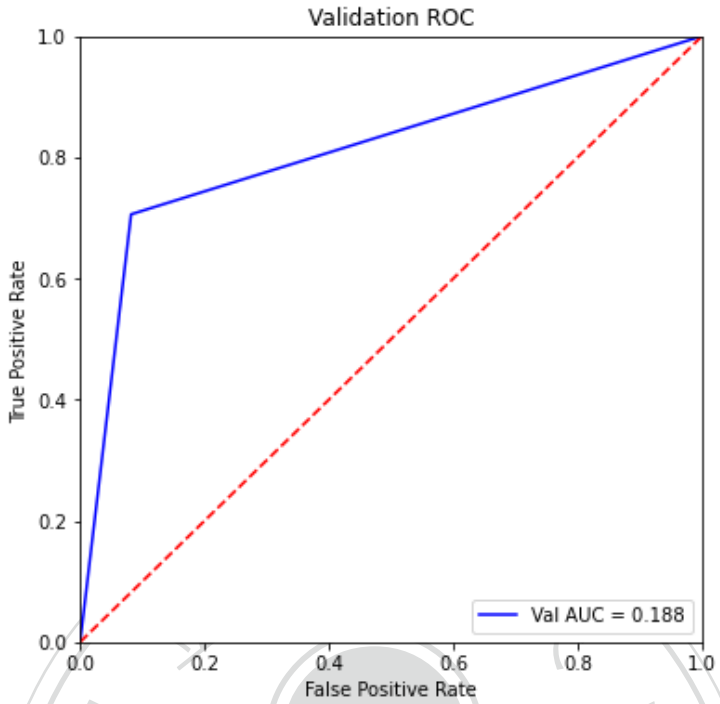

圖 4-7　XGBoost 採用 SMOTE 演算法不同門檻值準確度

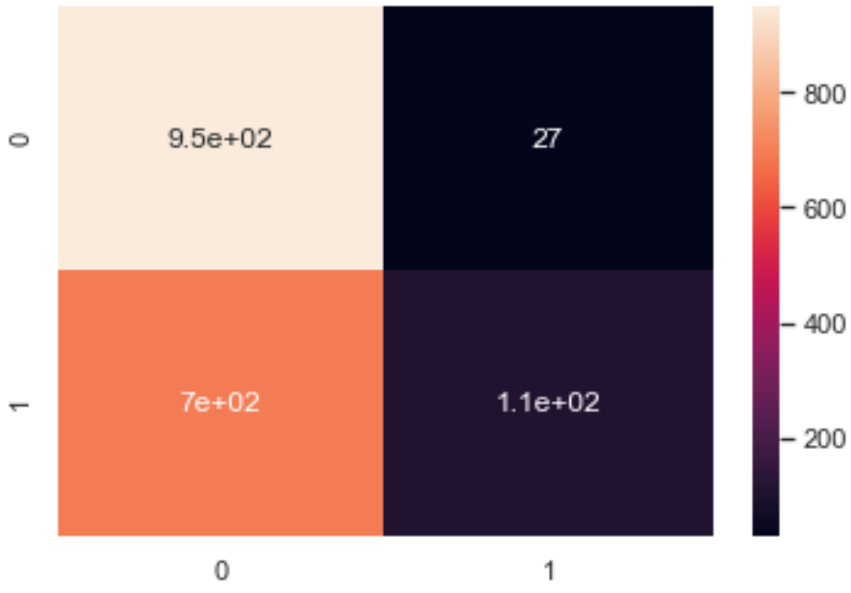

圖 4-8　Logit model 混淆矩陣採用 SMOTE 演算法

DOI:10.6814/NCCU202100708

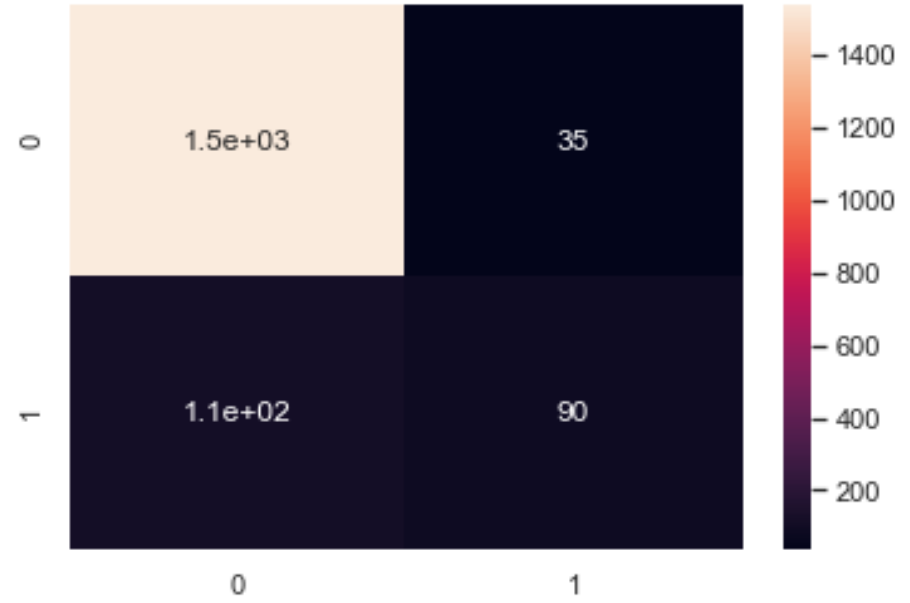

圖 4-9　Random Forest 混淆矩陣採用 SMOTE 演算法

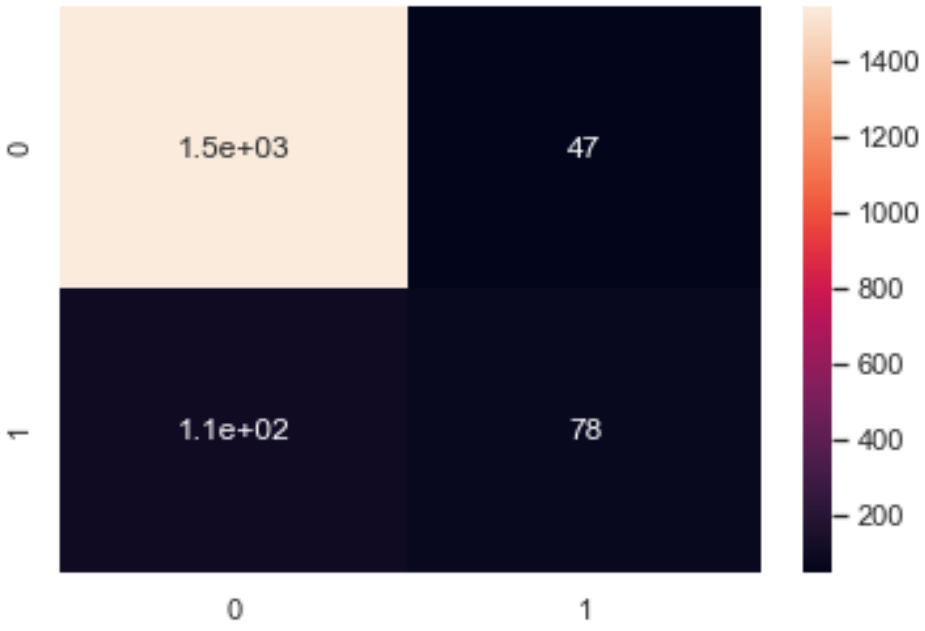

圖 4-10　XGBoost 混淆矩陣採用 SMOTE 演算法

DOI:10.6814/NCCU202100708

而在後續研究中，使用了 LDA 主題模型來捕捉 10-k 文本之文字資訊，從字雲的圖片中我們可以看到在 10-k 文本中最常出現之字詞為何，也可以用來判斷進一步需要刪除之字詞為何。

圖 4-11　LDA 模型 字雲

本研究中，LDA 的主題模型中的主題參數，將其設為 3、5、10、20、30 個，以這 5 個主題參數來觀察各文本在各主題中所佔之比重，並將各主題參數之結果加上財務數據代入訓練模型。

property	interest	lease	tax	expense
development	agreement	clinical	drug	research
service	revenue	cost	market	asset
loan	bank	security	loss	deposit
technology	system	end	result	include

圖 4-12　LDA 捕捉到主題詞彙

DOI:10.6814/NCCU202100708

從下方圖中可看到各主題參數之數據加入財務數據代入後在 Logit model 方面在 topics 為 3、20、30 時 AUC 分數會有些微提升，而 topics3 之所以提升分數跟 Logit 本身對特徵抓取較為敏感，過多的細微參數可能會使 Logit model 無法判斷違約樣本及非違約樣本，而 Random Forest 和 XGBoost 模型中，無論是 topics3、5、10、20、30，其 AUC 分數皆高於原始模型，而除了 topics3 以外其餘參數模型之 Precision、Recall、Score、AUC 分數皆高於原始模型。

表 4-4 原始財務數據加入 LDA 主題模型(topic 為主題數)模型準確度

模型	Precision	Recall	F1-Score	AUC 分數
Logit model(3)	0.54	0.64	0.46	0.6398
Random Forest(3)	0.69	0.78	0.72	0.7787
Xgboost(3)	0.71	0.83	0.75	0.8261
Logit model(5)	0.46	0.50	0.48	0.5000
Random Forest(5)	0.86	0.72	0.77	0.7189
Xgboost(5)	0.85	0.75	0.79	0.7453
Logit model(10)	0.46	0.50	0.48	0.5000
Random Forest(10)	0.89	0.71	0.77	0.7127
Xgboost(10)	0.88	0.80	0.84	0.8019
Logit model(20)	0.63	0.50	0.49	0.5034
Random Forest(20)	0.91	0.70	0.77	0.7016
Xgboost(20)	0.86	0.76	0.80	0.7610
Logit model(30)	0.67	0.52	0.53	0.5250
Random Forest(30)	0.92	0.71	0.77	0.7099
Xgboost(30)	0.86	0.76	0.8	0.761

DOI:10.6814/NCCU202100708

而加入 SMOTE 演算法後可以發現各主題參數下使用 SMOTE 演算法，其 AUC 分數皆有相當大程度的提升，尤其是 Logit model 上升幅度最大，而 Recall、F1-Score 也有不同程度之提升，但 Precision 值呈現些微下降之狀態。

表 4-5 原始財務數據加入 LDA 主題模型和 SMOTE 演算法模型準確度（topic 為主題數）

模型	Precision	Recall	F1-Score	AUC 分數
Logit model(3)	0.56	0.71	0.52	0.7105
Random Forest(3)	0.73	0.80	0.76	0.7974
Xgboost(3)	0.73	0.83	0.77	0.8319
Logit model(5)	0.56	0.71	0.77	0.7082
Random Forest(5)	0.77	0.81	0.79	0.8089
Xgboost(5)	0.79	0.84	0.81	0.8375
Logit model(10)	0.57	0.75	0.55	0.7490
Random Forest(10)	0.80	0.83	0.81	0.8280
Xgboost(10)	0.82	0.87	0.84	0.8683
Logit model(20)	0.56	0.72	0.54	0.7191
Random Forest(20)	0.83	0.81	0.82	0.8051
Xgboost(20)	0.82	0.83	0.82	0.8267
Logit model(30)	0.57	0.73	0.55	0.7321
Random Forest(30)	0.85	0.82	0.84	0.8229
Xgboost(30)	0.85	0.83	0.84	0.8343

DOI:10.6814/NCCU202100708

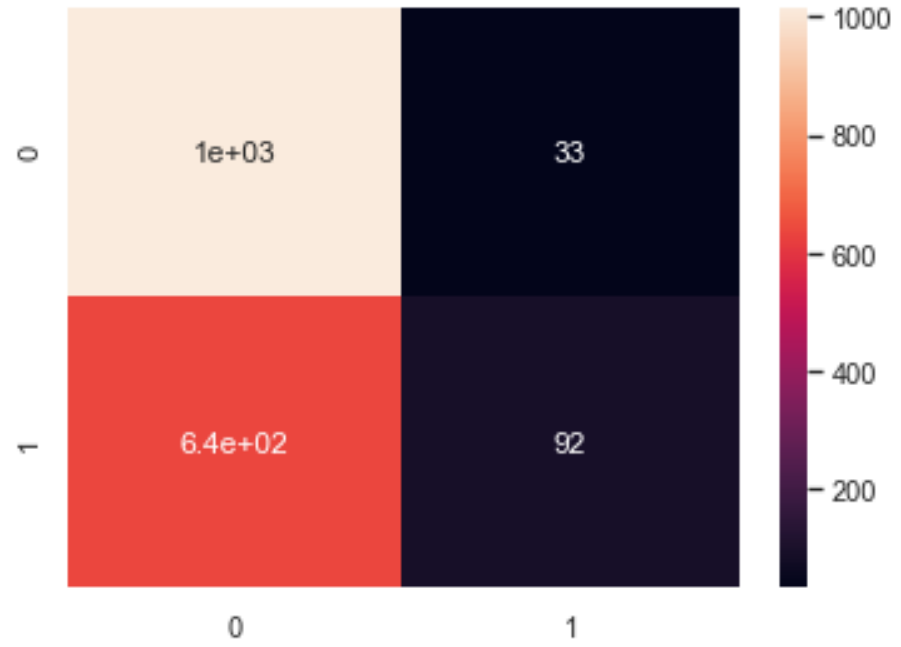

圖 4-13　Logit model 加入 SMOTE 演算法和主題參數混淆矩陣

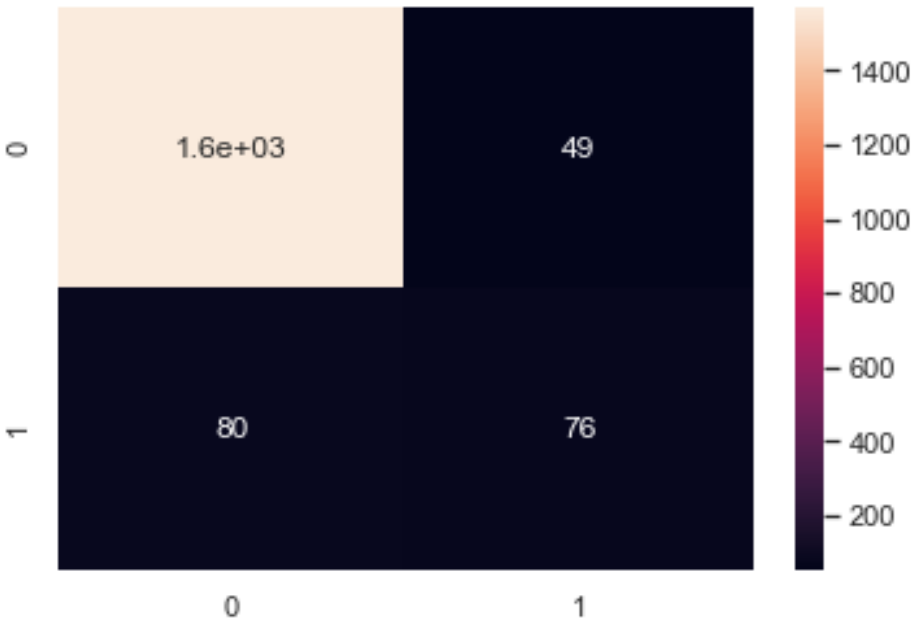

圖 4-14　Random Forest 加入 SMOTE 演算法和主題參數混淆矩陣

DOI:10.6814/NCCU202100708

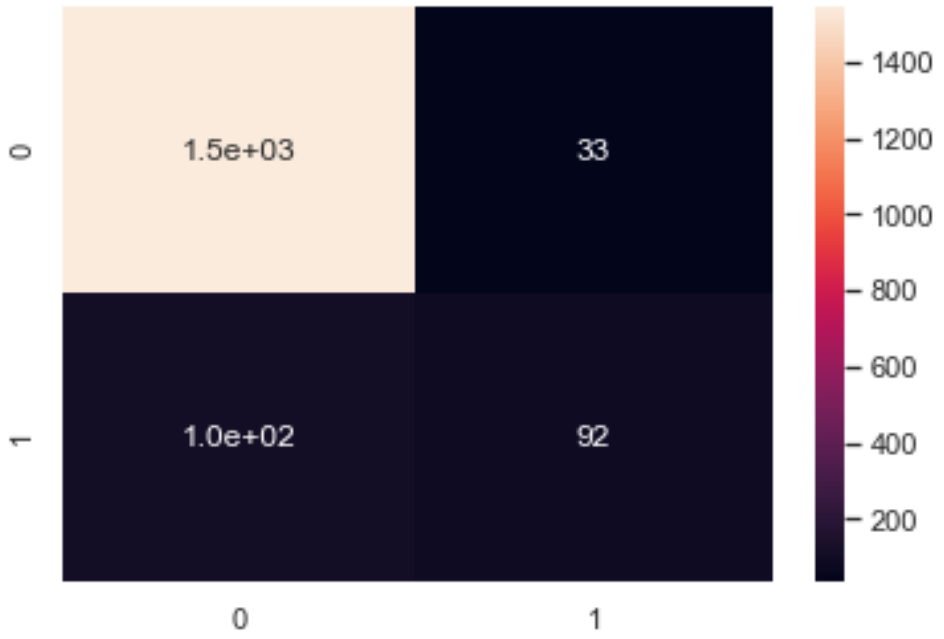

圖 4-15　XGBoost 加入 SMOTE 演算法和主題參數混淆矩陣

而加入深度學習 bert 處理後的主題模型，可發現在 XGBoost 模型中，除了主題參數 10 外的準確度皆較 LDA 模型來得好，但可發現雖然其預測非違約樣本的準確度極高，預測違約樣本的準確度相對較低，而在 Logit 和 Random Forest 模型中，主題參數 10、20、30 其準確度則都較 LDA 模型低，而主題參數 3、5 較 LDA 模型高。

DOI:10.6814/NCCU202100708

表 4-6 原始財務數據加入 BERTopic 主題模型模型準確度

(topic 為主題數)

模型	Precision	Recall	F1-Score	AUC 分數
Logit model(3)	0.46	0.5	0.48	0.5
Random Forest(3)	0.89	0.74	0.78	0.7536
Xgboost(3)	0.84	0.78	0.81	0.7802
Logit model(5)	0.46	0.5	0.48	0.5
Random Forest(5)	0.87	0.71	0.77	0.7145
Xgboost(5)	0.87	0.78	0.82	0.7825
Logit model(10)	0.96	0.5	0.49	0.5038
Random Forest(10)	0.86	0.69	0.75	0.6930
Xgboost(10)	0.84	0.78	0.81	0.7789
Logit model(20)	0.96	0.5	0.49	0.5038
Random Forest(20)	0.87	0.69	0.72	0.6896
Xgboost(20)	0.83	0.79	0.82	0.7876
Logit model(30)	0.72	0.5	0.48	0.5037
Random Forest(30)	0.92	0.66	0.78	0.6599
Xgboost(30)	0.85	0.8	0.78	0.7597

DOI:10.6814/NCCU202100708

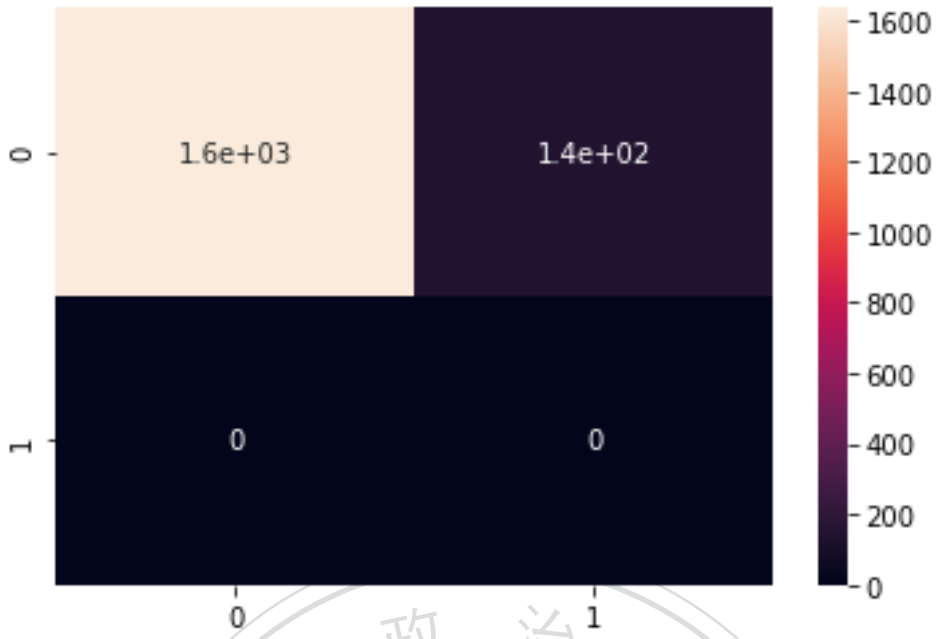

圖 4-16　Logit model 加入 Bertopic 主題參數之混淆矩陣

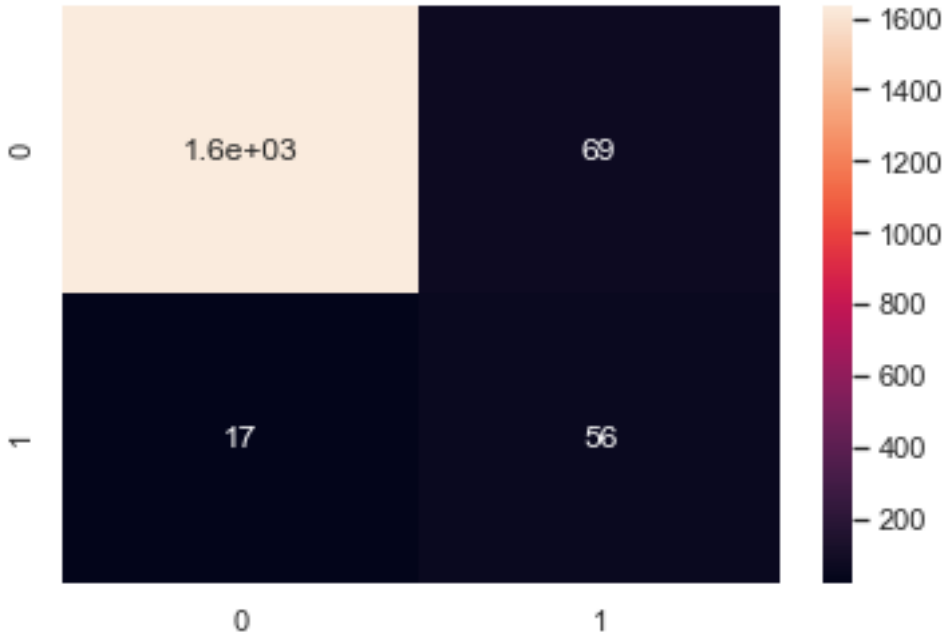

圖 4-17　Random Forest 加入 Bertopic 主題參數之混淆矩陣

DOI:10.6814/NCCU202100708

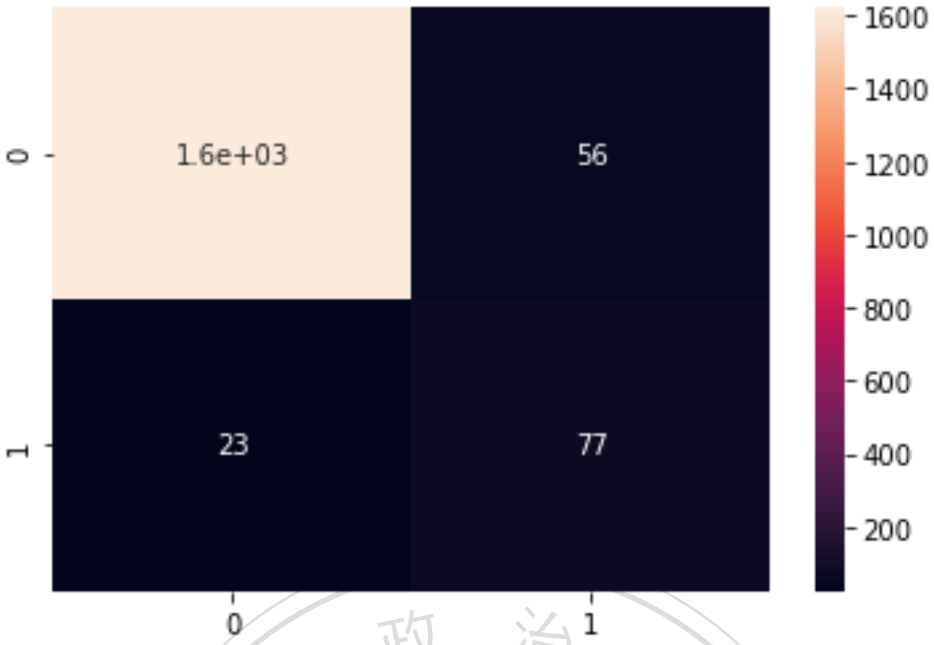

圖 4-18　XGBoost 加入 Bertopic 主題參數之混淆矩陣

加入 SMOTE 演算法後可以發現各主題參數及演算法組合之 AUC 分數皆有明顯增加，而主題參數 5 之 XGBoost 模型的 AUC 分數則是全部模型最高者，但主題參數 30 之 Random Forest 的 AUC 分數反而較原始財務模型低，而主題參數較少之模型表現則較主題參數較多之模型無論是 AUC 分數、Precision、F1-Score 、Recall 皆較為優秀。

DOI:10.6814/NCCU202100708

表 4-7 原始財務數據加入 BERTopic 主題模型和 SMOTE 演算法模型準確度 (topic 為主題數)

模型	Precision	Recall	F1-Score	AUC 分數
Logit model(3)	0.55	0.72	0.48	0.7178
Random Forest(3)	0.73	0.85	0.78	0.8494
Xgboost(3)	0.74	0.86	0.78	0.8611
Logit model(5)	0.56	0.71	0.5	0.7141
Random Forest(5)	0.76	0.85	0.79	0.8469
Xgboost(5)	0.76	0.88	0.8	0.8774
Logit model(10)	0.57	0.72	0.52	0.7177
Random Forest(10)	0.77	0.81	0.79	0.8101
Xgboost(10)	0.77	0.84	0.8	0.8371
Logit model(20)	0.56	0.69	0.51	0.6924
Random Forest(20)	0.8	0.79	0.8	0.7905
Xgboost(20)	0.8	0.85	0.82	0.8461
Logit model(30)	0.57	0.72	0.53	0.7182
Random Forest(30)	0.86	0.73	0.78	0.7273
Xgboost(30)	0.82	0.81	0.82	0.8131

DOI:10.6814/NCCU202100708

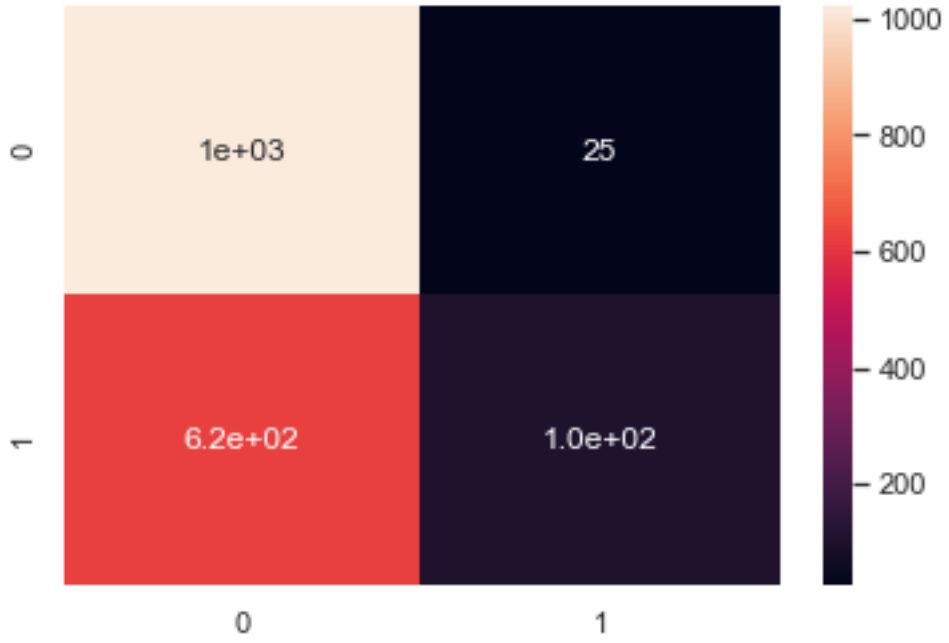

圖 4-19　Logit model 加入 SMOTE 演算法和 Bertopic 主題參數混淆矩陣

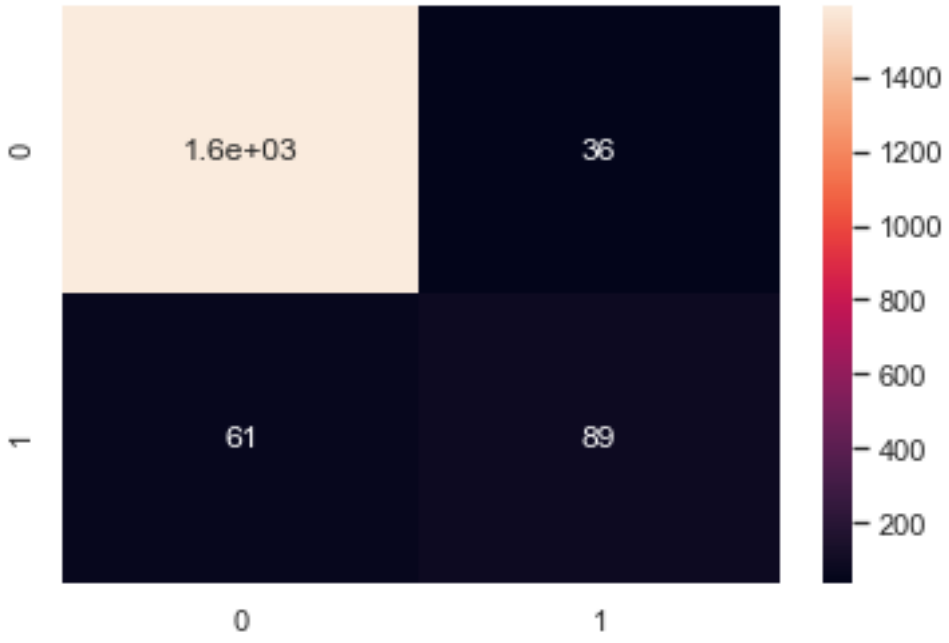

圖 4-20 Random Forest 加入 SMOTE 演算法和 Bertopic 主題參數混淆矩陣

DOI:10.6814/NCCU202100708

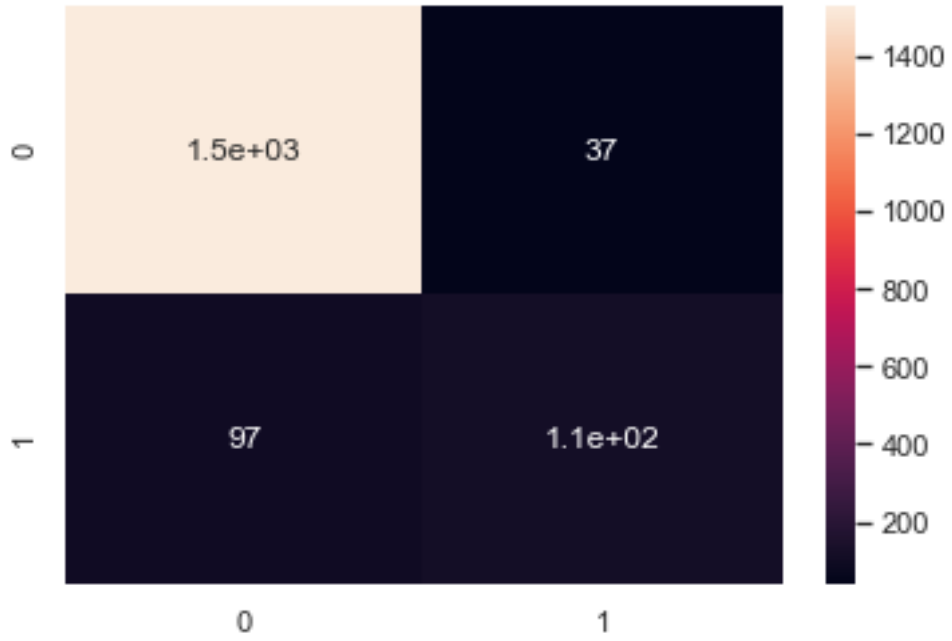

圖 4-21　XGBoost 加入 SMOTE 演算法和 Bertopic 主題參數混淆矩陣

DOI:10.6814/NCCU202100708

第五章 結論

從研究結果中可得知，加入主題模型參數時，結果與 Lopatta,Gloger and Jaeschke(2017)之結論相同，在各主題參數設定的情境下，其 AUC 分數、Precision、F1-Score 分數皆有所上升，Recall 分數雖有稍微下滑，但就測試集數量及未加入主題模型之原始財務數據模型的混淆矩陣來看，可能是由於單純就財務數據的模型其判別力並不夠，判別違約樣本能力較為薄弱，因此大多數樣本都被原始模型直接歸類為未違約樣本，才會導致加入主題模型時 Recall 分數稍微下滑。

而加入主題模型和 SMOTE 模型後可發現其 AUC 分數皆進一步向上提升，而 Precision 分數皆稍微下滑，Recall、F1-Score 皆上升，若以 AUC 方法作為評估模型總體績效的方式來看可發現評估效能是提升的，而 Precision 分數會些微下降從混淆矩陣來看是因為其訓練樣本經過 SMOTE 演算法後樣本數較為平均，在此狀況下訓練模型也較能觀察出真實情況下之模型預測能力，而從其主題捕捉到的字詞來觀察，也可看到每一個主題皆圍繞著產業或是風險主題，像

development	agreement	clinical	drug	Research

主題即可觀察到其主題是與醫藥生技有關。

加入深度學習之 BERTopic 模型後，可發現使用 XGBoost 演算法時，其 AUC 分數相對原始財務數據模型，皆有顯著之提升，但使用 Random Forest 演算法時，在 topic3、5、10 時 AUC 分數皆有上升，topic20、30 時 AUC 反而下降，從其深度學習的特性可發現當樣本不夠多時，萃取主題數多時有可能導致雜訊過多，而隨機森林分類可能無法萃取特徵值，所以準確度反而降低，而跟 LDA 主題模型比較起來，無論使用 Logit model、Random Forest、XGBoost 演算法，其 topic3、5 之 AUC 分數較高，topic10、20、30，AUC 分數反而有下降之趨勢，當使用 SMOTE 進行不平衡樣本處理後，則會得到跟 LDA 模型還有原始財務模型使用 SMOTE 演算法後類似之結果，其 AUC、F1-Score、Recall 分

DOI:10.6814/NCCU202100708

數皆上升，而 Precision 分數則些微下降。

從實證結果得知使用 SMOTE 演算法，則可以改善樣本不平衡之情形，使得公司違約預測模型之準確率有顯著之提升，在 SMOTE 演算法使用時，可彌補原違約樣本只占總樣本一成之缺陷，不管在 Logit 模型或 Random Forest 或 XGBoost 模型使用皆可以使 Recall、F1-Score 和 AUC 分數提升。

在研究最後也可得知，除了文本之文字資訊能對公司違約預測產生幫助外，運用機器學習之演算法如 Random Forest 和 XGBoost 相較於傳統 Logit model 在公司違約模型預測之準確率上，無論是使用傳統財務模型或加入 LDA 和 Bertopic 之主題參數，其 AUC、F1-Score、Recall、Precision 皆優於 Logit model，在多維度資料集的預測分類當中，無論 Logit 之閥值如何設置，其準確度很難優於 Random Forest 和 XGBoost，Logit model 其優點在於設置閥值時比較易於觀察到其模型與資料集預測能力之影響。

未來研究方向可分為樣本不平衡資料集處理，機器學習模型選擇，主題模型處理改善，公司違約模型與 CDS 商品之風險評估及幫助，從這幾個面向能使模型有更好地提升及運用方向，在樣本不平衡資料集處理方面，除了使用欠採樣演算法外，也可使用過採樣演算法來比較兩者在樣本不平衡資料集數據中改善之效果，而選擇更新或是更能進行非線性運算之機器學習演算法則能使公司違約預測模型更好地捕捉到公司違約預測數據之特徵，而主題模型則除了關鍵字捕捉外，也可以透過更詳細地分類，分出風險來源，而將公司違約預測模型運用至 CDS 實際案例中做評估，則可讓參考者更好地知道 CDS 違約之風險，並使投資者更準確地估計 CDS 商品投資之價值，也可更清楚財務數據及財報文字訊息對真實商品的影響及兩者相關性。

DOI:10.6814/NCCU202100708

參考文獻

ALTMAN, E. I. "Financial Ratios, Discriminant Analysis and Prediction of Corporate Bankruptcy." Journal of Finance 22 (September 1968): 589-610

Altman, E. I.; Haldeman, R.; and Narayanan, P. 1977. ZETA analysis: A new model to identify bankruptcy risk of corporations. Journal of Banking and Finance 10:29–54.

Black F, Scholes M (1973) The pricing of options and corporate liabilities. J Polit Econ 81:637–654; reprinted in Black F, Scholes M (2012) Financial Risk Measurement and Management, International Library of Critical Writings in Economics (Edward Elgar, Cheltenham, UK), Vol 267, pp 100–117.

Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. Journal of Machine Learning Research, 2003,3:993−1022. [doi: 10.1162/ jmlr.2003.3.4-5.993]

Computing, pp. 878-887, 2005.

C. Lin and Y. He. 2009. Joint sentiment/topic model for sentiment analysis. In Proceeding of the 18th ACM Conference on Information and Knowledge Management, pages 375–384.

Duan, J.-C.; J. Sun; and T. Wang. "Multiperiod Corporate Default Prediction - A Forward Intensity Approach Journal of Econometrics, 170 (2012), 1

H. Han, W.Y. Wang and B.H. Mao, "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning", *Proc. Int'l Conf. Intelligent*

H. He, Y. Bai, E.A. Garcia and S. Li, "ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning", *Proc. Int'l J. Conf. Neural Networks*, pp. 1322-1328, 2008.

J. R. Quinlan. Induction of decision trees. Machine Learning, 1:81-106, 1986.

L. Breiman. Random forests. Machine Learning, 45(1): 5–32, 2001

Lopatta, K., M. A. Gloger, and R. Jaeschke, 2017, Can language predict bankruptcy? The explanatory power of tone in 10-K filings, Accounting Perspectives 16, 315–343

Loughran, Tim, and Bill McDonald. 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. Journal of Finance 66(1):35–65..

Merton, R., 1974, On the pricing of corporate debt: The risk structure of interest rates, Journal of Finance 29, 449–470.

N. Peinelt, D. Nguyen and M. Liakata, "tBERT: Topic models and BERT joining forces for semantic similarity detection", *Proceedings of the Annual Conference of the International Speech Communication Association (ACL)*, pp. 7047-7055, 2020.

N.V. Chawla, K.W. Bowyer, L.O. Hall and W.P. Kegelmeyer, "SMOTE: Synthetic

DOI:10.6814/NCCU202100708

Minority Over-Sampling Technique", *J. Artificial Intelligence Research*, vol. 16, pp. 321-357, 2002.

Ohlson, James A., 1980, Financial ratios and the probabilistic prediction of bankruptcy, Journal of Accounting Research 18, 109–131.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41.

T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of SIGIR*, pp. 50-57, 1999.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 785–794. ACM, 2016.

Tomek, I. (1976). Two Modifications of CNN. IEEE Transactions on Systems, Man and Cybernetics, 6, 769–772.

Wilson, D. L. Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. IEEE Transactions on Systems, Man, and Communications 2, 3 (1972), 408–421.

DOI:10.6814/NCCU202100708