一、研究目的

利用財務特徵(X1-X95)預測公司是否破產(Y),比較不同的機器學習模型,若 Precision, Recall, Accuracy若未達0.9,則發展衍生變數使模式具有良好的預 測能力。

二、使用技術方法

改善目標類別不均衡問題:向下取樣(RandomUnderSampler)。

機器學習模型:決策樹、隨機森林、羅吉斯回歸、KNN。

評估模型正確率:混淆矩陣。

三、分析結果

• 原始資料的混淆矩陣

自變數X: X1~X95, 共95個自變數。

依變數y:Bankrupt?,共1個依變數。

訓練集:測試集=75:25 & 90:10。

• 向下取樣的混淆矩陣

自變數X: X1~X95, 共95個自變數。

依變數y:Bankrupt?,共1個依變數。

改善目標類別不均衡問題:向下取樣(RandomUnderSampler)。

訓練集:測試集=75:25 & 90:10。

備註:只有將測試集向下取樣,訓練集沒有向下取樣(採原資料集)。

此資料集在「Bankrupt?」欄位中比例懸殊太大,會造成混淆矩陣有過度 擬合(overfitting)的問題,故為了降低此問題,使用向下取樣來解決此 問題。

Bankrupt?	Frequence_是否破產	Percentage_是否破產
0.0	6599	0. 968
1.0	220	0. 032

備註:Bankrupt?=0.0→沒破產,Bankrupt?=1.0→有破產。

• 全部的混淆矩陣

○ 決策樹 (max_depth=2)

訓練集:測試集=75:25和90:10在經過資料處理和向下取樣後的混淆矩陣recall值皆有到1.0的準度。

若要使用決策樹模型,會建議使用向下取樣後的模型,且訓練集: 測試集=75:25或90:10。



cy 0. 97 5114 vg 0. 48 0. 50 0. 49 5114	precision	recall	f1-score	support
vg 0.48 0.50 0.49 5114				4941 173
rg 0.93 0.97 0.95 5114	0. 48 0. 93	0. 50 0. 97		5114 5114 5114

[[1658 0] [47 0]] ========				
	precision	recall	f1-score	support
0. 0 1. 0	0. 97 0. 00	1. 00 0. 00	0. 99 0. 00	1658 47
accuracy macro avg weighted avg	0. 49 0. 95	0. 50 0. 97	0. 97 0. 49 0. 96	1705 1705 1705

訓練集:測試集=90:10

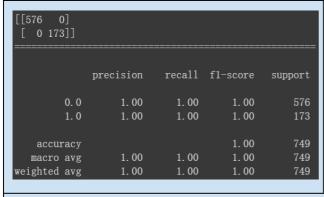
[[5923 14] [171 29]				
	precision	recal1	f1-score	support
0. 0 1. 0		1. 00 0. 14	0. 98 0. 24	5937 200
accuracy macro avg weighted avg	0.82	0. 57 0. 97	0. 97 0. 61 0. 96	6137 6137 6137

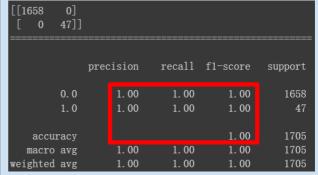
weighted av

[[659 3] [17 3]]				
	precision	recall	f1-score	support
0. 0 1. 0	0. 97 0. 50	1. 00 0. 15	0. 99 0. 23	662 20
accuracy macro avg weighted avg	0. 74 0. 96	0. 57 0. 97	0. 97 0. 61 0. 96	682 682 682

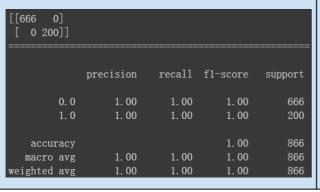
已處理

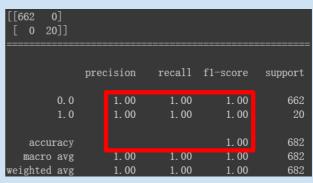
訓練集: 測試集=75:25





訓練集:測試集=90:10



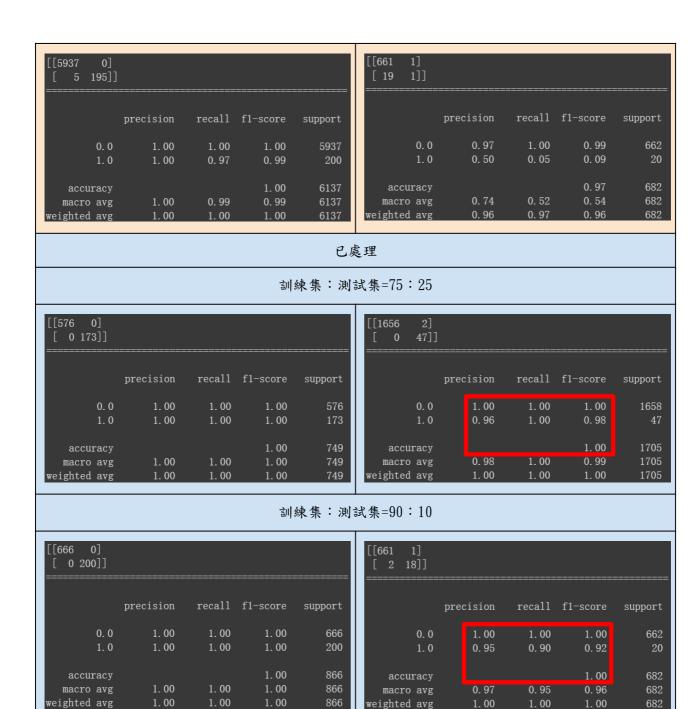


○ 隨機森林 (n_estimators=25)

訓練集:測試集=75:25和90:10在經過向下取樣後的混淆矩陣 recall值皆有到大於0.9的準度。

若要使用決策樹模型,會建議使用向下取樣後的模型,且訓練集: 測試集=75:25或90:10。





 羅吉斯回歸(solver='newton-cg') 訓練集:測試集=75:25和90:10在經過向下取樣後的混淆矩陣皆 有提高數值,但成果仍然不理想,所以不會建議使用羅吉斯回歸模型。



[[4926 15] [169 4]]				
	precision	recall	f1-score	support
0. 0 1. 0	0. 97 0. 21	1. 00 0. 02	0. 98 0. 04	4941 173
accuracy macro avg weighted avg	0. 59 0. 94	0. 51 0. 96	0. 96 0. 51 0. 95	5114 5114 5114

[[1647 [47	11] 0]]				
		precision	recall	fl-score	support
	0. 0 1. 0	0. 97 0. 00	0. 99 0. 00	0. 98 0. 00	1658 47
accui	racy			0. 97	1705
macro weighted		0. 49 0. 95	0. 50 0. 97	0. 49 0. 96	1705 1705

訓練集:測試集=90:10

[[5921 16] [195 5]]]] ======			
	precisio	n recall	f1-score	support
0. 1.			0. 98 0. 05	5937 200
accurac macro av	g 0.6		0. 97 0. 51	6137 6137
weighted av	g 0.9	0.97	0. 95	6137

	1] D]]] =====				
		precision	recall	f1-score	support
	0. 0 1. 0	0. 97 0. 00	0. 99 0. 00	0. 98 0. 00	662 20
accur macro weighted	avg	0. 49 0. 94	0. 50 0. 96	0. 96 0. 49 0. 95	682 682 682

已處理

訓練集:測試集=75:25

[[551 25] [145 28]]				
	precision	recall	f1-score	support
0. 0 1. 0	0. 79 0. 53	0. 96 0. 16	0. 87 0. 25	576 173
accuracy macro avg weighted avg	0. 66 0. 73	0. 56 0. 77	0. 77 0. 56 0. 72	749 749 749

[[1591 42	67] 5]] =====				
		precision	recall	f1-score	support
	0. 0 1. 0	0. 97 0. 07	0. 96 0. 11	0. 97 0. 08	1658 47
accur		0. 52	0. 53	0. 94 0. 53	1705 1705
weighted		0. 95	0. 94	0. 94	1705

訓練集:測試集=90:10

[[638 28] [175 25]]				
	precision	recall	f1-score	support
0. 0 1. 0	0. 78 0. 47	0. 96 0. 12	0. 86 0. 20	666 200
accuracy macro avg weighted avg	0. 63 0. 71	0. 54 0. 77	0. 77 0. 53 0. 71	866 866 866

[[642 20] [19 1]]				
	precision	recall	f1-score	support
0. 0 1. 0	0. 97 0. 05	0. 97 0. 05	0. 97 0. 05	662 20
accuracy			0. 94	682
macro avg weighted avg	0. 51 0. 94	0. 51 0. 94	0. 51 0. 94	682 682

o KNN (n_neighbors=3)

訓練集:測試集=75:25和90:10在經過向下取樣後的混淆矩陣皆有提高數值,但成果仍然不理想,所以不會建議使用KNN模型。

混淆矩陣									
訓練				測試					
未處理									
	訓練集:測試集=75:25								
[[4935 6] [149 24]] =			[[1648 10] [46 1]]]					
precisio				precision		f1-score	support		
0. 0 0. 9 1. 0 0. 8		0. 98 4941 0. 24 173	0. 0 1. 0	0. 97 0. 09	0. 99 0. 02	0. 98 0. 03	1658 47		
accuracy macro avg 0.8 weighted avg 0.9		0. 97 5114 0. 61 5114 0. 96 5114	accuracy macro avg weighted avg	0. 53 0. 95	0. 51 0. 97	0. 97 0. 51 0. 96	1705 1705 1705		
	訓練集:測試集=90:10								
[[5932 5] [168 32]] =======			[[659 3] [20 0]]						
precisio	n recall f1-	score support		precision	recall	f1-score	support		
0. 0 0. 9 1. 0 0. 8		0. 99 5937 0. 27 200	0. 0 1. 0	0. 97 0. 00	1. 00 0. 00	0. 98 0. 00	662 20		
accuracy macro avg 0.9 weighted avg 0.9		0. 97 6137 0. 63 6137 0. 96 6137	accuracy macro avg weighted avg	0. 49 0. 94	0. 50 0. 97	0. 97 0. 49 0. 95	682 682 682		
	訓練集:測試集=75:25								
[[559 17] [77 96]] ========			[[1470 188] [37 10]]]					
precision	n recall f1-s	score support		precision	recall	f1-score	support		
0. 0 0. 88 1. 0 0. 88		0. 92 576 0. 67 173	0. 0 1. 0	0. 98 0. 05	0. 89 0. 21	0. 93 0. 08	1658 47		
accuracy macro avg 0.8 weighted avg 0.8		0. 87 749 0. 80 749 0. 86 749	accuracy macro avg weighted avg	0. 51 0. 95	0. 55 0. 87	0. 87 0. 51 0. 91	1705 1705 1705		
訓練集:測試集=90:10									

[[634 32] [90 110]]				
	precision	recall	f1-score	support
0.0	0.88	0. 95	0. 91	666
1.0	0. 77	0. 55	0.64	200
accuracy			0. 86	866
macro avg	0.83	0. 75	0. 78	866
weighted avg	0.85	0.86	0.85	866

[[589 73] [15 5]]				
	precision	recall	f1-score	support
0. 0 1. 0	0. 98 0. 06	0. 89 0. 25	0. 93 0. 10	662 20
accuracy macro avg weighted avg	0. 52 0. 95	0. 57 0. 87	0. 87 0. 52 0. 91	682 682 682

四、結論

各項數值最高最優良的模型為有處理過資料的決策樹(max_depth=2)模型,再來為有處理過資料的隨機森林(n estimators=25)模型。

其他模型在經過資料處理後大多都有提高混淆矩陣的數值,但成果仍然不理想,所以不會建議使用。