

# Machine Learning approaches to predict season outcomes for the NBA

Group2

王勁程

黃章瑋

蔡欣緹

## 1. Research Question and Background

隨著運動數據科學的發展，單純的得分已不足以評估球隊競爭力。本研究旨在透過包含傳統數據與進階數據（Advanced stats）的全方位變項，建立預測模型，並探討不同演算法（Logistic, SVM, RF, CART）在預測準確度上之差異。基於此，我們提出以下研究問題：

- 在 2024-25 的 NBA 球季中，哪些球隊統計數據（如投籃命中率、進攻效率等）隊贏得比賽最具預測力？
- 我們能否利用機器學習模型，僅透過球隊整體統計數據精確預測單場比賽的勝負？

## 2. Data Sources

本研究使用 Python 第三方函式庫 `nba_api` 進行資料抓取。該工具直接呼叫 NBA 官方 API 端點（[stats.nba.com](https://stats.nba.com)），確保數據的準確性及即時性。

我們針對 2024-25 例行賽賽季，擷取了以下核心資料：

- Team stats：包含球隊的進攻、防守、投籃命中率等超過 30 項基礎及進階數據。
- Game logs：記錄每場比賽的對戰組合（Matchup）、主客場狀態以及最終勝負結果。
- Team info：用於建立球隊 ID 及縮寫（Abbreviation）的對照關係

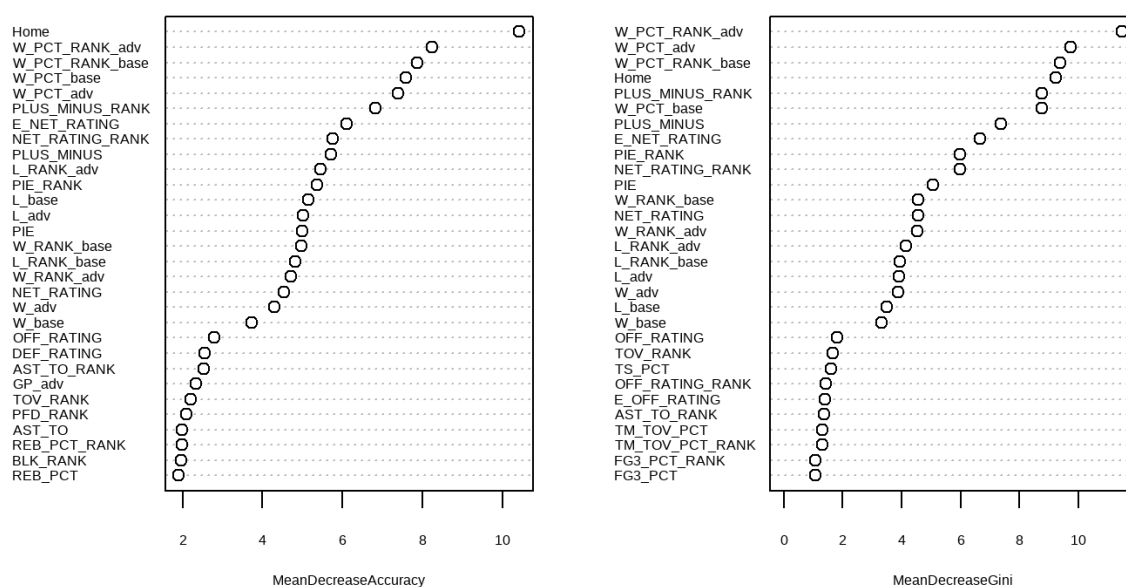
## 3. Data preprocessing

本研究首先整合多個 NBA 官方資料表，以建立逐場比賽層級的分析資料集。資料來源包含比賽結果（GameLogs）、球隊賽季整體表現（TeamStats）、球隊基本資訊（TeamInfo）以及球員個人表現（PlayerStats）。資料整合過程主要透過 `dplyr` 套件中的 `left_join` 函數，依據球隊識別碼（TEAM\_ID）與球隊縮寫（TEAM\_ABBREVIATION）進行串接，以確保各資料表能在同一分析架構下對齊。

此外，在比賽層級的資料處理中，我們自比賽對戰欄位（MATCHUP）中解析主客場資訊，將「@」符號轉換為二元變數 Home（主場 = 1，客場 = 0），以捕捉 NBA 比賽中顯著存在的主場優勢。同時，比賽結果（WL）被轉換為二元結果變數 WL\_num（勝 = 1，負 = 0），作為後續分類模型的應變數。

為了確保 Logistic regression 不會過擬和，但要給予機器學習模型夠多變數，我們經過初步 RF 建模篩選出重要變數，全變數 RF 建模結果如下：

隨機森林變項重要性排行



經過排除直接反映勝負的變數（如 W\_PCT\_RANK 即為賽季勝率排行）以及預測率過低的變項，最終篩選出以下變項作最後建模：

變項	說明	變項	說明
傳統數據		進階數據	
FGM, FGA, FG_PCT,	投籃命中、出手、命中率	AST_RATIO	每 100 回合共有多少次助攻
FG3M, FG3A, FG3_PCT	三分球命中、出手、命中率	REB_PCT	球隊搶下籃板總數占總籃板球機會的百分比
FTM, FTA, FT_PCT	罰球命中、出手、命中率	TM_TOV_PCT	球隊進攻回合結果為失誤的比例
OREB, DREB, REB	進攻籃板、防守籃板、總籃板	EFG_PCT	有效命中率，公式為： $((FGM + (0.5 * 3PM)) / FGA)$
AST	助攻	TS_PCT	真實投籃命中率，除考慮兩分球外還考慮了三 分球和罰球的價值，公式為 $PTS / [2 * (FGA + 0.44 * FTA)]$
TOV	失誤	PIE	球員貢獻值，衡量球員在比賽中相對於總統計 數據的整體統計貢獻。

STL	抄截	PACE	每 48 分鐘總共有多少進攻回合
BLK	阻攻	NET_RATING	淨效率，衡量球隊每百回合的淨勝分
BLKA	投籃被封阻數		
PF	個人犯規數		
PFD	被對手犯規數		
PTS	得分		

## 4. Data analysis

### 1. Logistic regression

羅吉斯迴歸是本研究採用的基準模型（Baseline model），用於建立球隊各項統計指標與獲勝機率之間的線性關係。

模型透過 logit 轉換將線性組合的預測值映射到  $[0, 1]$  區間，公式如下：

$$P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \dots + \beta_n X_n)}}$$

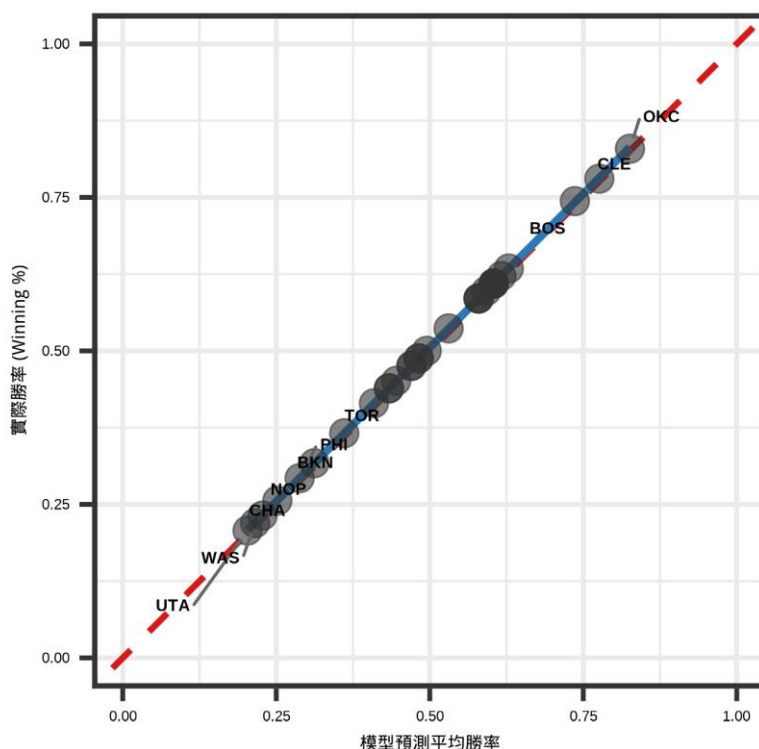
設定門檻值為 0.5，當模型輸出的預測機率大於 0.5 時，判定該場比賽為「勝」。

最終模型預測準確度為 0.6216，羅吉斯迴歸能提供直觀的解釋性，讓我們了解單項指標對贏球勝算比（Odds ratio）的貢獻。

最終各隊預測散點圖如下圖所示：

Logistic Regression: 預測 vs 實際勝率

紅色虛線: 完美預測 | 藍色實線: 模型擬合趨勢



## 2. SVM

為了捕捉球隊數據之間複雜的非線性互動關係（例如當籃板數領先但失誤率也同時過高的綜合影響），本研究導入了 SVM 模型。

SVM 的目標是在高維特徵空間中尋找一個「最佳超平面 (Optimal Hyperplane)」，使兩類樣本（勝 vs 負）之間的時間達到最大化。

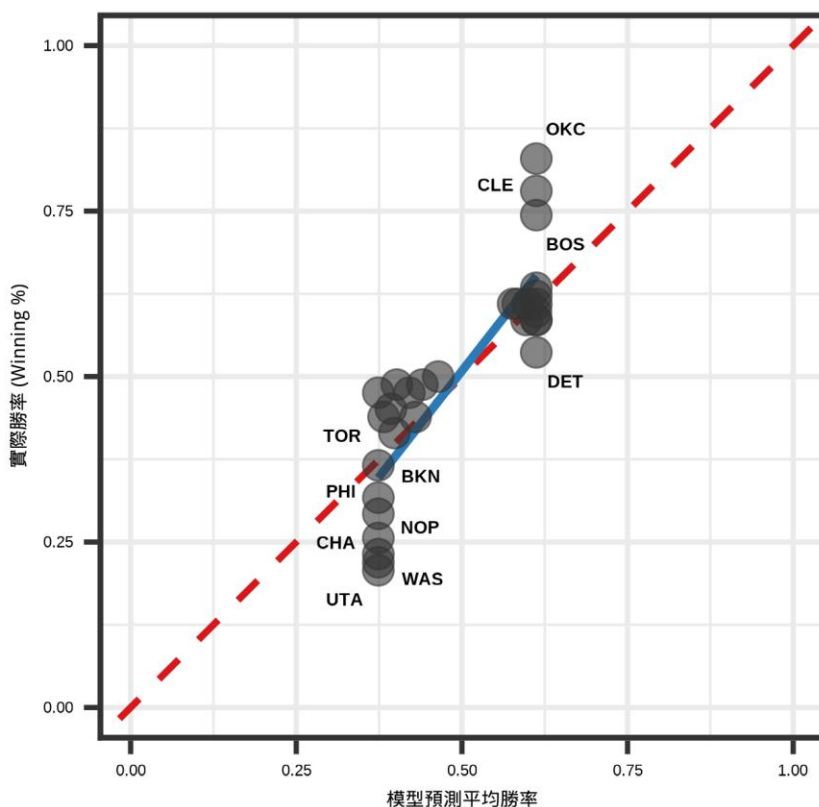
本研究採用「Radial Basis Function, RBF」作為 Kernel Function，使模型能夠處理非線性的決策邊界，將原始數據映射到更高維度的空間進行分類。由於 NBA 數據數量級差異巨大（例如總得分通常破百，而三分命中率僅為小數），本分析在訓練 SVM 模型時執行了標準化處理 (scale = TRUE)，確保模型不會被數值較大的變項所主導。

最終 SVM 模型預測準確度為 0.6259，SVM 在處理高維度數據時具有極強的泛化能力，預期在精準度上能較傳統線性模型有所提升。

最終各隊預測散點圖如下圖所示：

### SVM: 預測 vs 實際勝率

紅色虛線: 完美預測 | 藍色實線: 模型擬合趨勢



### 3. Random Forest

為了進一步捕捉 NBA 球隊統計指標之間可能存在的非線性關係與高階交互作用，本研究採用 Random Forest（隨機森林）模型進行預測分析。Random Forest 屬於集成式學習方法，透過對訓練資料進行重複抽樣（bootstrap sampling），建立多棵決策樹，並以投票方式整合各樹的分類結果，以降低單一模型過擬合的風險並提升整體預測穩定性。

在本研究中，Random Forest 模型以完整的數值型球隊與球員特徵作為輸入，並設定樹木數量為 500 棵，每次節點分裂僅考慮部分隨機抽取的變數（ $mtry = \sqrt{p}$ ），使模型能有效探索不同特徵組合下的預測結構。相較於 Logistic regression 與 SVM，Random Forest 不需事先假設變數與勝負結果之間的函數形式，亦不依賴資料線性可分的假設，特別適合用於處理本研究中高維度、變項來源多元的 NBA 資料。

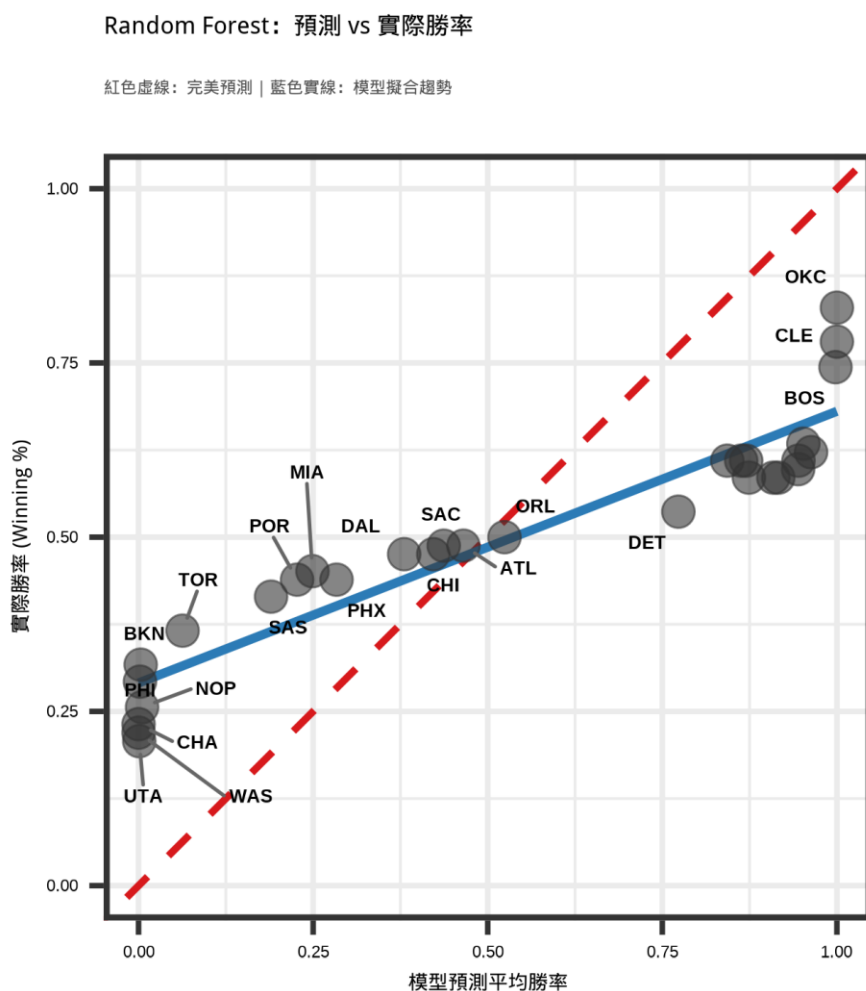
由 Random Forest 的「預測勝率 vs 實際勝率」散點圖可觀察到，模型整體呈現明顯的正向趨勢，顯示其能有效區分實力較強與較弱的球

隊。然而，相較於 Logistic regression 與 SVM，Random Forest 的預測結果呈現出「向平均值收斂」的現象：對於實際勝率極高的球隊（如 OKC、CLE），模型預測勝率略低於實際值；而對於實際勝率偏低的球隊，模型則傾向給予略高的預測勝率。

此現象反映 Random Forest 的核心特性之一，即透過多棵決策樹取平均來降低極端預測所造成的變異（variance reduction）。雖然這種保守特性有助於提升整體穩定度，但也可能使模型在預測聯盟頂尖或墊底球隊時出現系統性低估或高估。

在預測準確度方面，最終 Random forest 模型預測準確度為 0.6285，整體表現優於單一決策樹（CART）和 SVM 模型，顯示其在兼顧預測能力與模型穩定性方面具有良好表現。整體而言，Random Forest 能有效整合多層次球隊與球員資訊，適合作為本研究中評估 NBA 比賽勝負的主要非線性預測模型之一。

最終各隊預測散點圖如下圖所示：



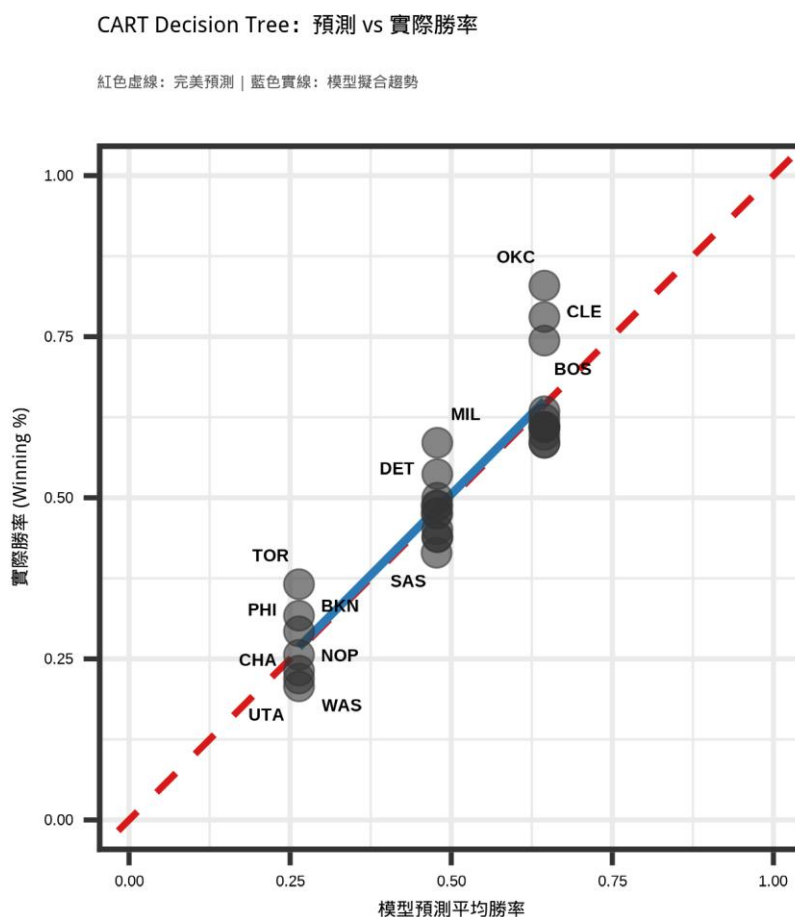
## 4. CART

本研究亦使用 CART (Classification and Regression Tree) 決策樹模型作為另一種非線性分類方法。CART 透過一連串的 if-then 分裂規則，將資料逐步劃分為勝率結構相對一致的子群體，使模型具有高度可解釋性，能清楚呈現哪些變項在不同決策層級中影響勝負判定。

由 CART 的「預測勝率 vs 實際勝率」散點圖可清楚觀察到階梯狀 (stepwise) 的分布結構，反映決策樹模型本質上僅能在有限的葉節點中產生離散化的預測結果。相較於其他模型，CART 對於不同實力層級的球隊能進行粗略分類，但在同一節點內的球隊，其預測勝率往往完全相同，限制了預測的精細程度。

在預測表現上，CART 模型與 SVM 的預測結果相近，為 0.6259，但較低於 RF，顯示單一決策樹容易受到資料切割方式影響而產生較高變異。然而，其高度可解釋的結構仍有助於理解模型決策邏輯，並作為後續 Random Forest 集成學習的重要基礎。

最終各隊預測散點圖如下圖所示：



## 5. Literature Review and Methodological Comparison

過去已有相當多研究嘗試利用機器學習方法預測 NBA 單場比賽的勝負結果，多數研究將問題設定為二元分類（Win / Loss），並以球隊層級或球員層級的歷史統計數據作為模型輸入變數。常見的方法包括 Logistic Regression、Support Vector Machine（SVM）、Random Forest 等監督式學習模型，其預測準確率多落在約 60%–70% 的區間，顯示在高度不確定且受隨機因素影響的運動賽事中，勝負預測本身具有一定的難度。

為了與既有研究進行方法論上的對照，本研究選擇 Horvat et al.（2023）發表於 *Symmetry* 期刊的一篇研究作為文獻回顧與比較對象。該研究同樣以 NBA 單場比賽勝負預測為研究目標，但在方法設計上與多數機器學習導向的研究明顯不同。

在問題設定上，Horvat 等人將每一場 NBA 比賽視為一個二元分類問題，與本研究相同，皆以「勝 / 負」作為預測目標。然而，在資料使用與模型建構策略上，該研究並未採用 Random Forest、SVM 等標準的機器學習分類器，而是嘗試透過資料驅動的效率指標與規則式方法進行預測。其核心動機在於探討，是否一定需要高度複雜的模型結構，才能在 NBA 勝負預測問題中取得合理的預測表現。

在資料層級方面，該研究主要使用球隊層級的歷史表現資料，例如進攻效率與防守效率等整體指標，而未納入球員層級或即時比賽資訊，與本研究以賽季累積球隊統計作為主要資訊來源的設計邏輯相近。不同之處在於，Horvat 等人進一步設計了一個延伸的球隊效率指標，特別強調兩支對戰球隊之間的「相對效率差異」，而非僅關注單一球隊的絕對表現水準。

在方法上，Horvat 等人引入「最佳時間窗（Optimal Time Window）」的概念，系統性地測試使用不同長度的歷史比賽資料進行預測，藉此評估資料時間範圍對模型表現的影響，而非假設使用越長期的歷史資料必然能提升預測效果。最終，該研究透過一個勝率函數（Win Function），將兩隊效率指標的差值轉換為勝負預測結果，使整體方法更接近資料驅動的規則式分類，而非傳統意義上的機器學習模型訓練流程。

在結果方面，Horvat 等人報告其模型的平均預測準確率約為 66%，在最佳設定下可達約 78%。值得注意的是，這樣的表現與許多使用 Random Forest 或 SVM 的研究結果相當，顯示即使不依賴複雜的機器學習模型，透過精心設計的

效率指標與資料驅動規則，仍能在 NBA 單場勝負預測問題中取得具競爭力的表現。

與該研究相比，本研究採用的是較為典型的機器學習框架，系統性比較 Logistic Regression、SVM、Random Forest 與 CART 四種模型在相同資料架構下的預測表現。雖然本研究最終模型的預測準確率同樣落在約 62%–63% 的區間，略低於 Horvat 等人在最佳情境下所報告的上限值，但兩者在預測水準上屬於相近範圍，顯示本研究結果與既有文獻具良好一致性。

整體而言，Horvat et al. (2023) 的研究提供了一個重要的比較視角，說明在 NBA 勝負預測問題中，模型複雜度並非唯一決定預測表現的關鍵因素。本研究則進一步從機器學習模型比較的角度，展示不同演算法在預測能力、穩定性與解釋性之間的取捨，並與效率導向、規則式方法形成方法論上的互補關係。此比較有助於更全面地理解 NBA 勝負預測問題中，不同分析策略各自的優勢與限制。

## 6. Conclusion and Discussion

本研究以 2024–25 賽季 NBA 例行賽為研究對象，整合球隊層級與球員層級之賽季統計資料，建構逐場比賽層級的分析資料集，並比較 Logistic Regression、Support Vector Machine (SVM)、Random Forest (RF) 及 CART 四種模型在單場比賽勝負預測上的表現。研究結果顯示，即使僅使用賽季累積的球隊與主要輪換球員統計指標，各模型仍能有效捕捉球隊實力差異，並對比賽結果提供具合理準確度的預測。

在模型比較方面，Logistic Regression 作為基準模型，具備良好的解釋性，使得各項表現指標對勝負機率的影響方向與相對重要性得以被清楚說明。然而，其線性假設亦限制了模型對於球隊效率、節奏與球員表現之間非線性交互關係的刻畫能力。相較之下，SVM 與 Random Forest 能夠處理更複雜的非線性結構，在整體預測表現上略優於 Logistic Regression，顯示機器學習方法在整合多維度籃球數據時具備一定優勢。

其中，Random Forest 展現出最穩定的整體表現，反映集成學習在降低模型變異與提升泛化能力上的優勢。然而，RF 預測結果亦呈現對極端強弱球隊勝率向聯盟平均值收斂的現象，顯示模型在追求穩定性的同時，可能犧牲對極端結果的敏感度。CART 模型則以其高度可解釋的樹狀決策結構，提供對模型判斷邏輯的直觀理解，雖然預測準確度相對較低，但在說明勝負判斷規則與輔助決策解讀上仍具有重要價值。

整體而言，本研究結果顯示，不同模型在「預測能力」與「解釋性」之間存在明顯取捨，模型選擇應依研究目的而定。對於重視解釋與變數影響分析的情境，傳統統計模型仍具不可取代的優勢；而在追求整體預測表現時，機器學習模型則能提供額外的改善。本研究結果顯示，在明確界定資料層級與研究假設的前提下，僅透過賽季層級的球隊與球員資訊，仍可對 NBA 單場比賽勝負提供具實證基礎的解釋與預測。

## 7. Limitations and Future Work

本研究仍存在若干限制，需於解讀研究結果時加以審慎考量，亦為後續研究提供明確的延伸方向。首先，在資料層級方面，本研究使用之球隊與球員變數皆為賽季累積統計指標，未能反映比賽當下的即時狀態變化，例如近期傷病情形、背靠背賽程所造成的疲勞效應、主力球員輪休策略，或短期狀態波動等因素。此一設計隱含球隊整體實力於賽季期間相對穩定的假設，因此本研究之模型較適合用於解釋與評估球隊的「長期實力水準」，而非作為即時賽前預測或臨場決策之工具。

其次，本研究僅使用單一賽季資料進行模型訓練與評估，未進行跨賽季或時間切割的驗證設計，可能使模型表現受到特定賽季環境或聯盟趨勢影響，進而高估其在其他賽季或未來比賽中的泛化能力。未來研究可考慮納入多賽季資料，並採用滾動視窗（rolling window）或時間序列交叉驗證（time-series cross-validation）方法，以更貼近實際預測情境，並提升模型穩定性與外推性。

在球員層級資料處理方面，雖然本研究已透過選取每隊上場時間前八名球員以近似主要輪換陣容，藉此引入球員影響因素，但此作法仍難以完整捕捉實際比賽中的戰術調整、角色轉換與特定對戰組合所產生的交互效果。鑒於球員間的場上互動與陣容搭配可能對比賽結果產生關鍵影響，後續研究可進一步引入加權球員影響指標，或使用 lineup-based 資料，以更細緻地刻畫比賽動態。

此外，在模型評估指標的選擇上，本研究主要以分類準確率作為比較不同模型表現的基準。雖然準確率能提供直觀的整體預測能力衡量，但其未能充分反映模型在不同勝率區間的預測信心與校準表現。未來研究可進一步納入如 AUC、Brier score 或機率校準曲線（calibration curve）等評估方式，以更全面地比較不同模型在風險評估與不確定性刻畫上的差異。

總體而言，儘管本研究仍有若干改進空間，但已建立一套結構清楚、方法一致且具可重現性的 NBA 單場比賽勝負預測流程。透過明確界定研究假設與分

析範圍，本研究不僅呈現不同模型在相同資料結構下的相對表現，也為未來結合時間動態資訊與更高解析度資料的延伸研究奠定基礎。

## 8. 參考文獻

[A Data-Driven Machine Learning Algorithm for Predicting the Outcomes of NBA Games](#)