# B5 - Winning In Basketball

DUSTIN BRUNNER[*], ETH Zürich, Switzerland
JONATHAN KOCH[*], ETH Zürich, Switzerland
LIULE YANG[*], ETH Zürich, Switzerland
TIMOTHÉ LABORIE[*], ETH Zürich, Switzerland
JAVIER SANGUINO, ETH Zürich, Switzerland

In this paper, we present a basketball dashboard designed specifically for basketball coaches and analysts to enhance their understanding of the crucial factors contributing to game outcomes. Our approach revolves around the utilization of a LightGBM model, which enables the prediction of winning odds for two teams engaged in a basketball match based on their boxscore statistics. To further enhance interpretability, we incorporate SHAP (SHapley Additive exPlanations) techniques, allowing for insightful explanations of the model's predictions. Additionally, we introduce a scatterplot visualization that positions teams in terms of their offensive and defensive abilities relative to one another, providing valuable insights into team performance. Through our basketball dashboard, coaches and analysts gain a comprehensive and interpretable toolset to aid in their strategic decision-making processes and overall game understanding.

CCS Concepts: • **Information systems** → *Data analytics*; • **Computing methodologies** → *Supervised learning by classification*; • **Human-centered computing** → **Information visualization**.

Additional Key Words and Phrases: interactive machine learning, data analytics, dashboard, basketball, interpretability, explainability

**ACM Reference Format:**
Dustin Brunner, Jonathan Koch, Liule Yang, Timothé Laborie, and Javier Sanguino. 2023. B5 - Winning In Basketball. In . ACM, New York, NY, USA, 6 pages. https://doi.org/XXXXXXX.XXXXXXX

[*] These authors contributed equally to this research.

## 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.

This project was developed as part of the lecture *Interactive Machine Learning: Visualization & Explainability* held by the *ETH AI Center* at ETH Zürich in the spring semester of 2023.

## 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 of 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 use 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 calibers 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

If you need to take off things, I would readapt this subsection to only show that the use case considered is (2) and not even mention (1). Maybe just mention later in future work that minimal adaptation would be needed for a use case like (1) 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. based on the kind of data we had readily available, which was more appropriate for game-based analytics rather than evaluating individual players. Fran as our primary potential user enforced this decision by agreeing that he would benefit greatly from such a tool in his day-to-day work.

### 2.3 Related Work

Predicting winning has been one of the most prolific perspectives to study the game of basketball using data. In fact, one of the first works [10], Oliver Dean proposed 4 factors to predict winning. The literature can be divided between pre-game prediction [1], [5] and in-game prediction [16], [9], [12]. To our knowledge there exist no previous work that applies explainability techniques in order to explain what wins games in basketball on this kind of models. Nevertheless, on other fields, it has ML models have been used to generate hypothesis between the input and output [11], [15].

In this work, we take this novel approach to merge the fields of user experience design, explainability, and basketball analytics. We have developed a user-friendly dashboard that promotes interactivity and explainability of the user with the ML model. This approach improved trust when making decisions in fields with a domain expert with no ML knowledge [14], [13].]

Our research builds upon prior work in the fields of user experience design, data science, and basketball analytics. Drawing inspiration from existing literature and industry best practices, we have developed a user-friendly dashboard that promotes ease of use and accessibility. By incorporating elements such as step-by-step tutorials and intuitive top-to-bottom design, we aim to enhance user experience and facilitate seamless interaction with the dashboard. Furthermore, we make sure of an advanced machine learning model called LightGBM to provide accurate predictions and valuable insights for strategy development.

## 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 focused on the **games** and **games_details** datasets since the goal was to use data from past games to predict the outcome of future matchups.

### 3.1 Preprocessing

To predict the outcome of games, we used the box score of all previous games to train our ML model. The raw box scores from the **games** dataset were augmented with additional data from the **games_details** dataset to gather additional statistics for each game. 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 across that season. Post-season games were ignored to account for a common baseline 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, etc.) instead. 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 front end technologies. The backend utilizes Flask, a lightweight web framework, to provide a low-latency interface through application programming interfaces (APIs) connecting the ML model and the data to the front end application. On the front end, 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 outcome

### 4.1 Machine Learning Pipeline

*4.1.1 Light Gradient-Boosting Machine (LightGBM).* In our ML pipeline, we use LightGBM [2] to train on the historical box score data of NBA games. LightGBM is a gradient-boosting framework that uses tree-based learning algorithms. We use the model to predict winning.

*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 and their effects. 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] for the LightGBM model and forceplot[8] for visualization.

### 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. 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 pre-computed 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 will impact the predicted outcome of the matchup. For a meaningful application, shooting percentage statistics and field goal made statistics are not used, which provides a realistic scenario for coaches and analysts to design their strategies. The box score statistics data chosen in the end 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).

*4.2.2 Winning Probability.* Our pre-trained ML model based on LightGBM 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 from the model. 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.

#### 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. When users manually adapt the box score from an existing team, we calculate the distance of this new, custom box score to all the 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, this 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, $OP = 100 * PTS/(possessions_{team} + mean(possessions_{opponents}))$. The Defensive Performance (DP) statistic is calculated as $DP = 100 * (BLK + DREB + STL)/(possessions_{team} + mean(possessions_{opponents}))$, 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.

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.

## 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 teaching assistant and all others 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] Feifang Hu and James V Zidek. 2004. Forecasting NBA basketball playoff outcomes using the weighted likelihood. *Lecture Notes-Monograph Series* (2004), 385–395.

[2] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems* 30 (2017).

[3] Nathan Lauga. 2022. NBA games data. https://www.kaggle.com/datasets/nathanlauga/nba-games

[4] NBA Media Ventures LLC. 2023. *NBA Advanced Stats*. Accessed: 2023-05-25.

[5] Bernard Loeffelholz, Earl Bednar, and Kenneth W Bauer. 2009. Predicting NBA games using neural networks. *Journal of Quantitative Analysis in Sports* 5, 1 (2009).

[6] Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence* 2, 1 (2020), 56–67.

[7] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).

[8] Scott M Lundberg, Bala Nair, Monica S Vavilala, Mayumi Horibe, Michael J Eisses, Trevor Adams, David E Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry Kim, et al. 2018. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature biomedical engineering* 2, 10 (2018), 749–760.

[9] Jason T Maddox, Ryan Sides, and Jane L Harvill. 2022. Bayesian estimation of in-game home team win probability for National Basketball Association games. *arXiv preprint arXiv:2207.05114* (2022).

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

[11] Pramila Rani, Changchun Liu, Nilanjan Sarkar, and Eric Vanman. 2006. An empirical study of machine learning techniques for affect recognition in human–robot interaction. *Pattern Analysis and Applications* 9 (2006), 58–69.

[12] Kai Song and Jian Shi. 2020. A gamma process based in-play prediction model for National Basketball Association games. *European Journal of Operational Research* 283, 2 (2020), 706–713.

[13] Lingyun Sun, Zhuoshu Li, Zhibin Zhou, Shanghua Lou, Wenan Li, and Yuyang Zhang. 2023. Towards the conceptual design of ML-enhanced products: the UX value framework and the CoMLUX design process. *AI EDAM* 37 (2023), e13.

[14] Sana Tonekaboni, Shalmali Joshi, Melissa D McCradden, and Anna Goldenberg. 2019. What clinicians want: contextualizing explainable machine learning for clinical end use. In *Machine learning for healthcare conference*. PMLR, 359–380.

[15] Hui-Xin Wang, Laura Fratiglioni, Giovanni B Frisoni, Matti Viitanen, and Bengt Winblad. 1999. Smoking and the occurence of Alzheimer's disease: Cross-sectional and longitudinal data in a population-based study. *American journal of epidemiology* 149, 7 (1999), 640–644.

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

Received 26 May 2023