Introduction

As big data and analytics continue to become more and more relevant in today's world, businesses are constantly searching for ways to utilize data to get ahead, and professional sports are no exception. According to Forbes, "the market for sports analytics is expected to reach $4 billion by 2022". Data analysis is being used across a variety of sports not only to improve winning percentage, but also to measure and drive customer engagement and expand corporate partnerships.

The National Basketball Association (NBA) generates over $7 billion in revenue annually, but even still not all teams are profitable. In fact, nine teams were operating at a loss in 2017. Like many businesses, NBA teams are investing in and using data and analytics to maximize efficiency and improve team performance, which has even garnered attention from academic institutions such as MIT and Harvard.

Because the objective of any sports team is to win, the use of data and modeling to predict the outcome of a game is perhaps the most obvious, but also has the potential to be the most useful. Over the past few decades, there have been numerous studies and methodologies using NBA game statistics to predict wins for teams. The methodologies vary from being straightforward and intuitive to complex and unique, demonstrating the ample opportunity that exists and the sustained relevancy of sports analysis over time.

Statistical Testing and Parametric Models

The data set used in this analysis contained NBA game data from 2014-2018 seasons with a number of in-game statistics for teams and their opponents. The target variable is "WINorLOSS", our objective being to build a model that would predict the categorical outcome of the game (win or loss). Some of the variables considered as predictors include: Home versus Away, Field Goals (2-point shots) and attempts, 3-point shots and attempts, made free throw percentage, assists, rebounds, blocks, steals, turnovers, fouls, and the same variables for the opponent. Most of the predictor variables are categorical in nature. In preparing the data to build a model, columns "X", "Team", "Game", "Date", and "Opponent" were all excluded as they did not contain information helpful in predicting our target (wins versus losses).

As part of our analysis, we created contingency tables that reflected win probabilities for home teams versus away teams (win and home, win and away, loss and home, loss and away) and ran a chi-square test to see if the probabilities were statistically significant. (Figure 1) In this case, the p-value was less than .05 so we could reasonably conclude that our two categorical variables "WINorLOSS" and "Home versus Away" were statistically related with each other. The most appropriate model type to use for the purpose of this analysis was the multiple logistic regression, because our target variable is categorical and binary in nature. To build the model, VIF analysis was used in order to eliminate variables one by one until all of the variables included were significantly significant. The first variable excluded with the highest VIF value was Opponent Field Goals, and a total of ten variables were excluded before all VIF values were less than 5. The variables were then assessed based on p-values and two more were eliminated based on having a p-value greater than .05 (Opponent blocks and steals). (Figure 2). For the purposes of comparison, the next step was to also create a model using stepwise regression.

(Figure 3). In comparing the results of Model 1 (using VIF analysis and p-value) and Model 2 (stepwise regression model), the variable with the strongest influence on winning remains constant- "FieldGoals". Each additional field goal for a team positively impacts the log odds of winning by 44.06. Similarly, assists also positively impact the log odds of winning, though to a lesser degree. Opposing team assists decrease the log odds of winning a game by .27. In comparing the AIC and BIC between Model 1 and Model 2, the stepwise regression model obtains a smaller value than Model 1. (Figure 4). However, because Model 2 does not include the "Home versus Away" variable, and earlier analysis already concluded this variable to be statistically significant, we decided that Model 1 is a better model.

Conclusions

Due to the model's relatively high accuracy (87%), it is reasonable to conclude this model can be useful for predicting wins and losses of NBA games. However, many of the observations drawn from the analysis are more or less intuitive to the average fan of basketball. For example, field goals being an important contributor to wins is logical, because more field goals made would mean a higher score, and the key to winning a basketball game (or any team sport for that matter) is to outscore the opponent. Another example would be assists, which have a smaller but still positive effect on winning probability, while opponent assists have a negative impact. Again, this makes sense, but perhaps a less obvious conclusion here is that due to the negative impact of an opponent assist being the stronger of the two, indications are that defense is more important than offense in this regard. By the similar logic, it is also not surprising that made 3-point shots by the opposing teams negatively impact a team's probability of winning, as more points for opponent make it increasingly difficult for a team to "keep up" and outscore the other team.

While this analysis may not be considered groundbreaking, it does reflect and validate the strategies NBA teams use today to try and win. For example, the observation regarding opposing teams 3-pointers correlates to the emphasis teams have on defending the 3-point shot, and also may explain why 3-pointers are attempted more in games now than they were ten years ago. Next steps for this analysis could be to focus on and expand upon the 3-point shot by getting a more detailed data set with locations of 3-pointers as well as which defensive schemes are most and least effective at defending a 3-pointer. This could help teams strengthen their defenses against the 3-pointer while simultaneously growing more efficient in 3-point shots taken, both of which would increase probability of winning, according to the model results.

Appendix

Figure 1

```
wlah_table <- prop.table(table(nba_raw_data$Home, nba_raw_data$WINorLOSS))
addmargins((wlah_table))
```

```
             L         W       Sum
Away 0.290752 0.209248 0.500000
Home 0.209248 0.290752 0.500000
Sum  0.500000 0.500000 1.000000
```

```
        Pearson's Chi-squared test

data:  nba_raw_data$WINorLOSS and nba_raw_data$Home
X-squared = 261.47, df = 1, p-value < 2.2e-16
```

Figure 2

```
#Based on VIF analysis, time to eliminate variables having VIF values greater than 5 one by one

fit_for_vif <- update(fit_for_vif,.~.-Opp.FieldGoals)
fit_for_vif <- update(fit_for_vif,.~.-FieldGoals)
fit_for_vif <- update(fit_for_vif,.~.-OpponentPoints)
fit_for_vif <- update(fit_for_vif,.~.-TeamPoints)
fit_for_vif <- update(fit_for_vif,.~.-Opp.FieldGoals.)
fit_for_vif <- update(fit_for_vif,.~.-Opp.3PointShots)
fit_for_vif <- update(fit_for_vif,.~.-X3PointShots)
fit_for_vif <- update(fit_for_vif,.~.-Opp.FieldGoalsAttempted)
fit_for_vif <- update(fit_for_vif,.~.-Opp.TotalRebounds)
fit_for_vif <- update(fit_for_vif,.~.-TotalRebounds)

vif(fit_for_vif)
summary(fit_for_vif)

#Lets eliminate variables having p-value greater than 0.05(Assuming 95% confidence)
fit_for_p <- update(fit_for_vif,.~.-Opp.Blocks)
fit_for_p <- update(fit_for_p,.~.-Steals)

summary(fit_for_p)
```

```
Call:
glm(formula = nba_data$WINorLOSS ~ Home + FieldGoalsAttempted +
    FieldGoals. + X3PointShotsAttempted + X3PointShots. + FreeThrowsAttempted +
    FreeThrows. + OffRebounds + Assists + Blocks + Turnovers +
    TotalFouls + Opp.3PointShotsAttempted + Opp.3PointShots. +
    Opp.FreeThrowsAttempted + Opp.FreeThrows. + Opp.OffRebounds +
    Opp.Assists + Opp.Steals + Opp.Turnovers + Opp.TotalFouls,
    family = binomial(link = "logit"), data = nba_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.4090  -0.3494  -0.0001   0.3549   3.2700

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -18.291116   0.875642 -20.889  < 2e-16 ***
HomeHome                  0.057187   0.069371   0.824   0.4097
FieldGoalsAttempted       0.035319   0.007387   4.781 1.74e-06 ***
FieldGoals.              44.063771   1.225381  35.959  < 2e-16 ***
X3PointShotsAttempted     0.067838   0.005408  12.543  < 2e-16 ***
X3PointShots.             6.725267   0.434758  15.469  < 2e-16 ***
FreeThrowsAttempted       0.121022   0.008571  14.120  < 2e-16 ***
FreeThrows.               5.970990   0.354521  16.842  < 2e-16 ***
OffRebounds               0.187867   0.012550  14.969  < 2e-16 ***
Assists                   0.061207   0.009312   6.573 4.93e-11 ***
Blocks                    0.198555   0.014541  13.655  < 2e-16 ***
Turnovers                -0.156837   0.015591 -10.059  < 2e-16 ***
TotalFouls               -0.087234   0.014076  -6.197 5.75e-10 ***
Opp.3PointShotsAttempted -0.012534   0.005090  -2.463   0.0138 *
Opp.3PointShots.         -14.268340   0.475298 -30.020  < 2e-16 ***
Opp.FreeThrowsAttempted   -0.129242   0.008444 -15.306  < 2e-16 ***
Opp.FreeThrows.           -6.278128   0.355388 -17.666  < 2e-16 ***
Opp.OffRebounds           -0.109036   0.009731 -11.205  < 2e-16 ***
Opp.Assists               -0.274905   0.009555 -28.771  < 2e-16 ***
Opp.Steals                -0.079730   0.018383  -4.337 1.44e-05 ***
Opp.Turnovers              0.208333   0.010252  20.322  < 2e-16 ***
Opp.TotalFouls             0.077212   0.013751   5.615 1.97e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 3

```
Call:
glm(formula = nba_data$WINorLOSS ~ FieldGoalsAttempted + FieldGoals. +
    X3PointShotsAttempted + X3PointShots. + FreeThrowsAttempted +
    FreeThrows. + OffRebounds + Assists + Blocks + Turnovers +
    TotalFouls + Opp.3PointShotsAttempted + Opp.3PointShots. +
    Opp.FreeThrowsAttempted + Opp.FreeThrows. + Opp.OffRebounds +
    Opp.Assists + Opp.Steals + Opp.Turnovers + Opp.TotalFouls,
    family = binomial(link = "logit"), data = nba_data)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-3.405  -0.348   0.000   0.354   3.264

Coefficients:
                         Estimate Std. Error z value             Pr(>|z|)
(Intercept)             -18.26314    0.87481  -20.88 < 0.0000000000000002 ***
FieldGoalsAttempted       0.03527    0.00739    4.77        0.000001805753 ***
FieldGoals.              44.08978    1.22501   35.99 < 0.0000000000000002 ***
X3PointShotsAttempted     0.06786    0.00541   12.55 < 0.0000000000000002 ***
X3PointShots.             6.73082    0.43484   15.48 < 0.0000000000000002 ***
FreeThrowsAttempted       0.12116    0.00857   14.14 < 0.0000000000000002 ***
FreeThrows.               5.96986    0.35438   16.85 < 0.0000000000000002 ***
OffRebounds               0.18837    0.01253   15.03 < 0.0000000000000002 ***
Assists                   0.06186    0.00928    6.67        0.000000000026 ***
Blocks                    0.19935    0.01451   13.74 < 0.0000000000000002 ***
Turnovers                -0.15728    0.01558  -10.09 < 0.0000000000000002 ***
TotalFouls               -0.08797    0.01405   -6.26        0.000000000379 ***
Opp.3PointShotsAttempted -0.01251    0.00509   -2.46                 0.014 *
Opp.3PointShots.         -14.27967    0.47519  -30.05 < 0.0000000000000002 ***
Opp.FreeThrowsAttempted  -0.12942    0.00844  -15.33 < 0.0000000000000002 ***
Opp.FreeThrows.          -6.28212    0.35534  -17.68 < 0.0000000000000002 ***
Opp.OffRebounds          -0.10972    0.00970  -11.31 < 0.0000000000000002 ***
Opp.Assists              -0.27568    0.00951  -28.99 < 0.0000000000000002 ***
Opp.Steals               -0.07932    0.01837   -4.32        0.000015761275 ***
Opp.Turnovers             0.20844    0.01025   20.33 < 0.0000000000000002 ***
Opp.TotalFouls            0.07799    0.01372    5.69        0.000000013031 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 4

	df <dbl>	AIC <dbl>
fit_for_p	22	5606
stepback_full	21	5605

2 rows

	df <dbl>	BIC <dbl>
fit_for_p	22	5765
stepback_full	21	5756

2 rows