元 智 大 學

資 訊 工 程 學 系

期 末 論 文

點對點借貸信用風險預測

Peer-to-Peer Lending Credit Risk Prediction

學 生：陳卿雅、郭羽蕎、李婕綾、謝宜庭

指導教授：簡廷因

中 華 民 國 　一一三　 年 六 月

點對點借貸信用風險預測

Peer-to-Peer Lending Credit Risk Prediction

學生：陳卿雅、郭羽蕎、李婕綾、謝宜庭

元 智 大 學

資 訊 工 程 學 系

期 末 論 文

# 摘 要

本研究探討了傳統借貸信用風險管理的挑戰，指出傳統方法依賴於信用評分、收入證明和財務報表，且重度依賴人工審核，容易導致不一致和偏見決策。此外，這些方法基於靜態模型和歷史資料，難以適應不斷變化的經濟條件和借款人行為。為了解決這些問題，提出建立一個可靠的信用風險預測模型，以提高信用決策的準確性和運營效率，並適應市場變化。

Peer-to-peer lending credit risk prediction

Student：陳卿雅、郭羽蕎、李婕綾、謝宜庭

Submitted to Department of Computer Science and Engineering

Yuan Ze University

# ABSTRACT

This study explores the challenges of traditional credit risk management in lending, highlighting that conventional methods rely heavily on credit scores, income proofs, and financial statements, as well as manual review processes, which can lead to inconsistent and biased decisions. Additionally, these methods are based on static models and historical data, making it difficult to adapt to changing economic conditions and borrower behaviors. To address these issues, the study proposes the development of a reliable credit risk prediction model to improve the accuracy of credit decisions and operational efficiency while adapting to market change.

目錄

書名頁 i

[摘 要 ii](#_Toc170037538)

[ABSTRACT iii](#_Toc170037539)

[一、介紹 1](#_Toc170037540)

[二、研究目的 1](#_Toc170037541)

[三、研究方法 1](#_Toc170037542)

[3.1 Original data 1](#_Toc170037543)

[3.2 Data Preprocessing 2](#_Toc170037544)

[3.3 Feature Selection 2](#_Toc170037545)

[3.4 Data Synthesis 2](#_Toc170037546)

[3.5 Credit Risk Prediction Model 2](#_Toc170037547)

[3.6 SHAP Explanation Model 3](#_Toc170037548)

[四、研究成果 3](#_Toc170037549)

[4.1 Data Preprocessing 3](#_Toc170037550)

[4.2 Feature Selection 5](#_Toc170037551)

[4.3 Data Synthesis 6](#_Toc170037552)

[4.4 Credit Risk Prediction Model & SHAP Explanation Model 6](#_Toc170037553)

[五、結果與討論 11](#_Toc170037554)

[六、參考文獻 12](#_Toc170037555)

# 一、介紹

貸款信用風險指的是借款人無法履行其財務義務的可能性，這種情況可能會導致貸款機構蒙受損失。了解和預測信用風險對於銀行和金融機構來說至關重要，它可以幫助這些機構有效評估和管理潛在的風險，從而制定更合理的貸款政策和信貸決策。

在傳統的貸款信用風險管理中，業界面臨諸多挑戰。首先，資料來源的限制性非常高，嚴重依賴於信用評分、收入證明和財務報表等傳統資料，這些資料往往無法涵蓋借款人的全面風險狀況，這種傳統方法經常錯過許多有價值的資訊，無法全面評估借款人的信用風險。

傳統方法大多依賴人工審核，不僅效率低下，而且容易受到個人主觀意識的影響，導致不同信貸官員或機構之間可能做出不一致且帶有偏見的決策。同時，大多數現有的風險管理模型是基於靜態規則和過去的歷史資料建立的，這使得它們無法適應經濟條件和借款人行為的快速變化。

本研究透過Lending Club(美國的一家點對點借貸公司)在Kaggle上所提供的數據集[1]，提取特徵並訓練模型，實現一個可靠的信用風險預測模型。

# 二、研究目的

本研究的目的是建立一個可靠的信用風險預測模型，以提高決策的準確性和效率，並能夠適應不斷變化的市場條件和借款人行為。讓這個模型克服傳統信用風險管理方法中的局限性，如依賴靜態模型、人工審查所帶來的不一致性和偏見，從而提供更精確和高效的風險管理工具。

# 三、研究方法

研究方法分為五部分：「Data Preprocessing」、「Feature Selection」、「Data Synthesis」、「Credit Risk Prediction Model」、「SHAP Explanation Model」，以下會針對每個部分依序做詳細的說明，另外會先介紹本研究的原始資料。

一張含有 文字, 字型, 便利貼 的圖片

自動產生的描述

圖 1、研究方法流程圖

## 3.1 Original data

原始資料大小為2260668筆\*145cols，是由Lending Club所提供的數據集[1]，裡面包含了2007-2015年之間所有貸款的完整貸款數據，像是當前貸款狀態和最新付款資訊，特徵的部分像是信用評分、財務查詢數量、地址等都包含在內，如下圖2。

一張含有 文字, 螢幕擷取畫面 的圖片

自動產生的描述

圖 2、原始資料樣貌

## 3.2 Data Preprocessing

首先會刪除欄位缺失值大於80%的欄位，再針對缺失值進行補值，像是fill zero、mean value、mode，以及針對有非數值型態的資料則會使用Label encoding或One hot encoding來將文字轉換成數字。

## 3.3 Feature Selection

本研究篩選特徵之方法採用RFECV，結合遞歸特徵消除(Recursive Feature Elimination)及交叉驗證(Cross Validation)，針對不同特徵子集合訓練模型後做評分。其特性會將資料拆分為K Folds並單獨執行RFE，直到保留一個特徵為止，並取得最佳特徵子集。在不斷遞迴的運算中，會自行消除效果不佳的特徵值，保留較佳的特徵值進行下一次運算，最終輸出模型效能最高的最佳特徵子集合。

## 3.4 Data Synthesis

合成資料採用Borderline SMOTE，其採樣過程會將少數類樣本分為三類，Safe、Danger和Noise，且只會在Danger樣本附近生成新的樣本，以此來平衡資料與增加界線的明確度。

## 3.5 Credit Risk Prediction Model

本研究共訓練了DNN、Random Forest、Logistic Regression、CatBoost以及XGBoost這5種模型，且將會以80%訓練、20%測試的資料比例去訓練模型，每個模型我們會分成兩個組別去實作，一組是沒有使用Borderline SMOTE&RFECV，另一組則有。此外，因為DNN有自行選擇特徵的能力，因此DNN分成有使用Borderline SMOTE以及沒有Borderline SMOTE這兩個組別，我們將模型分別實作兩組的用意是藉由沒有使用Borderline SMOTE&RFECV的組別來證明有使用的組別更能精確預測出借款人是否具有信用風險。

## 3.6 SHAP Explanation Model

最後會使用SHAP來對模型進行解釋，SHAP是基於賽局理論，將模型的預測解釋分析成每個特徵的貢獻，透過計算每個特徵的shapely value，藉此衡量該特徵對預測的貢獻度，使我們理解模型是如何做出決策，提升模型的透明度以及可信度。

# 四、研究成果

## 4.1 Data Preprocessing

#### 缺失值刪除

將欄位缺失值大於資料總筆數80%的欄位先刪除，刪除的欄位如圖3所示，刪除後資料欄位數量變為87個欄位，如圖4所示。

一張含有 文字, 螢幕擷取畫面, 字型, 功能表 的圖片

自動產生的描述

圖 3、欄位缺失值大於總筆數80%的欄位

一張含有 文字, 螢幕擷取畫面, 黑與白, 單色 的圖片

自動產生的描述

圖 4、去除缺失值之欄位樣貌

#### 非數值型資料處理

將非數值型的資料篩選出來分別作處理，所篩選出的欄位如圖5所示，會先針對目標欄位loan\_status，將值為Fully Paid跟Charged Off兩種值篩選出來，其餘的則是觀察其數值再決定做Label encoding或One hot encoding，處理後之資料數量為1303607筆\*101 cols，如圖6所示。

一張含有 文字, 螢幕擷取畫面, 功能表, 字型 的圖片

自動產生的描述一張含有 文字, 螢幕擷取畫面, 功能表, 字型 的圖片

自動產生的描述

圖 5、包含非數字值之欄位

一張含有 文字, 螢幕擷取畫面, 黑與白, 單色 的圖片

自動產生的描述

圖 6、處理非數值型後之資料

#### 缺失值補值

接下來針對欄位有缺失值的欄位補上平均值。

一張含有 文字, 螢幕擷取畫面, 功能表 的圖片

自動產生的描述一張含有 文字, 螢幕擷取畫面, 功能表 的圖片

自動產生的描述

圖 7、包含缺失值之欄位

## 4.2 Feature Selection

使用RFECV篩選出10個最佳特徵子集合，下方顯示最佳的10個特徵及其意義，並以圖表及表格之方式排列顯示其重要性：

* + loan\_amnt：借款人申請貸款之金額。如果在某個時間點，信貸部門減少了貸款金額就會反應在數值上。
  + funded\_amnt：當時承諾的貸款總額。
  + funded\_amnt\_inv：投資者當時對該貸款承諾的總金額。
  + term：貸款的還款次數，數值以月為單位，可以是36或60。
  + total\_rec\_prncp：迄今收到的本金。
  + total\_rec\_late\_fee：至今已追回的總遲繳金額。
  + recoveries：扣除回收後的費用。
  + last\_pymnt\_amnt：最後收到的付款總額。
  + num\_tl\_30dpd：目前逾期30天的帳戶數量(過去兩個月更新)。
  + debt\_settlement\_flag：標記已註銷的借款人是否正在與債務結算公司合作。

一張含有 螢幕擷取畫面, Rectangle, 正方形, 行 的圖片

自動產生的描述一張含有 文字, 螢幕擷取畫面, 字型 的圖片

自動產生的描述

圖8、最佳特徵特徵子集合

## 4.3 Data Synthesis

接下來將處理過後的資料切成訓練及測試兩種資料，比例為8:2 (1042885: 260722)，再將訓練資料做Borderline SMOTE，其總筆數由1042885至1667122筆。

一張含有 圓形, 圖表, 字型, 文字 的圖片

自動產生的描述 一張含有 圓形, 圖表, 字型, 文字 的圖片

自動產生的描述

圖9、訓練資料Borderline SMOTE前後對比圖

## 4.4 Credit Risk Prediction Model & SHAP Explanation Model

以下用表格分別呈現五個模型在有無使用RFECV及Borderline SMOTE的Confusion Matrix，並顯示其SHAP解釋圖。

#### **DNN**

表1顯示了無使用Borderline SMOTE的Confusion Matrix數值，可以看到模型雖然能準確識別是否具有信用風險的借款人，但相較於表2有使用Borderline SMOTE的Confusion Matrix數值，可以看出使用Borderline SMOTE後，模型能更有效地識別具有信用風險的借款者，因此在使用深度模型搭配Borderline SMOTE可以提高預測的準確性。

|  |  |  |  |
| --- | --- | --- | --- |
| Baseline Model  Confusion Matrix | | Predicted Condition | |
| Positive | Negative |
| Actual  Condition | Positive | 208372 | 19 |
| Negative | 316 | 52015 |

表1、Baseline Model Confusion Matrix

|  |  |  |  |
| --- | --- | --- | --- |
| Our Method  Confusion Matrix | | Predicted Condition | |
| Positive | Negative |
| Actual  Condition | Positive | 208039 | 352 |
| Negative | 20 | 52311 |

表2、Our Method Confusion Matrix

一張含有 文字, 螢幕擷取畫面, 行, 圖表 的圖片

自動產生的描述

圖 10、DNN SHAP

DNN SHAP圖說明：

* total\_rec\_prncp：借款人已經償還的本金總額。此數值越大，表示借款人還款能力強，具有較高的信用，越可以借錢；此數值越小代表借款人還款能力較弱，就具有信用風險。
* funded\_amnt：承諾的貸款總額。此數值越大，表示借款人承諾要償還的貸款總額很大，可能具有較高的信用風險；此數值越小，表示借款人承諾的貸款總額較小，具有較低的信用風險。
* funded\_amnt\_inv：投資者資助的貸款金額。此數值越大，表示承擔的貸款金額增加，因此可能增加欠款的風險；數值較小時，表示較少的資金被投資者資助，可能會降低借款人的還款壓力從而降低違約風險。
* last\_pymnt\_amnt：最近一次的付款金額。較高的付款金額表明借款人最近有良好的還款能力，降低違約風險；較低的付款金額可能表明還款能力差，增加違約風險。

#### **Random Forest**

表3顯示了無使用RFECV及Borderline SMOTE的Confusion Matrix數值，可以看到模型雖然能準確識別是否具有信用風險的借款人，但相較於表4有使用RFECV及Borderline SMOTE的Confusion Matrix數值，可以看出使用RFECV及Borderline SMOTE後，模型能更有效地識別具有信用風險的借款者，因此在使用REFCV搭配Borderline SMOTE可以提高預測的準確性。

|  |  |  |  |
| --- | --- | --- | --- |
| Baseline Model  Confusion Matrix | | Predicted Condition | |
| Positive | Negative |
| Actual  Condition | Positive | 208390 | 1 |
| Negative | 529 | 51802 |

表3、Baseline Model Confusion Matrix

|  |  |  |  |
| --- | --- | --- | --- |
| Our Method  Confusion Matrix | | Predicted Condition | |
| Positive | Negative |
| Actual  Condition | Positive | 208387 | 4 |
| Negative | 78 | 52253 |

表4、Our Method Confusion Matrix

一張含有 文字, 螢幕擷取畫面, 行, 字型 的圖片

自動產生的描述

圖11、Random Forest SHAP

Random Forest SHAP圖說明：

* total\_rec\_prncp：借款人已經償還的本金總額。此數值越大，表示借款人還款能力強，具有較高的信用，越可以借錢；此數值越小代表借款人還款能力較弱，就具有信用風險。
* funded\_amnt：承諾的貸款總額。此數值越大，表示借款人承諾要償還的貸款總額很大，可能具有較高的信用風險；此數值越小，表示借款人承諾的貸款總額較小，具有較低的信用風險。
* loan\_amnt：借款人申請貸款之金額，此數值越大，表示借款人需要的資金越多，較具有較高信用風險，反之數值越小，需要的資金較少，其信用風險越低。

#### **Logistic Regression**

表5顯示了無使用RFECV及Borderline SMOTE的Confusion Matrix數值，可以看到模型雖然能準確識別是否具有信用風險的借款人，但相較於表6有使用RFECV及Borderline SMOTE的Confusion Matrix數值，可以看出使用RFECV及Borderline SMOTE後，模型能更有效地識別具有信用風險的借款者，因此在使用REFCV搭配Borderline SMOTE可以提高預測的準確性，Logistic Regression也是本實驗中表現最佳的模型。

|  |  |  |  |
| --- | --- | --- | --- |
| Baseline Model  Confusion Matrix | | Predicted Condition | |
| Positive | Negative |
| Actual  Condition | Positive | 208280 | 111 |
| Negative | 523 | 51808 |

表5、Baseline Model Confusion Matrix

|  |  |  |  |
| --- | --- | --- | --- |
| Our Method  Confusion Matrix | | Predicted Condition | |
| Positive | Negative |
| Actual  Condition | Positive | 208391 | 0 |
| Negative | 7 | 52324 |

表6、Our Method Confusion Matrix

一張含有 文字, 螢幕擷取畫面, 行, 字型 的圖片

自動產生的描述

圖12、Logistic Regression SHAP

Logistic Regression SHAP圖說明：

* total\_rec\_prncp：借款人已經償還的本金總額。此數值越大，表示借款人還款能力強，具有較高的信用，越可以借錢；此數值越小代表借款人還款能力較弱，就具有信用風險。
* funded\_amnt：承諾的貸款總額。此數值越大，表示借款人承諾要償還的貸款總額很大，可能具有較高的信用風險；此數值越小，表示借款人承諾的貸款總額較小，具有較低的信用風險。
* recoveries：扣除回收後的費用。此數值越大，表示借款人扣除已還款之金額後還剩下大量的債務，就具有較低的信用；此數值越小，表示借款人已償還大量債務，就具有較高的信用。

#### **CatBoost**

表7顯示了無使用RFECV及Borderline SMOTE的Confusion Matrix數值，可以看到模型雖然能準確識別是否具有信用風險的借款人，但相較於表8有使用RFECV及Borderline SMOTE的Confusion Matrix數值，可以看出使用RFECV及Borderline SMOTE後，模型能更有效地識別具有信用風險的借款者，因此在使用REFCV搭配Borderline SMOTE可以提高預測的準確性。

|  |  |  |  |
| --- | --- | --- | --- |
| Baseline Model  Confusion Matrix | | Predicted Condiction | |
| Positive | Negative |
| Actual  Condiction | Positive | 208388 | 3 |
| Negative | 410 | 51921 |

表7、Baseline Model Confusion Matrix

|  |  |  |  |
| --- | --- | --- | --- |
| Our Method  Confusion Matrix | | Predicted Condiction | |
| Positive | Negative |
| Actual  Condiction | Positive | 208389 | 2 |
| Negative | 121 | 52210 |

表8、Our Method Confusion Matrix

一張含有 文字, 螢幕擷取畫面, 行, 字型 的圖片

自動產生的描述

圖13、CatBoost SHAP

CatBoost SHAP圖說明：

* total\_rec\_prncp：借款人已經償還的本金總額。此數值越大，表示借款人還款能力強，具有較高的信用，越可以借錢；此數值越小代表借款人還款能力較弱，就具有信用風險。
* funded\_amnt：承諾的貸款總額。此數值越大，表示借款人承諾要償還的貸款總額很大，可能具有較高的信用風險；此數值越小，表示借款人承諾的貸款總額較小，具有較低的信用風險。
* debt\_settlement\_flag：標記已註銷的借款人是否正在與債務結算公司合作。此數值越大，表示此債務人財務狀況可能較嚴峻，就具有較高的信用風險；反之此數值越小，就具有較低的信用風險。

#### **XGBoost**

表9顯示了無使用RFECV及Borderline SMOTE的Confusion Matrix數值，可以看到模型雖然能準確識別是否具有信用風險的借款人，但相較於表10有使用RFECV及Borderline SMOTE的Confusion Matrix數值，可以看出使用RFECV及Borderline SMOTE後，模型能更有效地識別具有信用風險的借款者，因此在使用REFCV搭配Borderline SMOTE可以提高預測的準確性。

|  |  |  |  |
| --- | --- | --- | --- |
| Baseline Model  Confusion Matrix | | Predicted Condiction | |
| Positive | Negative |
| Actual  Condiction | Positive | 208388 | 3 |
| Negative | 79 | 52252 |

表9、Baseline Model Confusion Matrix

|  |  |  |  |
| --- | --- | --- | --- |
| Our Method  Confusion Matrix | | Predicted Condiction | |
| Positive | Negative |
| Actual  Condiction | Positive | 208390 | 1 |
| Negative | 55 | 52276 |

表10、Baseline Model Confusion Matrix

一張含有 文字, 螢幕擷取畫面, 圖表, 行 的圖片

自動產生的描述

圖14、XGBoost SHAP

XGBoost SHAP圖說明：

* total\_rec\_prncp：借款人已經償還的本金總額。此數值越大，表示借款人還款能力強，具有較高的信用，越可以借錢；此數值越小代表借款人還款能力較弱，就具有信用風險。
* funded\_amnt：承諾的貸款總額。此數值越大，表示借款人承諾要償還的貸款總額很大，可能具有較高的信用風險；此數值越小，表示借款人承諾的貸款總額較小，具有較低的信用風險。
* debt\_settlement\_flag：標記已註銷的借款人是否正在與債務結算公司合作。此數值越大，表示此債務人財務狀況可能較嚴峻，就具有較高的信用風險；反之此數值越小，就具有較低的信用風險。

# 五、結果與討論

根據研究成果來看，在訓練模型時若有使用SMOTE以及RFECV這兩個技術，比起沒有使用的模型，不僅可以使計算量減少，還可以增加正確預測的數量，儘管在準確率或是recall的部分無法明顯看出，但Confusion Matrix還是說明了這種現象，由其是在Logistic Regression模型上結果最佳。在特徵貢獻方面，total\_rec\_prncp和funded\_amnt對模型的預測結果有明顯的影響，SHAP分析顯示，這兩個特徵在不同數據點上的影響差異較大，表明它們在影響貸款違約風險評估中扮演重要角色。

本研究搭配SMOTE以及RFECV技術去訓練多種模型，實現了一個可靠的信用風險預測模型。未來的研究可以進一步探索其他類型的機器學習模型和特徵工程技術，以不斷提高信用風險預測的準確性和適應性。對模型的解釋性進行深入分析，也是提升金融機構信用風險管理透明度和可靠性的關鍵步驟。

# 六、參考文獻

1. Lending Club Loan Data. Kaggle. 2021. https://www.kaggle.com/datasets/adarshsng/lending-club-loan-data-csv/data