

智慧型信用貸款 違約預測與評估系統

Smart Default Prediction and Evaluation System in Credit Loan

指導老師：黃登揚老師、蔡智勇老師（依姓氏筆畫排序）

報告人：李冠廷、楊紫瑄

小組成員：洪世國、李冠廷、楊紫瑄、鄒惠瑛、

黃彥翔、王芳瑜、王天舟、蘇子芸（依學號排序）

目錄 Table of Contents

0 序章

1 資料集介紹

2 資料預處理

3 機器學習

4 總結

5 信用評級檢測互動

0

序 章

Prologue

痛點 Pain Points

1. 對於非傳統意義認定且具有還款能力的借款人，金融機構錯估其違約風險較高，以致拒絕提供信貸
2. 如何以有限非平衡數據對歷史借款人的提交資料進行有效分析
3. 對於未曾與銀行往來的客戶，缺乏歷史往來紀錄，難以判斷違約機率

解決方法 Resolutions

1. 運用AI來更精準預測客戶是否會違約
2. 擴大服務至傳統認為無法放貸的客戶，找出「有收入」且「有能力還款」的潛在客戶，並針對不同類別客戶推行有效信貸方案。

系統架構

Analysis Process

資料檢視

資料清洗

第一波模型結論

機器學習

第二波模型結論

結論與展望

kaggle



colab



python™

NumPy

pandas

matplotlib



seaborn



python™

NumPy

pandas

matplotlib



seaborn



AMCHARTS

Flask
web development,
one drop at a time

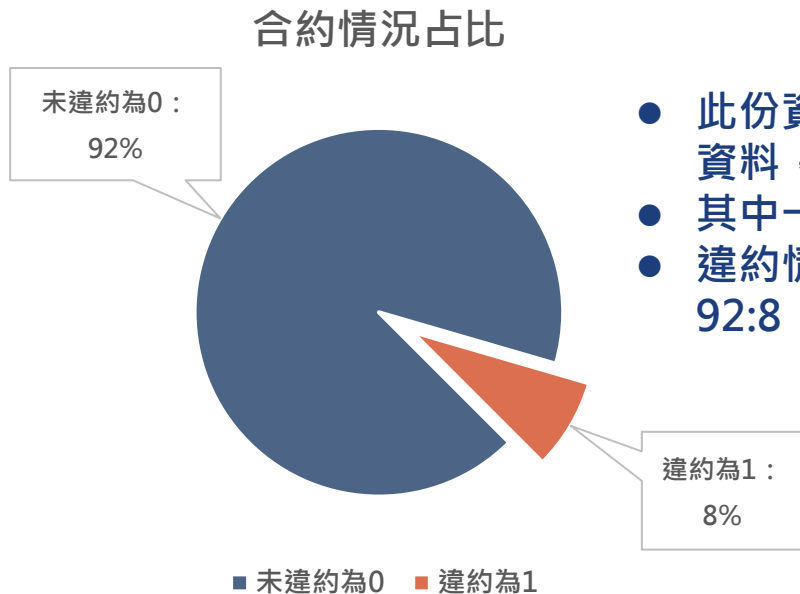


1

資料集介紹

Introduction to Dataset

資料檢視 Data Examination



- 此份資料集共有30萬筆的資料，122個欄位
- 其中一欄為違約情況註記
- 違約情況註記資料比例為92:8，資料極度不平衡

資料檢視

Data Examination



個人資訊

貸款ID
行業別
客戶職業
客戶性別
客戶年齡(天)
客戶就業天數
最高教育水平
客戶收入類型
客戶收入(年薪)
客戶的婚姻狀況
客戶扶養孩子人數
客戶家庭成員數目
申請貸款陪同者身分



個人資產

是否有車子
客戶車齡
是否有房子



貸款情況

目標變數
貸款類型
貸款額度
每期應付貸款
消費貸款金額
申請貸款時間(星期)
申請貸款時間(小時)



居住資訊

居住情況
居住區域人口狀況
居住行政區域評級
居住城市評級
客戶居住建築物的相關資訊



申請文件資料

聯絡地址
貸款所需其他文件



信用資訊

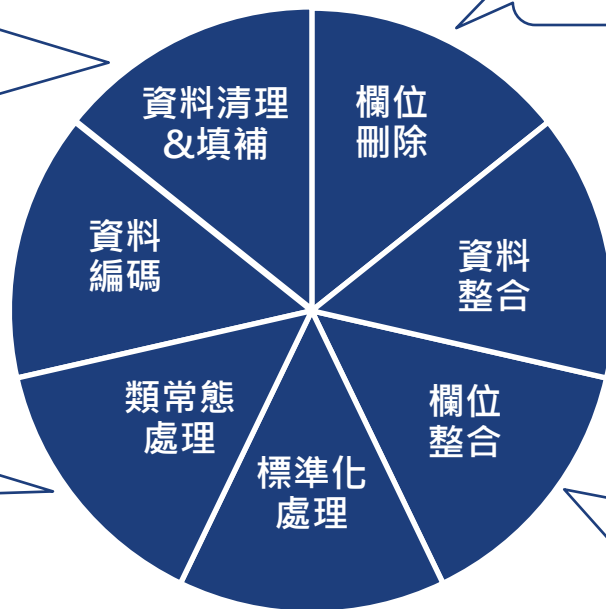
客戶逾期30天次數
客戶違約30天次數
客戶逾期60天次數
客戶違約60天次數
客戶徵信次數

將欄位分為六大類

2

資料預處理 Data Preprocessing

資料處理 Data processing



- 將徵信次數、車齡、建築物評分資訊的null值填補為0
- 類別欄位的null值超出10%的以其他字串(other)填補
- 其餘欄位的null值分別使用中位數(數值)、眾數(類別)填補

- X變數間相關性大於0.8的刪除

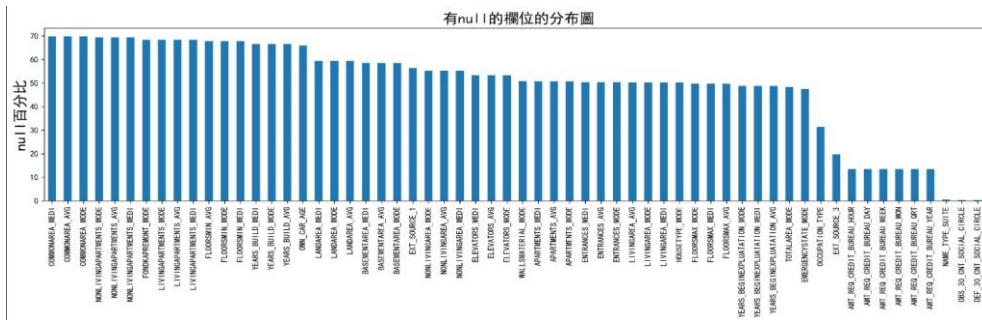
- 將申請貸款的時間轉為工作日、週末日
- 將年齡轉為以五年為一組的分類

- 將建築物訊息評分資訊合併成一個欄位，變為對整個房子的評分欄位

- 將偏度 >0.5 做類常態處理(對數、平方根、倒數轉換)

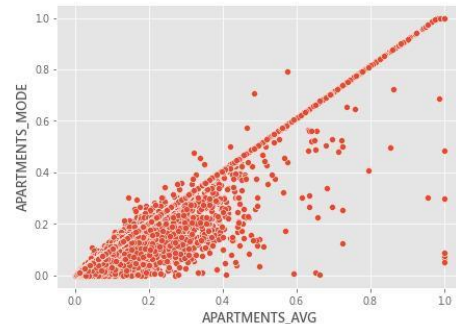
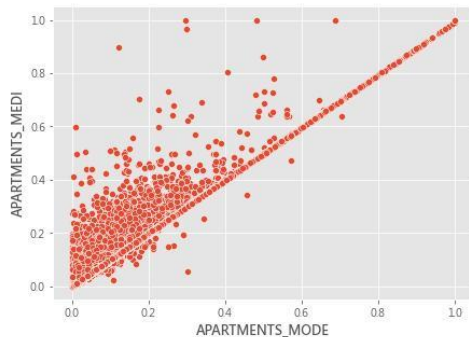
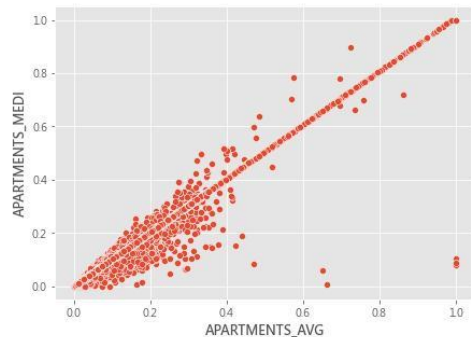
資料填補 Data Filling

1. 將徵信次數、車齡、建築物評分資訊的null值填補為0
2. 類別欄位的null值超出10%的以其他字串(other)填補
3. 其餘欄位的null值分別使用中位數(數值)、眾數(類別)填補



欄位清理 Column Cleaning

- 將X變數間的相關性 >0.8 的做欄位刪除
- 散佈圖呈左下往右上的直線趨勢時，則兩變數之間存在正相關



3

機器學習 Machine Learning

上採樣前後Oversampling before & after

資料集 : CreditV2-2

Before

| 模型 | 訓練(整體) | 測試(整體) | recall(y=1) | F1 score(y=1) |
|---------|--------|--------|-------------|---------------|
| DNN | 92% | 91.9% | 0 | 0 |
| KNN | 92.18% | 91.46% | 0.01 | 0.03 |
| 隨機森林 | 91.91% | 91.96% | 0 | 0 |
| 貝式分類器 | 89.53% | 89.63% | 0.1 | 0.13 |
| XGBoost | 91.94% | 91.99% | 0.01 | 0.02 |

After

| | | | | |
|---------|--------|--------|------|------|
| DNN | 68.65% | 66.79% | 0.62 | 0.24 |
| KNN | 95.97% | 79.42% | 0.22 | 0.15 |
| ★ 隨機森林 | 66.21% | 65.39% | 0.68 | 0.24 |
| 貝式分類器 | 61.44% | 57.19% | 0.67 | 0.20 |
| XGBoost | 65.49% | 67.19% | 0.63 | 0.24 |

資料清洗—集群&欄位合併

Data Scrubbing-Cluster & Column Merging

貸款金額
每期應繳金額
消費金額

貸款資訊

註冊日期的
變更天數
身分證件的
變更天數
聯絡方式的
變更天數

變更資訊

年收入
年齡
就業天數

收入狀態

建築物所在區
域評分的平均
數欄位

房屋評分

一小時、
一天、一週、
一月、一季、
一年的信用查
詢次數

信用查詢次數

集群前後 Cluster before & after

Before

資料集：CreditV2-2、CreditV3-3

| 模型 | 訓練(整體) | 測試(整體) | recall(y=1) | F1 score(y=1) |
|---------|--------|--------|-------------|---------------|
| DNN | 68.65% | 66.79% | 0.62 | 0.24 |
| 隨機森林 | 66.21% | 65.39% | 0.68 | 0.24 |
| 貝式分類器 | 61.44% | 57.19% | 0.67 | 0.20 |
| XGBoost | 65.49% | 67.19% | 0.63 | 0.24 |

After

| 模型 | 訓練(整體) | 測試(整體) | recall(y=1) | F1 score(y=1) |
|-----------|--------|--------|-------------|---------------|
| DNN | 67.04% | 76.31% | 0.55 | 0.27 |
| 隨機森林 | 71.62% | 69.92% | 0.64 | 0.25 |
| 貝式分類器 | 64.36% | 61.92% | 0.67 | 0.22 |
| ★ XGBoost | 69.19% | 68.85% | 0.67 | 0.26 |

特徵工程--方法一 (Wrapper)

Feature Engineering

將集群後的35個欄位，結合特徵工程挑選17個特徵去做測試
下表為使用XGBoost模型的測試結果：

| 資料集 | 訓練(整體) | 測試(整體) | recall(y=1) | F1 score(y=1) |
|----------------|--------|--------|-------------|---------------|
| creditV3-4(35) | 69.19% | 68.85% | 0.67 | 0.26 |
| creditV3-4(17) | 77.12% | 73.10% | 0.61 | 0.26 |

特徵工程--方法二 (SelectKBest)

Feature Engineering

將集群後的35個欄位，結合特徵工程分別挑選15、7、5個特徵去做測試
下表為使用XGBoost模型的測試結果：

| 資料集 | 訓練(整體) | 測試(整體) | recall(y=1) | F1 score(y=1) |
|----------------|---------------|---------------|-------------|---------------|
| creditV3-4(35) | 69.19% | 68.85% | 0.67 | 0.26 |
| creditV3-4(15) | 70.25% | 68.75% | 0.64 | 0.25 |
| creditV3-4(7) | 57.44% | 52.65% | 0.6 | 0.17 |
| creditV3-4(5) | 56.29% | 47.75% | 0.65 | 0.17 |

特徵工程--方法三 (Feature Importance)

Feature Engineering

將集群後的35個欄位，結合特徵工程挑選5個特徵去做測試
下表為使用XGBoost模型的測試結果：

| 資料集 | 訓練(整體) | 測試(整體) | recall(y=1) | F1 score(y=1) |
|----------------|--------|--------|-------------|---------------|
| creditV3-4(35) | 69.19% | 68.85% | 0.67 | 0.26 |
| creditV3-4(5) | 73.31% | 71.98% | 0.65 | 0.27 |

僅需少數欄位即可達到與原先資料相似的精準度效果

集成學習--投票法

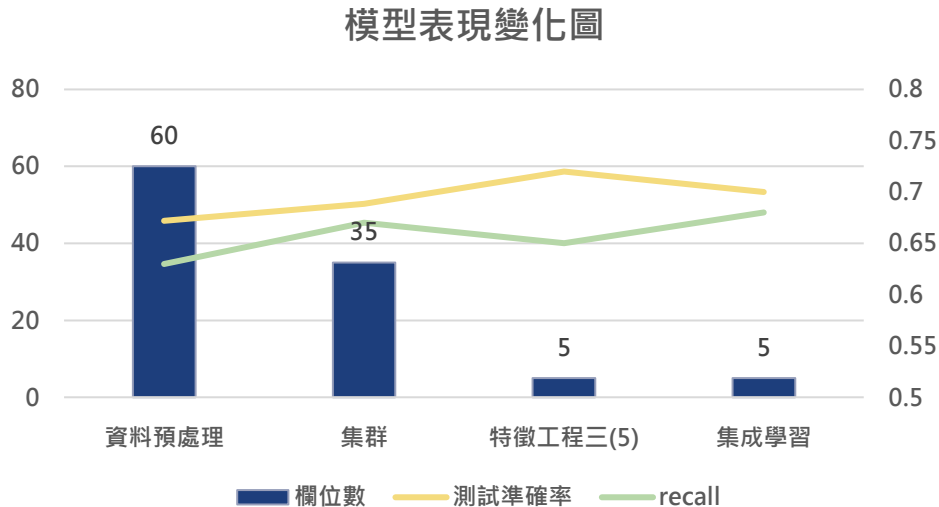
Ensemble Learning--Voting

將選擇後的五個特徵，使用集成學習來整合多個模型，解決單一模型高偏誤或高變異引起的問題，測試結果和特徵工程(5)相比，召回率(recall)有所提升
下表為使用集成學習的測試結果：

| 資料集 | 訓練(整體) | 測試(整體) | recall(y=1) | F1 score(y=1) |
|-----------------|--------|--------|-------------|---------------|
| creditV3-4(5) | 73.31% | 71.98% | 0.65 | 0.27 |
| creditV3-4(集成5) | 68.00% | 70.00% | 0.68 | 0.26 |

最終冠軍 Champion

最終，本專案選用**XGBoost**來進行對信用貸款違約率的預測，下圖為經過清理、機器學習後，欄位數及模型表現的變化趨勢圖



4

總結 Conclusion

管理意涵

Conclusion in Management

關鍵決策因子：

教育類別、信用機構評分、受雇年數、所在居住地區評分

相關應用：

本業(金融業)：信用違約預測對於相關業務運行之幫助與增值

延伸應用：「詐欺檢測」、「不良率檢測」、「預防醫學」

成員與分工

Work Distribution

| 姓 名 | 專 長 | 專案分工 |
|-----|---------------------------------------|-------------------|
| 洪世國 | 系統分析、程式設計、專案管理、AI應用、資料科學、模型、架構 | 特徵工程、模型、程式開發、任務分派 |
| 李冠廷 | Python、MySQL、數據分析、資料清洗、資料科學、演算法 | 資料清理、特徵工程、程式開發、簡報 |
| 楊紫瑄 | 數據分析、統計應用、Python、MySQL、HTML | 簡報製作、模型測試、網站前端、簡報 |
| 鄒惠瑛 | 外文（英日德法）、活動策劃、計畫管理 | 簡報製作、文稿編修 |
| 黃彥翔 | Python、MySQL、HTML(CSS)、數據分析、資料科學、資訊管理 | 網站前端、模型測試 |
| 王芳瑜 | Python、HTML、MySQL、資訊管理 | 簡報製作、模型測試、網站前端 |
| 王天舟 | 團隊協調、專案管理、數據分析、系統分析、算法模型、資料科學 | 組長、任務協調、模型測試 |
| 蘇子芸 | 財務金融、統計科學、品牌行銷、Python、MySQL | 簡報製作、模型測試 |

5

信用評級檢測互動 Interactive Default System

視覺呈現

Visual Present



掃描進互動式網站

前言

資料集介紹

資料預處理

機器學習

模型選定

總結

信用評級檢測互動區

智慧型信用貸款違約預測與評估系統

系統開發源起

現代金融機構信用風險管理之基礎和重要環節在於核准信用貸款前能有效評價和識別借款人潛在信用違約風險。計算借款人的信用違約機率，進而對借款人進行有效風險識別。傳統上，個人信用貸款的評估發放主要是基於個人資產、薪資、現金流、工作、搬遷紀錄和過去信貸記錄結合，放款機構透過以上資料辨識他們是否有拖欠償還貸款造成違約的風險。假如某個工作或行業人士需要頻繁搬遷，放款機構一般會將此類經常性的流動視為不穩定的指標，並可能會對提供信貸予這些人士持謹慎或拒絕態度。

是以，金融機構在提供客戶貸款方案時，經常面臨如下痛點：

- 1.對於非傳統意義認定且具有還款能力的借款人，金融機構錯估其違約風險較高，以致拒絕提供信貸；
- 2.如何以有限非平衡數據對歷史借款人的提交資料進行有效分析；
- 3.對於未曾與銀行往來的客戶，缺乏歷史往來紀錄，難以判斷違約機率。

為解決前述困難，本組提供以下方法，幫助金融機構提升判定客戶違約與否之正確率，達成提供更好客戶服務與增加金融機構利潤之雙贏目標。

智慧型信用貸款違約預測與評估系統

前言

資料集介紹

資料預處理

機器學習

總結

信用評級檢測互動區

信用評級檢測

測驗看看，您的客戶是否有可能違約？

教育類別：請下拉選擇 ▾

信用機構一評分：請填入0~1數值(可輸入小數)

信用機構二評分：請填入0~1數值(可輸入小數)

受雇年數：請填入整數

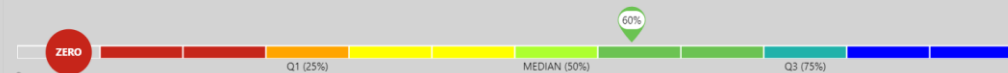
所在居住地區評分：請下拉選擇 ▾

點我做評分 →

此顧客是否會違約：會違約



此顧客的第一份信用機構評分：0.6131



此顧客的第二份信用機構評分：0.1314



此顧客的教育程度為：國高中

