108學年度專題報告競賽

題目：P2P違約風險預測

系所班別：統計系三年級

姓名學號：吳啓玄(410678011)

張逸昇(410678040)

王薇淳(410678049)

陳怡升(410678052)

胡郁唯(410678056)

報告日期：2020/06/05

P2P 違約風險預測

吳啓玄，張逸昇，王薇淳，陳怡升，胡郁唯

May 29, 2020

摘要

金融與科技的結合，產生新的借貸方式：P2P 網路借貸。本研究旨在預測 P2P 借貸的違約與否，使投資人可以做參考，進而降低違約風險。本研究採用 Lending Club 2016 年至 2019 年的最新公開資料進行預測，而基於投資人在放款前所獲得的資訊量並不多，因此本研究之變數篩選也依據投資人的角度，選取在未投資前可以獲得的資訊作為變數，亦即利用較少的變數，使投資人只需觀察少數變數便能判斷結果，並利用羅吉斯迴歸與隨機森林建模，找出個別模型中的重要變數，再相互比較兩模型之間的差異。在羅吉斯迴歸模型與隨機森林模型中，都發現利率為主要的解釋變數，利率越高越容易產生違約；反之，利率越低則違約的機率也會隨之下降。本研究最終希望能以 APP 的方式呈現結果，使投資人能更快速且方便的以此做為投資依據，而 APP 之模型將利用羅吉斯迴歸模型進行預測，因羅吉斯迴歸模型的整體表現優於隨機森林模型，所以會選擇以羅吉斯迴歸模型完成 APP 之建構。

目錄

圖目錄

表目錄

1 緒論

1.1 研究背景及動機

隨著世界快速的經濟發展，貸款成為資金週轉的主要方式之一，不僅能舒緩資金壓力，亦能使資金更有效率地被運用。除此之外，貸款同時也是傳統商業銀行的主要業務。由於網路的興起，使人們藉此來完成許多事情，甚至發展出網路借貸，提供人們另一種獲取資金的渠道。

近十年來，P2P 網路借貸市場快速成長，吸引許多的資金需求者、小型投資人甚至是機構投資人。目前全球有超過 2500 家的 P2P 借貸平台，僅中國就囊括了絕大部分，擁有超過 2000 家的 P2P 借貸平台，致使中國成為 P2P 借貸發展最為迅速之國家。(吳志龍, 2019)

P2P 借貸主要是建構於群眾籌資的想法上，資金需求者能以不同目的在上面申請小額貸款，而投資人則依據所提供的資料來決定是否要貸款以及決定提供貸款金額。然而網路借貸平台並無特定法律給予保障及規範，也沒有金融機構之介入，因此投資人很有可能須承擔無法回收貸款金額的龐大風險。而本研究欲探討 P2P 網路借貸之違約預測，並依據貸款人以及借貸平台介面上所提供的資訊，給予投資人適當的建議，進而降低投資人遇到貸款人違約的可能性。

1.2 研究目的

基於上述研究動機，本研究將使用隨機森林模型對個人信用進行評估，以檢測隨機森林模型對於個人信用評估預測的可行性，又由於傳統商業銀行大部分是使用羅吉斯迴歸模型做違約預測，因此也將使用羅吉斯迴歸模型進行預測。用測試集對模型預測能力進行評估，並以測試集進行模型預測，再相互比較兩模型之差異。而本研究以下列三項作為研究目的:

1. 使用羅吉斯迴歸模型對個人信用進行評估，以檢驗該模型對於個人信用評估預測的能力。
2. 運用隨機森林模型對個人信用進行評估，以檢測該模型對於個人信用評估預測之可行性。
3. 比較兩模型間的準確率及模型效果

最後以投資方的角度做出總結，投資人應該依據哪些重要變數來降低違約風險。

2 文獻回顧

2.1 P2P 借貸之概念與背景

P2P 網路借貸（Peer to Peer Lending，簡稱 P2P Lending），是一種新興的金融科技行業，屬於民間的借貸媒合平台。由於網路平台的交易成本低，不但能使貸款人以低於銀行的利息進行借貸，亦可提供投資人擁有比傳統商業銀行更高的報酬率。受到外部環境之影響，傳統商業銀行對於放款業者的態度漸趨保守，使中小型業者的借貸需求無法被滿足。在種種原因的推波助瀾之下，網路借貸因而隨之崛起。相較於傳統商業銀行，P2P 網路借貸擁有以下三點特色：

1. 不具有實體交易平台，因此更能夠節省成本及突破時間、地點的限制。
2. 以小額或短期借貸為主，將商業銀行無法填補的資金缺口得以彌補。
3. 可提供投資人更高的報酬率，貸款人相對較低的利率。

然而優劣參半，P2P 網路借貸強調沒有金融機構的介入，相較於傳統商業銀行便擁有更高的違約風險，投資人往往需要擔心投資的金額是否能夠全數回收，故投資人應更加有效率地投資，給予相對安全的貸款人資金，才不會造成網路借貸之倒閉風波。P2P 網路借貸的存在同時也衝擊著傳統商業銀行，因此政府應在兩者間取得平衡，給予投資人及貸款人適當的權利與義務，以下將以美國及台灣為例。

- 美國

  因剛開始網路借貸平台尚未發展成熟，所以美國網路借貸平台的違約事件相當頻繁，隨後因美國證券交易委員會之介入，要求其商品註冊為證券，促使訊息公開、透明化，同時也要求業者須定期更新相關資訊，使得網路借貸違約事件不再頻繁。(Jin et al., 2018)

  美國存有眾多的網路平台，其中又以 Lending Club 最廣為大眾所知。Lending Club 為全球第三家互聯網借貸公司，其能利用網路搜集資料及大數據分析，建立線上即時借貸風險評估機制，並能更準確地預測用戶的付款行為和壞帳率。(殷麗萍, 2014) 因此，這也成為 Lending Club 能在近期併購美國網路銀行 Web Bank 成功的原因。

- 台灣

  台灣的 P2P 網路借貸目前僅有 5 家，其中知名的網路借貸平台為鄉民貸、LnB

信用市集。相較於美國，台灣的 P2P 網路借貸較晚起步。而至今仍不盛行的原因主要是受限於政府的法規與主管機關，再加上我國金融服務成熟，金融普及性高，且銀行逐年提高網路銀行業務，以及一般民眾與中小企業對 P2P 借貸需求不大。(劉采薇 et al., 2017) 因此 P2P 網路借貸在台灣只剩下補充傳統銀行放款，或投資人投資管道的不足之功用。

2.2 違約預測之文獻回顧

目前台灣已有不少論文探討過 P2P 網路借貸的違約預測，大部分的文獻都是挑選兩個模型相互做比較，一種是透過統計方法的羅吉斯迴歸分析，另一種分析方法則是機器學習，例如：隨機森林模型、類神經網路、支援向量機等。以下將針對兩篇文獻做探討：

1. 機器學習在 P2P 借貸信用風險之應用 (陳勃文, 2018) 此研究利用兩種機器學習方法: 支援向量機及類神經網路和一種統計方法: 羅吉斯迴歸模型，預測 P2P 網路借貸預測之違約，並著重比較類神經網路和羅吉斯迴歸模型之差異。此研究運用 Lending Club 2011 年 1 月至 2014 年 12 月之資料，進行篩選，最後共選出 14 個解釋變數進行建模，再將資料分為訓練集與測試集，並使用欠抽樣方法讓資料平衡。在類神經網路中，其訓練次數選 200 次、激發函數選取雙取正切函數、批次樣本數選 70、並選定隱藏層 1 層、隱藏層神經元數為 8 進行模型之最終訓練，其總體準確率有 61%，而在羅吉斯迴歸模型中，總體準確率有 57%，因此可以證明類神經網路在違約貸款的預測問題上比羅吉斯迴歸模型好。

2. 基於隨機森林模型下 P2P 網路借貸違約預測 (吳志龍, 2019) 此研究使用 Lending Club 2018 年度的貸款資料進行實證分析，選出 19 個解釋變數進行模型建構，再將資料分為訓練集和測試集，各別對隨機森林模型與羅吉斯迴歸模型建模。此研究之隨機森林模型決策數個數為 800 棵，每棵決策數的特徵值設定為 3，在未平衡資料前之總體準確率高達 74%，但是對於違約部分的預測正確率只有 16%，因此欲平衡資料，運用 Smote 演算法後，其總體準確率高達 83%，而對於違約部分的預測正確率也高達 81%。而在羅吉斯迴歸未平衡資料前，總體準確率也有 74%，但是對於違約部分的預測正確率只有 16%，然而在平衡資料後，其總體準確率掉到剩 64%，而違約部分的預測正確率提升到 62%。根據上述結果，不論是平衡前還是平衡後，隨機森林模型之準確率皆優於羅吉斯迴歸模型。

3 研究方法

本研究先是使用隨機欠抽樣的方法將資料平衡，以提升敏感度，再以羅吉斯迴歸模型及隨機森林來預測貸款人是否違約，找出各模型的重要變數，經由 10-fold cross validation 得到平均的模型準確率、特意度及敏感度，比較機器學習模型與一般統計模型在此分類問題上的表現是否有顯著差異。

3.1 皮爾森相關係數

皮爾森相關係數主要衡量兩個連續變數之間是否有線性關係，若兩個變數為正相關，當一方提升，另一方也會隨之提升；反之，若兩個變數為負相關，當一方提升，另一方會下降。式 (3.1) 為皮爾森相關係數的公式。

$$\begin{aligned} r(x,y) &= \frac{COV(x,y)}{s_x s_y} \\ &= \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\sum_{i-1}^{n}(y_i-\overline{y})^2}} \end{aligned} \tag{3.1}$$

相關係數 0.3 以下為低相關，0.3 到 0.7 為中等相關，0.7 以上為高度相關，若變數之間的相關性太高，有可能會導致模型預測的準確率降低。

3.2 不平衡資料：隨機欠抽樣

將不平衡資料處理為平衡資料的方法有許多，大致可分為兩類：欠抽樣與過抽樣。欠抽樣為合理的刪減多數類的樣本；反之，過抽樣為合理的增加少數類的樣本。通常樣本數多的資料較適合欠抽樣，而樣本數少的資料較適合過抽樣。由於本研究的樣本數較多，因此採用隨機欠抽樣。隨機欠抽樣以隨機抽取的方式來達到減少多數類樣本，使資料平衡。

3.3 羅吉斯迴歸

羅吉斯迴歸早由 (Berkson, 1944) 提出，但最早由 (Ohlson and Zhang, 1998)，其優點在於羅吉斯迴歸模型在研究分類問題上運算量小，模型易於解釋，其解釋力強又具穩定性，且發現在個人信用違約預測上具備良好的預測能力，所以大部分的商業銀行常使用羅吉斯迴歸模型對於客戶個人信用風險預測。使用羅吉斯迴歸模型能依據解釋變數變數估計事件發生的可能性，而在評估預測結果優劣時，會運用最大機率法則將每個樣本的數值代入迴歸模型中，如果機率發生大於或等於 0.5，表示模型

預期該事件較容易發生；反之，小於 0.5 則事件不輕易發生。
當反應變數 Y 為二元資料型態（違約與不違約），解釋變數 X 可為連續或類別資料型態。在羅吉斯迴歸模型之數學式中，則表示為：

$$log(\frac{\pi(x)}{1-\pi(x)}) = \alpha + \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \tag{3.2}$$

違約的機率為 $\pi(x)$，不違約的機率為 $1-\pi(x)$，勝算（Odd）為 $\frac{\pi(x)}{1-\pi(x)}$
由式 (3.2) 可得違約機率為：

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)}{1 - \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)} \tag{3.3}$$

3.4 隨機森林

3.4.1 決策樹

介紹隨機森林模型之前，將先帶讀者了解隨機森林模型的基底，決策樹。決策樹從根部開始，在每個節點依據 Gini Impurity 或是 Entropy（熵）所計算出的資訊增益（Information Gain）來分割母體，最後產生出許多末端節點，亦即分類結果。本研究使用 Gini Impurity，計算公式為式 (3.4)，C 是反應變數的類別，例如：違約與不違約，則 $C = 2$，p_i 是第 i 個類別佔的比例。決策樹的優點是可解釋度高，並可呈現整棵樹，方便讓非統計專業人士理解資料的分類規則與過程。然而其一大缺點是在訓練模型時容易產生過度配適的問題，模型往往會學習資料中的一些特殊規則，造成模型在測試集擁有極高準確率，但在預測新的資料時，表現卻大幅降低。(Grus, 2019)

$$Gini = \sum_{i=1}^{C} p_i(1-p_i) = 1 - \sum_{i=1}^{C} p_i^2 \tag{3.4}$$

3.4.2 模型簡介

隨機森林是被廣泛使用的一個分類模型，由許多棵決策樹（Decision Tree）所組成，其優點為準確度高、學習過程快速且可以處理大量變數。為了解決單棵決策樹過度配適的問題，隨機森林集合多顆決策樹，每棵樹在每個節點的分類都有自己的看法，這些樹中佔多數的看法就成為隨機森林在該次分類的最終結果，而在大數法則下，每棵樹的結果會趨於一致。假設有 500 棵樹，在某個節點中有 400 棵樹認為該筆資料應被分到類別 A，100 棵樹認為該筆資料應被分到類別 B，其表現就像投票一樣，會依據多數決來決定最終結果，因此隨機森林在此節點會將該筆資料分

為類別 A。這種集合許多相同模型的結果所做的預測，在機器學習中稱作集成學習（Ensemble Learning），能夠獲得比使用單個模型更高的預測表現，隨機森林透過集成許多棵決策樹來平衡掉過度配適的問題，讓模型在預測新的資料時仍能維持準確率。

3.4.3 模型運作

接著將詳細介紹隨機森林中的單棵決策樹是如何運作。隨機森林之所以能獲得比單棵決策樹更好的預測結果，主要原因是隨機性。隨機性使得模型中決策樹間的多樣性高，多樣性高代表每棵樹對分類結果都有不同看法，而隨機森林綜合這些不同看法得出的最終結果，也較不容易出現將資料分到特定類別的問題。隨機森林中的隨機性，亦即決策樹的多樣性，有兩個來源：

1. 訓練單棵決策樹的資料
   假設原始資料有 N 筆，每棵決策樹並不會使用全部 N 筆資料來生長，而是以拔靴法（Bootstrap Method）隨機抽取當中的部分資料訓練該棵決策樹，讓每棵決策樹運用到不同的資料，使得決策樹的多樣性以及隨機森林的隨機性更高。

2. 預測變數
   假設原始資料有 P 個預測變數，每棵決策樹也不會使用到所有預測變數來生長，而是在各個節點隨機抽取 m 個預測變數來生長該棵樹，例如：$m = \sqrt{P}$ ，使得每次的分類都是根據不同的資料特徵進行，以增加決策樹的多樣性及隨機森林的隨機性 (Ho, 1998)。

以上來源可以使隨機森林平衡單棵決策樹所造成的偏誤，藉此提升模型預測準確率。而 Bootstrap Sample 的樣本數、抽取的預測變數數目、決策樹的棵樹，皆是隨機森林模型中非常重要的參數，本研究也將在後文比較不同的參數值是如何影響模型的預測準確率。

3.4.4 Out of Bag Error

Out of Bag Error，縮寫為 OOB Error，是隨機森林中一種衡量預測誤差的指標。在前文中提及，假設原始資料有 N 筆，每棵決策樹並不會使用全部 N 筆資料來生長，而是以拔靴法隨機抽取當中的一部分資料訓練該棵決策樹。因此，抽取特定的樣本 n_i，亦將未使用之 n_i 訓練的決策樹組成小隨機森林，來預測 n_i，其預測誤差就是 n_i 在這個小隨機森林的 OOB Error，e_i。把所有的 n_i 對應到 e_i 做平均，就是整個

隨機森林的 OOB Error，原理相似於交叉驗證。本研究也將透過 OOB Error 對其他隨機森林的參數作圖，了解參數值的變化是如何影響模型的預測準確率。

3.5 交叉驗證

本研究在交互驗證時使用 k-fold Cross-validation，步驟如下：

1. 將資料隨機平均分成 k 個集合。
2. 令其中一個集合為測試集資料，剩餘 $k-1$ 個集合作為訓練集資料
3. 重複步驟二直到每個集合都被當作測試集為止。

3.6 混淆矩陣與衡量指標

混淆矩陣為二元分類常見的工具，利用矩陣來呈現預測結果與真實結果，經由矩陣的內容去計算衡量指標，由 3.1 可知，TP(True Positive) 是預測為違約且真實是違約，TN(True Negative) 是預測為不違約且真實是不違約，FP(FalsePositive) 是預測為違約但真實是不違約，FN(False Negative) 是預測為不違約且真實是違約，亦可以說是沒能預測出來的違約數量樣本。而本研究使用三個指標來判斷模型的優劣：

1. 準確度（Accuracy）：整體的準確率，計算公式為 $\dfrac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$
2. 敏感度（Sensitivity）：亦稱為 True Positive Rate，計算公式為 $\dfrac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$
3. 特異度（Specificity）：亦稱為 True Negative Rate，計算公式為 $\dfrac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}$

表 3.1: 二元混淆矩陣

		真實情況	
		違約	不違約
模型預測	違約	TP	FP
	不違約	FN	TN

4 研究結果

4.1 資料處理與變數篩選

本研究使用 Lending Club 公布於官方網站的公開資料，採用的年份為 2016 年至 2019 年，共 1891262 筆的資料，151 個變數。在經過以下步驟的處理後，最終剩下 823074 筆資料，17 個變數。

1. 刪除未到期的資料

   在此公開資料中，可依照借貸狀態（loan_ status）分為七個選項，包含正在還款（Current）、處於寬限期內（In Grace Period）、已付清（Fully Paid）、逾期 16-30 天（Late 16-30 days）、逾期 31-120 天（Late 31-120 days）、違約（Default）、呆帳（Charged Off），其中正在還款（Current）及處於寬限期內（In Grace Period）為未到期貸款，其餘都是已到期貸款。由於無法確認未到期之貸款在將來會不會違約，因此本研究不探討未到期的貸款，而已到期貸款共有 889580 筆。

2. 變數選擇

   本研究希望能給予投資者適當的建議，讓投資者減少遇到貸款違約的機率，因此將採用借貸平台上可觀察到的變數，共 19 個解釋變數。

3. 相關性檢定

   本研究在參考皮爾森相關係數，比較膨脹係數因子（vif），選擇將 funded_ amnt、fico_ range_ high 刪除，詳細的相關係數圖請參考附錄 A 。

4. 異常值、遺失值處理

   在初步挑選變數過後，發現四筆欄位錯誤的資料，而有遺失值的資料佔總資料的比例小且不好補值，因此將有遺失值與欄位錯誤的資料直接刪除。

5. 量化資料

   本研究的反應變數為違約狀態（loan_ status），將違約（Default）、呆帳（Charged Off）、逾期 16-30 天（Late 16-30 days）與逾期 31-120 天（Late 31-120 days）定義為違約，已付清（Fully Paid）定義為不違約，為了消除變數自身變異大小和數值大小的影響，因此將變數做標準化。

6. 平衡資料

   本研究資料未違約之貸款共 636403 筆，佔 77.32%，違約之貸款共 186671 筆，

佔 22.68%，為不平衡資料，為了避免敏感度過低，本研究針對訓練集利用抽樣方法將不平衡資料調整為平衡資料。總資料筆數相當龐大，因此採取隨機欠抽樣（Random Undersampling）的抽樣方法，將訓練集抽至 1：1。

經過資料前處理後，最終剩下 823074 筆資料，貸款狀態（loan_ status）為反應變數，其餘 17 個解釋變數請見表 4.1 。

表 4.1: 變數解釋

變數名稱	變數說明	變數名稱	變數說明
loan_ amnt	貸款金額	purpose	貸款目的
term	貸款週期	grade	LC 分配貸款等級
int_ rate	貸款利率	fico_ range_ low	信用區間的最低值
delinq_ 2yrs	過去 2 年內逾期 30 天以上的欠款次數	inq_ last_ 6mths	過去 6 個月內，貸款人的信用檔案被查詢之次數
emp_ length	工作年資	open_ acc	貸款人信用貸款額度
home_ ownership	住房持有狀況	revol_ bal	總貸款週轉餘額
annual_ inc	年收入	revol_ util	循環利用率
verification_ status	收入是否認證	total_ acc	信用檔案中當前信用額度
dti	負債收入比		

4.2 探索性資料分析

本研究的應變數為貸款狀態，0 為不違約、1 為違約，共 17 個解釋變數，由表 4.2、4.3，可初步了解各個變數的分布狀態。

表 4.2: 連續型變數

變數名稱	不違約		違約		總和	
	平均數	標準差	平均數	標準差	平均數	標準差
年收入	82736.00	81027.07	76937.00	70185.92	81421.00	78736.80
貸款金額	14294.00	9326.84	16530.00	9603.95	14801.00	9436.94
貸款利率	12.46	4.87	15.85	5.49	13.23	5.22
工作年資	6.95	3.74	6.66	3.76	6.88	3.75
週轉額度	16285.00	23864.39	15294.00	19519.53	16060.00	22954.96
負債收入比	18.23	12.25	20.17	14.19	18.67	12.74
循環利用率	46.25	24.97	49.94	24.42	47.09	24.89
信用貸款額度	11.74	5.79	11.97	5.94	11.80	5.82
當前信用額度	24.39	12.22	23.56	12.25	24.20	12.23
六個月查詢次數	0.52	0.82	0.67	0.92	0.56	0.84
信用區間的最低值	701.40	34.76	691.20	28.43	699.10	33.70

表 4.3: 類別型變數

變數名稱	類別	行占比（%）		類別占比（%）
		不違約	違約	
貸款週期	36months	81.19	62.99	77.06
	60months	18.81	37.01	22.94
LC 貸款等級	A	22.61	6.59	18.98
	B	32.11	21.28	29.66
	C	28.12	34.54	29.58
	D	11.90	22.48	14.30
	E	3.84	9.95	5.22
	F	1.41	5.17	2.26
住房持有狀況	any	0.11	0.00	0.11
	mortgage	51.12	42.08	49.06
	own	11.46	11.36	11.44
	rent	37.32	46.46	39.39
收入認證	Not Verified	34.62	26.18	32.70
	Source Verified	42.44	43.50	42.68
	Verified	22.94	30.32	24.62
貸款目的	car	1.25	0.82	1.15
	credit card	21.66	18.82	21.01
	debt consolidation	55.12	58.91	55.98
	home improvement	7.56	6.30	7.27
	major purchase	2.53	2.48	2.52
	medical	1.37	1.37	1.37
	small business	0.96	1.65	1.11
	other	9.57	9.65	9.59
逾期次數	0	81.04	79.55	80.70
	1	12.63	13.33	12.79
	2	3.62	3.95	3.70
	3	1.34	1.57	1.39
	> 4	1.37	1.60	1.42

4.3 羅吉斯迴歸

本研究期望能給予投資人最少的資訊便能獲得最大化的報酬，因此依照不完全相同解釋變數建構的三種迴歸模型。模型一以借貸利率（int_ rate）、借貸人等級（grade）、貸款目的（purpose）以及借貸金額（loan_ amnt）為解釋變數；模型二以借貸利率（int_ rate）、借貸人等級（grade）、貸款目的（purpose）、信用範圍最低值（fico_ range_ low）、貸款週期（term）以及借貸金額（loan_ amnt）為解釋變數；模型三則為使用所有變數。本研究針對模型之準確率、敏感度與特異度進行模型評估，表 4.4 為三個模型經過交互驗證的平均結果。各模型交互驗證的預測結果，請見附錄 B。

表 4.4: 三個迴歸模型的比較

模型	準確度		敏感度		特異度	
	平均數	標準差	平均數	標準差	平均數	標準差
模型一	0.622	0.0180	0.671	0.0017	0.607	0.0023
模型二	0.635	0.0015	0.655	0.0027	0.628	0.0020
模型三	0.657	0.0016	0.648	0.0029	0.659	0.0023

最終選定模型三與隨機森林做比較，因其預測率與特異度皆較高，整體來說，模型三為最適當的選擇。但若想做相對保守的投資，可選擇模型一，使用較少的變數便能得到較高的違約預測能力，除此之外，三種模型無論是準確度、敏感度及特異度皆高於 60%，因此這是值得參考的模型，但卻無法完全依照此模型作為投資依據。

4.4 隨機森林

4.4.1 決策樹個數選擇

隨機森林模型中，決策樹個數是此模型的主要參數，當決策樹個數增加到一定數量，模型的預測錯誤率會逐漸下降，並接近平穩且不再有大幅度變動的趨勢，因此找到使模型預測錯誤率達到平穩的最少決策樹個數，便是最合適的決策樹個數的主要方式，此方式不僅減少了不必要的運算時間，也讓模型的準確率不會有所下降。本研究使用 50 到 300 棵的決策樹，觀察訓練資料的袋外錯誤率（OOB Error），藉此挑選合適的決策樹個數。

由表 4.5 可發現，OOB Error 隨著決策樹個數增加，有逐漸下降且接近平緩的趨勢，因此 200 棵決策樹似乎是隨機森林模型的可用個數，但欲更加確定 200 棵決策

樹是否為合適選擇，因此進一步觀察圖 4.1 隨機森林模型決策樹個數與 OOB Error 之關係曲線。

表 4.5: 決策樹個數與訓練集袋外預測錯誤率的關係

決策數個數	OOB Error		OOB NonDefault Error		OOB Default Error	
	平均數	標準差	平均數	標準差	平均數	標準差
50	0.366	0.0011	0.368	0.0013	0.364	0.0014
100	0.355	0.0010	0.365	0.0013	0.345	0.0012
150	0.351	0.0008	0.364	0.0010	0.338	0.0012
200	0.349	0.0007	0.363	0.0012	0.334	0.0010
250	0.348	0.0007	0.363	0.0010	0.332	0.0011
300	0.347	0.0008	0.363	0.0011	0.330	0.0014

由於表 4.5 確認了交互驗證中 OOB Error 之標準差極小，表示資料隨機效果良好，因此使用交互驗證裡的第一次訓練資料，訓練集為 Fold2 到 Fold10 的所有資料，即可繪製隨機森林模型決策樹個數與 OOB Error 之關係曲線圖（圖 4.1）並代表整體模型決策樹個數與 OOB Error 之關係。

由圖 4.1 可發現，當決策樹個數為 100 到 200 時，OOB Error 便達到平緩，僅剩訓練集袋外違約預測錯誤率（OOB Default Error）有微小的下降趨勢，因此最終本研究的隨機森林模型使用 200 棵決策樹。

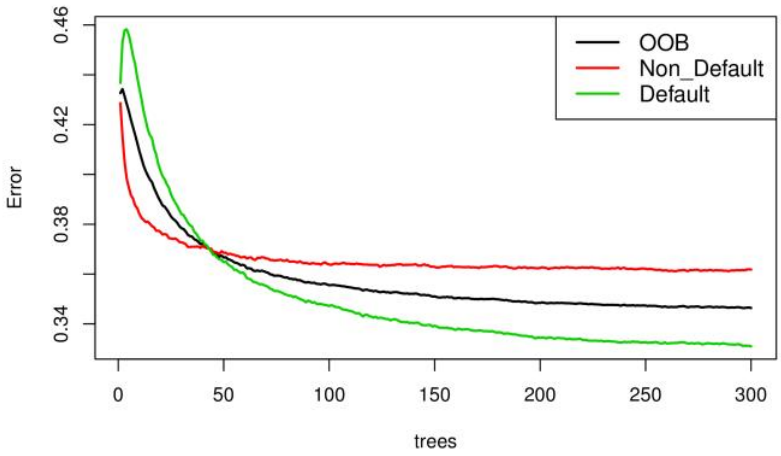


圖 4.1: 決策樹個數與 OOB Error 之關係曲線圖

4.4.2 特徵數量選擇

特徵數量亦是影響隨機森林模型的重要參數，使用較多的特徵數量，並不會讓模型準確率增加。本研究共有 18 個特徵，由圖 4.2 可知，在使用 2 個特徵時，OOB

Error 為最小，因此隨機森林模型中的每個決策樹分枝使用 2 個變數會是建議的選擇，上述的表 4.5 、圖 4.1 皆是使用 2 個特徵數量。

從圖 4.2 可知，隨著特徵數量的增加，OOB Error 有逐漸增加的趨勢，因此本研究使用的特徵數量為 4。

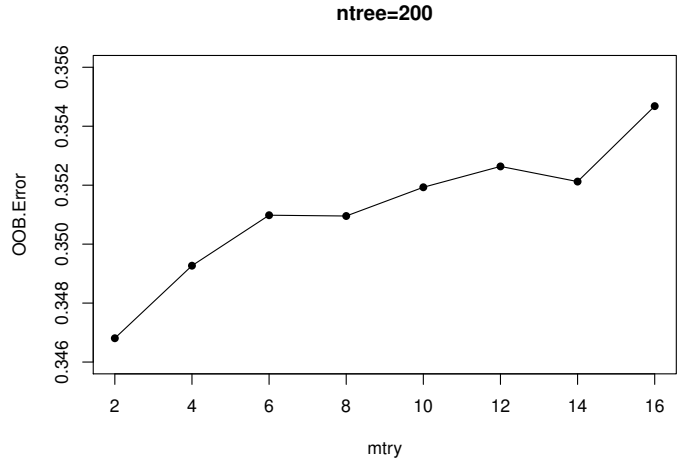


圖 4.2: 特徵數量與訓練集袋外預測誤差之關係曲線圖

4.4.3 隨機森林模型預測結果

根據上述分析，使用決策樹個數 200，特徵數量 4，作為訓練隨機森林模型的參數，找出重要變數，並利用 10-fold Cross-Validation 針對模型之準確率、敏感度與特異度進行評估。Mean Decrease Gini 表示 Gini 係數減少的平均值，可以作為衡量變數重要性的指數，在此僅列出前 10 名。由表 4.6 可發現，貸款利率為隨機森林模型最重要的變數，次之為貸款人債務收入比與貸款人信貸周轉總額。隨機森林的平均準確率為 65%，而平均敏感度 66.9% 相較平均特異度 64.4% 高出 2.5%，更多詳細的結果請見附錄 C。由上述可推論出，隨機森林模型對於預測違約的預測率會比預測不違約的預測率要來的更準確，這說明隨機森林是一個對於投資者較保守的模型，因其有較高的違約預測能力。另外，本研究的隨機森林模型之敏感度與特異度皆高於五成，因此是一個值得 Lending Club 投資者參考的模型，但因無法精準計算報酬，所以無法作為百分之百獲利的模型。

表 4.6: 隨機森林模型重要變數

變數	Mean Decrease Gini
貸款利率	19877.234
負債收入比	16773.698
週轉餘額	16416.461
循環利用率	15921.655
年收入	14589.895
貸款金額	13550.139
當前信用額度	12707.153
信用貸款額度	10582.585
信用區間最低值	10524.542
工作年資	7504.795

4.5 羅吉斯迴歸與隨機森林的比較

由表 4.7 可知，羅吉斯迴歸在準確率與特異度方面較隨機森林顯著，而敏感度則是隨機森林較羅吉斯迴歸顯著，但從表 4.7 的模型評估之平均值可推論出，兩模型間並不具有太大的差異，因此利用 ROC 曲線進一步比較模型。

表 4.7: 羅吉斯迴歸與隨機森林的 T 檢定

比較項目	羅吉斯迴歸			隨機森林			p 值
	次數	平均數	標準差	次數	平均數	標準差	
準確度	10	0.657	0.0016	10	0.646	0.0023	< 0.0001
敏感度	10	0.648	0.0029	10	0.675	0.0033	< 0.0001
特異度	10	0.659	0.0023	10	0.640	0.0011	0.0009

由圖 4.3 可看出，兩模型之 ROC 曲線幾乎呈現貼合狀態，代表模型間實際上並無太大的差距，接著考量到模型運算效率，羅吉斯迴歸的時間複雜度為 O(1)，代表無論輸入多少筆資料，都會在同一個時間內完成，但隨機森林的時間複雜度為 O(M*N*log(N))，其中 M 為決策樹的樹木，N 為樣本數目，模型的運算效率會受到 M 與 N 的影響，由於本研究樣本數目非常之龐大，羅吉斯迴歸會比隨機森林快速許多，因此將選擇羅吉斯迴歸為較佳的預測模型。

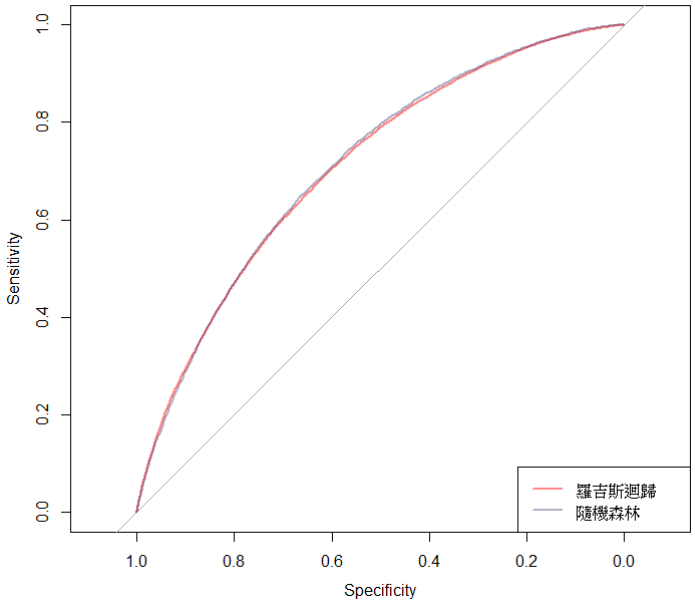


圖 4.3: 羅吉斯迴歸與隨機森林的 ROC 曲線圖

5 結論與未來展望

相較於國內其他論文，本研究運用較長的期間進行預測，且為了更符合投資者的需求，本研究只使用投資人在未投資前所能夠獲得的資訊進行預測，亦即用較少的變數來建構模型，因此整體準確率相較於文獻略為下降。在羅吉斯迴歸模型中，平均準確率約為 65%，隨機森林模型中，平均準確率約為 64%，然而此結果與先前的文獻回顧有明顯不同的差異，其原因在於本研究解釋變數中的變異大且噪聲變量多，導致羅吉斯迴歸之準確度會優於隨機森林。在羅吉斯迴歸模型一中，投資人可以參考利率、借貸金額、貸款目的與信用等級作為投資的依據，而在隨機森林模型中，本研究使用 200 棵決策樹，特徵數量 2 為隨機森林的參數，投資人可以參考利率、貸款人信貸周轉總額與債務收入比。最後考量準確率和運算時間，而選定羅吉斯迴歸模型為本研究之最終模型，也希望在未來能夠發展出 APP，並利用本研究之模型，使投資人更快速地觀察貸款人之資訊及果斷的投資，甚至能夠更有效地回收貸款金額。

總體經濟與個體經濟是息息相關的，也許能從總體經濟指標，例如：失業率、初次請領失業救濟金人數等，觀察違約率的變動，讓投資人在適當的時間點進行投資。由於 Lending Club 提供的資訊並沒有明確定義違約日，因此無法進行，而本研究建議 Lending Club 可以提供相關變數，以利於日後的研究。另外，本研究未嘗試其他集成學習方法，建議未來可以使用，並相互比較，以便於找出最適合的模型。

而礙於研究者並非美國國民，無法取得 Lending Club 貸款的即時資料，若能取得資料，APP 會更符合 Lending Club 投資人的需求。從各國家 P2P 平台發展得知政府若能適時的管控，會使網路借貸運作更加順利，因此建議台灣政府可以密切關注 P2P 網路借貸。最後，希望台灣 P2P 平台可以提供相關借貸資料供研究使用，亦或提供相關的 APP 給投資人參考。

參考文獻

Berkson, J. (1944). Application of the logistic function to bio-assay. *Journal of the American statistical association*, 39(227):357–365.

Grus, J. (2019). *Data science from scratch:first principles with python*. O'Reilly Media.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, 20(8):832–844.

Jin, B.-H., Li, Y.-M., and Liu, T.-W. (2018). Feasibility and development analysis of p2p online lending platforms in taiwan. In *World Conference on Information Systems and Technologies*, pages 82–91. Springer.

Ohlson, J. A. and Zhang, X.-J. (1998). Accrual accounting and equity valuation. *Journal of Accounting Research*, 36:85–111.

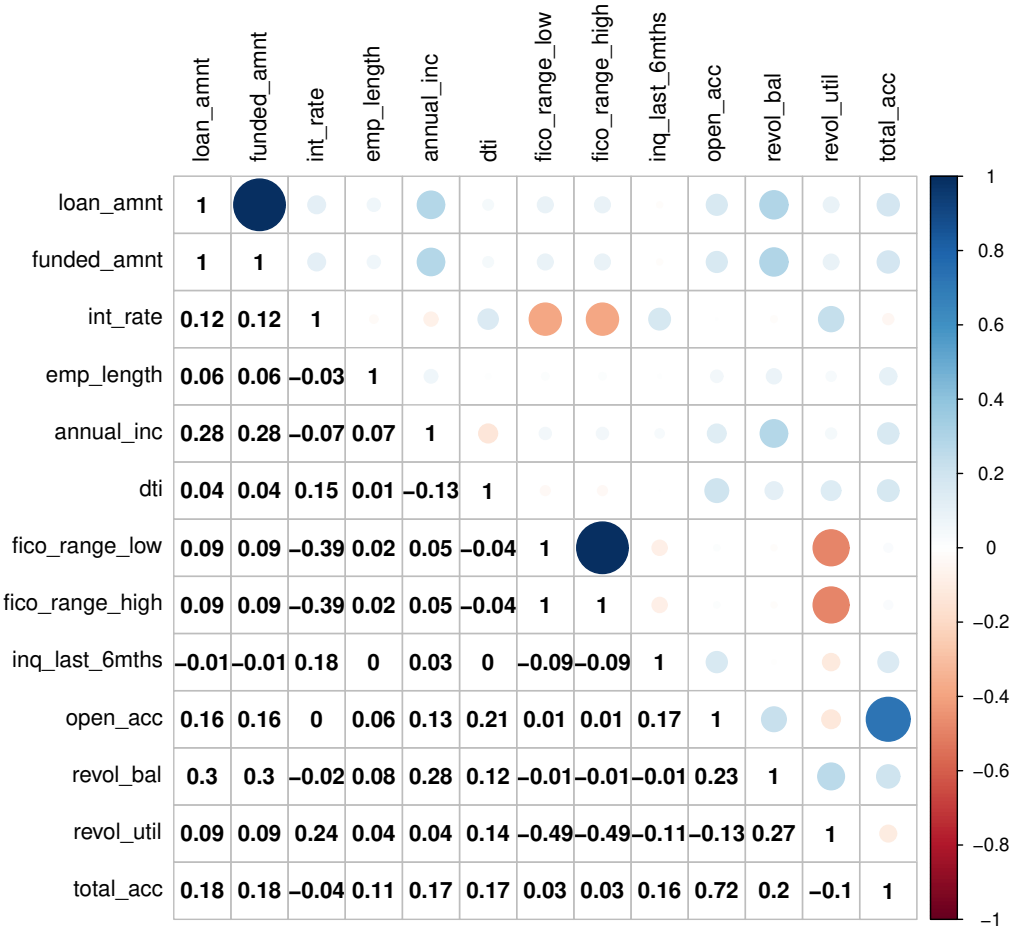
劉采薇, 李永銘, 陳姵樺, et al. (2017). *P2P* 網路借貸平台於臺灣發展趨勢研究. PhD thesis.

殷麗萍 (2014). Lending club 如何橫掃美國銀行. 中外管理, (11):34–35.

陳勃文 (2018). 機器學習在 p2p 借貸信用風險模型之應用: 以 lending club 為例. 政治大學金融學系學位論文, pages 1–41.

吳志龍 (2019). 基於隨機森林模型下 p2p 網路借貸違約預測. 政治大學金融學系學位論文, pages 1–41.

附錄 A：皮爾森相關係數圖



附圖 A.1: 皮爾森相關係數圖

附錄 **B**：羅吉斯迴歸各模型預測結果

附表 B.1 為模型一交互驗證的預測結果，以貸款利率、貸款等級、貸款目的以及貸款金額為解釋變數；附表 B.2 為模型二交互驗證的預測結果，其中解釋變數為模型一的解釋變數加上信用區間的最低值、和貸款週期；附表 B.3 為模型三交互驗證的預測結果，使用所有變數。

附表 B.1: 羅吉斯迴歸模型一的預測結果

10-fold Cross Validation	準確度	敏感度	特異度	AUC
test data = Fold 1	0.621	0.670	0.605	0.687
test data = Fold 2	0.622	0.668	0.608	0.689
test data = Fold 3	0.621	0.674	0.605	0.691
test data = Fold 4	0.622	0.669	0.608	0.690
test data = Fold 5	0.619	0.669	0.603	0.685
test data = Fold 6	0.623	0.67	0.609	0.690
test data = Fold 7	0.620	0.671	0.604	0.687
test data = Fold 8	0.624	0.672	0.609	0.690
test data = Fold 9	0.623	0.672	0.608	0.691
test data = Fold 10	0.624	0.671	0.610	0.689
平均數	0.622	0.671	0.607	
標準差	0.018	0.0017	0.0023	

附表 B.2: 羅吉斯迴歸模型二的預測結果

10-fold Cross Validation	準確度	敏感度	特異度	AUC
test data = Fold 1	0.621	0.670	0.605	0.692
test data = Fold 2	0.622	0.668	0.608	0.695
test data = Fold 3	0.621	0.674	0.605	0.695
test data = Fold 4	0.623	0.669	0.608	0.696
test data = Fold 5	0.619	0.669	0.603	0.685
test data = Fold 6	0.623	0.670	0.609	0.695
test data = Fold 7	0.620	0.671	0.604	0.693
test data = Fold 8	0.624	0.672	0.609	0.696
test data = Fold 9	0.623	0.672	0.608	0.697
test data = Fold 10	0.624	0.671	0.610	0.695
平均數	0.635	0.655	0.628	
標準差	0.0015	0.0027	0.002	

附表 B.3: 羅吉斯迴歸模型三的預測結果

10-fold Cross Validation	準確度	敏感度	特異度	AUC
test data = Fold 1	0.655	0.646	0.658	0.709
test data = Fold 2	0.657	0.647	0.660	0.712
test data = Fold 3	0.657	0.654	0.658	0.713
test data = Fold 4	0.656	0.648	0.658	0.712
test data = Fold 5	0.654	0.647	0.656	0.706
test data = Fold 6	0.660	0.644	0.664	0.712
test data = Fold 7	0.656	0.649	0.658	0.709
test data = Fold 8	0.658	0.646	0.661	0.713
test data = Fold 9	0.658	0.650	0.661	0.714
test data = Fold 10	0.656	0.644	0.659	0.709
平均數	0.657	0.648	0.659	
標準差	0.0016	0.0029	0.0023	

附錄 C：隨機森林模型預測結果

附表 C.1 為隨機森林模型交互驗證的預測結果。

附表 C.1: 隨機森林模型預測結果

10-fold Cross Validation	準確度	敏感度	特異度	AUC
test data = Fold 1	0.653	0.672	0.648	0.714
test data = Fold 2	0.648	0.672	0.641	0.715
test data = Fold 3	0.649	0.670	0.643	0.711
test data = Fold 4	0.649	0.672	0.643	0.713
test data = Fold 5	0.649	0.667	0.644	0.712
test data = Fold 6	0.649	0.669	0.643	0.717
test data = Fold 7	0.647	0.671	0.640	0.710
test data = Fold 8	0.651	0.663	0.647	0.715
test data = Fold 9	0.651	0.671	0.646	0.711
test data = Fold 10	0.651	0.663	0.648	0.713
平均數	0.650	0.669	0.644	
標準差	0.0019	0.0035	0.0028	