B5: Winning in Basketball

Dustin Brunner* ETH Zürich, CS
Jonathan Koch† ETH Zürich, ITET
Timothé Laborie‡ ETH Zürich, CS
Liule Yang§ ETH Zürich, MATH

ABSTRACT

In this paper we present an innovative approach that utilizes interactive machine learning techniques to empower basketball coaches and analysts with a user-friendly dashboard for enhanced strategy development.

We use data from previous basketball seasons to predict the winning probabilities of matchups between two teams. We also provide features to analyse the matchups further and to conduct what-if analyses based on changes in the input data. The results of these machine learning (ML) backed features are then explained using explainability methods such as SHAP (SHapley Additive exPlanations).

This tool is deployed as an interactive web application dashboard that is built with state-of-the-art visualisation techniques and a big emphasis on usability and interactivity. The user workflow is designed to be user-friendly and self-explanatory.

Our target audience includes basketball coaches and analysts at various levels who seek comprehensive and intuitive tools for strategy design, optimizing gameplay, and improving performance.

A combination of backend and frontend technologies is involved in the implementation of the dashboard, utilizing Flask, React, and D3.js. The ML pipeline employs a LightGBM (light gradient-boosting machine) model for training and SHAP values for interpretability. The dashboard components include interactive box score data, winning probability predictions, analysis of similar previous matchups, feature importance visualization, as well as a league overview.

Index Terms: Human-centered computing—Information visualization; Information systems—Information systems applications—Decision support systems

1 INTRODUCTION

In the realm of sports, the strategic decisions made by coaches and analysts play a crucial role in the success of a team. With the advent of modern data science technologies, there is a growing opportunity to leverage these advancements to aid in the design and optimization of team strategies. In this paper, we present an innovative approach that utilizes interactive machine learning (IML) techniques to empower basketball coaches and analysts with a user-friendly dashboard for enhanced strategy development.

*e-mail: brunnedu@ethz.ch
†e-mail: jokoch@ethz.ch
‡e-mail: tlaborie@ethz.ch
§e-mail: liulyang@ethz.ch

2 BACKGROUND AND MOTIVATION

The primary objective of this project is to harness the power of data science in the context of basketball strategy analysis and to make these capabilities available to coaches and analysts in an interactive and intuitive manner. To achieve this, statistical data from previous basketball seasons is used to train a machine learning (ML) model to predict the outcome of future matchups between two teams. An interactive dashboard implemented as a web application then acts as an interface for users to visualize, understand, and interact with the results. This dashboard has been developed with the principles and good practices of interactive machine learning and visualization in mind, such that users with little technical knowledge are able to benefit from it.

Our motivation stems from the desire to bridge the gap between data-driven analysis and practical implementation, ultimately empowering teams to optimize their gameplay and improve performance.

2.1 Target Audience and Users

The target audience for our dashboard is basketball coaches and analysts who are seeking a comprehensive and intuitive tool to aid in strategy design. This includes professionals at various levels, ranging from grassroots basketball programs to elite leagues, who are keen on utilizing data-driven insights to gain a competitive edge. By catering to this audience, we aim to democratize the benefits of advanced analytics and enable teams of all calibres to optimize their performance on the court. During the project development, we conducted interviews with a potential target user, Fran Camba Rodríguez, who is a data analyst at Obradoiro CAB. Obradoiro is a professional basketball team competing in the highest Spanish basketball league.

2.2 Use Cases

We worked out two different use cases where this kind of analytics tool would provide the maximum amount of value to users:

1. In between seasons is when teams and especially their management can have the biggest impact on the long-term success by signing new players and retaining existing key players. Data

and analytics could be leveraged to better understand the impact of potential new as well as existing players in order to optimize signing strategies for the upcoming season.

2. As soon as a game is over during the season, analysts have to start preparing for the next game. This involves working out specific strategies for the upcoming matchup and understanding the strengths and weaknesses of the opposing team as well as working out the chances of winning against it. This could again be facilitated by leveraging data and analytics.

We decided to focus on the second of these two use cases. Fran as our primary potential user encouraged this decision by telling us that he would benefit greatly from such a tool in his day-to-day work.

2.3 Related Work

Plenty of work has already been done in the realm of sports analytics, particularly in the context of basketball. Various researchers have explored different aspects of basketball analytics, including pre-game prediction and in-game prediction.

In one of the earliest contributions to basketball analytics, Oliver Dean proposed the concept of "four factors" to predict winning in basketball games [10]. These four factors include shooting efficiency, turnover rate, offensive rebounding rate, and free throw rate. Dean's work laid the foundation for understanding the key factors that contribute to team success in basketball.

Building upon Dean's work, researchers have delved into pre-game prediction models to forecast the outcomes of basketball matchups. Hu et al. developed a methodology using the weighted likelihood to predict the winners of NBA games [1]. Loeffelholz et al. explored the use of neural network models to forecast the results of NBA basketball games [5]. These pre-game prediction models aim to provide insights into the expected outcomes before the games take place.

In addition to pre-game prediction, researchers have also focused on in-game prediction models that provide real-time forecasts during basketball games. Westfall et al. developed a real-time graphical plot showing the score difference over time in basketball games, enabling easy identification of items of interest such as largest leads or lead changes [14]. Maddox et al. developed a Bayesian framework for in-game prediction of the outcome of NBA games [9]. Song et al. published an in-play prediction model based on the gamma process for the scoring processes of NBA games [11].

We extend this existing work by applying explainability methods to enhance the interpretability of our prediction results. Explainability methods, such as SHAP (SHapley Additive exPlanations), provide insights into the contribution of different features or variables in predicting outcomes. This approach has been shown to increase trust and understanding when making decisions based on machine learning models, especially for users with domain knowledge but a limited background in ML [13], [12].

While the application of explainability techniques in sports analytics is a relatively new area of research, it holds significant potential for providing coaches and analysts with transparent insights into the factors driving team performance. By integrating explainability methods into our interactive machine learning dashboard, we aim to empower basketball coaches and analysts with comprehensible and actionable information for strategy development and performance improvement.

Overall, the combination of our interactive machine learning techniques, explainability methods, and user-friendly dashboard design builds upon and extends the existing body of research in sports analytics, offering a valuable tool for basketball strategy design and optimization.

3 Data

The tool that we developed is based on historical gameplay data from the National Basketball Association (NBA) with all games starting from the 2004 season up to December 2020. This data is freely available from the *NBA Advanced Stats* website [4] and has been compiled into a Kaggle dataset [3], which was used for this project. There are five different datasets in the Kaggle repository:

- **games**: all games with the date, teams, and some aggregated details like the number of points, etc., also called box score
- **games_details**: details of games with all statistics of players for a given game
- **players**: players details
- **ranking**: league standings for each day
- **teams**: team details

In our project, we mainly used the **games_details** dataset containing box score data of each player for each game, the **games** dataset for information on which teams played against each other on which date and the **teams** dataset for meta-information on the teams.

3.1 Preprocessing

To predict the outcome of games, we used the box score of previous games to train our ML model. In order to represent the current ability of each team as well as possible, we only used the previous season's data, which was aggregated to represent the average box score of that team across that season. More specifically, we aggregated the per player per game box scores from the **games_details** dataset over all players in a given team as well as over all games this team played during the regular season. Postseason/playoff games were ignored in order to have a common baseline in the number of games. Additionally, we discarded all the box scores referring to "made" quantities (e.g. Field Goals Made, Free Throws Made, etc.) and only kept the "attempted" quantities (e.g. Field Goals Attempted, Free Throws Attempted etc.). This decision was based on the high correlation between "made" and "attempted" quantities and the fact that "attempted" quantities could be influenced more directly by a coach. All data was precomputed such that the web application could access it with low latency and no computation-heavy aggregation on the fly.

4 Implementation

The technical implementation of our dashboard involves a combination of backend and frontend technologies. The backend utilizes Flask, a lightweight Python web framework, to provide a low-latency interface through application programming interfaces (APIs) connecting the ML model and the data to the frontend application. On the frontend, we leverage React and D3.js to create an interactive and visually appealing user interface that allows coaches and analysts to explore and analyze matchups between teams and their predicted outcomes.

4.1 Machine Learning Pipeline

4.1.1 Light Gradient-Boosting Machine (LightGBM)

As previously mentioned we predict the winning odds of two teams in a matchup based on their aggregate box scores over the previous season. This task can be formulated as a tabular classification task with numeric features. Hence, we decided to use a LightGBM [2] model for this task. LightGBM is a gradient-boosting framework that achieves state-of-the-art performance on tabular classification tasks by utilizing a tree-based learning algorithm with a focus on efficiency and speed. This comes with the added benefit that the LightGBM model also allows for very fast inference, which is crucial

for maintaining low latency, a key requirement for our application. We trained the model on all regular season games from the years 2017, 2018, 2020 and 2021, excluding the 2019 season due to irregularities caused by Covid-19. To evaluate the model's performance, we conducted 5-fold cross-validation, providing an estimate of the model's out-of-sample accuracy. The average validation accuracy achieved was 56%, which surprisingly highlights the difficulty in reliably predicting game outcomes based solely on box score statistics.

4.1.2 SHapley Additive exPlanations (SHAP)

To interpret the LightGBM model, we used SHAP (SHapley Additive exPlanations) values [7], a unified measure of feature importance that is rooted in game theory. SHAP assigns each feature an importance value for a particular prediction based on how much each feature contributes to moving the model output from the baseline prediction, which facilitates the understanding of how the model arrives at its predictions. In particular, we used the implementation of TreeExplainer [6] which is suited best for a decision-tree-based model such as our LightGBM model. For visualizing the individual contributions of each feature on a given prediction we used the force plot [8] visualization method. This approach enables us to better comprehend how the model generates its predictions. However, it is crucial to exercise caution when interpreting predictive models in search of causal insights. Predictive models usually only rely on correlations within the data, which may not necessarily represent causal effects. Therefore, it is essential to avoid drawing definitive causal conclusions based solely on predictive model interpretations. Our interactive dashboard aims to address this concern by having the user take the role of the ultimate decision-maker, leveraging their domain knowledge and expertise.

4.2 Dashboard Components

4.2.1 Interactive Box Score Data

The core piece of the dashboard is an interactive parallel coordinates plot of the box score data of both the home and the away team. Importantly, each separate dimension on the x-axis also serves as a slider that can be moved around freely. When choosing an existing team to start an analysis, the precomputed box score data for this team is displayed in the parallel coordinates plot. However, using the sliders users are able to change the box score data and thereby can simulate what-if scenarios to see how small changes in the box score would impact the predicted outcome of the matchup. The adjustable box score statistics are assists (AST), blocks (BLK), defensive rebounds (DREB), 3-point attempts (FG3A), field goal attempts (FGA), free throw attempts (FTA), offensive rebounds (OREB), steals (STL), and turnovers (TO). In order to provide a reference, we also show the box scores of the 4 teams with the most similar box scores in the background of the parallel coordinates plot.

Also, to inform the users whether they adjust the boxscore data reasonably, the number of possessions [10] for each custom team is calculated accordingly. For each team,

$$Possession = FGA - OREB + TO + 0.4 * FTA$$

4.2.2 Winning Probability

Our pre-trained LightGBM model is used to infer the winning chances of the two teams based on the provided box score data. With every change to the sliders of the box scores, an API call is made to the backend to infer the new winning probabilities based on the updated box scores. This is the main metric for users that they can use to understand how changes in the box score will influence the predicted outcome of a game. The winning probabilities for both teams will always add up to 100% as there is no possibility for a draw in basketball.

4.2.3 Similar Matchups

Though we are providing predictions for the outcome of future games, it is also helpful for users to be able to look back at past games of teams to see how they previously performed. This can also help the user to assess the reliability of the model's prediction and verify it. When users manually adapt the box score of an existing team, we calculate the distance of this new, custom box score to the box scores of all existing teams and provide the user with the closest matching one. That way, even with a custom box score, users are able to look at the closest matching historic matchups and what the outcomes of those games were.

4.2.4 Feature Importance

We use SHAP as an explainability tool for users to better understand why our ML model provides the result it does. The contribution of each input feature to the output is displayed graphically and intuitively, even without any technical knowledge of the underlying theory. Like all the other elements, the SHAP plot also dynamically adapts whenever users make any changes to the input box score data.

4.2.5 League Overview

In the league overview, the offensive performance and defensive performance of a team are calculated and visualized to show how the team performs in the league compared to all other teams. To represent offensive performance, an Offensive Performance (OP) statistic is calculated. The calculation of OP is similar to the calculation of offensive rating. For each team,

$$OP = 100 * \frac{PTS}{possession_{team} + mean(possession_{opponents})}$$

The Defensive Performance (DP) statistic is calculated as

$$DP = 100 * \frac{BLK + DREB + STL}{possession_{team} + mean(possession_{opponents})}$$

which represents how well the team is guarding other teams on average.

4.3 User Workflow

Our typical user is generally not expected to have a technical background nor necessarily an interest to get deeper into it. Instead, they require an easily understandable tool that is self-explanatory but can also provide them with background knowledge if desired.

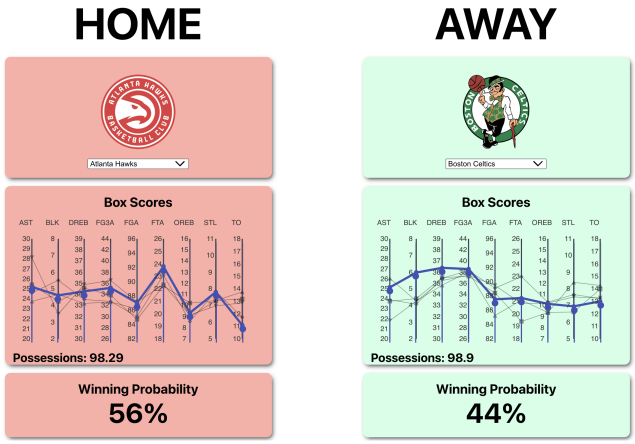

Figure 1: These are the top components of the dashboard. Including the team selection, the parallel coordinates plot of the box scores and the predicted winning probabilities.

For this reason, we have designed the workflow of our application such that first-time users do not see any of the functionality in the beginning except for the dropdown to select the two teams to start an analysis. Only after they have done that will the results show and users are able to interact with the box scores to experiment and explore. There is a tutorial that runs when first opening the application, explaining its features in a concise manner. If desired, help buttons provide deeper background and information about each of the application's components.

The usual workflow for our dashboard is as follows:

1. Select a matchup by choosing a home and away team from the dropdown menus in the top component.
2. Inspect the box scores of the teams and compare them visually.
3. Evaluate the predicted winning probabilities of the teams.
4. Adjust the teams' box scores by moving the interactive box score sliders.
5. Perform further analysis of the predicted winning odds:
   (a) Inspect previous matchups of the selected teams (if the box score wasn't adjusted) or of similar teams (if the box score was adjusted).
   (b) Inspect the SHAP feature importance plot to better comprehend the model's prediction. This includes looking for which features contributed to increasing the predicted winning probability of the home team and which features contributed to lowering it and by how much.
   (c) Inspect the League Overview to get a better understanding of where the selected or custom teams lie within the whole league when it comes to offensive and defensive capabilities.
6. Continue analysis by going back to step 4.

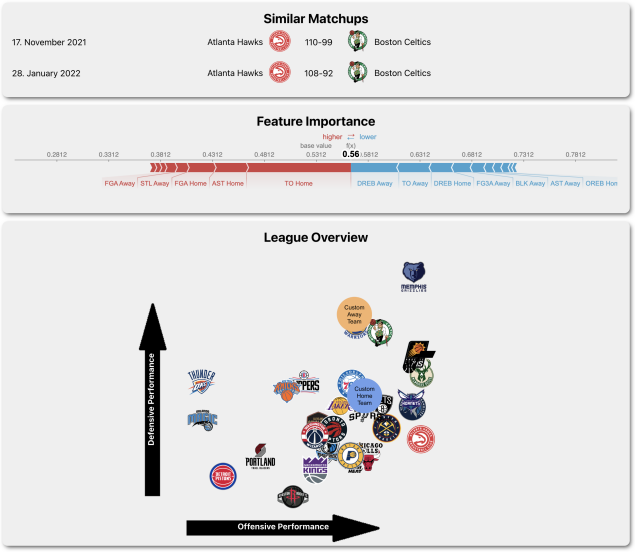

Figure 2: These are the bottom components of the dashboard. Including the display of similar previous matchups, the SHAP feature importance plot and the League Overview scatter plot.

5 Future Work and Conclusion

In conclusion, our dashboard represents a significant step forward in empowering basketball coaches and analysts with data-driven decision-making capabilities. While this paper presents the current implementation and capabilities of our system, there are several avenues for future work. This includes creating a data update pipeline using external APIs to improve the model's accuracy and incorporating data from different leagues to cater to a broader range of users. Additionally, future iterations could focus on developing more detailed player analysis and trend analysis for individual teams. Through ongoing research and development, we aim to continually enhance the capabilities of our dashboard, ultimately enabling teams to unlock their full potential on the basketball court.

Acknowledgments

We would like to express our gratitude to Javier Sanguino as our project supervisor as well as to Dr Mennatallah El-Assady and everyone else involved in offering the XAI lecture at ETH. Thanks are also due to Fran Camba Rodríguez, who as an actual basketball coach provided us with first-hand user feedback. His and Javier's insights, support, and expertise have been invaluable in bringing this project to fruition.

References

[1] F. Hu and J. V. Zidek. Forecasting nba basketball playoff outcomes using the weighted likelihood. *Lecture Notes-Monograph Series*, pp. 385–395, 2004.

[2] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

[3] N. Lauga. Nba games data, 2022.

[4] N. M. V. LLC. Nba advanced stats. `https://www.nba.com/stats`, 2023. Accessed: 2023-05-25.

[5] B. Loeffelholz, E. Bednar, and K. W. Bauer. Predicting nba games using neural networks. *Journal of Quantitative Analysis in Sports*, 5(1), 2009.

[6] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee. From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2(1):56–67, 2020.

[7] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[8] S. M. Lundberg, B. Nair, M. S. Vavilala, M. Horibe, M. J. Eisses, T. Adams, D. E. Liston, D. K.-W. Low, S.-F. Newman, J. Kim, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature biomedical engineering*, 2(10):749–760, 2018.

[9] J. T. Maddox, R. Sides, and J. L. Harvill. Bayesian estimation of in-game home team win probability for national basketball association games. *arXiv preprint arXiv:2207.05114*, 2022.

[10] D. Oliver. *Basketball on paper: rules and tools for performance analysis*. Potomac Books, Inc., 2004.

[11] K. Song and J. Shi. A gamma process based in-play prediction model for national basketball association games. *European Journal of Operational Research*, 283(2):706–713, 2020.

[12] L. Sun, Z. Li, Z. Zhou, S. Lou, W. Li, and Y. Zhang. Towards the conceptual design of ml-enhanced products: the ux value framework and the comlux design process. *AI EDAM*, 37:e13, 2023.

[13] S. Tonekaboni, S. Joshi, M. D. McCradden, and A. Goldenberg. What clinicians want: contextualizing explainable machine learning for clinical end use. In *Machine learning for healthcare conference*, pp. 359–380. PMLR, 2019.

[14] P. H. Westfall. Graphical presentation of a basketball game. *The American Statistician*, 44(4):305–307, 1990.