# 財務工程期末專題

## 違約率預測與信用金融


組員:

科管 21 106070025 魏麗容

科管 21 106070012 周久筠

科管 21 106070020 何羿樺

資工碩一　王傳鈞

# 目錄 CONTENT

## 一、前言

馬雲以信用體系起家，點出金融的本質即為信用，因此藉助科技的力量，使無形、難以量化的信用轉為財富才是金融的烏托邦。傳統銀行體系看待借貸及兩融交易，圍繞著抵押、擔保為主的「當鋪思想」，透過有形的資產去換得信用，以規避交易對手的違約風險。

然去中心化與大數據金融的風潮興起，皆促成傳統「當鋪思想」得以轉變為基礎信用體系，使信用成為財富。近年來，台灣正迅速發展的 P2P 借貸即屬一例：資金需求方得不直接透過銀行，而透過去中心化的系統由投資方直接放款給資金需求方。而促成投資方決定是否投資的關鍵，便是透過大數據分析資金需求方的背景、交易等日常紀錄，預測出的信用分數去做判斷的基準，當然亦是由此防範信用風險。故為使得信用體系得以完善，準確且快速的預測系統是關鍵，此即為本文探討之主軸。

## 二、研究動機與目的

本組專題發展的契機在於學期中參與的一場競賽，由 AI 金融科技協會舉辦的 AI Fintech 賽事。在競賽過程中，使本組認知到信用金融發展的潛能與必然，並盡力運用 Machine Learning 的技術提升預測的準確度。至於本文之目的，一是為記錄本組在 trial and error 中尋出本組之最佳解的過程，提供更多讀者去進行進階的運用；二是期望能對信用金融的未來發展作引言，也提出本組對此項技術可用於某些金融領域的設想。

## 三、研究方法

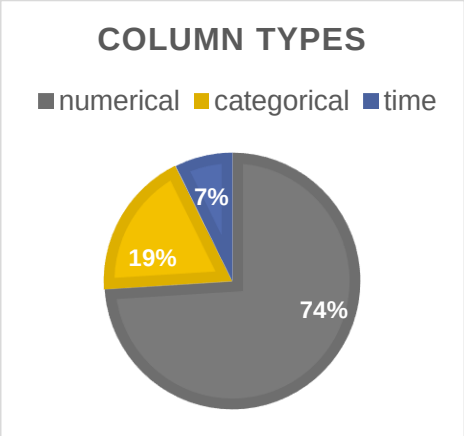
(一)資料前處理

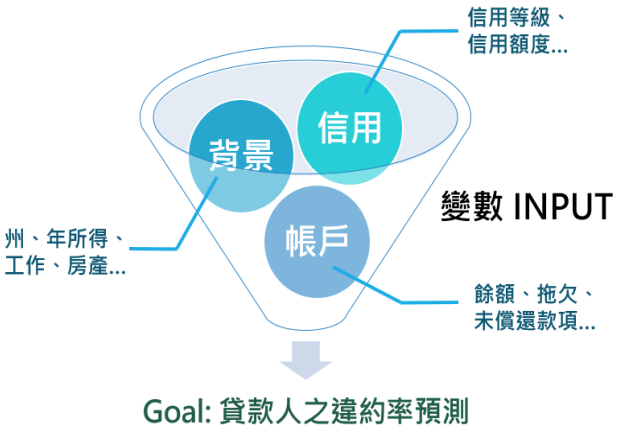
本文的數據資料是取自 AI Fintech 競賽實作組提供的 P2P 信用借貸的之去識別化數據資料，共有 x_train、y_train、x_test、y_test 四件檔案，其中 x_train 的數據規模為(2132800, 150)：由於數據規模過於龐大致 ram 無法成功負荷，同時在檢閱資料時，亦可發現數據頗髒，故在進行 ML 模型預測前，必然先對數據資料進行前處理。以下討論皆是針對 x_train 資料進行前處理的過程：

1. Column 處理

（1）Column 樣態

X_train 中包含的資料型態，共可分作連續型(Numerical Data)、類別型(Categorical Data)與時間資料(Time)，占比約如下圖所示約佔 7 : 2 : 1。另 150 種變數若作意義上的區分，可大致分作貸款人(資金需求方)的背景資訊、信用紀錄、與所持有或曾經持有之帳戶資訊；其中又可依貸款種類細分成單人/多人貸款、一般貸款/分期/循環信用額度支付、一般貸款人/次級貸款人/其他(如:紓困計畫下的貸款)，依循著不同的貸款人對於違約率預測的解釋力有些會明顯出現差異。




（2）Drop Certain Columns and Rows

排除對應變數解釋力不高的變數，能在不影響模型準確度的同時，提高模型的效率。另由於實務應用的考量，過多的變數量可能使載具無法承受、時間上也會消耗許久，故割捨掉部分變量是實踐化違約率預測的重要一步。

■ 將 NA 值過多的欄列 drop

由於數據中約有 200 萬筆資料，故我們將變量數據少於 50%的資料(NA 值大於 100 萬筆)先挑選出來(約佔 3 成)，並非直接刪除，而是兼具衡量變數意義與下列幾點綜合判斷後才作出決定。

註: >100 萬筆:0~100 萬:0 筆 = 3:2:5

■ 將常理認與違約率(Y)無關的欄列 drop

資料裏貸款人背景資訊中，存有一些方便中介機構或系統驗證所需的個人化資訊，如識別證號碼、Member ID、郵編號碼、產品編號等資訊。而顯然這些欄位並無助於違約率預測的判斷，故可先行排除。

■ 將難以數字化的 data drop

諸如貸款原因、紓困計畫申請理由與貸款描述等資料，由於每筆資料皆因貸款人不同而有不同的呈現方式(屬於開放式回覆的類型)，故顯然非屬機器能夠辨別之語言。故縱這些資料應對違約率預測有顯著的貢獻，然因此牽涉到 Natural Language Processing 的技術，不為本文探討的重心，故也先排除在 data set 外。

■ 將共線性過高的欄列擇一 drop

在模型中，倘若任兩或多個自變數間存在過高的共線性，可能會使模型預測出的結果與係數產生偏誤，故下表為 x_train 中各變量間的 correlation matrix，並依共線性程度的不同，顯示不同的顏色深淺。此次挑選亦須綜合變數意義判斷，對共線性過高的判斷依準先以相關係數高於 0.5 挑出。

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | | loan_amnt | funded_amnt | nded_amnt_i | installment | annual_inc | dti |
| 2 | loan_amnt | 1 | 0.99999949 | 0.99999417 | 0.94298754 | 0.19773912 | 0.05022713 |
| 3 | funded_amnt | 0.99999949 | 1 | 0.99999468 | 0.94298817 | 0.19773935 | 0.05022683 |
| 4 | funded_amnt_inv | 0.99999417 | 0.99999468 | 1 | 0.94291694 | 0.19774424 | 0.05021538 |
| 5 | installment | 0.94298754 | 0.94298817 | 0.94291694 | 1 | 0.19171744 | 0.05264408 |
| 6 | annual_inc | 0.19773912 | 0.19773935 | 0.19774424 | 0.19171744 | 1 | -0.08609204 |
| 7 | dti | 0.05022713 | 0.05022683 | 0.05021538 | 0.05264408 | -0.08609204 | 1 |
| 8 | delinq_2yrs | -0.01762024 | -0.01761991 | -0.01764829 | -0.00300421 | 0.02551699 | -0.01850268 |
| 9 | fico_range_low | 0.11661966 | 0.11661988 | 0.11669953 | 0.05566632 | 0.03215376 | -0.00459086 |
| 10 | fico_range_high | 0.11661898 | 0.1166192 | 0.11669885 | 0.05566625 | 0.03215347 | -0.00459373 |
| 11 | inq_last_6mths | -0.03291278 | -0.03291215 | -0.0329553 | -0.00965823 | 0.02214352 | -0.00941374 |
| 12 | mths_since_last_delinq | -0.00216082 | -0.00216192 | -0.00212581 | -0.0160767 | -0.03373003 | 0.01850664 |
| 13 | mths_since_last_record | 0.04692558 | 0.04692558 | 0.04702316 | 0.02576205 | -0.03056862 | 0.06993983 |
| 14 | open_acc | 0.17647036 | 0.17647098 | 0.17645239 | 0.167348 | 0.1014118 | 0.16463379 |
| 15 | pub_rec | -0.06461364 | -0.0646133 | -0.0646429 | -0.052201 | -0.00847713 | -0.02768237 |
| 16 | revol_bal | 0.32485875 | 0.32485877 | 0.32485086 | 0.31304195 | 0.20037621 | 0.10011008 |
| 17 | total_acc | 0.18949621 | 0.18949604 | 0.18947448 | 0.16970981 | 0.11914986 | 0.13306554 |

2. NA 值處理
3. 類別型資料轉換
4. Over-sampling 的處理

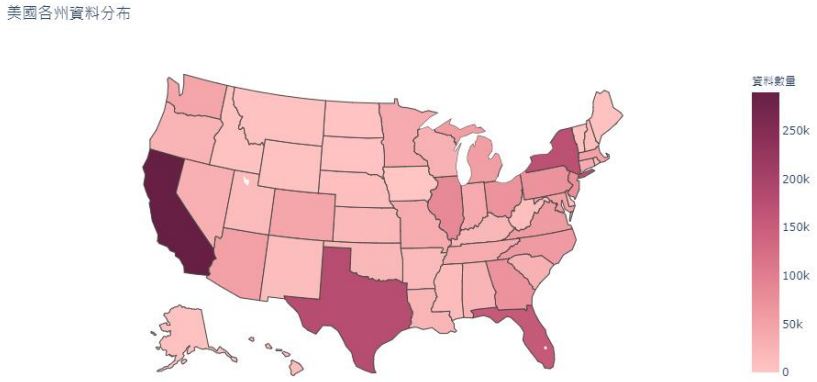
(二)模型建構與預測

## 四、實證結果與探討

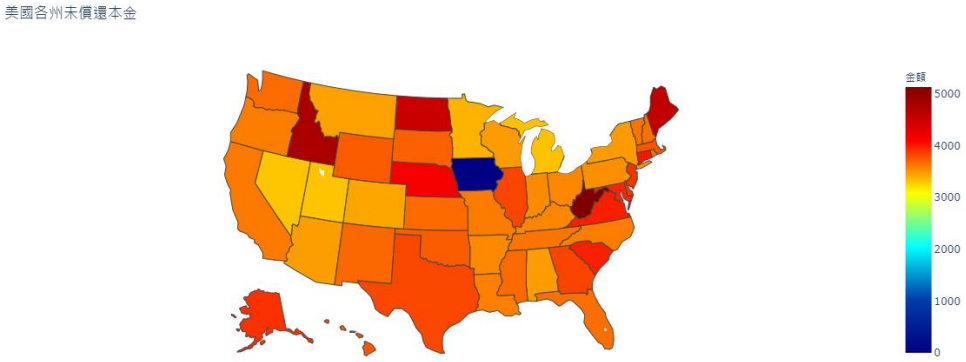
(一) AI-Fintech 競賽預測表現

(二) EDA 分析

選擇幾項具有代表性且資料較完整的變數：['state' 資料所屬州, 'grade' 信用等級, 'delinq' 欠款金額, 'loan' 借款金額, 'fico_low' 貸款發放時的信用分數所屬下限, 'remain' 未償還本金]，經過整理後有 50 萬筆左右的資料。



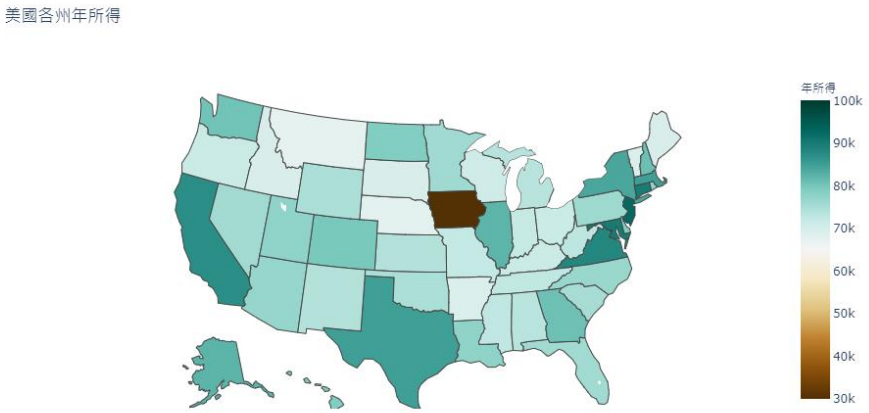

從這筆資料在美國各州的分佈可以看出其非常不均勻的情況，在加州(CA)的資料數量有高達 29 萬筆，在愛荷華州(IA)卻僅僅只有一筆，而且這一筆在後面的討論可以看到它非常的不正常，在美國的中北部地區都有資料相當不足的問題，由於視覺效果的關係，可能會有點不明顯，但是滑鼠游標移動到蒙大拿(MT)、愛達荷(ID)或是阿拉斯加(AK)等區域，它們的資料數量甚至只有幾千筆，但是在後面的交易數量上卻比平均要高，可能是該銀行在收集資料的時候不夠隨機分佈，它的資料可能可以去推估是否為 influential point 或 outliers.



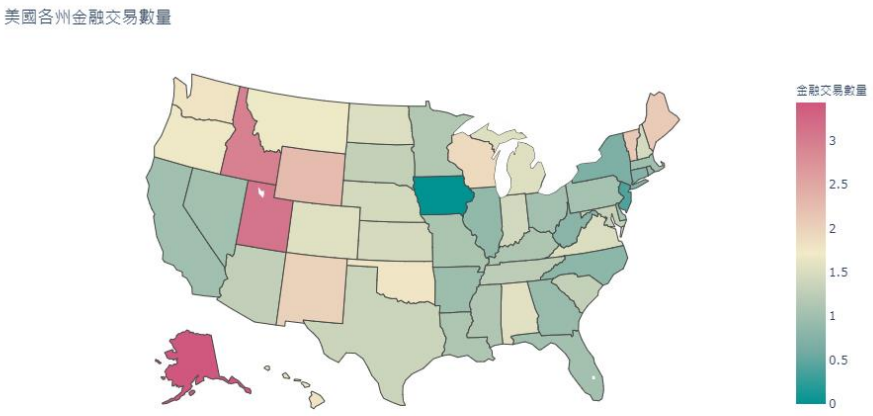

西維吉尼亞州盛產煤礦。主要產業為能源和製造業、天然氣豐富，且是全美硬木主要供應來源。農業是北達科他州的最大產業，此外還有煉油、食品加工及技術工業也是主要產業。緬因州主要出產農產品和漁業；工業方面，旅遊業和戶外娛樂是緬因州經濟的重要組成部分，包括打獵、釣魚、雪地車、滑雪、野營、徒步等。

從這張圖我們可以看出在 WV(西維吉尼亞)州的未償還比例最高，接著分別是(ND)北達科他和(ME)緬因州；而 IA(愛荷華)州因為資料較少，故在圖上顯示的值最低。根據產業以及經濟分析，這些未償還本金數目最高的幾個州皆有從事工業、能源業和農業，而這些

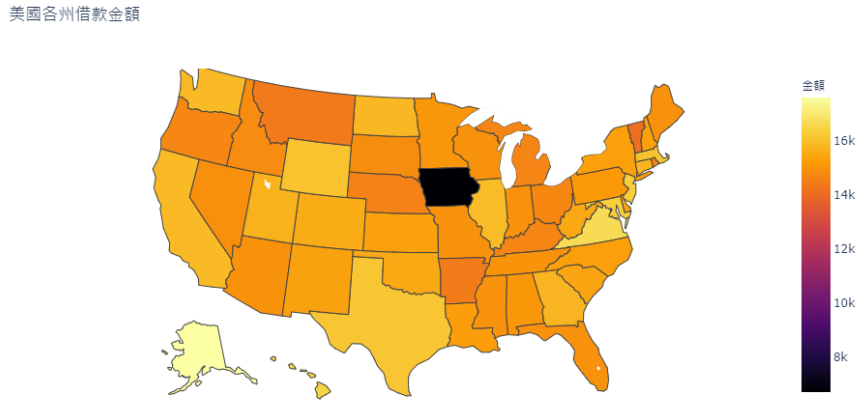
產業經常需要在初期投入大量資本，但無法在短時間內大量的賺回本金還款，因此在未償還本金方面會有較高的占比。



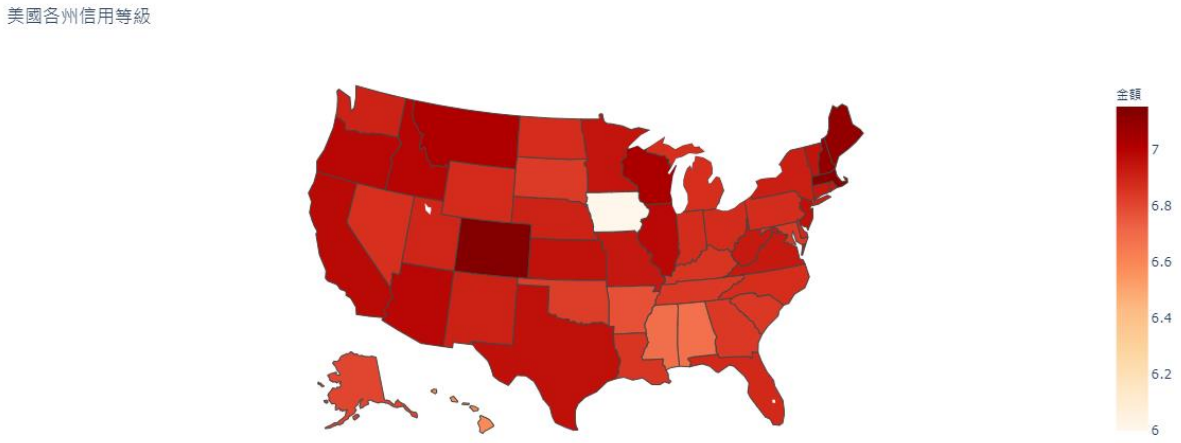

康乃狄克(CT)， 紐澤西(NJ)， 德州(TX)， 加利福尼亞(CA)， 馬里蘭(MD)，維吉尼亞州(VA)為收入較高的州，對比未償還本金的各州分布圖，可以發現這幾周的未償還本金的比例也是偏高，由於收入更高，這些州的居民更有可能借款或是融資並進行較大金額的投資(EX.房地產)，而若要進行這些投資就會時常需要投入資金，並且還款時間較長，故會有較高的未償還本金。若是繼續對照信用評等的圖，可以發現這幾州的信用評等分數也是更高的，可以推測高收入是銀行有信心在這些州拿回借的本金和利息的原因。



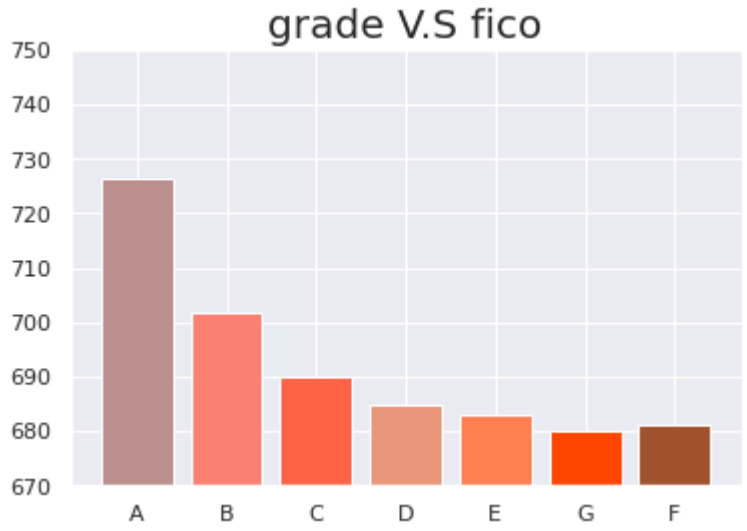

猶他(UT)、愛達荷(ID)、阿拉斯加州(AK)三個州為金融交易數量較多的州，對照各州的年所得，可以發現年所得平均居於中上，但交易次數較多，可能是次數較多但較小金額的交易，推測是小額的投資或是商業活動，可能是新興的商業發展區域，故在未來或許會有更大的發展潛力。對比上述所得較高的州，康乃狄克(CT)， 紐澤西(NJ)， 德州(TX)， 加利福尼亞(CA)， 馬里蘭(MD)，維吉尼亞州(VA)，在這裡的交易次數卻較低，故推測可能是商業活動已趨於成熟，居民傾向花更多錢置產。



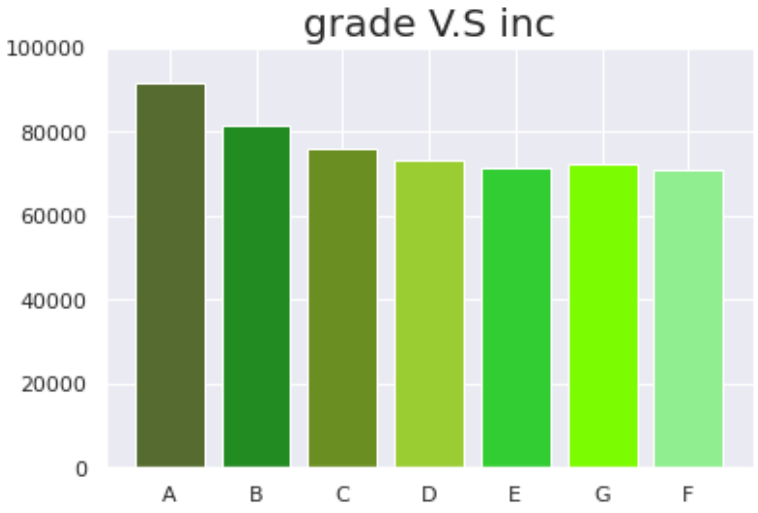

最顯眼的就是愛荷華州(IA)的那塊黑色，但是由於資料量過低的關係，在此先忽略不計，圖中數據是以平均的方式去計算借款金額，想要知道借款金額的平均量額是否會因為地區而有所不同，圖中越淺黃色的借款金額就越大，越深色的金額就越小。
最有趣的是阿拉斯加(AK)的金額 17.6K 明顯的高於其他州，遠高於第二名維吉尼亞州(VA)的 16.66K，而佛蒙特(VT)的借款平均金額為全美最低的，可以在後面討論借款金額和年所得的關係，但是圖中沒有借款金額明顯的地域關係。



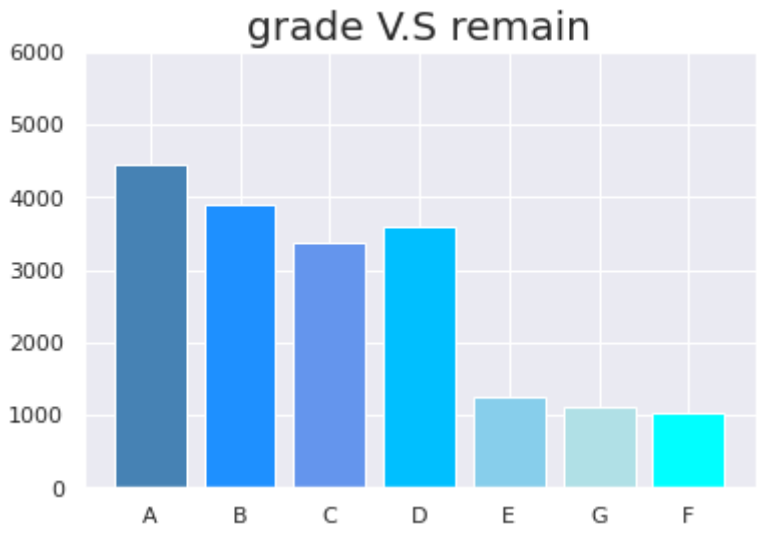

越深紅色的州就有越高的平均信用等級，所以在美國中部的科羅拉多(CO)其平均信用等級為 7.14 全美國最高，而美國東北角的幾個州像是麻州(MA)、新罕布夏(NH)和緬因(ME)也有較高的平均信用等級，阿拉巴馬(AL)、密西西比(MS)和阿肯色(AR)則是有稍低的平均信用等級。



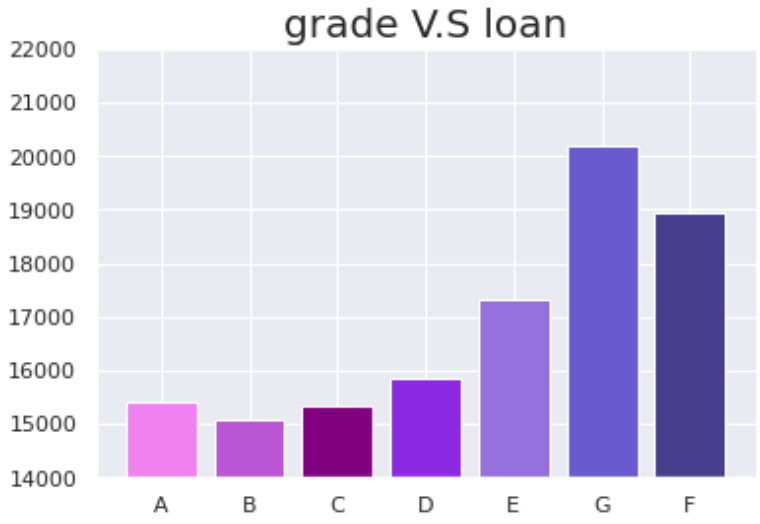

這張圖的 x 軸為信用等級，而 y 軸為借款人在貸款發放時的信用分數所屬下限，且多分布於 200 到 800 之間，可以直覺的理解信用等級越高，信用分數會越高，而核貸時的分數下這張圖的 x 軸為信用等級，而 y 軸為借款人在貸款發放時的信用分數所屬下限，且多分布於 200 到 800 之間，可以直覺的理解信用等級越高，信用分數會越高，而核貸時的分數下限也就會越高，但是 F 的值卻稍微高於 G，是個值得探討原因的部分。fico 這個變數最有趣的地方在於它和比賽中要預測的違約率在統計上有相當高的關聯性。




從此圖可以清楚看出信用等級和收入成正比，因為收入較高，貸款方會更願意相信解款人的償債能力，並且若對照上述的各州信用等級圖和收入圖，我們同樣可以發現收入較高的州會有較高的信用。




從圖上可以看出若有愈高的信用評等，則未償還本金的金額愈高，這可以解釋為因為借款人較有信用，故銀行願意讓他延長還款時間，未償還的金額也就越多。但由圖可知信用評等 D 的欠款金額高於 C，推測可能是因為信用較差而導致欠款較高。在 E、F、G 三個信用等級之下，因信用更差，故銀行限縮其還款期限，導致未償還本金較少。




上圖為信用等級和貸款金額之間的關係，特別是這張圖的值呈現正向的關係，信用等級越低，貸款金額的平均金額就越高，在 G 等級有最高的借款平均金額，初步評估是因為由於利率和對金錢需求的關係，由於無法得知信用等級實際運算的公式，所以評估為 A 等級的人可能不會是有大量借款需求的用戶，雖然信用等級越低貸款利率的利率越高，但是高風險用戶仍願意貸更多的款項。

（三）信用金融應用設想

1. 發想處

此次競賽的數據龐大而雜，資料中還存在許多 NA 值，導致在預測信用違約率的難度上升許多，預測的準確度亦會有所爭議。由此，數據的多寡顯然對貸款人的信用等級預測準確度有顯著的重要性。

有鑑於此，如何將平台與金融機構得到的數據量增加為一重要之課題，而參照日前開放銀行的資訊授權架構（亦即將資訊自主權歸還於客戶，唯有經客戶親自的授權，各類合作企業才得藉開放 API 共享數據，進行更多創新）的概念，本組認為此可將此授權機制輔以獎勵措施，以獲取更多類型的數據，從而進行進階的分析與預測。

2. 受益方

（1）投資者: 若個人願意將數據資料、交易資訊一部或全部的授權於平台或機構，便得享有更多金融服務的體驗與相關優惠抵減。

（2）平台或金融機構: 藉個人交易資訊與行為數據的授權，不僅個人信用等級預測的準確度上升，也使機構對該同類型(對個人的基本資料進行分類)的族群得以生成更準確的違約率預測。

## 五、結論