IDS 702 Final Project Report

How to Win NBA Betting

Leon Zhang

Due date: Nov 24th 2020

1 Summary

Betting has become a hot topic in NBA. The overall goal of this project is to explore the best set of predictors a bettor should consider when placing moneyline bets. The report answers three major questions: Can we quantify the effect of winning/losing streak on a team's chance of winning? To what degree does "back to back" games influence a team's winning outcome? What in-game statistics affects a team's odds of winning the most?

A multi-level (NBA teams) logistic regression model was built to explore the relationship between predictors and team performance. It was found that back to back games significantly lowers the chance of winning.Streak is not that of a big win condition. Home court advantage is a huge factor and assist, turnover, record are pretty useful predictors as well.

In later sections, each of the points above is discussed in detail. In addition, the report also talks about the rational of EDA, model building, selection and limitations.

2 Introduction

In 2019, the National Basketball Associations, NBA, announced to legalize sports betting across United States. This brings the field of data science with tremendous opportunities in the sports betting market.

There are many advantages in using statistical modeling in sporting betting because it offers a quantitative insight of expected returns. For this project, we will find out what factors is worth taking into considerations when placing NBA bets.

There are three common types on bet: moneyline: which team wins the game, spread: how many points a team will win or lose by, and totals (over/under): the total points scored by both teams. To narrow down the scope, this analysis will be only focusing on the moneyline bets.

The overall goal of this project is to explore the best set of predictors a bettor should consider when placing moneyline bets.Given the scope of this project, I want to answer three major questions:

1. Can we quantify the effect of winning/losing streak on a team's chance of winning?
2. To what degree does "back to back" games influence a team's winning outcome?
3. What in-game statistics affects a team's odds of winning the most?

3 Data

Since we are doing pregame betting, we want our predictors to be anything we know before the game starts. It was difficult to find a data set that includes all the information I desire, so I created my own data set from multiple sources. The dataset consists of nearly 10k rows that records every game for every NBA team from 2014 to 2018. The final columns of the dataset are listed with

description below:

Game: The i th game of a particular season
Team: The team which played the game
Opponent: The opponent which the team played against
Date: Date of the game which was played
Home: Is the game is played on the team's home court? 1 for yes
is_b2b: Is this game a back to back game for the team? 1 for yes
is_b2b_opp: Is this game a back to back game for the team's opponent? 1 for yes
streak: Current winning/losing streak of the team. >0 for winning; <0 for losing.
opp_streak: Current winning/losing streak of the opponent. >0 for winning; <0 for losing.
record: Total wins - total loses of the team.
opp_record: Total wins - total loses of the opponent.
others: All the average game stats for each team, such as total score, rebound, assist, etc. from the past games in the season.

For EDA, I explored the effect of each predictor against the response variable. The most notable finding was that "home court", "total points", "defensive rebound". "assist", "records", and streak has a strong correlation with winning. I also found that "back to back" and "turnover" has a strong correlation with losing. This all make intuitive sense and suggest we should continue exploring these predictors in our model. The promising continuous predictors are plotted below in figure 1 and the promising categorical predictors are shown in the contingency tables.

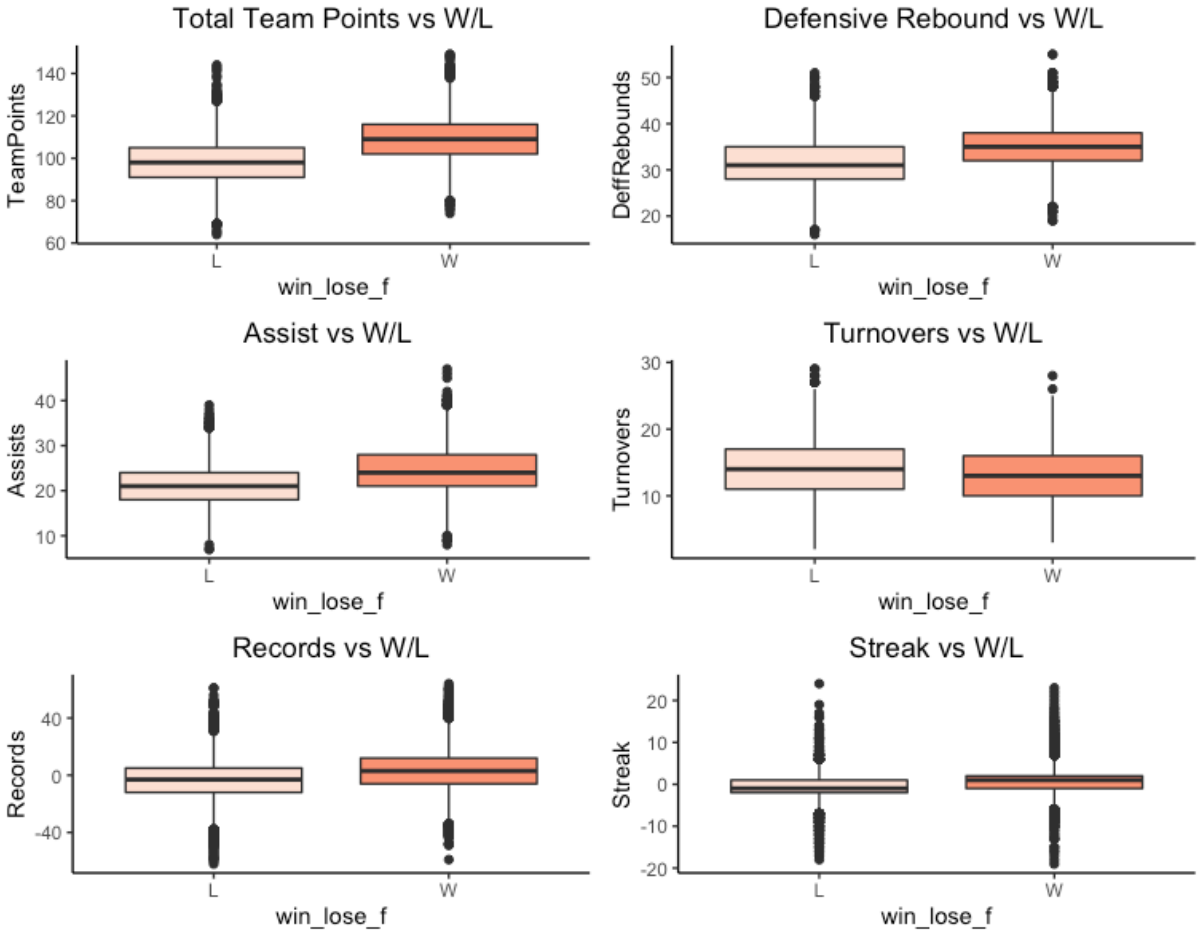

Figure 1: Promising Predictors Discovered vs Win/Lose

```
Is Back to Back Game vs Win/Lose

            is_b2b_f
win_lose_f          0        1
         L  0.4850256 0.556546
         W  0.5149744 0.443454


Home Game vs Win/Lose

            home_f
win_lose_f       Away     Home
         L  0.5838926 0.415837
         W  0.4161074 0.584163
```

After checking multi-collinearity I found out that total points, field goal and field goal percent are highly correlated so I decided to only use total points among the three. Finally, to reduce the number of predictors for each statistics in modeling I took the difference between the two teams, this also does some sort of mean centering to the data.

4 Model

I built a multi-level logistic regression model. The levels are the different teams in the NBA. I used AIC, BIC and stepwise metric to select the best model and the final model is shown below:

```
Final.Model <- glmer(win_lose_f~
                     Cumu.TeamPoints.Diff+
                     Cumu.Assists.Diff+
                     Cumu.DeffRebounds.Diff+
                     Cumu.Turnovers.Diff+
                     Cumu.Streak.Diff+
                     Cumu.Records.Diff+
                     home_f+
                     is_b2b_f+
                     is_b2b_opp_f+
                     (1|team_f)
                     ,family=binomial(link=logit),data=nba_subset)
```

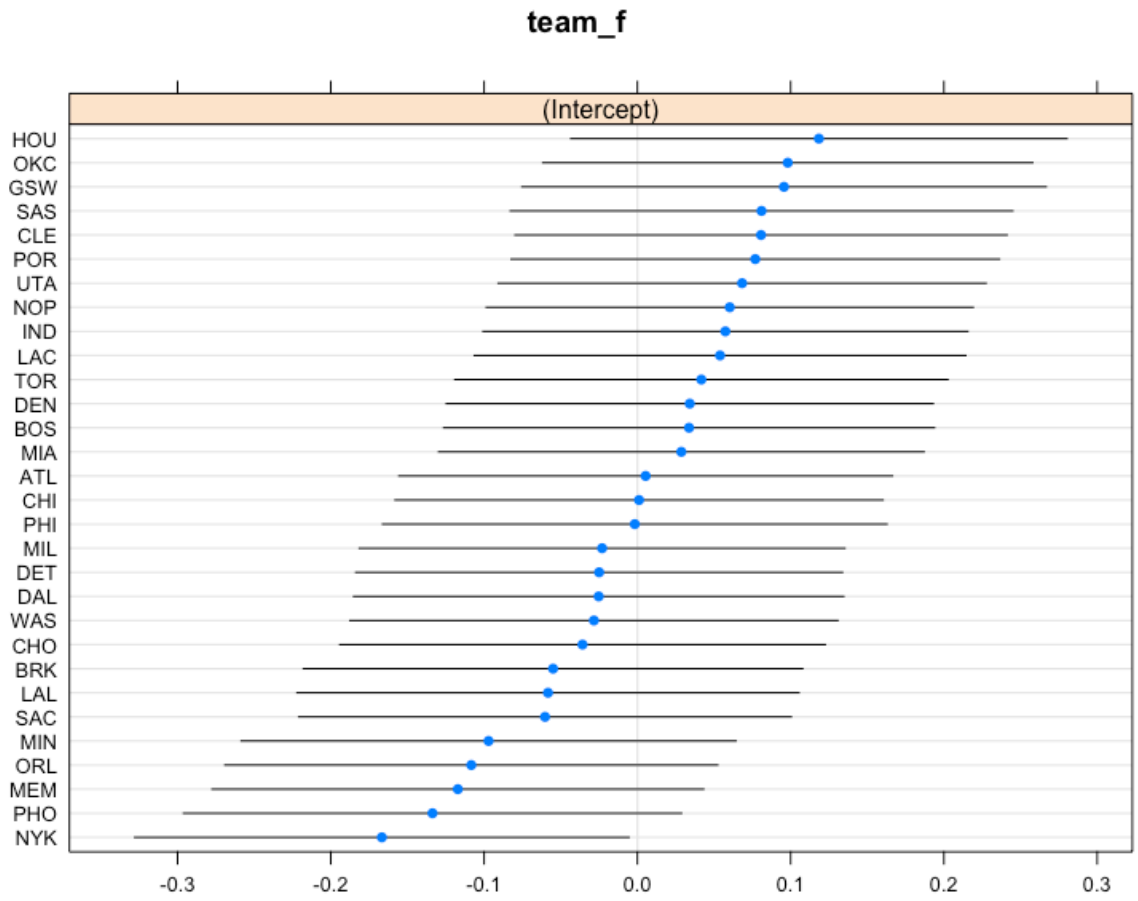

Figure 2: Dot Plot of all the NBA Teams from the Final Model

Summary of the Final Model

	Dependent variable:
	Win or Lose
Cumu.TeamPoints.Diff	0.007
	(0.009)
Cumu.Assists.Diff	0.034**
	(0.013)
Cumu.DeffRebounds.Diff	0.024***
	(0.006)
Cumu.Turnovers.Diff	-0.043**
	(0.017)
Cumu.Streak.Diff	0.022***
	(0.006)
Cumu.Records.Diff	0.030***
	(0.002)
home_fHome	0.732***
	(0.047)
is_b2b_f1	-0.188***
	(0.057)
is_b2b_opp_f1	0.183***
	(0.057)
Constant	-0.364***

```
                                  (0.042)
-------------------------------------------------
Observations                       9,240
Log Likelihood                  -5,660.890
Akaike Inf. Crit.               11,343.780
Bayesian Inf. Crit.             11,422.230
=================================================
Note:                  *p<0.1; **p<0.05; ***p<0.01
```

In addition, I checked the model assumptions: We are confident that our model agrees with linearity, independence and normality assumptions as the points are scattered in the residuals vs variable plot and the line is straight in the residuals vs fitted and QQ-plot.

5 Conclusions

First of all, the dot plot of teams is interesting to look at. It is surprisingly accurate we can see the top 5 teams from 2014-18 are rockets, thunder, warriors, spurs and cavaliers. They were indeed very good teams. And if we look at the bad teams: Knicks, suns, grizzlies, magic, and Timberwolves were pretty bad during that time.

To answer the inference questions, I place the model summary in the model summary section for reference.

For the first question, quantify the effect of streak on a chance of winning. If we interpret the coefficient of streak difference, we can see every increment more than the opponent will increase the chance of winning by 2%, with a p-value of less than 0.05 meaning the predictor is significant at 95 percent confidence level.

For the second question, to what degree does "back-to-back" games influence a team's winning outcome. I found that playing a back to back game decreases the chance of winning by 30.7%. Although is p-value is 0.057, which is above 0.05, is is still convincing that back-to-back games have a great influence on the team's performance.

Finally, to find what game statistics affect a team's of winning the most, I discovered that every single assist earned more than opponent increases chance of winning by 3%. Turnover decrease by 4%, record increases by 3% and home court advantage increases more than 100%.

I would like to acknowledge that this study exists few limitations. First of all, Although I took the difference between team and opponent team, which does some degree of mean centering, I did not normalize the predictors, which may cause some issue since some statistics are such as team points tends to have larger value that others such as assists. Secondly, there is symmetry in the dataset meaning there are two games representing the same information just from the opponent's perspective. This will lead to redundancies in the dataset. And lastly, in future I wish to build a betting bot from this study. It would be better if I take into the return of a bet into considerations. To make this better, I change by response variable into prediction probability * return.