MDML Final Project Written Report

Frank Jiang, Diana Liang, Sue Shen

12/9/2019

#Executive Summary

In our project, we use data that are scraped from basketball reference websites and pro sports transactions joining the player data from the Kaggle. After performing regression analysis for win and loss rates associated with home or away games, we can carefully conclude that a team has a higher win rate at home games. Surprisingly, we found out that the attendance of the audience doesn't affect the team's winning rate. This might due to the team's rules and regulations and professional contract. On a team level, offensive wise, we can conclude that a team performs better at home games with higher field goal percentage, higher 3 pointer percentage, higher true shooting percentage, and higher efficiency field goal percentage. Defensively, in general, a team performs significantly better at home game too with less personal fouls, higher total steals, and higher total blocks. Since rebound is considered more important in the game basketball, we look at it separately and discovered that a team averages a higher total of both offensive and defensive rebounds in a home game than away game. Also, starters seem to play slightly longer at home games than away games. We also fit the logistic regression model with machine learning to predict the winning rate for home/away games. The prediction indicates that a team potentially has a higher winning rate at home game than away game. We also fit a random forest model with machine learning using an injury report to predict player injury. Our model achieves an accuracy of 70% when the threshold equals 0.5. The importance plot shows that minutes of playing time on average before this game, average offensive ratings of the player before this game, and the average points before this game are the top three most important factors that can predict player's injury.

#Introduction

After watching last year's NBA's final, with all the injury that Golden state warrior is suffering, the game seems to become less exciting than it should be. Even so, the finals are always controversial with lots of complaining about home advantage. Therefore, when we first start this machine learning project, We become interested in discovering if home advantage exists. More specifically, how does home advantage impact the way a team performs. And also if there is a way that we could predict a player's injury or even analyze a little bit what could be correlated to a player's injury with the help of machine learning. Nowadays, the NBA collects a large amount of data for every game every year in a very detailed manner. This provides us the data to analyze and train a model for machine learning. 'Preventing in-game injuries for NBA players' by Talukder, Vincent, Foster, etc inspires us on how we could potentially use specific game data into constructing the injury prediction model. The final goal is to potentially give some advice to the General Manager of an NBA team about how to avoid player injury. In general, we focus our research on two aspects. One, the impact of home advantage to win rate of a game. Is there any difference in winning probability between home matches and away matches? If there is, what are the underlying factors influencing team performance? Two, develop a model to predict a player's injury. The research goal is to predict an individual player's risk of injury and identify impactful factors increasing the risk of injury.

#Method

##Data Scraping

For our home/away game analysis, we need to collect data that includes specific in-game statistics by team, matched with each game's result. The outcome of NBA stats is the win or loss of the match and indicators include whether home or away, number of audiences, distance from home city, field goal %, three-point %, rebounds, assists and other team stats in the match. There will be at least 2 season stats scaped using 'rvest' in R. Data are parsed from *https://www.basketball-reference.com/leagues/NBA_2018_games-october.html* to summarise game results and the number of the audience attended to discover if there exists a correlation between these two. Also, we scape data from *https://www.basketball-reference.com/boxscores/*

201910220LAC.html including both basic and advanced statistics for each game, for the visitor team and home team. For each match, there are four tables. Therefore, we join the data into a two-level data table matching the basic and advanced statistics with game results. To avoid the self-correlation between win and loss of different teams that are playing against each other, we treat each game results independently to solve the issue of 'game theory'. For our injury model and prediction, we parsed data from *https://www.prosportstransactions.com/basketball/Search/Search.php* for exploratory research. We also scraped data from *https://www.kaggle.com/ghopkins/nba-injuries-2010-2018/version/1* to join with previous game statistics to serve as training and testing dataset for our model construction and prediction.

##Analysis

For the home and away game's difference in winning probability and the possible reasons underlying, different analysis is performed using the dataset collected. For the data to be able to be fitted into different models, the data are transformed from the basic and advanced statistics collected previously to team stats and join it with the team game result stats to create a dataset with relevant variables. Two models are fitted both for home matches and away matches respectively. Specifically, in the model for home matches, we will include all the home matches for all the NBA teams, using win or loss as the outcome and team stats for the home team as well as the number of audience as the independent indicators.

$$Pr(Y_H = 1| \,..) = logit^{-1}(\beta_{H1}X_1 + ... + \beta_{Hk}X_k + \beta_{HN}NumAudience + \epsilon_H),\ \epsilon_H \sim N(0, \sigma_H^2),\ indep.$$

While in the model for away matches, we will include all the away matches for all the NBA teams, using win or loss as the outcome and team stats for the away team as well as the distance from home city, number of audience as the independent indicators.

$$Pr(Y_A = 1| \,..) = logit^{-1}(\beta_{A1}X_1 + ... + \beta_{Ak}X_k + \beta_{AN}NumAudience + \beta_{AD}Distance + \epsilon_A),\ \epsilon_A \sim N(0, \sigma_A^2)\ indep.$$

Model selection and benefits will be discussed later in this writeup.

Some exploratory research is done to help understand the relationship between different factors and winning rates. Based on the results of those exploratory results, We can analyze the difference of winning rates in home-game by fitting the number of audiences attended, field goal percentage, 3 pointer field goal percentage, total offensive rebound, total defensive rebound, total personal fouls, total steals, total blocks, defensive ratings, and average playing time of the starters into the home game model. And all the above variables plus the distance into the away game model. Logistic regression is chosen as the method for machine learning and prediction. AUC will be calculated to evaluate the accuracy of the model. Visualization is created to analyze the predicted probability of winning and losing in a home or away game using our model.

The coefficient of the model shows whether the distance and the number of audiences can impact the winning probability. We also run a T-test for further investigation on the difference in team stats between home and away matches such as field goal percentage, three-point percentage, rebounds, assists. Lastly, we plotted the predicted probability of winning to decide if the game outcome changes when a player plays away by comparing the slope.

For our injury prediction analysis, specific in-game data and player injury are joined by players, making each row represent a player in a specific game. For each player in each game, the following variables are created: Mean minutes on court in corresponding season before this match excluding absence, Mean of points and other stats before this match, Total matches attended before this match, Minutes on court in last five matches respectively, Mean of points and other stats in last five matches respectively, Home or away for last five matches respectively, Home or away for this match, Ranking of the opponent for this match and Whether got injury or not. Data are split into a training set and testing set. Training sets are used with random forest to build the prediction model. Then, testing data are used to predict individual player's injuries. An importance plot is constructed similarly to the 'Preventing in-game injuries for NBA players' paper by Talukder, Vincent, Foster suggested. Influencing factors are identified by considering p-value and standardized coefficients.

##Model Selection There are two main benefits of developing two models for home and away matches respectively instead of using the home or away as a dummy variable. Firstly, the result of a match for two teams are paired, one wins, then the other loses, which will violate the assumptions of independent epsilons in regression. So with two models, we can meet the assumption that both models have independent errors. Secondly, Using home or away as a dummy variable means the model only allows difference in intercepts, but the difference in slopes is also likely to exist. Using two models, coefficients are more flexible. Then we will split data to training and testing datasets, fit models using logistic regression as stated above and calculate accuracy, precision, recall, and AUC to measure the performance of models.

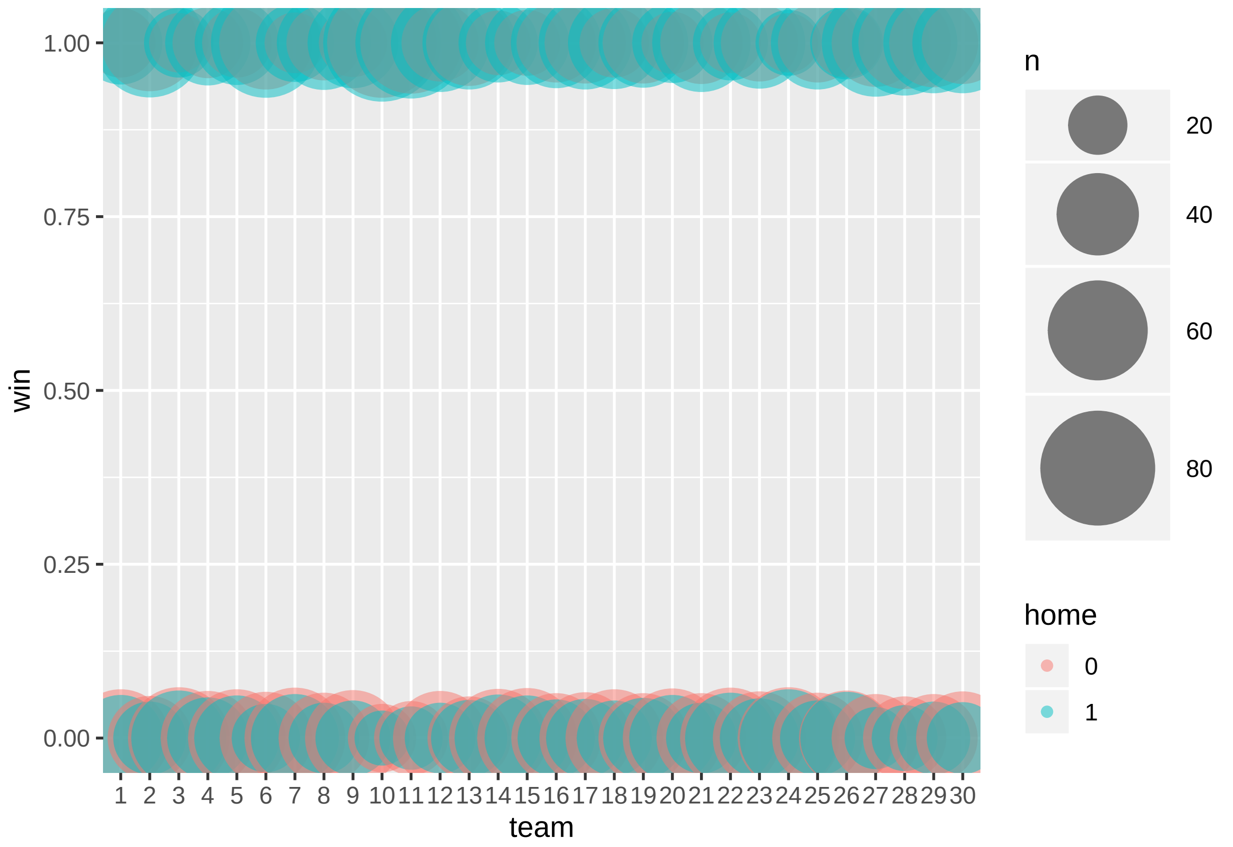
There are two main reasons to choose logistic regression as the method for machine learning to predict winning probability. One is the fact that we are fitting the data from 2018-2019 to predict and explain the winning probability of different teams on an individual home/away game in 2018-2019. So using logistic regression to fit this model will have a generally high AUC(accuracy under the curve) score. This implies that the model is accurate enough to explain the underlying reasons for a possible higher winning rate at a home game. The other reason is that logistic regression is easier to interpret. There exist a coefficient and p-value to each variable that is used to fit the model. Therefore, it is easier to understand the impact of different variables on the winning rate at home/away games.

For the player injury data, we choose to apply a random forest to fit the model for prediction. The advantage is that random forest normally guarantees a relatively high AUC(area under the curve) score, which allows us to have higher accuracy for our future prediction on player's injury. However, it creates a problem. Random forest is not a straightforward model for interpretation. It is hard to understand and explain the impact of different factors/variables that are used to fit the model. Therefore, an important plot is created to understand how each variable contributes to the prediction of a player's injury. The importance plot uses p-value and standardized coefficient which compensates the issue of interpretability of using the random forest model.

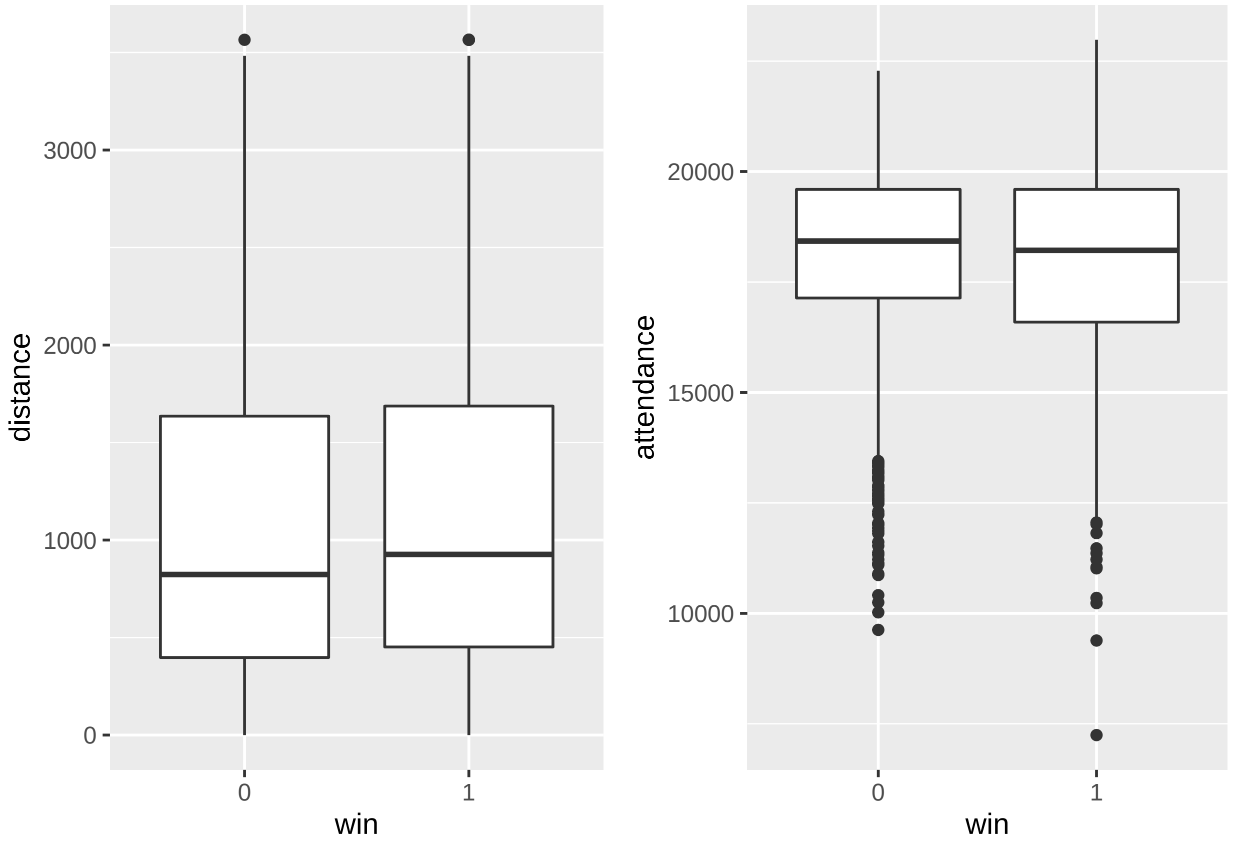
#Results

##Home and away matches the difference in winning probability and reasons

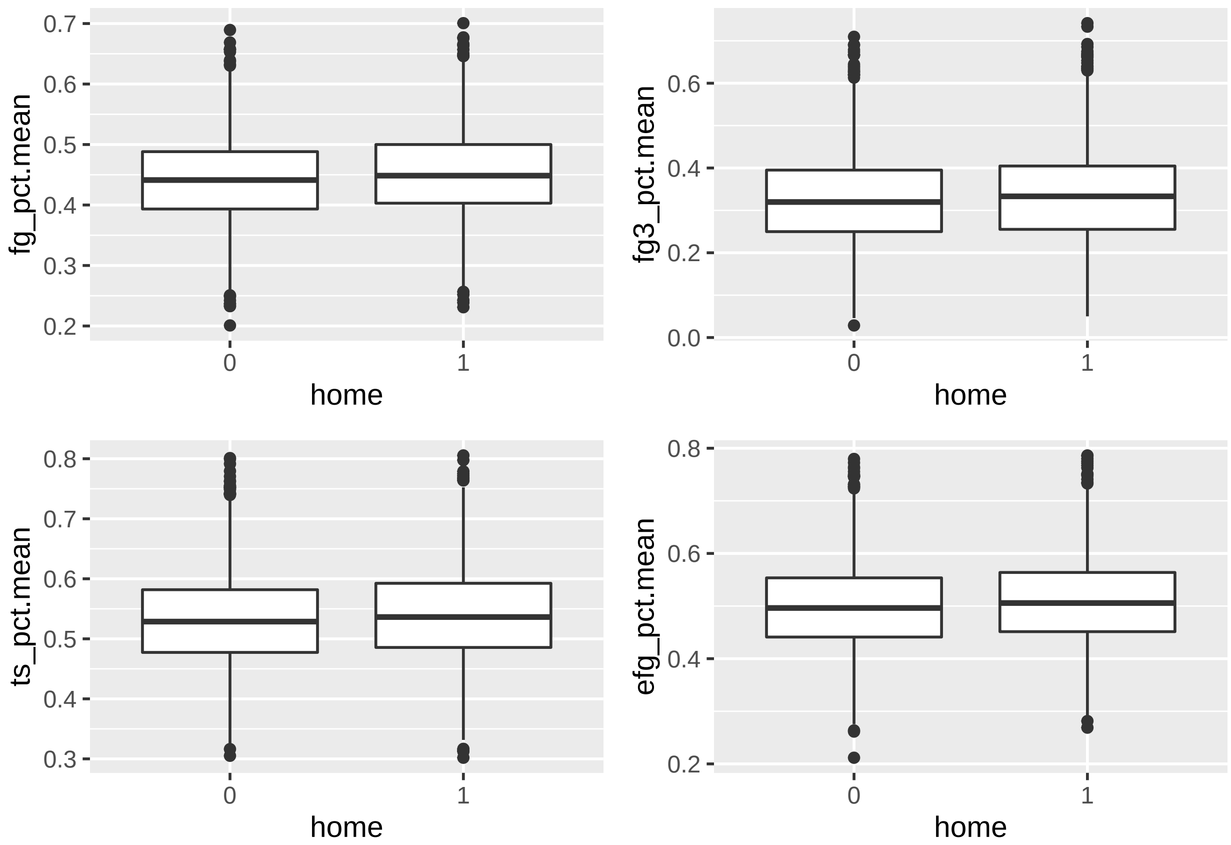
Several explortory research was conducted along with visualization to roughly understand the difference of home/away games in winning rates and variables that possibly reveals the underlying reasons.



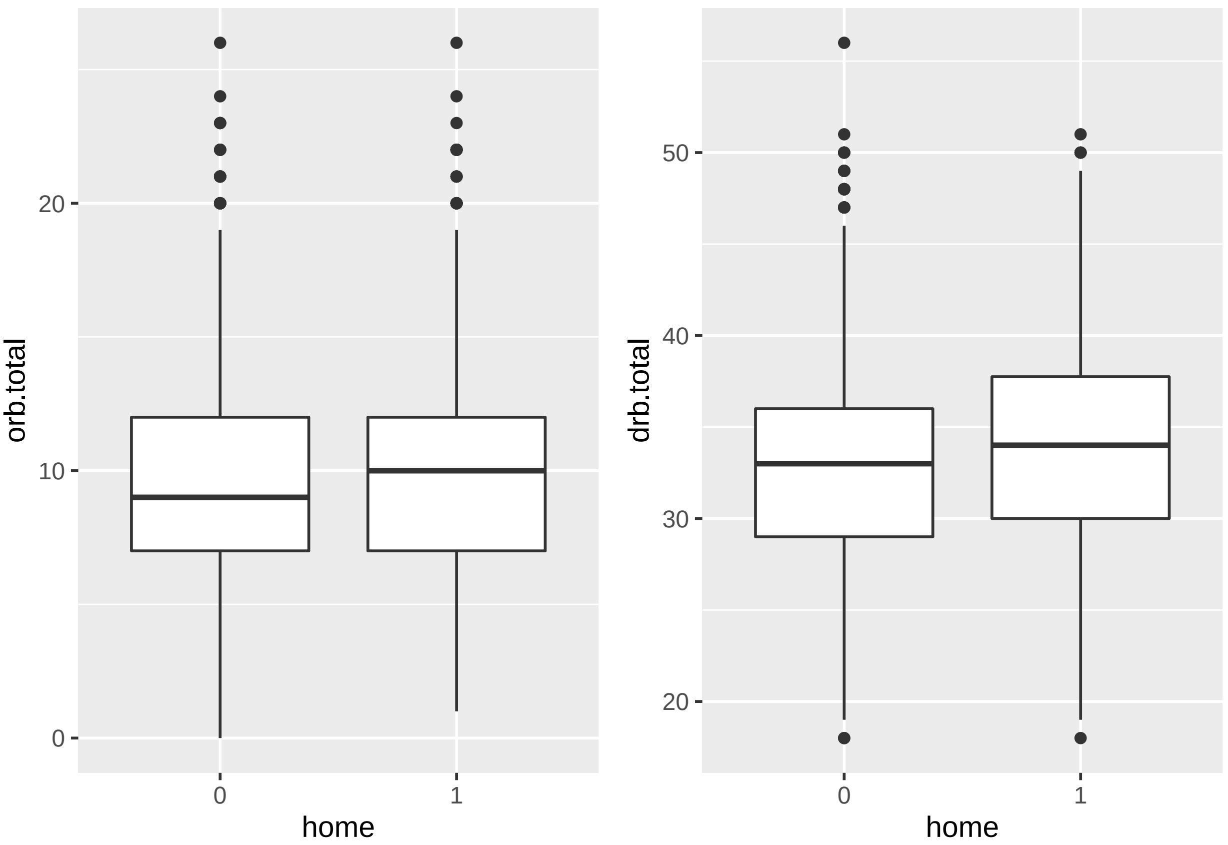
Graph 1 (win_loss_plot) is the visualization of win or loss for home or visitors team respectively with point size. The y-axis represents win or loss. Since the game results is a binary variable, therefore, all the point are distributed over the top and bottom of the graph. The x-axis represents a different team in the NBA(use number to represent rather than team name). Meanwhile, the home teams are represented by the blue dot and away teams are represented by the red dots. The area of the point represents the games a certain team won or lost during this season. From the graph, we can tell that when it is a win, the blue dot overshadows the red dot, implying that it is more likely to be the home team than the visiting team. Similarly, when it is a loss, the red dot overshadows the blue dot, implying that it is more likely to be the away team. This summarizes that there is a difference in the chance of winning for the home and away teams.



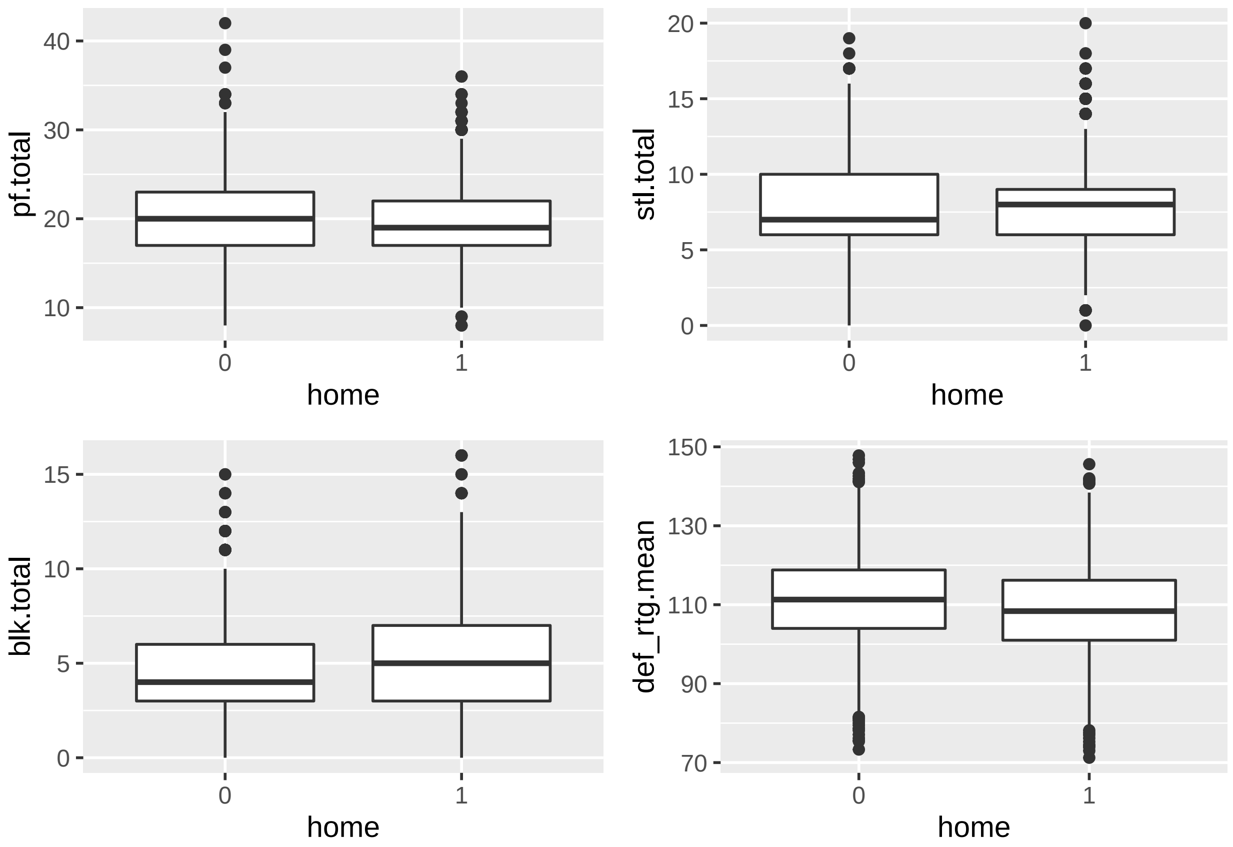
The distance of the away teams and attendance of the audience might have a possible influence on the results of the game as well. For example, if the New York Knicks is playing at Philadelphia which is 2-3 hours drive away, being an away team might influence them differently when they are playing at Los Angeles. Graph 2 above shows the potential impact the distance of the away team and the number of the audience attended. The y-axis of the graph both on the left and the right represents the distance, and the number of fans attended the game. The x-axis is the binary outcome of win or loss. To our surprise, the number of the audience attended has no or little impact on the result of the game. Team with longer travel distances might have a slightly better game result. However, this might happened due to the outliers of the data.



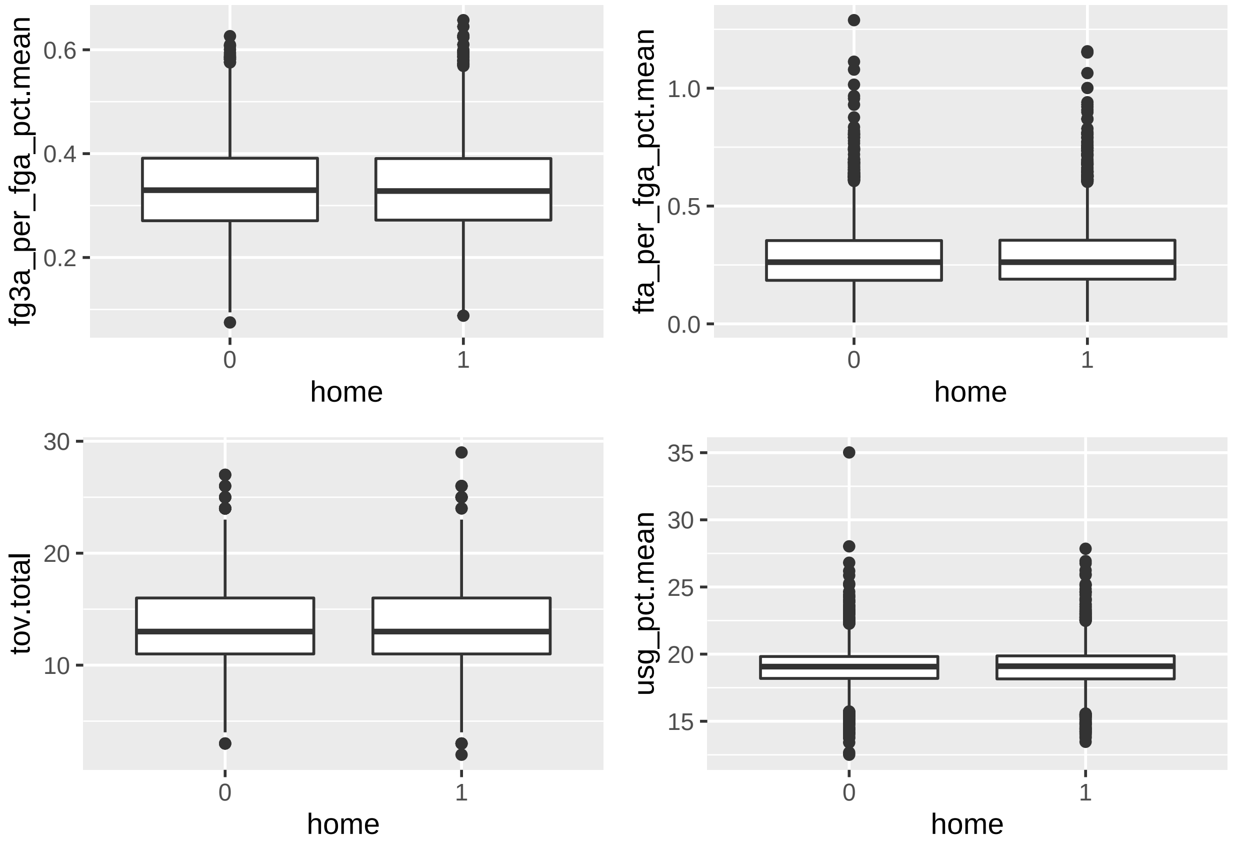
The following research looks in-depth into the in-game statistics. Graph 3 shows the difference in a specific category of shooting performance of the home and away teams. The Y-axis represents the average field goal percentage, 3 point field goal percentage, true shooting percentage, and effective field goal percentage per game respectively. The x-axis represents the home and away team. We can conclude that home teams seem to have a slightly higher average in all four categories. For example, we can tell that for true shooting percentage, home teams average around 55% while the away teams only average around 53%. a 2% difference might not seem like a big difference but for a basketball game, it could lead directly to the winning or losing of a team. Although we can't make any causal conclusion, we can say that there is a correlation between home/away team and their in-game shooting performance measured in average field goal percentage, 3 point field goal percentage, true shooting percentage, and effective field goal percentage per game.



Rebounds have a significant impact on a basketball game. The research investigates the rebounds category to understand the home/away team's performance. Graph 4 visualize the average of an offensive rebound and defensive rebound for home and away team. The Y-axis represents the average total of offensive rebounds per game and the average total of defensive rebounds per game. The x-axis represents the home/away status. The home team averages a total of 10 rebounds per game, whereas the away team averages a total of around 8 rebounds per game. For offensive rebound, the home team and away team average a total of around 34 and 33 rebounds respectively. The rebound category is also correlated with the home/away team in a certain way.



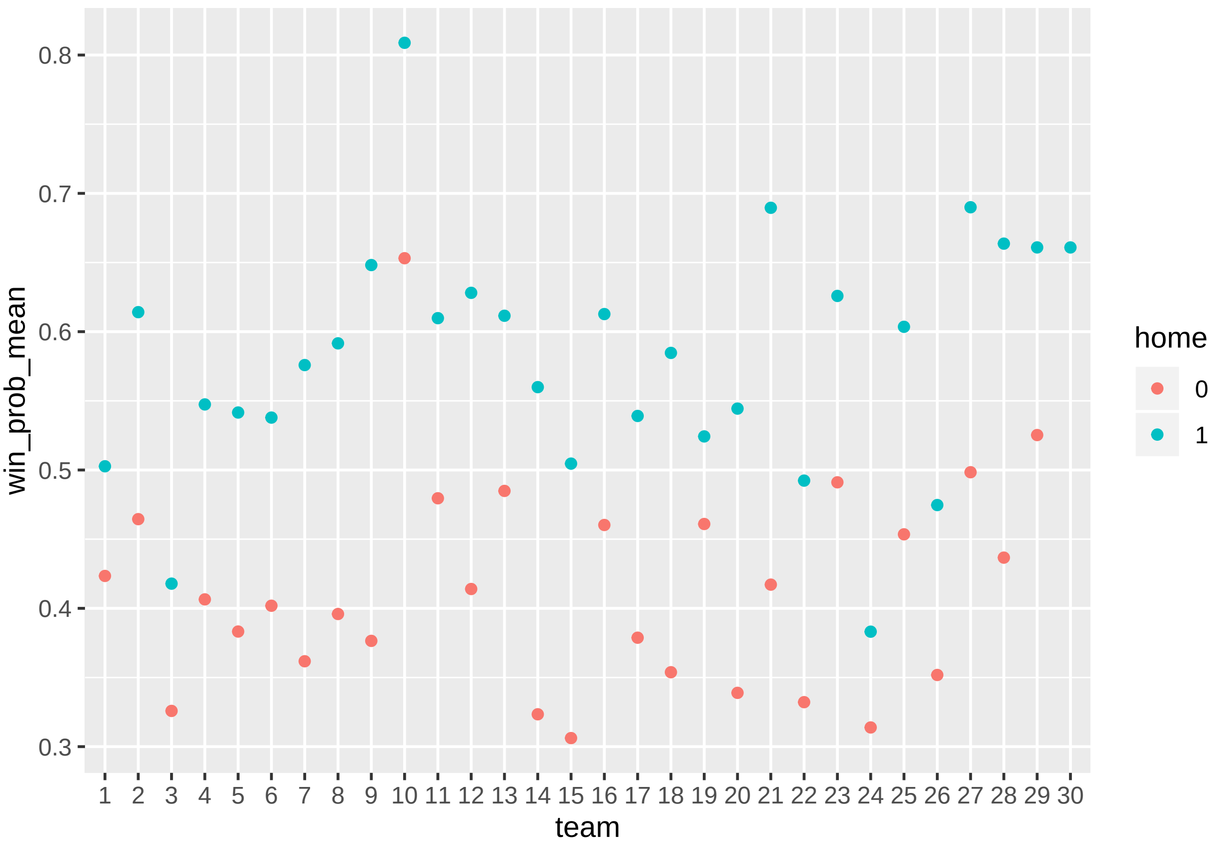
Next exploratory research looks at the defensive performance of a team under the influence of being the home/away team. Defensive performance is measured by total personal fouls per game, total steals per game, total block per game and average of defensive ratings(points allowed per game). Graph 5 shows that the home team averages around 8 steals and 5 blocks per game. The away team averages around 6 steals and 3 blocks per game. The home team also averages around 18 personal fouls and allow another team to score 100 points per game. The away team averages around 20 personal fouls and allow the other team to score around 110 points per game. We can tell that the home team has a better defensive performance in all four categories than the away team.



It also occurred in the analysis that game strategy might be different at home/away games. For example, a home team might play more aggressively by attempting more shots. Game strategy is measured by four different categories, 3 point shots attempted free throw attempts per game, total turnovers, usage percentage (‘an estimate of the percentage of team plays used by a player while he was on the floor’). From graph 6, we can tell that it is little or no influence of game strategy being the home/away team. There is no correlation between the home/away team and game strategy. Therefore, these variables measuring game strategies will not be fitted into the model for prediction.

The last exploratory research discusses the average minutes that starters play when the team is at home or play away. Graph 7 reveals that there is not a huge difference between the playing time of starters at home or away. However, our measure is the average playing time of starters. It cannot accurately describe the playing time of an individual start playing on a team. Also, there exist a lot of outliers. The average playing time of the starters might be able to explain the impact of playing home/away on a team’s star player to win or lose. Therefore, the average playing time of the starter variable will still be considered to fit into the model for prediction.

Two models for prediction are developed based on the exploratory research. One for home team and one for away team. For home team, Winning or loss as the binary outcome is fitted on audience attendance, field goal percentage, 3 point field goal percentage, total offensive rebound, total defensive rebound, total personal fouls per game, total steals per game, total blocks per game, defensive ratings (points allowed per game), average playing time of starters per game using logistic regression. The only difference between the home team and away team is that the away team also includes the distance of traveling to the model. The date is split with a ratio of 80% and 20% into training data and testing data. After calculation, the home team model has an AUC score of 89.32% and the away team model has an AUC score of 89.27%. Both models are accurate enough for prediction and interpretation.



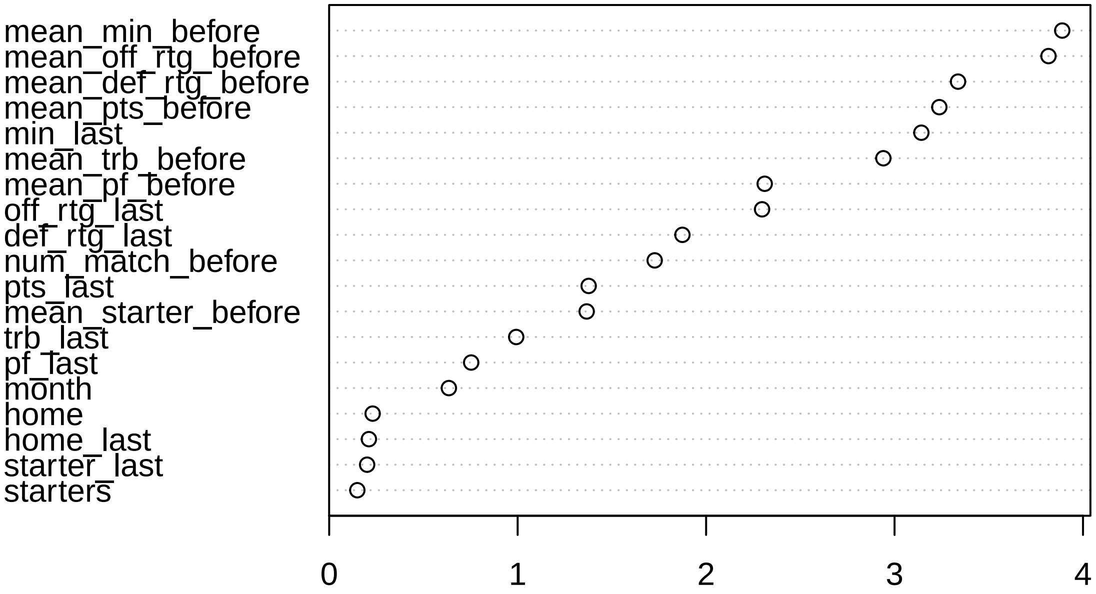
Graph 8 above is the final prediction of the probability of winning or losing based on the model built for the home and away team. The y-axis is the probability of winning. The x-axis represents different teams. Home and away teams are represented by the blue dots and red dots respectively. Almost for all the teams, play at home predicted to have a higher probability of winning than play away. A conclusion can be carefully drawn that there exists a correlation between the probability of winning and play at home/away. In other word, 'home advantage' does exist with statistical evidence.

##Injury prediction for individual players

Another research question we are interested in is to predict a player's injury based on his history of in-game statistics. Two levels of data are considered into building the model for prediction, the average statistics before this coming game such as the average points scored before this game, and statistics from the last game such as the minutes played last game. Datasets are split into 70% and 30% into training and testing data. The outcome of player's injury is fitted on playing home/away, whether he starts or not in the previous game, number of games he played before this game, number of being a starter before this game, average playing time before this game, average points per game before this game, the average of total rebounds, the average of personal fouls, the average of offensive ratings, the average of defensive ratings, if he is starter in the previous game, the minutes played in the previous game, the points he scored in the previous game, the total rebounds he got in the previous game, personal fouls in the previous game, offensive ratings in the previous game, defensive ratings in the previous game, and whether if he plays home/away in the previous game using random forest for prediction. The performance of the model is measured by the AUC score(area under the curve)

Graph 9 is the performance plot of the model under different thresholds. It shows that when the threshold is 0.5, the AUC score is around 62.98%. This is relatively speaking a good accuracy for injury prediction since injury depends on factors outside the basketball court as well. (For example, Ranjo Rondo once fractured his hands in the bathroom while taking a shower when he played for the Celtics)

However, unlike logistic regression, the random forest is hard to interpret.



Graph 10 is the importance plot to explain the impact of each variable in the injury prediction. The x-axis represents the important coefficient included the random forest model. Average offensive ratings and average playing time before both have an importance coefficient of over 3.5. The defensive ratings, minutes played in the last game, average points scored before and average total rebounds before have an importance coefficient of over 3. This reveals the factor that is most likely to predict a player's injury. It contributes to our recommendation for the General Manager for game strategies to avoid players' future injury later in this report.

#Discussions

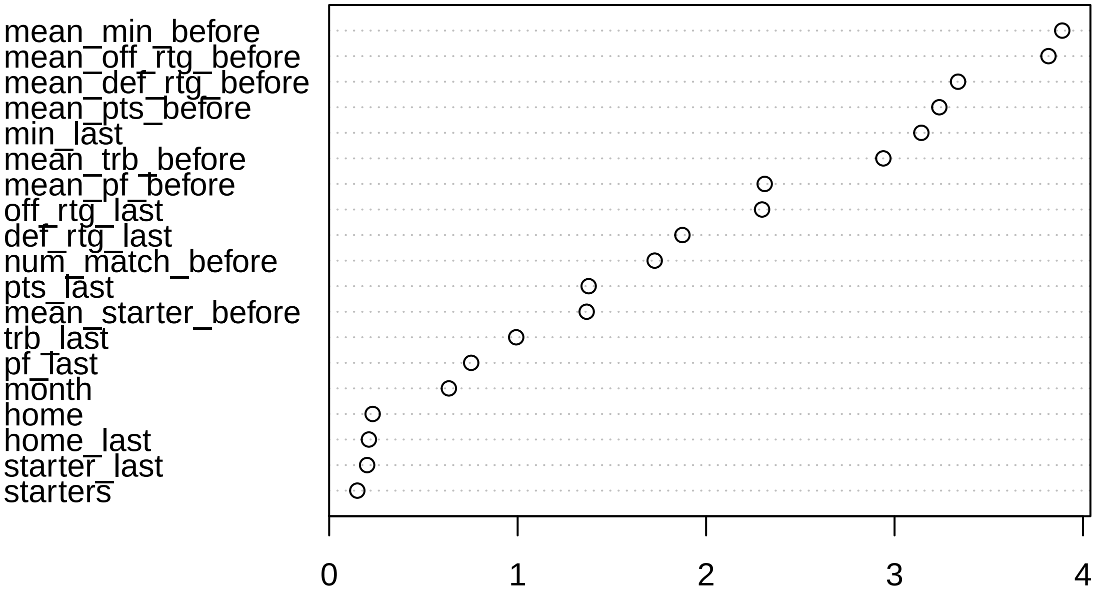
Most of the results are presented in a manner that already been discussed over its application. However, the limitation of the two algorithms we use in this research draws my attention. In 'Random Forests: An algorithm for image classification and generation of continuous fields data sets' by Ned Horning, it discusses the limitation and common issue of random forest. The paper writes that when we are using random forest on regression, the algorithm itself is likely to estimate a narrower range of value. It is likely to predict a lower value for the high value and higher value for the low value. The paper suggests that it is really important to ensure that the training data can 'cover the entire range of response data value'. In our case, the binary outcome is not affected by this. But if we are trying to predict the possibility of getting injured, that would be different.

As for logistic regression, one thing that we fail to accomplish is to perform a longitudinal study on the dataset to see the transition of the impact of each factor/variable to the probability of win on home/away teams. In our future research, we might separate data into different waves and potentially analyze the short-term effect of a certain game statistic influence's being the home or away team.

#Recommendation

Below is a sample suggestion to San Antonio Spurs' head coach and general manager on what they can do to avoid their players getting injured using our injury prediction model.

To R.C. Buford and Gregg Popovich If you can recall in 2016-2017, your all-star player Kawhi Leonard got injured on 2017/05/14. It significantly directly affected your team's chance to compete in the playoffs. It also directly leads to the rest of Kawhi Lenoard over a season and eventually trade to Toronto Raptors. He then won the NBA finals in 2018-2019 and final MVP. We collect and analyze in-game statistics along with the injury data to build a model with the help of machine learning for injury prediction. We can follow Kawhi Lenoard's event and explain to you what we learn from this research. More specifically, what are some suggestions to avoid a player's injury using this model. In general, we built a model using the random forest, a algorithm of machine learning and able to generate a model for injury prediction of an individual player with his previous game statistics such as average playing time before this game, average points per game, etc. We can create an importance plot as the following graph indicates the importance of the different factors in predicting a player's injury.



We can tell that four most factors that indicate that players might get an injury in the next game are Average offensive ratings before this game, average playing time before this game, the average defensive ratings before this game, and minutes played in the last game. If we look at Kawhi Lenoard's case, we can easily find out that on 2017/05/09, the game before he got injured, his offensive ratings and defensive ratings reaches a high point of 123.71 and 103.6. Meanwhile, his average playing time reaches close to a new high of 33.82 mins. He also plays 38.3 minutes in the previous game, which is a new high in the season. He almost reaches a high point in all four most important indicators for predicting that a player might get an injury. With our model, we might be able to provide some suggestions in avoiding a player's injury in the future.

If a player reaches a high point comparing to his previous statistics in the indicators shown in the importance plot, then we need to be careful that this player might get injured. However, as a team, we understand that he still needs to play for the team to win but some small changes might be able to avoid him from getting injured. The most obvious suggestion is to reduce his playing time, even by 5 minutes can dramatically impact the results of getting injured or not because of the high important coefficient of minutes played in the previous game. Another suggestion might not be so straightforward. The total rebounds and personal fouls have a significant influence on the prediction of a certain's injury. Therefore, if a player is predicted from the model that he is likely to be injured in the next game. Assign a player that doesn't require a lot of effort in defense to him might be able to help him to avoid injury. At the same time, it is interesting to see that play as a starter or substitute does not have a major influence on the injury prediction of a player. Therefore, it is a common misconception that a substitute player or a certain player as a substitute instead of a starter this game can help him to avoid getting injured.

#Reference

Glossary. (n.d.). Retrieved from *https://www.basketball-reference.com/about/glossary.html.*

Talukder, H., Vincent, T., Foster, G., & Hu, C. (n.d.). Preventing in-game injuries for NBA players. Retrieved from *http://www.sloansportsconference.com/wp-content/uploads/2016/02/1590-Preventing-in-game-injuries-for-NB pdf*

Horning, N. (2010, December). Random Forests: An algorithm for image classification and generation of continuous field data sets. In Proceedings of the International Conference on Geoinformatics for Spatial Infrastructure Development in Earth and Allied Sciences, Osaka, Japan (Vol. 911).

Ten Have, T. R., Reboussin, B. A., Miller, M. E., & Kunselman, A. (2002). Mixed-effects logistic regression models for multiple longitudinal binary functional limitation responses with informative drop-out and confounding by baseline outcomes. Biometrics, 58(1), 137-144.

#Data source:

https://www.basketball-reference.com/leagues/NBA_2018_games-october.html https://www.basketball-reference.com/boxscores/201910220LAC.html https://www.prosportstransactions.com/basketball/Search/Search.php https://www.kaggle.com/ghopkins/nba-injuries-2010-2018/version/1