P2P平台智慧型投資決策系統

Smart P2P Investment Decision Support System

指導教師：蔡智勇、黃登揚

執行團隊：林盈均、周秀樺、黃品婷

許睿倩、張國屏、劉韋廷

楊孟儒、黃聖哲、康博鈞

LendingClub

執行團隊 MEET OUR TEAM

黃聖哲

Pandas, SQL, Hadoop

張國屏

Pandas, VTK, JS
Machine Learning

林盈均

PMP, SQL
Machine Learning

黃品婷

Linux, SQL, Hadoop

楊孟儒

Python, Hadoop,
Machine Learning

康博鈞

Pandas,SQL,ML

許睿倩

SQL, Python

周秀樺

Python, SQL
Oracle, D2K

劉韋廷

Pandas,SQL,Hadoop
Pro/E

摘要大綱

透過機器學習進行數據分析，以利助網絡借貸(peer to peer)的投資者提升投資獲利；亦可協助借貸中介平台管控借貸人

簡報流程：

背景介紹

系統架構

資料分析

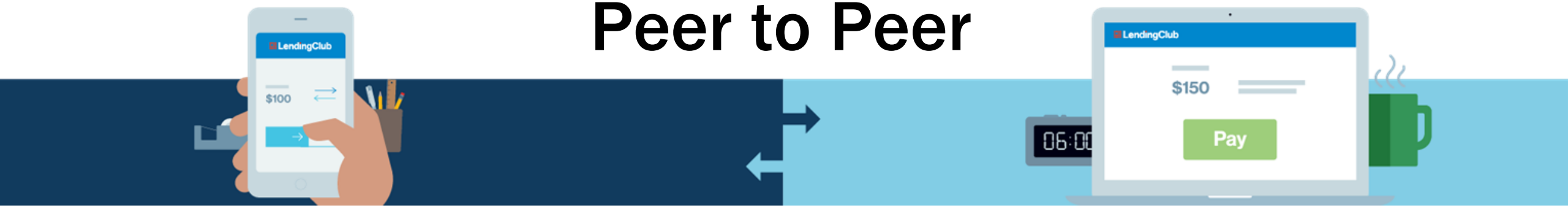
結論報告

背景介紹

background

網絡借貸

以網絡為管道，透過中介機構的媒合，出借人(投資客)與貸款人(借款人)實現直接借貸、資訊互動、資信評估等服務

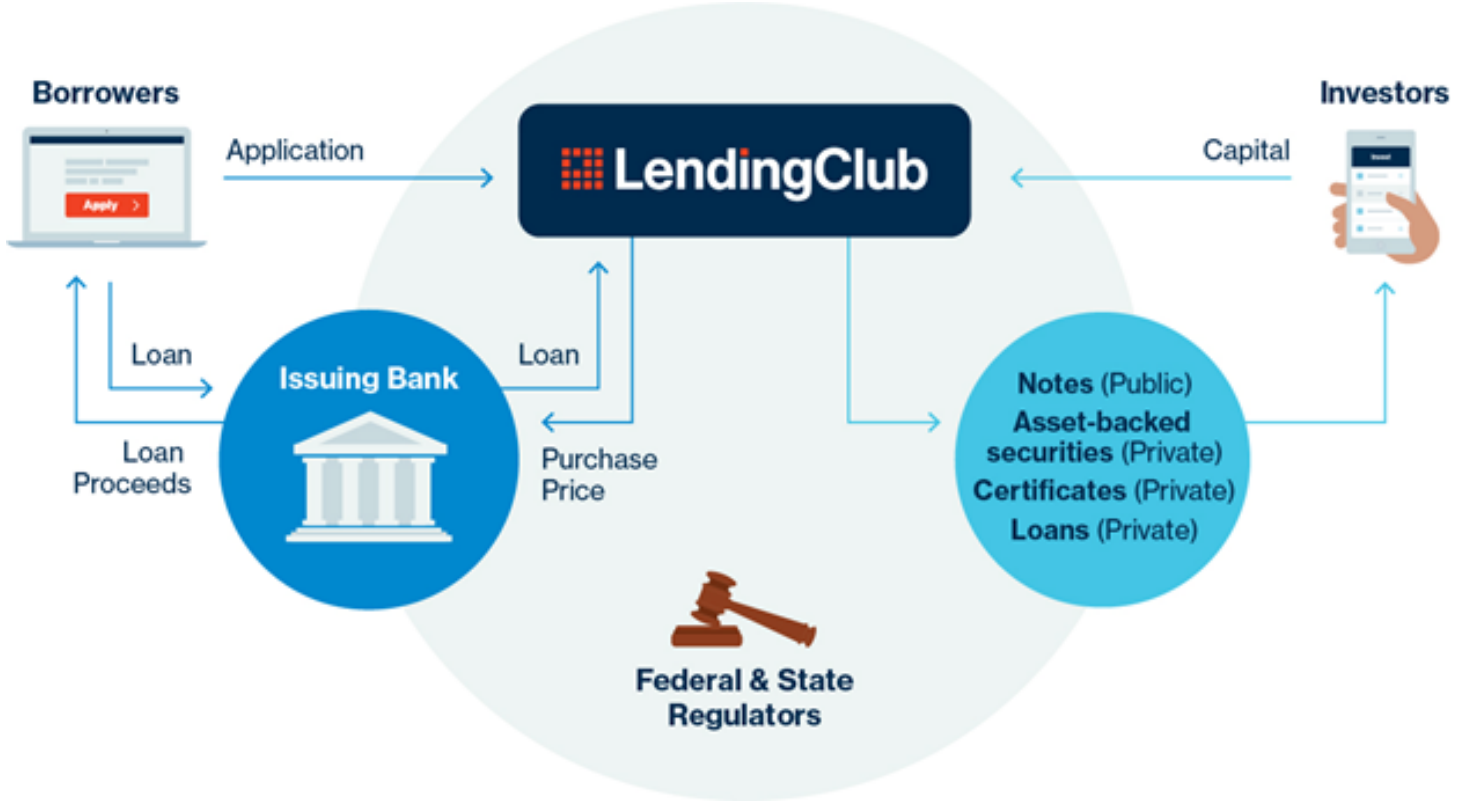


出借人(投資客)

貸款人(借款人)

借貸俱樂部 LendingClub

- 為美國一家提供線上平台作為交易服務媒介的公司。
- 依照債務人的信用資訊計算出每期應償還之利息及本金。
- 再將債務總額分割為小金額債券，供眾多投資人參酌選購。



當前挑戰

貸款人過高的呆帳率，導致約有34億美金的款項無法討回

	TOTAL ISSUED	CHARGED OFF (NET)
A	$6,321,329,325	$159,424,950
B	$9,401,557,325	$562,829,442
C	$9,773,557,750	$1,047,848,438
D	$5,096,087,575	$819,304,308
E	$2,365,386,850	$521,865,936
FG	$1,046,289,775	$306,242,270
All	$34,004,208,600	$3,417,515,344

總發行放款額

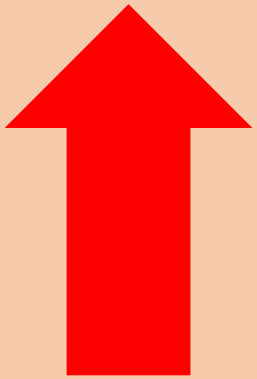
34 Billion

總呆帳額度

3.4 Billion

呆帳率

10%



當前挑戰

對於借貸俱樂部

1. 降低公司的信用程度
2. 減少投資者的投資比例
3. 降低公司與同業的競爭力

對於投資者

1. 增加高風險的違約放款交易
2. 增增加投資到呆帳客戶的機率

團隊目標

1.預測不良貸款

讓投資人能夠利用預測結果，避免不良貸款發生，降低投資風險。

2.分析惡意貸款的原因

讓平台管理者，能以預測結果，預防惡意貸款的發生，並減少後端催討的時間、相關人力的成本。

3.預測投資獲利率

以投資人的觀點，分析由不同的投資組合在外部經濟情況的變動下，投資獲利率之變化。

系統架構

system

資料處理流程

資料取得

LendingClub

USA.gov

kaggle

資料清洗

vmware

ubuntu

hadoop HDFS

HIVE

Apache HBase

資料儲存

MariaDB FOUNDATION

DB Browser for SQLite

SQLite

資料預處理

python

ANACONDA

jupyter

Apache Zeppelin

機器學習資料分析

pandas

NumPy

TensorFlow

scikit learn

Seaborn

matplotlib

資料視覺化

AMCHARTS

HTML

Flask web development, one drop at a time

資料處理流程


資料視覺化
機器學習資料分析
資料預處理
資料清洗
資料取得
相關係數分析
虛擬編碼處理
類別資料轉換
資料常規化
AMCHARTS
HTML
pandas
NumPy
python
TensorFlow
ANACONDA
scikit learn
jupyter
Seaborn
matplotlib
Flask
LendingClub
USA.gov
kaggle
vmware
HIVE
Apache HBase
SQLite
Apache Zeppelin

資料處理流程

資料取得

資料清洗

機器學習資料分析

預測	Logistic Regression, DNN, SVM
決策	Random Forest, J48, XGBoost
關聯	Aproni, FP-tree
集群	K-means

資料視覺化

資料清洗

原始資料觀察

- 欄位共154欄，總計2,146,835筆資料。
- 欄位空值率大於30%，予以刪除。
- 欄位空值率小於30%：
  a. 類別型態的筆數予以刪除(年資、職業等)
  b. 數值型態的筆數補以數值0或刪除該列

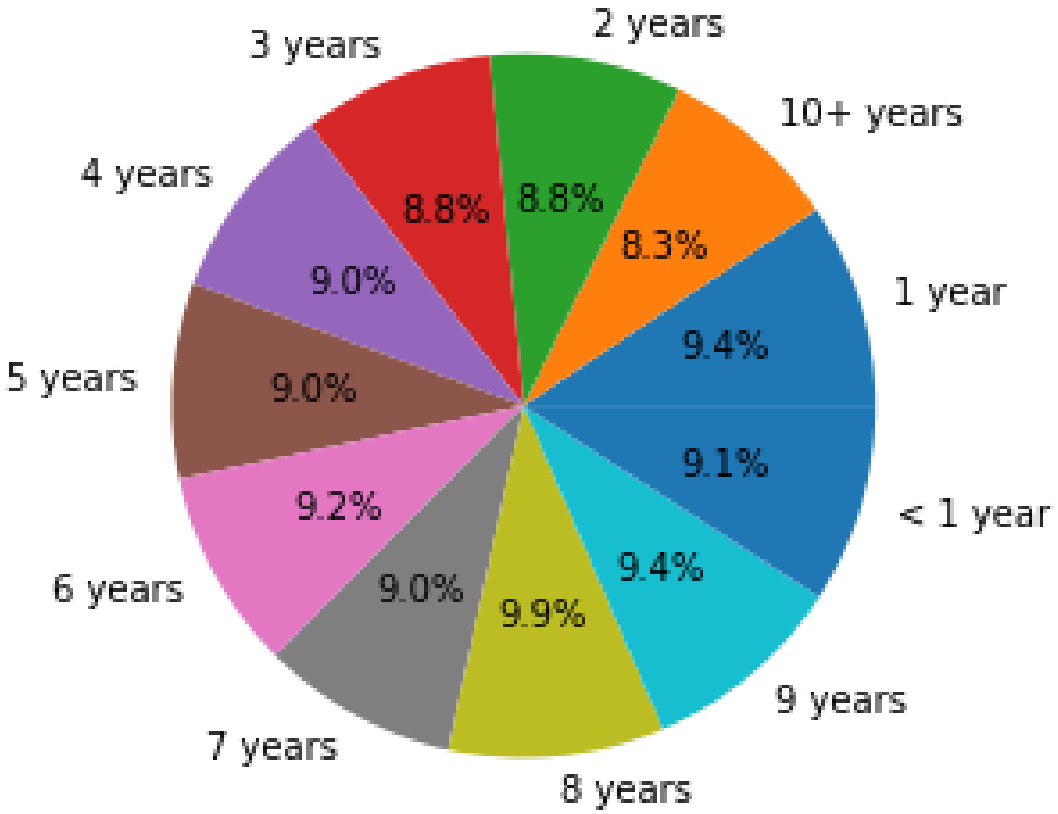
機器學習目標

- 預測不良貸款：
  原有8種貸款狀態；本專案分成好帳與呆帳2種
- 惡意呆帳原因：
  呆帳細分一般呆帳與惡意呆帳，以缺繳期間6期為界
- 預測投資獲利：
  利用投資人投資金額與投資收入計算出年獲利率

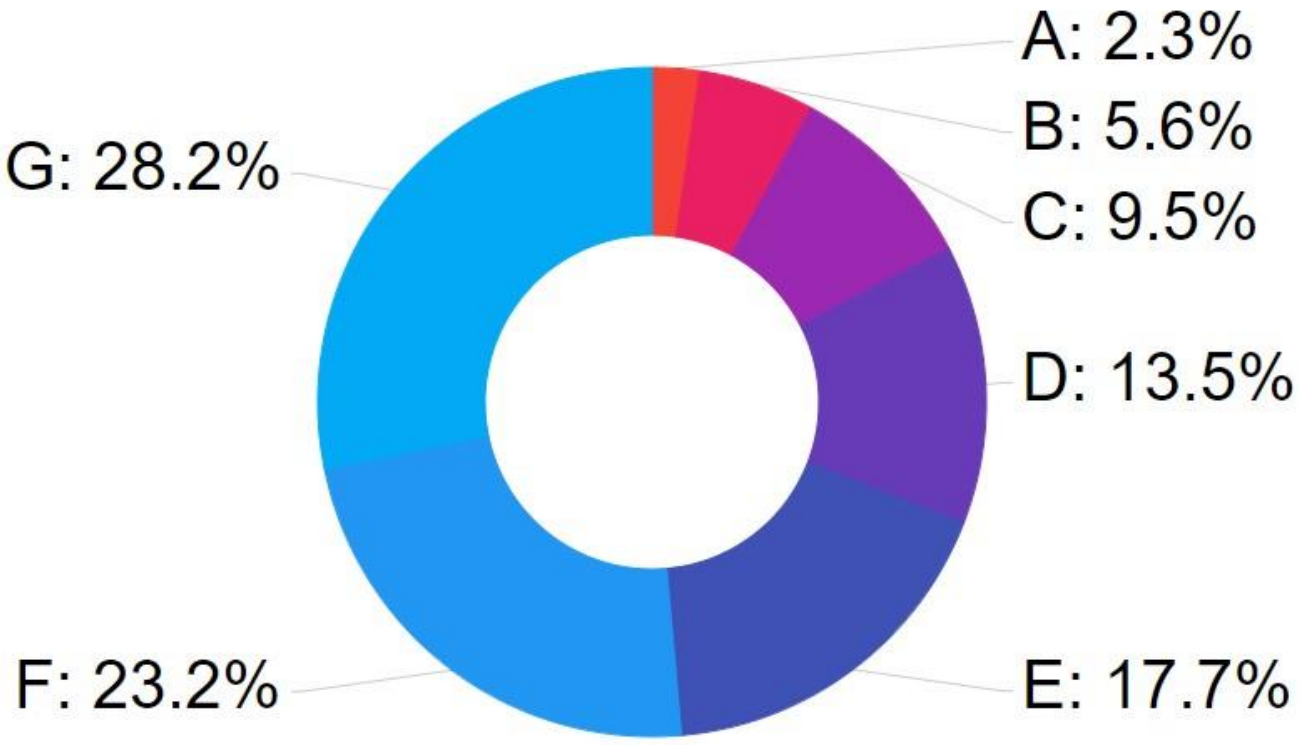
不良貸款

bad loan

不良貸款預測 欄位敘述性統計



工作年資



信用評等

不良貸款預測 欄位資料分類

除了作圖觀察，也採用屬性評估(Attribute Eval)的Ranker功能找出關聯性強的欄位。接著把關聯性低、發生在Loan_status之後的欄位刪除。

[年收入]採用美國稅率分級，共分成6種;
[信用評等]具有依序性，使用label encoder分7級

```
#data['annual_inc']
for i in range(len(data['annual_inc'])) :
    if int(data.loc[i,'annual_inc']) < 9700 :
        data.loc[i,'annual_inc'] = 0
    elif int(data.loc[i,'annual_inc']) >= 9700 and int(data.loc[i,'annual_inc']) < 39475 :
        data.loc[i,'annual_inc'] = 1
    elif int(data.loc[i,'annual_inc']) >= 39475 and int(data.loc[i,'annual_inc']) < 84201 :
        data.loc[i,'annual_inc'] = 2
    elif int(data.loc[i,'annual_inc']) >= 84201 and int(data.loc[i,'annual_inc']) < 160725 :
        data.loc[i,'annual_inc'] = 3
    elif int(data.loc[i,'annual_inc']) >= 160725 and int(data.loc[i,'annual_inc']) < 204100 :
        data.loc[i,'annual_inc'] = 4
    else :
        data.loc[i,'annual_inc'] = 5
```

[借貸目的]，[取款方式]，[貸款期]，[房屋所有權]，[財產驗證]採虛擬編碼

purpose_vacation	purpose_wedding	term_ 36 months	term_ 60 months
0	0	1	0
0	0	1	0
0	0	1	0
0	0	0	1
0	0	1	0

不良貸款預測 資料分析結果

演算法	訓練準確 (門檻:1)	測試準確 (門檻: 0.88)	猜對好帳機率 (門檻: 0.88)	猜對呆帳機率 (門檻: 0.12)
RandomForest	**1**	**0.71**	**0.71**	**0.67**
J48	**0.924**	**0.63**	**0.63**	**0.64**
logistic regression	**0.7**	**0.7**	**0.71**	**0.63**
multilayer perceptron	**0.81**	**0.58**	**0.57**	**0.68**
LibSVM	**1**	**0.13**	**0**	**1**

惡意貸款

worst Loan

惡意貸款-資料觀察、清洗

annual_inc	loan_amnt		purpose	int_rate	dti	delinq_2yrs	grade	il_util	percent_bc_gt_75	pub_rec	all_util
NaN	NaN	unique	14	NaN	NaN	NaN	7	NaN	NaN	NaN	NaN
NaN	NaN	top	debt_consolidation	NaN	NaN	NaN	B	NaN	NaN	NaN	NaN
NaN	NaN	freq	5807	NaN	NaN	NaN	2945	NaN	NaN	NaN	NaN
7.641296e+04	14446.297500	mean	NaN	13.268677	18.417900	0.314800	NaN	24.129600	43.886200	0.224400	22.536000
5.928890e+04	8748.034331	std	NaN	4.829080	14.218758	0.863467	NaN	36.393148	36.488332	1.043721	31.257738
1.000000e+03	1000.000000	min	NaN	5.310000	1.000000	0.000000	NaN	0.000000	0.000000	0.000000	0.000000
4.700000e+04	8000.000000	25%	NaN	9.750000	12.000000	0.000000	NaN	0.000000	0.000000	0.000000	0.000000
6.500000e+04	12000.000000	50%	NaN	12.740000	18.000000	0.000000	NaN	0.000000	40.000000	0.000000	0.000000
9.140850e+04	20000.000000	75%	NaN	16.020000	24.000000	0.000000	NaN	60.000000	75.000000	0.000000	51.000000
3.120000e+06	40000.000000	max	NaN	30.990000	818.000000	20.000000	NaN	158.000000	100.000000	86.000000	136.000000

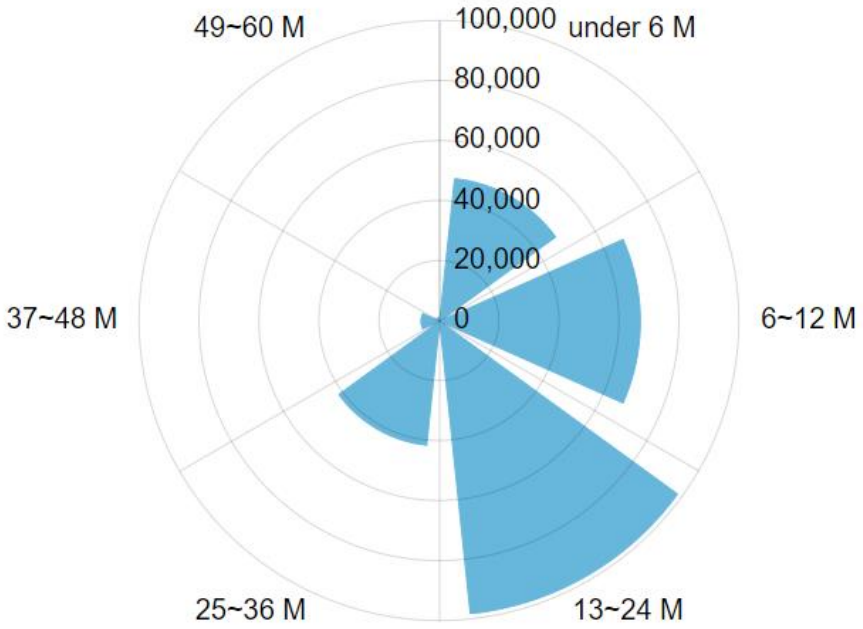
正規化Log10

補空值, 數值類別化, 分級、分群

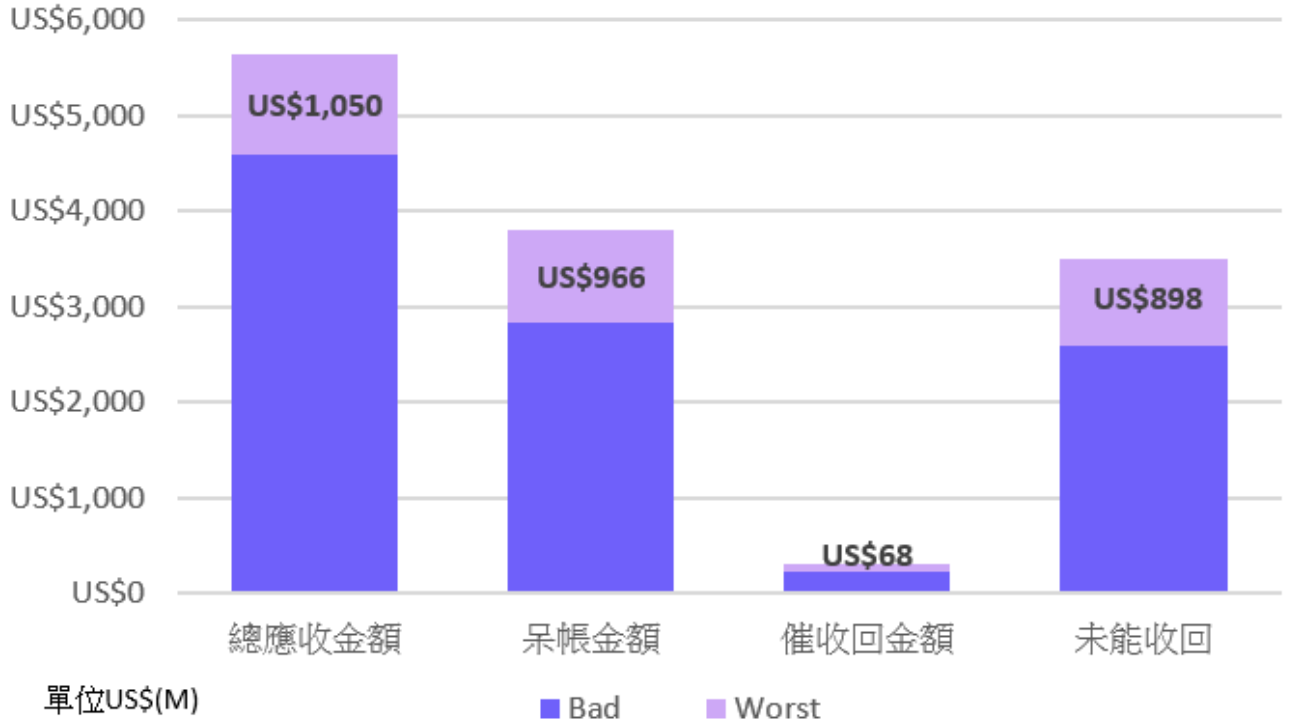
- annual income
- loan amount
- purpose
- interest rate
- dti
- delinquent in 2 yrs
- grade
- il util
- percent bc gt 75
- public record

惡意貸款原因 欄位敘述性統計

呆帳期分佈圖

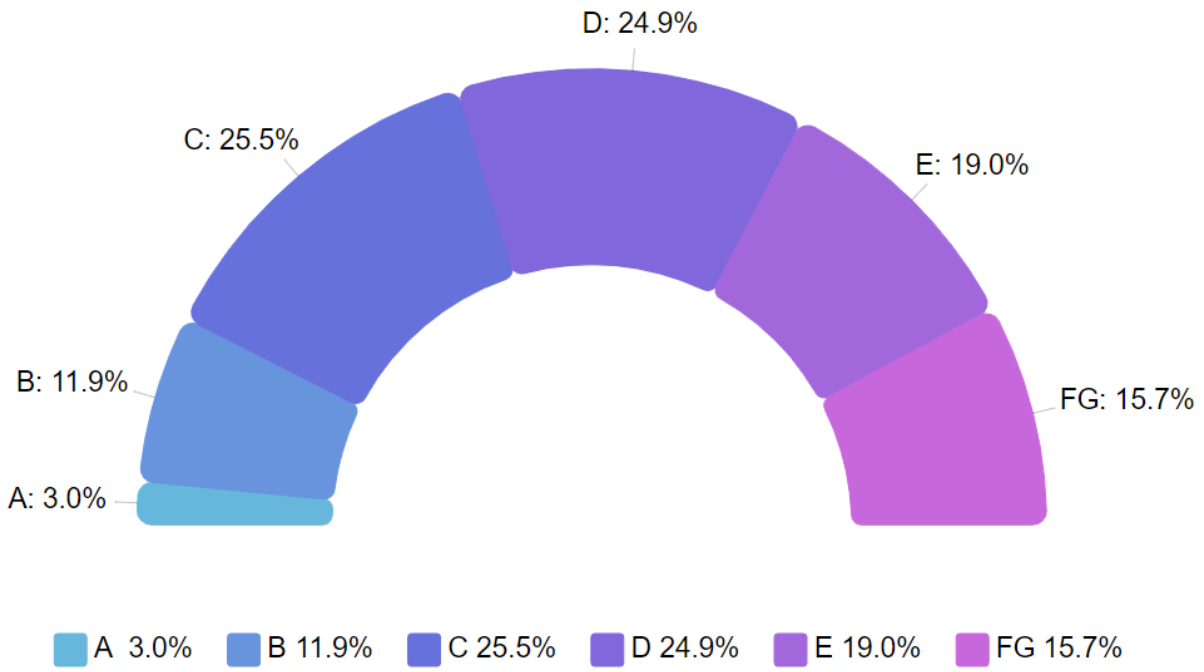


惡意呆帳損失

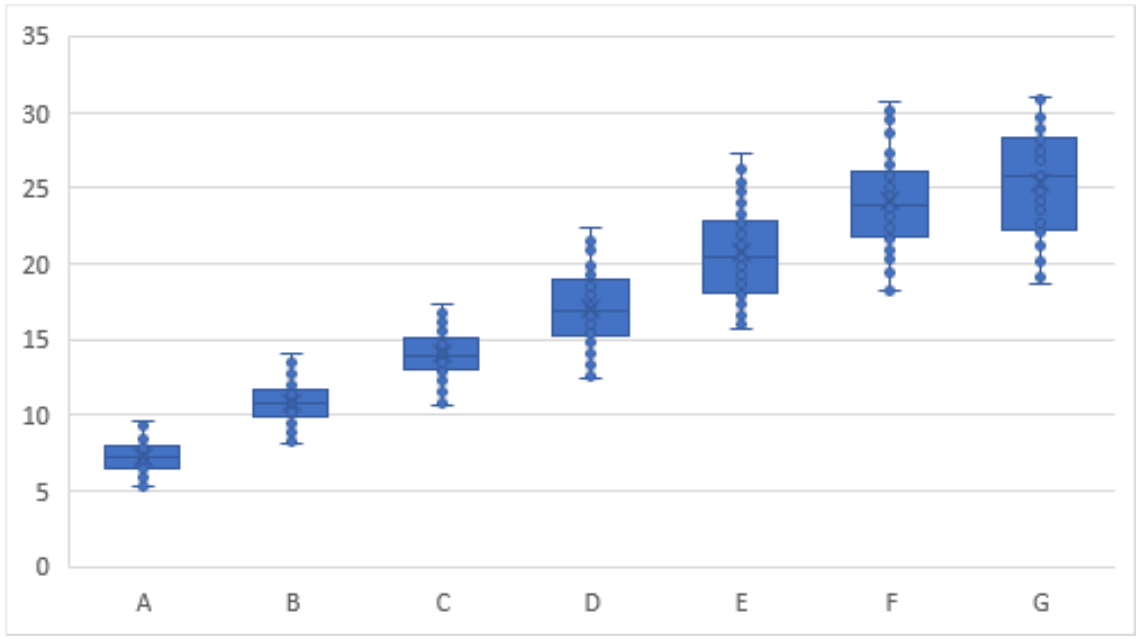


惡意貸款原因 欄位敘述性統計

惡意呆帳金額by Grade

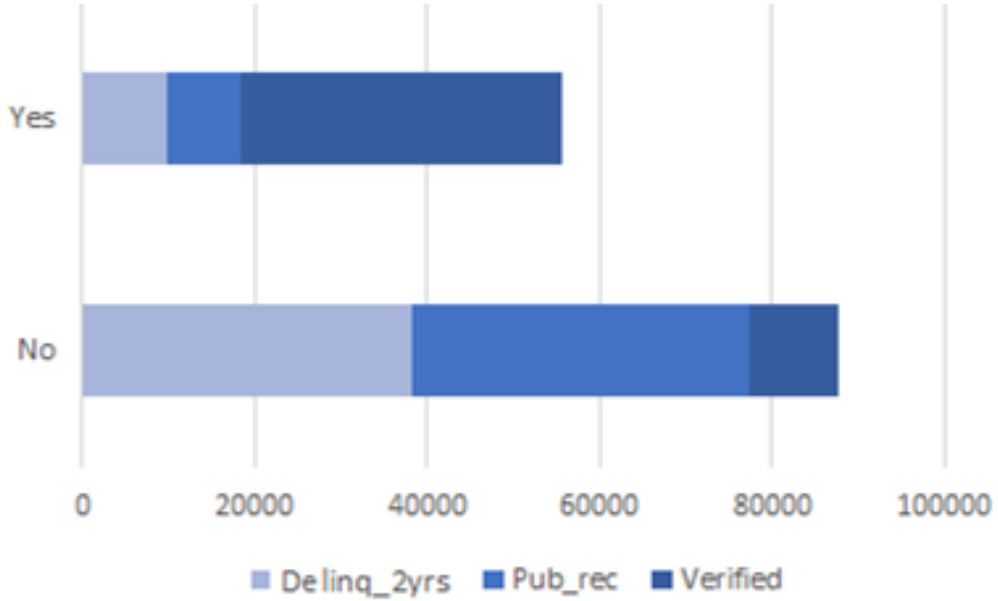


利息分佈圖



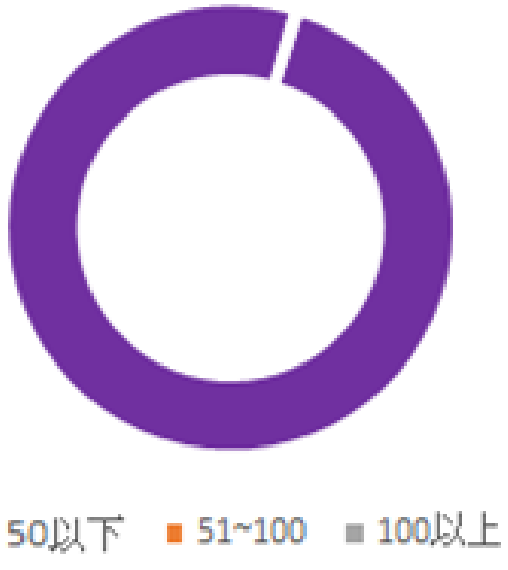
惡意貸款原因 欄位敘述性統計

兩年延期付款次數、公共記錄、收入驗證



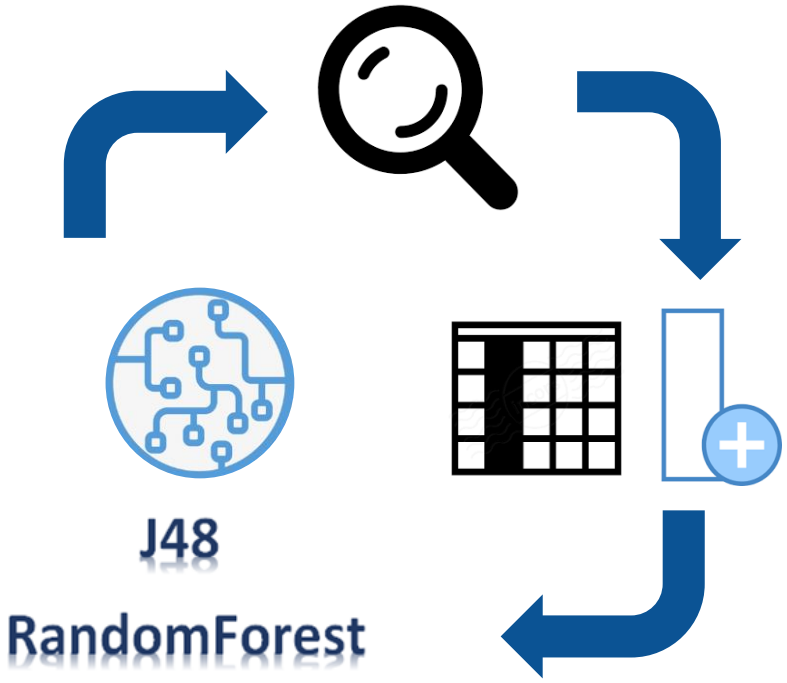
銀行帳戶使用率

收入負債比



惡意貸款-機器學習

檔名	檔名	模型	參數	訓練	測試	RMK	BAD	Worst
BWtrain(77)v2	BWtest(77)v2	RandomForest	default	100.0%	70.7%			
BWtrain(77)v2	BWtest(77)v2	RandomForest	iteration 150	100.0%	73.0%			
BWtrain(77)v2	BWtest(77)v2	RandomForest	iteration 200	100.0%	72.3%			
BWtrain(77)v2	BWtest(77)v2	RandomForest	iteration 180	100.0%	71.7%			
BWtrain(77)v2	BWtest(77)v2	RandomForest	iteration 160	100.0%	72.3%			
BWtrain(77)v2	BWtest(77)v2	RandomForest	iteration 140	100.0%	72.3%			
BWtrain(77)v2	BWtest(77)	J48	default	96.4%	56.7%		0.508	0.833
BWtrain(77)v2	BWtest(77)	RandomForest	default	100.0%	51.0%		0.463	0.722
BWtrain(77)v2(d)	BWtest(77)v2(d)	J48	default	85.0%	70.3%		75%	0.5
BWtrain(77)v2(d)	BWtest(77)v2(d)	J48	default	84.4%	67.7%	remove purpose	0.703	0.556
BWtrain(77)v2(d)	BWtest(77)v2(d)	J48	default	84.9%	70.3%	recover purpose, remove RR_max	0.748	0.5
BWtrain(77)v2(d)	BWtest(77)v2(d)	RandomForest	default	100.0%	71.0%		0.748	0.537
BWtrain(77)v2(d)	BWtest(77)v2(d)	RandomForest	default	100.0%	74.0%	removed in_fi	0.785	0.537
BWtrain(77)v2(d)	BWtest(77)v2(d)	J48	default	84.6%	76.7%		0.85	0.389
BWtrain(77)v2(d)	BWtest(77)v2(d)	J48	default	85.6%	67.0%	removed never deliquent	0.724	0.426



惡意貸款原因 資料分析結果

規則	信賴程度	準確率
all_util > 1, dti<=1 (所有帳戶使用率高，然而債務比低者)	13%	92%
all_util > 1, dti>1, int_rate > 3 (所有帳戶使用率高，債務比高者，且利率高者)	19%	85.9%

投資報酬

Investment

預測投資獲利率-合併外部資料表

dti	term	Date	ECRI	Effective Federal Funds Rate	1 Year Treasury Rate	AVG. INTEREST RATE	ADJ. NET ANNUALIZED RETURN1
16	36 months	17-Oct	-0.0517	0.14	0.37	0.0689	0.0457
23	36 months	15-Aug	-0.9391	0.37	0.59	0.0681	0.04
6	36 months	16-Jul	-0.1382	0.12	0.28	0.0698	0.0444
8	36 months	14-Jan	-0.387	0.34	0.53	0.067	0.0419
25	36 months	15-Jan	0.0705	0.08	0.11	0.0788	0.0541
17	36 months	15-Feb	0.0399	0.07	0.13	0.0788	0.0541
10	36 months	15-Oct	-0.1126	0.14	0.16	0.0777	0.0557
13	36 months	17-Jun	-0.1382	0.12	0.28	0.0698	0.0444
13	36 months	16-Oct	-0.1517	0.12	0.48	0.0685	0.0444
40	36 months	16-Oct	-0.4566	0.24	0.54	0.0685	0.0444
15	36 months	17-Dec	-0.1517	0.12	0.48	0.0685	0.0444
7	36 months	14-May	-0.1475	0.38	0.66	0.067	0.0419
21	36 months	12-Oct	-0.1011	0.11	0.23	0.0718	0.0473
24	36 months	13-Nov	0.1734	1.69	2.27	0.0672	0.0372
19	36 months	17-Aug	-0.387	0.34	0.53	0.067	0.0419
18	36 months	16-Oct	0.0073	0.11	0.22	0.0718	0.0473
21	36 months	13-May	-0.0721	0.08	0.11	0.0719	0.0408
27	36 months	18-Oct	0.0089	0.09	0.11	0.073	0.0502
6	36 months	15-Feb	0.0089	0.09	0.11	0.073	0.0502

ECONOMIC

LC Performance

預測投資獲利率-資料觀察、清洗

	delinq_amnt	installment	annual_inc	Effective Federal Funds Rate	1 Year Treasury Rate	ADJ. NET ANNUALIZED RETURN1
mean	12.7671	443.259244	7.611808e+04	0.354720	0.539458	0.055832
std	710.3572	262.863906	5.101982e+04	0.431485	0.528749	0.016472
min	0.0000	30.120000	2.400000e+03	0.070000	0.100000	-0.090100
25%	0.0000	250.240000	4.505625e+04	0.110000	0.150000	0.046800
50%	0.0000	382.170000	6.500000e+04	0.140000	0.370000	0.053700
75%	0.0000	589.220000	9.200000e+04	0.390000	0.650000	0.064300
max	65000.0000	1501.000000	1.320000e+06	4.760000	3.500000	0.112800

正規化 Log10

補空值、數值類別化、分級、分群

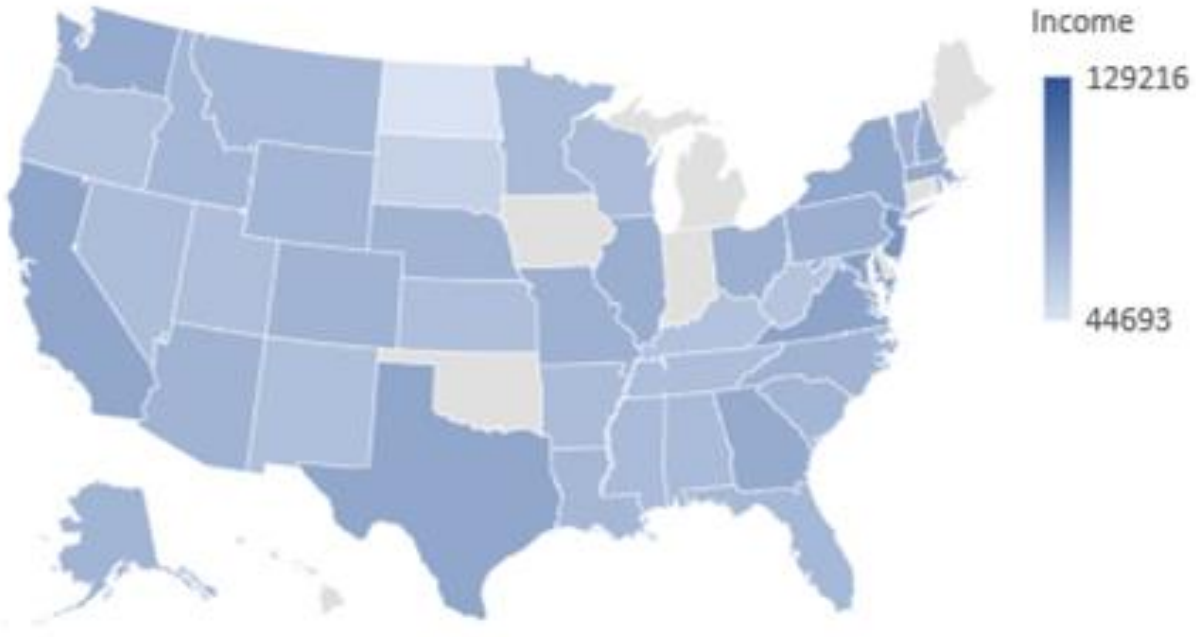
預測投資獲利率-增加Y欄、四分位

投資收入	投資金額	YEAR	NAR
20,215.8	20,000	3	0.03596484375
2,675.83	2,675	3	0.001034365264797507
15,063	15,000	3	0.014
8,053.55	8,000	3	0.022312418619791666
5,013.65	5,000	3	0.009099934895833333
13,020.7	12,950	3	0.018200762146074648
10,014.3	10,000	3	0.004763346354166666
9,502.85	9,500	3	0.0009998629385964911

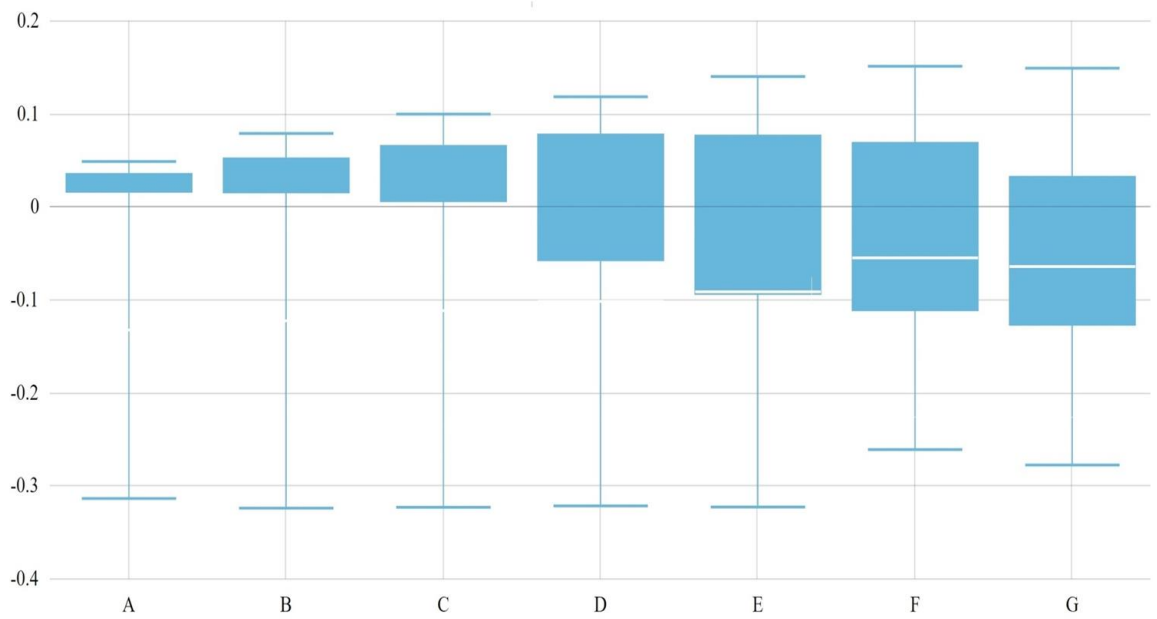
$$\frac{\text{投資收入-投資金額}}{\text{投資金額*YEAR}}$$

預測投資獲利 敘述性統計

美國各州平均年收入

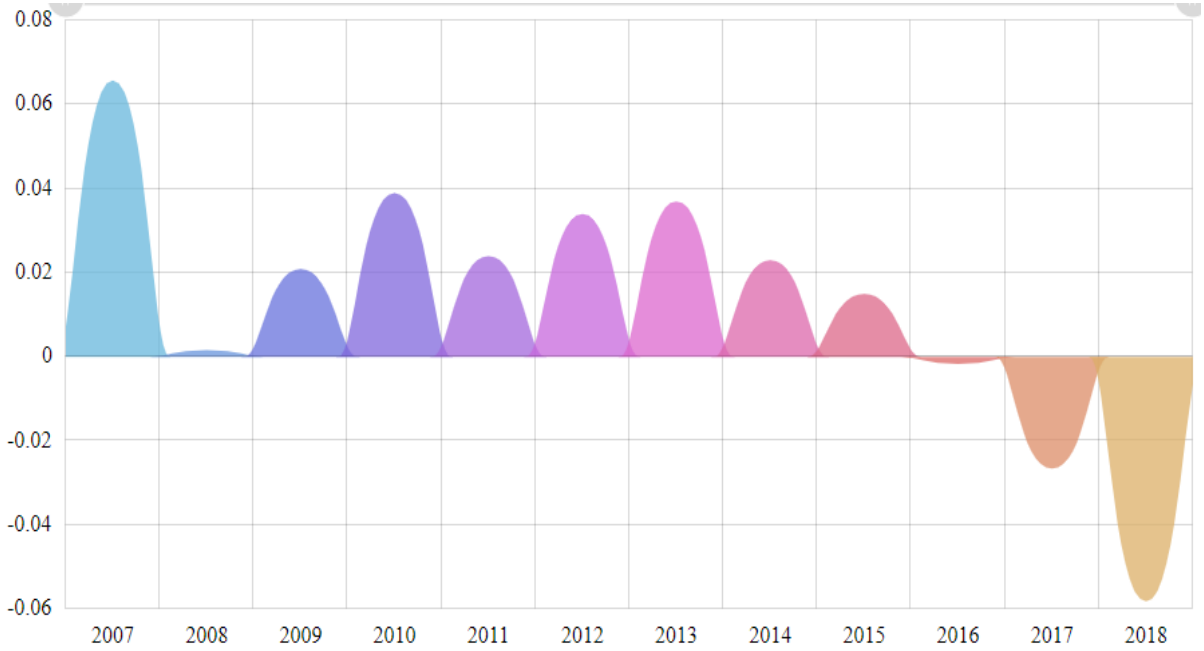


信用等級獲利率



預測投資獲利 敘述性統計

每年平均NAR



美國經濟同時指標

預測投資獲利 資料分析結果

規則	信賴程度	準確率
Effective Federal Funds Rate: > 0.3752 低報酬	31.70%	80.90%
Effective Federal Funds Rate: < 0.3752 grade B 高報酬	19%	77.84%
Effective Federal Funds Rate: < 0.3752 grade C 高報酬	21.56%	75.67%
Effective Federal Funds Rate: < 0.3752 grade D 高報酬	12.57%	70.60%

預測投資獲利 資料分析結果

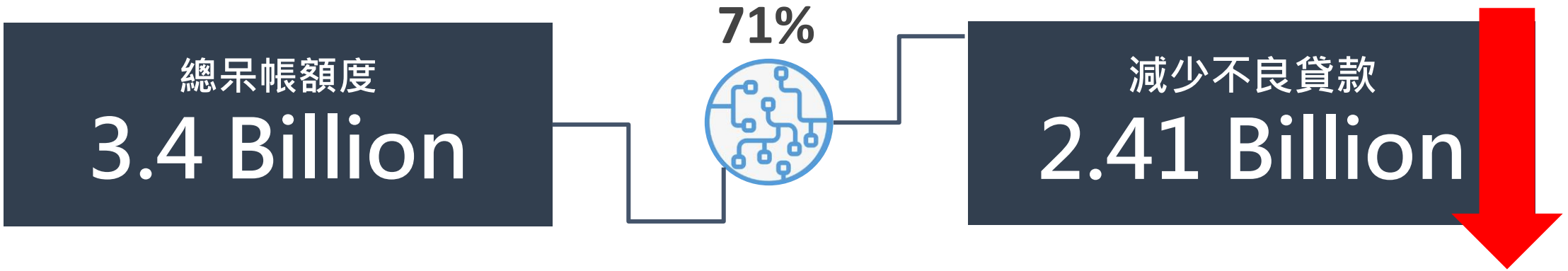
結論報告

conclusion

結論

1. 透過隨機森林演算法與機器學習，我們預測呆帳的準確率比以往提高 5倍(約7成)，可幫借貸俱樂部減少約25億的損失

1. 不同於以往呆帳率發生在以還卡債、以債還債為目的；本團隊發現呆帳率發生在以"臨時"急需"大筆現金"、"一次性之日常消費"為目的的族群



結論

透過決策樹有9成2的準確率，分辨出繳款未滿六個月的族群的決策因子。
建議平台在當下列情況發生時。增加檢核方式，以減少惡意貸款的借款。

1. 帳戶使用率高，債務比低者
2. 帳戶使用率高，債務比高者，且利率高者

不動產、房屋型態

提升FICO分數

延長查核年限

加權決策因子後
付予新的等級

結論

透過決策樹有8成準確率，分辨出投資報酬高、低。

1.當聯邦基準利率偏高時，投資風險偏高，建議投資其他標的
2.當聯邦基準利率偏低時，推薦投資人投資B、C、D等級的標的

Effective Federal Funds Rate: > 0.3752

謝謝聆聽