Data 102: Data, Inference, and Decisions
Fall 2023 — Final Project Report

Carl Conste, Emily Chin, Joshua Sengvongdeuane, Osmaan Mysorewala

University of California, Berkeley

December 11, 2023

Contents

1 Introduction

Our group's final project uses the NBA data set that compiles the statistics of all NBA games from 2003-2022.

1.1 Background

The National Basketball Association (NBA) is a basketball league comprised of 30 professional teams. These 30 teams are divided into two "conferences", the Western Conference and the Eastern Conference. There are 15 teams in each conference and the winners of each conference will proceed to the Finals as they seek to win the championship. The statistics of each of these games are recorded. Specifically, the years in which the statistics of these games, teams, and players are recorded range from 2003 to 2022. In an NBA season, there are typically 82 games played by each team.

1.2 Data Overview

Our data was generated through the NBA statistics website by the National Basketball Association. Statistics of the players, teams, and individual games are all recorded. The data set we are working with is a census because it includes data on the entire population, which are the NBA players and teams between the years 2003-2022. These participants were all aware of their data being collected for each game, as it is stated in their contracts.

In regards to the granularity of the data, our data has a high level of granularity as we have data for each player. There were four csv files we utilized in our models.

1. *games.csv*
    - This data contains all the games from 2003 - 2022 and includes game IDs, points, home team, visiting team, and other details.
    - Each row represents a game that was played from 2003-2022.
2. *games_details.csv*
    - This data set contains all the details from each game between the years 2003 to 2022. It also includes all the statistics of players per game such as points, blocks, assists, and score fluctuation.
    - Each row in this data set represents a player for a team and for a given game.
3. *ranking.csv*
    - This data set includes the rankings of each team divided by the Western Conference and Eastern Conference. We utilized the win percentage column in this data set.
    - Each row represents a team's ranking on a given day during the years 2003 to 2022.
4. *teams.csv*
    - This data set contains the data for each team, such as their ID, and location.
    - Each row in this data set represents a team in the NBA.

1.3 Data Limitations

It is worth noting that there were seasons the NBA games were disrupted due to outside factors. For instance, in 2011, the NBA was shut down due to a lockout of the players by the NBA's owners, which greatly decreased the amount of games played as many were canceled. In the second half of 2019, COVID-19 affected the 2019-2020 NBA season. What was left of the ongoing season was postponed indefinitely and many games were cancelled.

Because of these irregularities in our data, we excluded the years 2011, 2019, and 2020 from our analysis.

1.4 Data Cleaning and Pre-Processing

The tables we mainly used were games.csv, games_details.csv, ranking.csv, and teams.csv.

For our Non-parametric Methods and GLMs, the most important table that consisted of all the overall stats needed were called merged_stats_home, and merged_stats_away. We created them in a step-by-step process starting with taking the games table and grouping by the season and home_team_id/away_team_id respectively while aggregating on the mean. Then we grouped our ranking table by the team_id and the standingsdate also while aggregating by mean. Then we filtered the table to only have stats for when a team had reached 82 games (this unfortunately also eliminated the 2011, 2019 and 2020 seasons for reasons mentioned above). The next step is to take the team name for each team_id from game_details and merge that with our grouped ranking table. With these four tables, we merged on the season and team_ids to create the two main tables necessary.

For our Bayesian Hierarchical Model, we merged games_details.csv and games.csv to get the points scored by a team by season, and ranking.csv to get the win rate of a team by season.

For the rankings.csv file, the format of the column for season id was formatted as a string whereas for all the other tables, it was formatted as an int. We had to parse out the string for each season due to potential mismatching of season ids during merging and convert them to integer values.

2 Research Questions

In this project, we proposed two research questions to explore with our data:

1. Can we estimate the win rate of a team in the upcoming season based on previous performances (points scored, win rate, lose rate, etc.)?
2. How does the exposure of a team to an injury affect the amount of points they score?

For our first research question, the real-world question we could answer is what certain statistics are most indicative of a team's success.

OLS is a good fit for this as the model summary will display the coefficients, which can be used to determine which game statistics have the highest predictive importance.

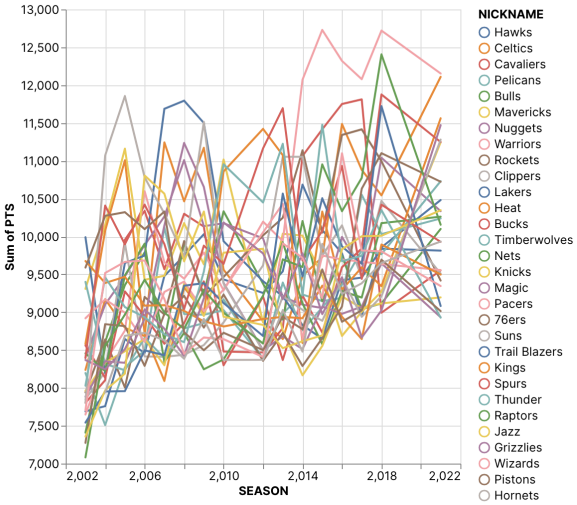
For our second research question, the real-world question we could answer is how significant injuries impact team performance and how teams can mitigate the severity of a player's injury.

Using Bayesian Hierarchical Modeling is a good fit for our second research question because we do not have access to data regarding the difference in team performance if a player has a significant injury. However, we do have access to samples of team performance. From the samples, we can make assumptions about our data and use Bayesian sampling to estimate the difference in team performance.

One of the limitations of Bayesian Hierarchical Modeling is selecting the right distribution for the variables. If our variables and their distributions aren't selected correctly, our model won't produce accurate results.

3 EDA

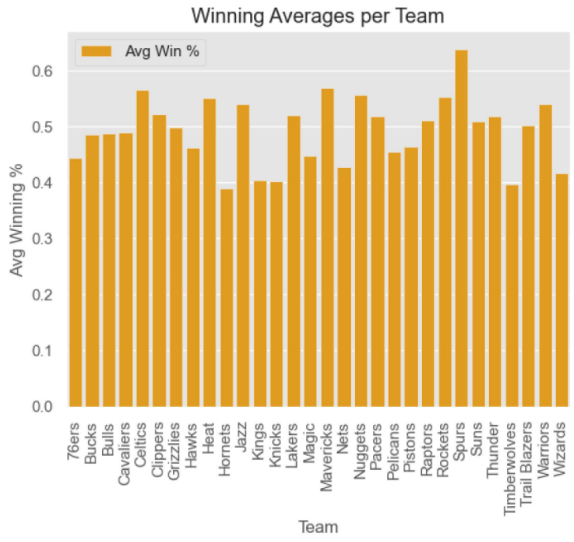
For our first research question, because we're trying to predict the win rate of a team in the upcoming season based on previous team performances, we wanted to initially see the trends of each performance metric from the prior seasons. Looking at the trends would allow us to select for the best model that would fit our data.



Looking at this data, we can see there is seasonality in a team's total number of points scored within a season. Additionally, a majority of teams did not exhibit linear performance when it came to different seasons.

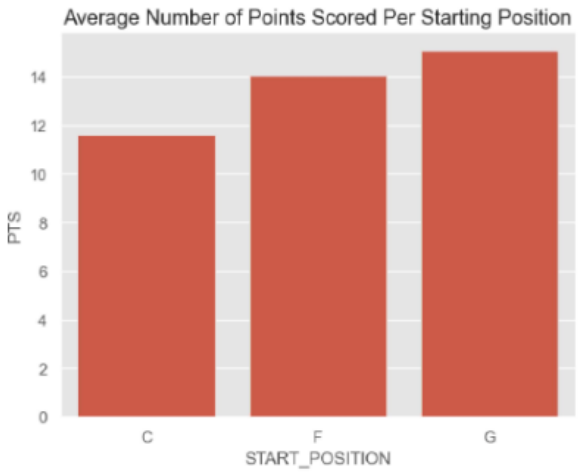
Because of this, we wanted to use a non-parametric model that would best capture that non-linearity in our data. With that, we used a Random Forest Regressor for one of our Regression models.

Next, we wanted to look at the average win percentage by team across all seasons and see if there was any significant difference between them. This would be crucial in understanding if there was any difference in win rate between groups because if there was not, using a model would not be worthwhile.



In the above histogram, we observe at first glance that the San Antonio Spurs have the highest average win rate, while the Charlotte Hornets have the lowest average win rate. Seeing this visualization, we wanted to know if our models would be able to accurately predict that these teams have their respective average win rates.

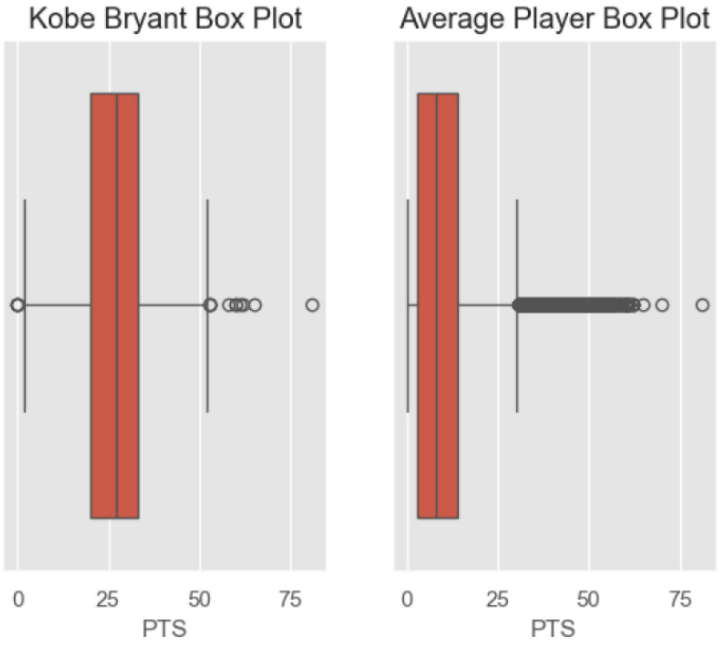
Next, we wanted to consider and explore variables to use for our regression models for predicting win rate. We thought that start position point contribution would be a good predictor for win rate as it gives good insight into a player's performance and contribution towards their team.



In our above bar chart, we noticed that certain positions score more points for their teams. Specifically, we noticed that guards scored more points, on average. From this, we concluded that guards would have the highest contribution towards a team's performance.

For our Bayesian models, this would mean that when guards are out due to injury, a team's scoring potential likely drops, which would show that an injury to a guard may be more important than a center being injured.

Lastly, we wanted to look at how an individual player's points score distribution compared to the overall average points distribution for all players. This would allow us to see which players on average were making more points for their team and vice versa, and would be considered important players.



In this example, we looked and compared the distribution of points Kobe Bryant has scored in comparison to the average player. With Kobe Bryant having a higher average in comparison to the box plot for the average player, it suggests that some players contribute more towards their team's performance. In this case, it would be regarding total points scored.

If players such as Kobe Bryant were to not play, there would be a much bigger impact on a team's performance compared to an average player. Because of this trend, we wanted to explore how big of an impact the absence of these players corresponds to.

This idea guided our research question as well as a general idea for some of the parameters for our model.

4 Methods

We decided to go with Options B (Bayesian Hierarchical Modeling) and Options C (GLM and Non-Parametric Method) for our method selection. Bayesian Hierarchical Modeling would allow us to evaluate the unknown variables we are trying to infer for the number of points a player will score for the next season. Using and predicting with GLMs and Non-Parametric methods would be used to minimize the bias-variance trade off and least squares regressor.

4.1 GLMs and Non-Parametric Method

Our first question aims to predict a team's win percentage using key performance indicators, specifically points, steals, defensive rebounds, offensive rebounds, and assists. These variables are selected based on their relationship with a higher win percentage, reflecting indicators of strong team performance.

For the modeling framework, we adopted an Ordinary Least Squares (OLS) Generalized Linear Model (GLM). The choice of OLS is motivated by the continuous nature of the dependent variable (win percentage), making it unsuitable for models such as Poisson or Negative Binomial Regression designed for discrete outcomes. Logistic regression, tailored for binary classification, is also ruled out given our prediction of a continuous variable.

Our assumptions regarding the absence of linear relationships between predictor variables (no multicollinearity) is significant for the validity of the OLS model.

In exploring non-parametric methods, we implemented a Random Forest Regressor. We chose this specific model as it utilizes multiple decision trees to mitigate overfitting and reduces variance by limiting the depth of splits. We limited the depth to 2 as this is 1/3 of the amount of parameters we plan to use. Careful consideration in feature selection is crucial for our model, as redundant variables such as field goals and points made are purposely excluded to ensure independence and optimize model performance.

After observing the data, we saw it listed which team was the home team and which was the away team. As a result, we decided to split the data into games where the teams were the home teams and games where the teams were the away teams. We assumed that being the visiting team could affect the players' mentality and performance.

The features that we used for the Random Forest Regressor model were points, field goal percentage assists, rebounds, free throw percentage, and three point shot percentage. We are using assists, rebounds, free throw percentage, and three point shot percentage because they affect the performance of the players on the team while not having any dependency on each other. The target feature that we are trying to predict for this random forest model is the win percentage of each team.

We assessed the Random Forest model using the Root Mean Square Error (RMSE). In the case of the Ordinary Least Squares (OLS) Regression, we used the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) for performance evaluation.

This research question strives to contribute a comprehensive analysis to the field of sports analytics, highlighting factors influencing a team's win percentage. Our approach to model selection and feature incorporation emphasizes the validity of our predictive modeling methods.

4.2 Bayesian Hierarchical Modeling

In our Bayesian model, we predicted the performance of each team based on whether or not they had a significant injury, and how it influenced the team's score. We broke up our data by seasons, creating a model for each relevant season.

We defined performance as a team's average score fluctuation value for the entire season, where it represents the net points scored by a team when certain players are playing on average. Higher values represent a higher number of net points scored on average when certain players are on the field, and vice versa.

We considered an injury 'significant' if it keeps a player off the court for more than half of the season. For our model, we are making the assumption that any significant injury rate on a team would be 25%.

Selection of 25% as a parameter in our Bernoulli prior distribution was motivated by our player-specific EDA plot for a player's total points scored in one game. We considered that an injury to a player with high amounts of contribution to their team, such as Kobe, while rather impactful, wouldn't have that high of a frequency.

To construct our graphical model, we took into account the unobserved variables as our parameter of interest that followed a prior distribution. In this case, our parameter of interest would be whether or not a team had a serious injury, along with the performance fluctuations based on players being on or off the court.

The variable whether or not a team had a significant injury is distributed by:

$I_i ... I_n \sim Bernoulli(p)$: whether the i^{th} team has an injury, for n total teams in one season

We opted for a *Bernoulli* distribution for our I_n to differentiate team performance metrics based on the presence or absence of significant injuries.

Next, our variable for performance fluctuations was represented by q_0 and q_1, with distributions:

$q_0 \sim Normal(10, 6)$: score fluctuation if no significant injury
$q_1 \sim Normal(10, 8)$: score fluctuation if significant injury

We chose a Normal Distribution because we had no information about the potential values. Additionally, we assumed that the score fluctuation would be symmetric in that if a team scores a point, the opposing team loses a potential point they could have scored.

For our parameters of q_0 and q_1, we decided to use 10 as the mean to shift our distribution to the right to account for potential negative values of score fluctuations. This is because we used the samples from q_0 and q_1 as our parameters for our observed variables and their respective distributions, which only take in positive parameters. So for our analysis, a parameter of 10 really means that on average teams score a net point of 0 points when certain players are playing.

Additionally, we assumed that teams who suffered a significant injury have the same average score fluctuation value, but with more variance due to the difference in team composition having a much more variable effect on their performance.

As for our observed data/response variables, we chose the total points scored by the i^{th} team and can be denoted as:

$P_i ... P_n$: average points scored per game for the i^{th} team

The other response variable will be the win rates for the i^{th} team:

$$X_i ... X_n: \text{ win rate for the } i^{th} \text{ team}$$

In Bayesian terms, our likelihoods would be:

$$\mathbb{P}(X|I_i, q0, q1) \sim Beta(\alpha, \beta)$$
$$\mathbb{P}(P_i|I_i, q0, q1) \sim Poisson(\mu)$$

with parameters:

$$\alpha = q[I], \text{ where q[I] is sampling}$$
$$\beta = 10$$
$$\mu = q[I]$$

And our priors would be represented as:

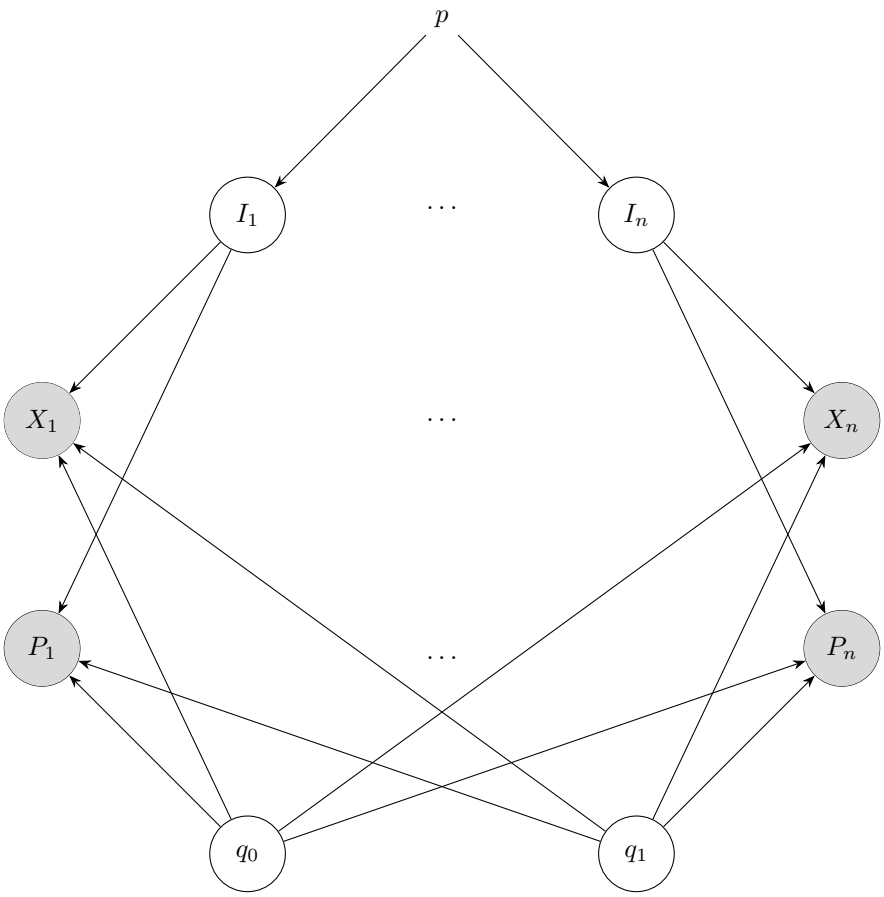
$$\mathbb{P}(I_i, q0, q1)$$

We chose a Poisson distribution to model our P_n variable because the points a team scored per game are discrete, positive values. Additionally, the Poisson Distribution models count, and in this case, the points can be represented as counts.

We chose a Beta distribution to model our X_n variable because a team's win rate can take on the values between 0 and 1, and the support of a beta distribution is between 0 and 1 as well. We said our beta parameter for the Beta distribution

We set $\beta = 10$ due to our performance fluctuations being centered at $\mu = 10$. This is because when our sample $q[I] = 10$, our sample is saying that on average since a team has a net point score of 0, their win rate distribution would be $Beta(10, 10)$, centered around a win rate of 0.5. When the α values are higher than β, meaning their net point average is greater than 0, the distribution of their win rate shifts to the right, increasing the probability of having a higher win rate.

Our graphical model has been drawn out below:


p
I_1
$\cdots$
I_n
X_1
$\cdots$
X_n
P_1
$\cdots$
P_n
q_0
q_1

5 Implementations

5.1 GLM and Non-Parametric Methods

5.1.1 Function Creation

Since we wanted to be able to filter by year and a team's home and away games for our OLS model and the Random Forests, we needed to create functions for each. These functions both take in a year, and whether the data set to include is the merged_stats_home or merged_stats_away. For the OLS model, we are able to simply filter by year then take the statistics we want to use to predict Win Percentage and fit the model/print the summary instantly. The Random Forest function has a bit more nuance, however. We started with the same process as our OLS, which means we filter by the year. Before fitting, however, we also used the train_test_split functionality from sklearn. This means that when we later calculate the RMSE, we have it for both training and test data sets to make sure that we are not overfitting.

5.1.2 Results

With the functions that performed OLS and Random Forest Regression for specific years in hand, now we wanted to choose specific time intervals that would be of the highest importance. Thus we chose 2003, 2021, and over all years as our 3. The reason we chose those specific years to create the visualizations is because 2003 is the earliest NBA season that our data set goes back and 2021 is the most recent complete NBA season it has recorded. Thus we can see if our accuracy differs in the 18 years between the earliest and latest season. Having all data we can see if that is more accurate than just any specific year.

We started by getting the OLS of 2003 home and away games to see the importance of the features we chose (free throws, assists, rebounds, etc.) during this time. Then, we also used OLS on the 2021 home and away games data we created. One thing that is of immediate notice is the difference in importance that the OLS puts on the FG3_PCT (three point percentage) in 2021 vs 2003. In 2021, we can see the coefficient at least for home is 4.850, whereas in 2003 it is -0.8745. This massive difference shows that 3 point shooting is a large reason why teams win in 2021, but in 2003, a good 3 shooting percentage did not matter. This was initially surprising, but we realized that basketball has recently been driven on 3 point shooting compared to 18 years ago, so this is pretty interesting, but understandable to see. Discoveries like these is why we wanted to perform OLS on such far apart specific years. (Figure 1 shows our OLS summaries)

We decided to use the Random Forest Regressor model for our non-parametric method to specifically predict the win percentage of each team in 2003 and 2021, while also creating visualizations to compare the difference between the actual win percentage of the team with the predicted win percentage of the team.

In Figure 2 below, we are using seaborn barplots to visualize the actual win percentage and predicted win percentage for each team in the training set when they are the home and away teams in the 2003 and 2021 games respectively. What we can observe from the visualizations is the predicted win percentages are, for the most part, close to the actual win percentage of the team. There are indeed some predicted percentages that are not as close to the actual win percentage, but for the majority of the teams, it is quite accurate.

OLS Regression Results

Dep. Variable:	W_PCT	R-squared (uncentered):	0.966
Model:	OLS	Adj. R-squared (uncentered):	0.958
Method:	Least Squares	F-statistic:	110.2
Date:	Mon, 11 Dec 2023	Prob (F-statistic):	9.03e-16
Time:	05:01:38	Log-Likelihood:	27.145
No. Observations:	29	AIC:	-42.29
Df Residuals:	23	BIC:	-34.09
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
PTS_home	-0.0087	0.007	-1.257	0.221	-0.023	0.006
FG_PCT_home	4.3385	1.574	2.756	0.011	1.082	7.595
AST_home	0.0286	0.011	2.691	0.013	0.007	0.051
REB_home	-0.0040	0.010	-0.418	0.680	-0.024	0.016
FT_PCT_home	-1.0104	0.663	-1.523	0.141	-2.383	0.362
FG3_PCT_home	-0.8745	1.206	-0.725	0.476	-3.370	1.621

(a) 2003 Home Games

OLS Regression Results

Dep. Variable:	W_PCT	R-squared (uncentered):	0.944
Model:	OLS	Adj. R-squared (uncentered):	0.929
Method:	Least Squares	F-statistic:	64.45
Date:	Mon, 11 Dec 2023	Prob (F-statistic):	3.14e-13
Time:	05:01:38	Log-Likelihood:	19.712
No. Observations:	29	AIC:	-27.42
Df Residuals:	23	BIC:	-19.22
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
PTS_away	-0.0109	0.011	-0.969	0.343	-0.034	0.012
FG_PCT_away	2.7102	2.625	1.033	0.313	-2.720	8.140
AST_away	0.0260	0.021	1.252	0.223	-0.017	0.069
REB_away	0.0027	0.013	0.201	0.842	-0.025	0.030
FT_PCT_away	-0.2397	0.801	-0.299	0.767	-1.896	1.417
FG3_PCT_away	-0.4131	1.263	-0.327	0.747	-3.025	2.199

(b) 2003 Away Games

OLS Regression Results

Dep. Variable:	W_PCT	R-squared (uncentered):	0.950
Model:	OLS	Adj. R-squared (uncentered):	0.937
Method:	Least Squares	F-statistic:	75.64
Date:	Mon, 11 Dec 2023	Prob (F-statistic):	2.13e-14
Time:	05:48:05	Log-Likelihood:	21.986
No. Observations:	30	AIC:	-31.97
Df Residuals:	24	BIC:	-23.57
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
PTS_home	-0.0195	0.015	-1.329	0.196	-0.050	0.011
FG_PCT_home	1.8335	3.406	0.538	0.595	-5.197	8.864
AST_home	0.0116	0.019	0.610	0.547	-0.028	0.051
REB_home	0.0123	0.015	0.830	0.415	-0.018	0.043
FT_PCT_home	-0.8335	0.774	-1.077	0.292	-2.430	0.763
FG3_PCT_home	4.5850	2.003	2.289	0.031	0.450	8.720

(c) 2021 Home Games

OLS Regression Results

Dep. Variable:	W_PCT	R-squared (uncentered):	0.944
Model:	OLS	Adj. R-squared (uncentered):	0.930
Method:	Least Squares	F-statistic:	67.81
Date:	Mon, 11 Dec 2023	Prob (F-statistic):	7.30e-14
Time:	05:48:06	Log-Likelihood:	20.435
No. Observations:	30	AIC:	-28.87
Df Residuals:	24	BIC:	-20.46
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
PTS_away	-0.0030	0.012	-0.238	0.814	-0.029	0.023
FG_PCT_away	0.6331	2.754	0.230	0.820	-5.052	6.318
AST_away	0.0031	0.018	0.168	0.868	-0.035	0.041
REB_away	-0.0121	0.013	-0.914	0.370	-0.039	0.015
FT_PCT_away	0.1516	1.079	0.141	0.889	-2.074	2.378
FG3_PCT_away	2.5127	2.670	0.941	0.356	-2.997	8.023

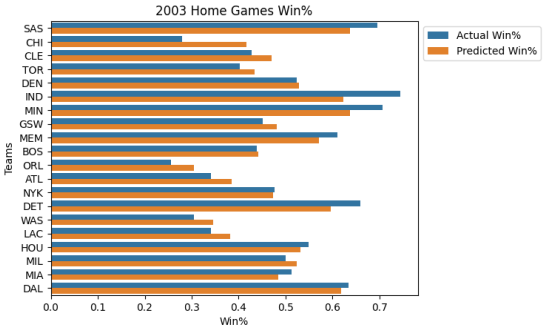
(d) 2021 Away Games

Figure 1: OLS Prediction Summaries

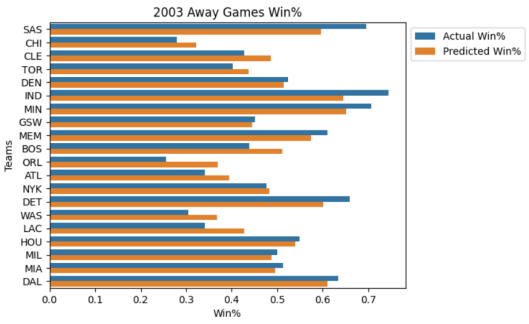
In Figure 3 above, we can tell that these two graphs where we use the data for all years, are significantly more accurate than the graphs for just the 2003 and 2021 seasons. By incorporating data from all years, we're tapping into the wealth of patterns and correlations that might not be immediately apparent in the limited snapshots of individual seasons. These broader datasets, which we created from grouping merged_stats_home and merged_stats_away on the team_ids, enable us to generate these graphs where the bars are closely packed together.

In both 2003 vs 2021 there are differences in the OLS model summaries. The parameters change dramatically (especially FG3_PCT which we talked about above). Even for home vs away games, the OLS model summaries have largely different parameters. For example, there is more negative weight put on a team's FT_PCT at home which could be because free throws at home games are more common.

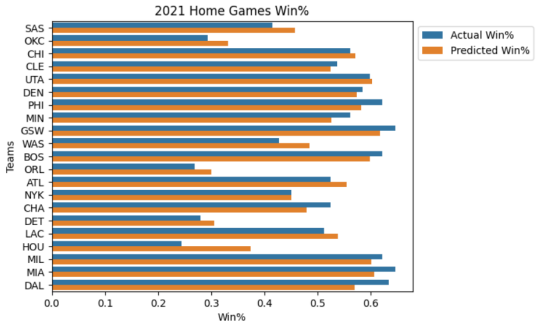
While we thought there would also be a major difference in how our Random Forest predicts a



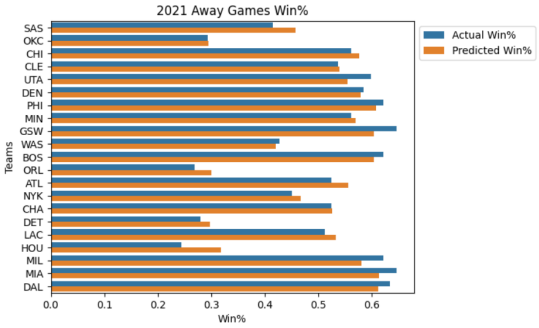
(a) 2003 Home Games



(b) 2003 Away Games

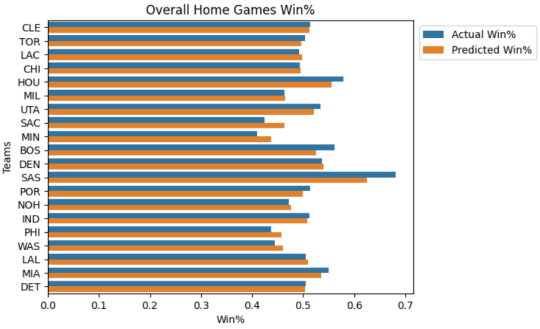


(c) 2021 Home Games

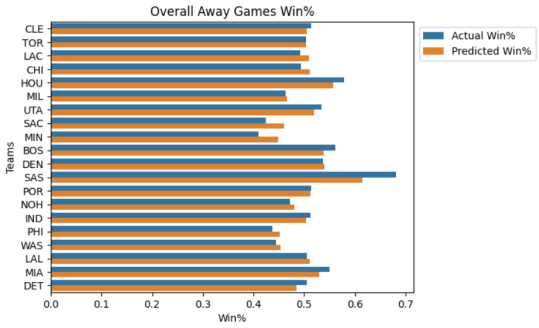


(d) 2021 Away Games

Figure 2: Year Specific Win% Graphs



(a) Overall Home Games



(b) Overall Away Games

Figure 3: Overall Win% Graphs

team's win percentage when they are playing at home vs. away, we see that the predictor is fairly similar (which differs from OLS). In both the home and away graph, the Actual Win% represents the team's Win% for the season overall home and away games. The Predicted Win% only uses the team's predicted home win% in the home charts and vice versa for the away charts. We expected

that the predicted Win% should be higher when we only use home games and should be lower when we only use away games. Instead, we see that both vary from being above or below for different teams and are also fairly accurate.

Back to the OLS model again, we are a little uncertain about how our GLM used the points variable. In Figure 1, we can see that for all the model summaries, the magnitude of the points variable is close to 0 as portrayed by the value in the coefficient column of the OLS summary. This was surprising as we thought that the amount of points a team scores would be the most important coefficient. Another uncertainty is how immense the magnitude of the FG3_PCT (three-point shot percentage) was, as mentioned about above.

5.1.3 Discussion

Because our OLS model uses AIC/BIC to determine how well the model did, and we calculated the RMSE for our Random Forest Regressor, it is hard to compare the two directly.

However, by looking at the graph we have created above, we can tell that, especially for our Overall Win%, the bars are very close and thus it fits the data well. When we also take the RMSE we have calculated into account, we can easily tell it has performed exceptionally. Our test RMSE for the Overall Win% of home games was 0.055. Thus we believe that the better model out of the two is the Random Forest.

We believe that we can use the Random Forest Regressor model on future data for upcoming NBA seasons as well. We saw that our Random Forest model did quite well in predicting the win percentages in 2003 and 2021 as we can observe in their respective bar plots in Figure 2, so, if we applied this model to future NBA data, then this model will most likely be able to predict their win percentage.

An interpretation we made from the OLS model is the three-point shot percentage has a greater importance on the win percentage of a team during a season in more recent years. When observing the OLS summary for 2003, we saw that the three-point shot percentage coefficient was quite low in comparison to 2021, which we believe is due to the greater importance players have put on shooting three-pointers. The way the game has changed since 2003 as so much time has passed.

Data that could have been useful for improving our models would simply be more statistics for each team. All the ones we have right now are important, but things like turnovers could also be helpful. Another thing to note is that we have many offensive statistics (points, assists, etc.) but defensive statistics such as steals or blocks are overlooked and would make our predictions better.

The uncertainty in our results is qualitatively high due to the data we have access to. The dataset size includes data from 2003 to 2021, but it could have included data from before 2003 as the NBA started much earlier. We could have the finished data for 2022 too as well. Additionally, some noisy data that may have been included is due to possible errors in the collection of this data as there may have been mistakes in statistical reporting.

5.2 Bayesian Hierarchical Model

Similarly to the OLS and Non-Parametric models, we decided to focus on the 2003 and 2021 seasons to compare the difference between the beginning of our data set and the most recent complete data set.

5.2.1 Visualizations

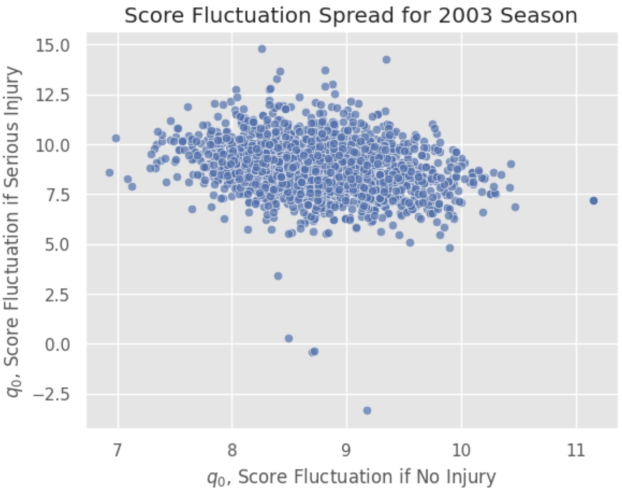


Figure 4: Score Fluctuations - Year 2003

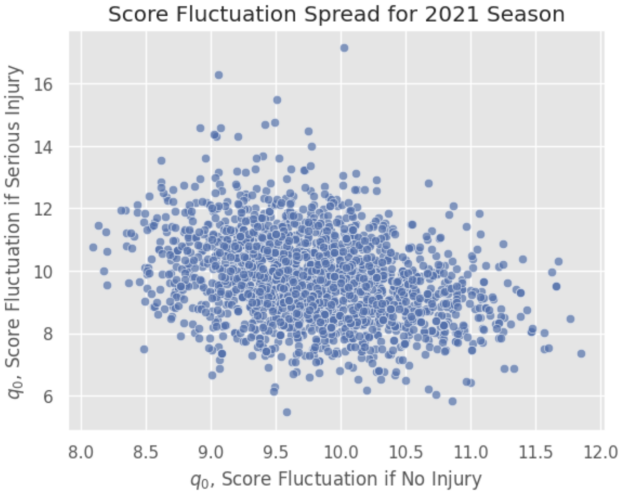


Figure 5: Score Fluctuations - Year 2021

5.2.2 Interpretations

As seen in Figures 4 and 5 above, the scatter plots show a greater spread in average score fluctuation for teams that were classified to have had a significant injury. For teams that were not classified as having a significant injury, their spread was not as significant.

To compare the samples, we looked at their medians, as the means would be rather skewed. When we looked at the median sample values for each q_0 and q_1 for both 2003 and 2021, we saw that q_0 values were greater than q_1 in 2021 and less than in 2003:

```
2003, Uninjured (q0): 8.719547561855169, Injured (q1): 8.818698873545376
2021, Uninjured (q0): 9.727016589549532, Injured (q1): 9.691724730814807
```

From this, we concluded that on average, teams without significant injuries have higher average score fluctuation values, meaning they have a higher amount of net points scored.

Calculating our credible interval, we set a 95% interval for both 2003 and 2021 samples. For 2003 with q_0, we calculated that the lower bound was 7.65 and the upper bound was 9.8. Because the range is relatively tight, we said that the estimate for the true parameter of q_0 was relatively precise and the true parameter does not vary that much. For q_1, the lower interval was 6.72 and the upper interval was 11.81. Because this range is a lot wider than the estimates for q_0 due to the unpredictability of a new player on the field, we said that our estimate for q_1 is not as precise as the estimate for q_0, and the true parameter may vary much more.

For 2021, we observed the same patterns. The intervals for q_0 for the lower and upper bound were 8.69 and 10.95, respectively, q_1 was 7.18 and 12.84, respectively.

6 Conclusions

The use of Bayesian Hierarchical Modeling, Generalized Linear Models, and Non-Parametric methods assisted us in predicting factors that affect the teams' performance, as well as quantifying the effect of variables not present in our data.

6.1 Key Findings

6.1.1 GLMs

The way the game is played in the NBA has changed throughout the years, which can be seen in the greater importance the three-point shot has when comparing it in 2003 with 2021.

While the findings provide valuable insights into NBA team performance, their applicability might be confined to basketball or similar sports due to the specificity of the data set and the dynamics inherent to professional sports leagues. However, the methods employed could also be used in other sports data sets with minor changes, and the Non-Parametric approaches could be useful in analyzing performance-driven industries.

The findings highlight the necessity for continuous adaptation in basketball strategies. Teams and coaching staff could use insights on evolving performance indicators, such as the growing significance of three-point shooting, to inform training drills, player recruitment, and game strategies. By adapting to the way the game has changed in more recent times, the teams that adapt well to the times will have a greater advantage in winning.

6.1.2 Bayesian Hierarchical Model

We were able to see the different spreads of the average point fluctuations based on whether a team had a severe injury or not. From our plots, we concluded that teams with any sort of significant injury have a wider spread in average point fluctuation and teams with no significant injury have a more concentrated spread. The interpretation of this data suggests that injuries have an impact on how well a team performs.

The effect of a player substitution varies if a team experiences a significant injury to their roster. The team can either perform better if certain players are benched, allowing for the starters to maximize their performance, or they can perform worse if their star players are benched and more rookies are put into play.

As for a team with no significant injuries, the data suggests that their performance stays more consistent and concentrated in a certain interval compared to teams with higher injury rates. Having a team that puts in the same players each game should have similar performance, and have a little change in their point rates. After analyzing the results of our model, we have credible results to justify team injury having an impact on performance fluctuation.

Concerning injuries, teams could implement injury prevention strategies or optimize player rotations to mitigate the impact of significant injuries on performance.

6.2 Limitations

In our data sets, we did not have access to turnover differential. We believe this would be an important metric to include since it reflects a team's efficiency and success. A lower turnover rate signifies more possession and scoring opportunities, contributing to offensive effectiveness. A

positive differential indicates defensive strength in creating turnovers and disrupting opponents. This metric provides insights into ball control, team cohesion, and predictive power, influencing scoring disparities and game dynamics.

6.3 Future Studies

There are some ideas on how we could build on top of this study. An extension is to use a dataset with more statistics rather than just descriptive statistics and see if that can predict certain team's Win% even better. We could also somehow obtain data on the referee that is in each game and the number of foul calls each ref gives out, because the foul calls referees make can affect who wins the game and the teams' performance.