依據擊球仰角和擊球初速預測擊球員的安打類型

學生:魏嗣宸

指導老師:呂明頴

一. 介紹

棒球運動主要風行於中北美及東亞地區，當中包括美國、日本、韓國、台灣、多明尼加、委內瑞拉、古巴、波多黎各等。隨著科技的進步，現今運動員的訓練方式也和過往不同，從一名棒球員身上可取得到的資訊也越來越多，舉凡棒球員的擊球初速和仰角，投手的投球的轉速等等．而這些資訊除了球隊到球員皆可廣泛運用外，身在場外的我們這些球迷也可以透過這些資料來做分析和預測．本篇研究將以美國職棒大聯盟 statcast 的資料來預測球員的打擊表現．

Statcast 是由 美國職業棒大聯盟(以下簡稱 MLB)透過比賽數據去制定出的一系列指標，作為主播與球評在報導，判定分析球員運動能力的數據。這樣的數據也讓比賽播報上多了更多的趣味性，也讓球迷不再是透過猜測，而是透過科學化數據來討論球員的表現。MLB 於 2015 年在全美 30 座球場啟用 Statcast 系統後，球賽開始變得不一樣。近幾年提起棒球比賽與球員表現時，大家討論的不只是球速，也開始提到轉速、共軌效益、Poptime、擊球仰角等等新一代的科學數據。而在棒球訓練與傷後恢復也加

入了更多運動科學數據的協助，因此如何有效評估一個選手的能力已經變得和以前大不相同。

二. 文獻探討

現今 statcast 系統的出現改變了大聯盟球探觀察重點、數據的應用，連比賽本身都改變很多。而棒球界球風的改變是由上到下的，這意味著大聯盟怎麼打，小聯盟就會怎麼效法，並連帶影響到業餘棒球或更基層的棒球。現已有許多使用 python 來做分析的文章或文獻，大多題目皆各式各樣，而最常使用的預測模型有兩個:隨機森林和 Linear Regresion

三. 研究方法

本研究將使用隨機森林來預測球員的安打

a. 資料來源

使用 MLB 的 statcast 系統的資料，並使用 Python 的套件 Pybaseball 可直接取得該資料

```
In [1]: from pybaseball import statcast
```

```
In [2]: hitter_stats = statcast(start_dt='2016-04-01', end_dt='2019-10-04')
```

	pitch_type	game_date	release_speed	release_pos_x	release_pos_z	player_name	batter	pitcher	events	des
582	KC	2019-10-04	81.7	-2.51	6.19	Melancon, Mark	543939	453343	strikeout	swinging_strike
604	KC	2019-10-04	82.0	-2.67	6.27	Melancon, Mark	543939	453343	NaN	swinging_strike
613	KC	2019-10-04	80.8	-2.49	6.24	Melancon, Mark	543939	453343	NaN	call
634	KC	2019-10-04	81.2	-2.54	6.25	Melancon, Mark	543939	453343	NaN	
650	KC	2019-10-04	81.1	-2.67	6.23	Melancon, Mark	543939	453343	NaN	blo

b. 欄位選取

欄位名稱	說明
Pitch_type	球種簡寫
Pitch_name	球種名稱
events	每一顆球的結果
description	球場上狀況描述
bb_type	打者將球打出去後的形式
Release_speed	投手球速
Launch_speed	打者擊球初速
Launch_angle	打者擊球仰角
Hit_distance_sc	打者擊球距離

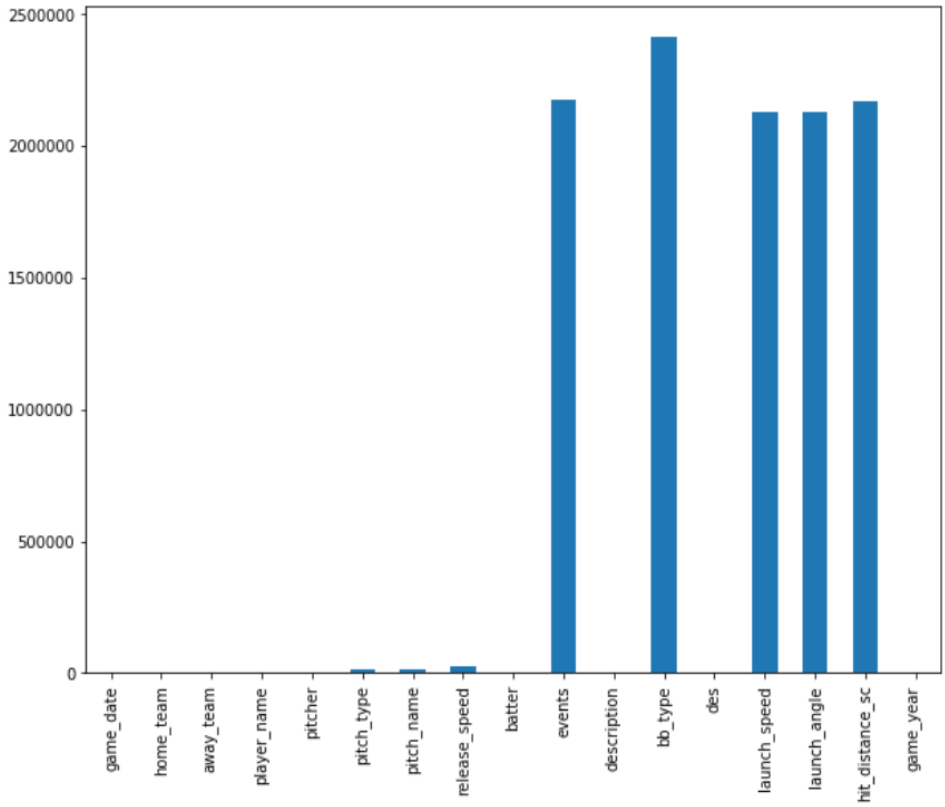
上表為幾個重要欄位，其他欄位則有 batter(打者名字),pitcher(投手名字),game_date(比賽日期)...等等

```
baseball_cols = ['game_date', 'home_team', 'away_team', 'player_name', 'pitcher', 'pitch_type', 'pitch_name', 'release_speed',
                 'batter', 'events', 'description', 'bb_type', 'des', 'launch_speed', 'launch_angle', 'hit_distance_sc','game_year']

hitter = hitter_stats[baseball_cols]
```

(上圖為所有欄位)

c. 資料處理



(圖一)

上圖(一)為資料個欄位的空值狀況，其中 bb_type 有 2410574 個空值

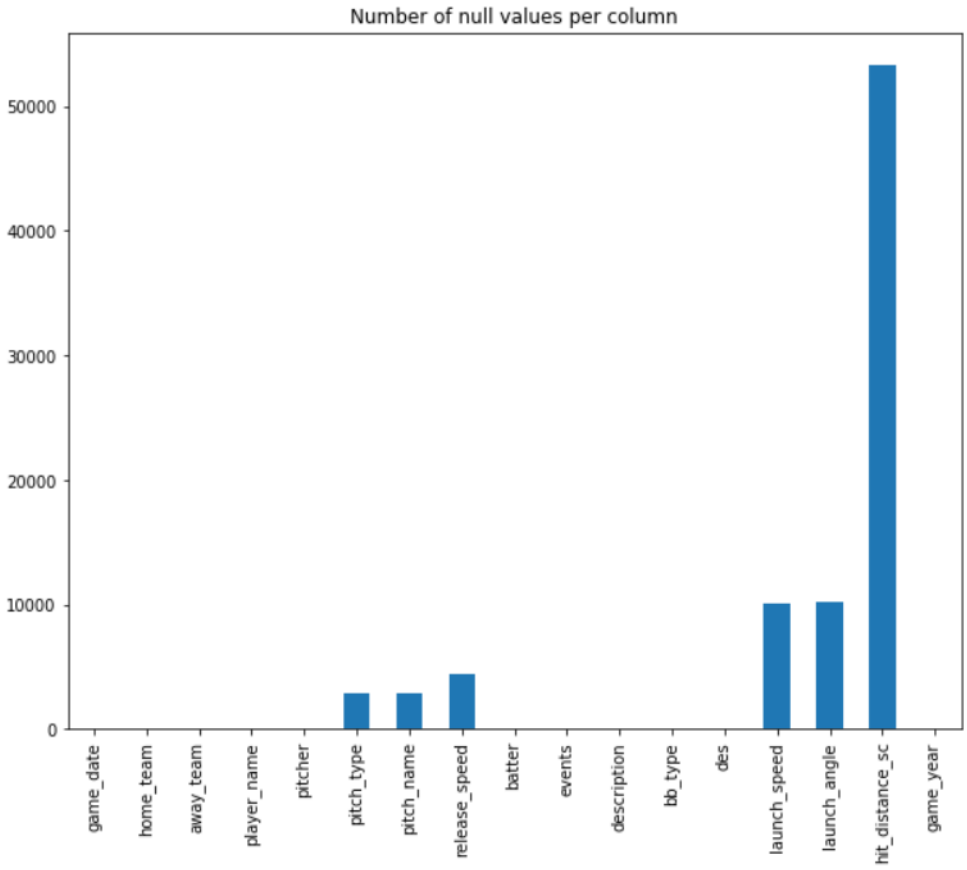
檢驗 bb_type 空值得原因(圖(二))，發現其內容大多為壞球或者是好球，也就代表說 bb_type 會沒有資料的原因是因為其大多為打者這一球並無打擊結果，可能是因為揮空或打成界外又或者是單純看球近來沒出手...等等

```
hitter[null_batted_ball_type].description.value_counts(dropna=False)
```

```
ball                        983382
foul                        511921
called_strike               489350
swinging_strike             288275
blocked_ball                 68740
foul_tip                     24678
swinging_strike_blocked      23297
foul_bunt                     7839
hit_by_pitch                  7408
intent_ball                   3325
missed_bunt                   1711
pitchout                       448
bunt_foul_tip                  153
hit_into_play                   42
swinging_pitchout                4
foul_pitchout                    1
Name: description, dtype: int64
```

(圖二)

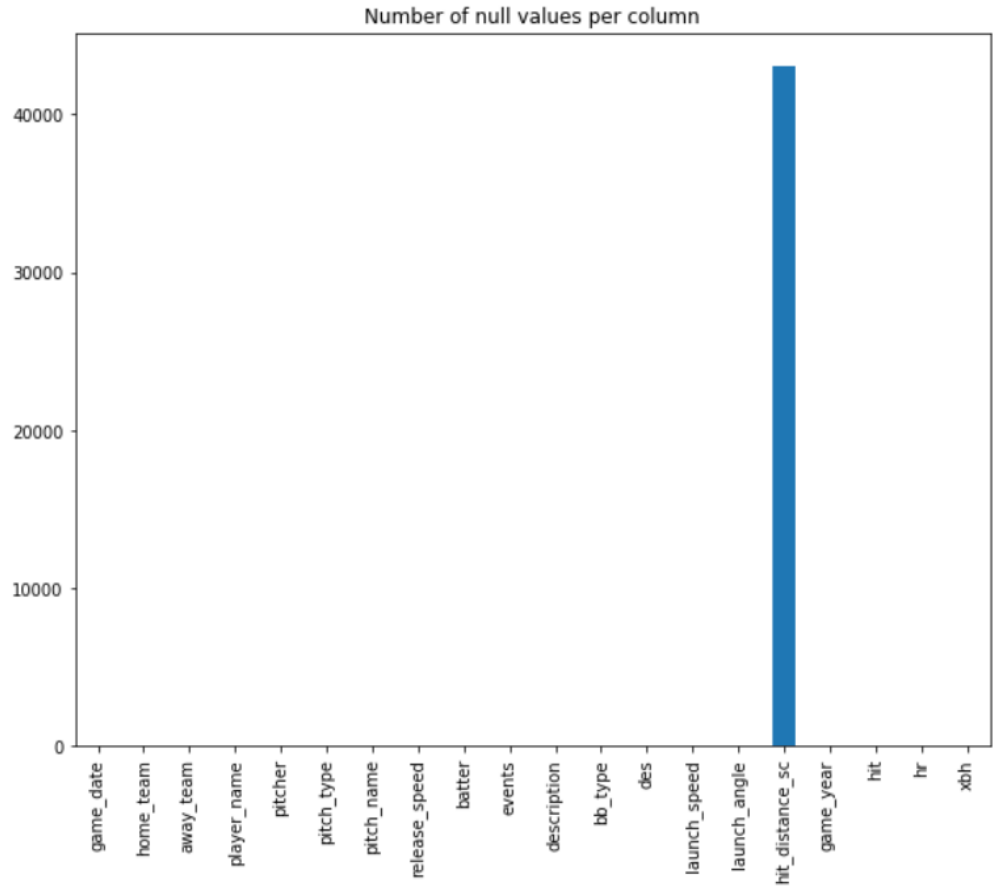
刪除完 bb_type 的空值後，仍有些許空值(圖三)



(圖三)

剩餘的這些是直接做刪除，因為這些空值刪除後並不會影響到之後的模型預測．

最後在清理完所有空值後的樣貌如圖四



(圖四)

當資料清理完後，發現 hit_distanse_sc 仍有許多空值，但這些資料仍需要被保留，根據 statcast 對擊球距離資料的定義，距離在 5 英尺以下在資料裡都會被設為 0，因此剩餘的這些空值實際上是有擊球結果的，但是這些球擊球距離都小於 5 英尺．

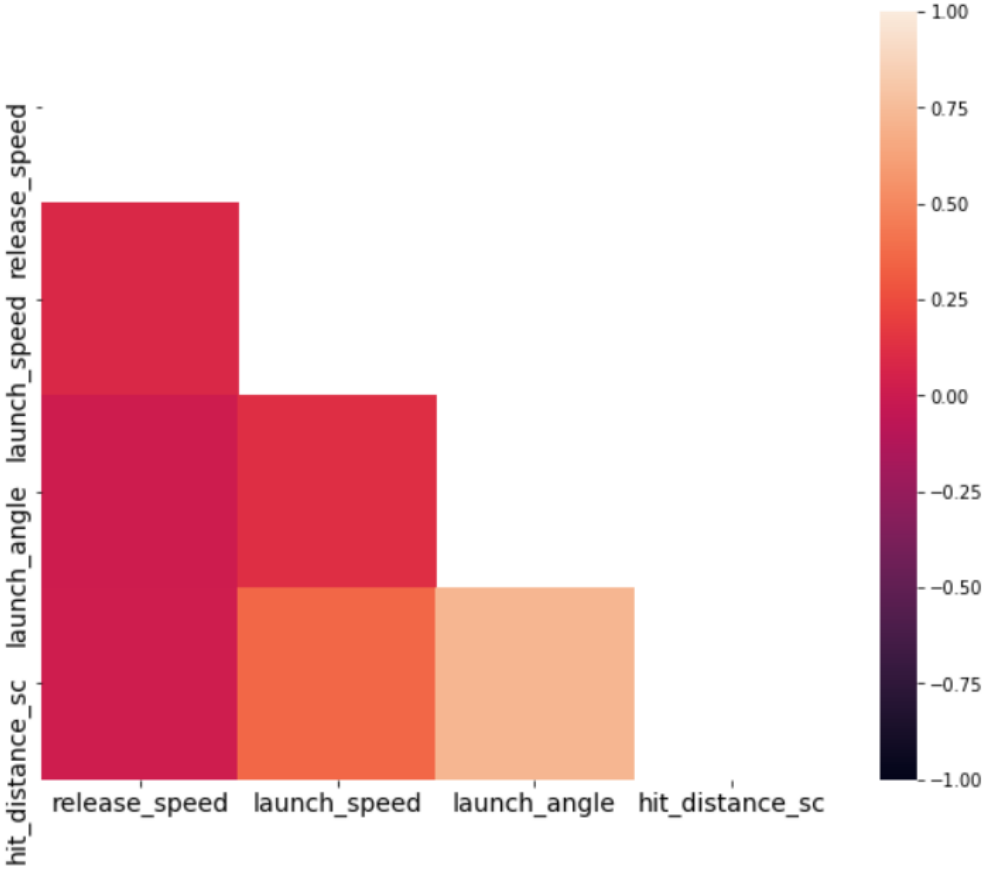
最後，為了之後預測的關係，需先將依些資料轉為數值型

1. 球種(每一個球種都為他編號 1~9)

2. 安打(一壘安打,二壘安打,三壘安打,全壘打)

d. 視覺化

(1)相關係數圖

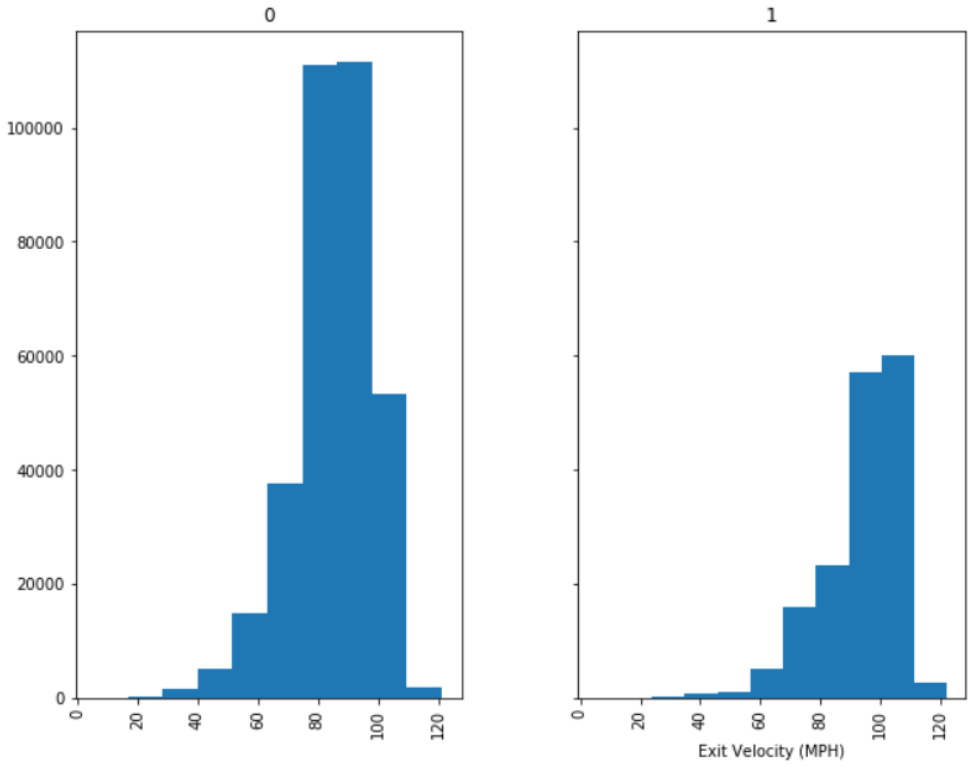


圖五

上圖(五)中，下方欄位由左至右分別為[投手出手後球速],[擊球速度],[擊球仰角]以及[擊球距離]，這 4 項將會是這整個專題中重要的 4 個欄位，因此透過先關係

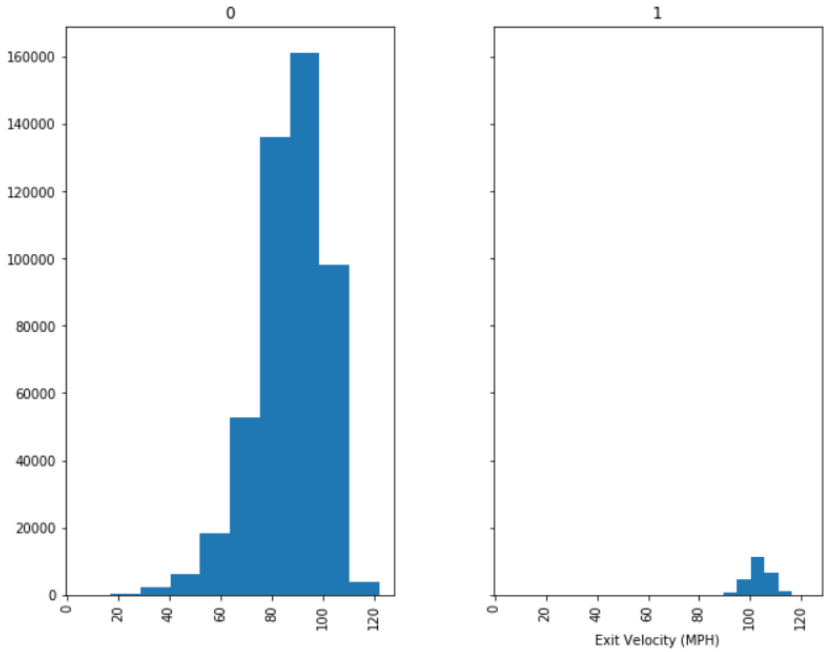
樹的熱圖，可以看出要有好的擊球距離勢必要有強力的擊球初速，但是同時也必須要有適當的擊球仰角來配合讓球飛行的更遠．

(2)擊球初速



圖(六)

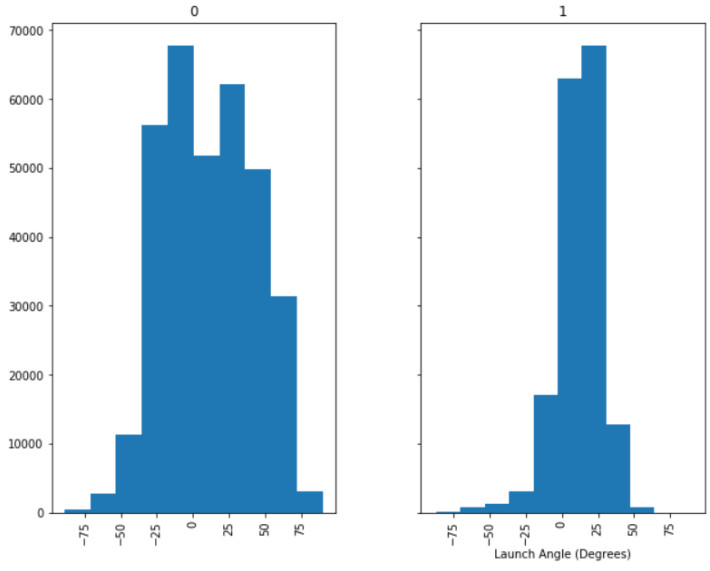
圖(六)中，左邊縱軸皆為數量，橫軸為擊球初速，上頭的 1/0 為安打/非安打，可以看出擊球初速即便在快也不一定會形成安打．



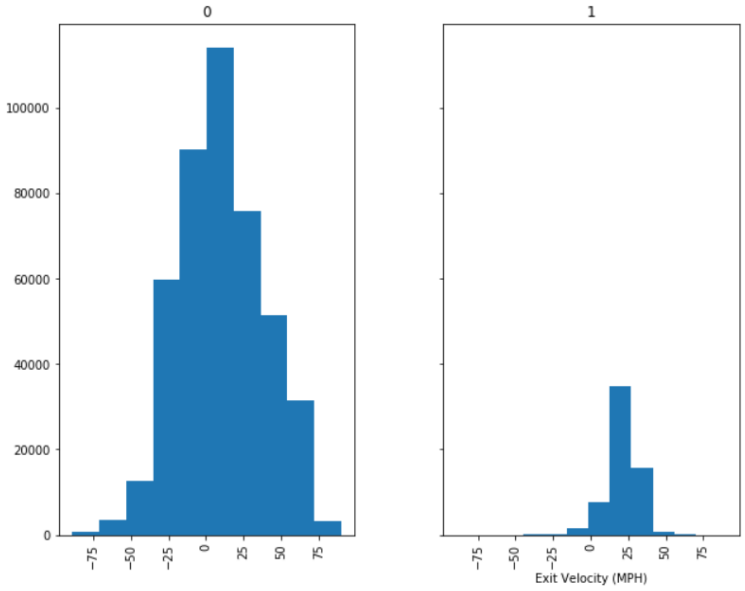
圖(七)

圖(七)中，左邊縱軸皆為數量，橫軸為擊球初速，上頭的 1/0 為全壘打/非全壘打，和圖(六)互相比對可以看出擊球初速慢也有可能幸運成為安打，然而如果今天想打全壘打，即球初速就必在約 100mph 左右

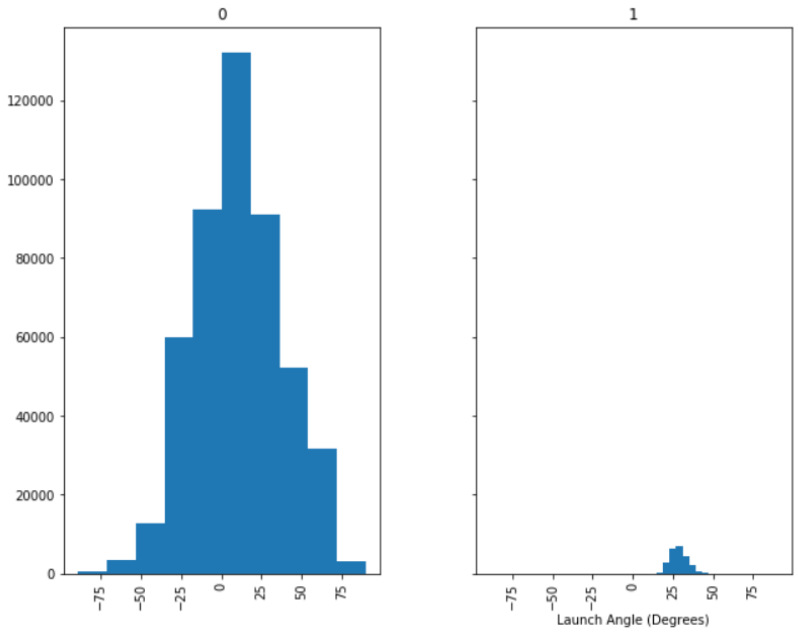
(3)擊球仰角



圖(八)



圖(九)



圖(十)

圖(八)~圖(十)縱軸為數量，橫軸為擊球仰角

圖(八):安打/擊球仰角 左:非安打 右:安打

圖(九):長打/擊球仰角 左:非長打 右:長打

圖(十):全壘打/擊球仰角 左:非全壘打 右:全壘打

根據 MLB 對各個擊球型態之仰角的定義，(如下表)

擊球形態	仰角範圍
滾地球	10 度以下
平飛球	10 度~25 度
飛球	25 度~50 度
沖天炮	50 度以上

另外 8 度~32 度的擊球仰角通常在這角度出去的都能夠非常強勁，因此這個角度區間又被稱作**甜蜜點（sweet spot）**

從圖(八)，可以發現各種各種擊球形態的擊球都有可能打出安打，但是換成只看長打時(圖九)，因為擊出長打通常需要強勁的擊球，因此其仰角大多落在 25 度左右(落在甜蜜點的範圍內)

最後，若只單看全壘打的話，因全壘打大多為飛球和平飛球，因此可以看到圖(十)的仰角大多座落在 25 度左右但 50 度以下．

e. 模型使用與預測

本研究將使用 Random Forest 預測球員擊出的安打和其安打類型．

訓練資料:70% 測試資料:30%

第一輪:預測是否為安打

150972 筆資料，並將預測為安打的資料篩選出來(45292 筆)至下一輪

第二輪:預測安打 是否為長打

將預測為長打的資料篩選出來(共 13588 筆資料)

第三輪:預測這些安打的擊球距離

最後的 13588 筆的長打資料，用來預測其是否為全壘打

四.結果

```
Training accuracy score: 0.7898
Testing accuracy score : 0.7914
```

第一輪

```
Training accuracy score: 0.9271
Testing accuracy score : 0.9281
```

第二輪

前兩輪在整體預測表現上都很正常

```
Training accuracy score: 0.9668
Testing accuracy score : 0.9628
```

但到第三輪時，因資料(13588 筆)漸漸變少，導致 overfitting

五.結論與討論

在第三輪時，資料變的過少導致不斷 overfitting，這邊我認為是在整個選擇欄位的過程中，考慮到的欄位太少了，導致整個預測的結果不如預期．

在這整個預測過程中，我認為還有非常多可以放進來討論的因素，

例如:

1. 預測模型僅考慮擊球初速仰角和投手球速，但未考慮許多球場因素
2. 球員的傷病問題或舊傷問題，是否會影響其安打方面的表現

六.參考文獻

https://www.sportsv.net/articles/84644?page=2

https://zh.wikipedia.org/wiki/%E9%9A%8F%E6%9C%BA%E6%A3%AE%E6%9E%97

https://nkoenig06.github.io/performance-baseball.html

https://jamesrledoux.com/projects/open-source/introducing-pybaseball/