Predicting NBA Outcomes

Members: Rosario Chiovaro, Nathan Davis, Daniel Eschman, Jill Mantel, and Vishnu Nistala

Introduction (motivation)

As avid sports enthusiasts, the members of this group brainstormed a simple question, "What is the most impactful sports-related question that we could effectively answer?" We mulled long on the realities of cost, effort, and time before concluding that predicting total wins or season rank for a given NBA team for their current season would be a worthy endeavor. For NBA teams, season performance may be the most crucial factor in the process of determining how they should arbitrate, sign free agent players, and make trades with other teams. For instance, a team projected to win around 45 games and be on the edge of the playoffs may be more likely to swing a major deal for an impact player to push them higher in the standings, compared to a 65-win team which is already a well-oiled machine. Furthermore, a team's wins in the current year intuitively have a linear relationship with points scored in the current year and points scored by this year's opponents. If our research is convincing it could have an impact on analytics research and development in some organizations.

Problem Definition (Provide a precise formal problem definition, in addition to the jargon free version Heilmeier question #1)

Though teams commonly compete to win a championship, success in a playoff scenario is often random and unpredictable. Many of the decisions made by major league sports teams would be made much simpler if decision makers knew precisely how many regular season games the team would win during the current season depending on the decisions made. At the beginning of this project, we hypothesized that an NBA team's number of wins can be accurately predicted as a function of various performance measures gathered from the current season. As such, teams should spend time attempting to predict wins in the regular season, using simple and explainable methods, instead of attempting to predict a championship victory.

As the project continued, we expanded the idea of team success to include team rank for a given season as an alternative to total wins. While predicting a ranking instead of wins requires less precision, it still ultimately reflects how well a team performed in comparison with other teams. Such a prediction could therefore offer decision makers similar details to inform team building.

Proposed Method

Intuition (why is it better than the state of the art)

Many sports teams around the world are increasingly relying on analytics to evaluate players and project their season performance. This analysis is primarily limited to a focus on the likelihood of making the playoffs or likelihood of winning a championship. In contrast, our approach focuses on projecting total wins or projecting team rank for a given season, as opposed to success in the playoffs or championship games. Additionally, our approach utilizes a straight-forward linear regression model. This model has the advantage of being less complex than more common ensemble methods while also offering a more explainable relationship between attributes and responses. Therefore, an organization may better understand what it would take to adjust their performance as needed.

Detailed Description of approaches: algorithms, UI, etc.

Our algorithms utilize linear regression and are trained on numerous factors in team statistics that we determine are most indicative of success. Namely, we found wins were best predicted as a combination of assists, defensive rebounds, field goal percentage, field goals, free throw attempts, free throws, opponent field goal attempts, opponent two-point field goal attempts, points, rank, total rebounds, and two-point field goal percentage. While the group intended to transform these statistics using principal component analysis, the effects of such a transformation were found to be negligible while increasing overall complexity. Instead, these specific factors were chosen by calculating Pearson correlation coefficients for each. By leveraging Tableau, users can easily interact with our predictions in a familiar and visually appealing environment.

Our process begins with a data gathering and cleaning phase. We pull data utilizing the Basketball Reference branch of the Sports Reference API. We also have a downloaded file of final standing with total wins per team, that we will use to get the total number of wins per team. Upon completing the data cleaning and manipulation we finalize data frames that we then ran data analysis on. Our complete data frame that we used for training and testing consisted of 54 columns and 90 rows. Our post-Pearson tables consisted of 14 columns for rank and the wins prediction utilized 9 columns (both still using all 90 rows). Those 90 rows were split into training and test data sets. Our final test set for the 2020 season is 30 rows (one per team). All our data took up approximately 51.6 KB of disk space.

Our algorithms and models are trained on multiple years of team statistics to achieve the goal of predicting total team wins and final ranks in each season. Using those prediction outcomes, we create visualizations in Tableau, which offers users a UI to see the predictions compared to the true outcomes.

Experiments / Evaluation

Initial experiment design, as stated in our progress report: “We will be fitting a model on a small subset of data (one or two regular seasons) using linear regression. Then after fitting the model, we will evaluate the accuracy using the following season. For example, we can train the model on the 2018 and 2019 seasons, then use the model to predict the 2020 season. Given the 2020 season has already happened, we would be able to compare our model results with true results. Similarly, we can also experiment to see if regular season statistics can be used to predict playoff results.”

When conducting these experiments in practice, it became clear that additional experiments would be undertaken to better predict wins throughout the season. As such, we would test model accuracy using all available factors against model accuracy using only factors that highly correlate with wins.

When it became clear that predicting wins would prove difficult, the group decided it may be best to conduct another additional experiment. The goal of this last experiment was to find if predicting overall team rank is easier than predicting team wins. In practice, such a prediction would offer decision makers similar information on how well their team will perform in comparison to other teams in the league, though less precision may be required.

Description of the testbed: list of question experiments should answer

Can you accurately predict an NBA team’s win total given team and player statistics from previous seasons?

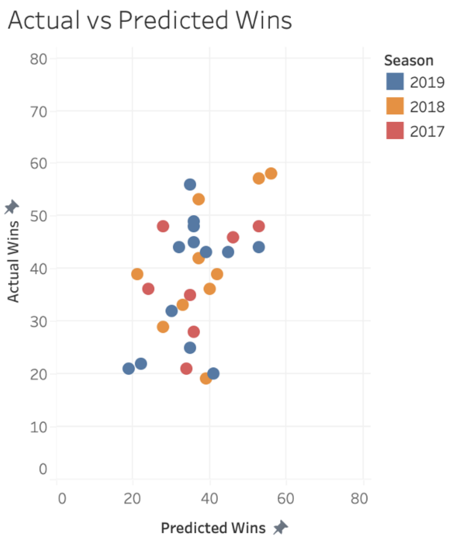
Is a simpler and more explainable model more effective than a model with many attributes?

Is predicting regular season wins easier than predicting a championship win?

Can total wins or team rank be more easily predicted?

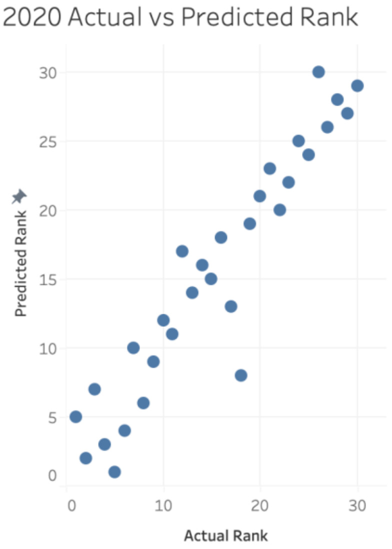
Description of the experiments; Observations

As we began testing models to predict the number of wins and rank, using two previous seasons as a train set, and one as a test, we did not get accurate results. Specifically, the results were correct less than 5% of the time. This was the case using all variables, and only using the high correlation variables found when calculating Pearson correlation coefficients. When we discussed options on why this may be happening, such as if we needed more test data or if there may have been another reason. We decided that we did not need to bring in more seasons, and this was likely due to the roster turnover. Therefore, we changed our approach by splitting the data randomly.



Following the change in approach, our model became more accurate, though not as accurate as we had hoped. Using the high correlation variables, the number of wins was correct 33.3% of the wins were within 3 of the actual value. That was still higher than using all the variables, which got an accuracy of 20% within 3 wins. Clearly our experiments for predicting the total number of games won were not accurate. Since our primary goal was to predict team performance compared to other teams, we turned more directly to that.

As for predicting rank, we were much more successful with that model. 71.6% of our ranks were correctly predicted within 3 ranks. As you can see in the visualization on the right, we have a linear representation of actual vs predicted wins. Unfortunately, the same cannot be said for our wins count prediction.



As a final experiment, we ran our models on future data to see if past seasons could predict future outcomes. Our wins model was only 3.3% accurate, and our ranks model had accuracy of 6.6%. This does not bode well for past statistics predicting future outcomes. This, again, is likely due to roster turnover. However, even given the low accuracy of the ranks, there is a clear linear pattern in the plot graph

Users of our interface can interact with the predictions we have prepared for multiple seasons across all teams. The interactive experience in Tableau offers a familiar and pleasant interface where they can think through the predictions in detail. A limited information-to-ink ratio appropriate color choice, and special consideration to chart choices would allow decision makers with this information to clearly understand the results.

Conclusions and discussion

Unlike we hypothesized, this group has found that an organization's wins cannot be accurately predicted as a linear function of various statistics gathered throughout the current season. However, we did find that predicting season rank to be more effective than predicting wins. Further, a simpler model did, in fact, prove more accurate than our model that utilized as many statistics as we had access to.

Since our hypothesis can be safely rejected, this group no longer would feel comfortable advising decision makers to chase wins rather than championships. Though this philosophy has not been proven not to work by these experiments, our conclusions do support the use of a predictive model in the executive decision-making process. Organizations that hope to improve total wins should seek players that produce statistics which positively correlate with wins. Rebounds, free throw attempts, and field goals are of greatest importance in this regard.

Statement of Efforts

All team members contributed equally to the high-level design of this project with Rosario also contributing data collection and cleaning code, Nathan also contributing interactive visualizations in Tableau and assisting in model creation, Daniel also contributing authorship of written materials, Jill also contributing model implementation, and assisting in data collection and visualizations, and Vishnu also contributing poster design.

Literature Survey

(1) This article is useful because it studied the same idea as we are looking into, but for a different sport. The data in this study was all collected at the beginning of a season, and while it was accurate, we could improve by seeing how the predictions are affected as more data is added throughout the season.

(2) This article focuses on basketball players. This paper could be useful for our project, because it does predict events happening like we will. But there are inconsistencies in the model results received that we could improve upon. For instance, their model does not take foul shots into consideration.

(3) This article focuses on interactive searches using trajectories instead of words to help find specific plays. While this paper is much more advanced than what we will be able to do in the short amount of time we have, it does consider other statistics that could be taken into consideration when looking at schedules.

(4) This article presents a data mining approach to predicting outcomes of head-to-head games rather than player and team-based stats. It may be useful for our project because rosters of NBA teams vary from year to year, you cannot rely completely on team-based statistics to forecast results.

(5) This article presents an approach of predicting the outcome of hockey games utilizing various statistical and machine learning methods upon player and team performance metrics. It will be useful in helping to identify what makes a metric indicative of wins and losses. A potential shortcoming will be the fact that this study was done on hockey not basketball.

(6) This study analyzes the correlation between individual player statistics and overall team performance at the NBA level. This article should be particularly useful as we try to answer the same question via similar methods. A limitation to this study is that the model only takes one input (player efficiency rating).

(7) Heumann describes a common way that teams project their wins: as a function of scores for and against. Their described methods of predicting wins could prove to be a valuable baseline by which we hope to improve upon.

(8) Akhil Nimmagadda focuses on how statisticians predict scores during a cricket match. The methods described in this article, namely Multiple Variable Linear Regression, seem to show promise in predicting scores. They may be useful for predicting the final score of games which could predict wins and losses.

(9) Chu clearly identifies the question that team managers ask of analytics – will spending money on an analytics team lead to team success in playoffs or championship games? The conclusion reached demonstrates that championships are hard to predict, but wins may not be. If predicting wins is more accurate, this could be an improvement.

(10) Yamamoto expands on the Pythagorean Expectation developed by Bill James, and its adaptation to predicting the standings in a Japanese soccer league. Their described methods could be useful in adapting Bill James' original model to the NBA.

(11) Geng and Hu propose a genetic programming approach to predict the outcome of the Playoffs using the regular season performance statistics of each team. This may be useful using past regular season performance statistics to predict the results of an upcoming season if a player is traded to another team.

(12) Neal looks to predict batting averages for Major League Baseball for the second half of the season using the first half of the season as a predictor. The authors find linear models to be superior to Bayesian estimators for both batting average and on-base percentage.

(13) Sanchez uses supervised regression machine learning analysis to predict individual win shares (estimated number of wins a player produces for their team) of NBA players using other metrics. The methods described could be used to predict which players on which teams contribute most to wins / which players most increase a team's probability of winning any given game.

(14) Kvam and Sokol present a combined logistic regression/Markov chain model for predicting the outcome of NCAA tournament games given only basic input data. The model has been significantly more successful than the other common methods over the past six years' tournaments. Their methods could be particularly useful in helping us create a more accurate model to predict NBA games.

(15) A logistic regression model is built to predict matches results of Barclays' Premier League season 2015/2016 for home win or away win and to determine what are the significant variables

to win matches. Despite it being based on a different sport, the methodology used to make predictions can be a particularly useful reference for us to improve our prediction quality.

Citations

1. Apostolou, Konstantinos, & Tjortjis, Christos. (2019). Sports Analytics algorithms for performance prediction. 2019 10TH INTERNATIONAL CONFERENCE ON INFORMATION, INTELLIGENCE, SYSTEMS AND APPLICATIONS (IISA), 469–472.
2. Oldham, Matthew & Crooks, Andrew. (2019). Drafting Agent-Based Modeling Into Basketball Analytics. 1-12. 10.23919/SpringSim.2019.8732893.
3. Sha, Long & Lucey, Patrick & Yue, Yisong & Wei, Xinyu & Hobbs, Jennifer & Rohlf, Charlie & Sridharan, Sridha. (2018). Interactive Sports Analytics: An Intelligent Interface for Utilizing Trajectories for Interactive Sports Play Retrieval and Analytics. ACM Transactions on Computer-Human Interaction. 25. 1-32. 10.1145/3185596.
4. Carson K. Leung, Kyle W. Joseph, Sports Data Mining: Predicting Results for the College Football Games, Procedia Computer Science, Volume 35, 2014, Pages 710-719, ISSN 1877-0509
5. Wei Gu, Krista Foster, Jennifer Shang, Lirong Wei, A game-predicting expert system using big data and machine learning, Expert Systems with Applications, Volume 130, 2019, Pages 293-305, ISSN 0957-4174
6. Yang, Yuanhao (Stanley), Predicting Regular Season Results of NBA Teams Based on Regression Analysis of Common Basketball Statistics, University of California at Berkeley, 2015
7. Heumann, Jay. ‘An Improvement to the Baseball Statistic “Pythagorean Wins”’. 1 Jan. 2016 : 49 – 59.
8. Akhil Nimmagadda, Nidamanuri Venkata Kalyan, Manigandla Venkatesh, Nuthi Naga Sai Teja, Chavali Gopi Raju. "Cricket score and winning prediction using data mining." International Journal of Advance Research, Ideas and Innovations in Technology 3.3 (2018). www.IJARnD.com.
9. Chu, David P. and Wang, Cheng W. ‘Empirical Study on Relationship Between Sports Analytics and Success in Regular Season and Postseason in Major League Baseball’. 1 Jan. 2019 : 205 – 222.
10. R. Yamamoto and Y. Yamamoto, "Predicting the J1 League Standings Using Pythagorean Expectations," 2021 19th International Conference on ICT and Knowledge Engineering (ICT&KE), 2021, pp. 1-4, doi: 10.1109/ICTKE52386.2021.9665702.
11. S. Geng and T. Hu, "Sports Games Modeling and Prediction using Genetic Programming," 2020 IEEE Congress on Evolutionary Computation (CEC), 2020, pp. 1-6, doi: 10.1109/CEC48606.2020.9185917.
12. Neal, Dan, Tan, James, Hao, Feng and Wu, Samuel S. "Simply Better: Using Regression Models to Estimate Major League Batting Averages" Journal of Quantitative Analysis in Sports, vol. 6, no. 3, 2010. https://doi.org/10.2202/1559-0410.1229
13. Sanchez, Oscar. “Basketball Analytics: Predicting Win Shares - Towards Data Science.” Medium, 7 Dec. 2021, towardsdatascience.com/basketball-analytics-predicting-win-shares-7c155651e7cc.
14. Kvam, Paul, and Joel S. Sokol. “A Logistic Regression/Markov Chain Model for NCAA Basketball.” Naval Research Logistics, vol. 53, no. 8, 2006, pp. 788–803. Crossref, https://doi.org/10.1002/nav.20170.
15. Prasetio, Darwin, and Dra. Harlili. “Predicting Football Match Results with Logistic Regression.” 2016 International Conference On Advanced Informatics: Concepts,

Theory And Application (ICAICTA), 2016. Crossref, https://doi.org/10.1109/icaicta.2016.7803111.