1092_0701專題報告

ATP網球比賽預測

S07240013 簡睿德

準備流程

目的

收集來自2019年和2020年French Open官網提供的數據共254場比賽，分別先利用Probit模型預測(今年法網前五輪)球員NADAL比賽的勝負，再用RFECV特徵選取提升原先模型的精準度，之後再預測一次。

會針對RAFAEL NADAL進行預測是因為他過去再法國網球公開賽都有良好的成績，加上他分別拿下19年和20年的冠軍，比賽的場次和數據量與其他人相比之下多了一些，參考性也高了一點。

Association of Tennis Professionals (ATP)

職業網球聯合會

ATP世界巡迴賽		
賽事類別	**賽事數量**	**冠軍所獲積分**
大滿貫賽事	4	2000
ATP世界巡迴賽總決賽	1	1100~1500
ATP世界巡迴賽1000大師系列賽事	9	1000
ATP世界巡迴賽500系列賽事	13	500
ATP世界巡迴賽250系列賽事	39	250
ATP挑戰賽	178	80~125


澳洲網球公開賽(慢速硬地賽事)　比賽時間:每年1月的最後兩個星期


法國網球公開賽 (紅土賽事)　比賽時間:每年5月中到6月初

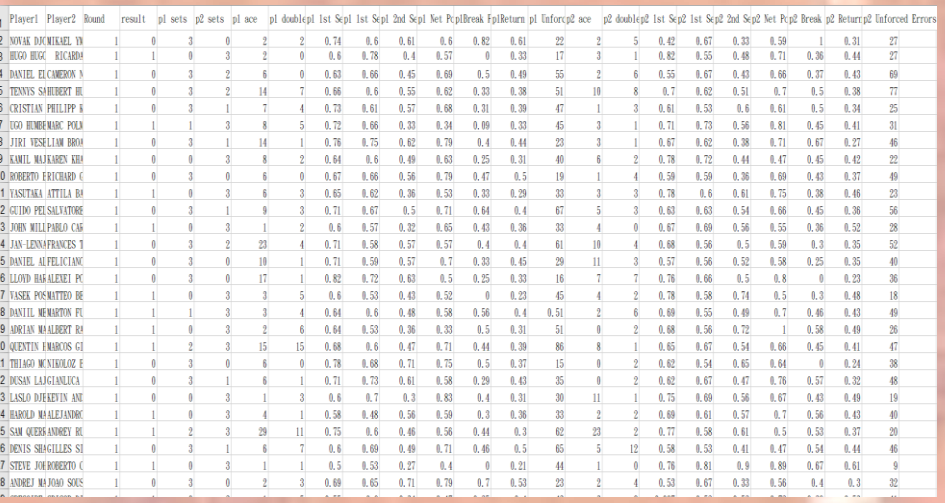

溫布頓網球錦標賽(草地賽事)　比賽時間:每年6月底至7月初


美國網球公開賽(快速硬地賽事)　比賽時間:每年8月底至9月初

資料敘述

原先EXCEL檔收集19、20年各127場比賽，22種的變數和球員名子，但為了能簡單順利地讀進python，因此將變數改為11種以及把原先每個變數的名稱只取每個單字的第一個字母，方便查看，然而在存成csv檔。

詳細的變數名詞介紹看下一頁



Player1	Player2	Round	result	p1 sets	p2 sets	p1 ace	p1 double	p1 1st Sep	p1 1st Sep	p1 2nd Sep	p1 Net Pcp	p1 Break Fp	p1 Return p	p1 Unforc	p2 ace	p2 double	p2 1st Sep	p2 1st Sep	p2 2nd Sep	p2 Net Pcp	p2 Break	p2 Return	p2 Unforced	Errors
NOVAK DJ	MIKAEL YM		1	0	3	0	2	2	0.74	0.6	0.61	0.6	0.82	0.61	22	2	5	0.42	0.67	0.33	0.59	1	0.31	27	
HUGO HUGC	RICARDO		1	1	0	3	2	0	0.6	0.78	0.4	0.57	0	0.33	17	3	1	0.82	0.55	0.48	0.71	0.36	0.44	27	
DANIEL EL	CAMERON N		1	0	3	2	6	0	0.63	0.66	0.45	0.69	0.5	0.49	55	2	6	0.55	0.67	0.43	0.66	0.37	0.43	69	
TENNYS SA	HUBERT HU		1	0	3	2	14	7	0.66	0.6	0.55	0.62	0.33	0.38	51	10	8	0.7	0.62	0.51	0.7	0.5	0.38	77	
CRISTIAN	PHILIPP K		1	0	3	1	7	4	0.73	0.61	0.57	0.68	0.31	0.39	47	1	3	0.61	0.53	0.6	0.61	0.5	0.34	25	
UGO HUMB	MARC POLM		1	1	1	3	8	5	0.72	0.66	0.33	0.34	0.09	0.33	45	3	1	0.71	0.73	0.56	0.81	0.45	0.41	31	
JIRI VESE	LIAM BROI		1	0	3	1	14	1	0.76	0.75	0.62	0.79	0.4	0.44	23	3	1	0.67	0.62	0.38	0.71	0.67	0.27	46	
KAMIL MA	FABIEN KRI		1	0	0	3	8	2	0.64	0.6	0.49	0.63	0.25	0.31	40	6	2	0.78	0.72	0.44	0.47	0.45	0.42	22	
ROBERTO	ERICHARD C		1	0	3	0	6	0	0.67	0.66	0.56	0.79	0.47	0.5	19	1	4	0.59	0.59	0.36	0.69	0.43	0.37	49	
YASUTAKA	ATTILA BA		1	1	0	3	6	3	0.65	0.62	0.36	0.53	0.33	0.29	33	3	3	0.78	0.6	0.61	0.75	0.38	0.46	23	
GUIDO PEL	SALVATORE		1	0	3	1	9	3	0.71	0.67	0.5	0.71	0.64	0.4	67	5	3	0.63	0.63	0.54	0.66	0.45	0.36	56	
JOHN MIL	PABLO CAS		1	1	0	3	1	2	0.6	0.57	0.32	0.65	0.43	0.36	33	4	0	0.67	0.69	0.56	0.55	0.36	0.52	28	
JAN-LENN	FRANCES T		1	0	3	2	23	4	0.71	0.58	0.57	0.57	0.4	0.4	61	10	4	0.68	0.56	0.5	0.59	0.3	0.35	52	
DANIEL AL	FELICIANC		1	0	3	0	10	1	0.71	0.59	0.57	0.7	0.33	0.45	29	11	3	0.57	0.56	0.52	0.58	0.25	0.35	40	
LLOYD HA	ALEXEI PO		1	0	3	0	17	1	0.82	0.72	0.63	0.5	0.25	0.33	16	7	7	0.76	0.66	0.5	0.8	0	0.23	36	
YASEK PO	MATTEO BE		1	1	0	3	3	5	0.6	0.53	0.43	0.52	0	0.23	45	4	2	0.78	0.58	0.74	0.5	0.3	0.48	18	
DANTIL M	MARTON FU		1	1	1	3	3	4	0.64	0.6	0.48	0.58	0.56	0.4	0.51	2	6	0.69	0.55	0.49	0.7	0.46	0.43	49	
ADRIAN MA	ALBERT RA		1	1	0	3	2	6	0.64	0.53	0.36	0.33	0.5	0.31	51	0	2	0.68	0.56	0.72	1	0.58	0.49	26	
QUENTIN	EMARCOS GI		1	1	2	3	15	15	0.68	0.6	0.47	0.71	0.44	0.39	86	8	1	0.65	0.67	0.54	0.66	0.45	0.41	47	
THIAGO M	NIKOLOZ B		1	0	3	0	6	0	0.78	0.68	0.71	0.75	0.5	0.37	15	0	2	0.82	0.54	0.65	0.64	0	0.24	38	
DUSAN LA	JGIANLUCA		1	0	3	1	6	1	0.71	0.73	0.61	0.58	0.29	0.43	35	0	2	0.82	0.67	0.47	0.76	0.57	0.32	48	
LASLO DJ	KEVIN ANE		1	0	0	3	1	3	0.6	0.7	0.3	0.83	0.4	0.31	30	11	1	0.75	0.69	0.56	0.67	0.43	0.49	19	
HAROLD MA	ALEJANDRC		1	1	0	3	4	1	0.58	0.48	0.56	0.59	0.3	0.36	33	2	2	0.69	0.61	0.57	0.7	0.56	0.43	40	
SAM QUER	ANDREY RU		1	1	2	3	29	11	0.75	0.6	0.46	0.56	0.44	0.3	62	23	2	0.77	0.58	0.61	0.5	0.53	0.37	20	
DENIS SH	AGILLES SI		1	0	3	1	6	7	0.6	0.69	0.49	0.71	0.46	0.5	65	5	12	0.58	0.53	0.41	0.47	0.54	0.44	46	
STEVE JO	ROBERTO C		1	1	0	3	1	1	0.5	0.53	0.27	0.4	0	0.21	44	1	0	0.76	0.81	0.9	0.89	0.67	0.61	9	
ANDREJ M	JOAO SOUS		1	0	3	0	2	3	0.69	0.65	0.71	0.79	0.7	0.53	23	2	4	0.53	0.67	0.33	0.56	0.4	0.3	32	

修改前

	result	p11spw	p11sp	p12spw	p1net	p1error	p21spw	p2sp	p22spw	p2net	p2error
2	0	0.74	0.6	0.61	0.6	22	0.42	0.67	0.33	0.59	27
3	1	0.6	0.78	0.4	0.57	17	0.82	0.55	0.48	0.71	27
4	0	0.63	0.66	0.45	0.69	55	0.55	0.67	0.43	0.66	69
5	0	0.66	0.6	0.55	0.62	51	0.7	0.62	0.51	0.7	77
6	0	0.73	0.61	0.57	0.68	47	0.61	0.53	0.6	0.61	25
7	1	0.72	0.66	0.33	0.34	45	0.71	0.73	0.56	0.81	31
8	0	0.76	0.75	0.62	0.79	23	0.67	0.62	0.38	0.71	46
9	0	0.64	0.6	0.49	0.63	40	0.78	0.72	0.44	0.47	22
10	0	0.67	0.66	0.56	0.79	19	0.59	0.59	0.36	0.69	49
11	1	0.65	0.62	0.36	0.53	33	0.78	0.6	0.61	0.75	23
12	0	0.71	0.67	0.5	0.71	67	0.63	0.63	0.54	0.66	56
13	1	0.6	0.57	0.32	0.65	33	0.67	0.69	0.56	0.55	28
14	0	0.71	0.58	0.57	0.57	61	0.68	0.56	0.5	0.59	52
15	0	0.71	0.59	0.57	0.7	29	0.57	0.56	0.52	0.58	40
16	0	0.82	0.72	0.63	0.5	16	0.76	0.66	0.5	0.8	36
17	1	0.6	0.53	0.43	0.52	45	0.78	0.58	0.74	0.5	18

修改後

比賽規則:

1. 一場球賽分別有五盤制或三盤制，五盤制者先勝三盤者為勝，三盤制則先勝二盤者為勝
2. 一盤(set)球賽有13局，先勝六局者為勝，如打成五局比五局平手，須打至有一方贏七局，其中如果是先打成六局比六局平手時，最後一局，也就是第13局也稱為決勝局

名詞解釋:

Round(輪)	第一輪為64強，第二輪為32強，後面以此類推，最後的冠軍賽則為第7輪
Result(結果)	玩家一(p1)贏表示0,玩家二(p2)贏表示1
Ace球	發球一方，將球發在有效區內，但接發球者沒接到球，直接得分的發球
Double Faults(雙發失誤)	在一分之中連續兩次發球失誤
1st Serve Percentage(一發進球率)	$\frac{\text{第一次發球成功直接進入發球區的球數}}{\text{一發進球球數}} \times 100\%$
1ST SERVE POINTS WON(第一發贏球率)	$\frac{\text{第一次發球贏球球數}}{\text{第一次發球成功的球數}} \times 100\%$
2ND SERVE POINTS WON(第二發贏球率)	$\frac{\text{第二次發球贏球球數}}{\text{第二次發球成功的球數}} \times 100\%$
NET POINTS WON(網前得分率)	$\frac{\text{上網前得分的球數}}{\text{上網前的球數}} \times 100\%$
Break Points Converted (把握破發點成功率)	$\frac{\text{成功破發球的球數}}{\text{可破發球數}} \times 100\%$
Return Points Won (接發球贏球率)	$\frac{\text{接球方贏得回發球的球數}}{\text{接球方回擊發球的球數}} \times 100\%$
UNFORCED ERRORS(非受迫性失誤)	在沒有受到對手壓力情況下，自己犯下的失誤

Probit模型

- Probit 模型以常態的累積機率密度函數(standardized cumulativenormal function)來取代原來的線性模型。透過Probit 迴歸分析所算出來的結果變數會是二元(yes,no)或是多元的離散資料。

- 在二元應變數當中，可使用許多種類的分布， 其中logistic 分佈通常被運用在非常態的分佈狀況下，而probit 則運用在標準常態分佈(normal distribution)

- probit 模型的方法先從「猜測」參數的值開始，然後會經過重複的猜測與改良而找出最好的估計值。

$$X_{i+1}=AX_i + B_1u_{1i} + B_2u_{2i} + \cdots\cdots + B_ku_{ki} + \varepsilon_i \quad , i = 1,2,\cdots,n$$

A、B 均為迴歸係數，A 為一常數

模型summary()

Current function value: 0.332528
Iterations 7

Probit Regression Results

Dep. Variable:	result	No. Observations:	254
Model:	Probit	Df Residuals:	243
Method:	MLE	Df Model:	10
Date:	Fri, 11 Jun 2021	Pseudo R-squ.:	0.5201
Time:	13:21:18	Log-Likelihood:	-84.462
converged:	True	LL-Null:	-175.99
Covariance Type:	nonrobust	LLR p-value:	5.443e-34

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.7987	2.650	0.301	0.763	-4.394	5.992
p11spw	-10.3935	1.725	-6.027	0.000	-13.774	-7.013
p11sp	-0.6690	1.774	-0.377	0.706	-4.145	2.807
p12spw	-0.0692	0.074	-0.930	0.352	-0.215	0.077
p1net	-2.5785	1.035	-2.491	0.013	-4.607	-0.550
p1error	0.0211	0.008	2.630	0.009	0.005	0.037
p21spw	3.1408	0.947	3.318	0.001	1.285	4.996
p2sp	2.7790	1.903	1.460	0.144	-0.951	6.508
p22spw	4.4375	1.326	3.346	0.001	1.838	7.037
p2net	2.6608	1.149	2.317	0.021	0.410	4.912
p2error	-0.0103	0.008	-1.229	0.219	-0.027	0.006

$x_{i+1} = 0.7987 - 10.3935 * p1spw - 0.669 * p11sp - 0.0692 * p12spw - 2.5785 * p1net + 0.0211 * p1error + 3.1408 * p21spw + 2.779 * p21sp + 4.4375 * p22spw + 2.6608 * p2net - 0.0103 * p2$error

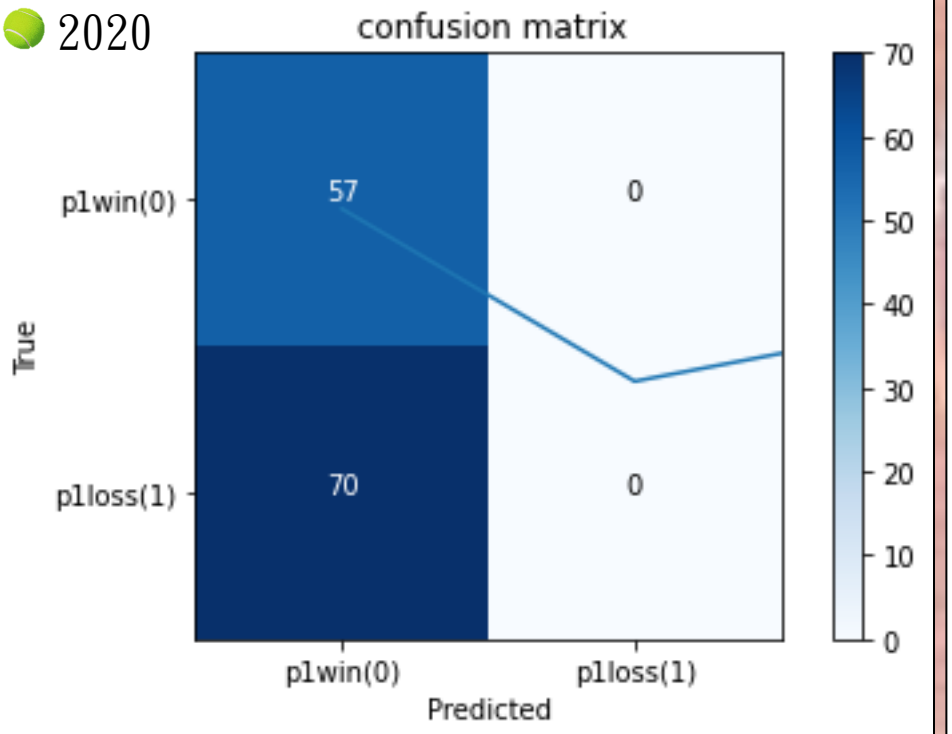
原本結果 vs 訓練結果

Index	result
0	0
1	0
2	1
3	1
4	1
5	0
6	0
7	1
8	0
9	1
10	0
11	1
12	0
13	1
14	0
249	0
250	1
251	0
252	1
253	1

Index	0
0	0
1	0
2	0
3	0
4	0
5	0
6	0
7	0
8	0
9	0
10	0
11	0
12	0
13	0
14	0
249	0
250	0
251	0
252	0
253	0

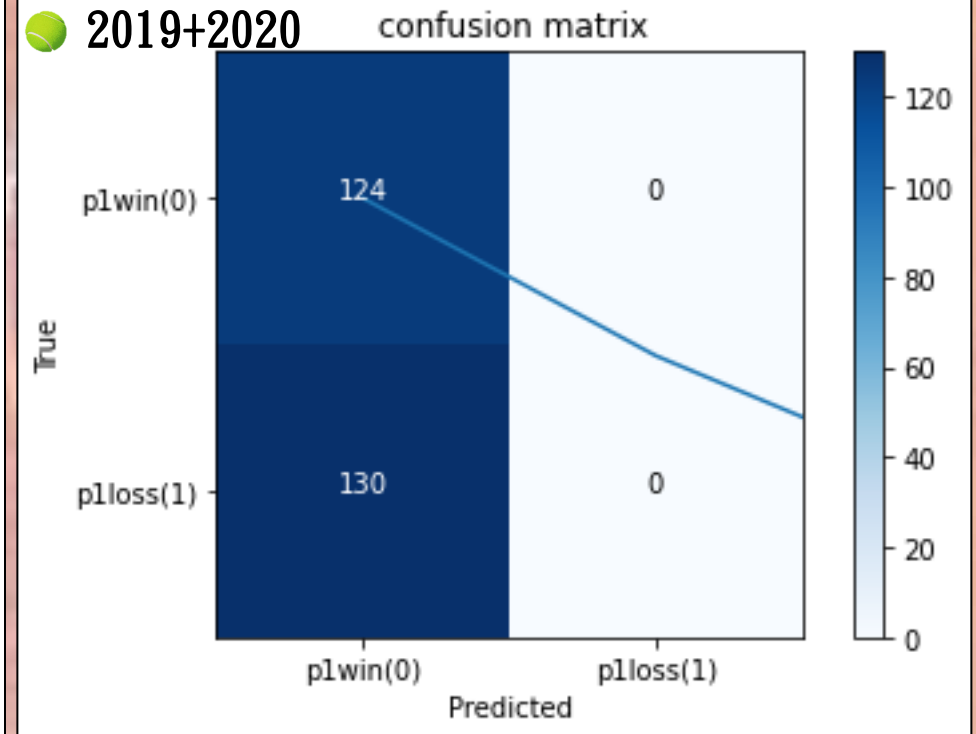
混淆矩陣

2020



Test accuracy = 0.4488188976377952

2019+2020



Test accuracy = 0.4881889763779528

特徵選取--------RFE遞歸特徵消除法

(Recursive Feature Elimination)

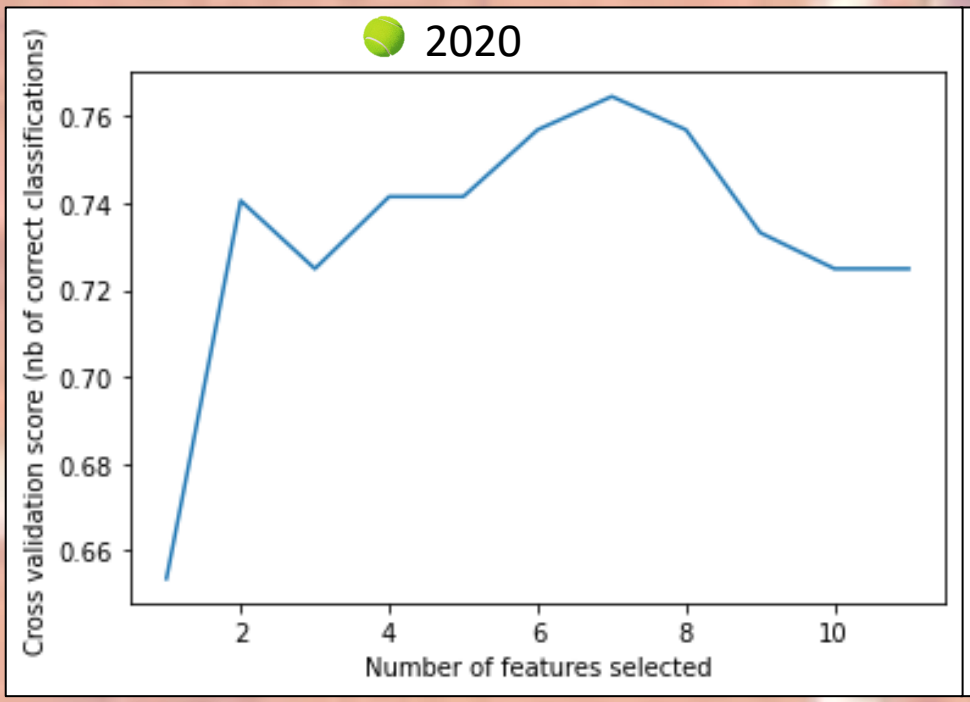
特徵選取

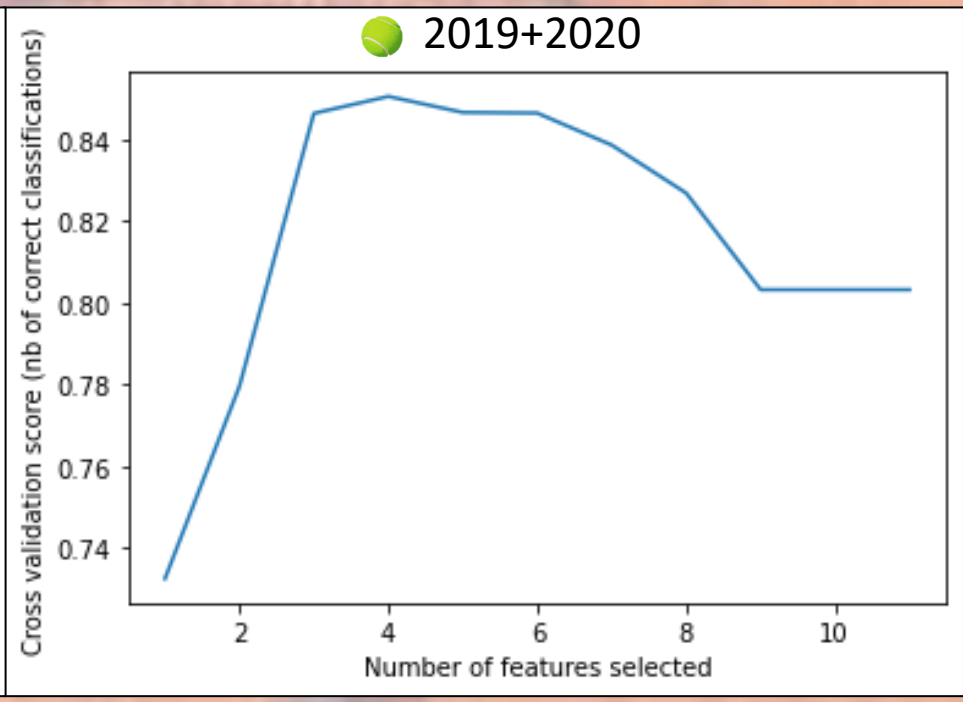
- 簡化模型
- 縮短訓練時間
- 改善通用性、避免過擬合

- 它擬合模型並移除較弱的特徵，直到達到指定的特徵數量。 coef_ 屬性或者feature_importances_並且通過在每個循環中遞歸消除少量特徵，RFE 試圖消除模型中可能存在的依賴關係和共線性。
- RFECV在一個交叉驗證的循環中執行RFE來找到最優的特徵數量

Index	p11spw	p11sp	P1net	P21spw	P2sp	P22spw	P2net
0	0.74	0.6	0.6	0.42	0.67	0.33	0.59
1	0.6	0.78	0.57	0.82	0.55	0.48	0.71
2	0.63	0.66	0.69	0.55	0.67	0.43	0.66
3	0.66	0.6	0.62	0.7	0.62	0.51	0.7
4	0.73	0.61	0.68	0.61	0.53	0.6	0.61
5	0.72	0.66	0.34	0.71	0.73	0.56	0.81
6	0.76	0.75	0.79	0.67	0.62	0.38	0.71
7	0.64	0.6	0.63	0.78	0.72	0.44	0.47
8	0.67	0.66	0.79	0.59	0.59	0.36	0.69

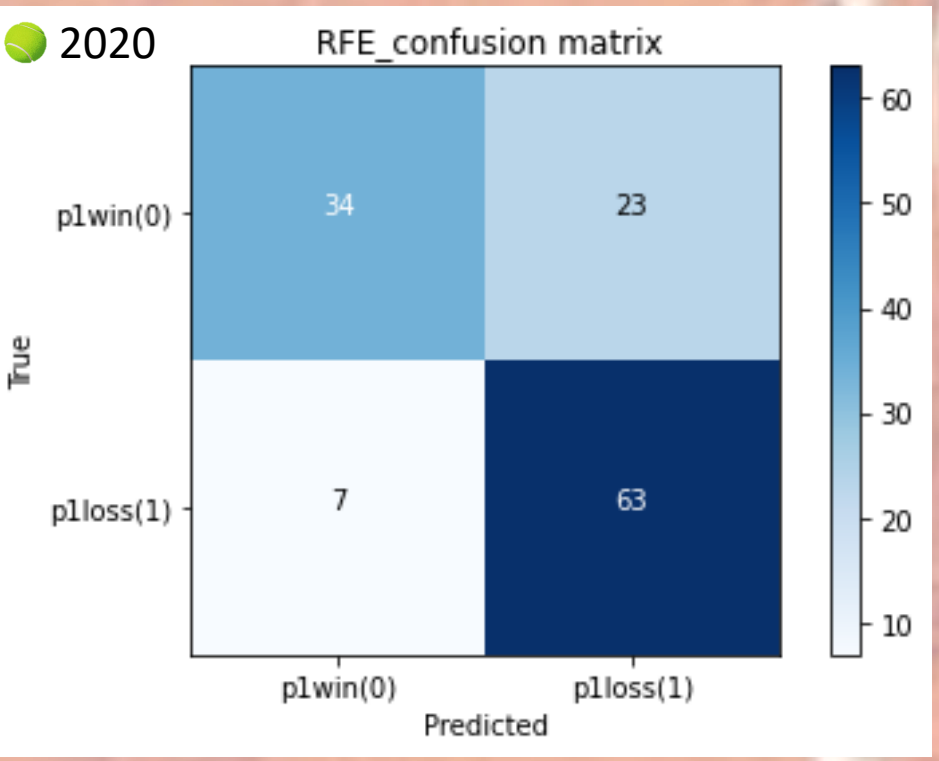
Index	p11spw	p1net	p21spw	p22spw
0	0.82	0.86	0.59	0.4
1	0.68	0.67	0.51	0.36
2	0.72	0.44	0.71	0.58
3	0.55	0.47	0.84	0.69
4	0.7	0.61	0.72	0.52





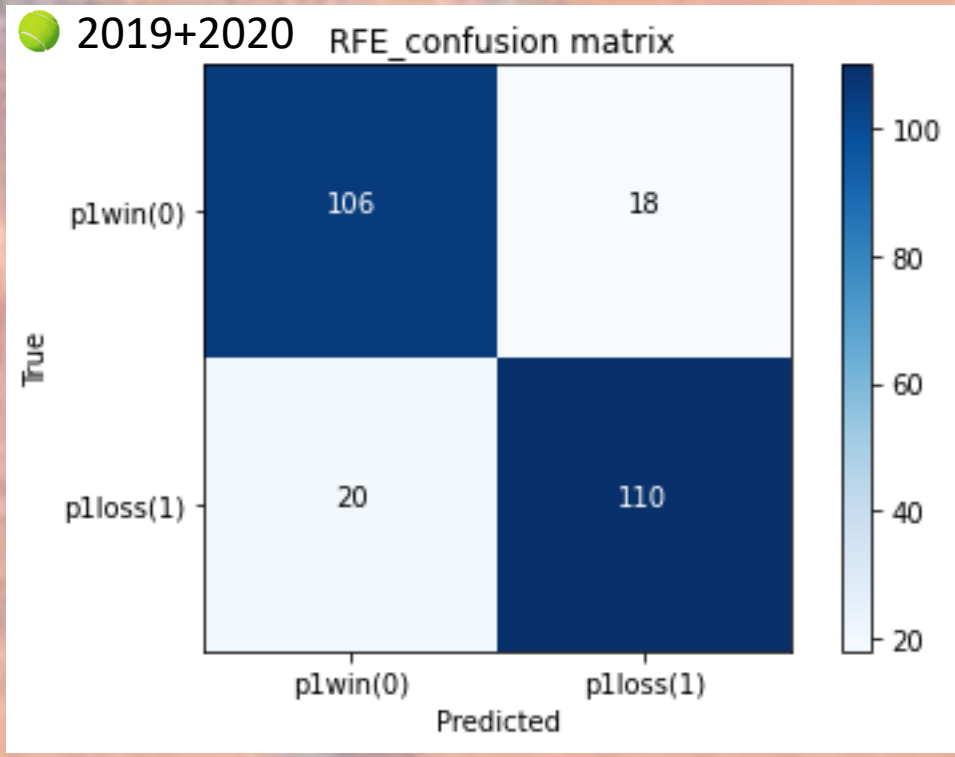
選取後的混淆矩陣

2020



Test accuracy = 0.763779527559055

2019+2020



Test accuracy = 0.8503937007874016

- 利用前面2019+2020數據訓練的模型，在predict後面輸入相關的數據來預測今年nadal的比賽結果。
  輸入的相關數據為網站提供前五輪個別的賽後統計結果

- 第六輪(平均)是以nadal前五場的數據取平均和Novak Djokovic前五場的數據取平均來當輸入值
  第六輪(去年)則是他們去年的對戰數據來當輸入值

	Probit預測結果		RFECV預測結果	
	預測	實際	預測	實際
第一輪	贏	贏	贏	贏
第二輪	贏	贏	贏	贏
第三輪	贏	贏	贏	贏
第四輪	贏	贏	贏	贏
第五輪	贏	贏	贏	贏
第六輪(平均)	贏	?	贏	?
第六輪(去年)	贏	?	贏	?

結論與心得

從這次的報告結果看來，一開始probit訓練的準確率雖然不高，但是經過特徵選取後準確率好了很多，然而變數的數量多，它的準確率不一定會比較高，還是要看哪些變數的效果比較好，而數據量多訓練出來的準確率也會比較高。

最後要預測下一場比賽時，要找兩個球員相關的數據帶入predict，但發現要找到合適的數據卻找不太到，可能要以手邊有的數據去做調整。由於網球比賽是單人或雙人的項目，每一年比賽的場次跟其他運動項目比較少，兩個相同的人過去的參考對戰次數相對的比較少，加上網球數據在近幾年才開始有在統計整理，一旦要用到之前的數據，能用的範圍就縮小了很多，預測出來的參考性也比較低。

參考資料

- https://en.wikipedia.org/wiki/2019_French_Open_%E2%80%93_Men%27s_Singles
- https://www.rolandgarros.com/en-us/matches?finished=true&tournamentDay=20190527&type=SM
- https://zh.wikipedia.org/wiki/%E7%B6%B2%E7%90%83%E8%A1%93%E8%AA%9E%E5%88%97%E8%A1%A8
- https://www.scoreboard.com/en/tennis/atp-singles/french-open/results/
- 頂尖職業網球男子單打選手攻守數據分析,李正安
- Probit迴歸模型與羅吉斯迴歸模型預測理論之研究-以美國職棒大聯盟為例，楊意婷
- 以Probit迴歸模型預測NBA籃球比賽結果,潘彥甫