FDA HW 3-2 Report

F74056255 陳郁明

1. Dataset：Dota2 Game Result
2. 目標：預測遊戲勝利隊伍(Radiant or Dire)
3. 資料分析：

	0	1	2	3	4	5	6	7	8	9	...	107	108	109	110	111	112	113	114	115	116
0	-1	223	2	2	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	1	152	2	2	0	0	0	1	0	-1	...	0	0	0	0	0	0	0	0	0	0
2	1	131	2	2	0	0	0	1	0	-1	...	0	0	0	0	0	0	0	0	0	0
3	1	154	2	2	0	0	0	0	0	0	...	-1	0	0	0	0	0	0	0	0	0
4	-1	171	2	3	0	0	0	0	0	-1	...	0	0	0	0	0	0	0	0	0	0

5 rows × 117 columns

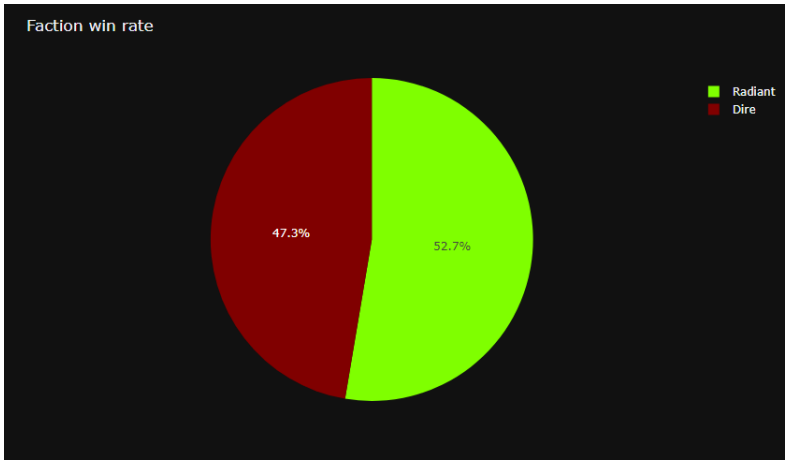
原始資料前四行分別是：勝利隊伍(1 or -1)、地區、遊戲模式、大廳類別，後面 113 行是各個英雄在該場遊戲有無被選擇(1 or -1 or 0)，共計 92650 場遊戲，無缺值。

由於原始資料我覺得不易觀看和處理，因此進行處理變成以下結構：

	winner	cluster	mod	lobby	r_1	r_2	r_3	r_4	r_5	d_1	d_2	d_3	d_4	d_5
0	-1	223	2	2	10	14	25	28	32	18	22	38	74	88
1	1	152	2	2	4	14	26	27	71	6	21	35	93	98
2	1	131	2	2	4	22	25	32	59	6	20	46	72	93
3	1	154	2	2	17	35	54	62	95	7	23	42	47	104
4	-1	171	2	3	16	31	36	44	73	6	9	11	29	86

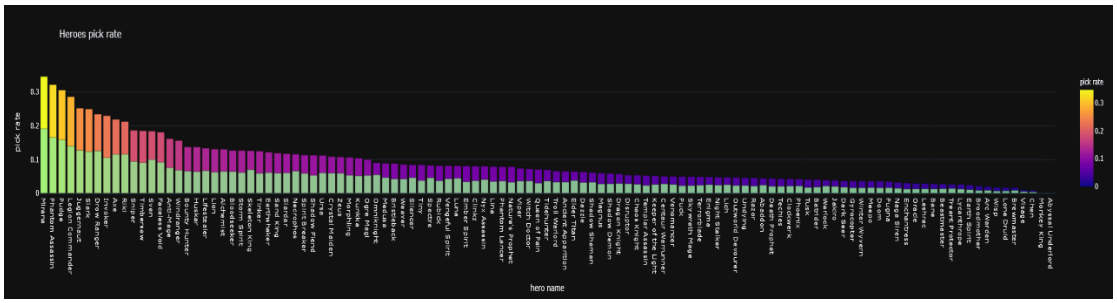
後面 10 行分別是兩隊(Radiant & Dire)各 5 位玩家，在該場遊戲選擇的英雄編號。

- 陣營勝率



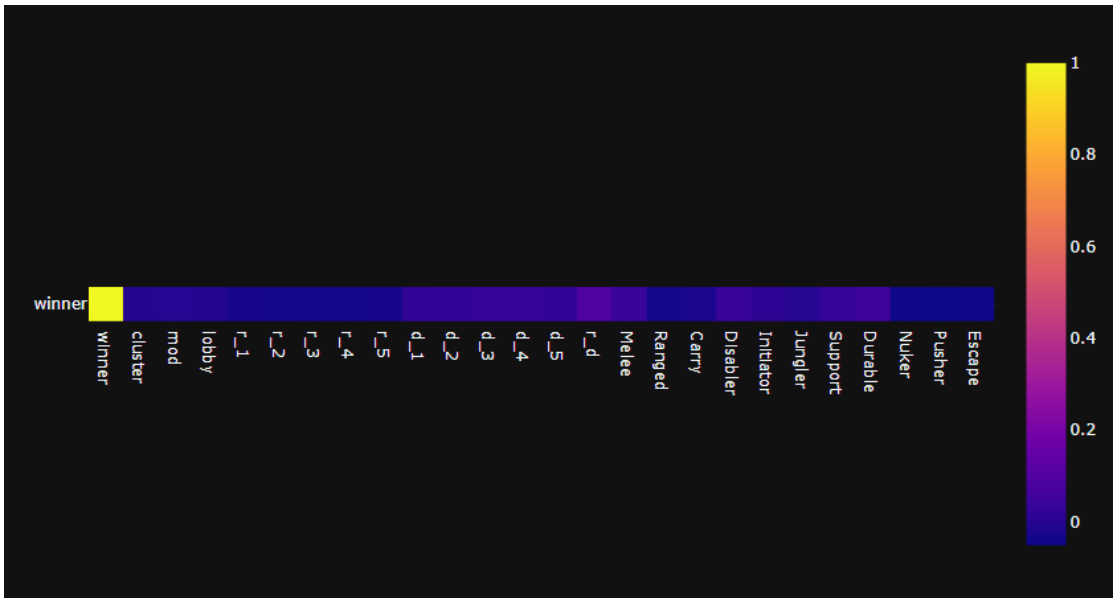
Radiant 的勝率略高於 Dire

- Hero pick rate & win rate



總計九萬多場遊戲中，大部分的英雄出場率都不高，在熱門英雄中並沒有勝率特別突出的腳色(> 60%)，因此理論上並沒有能憑一己之類扭轉戰局的英雄。

- 相關係數



其中以 r_d 欄位以及幾個英雄 tag 的欄位較具影響(稍後解釋後面幾個欄位)

4. 實驗

- Baseline

訓練資料：X = “cluster” ~ “d_5” , y = “winner”

模型：AdaBoostClassifier(n_estimators = 100)，

驗證：Kford(n_splits = 10)

結果：

  - 平均驗證(valid)準確度 = 0.5534700485698867
  - 平均測試(test)準確度 = 0.5560520691665047

- 結果分析

準確度不高的原因可能是訓練資料不夠有意義，單靠兩隊英雄陣容難以預測勝

利隊伍，這和之前的勝率分析相關，單位英雄對結果的影響不是決定性的，其他欄位如地區或遊戲模式等等也不夠有意義。

- 改進

- 新增 r_d 欄位：
  英雄相對勝率 rwr = pick rate * win rate
  r_d = (Radiant 方的 rwr 總和) – (Dire 方的 rwr 總和)
- 新增數個英雄 tag 欄位：
  每個英雄都會有幾個相關的 tag，代表英雄的屬性或特色，
  以” Anti-Mage”為例，他有"Melee","Carry","Escape","Nuker"四種 tag。
  而新增的欄位以”Melee”舉例，代表 Radiant 方有”Melee”標籤的英雄數扣掉 Dire 方有同樣標籤的英雄數，
  這些跟 tag 有關的欄位可以說明某些特色(例如近戰或遠程)對最終戰果的影響。

- 改進結果

使用與 Baseline 相同的模型與驗證方法，
改變 X = [“r_1” : “d_5”, “Melee”, “Ranged”, “Carry”, “Disabler”, “Durable”, “Nuker”, “Pusher”, “Escape”]，
最終結果：

- 平均驗證(valid)準確度 = 0.5672746896923908
- 平均測試(test)準確度 = 0.5747037108995532