Kobe Bryant's Shot Selection

統計107 張尹甄 統計107 鄭凱元

統計107 蘇東洵 統計107 張鈺欣

1. 報告大綱

這筆資料裡包含了Kobe Bryant在其20年職業生涯中的每個出手的位置與情況，而我們要做的事是預測Kobe Bryant是否有投進。

1. 研究動機

籃球場上的每個進攻與防守回合，皆與最終的勝利息息相關，要如何最大化己方的進攻效率與如何限制對方的得分會是每場比賽需要做的事情，因此若能從球員的表現從而預測其接下來的行為，將會有效幫助球隊制定戰術、贏得比賽！

我們選擇了NBA退役傳奇球星Kobe Bryant來當我們了研究對象，透過分析Kobe Bryant投籃的背景資料去探討哪些條件會影響出手的投進與否，以提供資訊去預測Kobe Bryant是否進籃，並參考其分析方式套用在其他球員身上，預測其賽季表現。

1. 資料型態

資料中一共有25個變數，可大致分為編號、時間、空間與其他四大類。

|  |  |  |
| --- | --- | --- |
|  | 類別型 | 數值型 |
| 編號類 | game\_event\_id / game\_id / team\_id / shot\_id | |
| 時間類 | period / playoffs / season / game\_date | minutes\_remaining / seconds\_remaining |
| 空間類 | shot\_zone\_area / shot\_zone\_basic / shot\_zone\_range | lat / lon / loc\_x / loc\_y / shot\_distance |
| 其他 | action\_type / shot\_made\_flag / combined\_shot\_type / matchup / shot\_type / team\_name / opponent |  |

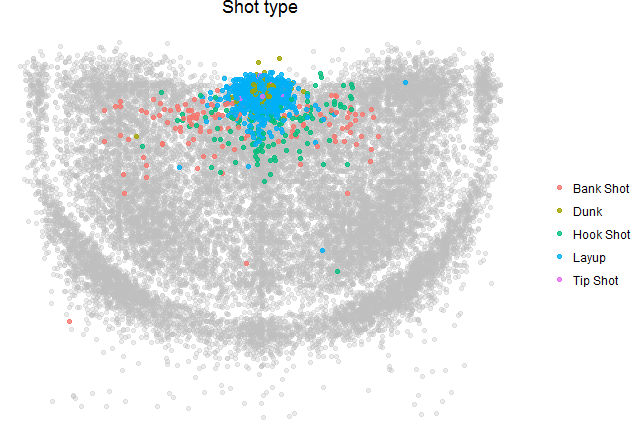
各個變數的意思分別如下表：

|  |  |
| --- | --- |
| 變數 | 定義 |
| shot\_id | Unique ID of a shot. |
| action\_type | The action type of a shot. (Total of 57 different types) |
| lat | Latitude of a shot. |
| lon | Longitude of a shot. |
| loc\_x, loc\_y | The coordinates on the court where the shot took place. |
| minutes\_remaining | Minutes remaining in that quarter. |
| period | Period = Quarter = 1, 2, 3, 4, 5, 6, 7 |
| playoffs | Whether or not the game is in the playoff. (0 = yes, 1 = no) |
| season | Season of the game. (Total of 20 different seasons) |
| seconds\_remaining | Seconds remaining in that quarter. |
| shot\_distance | Distance to the basket. |
| shot\_zone\_area | The zone where the shot take. (Total of 6 different zone areas) |
| game\_date | Date of the game. |
| opponent | Opponent of Lakers. (Total of 33 different opponents) |
| matchup | Whether or not the game is in home field. (@ = away, vs = home) |
| team\_name | The team name Kobe belong |
| team\_id | Unique ID of team(Lakers). |
| shot\_zone\_range | The distance where the shot take. (Total of 3 different range) |
| shot\_zone\_basic | The zone where the shot take. (Total of 7 different basic) |
| shot\_type | The type of a shot. |
| game\_id | Unique ID of game. |
| game\_event\_id | Unique ID of game event. |
| combined\_shot\_type | Combine similar shot type. (Total of 6 different level) |
| shot\_made\_flag | Whether or not the shot is in. (0 = missed, 1 = made) |

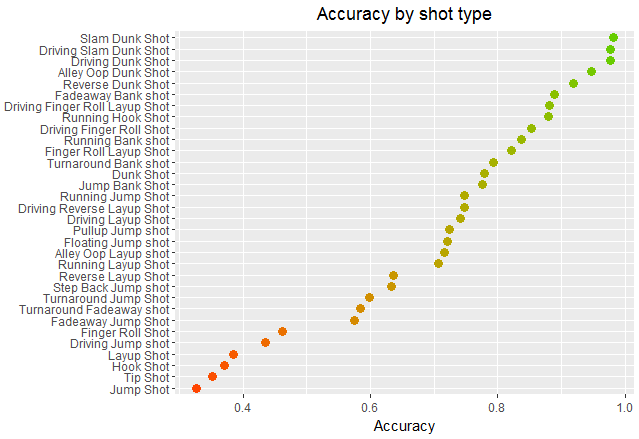
我們將用其餘24個變數去預測shot\_made\_flag這個變數，並想辦法提升它的準確率。在training data中，shot\_made\_flag分為兩種：0為沒有投進籃框，而1為有投進籃框。但在這次的Kaggle競賽上是用Log Loss來判斷模型的好壞，故我們預測出的結果可以是0到1之間的小數。

1. 敘述統計

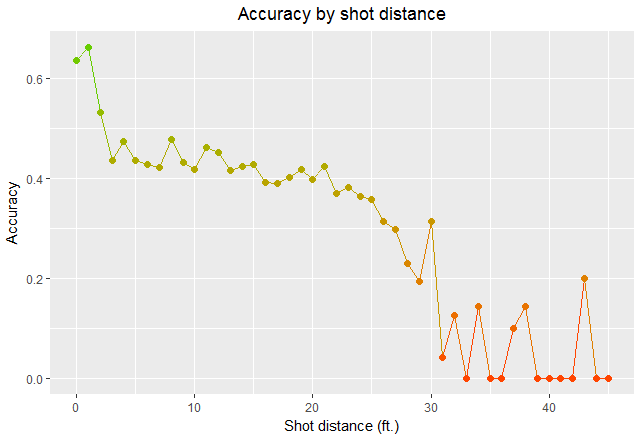
下圖為Kobe Bryant 生涯中所有出手的出手方式，其中灰色的部分為jump shot，由此圖可以明顯看出jump shot為Kobe Bryant 20年生涯中最常使用之出手方式。



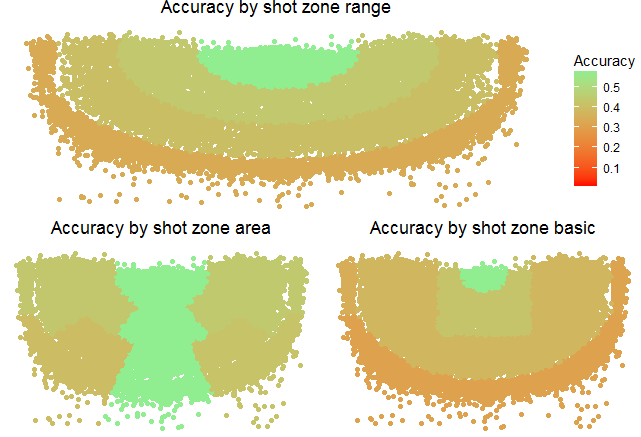
下圖為各出手方式之命中率，可以粗略的看出dunk shot之命中率在所有出手方式中算是名列前茅，而出手位置離籃框的遠近將會影響其出手方式，故接下來來看出手距離與命中率的關係。



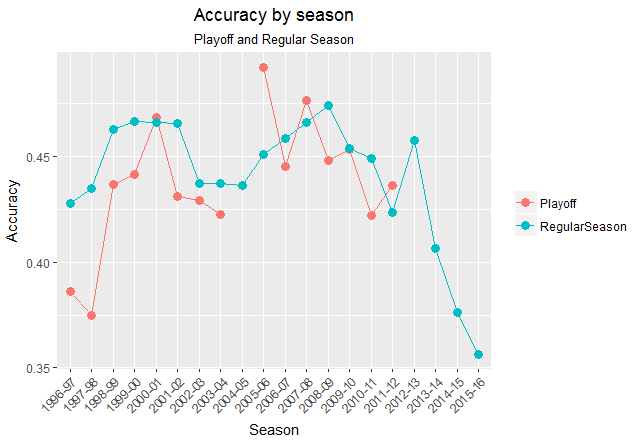
下圖為與籃框距離遠近之命中率，我們可以很清楚的看出，離籃框越近，Kobe Bryant 出手的命中率越高。



下圖為Kobe 在不同區域時出手時的命中率，可以看出除了離籃框越近、命中率越高之外，Kobe 在面對籃框正中間的區域的命中率高於面對籃框左右兩側，而面對籃框右側之命中率又略高於面對籃框左側。



下圖為Kobe Bryant 生涯中20個不同賽季的命中率，圖中藍色的部分為例行賽之命中率，紅色的部分則是季後賽之命中率。從例行賽的命中率可以看出，即使如Kobe Bryant 般偉大的籃球員，依然無法抵擋時間老人的摧殘，在接近生涯晚期的幾年中，命中率大幅的衰退。



1. 資料預處理

資料中一共有25個變數，可大致分為編號、時間、空間與其他四大類，故我們一開始便先將大部分編號類的變數都先丟棄，只留shot\_id來當label。

利用fread指令讀取資料時，將string的變數先轉成factor，再將其他numeric的變數如period與playoff等類別性質的變數再個別轉成factor。

由於空間類的變數有不少個，且彼此之間有一定的相關性，所以我們將性質重複的變數也丟棄。

時間類變數的部分，我們將minutes\_remaining乘60與seconds\_remaining相加，令其為一新的變數。

最後是將opponent變數轉換成”Guest\_Home”的新變數，表示為該比賽為主場或客場。

1. 模型

在這次報告中，我們使用了KNN, Logistic Regression,和XGBoost來配適這筆資料。各模型預測的效果好不好是先使用cross-validation來檢測模型預測的結果，再將預測出的結果上傳至Kaggle的比賽網站去測Log Loss。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Method to predict** | **Accuracy**  **(CV)** | **Logloss**  **(Kaggle)** | **Kaggle上排名** |
| **Model 1** | KNN | 0.5962249 | 18.88624 | 1115/1117 |
| **Model 2** | Logistic Regression | 0.6892392 | 0.61406 | 611/1117 |
| **Model 3** | XGBoost | 0.6825794 | 0.60542 | 352/1117 |
| **Model 4** | Ensemble | -- | 0.60625 | 381/1117 |

* Logistic Regression

由邏輯斯迴歸之anova表(詳見附錄)可以看出，在許多顯著的出手選擇中，其預估之參數為負數，比如說hook shot之係數為-3.319，可以解讀為若Kobe Bryant使用此種出手方式，則會顯著降低他的命中率，所以如果要提升命中率，Kobe Bryant可以考慮減少報表中為顯著但係數為負數的出手方式。

而在shot zone area的部分，在Center, Left Side Center, Left Side, Right Side Center以及Right Side皆會顯著提升Kobe Bryant的命中率，所以可以多嘗試在這些shot zone area的出手以提高進攻效率。

最後是不同賽季的變數，可以看到在最後兩年的賽季變數很顯著的降低命中率，也可以從此看出時間流逝對運動員的生涯表現來說有很大的影響，再偉大的運動員也抵擋不住時間所帶來的退化。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | P-value |  |
| action\_type\_Hook Shot | -3.319 | -0.5209 | 1.88 x 10-10 | \*\*\* |
| shot\_zone\_area\_Center(C) | 3.799 | 1.018 | 0.000190 | \*\*\* |
| shot\_zone\_area\_Left Side Center(LC) | 3.779 | 1.018 | 0.000206 | \*\*\* |
| shot\_zone\_area\_Left Side | 3.634 | 1.023 | 0.000381 | \*\*\* |
| shot\_zone\_area\_Right Side Center(RC) | 3.969 | 1.018 | 9.69 x 10-5 | \*\*\* |
| shot\_zone\_area\_Right Side(R) | 3.802 | 1.022 | 0.0002 | \*\*\* |
| season\_2014-15 | -0.3809 | 0.111 | 0.000602 | \*\*\* |
| season\_2015-16 | -0.5742 | 0.09792 | 4.52 x 10-9 | \*\*\* |

* KNN

起初建立KNN模型時，用CV預測出來的準確率就有六成左右。但由於KNN預測出的結果為類別型資料，即0(沒投進)與1(有投進)，而Kaggle評估結果準確程度的基準為Log Loss，故用KNN預測出來的結果計算的Log Loss會很大。因此，我們決定不再優化我們的KNN模型。

* XGBoost

在 R 裡面，使用的是套件xgboost ，步驟如下:

1. 將資料格式(Data.frame)，用xgb.DMatrix()轉換為 xgboost 的稀疏矩陣
2. 設定xgb.params，也就是 xgboost 裡面的參數
3. 使用xgb.cv()，tune 出最佳的決策樹數量(nrounds)
4. 用xgb.train()建立模型
5. 利用把cv.model中的best\_iteration取出來，這是 xgboost 內部會自動判斷最好的nrounds是多少，其方法是觀察 Train 跟 Validation 之間的表現差異

* Ensemble

將Logistic 及 XGBoost 模型跑出的最好結果進行ensemble，但我們發現ensemble後的結果皆不如XGBoost單一模型跑出的結果。因此最後還是採用XGBoost為最終模型。

1. 結論

一個人出手的投進與否，將會受到許許多多因素所影響，所以並不是那麼容易去做完美的預測，但仍然能夠依據各種條件提供資訊去粗略的判斷這球是否會投進，再依當天的情況去針對不同球員執行不同的防守策略。

在NBA裡不時會看到比賽進行中的暫停或是中場休息時間時，教練團會依據對方或己方球員的表現去制定接下來的戰術，所以如果能夠透過統計分析更有依據的去分析對手的行為，能夠有效協助教練團去下達在進攻或防守上的指令。

1. 附錄 – Logistic Regression Anova Table

