Logistic Regression Models Analyzing Likelihood of Winning

2024-05-15

Abstract

This analysis investigates the key factors influencing the likelihood of winning NBA games, with a specific focus on metrics such as home court advantage, rebounding, three-pointers made and attempted, turnovers, and shooting percentages.

Utilizing logistic regression models and a correlation matrix heat map, the study reveals significant predictors of game outcomes. Field goal percentage and three-point made advantages are the strongest predictors of margin of victory, underscoring the importance of shooting efficiency. Additionally, turnover and rebound advantages positively impact winning chances.

Despite a slight negative correlation between three-point attempts and winning probability, the study highlights a notable rise in three-point attempts over the years, prompting a deeper exploration of evolving NBA strategies. This comprehensive analysis aims to provide valuable insights into the effectiveness of contemporary basketball strategies, informed by historical data and statistical trends.

Limitations

It is important to acknowledge that the dataset has inherent limitations, including incomplete data for certain years and potential biases in data collection methods. Additionally, changes in game rules, such as the introduction of the three-point line in 1979, could influence the trends observed in the data.

```
library(DBI)
library(RSQLite)
library(ggplot2)
library(ggcorrplot)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
db_path <- "/Users/colinhadden/Downloads/archive/nba.sqlite"

con <- dbConnect(RSQLite::SQLite(), dbname = db_path)

tables <- dbListTables(con)

gameTable <- dbGetQuery(con, "SELECT * FROM game")
```

This setup allows for subsequent data analysis and visualization tasks on the game table data from the NBA SQLite database.

Home Court Advantage

```
home_court_adv_query <- dbGetQuery(con, "
SELECT
    wl_home AS win_loss,
    CASE
        WHEN wl_home = 'W' THEN 1
        ELSE 0
    END AS win_loss_numeric,
    1 AS is_home
FROM game
UNION ALL
SELECT
    wl_away AS win_loss,
    CASE
        WHEN wl_away = 'W' THEN 1
        ELSE 0
    END AS win_loss_numeric,
    0 AS is_home
FROM game
")

home_court_adv_model <- glm(win_loss_numeric ~ is_home, data = home_court_adv_query, family = binomial()

summary(home_court_adv_model)
```

```
##
## Call:
## glm(formula = win_loss_numeric ~ is_home, family = binomial(),
##     data = home_court_adv_query)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3887  -0.9798  -0.9798   0.9799   1.3887
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.484269   0.008033  -60.29   <2e-16 ***
## is_home      0.968410   0.011360   85.25   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 182154  on 131395  degrees of freedom
## Residual deviance: 174672  on 131394  degrees of freedom
## AIC: 174676
##
## Number of Fisher Scoring iterations: 4
```

The logistic regression model indicates that playing at home significantly increases the likelihood of winning, with the is_home coefficient (0.968410) being highly significant ($p < 2e\text{-}16$). The positive estimate suggests

that home teams have a strong advantage, as the odds of winning are substantially higher when playing at home compared to playing away.

Rebounding

```
rebounding_adv_query <- dbGetQuery(con, "SELECT game_id,
                    wl_home,
                    reb_home,
                    reb_away,
                    (reb_home - reb_away) AS rebound_adv,
                    CASE
                        WHEN wl_home = 'W' THEN 1
                        ELSE 0
                    END AS win_loss_numeric
                    FROM game
                    WHERE reb_home IS NOT NULL and reb_away IS NOT NULL")

rebounding_adv_model <- glm(win_loss_numeric ~ rebound_adv, data = rebounding_adv_query, family = binom

summary(rebounding_adv_model)
```

```
##
## Call:
## glm(formula = win_loss_numeric ~ rebound_adv, family = binomial(),
##     data = rebounding_adv_query)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5752  -1.1382   0.6593   0.9628   2.3734
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 0.350790   0.009793   35.82   <2e-16 ***
## rebound_adv 0.088731   0.001209   73.40   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 66983  on 49966  degrees of freedom
## Residual deviance: 60476  on 49965  degrees of freedom
## AIC: 60480
##
## Number of Fisher Scoring iterations: 3
```

The logistic regression model shows that rebound advantage significantly increases the likelihood of winning, with the `rebound_adv` coefficient (0.088731) being highly significant ($p < 2e\text{-}16$). The positive estimate indicates that as the rebound advantage increases, the odds of winning also increase, underscoring the importance of rebounding in determining game outcomes.

Three-pointers made

```
fg3m_adv_query <- dbGetQuery(con, "SELECT game_id,
                    wl_home,
```

```
                    fg3m_home,
                    fg3m_away,
                    (fg3m_home - fg3m_away) AS fg3m_adv,
                    CASE
                        WHEN wl_home = 'W' THEN 1
                        ELSE 0
                    END AS win_loss_numeric
                    FROM game
                    WHERE fg3m_home IS NOT NULL AND fg3m_away IS NOT NULL")

fg3m_adv_model <- glm(win_loss_numeric ~ fg3m_adv, data = fg3m_adv_query, family = binomial())

summary(fg3m_adv_model)
```

```
##
## Call:
## glm(formula = win_loss_numeric ~ fg3m_adv, family = binomial(),
##     data = fg3m_adv_query)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2647  -1.2546   0.7939   0.9929   1.9408
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 0.450802   0.009232   48.83   <2e-16 ***
## fg3m_adv    0.135572   0.002587   52.40   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 70312  on 52479  degrees of freedom
## Residual deviance: 67242  on 52478  degrees of freedom
## AIC: 67246
##
## Number of Fisher Scoring iterations: 4
```

The logistic regression model indicates that the number of three-pointers made advantage `fg3m_adv` significantly increases the likelihood of winning, with the `fg3m_adv` coefficient (0.135572) being highly significant ($p < 2e\text{-}16$). The positive estimate suggests that as the three-point made advantage increases, the odds of winning also increase, highlighting the critical role of three-point shooting in determining game outcomes.

Three-pointers attempted

```
library(DBI)
fg3a_adv_query <- dbGetQuery(con, "
SELECT game_id,
       wl_home,
       fg3a_home,
       fg3a_away,
       (fg3a_home - fg3a_away) AS fg3a_adv,
       CASE
```

```
           WHEN wl_home = 'W' THEN 1
           ELSE 0
       END AS win_loss_numeric
FROM game
WHERE fg3a_home IS NOT NULL AND fg3a_away IS NOT NULL
")

fg3a_adv_model <- glm(win_loss_numeric ~ fg3a_adv, data = fg3a_adv_query, family = binomial())

summary(fg3a_adv_model)
```

```
##
## Call:
## glm(formula = win_loss_numeric ~ fg3a_adv, family = binomial(),
##     data = fg3a_adv_query)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5468  -1.3468   0.9742   1.0083   1.2016
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.422512   0.009439  44.762   <2e-16 ***
## fg3a_adv    -0.010892   0.001211  -8.996   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 63128  on 47014  degrees of freedom
## Residual deviance: 63047  on 47013  degrees of freedom
## AIC: 63051
##
## Number of Fisher Scoring iterations: 4
```

The logistic regression model shows that the number of three-pointers attempted advantage `fg3a_adv` has a small but significant negative effect on the likelihood of winning, with the `fg3a_adv` coefficient (-0.010892) being highly significant ($p < 2e\text{-}16$). This negative estimate suggests that attempting more three-pointers than the opponent is slightly associated with a lower probability of winning, indicating that simply attempting more three-pointers without making them may not be an effective strategy for securing victories.

Data Extraction and Cleaning

```
library(ggcorrplot)
library(psych)
```

```
##
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
full_data_query <- "
SELECT
    game_id,
    reb_home - reb_away AS rebound_adv,
    tov_away - tov_home AS tov_adv,
    fg3a_home - fg3a_away AS fg3a_adv,
    fg3m_home - fg3m_away AS fg3m_adv,
    fg_pct_home - fg_pct_away AS fg_pct_adv,
    fg3_pct_home - fg3_pct_away AS fg3_pct_adv,
    ft_pct_home - ft_pct_away AS ft_pct_adv,
    pts_home - pts_away AS margin_of_victory
FROM
    game
WHERE
    reb_home IS NOT NULL AND reb_away IS NOT NULL
    AND tov_home IS NOT NULL AND tov_away IS NOT NULL
    AND fg3a_home IS NOT NULL AND fg3a_away IS NOT NULL
    AND fg3m_home IS NOT NULL AND fg3m_away IS NOT NULL
    AND fg_pct_home IS NOT NULL AND fg_pct_away IS NOT NULL
    AND fg3_pct_home IS NOT NULL AND fg3_pct_away IS NOT NULL
    AND ft_pct_home IS NOT NULL AND ft_pct_away IS NOT NULL
    AND pts_home IS NOT NULL AND pts_away IS NOT NULL
"

full_data <- dbGetQuery(con, full_data_query)

# Remove duplicates
full_data <- full_data %>% group_by(game_id) %>% slice(1) %>% ungroup()
```

Calculating Descriptive Stats

```
descriptive_stats <- describe(full_data[, c("margin_of_victory", "rebound_adv", "tov_adv", "fg3a_adv",
print(descriptive_stats)
```

```
##                   vars     n  mean    sd    min   max  range   se
## margin_of_victory    1 46316  3.32 13.38 -68.00 73.00 141.00 0.06
## rebound_adv          2 46316  1.59  9.16 -39.00 41.00  80.00 0.04
## tov_adv              3 46316  0.42  5.16 -22.00 23.00  45.00 0.02
## fg3a_adv             4 46316 -0.02  7.84 -42.00 44.00  86.00 0.04
## fg3m_adv             5 46316  0.13  4.09 -19.00 24.00  43.00 0.02
## fg_pct_adv           6 46316  0.01  0.08  -0.32  0.34   0.66 0.00
## fg3_pct_adv          7 46316  0.01  0.21  -1.00  1.00   2.00 0.00
## ft_pct_adv           8 46316  0.00  0.14  -0.63  0.78   1.41 0.00
```

The descriptive statistics reveal that while the mean advantages in various metrics are generally small, the standard deviations and ranges indicate significant variability across games. This suggests that while some metrics like rebounding and turnovers show consistent advantages for the home team, others like shooting percentages have more balanced distributions, emphasizing the unpredictable nature of basketball game outcomes.

These insights set the stage for further analysis using correlation and regression models to understand the impact of these metrics on the margin of victory.

Regression Model Summary

```r
# Fit a linear regression model
nba_model <- lm(margin_of_victory ~ rebound_adv + tov_adv + fg3a_adv + fg3m_adv + fg_pct_adv + fg3_pct_

model_summary <- summary(nba_model)

print(model_summary)
```

```
##
## Call:
## lm(formula = margin_of_victory ~ rebound_adv + tov_adv + fg3a_adv +
##     fg3m_adv + fg_pct_adv + fg3_pct_adv + ft_pct_adv, data = full_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.6030  -2.5980   0.0129   2.5865  24.1375
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.387507   0.018716  20.704   <2e-16 ***
## rebound_adv   0.596602   0.002194 271.914   <2e-16 ***
## tov_adv       1.167749   0.003766 310.087   <2e-16 ***
## fg3a_adv     -0.050470   0.004449 -11.343   <2e-16 ***
## fg3m_adv      0.952126   0.010165  93.671   <2e-16 ***
## fg_pct_adv  110.100663   0.265181 415.190   <2e-16 ***
## fg3_pct_adv  -0.175231   0.136461  -1.284    0.199
## ft_pct_adv   14.946370   0.131369 113.774   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.897 on 46308 degrees of freedom
## Multiple R-squared:  0.9152, Adjusted R-squared:  0.9151
## F-statistic: 7.136e+04 on 7 and 46308 DF,  p-value: < 2.2e-16
```

The linear regression model reveals significant factors influencing the margin of victory in basketball games.

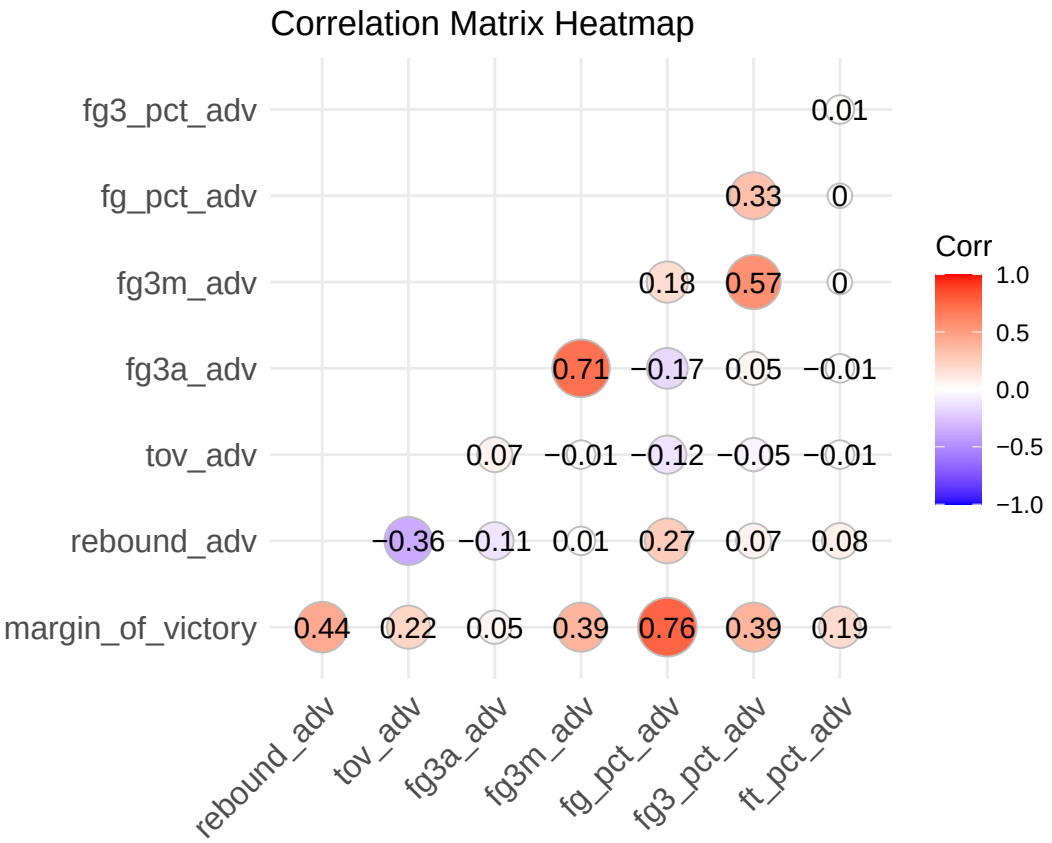
Key predictors include rebound advantage (Estimate: 0.596602), turnover advantage (Estimate: 1.167749), and three-point made advantage (Estimate: 0.952126), all positively affecting the margin. Conversely, three-point attempts advantage has a slight negative effect (Estimate: -0.050470). Field goal percentage advantage (Estimate: 110.100663) is the strongest predictor, emphasizing shooting efficiency's critical role. Free throw percentage advantage (Estimate: 14.946370) also significantly impacts the margin.

The model's high R-squared value (0.9152) indicates these variables explain most of the variance in game outcomes, highlighting their importance in determining victory.

Correlation Heat Map

```r
cormatrix <- cor(full_data[, c("margin_of_victory", "rebound_adv", "tov_adv", "fg3a_adv",
                               "fg3m_adv", "fg_pct_adv", "fg3_pct_adv", "ft_pct_adv")], use = "complete

ggcorrplot(cormatrix,
           method = "circle",
           type = "lower",
           lab = TRUE,
           title = "Correlation Matrix Heatmap")
```



The correlation matrix heat map reveals that field goal percentage advantage (0.76) and three-point made advantage (0.71) are the strongest predictors of margin of victory, highlighting the importance of shooting efficiency. Turnover advantage (0.39) and rebound advantage (0.44) also positively impact the margin, indicating that controlling turnovers and rebounds is crucial.

Three-point attempt advantage shows a slight negative correlation (-0.05), suggesting that simply attempting more three-pointers is ineffective. But looking at the modern landscape of the NBA, how could this be the case?

Average Three-Point Attempts Per Game by Year

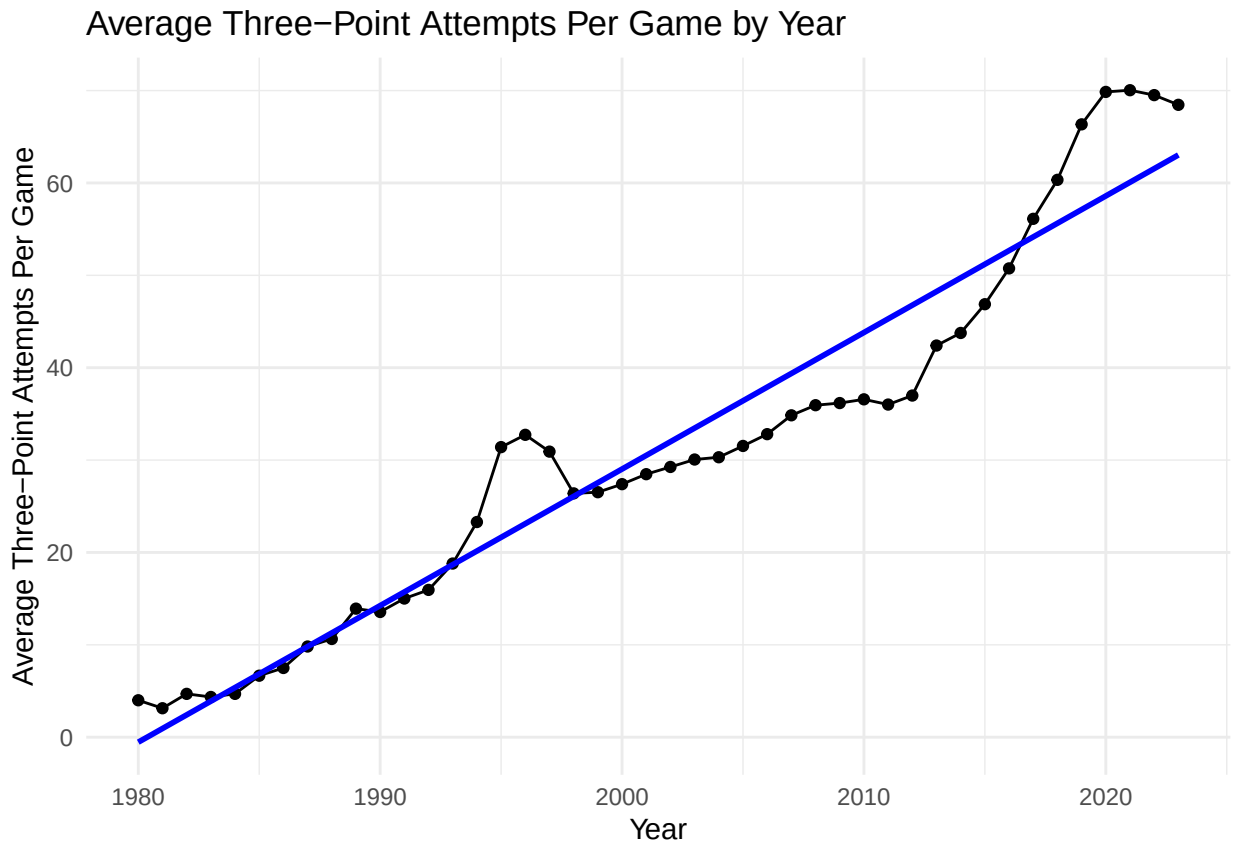
```
fg3a_per_game_query <- "
SELECT
    strftime('%Y', game_date) AS year,
    AVG(fg3a_home + fg3a_away) AS fg3a_per_game
FROM
    game
WHERE
    fg3a_home IS NOT NULL
    AND fg3a_away IS NOT NULL
    AND strftime('%Y', game_date) > '1979'
GROUP BY
    year
"

fg3a_per_game <- dbGetQuery(con, fg3a_per_game_query)
```

```r
fg3a_per_game$year <- as.numeric(fg3a_per_game$year)

# Plot the data
ggplot(fg3a_per_game, aes(x = year, y = fg3a_per_game)) +
    geom_line() +
    geom_point() +
    geom_smooth(method = "lm", se = FALSE, color = "blue") +
    labs(title = "Average Three-Point Attempts Per Game by Year",
         x = "Year",
         y = "Average Three-Point Attempts Per Game") +
    theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



As the plot vividly illustrates, the NBA has experienced a dramatic rise in the average number of three-point attempts per game over the years.

This apparent contradiction invites a deeper reflection on the evolving strategies of the game. Why has there been such a pronounced shift towards more three-point attempts despite their questionable direct impact on winning? This question underscores the necessity for a comprehensive analysis of all relevant statistics over the years.

By examining factors such as field goal percentage, turnover advantage, and rebounding, we can better understand the complex interplay of these elements and how they have influenced the game's outcomes in the context of an ever-evolving strategic landscape. This holistic analysis can provide valuable insights into the effectiveness of contemporary basketball strategies and inform future approaches to the game.