NBA Team Rankings: A Visualization

Team 32: Hari Kumaran, Mridul Gupta, May Kalnik, Pujith Veeravelli, Parth Adhia

Introduction

In our visualizations, we aim to give basketball fans a better understanding of the relative strengths of different basketball teams, specifically NBA teams. Standings based on win-loss records may not give a fan the full picture. For example, it does not show how much better one team is than another, and may be misleading, as some teams may have played against easier teams or are only good during home games. As shown in a paper by David Aldous [Parth 1], a more complete rating system similar to the one used to rate Chess players (generally known as Elo) may be used to show the relative strengths of basketball teams.

Problem Definition

Our objective is to develop a predictive model that accurately quantifies the relative strengths of NBA teams, taking into account not just their win-loss records but also various factors influencing game outcomes. This model aims to calculate the projected margin of victory for any given game, incorporating variables such as team Elo ratings and home-court advantage. The model's predictions will be based on historical data from the FiveThirtyEight NBA dataset, offering both probabilistic outcomes of games and a detailed rating system for team comparisons.

Literature survey

Our research showed us what the current innovations for basketball predictions were. Firstly [May 1] details the 3 current sports data types: data per game, prediction data per game play and data per player. As the aforementioned Elo ratings fit mostly into the last group, we can infer that we should be making expected performance visualizations. While the paper gives good insight into what types of visualizations we should be making, it doesn't detail anything about their implementations or feasibility. There are some existing systems, like the HoopInSight system detailed in [Hari 1], that have enhanced the capabilities of shot maps (graphical representations where shots were taken and made by each team) by implementing visualizations that use novel spatial comparisons.

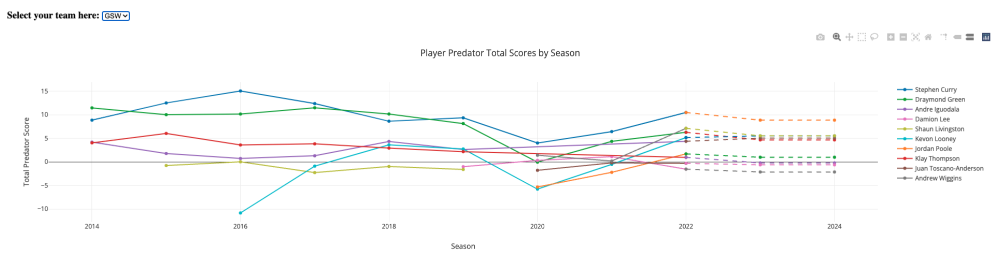
While there were many good existing programs, there was still room for improvement in terms of visualization. Having a solid visualization for NBA predictions is useful for athletes themselves, as seen in [May 2]. Research has found that proper visualizations are pertinent to athletes in improving their skills and motivation. Furthermore, these visualizations, as described in [Hari 2], support sports journalists heavily as they can leverage the means of rapid data exploration through these visualizations throughout the season. This tool is a great help to coaching staff and analysts as demonstrated by [Pujith 3]. A good visualization and algorithm would also help people who bet on the NBA, as knowing which teams are ranked higher than others would aid in making betting decisions [Pujith 1]. Based on this research we knew that creating a quality visualization with good innovation could make a very positive impact on the basketball community.

If successful, our visualization will give clear, unambiguous prediction visualizations that are easy to interpret by coaches, athletes, news reporters and betters alike. The complexity of our visualizations could become a risk and we need to ensure that our plots don't require significant effort to understand and don't present any ambiguity, as shown in [Mridul 1]. Using composite metrics, as presented in [Mridul 3], is a great tool to increase generalizability and readability. However, the use of AI/ML to predict changes in ELO rankings and win probability is a great payoff, as it helps address the problem of data scarcity and adds a potential for real-time analysis, as described by [Mridul 2 and May 3].

Proposed Methods

1. Linear Regression Predictions

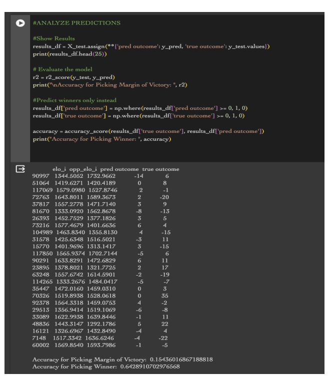
We used Linear Regression to predict the performance (predator scores) of players in each team in the upcoming seasons, 2023 and 2024. This method improves upon the current state of the visualization because it integrates an interactive tool with a predictive model, offering stakeholders insights into not just the past performance but also their future potential. The interaction is dynamic where users can select teams to tailor the visualization according to their needs, making this approach superior to current static models.

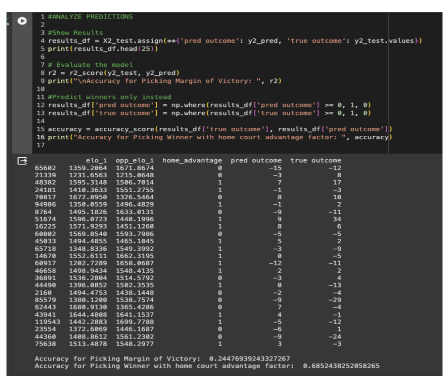


Our approach combines predictive power with user interface design. The backend utilizes a Python Flask server to handle requests and serve the data. The predictive model was tested on both Linear Regression and Random Forest Regressor models implemented using SciKit Library, and Linear Regression was selected based on better accuracy in cross validation tests. We used the players raptor_offense, raptor_defense, war_total, possession minutes, minutes played metrics to estimate their predictive score. The frontend visualization is supported in Plotly.js, where the chart updates to display lines representing each player's performance over the seasons, with future season predictions distinguished by dashed lines. There is also a tooltip feature allowing users to view the exact score and season. This informative visualization helps users make decisions based on the predicted future performance of the players, and enhances their engagement with the data.

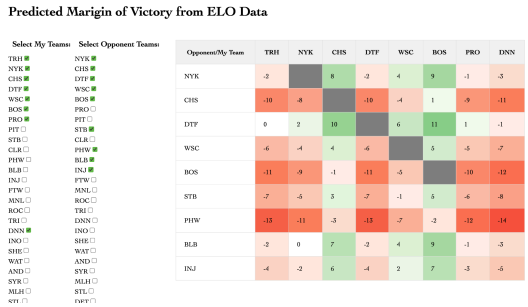
2. Table Model

We used a linear Regression to Predict Margin of Victory (MOV) from ELO Data. We were able to achieve a 69% accuracy for the Binary Classification of Winner/Loser. We then took the ELO data of the most recent game for each team and used that to predict MOV of the next game for all possible team pairings.





We used D3 to create a new visualization from scratch. Our Visualization shows a table that has every possible match up between teams. The columns contain the home team and the rows are their opponents. The data in the cell contains the predicted number of points the home team will score in relation to the opponent team. This visualization is innovative because of its coloring. The cells are colored so that more negative values are deeper red and more positive values are deeper green. This gives the user instant and intuitive information at a glance without having to read each individual number. Because there are over 40 teams, we wanted to make the table more manageable so as not to overwhelm the user. So the user can select which teams they want to see in the table.



3. Player Clustering

The player clustering method aims to categorize NBA players into distinct groups based on their performance metrics, particularly focusing on RAPTOR scores. By clustering players with similar playing styles and skill sets, we can uncover valuable insights into team composition, player roles, and decision making in basketball. Here is how we implemented this:

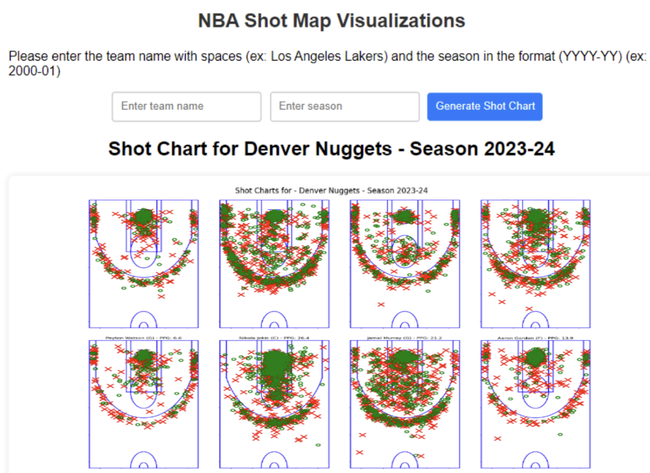
- Selected relevant features from the RAPTOR dataset, such as raptor_offense, raptor_defense, and other performance metrics
- Normalized the data to ensure that each attribute contributes equally to the distance calculations
- Researched and tested the various clustering methods and found that K-means worked the best
    - Used methods elbow method and silhouette analysis to determine the best number of clusters and applied the clustering algorithm to the pre-processed data
    - Found that the optimal number was 3, clustering players into: Balanced, Offensive, and Defensive types

Similarly, for visualizations, we used matplotlib to create PCA and TSNE plots, with each cluster being represented by a different color. In our front end, a user selects a particular NBA season, and then the cluster assignments, statistics, and graphs are displayed. Overall, the novelty of this innovation lies in its ability to categorize players based on advanced performance metrics like RAPTOR scores, which consider both box score data and on/off-court impact. Previously, clustering has only been done based on traditional statistics. Furthermore, our clustering results help the casual NBA follower understand how a player plays. It allows them to see whether their favorite players have more of an impact on offense or defense.

4. Shot Map Visualizations

Another innovation that we improved upon was player shot maps. We were previously able to generate individual NBA player shot maps with the NBA API, as shown in the progress report. We've now expanded on the feature to show shot maps for players on a team by team basis (under a specific criteria described below). Our intuition as to why this is better than the state of the art is its simplicity geared towards a casual audience. Current shot map visualizations, like StatMuse, often provide complex shot distributions relative to field goal percentage or league averages that are geared towards a non-casual audience. Our shot maps not only provide a simple means of understanding where shots are being taken and made, but they display the essential information (key players, position, points per game) towards understanding a team's information. It's easy for this audience to understand offensive preferences by players and teams without much complexity.

As previously mentioned we utilized the NBA API towards retrieving player information given a specific team. Specifically, we filtered players on a team on the criteria of having played at least 41 games (half the season games) and scoring at least 5 points per game. We then utilized the NBA API again towards retrieving the shot map data for these players while utilizing matplotlib towards actually constructing the visualization for the court, the makes (green circles), and misses (red X's). Finally, for the complete visualization we utilized Flask (which hosted our front-end HTML template), in displaying our visualizations pictured below:



Experiments/Evaluation

1. Linear Regression Predictions
Testbed:
- How accurately can we predict future performance (predator_total score) for seasons 2023 and 2024 based on historical data?
- How do the different algorithms (Linear Regression vs. Random Forest) compare in terms of prediction accuracy for this particular dataset?
- How does interactivity affect the user's ability to understand data and make informed decisions?

Experiments/Observations:
- Top feature selection
  - Experiment: Compare prediction results with various player metrics
  - Observations: We experimented with different metrics and finalized on raptor_offense, raptor_defense, war_total, possession minutes, minutes played to predict the final score since they gave the least MSE score for predictions
- Best Predictive Model selection
  - Experiment: Compare accuracy and MSE for Linear Regression and Random Forest
  - Observation: Linear Regression model resulted in an MSE score of 1.3733e-18 while Random Forest Regression resulted in an MSE score of 1.6757
- Impact on user intractability:
  - Experiment: Assess the visualization's effectiveness in engaging users
  - Observation: Users were particularly engaged by the interactive dropdown and specific player selection feature on the line chart that enables them to select individual or group of players. Additionally, they were able to explore future predictions that aren't apparent from the static dataset.

2. Table Model:
Testbed:

- How easily can Margin of Victory between two teams be viewed?
- How good is ELO data as a predictor for Margin of Victory?
- Would this visualization be useful in quickly filling out a bracket (for sports fans)?

Experiments/Observations

- Intuitiveness:
  - Experiment: We completed a user study to see how intuitive it was to use this program
  - Observations: Users in the target demographic found the table very easy to use. They immediately knew what the checkboxes were as well as what the table and its colors represented.
- Accuracy:
  - Experiment: We used Sklearn metrics to evaluate the results of the model
  - Observations: ELO data is good at predicting the winner, having an accuracy of about 70%. However it is less good at predicting the exact MOV. R2 analysis revealed accuracies of about 20%
- Filling out a Bracket Quickly:
  - Experiment: We asked sports fans if this type of visualization would be useful in filling out a bracket for tournaments like March Madness
  - Observations: Fans of basketball agreed that this could be a useful tool for filling out brackets, however, they would likely want to use their qualitative knowledge of the players along with the statistical data represented here to make predictions. Because the model was decent at predicting the winners, the users said the visualization was useful.

3. Player Clustering
Testbed:

- Cluster Cohesion and Separation: How is the cohesion within clusters and separation between clusters?
- Impact of Feature Selection: What is the effect of different feature selections on clustering performance?
- User Engagement: Are the clustering results engaging and interpretable?

Experiments/Observations:

- Cluster Cohesion and Separation:
  - Experiment: Compute silhouette score and WCSS for varying numbers of clusters to identify the optimal number of clusters.
  - Observation: A silhouette score close to 1 indicates that most samples are correctly assigned to their clusters, while a value near -1 suggests incorrect clustering. I observed best results with n = 3.
- Impact of Feature Selection:
  - Experiment: Compare clustering performance using different subsets of performance metrics.
  - Observation: Analyzing clustering performance with different feature subsets provided insights into the importance of individual metrics. With the current subset of features (raptor_box_offense, raptor_box_defense, raptor_onoff_offense, etc), I found that accuracy metrics were maximized and the cluster structures/boundaries looked the most clear.
- User Engagement:
  - Experiment: Get user feedback on clustering results through our interactive user interface
  - Observation: Users noted exemplary comprehension of cluster interpretations, ease of navigation within the visualization, as well as some actionable insights (like displaying statistics).

4. Shot Map Experimentations
Testbed:

- How easily can a non-casual audience draw conclusions from shot maps?
- What is the ideal number of players to display on screen for a given team?
- What criteria determines if a player's shot map should be displayed for a given team?

Experiments/Observations:

- Impact of Player Filtration:
  - Experiment: Compare visualization quantity and quality based on filtering criteria
  - Observation: Certain teams barely had any players to display if the filtering criteria was too strict (greater than 10 ppg and 75% of games played). Alternate criteria apart from scoring often resulted in inconsistent shotmaps in terms of the numbers presented for a given team. We wanted the range to be somewhere between 5-10 players for a non-casual audience to focus on.
- Impact on User Knowledge:
  - Experiment: Measure user assessment of player's scoring threats relative to position on the court
  - Observation: Users noted how certain players exhibited scoring at all three levels (post, mid-range, 3-point), while other users noted player preferences on shot selections (left corner vs right corner for example). They were able to perceive how coaches and strategies could be developed for or against teams using these visualizations.

Conclusions and Discussion:

Together, our methods to analyze and visualize the data allow NBA viewers to get a better and more thorough understanding of the relative strength of different players and teams. The linear regression models created enable users to not only understand current rankings and strengths of different teams, but also the projected standings of teams later in the season based on their future schedule. The table model utilized information and data from sources such as the linear regression model to help users quickly visualize the probability of a team winning a certain game or the statistics of different players on a certain team. The visualizations in the table model give users access to a variety of different information but in a concise and intuitive manner. The player clustering model is another visualization that enables one to see the true effectiveness of players based on different methods. Many times, surface statistics such as total points and shot percentage can be misleading on how valuable a player actually is. Lastly, the shot map experimentations delve deeper into individual player performance to help users to visualize where players may succeed or struggle the most.

Most of the limitations we faced were due to the dataset we used. While the dataset was large and in-depth, there were certain data points that could have led to more effective results. For example, team statistics such as when star players were injured would have allowed us to use more complex models to predict future team and player performance. In addition, even though a player missed a certain shot, they may have gotten fouled on that shot, and having that data would allow our shot map experimentations to be more complete. Overall, future continuations of the models in this paper include utilizing more complex prediction models and gathering more data points. For the project, all team members contributed a similar amount of effort.

Team Effort:

All team members have contributed a similar amount of effort

Citations

Aldous, D. (2017). Elo Ratings and the Sports Model: A Neglected Topic in Applied Probability?. In *Statistical Science*(Vols. 32, pp. 616-629). Statistical Science. https://doi.org/10.1214/17-STS628 [Parth 1]

Boice, Jay, and Ryan Best. “2022-23 NBA Predictions.” *FiveThirtyEight*, 14 June 2023, projects.fivethirtyeight.com/2023-nba-predictions/games/. [dataset]

Bhatt, V., Aggarwal, U., & Kumar, C. V. (2022, December 23). Sports Data Visualization and Betting. 2022 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON). https://ieeexplore.ieee.org/abstract/document/10083831 [Pujith 1]

Casals, M., & Daunis-i-Estadella, P. (2023). Violin Boxplot and enhanced radar plot as components of effective graphical dashboards: An educational example of sports analytics. International Journal of Sports Science & Coaching, 18(2), 572-583. https://doi.org/10.1177/17479541221099638 [Mridul 1]

Cossich, V. R., Carlgren, D., Holash, R. J., & Katz, L. (2023). Technological Breakthroughs in sport: Current practice and future potential of artificial intelligence, virtual reality, augmented reality, and modern data visualization in performance analysis. *Applied Sciences*, *13*(23), 12965. https://doi.org/10.3390/app132312965 [May 3]

Du, M., & Yuan, X. (2020). A survey of competitive sports data visualization and visual analysis. In *Journal of Visualization*(Vols. 24, pp. 47-67). Journal of Visualization. https://doi.org/10.1007/s12650-020-00687-2 [Parth 2]

Fu, Y., & Stasko, J. (2023). HoopInSight: Analyzing and Comparing Basketball Shooting Performance Through Visualization. In *IEEE Transactions on Visualization and Computer Graphics*(Vols. 30, pp. 858-868). IEEE Transactions on Visualization and Computer Graphics. https://doi.org/10.1109/TVCG.2023.3326910 [Hari 1]

Fu, Y., & Stasko, J. (2022). Supporting Data-Driven Basketball Journalism through Interactive Visualization. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3491102.3502078 [Hari 2]

Hung, S., Chen, G., Xu, J., & Gao, X. (2018, July 1). Basketview: Proceedings of the 2018 International Conference on Data Science and Information Technology. ACM Other conferences. https://dl.acm.org/doi/pdf/10.1145/3239283.3239298 [Hari 3]

Kovalchik, S. (2020). Extension of the Elo rating system to margin of victory. In *International Journal of Forecasting*(Vols. 36, pp. 1329-1341). International Journal of Forecasting. https://doi.org/10.1016/j.ijforecast.2020.01.006 [Parth 3]

Legg, P.A., Chung, D.H.S., Parry, M.L., Jones, M.W., Long, R., Griffiths, I.W. and Chen, M. (2012), MatchPad: Interactive Glyph-Based Visualization for Real-Time Sports Performance Analysis. Computer Graphics Forum, 31: 1255-1264. https://doi.org/10.1111/j.1467-8659.2012.03118.x [Pujith 2]

Liu, A., Mahapatra, R. P., & Mayuri, A. V. R. (2021, October 29). Hybrid design for sports data visualization using AI and Big Data Analytics - Complex & Intelligent Systems. SpringerLink. https://link.springer.com/article/10.1007/s40747-021-00557-w [Pujith 3]

Perin, C., Vuillemot, R., Stolper, C.D., Stasko, J.T., Wood, J. and Carpendale, S. (2018), State of the Art of Sports Data Visualization. Computer Graphics Forum, 37: 663-686. https://doi.org/10.1111/cgf.13447 [May 1]

Sarlis, V., & Tjortjis, C. (2020). Sports analytics - Evaluation of basketball players and team performance. In *Inf. Syst.*(Vols. 93, p. 101562). Inf. Syst. https://doi.org/10.1016/j.is.2020.101562 [Mridul 3]

Stusak, S., Tabard, A., Sauka, F., Khot, R. A., & Butz, A. (2014). Activity sculptures: Exploring the impact of physical visualizations on running activity. *IEEE Transactions on Visualization and Computer Graphics*, *20*(12), 2201–2210. https://doi.org/10.1109/tvcg.2014.2352953 [May 2]

Terner, Z., & Franks, A. (2021). Modeling player and team performance in basketball. *Annual Review of Statistics and Its Application*, *8*(1), 1–23. https://doi.org/10.1146/annurev-statistics-040720-015536 [Mridul 2]