資料基本敘述

資料共 916567 筆，包含 2007~2017 的資料，2007~2011 年共 24972 筆、2012~2013 年共 175621 筆、2014 年 202495 筆、2015 年 269927 筆、2016 年 175406 筆、2017 年 68146 筆。下圖為各變數之基本統計敘述：

	annual_inc	emp_length	dti	delinq_2yrs	loan_amnt	term
Mean	77151	5.98	17.99	0.32	14568.37	41.89
Std	67617	3.68	8.71	0.88	8586.4	10.33
Min	100	0	-1	0	500	36
25%	47600	3	11.75	0	8000	36
50%	65000	6	17.46	0	12300	36
75%	92000	10	23.77	0	20000	36
Max	9550000	10	999	39	40000	60

loan_stat

Fully Paid	731079
Charged Off	185462
Default	26

grade

A	153113
B	264371
C	257999
D	140314
E	69203
F	24793
G	6774

inq_last_6mths

0	509275
1	253607
2	99072
3	38881
4	10931
5	3870
6	870
7	43
8	17

purpose

debt_consolidation	543275
credit_card	198919
home_improveme	56934
other	49008
major_purchase	19128
small_business	10771
medical	9592
car	9539
moving	6283
vacation	5729
house	4568
wedding	1927
renewable_energy	626
educational	268

變數轉換

loan_stat

Fully_paid -> 0

Defult, Charged-Off -> 1

Grade

A, B, C, D, E, F, D -> 1, 2, 3, 4, 5, 6, 7

Purpose

debt_consolidation -> 1

Other -> 0

各年分 LoanStat 比例

2007~2011

0	21483	86%
1	3489	14%

2012~2013

0	147716	84%
1	27905	16%

2014

0	165488	82%
1	37007	18%

2015

0	207879	77%
1	62048	23%

2016

0	131142	75%
1	44263	25%

2017

0	57370	84%
1	10776	16%

方法一：

原始資料共有 916567 筆，包含 2007~2017 年的資料，此方法依照原本數據排列順序，以及隨機打亂數據分別做測試。將資料切分為 8:2 做為訓練與測試集，訓練集共 733252 筆資料，測試集共 183314 筆資料。 為了解決數據不平衡，對訓練集使用 SMOTE 重新採樣，以改善原始數據過採樣的問題。
預測模型使用 Random Forest，測試過後訓練結果有過擬合的問題，因此模型參數 n_estimators 和 max_depth 不宜過大，分別設定為 100 和 18。

方法一訓練結果：依照原本數據排列順序

	Score
0_precision	0.80
0_recall	0.96
1_precision	0.43
1_recall	0.11
accuracy	0.78
預測 0 比例	94.3%
預測 1 比例	5.7%

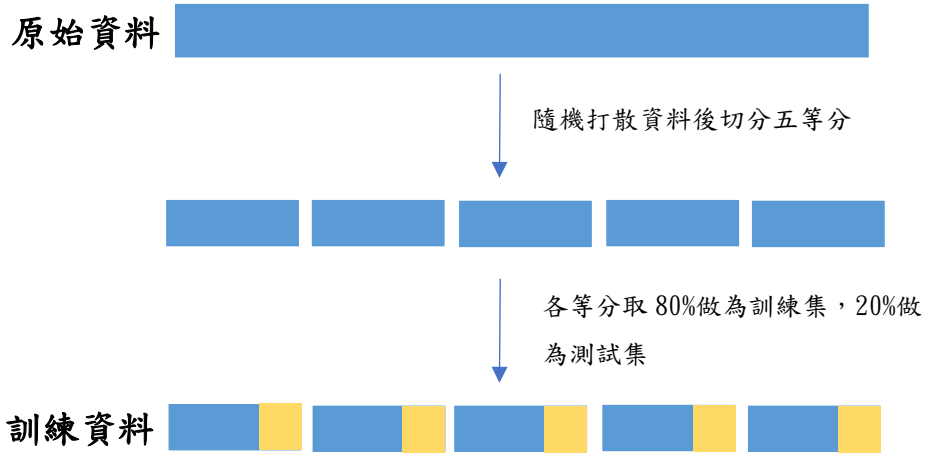
方法一訓練結果：將資料隨機打散

此部分將原始資料隨機打散

	Score
0_precision	0.82
0_recall	0.95
1_precision	0.46
1_recall	0.16
accuracy	0.79
預測 0 比例	93 %
預測 1 比例	7%

方法二：K-fold K=5

此方法將原始 916567 筆資料隨機打散，之後將資料平均切成五等分，每等分共有 183313 筆資料。切分完五等分後再依照 8:2 的比例將各等分切分為測試集與訓練集，每等分各有 146650 筆訓練資料，36663 筆測試資料。過採樣處理、模型、模型參數皆與方法一使用相同的方法。



方法二訓練結果

	1st fold	2st fold	3st fold	4st fold	5st fold	平均
0_precision	0.85	0.84	0.78	0.76	0.86	0.818
0_recall	0.98	0.95	0.92	0.95	0.99	0.958
1_precision	0.35	0.43	0.5	0.52	0.32	0.424
1_recall	0.07	0.16	0.23	0.16	0.03	0.13
accuracy	0.83	0.81	0.74	0.74	0.86	0.796
預測 0 比例	96.8%	93.3%	88.4%	92.1%	98.8%	
預測 1 比例	3.2%	6.7%	11.6%	7.9%	1.2%	

方法三：Time split

此方法依照原本數據排列順序，切分測試與訓練集。在切分資料時，使用前一年度做為訓練集，後一年度做為測試集，並且將訓練與測試樣本比例設為8:2，因此每一年度的樣本皆不相同。2007~2011 的樣本在此訓練方法中過少，因此不會將此區間的資料納入分析。

過採樣處理、模型、模型參數皆與方法一使用相同的方法。

切分為四個時間區間

1. 訓練集：2013 年共 175621 筆　測試集：2014 年共 43905 筆
2. 訓練集：2014 年共 202495 筆　測試集：2015 年共 50624 筆
3. 訓練集：2015 年共 269927 筆　測試集：2016 年共 67481 筆
4. 訓練集：2016 年共 175406 筆　測試集：2017 年共 43851 筆

方法三訓練結果

	2013	2014	2015	2016	平均
0_precision	0.82	0.76	0.77	0.83	0.795
0_recall	0.98	0.95	0.93	0.94	0.950
1_precision	0.44	0.51	0.5	0.37	0.455
1_recall	0.06	0.15	0.19	0.16	0.140
accuracy	0.81	0.74	0.74	0.80	0.773
預測 0 比例	97.5%	92.6%	90.4%	91.9%	
預測 1 比例	2.5%	7.4%	9.6%	8.1%	

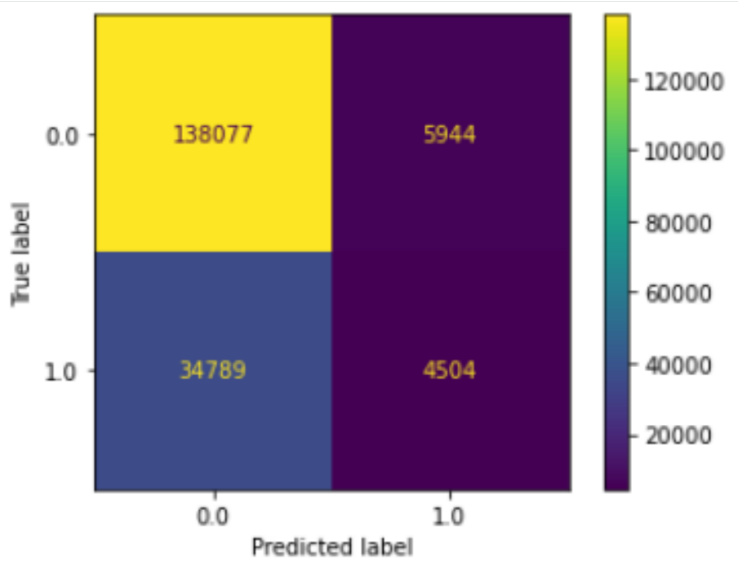
三種方法平均分數

	方法一	方法二	方法三
0_precision	0.80	0.818	0.795
0_recall	0.96	0.958	0.950
1_precision	0.43	0.424	0.455
1_recall	0.11	0.13	0.140
accuracy	0.78	0.796	0.773

混淆矩陣、各區間分數表

方法一

	precision	recall	f1-score	佔比
0	0.80	0.96	0.87	94.3%
1	0.43	0.11	0.18	5.7%

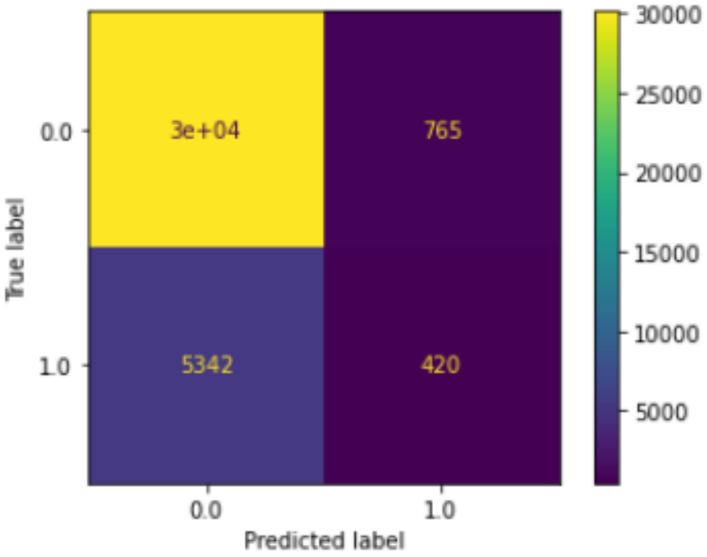


方法二

1st fold	precision	recall	f1-score	佔比
0	0.85	0.98	0.91	96.8%
1	0.35	0.07	0.12	3.2%

train_accuracy: 0.908

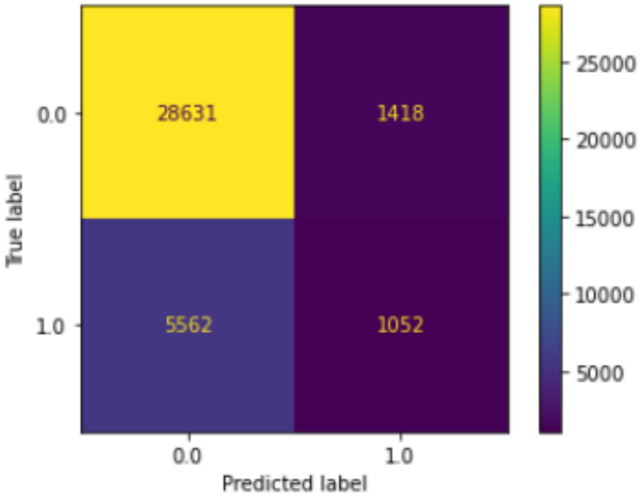
test _accuracy: 0.833



2st fold	precision	recall	f1-score	佔比
0	0.84	0.95	0.89	93.3%
1	0.43	0.16	0.23	6.7%

train_accuracy: 0.902

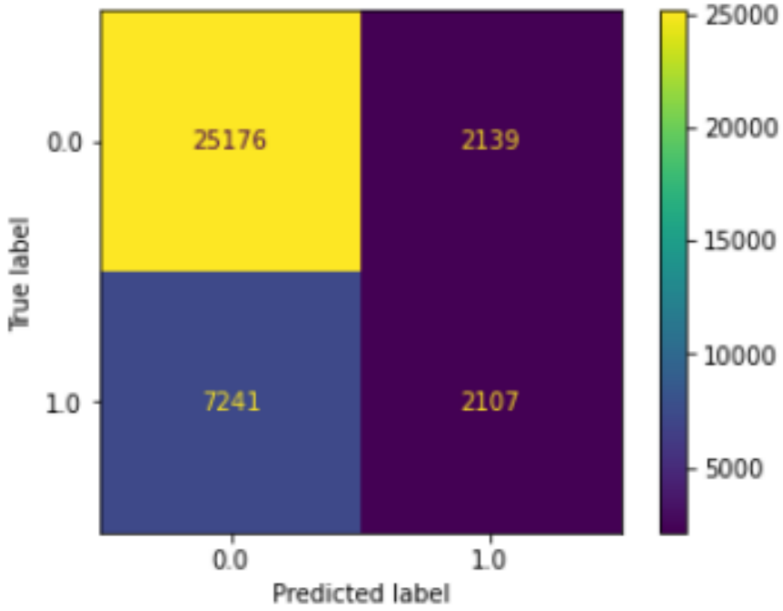
test _accuracy: 0.81



3st fold	precision	recall	f1-score	佔比
0	0.78	0.92	0.84	88.4%
1	0.50	0.23	0.31	11.6%

train_accuracy: 0.888

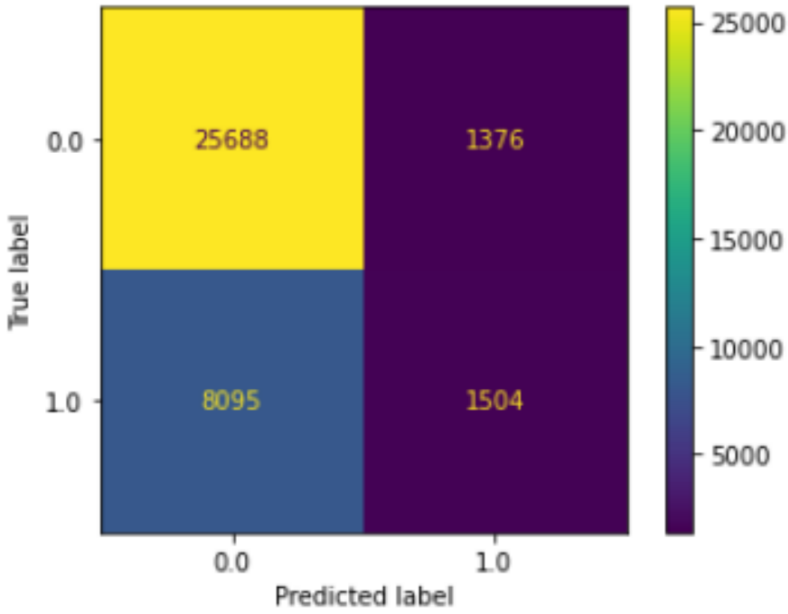
test _accuracy: 0.744



4st fold	precision	recall	f1-score	佔比
0	0.76	0.95	0.84	92.1%
1	0.52	0.16	0.24	7.9%

train_accuracy:0.889

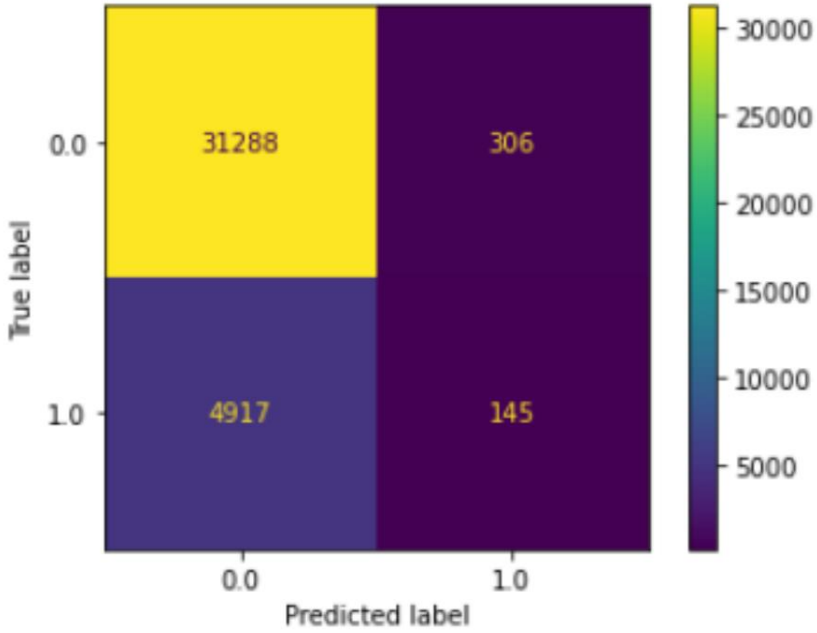
test _accuracy: 0.742



5st fold	precision	recall	f1-score	佔比
0	0.86	0.99	0.92	98.8%
1	0.32	0.03	0.05	1.2%

train_accuracy:0.878

test _accuracy: 0.858

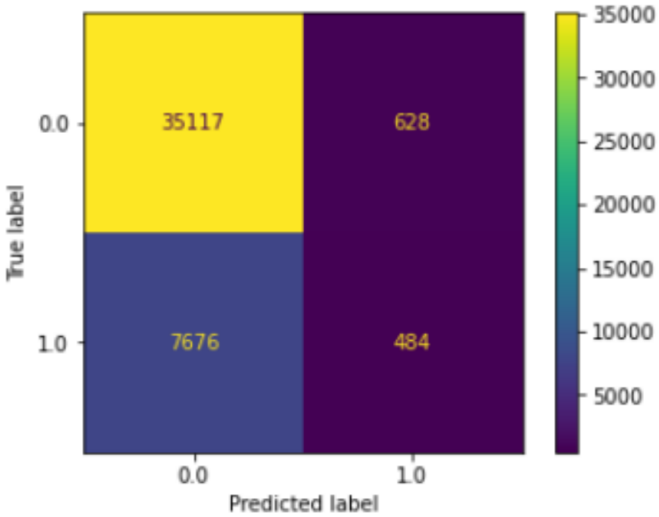


方法三

2013	precision	recall	f1-score	佔比
0	0.82	0.98	0.89	97.5%
1	0.44	0.06	0.10	2.5%

train_accuracy: 0.9

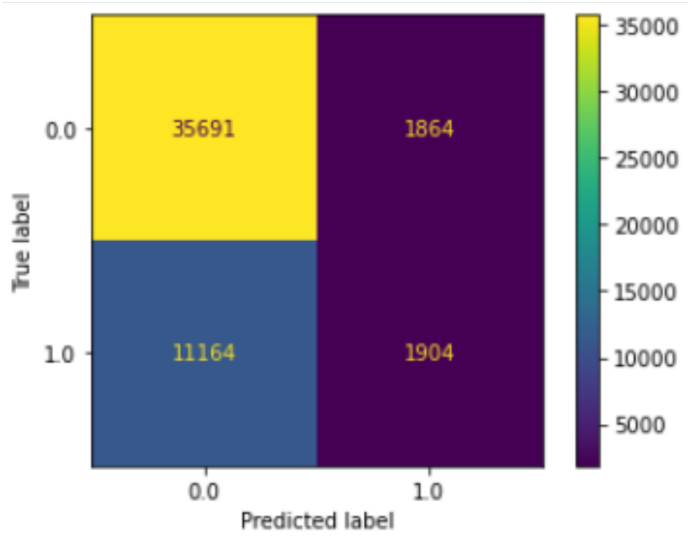
test _accuracy: 0.811



2014	precision	recall	f1-score	佔比
0	0.76	0.95	0.85	92.6%
1	0.51	0.15	0.23	7.4%

train_accuracy: 0.895

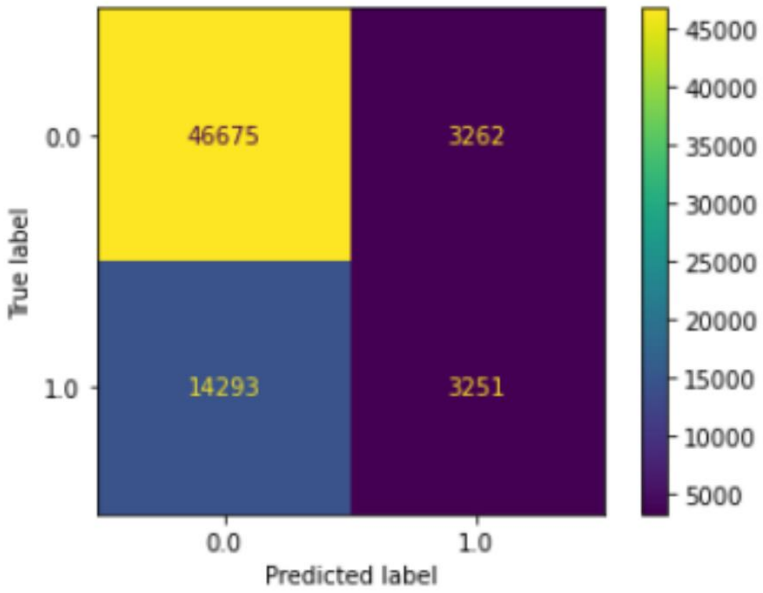
test _accuracy: 0.743



2015	precision	recall	f1-score	佔比
0	0.77	0.93	0.84	90.4%
1	0.50	0.19	0.27	9.6%

train_accuracy: 0.864

test _accuracy: 0.74



2016	precision	recall	f1-score	佔比
0	0.83	0.94	0.88	91.9%
1	0.37	0.16	0.23	8.1%

train_accuracy: 0.87

test _accuracy: 0.796