PLG+賽事的分析及預測

第五組

E44096021 林承慶

B54091232 李軒宇

**第一章 緒論**

**1.1研究背景與動機**

Plg+每一季的總冠軍歸屬相信是每一個熱愛籃球的球迷都十分關注的話題，哪支球隊最終能夠出線也是球迷間討論十分激烈的話題，但是要準確的預測冠軍卻不是一間容易的事情。通過大數據分析課程，我們有了一種較為可靠的方法進行分析和預測。

我們希望可以建立一個可靠的數據分析預測模型，並帶有一定的準確性，讓球迷可以有一個討論的依據。我們希望模型能夠達到以下目標：

1. 提供準確的預測結果：通過運用大數據分析和機器學習技術，我們期望能建立一個具有高準確性的模型，能夠預測P. LEAGUE+賽季的勝率以及對戰的勝敗。
2. 探索關鍵指標和特徵：我們可以蒐集並分析豐富的數據，包括球隊的「勝負記錄、得分表現、籃板數據、助攻、犯規、失誤等等…」(我們用的資料)等等。並利用這些數據嘗試建立可靠的模型和算法，來找出不同因素之間的關聯性和趨勢。透過深入的數據分析，我們希望發現隱藏的模式和特徵，並進一步預測贏球隊伍。
3. 考慮歷史記錄和趨勢：我們將納入球隊的歷史記錄以及賽季中的表現，以了解長期趨勢和球隊間的對比。透過分析歷史數據，我們可以發現潛在的模式和趨勢，並將其納入我們的預測模型中，進一步提高預測準確性。
4. 推動大數據在體育預測領域的應用：本研究的成果將不僅僅適用於P. LEAGUE+，還可以應用於其他體育聯賽和比賽。我們的目標是推動大數據分析在體育預測領域的應用和發展，為相關行業的决策制定者提供更多有價值的洞察力和決策支持。

**1.2研究目的**

本研究目的有以下兩點：

1. 嘗試利用球隊的平均數據(得分表現、籃板數據、助攻、犯規、失誤等等…)來判斷該隊在賽季中的勝率
2. 選出今年的冠亞軍隊伍(國王、勇士)的賽季中的對戰紀錄加上冠軍戰的戰績去分析，並且預測兩隊對戰的勝敗。

**第二章 研究過程**

**2.1資料來源與處理**

1. 到PLG+的網站讀取戰績的部分，並且將數據填到EXCLE上面做備用，命名為game\_stats.csv。
2. 一張含有 文字, 螢幕擷取畫面, 字型, 數字 的圖片

   自動產生的描述一張含有 文字, 螢幕擷取畫面, 字型 的圖片

   自動產生的描述查詢國王以及勇士的對戰場次，將數據填到EXCLE中，並且因為其中的贏球隊伍，主客場隊伍都是使用中文呈現，我們利用資料處理的方法變成虛擬變數，以利後面的分析。並且創建新的欄位，來判斷是主場球隊贏球還是客場球隊獲勝，命名為Win\_stats.csv。

**2.2研究方法**

**第一部分:賽季勝率預測**

一張含有 文字, 字型, 筆跡, 螢幕擷取畫面 的圖片

自動產生的描述一張含有 文字, 螢幕擷取畫面, 字型 的圖片

自動產生的描述 首先使用迴歸分析模型，把球隊的攻守數據進行分析，進攻面我們選了兩分球命中率以及三分球命中率數據來分析其賽季勝率，防守面我們則選擇對手的兩分球命中率以及三分球命中率來做分析，計算其R平方值，但當選擇到三個數據去分析的時候R平方值很容易會直接是1，故以分析兩個X變數為主，由此來了解模型是否合適，並且將MSE以及RMSE算出來，以防過度擬合的狀況發生，最後檢查是否有容忍值大於0.1的項目，要將其剔除。

如果為進攻數據就將’對手兩分%’改成’兩分%’，’對手三分%’改成’三分%’。

**第二部分:單場數據預測勝敗**

一張含有 文字, 螢幕擷取畫面, 字型, 文件 的圖片

自動產生的描述利用決策樹分析，選出其中兩個方法，一是用資料增益比的方法，二是用gini分析的方法，先將要選出來做資料增益比以及gini分析的項目給選出來。

一張含有 文字, 螢幕擷取畫面, 字型, 文件 的圖片

自動產生的描述

其中因為我希望可以做出二維的圖，所以在選擇屬性的時候選擇了兩個，當地一個屬性選出後在從剩下的屬性中選出第二合適的。

把目標設定好創建決策樹，其中印出錯誤的個數，並且計算預測的正確率，對比其中使用資料增益比方式的預測以及使用gini方式的預測。

**第三章 研究結果**

**第一部分:賽季勝率預測**

**防守面:**

一開始先以球隊防守面的數據做分析，做對手兩分命中率以及對手的三分命中率，可以從R平方值看出來，其值高達0.98可以知道球隊如果可以將對手的命中率有效降低，對於勝率是有很直接的影響，對於建模預測也很好，其MSE值也很小，表示模型適合，最後看到是否有超過容忍值，也是沒有，很好的說明

一張含有 文字, 螢幕擷取畫面, 行, 圖表 的圖片

自動產生的描述一張含有 文字, 字型, 螢幕擷取畫面 的圖片

自動產生的描述了正確的數據選擇。最後使用繪圖的方式來看預測值與實際值的誤差。

做出來的預測值跟實際值還是有一段落差，沒辦法很好的去精準預測出賽季的勝率。

**進攻面:**

接著看到進攻數據的部分，也可以看到R平方值很高，但是略低於防守數據的部分，其MSE及RMSE也都跟防守數據類似，說明了進攻的部分也跟模型有很好的適配性。

利用進攻數據做出來的預測值跟實際值的誤差比較起防守數據時做出來的誤差是比較小的，可以看出來利用進攻數據做進一步的預測會是更好的選擇。一張含有 文字, 螢幕擷取畫面, 行, 圖表 的圖片

自動產生的描述

**第二部分:單場數據預測勝敗**

一張含有 文字, 字型, 白色 的圖片

自動產生的描述先將剛剛的合適屬性判斷出來:

**Entropy部分:**

一張含有 文字, 螢幕擷取畫面, 圖表, 行 的圖片

自動產生的描述 利用剛剛的一開始判斷出來的屬性，去建立決策樹，再利用繪圖的方式將兩個被選出來的屬性做XY軸，看是否可以藉由判斷這些項目來預測出球隊的輸贏，藍紫色的部分是主場球隊贏球，而黃色的部分是客場球隊贏球。

預測的正確率為0.6，發生了兩個錯誤的判斷，可以從這個圖看出來，當黃色客場的球隊的兩分球命中率超過大約五成的時候並且不讓主場球隊拿出三分球超過4成的命中率的時候，就有很高的機率可以贏球，而當藍紫色主場球隊的三分球命中率達到四成以上的時候更有著100%的勝率(這部分有牽涉到場次較少的問題故勝率到100%)。

做出來的G一張含有 文字, 螢幕擷取畫面, 字型, 行 的圖片

自動產生的描述raphviz圖可以看出來以客隊兩分球命中率使否在48%以上，有兩個點位錯誤就是在上上個圖中可以看到最低那一點以及最靠右邊那點，一刺客隊以不到四成的命中率拿下比賽，另外一次客隊有著高於六成的兩分命中率，卻因為主隊有超過四成的三分命中率而敗北。

一張含有 文字, 螢幕擷取畫面, 圖表, 繪圖 的圖片

自動產生的描述**GINI分析部分:**

一張含有 文字, 螢幕擷取畫面, 字型, Rectangle 的圖片

自動產生的描述跟利用entropy的方法一樣做出了0.6的預測準確度，同樣是發生了兩次的錯誤預測，而從圖中可以看到藍紫色的部分，也就是主場球隊的部分，在得分低於約90分的時候基本上勝率比較低，而在黃色客場贏球的部分則是，如果客場球隊可以在比賽中拿出超過0.33左右的三分球命中率且限制主場球隊不拿下超過110的得分，那就可以很好的判斷出客場球隊的勝利。

在圖中看出來先以主隊得分是否未達到89分以上，未達到即是客隊獲勝，接著判斷客隊三分球命中率是否未達31%，未達到即是主隊獲勝，接著判斷主隊分數是否未達到111.5分，達到即主隊勝，接著判斷主隊得分是否未達到97分，達到即主隊勝，最後再利用判斷客隊三分球命中率來分出最後兩個，整個決策數長得很高，比較起使用Entropy的方法並不是那麼有效率。

**第四章 結論**

**第一部分:賽季勝率預測**

不管事從進攻面數據還是防守面數據都可以看出來對於球隊勝率來說都有是重要的因素，但其中進攻面的預測數值又比起防守數值更準確，其原因有二:

(1)以籃球這項運動來講，它的進攻的方式會比防守來的更複雜且豐富，也因為這樣，如果可以得到較高分的球隊即獲勝(2)在球場中一般來說是以進攻較快較猛的球隊在控制整場的節奏，進攻較好的球隊會直接的讓防守變得簡單，因為

會有足夠的時間進行回防等等，所以給予進攻數據去做預測會有比較準確的結果。

**第二部分:單場數據預測勝敗**

目前利用決策樹所求出來的預測單場勝敗能夠做到0.6的預測正確率，目前也無法只用兩項數據就把勝敗以更好的準確率預測出來，籃球比賽不易於僅看幾個數據就可以判斷精確，因為兩端的數據是互相影響的，跟球隊對球隊的戰術配置也有關聯性。

**第五章 後續研究建議**

1. 以賽季勝率的預測來說，可以納入更進階的數據，像是進攻效率值，防守效率值，節奏值，以及一些跟得分較無直接關係的籃板助攻等等…。
2. 以單場數據去預測勝敗的部分，之後應要要用到其他的機器學習模型或是將兩個目標轉為三個，四個等更多的目標…，並且使用網站爬蟲，抓取更大的樣本數(比賽場次)，來做分析。
3. 這樣的機器學習不只是可以預測勝敗，還可以加入球局的賭盤預測，加入賭盤的讓分機制以後，去讀取過往的賭盤數據做分析，應該也可以做出不錯的賭盤預測效果。
4. 抓取的樣本數到一定的量以後，並且學習更深的機器學習方法，應該可以進一步預測系列賽的勝敗(7戰4勝)，但這個的場外因素有更多，也考驗每場球之後教練的調度，以及對對手的針對性策略，也同時增加預測難度。

**第六章 Q&A**

1. 為甚麼沒有想做NBA，那有更多更廣且好用的數據?
2. 目前在網路上充斥著許多的大數據分析的資料，很容易可以找到有人分析NBA的部分，且最近PLG在打冠軍賽，十分激烈，故想實作看看。
3. 為什麼是選擇使用用決策樹做分析?
4. 因為每一場球的數值差距比較大，我們覺得用決策樹可以比較好的分出結果。
5. 有想過要做預測冠軍的部分嗎?
6. 是有的，但因為預測冠軍的變數會變得很多，並不只是單場的表現，還包含到甚至是心理方面的狀況，每一場球的戰術策略可能也都不同。這個會放在後續研究建議。