機器學習應用於中華職棒球員薪資評估

學生 ： 林元慈

目錄

一、 背景介紹

中華職業棒球大聯盟（簡稱中華職棒、中職、CPBL）是中華民國成立最早、以及目前唯一的職業棒球聯盟，前身是1989年創立的「中華職業棒球聯盟」，創始球隊為統一獅、味全龍、三商虎、兄弟象，2003年與台灣職棒大聯盟合併後改為現名。目前有統一獅、中信兄弟、Lamigo桃猿、富邦悍將4支球隊，目前競賽場地則主要分布於臺南市南區、臺中市北屯區、桃園市中壢區、新北市新莊區等地區，原本較偏向都市巡迴，但近年來開始改走屬地主義。

自2009年起（職棒二十年），因應「中信鯨」、「米迪亞暴龍」解散影響，球隊從6隊減至4隊，導致當時轉播的「緯來電視網」有意調整轉播權利金。中華職棒各球團為了維持現行權利金，因此決議將「每隊一季100場」提高至「每隊一季120場」，全年一隊總共240場的例行賽。而2007年（職棒十八年）本要取消和局制，改成每一戰都必須打到分出勝負為止，與美國大聯盟看齊。但後來考慮到球員數不足的問題，改為「最多12局但不限時間」，避免球隊在時間限制下故意拖延比賽。

最近中華職棒球員工會公布近年各隊本土球員薪資資料，今年全聯盟平均月薪為臺幣15.6萬，又以中信兄弟平均17.7萬最高，第二到四名分別為富邦悍將15.6萬、統一獅15.1萬元以及Lamigo桃猿的13.9萬，聯盟平均薪資比去年成長2.68%。

各隊團隊總薪資的部分，中信兄弟同樣以1億2285萬居冠，接下來分別為富邦悍將的1億830萬、統一獅9589.2萬和Lamigo桃猿的9186萬。其中，Lamigo薪資漲幅10.54%為四隊最高，獅隊則是唯一薪資負成長的球隊，漲幅-6.91%。據了解中職球員工會自2013年開始進行全面的薪資普查，薪資金額是由球員自行填寫的會費資料反推而來，並不是直接得自於球團，可以看出許多資料要完整獲得並不容易。

過去薪資一直是中職較不透明的一塊，直至近年還是有球隊”選擇性”的公布本土球員薪資，態度遮遮掩掩、害怕被比較，引來不少批評。現在薪資一定程度攤在陽光下後，也能迫使球團用更負責任的態度面對外界檢視。

過去球團與球員簽訂薪資時，主要是”主觀的”基於過去的成績，例如打者的打擊率、打點等等數據，當然由於受到人為決定的影響，或許一個球員所獲得的薪資會有過高過低的情況發生。為了維護各球員所能獲得薪資的公平性以及讓球團在與球員簽訂薪資時，能夠有更有效率的系統保證其公平性，因此

本文採用了機器學習的手法，希望能透過歷史球員數據訓練出一套合適的模型，並基於球員上季數據及此模型，推估出其在下一季能獲得的薪資水準，也給予球團與球員在了解其身價上有所幫助。

二、 應用手法

本文透過機器學習中的回歸模型，再配合中華職棒 2010-2017 年的球員數據來去訓練此模型，最後再用 2018 年個球員的表現投入以訓練完成的模型之中，來預測其今年的表現能使其在下一季獲得多少薪資水準。最後透過 Android APP 來展示各球員下一句獲得的薪資水準。而其中運用到的手法分別說明如下。

1. 球員數據（Python + BeautifulSoup）

再運用機器學習之前，其實花費最多時間的並不是建立模型，而是獲取資料並且加以整理成可以被使用的過程。本次研究使用 Python 為環境語言，配合 BeautifulSoup 建立一個可以自動從網頁抓取所需數據的程式。

而抓取的數據為中華職棒各打者的打擊率、上壘率、長打率等等，各投手的防禦率、自責率、好壞比等等。抓取的數據為中職 2010-2018 年份，其中 2018 年的數據為用來推估下一季球員薪資，而 2010-2017 則為用來配合訓練模型。

2. 機器學習（Python + Scikit-learn）

透過 BeautifulSoup 抓取到的球員數據，加以整理過後就可以用來投入訓練模型之中。機器學習的部分依然使用 Python 做為環境語言，訓練模型的建立則是透過 Scikit-learn。
本次訓練模型採用機器學習中的回歸模型，訓練樣本如上一小節所述，由於投手與打者的數據並不相同，因此模型會分為投手以及打者兩個模型。我們所希望預測出的即為各球員的薪資水準，而最後的結果則會再下一章中展示。

3. 資料庫（MySQL）

透過模型所獲得的各球員薪資水準則會被儲存於雲端資料庫 MySQL 之中，並透過下一小節的 APP 來展示各球員薪資水準。

4. 手機 APP（Android Studio）

本次開發 APP 採用 Android Studio，並配合 php 來從 MySQL 之中抓取要展示的數據，並將各球員依照其所屬球隊分類，讓使用者可以更輕易的找到自己所希望查詢的球員。

三、 成果

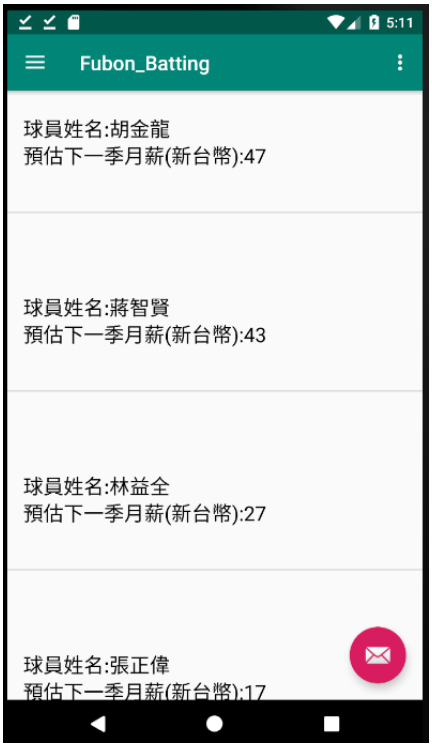
透過 Python + BeautifulSoup 所抓取的球員數據如下，

	NAME	G	GS	GR	IP	ERA	WHIP	FIP	ERA+	BB%	K%	BABIP	LOB%	DER	DIP	DIPS ERA	DIPS WHIP	WAR
0	伍鐸	28	27	1	169.0	3.25	1.21	3.68	131.6	3.59	18.10	0.323	66.2	56.83	173.2	2.95	1.10	4.04
1	羅力	26	26	0	161.0	3.47	1.20	3.58	127.0	2.41	23.64	0.351	70.3	52.24	169.6	2.77	0.99	4.09
2	艾迪頓	27	22	5	160.1	3.48	1.33	4.00	126.8	7.41	20.35	0.326	71.2	59.10	166.1	3.20	1.21	3.06
3	克恩三世	25	25	0	156.1	3.28	1.41	4.43	131.0	7.50	16.49	0.335	74.3	64.68	158.2	3.68	1.30	2.59
4	羅里奇	26	26	0	156.0	3.17	1.21	3.77	133.3	4.10	20.94	0.318	67.9	54.59	164.5	2.99	1.09	3.64
5	萊福力	28	20	8	151.0	4.05	1.44	4.30	114.8	6.31	18.31	0.347	67.4	61.05	157.0	3.50	1.22	2.27
6	瑞安	28	26	2	148.1	4.31	1.39	4.07	109.3	5.20	21.71	0.350	63.7	57.80	161.0	3.23	1.13	2.89
7	尼克斯	26	25	1	142.2	3.72	1.46	4.63	121.7	7.72	16.40	0.329	72.9	62.76	146.7	3.85	1.32	2.01
8	王溢正	22	22	0	139.2	4.25	1.35	3.91	110.6	5.93	16.97	0.320	60.5	52.25	147.2	3.19	1.20	3.05
9	施子謙	19	19	0	107.1	3.86	1.60	4.64	118.8	6.83	10.97	0.354	74.7	49.18	112.6	3.93	1.36	1.54
10	布魯斯	20	20	0	102.1	5.72	1.62	4.91	79.7	8.58	19.31	0.368	62.7	48.69	108.4	4.04	1.31	1.20
11	江辰晏	19	17	2	96.1	3.46	1.30	4.67	127.2	8.52	19.71	0.298	73.7	41.16	96.3	3.85	1.29	1.26
12	陳琥	27	15	12	94.2	4.66	1.50	4.95	102.0	9.98	19.27	0.313	64.4	44.28	101.3	3.93	1.36	0.66
13	史博威	13	13	0	77.1	5.12	1.58	5.35	92.3	9.30	15.70	0.376	65.6	39.31	78.5	4.51	1.43	0.59
14	賽格威	14	10	4	77.0	7.13	1.69	6.59	50.0	8.29	15.19	0.341	61.0	48.57	80.9	5.40	1.47	-0.47
15	黃亦志	22	11	11	76.1	5.66	1.73	5.05	80.9	7.37	10.20	0.365	64.3	38.57	81.2	4.28	1.42	0.41
16	馬丁尼茲	17	16	1	75.2	7.73	1.76	5.59	37.4	7.80	14.76	0.361	55.6	42.26	82.6	4.61	1.40	0.36
17	林樺慶	14	13	1	70.2	5.60	1.78	7.25	82.2	9.39	10.91	0.322	70.7	49.75	72.7	6.16	1.64	-0.68
18	鄭凱文	56	0	56	64.2	4.18	1.28	3.96	112.1	4.23	19.37	0.322	60.8	24.37	70.0	3.13	1.12	0.71
19	楊志龍	23	14	9	64.0	4.50	1.17	4.53	105.3	4.44	25.56	0.316	67.7	25.92	66.1	3.53	1.07	0.74

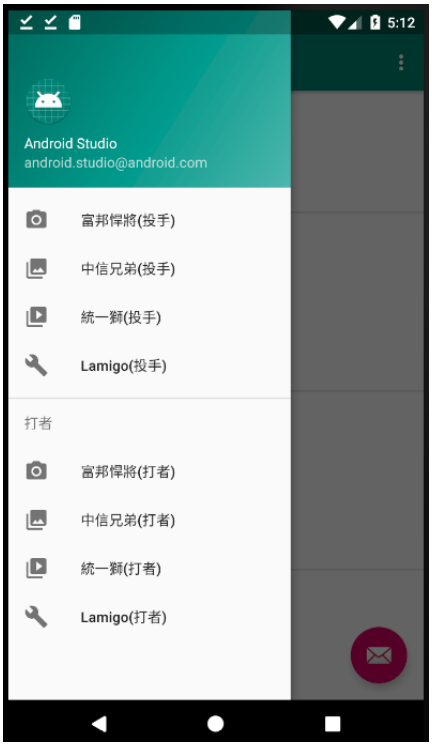
由於資料眾多，因此並不逐項展示，而將 2010-2017 年中職各球員數據投入模型之中訓練模型後，再使用 2018 年球員數據來去預測其再 2019 年所能獲得的薪資水準，而所能獲得的薪資水準則在下圖中最右欄，

	NAME	G	GS	GR	IP	ERA	WHIP	FIP	ERA+	BB%	...	BABIP	LOB%	DER	DIP	DIPS ERA	DIPS WHIP	WAR	Score	Rank	Salary
0	伍鐸	28	27	1	169.0	3.25	1.21	3.68	131.6	3.59	...	0.323	66.2	56.83	173.2	2.95	1.10	4.04	4.771412	99	71.824654
1	龍力	26	26	0	161.0	3.47	1.20	3.58	127.0	2.41	...	0.351	70.3	52.24	169.6	2.77	0.99	4.09	5.038575	100	73.792975
2	艾迪頓	27	22	5	160.1	3.48	1.33	4.00	126.8	7.41	...	0.326	71.2	59.10	166.1	3.20	1.21	3.06	4.082329	96	63.078173
3	克恩三世	25	25	0	156.1	3.28	1.41	4.43	131.0	7.50	...	0.335	74.3	64.68	158.2	3.68	1.30	2.59	3.303817	93	52.091598
4	羅里奇	26	26	0	156.0	3.17	1.21	3.77	133.3	4.10	...	0.318	67.9	54.59	164.5	2.99	1.09	3.64	4.584439	98	69.382115
5	黃福力	28	20	8	151.0	4.05	1.44	4.30	114.8	6.31	...	0.347	67.4	61.05	157.0	3.50	1.22	2.27	3.387005	94	62.176385
6	瑞安	28	26	2	148.1	4.31	1.39	4.07	109.3	5.20	...	0.350	63.7	57.80	161.0	3.23	1.13	2.89	4.151891	97	65.580806
7	尼克斯	26	25	1	142.2	3.72	1.46	4.63	121.7	7.72	...	0.329	72.9	62.76	146.7	3.85	1.32	2.01	2.946282	92	45.532564
8	王溢正	22	22	0	139.2	4.25	1.35	3.91	110.6	5.93	...	0.320	60.5	52.25	147.2	3.19	1.20	3.05	3.830781	95	62.605927
9	施子謙	19	19	0	107.1	3.86	1.60	4.64	118.8	6.83	...	0.354	74.7	49.18	112.6	3.93	1.36	1.54	2.355623	91	43.295423
10	布魯斯	20	20	0	102.1	5.72	1.62	4.91	79.7	8.58	...	0.368	62.7	48.69	108.4	4.04	1.31	1.20	2.304287	90	41.355705
11	江辰晏	19	17	2	96.1	3.46	1.30	4.67	127.2	8.52	...	0.298	73.7	41.16	96.3	3.85	1.29	1.26	2.283189	89	40.998414
12	陳琥	27	15	12	94.2	4.66	1.50	4.95	102.0	9.98	...	0.313	64.4	44.28	101.3	3.93	1.36	0.66	2.175009	88	38.955619
13	史博威	13	13	0	77.1	5.12	1.58	5.35	92.3	9.30	...	0.376	65.6	39.31	78.5	4.51	1.43	0.59	1.506764	74	27.789883
14	賽格威	14	10	4	77.0	7.13	1.69	6.59	50.0	8.29	...	0.341	61.0	48.57	80.9	5.40	1.47	-0.47	0.686754	10	7.000000
15	黃亦志	22	11	11	76.1	5.66	1.73	5.05	80.9	7.37	...	0.365	64.3	38.57	81.2	4.28	1.42	0.41	1.519857	76	28.250195
16	馬丁尼茲	17	16	1	75.2	7.73	1.76	5.59	37.4	7.80	...	0.361	55.6	42.26	82.6	4.61	1.40	0.36	1.525383	77	28.819455
17	林樺慶	14	13	1	70.2	5.60	1.78	7.25	82.2	9.39	...	0.322	70.7	49.75	72.7	6.16	1.64	-0.68	0.295241	3	7.000000
18	鄭凱文	56	0	56	64.2	4.18	1.28	3.96	112.1	4.23	...	0.322	60.8	24.37	70.0	3.13	1.12	0.71	1.700949	81	33.967743
19	楊志龍	23	14	9	64.0	4.50	1.17	4.53	105.3	4.44	...	0.316	67.7	25.92	66.1	3.53	1.07	0.74	2.154115	87	38.195601

為了讓使用者能更方便的獲得下一季的薪資水準，因此採用 APP 的形式來展示最終成果，如下圖，



而球員則會依照其所屬球團以及打者、投手來做分類，方便使用者查詢。



四、　結論

目前中華職棒大數據應用相較於國外如美國還相當不成熟，美國職棒除了各個球團應用大數據於薪資水準、評斷球員好壞、預測球賽走向等等，棒球數據的應用許多學校尚且有專門開設的課程，甚至碩博士專門研究此方面，棒球大數據應用已經開始深入於教育之中，因此中華職棒若要能好好的應用棒球數據，相較於別的國家依然有很長的一段路要走。

本此成果為參考球員當年度數據，判定其表現來給予下一季合理的薪資水準，消除了球員的主觀因素，以客觀的角度(球員表現數據)來判斷其薪資，但球員的價值有時候可能是在球場之外，如名聲或能帶給球團的效益，因此這些比較主觀的部分目前並無法被考慮近結果之中，不過大多數球員也確實預測出了不錯的水準，許多眾所皆知的強將都預測到了不錯的薪資水準。

五、 後續發展

目前投手的回歸模型有著不錯的相關係數，因此目前自變項的部分可能不需要做太多的調整，但打者的模型相關係數並非非常理想，後續可以進一步篩選所需的自變項，並且可以加入守備方面的數據，畢竟守備是棒球中非常重要的一環，打者成績並不可以只參照打擊數據。

再 APP 方面，除了介面可以更加美化以外，球員部分可以進一步加入完整的數據，並非只有姓名以及預測的薪資水準，理想之下也可以加入球員照片辨識系統，球迷在看到一切不認識的新秀時，可以直接拍照讓系統辨識球員相關資訊等等。