NBA Games Data

Predicting the Characteristics of Basketball Winners

Minh Bui
Applied Computer Science
University of Colorado Boulder
Boulder, Colorado, USA
mibu4178@colorado.edu

Bryant Hicks
Applied Computer Science
University of Colorado Boulder
Boulder, Colorado, USA
brhi6213@colorado.edu

Cassandra Cohen
Applied Computer Science
University of Colorado Boulder
Boulder, Colorado, USA
joco3931@colorado.edu

ABSTRACT

Our project hopes to answer the simple question: what is the best way to win basketball games in the NBA? This is not a question of what playstyle is most aesthetically appealing to athletes and fans, as that is entirely subjective. In competitive sports, the best way to play is the strategy that wins you the most games, not the playstyle that you personally prefer. Professional sports is the closest we may ever get to a true meritocracy, so we intend to discover the objectively best way to play basketball according to wins and losses.

CCS CONCEPTS

• Data Mining

• Machine Learning

• Applied Computer Science

KEYWORDS

Basketball, NBA, Statistical Analysis

ACM Reference format:

Minh Bui, Bryant Hicks and Cassandra Cohen. 2022. NBA Games Data: Predicting the Characteristics of Basketball Winners.

1 Problem Statement/Motivation

All NBA teams face the same problem: how do we win more games? In order to answer this question, we want to build a statistical model that can predict win rates in the NBA with the greatest accuracy possible. Our group intends to accomplish this by utilizing data mining tools such as market basket analysis, Apriori, and Bayesian classification. By searching for the stats that correlate highest with win rates, it is our hope that we will find the strongest predictors of winning basketball.

Throughout the NBA's history, there have existed platitudes like "defense wins championships" and "you can't live and die by the three." These beliefs are based on subjective opinions, not verifiable research. Anecdotal evidence is not statistical data, so we intend to either prove or debunk some of the NBA's most commonly held wisdom.

We hope to make novel discoveries within the NBA's team and player statistics. They say that the best predictor of future behavior is past behavior. To predict who will win the most games in the future, we need to discern which teams and players have won the most games in the past. This information will also help us answer the question of who are the most dominant players and teams?

Everyone has a different argument for who is the Greatest of All Time, but our group hopes to answer that question using empirical data rather than personal opinion.

2 Literature Survey

The NBA itself has conducted extensive research on this subject. The league maintains a vast database of game statistics that can be accessed on their official website: https://www.nba.com/stats [1]

The following is an article citing how data analytics changed the way basketball is played: https://d3.harvard.edu/platform-digit/submission/how-data-analytics-is-revolutionizing-the-nba/ [2] The article talks about how players like Stephen Curry use shooting charts to determine the best locations on the court to take shots from. Curry and the Golden State Warriors are major contributors to the NBA's three point revolution. The article also talks about how teams use fatigue data to predict and prevent player injury.

The next article is about the importance of injury prediction in the National Basketball Association: https://sportsanalytics.berkeley.edu/articles/is-injury-prediction-the-next-moneyball.html [3] A great deal of games are lost every year due to injuries, so teams invest heavily in preventing histories and maintaining player health. Whether in winning championships or improving athlete safety, the NBA is constantly utilizing advanced statistical analytics to make the game better.

3 Proposed Work

Our data collection process has yielded a dataset that contains data on basketball games taking place between 2004 to 2020. The dataset is split up into multiple csv files, so we will need to integrate all of it into a single data frame. The data cleaning process will require us to remove irrelevant data, such as the date on which games were played. We will also not consider player statistics for athletes who participated in an inconsequential amount of games, so as to avoid outliers. There are several benchwarmers throughout the NBA who could become outliers for stats like net plus/minus. We will not consider these players in our data analysis, as there's no reason to inflate the relevance of an individual who hardly played in any games.

We intend to perform exploratory data analysis to calculate correlations between win percentage and every statistical category in our datasets. We will generate graphs comparing stats for players and teams side by side. Our group shall also produce correlation heat maps to demonstrate the impact of each unique stat on win rate.

Our intention is to perform market basket analysis to discover which sets of players competed in the most games together. This information will help us determine which teammates worked best together. It is also a commonly held belief within the NBA that roster consistency is one of the keys to winning basketball. We hope to discover if this is true, or if it's possible to win consistently with high roster turnover.

We plan to use machine learning methods, like Bayesian classification and K nearest neighbor, to create a predictive model of wins in the NBA. We will need to control for both overfitting and underfitting in order to optimize the performance of our model.

Our proposed work is different from the research cited in our literature survey, because it will not be specialized to any one team or player. When NBA franchises study analytical data, they are doing so with the sole purpose of improving their own team. They are searching for information that may help them make player trades or sign free agents. The teams look for knowledge that can help them change their coaching strategies. Athletes use their own player stats to prove their credentials when negotiating their contracts. Sports media considers statistical data when voting on awards such as Most Valuable Player.

Our project will be different because we seek to discover basketball truths without bias. It doesn't matter to us who the best player is or who's the best team. We simply wish to learn the answer to those questions and the reasons why the winning teams and players are so successful. With this knowledge, our hope is to construct a statistical model that will predict the future success of NBA teams with the greatest possible accuracy.

4 Data Set

Our data set is obtained from Kaggle, here: https://www.kaggle.com/datasets/nathanlauga/nba-games?resource=download [4]
The data set contains details on games played between 2004 and 2020. The data includes hundreds of thousands of rows detailing player stats posted in every game. There is also data on the winning team of each game played during the 2004-2020 time frame. This data is organized by home and away teams, so we can also track the home and road records of every team. One particular point of interest will be the home record of the Denver Nuggets due to the high altitude that they play at. All data has been downloaded and is currently saved on a private computer.

5 Evaluation Methods

The results of our data mining project will be evaluated based on the predictive accuracy of our machine learning model. This can be measured using training and test datasets to guard against both overfitting and underfitting. We can also evaluate our statistical model by using it to predict seasons outside of the 2004-2020 timeline. In particular, the full 2021-2022 NBA season has been played since the publication of the original data set. This recent season can provide an evaluation of the predictive capabilities of our data model.

6 Tools

- Jupyter Notebook
- Python libraries (e.g. pandas, numpy)
- Structured Query Language
- Github

7 Milestones

We plan to complete the exploratory data analysis by the time of our next project progress report, as well as to have performed market basket analysis using the Apriori algorithm. After this, we intend to implement more machine learning techniques, like Bayesian classification, to build our predictive data model.

7.1 Milestones Completed

We have accomplished our intended goal of exploratory data analysis. Our group plotted graphs displaying the win/loss records of every team and their point differential. We also created bar graphs comparing the plus/minus of top players versus their individual points scored.

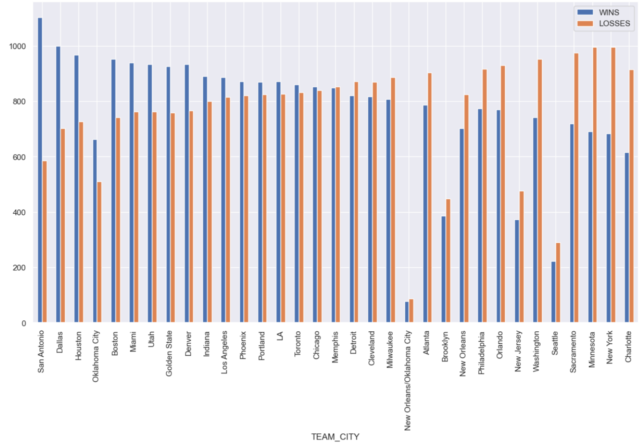
Our team has performed Apriori market basket analysis according to plan. We learned that LeBron James has played in more NBA games than any other active player by a wide margin. Subsets of teammates who played many games together consisted mostly of members of the San Antonio Spurs.

7.2 Milestones Todo

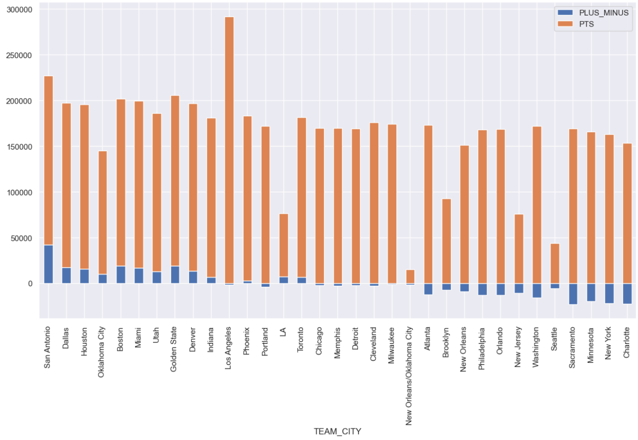
Our group still plans on integrating the player data so that we can track wins and losses by individual players as well as by team. We are still planning to implement a machine learning model that can predict the future performances of teams and players. We hope to evaluate the performance of our model by testing it on recent NBA seasons that were not included in the original database.

8 Interim Results

Below is a graph of every team in the NBA charting their total wins and losses over the recorded time period.



The stacked bar graph below shows every team's total points scored in orange versus their net point differential in orange.

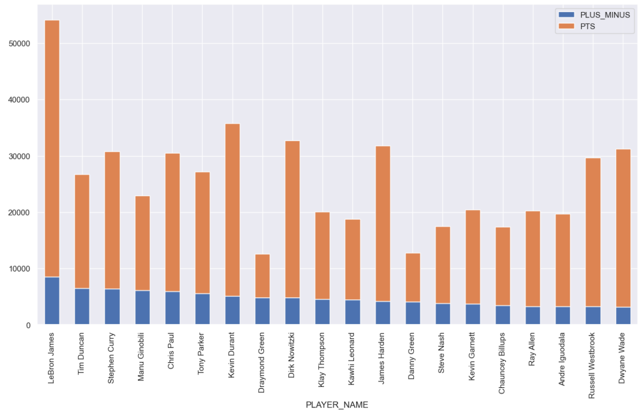


Both of the graphs shown above are sorted by winning percentage, so it's obvious why the teams on the left have more wins and the teams on the right recorded more losses. The second graph clearly shows how net point differential is a far superior predictor of win rates compared to gross points scored.

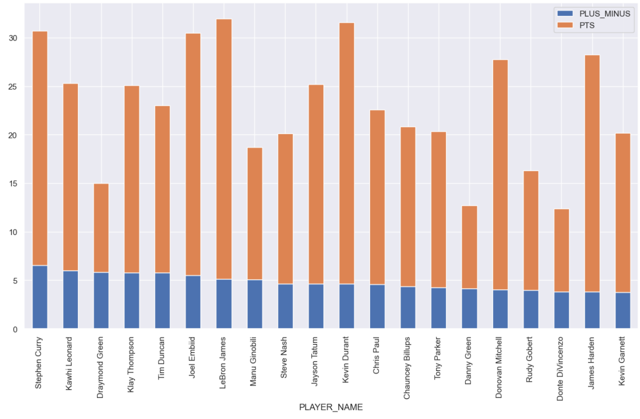
The Los Angeles Lakers are an outlier that proves the old adage that "Defense wins championships." The Lakers scored the most points out of any team in the NBA by quite a wide margin. Despite the productivity of their offense, their team's net plus/minus is still negative because their defense was so porous that opposing teams still managed to outscore them on average.

Another strange outlier is the New Orleans/Oklahoma City team. This franchise used to be called the New Orleans Hornets before relocating and becoming the Oklahoma City Thunder. The city of New Orleans got a new team called the Pelicans. Once the team name Hornets was unused, the Charlotte Bobcats rebranded themselves the Charlotte Hornets. Due to the convoluted way in which franchise names changed hands between these three teams, the data tracking those teams during the transition became very niche.

The plot below lists the top 20 players according to their total plus/minus. Each player's overall plus/minus is depicted in blue and their total points scored is shown in orange.



The following graph shows the top 20 ranking listed in order of average plus/minus and points. Mean plus/minus is shown in blue while points scored per game is displayed in orange.

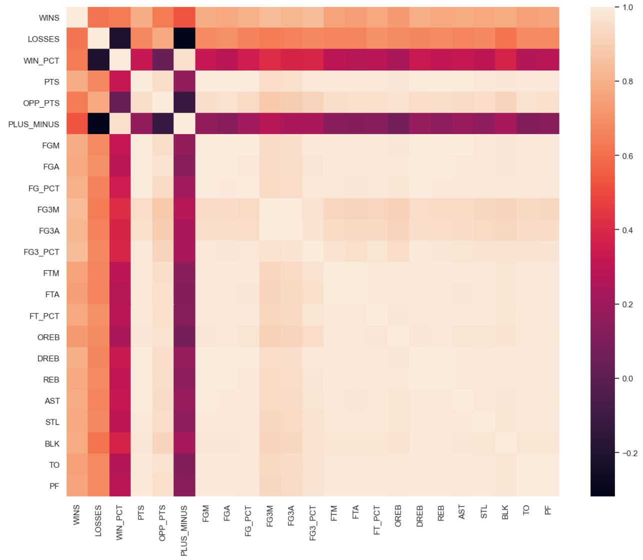


The average plus/minus rankings required data cleaning in order to obtain an accurate listing. When we attempted to generate the list involving every player in the NBA, we found that the top players were random players who no one had ever heard of. It was obvious that the best players in the NBA couldn't be complete unknowns, so we quickly realized our error. By tracking games played according to player minutes, our group found that the highest plus/minus averages belonged to athletes who played in very few games. Due to their low amount of games played, the plus/minus per game average became artificially inflated. To fix this problem, we calculated the average number of games played by athletes across the entire league. Any players whose total games played came below the threshold of league average were expunged by our data cleaning process.

Many of the same players appear in both top 20 lists. One notable difference in the two rankings is how high the Golden State Warriors ranked in the average plus/minus metric. Steph Curry, Draymond Green, and Klay Thompson occupy 3 of the top 4 spots in the rankings. All three of these athletes are also in the total plus/minus rankings, but Curry is the only one within the top 7 of that chart.

The total plus/minus rankings were dominated by the San Antonio Spurs. Their Big 3 of Tim Duncan, Manu Ginobili, and Tony Parker are all within the top 6. Duncan ranked second only to LeBron James who had the highest net plus/minus and the most points scored of anyone in the entire league. LeBron represents an outlier in the first graph because his total overall stats overwhelm everyone else's. But when sorted by average plus/minus and points score per game, LeBron's dominance is brought closer to Earth as his statistics become comparable to those of players like Steph Curry and Kevin Durant.

Below is a heatmap showing the correlation between winning percentage versus every over stat in the database.



The top 3 predictors are rather obvious: winning percentage itself, team plus/minus, and win total. The next three top correlated stats are particularly interesting because they're all 3 point shooting statistics. Three points made, 3 point percentage, and 3 pointers attempted are some of the most strongly positive correlations with win rates. This correlation matrix provides statistical evidence against the old fashioned notion that "Jump shooting teams can't win championships."

The results of our Apriori market basket analysis showed that LeBron James is the individual who appeared in the most NBA games by far with a support of 6.63%. By comparison, the next player on the list was Andre Iguodala with a support of 5.76%. LeBron's support is almost a full percentage point higher than the second ranked athlete, demonstrating just how long and prolific his career has truly been.

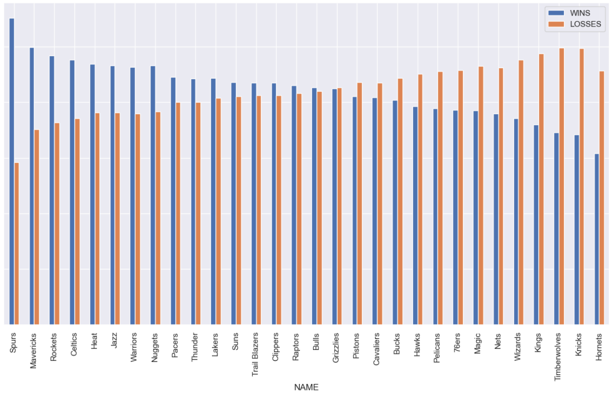
When looking at sets of teammates who played together in a great number of games, LeBron James ironically doesn't appear at all on the list. He has won 4 championships with 3 different teams, which means he hasn't shared the court with any one particular teammate for long. LeBron's former teammate, Dwyane Wade, played far more games with Udonis Haslem than he ever did with James. LeBron James is widely considered the greatest player of our era due to his ability to win championships at the highest level regardless of who his teammates are.

Our market basket analysis showed that the Big 3 of the San Antonio Spurs played by far the most games together as teammates. In both 2 player and 3 player sets, different combinations of Duncan, Ginobili, and Parker yielded the highest support and confidence numbers. The longevity of the Spurs dynasty provides evidence for the belief that roster stability can produce sustainable success for a team. San Antonio won 5 championships during the Tim Duncan era. They also won an astonishing 65% of the games they played during the timeline of our dataset.

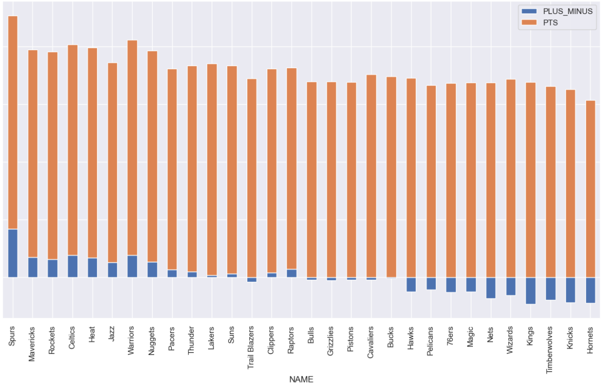
9 Final Results

To perform the remainder of our analysis, we needed to integrate the data that was originally contained in multiple csv tables, so that players could also be sorted by game wins in addition to teams. The data integration was achieved by importing the original csv data tables into a SQLite database. Once the SQL database was created, we were most notably able to perform join operations between the tables to track game wins to specific players among other things.

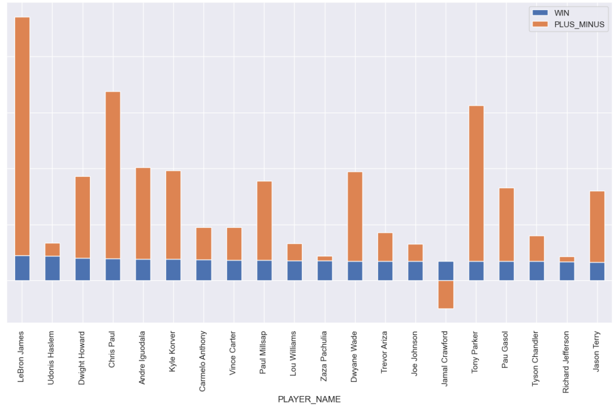
Below is an updated graph of the plot which tracks wins and losses for every team in the NBA.



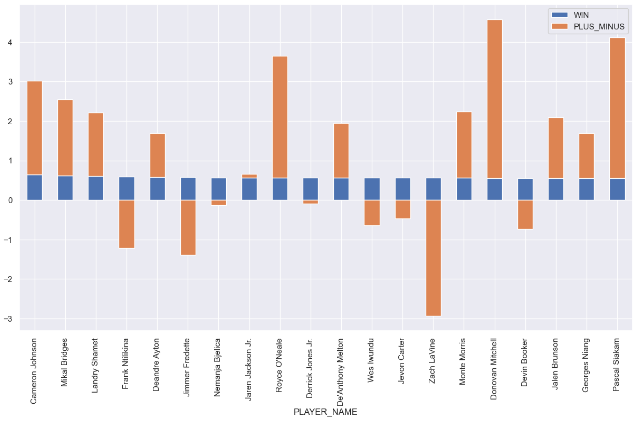
Los Angeles and New Orleans/Oklahoma City are no longer outliers because after the data integration the teams were now being tracked by Team ID instead of city. The previous numbers for Los Angeles were exaggerated due to the fact that LA has two teams: the Lakers and Clippers. The team that transitioned from New Orleans to Oklahoma City is now being accurately tracked by its unique ID. The result is our graph now follows a much more normal distribution fitting a bell curve. Below is a graph showing every team's points scored versus their plus/minus.



The teams are sorted in order of highest winning percentage, so it becomes clear from this graph that a team's net point differential holds a strong correlation with its win rate. Strong teams like the Spurs have a highly positive plus/minus while weak teams like the Knicks have extremely low negative plus/minus. Next is an image of our graph depicting a ranking of the players with the top 20 win totals.



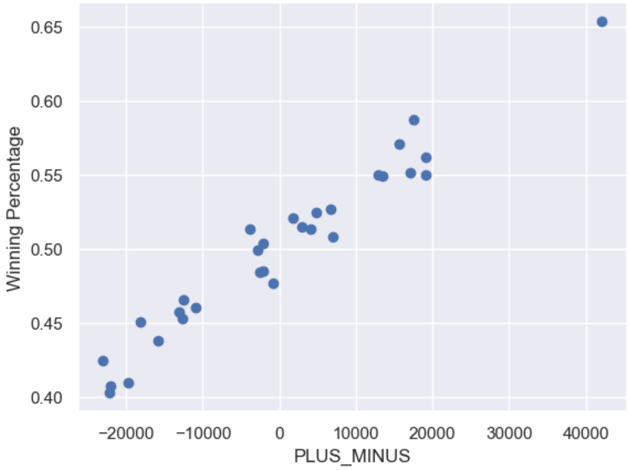
Surprisingly, this chart isn't nearly as useful as our previous plot that ranked players according to total plus/minus. That top 20 list consisted of athletes who are some of the best in the NBA. This win total ranking somehow has Zaza Pachulia ranked above Dwyane Wade. There is no way that Pachulia is a better player than Wade, yet he apparently won more games than Wade over the course of our timeframe. The graph below depicts a similar ranking of players comparing average win rates to net plus/minus.



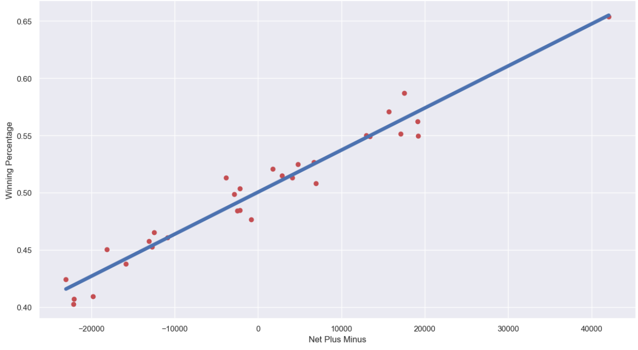
This top 20 list makes absolutely no sense at all to any knowledgeable basketball fan. LeBron James is nowhere to be found on this list despite being on every ranking prior to this one. Frank Ntilikina is ranked fourth according to average wins per game, but no one even knows who he is. The plot also tracks each player's mean plus/minus in orange and several of these players have negative scores. Zach LaVine in particular has a net plus/minus of almost -3 which means that when he's on the floor, opposing teams are outscoring his team by almost 3 points.

It was extremely interesting to learn that a player's plus/minus score is actually a better indicator of his true ability than the amount of games he's won. Intuitively, basketball teams value winning above all else, so they place greater emphasis on win totals than point differential. Yet when measuring an athlete's worth to his team, our data has made it clear that plus/minus is a much better indicator than wins.

The reason for this is the fact that wins can be accumulated by players who aren't making meaningful contributions, but plus/minus only measures the team's performance during the player's minutes. For example, Zaza Pachulia was able to win a lot of games simply due to the fact that he happened to play for the Golden State Warriors. Dwayne Wade, on the other hand, has a far superior plus/minus because when he's on the court, his team is outscoring their opponents by a significant margin. Shown below is a scatter plot mapping the relationship between plus minus and winning percentage.

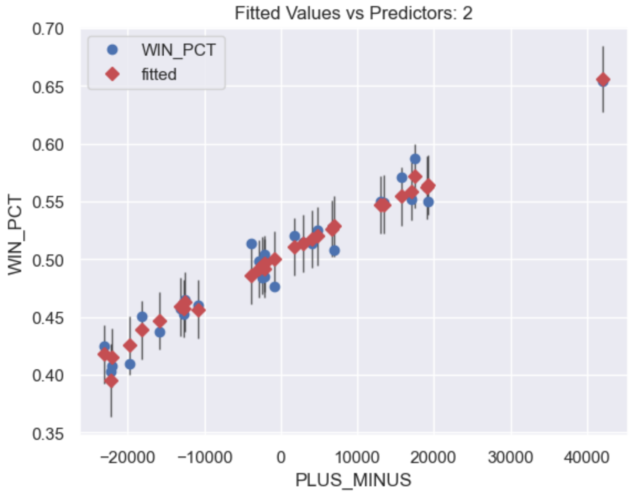


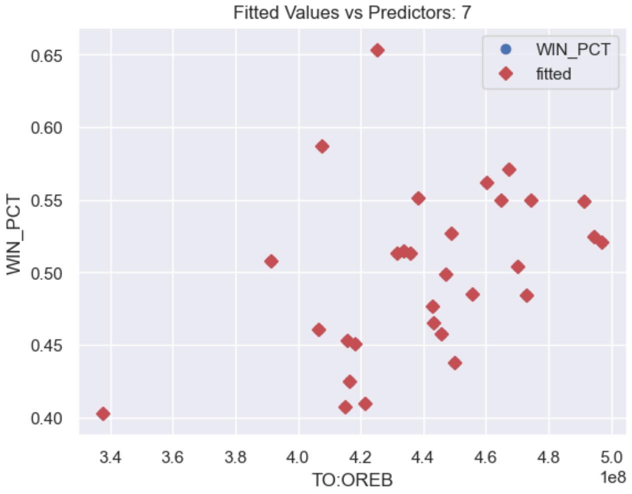
It's easy to see that there's a positive linear correlation between net point differential and win rates. To find more correlations, our group used decision trees to calculate the information gain of wins split against every stat in our database. Plus/minus provided the highest information gain by a wide margin. All other statistics had an average information gain of about 0.163, while plus/minus had a score of 0.2. This gave us further evidence that net plus/minus is a powerful predictor of wins in the NBA. Displayed next is a visualization of our linear regression model using plus/minus as a predictor of winning percentage.



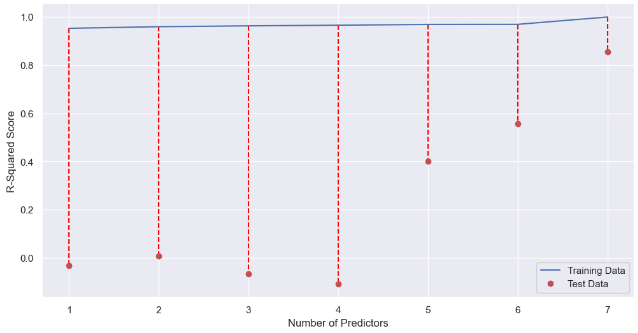
The blue line represents our model's prediction while the red dots show the true data points. The dots of this graph are the same as the scatter plot shown previously. We ran linear regression models for every basketball statistic to see how they compare to plus/minus as a predictor. The R-Squared score for plus/minus came in at an astonishing 0.95 while the next best predictor was opponents' points at 0.77. Plus/minus yielded the highest R-Squared score by a massive margin, but it was interesting to learn that opponent's points scored was the next best predictor. The fact that a defensive stat had the second highest R-Squared provides further evidence in support of "defense wins championships."

Next we did multilinear regression using multiple predictors simultaneously instead of just one. We found that the model could predict the original data with near perfect accuracy with the 7 following predictors: plus/minus, turnovers, offensive rebounds, blocks, free throw percentage, free throw attempts, and steals. Below are graphs of the 2 predictor model compared to the final 7 predictor model.



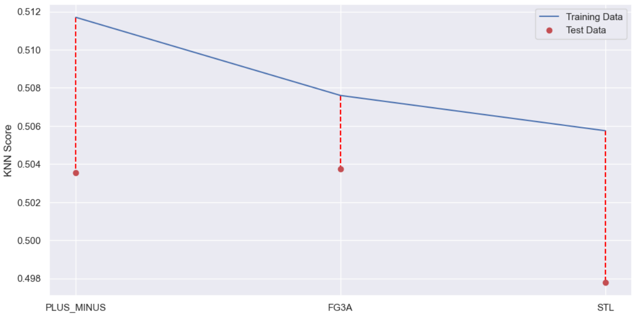


The blue dots represent the true values, while the red dots display the model's prediction, and black lines help visualize the difference between prediction and truth. The 2 predictor model came fairly close to predicting the real data, but the 7 predictor model fitted the true data almost perfectly. To evaluate our model we generated test data from half a million randomly selected samples from the original data. The following is a graph showing the results of our data testing.



The blue line represents our multilinear model's original training score and the red dots show the test results with dotted lines to help visualize the difference. Our results varied because the test datasets were generated randomly, but many of our tests turned out like this one. 7 predictors generally yielded the highest test scores, while models using fewer predictors tended to produce lower test scores. This would suggest that models based on a small number of predictors are underfitted. Complex models with many predictors seem to fit the random test data better.

Our next step was to create a KNN classification model and observe its performance compared to our linear regression models. KNN models take much longer to compute, so we had to create training data sampled from the original data. We took 20,000 random samples of the source data because that amount gave us fairly consistent results while allowing KNN to run in a reasonable time. We used the same number of random samples to create a separate test dataset to evaluate our classification model and the results are shown below.



Like the previous plot, the blue line represents the training score and red dots are test scores. In this instance, plus/minus yielded the highest training score, but 3 point attempts produced the best test score. Exact scores will vary because both the training and test sets were generated randomly. A general trend in the scores that we found was that compared to our regression models, KNN classification yielded worse training scores but better test scores. The results of our machine learning models suggest that linear regression with a high number of predictors is more accurate, but KNN classification is less susceptible to overfitting and underfitting.

10 **Applications**

The knowledge gained from our data mining can be applied to the NBA in terms of player signings and coaching strategy. Teams will often give big contracts to athletes with strong offensive statistics because most teams want players who can help them score. Many of these high scoring athletes turn out to be overpaid because they have other weaknesses in their game, like a lack of defense. With our data, teams might start prioritizing players with strong plus/minus because that stat is the best predictor of how much the individual helps his team win. Bench rotations and on court play calling may also change because coaches might start prioritizing defense more, instead of focusing only on offensive production.

ACKNOWLEDGMENTS

Thanks to the NBA for providing such in depth detail on their games data. Thank you to Kaggle for producing the datasets we are using for our project. We also want to acknowledge our classmate, Charles Dorman, for sending us the two basketball statistics articles cited in the Literature Survey section.

REFERENCES

[1] NBA. 2022. https://www.nba.com/stats

[2] Petra. 2020. *How data analytics is revolutionizing the NBA.* https://d3.harvard.edu/platform-digit/submission/how-data-analytics-is-revolutionizing-the-nba/

[3] Dillain Saparamadu. 2020. *Is Injury Prediction the Next Moneyball?* https://sportsanalytics.berkeley.edu/articles/is-injury-prediction-the-next-moneyball.html

[4] Nathan Lauga. 2022. *NBA Games Data: Dataset with all NBA games from 2004 season to December 2020.* https://www.kaggle.com/datasets/nathanlauga/nba-games?resource=download