Winning Rate of HearthStone Decks

陳柏瑜, 高翊傑

ccClub2020 Fall

NTU Econ

2020.12.23

Outline

Background

Hearthstone

Hearthstone

Hearthstone

Hearthstone

- 職業：
  1. 惡魔獵人、德魯伊、獵人、法師、聖騎士、牧師、盜賊、薩滿、術士、戰士
- 牌組：上千副乃至於更多（每三十張牌 (cards) 組成一副牌組 (Deck)）
- 常見類型：快攻、節奏、生物鋪場、死聲、Buff、一波帶走…等

Core Question

Winning Rate of a deck

- 給定任意一副牌組（i.e. 30 張卡牌），這副牌牌組的勝率是多少？

考量其他影響勝率的因素：

- 職業
- 牌組造價（魔塵）
- 牌組節奏（一局遊戲的時長）
- 牌組遊玩局數（熱門程度）

Data

HSReplay.net

在總共十個職業中，根據每個職業、在每週不定期紀錄以下資訊：

- 紀錄至少 36 副牌組的勝率
- 紀錄每副牌組的組成（其中 30 張卡分別是哪些卡）
- 每副牌組的造價（魔塵花費）
- 每副牌組節奏（一局遊戲的時長）
- 每副牌組的遊玩局數（至少對局 200 場以上）

其中，我們自 2020-12-06 起至 2020-12-24 每天爬取該網站，取得每副牌組的勝率紀錄

Scraped Data

以 dictionary of dictionary 儲存牌組資訊

```
{'URL': 'https://hsreplay.net/decks/0tCHPvwBZXSS375UZDVxO/',
 'Deck_Name': '炸彈戰',
 'Deck_Composition': {57718: {'card_name': '劍盾合璧',
   'IsDuo': 1,
   'card_URL': 'https://hsreplay.net/cards/57718/'},
  511: {'card_name': '升級！',
   'IsDuo': 1,
   'card_URL': 'https://hsreplay.net/cards/511/'},
  636: {'card_name': '旋風斬',
   'IsDuo': 'NaN',
   'card_URL': 'https://hsreplay.net/cards/636/'},
  546: {'card_name': '盾牌猛擊',
   'IsDuo': 1,
   'card_URL': 'https://hsreplay.net/cards/546/'},
  56512: {'card_name': '海盜庫房',
   'IsDuo': 1,
   'card_URL': 'https://hsreplay.net/cards/56512/'},
  56504: {'card_name': '劍刃風暴',
   'IsDuo': 1,
   'card_URL': 'https://hsreplay.net/cards/56504/'},
  59411: {'card_name': '巴羅夫領主',
   'IsDuo': 'NaN',
   'card_URL': 'https://hsreplay.net/cards/59411/'},
  1023: {'card_name': '盾牌格擋',
   'IsDuo': 1,
```

Distinct Cards

在標準環境中對戰的所有牌組共有 973 張相異的卡牌，其中有 226 張中立卡牌

	Card_ID	Card_Name	IsNeutral
0	8	精神控制	0
1	22	心靈之怒	0
2	23	末日災厄	0
3	28	柯爾克隆精英	0
4	30	思想竊取	0
...	...	...	...
968	61842	瑪克希瑪·布萊頓海默	0
969	61846	請勿餵食動物	0
970	61884	舞台鬼母	0
971	61898	蠕動懼怪	1
972	62049	安全檢查員	1

973 rows × 3 columns

Tidy Deck Data

	win_rate_RE	time_duration_RE	game_count_RE	dust_cost	Warrior	Warlock	Shaman	Rogue	Priest	Paladin	...	Card_1035	Card_1029
0	61.1	5.6	130000	1840	0	0	0	0	0	0	...	0	0
1	59.2	7.5	60000	6000	0	0	0	0	0	0	...	0	0
2	62.6	5.9	47000	3760	0	0	0	0	0	0	...	0	0
3	59.4	7.3	16000	4360	0	0	0	0	0	0	...	0	0
4	50.6	7.4	15000	0	0	0	0	0	0	0	...	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
16721	61.1	10.0	250	10520	1	0	0	0	0	0	...	0	0
16722	57.5	8.5	240	12120	1	0	0	0	0	0	...	0	0
16723	52.3	9.6	240	13460	1	0	0	0	0	0	...	0	0
16724	54.8	11.7	200	11800	1	0	0	0	0	0	...	0	0
16725	55.9	11.7	200	16600	1	0	0	0	0	0	...	0	0

16726 rows × 990 columns

Tidy Deck Data

- 共有 16726 筆觀察值（不盡相同勝率的 deck），990 個 feature
- dummy variables 中有 973 個為個別卡牌、10 個為職業
- variable of interest 為 win rate
- time duration, dust cost 皆為 continuous variable
- game count 為 discrete count data
- Deck URL、Deck ID 分別為 deck 對應的網址以及 ID
- Deck Name 為牌組名稱（不唯一，不同組成的牌組可能有相同名稱）

Model

Random Forest

相較於傳統的迴歸模型，決策樹模型可以給我們在「是否選擇把某張卡放進牌組裡」一個更好的詮釋

我們將提供以下模型的估計：

- OLS result
- Random Forest
- Random Forest with PCA dimension reduction

Result

Random Forest

相較於傳統的迴歸模型，決策樹模型可以給我們在「是否選擇把某張卡放進牌組裡」一個更好的詮釋

我們將提供以下模型的估計：

- OLS
- LASSO
- Random Forest
- Random Forest with PCA dimension reduction

Before Random Forest: OLS

An OLS Model without specific cards: sketch the big picture of the meta

對於勝率有正相關的大致指標包含：

- 節奏快的牌組（天下武功，唯快不破？!)
- 造價高的牌組（有傳說、史詩）
- 熱門的牌組（比較多人愛抄）
- 天梯上比較常見的職業：惡魔獵人、獵人、聖騎士

有較高的勝率

Before Random Forest: OLS

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       6.158e+01  3.228e-01 190.791  < 2e-16 ***
time_duration_RE -7.504e-01  4.453e-02 -16.850  < 2e-16 ***
game_count_RE     1.221e-04  1.273e-05   9.591  < 2e-16 ***
dust_cost         8.286e-05  1.901e-05   4.359 1.31e-05 ***
Warrior          -3.756e+00  2.544e-01 -14.766  < 2e-16 ***
Warlock          -6.240e+00  2.437e-01 -25.611  < 2e-16 ***
Shaman           -1.046e+00  2.002e-01  -5.224 1.77e-07 ***
Rogue            -3.385e+00  2.232e-01 -15.169  < 2e-16 ***
Priest           -6.962e+00  2.748e-01 -25.335  < 2e-16 ***
Paladin           2.436e+00  2.178e-01  11.183  < 2e-16 ***
Mage             -5.185e+00  2.340e-01 -22.159  < 2e-16 ***
Hunter            6.742e-01  2.185e-01   3.086  0.00203 **
Druid            -8.951e+00  2.453e-01 -36.481  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.577 on 16713 degrees of freedom
Multiple R-squared:  0.2693,    Adjusted R-squared:  0.2688
F-statistic: 513.3 on 12 and 16713 DF,  p-value: < 2.2e-16
```

Before Random Forest: OLS after control

```
Coefficients: (153 not defined because of singularities)
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.955e+01  1.422e+01   2.781 0.005429 **
time_duration_RE -7.472e-01  4.588e-02 -16.284  < 2e-16 ***
game_count_RE    -8.504e-08  1.849e-05  -0.005 0.996331
dust_cost         1.623e-04  2.258e-05   7.190 6.76e-13 ***
Warrior           1.891e+01  1.472e+01   1.284 0.199017
Warlock           1.425e+01  1.588e+01   0.898 0.369394
Shaman            1.195e+01  1.481e+01   0.807 0.419490
Rogue             5.301e+00  1.613e+01   0.329 0.742338
Priest            4.645e+00  1.588e+01   0.293 0.769836
Paladin           2.058e+01  1.924e+01   1.070 0.284610
Mage              1.511e+01  1.838e+01   0.822 0.411254
Hunter            2.547e+01  1.528e+01   1.667 0.095584 .
Druid             9.443e+00  1.461e+01   0.646 0.518069
Card_999         -1.170e+01  2.312e+00  -5.060 4.24e-07 ***
Card_997         -1.739e+00  1.330e+00  -1.308 0.190915
Card_995         -4.646e+00  4.247e+00  -1.094 0.273966
Card_994         -9.416e-01  8.283e-01  -1.137 0.255628
Card_985         -5.910e-01  5.312e+00  -0.111 0.911424
```

Before Random Forest: OLS after control

在以放入哪些卡牌作為控制變數後，可以看見仍是

- 節奏快的牌組（天下武功，唯快不破？!)
- 造價高的牌組（有傳說、史詩）

有較高的勝率

當然，此時我們便面對了 the curse of high dimensionality

Before Random Forest: LASSO

一個自然的解決方式是用 penalty regression；透過 cross validation 選擇 tuning parameter λ 後，便可得到：

```
 [1] win_rate          4.67218665473637
 [2] time_duration     0.000156656548294709
 [3] Card_62049        0.832009763102095
 [4] Card_60280        2.55282958928911
 [5] Card_59658        0.15411656900678
 [6] Card_59646        -1.49511713454943
 [7] Card_59395        4.34942441528714
 [8] Card_59039        1.76100661192004
 [9] Card_59036        1.78894309462431
[10] Card_59035        1.48871453526339
[11] Card_58973        2.26300441482049
[12] Card_55429        2.16930380979627
[13] Card_55401        1.09320992464284
[14] Card_53771        1.78211572392716
[15] Card_48           0.78900693038165
```

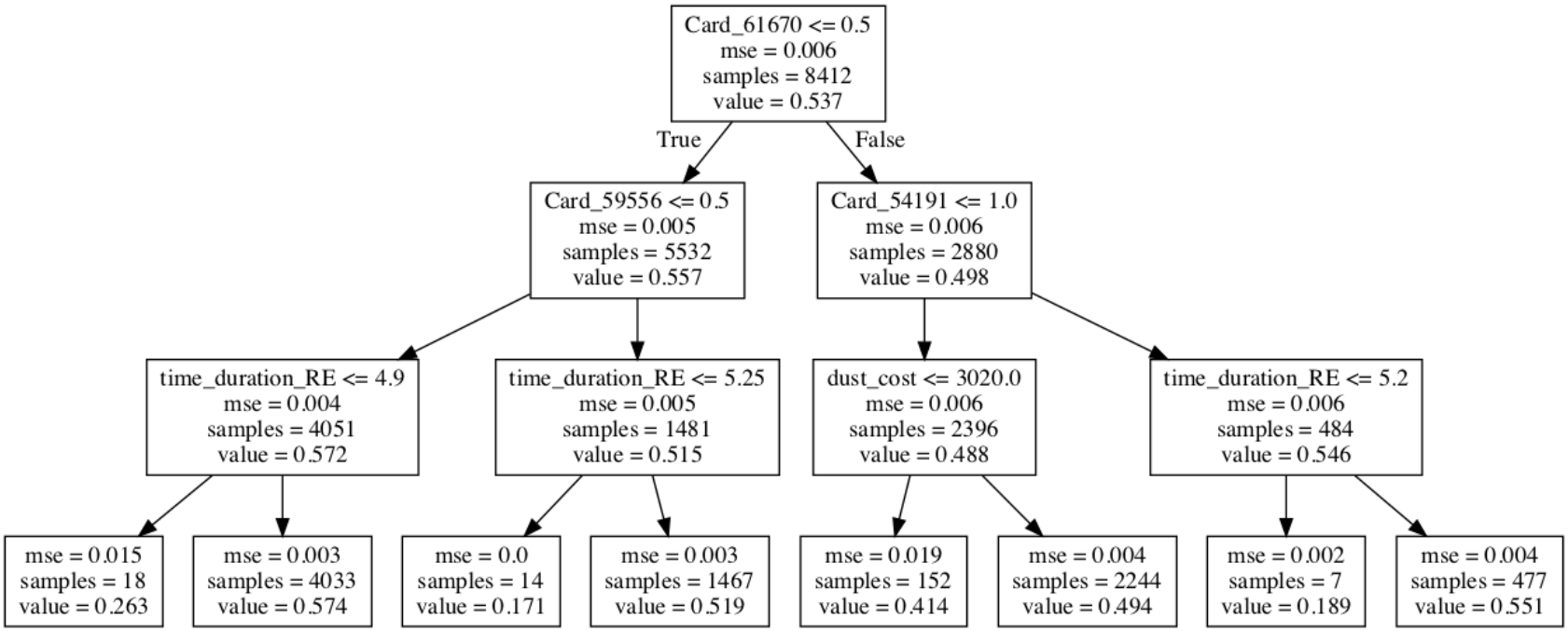
Before Random Forest: LASSO

翻譯過來就是：

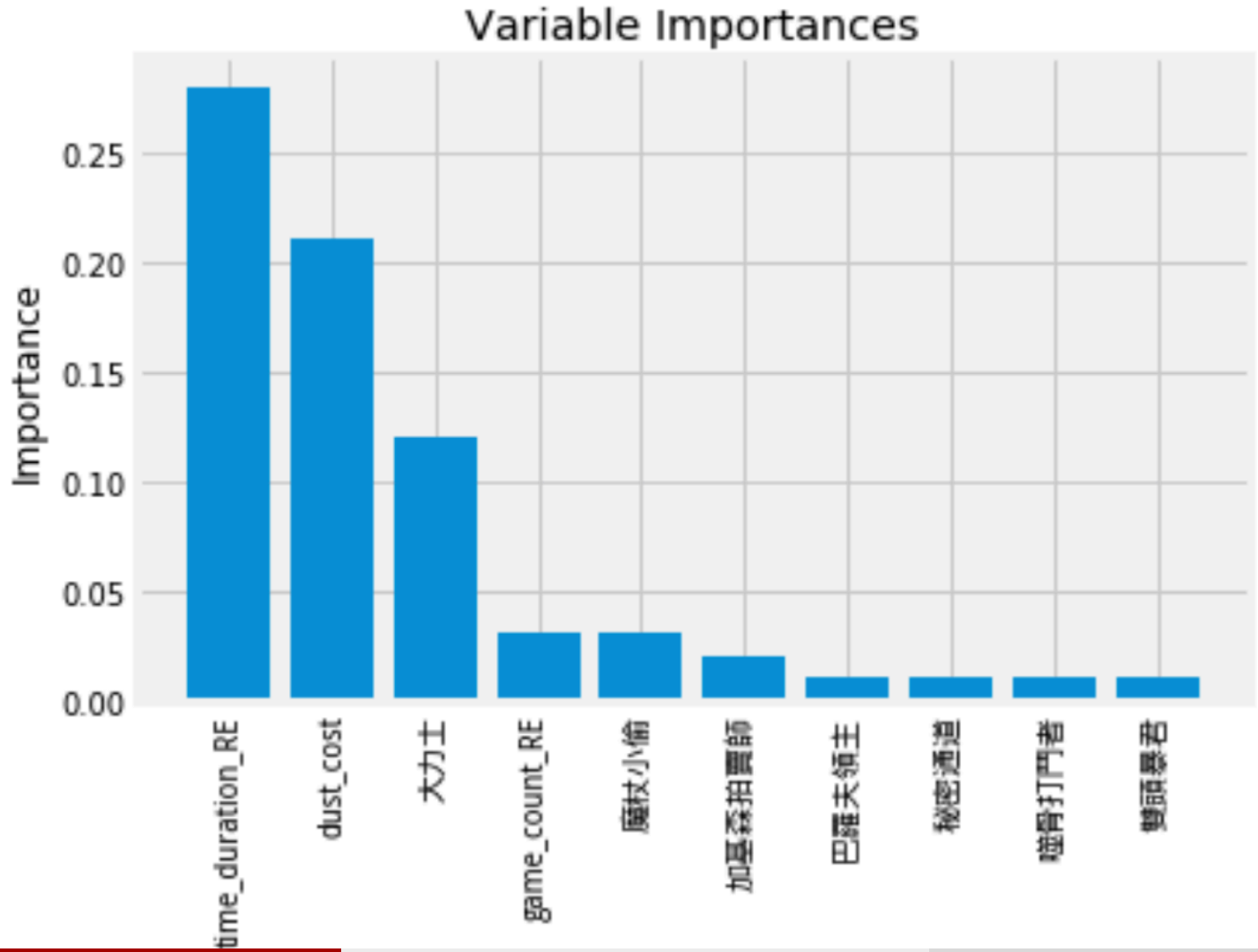
[1]	win_rate	4.67218665473637
[2]	time_duration	0.00015665654829 4709
[3]	安全檢查員	0.832009763102095
[4]	盤牙督軍	2.55282958928911
[5]	維克圖斯	0.154116569006 78
[6]	魔杖師	-1.49511713454943
[7]	模範生史黛琳娜	4.34942441528714
[8]	愛玩筆的學生	1.76100661192004
[9]	討人厭的教師	1.78894309462431
[10]	導覽員	1.48871453526339
[11]	火爆的赤紅龍人	2.26300441482049
[12]	閃避飛龍	2.16930380979627
[13]	盜匪傘兵	1.09320992464284
[14]	托爾托朝聖者	1.78211572392716
[15]	虛無行者	0.78900693038165

Random Forest

透過 Random Forest，我們得到：



Random Forest



Random Forest

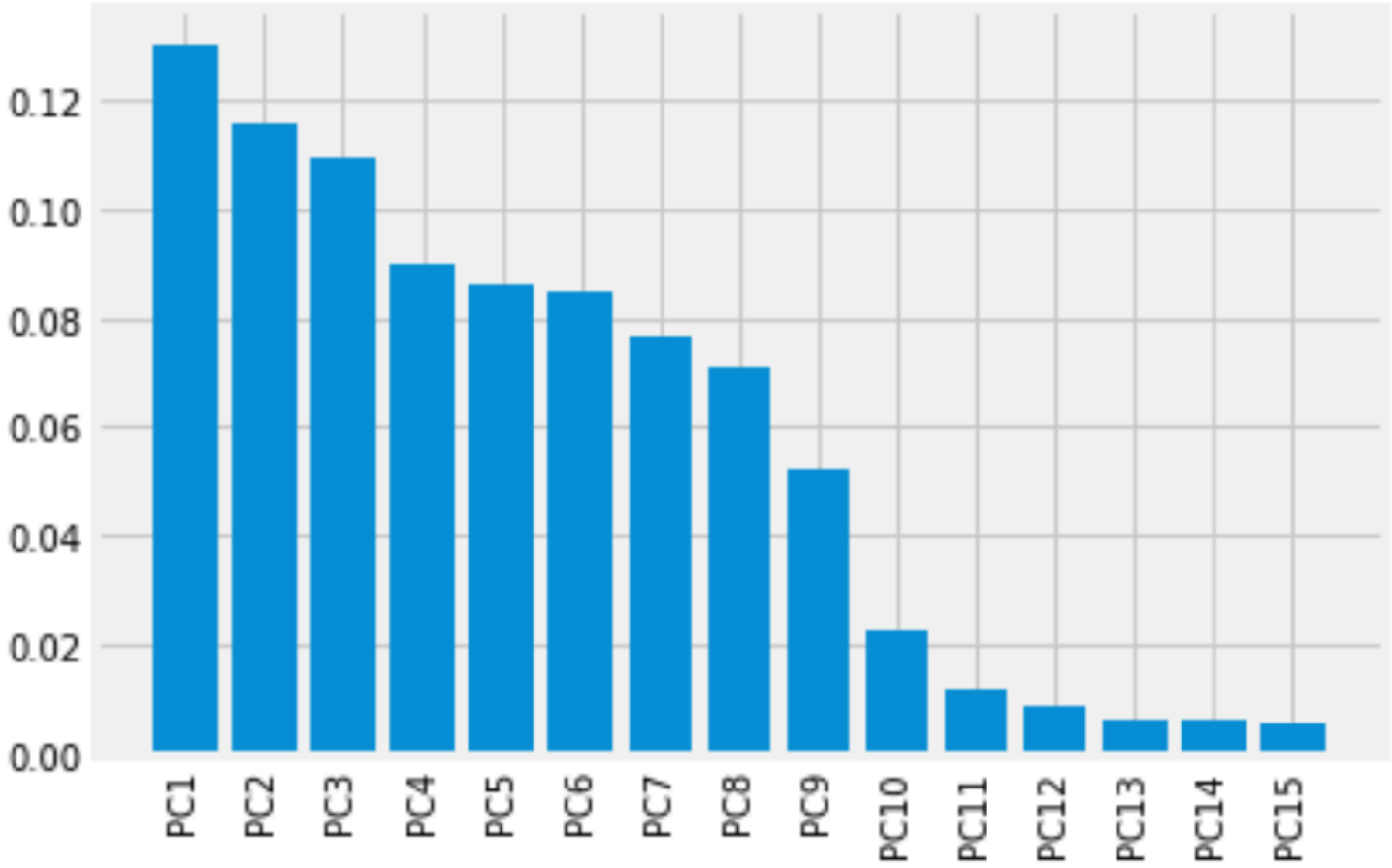
然而我們仍然擔心會有 the curse of high dimensionality 的問題，

因此針對所有的卡牌（共 973 張）做 PCA，即：

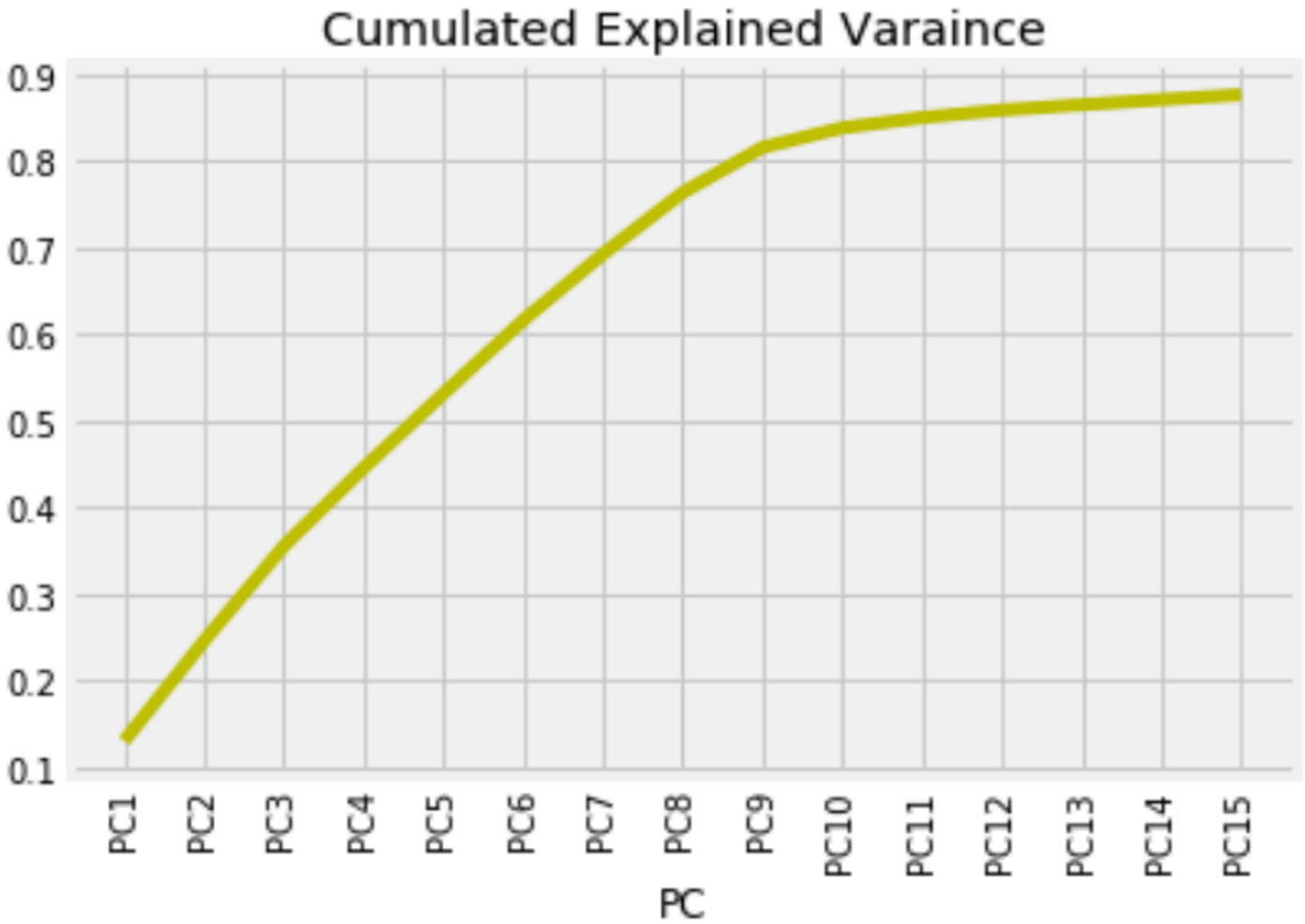
夠過 PCA 將 973 張卡牌做 dimension reduction，期望能得到

extracted information

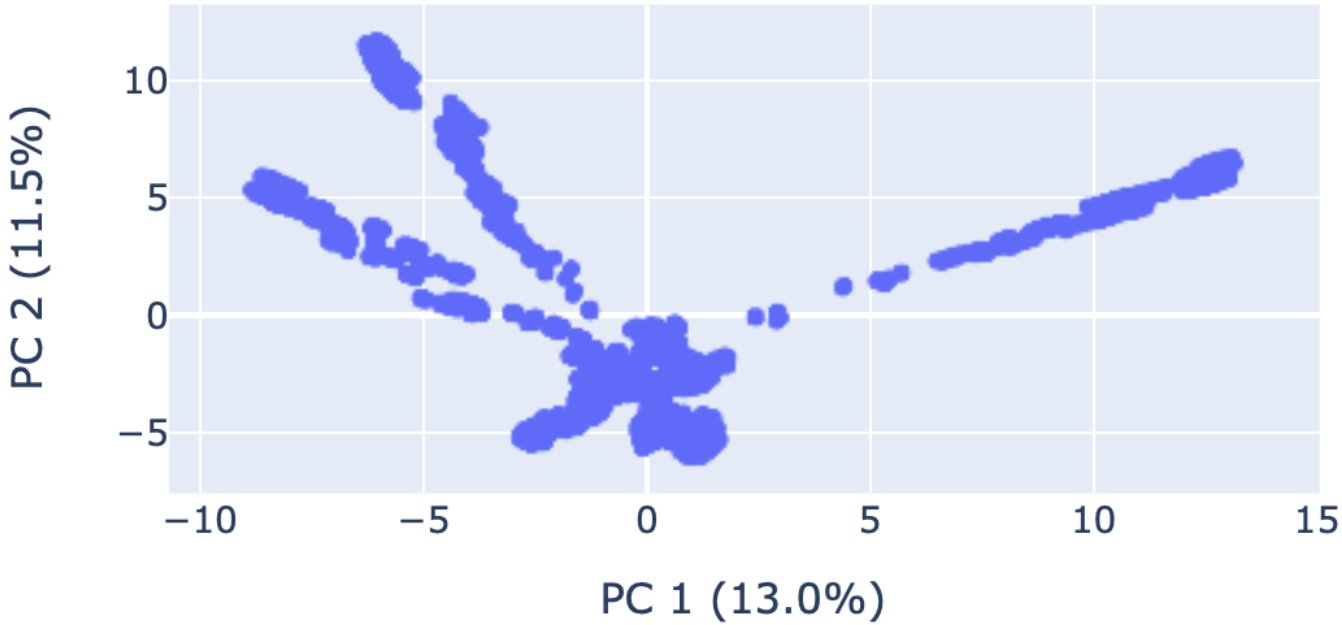
Random Forest with PCA



Random Forest with PCA

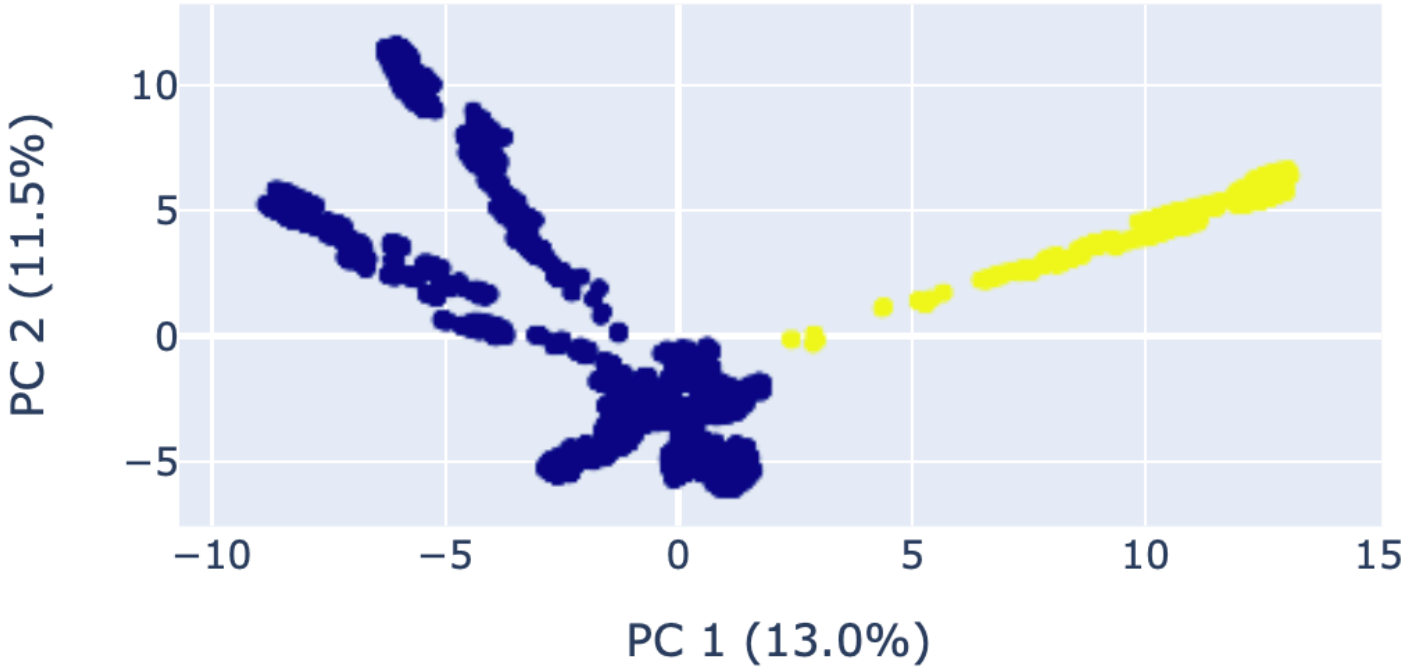


Random Forest with PCA

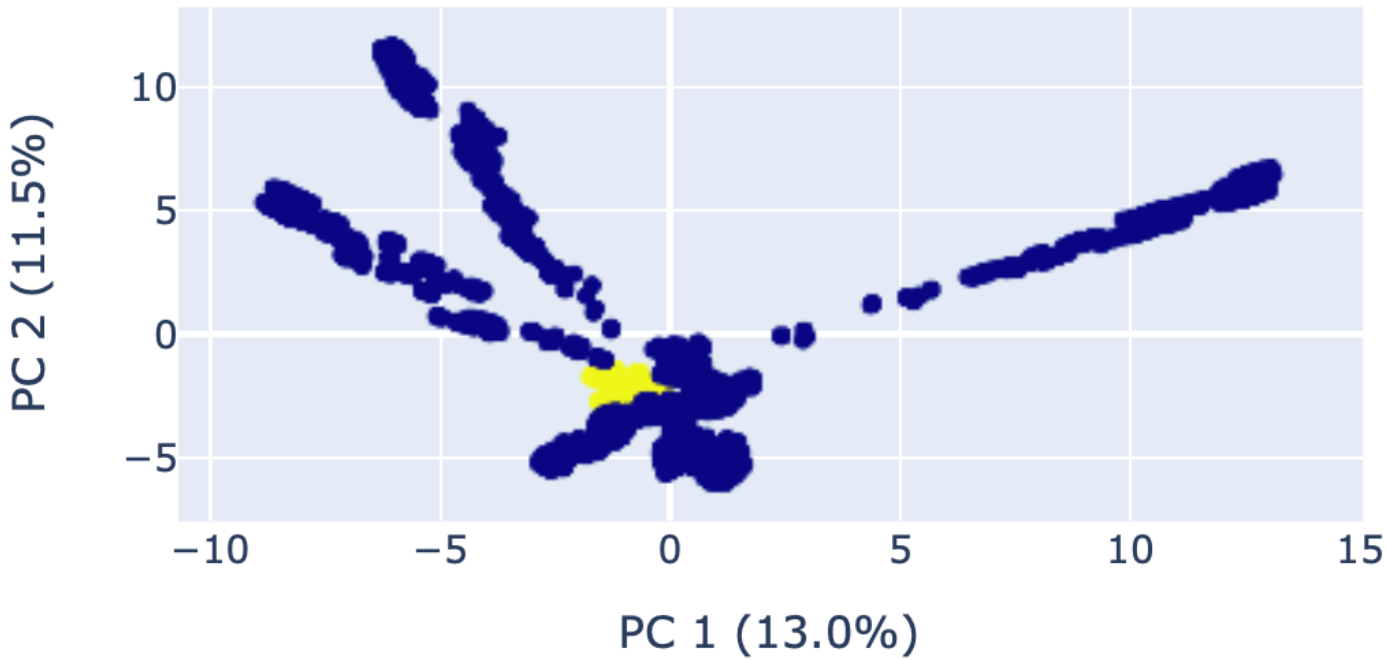


放射狀？

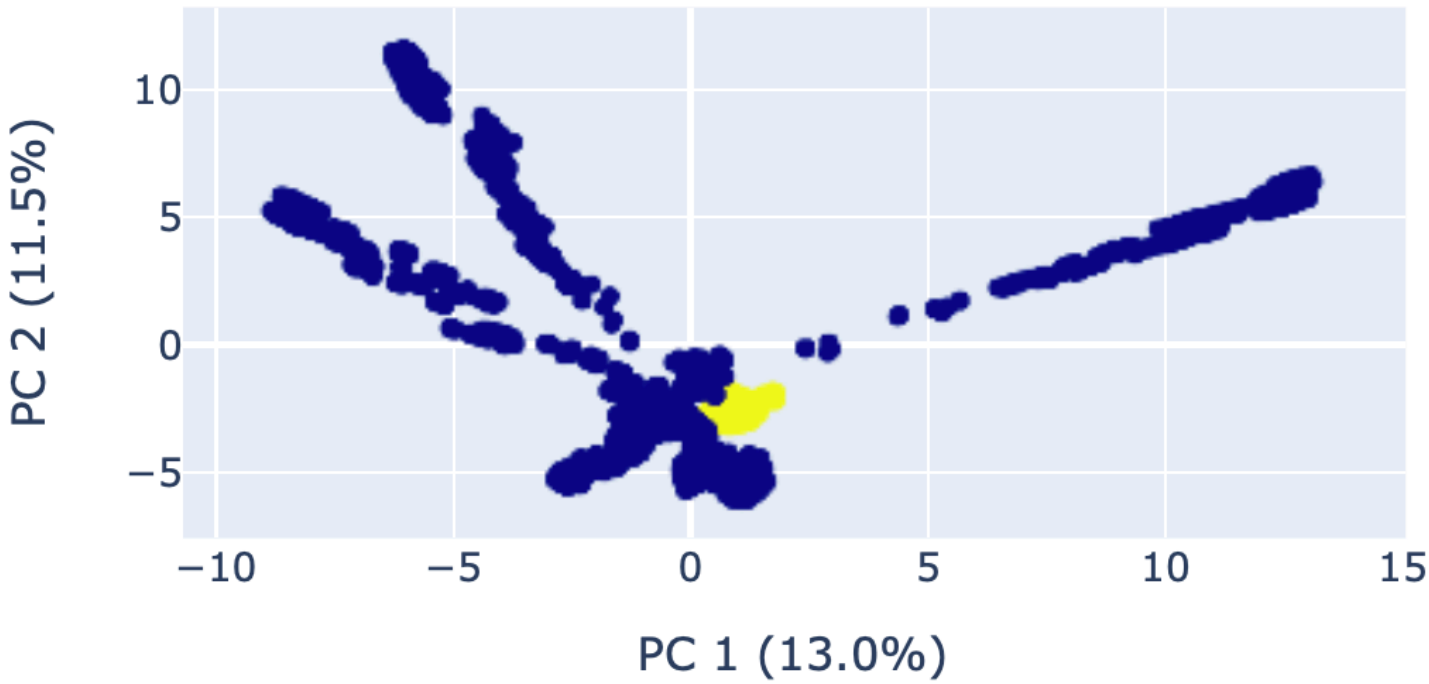
Random Forest with PCA: 職業 dependent 惡魔獵人



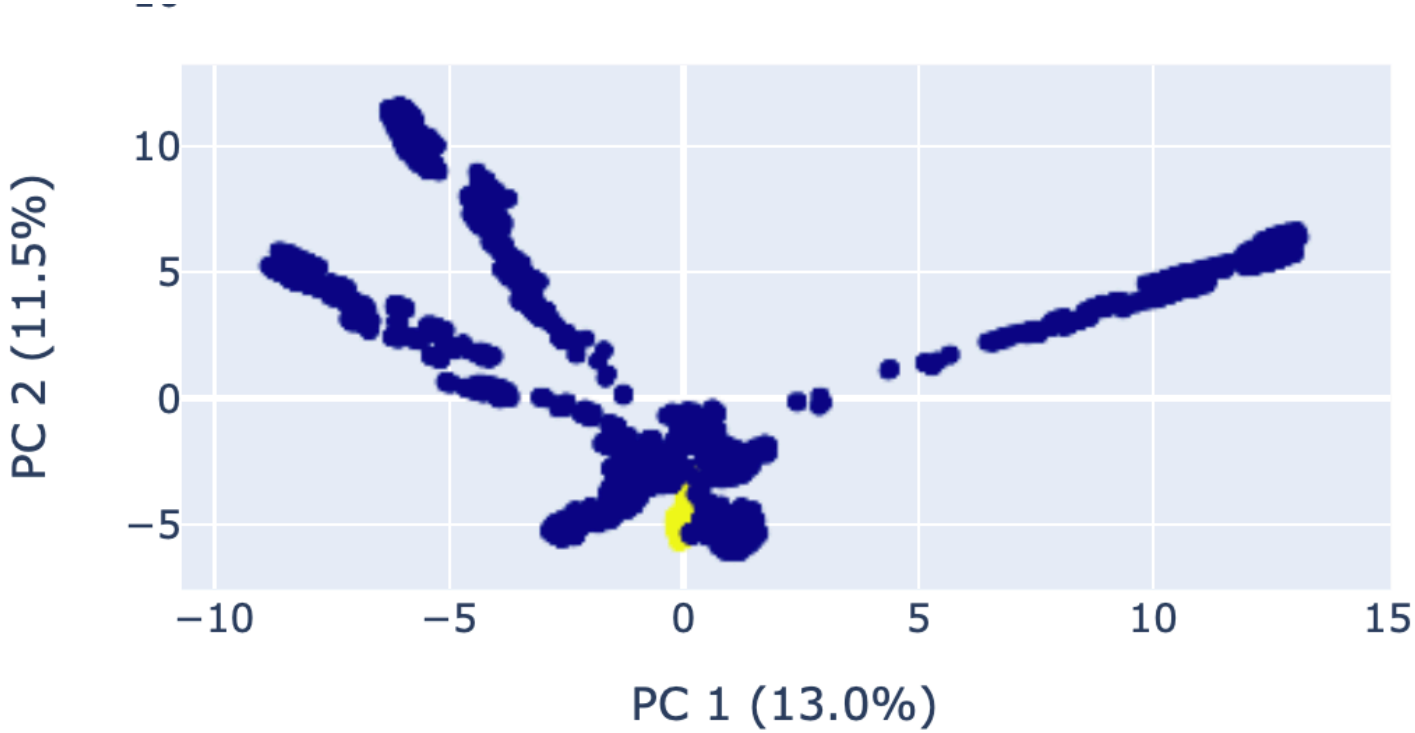
Random Forest with PCA: 職業 dependent 德魯伊



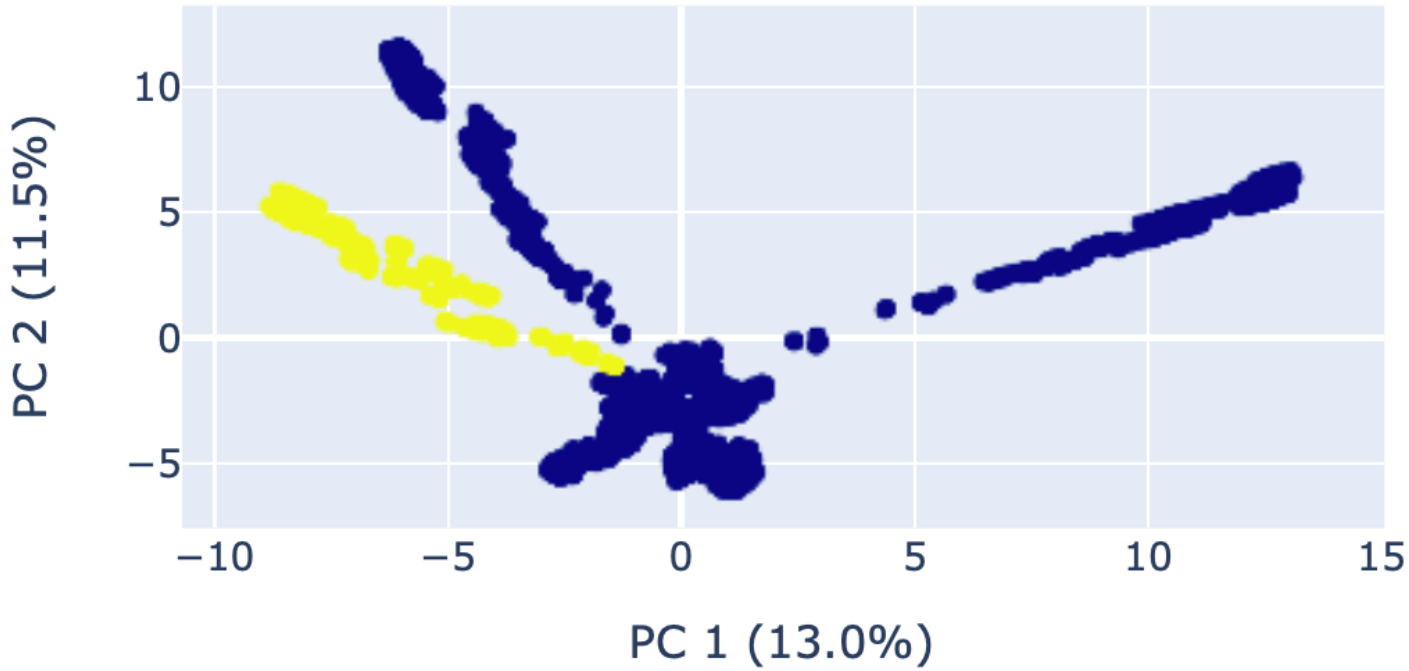
Random Forest with PCA: 職業 dependent 獵人



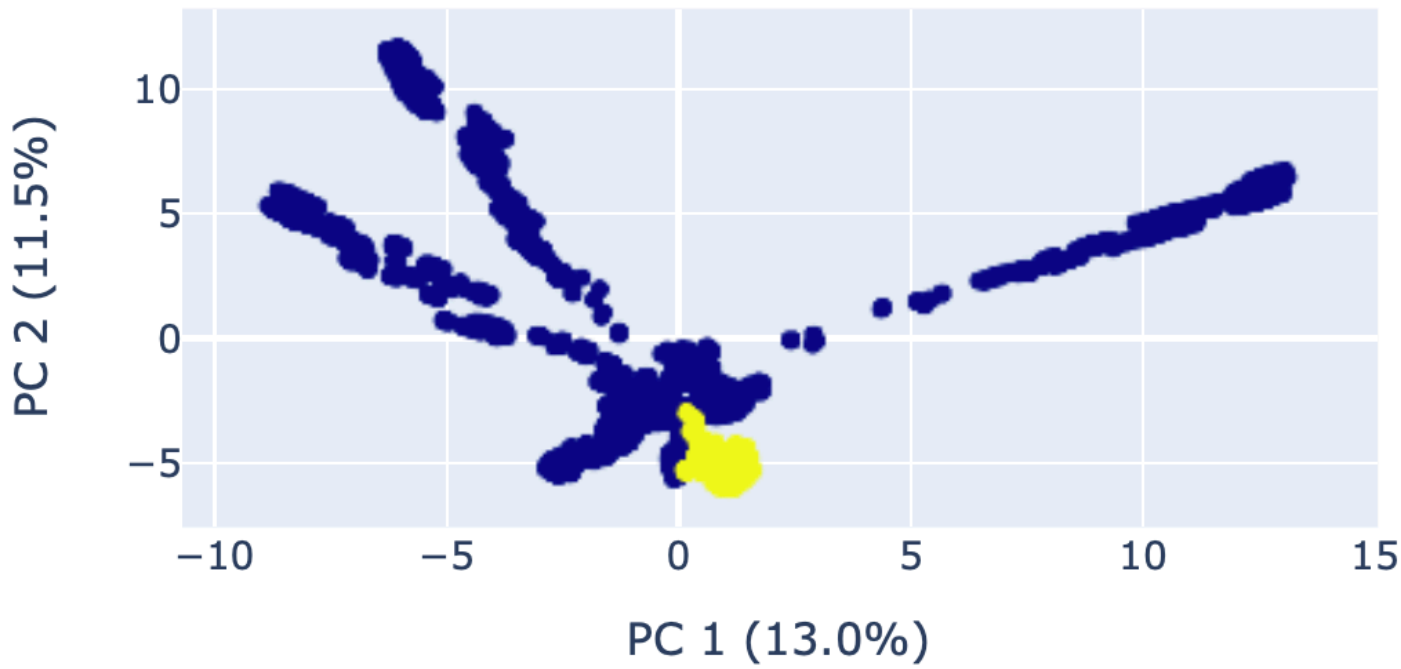
Random Forest with PCA: 職業 dependent 法師



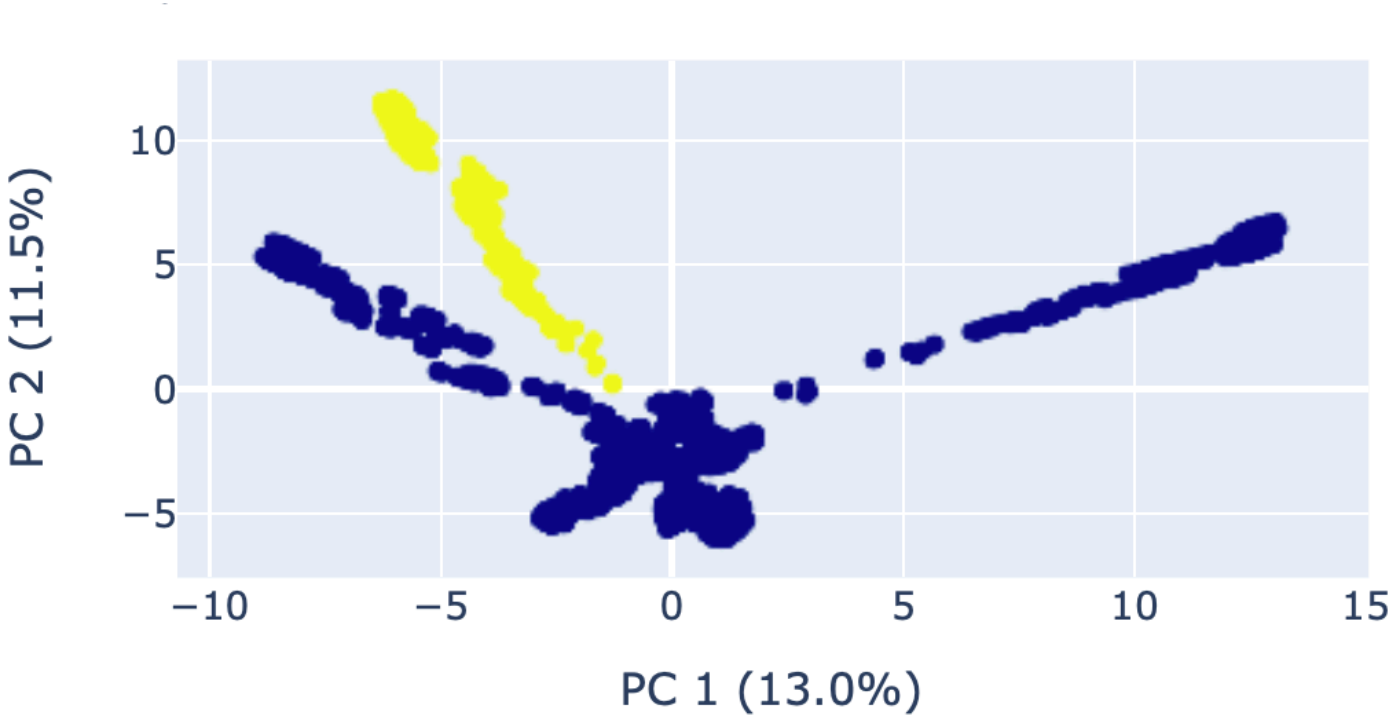
Random Forest with PCA: 職業 dependent 聖騎士



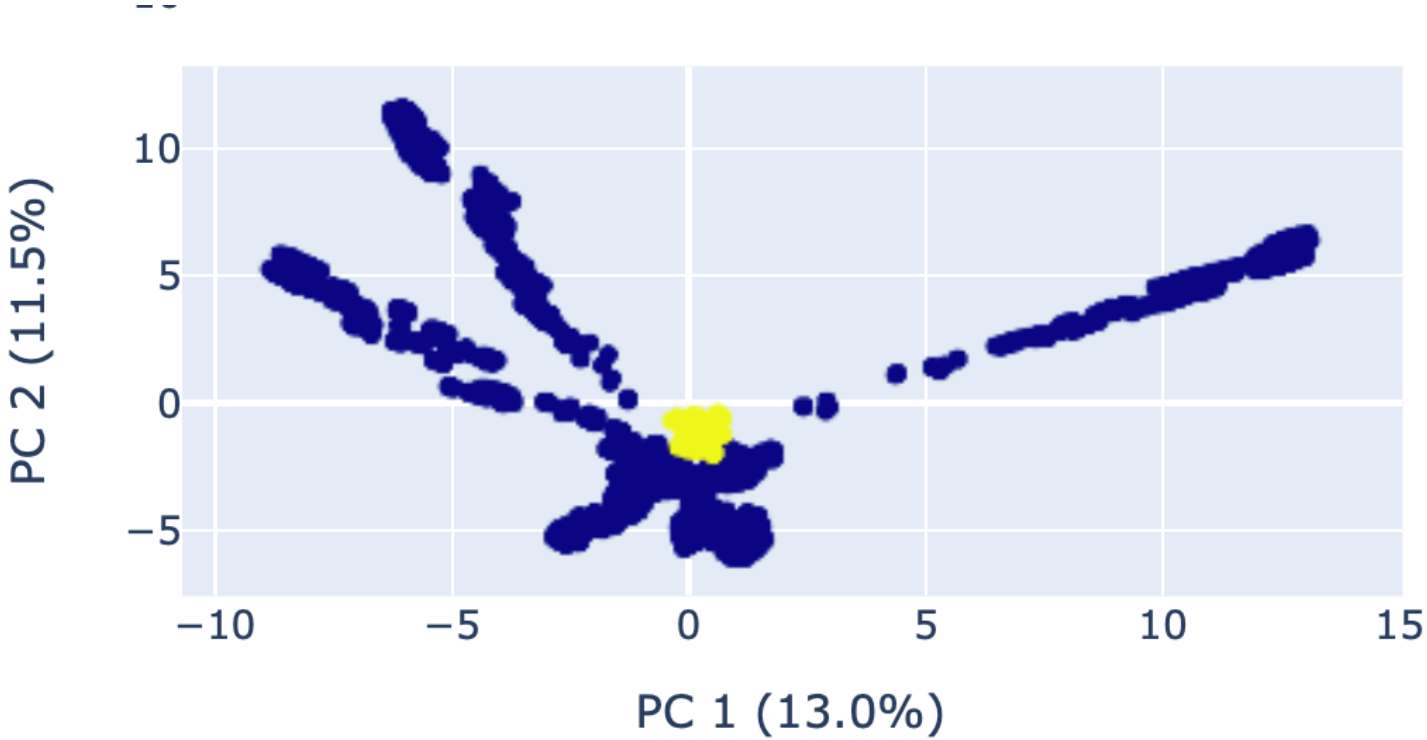
Random Forest with PCA: 職業 dependent 盜賊



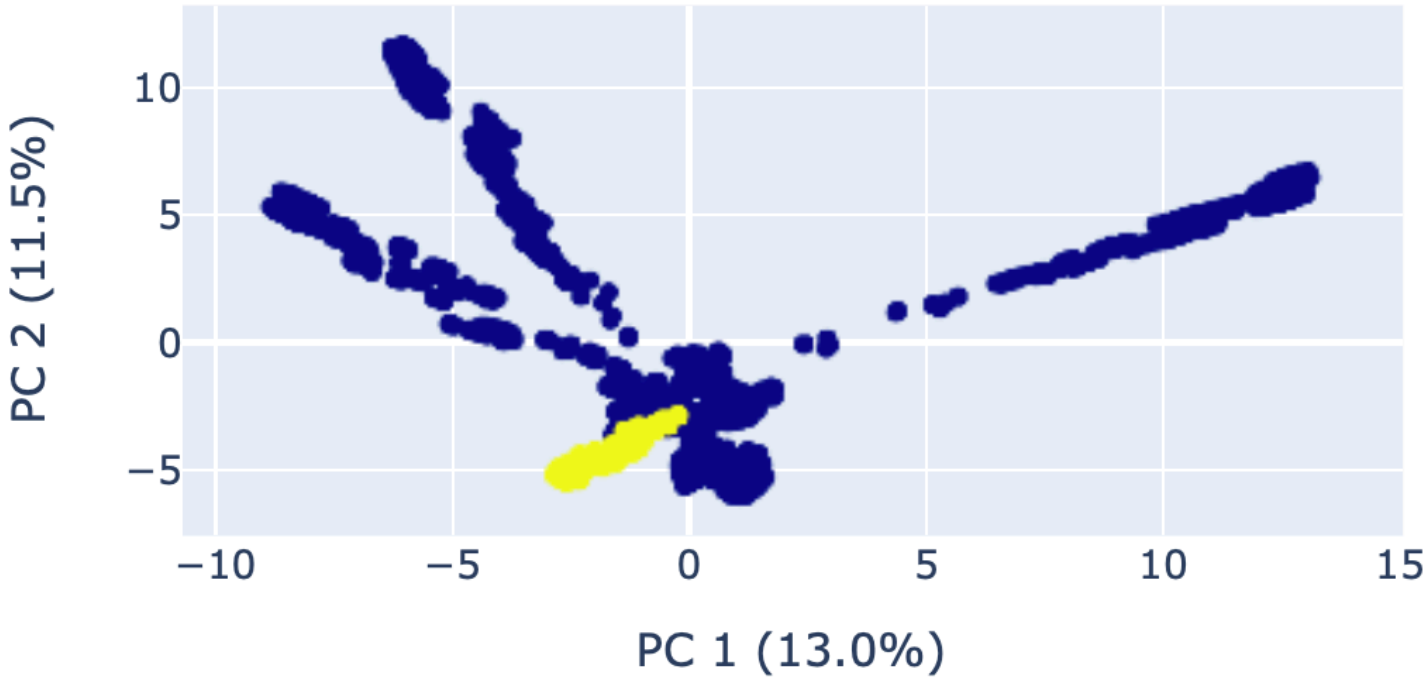
Random Forest with PCA: 職業 dependent 薩滿



Random Forest with PCA: 職業 dependent 術士

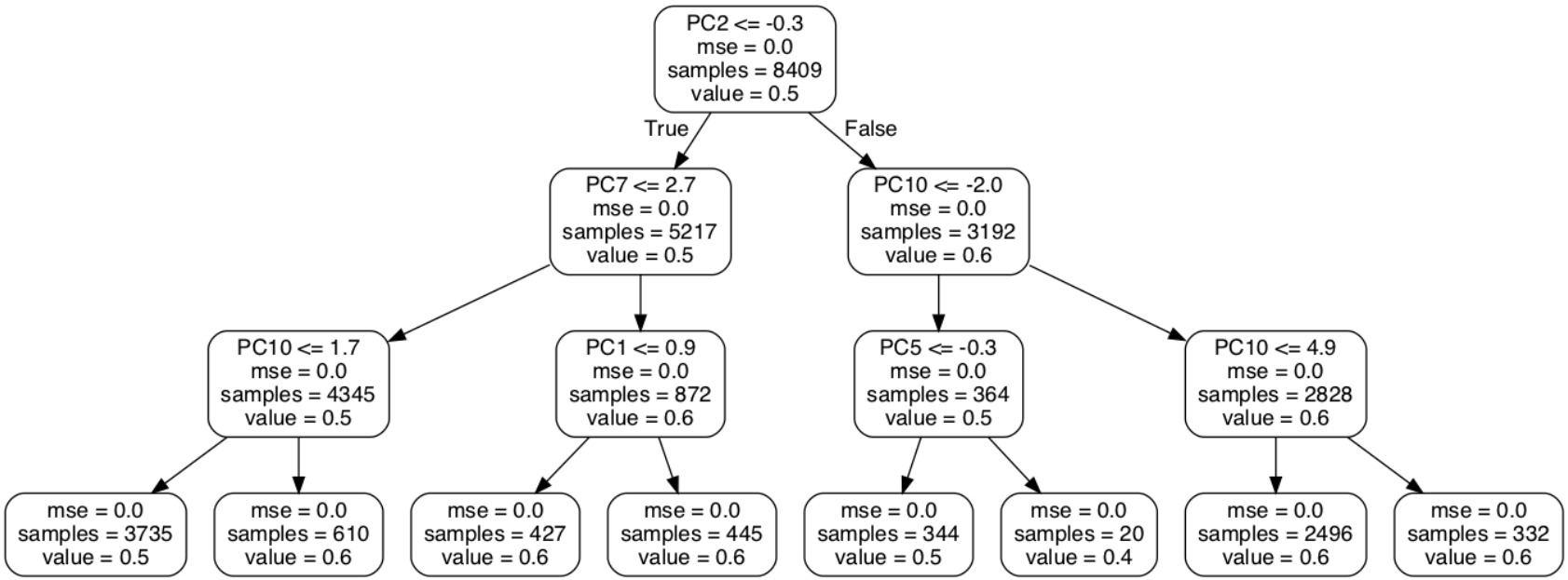


Random Forest with PCA: 職業 dependent 戰士



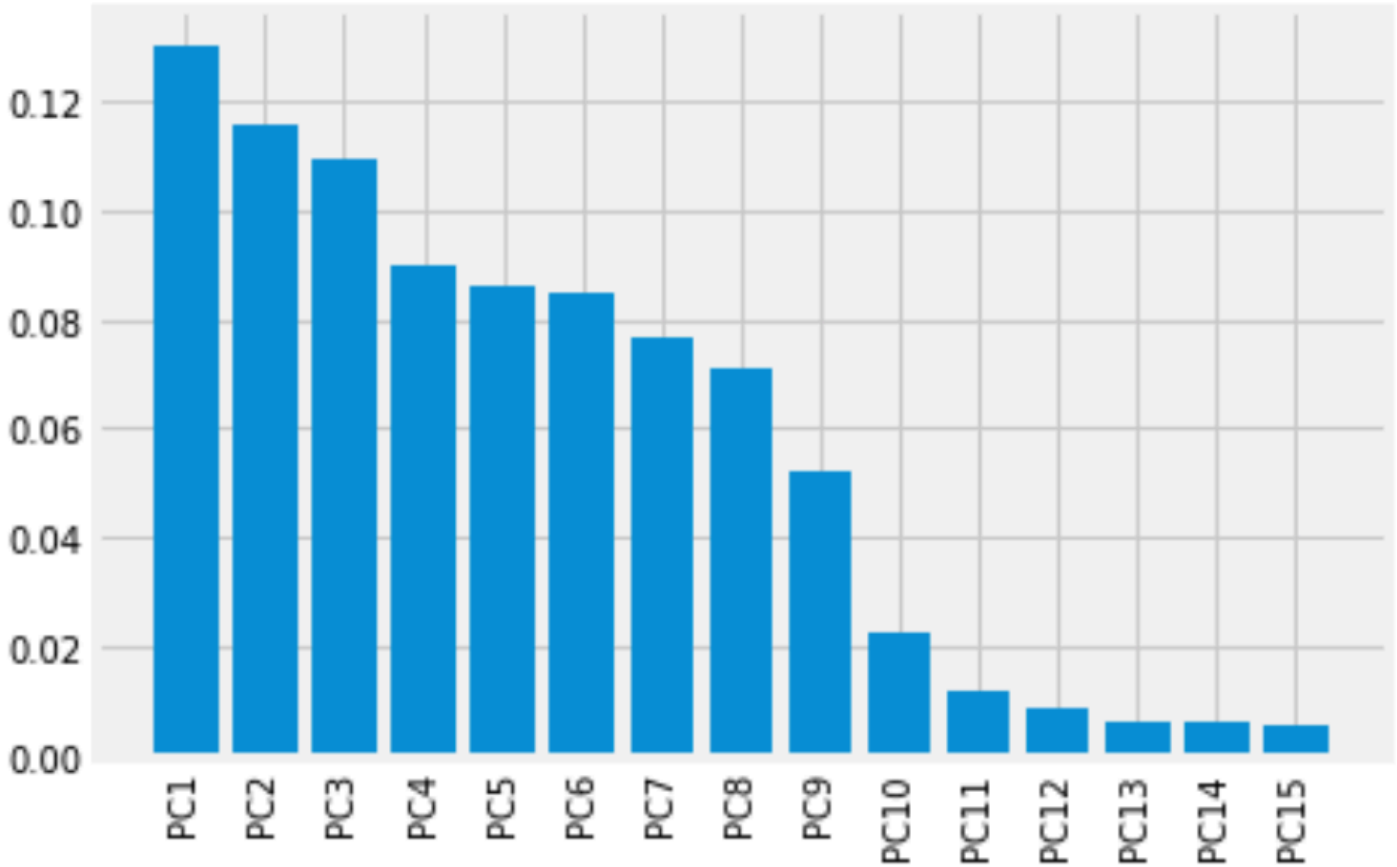
Random Forest with PCA

Hard to explain...



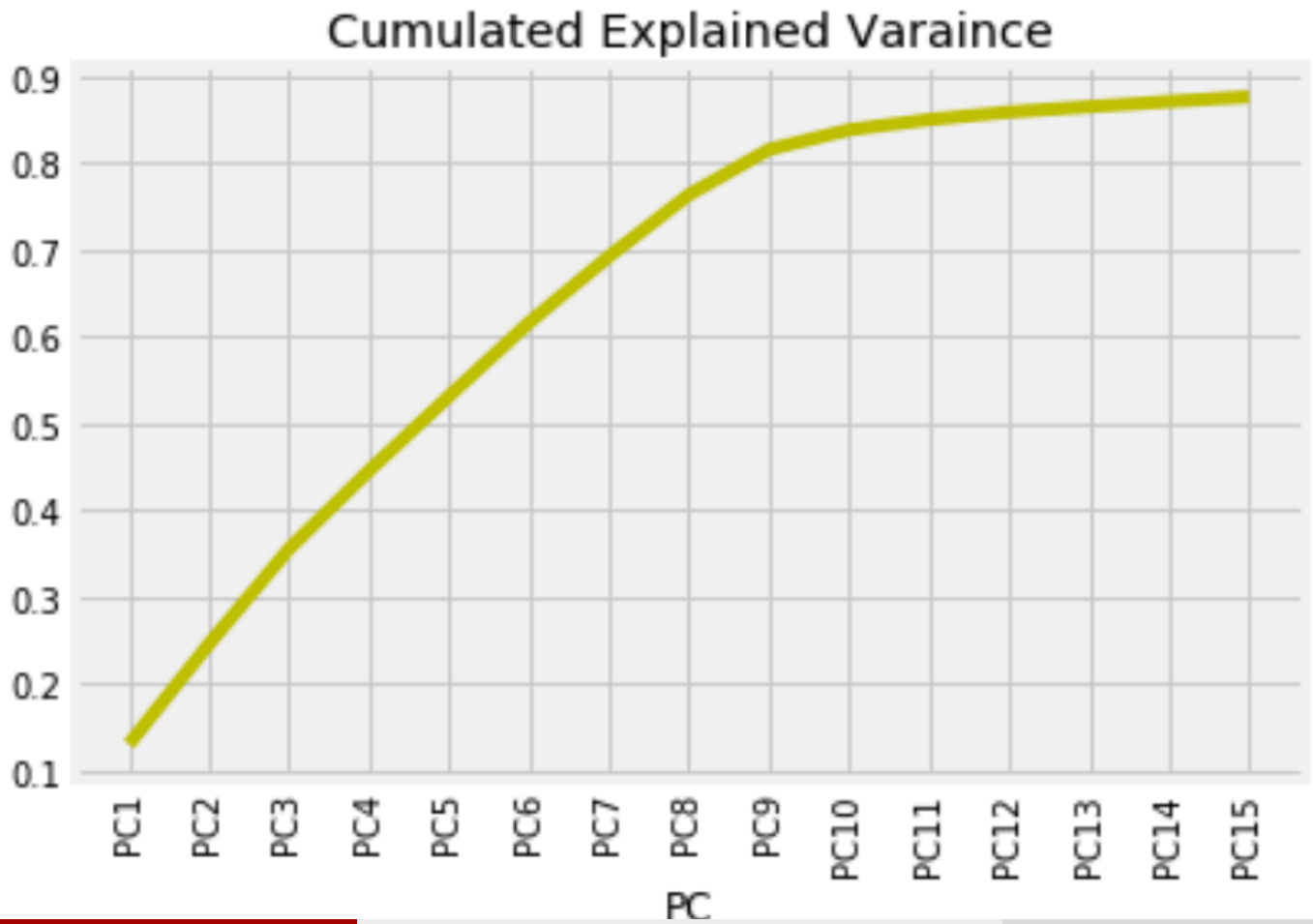
Random Forest with PCA: For Neutral Cards

To avoid mixing up 職業 and 職業專屬卡牌



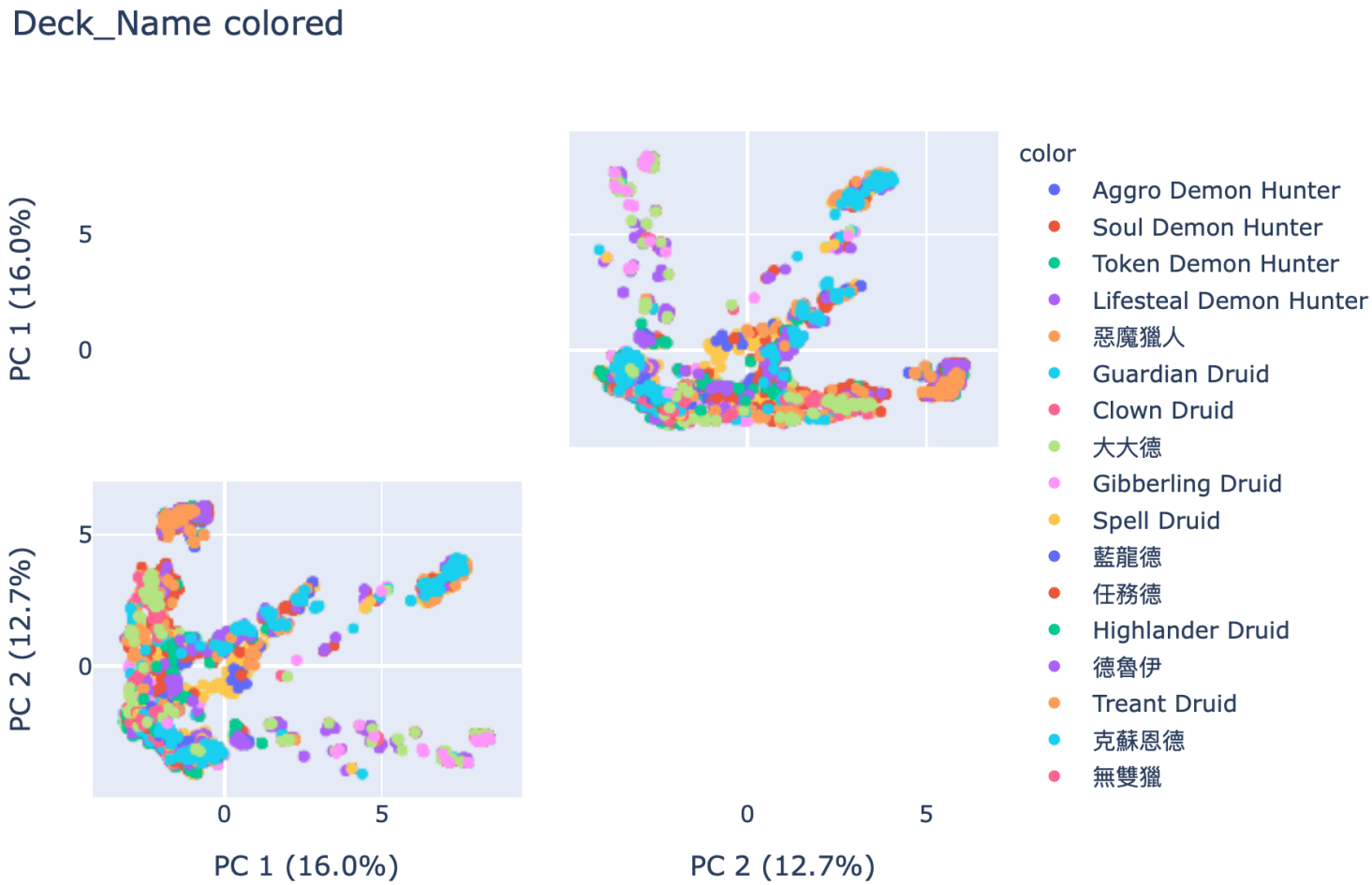
Random Forest with PCA: For Neutral Cards

To avoid mixing up 職業 with 職業專屬卡牌



Random Forest with PCA: For Neutral Cards

Plot the deck names

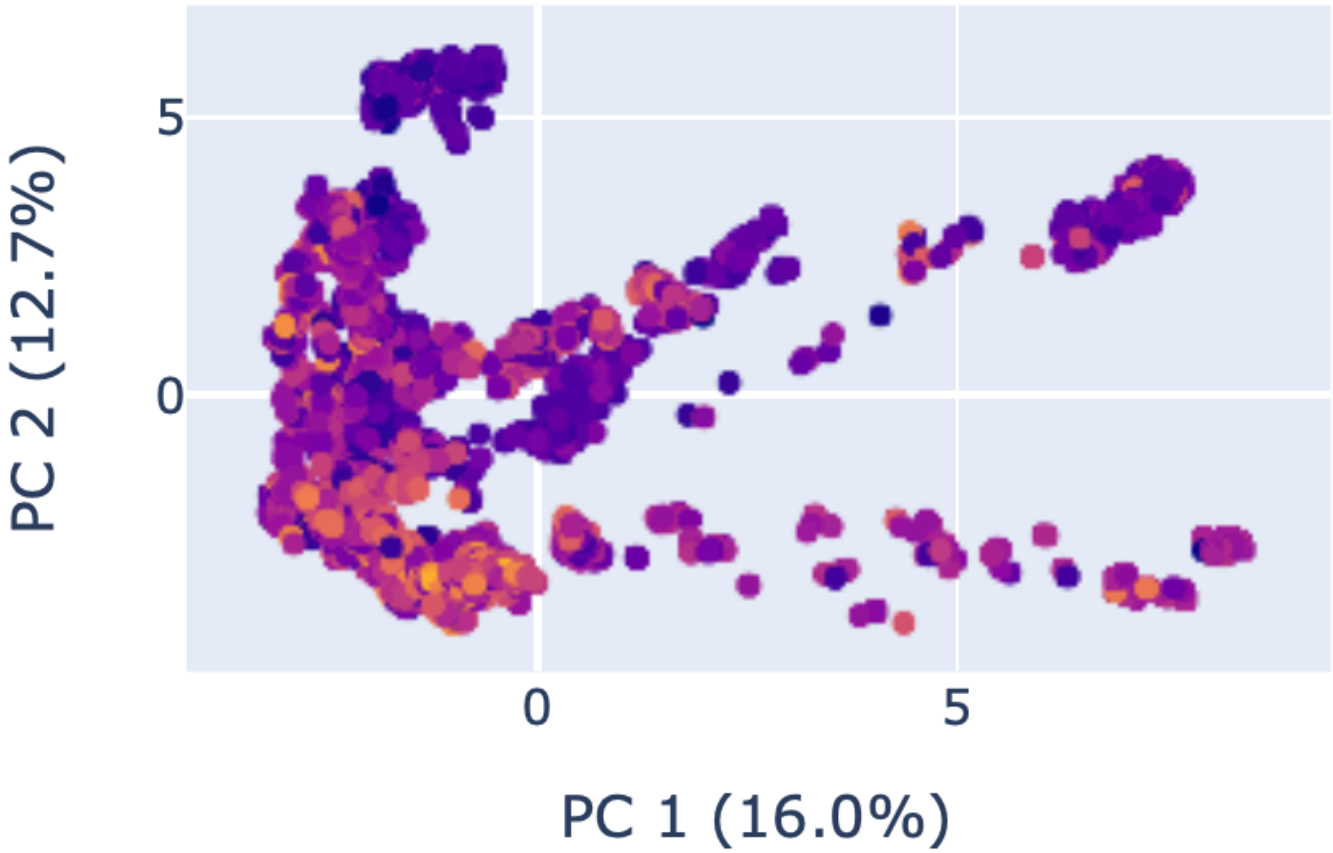


Random Forest with PCA: For Neutral Cards

Check the gif

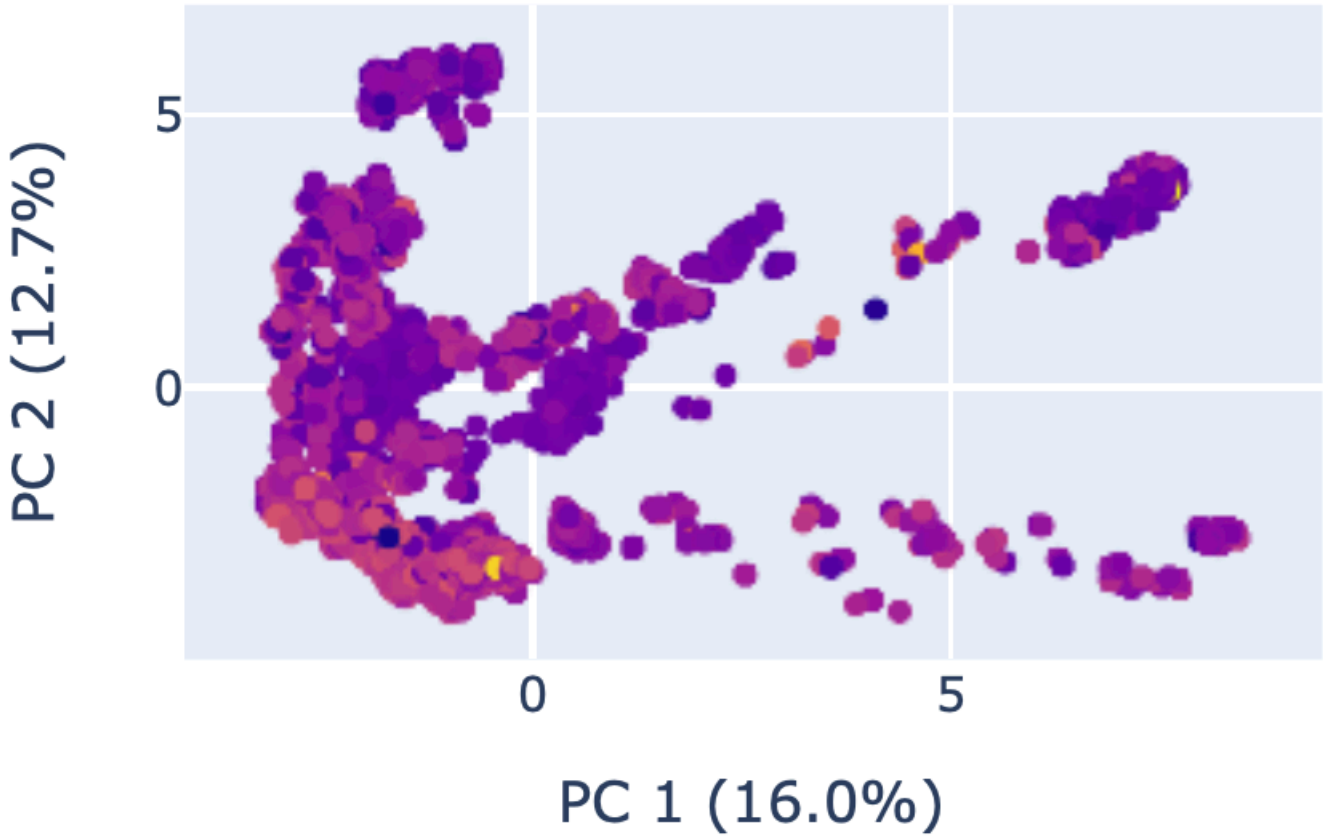
Random Forest with PCA: For Neutral Cards

labeling dust cost

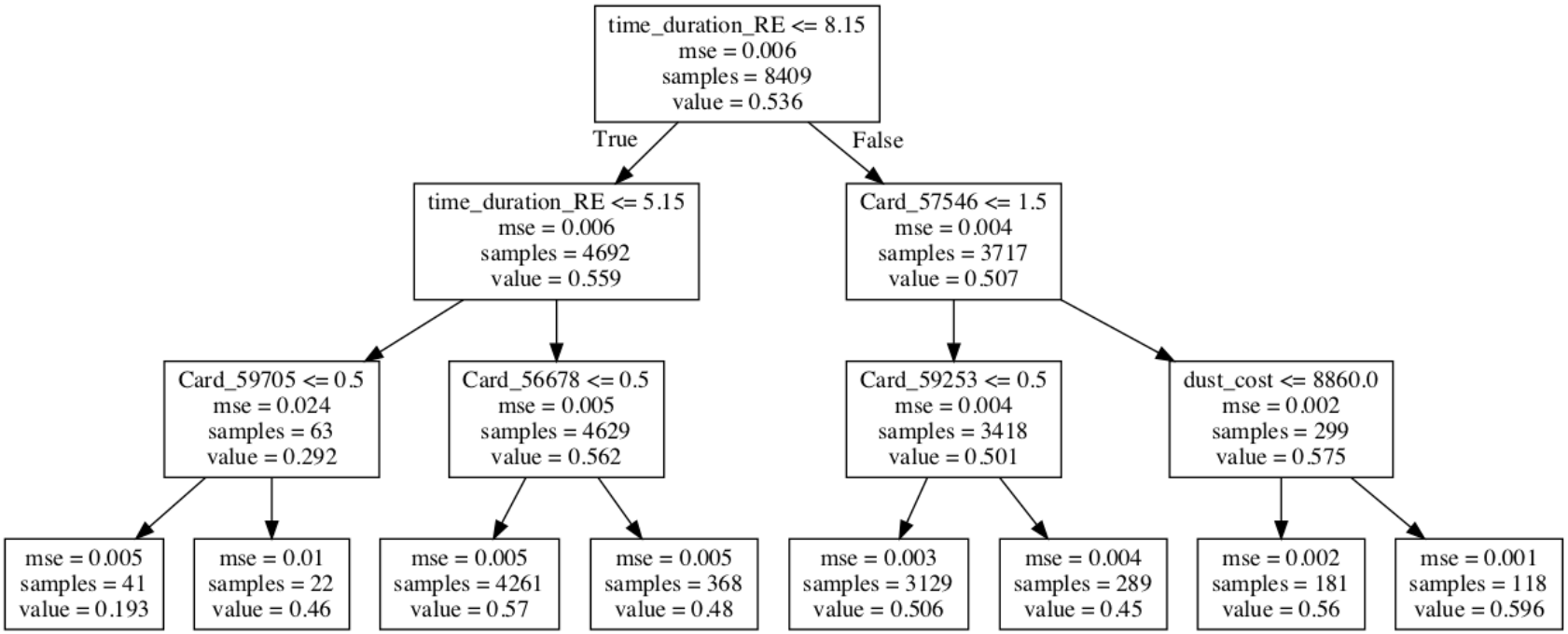


Random Forest with PCA: For Neutral Cards

labeling time duration



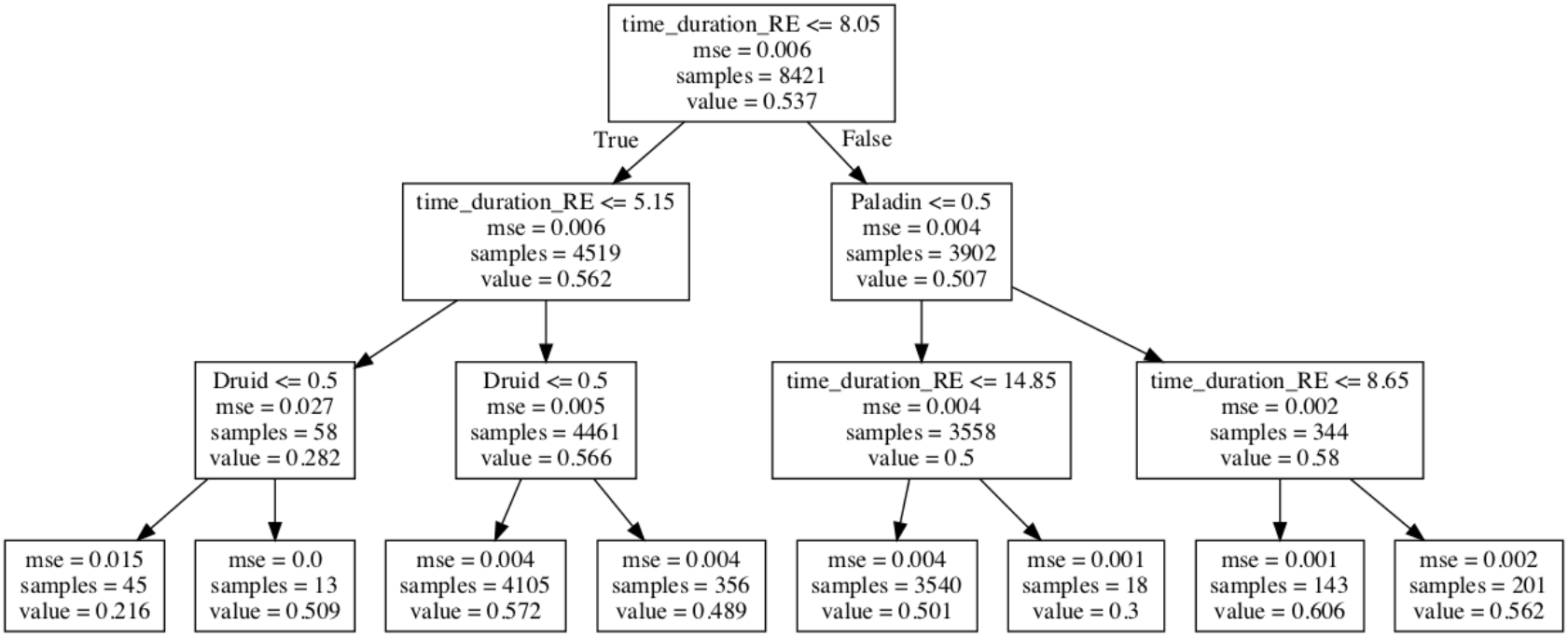
Random Forest with PCA: For Neutral Cards



Random Forest with PCA: For Neutral Cards

Variable: time_duration_RE	Importance: 0.6555
Variable: Card_57546	Importance: 0.1152
Variable: Card_56262	Importance: 0.0663
Variable: Card_467	Importance: 0.0366
Variable: Card_56678	Importance: 0.0308
Variable: Card_59705	Importance: 0.0231
Variable: Card_59253	Importance: 0.0181
Variable: Card_59553	Importance: 0.0142
Variable: Card_59026	Importance: 0.0072
Variable: dust_cost	Importance: 0.0045
Variable: Card_52810	Importance: 0.0029
Variable: Card_59539	Importance: 0.0025
Variable: Card_56680	Importance: 0.0022
Variable: Card_56686	Importance: 0.0021
Variable: Card_55037	Importance: 0.0019
Variable: Card_60023	Importance: 0.0014
Variable: Card_56687	Importance: 0.0013
Variable: Card_55262	Importance: 0.0013
Variable: Card_59001	Importance: 0.0012
Variable: Card_51790	Importance: 0.0012
Variable: Card_59450	Importance: 0.0012
Variable: Card_836	Importance: 0.0012
Variable: Druid	Importance: 0.0011
Variable: Card_61461	Importance: 0.001

Random Forest with PCA: For Neutral Cards



Random Forest with PCA: For Neutral Cards

```
Variable: time_duration_RE     Importance: 0.6972
Variable: Druid                Importance: 0.1849
Variable: Paladin              Importance: 0.1132
Variable: dust_cost            Importance: 0.002
Variable: game_count_RE        Importance: 0.0017
Variable: PC1                  Importance: 0.0009
Variable: PC2                  Importance: 0.0
Variable: Warrior              Importance: 0.0
Variable: Warlock              Importance: 0.0
Variable: Shaman               Importance: 0.0
Variable: Rogue                Importance: 0.0
Variable: Priest               Importance: 0.0
Variable: Mage                 Importance: 0.0
Variable: Hunter               Importance: 0.0
Variable: Demon Hunter         Importance: 0.0
```

Conclusion

Conclusion

- A baseline model (only using a fixed sample mean to predict the winning rate) predicts a fixed 53% winning rate.
- Even a OLS model induce that same conclusion that time duration is most important feature.
- A certain card may play an important role in a deck, but the composition plays a more important role.
- The cost of a deck matters, but not a lot.

Further Topic

Next Step...

經過以上分析，我們仍有以下幾處問題尚待解決：

- A better model selection decision for a tree-based model
- the validity of a continuous / discrete variable in Random Forest
- Random Forest with PCA dimension reduction
- There may exist a more reliable model for modeling the distribution of the winning rate