Final Project, Fall 2023

Samuel Edwards, Kendall Schroth, Kachun Cheung, Eduardo Hinojosa

Monday December 11, 2023 at 11:59 PM

Contents

1 Data Overview

Our data is a census of all NBA games from the years 2004 to 2022. The data downloaded from Kaggle provides 5 sets of data that show the statistics of all games during the given years.

1. The **Games** data set has data from all games from 2004 season to 2022 season, with basic information about team statistics like home and away teams and number of points.
2. The **Games Details** data set goes into further detail of each game giving all statistics of players for a given game.
3. The **Players** data set lists the players in the NBA for a specific season and the team they played for in that respective year
4. The **Teams** data set which elaborates on the player and team id used to identify individuals in the Games and Games Details data sets.
5. The **Ranking** data set that gives a ranking of NBA teams on a given day.

Since our data is a census of all NBA games and players, there is no systematic exclusion of certain groups. Put differently, every player on every team in the seasons represented in the data set is included. Due to the nature of basketball statistics and their effect on a player's minutes played in a game, the players are well aware that every play they make is recorded as some statistic. The diverse collection of data sets allows for various types of analysis. After some exploratory data analysis, we decided to engineer a data set with individual player statistics per game and another looking at team statistics per game. More specifically, each instance in the data set for Plus/Minus prediction represents the statistics of a player for a given game. This will mean that our interpretation of the prediction will only apply to that specify player. Similarly, for the Away game and Free throw percentage data set, each row represents a game and the corresponding home and away team's free throw percentages. This leads to an interpretation on a game by game basis. When we aggregate these games statistics, we will be interpreting the aggregation of each game in one season. When analyzing our findings from the team statistics, we must remember that in any game, a team can include different players throughout the seasons; players are traded throughout the season, and injuries prevent players from playing in certain games.

Our study examines player and team statistics to identify if a mental effect exists for a free throw during an away game. Therefore, we assume that the team mindset remains relatively similar whether or not a star player sits on the bench. When looking at the player statistics and their effect on plus/minus, we will need to remember that each player has a different purpose on the court. This means that a defensive post player making a turnover has a much different effect on the pace of a game than a point guard making a turnover.

The classic topic of referee bias must be discussed when using any sports data. Since our study looks explicitly at free throws resulting from a foul called by the referee, the question of whether referee bias exists must be considered. The two main types of referee bias are "Superstar Bias," and "Home Team Bias." "Superstar Bias" is the belief that referees tend to favor calling fouls for better performing, more popular players, and "Home Team Bias" is the belief that the local team crowd influences referees to call fouls in favor of the home team. Many studies have been conducted that contradict one another on their findings. One such study, published by Konstantinos Pelechrinis states that the" Home Team Bias" is mainly pronounced during playoff games [7]. The study also finds the Superstar Bias to have conflicting positive and negative effects on teams. Another study by Christian Deutscher finds no referee bias in terms of both home team and superstar bias when looking at calls made in the final two minutes of a game, again commenting that playoff games should be assessed separately [5]. Since the studies suggest conflicting evidence on referee bias as having an actual bias towards our data, we determined this will not significantly affect our research. Other common types of bias–selection bias, measurement error, convenience sampling–do not apply either. We use the entire season's data and know that NBA statistics, such as free throw percentages, are concrete.

There was no need to adjust for differential privacy because inherently, the statistics of the NBA are not private, and we were provided with all of the features needed; there was a significant amount of data-cleaning and engineering necessary to apply to the data set.

After looking into the rows with missing data, we determined that most columns with missing data come from players who do not have enough playing time to form a concrete Plus/Minus. Whether they do not play in every game or do not have enough minutes in a game, we decided these players were not significant enough to be included in our study. We dropped any rows with a NULL Plus/Minus values because of their statistics' low impact on the game at hand. While we encountered no missing values when looking at team free throw percentages, we did, however, clean the data set for the free throw percentages to drop any playoff games from our study. Our reasoning is that Home and Away games are not randomly determined for playoff games as they are the rest of the season. The playoff games give a team home-court advantage, meaning they play more games at home if they are ranked higher. We also found uncertainty in referee bias in playoff games specifically. Therefore, we decided not to include playoff games in the study to maintain randomness assumptions. To determine which games to exclude, we used a data set from "Basketball Reference," a website that provides historical data on the NBA [1].

2 Research Questions

In this paper, we will address two questions: Can we predict the plus-minus score by looking at their performance during that game, and does playing in an opponent's stadium impact a team's free throw percentage? By creating a method to evaluate Plus/Minus, teams can see the importance of specific strengths for players and how they affect the game's pace. By evaluating the impact of the stadium location on a team's performance, the NBA could decide how free throws should be carried out or can choose set standards for audience etiquette.

2.1 Predictions

In our Prediction research, we are trying to find the outcome of a player's performance based on a series of factors that are recorded throughout the game and season. In other words, we are looking at the multiple variables encompassing a player's performance and trying to arrive at an accurate conclusion. As a result, this structure of questions is ideal for a Random Forest Regressor that utilizes various layers to arrive at a prediction. In addition, the Ordinary Least Squares GLM was another prediction method that fits our continuous Plus/Minus variable, assuming that this is a linear relationship between Plus/Minus and the other variables. As with any GLM, this method might only show the best results if the relationship between our variables and the Plus/Minus scores is linear. If this relationship is not linear, it will not show accurate results, ultimately making our predictor unnecessary. The random forest regressor also has limitations with the interpretation. Since it is not easy to interpret each variable's effects on the prediction, utilizing the predictor to gain information about the value of specific statistics will not be helpful.

2.2 Causal Inference

To uncover the impact of stadium location on a team's performance, we will be using causal inference to asses how, as a treatment, playing in an opponent's stadium might impact the team's free throw percentage, being the outcome in this method. We chose this method because it is used to answer what we want to know: does playing an away game *cause* a team to perform worse? This method will not only show us if there is a causal relationship but also give us an idea of how much of an impact our treatment will have on the outcome. NBA officials could then use this quantified relationship to decide where they deem fit. A common concern with causal inference problems is addressing all variables that that may subliminally affect the treatment and/or the outcome. We would encounter a limitation if we fail to uphold the assumptions of a Stable Unit Treatment Value Assumption; having interference of treatments or different versions of treatments. We would have a hard time justifying the application of our causal question to a season such as the 2020 Bubble, which had each team play in the same arena, with no audience. In this case, there are no away games to observe and no way of using matching to evaluate Individual Treatment Effect, making our question and method inapplicable.

3 Exploratory Data Analysis (EDA)

3.1 Predictions EDA

To see how the variables independently effect the plus-minus score of a player, we plotted each feature against plus-minus as seen in Figure 3.1.1, where the features are (per player, per game):

- Field Goal Percentage (FG_PCT)
- Three Point Field Goal Percentage (FG3_PCT)
- Free Throw Percentage (FT_PCT)
- Offensive Rebounds (ORB)
- Defensive Rebounds (DRB)
- Rebounds (RB)
- Assists (AST)
- Steals (ST)
- Blocks (BLK)
- Turn Overs (TO)
- Personal Fouls (PF)
- Points (PTS)
- Time played in seconds (SECS)

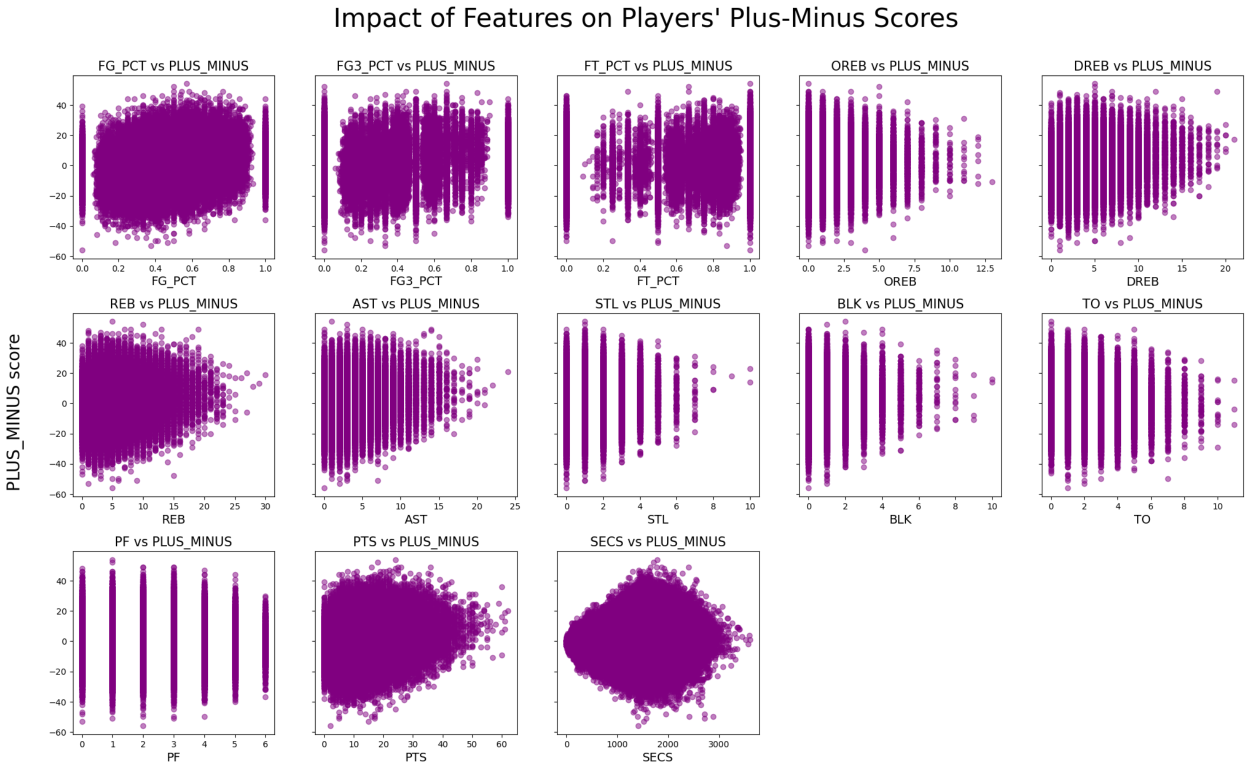


Figure 3.1.1

It is apparent in the visualizations that FG_PCT, FG3_PCT, and PF have little effect on the plus-minus of the players. However, all other variables show a trend; when approaching larger values of the respective variables, the plus-minus tapers to a smaller value. These trends are surprising, given the theoretical importance of those variables in assessing a player's plus-minus score. It can then be interpreted that FG_PCT, FG3_PCT, and PF are not as important features when making predictions on players' plus-minus scores. The comparison of SECS and plus-minus is an exception among the variables. It shows that players with longer and shorter playing times tend to have a lower plus-minus score.

Another idea that we felt required further exploration is if starting players produce higher plus-minus scores for that game. The natural assumption here is that starters tend to be the better-performing players on a team, contributing positively the most, earning them a spot in the starting five. To visualize this, a binary response was used (0 for a negative plus-minus score; 1 for a positive plus-minus score in the graph for Figure 3.1.2; the positive and negative plus-minus distributions are relatively equal. This went against our expectations and showed little to no impact on the plus-minus given whether the player was starting. It is concluded that being a starting player is not an essential feature when predicting the plus-minus score.

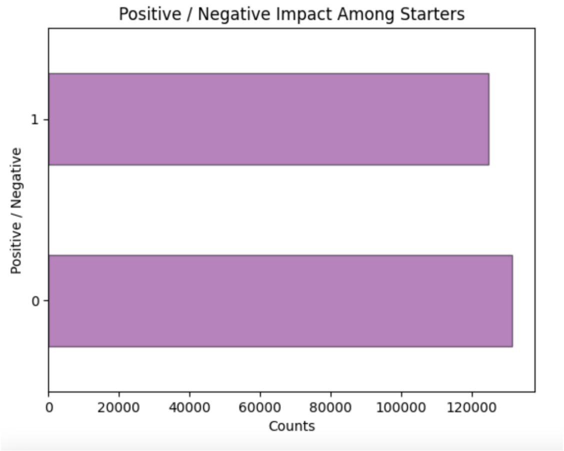


Figure 3.1.2

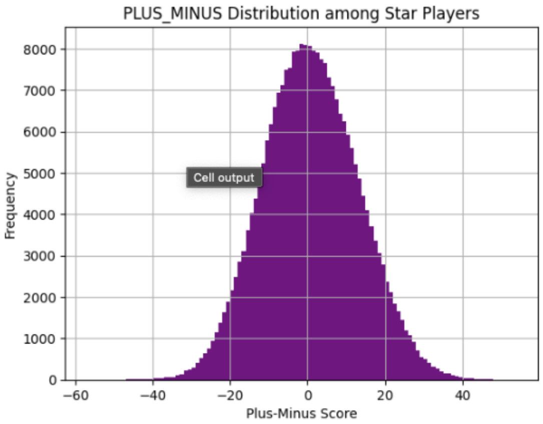


Figure 3.1.3

3.2 Causal Inference EDA

For our causal question, we decided to use home and away free throw percentages, averaged per team, to see if there is a causal relationship between an Away game and the average team free throw percentage. When comparing free throw percentages across teams in the NBA, we noticed a stark difference between the free throw percentages of the Sacramento Kings in 2018. In the 2017-2018 NBA season, the Kings successfully made, on average, 76.0% of their free throws at home games but only 71.1% when playing away. This difference of 4.9% seems significant compared to the league's average difference in free throw percentage of home and away games at 0.15% . This led to the exploration of whether there is a causal relationship between Away Games and Free Throw Percentage for the 2017-2018 Sacramento Kings.

This question is further motivated by Figure 3.2.1, which shows the distribution of free throw percentages for home and away games. Visually, you can better understand the difference in home and away free throw percentage distributions. The away games are centered around a much smaller percentage than the home games. We can also see that while both distributions have a large portion of their data in the same area as the other distribution, the home games have several games with percentages much higher than the large portion of the data. The opposite can be said for away games.

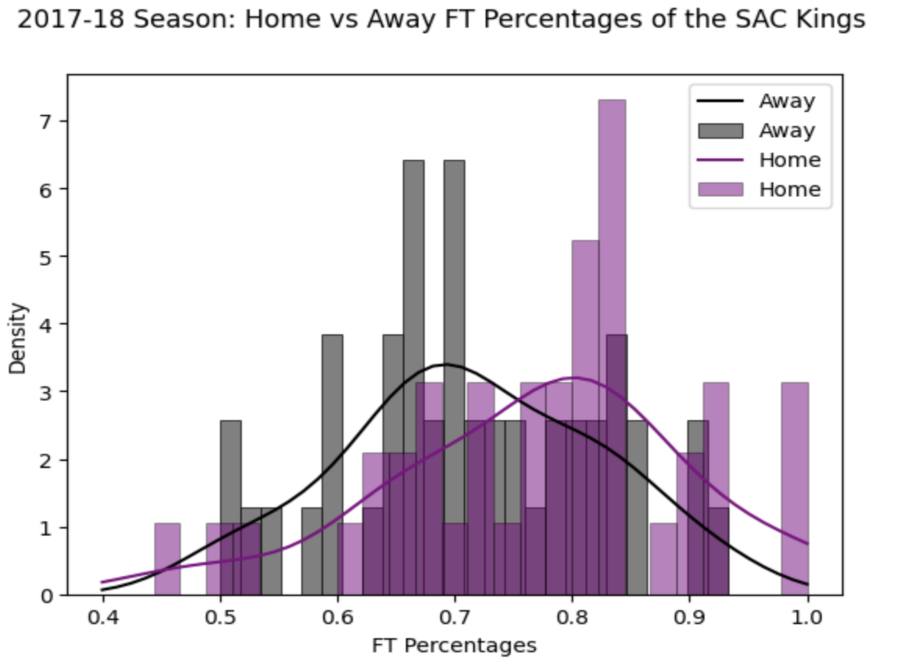


Figure 3.2.1

Because of the causal nature of our question, we also considered what confounding variables might affect the result. While many things correlate with free throw percentage, few variables affect home and away games. The exception to this assumption is playoff games. We excluded playoff games from our study because home and away games are decided by team ranking. To verify that home and away games have a random distribution, we visualized the timeline of regular season games, indicating whether or not it was a home game, as seen in Figure 3.2.2. While there are strings of 5-6 games that are all away or all at home, the main distribution of games and the distribution of the strings of games are evenly distributed.

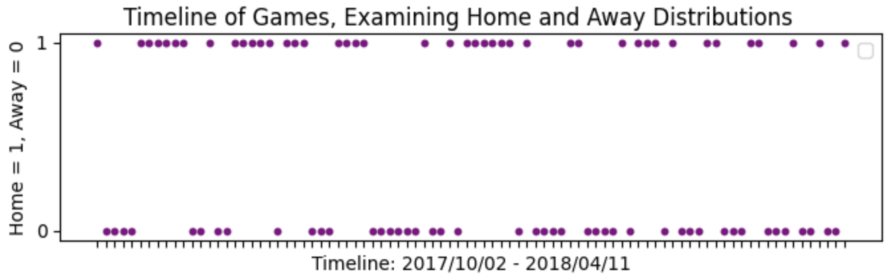


Figure 3.2.2

Another potential confounding variable we considered was the effect of longer loss streaks on our treatment and outcome. A loss streak is the amount of consecutive losses leading up to a game. Since a predetermined schedule determines who a team plays, if a team coincidentally has a sequence of games against the top teams in the league, this team could accumulate a large loss streak. In addition, this schedule determines when teams play each other and whether it is a home or away game; there could be an effect on through-loss streaks. To ensure this is not the case, we visualized the schedule of games on a timeline, comparing free throw percentages for each game on the Y-axis. To visualize the corresponding loss streak for each game, the dot's color denotes how long the loss streak is. In other words, the darker the dot is, the longer the loss streak. As seen in Figure 3.2.3, no visible trend shows that a long loss streak lowers the free throw percentage in a game, confirming that there is no "spillover effect" from one game to the next. If there was, we would observe darker dots appearing lower on the y-axis and further to the right. Nonetheless, after considering these two visualizations, we prove that Away Games have no confounding variables when excluding playoff games.

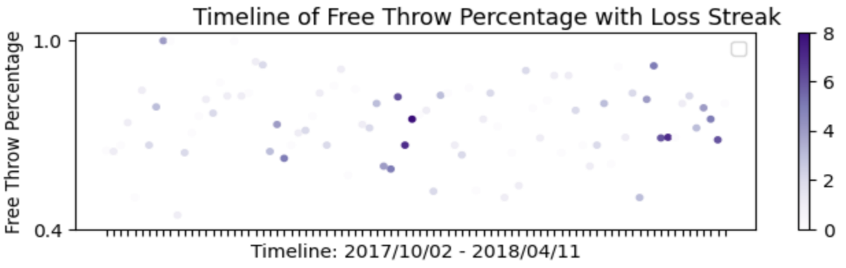


Figure 3.2.3

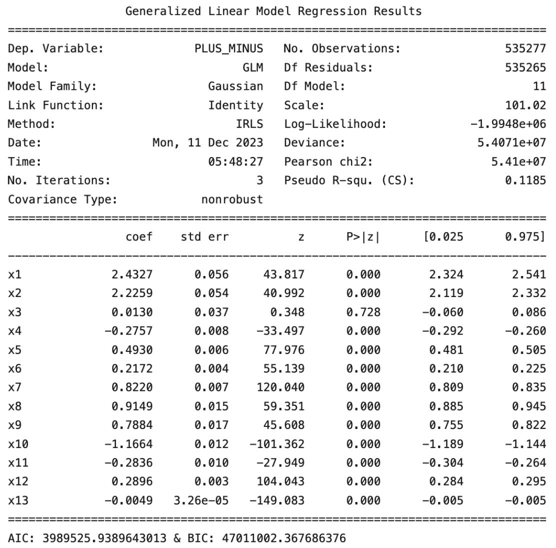
4 Prediction with GLMs and Nonparametric methods

4.1 Methods

We attempt to predict the Plus/Minus score using the features: Field Goal Percentage (FG_PCT), Three Point Field Goal Percentage (FG3_PCT), Free Throw Percentage (FT_PCT), Offensive Rebounds (ORB), Defensive Rebounds (DRB), Rebounds (RB), Assists (AST), Steals (ST), Blocks (BLK), Turn Overs (TO), Personal Fouls (PF), Points (PTS), Time played in seconds (SECS). The Plus/Minus statistic measures how much a player's team scores versus the other team's score, when that player is on the court. It essentially is a measure of their impact on the game. We chose to use the features above because these statistics are often used to determine the value of a player, thus should be a good predictor for the player's impact on the game. To make these predictions, we chose to use Linear Regression and took a Frequentist approach. Linear Regression is able to cover the whole range of our possible outcomes since the plus-minus scores can be either positive or negative; this is under our assumption that there is a linear relationship between features and the dependent variables. To ensure the best results, we compared two models: GLM with Identity Link Function and Ordianry Least Squares (OLS) regression. We compared their respective Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), metrics of how well a model fits the data given its complexity, to decide which is best to use. Our results showed that the OLS Regression model yielded a better model due to it having a lower AIC and BIC than its counterpart (Figures 4.1.1 and 4.1.2).

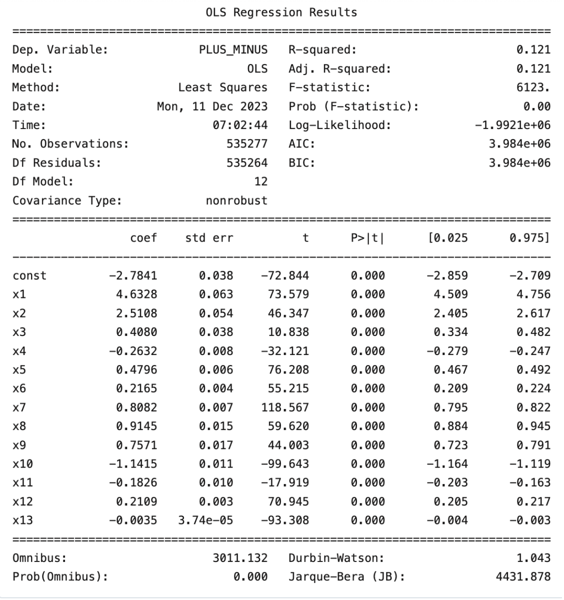
The non-parametric model we chose to use is the Random Forest Regressor. This model is great at making predictions because we set n estimators = 10 to combine the output of multiple decision trees to make a final prediction. It is generally more accurate because it reduces overfitting. We are also working under the assumption that the data is normally distributed as shown in our EDA in Figure 3.1.3.

With both of our chosen models, the MSE of each model will be used to portray their performance in predicting the plus-minus scores; the model with the lowest Mean Squared Error (MSE) will be evaluated as the best model.



```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:             PLUS_MINUS   No. Observations:               535277
Model:                            GLM   Df Residuals:                   535265
Model Family:                Gaussian   Df Model:                           11
Link Function:               Identity   Scale:                          101.02
Method:                          IRLS   Log-Likelihood:            -1.9948e+06
Date:                Mon, 11 Dec 2023   Deviance:                   5.4071e+07
Time:                        05:48:27   Pearson chi2:                 5.41e+07
No. Iterations:                     3   Pseudo R-squ. (CS):             0.1185
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
x1             2.4327      0.056     43.817      0.000       2.324       2.541
x2             2.2259      0.054     40.992      0.000       2.119       2.332
x3             0.0130      0.037      0.348      0.728      -0.060       0.086
x4            -0.2757      0.008    -33.497      0.000      -0.292      -0.260
x5             0.4930      0.006     77.976      0.000       0.481       0.505
x6             0.2172      0.004     55.139      0.000       0.210       0.225
x7             0.8220      0.007    120.040      0.000       0.809       0.835
x8             0.9149      0.015     59.351      0.000       0.885       0.945
x9             0.7884      0.017     45.608      0.000       0.755       0.822
x10           -1.1664      0.012   -101.362      0.000      -1.189      -1.144
x11           -0.2836      0.010    -27.949      0.000      -0.304      -0.264
x12            0.2896      0.003    104.043      0.000       0.284       0.295
x13           -0.0049   3.26e-05   -149.083      0.000      -0.005      -0.005
==============================================================================
AIC: 3989525.9389643013 & BIC: 47011002.367686376
```

Figure 4.1.1



```
                            OLS Regression Results
==============================================================================
Dep. Variable:             PLUS_MINUS   R-squared:                       0.121
Model:                            OLS   Adj. R-squared:                  0.121
Method:                 Least Squares   F-statistic:                     6123.
Date:                Mon, 11 Dec 2023   Prob (F-statistic):               0.00
Time:                        07:02:44   Log-Likelihood:            -1.9921e+06
No. Observations:              535277   AIC:                         3.984e+06
Df Residuals:                  535264   BIC:                         3.984e+06
Df Model:                          12
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -2.7841      0.038    -72.844      0.000      -2.859      -2.709
x1             4.6328      0.063     73.579      0.000       4.509       4.756
x2             2.5108      0.054     46.347      0.000       2.405       2.617
x3             0.4080      0.038     10.838      0.000       0.334       0.482
x4            -0.2632      0.008    -32.121      0.000      -0.279      -0.247
x5             0.4796      0.006     76.208      0.000       0.467       0.492
x6             0.2165      0.004     55.215      0.000       0.209       0.224
x7             0.8082      0.007    118.567      0.000       0.795       0.822
x8             0.9145      0.015     59.620      0.000       0.884       0.945
x9             0.7571      0.017     44.003      0.000       0.723       0.791
x10           -1.1415      0.011    -99.643      0.000      -1.164      -1.119
x11           -0.1826      0.010    -17.919      0.000      -0.203      -0.163
x12            0.2109      0.003     70.945      0.000       0.205       0.217
x13           -0.0035   3.74e-05    -93.308      0.000      -0.004      -0.003
==============================================================================
Omnibus:                     3011.132   Durbin-Watson:                   1.043
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             4431.878
```

Figure 4.1.2

4.2 Results

All things considered, we looked at the performance of OLS and Random Forest by comparing the mean squared errors against each other to identify which approach produces more accurate predictions on average.

- **OLS Regression MSE**: 99.03716892224078
- **Random Forest Regressor MSE**: 110.38445486311059

Using this information we can identify that Linear Regression produced the smaller MSE than Random Forest Regressor and in turn we chose this model to predict PLUS/MINUS. To show the uncertainty of our GLM model, we graph its fitted values against the observed values in the data set (Figure 4.2.1). We

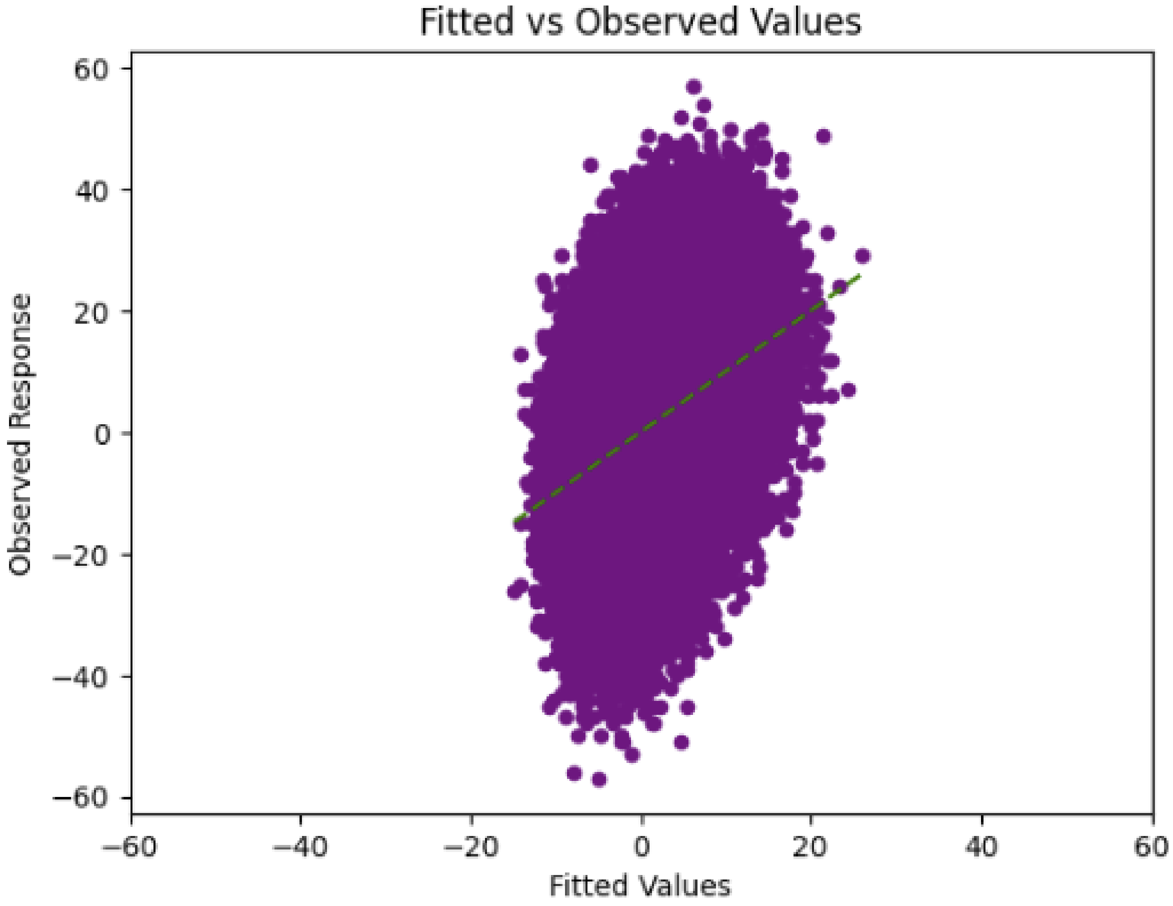


Figure 4.2.1

can see that the fitted values fall in a much smaller range (-15, 25) than the range covered by the observed values (-60, 60). This observation indicates that the GLM model could be experiencing a rather large level of uncertainty due to underfitting or uncaptured variability within the data.

4.3 Discussion

Although we determined the Linear Regression model best predicted the unknown plus-minus dataset, we need more confidence in applying this to future datasets. Because our Mean Squared Error is near or above 100 for each regressor, we can not confidently say that either regressor fits the model well. One of our model's significant limitations is the potential for improvement in using standard Plus/Minus to measure player performance. For example, it does not account for competition or other players on the team. Adjusted Plus/Minus (APM) and Regularized Adjusted Plus Minus (RAPM) would help improve our model.

Dan Rosenbaum, a former Executive Director of Basketball Strategy and Analytics with the Atlanta Hawks, an analytics consultant with the Cleveland Cavaliers, and an assistant professor at the University of North Carolina at Greensboro, articulated and defined Adjusted Plus/Minus in 2004. According to his paper, ”Measuring How NBA Players Help Their Teams Win,” [3] He wrote, ” these ”unadjusted” plus/minus ratings do not measure the value of a player per se; they measure the value of the player relative to the players that substitute in for him. In addition, there are differences in the quality of players that players play with and against. A weak starter on a team with exceptionally good starters (relative to bench players) will generally get an unadjusted plus/minus rating – regardless of their actual contribution to the team.” This is likely contributing to our high uncertainty. Because there is too much variation in the effect that other players on the court have, it is hard to predict using only the individual's statistics.

He suggested, ”a better measure of player value would ”adjust” these plus/minus ratings to account for the quality of players that a given player plays with and against. In addition, it would account for home-court advantage and clutch time/garbage time play. Thus, unlike in unadjusted plus/minus ratings, these ”adjusted” plus/minus ratings do not reward players simply for being fortunate to be playing with teammates better than their opponents.”

APM is a variation of the standard Plus/Minus stat, and ”tries to figure out how these players work together to reach the margin on the court during that time, while also accounting for home court advantage.” [4] RAPM is a Bayesian technique combining the data with theoretical beliefs regarding reasonable, large data ranges for the parameters to produce more accurate models. According to the NBAstuffer analytics website [6] , ”RAPM is about twice as accurate as an APM using standard regression and using 3 years of data, where the weighting of past years of data and the reference player minutes cutoff has also been carefully optimized.”

5 Causal Inference

5.1 Methods

For our Causal Inference experiment, we have the treatment as an NBA team playing the opponent's stadium, and the outcome we are measuring is the Free Throw Percentage (FT_PCT) of the visiting team (i.e. measuring the FT_PCT of the Sacramento Kings when they are playing in an away stadium) (Figure 5.1.1).

While researching for this experiment, we learned that the NBA uses a formula when formulating the schedule; this includes looking at arena availability and travel costs. But with every season, each team is guaranteed to play all other teams **at least** 2 times, one home and one away [2]. Because of this, there is no advantage for teams, and it eliminates the possibility of schedule assignment being a confounding variable. Furthermore, no variables affect both the FT_PCT and whether or not a team will be playing an away game. These conditions make this a randomized experiment, which implies that unconfoundedness holds for this research question. The only possible confounder is whether or not a team makes it to the playoffs (winning a match will determine if you play the next team and also determines the location where you will be playing), but this is accounted for by excluding all playoff games from the data set used in the experiment.

Under the assumption of this being a randomized experiment, we are able to use Simple Difference in Mean Outcomes (SDO) to calculate the causal effect of game location on the visiting team's FT_PCT. This also ensures that the SDO will be an unbiased estimator of the Average Treatment Effect (ATE) [8]. We can also assume that there is no interference between the treatment of one unit and the outcome of another unit after visualizing free throw percentages compared to the length of the prior Loss Streak. Lastly we can assume that there is no treatment inconsistency with different versions of the treatment because we use the same season and team to ensure the treated unit remains the same throughout our study.

We could not find or think of any possible colliders when further researching the data set, treatment, and outcome. This is because there is nothing that our treatment affects that our outcome also affects. The game's location is set in stone before the season begins, and the FT_PCT is independent for each player and each game.

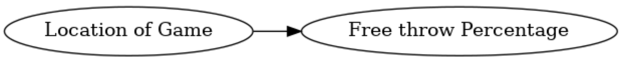


Figure 5.1.1

5.2 Results

After calculating the ATE for all teams, we plotted them on a histogram and compared the ATE of the Sacramento Kings to the rest of the teams (Figure 5.2.1). The Kings had one of, if not the lowest, ATEs amongst all the teams at an ATE ≈ -0.061, while the average ATE amongst all teams is approximately 0. This result tells us that playing in a visiting stadium has minimal impact on the FT_PCT of the visiting team. To put the Kings ATE in perspective, out of 25 free throws, a team might miss 1.5 free throws in a game due to being a visiting team. Overall, this effect is insignificant enough to consider when a team travels.

Regarding uncertainty, we only focused on one team and one season. This season could have been an outlier among the other seasons. Our initial hypothesis was that playing at a visiting stadium would notably impact the visiting team's FT_PCT, but these findings prove otherwise.

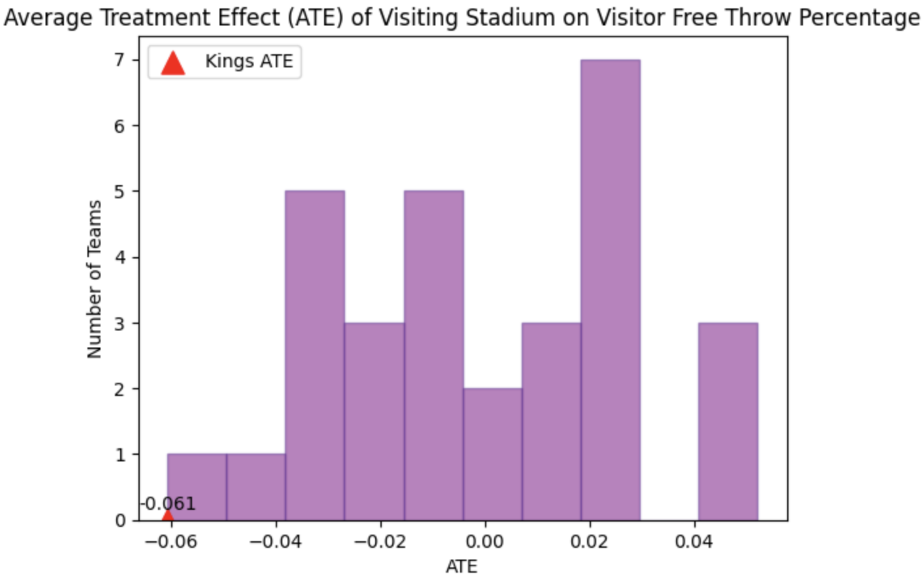


Figure 5.2.1

5.3 Discussion

Though we obtained results that showed the causal relationship between our treatment and the outcome, there still needs to be concern with the method used. We chose this specific team (Sacramento Kings) and season (2017) because of their difference in home and away performance in free throws compared to other teams and seasons. Our limitation in what seasons we were given in the data has the potential to leave out substantial evidence that could lead to unexpected results. Also, this data set does not provide information about the referees, which could be helpful to examine potential referee bias.

Having a more extensive time range of data and information about the referee (e.g., hometown, years of experience, any previous sanctions) would allow us to rule out these limitations that reduce our confidence in our results. Data from more seasons would give us more example games to compare our results. Having the additional referee data would solidify the referee's credibility and allow us to make assumptions about the referee or if referee bias is prevalent when looking at the causal relationship.

Ultimately, our findings show that no significant causal relationship exists between a team playing at the opponent's stadium and how that impacts their free throw performance. Given the data we have, we are confident with these results. We looked at the most extreme case of a team's performance in free throws, and even then, there were no signs that the treatment impacted the outcome, or at least significant enough for a team to worry about.

6 Conclusions

While we did find the best possible regressor, for the data available to us, the OLS regressor had a significant amount of uncertainty. For this reason, we do not recommend a Linear or Random Forest model for the Plus/Minus prediction. Additionally, we were unable to find significant causality between the Kings Away game and their free throw percentages respective to their home free throw percentages. We found that the Kings would only score roughly 1.5 points less on average in an away game.

Our Prediction on Plus/Minus can be applied to any player that we have the correct statistics on, therefore it is extremely generalizable to all NBA players. However, our Away game and free throw percentage causality question contains results only applicable to the Sacramento Kings in the 2017-2018 season. We could attempt to apply our approach to other teams, but we would likely find similar lack of causality, especially without preliminary statistics, such as the -4.9 percent difference that we found.

The results of our prediction regressors have low confidence, therefore we cannot suggest that a linear relationship be applied to Plus/Minus statistics or that the Random Forest Regressor be used to interpret the effect of other statistics. Ultimately, we call to action the NBA officials to establish a more comprehensive approach to quantify the Plus/Minus of a player.

Because our Away game and Free throw percentage causality found such a small impact on the game, our only call to action would be to allow the fans of the home team to cheer as they please. Home team advantage is a key part to fan participation in sports, and since it appears to not have much impact on the game, there is not much reason to not let the fans have their fun.

We did have the need to merge any data source because of the comprehensive data provided by Kaggle. Because of the conflicting research on whether or not referee bias exists, it was hard to quantify any bias that specific referees might have on a given game. This could have an affect on the Plus/Minus of a game, if personal fouls had a big affect on the prediction. It could also have any affect on the number of free throws taken, and who is taking them in a game based on home bias and superstar bias.

To build upon our Plus/Minus Prediction, a future study could use our predictors to help aid in the draft pick as well as trades. By using the predictor we created and including more information on how a player's strengths works with other players, it could be used to see if a particular player would play well with a team. As for the Away Games and Free throw Percentages question, other studies could use our study as a stepping stone to other ways a home and away game might affect the game. For example, using audience attendance in the most popular stadiums could be a new approach in finding causality between stadiums and free throw percentages.

Overall, we learned that statistics on NBA games and players can be misleading in their effect on the game. While of course the free throw percentages and the Plus/Minus of a game are statistics that have value, and should be considered is evaluating the overall performance of a player or game. These considerations should also be taken lightly. We have proved that a -4.9% difference in home free throw percentages overall and away free throw percentages overall does not have as much causal impact on the game as one might think. Similarly, we saw that Plus/Minus has much less to do with the statistics of a player and more to do with unquantifiable aspects of the game, such as pacing and who a player is on the court with. Altogether, these statistics, and likely other statistics as well, can be helpful in quantifying success, they should only be used as an aid to the qualitative aspects of the game that cannot be captured in quantitative statistics.

7 References

References

[1] Nba playoffs series page. URL https://www.basketball-reference.com/playoffs/series.html. Accessed: February 13, 2024.

[2] D. Arlauckas and A. Hall. How is the nba schedule made? rules and formula, October 20 2023. URL https://en.as.com/nba/how-is-the-nba-schedule-made-rules-and-formula-n-2/. Diario AS.

[3] I. Barzilai. Regularized adjusted plus-minus (rapm) - 82games.com. URL https://www.82games.com/barzilai2.htm.

[4] BasketballStat. Regularized adjusted plus-minus (rapm), 2019. URL https://basketballstat.home.blog/2019/08/14/regularized-adjusted-plus-minus-rapm/.

[5] C. Deutscher. No referee bias in the nba: New evidence with leagues' assessment data. *Journal of Sports Economics*, 16(1):91–96, January 1 2015.

[6] NBAstuffer. Regularized adjusted plus-minus (rapm) - nbastuffer.

[7] K. Pelechrinis. Quantifying implicit biases in refereeing using nba referees as a testbed. *Scientific Reports*, 13:4664, 2023. doi: 10.1038/s41598-023-31799-y.

[8] R. Sridharad. Lecture 16: Causal inference, part 2: Potential outcomes and randomized experiments, 2023. URL https://data102.org/fa23/. University of California, Berkeley Data, Inference, and Decisions.