NOVA IMS Information Management School

MDSAA

Mestrado em

Data Science and Advanced Analytics

LEVERAGING PLAYER-LEVEL INJURY RISK AND CLUTCH PERFORMANCE TO IMPROVE INDIVIDUAL FORECASTS AND TEAM-LEVEL NBA GAME OUTCOME PREDICTION

João Tomás Costa Cristo

Master Thesis submitted in partial fulfilment of the requirements for the degree of Master's in Data Science and Advanced Analytics

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

LEVERAGING PLAYER-LEVEL INJURY RISK AND CLUTCH PERFORMANCE TO IMPROVE INDIVIDUAL FORECASTS AND TEAM-LEVEL NBA GAME OUT COME PREDICTION

by

João Tomás Costa Cristo

Master Thesis presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics, with a specialization in Business Analytics

Supervised by

Márcia Lourenço Baptista, PhD, NOVA Information Management School

July, 2025

STATEMENT OF INTEGRITY

I, João Cristo, hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledge the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Lisbon, 20250701
João Tomás Costa Cristo

Leveraging Player-Level Injury Risk and Clutch Performance to Improve Individual Forecasts and Team-Level NBA Game Outcome Prediction

This document was created with the (pdf/Xe/Lua)LaTeX processor and the NOVAthesis template (v7.3.9) (Lourenço, 2021).

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Professor Márcia Baptista, for her invaluable guidance, insightful feedback, and unwavering support throughout this project. My heartfelt thanks go to my sister, Mariana Cristo, whose constant encouragement and wise advice sustained me during the most challenging moments. Finally, and in the words of Snoop Dogg: "Last but not least, I wanna thank me. I wanna thank me for believing in me, I wanna thank me for doing all this hard work, I wanna thank me for having no days off, I wanna thank me for for never quitting."

”

“You cannot teach a man anything; you can only help him discover it in himself.”

— **Galileo**, Somewhere in a book or speach
(Astronomer, physicist and engineer)

Abstract

Injury risk, player performance, and game outcomes are traditionally modeled as independent domains within basketball analytics. However, their interaction may reveal latent dynamics that influence both short and long-term team success. This dissertation explores the development of an integrated modeling pipeline for the National Basketball Association (NBA), where injury forecasting, individual point prediction, and team outcome classification are structured as modular yet interconnected stages. This research pursues three objectives: (1) to assess the feasibility of short-term injury prediction using structured player data; (2) to evaluate how injury risk and the use of clutch variables affects individual scoring performance; and (3) to determine the predictive power of integrating injury and performance forecasts into game outcome classification. A set of machine learning models, including XGBoost, LightGBM, Random Forest, and Mixture Density Networks, are applied across tasks and evaluated using classification and regression metrics. Feature importance and SHAP value analyses are used to ensure interpretability throughout the pipeline. The findings confirm that short-term injury risk can be predicted with meaningful recall, that such risk negatively impacts expected point production, and that integrating injury and performance data significantly improves the accuracy of team-level forecasts. These results underscore the practical and analytical value of embedding injury-aware intelligence into basketball prediction systems, contributing to the growing body of work in interpretable sports analytics.

Keywords: injury forecasting, basketball analytics, player performance, player clutchness, game outcome prediction

Sustainable Development Goals (SDG):

Resumo

O risco de lesão, o desempenho dos jogadores e os resultados dos jogos têm sido tradicionalmente analisados como áreas separadas no estudo do basquetebol. Contudo, a sua interação pode revelar dinâmicas latentes que influenciam o sucesso da equipa a curto e longo prazo. Esta dissertação explora o desenvolvimento de um pipeline de modelação integrado para a Associação Nacional de Basquetebol (NBA), onde a previsão de lesões, a previsão individual de pontos e a classificação de resultados das equipas são estruturadas como etapas modulares embora interligadas. A investigação tem três objetivos: (1) avaliar a viabilidade da previsão de lesões a curto prazo com base em dados estatísticos sobre os jogadores;

(2) analisar de que forma o risco de lesão e a utilização de variáveis clutch influenciam o desempenho individual dos jogadores;

(3) determinar o impacto preditivo da integração de previsão de lesões e do desempenho individual na classificação dos resultados das equipas.

Um conjunto de modelos de aprendizagem automática, incluindo XGBoost, LightGBM, Random Forest e Mixture Density Networks, é aplicado a todas as tarefas e avaliado com métricas de classificação e regressão. Análises de importância de características e de valores SHAP são utilizadas para garantir interpretabilidade ao longo de todo o pipeline. Os resultados confirmam que o risco de lesão a curto prazo pode ser previsto com recall significativo, que esse risco impacta negativamente a produção esperada de pontos e que a integração de dados de lesões e de desempenho melhora significativamente a precisão das previsões a nível de equipa. Estes resultados sublinham o valor prático e analítico da incorporação da previsão de lesões no basquetebol, , contribuindo para o crescente corpo de trabalho na análise desportiva interpretável.

Palavras-chave: previsão de lesões, análise ao basquetebol, desempenho dos jogadores, previsão de resultados de jogos

Table of contents

List of Figures

List of Tables

Chapter 1

Introduction

Predictive modeling has become an integral component of performance optimization and strategic decision-making in professional sports (e.g., Bunker & Thabtah, 2019; Gudmundsson & Horton, 2017; Rein & Memmert, 2016). In the National Basketball Association (NBA), the growing availability of structured data – ranging from basic in-game statistics and player tracking to injury records and detailed game logs – has enabled the development of increasingly sophisticated models capable of forecasting individual and team performance. These tools are now widely used by teams, analysts, and commercial entities to inform tactical adjustments, manage player workloads, and optimize season-long planning (Steffen, 2022; Horvat et al., 2020).

Despite substantial methodological progress, existing models often remain narrowly focused, frequently overlooking critical contextual factors such as injury status, fatigue, and performance under pressure (Genoud, 2024; Kalkhoven et al., 2024; Sarlis et al., 2024; Groll et al., 2019). Most limit their scope to predicting single-game outcomes or point spreads, frequently omitting key contextual factors that shape real-world team trajectories. Among the most influential and yet underrepresented variables are injury risk, player availability, and clutch performance – defined as the ability to sustain or improve performance under high-pressure game scenarios (Sarlis et al., 2024; Mehrasa et al., 2018). These dimensions are not peripheral; they are central to understanding competitive dynamics in the NBA, where team success often hinges on the health and reliability of core players across an 82-game regular season and subsequent playoffs. Models that fail to account for these elements tend to overestimate performance continuity and underestimate volatility, ultimately reducing their utility for long-term forecasting or strategic decision support (Genoud, 2024; Kalkhoven et al., 2024).

This dissertation aims to address this gap by investigating how the integration of injury forecasting, clutch performance metrics, and contextual game variables can enhance the predictive power and realism of NBA outcome models. The central research question guiding this work is: how can player-level injury risk and clutch performance forecasting be used to improve individual performance predictions and,

in turn, enhance team-level NBA game outcome models?

To address this question, this study pursues three objectives. The first is to develop a predictive model capable of achieving over 80% accuracy in forecasting the winner of an NBA game. This involves constructing a multi-source dataset that integrates official NBA statistics, game schedules, biographical data, shot log events, and historical injury records. The second objective is to understand the long-term impact of player injuries on team success. To this end, the injury data is structured to capture return dates, games missed, and recovery durations, enabling analysis of how prolonged absences affect cumulative team performance. The third and final objective is to investigate the influence of a player's clutchness on team victories. This is operationalized through the engineering of shot-based features that quantify shooting behavior in critical moments, allowing for the measurement of player impact in decisive phases of the game.

The methodological approach is modular. Injury risk is modeled independently using contextual features such as age, physical profile, recent workload, and injury history (Bai & Yang, 2024). Player performance is modeled through shot log data with an emphasis on clutch moments, where outcomes are often decided (Kambhamettu & Shrivastava, 2024). These elements are then integrated into a team game-level prediction pipeline that accounts for contextual factors such as rest days and travel schedules (Zhao et al., 2023). The models are designed to operate in a chronologically consistent manner, ensuring that no future information is leaked into past predictions. To support interpretability, the framework incorporates explainable machine learning techniques such as SHAP, allowing for transparent evaluation of feature contributions (Lundberg & Lee, 2020).

A notable feature of this study is the combination of a large dataset – over 32,000 games – with a set of individual player advanced variables. While prior research varies widely in terms of dataset size and chosen features, this approach uniquely brings together both scale and context-specific indicators, such as shot location and timing.

This dissertation is structured as follows: chapter 2 reviews the existing literature on predictive modeling in basketball, with a focus on injury forecasting, clutch performance analysis, and the integration of contextual variables in outcome prediction; chapter 3 outlines the methodological framework, including data sources, preprocessing techniques, feature engineering, and model development; chapter 4 presents the results, analyzing model performance and examining the influence of injury risk and clutchness on game outcome predictions and chapter 5 presents a critical discussion of the results, analyzing model performance and evaluating the contribution of injury risk and clutchness features to predictive accuracy.

Chapter 2

Literature Review

This chapter presents the theoretical foundations and empirical evidence supporting the use of predictive modeling in sports, with a focus on basketball. It is structured into three main sections. The first explores the historical evolution of predictive modeling across sports. The second narrows the scope to predictive modeling in basketball, emphasizing current limitations and advancements. The final section addresses recent developments in injury forecasting, highlighting their relevance to performance modeling and game outcome prediction.

2.1 Predictive Modeling in Sports

Predictive modeling in sports emerged from sabermetrics in baseball, where statistical innovations provided teams with competitive advantages by identifying undervalued players (James, 1982; Lewis, 2004). This marked the transition from intuition-based scouting to evidence-driven decision-making.

Modern predictive modeling is composed of three components: outcome forecasting, supervised learning from labeled data, and interpretability frameworks such as SHAP. Together, these components allow teams to understand complex datasets, extract actionable insights, and inform strategy.

These methodologies have been adopted across various sports. In football, Random Forests and Gradient Boosting have shown improved accuracy over traditional methods (Groll et al., 2019). In American football, Hidden Markov Models and artificial neural networks have demonstrated strong performance in modeling temporal play sequences (Ötting, 2021; Purucker, 1996).Ensemble methods like RobustBoost achieved high accuracies in ice hockey, even in the presence of sparse data (Gu et al., 2019).

Despite these advances, challenges remain constant across sports. These include the availability of high-quality data, variability in external conditions (e.g., weather or scheduling), and the complexity of human factors like fatigue and psychology. Explainable AI techniques, particularly SHAP (Lundberg & Lee, 2020), have been developed to enhance model transparency and support practical decision-making.

2.2 Predictive Modeling in Basketball: Capabilities and Gaps

Basketball provides a favorable context for predictive analytics due to the availability of granular data and the game's structured format. Early approaches used logistic regression on basic box score data and achieved approximately 75% accuracy during regular seasons, but their performance dropped in the playoffs due to the lack of contextual variables (Loeffelholz et al., 2009).

The integration of machine learning models such as Random Forests, SVMs (Support Vector Machine), and neural networks expanded the range of features considered and improved accuracy (Horvat et al., 2020). Random Forests and XGBoost remain widely used due to their ability to model non-linear relationships and rank feature importance. For example, XGBoost outperformed logistic regression in injury prediction tasks, achieving AUC-ROC scores above 0.84 (Lu et al., 2022).

Recent developments have focused on temporal and context-aware models. LSTM (Long Short-Term Memory) networks have been used to model shot sequences and evaluate team decision-making (Kambhamettu & Shrivastava, 2024). These architectures allow models to capture temporal dependencies that are not observable in static features.

Contextual modeling efforts such as those by FiveThirtyEight, which combine Elo ratings with RAPTOR metrics, demonstrate the value of integrating contextual information like travel schedules and playoff seeding (Silver, 2015). However, their predictive performance remains limited – reportedly around 65.1% accuracy – primarily due to the exclusion of dynamic game events and health-related disruptions (Sarlis et al., 2024; Genoud, 2024).

Studies have also emphasized the importance of clutch performance – defined as efficiency in high-pressure moments – as a critical factor influencing both individual and team success (Sarlis et al., 2024; Skinner, 2012; Chang et al., 2021). These indicators are often overlooked in general predictive frameworks, despite their demonstrated explanatory power.

Game simulations, such as Monte Carlo and Markov models, serve as complementary methods to estimate outcome probabilities. While effective under static assumptions, they struggle with adapting to dynamic elements like injuries or sudden tactical shifts (Steffen, 2022).

Deep learning models based on player tracking data have reached accuracies above 80% by evaluating spatial dynamics and possession patterns (Mehrasa et al., 2018). Graph Neural Networks (GNNs) have recently emerged as a method to model player interactions within teams, achieving accuracies around 85% in game predictions (Zhao et al., 2023). However, most models remain limited by their insufficient treatment of injury-related availability.

Despite growing methodological sophistication, most prior studies rely on limited datasets – often focusing on a few seasons or a restricted number of observations such

as game logs or injury episodes. For instance, some models train on fewer than 2,000 games (Eppel et al., 2022). This restricted scope constrains model generalizability and undermines robustness across player types and game contexts.

In contrast, the present dissertation leverages a comprehensive dataset of over 32,000 NBA games across 28 seasons (1996–2024), used to train both injury prediction models and performance forecasting systems. This broader empirical foundation integrates multiple data modalities – including biometric tracking, contextual game metadata, and historical player load metrics – providing a more generalizable and analytically robust framework.

2.3 Predictive Modeling for Injury Forecasting in Basketball

Injury forecasting has become an emerging research domain due to its relevance for performance management, roster strategy, and financial decision-making (Kalkhoven et al., 2024; Genoud, 2024; Colby et al., 2017). Injuries to key players directly impact game outcomes and team trajectories (Genoud, 2024).

Traditional injury modeling approaches relied on historical exposure metrics such as minutes played. While informative, these methods lack sensitivity to biomechanical stress and real-time physiological strain (Byman, 2023). The use of wearable devices and tracking technologies has improved granularity by capturing metrics like jump intensity, acceleration, and deceleration (Bai & Yang, 2024).

Causal modeling frameworks have been proposed to improve interpretability. Kalkhoven et al. (2024) used causal diagrams to map the interaction between workload, recovery time, and injury likelihood. This approach helps minimize confounding and supports the design of personalized training plans.

Clustering methods such as DBSCAN have been applied to identify injury risk patterns by demographic and positional groupings. These methods have highlighted that nearly half of all injury-related costs are associated with musculoskeletal conditions (Sarlis & Tjortjis, 2024).

Machine learning methods, including Random Forests, XGBoost, and neural networks, remain dominant in injury forecasting due to their robustness and flexibility. Nonetheless, the field faces systemic challenges. First, data consistency is limited across franchises. Second, differences in player physiology require individualized modeling approaches. Third, contextual factors such as travel fatigue and psychological stress are rarely quantified or integrated (Musat et al., 2024).

Finally, the lack of interpretability in deep learning models limits their practical deployment. Techniques such as SHAP help address this by identifying the most influential features in injury risk predictions, particularly those related to workload, fatigue, and prior recovery duration.

Chapter 3

Methodology

The chapter begins by detailing the data sources and preprocessing steps, including feature engineering focused on temporal alignment and performance indicators. It then presents the rationale for using Random Forest, XGBoost, and LightGBM, based on their robustness to real-world data and established use in sports analytics.

Each modeling task is described in terms of dataset construction, model training, and evaluation, with specific metrics aligned to each objective – such as recall for injury forecasting and accuracy for game classification. Throughout, the pipeline ensures temporal integrity and avoids data leakage by using only pre-game information.

This methodology sets the foundation for evaluating how injury risk and performance trends contribute to accurate, actionable predictions in the basketball domain.

3.1 Data Collection

This study analyzes NBA data from the 1996–2024 seasons. Data was collected from the NBA Official API and two public sources: Basketball Reference and the Basketball Transactions Archive. The integrated dataset includes performance metrics, contextual game variables, and injury records. This subsection details the collection and composition of each dataset used.

The NBA Official API provided four core datasets: box score data, game schedule data, player biostatistics, and shot logs.

Box score data includes basic player-level statistics such as minutes played, points, rebounds, and assists. It covers all NBA game events (regular season, playoffs, preseason, and All-Star events) from 1996 onward. Regular season data offers greater volume and temporal consistency across players and teams, providing a more reliable basis for training and evaluating predictive models. Playoff data is inherently selective, typically involving only the top-performing teams and players. This creates a skewed sample where role players receive fewer minutes, star players experience increased workloads, and roster rotations are shortened. These shifts distort standard performance distributions and might introduce structural bias into model training. In

addition, playoff games are characterized by elevated psychological and tactical pressure, which can significantly alter individual behavior and team strategies. Preseason and All-Star games were also excluded, as they do not reflect competitive gameplay conditions, typical player rotations, or standard tactical intensity. Since the objective of this dissertation is to develop a broadly applicable framework for predicting NBA game outcomes and injury risks, limiting the dataset to regular season games ensures that models are trained on representative, evenly distributed data that reflects typical gameplay and player usage patterns over the course of a season. Games were filtered out using the GameID format, which encodes the NBA game event type.

Season-level player aggregates were also retrieved to capture longitudinal trends. This included total and ranking metrics per season and per player. Seasonal ranking metrics represent each player's ordinal position in specific statistical categories (e.g., points scored), with lower ranks indicating higher performance within that season. Derived features such as`last_season_avg_points` served as historical baselines for modeling consistency and reputation.

Game schedule data was used to identify home teams and assign calendar dates to each GameID. These timestamps were critical for generating time-sensitive features – such as rest periods, back-to-back games, and injury recovery timelines – particularly because most other datasets lacked explicit game date information and could not support temporal alignment independently. Notably, the game schedule dataset required no preprocessing, as it was already well-structured and consistent with the official league calendar, allowing for direct integration into the pipeline.

Biostatistics data included static attributes such as player height, weight, and birthdate. These values were used to derive contextual features such as age at each game. Age is a known factor influencing both injury risk and performance variability.

Shot logs contained high-dimensional data for every shot attempt from 1996–2024. Each entry recorded timestamp (period, minutes, seconds), spatial coordinates (LOC_X, LOC_Y), and shot type (Lay-up, mid-range jumper, 3pt shot attempt, among others). This dataset enabled the construction of advanced shot-based metrics, including zone-specific shooting percentages and variability in shot selection. It also supported the quantification of clutchness, defined in this study as performance during the final five minutes of games with a score differential of five points or fewer (NBA.com, n.d.; Engelmann & Skinner, 2020).

Injury records were not available through an official structured database. Data was sourced from the Pro Sports Transactions Archive, a historical repository of NBA injury events. Due to the site's unstructured format, a Robotic Process Automation (RPA) pipeline was developed using Power Automate Desktop to extract and structure the data. The resulting injury dataset included more than 60,000 rows and five columns: player name, entry and exit dates from the injury list, a free-text injury description, and injury date.As no unique player identifiers were provided, a name-matching procedure was applied to link each entry to a corresponding PlayerID from the NBA

API. Although imperfect, this method allowed partial integration with performance and schedule datasets. The processes used to classify injuries by type, affected body part, and estimated recovery duration are described in the following sections.

3.2 Preprocessing and Feature Engineering

This section details the preprocessing steps and feature engineering procedures applied to each dataset. The objective was to transform raw data into a structured, chronologically consistent format suitable for predictive modeling. Each subsection describes the operations performed on one dataset, followed by the derived features and their relevance to the overall framework.

3.2.1 Injury Dataset

The injury dataset required extensive filtering to isolate confirmed injury events from unrelated status updates. Many records in the raw data included vague or non-medical labels such as "day-to-day," "evaluation," or absences for personal reasons, which were excluded from analysis (for instance, funerals).

The primary source of injury detail was a free-text comment field. This field was processed using a custom regular expression pipeline, developed after exploratory text analysis. Two key features were extracted: injury type, classified into eleven categories (e.g., Sprain, Fracture, Generic Injury) and affected body part, initially mapped into sixteen anatomical regions and later grouped into five high-level zones (Lower Body, Arms, Head/Neck, Torso, and Other). This structure enabled consistent encoding of injuries and improved downstream model interpretability.

To support temporal modeling, new variables were added: injury start and return dates. These allowed the computation of recovery duration (in days) and games missed, serving as quantitative measures of injury impact on player availability and team performance.

3.2.2 Shot Logs Dataset

The shot logs dataset was used to generate temporally and spatially aware features for modeling player behavior and game context. Each record contained shot coordinates, timestamps, shot type, and contextual metadata. Feature construction followed strict anti-leakage procedures: all statistics were calculated using only data available prior to each game. This was enforced using the `shift()` operation to maintain chronological integrity.

Features were grouped into five categories: shot efficiency metrics, time-related shot metrics, location-based shot metrics, shot type and style diversity metrics, and clutch-specific indicators.

Shot efficiency metrics quantified success rates by court zone and shot type. Examples include `clutch_fg_pct_alltime` and `rolling_corner_3_ratio_global`, which measure both conversion efficiency and usage frequency in distinct shooting contexts.

Time-related shot metrics captured temporal shot distribution, such as `rolling_q4_shots_global` and `rolling_ratio_q1_global`, which track shooting behavior across quarters and serve as proxies for stamina and usage trends.

Location-based shot metrics described spatial shooting tendencies. Zone-specific ratios (e.g., `rolling_zone_usage_ratio_restricted_area_global`) captured preferred shooting areas. Entropy-based metrics (e.g., `rolling_zone_entropy_global`) measured shot dispersion and variability, while `rolling_game_shot_variability_x_global` and assessed spatial consistency.

Shot type and style diversity metrics assessed offensive capacity, including `rolling_action_type_entropy_alltime`, which captured how varied a player's shot-generating actions are. These metrics quantify whether a player relies on a narrow or diverse set of offensive tools such as layups, dunks, or long-range shots.

Clutch-specific indicators isolated performance under pressure. Variables such as `rolling_clutch_fg_pct_alltime` and `rolling_clutch_shots_made_alltime` focused on the final five minutes of games with a score difference of five points or fewer. These features provide key insights into player reliability in critical moments.

3.2.3 Box Score Dataset

The box score dataset preprocessing included removing ten rows with negative minutes values, as such values are not valid in a physical context. Additionally, due to extraction duplication, all records appeared twice and were deduplicated using a composite key formed from GameID, PlayerID, and TeamID.

3.2.4 Schedule Dataset

The schedule dataset was structurally clean and required no correction. It was used to extract game date and home team identifiers. These variables were essential for time alignment across datasets and for deriving contextual game features such as rest days and home-court advantage.

3.2.5 Biostatistics Dataset

Missing values were present only in the birthdate field. To resolve this, an RPA script was developed using Power Automate Desktop. The workflow retrieved missing dates from the Basketball Reference archive and inserted them into the dataset. This step ensured the accuracy of derived features such as player age at each game instance.

3.3 Model Selection Justification

The main predictive models selected for this study – Random Forest, XGBoost, and LightGBM – were chosen for their strong performance on structured, tabular data and their widespread use in both academic and applied machine learning research (Zhou, 2021; Chen & Guestrin, 2016; Ke et al., 2017). These models are tree-based ensemble methods known for their robustness to missing values and outliers. Specifically, XGBoost and LightGBM can handle missing data natively by learning optimal split directions during training, while Random Forest requires imputation but is generally tolerant of moderate data quality issues. All three are well-suited to real-world datasets where noise and irregularities are common. They are also relatively easy to implement and tune compared to deep learning architectures, making them ideal for projects with time or resource constraints. Additionally, their interpretability via feature importance rankings and compatibility with explainability tools like SHAP (Lundberg & Lee, 2020) further support their use in research that values both performance and transparency.

Despite their strengths, these models are not without limitations. Tree-based ensembles can become computationally expensive with very large datasets or excessive feature dimensionality and may overfit if hyperparameters are not carefully tuned. Moreover, while robust to feature outliers, they can still be sensitive to extreme values in the target variable – a relevant consideration in regression settings such as player point prediction. Finally, their performance may plateau on tasks involving highly sequential or temporal dependencies, where deep learning models (e.g., RNNs, Transformers) may be more appropriate (Lim et al., 2021). However, given the structured nature of the available data, the project's timeline, and the goal of building a transparent and practical modeling pipeline, these models offer a strong balance between accuracy, efficiency, and interpretability.

3.4 Injury Predictive Model

3.4.1 Baseline Dataset Construction and Missing Data Handling

To support injury forecasting, a unified contextual dataset was constructed by merging the box score, schedule, and biostatistics datasets. The merge was performed using the shared keys `GameID` and `PlayerID`, resulting in a consolidated dataset containing 975.513 entries and 43 variables. Redundant or analytically irrelevant columns such as `TEAM_CITY` and `TEAM_ABBREVIATION` were excluded. To ensure temporal and competitive consistency, only regular season games were retained. Matches from preseason, All-Star events, and other exhibitions were filtered out using metadata embedded in the `GameID`. An integrity check on team participation revealed 2.675 malformed game records, each associated with either only one team or more than two teams linked to the same `GameID`. These records were discarded to preserve the assumption of valid two-team configurations per match.

Missing values were handled according to column semantics. For biometric data (`HEIGHT`, `WEIGHT`), missing entries were imputed using the global column mean. For performance statistics (e.g., points, assists, rebounds), rows with null or zero minutes values were interpreted as instances where the player did not participate, and all stats were set to zero accordingly. Residual missing performance values were imputed using player-specific averages across other games.

Player positions were assigned using the most frequent value when available. For players with missing position data but known height and weight, positions were estimated by calculating the Euclidean distance between the player's biometric profile and the average profile of known positions.

The `PLUS_MINUS` column required special handling. Missing values in this variable are not data errors but reflect player inactivity. Because any recorded value – positive, negative, or zero – has clear interpretive value, null entries were preserved to signify non-participation and will be handled accordingly during modeling.

3.4.2 Feature Engineering

Additional features were engineered to enhance temporal and contextual representation. Player age was calculated using birthdate and game date. The `overtime_flag` variable was introduced by computing the total team minutes per game. Since a regulation NBA game consists of 48 minutes with 5 players on the court per team (i.e., $48 \times 5 = 240$ total minutes), any value exceeding 242 minutes–allowing for a minimal buffer–was classified as overtime.

Rest days were calculated by measuring the interval since a player's last game. Although initially flagged as a potential data leakage risk, this feature was retained since it is computed only after prior game completion and is not influenced by future injury events. Missing rest values (approximately 1%) were imputed using the column median.

A binary indicator was also created to flag back-to-back games, a known contributor to fatigue and injury risk (Bird et al., 2021). Further injury-linked variables were derived by aligning historical injury logs with each `GameID` – `PlayerID` pair, capturing the number of prior injuries, the last injury type and affected body part, and the most recent recovery duration. These values were strictly lagged to avoid future information leakage.

Player experience features included the cumulative game count (career total) and the season game number for each record. Additionally, the number of games missed due to injury was computed both seasonally and historically. Temporal integrity was ensured using `shift()` operations, restricting features to past data.

Categorical features – such as position, last injury type, and body part – were one-hot encoded. Month of play was also included as a calendar signal. To account

for short- and medium-term player load, rolling sums of minutes played and game appearances were computed over 5-, 10-, and 20-game windows.

For the `last_recovery_duration` field, missing values were filled using the column median to mitigate sparsity without introducing distortion.

3.4.3 Target Variable Construction

The target variable modeled a binary injury-forecasting task in two forms: time-based and game-based. To generate the time-based target, each player – game record was matched with the subsequent injury event. If an injury occurred within 5, 7, 15, or 30 days following the game, the label was assigned a value of 1; otherwise, it was set to 0.

To construct the game-based target, temporary synthetic entries were generated to represent missed games, since injured players do not appear in official box scores. These placeholder rows, marked with `is_added = True`, contained zeroed performance statistics and were used solely for labeling purposes. For each real game, the model scanned the subsequent N games, and if any matched a synthetic injury entry, the instance was labeled as positive. This approach enabled the model to learn from realistic patterns of player absence while maintaining chronological consistency, without permanently altering the original dataset.

Both target types were calculated only over real game records when evaluating class balance.

3.4.4 Model Training and Evaluation Metrics

To prevent data leakage, the dataset was sorted chronologically by game date throughout training and evaluation. Random shuffling was explicitly avoided to ensure that the models did not inadvertently learn patterns from future injury events, thus preserving temporal integrity.

Feature selection was conducted using a LightGBM pipeline with default parameters. Variables with importance scores below 0.01 were excluded to reduce dimensionality while retaining the most relevant predictors.

The final set of features was used to train three ensemble-based classifiers: Random Forest, LightGBM, and XGBoost. These models were selected based on their strong performance in structured data tasks, resilience to noise, and ability to handle moderate levels of missingness and outliers (see Section 3.3). Temporal deep learning models – including DeepSurv and Temporal Fusion Transformers – were also explored for their potential to capture sequential dependencies. However, they were ultimately excluded due to limited training time and unstable preliminary performance.

Given the significant class imbalance in the dataset, further explored in chapter 4 (table 4.2), with positive labels ranging from 5% to 27%, depending on the prediction window, multiple balancing techniques were evaluated. These included SMOTE (Synthetic Minority Oversampling Technique), noise-based augmentation, and random

undersampling. However, none of these methods produced consistent improvements in validation performance and were therefore excluded from the final modeling pipeline.

Model evaluation prioritized Recall as the primary performance metric, reflecting the high cost of false negatives in the context of injury forecasting. Missing an injury prediction may result in avoidable playtime loss, disruption of lineup planning, and increased risk to athlete health. While Precision is also important – particularly to avoid excessive caution or misallocated rest – it was considered secondary in this setting. Additional evaluation metrics included Accuracy, F1 Score, and AUC, offering a comprehensive assessment of model performance across different trade-offs between sensitivity and specificity.

3.5 Player Points Prediction Model

3.5.1 Dataset Construction

The final dataset for predicting player points was assembled through a multi-stage integration process. It extended the baseline injury dataset (without direct injury-related features) by incorporating contextual game-level features and advanced shot log statistics. Prior to merging, each dataset underwent targeted preprocessing to enhance its informational granularity.

The season-level statistics dataset – summarizing cumulative performance per player per season – was first processed to generate typological indicators based on rank-based metrics. These ranks reflected a player's relative standing across core statistics and were combined into derived features capturing archetypal play styles, such as Stretch Big Score and Three-and-D Score. Although these features were ultimately excluded from the final model, they illustrate the potential value of role-based player segmentation.

Historical season averages, such as last-season and three-season rolling averages (e.g., `last3seasonAVGPTS`), were computed and aligned with the temporal index of each game record. These features helped capture stable performance baselines and trends in scoring behavior.

The shot log dataset was also incorporated to enrich the dataset with advanced shot-level indicators. Features extracted included zone-specific shooting efficiency, shot timing, and clutch performance metrics.

Rolling aggregates of key performance statistics – points, assists, rebounds – were computed over the previous 5, 10, and 20 games to reflect form and momentum. Contextual load indicators such as overtime frequency, cumulative overtime minutes, and recent game counts were also added. Binary flags for back-to-back game participation were included to capture fatigue-related factors.

In addition to raw statistics, derived indicators were constructed to represent advanced traits, including scoring efficiency ratios, assist-to-turnover balances, and composite activity indices that integrate blocks, rebounds, and steals. These aimed to reflect latent performance aspects such as intensity, rhythm, and versatility.

To ensure temporal validity and prevent data leakage, all engineered features were derived strictly from data available prior to the prediction game. Columns containing statistics from the target game were removed, and `.shift()` was applied to all rolling windows to preserve chronological consistency.

Missing values – primarily from early-season games where rolling windows were incomplete – were imputed using backward fill (`bfill`), assuming that recent past context is the best proxy in the absence of long-term history.

After all preprocessing and alignment steps, the final merge was executed. Highly correlated feature pairs (Pearson $r > 0.85$) were examined for redundancy, and the less target-relevant variable was dropped. The resulting matrix contained 540,140 rows and 166 columns, representing a temporally structured and context-rich player-game feature set.

3.5.2 Model Training and Evaluation

A two-stage evaluation framework was implemented to assess the model's performance in predicting player scoring outcomes. In the baseline phase, predictions were generated using only box score statistics and contextual features, excluding any injury-related information. This allowed for the assessment of the model's standalone predictive power. In the enhanced phase, injury risk estimates and historical injury indicators were reintroduced to quantify their added value in forecasting individual performance.

Feature selection was carried out using an XGBoost Regressor trained on the full feature set, excluding the target and identifier columns. Feature importances were computed and ranked, and a threshold of 0.00251 was applied. This threshold was selected empirically, with the goal of retaining a wide range of potentially informative predictors while discarding features with minimal contribution. Given the exploratory nature of this stage and the emphasis on model interpretability and inclusiveness, the chosen cutoff prioritized feature richness over aggressive dimensionality reduction.

In the second stage of modeling, a Mixture Density Network (MDN) was also trained using the injury-related features, serving as a complementary approach to the XGBoost regressor. Unlike traditional point-estimate models, the MDN outputs a full conditional probability distribution for the target variable – in this case, points scored – enabling a more nuanced understanding of prediction uncertainty. This architecture is particularly suited for complex, multimodal scenarios such as NBA player performance, where variability arises from a mix of contextual, physical, and tactical factors. The MDN's results were later compared against those of the XGBoost model to evaluate not only point accuracy but also the ability to characterize uncertainty in performance forecasts.

Model performance was evaluated using two standard regression metrics: the Coefficient of Determination (R^2) and the Mean Absolute Error (MAE). R^2 measures the proportion of variance in the target variable (points scored) that is explained by the

model, with higher values indicating stronger explanatory power. MAE quantifies the average magnitude of prediction errors in the same units as the target, offering a direct interpretation of practical accuracy. Together, these metrics provide a complementary view of model quality – R^2 reflecting overall fit, and MAE capturing real-world prediction accuracy.

3.6 Game Outcome Predictive Model

3.6.1 Dataset Construction and Target Definition

To model NBA game outcomes, a structured dataset was developed by integrating multiple feature sources. These included selected variables from the injury prediction model, the most predictive features from the player points model, team-aggregated shot log metrics, season-level team statistics and advanced team metrics. To evaluate the contribution of injury-related and player scoring prediction features, initial models were trained without these variables. Comparative performance analyses were then conducted after incorporating them, allowing for the assessment of their incremental predictive value.

The merging process produced a high-dimensional and context-aware representation of team matchups, suitable for binary classification tasks.

Additional engineered features were incorporated to capture strategic dimensions of gameplay. Offensive indicators such as composite metrics based on points scored, assists, field goal percentage, and turnovers were created to reflect team scoring efficiency. Defensive metrics, including blocks and steals per game, were used to model assertiveness on the opponent's end. Momentum proxies were derived from recent win sequences, and a composite team strength score was constructed by combining multiple efficiency and consistency indicators into a single interpretable value. All features were generated using pre-game information only to ensure full chronological integrity and prevent future data leakage.

The target variable was constructed by comparing the total points scored by each team in a given match. Each game was represented twice – once from the home team's perspective and once from the away team's – allowing the definition of a binary variable `WIN_HOME_TEAM`, set to 1 if the home team outscored the visitor. This framing aligns with traditional outcome prediction tasks in professional basketball, where home-court advantage is a well-documented phenomenon (Carron, A. V., Loughhead, T. M., & Bray, S. R. 2005). In the dataset of 32,129 games analyzed, home teams won approximately 58% of the time, reflecting this established trend.

To reflect comparative dynamics between teams rather than absolute performance, the final dataset was structured as the difference between home and away team values across all numeric and Boolean features. This formulation captures relative strength and situational advantage in each matchup, enhancing the model's ability to generalize across diverse pairings (Bunker & Thabtah, 2019).

Time-aware variables were also introduced, such as the number of games won in the last five or twenty contests, computed using `.shift()` operations to prevent leakage. The complete feature set initially comprised 293 variables. To reduce dimensionality and mitigate multicollinearity, feature pairs with a Pearson correlation coefficient above 0.85 were evaluated. In each highly correlated pair, the feature less predictive of the target was removed, resulting in the exclusion of 95 features and a refined input set.

3.6.2 Model Selection and Evaluation

To predict game outcomes, three ensemble-based classification algorithms were employed: XGBoost, LightGBM, and Random Forest. These models were selected for their strong performance on structured datasets, ability to capture non-linear interactions, and proven success in similar sports analytics applications (see Section 3.3). Each model was trained using the top 30 most informative features, selected based on their contribution to validation performance as determined by feature importance rankings.

Model performance was evaluated using four standard classification metrics: Accuracy, Precision, Recall, and ROC AUC. Among these, Accuracy was prioritized as the primary evaluation criterion given the binary and relatively balanced nature of the task – predicting whether the home or away team would win. Accuracy directly reflects the model's effectiveness in selecting the correct outcome and is appropriate in this context where class imbalance is minimal. Precision and Recall offer insight into the model's behavior in edge cases, while ROC AUC provides a threshold-independent view of overall discriminatory ability.

Chapter 4

Modeling Results and Interpretability

This chapter presents the results obtained from the three core components of the modeling pipeline: injury forecasting, player performance prediction, and team game outcome classification. Each section follows a structured format, beginning with the evaluation of model performance using standard classification and regression metrics, followed by an interpretability analysis based on SHAP values. Particular attention is given to the integration of injury risk and clutch performance variables across stages of the pipeline. This allows for a systematic examination of how upstream predictions – such as individual scoring output and player availability – contribute to downstream forecasting tasks at the team level. Throughout the chapter, results are interpreted in relation to the research question and evaluated in terms of both statistical performance and practical relevance.

4.1 Modeling Injury Risk in NBA Players

To evaluate the effectiveness of injury forecasting strategies, a total of eight binary classification targets were tested. These included both time-based outcomes (e.g., whether a player will be injured within 5, 7, 15, or 30 days) and game-based outcomes (e.g., within 1, 3, 5, or 10 games). For each target, three models – Random Forest, XGBoost, and LightGBM – were independently trained using a consistent set of features and a standardized evaluation framework. The classification metrics, averaged across multiple runs, are reported in Table 4.1.

Target	Model	Accuracy	Precision	Recall	F1-score	ROC AUC
WillBeInjured_1games	RandomForest	97%	8%	9%	9%	68%
	XGBoost	97%	8%	11%	9%	71%
	LightGBM	95%	5%	13%	7%	68%
WillBeInjured_3games	RandomForest	64%	11%	57%	19%	65%
	XGBoost	58%	11%	65%	18%	66%
	LightGBM	57%	10%	65%	18%	65%
WillBeInjured_5games	RandomForest	58%	17%	64%	27%	65%
	XGBoost	57%	17%	66%	28%	65%
	LightGBM	55%	17%	67%	27%	64%
WillBeInjured_10games	RandomForest	56%	30%	70%	42%	65%
	XGBoost	57%	30%	67%	42%	65%
	LightGBM	54%	30%	71%	42%	64%
WillBeInjured_5days	RandomForest	56%	12%	68%	20%	67%
	XGBoost	51%	11%	75%	20%	67%
	LightGBM	50%	11%	76%	20%	68%
WillBeInjured_7days	RandomForest	55%	16%	71%	25%	67%
	XGBoost	53%	16%	73%	26%	67%
	LightGBM	49%	15%	77%	25%	67%
WillBeInjured_15days	RandomForest	54%	28%	75%	41%	68%
	XGBoost	53%	28%	78%	41%	68%
	LightGBM	50%	27%	81%	40%	68%
WillBeInjured_30days	RandomForest	58%	43%	78%	55%	70%
	XGBoost	58%	43%	80%	56%	70%
	LightGBM	56%	42%	81%	55%	69%

Table 4.1: Classification metrics for eight injury forecasting targets across three models.

All injury prediction targets present highly imbalanced classification problems (Table 4.2), with injury events representing a small minority of cases. As a result, models tend to favor recall over precision–a common trade-off in risk-sensitive domains where failing to detect a true event is more costly than issuing a false alarm.

Target	Positivos	Total	Pct (%)
WillBeInjured_1games	45984	794860	5.79
WillBeInjured_5days	46377	794860	5.83
WillBeInjured_7days	64752	794860	8.15
WillBeInjured_3games	79444	794860	9.99
WillBeInjured_5games	109372	794860	13.76
WillBeInjured_15days	126963	794860	15.97
WillBeInjured_10games	173420	794860	21.82
WillBeInjured_30days	214641	794860	27.00

Table 4.2: Distribution of positive and total observations for each injury prediction target.

As expected, broader targets – such as predicting injury over the next 30 days or 10

games – yielded stronger scores across most evaluation metrics. This trend reflects the increased number of positive cases in longer windows, which facilitates generalization. However, these broader targets are less useful for short-term decision-making. While a 30-day forecast might help with long-range planning, it lacks usefulness for day-to-day roster and minute management.

Given this trade-off between predictive strength and operational utility, the target `WillBeInjured_3games` was selected for downstream modeling. This formulation strikes a practical balance: with a recall of 65% and acceptable specificity, it aligns well with short-term performance forecasting and coaching workflows. Notably, when multiple injury targets were tested as inputs in the point prediction model, `WillBeInjured_3games` consistently delivered the highest predictive gain.

Although the model performs well in identifying non-injury cases (class 0), its relatively low precision (11%) for injury predictions highlights a tendency to overpredict injuries. This is an expected behavior in imbalanced datasets and may be acceptable in contexts where precaution is preferred. Encouragingly, the recall of 65% means the model correctly flags nearly two-thirds of players who will indeed suffer an injury within the next three games.

The confusion matrix below supports this interpretation, with a high number of false positives (e.g., 51,188 predictions of injury that did not materialize). While this reflects a conservative prediction strategy, it can still be actionable in settings where anticipating injury – even at the cost of some false alarms – is preferable to being unprepared.

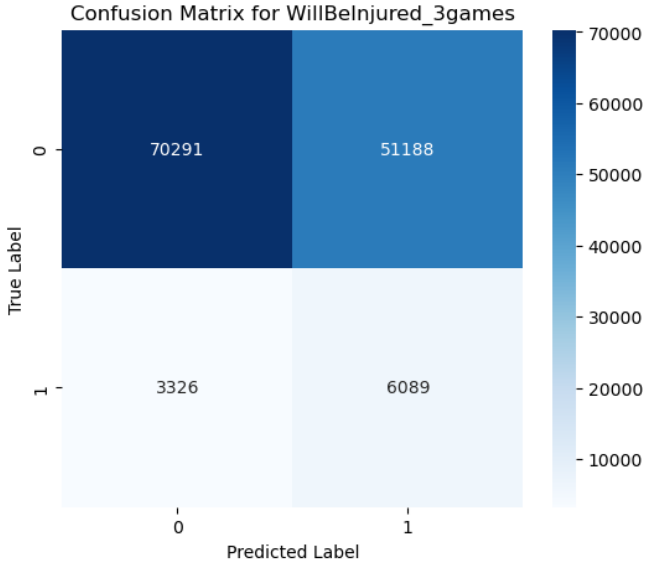

Figure 4.1: Confusion matrix for `WillBeInjured_3games` predictions.

4.1.1 SHAP Analysis and Feature Interpretability

To better understand the inner workings of the model, SHAP (SHapley Additive exPlanations) values were computed. These provide consistent, locally accurate insights into each feature's contribution to the prediction. In the SHAP plot (Figure 4.4), each dot represents a player – game instance. The color indicates the feature value (blue = low, red = high), while the position on the x-axis shows the direction and magnitude of its impact on the prediction: points to the right indicate a positive contribution to the predicted target, and points to the left indicate a negative contribution. The horizontal spread reflects how much each feature contributes across different cases, with wider dispersion meaning greater variation in its influence.

Among all variables, `total_previous_injuries` was the most impactful feature. As expected, higher values were strongly associated with increased injury risk, confirming prior research that injury history is one of the strongest predictors of future injury. The variable `PLAYER_GAME_NUMBER_ALLTIME`, which tracks total career games played, exhibited a negative association with injury risk. Though initially counterintuitive, this likely reflects survivorship bias–only durable players tend to accumulate high game counts over time.

`rolling_games_missed_season` - a proxy for recent injury-related absences - also contributed meaningfully to predictions. Players who had missed multiple recent games were more likely to be flagged as injury-prone. A similar trend was observed with `rolling_minutes_20g`, which represents the total minutes played by a player over the last 20 games. When this value is higher, the model predicts a greater likelihood of injury–suggesting that accumulated short-term workload contributes to physical strain and increases vulnerability.

`PLAYER_GAME_NUMBER_SEASON` also showed a positive correlation with injury risk. This is consistent with the idea that accumulated in-season exposure contributes to fatigue and physical breakdown. Supporting this, player availability-related variables (like `isStarter` and `Position_BENCH`) also indicated greater injury likelihood among players with higher roles and responsibilities–matching expectations in the sports medicine literature (Drew & Finch, 2016).

Other useful signals included `RECOVERY_TIME_LAST`, where longer recovery periods from the most recent injury were associated with slightly reduced risk, possibly due to better rest management or conservative reintegration practices. Age, interestingly, showed a mild inverse association with injury risk, likely due to sample structure: older players who remain in the league tend to be more robust, having avoided early career derailment due to injuries. Additionally, as illustrated in the distribution chart, the number of players decreases significantly with age, further reinforcing this selective effect. To partially mitigate the imbalance in age representation and enhance interpretability, a dedicated age group (35+) was created, as shown in Figure 4.3.

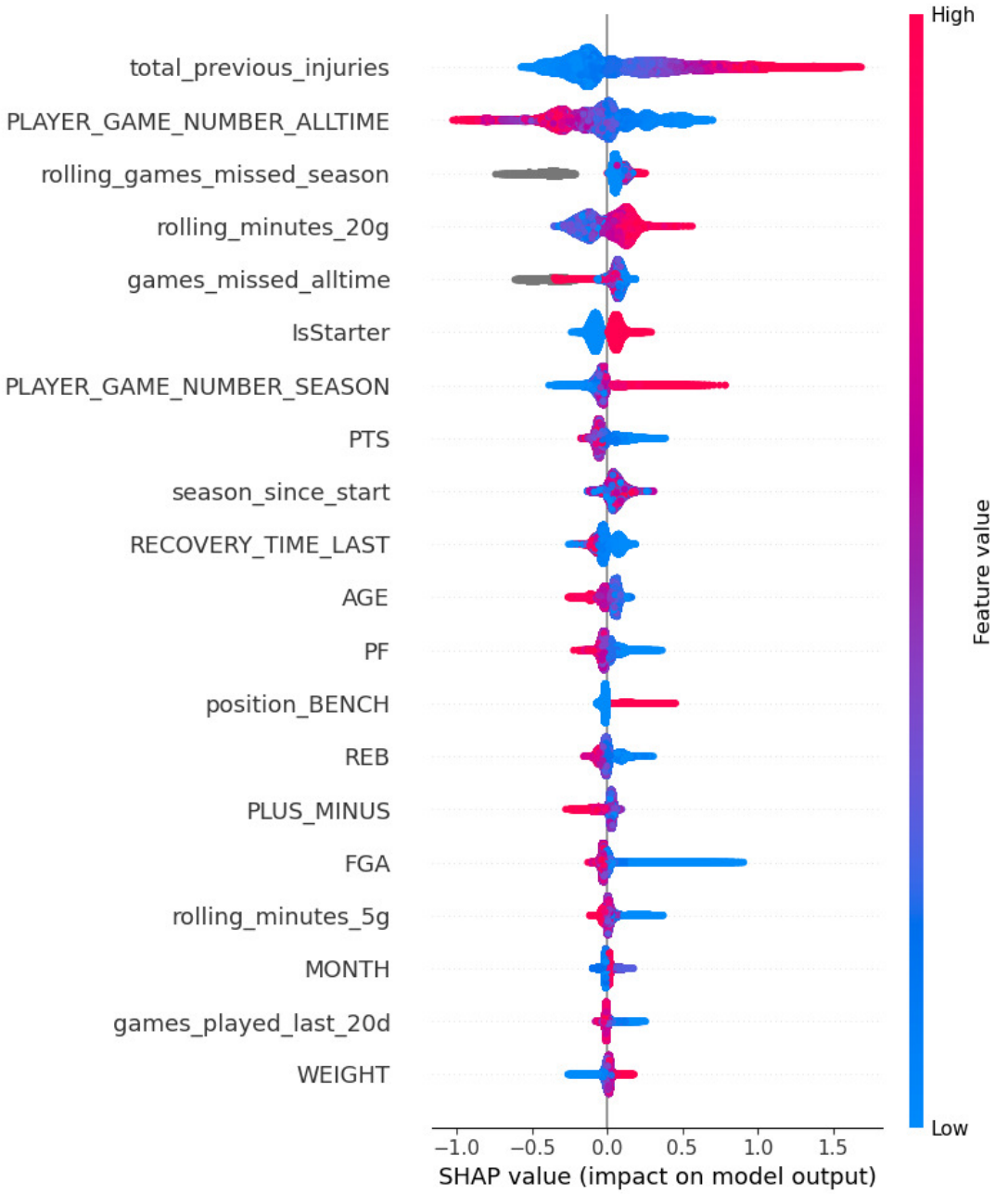

Figure 4.2: SHAP summary plot for Injury model

Together, these results confirm the model's ability to learn meaningful and interpretable relationships between physical, historical, and contextual factors and short-term injury risk. The combination of clinical logic, recent performance, and usage dynamics provides a robust basis for use in downstream decision-making–from lineup rotation to performance forecasting.

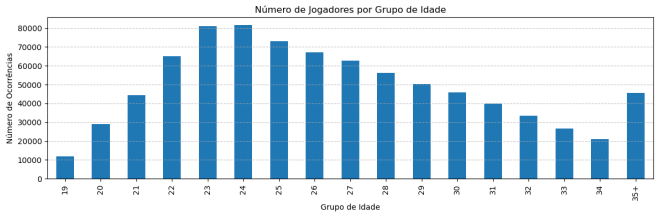

Figure 4.3: Distribution of player ages

4.2 Forecasting Player Scoring Performance

The XGBoost baseline model, trained without incorporating any injury-related features, achieved a Mean Absolute Error (MAE) of 3.14 points and a Coefficient of Determination (R^2) of 72% when evaluated on the full dataset. These results indicate that the model captures a substantial portion of the variance in individual point production, although some prediction error remains – an expected outcome given the inherent complexity and unpredictability of in-game basketball performance.

To account for the diversity of scoring roles and styles across players, a segmentation strategy was tested. Players were grouped into scoring terciles – low, medium, and high – based on their historical average points per game. This stratification aimed to assess whether specialized models could yield better performance within more homogeneous scoring profiles. The results are presented in Table 4.3:

Scoring Group	MAE	R^2	N Samples
Low	1.43	28%	214,655
Medium	1.31	16%	148,015
High	3.25	48%	177,470

Table 4.3: Model performance metrics stratified by player scoring tercile.

As expected, the model performed best among high scorers in terms of R^2, reflecting its ability to explain more variance within this group. However, it also exhibited the highest MAE, consistent with the greater performance volatility often observed among high-usage players. Conversely, low-scoring players – who tend to have more stable and predictable roles – resulted in lower MAE values, although with modest explanatory power due to the narrower range of scoring outcomes. These findings confirm that stratifying predictions by scoring profile is an effective approach for improving both interpretability and accuracy across player types.

To better understand which features drove model performance, SHAP values were computed for the baseline model.

The most influential variable was `hot_streak_flag_season`, which captures whether a player had made at least three consecutive field goals in recent games. This feature was averaged across each game and interpreted non-linearly by the model. Moderate

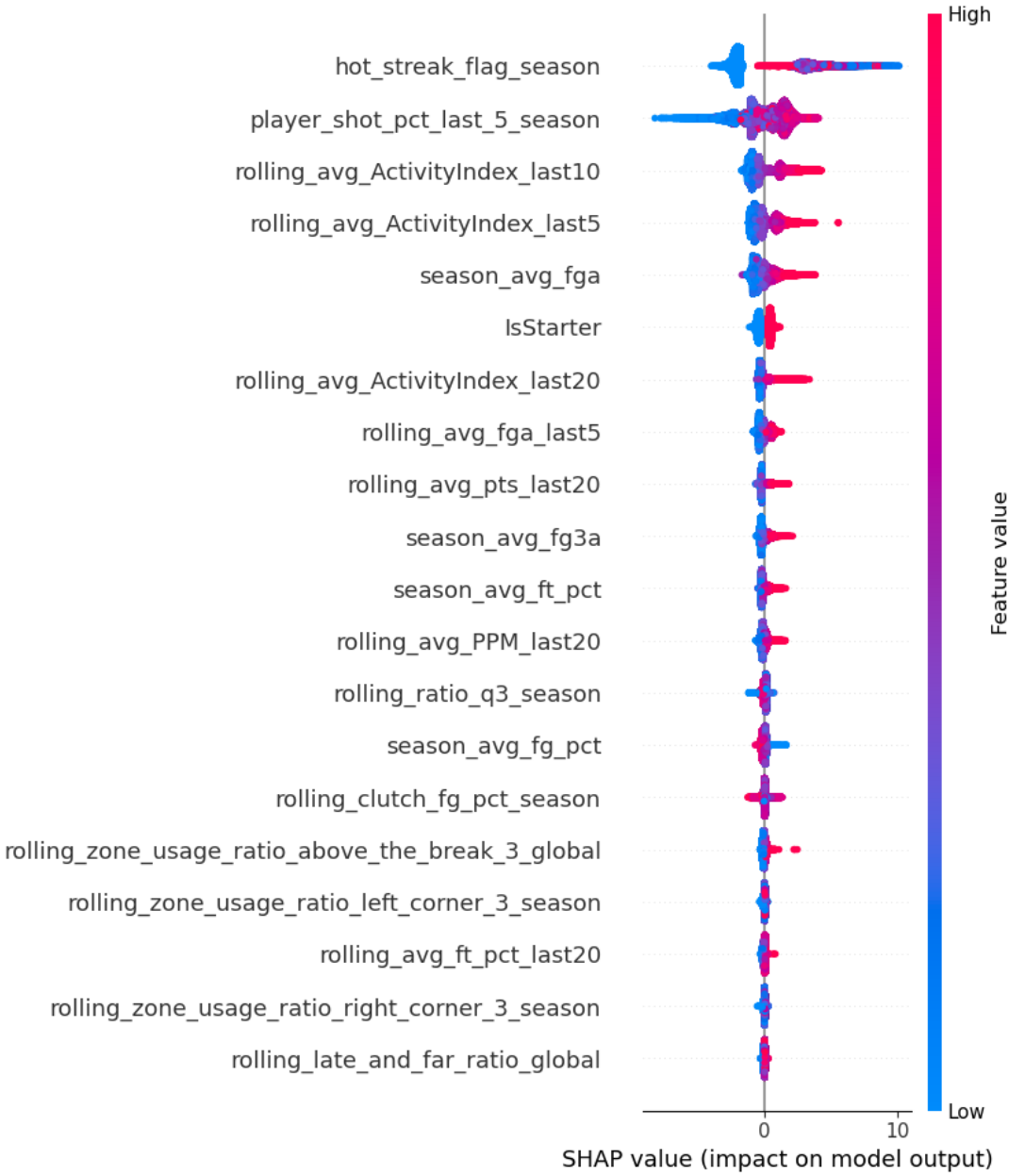

Figure 4.4: SHAP summary plot for points prediction model

values – typically between 0.25 and 0.50 – were associated with higher scoring outcomes, suggesting that sustained but controlled momentum is positively associated with offensive productivity.

Other key variables included `player_shot_pct_last_5_season`, which reflects long-term shooting efficiency, and the `rolling_avg_ActivityIndex` over the past 5 and 10 games, a composite metric designed to capture recent engagement and workload. These features underscore the importance of both historical consistency and short-term form in shaping point production.

Additional influential predictors were `season_avg_fga` and `season_avg_fg3a`, which

track field goal and three-point attempt averages, respectively. These proxies for shot volume and offensive role were positively associated with higher scoring predictions. Likewise, `usage_index`, a measure of a player's share of team possessions, contributed meaningfully to the model's output – as expected, players with higher usage rates typically accumulate more points.

Finally, rolling metrics such as `rolling_avg_pts_last20` and `rolling_avg_PPM_last20` (points per minute) emerged as important contributors. These indicators capture recent scoring rhythm and consistency, reinforcing the model's ability to integrate short-term dynamics into its predictions.

Taken together, the results of this section demonstrate that player scoring output can be forecasted with reasonable accuracy using a combination of structural variables (e.g., usage and shot volume), performance trends (e.g., streaks and rolling averages), and long-term indicators (e.g., historical shooting percentage). These features collectively form a robust foundation for integrating player performance forecasts into higher-level models of team behavior and outcomes.

4.3 Predicting Team Outcomes

The game outcome prediction model, developed using both `XGBoost` and `LightGBM` classifiers, achieved strong baseline performance. Both algorithms reached an overall accuracy of 79%, a notable result considering that these models were trained without access to injury-related data or the outputs from the preceding models. Instead, they relied exclusively on player-level and team-level contextual features and historical performance indicators.

The classification report for the `LightGBM` model revealed balanced performance across classes. For games predicted as losses (label 0), the model reached a precision of 81% and a recall of 85%. For predicted wins (label 1), precision was 77% and recall was 71%. The overall F1-score was 0.79, consistent with the model's overall accuracy. Additionally, the ROC AUC score reached 87%, indicating that the model is capable of discriminating between winning and losing outcomes across a broad range of probability thresholds – a desirable property in high-stakes applications such as professional sports forecasting.

To gain insight into the model's decision process, SHAP value analysis was conducted. Figure 4.5 displays the SHAP summary plot for the LightGBM game outcome model.

Among the most impactful features, `diff_3pt_zone_fg_pct_above_break`, `diff_paint_fg_pct`, and `diff_shot_time_ratio_q4` emerged as key contributors.
High values of `diff_3pt_zone_fg_pct_above_break` were associated with an increased probability of winning, indicating that outperforming the opponent in long-distance shooting from the top of the arc substantially improves competitive advantage. Similarly, higher `diff_paint_fg_pct` values were positively linked to win probability,

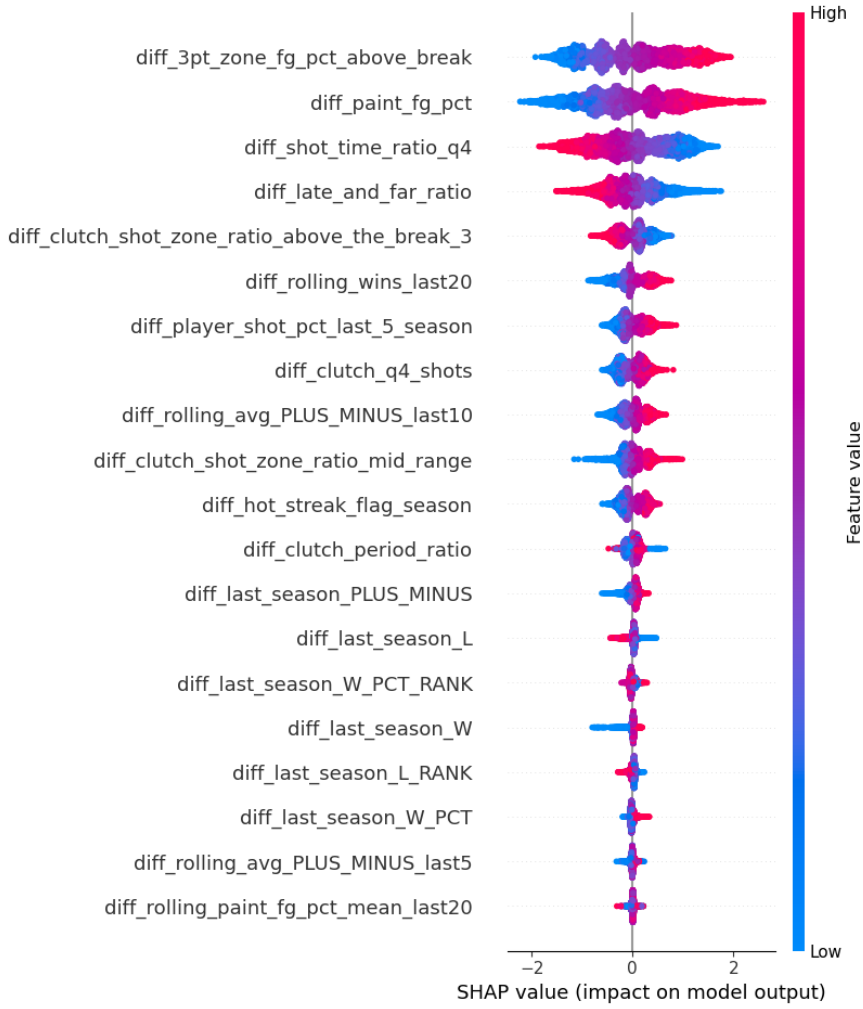

Figure 4.5: SHAP summary plot for the game outcome prediction model.

reflecting the importance of interior efficiency, physical dominance, and smart shot selection.

In contrast, elevated values of `diff_shot_time_ratio_q4` had a negative impact on win probability. This feature measures the frequency of late shot-clock attempts in the fourth quarter relative to the opponent; higher values – often indicative of rushed or poorly organized possessions – decrease the likelihood of victory.

Clutch-related variables also played a significant role. `diff_clutch_q4_shots`, which measures clutch-time shot volume, was positively associated with win probability, suggesting that generating more clutch opportunities enhances a team's chances in close contests. `diff_clutch_shot_zone_ratio_mid_range` indicated that selective mid-range attempts in clutch moments can also boost win probability. Conversely, higher values of `diff_clutch_period_ratio`, capturing time management during clutch periods, were linked to lower win probabilities, underscoring the importance of efficient late-game execution. This variable reflects the overall difference between the two teams in the proportion of total game time spent in clutch situations, highlighting

disparities in game control.

Taken together, these results highlight the critical role of both spatial execution and temporal dynamics – especially under pressure – in determining NBA game outcomes. The model's identification of high-leverage behaviors such as late shot-clock execution, effective clutch-time shot selection, and interior efficiency underscores the relevance of these factors for predictive success.

4.4 From Player to Team: Integrating Injury and Clutch Metrics into NBA Game Forecasting

4.4.1 Integrating Injury Forecasts into Player Performance Modeling

To evaluate the added value of incorporating injury-related information into player-level performance forecasting, a second version of the point prediction model was developed. This enhanced model integrated outputs from the injury prediction module – specifically, the probability of injury within the next three games (`prob_WillBeInjured_3games`) and the corresponding binary flag (`WillBeInjured_3games`) – along with six additional features representing historical injury patterns and player availability. In parallel, a Mixture Density Network (MDN) was also trained using the same feature set to serve as a comparative benchmark. However, the MDN underperformed relative to the tree-based models, yielding a Mean Absolute Error (MAE) of 3.17 and an R^2 of 72%, suggesting that its added probabilistic flexibility did not translate into improved point accuracy in this specific context. Compared to the XGBoost baseline model – which achieved an MAE of 3.14 and an R^2 of 72% – the enhanced version obtained an MAE of 3.06 and an R^2 of 73%. While the improvement may appear modest, it reflects meaningful progress in capturing short-term availability risks and accumulated physical strain, both of which influence player performance.

4.4.2 Explaining Points Outcome Predictions Using SHAP

Several injury-related variables emerged as significant contributors to prediction accuracy. The feature `prob_WillBeInjured_3games` showed a clear negative relationship with expected point output: as short-term injury risk increased, projected scoring decreased. This relationship is intuitive, as injury-prone players are more likely to experience reduced playing time, assume less demanding roles, or be excluded from game rotations. In contrast, the cumulative feature `total_previous_injuries` displayed a positive association with scoring, contrary to initial expectations. This pattern likely reflects a survivor bias – only players capable of sustaining performance despite prior injuries remain active in the rotation, thereby skewing the relationship in a positive direction.

Experience-related features also played a notable role. `PLAYER_GAME_NUMBER_ALLTIME`

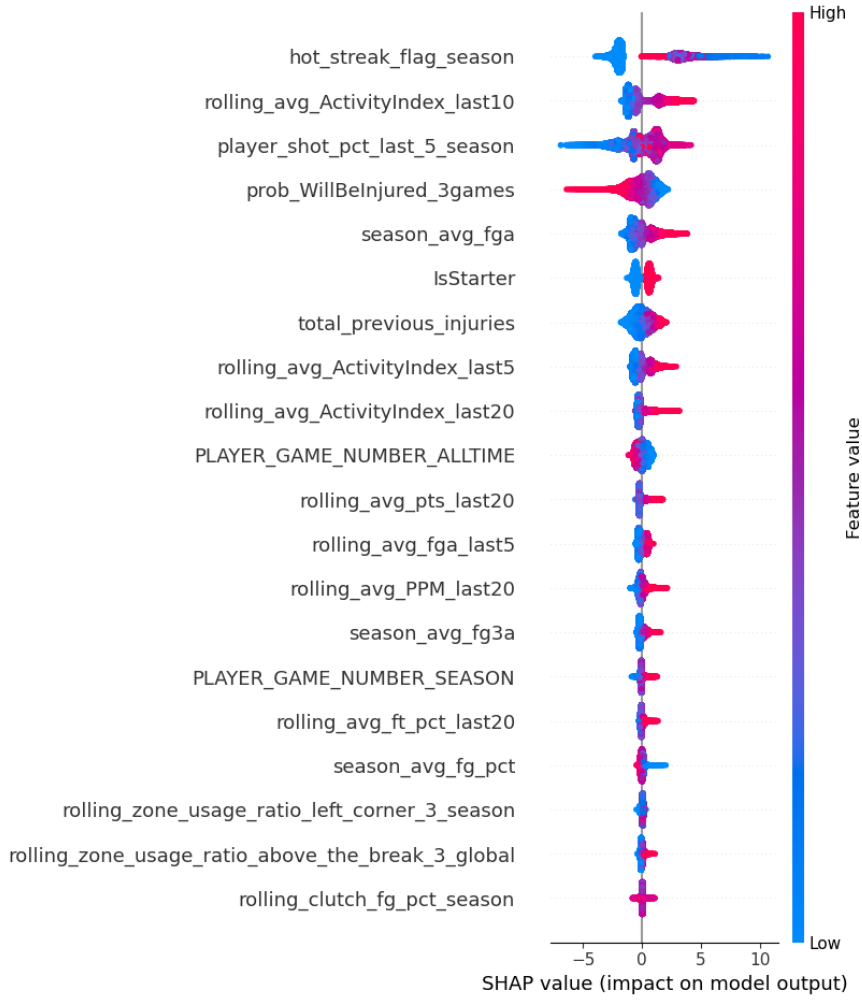

Figure 4.6: SHAP summary plot for the enhanced points prediction model, including injury-related features.

demonstrated a negative correlation with point predictions, possibly capturing the scoring decline typically observed in players beyond their peak years.
Conversely, `PLAYER_GAME_NUMBER_SEASON` was positively associated with performance, suggesting that continued participation throughout the season supports rhythm, role stability, and offensive productivity.

Importantly, several core features retained high importance even after the integration of injury-related information.
Variables such as `hot_streak_flag_season`, `rolling_avg_ActivityIndex_last10`, and `player_shot_pct_last_5_season` remained among the most influential, underscoring the lasting predictive value of momentum, engagement, and long-term efficiency. The injury features thus complemented – rather than displaced – the core drivers of scoring performance, improving predictive accuracy through the addition of contextually meaningful information.

4.4.3 Integrating Injury Risk and Performance Predictions into Team Outcome Forecasting

To further extend the pipeline, the final model incorporated outputs from both the point prediction and injury forecasting components into the game outcome classifier. At this stage, features were aggregated at the team level, including average `prob_WillBeInjured_3games`, `total_previous_injuries`, `PLAYER_GAME_NUMBER_SEASON`, and `PLAYER_GAME_NUMBER_ALLTIME` across all expected players in a given matchup, along with the team-level sum of predicted points (`PTS_pred`). These inputs collectively capture short-term health risk, long-term durability, player exposure, and overall expected team performance.

Both XGBoost and LightGBM classifiers were tested in this setting. While both outperformed their baseline counterparts – which had no access to individual health or performance forecasts – LightGBM achieved superior results, reaching an accuracy of 84% and a ROC AUC score of 0.915. These represent clear improvements over the baseline model (79% accuracy and 0.87 ROC AUC), especially in games where outcomes are less certain.

The LightGBM classification report showed balanced performance across both outcome classes. For predicted losses (label 0), the model achieved a precision of 85% and a recall of 89%. For predicted wins (label 1), it achieved a precision of 83% and a recall of 77%. These improvements in both precision and recall over the baseline indicate that the integration of health and performance forecasts strengthens the model's ability to differentiate winning from losing teams, particularly in close or competitive contexts.

4.4.4 Explaining Team Outcome Predictions Using SHAP

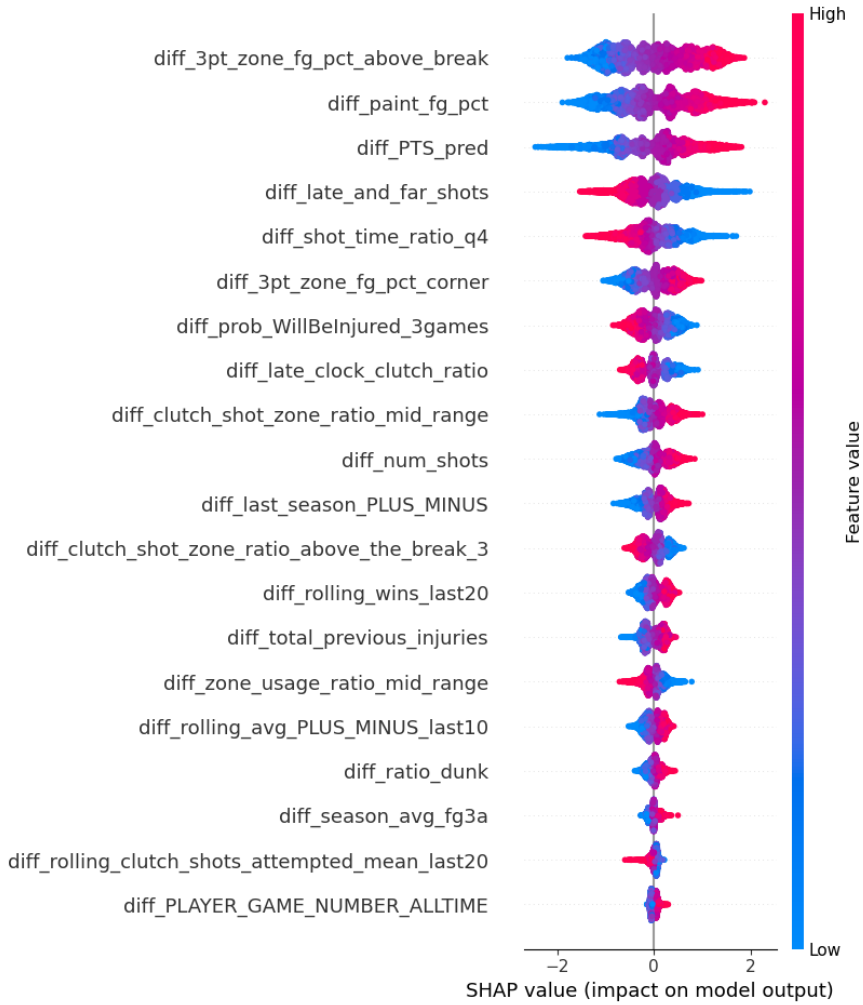

Figure 4.7: SHAP summary plot for the enhanced game outcome prediction model, including injury-related and predicted-points features.

SHAP value analysis revealed that `diff_PTS_pred`, the difference in total predicted points between teams, was among the top three most important features. This finding reinforces the high predictive value of individual scoring estimates in determining team-level outcomes and validates the modular architecture of the pipeline, in which player-level regression outputs are used to inform downstream classification.

Another highly influential variable was `diff_prob_WillBeInjured_3games`, representing the difference in average short-term injury risk between opposing teams. Higher values – indicating greater overall fragility or fatigue on one side – were consistently associated with a lower probability of winning. This is consistent with domain expectations, as teams with more at-risk players are less likely to execute efficiently on either end of the court. SHAP distribution plots further confirmed this relationship, showing that elevated injury probabilities (visualized as red markers to the right) contributed negatively to win likelihood.

Interestingly, higher values of `diff_total_previous_injuries` were associated with increased win probability. While this might seem counterintuitive, it likely reflects the presence of veteran players with extensive playing histories and proven resilience. These athletes, despite prior injuries, may provide leadership, tactical intelligence, and consistent performance that positively influence game outcomes.

The feature `PLAYER_GAME_NUMBER_ALLTIME` showed a similar pattern. Rather than signaling accumulated wear, high values tended to align with increased win probability. This likely captures the effect of career longevity as a proxy for player reliability, skill, and adaptability – factors particularly important in competitive scenarios.

Clutch-related features also maintained strong importance in the final model. The variable `diff_late_clock_clutch_ratio` measured the frequency with which teams were forced into late shot-clock situations during clutch moments. Higher values were negatively associated with win probability, suggesting that poor time management and rushed possessions under pressure decrease a team's chances of success.

Similarly, `diff_clutch_shot_zone_ratio_mid_range` captured the differential reliance on mid-range attempts during high-stakes periods. Contrary to common strategic assumptions that mid-range shots less efficient, the model identified that teams with greater relative use of this zone in clutch moments tended to perform better, suggesting that effective mid-range play can be a valuable asset in specific situations.

In contrast, `diff_clutch_shot_zone_ratio_above_the_break_3` showed a negative contribution to win probability when values were high. Although shots from above the arc can be valuable in late-game scenarios, excessive reliance on this area during clutch moments may reflect forced or low-percentage attempts. The model interpreted this pattern as detrimental, suggesting that more balanced or efficient shot selection strategies are preferable under pressure.

Finally, `diff_rolling_clutch_shots_attempted_mean_last20`, which tracked how often players had taken clutch shots in recent games, was positively linked to winning outcomes. This suggests that teams with individuals regularly trusted in high-pressure moments tend to perform better when it matters most.

Taken together, these patterns highlight the model's ability to capture not only structural factors like experience and injury history, but also nuanced behavioral indicators such as clutch execution. These insights further support the practical relevance of the integrated pipeline for forecasting game outcomes in professional basketball.

Chapter 5

Discussion

This chapter discusses the results obtained from the three key stages of the predictive pipeline: injury risk estimation, player scoring forecasting, and team-level outcome classification. The analysis moves beyond raw performance metrics to interpret how the models interact, with a focus on their modular integration and strategic relevance. Each section highlights the implications of model behavior, supported by SHAP-based interpretability, and explores domain-specific findings such as survivorship bias, the impact of clutch behavior, and the role of injury forecasts in shaping player and team performance. Cross-model insights are synthesized to evaluate the added value of structuring predictive tasks in a layered, interdependent manner. Finally, limitations and future directions are discussed, outlining paths for refining predictive accuracy and operational deployment in real-world basketball analytics.

5.1 Injury Risk Prediction

The injury forecasting task sought to identify short-term injury risk among NBA players using historical performance, workload, and contextual variables. Across eight binary targets, results confirmed the inherent difficulty of the task, driven by the strong class imbalance and noisy nature of injury occurrence. Nonetheless, all models performed above chance, with LightGBM slightly outperforming both Random Forest and XGBoost on most targets. The chosen formulation – `WillBeInjured_3games` – struck a balance between operational relevance and predictive strength, achieving a recall of 65% and an ROC AUC above 65%.

These results align with findings in the sports science literature that suggest previous injury, short-term workload, and game exposure are reliable, although imperfect, predictors of future injury (Drew & Finch, 2016). The SHAP analysis supported this view, highlighting `total_previous_injuries`, `PLAYER_GAME_NUMBER_SEASON`, and `rolling_minutes_20g` as key predictors. Interestingly, `PLAYER_GAME_NUMBER_ALLTIME` had a negative association with injury risk, likely due to survivorship bias: only the most physically resilient players accumulate long careers. This confirms the importance

of accounting for sampling structure in interpreting model outputs.

Despite low precision, which indicates a tendency to overpredict injuries, the model fulfills its goal as a risk-screening tool. In contexts where precaution is valued, such as roster management or rest scheduling, high recall is more desirable than conservative specificity. These findings support the use of probabilistic injury flags as input for downstream tasks, especially in settings where cost asymmetry favors overprotection.

5.2 Player Scoring Forecasting

The scoring prediction model addressed the central research question of whether NBA player performance, in terms of points scored, could be accurately estimated from contextual and injury-historical data. The baseline XGBoost model achieved an MAE of 3.14 and an R^2 of 72%, confirming that a large portion of the variance in player output can be explained through well-designed features. Stratified analysis by scoring terciles revealed that modeling scoring behavior by player type improves both interpretability and MAE, especially for lower-volume scorers.

SHAP analysis revealed that short-term momentum (`hot_streak_flag_season`), long-term efficiency (`player_shot_pct_last_5_season`), and recent engagement (`rolling_avg_ActivityIndex`) were consistent drivers of point production.

In the second modeling phase, the inclusion of injury-related features (e.g., `prob_WillBeInjured_3games`, `total_previous_injuries`, etc.) improved predictive accuracy (MAE 3.06, R^2 73%). This suggests that injury risk carries predictive value even before it materializes into missed games, potentially affecting player confidence, minutes, or tactical deployment. Notably, the relationship between `total_previous_injuries` and scoring was positive, possibly due to survivor bias. To benchmark this improvement, an MDN was also tested on the same enhanced feature set but underperformed relative to XGBoost, suggesting that tree-based methods remain more suited to this structured tabular context.

5.3 Team Outcome Classification

Building on the player-level forecasts, the team outcome model integrated predicted points and aggregated injury risk into a downstream classification task. LightGBM outperformed XGBoost with an accuracy of 84% and ROC AUC of 91.5%, compared to 79% and 87% for the baseline model. This demonstrates the value of modular model design: upstream predictions significantly improved downstream performance.

Feature importance analysis confirmed the strong predictive power of `diff_PTS_pred`, the difference in predicted total points between teams. This validates the integration of scoring forecasts into broader outcome modeling and shows that individual performance estimates aggregate meaningfully into team-level signals. Injury-related variables also ranked highly. In particular, `diff_prob_WillBeInjured_3games` was

negatively correlated with win probability, indicating that teams with higher average injury risk among active players were less likely to win.

Clutch-related variables continued to show a strong predictive value. While some findings, such as the negative contribution of `diff_late_clock_clutch_ratio`, aligned with expectations, others like `diff_clutch_shot_zone_ratio_above_the_break_3` revealed counterintuitive patterns. Despite being a popular clutch shooting area, higher reliance on this zone negatively impacted win likelihood, perhaps reflecting contested or suboptimal shot creation under pressure. These insights illustrate the importance of not only where teams shoot but how and when they generate those opportunities.

5.4 Cross-Model Insights and Strategic Implications

The modular pipeline introduced in this study, in which player-level injury and performance predictions inform team-level outcome forecasts, reflects a practical and scalable approach to sports analytics. Each layer of modeling added interpretable value: injury forecasts flagged health risks, scoring models predicted contributions, and the classifier integrated these to evaluate game outcomes. Together, they addressed the core research question: can individual-level dynamics improve the forecasting of team results?

Findings confirm that incorporating health and clutch related variables improves prediction accuracy in a statistically and operationally meaningful way. More broadly, this supports the thesis objective of building interpretable, decision-oriented tools in a high-variance sports context.

Future work could explore the use of these forecasts in real-time applications, such as load management, betting markets, or opponent scouting. To address the persistent challenge of class imbalance in injury forecasting, future studies should also focus on advanced sampling techniques – including undersampling of the negative class and data augmentation strategies for the positive (injury) class – to enhance model generalization. Incorporating additional contextual variables, such as whether a player was recently traded, may further refine point prediction models, as changes in teammates, systems, and usage patterns can significantly affect individual scoring dynamics. Moreover, future models could benefit from segmenting players into scoring terciles and training specialized regressors for each group, which may improve predictive accuracy by capturing distinct scoring profiles. Despite the known limitations of noisy data and unobserved variables, the pipeline demonstrates that machine learning can extract strategic signal from the chaos of sport.

Chapter 6

Conclusion

This dissertation set out to explore whether integrating injury forecasting, player performance prediction, and contextual team-level variables could improve the accuracy and interpretability of game outcome predictions in the NBA. By designing a modular pipeline that systematically connected these three predictive layers, the study addressed the central research objective of enhancing basketball forecasting through multi-level, interpretable machine learning models.

The findings demonstrate that injury risk can be meaningfully predicted using structured features derived from player exposure, workload, and historical health records. Although the precision of these models was limited due to the class imbalance inherent in injury data, their recall was sufficiently high to justify their use in risk-sensitive decision-making settings. The feature interpretability analysis revealed that short-term workload and prior injuries are reliable indicators of vulnerability, aligning with the existing literature in sports science and medicine.

Player performance, measured through individual point forecasts, was shown to be predictable with reasonable accuracy. The baseline XGBoost model, trained solely on performance and contextual indicators, captured a significant portion of scoring variance. When injury-related features were introduced, a modest but consistent improvement in predictive accuracy was observed. This suggests that short-term health risk influences not only availability but also individual performance – an insight with practical implications for load management, lineup planning, and sports betting models. Notably, clutch-related variables – such as late-game shot timing and zone-specific efficiency – emerged as key predictors in both the point prediction and game outcome models, reinforcing their central role in performance analysis and validating one of the central hypotheses of this study.

At the team level, outcome prediction models reached strong baseline performance even without incorporating upstream outputs. However, when point predictions and injury risks were aggregated and fed into the final classifier, performance improved further. Notably, SHAP analysis highlighted the central importance of predicted scoring differential and average injury risk in shaping team outcomes. These results validate

the pipeline's architecture: each component – inury risk, individual performance, and team dynamics – contributes uniquely to the final prediction, and their integration improves both accuracy and interpretability.

This work contributes to the literature in several ways. First, it demonstrates the feasibility and value of combining traditionally separate modeling tasks into a cohesive, modular architecture. Second, it emphasizes the relevance of explainable AI in sports analytics, showing that high-performing models can also yield actionable insights. Third, it reinforces findings from prior studies regarding the importance of workload, health, and spatial-temporal behaviors in determining basketball outcomes, while extending these insights through quantitative validation in a predictive context.

Several limitations must be acknowledged. The models were trained and evaluated on historical data, and their deployment in live, forward-looking scenarios would require careful adaptation. The use of publicly available injury records, while practical, may miss subtle or undisclosed health conditions that affect performance. Additionally, while SHAP values offer transparency, they do not capture all forms of interaction between features, nor do they guarantee causal interpretability.

In sum, this dissertation advances a structured and interpretable approach to NBA outcome forecasting, offering a proof of concept for how multi-level modeling and injury-informed analytics can enhance both the precision and relevance of basketball predictions. The pipeline's modularity, empirical grounding, and explanatory clarity position it as a meaningful step toward more holistic and intelligent sports analytics systems.

Bibliography

Bai, Y., & Yang, X. (2024).Prediction and treatment of joint injuries in basketball training based on improved regression algorithm from the perspective of sports biomechanics. *Molecular & Cellular Biomechanics*, *21*(3), 258.https://doi.org/10.62617/mcb258

Bird, S. P., Smith, A., & Johnson, R. (2021). Urgent wake up call for the NBA. Journal of Clinical Sleep Medicine, 17(10), 2000–2005. https://doi.org/10.5664/jcsm.8938

Bunker, R., & Thabtah, F.(2019).A machine learning framework for sport result prediction. *Applied Computing and Informatics*, *15*(1), 27–33. https://doi.org/10.1016/j.aci.2017.09.005

Byman, M. (n.d.). *Building a statistical learning model for evaluation of nba players using player tracking data*. https://www.ramapo.edu/dmc/wp-content/uploads/sites/361/2023/05/MSDS-Byman.pdf

Drew, M. K., & Finch, C. F. (2016). *The Relationship Between Training Load and Injury, Illness and Soreness: A Systematic and Literature Review* .https://pubmed.ncbi.nlm.nih.gov/26822969/

Genoud, M.(2024). *Statistics and machine learning in sports injury prevention* https://bodai.unibs.it/bdsports/wp-content/uploads/sites/2/2024/07/Bachelor-thesis-Melissa-GENOUD.pdf

Groll, A., Ley, C., Schauberger, G., & Van Eetvelde, H. (2019). A hybrid random forest to predict soccer matches in international tournaments. *International Journal of Forecasting*, *34*(1), 17–28. https://doi.org/10.1515/jqas-2018-0060

Horvat, T., Havaš, L., & Srpak, D. (2020). The impact of selecting a validation method in machine learning on predicting basketball game outcomes. *Symmetry*, *12*(3), 431.https://www.mdpi.com/2073-8994/12/3/431/pdf

Kambhamettu, A. R., & Shrivastava, A.(2024).Quantifying nba shot quality:A deep network approach. *Proceedings of the 7th ACM SIGAI Conference*. https://dl.acm.org/doi/pdf/10.1145/3689061.3689068

Kilcoyne, S. (2020). *The decline of the mid-range jump shot in basketball: A study of the impact of data analytics on shooting habits in the nba* https://digitalcommons.bryant.edu/cgi/viewcontent.cgi?article=1034&context=honors_mathematics

Lewis, M. (2004). *Moneyball: The art of winning an unfair game*. W.W. Norton & Company. https://onlinelibrary.wiley.com/doi/10.1002/mde.1220

Lim, B., Arık, S. Ö., Loeff, N., & Pfister, T.(2021).Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, *37*(3), 1745–1759. https://doi.org/10.1016/j.ijforecast.2021.03.012

Zhao, K., Du, C., & Tan, G. (2023). Enhancing basketball game outcome prediction through fused graph convolutional networks and random forest algorithm. *Entropy*, *25*(5), 765.https://doi.org/10.3390/e25050765

Lu, Y., Pareek, A., Lavoie-Gagne, O. Z., Forlenza, E. M., Patel, B. H., Reinholz, A. K., Forsythe, B., & Camp, C. L.(2022).Machine learning for predicting lower extremity muscle strain in national basketball association athletes. *Orthopaedic Journal of Sports Medicine*, *10*(7). https://doi.org/10.1177/23259671221111742

Lundberg, S. M., & Lee, S.-I.(2020).A unified approach to interpreting model predictions. *Nature Machine Intelligence*, 2, 56–67. https://dl.acm.org/doi/pdf/10.5555/3295222.3295230

Mehrasa, N., Zhong, Y., Tung, F., Bornn, L., & Mori, G.(2018). Learning Person Trajectory Representations for Team Activity Analysis. https://doi.org/10.48550/arXiv.1706.00893

Musat, C. L., Mereuta, C., Nechita, A., Tutunaru, D., Voipan, A. E., Voipan, D., Mereuta, E., Gurau, T. V., Gurău, G., & Nechita, L. C. (2024). Diagnostic applications of ai in sports: A comprehensive review of injury risk prediction methods. *Diagnostics*, *14*(22), 2516. https://doi.org/10.3390/diagnostics14222516

Ötting, M. (2021). Predicting play calls in the national football league using hidden markov models. *IMA Journal of Management Mathematics*, *32*(4), 535–545. https://doi.org/10.1093/imaman/dpab005

Sandri, M., & Zuccolotto, P. (2020). Markov switching modelling of shooting performance variability and teammate interactions in basketball. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *69*(5), 1337–1354. https://doi.org/10.1111/rssc.12442

Sarlis, V., Gerakas, D., & Tjortjis, C. (2024). A data science and sports analytics approach to decode clutch dynamics in the last minutes of nba games. *Machine Learning and Knowledge Extraction*, *6*(3), 102. https://www.mdpi.com/2504-4990/6/3/102

Steffen, P. (2022). *S*tatistical modeling of event probabilities subject on a sports bet: Theory and applications to soccer, tennis, and basketball https://theses.hal.science/tel-03891393/document

Annexs

Outlook

RE: NOVA IMS | Ethics Committee - NEED REVIEW

De Ethics Committee <ethicscommittee@novaims.unl.pt>
Data ter, 08/04/2025 23:52
Para Márcia Lourenço Baptista <m.baptista@novaims.unl.pt>; Joao Cristo <20230470@novaims.unl.pt>
Cc Ethics Committee <ethicscommittee@novaims.unl.pt>

Dear João Cristo,
Dear Professor Márcia Baptista,

Thank you for filling in the Research Ethics Checklist. After reviewing your request, you can proceed with the study as we do not foresee any major ethical concerns with the project.

Project No.: **DSCI2025-3-212833**

Project Title: **Forecasting NBA Championship Outcomes: The Role of Player Injury Records**

Principal Researcher: **João Cristo**

according to the regulations of the Ethics Committee of NOVA IMS and MagIC Research Center this project was considered to meet the requirements of the NOVA IMS Internal Review Board, being considered **APPROVED** on 08/04/2025.

It is the Principal Researcher's responsibility to ensure that all researchers and stakeholders associated with this project are aware of the conditions of approval and which documents have been approved.

The Principal Researcher is required to notify the Ethics Committee, via amendment or progress report, of

- Any significant change to the project and the reason for that change;
- Any unforeseen events or unexpected developments that merit notification;
- The inability of the Principal Researcher to continue in that role or any other change in research personnel involved in the project.

Lisbon, 08/04/2025
NOVA IMS Ethics Committee
ethicscommittee@novaims.unl.pt

Cristina Oliveira
Gestora executiva do centro de investigação MagIC | *Executive manager of the Information Management Research Center (MagIC)*
Find out more about our research at https://magic.novaims.unl.pt/en/
Team member of RM Roadmap - Co-creating the future of Research Management (https://rmroadmap.eu/)
https://orcid.org/0000-0002-0887-7961

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade NOVA de Lisboa