國立政治大學統計學系碩士班

碩士論文

P2P 借貸中借款人特徵與貸款表現關係之實證研究：以 Lending Club 和機器學習方法為例

An Empirical Study of the Relationship between Borrower Characteristics and Loan Performance in Peer-to-Peer Lending: Evidence from Lending Club and Machine Learning Techniques

指導教授：林士貴 博士
翁久幸 博士

研究生：陳槐廷

中華民國 112 年 07 月

摘要

本研究採用 P2P 平台的資料，相較於過往的文獻僅討論小型企業貸款，本研究將全面探討各種貸款目的下的貸款表現，並從借款者和投資者兩種不同角度進行分析，這包括債務整合、小型企業貸款以及信用卡等貸款類型，最後也透過機器學習的方法建構違約及貸款率模型。P2P 借貸平台中的借款者和投資者方面的重要變數包括借款金額、工作年限、年收入、債務收入比、循環信貸餘額等，透過提高借款人的信用特徵和降低投資者的風險意識，可以促進借款人的貸款通過率，並增加投資者對借款人的信任程度，在特定貸款目的（如教育、婚禮等）下，借款金額可能較低，因為這些目的不具備賺錢的能力，可能會增加投資者的風險意識，最後在預測貸款率及違約狀態模型中，XGBoost 表現最佳。

關鍵字：P2P 借貸平台、貸款目的、貸款率、違約狀態、機器學習

Abstract

This study utilizes data from a P2P platform. In comparison to previous literature that solely focused on small business loans, this research comprehensively discusses the loan performance across various loan purposes, exploring them from both borrower and investor perspectives. This includes different types of loans such as debt consolidation, small business loans, credit card loans, and more. Additionally, machine learning methods are employed to construct Loan Status and Funded Ratio models. Key variables from the borrower and investor aspects in the P2P lending platform include loan amount, years of employment, annual income, debt-to-income ratio, revolving credit balance, among others.By enhancing the credit characteristics of borrowers and reducing investors' risk perceptions, it is possible to promote higher loan approval rates for borrowers and increase investors' trust in borrowers. For specific loan purposes, such as education or weddings, the loan amounts may be lower as these purposes may not have revenue-generating potential, which could raise investors' risk awareness.Finally, in predicting loan rates and default status models, XGBoost outperformed other methods.

Keywords: P2P platform、Purpose、Funded Ratio、Loan Status、Machine Learning

目次

圖目錄

表目錄

第一章　緒論

大數據 5V 為容量 (Volume)、速度 (Velocity)、多樣性 (Variety)、準確性 (Veracity)、價值 (Value) 在近年來逐漸成為各行各業重要的議題。在資料庫技術的發展之下，政府機關以及企業得以將大量資料存放在資料庫中，為後續的大數據分析奠定基礎，而隨著資料探勘技術的進步，如迴歸分析、分群、決策樹等方法，可以將結構化以及非結構化資料轉換為所需的資訊，找到其中的模式及關聯性並加以解釋並預測。

然而，隨著資料量逐漸增加，在許多公司中卻面臨著運行資料的設備不足的問題，於是在 2000 年初，Amazon、Google 等公司推出雲端運算的服務，不僅能減少公司的成本，還可以進行大量的運算，為大數據的應用帶來更多的可能性。

在成本降低及運算增強之下，機器學習及人工智慧逐漸普及，透過不斷將資料丟入模型並學習的方式來建構所需的模型，並在各行各業中產生了極大的影響。大數據的價值也逐漸被挖掘出來，例如市場營銷、醫療保健、金融等領域都可以透過大數據分析找出潛在商機或解決問題。

總之，大數據的 5V 已經成為當今世界上不可或缺的一部分，透過不斷的技術發展以及資源的整合，大數據將會在未來繼續發揮更大的作用，進而走向更加智慧、高效的社會。

金融科技，即利用科技技術使金融服務更加智能和便利，旨在實現普惠金融的目標。大數據作為一個強大的資訊來源和分析工具，在金融科技領域中發揮著重要作用，金融科技的出現對傳統金融業產生巨大的影響，並沿生了許多新的商業模式，例如，移動支付、線上保險、風險評估等領域在金融科技的推動下有顯著的發展，大數據技術在這些領域中的應用為金融業提供了更為精準的客戶洞察、更高效的風險管理以及更創新的產品設計。

隨著資料型態的變化，金融科技領域不僅關注傳統的結構型資料，也開始將目光轉向非結構型資料，如文本、圖像和語音等。這些非結構化資料的應用為金

融服務帶來了更豐富的資訊來源，有助於金融機構更全面地了解客戶需求和市場趨勢。

2008 年至 2009 年金融危機爆發，各國的監管標準更加嚴格，使得傳統銀行的標準提高，且從傳統銀行中取得貸款變得困難; 在這其中 P2P 平台以便利性及小額貸款這些優點，開始成為趨勢，在不使用傳統金融機構下透過網路平台的方式，將投資者的資金集中並放貸給借款人，如美國的 UPST、SoFi、Lending Club。P2P 平台主要的目的是為那些在傳統金融機構無法照顧到的客群提供更便捷、低成本的借貸服務。在提供多樣化且全面的金融服務下達到普惠金融的目標。因為 P2P 平台在去中心化下，資訊可以更加的透明、公正，且減少中間的成本並提高金融服務的效率及可信度。

在 P2P 平台中借款人根據自己的一些特徵，如信用歷史、負債比率、年收入等等是否能在投資者得到信任並得到完全貸款，即申請貸款額與投資者給予資金相等; 投資者在觀察借款人出現甚麼特徵時會出現違約的狀況，那投資者該如何降低這些風險，在過去研究中，在完全貸款方面，Adam 等人 (2018) 利用 Tobit regression 發現貸款描述拼字錯誤會降低貸款提供資金的可能性，Lin 和 Wang(2020) 利用 Logistic regression 發現借款人年收入、逾期記錄對貸款人投資意願的影響。在違約方面，Carlos 等人 (2015) 以卡方檢定方式觀察貸款目的與違約的關係發現小型企業有顯著關係，Yu Jin 和 Yudan Zhu (2015) 以決策樹中的重要性作為選取變數的標準，貸款期限、年收入、貸款金額、債務收入比、信用等級和循環線利用率在貸款違約中起著重要作用。

P2P 借貸平台作為一種創新金融模式，與傳統金融機構相比，更加重視投資者與借款人之間的關係。在 P2P 借貸平台上，兩個核心問題成為研究重點：首先，投資者希望了解借款人需要具備哪些特徵才能成功獲得貸款；其次，投資者關注在哪些特徵下的借款人可能較易違約。通過深入分析這些議題，P2P 借貸平台能夠更好地為投資者和借款人提供優質的金融服務，降低風險，並提升整體市場效率。

第二章　文獻回顧

自從互聯網出現以來，P2P 借貸作為一種更常見的籌資來源，因為網站使其更容易觸及更廣泛的受眾，借款者從通過大量的投資者來集資來獲取資金，這些平台使借款人可以避開傳統銀行，直接與個人投資者合作，並以雙方有利的方式還進行貸款。例如 Lending Club 創建自己的信貸市場，跟以往傳統銀行不同的地方，將借款人和投資者直接連接在一起，為借款人提供了額外的籌資途徑。借款人不再依賴於傳統銀行的審核和批准程序，而是可以透過 P2P 借貸平台與個人投資者直接接觸。同時，這種模式也為投資者提供了更多的投資選擇，他們可以根據自己的風險偏好和投資目標選擇合適的借款人和項目，並設定利率和貸款條款以獲得回報，這種去中心化的配對模式為借款人和投資者帶來了更大的靈活性和自主性，同時也促進了更多的資金流動和經濟活動。

Nowak et al. (2018) 主要目標是探討文字描述對於貸款目的為小型企業貸款的影響，利用文字分析方法來評估借款請求的文字描述對於小型企業貸款獲得籌資的機會的影響。研究者分析了貸款描述的各種特徵，如字符數量、平均詞長、語言多樣性以及拼寫錯誤的存在，並考慮了貸款描述中使用的具體詞語對於貸款籌資的概率的影響。結果顯示，文字描述可以用來預測哪些小型企業貸款最有可能獲得投資者的籌資支持，文字描述可以幫助這些風險較高的借款者向投資者傳達有關自身和項目的有用訊息，這項研究對於小型企業在制定策略時給了重要的參考建議，以這項研究為出發點，後續以各個貸款目的進行後續分析，並給予借款者意見，進而獲取較高的貸款率。

在 P2P 平台中，借款者跟投資者兩者的信任關係固然重要，Chen et al. (2014) 結果指出借款人的信任和中介的信任都是影響投資者投資的重要因素。然而，借款者的信任更為關鍵，不僅能比中介的信任更有效地促進投資者的投資意願，還會影響投資者對中介信任的影響。為了建立投資者的信任，借款者應提供高質量的貸款訊息，關於借款者該如何表現自身的個人信用、個人資訊、貸款特徵等等對於提升投資者的信任是不可或缺的，Han et al. (2018) 的研究發現在人人貸中利率、貸款金額、貸款期限與貸款成功有顯著關係。當借款者的貸款金額越高、貸款成功的可能性就越低。

在金融危機爆發後，許多消費者和企業能從傳統銀行獲得的貸款減少，而 P2P 平台提供將尋求貸款的個人與期望借款並通過在線支付快速完成借款的個人聯繫起來。Jin and Zhu (2015) 的研究提到借款者跟投資者的訊息不對稱，可能會導致風險增加。而投資者可能只考慮高利率，而不考慮借款者的信用狀況。這裡使用決策樹、神經網路、支持向量機等模型發現款期限、年收入、貸款金額、債務收入比、信用等級和循環信用使用率是貸款違約的決定因素，Lin et al. (2017) 研究指出具有低違約風險的借款人具備以下特徵: 長時間工作、穩定的婚姻狀態、高教育水平、在大公司工作、低月供、低貸款金額、低債務收入比和無違約歷史，使得投資者可以更重視這些變數以降低投資風險。

Ma et al. (2018) 透過建立 LightGBM 和 XGBoost 模型，發現在 P2P 信用違約模型中，借款金額、貸款利率和信用評級是非常重要的因素。較低的信用評級導致更高的貸款利率和貸款金額，同時也增加了借款人違約的風險。此外，年收入、每月還款金額和每月負債收入比例也對於借款人的還款能力起到關鍵的作用。年收入代表著個人的經濟實力，高收入能夠增加借款人按時還款的能力，而每月還款金額高於每月負債收入比例則意味著借款人有足夠的資金來應對房貸等優先支付。Serrano-Cinca et al. (2015) 的研究指出利率越高，違約機率就越高，借款人的特徵，如年收入、目前住房狀況、信用歷史和借款人負債情況，都是相關的變量。因此，這些因素對於預測 P2P 借款人的違約風險至關重要。

Zhou et al. (2019) 這篇文章的主要研究目標是提出一種基於決策樹的異質集成學習違約預測模型，用於預測 P2P 借貸的客戶的違約概率。在這個模型中，他們使用了 GBDT、XGBoost 和 LightGBM 作為異質集成學習的個體分類器。此外，他們還應用了一種簡單且高效的線性加權集成策略來預測 P2P 借貸客戶的違約概

率，並取得了最佳的預測結果。他們進一步驗證了一個不平衡的、高維度的、稀疏的 P2P 借貸信用數據集，證明了基於決策樹的異質集成學習模型是解決上述問題的出色解決方案。

第三章　研究方法

第一節　邏輯斯迴歸

此方法針對是否違約、完全還款，建立模型應變數為是否違約其中 1 為違約、0 為完全還款，資料集 $(x_1, y_1), ..., (x_k, y_k), x_i \in \mathbb{R}^d, i = 1, 2, ..., k, y_i \in \{0, 1\}$, 其中 $y_i = 1$ 分別為違約發生，用線性迴歸表現機率:

$$g(x) = \ln\left(\frac{P(y=1|x_1, x_2, \ldots, x_k)}{P(y=0|x_1, x_2, \ldots, x_k)}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k \tag{3.1}$$

這個函數通常被稱為"Logistic 函數" 或"Sigmoid 函數"。在多元邏輯回歸中，我們將二元分類擴展為多元分類，並使用這個函數來計算觀測樣本屬於不同類別的機率。可得到以下算式:

$$\pi(x) = P(y=1|x_1, x_2, \ldots, x_k) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k}} \tag{3.2}$$

使用最大概似估計法來估計邏輯斯迴歸模型的參數可得概似函數:

$$L(\theta) \;\; = \;\; \prod_{i=1}^{n} \pi(x_i)^{y_i}(1-\pi(x_i))^{1-y_i} \tag{3.3}$$

接著透過 Hessian 矩陣的反矩陣來估計標準誤，矩陣中 $H_i j$ 為負對數概似函數對係數 β_i 和 β_j 的二階導數，Hessian 矩陣的反矩陣的對角線元素的平方根為係數的標準誤。

第二節　決策樹

決策樹是一種非監督式學習方法，通常用來分類或者是回歸，透過將數據集分成多個子集來進行分類，並通過對屬性值的二元劃分來決定每個分割點。決策樹演算法的分類依據通常使用熵或基尼指數來衡量每個分割點的純度。熵和基尼指數都是衡量樣本集合純度的指標，它們都可以用來選擇最佳的分割點。CART(Classification and Regression Tree) 演算法以吉尼係數作為選擇屬性的依據。資料集合 S 包含 n 個類別，將吉尼係數定義為:

$$Gini(S) = 1 - \sum_{j=1}^{n} {p_j}^2 \tag{3.4}$$

其中 p_j 為在 S 中屬於類別 j 的機率

以屬性 A 將資料集合 S 切分為 S_1 和 S_2，則此吉尼係數為:

$$Gini_A(S) = \frac{|S_1|}{|S|} Gini(S_1) + \frac{|S_2|}{|S|} Gini(S_2) \tag{3.5}$$

其中 S_1 和 S_2 為屬性 A 所構成的兩個子集合。最後以 $Gini_A(S)$ 最小作為分割屬性。

第三節　隨機森林

隨機森林是由 Breiman 在 2001 年提出，是基於決策樹的組合驗算法，隨機森林的演算法是由 CART 建構透過 Bagging 建立無相關的樹最後再平均每棵樹的預測結果。假設有一個數據集 $D = (x_i, y_i)_{i=1,...,n}$，其中 $x_i = (x_i^1, x_i^2, ..., x_i^m)$ 為包含 m 個特徵的隨機向量，而 $y_i \in \mathbb{R}$，且可以為數值或二元。

從 D 進行抽取放回，隨機選取 $|D|$ 個樣本，為一組 D_t^*，接著依以下步驟 1. 隨機從 m 個特徵中選取 k 個特徵，2. 在選取的 k 個特徵中選取較佳的特徵，3. 將節點分割為兩個子節點，利用以上 3 個步驟以 Bootstrap 抽取組成一個隨機森林 T_b，最後形成樹集合 ${T_b}_i^B$。

在預測回歸問題及分類問題公式如下:

$$\hat{f}^B(X) = \frac{1}{n}\sum_{b=1}^{B} T_b(x) \tag{3.6}$$

其中 B 為決策樹的數量、$T_b(x)$ 為第 b 個隨機森林中的預測結果。接著令 $\hat{C}_b$ 為第 b 個隨機森林的類別預測，分類公式如下:

$$\hat{C}^B = majorityvote\ \hat{C}_b(x)_1^B \tag{3.7}$$

隨機森林通常使用多數投票法來決定最終類別，即被最多數樹分類的類別被選為最終的分類結果。對於回歸問題，則通常使用平均法，即所有樹預測的平均值作為最終的預測結果。

第四節　LGBM

Ke et al. (2017) 研究說明 LGBM 是一個提升集成的模型，LGBM 提升了 Gradient Boosted Decision Tree(GBDT) 在運算時間加速及內存消耗緩解的能力，同時保留了高準確度，在傳統的 GBDT 當數據量提升時，模型的準確度會降抵且預測的速度會逐漸減慢。LGBM 模型採用直方圖的方法來降低高維度數據的負擔，並能夠加快計算過程，以避免模型過度擬合。將連續的浮點型特徵值數量級化為 l 個整數，同時生成一個具有深度限制和 k 寬度的直方圖。將原本的連續特徵值進行數量級化，並利用直方圖來進行數據結構的表達，這樣能夠控制模型的複雜度，也能夠更快地進行數據的處理和模型的訓練。LGBM 的目標函數如下:

$$Obj^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{i=1}^{t} \Omega(f_i) \tag{3.8}$$

其中 $\Omega(f_t)$ 為表示第 i 棵樹的複雜度。樹的複雜度可以由樹的葉子數量和樹的權重之平方和決定。、y_i 是實際值，而 $\hat{y}_i^{(t)}$ 是我們的模型在 t 時刻對第 i 個資料點的預測值。

$$Obj^{(t)} \cong \sum_{i=1}^{n}\left[l\left(y_i, \hat{y}_i^{(t)}\right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \Omega(f_t) \tag{3.9}$$

其中 $g_i = \partial_{\hat{y_i}^{(t-1)}} l\left(y_i, \hat{y_i}^{(t-1)}\right)$、$h_i = \partial^2_{\hat{y_i}^{(t-1)}} l\left(y_i, \hat{y_i}^{(t-1)}\right)$ 分別為一階導數和二階導數。使用 n 個樣本遍歷所有葉子節點，並計算出最終的目標函數如下:

$$Obj^{(t)} = \sum_{j=1}^{s} \left[G_j w_j + \frac{1}{2}\left(H_j + \lambda\right) w_j^2\right] \tag{3.10}$$

其中 $G_j = \sum_{i \in I} g_i$,$H_j = \sum_{i \in I} h_i$ 分別為在葉節點 j 的所有樣本上，損失函數的一階導數和二階倒數的總和，且 I 是落在葉子節點 j 中的樣本集合，λ 是一個正則化參數，用於防止過度擬合。

第五節　XGBoost

XGBoost 是由Chen and Guestrin (2016) 提出，可以解決大多數的迴歸及分類問題，是目前速度快且集成速度好的決策樹演算法，而集成學習是將多種學習模型組合，藉此或的更好的結果及泛化能力。採用 CART 決策樹演算法，由多個相關決策樹決定，從下一個決策樹的輸入樣本將與前一個決策樹的訓練和預測結果相關聯。XGBoost 模型由 K 個 CART 所組成，模型如下:

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), f_k \in F \tag{3.11}$$

其中 $\hat{y}_i$ 為預測值，$f_k(x_i)$ 為第 i 個樣本在第 k 顆樹的分數，K 為樣本總數。XGBoost 模型目的是訓練參數使目標函數最小化，目標函數為損失函數和正則項之和，目標函數如下:

$$L(\phi) = l(\phi) + \Omega(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{3.12}$$

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2$$

其中 y_i 為觀測值，l 為損失函數用來衡量 y_i 和 $\hat{y}_i$ 的差異，ω 為懲罰項避免過度擬和，γ 為每片葉子的複雜度，T 為決策樹葉子總數，λ 主要控制樹的葉子節點的權

重，進而影響模型的複雜度，ω_j 為第 j 個葉子上的分數。

$$L^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i + f_t(x_i)\right) + \Omega(f_t) \tag{3.13}$$

將(3.13)經由泰勒展開式進行近似可得(3.14)

$$L^{(t)} \cong \sum_{i=1}^{n} \left[l(y_i, \hat{y}_i{}^{(t-1)}) + g_i f_i(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \tag{3.14}$$

其中 $g_i = \partial_{\hat{y}_i{}^{(t-1)}} l\left(y_i, \hat{y}_i{}^{(t-1)}\right)$、$h_i = \partial^2_{\hat{y}_i{}^{(t-1)}} l\left(y_i, \hat{y}_i{}^{(t-1)}\right)$ 分別為一階導數和二階導數，接著我們將(3.14)經由簡化後得到(3.15):

$$L^{(t)} = \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \tag{3.15}$$

接著將(3.15)中 Ω 展開後可得到(3.16)，如下:

$$L^{(t)} = \sum_{j=1}^{T} \left[\left(\sum_{i \in I_j} g_i \right) \omega_j + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) \omega_j^2 \right] + \gamma T \tag{3.16}$$

經由(3.17)計算葉子節點 j 的最佳權重 ω_j^*

$$\omega_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \tag{3.17}$$

接著通過計算得到相應的最佳值:

$$L^{(t)} = -\frac{1}{2} \sum_{j=1}^{T} \frac{\left(\sum_{i \in I_j} g_i\right)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T \tag{3.18}$$

第四章　實證結果

第一節　資料描述及資料處理

本研究資料來自於美國的 Lending Club，原資料為 2007 年至 2018 年總筆數共 2142021 筆，每筆貸款資料中包含數值、類別及貸款目的描述，其中包含借款人個人資料以及歷史信用紀錄。

因為資料中存在遺失值、部分變數並非為數值無法做計算、建立違約、完全還款、完全貸款及部分貸款之類別變數和連續變數，所以進行以下資料清洗之步驟如下:

1、建立貸款狀態、獲得貸款、貸款率變數

2、將利率、循環帳戶中的字串符號轉為數值及工作年限轉為以年為單位

3、借款人居住地點分為東南西北、借款訊息有無填寫

4、將地址、目的 (債務、信用卡、裝潢)、房子持有狀況、申請類型、債務結算、支付方法、支付計畫、有無填寫以 one-hot encoding 轉為虛擬變數

5、將缺少 10 以上資料的變數刪除

6、如果該變數的值皆相同則刪除

7、以 VIF=10 為判斷是否有共線問題，超過則將該變數刪除

經由資料處理完，資料總筆數共 949718 筆，47 個變數，並在後續實證分析的部分會把資料切分為 7:3，分別為訓練集和測試集。根據圖4.1(1: 違約、0: 完全還款) 和圖4.2可得知其違約狀態、完全貸款、部分貸款情況，在 P2P 平台中借款人提供借款金額、期限、利率等訊息，會影響投資者在看到借款人的訊息下願不願意匹配及提供借款人所提供的借款金額，而借款人該如何提高自身的貸款通過率是一個重要的問題。

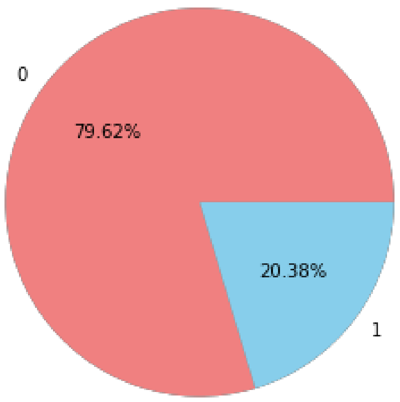


圖 4.1: 貸款狀態圓餅圖

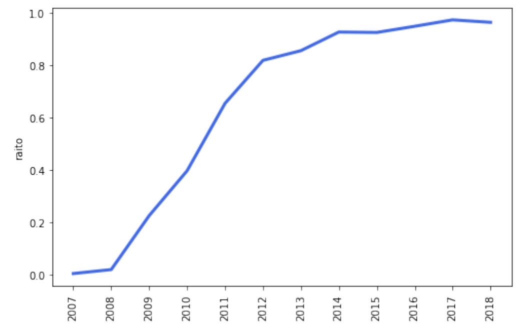


圖 4.2: 每年平均貸款率

接著因我們在意在各個貸款目的下所表現的特徵，所以先以各個貸款目的下為群組的敘述統計下進行初步的判斷。

根據表4.1的敘述統計來做初步的判斷，在以教育為主的貸款目的下，貸款率是偏低的，原因可能是因為在工作年限比較低的情況下，投資者可能比較不願意提供全額的借款給借款人，在沒有良好的收入，為了降低風險可能會降低借款的金額。

在小型企業當中可能因為企業中的計畫不確定性導致風險變高，教育和婚禮則與個人的年收入相關且並不會增加收入，這可能是投資者傾向提供低於借款人所提供的金額的原因。在貸款金額、工作年限、債務收入比、年收入、曾經欠款的總收集金額在敘述統計中可以看出在不同貸款目的下有顯著的差異，這可能會是後續再做迴歸分析時所在意的變數。

表 4.1: 各個貸款目的下的平均值

	car	credit_card	debt_consolidation	educational	home_improvement	house	major_purchase	medical	moving	other	renewable_energy	Sall_business	vacation	wedding
loan_amnt	8777.485	14697.005	15181.446	6796.319	14297.452	15549.542	11844.288	9142.711	8012.929	9975.940	9907.453	15767.389	6272.646	10442.805
term	3.360	3.412	3.537	3.098	3.512	3.525	3.415	3.317	3.234	3.334	3.318	3.487	3.137	3.326
int_rate	0.122	0.120	0.138	0.117	0.130	0.152	0.128	0.144	0.154	0.148	0.153	0.160	0.141	0.141
emp_length	5.264	5.763	6.066	3.546	6.416	5.426	5.411	5.780	4.300	5.773	5.993	5.476	6.178	4.516
annual_inc	68503.060	75260.918	74407.451	53505.204	91224.707	83501.841	77346.319	73801.482	69306.456	71512.952	72583.006	90516.174	69615.369	69337.643
dti	14.321	18.223	18.701	11.232	15.698	14.034	14.439	17.363	16.542	16.665	15.823	14.111	17.098	13.943
delinq_2yrs	0.288	0.256	0.307	0.175	0.353	0.343	0.297	0.346	0.331	0.319	0.276	0.314	0.300	0.233
inq_last_6mths	0.722	0.640	0.689	1.107	0.814	0.881	0.764	0.726	0.734	0.704	0.797	0.915	0.749	0.899
pub_rec	0.190	0.175	0.204	0.040	0.261	0.206	0.188	0.223	0.173	0.205	0.256	0.238	0.203	0.052
revol_bal	10465.369	18909.737	15921.031	9508.224	14834.980	10944.094	10796.265	12752.165	10429.519	11966.888	12303.728	14091.970	10986.423	10439.848
revol_util	0.395	0.557	0.530	0.389	0.424	0.362	0.379	0.459	0.475	0.463	0.472	0.438	0.453	0.469
last_pymnt_amnt	3775.609	6069.512	6553.679	1202.722	6759.462	7570.250	5047.369	3901.848	3296.922	4282.997	3918.455	5092.749	2582.835	2773.470
collections_12_mths_ex_med	0.017	0.012	0.015	0.000	0.016	0.019	0.015	0.016	0.017	0.017	0.017	0.016	0.014	0.000
acc_now_delinq	0.004	0.003	0.005	0.000	0.005	0.004	0.006	0.006	0.005	0.004	0.004	0.004	0.002	0.000
tot_coll_amt	771.534	205.376	220.883	241.972	298.553	245.917	264.837	305.866	211.163	416.032	206.121	295.369	249.582	184.828
total_rev_hi_lim	28235.770	35779.507	31875.402	32503.130	35042.651	31427.917	30749.126	29142.718	24520.805	27555.161	28204.622	32289.443	26580.866	29553.845
acc_open_past_24mths	4.353	4.488	4.851	4.732	4.991	4.741	4.518	4.677	4.361	4.563	4.784	4.400	5.008	4.356
avg_cur_bal	12575.466	13357.271	13211.526	13606.442	19964.389	13489.743	13104.425	13388.630	10583.569	12806.174	12557.533	14291.145	11859.762	12371.061
bc_open_to_buy	11480.060	10685.846	9645.547	10214.344	13085.189	13964.929	13096.902	9930.859	8464.133	9503.623	9710.779	11660.622	9110.020	9263.741
bc_util	48.421	63.486	61.258	59.780	49.478	43.161	45.649	54.929	56.567	54.581	56.088	52.035	54.024	59.547
chargeoff_within_12_mths	0.009	0.007	0.009	0.000	0.011	0.010	0.008	0.010	0.008	0.009	0.006	0.009	0.010	0.001
delinq_amnt	15.552	8.607	13.171	0.000	14.825	9.363	19.081	15.922	21.303	16.255	0.372	14.418	10.990	0.164
mo_sin_old_il_acct	119.096	125.342	126.043	125.622	129.337	119.961	119.503	127.179	118.983	120.813	117.699	122.739	119.008	120.001
mo_sin_old_rev_tl_op	167.437	182.009	180.115	179.897	190.141	165.090	167.225	185.636	157.982	172.000	168.178	173.008	169.336	167.436
mo_sin_rcnt_rev_tl_op	13.787	13.064	12.556	12.833	13.116	12.980	13.776	13.559	14.561	13.843	13.271	13.805	12.618	12.816
mo_sin_rcnt_tl	8.295	8.287	7.507	7.693	7.074	7.762	8.047	7.713	8.262	7.896	7.481	8.102	7.006	7.800
mort_acc	1.527	1.704	1.717	1.715	2.368	1.628	1.496	1.595	1.005	1.425	1.465	1.642	1.357	1.298
mths_since_recent_bc	24.159	22.736	23.079	23.384	24.787	23.387	24.465	27.342	25.635	26.226	23.381	25.236	24.366	24.776
num_accts_ever_120_pd	0.549	0.401	0.503	0.503	0.639	0.634	0.542	0.593	0.548	0.577	0.535	0.569	0.578	0.460
num_actv_bc_tl	3.061	3.976	3.637	3.615	3.217	3.033	3.000	3.092	2.915	3.035	3.295	3.277	3.131	3.394
num_bc_sats	4.266	5.103	4.721	4.727	4.545	4.401	4.305	4.146	3.824	4.053	4.368	4.481	4.171	4.333
num_bc_tl	7.604	8.708	8.305	8.285	8.298	7.683	7.542	7.618	6.780	7.220	7.767	7.828	7.580	7.936
num_il_tl	7.967	8.275	8.788	8.772	8.845	8.383	8.283	8.731	9.353	8.230	7.830	8.283	8.601	8.301
num_sats	10.322	11.844	11.762	11.719	11.395	10.756	10.502	10.997	10.593	10.550	10.623	10.676	10.722	11.061
num_tl_30dpd	0.003	0.002	0.003	0.003	0.004	0.003	0.003	0.004	0.004	0.003	0.005	0.003	0.001	0.002
num_tl_90g_dpd_24m	0.102	0.070	0.085	0.084	0.100	0.118	0.097	0.102	0.109	0.096	0.085	0.099	0.092	0.082
num_tl_op_past_12m	2.066	2.055	2.261	2.213	2.412	2.275	2.190	2.261	2.077	2.198	2.330	2.123	2.452	2.114
pct_tl_nvr_dlq	93.900	95.254	94.366	94.377	93.313	93.445	93.974	93.656	93.640	93.512	93.966	93.569	93.912	94.125
percent_bc_gt_75	33.105	48.564	46.885	45.310	34.011	27.580	30.686	41.430	42.995	40.699	43.477	36.017	40.812	44.961
tax_liens	0.038	0.036	0.043	0.000	0.060	0.044	0.042	0.043	0.039	0.047	0.069	0.093	0.038	0.008
total_bc_limit	19256.973	24795.665	20806.184	21478.228	21995.990	21419.678	20758.199	18392.278	16107.871	17392.475	18365.496	21321.406	16454.005	18890.398
earliest_cr_to_issue	14.710	16.217	16.115	12.135	17.108	14.915	14.836	16.618	13.989	15.402	15.056	15.130	15.093	12.198
last_cr_pull_to_issue	-2.517	-2.430	-2.431	-4.920	-2.234	-2.076	-2.318	-2.218	-2.316	-2.264	-2.554	-2.695	-2.257	-4.029
default	1536	36070	118860	56	10854	1232	3948	2460	1627	11747	166	3500	1205	278
funded_ratio	0.995	0.998	0.998	0.854	0.997	0.996	0.995	0.996	0.996	0.995	0.992	0.986	0.997	0.975
N	10441	205222	554081	326	60323	5783	21017	10759	6729	54421	691	11457	6192	2276

第二節　借款者方面

在借款者方面，我們首先使用迴歸分析探討不同貸款目的下對於貸款率的影響。我們解釋了相關變數並提出本研究認為借款人應考慮的變數，以獲得投資者的信任。接著，我們採用向後選擇的方法選擇變數，並利用貝葉斯資訊準則（BIC）進行模型選擇，並使用線性和非線性方法進行模型擬合，最終選擇具有最低均方根誤差（RMSE）的最佳模型。

2.1　各貸款目的下對於貸款率之迴歸分析

我們根據表4.2在不同貸款目的下，研究借款人能透過哪些個人資料和信用紀錄來提高貸款率。我們使用迴歸分析，將貸款率作為應變數，借款人的個人資料和信用紀錄作為自變數。我們觀察在各個貸款目的下，哪些變數對於不同族群的借款人能夠帶來正向效果。這樣修改後，表達更清晰地描述了借款者方面的研究內容，包括使用迴歸分析來探討不同貸款目的下對貸款率的影響，以及研究借款人應考慮的變數以獲得投資者的信任。同時，強調了使用向後選取和 BIC 進行模型選擇，並使用線性和非線性方法進行模型擬合的步驟。最後，指出在不同貸款目的下，分析哪些個人資料和信用紀錄能對貸款率產生正向影響。

對上述內容進行整理，在借款人選擇貸款目的時，例如家裡整修、醫療、搬家、旅遊、婚禮等，貸款金額這個變數呈現負向的相關性。這可能是因為這些目的並不具有賺錢的能力，對於投資者而言，這樣的貸款可能帶來較高的風險。另一方面，工作年限這個變數在各個貸款目的下呈現正向相關。擁有穩定工作對於投資者而言是一個較好的特徵，這代表借款人有較高的還款能力，也能降低投資者的風險。每月總債務支付額與總債務義務計算的比率這個變數在各個貸款目的下呈現正向相關。這表示借款人有良好的還款狀態，對於投資者而言也是較好的表現。然而，在教育這個貸款目的下，相關性並不顯著。對於未曾出過社會的借款人而言，他們可能沒有穩定的工作和收入，這使得他們的還款能力和信用風險難以評估。因此，投資者可能會選擇提供較低的貸款金額給這些借款人。綜合以上所述，投資者在考慮給予貸款時，通常會考慮借款人的貸款金額、工作年限、每月債務支付額和總債務比率等因素，以評估借款人的還款能力和信用風險。

表 4.2: 對借款者之迴歸分析結果

	all	car	credit_card	debt_consolidation	educational	home_improvement	house	major_purchase	Medical	moving	other	renewable_energy	small_business	vacation	wedding
intercept	(+)***	(+)***	(+)***	(+)***	(+)	(+)***	(+)***	(+)***	(+)***	(+)***	(+)***	(+)***	(+)	(+)***	(+)***
loan_amnt	(+)***	(-)***	(+)***	(+)***	(+)	(-)	(-)***	(-)	(-)	(-)***	(+)***	(+)***	(+)	(-)	(-)***
term	(+)***	(+)*	(+)	(+)***	(+)	(+)*	(+)***	(+)	(+)	(+)*	(+)	(+)***	(+)	(+)*	(+)***
int_rate	(+)***	(+)**	(+)	(+)***	(+)	(+)***	(+)***	(+)	(+)***	(+)**	(+)	(+)***	(+)	(+)***	(+)***
emp_length	(+)***	(+)	(+)***	(+)***	(+)	(+)***	(+)***	(+)	(+)	(+)	(+)***	(+)***	(+)	(+)***	(+)***
dti	(+)***	(+)***	(+)***	(+)***	(+)	(+)***	(+)***	(+)	(+)*	(+)***	(+)***	(+)***	(+)	(+)***	(+)***
inq_last_6mths	(-)***	(-)**	(-)	(-)**	(+)	(-)***	(-)***	(-)	(-)**	(-)**	(-)	(-)**	(+)	(-)***	(-)***
pub_rec	(+)***	(+)	(+)*	(+)**	(-)	(+)***	(+)***	(+)	(+)	(+)	(+)*	(+)**	(-)	(+)***	(+)***
revol_bal	(-)***	(-)*	(-)***	(-)***	(+)	(-)***	(+)***	(+)	(+)	(-)***	(-)	(-) ***	(+)	(-)*	(+)
revol_util	(+)***	(+)***	(+)	(+)*	(-)	(+)***	(+)***	(+)***	(+)	(+)***	(+)*	(+)***	(+)***	(+)***	(+)
last_pymnt_amnt	(-)***	(-)	(-)***	(-)***	(-)	(-)*	(+)***	(-)	(-)*	(-)	(-)	(-) **	(+)	(+)	(-)*
total_rev_hi_lim	(+)***	(+)	(+)***	(+)***	(+)	(+)***	(-)***	(-)	(-)	(+)***	(+)	(+)***	(-)	(+)	(-)
acc_open_past_24mths	(+)***	(+)*	(-)***	(-)***	(+)	(-)*	(-)***	(-)*	(-)	(-)***	(+)	(-) ***	(-)	(-)	(-)
bc_util	(-)***	(+)***	(+)***	(+)***	(+)	(-)***	(-)***	(-)***	(-)	(-)***	(-)*	(-) ***	(-)***	(-)***	(-)
mo_sin_old_il_acct	(-)***	(+)*	(-)***	(-)***	(-)	(-)	(-)***	(-)*	(-)	(-)***	(-)	(-)	(-)	(-)	(-)
mo_sin_old_rev_tl_op	(-)***	(+)***	(-)***	(-)***	(+)	(-)**	(+)***	(-)***	(-)***	(-)***	(-)	(-) ***	(-)*	(-)**	(-)***
mo_sin_rcnt_rev_tl_op	(-)***	(+)	(-)	(-)	(+)	(-)	(-)***	(-)	(+)	(-)	(+)	(-)	(-)	(-)	(+)
num_bc_sats	(-)***	(-)	(-)	(-)***	(-)	(-)***	(-)***	(-)	(-)	(-)**	(-)	(-)	(-)	(-)	(-)
num_bc_tl	(+)***	(+)*	(+)***	(+)***	(-)	(+)***	(+)***	(+)*	(+)	(+)***	(-)	(+)*	(+)	(+)	(+)
total_bc_limit	(-)***	(-)	(-)*	(-)***	(-)	(-)**	(+)***	(-)	(-)	(-)***	(-)	(-)	(-)	(+)	(-)
earliest_cr_to_issue	(+)***	(+)***	(+)***	(+)***	(-)	(+)***	(-)***	(+)***	(+)***	(+)***	(+)	(+)***	(+)*	(+)**	(+)***
last_cr_pull_to_issue	(+)***	(+)***	(+)***	(+)***	(+)	(+)***	(+)***	(+)**	(+)***	(+)***	(+)	(+)***	(+)***	(+)***	(+)***
Cash	(+)***	(+)	(+)***	(+)***	(+)	(+)***	(-)***	(+)***	(+)*	(+)***	(-)	(+)*	(+)	(+)	(+)*
N	**949718**	**10441**	**205222**	**554081**	**326**	**60323**	**5783**	**21017**	**6729**	**54421**	**691**	**11457**	**6192**	**2276**	**6729**
Adj R^2	**0.023**	**0.027**	**0.018**	**0.022**	**0.022**	**0.025**	**0.028**	**0.029**	**0.022**	**0.041**	**0.038**	**0.029**	**0.084**	**0.018**	**0.048**

2.2 借款者在貸款率下實證分析

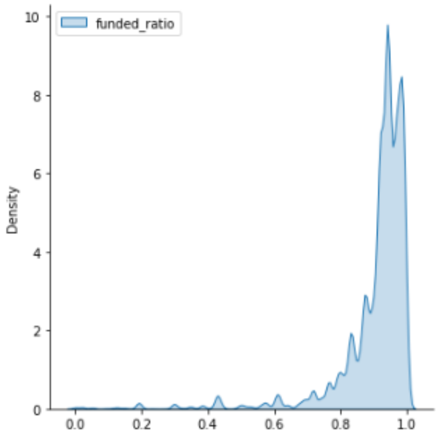
以貸款率作為應變數，將向後選取的變數，包含貸款金額、利率、工作年限等等共 22 個變數，這裡以研究方法中提到的決策樹、隨機森林、LGBM、XGBoost 進行預測，以 RMSE 作為判斷標準，在圖4.6顯示了在 XGBoost 模型中的 Log Loss 隨著迭代次數的變化。從圖中可以觀察到，在迭代次數增加的過程中，Log Loss 的值逐漸趨近於 0。這表示模型在訓練數據上取得了極好的擬合效果，能夠準確地預測訓練數據的標籤。這樣的結果表明模型能夠有效地捕捉到訓練數據中的模式和特徵，並能夠對新的數據進行準確的預測。在表4.3中的測試集可以發現在 XGBoost 中 RMSE 為 0.005，相較於其他方法更能預測貸款率。在圖4.3的密度圖中可以發現 XGBoost 的預測結果相較於決策樹的預測結果是比較接近實際值。

在迴歸分析中貸款金額、利率、債務比、工作年限、年收入等變數對於貸款率有正相關且為顯著變數，而在圖4.7中可以發現在 Feature Importance 前幾個是與迴歸分析的結果是一致的，對比前面在文獻回顧中Han et al. (2018) 提到利率、貸款金額、貸款期限與貸款成功有顯著關係是符合的。

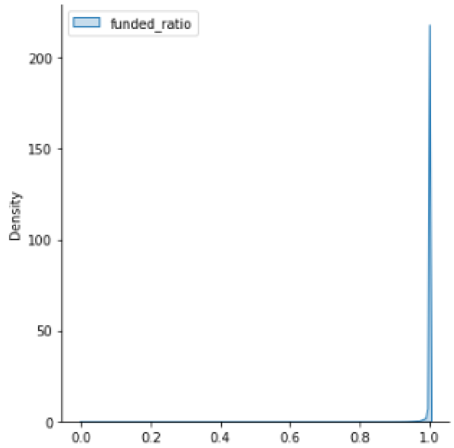
貸款申請的金額越高對貸款率越高，貸款金額較高的借款者往往有明確的貸款用途，關於債務合併以及房子在貸款目的中有比較高的貸款申請金額，這些項目可能帶來更高的經濟效益，使投資者更願意支持這些借款者。在循環利率越高，可能擁有穩定且可靠的收入來源。這讓借款者更有可能按時還款，降低了投資者所承擔的風險。

表 4.3: 機器學習預測結果

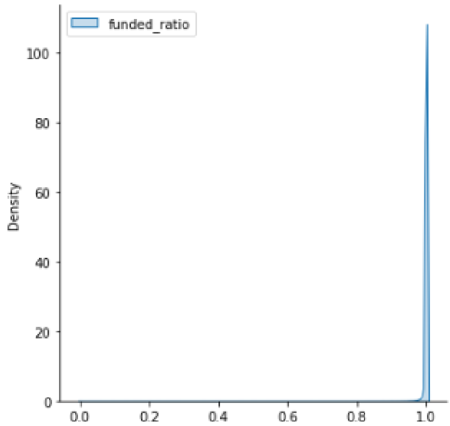
Decision Tree	Train RMSE	0.021	LGBM	Train RMSE	0.0005
	Test RMSE	0.022		Test RMSE	0.0007
Random Forest	Train RMSE	0.0001	XGBoost	Train RMSE	0.0004
	Test RMSE	0.0006		Test RMSE	0.0005



(a) 決策樹預測結果



(b) XGBoost 預測結果



(c) 實際值

圖 4.3: 決策樹及 XGBoost 預測和實際值密度圖

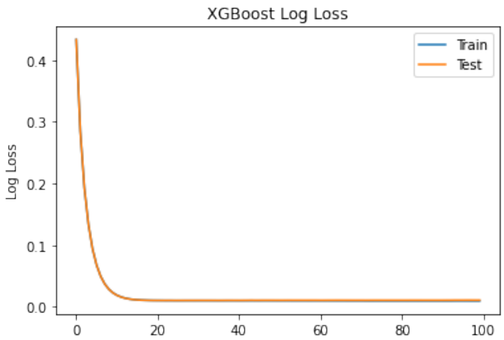


圖 4.4: Log-Loss 圖

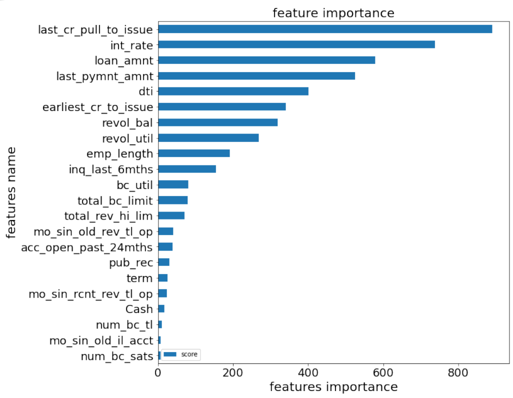


圖 4.5: Feature Importance

第三節　投資者方面

在投資者方面，在 P2P 平台中投資者在尋找投資對象時在看到借款者的個人訊息以及信用訊息，該如何針對這些特徵進行挑選以避免選擇高風險之借款者且降低違約風險。首先先針對各個貸款目的下的借款者的特徵進行邏輯斯迴歸分析，對於結果進行說明，分析在各貸款目的下出現什麼特徵時會讓違約的狀況提升，以讓投資者避免此情形。接著以線性和非線性模型進行預測，以提供投資者最佳模型，讓投資者在配對借款者時能降低違約風險。

3.1　各貸款目的下對於貸款狀態之邏輯斯迴歸分析

這裡以貸款狀態作為應變數，以向後選取的方式選取了 31 個變數，根據表4.6影響違約的大多數變數與過去文獻基本一致。例如Serrano-Cinca et al. (2015) 發現借款人的年收入和信用歷史對借款人最終是否違約有影響，年收入越高，違約率越低。Jin and Zhu (2015) 發現貸款期限、年收入、貸款金額、債務收入比、信用等級和循環信用使用率皆是影響違約的重要因素。

在表4.6可以發現以總體來看可以發現在各變數對於貸款狀正向和負向大致上和各個貸款目的相同，所以這裡可以發現投資者對於各個貸款目的下選擇借款者的訊息是一致的。

當工作年限、年收入越高時，會比較容易不違約，表示當借款人擁有一個穩定的工作及更高的收入，更容易完成還款，在對於投資者來說更是降低違約風險的參考變數之一。債務收入比、循環信貸使用率越高，可能越容易造成違約的風險，可能會讓投資者認為借款人利用比較大比例的收入去支付他的債務以及較高的信貸金額相對於所有可用的循環信貸的比例，導致借款人無法還款而增加違約的風險。貸款利率和貸款期限越高越容易違約，表示在借款人因為更高的利息和更長的還款時間而增加了本身的財務負擔，可能在面臨還款時遇到更大的困難。

而在工作年限中再生能源和婚禮這兩個貸款目的表示會提高違約率，婚禮貸款的特點是一次性的大筆支出，並且這種支出通常無法產生經濟回報，婚禮支出可能超過預算，導致借款人面臨償還壓力。

再生能源貸款通常用於購買和安裝太陽能板或風力渦輪機等設備。雖然這種投資長期來看可能會為借款人節省能源成本，甚至可能產生能源銷售的收入，但

這種收益可能需要一段時間才能實現。因此，如果借款人的財務狀況不穩定，或者他們對此類投資的預期回報過於樂觀，那麼他們可能會面臨還款困難。

表 4.4: 對投資者之邏輯斯迴歸分析結果

	all	car	credit_card	debt_consolidation	home_improvement	house	major_purchase	Medical	moving	other	small_business	vacation	educationl	renewable_energy	wedding
intercept	(-)***	(-)***	(-)***	(-)***	(-)***	(-)***	(-)***	(-)***	(-)***	(-)***	(-)***	(-)***	(+)	(-)	(-)
loan_amnt	(+)***	(+)***	(+)***	(+)***	(+)***	(+)***	(+)***	(+)***	(+)***	(+)***	(+)***	(+)***	(+)**	(+)***	(+)***
term	(+)***	(+)***	(+)***	(+)***	(+)***	(+)***	(+)***	(+)***	(+)***	(+)***	(+)***	(+)***	(+)	(+)**	(+)
int_rate	(+)***	(+)***	(+)***	(+)***	(+)***	(+)***	(+)***	(+)***	(+)**	(+)***	(+)***	(+)	(+)	(-)	(+)***
emp_length	(-)***	(-)***	(-)***	(-)***	(-)***	(-)	(-)	(-)	(+)	(-)***	(-)***	(-)	(-)	(+)	(+)
annual_inc	(-)***	(+)	(-)	(-)***	(-)***	(-)	(-)*	(-)	(-)***	(-)**	(-)	(-)***	(-)	(-)*	(-)
dti	(+)***	(+)***	(+)***	(+)***	(+)***	(+)*	(+)***	(+)***	(+)*	(+)***	(+)***	(+)***	(-)	(+)	(+)
delinq_2yrs	(+)***	(+)**	(+)***	(+)***	(+)***	(+)**	(+)**	(+)*	(+)	(+)***	(+)**	(+)	(-)	(+)**	(-)
inq_last_6mths	(+)***	(+)	(+)***	(+)***	(+)***	(+)	(+)**	(+)*	(+)***	(+)***	(+)*	(+)*	(+)	(+)	(+)**
pub_rec	(+)***	(+)**	(+)***	(+)***	(-)	(+)**	(+)	(+)	(+)	(+)	(+)	(+)	(-)	(-)	(-)
revol_util	(+)***	(+)**	(+)***	(+)***	(+)***	(+)*	(+)	(+)	(+)	(+)***	(+)	(+)**	(+)	(+)	(-)
last_pymnt_amnt	(-)***	(-)***	(-)***	(-)***	(-)***	(-)***	(-)***	(-)***	(-)***	(-)***	(-)***	(-)***	(-)**	(-)***	(-)***
collections_12_mths_ex_med	(+)***	(-)	(+)	(+)***	(+)*	(+)	(+)**	(+)	(+)	(+)***	(+)	(+)	(-)	(+)	(+)
acc_open_past_24mths	(+)***	(+)*	(+)***	(+)***	(+)***	(+)***	(+)	(+)**	(+)**	(+)***	(+)***	(+)	(-)	(+)*	(+)
avg_cur_bal	(-)***	(-)*	(-)***	(-)***	(-)***	(-)	(-)**	(-)***	(-)	(-)***	(+)	(+)	(-)	(+)	(+)
bc_open_to_buy	(-)***	(-)	(-)*	(-)***	(-)***	(-)	(-)	(-)	(+)	(-)**	(-)**	(-)	(+)	(-)	(-)
bc_util	(-)***	(-)	(-)***	(-)***	(-)***	(-)	(-)	(+)	(-)	(-)**	(-)*	(-)**	(+)	(+)	(-)
mo_sin_old_rev_tl_op	(-)***	(-)	(-)***	(-)***	(-)***	(-)	(-)	(+)	(-)	(+)	(-)	(-)	(-)	(-)	(-)
mort_acc	(-)***	(-)*	(-)***	(-)***	(-)***	(-)***	(-)***	(-)	(+)	(-)***	(-)	(-)	(-)	(-)	(-)
mths_since_recent_bc	(-)***	(-)*	(-)***	(-)***	(-)	(-)	(-)**	(-)	(-)	(-)	(-)	(-)***	(+)	(+)	(-)
num_accts_ever_120_pd	(+)***	(+)	(+)**	(+)***	(-)	(-)	(+)*	(+)	(+)*	(+)	(-)	(-)	(+)	(-)	(-)
num_bc_sats	(+)***	(+)	(+)***	(+)***	(+)**	(-)	(+)**	(+)*	(+)	(+)**	(+)	(+)	(+)	(+)	(-)
num_bc_tl	(-)***	(-)	(-)	(-)**	(-)*	(-)	(-)**	(-)**	(-)	(-)*	(-)**	(-)	(+)	(-)	(+)**
num_il_tl	(-)***	(-)	(-)***	(-)***	(-)	(+)	(-)	(-)	(-)*	(-)**	(-)	(+)	(+)	(-)	(+)
num_sats	(-)***	(-)	(-)***	(-)***	(-)	(-)	(-)	(-)	(+)	(-)**	(+)	(-)	(+)	(-)	(-)
num_tl_op_past_12m	(+)***	(+)	(+)***	(+)***	(+)	(-)	(+)	(+)	(+)	(+)***	(-)	(+)*	(-)	(-)	(-)
percent_bc_gt_75	(+)***	(+)*	(+)***	(+)***	(+)***	(+)	(+)	(+)	(+)	(+)	(+)	(+)	(+)	(-)	(+)
tax_liens	(-)***	(-)	(-)***	(-)***	(+)	(-)	(+)	(-)	(-)	(-)	(-)	(+)	(+)	(-)	(-)
total_bc_limit	(-)***	(-)	(-)***	(-)***	(+)	(-)	(-)	(+)	(-)	(-)	(+)	(+)	(-)	(+)	(+)
last_cr_pull_to_issue	(+)***	(+)***	(+)***	(+)***	(+)***	(+)***	(+)***	(+)***	(+)***	(+)***	(+)***	(+)***	(-)***	(+)***	(-)***
Cash	(+)***	(+)	(+)***	(+)***	(+)***	(-)***	(+)	(+)*	(+)	(+)	(-)***	(+)**			
DirectPay	(+)***	(+)	(+)***	(+)***	(+)*	(-)	(+)	(+)	(+)	(+)	(-)	(+)			
N	949718	10441	205222	554081	60323	5783	21017	10759	6729	54421	11457	6192	326	691	2276
Adj R^2	0.547	0.464	0.537	0.568	0.554	0.543	0.534	0.501	0.489	0.492	0.448	0.51	0.271	0.514	0.303

3.2 投資者在貸款狀態下實證分析

以貸款狀態作為應變數，將向後選取的變數，包含貸款金額、利率、工作年限等等共 31 個變數，這裡以研究方法中提到的迴歸、決策樹、隨機森林、LGBM、XGBoost 進行預測，以 F1-score 最為評估標準，根據表4.5藉由4.3計算出 F1-score，比較各模型。在圖4.6顯示了在 XGBoost 模型中的 Log Loss 隨著迭代次數的變化。從圖中可以觀察到，在迭代次數增加的過程中，Log Loss 的值逐漸趨近於 0。這表示模型在訓練數據上取得了極好的擬合效果，能夠準確地預測訓練數據的標籤。這樣的結果表明模型能夠有效地捕捉到訓練數據中的模式和特徵，並能夠對新的數據進行準確的預測。

表 4.5: 混淆矩陣

		實際	
		違約	完全還款
預測	違約	TP	FP
	完全還款	FN	TN

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4.1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4.2}$$

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.3}$$

從表4.6, 表4.7, 表4.8, 表4.9, 表4.10計算出各模型在訓練集的 F1-score，在邏輯斯迴歸中 F1-score 為 0.267, 在決策樹中 F1-score 為 0.46, 在隨機森林中 F1-score 為 0.56, 在 LGBM 中 F1-score 為 0.858, 在 XGBoost 中 F1-score 為 0.879，以 XGBoost 的表現最好。

表 4.6: 邏輯斯迴歸預測結果

訓練集		真實			測試集		真實		
		違約	完全還款	全部			違約	完全還款	全部
預測	違約	65413	291050	356463	預測	違約	28304	124902	153206
	完全還款	69901	238438	308339		完全還款	29921	101789	131710
全部		135314	529488	664802	全部		58225	226691	284916

表 4.7: 決策樹預測結果

訓練集		真實			測試集		真實		
		違約	完全還款	全部			完全貸款	不完全貸款	全部
預測	違約	42448	4689	47137	預測	違約	18223	2009	20232
	完全還款	92866	524799	617665		完全還款	40002	224682	264684
全部		135314	529488	664802	全部		58225	226691	284916

表 4.8: 隨機森林預測結果

訓練集		真實			測試集		真實		
		違約	完全還款	全部			違約	完全還款	全部
預測	違約	55796	4720	60516	預測	違約	23789	2049	25838
	完全還款	79518	524768	604286		完全還款	34436	224642	259078
全部		135314	529488	664802	全部		58225	226691	284916

表 4.9: LGBM 預測結果

訓練集		真實			測試集		真實		
		違約	完全還款	全部			違約	完全還款	全部
預測	違約	110559	11594	122153	預測	違約	47566	5112	52678
	完全還款	24755	517894	542649		完全還款	10659	221579	232238
全部		135314	529488	664802	全部		58,225	226,691	284,916

表 4.10: XGBoost 預測結果

訓練集		真實			測試集		真實		
		違約	完全還款	全部			違約	完全還款	全部
預測	違約	116800	9302	126102	預測	違約	49218	4597	53815
	完全還款	18514	520186	538700		完全還款	9007	222094	231101
全部		135314	529488	664802	全部		58,225	226,691	284,916

Jin and Zhu (2015) 提到在、年收入、貸款金額、債務收入比和循環信用使用率皆是影響違約狀態的顯著變數，Serrano-Cinca et al. (2015) 提到年收入、信用歷史和借款人負債情況都是影響的變數。在表4.4中可以看到在利用邏輯斯迴歸對於貸款狀態下進行分析可得知貸款金額、利率、工作年限、債務收入比、總信貸循環餘額、過去 2 年借款人信用中逾期 30 天以上的拖欠次數等等皆為顯著變數。與過往的文獻一致。接著因為在非線性的情況下 XGBoost 的表現狀況最好，所以把邏輯斯迴歸中的變數由 XGBoost 進行配適，最後看到圖4.5中的 Feature Importance 可以發現在前幾個變數中出現的變數像是貸款金額、利率、年收入、債務收入比、信用卡總信用額度都屬於比較重要的變數。

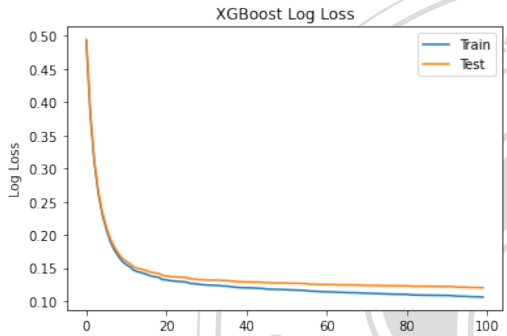


圖 4.6: Log-Loss 圖

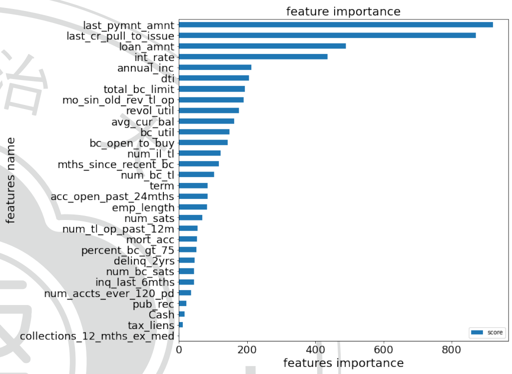


圖 4.7: Feature Importance

第五章　結論及未來展望

第一節　結論

從借款人和投資者的角度來看，有幾個關鍵的變數會影響他們的決策。對於借款人來說，如果他們希望獲得更高的貸款金額，他們可以提高他們的工作年限、年收入、降低債務收入比以及減少循環信貸使用率，這些因素能夠增加借款人獲得貸款的機會。然而，對於特定的貸款目的，例如教育、婚禮等，借款金額可能較低，因為這些目的並不具備賺錢的能力，這可能會增加投資者對風險的關注。

而對於投資者來說，他們關注的變數包括借款人的工作年限、年收入、債務收入比和循環信貸餘額等。高工作年限和較高的年收入對於投資者來說是重要的參考因素，因為這表示借款人具有較穩定的收入來源和還款能力。然而，高債務收入比和循環信貸使用率可能會增加投資者的違約風險意識，因為這意味著借款人可能已經承擔較大的財務壓力和債務負擔。

因此，借款人和投資者在考慮貸款時都應該注意這些關鍵變數，借款人可以通過提高收入、降低債務比例和循環信貸使用率來提高他們的貸款機會，而投資者則應該關注借款人的工作年限、年收入、債務收入比和循環信貸餘額等因素，以便做出明智的投資決策。

第二節　未來展望

在一些特定項目中像是教育、婚禮等，因為不具備賺錢的能力，所以可能比較不會吸引到投資者的關注，我們可以探索創新的投資模型和策略，例如引入社會影響投資或可持續發展目標等。這樣的投資模型可以使投資者更關注項目的社會效益和長期可持續性，而非僅僅追求短期經濟回報。希望在未來可以期待 P2P 借貸市場在吸引投資者、提供多樣化的投資機會以及促進可持續發展方面發揮更大的作用。這需要行業各方的共同努力，包括平台運營商、投資者、借款人和政府監管機構，以確保市場的可持續性和長期發展。

附錄 A

表 A.1: 借款者個人資料、信用紀錄、屬性、應變數之描述

變數名稱	描述
應變數	
loan_status(model_1)	貸款狀態 (0 表示全額還款，1 表示違約)
funded ratio(model_2)	借款人最後獲得的金額和一開始申請的金額之比率
貸款的屬性	
loan_amnt	借款人申請的貸款金額
term	貸款還款次數
借款人個人資料	
annual_inc	借款人的年收入
avg_cur_bal	所有帳戶的平均當前餘額
dti	借款人的每月總債務支付額與總債務義務計算的比率
emp_length	就業年限
revol_bal	總信貸循環餘額
acc_now_delinq	借款人現在拖欠的賬戶數量
delinq_amnt	借款人現在拖欠的賬戶所欠的逾期金額
num_bc_tl	銀行帳戶數量
percent_bc_gt_75	所有銀行帳戶大於限額的 75%
tax_liens	稅收留置權數
total_rev_hi_lim	總循環高信用/信用額度
revol_util	循環額度利用率
借款人的信用記錄	
inq_last_6mths	過去 6 個月的查詢數量
chargeoff_within_12_mths	12 個月內的註銷次數
collections_12_mths_ex_med	12 個月內的收集次數，不包括醫療收集
delinq_2yrs	過去 2 年借款人信用中逾期 30 天以上的拖欠次數
mo_sin_old_il_acct	自最早的銀行分期付款帳戶開設以來的月數
mo_sin_old_rev_tl_op	自最早的循環帳戶開設以來的月份
mo_sin_rcnt_rev_tl_op	最近一次開立循環帳戶後的月數
mo_sin_rcnt_tl	自最近開戶以來的月份
mths_since_recent_bc	自最近開設銀行帳戶以來的月數
mths_since_recent_inq	自最近一次查詢以來的幾個月
num_accts_ever_120_pd	逾期 120 天或以上的帳戶數量
num_tl_120dpd_2m	目前逾期 120 天的帳戶數量
num_tl_30dpd	當前逾期 30 天的帳戶數量
num_tl_90g_dpd_24m	過去 24 個月逾期 90 天或以上的帳戶數量
pct_tl_nvr_dlq	從未拖欠的交易百分比
pub_rec_bankruptcies	公共記錄破產數量
tot_coll_amt	欠款總額
earliest_cr_to_issue	借款人最早的信用額度開始向借款人發放貸款的月數
last_pymnt_amnt	最後收到的總還款金額
last_cr_pull_to_issue	從平台上提取這筆貸款給借款人發放這筆貸款的月數

參考文獻

Chen, D., Lai, F., and Lin, Z. (2014). A trust model for online peer-to-peer lending: a lender ' s perspective. *Information Technology and Management*, 15:239–254.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Han, J.-T., Chen, Q., Liu, J.-G., Luo, X.-L., and Fan, W. (2018). The persuasion of borrowers ' voluntary information in peer to peer lending: An empirical study based on elaboration likelihood model. *Computers in Human Behavior*, 78:200–214.

Jin, Y. and Zhu, Y. (2015). A data-driven approach to predict default risk of loan for online peer-to-peer (p2p) lending. In *2015 Fifth international conference on communication systems and network technologies*, pages 609–613. IEEE.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.

Lin, X., Li, X., and Zheng, Z. (2017). Evaluating borrower ' s default risk in peer-to-peer lending: evidence from a lending platform in china. *Applied Economics*, 49(35):3538–3545.

Ma, X., Sha, J., Wang, D., Yu, Y., Yang, Q., and Niu, X. (2018). Study on a prediction of p2p network loan default based on the machine learning lightgbm and xgboost algorithms according to different high dimensional data cleaning. *Electronic Commerce Research and Applications*, 31:24–39.

Nowak, A., Ross, A., and Yencha, C. (2018). Small business borrowing and peer-to-peer lending: Evidence from lending club. *Contemporary Economic Policy*, 36(2):318–336.

Serrano-Cinca, C., Gutiérrez-Nieto, B., and López-Palacios, L. (2015). Determinants of default in p2p lending. *PloS one*, 10(10):e0139427.

Zhou, J., Li, W., Wang, J., Ding, S., and Xia, C. (2019). Default prediction in p2p lending from high-dimensional data based on machine learning. *Physica A: Statistical Mechanics and its Applications*, 534:122370.