MIS Project Report

team 3

指導教授：魏志平 曹承礎

組員：余知諺 黃思凱 陳映樵 高宗毅 林鼎崴

一、 實驗動機

在美國職籃（NBA）中球員的薪資一直都是判斷一個球員價值最直接的指標，明星級的球員和一般球員的薪資差距非常大。我們發現，球員的表現，與在賽季初簽約的約定薪資，並不是永遠反映出球員的價值，例如：現為底特律活塞對的羅斯（Derrick Rose)，2019年的年薪為750萬美元，因為前幾個賽季的傷病，使得薪水在過去幾年調降非常多，但在2018年賽季，羅斯能夠交出18分與4.3助攻的成績，而且在球迷的心目中，羅斯的價值遠大於所得的年薪。

所以我們的發想，是找出球員的薪水是否和球員的表現及價值符合，以及某些球員薪水被高估或低估的原因為何。以此作為出發點，我們在Kaggle上找到了包含球員各項數據的資料集，以及球員在社群網站影響力的資料集，以這兩個資料集來實作分析。

我們希望能夠透過這次的實驗，找出被影響薪資的特徵，進而分析出被高估或低估的球員。這個分析，能夠應用的領域也很有拓展性，例如：公司員工的表現與薪資是否有被高估或低估、商品的價格該怎麼去符合商品的價值且能夠獲得利潤，我們認為所做的分析，是有能夠延伸應用到不同的領域的潛力的。

二、 文獻探討

NBA球員的薪水主要是以球員表現與個人素質作為判斷依據，此外還有許多其他因素可能間接影響到薪資。數據與先前研究顯示，通常在簽長期合約前，為了讓自身的價值能被看見，證明自己有這份能力，球員的表現會有大幅度的提升，但一旦合約確定了，球員的表現會開始呈現下滑的趨勢。而合約的激勵效應則會影響整個球隊的表現，球隊的獲勝次數與球員合約到期數量呈現正相關，但與剛簽訂多年合約的球員人數呈現負相關[1]。另一份研究則顯示，被低估的球員，也就是預測薪水比實際獲得薪水要來得高的人，通常是新進球員，因而無法獲得太高的薪資上限，而被高估的球員，也就是預測薪水比實際獲得薪水來的低的人，通常是巨星、超級巨星等人物。然而，高估球員較多的球隊卻比起低估球員較多的球隊擁有更高的勝率[2]。

雖然在用nba球員表現預測薪水這方面已有不少的研究，然而先前模型的預測結果並不盡理想，且先前研究並未對薪水的落差提出高估低估的概念，並探討這些落差出現的原因，因此我們希望在這次專案中補強以上的部分。另外，由於我們資料有球員的群影響力，因此找出社群影響力(可能與廣告利益相關)與薪水高低的相關性也是我們此篇報告的特色之一。

我們的資料中總共包含了數十項特徵，其中包括相關、無關與冗餘三個部分，存在冗餘的原因是因為一個相關特徵在存在與之高度相關的另一個相關特徵時可能變成多餘的存在，因此看完先前針對此問題之研究後，我們決定使用特徵選擇來篩選需要的特徵。在機器學習和統計中，特徵選擇可用來簡化模型、減少訓練時間、避免維數災難與過度擬合的發生，主要分成包裝法、過濾法與嵌入法三種類別。包裝法指的是用每個特徵子集來訓練一個新模型並測試錯誤率來給予特徵子集評分，計算量龐大。

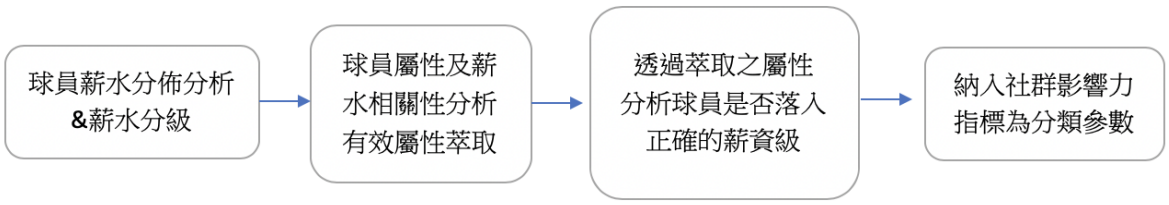
過濾法採用代理指標，而不是以特徵子集的錯誤率評分，常見指標如相關係數、卡方檢定、方差選擇等，通常預測能力較包裝法差，但優點在於計算速度較快，有時也可為包裝法做預處理的動作。嵌入法指的是先計算出特徵間的權重關係，並以此為基準來篩選特徵，複雜度介於包裝法與過濾法間[3]。

三、 研究模型

我們希冀以如下的流程完成本次的實驗。我們的實驗目的包括:

- 分析球員薪水分布並將薪水分級
- 球員屬性與薪水高低相關性分析
- 透過使用萃取出之屬性分析選手是否落入非自己目前之薪資級
- 融入社群影響力參數進行球員薪資分類

由於前述之結果是有先後之關係，因此我們會以下列之流程圖完成此實驗:



針對上述之流程會再有亙深入之解釋如下:

1. 分析球員薪水分布並將薪水分級

在此階段的研究，我們希望可以更了解手上的資料。需要討論的問題包括：資料分布情形及資料的完整度。針對資料分布情形，做更嚴謹的分類。希望可以將薪水層級分為：高、中、低，三類薪水層級。

目前討論中可行的分類方法有:

	低	中	高
方法一	25%以下	25-75%	75%以上
方法二	薪水<mean-1*std	mean-1*std <薪水<mean+1*std	薪水>mean+1*std

初步的想法是將薪水是以10million為一個單位來區分出薪水的等級，希望能反映出球員薪水高低與數據表現是成正比的，但仔細看過數據集後發現，很多在聯盟中非一線的球員也有相當高的薪水(可能為溢價的合約)，意即這位並不屬於全明星的行列，但卻有著與全明星的球星差不多的薪水，這可能會成為之後model判斷時的一種bias，而另一種bias是有些很年輕的球員，在生涯初期就打出全明星數據，但由於還處在生涯的第一或第二張合約，所以還在領新秀合約或生涯初期的薪水，這兩種bias可能會使得accuracy不夠準確。這兩種情況也多存在於10million以上的球員，因此我們將原本的分類法改分成三類分別是10million以上，1~10million，以及1 million以下，按照四分位數的結果這三類也分別大約對應到前25%、前25%~前75%(共50%)、以及後25%。另外一種我們想到的分類方法是將薪水做標準化，範圍在正負一個標準差內。當有更深入的研究薪水分佈後，會在定義出。

2. 球員屬性與薪水高低相關性分析

此階段希望透過使用特徵選取的方向。雖然是一個中間的過程，但是對於第三部模型預測薪資級有幫助外，對於理解NBA薪水的核定標準有更深入的認識，如：薪資高球員通常都有的特質。這些副產品對於我們了解NBA給新標準有更深入的了解。

3. 透過使用萃取出之屬性分析選手是否落入非自己目前之薪資級

最後，我們會運用前面步驟萃取之屬性，以及分級之薪資，更深入的透過萃取的屬性做分類到分級的薪資級中。從這個過程，我們可以得到球員薪水多給或少給的判斷。進而探討其背後之原因。我們在此階段會嘗試多種模型來進行分類，包括：Naïve Bayes, Decision Tree 到複雜的類神經網路來完成。並且比較何模型比較適合做今天的任務。最後，會使用10-fold cross validation或leave one out 方式評斷模型的好壞。

4. 納入社群影響力指標為分類參數

建立在前面的實驗結果，我們會使用社群影響力資料(twitter轉發及twitter favorite數) 的資料加入當作參考當作分類的依據。會有這樣的想法是希望可以透社群影響立為是否為球星的指標，進而判斷是否影響其為高薪或低新的影響因素。

四、 實驗

1. 資料集描述

本次專案的資料集取自於kaggle (https://reurl.cc/Obo7QX)，資料集名稱為nba_2017_nba_players_with_salary.csv(如下圖)，裡面共有342筆資料描述335位NBA球員各項表現數據及他在該年所領的薪水，每筆資料共有39個欄位。

Unnamed: 0	Rk	PLAYER	POSITION	AGE	MP	FG	FGA	FG%	3P	...	GP	MPG	ORPM	DRPM	RPM	WINS_RPM	PIE	PACE	W	SALARY_MILLIONS
0	1	Russell Westbrook	PG	28	34.6	10.2	24.0	0.425	2.5	...	81	34.6	6.74	-0.47	6.27	17.34	23.0	102.31	46	26.50
1	2	James Harden	PG	27	36.4	8.3	18.9	0.440	3.2	...	81	36.4	6.38	-1.57	4.81	15.54	19.0	102.98	54	26.50
2	3	Isaiah Thomas	PG	27	33.8	9.0	19.4	0.463	3.2	...	76	33.8	5.72	-3.89	1.83	8.19	16.1	99.84	51	6.59
3	4	Anthony Davis	C	23	36.1	10.3	20.3	0.505	0.5	...	75	36.1	0.45	3.90	4.35	12.81	19.2	100.19	31	22.12
4	6	DeMarcus Cousins	C	26	34.2	9.0	19.9	0.452	1.8	...	72	34.2	3.56	0.64	4.20	11.26	17.8	97.11	30	16.96
5	7	Damian Lillard	PG	26	35.9	8.8	19.8	0.444	2.9	...	75	35.9	4.63	-1.49	3.14	10.72	15.9	99.68	38	24.33
6	8	LeBron James	SF	32	37.8	9.9	18.2	0.548	1.7	...	74	37.8	6.49	1.93	8.42	20.43	18.3	98.38	51	30.96
7	9	Kawhi Leonard	SF	25	33.4	8.6	17.7	0.485	2.0	...	74	33.4	5.83	1.25	7.08	15.53	17.4	95.79	54	17.64
8	10	Stephen Curry	PG	28	33.4	8.5	18.3	0.468	4.1	...	79	33.4	7.27	0.14	7.41	18.80	15.1	105.08	65	12.11
9	11	Kyrie Irving	PG	24	35.1	9.3	19.7	0.473	2.5	...	72	35.1	4.35	-2.30	2.05	8.28	13.5	99.12	47	17.64

2. 資料前處理

資料前處理包含刪除不相關features、刪除重複值、薪水label標記、資料標準化四個步驟。在前兩步驟中，我們將明顯無法預測薪水之欄位刪除，接著我們發現資料集中有重複的球員，因此將重複資料刪除。第三步，我們將薪水依據它的分布分成三個類別，將問題簡化為3個class的預測問題。最後，由於每個欄位值差距及維度不同，若直接使用原始資料進行模型訓練將得到不好的效果，因此我們將所有features進行標準化，提高後面模型預測準確率。

2.1 刪除不相關features

在原資料集，薪水除外的剩下38個features中，有以下幾個欄位明顯無法預測薪，如Unnamed: 0、Rk及MP(由於資料集中已有MPG-Minute Per Game與MP重複)，因此我們將以上三個欄位刪除。另外，球員名字、球場位置與所屬球隊也非球場表現相關數據，不過這三個欄位可以協助之後實驗結果的分析，因此我們將此三個欄位另外儲存。最後，留下來預測薪水的features共有32個(38-6)，這32個features就是之後模型的input值。

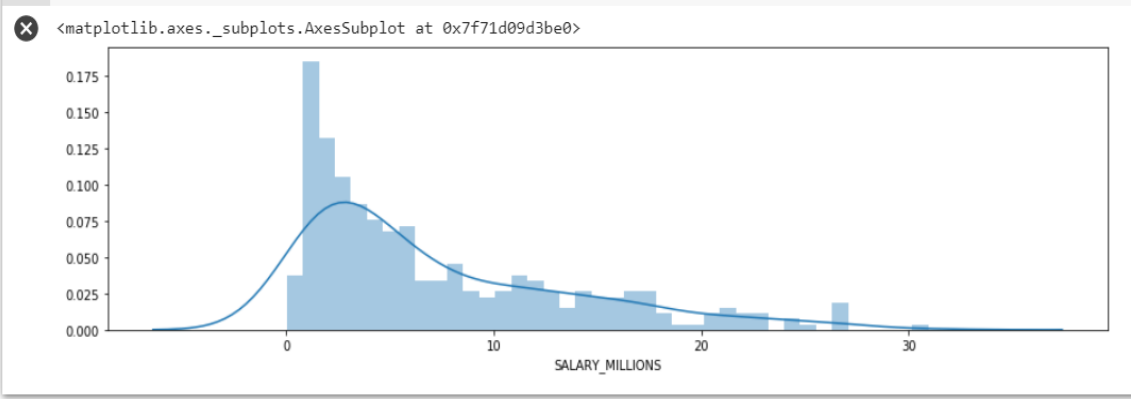
2.2 刪除重複值

原資料集共有342筆資料，但其中有幾筆資料完全相同，因此我們刪除掉重複球員資料。刪除後共有335筆unique value。

2.3 薪水label標記

在薪水的預測中若沒有將薪水分類，將會有過多的可能值不利於模型學習，因此必須找到薪水分類的方法集切割點，首先我們觀察薪水的分布(如下圖)，我們發現整體薪水分布呈現長尾狀，意即大部分的球員領的薪水較少，只有少部分的球員領高

薪。接著，我們看薪水的各項統計數字，最決定用第一四分位數與第三四分位數(2.29、11.235)做為分界，低於2.29的為第一類-低薪類，此類共有25%的球員;介於2.29與11.235之間的為第二類-中薪類，此類共有50%的球員;高於11.235的為第三類-高薪類，此類共有25%的球員。薪水label標記完後及為我們之後模型的output值。




```
In [15]: data['SALARY_MILLIONS'].describe()

Out[15]: count    335.000000
         mean       7.415313
         std        6.529251
         min        0.030000
         25%        2.290000
         50%        5.000000
         75%       11.235000
         max       30.960000
         Name: SALARY_MILLIONS, dtype: float64
```

2.4 資料標準化

標準化有非常多做法，我們使用sklearn.preprocessing.MinMaxScaler，此方法將所有數值投射到0-1的範圍，標準化後的資料集及為模型的訓練及測試資料(如下圖)。

	PLAYER	TEAM	POSITION	AGE	FG	FGA	FG%	3P	3PA	3P%	...	MPG	ORPM	DRPM	RPM
0	Russell Westbrook	OKC	PG	0.428571	0.990291	1.000000	0.566667	0.609756	0.72	0.343	...	0.910112	0.954701	0.347082	0.856858
1	James Harden	HOU	PG	0.380952	0.805825	0.780172	0.586667	0.780488	0.93	0.347	...	0.960674	0.923932	0.236419	0.759654
2	Isaiah Thomas	BOS	PG	0.380952	0.873786	0.801724	0.617333	0.780488	0.85	0.379	...	0.887640	0.867521	0.003018	0.561252
3	Anthony Davis	NO	C	0.190476	1.000000	0.840517	0.673333	0.121951	0.18	0.299	...	0.952247	0.417094	0.786720	0.729028
4	DeMarcus Cousins	NO/SAC	C	0.333333	0.873786	0.823276	0.602667	0.439024	0.50	0.361	...	0.898876	0.682906	0.458753	0.719041
5	Damian Lillard	POR	PG	0.333333	0.854369	0.818966	0.592000	0.707317	0.77	0.370	...	0.946629	0.774359	0.244467	0.648469
6	LeBron James	CLE	SF	0.619048	0.961165	0.750000	0.730667	0.414634	0.46	0.363	...	1.000000	0.933333	0.588531	1.000000
7	Kawhi Leonard	SA	SF	0.285714	0.834951	0.728448	0.646667	0.487805	0.52	0.380	...	0.876404	0.876923	0.520121	0.910786
8	Stephen Curry	GS	PG	0.428571	0.825243	0.754310	0.624000	1.000000	1.00	0.411	...	0.876404	1.000000	0.408451	0.932756
9	Kyrie Irving	CLE	PG	0.238095	0.902913	0.814655	0.630667	0.609756	0.61	0.401	...	0.924157	0.750427	0.162978	0.575899

3. 分類方法(classification)

我們採用三種分類分法來進行分類，分別是貝氏分類器、Support Vector Machine(SVM)、類神經網路(Neural Network，NN)，由於資料集較小，離群值較容易影響實驗結果，為了避免分類結果因為某些離群值而造成偏差，我們以100次實驗的平均值來當作最終實驗結果。

3.1 資料分割

由於資料量偏小，且分為三個類別，有別於使用一般的K-fold進行分類，我們採用StratifiedFold的方式進行分組，訓練與測試資料的比例為4 : 1，且每個類別都擁有固定的比例，以避免出現某個類別的資料太少或是太多的情況

3.2 對所有球員進行分類

a. 貝氏分類器

首先我們採用貝氏分類器來進行分類，貝氏分類器為一種建構簡易的分類方法，通常做為文本分類的研究基準，採用監督式學習的方式，不需要大量的訓練資料便能找出模型的相關參數。

我們將32項features當作訓練資料，並以薪水的label當作分類結果進行訓練，準確率達58.9%，以下為confusion matrix，縱軸為實際的薪水級距，橫軸為使用貝氏分類器預測出來的薪水級距。在低薪球員的部分，準確率達75.8%，有20.4%左右的人被預測為中薪，只有3.8%的人被預測為高薪；中薪球員的準確率只有49.4%，有32.7%的人被預測為低薪，17.9%的人被預測為高薪；高薪球員的準確率達62%，有32.8%的人被預測為中薪，還有5.2%的人被預測為低薪。

實際/預測結果	低薪	中薪	高薪
低薪	14.98975	4.015	0.74525
中薪	13.8575	21.006	7.6365
高薪	1.1145	7.04925	13.33625

b. SVM

SVM為一種常見的監督式學習分類器，首先通過非線性轉換將輸入空間轉到一個高維的空間，接著在這個高維空間中尋求最佳分類，可以解決非線性分類問題，且效果良好。

我們將32項features當作訓練資料，並以薪水的label當作分類結果進行訓練，準確率達69.68%，以下為confusion matrix，縱軸為實際的薪水級距，橫軸為使用貝氏分類器預測出來的薪水級距。在低薪球員的部分，準確率達55.4%，有42.2%的人被預測為中薪，還有2.4%的人被預測為高薪；中薪球員準確率高達86.8%，有5.75%的人被預測為低薪，7.45%的人被預測為高薪；高薪球員的準確率達49%，有51%的人被預測為中薪，沒有人被預測為低薪球員。

實際/預測結果	低薪	中薪	高薪
低薪	10.94975	8.32325	0.477
中薪	2.442	36.89175	3.16625
高薪	0	10.9845	10.5155

c. Neural Network

類神經網路以程式模擬大腦的學習過程，架構為輸入、隱藏與輸出三層，透過連線中權重的改變達到機器學習，建立分類機制。由於神經網路能夠擁有類似人一樣的具備簡單的判斷能力，這種方法比起邏輯學推理演算更具優勢，因此目前已被廣泛運用在各種領域之中。

我們將32種features當作訓練資料，並以薪水的label作為分類結果進行訓練，共有三層隱藏層，隱藏元分別為128、512、1024，activation為relu函數，dropout為0.1。測驗結果的準確率達74.2%，以下為confusion matrix，縱軸為實際的薪水級距，橫軸為使用貝氏分類器預測出來的薪水級距。在低薪球員的部分，準確率達63.9%，有33.3%的球原被預測為低薪，2.8%的球員被預測為高薪；中薪球員準確率80.8%，有6.5%的人被預測為低薪，12.7%的人被預測為高薪；高薪球員準確率70%，有28.5%的球員被預測為中薪，還有1.5%的人被預測為低薪。

實際/預測結果	低薪	中薪	高薪
低薪	12.62	6.58	0.55
中薪	2.73	34.36	5.41
高薪	0.33	6.11	15.06

d. 模型結論

從以上三種分類結果可以發現，在低薪球員預測的部分，貝氏分類器效果最好，Neural Network效果較差；在中薪球員預測的部分，SVM與Neural Network都有不錯的成效，而貝氏分類器表現極差；在高薪球員的部分，Neural Network效果最佳，SVM則表現極差。整體而言，Neural Network為表現最好的分類器，並且沒有在哪一類球員的分類上有極差的表現，因此我們最後決定以此作為後續實驗的分類器。

3.3 分別對各位置球員進行分類

前面的實驗我們都是將所有球員放進訓練與測試資料之中，對整體進行模型的訓練與預測，然而我們認為也許不同位置的球員，得分後衛(SG)、控球後衛(PG)、小前鋒(SF)、大前鋒(PF)、中鋒(C)，著重的features也許不同，假如我們全部放在一起實驗，彼此之前可能會互相干擾造成實驗誤差。因此我們決定分開探討各位置球員的薪水模型，觀察是否會影響模型的改變。

此實驗以32項features作為訓練資料，並以薪水的label當作分類結果進行訓練，使用Neural Network作為分類器，模型架構與前面一樣。從下表實驗結果可以發現，不論是在哪個位置，準確率與原有的74.2%比起來都有不小的提升。因此我們推測不同球員間可能各自著重於不同的features，我們將在下一個實驗來探討各位置球員與features的關聯性。

球員位置	SG	PG	SF	PF	C
準確率	82.1%	85.5%	81.6%	85.8%	78.1%

4. 特徵選取（Feature Selection）

4.1 對所有球員的分類進行特徵選取

a. 非神經網路

使用了Linear SVC的做法，並以sklearn的套件實作，經過實驗後發現在Naïve bayes 懲罰參數項設置成 0.05 會選出6個feature，這樣的參數設定會得到較佳的結果。而在SVM中，經過特徵選取，準確率反而下降，不管懲罰參數項設置成0.01、0.05、0.09或0.1都是下降的。所以我們認為在非神經網路中，因為資料量小，經過特徵選取反而降低了模型原有的能力。

(1) Naïve bayes沒經過特徵選取：

```
Train data:251.25 筆      test data:83.75 筆
accuracy : 0.5890
Confusion:
[[14.98975   4.015     0.74525]
 [13.8575   21.006     7.6365 ]
 [ 1.1145    7.04925 13.33625]]
```

Naïve bayes經過特徵選取：

```
trainDataLen: 251.25 testDataLen: 83.75
accuracy: 0.6392672555357595
confusion:
 [[13.25  5.5   1.  ]
 [ 8.25 27.5   6.75]
 [ 0.5   8.25 12.75]]
```

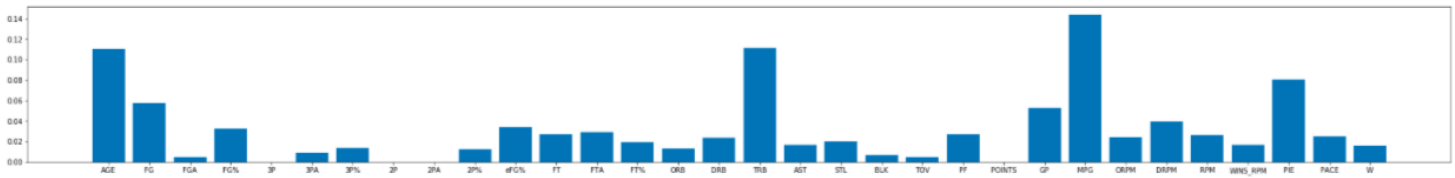
(2) SVM沒經過特徵選取：

```
Train data:251.25 筆      test data:83.75 筆
accuracy : 0.6968
Confusion：
[[10.94975   8.32325   0.477   ]
 [ 2.442    36.89175   3.16625]
 [ 0.       10.9845   10.5155 ]]
```

SVM經過特徵選取(懲罰參數 = 0.05)，和Naïve bayes：

```
trainDataLen: 251.25 testDataLen: 83.75
accuracy: 0.6864465294463177
confusion:
 [[1.013800e+01 9.298500e+00 3.135000e-01]
 [2.540750e+00 3.606050e+01 3.898750e+00]
 [7.500000e-04 1.020800e+01 1.129125e+01]]
```

b. 類神經網路



先使用DecisionTreeClassifier，呈現出特徵的重要性，並作視覺化：

之後，我們使用了兩個方法選取特徵：

（1）直接抽取前幾名的特徵，選取前6名最高的特徵進行預測。

(2) 使用ＰＣＡ降維，經過實驗後選取要輸出的主成份為 6 。

我們發現，使用ＰＣＡ的方法，模型的表現會比較好，我們推測是因為球員的薪水並不能用單純的幾個特徵就能夠決定，需要各種數據結合在一起才能夠表現出球員的價值，例如一些角色球員，只有在定點或罰球的得分率高，如果單看特定的數據表現，是無法決定薪資的。

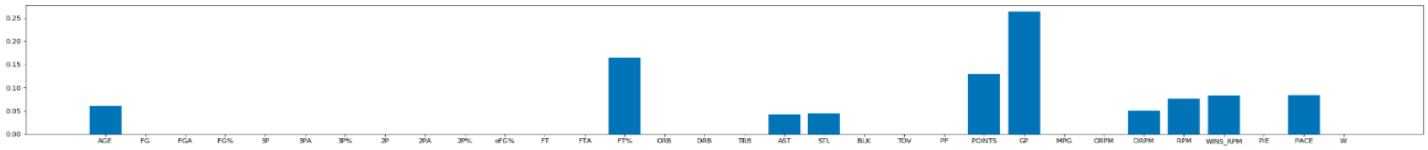
經過了特徵選取，我們發現ＮＮ加上ＰＣＡ的模型表現會最好，數據如下：

```
trainDataLen: 251.25 testDataLen: 83.75
Accuracy:82.5%
confusion:
 [[16.5   2.7   0.55]
 [ 2.55 35.5   4.45]
 [ 0.5   4.05 16.95]]
```

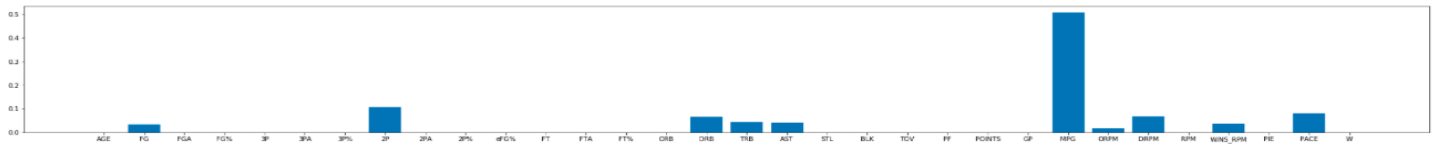
4.2 分別對各位置球員的分類進行特徵選取

經過了所有球員的實驗，我們分位置的實驗選擇使用ＮＮ加上ＰＣＡ模型。我們一開始的推測，是各個位置的球員薪水所注重的特徵是會不一樣的，經過了用DecisionTreeClassifier的視覺化後，發現確實如此，每個位置所注重的特徵是會不一樣的。

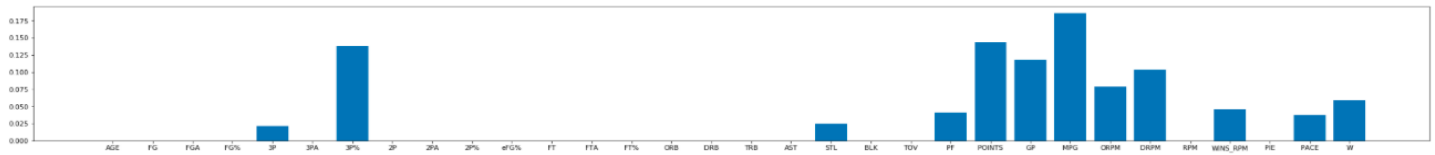
SG：



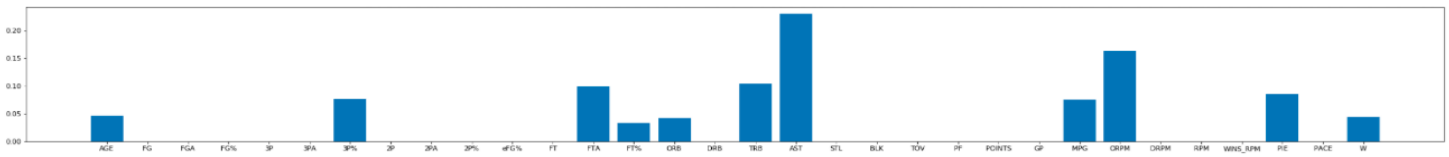
PG：



SF：

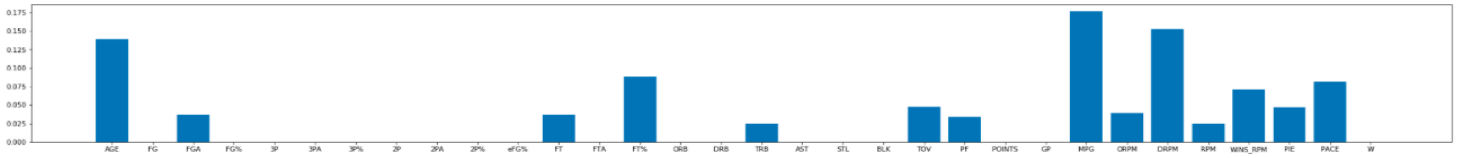


PF：



C：

4.3 納入社群影響力進行



延續前面實驗(3.2)的結論，我們可以建立在該段落中的方法，更納入社群媒體的資料來補強一些我們無法解釋的分類錯誤問題。

社群媒體資料包括兩個欄位，twitter_favorite_count(按讚數)以及twitter_retweet_count (轉發數)。使用此兩個數值的資料納入之前的model來進行球員的薪資評比的分類。

由於先前的實驗顯示出，有不同的分類器有不同的表現，其中又以採用neural network作為訓練模型的訓練有最高的accuracy。因此針對納入社群影響力分析的模型依然採用表現優異的的neural network的模型做訓練及預測，差別在於多納入的欄位(按讚數以及轉發數)。訓練結果如下，

trainDataLen: 178.5 testDataLen: 59.5

Accuracy:79.6%

confusion:

實際/預測結果	低薪	中薪	高薪
低薪	7.99	1.195	2.815
中薪	1.105	11.2175	3.4275
高薪	1.4875	2.1575	28.105

準確度從先前的82.5%下降到了到了準確度79.6的程度。由此可得知，社群媒體的影響力，對於球員薪水分類的任務可能是沒有那麼有影響力的。 對於這個現象，我們認為是由於我們所採納的數據並不具備那麼高度的代表性。也就是說按讚數跟轉發數，並不能代表一個球員是否為球星的指標。因此，社群媒體數據的那入在本實驗的表現並不理想。

五、 實驗結果與分析

根據跑出來的結果，我們的模型單從數據表現作薪水預測大致可以達到80%以上的準確率，為了瞭解剩餘不到20%的數據被分類錯誤的原因，我們對模型的預測結果進行統計與分析。由於單一次的測試可能會因為離群值影響使得模型產生偏差，進而使得預測出來的結果不準確，為了降低這方面的誤差，我們以表現最佳的Neural Network作為分類器，測試了100次並做統計，以下為實驗結果的部分截圖。PLAYER是球員姓名，predict為在100次中被預測成哪一類別的平均，real為實際資料中的類別，estimate為高估與低估的指標，0代表的是被球隊低估薪水的球員，1代表的是被球隊高估薪水的球員，times則是100次中被預測錯誤的次數。我們從統計數據中發現，其中這些不到20%的分類錯誤，有些可能是球隊單純高估或低估了這位球員的價值，但也有些情況可能是球員的價值無法只從數據面做預測的，例如當時市場的局勢、球隊薪資空間以及經營策略的考量、甚至給出的薪水也會受人情道義上的因素所影響。

Unnamed: 0	PLAYER	predict	real	estimate	times
0	Giannis Antetokounmpo	1.923077	1	0	78
1	George Hill	2.000000	1	0	100
2	Serge Ibaka	1.000000	2	1	82
3	Jamal Crawford	1.000000	2	1	53
4	Trevor Ariza	2.000000	1	0	100
5	Robin Lopez	1.000000	2	1	22
6	Trevor Booker	2.000000	1	0	93

1. 不太合理的薪水

1.1 Avery Bradely(低估)

Avery Bradely在進入聯盟初期還只是一個角色球員，在後衛的位置上也有著許多前輩包含Rajon Rondo、Ray Allen，但在一次大量傷兵潮的情況下，抓住了機會表現，不僅獲得教練的肯定，更在之後獲得穩定的上場時間，球場上的各項數據也越趨亮眼，尤其在防守端的表現相當突出。當時在2015年，塞爾提克隊總管和他簽了一份4年3200萬的合約，這份合約如果只依當時的數據來衡量的話，是會被評為高估的合約，因為Avery Bradley當時的價值僅在防守能力而已，但到了2017賽季時，整體數據上升至場均15.2分2.9籃板2.1助攻，同時在2017賽季更獲選為全聯盟年度防守一隊，而這樣的數據表現，球隊卻只需花費830萬，因此是筆相當划算的投資，但如果單以數據表現來判斷薪水的話，球隊確實低估了該球員在球隊的價值。

1.2 Chandler Parsons(高估)

於2011年踏入聯盟，在擔當休斯頓火箭隊的先發小前鋒的賽季中，Chandler Parsons的各項數據相比菜鳥時期都有著大幅進步，多次打破自己得分新高，刷新半場三分球命中的NBA紀錄，被聯盟視為小前鋒這個位置上的明日之星，後來轉隊至達拉斯

獨行俠後，雖因傷關係缺席了近40場比賽，僅留下14.7分、4.8籃板、2.6助攻的數據，仍得到了灰熊隊對其未來發展的期待，期待他能在傷癒復出後重拾以往的身手，仍給予他一份4年9480萬美元的頂級合約，但賽季開始後，大小傷痛不斷，到季中為止平均上場時間僅有20分鐘左右，數據也驟降至6.9分、2.9籃板，到後來甚至長期因傷缺陣，仍坐領高薪，被冠上「薪水小偷」的臭名，漸漸消失在球隊的主要輪替陣容中。不僅灰熊隊甩賣不了他，也沒有球隊願意吃下他的合約。這也反映了球隊有時候會為了網羅數據表現好的球員，多方球隊角逐的情況下便哄抬了該球員的價值，但往往這些球員場上的的價值明顯不符合這樣的身價，這樣的溢價合約也會在日後成為球隊薪資空間的一大痛點，也會造成球隊在自由市場補強戰力滿大的阻礙，這也是球隊高估其球員價值會有的情況和影響。

2. 可以合理解釋的薪水

這裡舉四個球員案例來分析：

2.1 Giannis Antetokounmpo(低估)

薪水被低估的可能原因: Giannis Antetokounmpo，目前25歲，算是相當年輕的球員，在年僅18、19歲時就在選秀會上被公鹿隊選中進入NBA聯盟，在短短的兩三年內不斷地精進自己的技術與身體素質，並在2016-2017賽季時，打出聯盟頂級的數據表現，同時也成為公鹿隊中的核心主將，因此在2016-2017的賽季後，球隊才決定與他續上一紙大合約，其合約內容為4年1億美元。但在2016-2017賽季時，Giannis Antetokounmpo領的還是初入聯盟的菜鳥時期合約的薪水，原先的這份薪水內容為期四年(由2013-14賽季效力至2016-17賽季)，Giannis Antetokounmpo每季有著25萬歐元(=277,225美元)的淨收入，該合約還包括了一個買斷條款，可以使其到NBA或歐洲

籃球聯賽的任一球隊比賽，也因此他在2013年球季後能夠參加NBA的選秀，所以如果單從該年數據表現來預測其薪水時，便會被判斷為低估。

2.2 Danny Green(低估)

薪水被低估的可能原因：聯盟數一數二的3D球員，雖然數據並無隊上球星亮眼，但場上防守貢獻度高，防守效率可以排在聯盟前段班，三分球的把握度也在聯盟前段水準，故從數據表現來看其薪水應該要屬於高薪類，但有鑑於2015-2016賽季恰好為隊伍上球星需要換約--Kawhi Leonard(5年9000萬)，身為隊伍中的球星的綠葉球員，自然優先被球團考慮的順位就會被往後延，因為這些市場上策略操作的因素，使得球隊基於薪水空間考量，選擇給予他屬於聯盟中等水準的薪水(4年4500萬)，這樣球隊也才能有足夠的錢去留住或續約下當時馬刺隊的未來巨星Kawhai Leonard，因此如果單從該年數據表現來預測Danny Green薪水時，就會被判斷為低估。

2.3 Manu Ginobili(高估)

薪水被高估的可能原因：Ginobili是眾人所皆知馬刺隊GDP王朝的核心成員，與隊友Tim Duncan, Tony Parker一同為馬刺奪下不少榮譽及奠定了馬刺三連霸的王朝。在2016年以前Ginobili最近一次合約的薪水為兩年570萬，這是一個屬於老將的合約，的確，在年紀增長，體能大不如前，無法像年輕時期球主宰球場的情況下，這份合約也確實合理，但2016-2017發生了一些事件--當初球隊卡在薪資空間限制: Tim Duncan尚未決定退役，而Boban Marjanovic還要續約的情況下，馬刺起初開給Ginobili的合約是一年300萬，但後來隨著Tim Duncan決定退役後，其之後的薪資可以用分三年支付，減少了薪資空間的占用。另外，Boban Marjanovic也被活塞隊以三年2100萬簽走，讓馬刺隊薪資空間得以獲得釋放，最後才為Manu Ginobili開出一年1400萬的合約，也算是一份感謝這位老將這多年對球隊的貢獻，為他獻上的最後的養老合

約，而也令大家不意外地，Manu Ginobili也於該年結束後選擇退役，結束其NBA籃球生涯，因此如果單從該年數據表現來預測其薪水時，會被判斷為高估。

2.4 Draymond Green(高估)

薪水被高估的可能原因：Draymond Green於2015年簽下5年8200萬美元的合約，球隊願意給出這麼高的薪水的理由可以從兩個理由來探討：

(1) Draymond Green表現很全面，在勇士隊中不僅扮演攻守轉換的樞紐同時也是陣中禁區防守的核心。

(2)同時他也是一名組織者，在觀賞勇士隊的比賽中，我們不難發現，很多時候他會代打場上組織者的角色，不管是利用無球的掩護，或是持球的組織分球，都能為隊友製造空檔及出手機會，也因此他被視為勇士隊兩連霸的核心和靈魂人物，Draymond Green連續3年入選全明星，並在2017年當選年度最佳防守球員。在這份合約的前三個賽季，據統計Draymond Green場均至少可以得到10分7籃板7助攻1火鍋1抄截，這樣的數據雖然全面但卻還不算是巨星的表現，仍在綠葉球員的數據表現範圍內，也由於他很多場上的貢獻是數據無法表現出來的，因此雖然Draymond Green的薪水是屬於高薪類，但如果單從數據表現去做分析的話，可能就無法準確預測初他在球隊的價值以及為何球隊要給它高薪的原因，因此才會被評為高估。

六、 總結

在美國職籃（NBA）中球員的薪資一直都是判斷一個球員價值最直接的指標，但是究竟是哪些數據影響了球員的薪資，我們想要了解哪些特徵對薪資是有影響的，以及目前球員的薪資有沒有符合該有的水準，所以進行了這次的實驗。

首先，經過實驗，我們發現球員的薪水是要將所有的數據結合起來看才有價值，對於各個位置（SG、PG、SF、PF、C）的球員，所注重的能力（特徵）是不一樣的，例如：得分後衛（SG），較注重的特徵是出場數(GP)、得分(Points)和罰球命中率(FP%)。而加入社群的影響後，模型的表現並沒有提升，代表球員的薪資最大部分還是由在賽場上的表現決定。

再來是目前球員的薪資有沒有符合該有的水準，我們將實驗中模型犯錯的資料抓出幾個進行分析，發現確實有數據以外的因素會影響到薪資，例如：為了簽約球星的薪資控管、老將球星的表現雖然不亮眼，但是對球隊的向心力以及球迷的支持度是有影響的。

經過實驗和分析，我們可以發現，球員的薪水大部分是由賽場上的表現來決定，但是因為NBA是一個廣大且複雜的市場，球隊的操作、球星的狀態等也是考量的因素，單看數據是無法100%決定球員的薪水的。目前我們的模型只單看各項數據預測薪水，在未來，可以加上球員在社群上的發言、加上球迷轉推的推特文章進行文字分析；除此之外，因為我們這次的資料集只有一個賽季的，可能導致模型不是非常準確和穩固，所以在未來我們可以收集更多賽季的數據資料，增加模型的穩固性。

最後，在未來這套預測薪水的系統發展成熟後，目標是能夠運用到各個市場上，例如：將員工的各項數據丟到模型中進行預測，不僅能夠讓員工能夠得到該有的薪資，也能夠增加薪水給付的公平性；在面試員工時就能夠依據求職者的各項能力預測求職者的價值，幫助公司的判斷。我們相信這是能夠對求職和公司內部的運作可以提升的系統。我們所做的是最基礎的模型，但是因為使用了公開資源很多的sklearn和keras，對於未來的可擴充性非常大，我們相信這次的實驗對於「價值分析」這個議題是有幫助的。

七、 參考文獻

[1] Stiroh, K. J. 2007. “PLAYING FOR KEEPS: PAY AND PERFORMANCE IN THE NBA” Economic Inquiry 45 (1): 145–161.

[2] Josh Rosson. “NBA Salary Predictions using Data Science and Linear Regression” https://towardsdatascience.com/nba-salary-predictions-4cd09931eb55

[3] Guyon I, Elisseeff A. An introduction to variable and feature selection, J. Mach Learn Res., 2003, vol. 3 (pg. 1157-1182)

[4] https://www.itread01.com/content/1543176315.html python資料預處理 ：資料降維

[5] https://scikit-learn.org/stable/modules/feature_selection.html scikit-learn 0.22

[6]https://www.sportsv.net/articles/13382 Kawhi Leonard新約五年117M

[7]https://basketball.fanpiece.com/SASpursBall 簽下佛心4年4500萬合約，Danny Green：只因馬刺是個Family

[8]https://kknews.cc/sports/958j4g5.html 字母哥，從貧民窟到百萬富翁，鋼鐵是怎樣煉成的

[9]https://zh.wikipedia.org/wiki/%E6%8F%9A%E5%B0%BC%E6%96%AF%C2%B7%E5%AE%89%E6%88%B4%E6%89%98%E6%98%86%E6%B3%A2#cite_note-17 維基百科--揚尼斯·安戴托昆波

[10]https://kknews.cc/zh-tw/sports/qj4r2o.html 吉諾比利1400萬？巔峰期也就這樣，馬刺瘋了還是被逼急了

[11]https://sports.ltn.com.tw/news/breakingnews/2533373 NBA》吉諾比利宣告退休 馬刺依鄧肯模式支付剩餘薪水

[12]https://basketball.biji.co 籃球筆記【專欄】一覽佛系合約 NBA十大超值球員